CHAPTER 11

Statistical Connectomics

Performing statistical inference on brain networks is an important step in many applications of **connectomics**. Inference might involve comparing groups of patients and healthy controls to identify which of the hundreds (or thousands) of connections comprising the **connectome** have been impacted by disease, or correlating measures of brain connectivity with some index of cognitive performance. Statistical connectomics is concerned with the methodologies used to perform this kind of inference.

In this chapter, we examine two aspects of connectomics that are important to the statistical analysis of populations of brain networks. The first is thresholding. Brain connectivity data are noisy, and it is sometimes desirable to apply a threshold to the **adjacency matrix** to distinguish true connections from spurious ones. As we will see, there are many methods for applying such thresholds and the specific technique chosen can have a major impact on the conclusions drawn from any statistical analysis that follows.

The second part of this chapter covers statistical inference on brain networks. The statistical testing of connectomic data falls within three broad categories: global or omnibus testing, mass univariate testing, and multivariate approaches. In global testing, we investigate between-group differences or associations using one or more of the global topological measures considered in other chapters. The global approach is simple but lacks specificity: global measures often provide useful summaries of network properties, but they cannot tell us whether the effects are indeed distributed throughout the brain, or confined to a specific subset of **nodes** and/or **edges**.

Mass univariate or *connectome-wide* hypothesis testing involves testing the same hypothesis over many different elements of a brain network. Typically, this is done either at the level of each region, in which case we test a hypothesis about some node-level property of interest; or at the level of each edge in the connectivity **matrix**, in which case we seek to identify connection-specific effects. Unlike global testing, these analyses allow for the localization of effects to specific brain regions (nodes) or neural connections (edges). Mass univariate testing across a family of nodes or edges naturally gives rise to a **multiple comparisons problem**, which we will see is particularly sizeable, even for

Fundamentals of Brain Network Analysis. http://dx.doi.org/10.1016/B978-0-12-407908-3.00011-X

coarse mappings of a connectome. The problem of multiple comparisons imposes a need for stricter significance thresholds when a family of independent hypotheses is tested simultaneously. We will describe various methods that can be used to correct for multiple comparisons when performing mass univariate testing with connectomic data. We will also see that multivariate methods such as machine learning have begun to gain traction in the field. These approaches seek to recognize patterns and high-level abstractions across multiple connections and nodes that discriminate between groups more accurately than any one measure alone.

How can statistical inference be performed on the topological measures considered in this book? This chapter seeks to address this question, focusing on global, connectome-wide, and multivariate analyses. The methods discussed in this chapter have emerged from the human connectomics literature, where access to large samples of data is readily available. Indeed, a primary goal of the large-scale Human Connectome Project is to understand how individual differences in brain connectivity relate to behavior (Van Essen et al., 2013). For this reason, the techniques presented in this chapter are largely focused on the human literature, although they are equally applicable to data acquired in other species.

11.1 MATRIX THRESHOLDING

As we saw in Chapter 3, it is common practice, particularly in the human connectomics literature, to apply a threshold to a connectivity matrix in order to remove noisy or spurious **links** and to emphasize key topological properties of the network. This step can complicate statistical analysis. For example, if we have two populations of networks—one from a patient group and one from a control group—which single threshold should we choose to make a fair comparison between them? This question is critical because many graph theoretical measures are dependent on the number of edges in the graph. We must therefore choose our threshold with caution. This section considers different approaches to threshold brain networks.

11.1.1 Global Thresholding

The simplest thresholding strategy is to apply a single, *global* threshold to all elements of the connectivity matrix. Elements below the threshold are set to zero. Surviving elements can either be set to one, resulting in a binary adjacency matrix, or they can retain their original connectivity weights if the goal is to then undertake an analysis of weighted networks. This threshold can be determined using one of two methods: weight-based or density-based thresholding.

Weight-based thresholding involves choosing a threshold value, τ, based on the weights in the connectivity matrix. We could choose a value for τ arbitrarily. For example, we could set $\tau = 0.25$ in a correlation-based network, meaning that any pair-wise correlations below this value will be set to zero. Alternatively, τ could be determined using statistical criteria, such as one based on statistical significance. For example, we could retain only correlation values within individual networks that satisfy an α-significance level.

One consequence of these weight-based approaches is that the **connection density** of each network may vary from network to network after the threshold has been applied. To illustrate this point, Figure 11.1a depicts **functional connectivity** data in a single patient with schizophrenia and a healthy control. In this case, the patient has, on average, lower connectivity compared to the control (Figure 11.1b). Assume that we now apply the same correlation threshold of $\tau = 0.20$ to both networks. The resulting adjacency matrix of the patient has a connection density of 53%, whereas the control matrix has a density of 75% (Figure 11.1c and d). This discrepancy arises because the higher average connectivity of the control network means that there are more connections satisfying the condition $w_{ij} > \tau$. Since most graph theoretic measures are sensitive to variations in the number of edges in a graph, any differences in density must be considered when comparing two or more networks.

Density-based thresholding explicitly addresses this problem. With this method, τ is allowed to vary from person to person to achieve a desired, fixed connection density κ. Consider our patient and control in Figure 11.1. Say that we threshold both networks to achieve a density of $\kappa = 0.20$ (i.e., 20% density). We need to set $\tau \approx 0.30$ for the patient and $\tau \approx 0.41$ for the control (Figure 11.1c and e). The value of τ for the control network is higher because the connectivity weights are, on average, higher in this network. Conversely, we use a lower value of τ when thresholding the patient network in order to achieve the same connection density as the control network. As a result, more low-value correlations, which may represent spurious connections, are included in the patient's thresholded network. One consequence of retaining a higher proportion of potentially spurious connectivity estimates is that the network **topology** will appear more random. Several brain disorders, including schizophrenia, have been shown to display a more random topology compared to controls, in the presence of mean functional connectivity reductions (Rubinov et al., 2009; Lynall et al., 2010; Stam, 2014). In such cases, it can be difficult to disentangle variations in connectivity weight from differences in network topology. The increased randomness in network topology seen in schizophrenia may therefore be a consequence of the density-based thresholding procedure itself; namely, the need to retain more low-weight edges in patient networks in order to achieve a particular connection density. These low-weight edges will

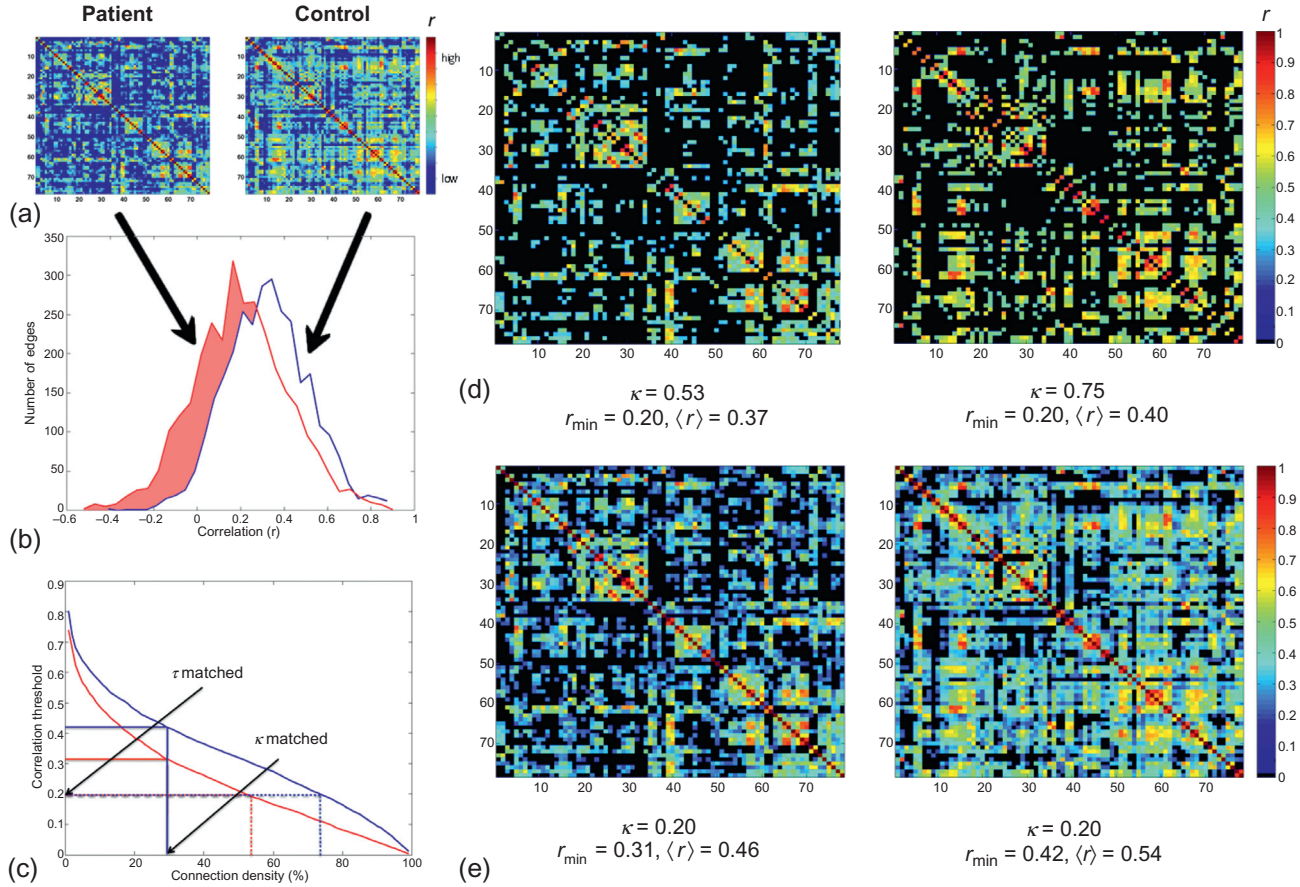

FIGURE 11.1 **Density-based and weight-based matrix thresholding.** Shown here is an example of the relative strengths and weakness of these two thresholding approaches. **(a)** Two individual functional connectivity matrices for 78-region networks estimated using **functional MRI**. One is from a patient with schizophrenia and the other is from a healthy control. The data are from the study reported by Fornito et al. (2011a). **(b)** The distribution of connectivity weights in each network. On average, connectivity is lower in the patient. **(c)** The relationship between τ and κ in these two networks. This plot illustrates how matching for τ results in networks with different connection densities while matching for κ requires us to apply a different weight threshold, τ, to both networks. **(d)** The thresholded adjacency matrices obtained for the patient and control after applying the same τ threshold. The minimum correlation value in the matrix, r_{min}, is the same, and the average correlation value, $\langle r \rangle$, across the two networks is comparable, but the connection density, κ, is different. **(e)** The thresholded adjacency matrices after applying the same κ threshold. The connection densities are the same, but the minimum and average correlations are different. *Figure adapted from Fornito et al. (2013) with permission.*

most likely be positioned randomly in the network. This problem affects both weighted and binary network analyses.

A possible alternative is to retain connections if they are statistically robust across the entire sample. For example, in a functional connectivity network, a one-sample *t*-test can be used to determine which correlations have a group mean that is significantly different from zero. This approach thus allows us to retain only those connections that are consistently identified across the sample, at some level of statistical confidence. In **diffusion MRI**, group-level thresholds have been shown to minimize the inclusion of false positive and false negative edges in the connectivity matrix (de Reus and van den Heuvel, 2013a). However, they also reduce inter-individual variability in network architecture, since the binary structure of the adjacency matrix is reduced to the connections identified with greatest consistency across the sample. As a result, any variations in network organization will be largely determined by differences in connectivity weight. Another limitation of group thresholds is that any reconstruction bias that consistently generates a false positive connection across the sample will also be declared significant. In other words, a one-sample *t*-test cannot suppress false positives that are consistent across a population.

The thresholding of connectivity matrices with signed edge weights (e.g., a correlation-based network) poses a special problem. In such cases, we must make a decision on how to treat negative weights. In general, a positive edge weight implies some degree of functional cooperation or integration between network nodes; a negative edge weight implies segregation or antagonism. If we wish to capture this qualitative distinction, we should threshold based on the absolute value of the correlation and then consider the polarity of the weights in subsequent analyses. This will ensure that both positive and negative edges are retained in the thresholded matrix. On the other hand, if we are only interested in positive connections, we can threshold the signed connectivity matrix (e.g., if our threshold is set to $\tau = 0.25$, any negative values will be set to zero).

The comparative strengths and limitations of weight-based and density-based thresholding highlight a conundrum. On the one hand, we know that topological properties vary as a function of the number of edges in a network, suggesting that we should use density-based thresholding. On the other hand, individual or group variations in connection density may themselves convey meaningful information, suggesting that any effect of these variations on network topology is real and should not be removed by density thresholding. This argument favors weight-based thresholding, but we know that comparing networks with different numbers of edges is problematic. The choice between the two methods depends on whether or not a difference in connection density is viewed as a confound. In some cases, it may be useful to understand how results vary across different thresholds selected using both approaches. Detailed

discussions of these issues are provided by Ginestet et al. (2011), van Wijk et al. (2010), and Simpson et al. (2013).

11.1.2 Network Fragmentation

One limitation of using a global threshold is that it can cause networks to fragment at sufficiently stringent thresholds. Recall from Chapter 6 that a network is **node-connected** when all of its nodes can be linked to each other by a **path** of edges, thus forming a single **connected component**. **Fragmentation** occurs when the network splits into disconnected components.

In the absence of significant pathology, we expect nervous systems to be node-connected networks (Chapter 6). Thresholding can violate this basic property. As we increase our threshold and use more stringent criteria for retaining links, the connectivity matrix will become sparser and begin to fragment. This fragmentation will affect the measures that we compute on the networks. For example, traditional measures of **path length** cannot be computed between disconnected components (Chapter 7).

One solution to this problem is to constrain the range of τ and/or κ values, such that only node-connected networks are analysed. With this approach, we do not examine any thresholds that result in fragmentation of the network. However, in noisy datasets it is sometimes difficult to ensure connectedness for all individual networks. One potential solution is to first compute the **minimum spanning tree** (MST) of the network (Box 6.1), and then add connections in decreasing order of weight until the desired connection density is achieved. This method will ensure that a **path** can be found between all pairs of nodes and that the thresholded network will be node-connected.

11.1.3 Local Thresholding

Global thresholding is a simple and intuitive approach, but it may miss important structure in the data. For example, networks with fat-tailed weight distributions such as the brain (Chapter 4) may possess interesting features that span multiple scales. These features will be overlooked when a global threshold is applied, since connections to nodes with low average connectivity will be penalized (Serrano et al., 2009). Local thresholding methods aim to address this problem by computing thresholds determined locally at the level of each node, rather than at the level of the entire network.

One such approach is the so-called *disparity filter* (Serrano et al. 2009). The disparity filter operates on each node independently. For a given node, we first normalize the weights of each of its edges by the **strength** of that node (Chapter 4). In particular, we divide the weight of an edge by the strength (i.e., the sum of all edge weights) of the node under consideration to yield

a node-wise normalized connectivity weight. The edges that account for a significant proportion of a given node's total connectivity with the rest of the network (i.e., its strength) are then identified with respect to a **null model** defining chance-expected fluctuations in regional weights. The null model for the disparity filter assumes that the node-wise normalized weights of a given node with binary **degree**, k, arise from random assignment from a uniform distribution. Since the node-wise normalized weights span the range [0, 1], the null model is given by distributing $k-1$ points with uniform probability in the unit interval, so that it is divided into k subintervals. The magnitude of these uniform subintervals is then compared to the node-wise normalized weights. Observed weights exceeding the chance-expected interval magnitudes for a given α-significance are declared statistically significant. Since weights are normalized on a node-wise basis, the new normalized edge values may be asymmetric, even in undirected networks. That is, the precise value of any particular normalized edge weight may vary depending on whether the strength of node i or j is used in the normalization. In such cases, the edge is retained if it meets the significance criterion for at least one of the two nodes to which it is attached.

An important property of the disparity filter is that it will only declare edges significant when there is considerable heterogeneity in the edge weights attached to a given node. To illustrate this point, take the following example. Consider a node in a correlation-based network with three edges, each of weight 0.70. The corresponding normalized weight of each edge will be $0.70/(3 \times 0.70) = 0.33$. If we then place $k-1=2$ points randomly on the unit interval, on average, the distance between consecutive points will be 0.33 which is identical to our observed normalized edge weights. In this case, none of the edges will be declared significant even though their weight is relatively high (on the scale of Pearson correlation). This example illustrates why the method is called a disparity filter—it is a method that is predominantly sensitive to large variations in edge weights at the local level of individual nodes. Nonparametric variants of this method have been proposed (Foti et al., 2011).

The disparity filter does not enforce a specific connection density on the data. Instead, the density is determined only by the proportion of edges that meet a given statistical significance. An alternative local thresholding method was proposed by Alexander-Bloch et al. (2010). Starting with the MST, they added edges incrementally based on the *k-nearest neighbor graph* (k-NNG). The k-NNG comprises edges linking each node to its k nearest **neighbors**—the k nodes most strongly connected to the index node. With this method, k can be gradually increased until a desired connection density is achieved (Figure 11.2a). However, the method cannot guarantee a precise match to a given target density, since the number of edges added with a single increment of k may be larger than one. An analysis of resting-state functional MRI connectivity networks in children with childhood-onset schizophrenia and healthy controls

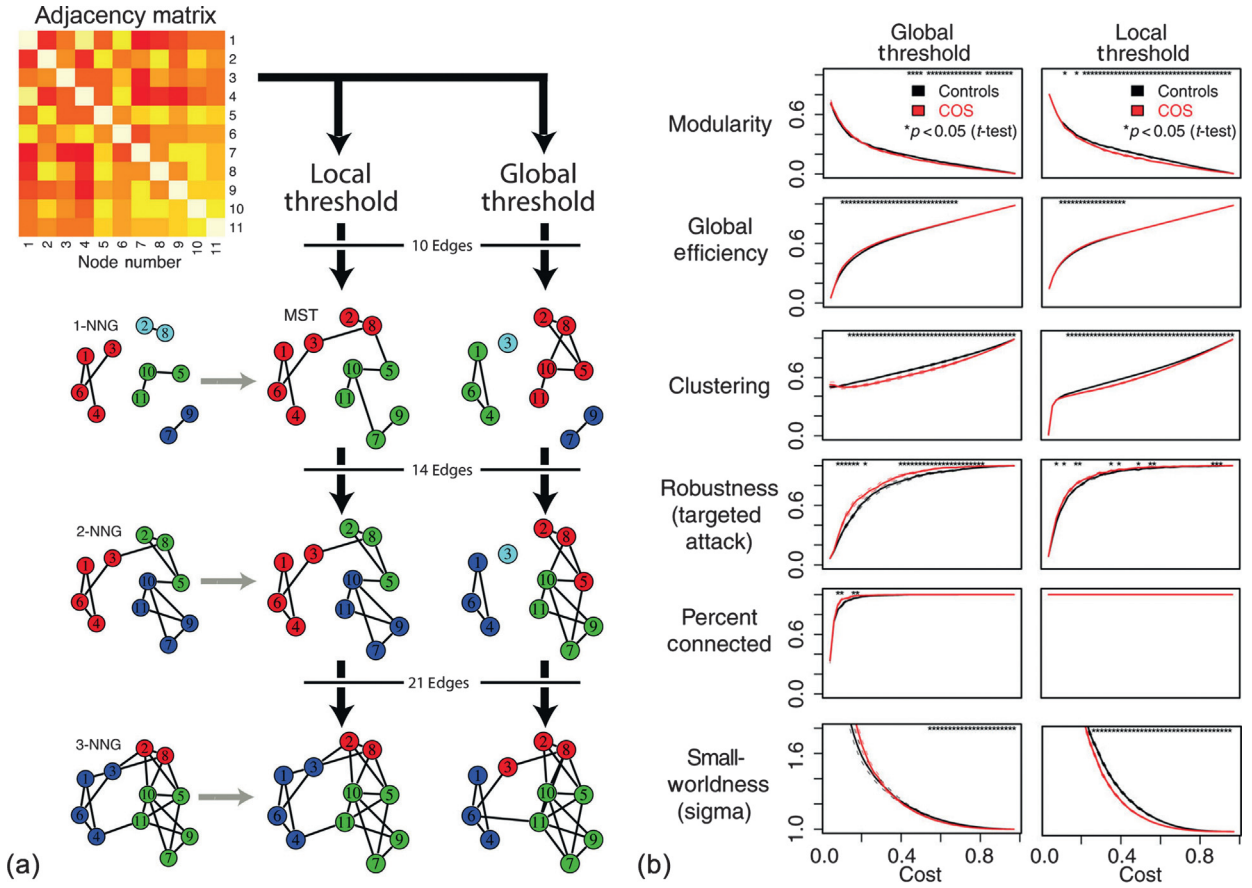

FIGURE 11.2 Comparison of global and local matrix thresholding. **(a)** An illustration of the difference between a local thresholding approach using *k*-nearest neighbor (*k*-NNG) graphs and global thresholding. The local threshold (left) first finds the MST of the graph, of which the 1-NNG is a subset. The parameter *k* is then increased to allow more edges into the network. For each *k*, we find the *k*-NNG graph, which includes each node's *k* most highly weighted connections. The global threshold (right) simply retains the most highly weighted edges in the graph at each threshold. **(b)** Different results can be obtained in comparisons of brain networks when these two thresholding methods are used. Shown here are differences in the topological properties of functional connectivity networks measured with functional MRI in people with childhood-onset schizophrenia (red) and healthy controls (black). The data are thresholded across a range of connection densities (**cost**). As shown by the asterisks, which represent group differences significant at $p < .05$, uncorrected, the two thresholding methods vary in their sensitivity to group differences depending on the specific topological metric and connection density considered. *Images reproduced from Alexander-Bloch et al. (2010) with permission.*

found that the results of this method sometimes diverged with those obtained using a global thresholding approach, depending on the specific metric and threshold used (Figure 11.2b). The advantage of the *k*-NNG method is that any group differences in network measures cannot be attributed to network fragmentation, which is a problem when applying a global threshold. The disadvantage of this method is that the local thresholding process can, in and of itself, introduce nontrivial topological structure. For example, there will never be a node of degree less than *k* if the edges of the *k*-NNG are added to the thresholded network.

11.1.4 Normalization with Reference Networks

Global and local thresholding methods share the common goal of distinguishing real from spurious connections. These methods alter the network's connection density, or the range of connectivity weights in the adjacency matrix. We have already seen that most measures computed in a thresholded matrix depend on these basic properties. It is therefore important to interpret measures computed in thresholded networks with respect to the connection density that has been imposed by the thresholding process.

One potential approach to deal with the dependence of network measures on connection density is to normalize the measure with respect to an appropriate benchmark network, such as a **lattice** or randomized network matched to the empirical network for size, connection density, and **degree distribution** (see Chapter 10). These benchmarks set a baseline expectation of how a given network property, *M*, varies as a function of connection density in the absence of structured topology. However, detailed analysis of this method by van Wijk et al. (2010) has demonstrated that normalization with respect to either lattice-like or randomized networks does not completely remove the threshold dependence of commonly used topological measures such as the **characteristic path length** and **clustering coefficient**. Moreover, they found that the threshold-dependence varied according to the specific measure and benchmark network that was used.

One method that van Wijk et al. (2010) endorsed as a reasonable solution to the problem of threshold dependence was to normalize the observed measures relative to the most extreme values that these measures can assume, given the connection density. For **small-world** networks such as the brain, suitable bounds for this range are provided by the values obtained in matched lattice and randomized graphs. We can thus formally define a normalized measure for any network property as,

$$M' = \frac{M - M_{\text{rand}}}{M_{\text{lattice}} - M_{\text{rand}}}, \tag{11.1}$$

where the subscripts 'rand' and 'lattice' refer to the values obtained in the appropriately matched random and lattice networks, respectively (Chapter 10). More

sophisticated methods for comparing networks with different connection densities, such as **exponential random graph models** (Box 10.2), hold promise but have not been extensively applied to the analysis of brain networks.

11.1.5 Integration and Area Under the Curve

It should now be clear that both local and global thresholding methods mandate selection of a particular threshold. The choice of threshold is rather arbitrary. Should we threshold to achieve a connection density of 5% or 10%, or should we only consider edges in a correlation-based network with a weight exceeding 0.5? To address the arbitrariness in this choice, we can repeat the desired statistical analysis across a range of thresholds, and examine the consistency of the findings. However, this method introduces a multiple comparisons problem, since we are computing a test statistic at many different thresholds. Generic methods for dealing with multiple comparisons are discussed in Section 11.2.

To sidestep the multiple comparisons problem, we can integrate a given network measure across the full range of thresholds analyzed. Analysis is then performed on the integrated measure, rather than on each of the many threshold-specific measures. The network measure of interest is integrated over a range of thresholds to yield the area under the curve (AUC), and then statistical testing is performed on the AUC (Bassett et al., 2006, 2012). The lower bound of this threshold range, κ_-, might be the point at which the network becomes node-connected, and the upper range, κ_+, might be the point at which the network is just short of being fully connected, or fails to show interesting topological structure, such as small-worldness (Lynall et al., 2010). With this approach, we start at κ_- and add edges, one at a time in decreasing order of weight (or some other criterion of priority) and compute the value of our topological metric, M_κ, at each connection density until we reach κ_+. This results in a curve that defines how M varies as a function of κ. We can then integrate these values of M_κ across all levels of κ to compute the area under this curve, and use this value in subsequent analyses. For example, we can imagine choosing a specific range of thresholds for the curves shown in Figure 11.2b, computing the area under those curves, and comparing the resulting values obtained for patients and controls.

This approach is similar in principle to the method for computing cost-integrated metrics favored by Ginestet et al. (2011). Specifically, these authors present theoretical arguments supporting inference on threshold-integrated metrics, showing that such measures are invariant to any monotonic remapping of the connectivity matrix. An example of a monotonic remapping is where the elements in one connectivity matrix are proportional to the corresponding elements in another connectivity matrix. Ginestet and colleagues prove that any threshold-integrated measure will be equal for these two matrices, which follows from the fact that a monotonic remapping does not affect the relative

ranking of the elements of the adjacency matrix. Importantly, this approach results in a single value summarizing the behavior of any topological measure across a broad range of thresholds, thus reducing the multiple comparisons problem and simplifying analysis. The lower and upper bounds of the threshold range, κ_- and κ_+, should be chosen carefully: analyzing a wide range of thresholds will generally obscure effects that are present within a specific regime, whereas choosing a range that is too narrow may not adequately sample the data or miss effects that are apparent at other densities.

Drakesmith et al. (2015) have extended the AUC approach with the inclusion of cluster-based **permutation testing**, where a cluster in this context is defined as a contiguous range of thresholds at which the size of an effect exceeds a critical threshold. In an approach called multi-threshold permutation correction (MTPC), they perform a statistical test at each threshold independently. A null distribution is then generated for the test statistic computed for each threshold using permutation testing (Box 11.1). Drakesmith and colleagues argue that a genuine effect should exceed the null distribution over a contiguous range of thresholds. They therefore identify clusters along the threshold axis at which the observed test statistics exceed a critical test statistic cut-off value. The size of each cluster is then measured by computing the AUC over the *circumscribed* range of thresholds defining the cluster. Note that this is in contrast to the conventional approach of integrating the network measure of interest over the *entire* range to yield the AUC (Bassett et al., 2006). The process of identifying clusters is repeated after permuting the data to build a null distribution for the size of the largest cluster under the null hypothesis (see Box 11.1). This null distribution can then be used to assess the significance of the clusters measured in the observed data. Figure 11.3 presents results obtained with the MTPC method. The approach provides greater sensitivity to detect between-group differences in some network measures when compared to the conventional AUC approach, although the improvement is marginal for others. As with the AUC approach, the MTPC may also better control for the presence of false positive connections arising from an imperfect connectome reconstruction process. Considering a range of thresholds ensures inference is not dependent on a particular threshold that results in many false positive connections.

11.1.6 Multiresolution Thresholding

The thresholding approaches that we have considered in this section typically apply a hard threshold, such that all matrix elements with $w_{ij} \leq \tau$ are set to zero and all elements $w_{ij} > \tau$ are retained for further analysis. Since the choice of τ is arbitrary, this approach may create an artificial distinction between matrix elements that are useful and not useful. To address this problem, Lohse et al. (2014) investigated three alternative methods that offer insights into the

BOX 11.1 PERMUTATION TESTS

Permutation tests, also known as *randomization tests* and *exact tests*, are a versatile class of statistical significance tests in which the null distribution of a test statistic is estimated empirically by repeatedly permuting data points between groups, with each permutation providing a new sample from the null distribution. Permutation tests were first described in the 1930s by one of the founders of statistical science, Sir Ronald Fisher, and the Australian mathematician Edwin Pitman (Fisher, 1935; Pitman, 1937). The ubiquity of inexpensive and fast computers in the last few decades has made the application of permutation tests a natural choice for a wide range of problems in statistical inference where the true distribution of a test statistic is unknown.

Permutation tests can be most easily understood in terms of a t-test. Generalization to other test statistics is usually straightforward, but not always. Suppose we measure the clustering coefficient (Chapter 8), for example, in two groups comprising n_1 and n_2 individuals, and we seek to test whether the sample means are significantly different, given an α-significance level. This can be achieved with a two-sample t-test, in which case we assume the t-statistic can be parameterized under the null hypothesis by a theoretical distribution known as the Student's t-distribution. However, the theoretical null distribution is sometimes unknown and might provide a poor fit for small sample sizes even if it is known.

A permutation test uses the observed dataset itself to estimate the null distribution empirically. In particular, we pool together the clustering coefficients from both groups and then compute the t-statistic for every possible way of dividing these pooled values into two groups of sizes n_1 and n_2. There are $(n_1 + n_2)!/(n_1!n_2!)$ such permutations. For example, if x_1, x_2, x_3 denote the network clustering values in one group, and y_1, y_2, y_3 denote the clustering values in the other group, then one of the 20 possible permutations is x_1, y_2, x_3 for one group and y_1, x_2, y_3 for the other. We compute a t-statistic comparing these permuted groups and then repeat the process again many times. The permuted t-statistics that we obtain from this procedure define an exact null distribution under some basic assumptions. A one-sided p-value is then given by the proportion of permuted t-statistics that are greater than or equal to the actual t-statistic. A two-sided p-value is given by the proportion of permuted t-statistics with absolute value greater than or equal to the absolute value of the actual t-statistic. When n_1 and/or n_2 are large, enumerating all possible permutations is intractable, and thus we can randomly sample a sufficient subset of permutations, usually between 5000 and 10,000, which is known as a Monte Carlo permutation test. Although we use parametric tests to quantify the difference in sample means, the p-values we compute with permutation tests are nonparametric. We can in principle use any measure to quantify the difference in sample means (e.g., percent difference), although there are good reasons to use a pivotal (normalized) quantity such as the t-statistic.

Permutation tests assume a technical condition known as **exchangeability**, which requires that the true null distribution is insensitive to permutations of the data, or equivalently, the data labels. For example, exchangeability is not satisfied in situations where the sizes and variances of the two groups are heterogeneous. Permutation tests have also been criticized for being computationally intensive, although nowadays this criticism is largely unfounded.

Permutation tests are a good way to account for dependencies between tests when performing mass univariate testing. Dependencies may be of a spatial or topological origin, or due to the transitive nature of the correlation coefficient (e.g., in functional connectivity networks). Under some basic assumptions, the (joint) null distribution estimated with a permutation test preserves any dependencies that are otherwise ignored by parametric approaches. To preserve dependencies, it is important to ensure that permutation is performed on all tests as a whole, rather than independently on each test.

A deeper treatment of permutation tests is provided by Good (1994) and Peasarin (2002). The work of Nichols and colleagues demonstrating the utility of permutation tests in traditional neuroimaging applications is also an invaluable resource (Nichols and Holmes, 2001; Nichols and Hayasaka, 2003).

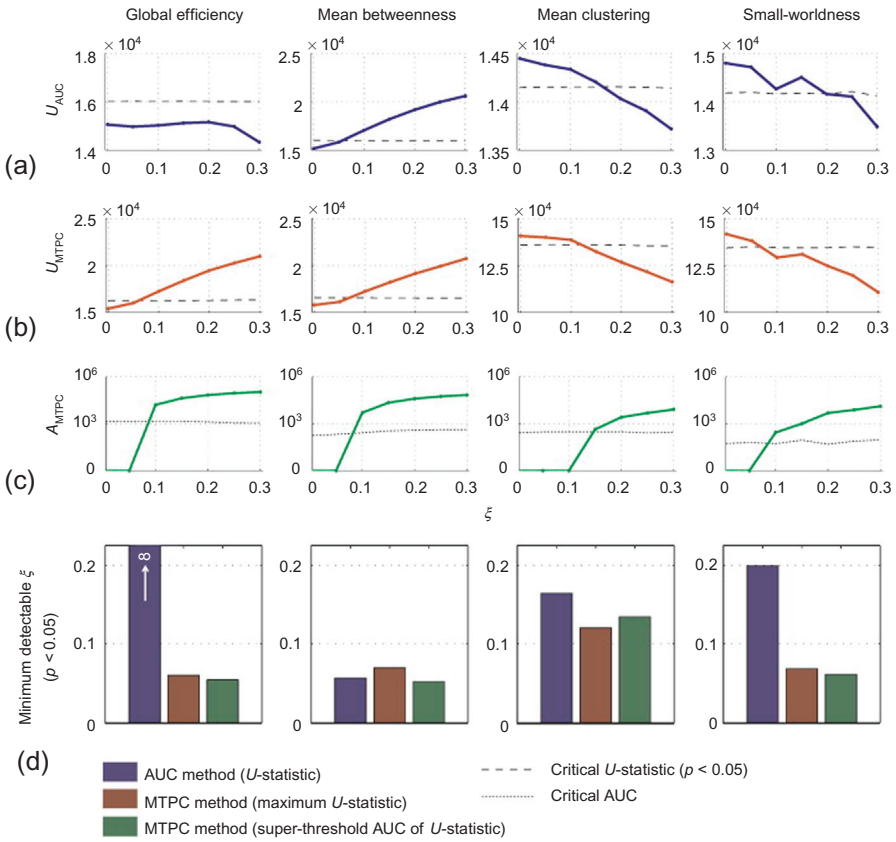

FIGURE 11.3 Comparison of multi-threshold permutation correction (MTPC) and the conventional area under curve (AUC) approach in detecting simulated between-group differences in four network measures. A difference between two populations of **structural connectivity** networks measured with diffusion MRI was simulated by removing interhemispheric streamlines from one of the populations. The proportion of streamlines removed was sampled from a half-normal distribution with standard deviation, ξ, and mean of zero. Increasing the standard deviation increased the magnitude of the between-group difference. The presence of a between-group difference in four network measures was then tested using the conventional AUC method **(a)**, MTPC based on the maximum test statistic observed across all thresholds considered **(b)**, and cluster-based MTPC **(c)**. The Mann-Whitney *U*-test was used in all cases. The vertical axis in panels **(a–c)** corresponds to the size of the test statistic. The minimum proportion of streamlines that needed to be removed for a significant difference to be declared (i.e., the smallest effect size required to reach significance) was then determined for the three methods **(d)**. The blue bar denotes AUC, brown denotes MTPC based on the maximum test statistic and blue is cluster-based MTPC. The MTPC methods offered improved sensitivity compared to the AUC method in most cases, although the improvement was marginal for the clustering coefficient and **betweenness**. *Images reproduced from Drakesmith et al. (2015) with permission.*

multiresolution structure of a brain network. One method involves first normalizing the weights of the adjacency matrix to the range [0, 1], and then raising the elements of the matrix to an arbitrary power r, such that $w_{ij} \rightarrow w_{ij}^r$. As r gets lower, the distribution of weights becomes more homogeneous until all elements have $w_{ij} = 1$ when $r = 0$. As $r \rightarrow \infty$, the distribution becomes more heterogeneous and the relative influence of strong connections is amplified. This method is suited to the analysis of weighted networks and is not a thresholding method, per se.

A second method investigated by Lohse et al. (2014) depends on the use of a multiresolution **module** decomposition of the network (see Chapter 9). With this approach, subsets of nodes within modules defined at particular scales of topological resolution are identified and their properties are investigated separately at each resolution. A third approach, termed *windowed thresholding*, involves independently assessing network properties within discrete windows of edge weights. A window is defined to span a limited range of edge weights. Edges with a weight that is spanned by the window are retained, while edges with a weight falling outside the window are set to zero. The length of the window dictates the number of edges that are retained and the position of the window determines the average weight of the retained edges. Examining network properties in different windows thus offers insight into the topological organization of connections within certain weight ranges. For example, we may wish to examine how the strongest connections are organized in the network, independently of the effects of weaker links. By construction, this method precludes an analysis of interactions between strong and weak links. Applying this method to functional connectivity networks measured with functional MRI in patients with schizophrenia and healthy controls, Lohse and colleagues only found between-group differences in windows with very low or high average percentile weight, but not for a broad range of intermediate threshold windows. This result suggests that topological disturbances in this group primarily affect the organization of very strong and very weak functional links.

11.2 STATISTICAL INFERENCE ON BRAIN NETWORKS

Connectivity weights and topological properties of brain networks can vary between different populations. Identifying differences in brain networks associated with brain disease, aging, and cognitive function (to name just a few) is an important application that requires statistical inference. In this section, we consider some of the main approaches to the statistical analysis of network measures.

11.2.1 Global Testing

Global, or omnibus testing, is the simplest kind of statistical inference that can be performed on brain networks. Global testing involves inference on one or more of the global topological measures considered in the earlier chapters, such as the average clustering coefficient, global **efficiency**, or Humphries' index of small-worldness (Humphries and Gurney, 2008). The exact network measure is not of concern to us here, since the basic approach is general. The goal is to determine whether one of these network parameters varies as a function of clinical diagnosis, age, cognitive ability, or some other characteristic of interest. Standard statistical tests can be employed to test for such associations, bearing in mind the usual caveats such as data normality, homogeneity of variance, and so on. Since it is often difficult to know the population distribution of most network measures a priori, the use of nonparametric statistics is recommended, as it offers a means for performing statistical analysis with a relaxed set of assumptions (Box 11.1).

In general, global testing is simple and offers insight into global network properties, but lacks specificity and may lack power when the effect of interest is limited to only a few connections or nodes. For example, if the clustering coefficient shows a difference at a single node in a network of hundreds of nodes, an omnibus test performed on the average clustering coefficient is unlikely to reveal a significant effect.

11.2.2 Mass Univariate Testing

Mass univariate testing is performed to localize effects to specific nodes or edges. It thus provides superior localizing power compared to global testing. Such an analysis allows us to move beyond global summary metrics to pinpoint which specific regions or connections might be driving the global alteration.

What exactly do we mean by mass univariate hypothesis testing? Mass univariate testing refers to the fact that a particular statistical test, such as a *t*-test, is performed independently across a large number of nodes or connections within a brain network. The *family* in the context of mass univariate testing defines the set of nodes or connections that we examine for an effect, which is often all of them. However, to reduce the number of multiple comparisons, it may be advantageous to constrain the family to a subset of connections or nodes based on prior beliefs.

When the family comprises nodes, we can perform inference on just about any node-specific measure, such as the node degree, local efficiency, or clustering coefficient. When the family comprises connections, we most often test for variations in connectivity strength (i.e., edge weights), or connection-specific topological properties, such as the edge betweenness centrality (Newman and Girvan, 2004). The particular measure is not of great concern to us here. The family might contain a mixture of nodes and connections, but this is rare.

Mass univariate testing yields a test statistic and corresponding p-value for each element in the family. Ginestet and Simmons (2011) refer to the family of test statistics as a statistical parametric network. The size of the family can range from hundreds to thousands, or even millions, depending on the size of the network. We denote the number of hypotheses in the family with J. We cannot simply reject the null hypothesis for all p-values that are less than a nominal significance of $\alpha = 0.05$, for example, since doing so can result in αJ false rejections, or *false positives*. This is an unacceptably large number of false positives. Usually, we must use a more stringent, or *corrected*, α-significance to ensure that the probability of one or more false positives across the entire family, known as the **familywise error rate** (FWER), is less than a prescribed level. This is known as the **multiple comparisons problem**.

There is a variety of old and trusted procedures for dealing with multiple comparisons. Bonferroni correction and control of the **false discovery rate** (FDR; Benjamini and Hochberg, 1995) are the most well-known. With the advance of statistical connectomics, more powerful correction procedures have been developed to exploit the spatial and topological properties of brain networks. These approaches can often provide greater statistical power than generic multiple comparison procedures such as the FDR. In other words, they provide us with greater power to detect true effects, while ensuring that the FWER is controlled at a prescribed level.

Most connectome-specific multiple comparison procedures have been designed for the case when the family comprises connections, since this is the most challenging case in terms of the scale of the multiple comparisons problem $\left(J \sim N^2\right)$. When the family comprises nodes, the multiple comparisons problem is substantially smaller $(J = N)$, and thus controlling the FDR may provide sufficient statistical power and is sometimes used in practice. The final goal of these methods is to allow the construction of a map depicting nodes and/or edges that show a particular effect of interest at a given level of statistical confidence. Interpretation of these maps can sometimes be complex, and is best done within an appropriate theoretical framework (Box 11.2).

11.2.3 Strong Control of Familywise Errors

The simplest and most conservative approach to the multiple comparisons problem is control of the FWER. This can be achieved with the classic Bonferonni procedure. To ensure a FWER no greater than a prescribed α, we reject the null hypothesis for all p-values satisfying,

$$p_j \leq \frac{\alpha}{J}, \quad j = 1, \ldots, J. \tag{11.2}$$

BOX 11.2 INTERPRETING CLINICAL DIFFERENCES IN BRAIN NETWORKS

Statistical connectomics is often used to characterize clinical differences between a group of patients with a given brain disorder and a healthy comparison group. For this purpose, the statistical methods discussed in this chapter offer a powerful framework for mapping abnormalities across the entire connectome. Once these maps are in hand, the major challenge is to interpret the findings in a way that clarifies disease mechanisms. Identifying a set of connections that is either increased or decreased in a given patient group may be useful, but the map will be most informative when it tells us something about the underlying disease process.

Interpretation of clinical differences in brain networks can be complicated by the complex pathological changes that often arise in brain disorders. For example, compared to healthy controls, patients with schizophrenia can show reduced structural connectivity, coupled with both increased and decreased functional connectivity (Skudlarski et al., 2010). Similarly, patients with multiple sclerosis, a disease that causes an inflammation and deterioration of structural connectivity, also show a complex pattern of increased and decreased functional connectivity (Hawellek et al., 2011). The simulation of large-scale brain dynamics with neural mass models (Deco et al., 2008) interconnected based on empirically derived structural connectivity networks suggests that distributed increases and decreases of functional connectivity can arise after the deletion of specific nodes (and their incident edges) of the structural network (Alstott et al., 2009; see also Figure 6.5). In each of these cases, a *reduction* in the integrity of the underlying structural network is associated with a complex combination of increases and decreases of functional connectivity.

We can reconcile these apparently contradictory findings by categorizing possible neural responses to insult into one of two broad classes: maladaptive and adaptive (Fornito et al., 2015). Figure 11.2.1 presents a schematic of how key examples of each class are expressed at the level of brain networks, and how they can be linked to underlying pathophysiological processes.

Maladaptive responses generally compound the effects of an initial insult. Three major examples are **diaschisis**, transneuronal degeneration, and dedifferentiation. Diaschisis was first described by von Monakow (1969) as the interruption in function of remote and otherwise intact areas that are connected to a lesioned site (Figure 11.2.1a). A classic example is crossed cerebellar diaschisis, where a unilateral lesion to the forebrain can result in a functional depression of the contralateral cerebellum (Gold and Lauritzen, 2002). These distal effects may arise from reduced afferent input to the intact area, or an impaired capacity for interregional synchronization following the lesion.

Transneuronal degeneration is the structural deterioration of these remote areas (Figure 11.2.1b). It may arise following prolonged exposure to aberrant signals emanating from the damaged site, which may lead to excitotoxic reactions in connected areas (Coan et al., 2014; Ross and Ebner, 1990). It may also be caused by deficits in axonal transport (Hirokawa et al., 2010) or the prion-like spread of pathological agents across synapses (Frost and Diamond, 2009).

Dedifferentiation refers to a breakdown of the brain's normally segregated and specialized processes (Li et al., 2001; Rajah, 2005; Figure 11.2.1c). For example, patients with schizophrenia show reduced activation of the canonical frontoparietal system during the performance of tests of executive function, coupled with increased activation in other areas (Minzenberg et al., 2009). Studies of functional brain network topology have also found evidence of reduced **modularity** in this patient group (Alexander-Bloch et al., 2010). This diffuse and desegregated pattern of activity may arise from aberrant plasticity, abnormal neurodevelopmental wiring, or altered modulatory influences from critical catecholamines such as dopamine, which may impact the fine-tuning and signal-to-noise ratio of neural information processing (Winterer and Weinberger, 2004).

Adaptive responses are changes in brain activity that are designed to maintain homeostasis and performance where possible. Three concepts that are important for these adaptive responses are compensation, neural reserve, and **degeneracy**. Compensation occurs when the activity of unaffected neural elements is increased to overcome deficits or dysfunction arising from an insult elsewhere in the brain (Figure 11.2.1d). For example, a unilateral stroke affecting motor cortex often results in increased activation of the motor area ipsilateral to the affected hand (Rehme et al., 2012). This recruitment of ipsilateral motor cortex has been shown to play a causal role in maintaining adequate motor performance (Johansen-Berg et al., 2002; O'Shea et al., 2007). Neural plasticity and cognitive flexibility facilitate the engagement of compensatory strategies.

Neural reserve is the amount of intact neural tissue within an affected system that is still able to support function. Reserve is implied if brain activity and performance are unaffected by an insult (Figure 11.2.1e). Degeneracy is the capacity of other neural elements to assume the function of the damaged system (Figure 11.2.1f; see also Chapter 8). The degeneracy of a nervous system depends on whether neurodevelopment and experience-dependent plasticity have sculpted network topology to enable distinct neural ensembles to support overlapping functions.

Continued

BOX 11.2 INTERPRETING CLINICAL DIFFERENCES IN BRAIN NETWORKS—CONT'D

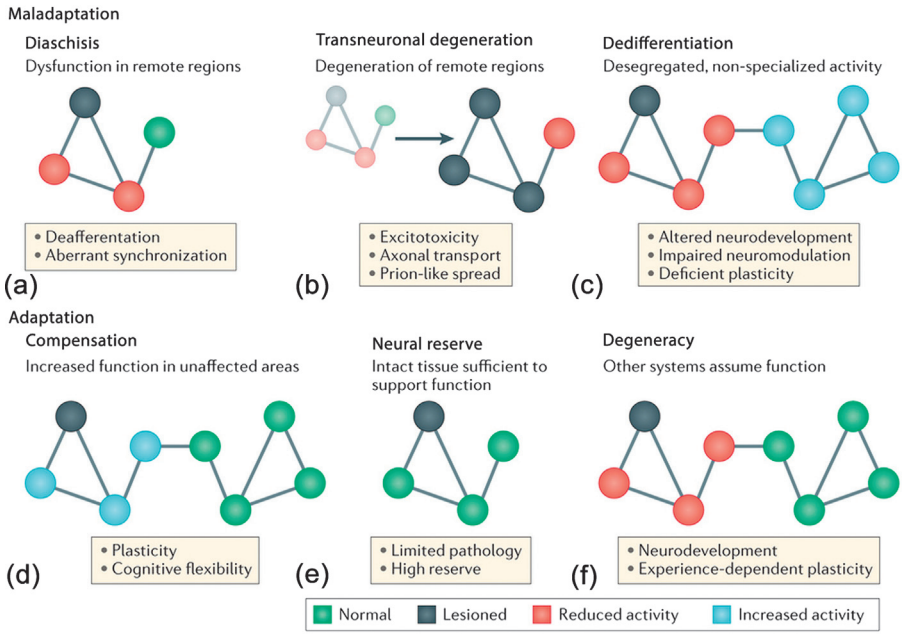

FIGURE 11.2.1 Maladaptive and adaptive responses to pathological perturbation of brain networks. This figure presents heuristic examples of how different neural responses to insult are expressed at the level of brain networks. Top row shows three examples of maladaptive responses. In each case, the behavioral output of the affected neural system will be impaired. Bottom row shows three examples of adaptive responses. In each case, the output of the system will be preserved as much as possible. Potential mechanisms underlying the changes are listed in the boxes under each graph. **(a)** In diaschisis, we see a reduction in the activity of structurally intact areas that are connected to a lesioned site. **(b)** In transneuronal degeneration, regions that are connected to a lesioned site show a structural deterioration over time (arrow). **(c)** Dedifferentiation occurs when the function of a system that is normally associated with a particular behavior is reduced, while the activity of other systems is increased. **(d)** Compensation occurs when other nodes of the affected neural system, or other neural systems, increase their activity to preserve behavior (the former is shown here). **(e)** If sufficient neural reserve is available following a lesion, there will be no behavioral impairment and no change in the activity of unaffected neural elements. **(f)** Degeneracy is evident when a second system assumes the function of the affected neural network, without any change in the activity of that second system (see Noppeney et al. (2004), for a discussion of other forms of degeneracy). Complex combinations of these responses may occur, and one may evolve into another over time. *Figure reproduced from Fornito et al. (2015) with permission.*

In this context, α/J is the *corrected* significance level that ensures the probability of making even *one* false rejection across the entire family of J tests is no more than α. The Bonferonni procedure follows from Boole's inequality, which asserts that the probability of rejecting the null hypothesis for at least one test in no more than the sum of the probabilities of rejecting each test.

The Sidak procedure (Sidak, 1967) is an alternative that offers a small gain in statistical power compared to the Bonferonni procedure, but requires that all the J tests are independent. With the Sidak procedure, we reject the null hypothesis for all p-values satisfying,

$$p_j \leq 1-(1-\alpha)^{1/J}, \quad j=1,\ldots,J, \tag{11.3}$$

where the right-hand side of this inequality comes from solving $\alpha = 1-(1-\alpha_1)^J$ for α_1, the uncorrected significance level. Note that $(1-\alpha_1)^J$ is the probability that the null hypothesis is accepted for all J tests, assuming independence between the tests, and thus $1-(1-\alpha_1)^J$ is the probability that the null hypothesis is rejected for at least one test.

The Holm-Bonferonni step-down procedure (Holm, 1979) is yet another alternative to the classic Bonferonni correction. Unlike the Sidak procedure, it does not assume independence between tests, although if the tests show positive dependence, the procedure becomes conservative. To ensure the FWER is no larger than a prescribed α, we identify the *smallest* j such that,

$$p_{(j)} > \frac{\alpha}{J+1-j}, \quad j=1,\ldots,J, \tag{11.4}$$

where $p_{(j)}$ denotes the jth smallest p-value. We then reject the null hypothesis for the $j-1$ smallest p-values; namely, the p-values that are smaller than $p_{(j)}$. All null hypotheses are accepted if the smallest p-value is greater than α/J. The Holm-Bonferonni procedure is universally more powerful than the classic Bonferonni correction.

11.2.4 Multiple Comparisons Under Dependence

Network measures computed across a family of nodes or connections are unlikely to be independent of each other. They usually show a positive dependence, although this issue has not been thoroughly studied and the general nature of any dependence is not known reliably. Positive dependence is especially evident in correlation-based networks. For example, if the neural activity at one node is disrupted, the functional connectivity (as measured by the correlation coefficient) between that node's activity and the activity at many other nodes may simultaneously show a disturbance. In this case, a test performed on each correlation coefficient will yield a set of test statistics that are positively

dependent. Under this kind of positive dependence, the probability that the null hypothesis is accepted for all J tests is greater than $(1-\alpha_1)^J$, and thus $1-(1-\alpha_1)^J$ becomes a conservative estimate of the true FWER. As a result, we lose sensitivity for declaring effects as significant when there are dependencies between the tests.

Accounting for correlated tests when correcting for multiple comparisons is a complicated issue. When the test statistics are independent, we might commit $x > 0$ false positives on average whenever a familywise error is committed. Under positive dependence, although familywise errors are rarer, due to the Bonferonni and Sidak procedures becoming conservative, they are each likely to encompass $x' > x$ false positives. In other words, familywise errors are rarer but their consequences are more severe under positive dependence. This is because a single disturbance at a given node or connection can spread to neighboring nodes and connections, causing multiple test statistics to simultaneously reach significance and thus giving rise to clusters of false positives. This effect can be quantified with the generalized FWER (Lehmann and Romano, 2005). Clarke and Hall (2009) show that when the test statistics have typically observed null distributions, clusters of false positives are no more common than in the independent case.

Multiple comparisons under dependence is an issue that has been tackled in statistical genetics. A common approach is to determine the effective number of independent tests, $J_{\text{eff}} \leq J$, and then perform the Sidak procedure, for example, such that the null hypothesis is rejected for all p-values satisfying,

$$p_j \leq 1-(1-\alpha)^{1/J_{\text{eff}}}, \quad j=1,\ldots J. \tag{11.5}$$

The effective number of tests can be estimated based on the **eigenvalues** (see Box 5.1) of the correlation matrix, since the eigenvalues quantify the extent of dependence between all pairs of tests (Li and Ji, 2005). Also see Leek and Storey (2008) for a related approach. Varoquaux et al. (2010) tackle the dependence issue in a fundamentally different way by representing edge-wise connectivity differences between two populations of networks as a series of independent and normally distributed perturbations of a group level covariance matrix.

In the connectomics literature, the dependence issue is either ignored, or resampling and permutation-based approaches are used. Permutation tests (Box 11.1) inherently ensure that any dependencies between tests carry through to the test statistic null distribution. Although generating empirical null distributions with permutation can be computationally burdensome, subsampling algorithms have been developed that improve computation times (Hinrichs et al., 2013). These factors, coupled with the ubiquity of parallel computing, make the application of permutation tests a natural and practical choice for many applications in statistical connectomics.

11.2.5 The False Discovery Rate

Bonferonni correction and related methods provide strong control of family-wise errors; that is, they ensure that the probability of a type I error does not exceed a desired α. However, they are generally considered too conservative for most practical applications in connectomics, particularly when the family is large. For example, if we have an undirected network of $N = 100$ nodes, $J = N(N-1)/2 = 4950$ (assuming a statistical test at each possible edge), and thus a typical FWER of $\alpha = 0.05$ mandates an uncorrected p-value of $0.05/4950 \approx 0.00001$. This uncorrected p-value corresponds to an effect size exceeding a Z-statistic of 4. Effects sizes of this magnitude may be rare in practice, leading to high false negative rates, or type II errors, where the null hypothesis is not rejected for many tests, even when it is false. It is for this reason that classic approaches to control the FWER in the strong sense are generally considered too conservative for performing inference on connectomic data.

Statistical power can be substantially improved if we are content to settle for a less conservative approach known as *weak* control of the FWER. Weak and strong control are the same when all the null hypotheses are true. However, with weak control, as soon as we reject the null hypothesis for at least one connection or node, control of the FWER is no longer guaranteed. In this case, we can be certain that an effect is present *somewhere* in the data, but we cannot be certain that we have localized that effect to a set of nodes or connections in a way that controls the FWER across the entire family of nodes or connections. Procedures that control the FWER in the weak sense usually rely on less conservative strategies for localizing effects to sets of nodes or connections. Control of the FDR is the most well-known such strategy.

The FDR is defined as,

$$\mathrm{FDR} = E\left[\frac{V}{\max(1, S+V)}\right], \tag{11.6}$$

where V and S are random variables denoting the total number of false positives, or *false discoveries*, and the total number of true positives, respectively. Therefore, $S+V$ is equal to the total number of discoveries, and thus $V/(S+V)$ is the proportion of false discoveries. The max in the denominator ensures that the FDR is zero when no discoveries are made at all. The FDR has been highly influential and was one the first alternatives to the FWER that gained broad acceptance across a range of scientific fields, including brain imaging (Genovese et al., 2002).

In Equation (11.6), E is used to denote the average value of a variable. The FDR is therefore the average proportion of false discoveries. What then do we mean by an FDR of 0.05, for example? Of all the null hypotheses we reject, on average,

we can expect 5% of them to be false positives. For instance, we might reject 100 null hypotheses, of which 5 are false positives, or equivalently, we might reject 20 null hypotheses, of which only one is a false discovery. In both cases, the FDR is 0.05. In this way, the FDR is adaptive in that the number of false discoveries (V) is normalized by the total number of discoveries ($S+V$). In other words, making one false discovery among 20 discoveries is probably acceptable, but most likely unacceptable among only two discoveries.

Controlling the FDR seeks to ensure that the proportion of false discoveries is on average less than a prescribed threshold α. Since we are controlling an average quantity, for any given experiment, the actual FDR may be lower or higher than α. However, if we perform many experiments, each time controlling the FDR at level α, we can be certain that the actual FDR averaged across these independent experiments is no more than α. In contrast, controlling the FWER guarantees the probability of even one false discovery is less than α for each experiment individually. Therefore, controlling the FDR provides greater power at the cost of an increased false positive rate.

Benjamini and Hochberg (1995) provide a well-known controlling procedure that ensures FDR $\leq \alpha$. We begin by sorting the p-values from smallest to largest, denoting the jth smallest p-value with $p_{(j)}$. We then identify the *largest* j such that,

$$p_{(j)} \leq \frac{\alpha j}{J}, \quad j = 1, \ldots, J, \tag{11.7}$$

and reject the null hypotheses corresponding to the p-values $p_{(1)}, \ldots, p_{(j)}$. Note that no rejections are made when $p_{(j)} > \alpha j/J$ for all $j = 1, \ldots, J$. Rearranging this expression as $Jp_{(j)}/j \leq \alpha$, we can see that $Jp_{(j)}/j$ is a bound on the expected number of false discoveries. This is known as the Benjamini and Hochberg (BH) procedure and was first described by Simes (1986) as a procedure to control the FWER in the weak sense.

Storey (2002) has proposed a modification of the BH procedure that is more powerful when the proportion of true null hypotheses, denoted with π_0, is small. In these cases, the BH procedure can be overly conservative, because it actually controls the FDR at level $\pi_0\alpha$, which is smaller than α if π_0 is small. Interested readers are referred to Storey (2002) for more details.

11.2.6 The Network-Based Statistic

The **network-based statistic** (NBS) is a network-specific approach to control the FWER in the weak sense when performing mass univariate testing on all connections in a network (Zalesky et al., 2010a). In analogy to statistical genetics, such analyses can be called connectome-wide analyses because they involve

performing a hypothesis test at each and every element of the connectivity matrix.

The NBS is used in settings where each connection is associated with a test statistic and corresponding p-value, and the goal is to identify groups of connections showing a significant effect while controlling the FWER. The approach is nonparametric and can be applied to both functional and structural brain networks.

The NBS seeks to exploit the topological characteristics of an effect to boost the power with which that effect can be detected. In particular, the NBS utilizes the fact that typical effects of interest in brain networks, such as the pathological consequences of disease or the changes in brain activity evoked by a cognitive task, are seldom confined to a single locus. Rather, they are distributed throughout the brain in a manner that is ultimately constrained by the anatomical network topology (Fornito et al., 2015). Therefore, any such effects on brain networks are likely to encompass multiple connections and nodes, which form interconnected subnetworks. This is akin to the cascade of failures seen in engineered interdependent networks, such as power grids and the Internet, where a failure or component breakdown in one node rapidly spreads to affect other connected elements (Buldyrev et al., 2010). A corollary of this observation is that interesting variations in network connectivity are more likely to span multiple edges, rather than be confined to individual connections. The NBS defines these interconnected subnetworks as connected components and ascribes a familywise error corrected p-value to each subnetwork using permutation testing. Recall from Chapter 6 that a connected component is a set of nodes for which a **path** can be found linking any pair of nodes in that set.

The NBS is somewhat analogous to cluster-based approaches developed for performing inference on statistical parametric maps in human neuroimaging (Bullmore et al., 1999; Nichols and Holmes, 2001). Instead of identifying clusters of voxels in physical space, the NBS identifies connected subnetworks in topological space. The size of a subnetwork is most typically measured by the number of edges that it comprises.

Figure 11.4 provides an overview of the main steps of the NBS. First, a univariate test statistic is independently computed for each and every connection. In this case, we depict a t-statistic that tests the null hypothesis of equality in the mean connectivity weight between two populations: a "control" population and a "patient" population (Figure 11.4a). The result is a matrix of test statistic values with the same dimensions as the connectivity matrix (Figure 11.4b, left). Note that for an undirected network, the adjacency matrix is symmetric about the matrix diagonal so we only need to compute statistics for the upper triangular part of the matrix. Next, we apply a primary, component-forming threshold to the matrix of test statistic values (Figure 11.4b, right). This thresholded

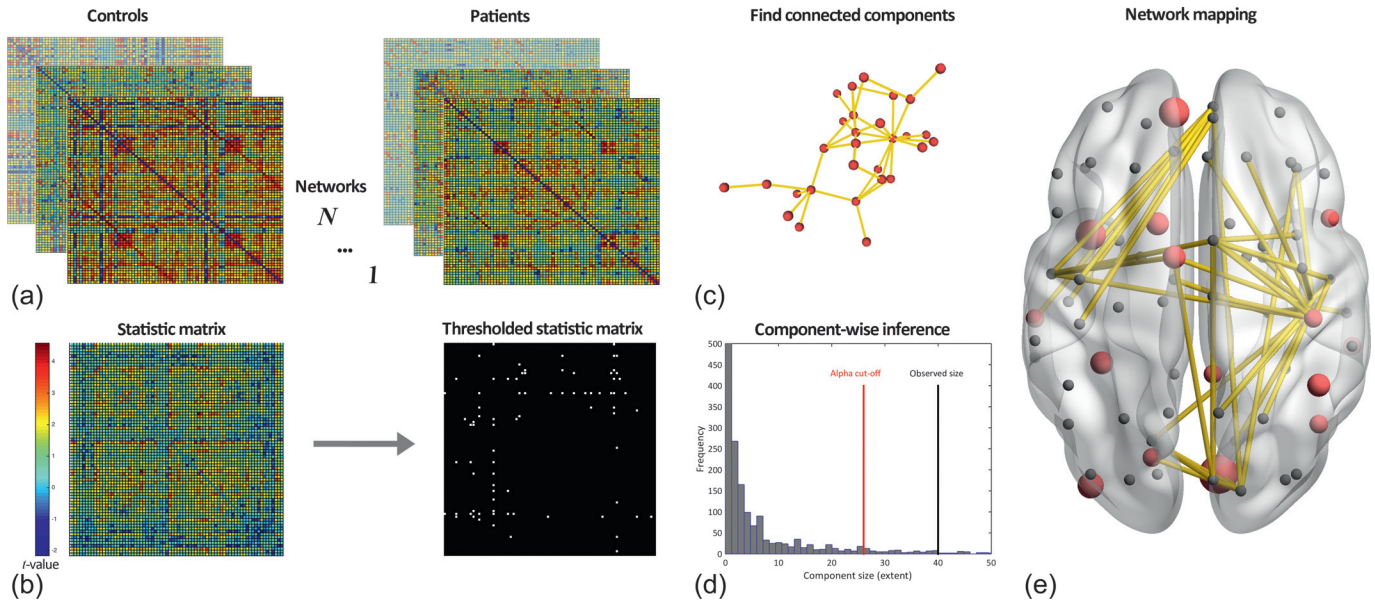

FIGURE 11.4 Key steps of the network-based statistic (NBS) analysis. The NBS methodology is illustrated with a comparison of functional connectivity networks in 15 healthy participants and 12 patients with chronic schizophrenia, as measured with resting-state functional MRI. Functional connectivity was measured for each pair of 74 anatomically defined regions using Pearson correlation between **wavelet** coefficients in the frequency interval $0.03 < f < 0.06$Hz. Details of the dataset are provided by Zalesky et al. (2010a). **(a)** We begin with two populations of connectivity matrices, one of controls (left) and one of patients (right). **(b)** A test statistic (in this case, a *t*-test) is computed at each and every matrix element, resulting in a matrix of statistic values (left). This matrix is then threshold using a primary, component-forming threshold to yield a thresholded and binarized statistic matrix (right). **(c)** The connected components of this thresholded, binarized statistic matrix are identified and the size of each (in terms of the number of links) is computed. Shown here is a topological projection of the largest connected component of the thresholded, binarized matrix depicted in the right panel of **(b)**. This component comprises 40 edges. **(d)** Data labels are permuted. In this case, the labels "control" and "patient" are randomly shuffled and reassigned to the connectivity matrices depicted in panel **(a)** and steps **(b–d)** are repeated. At each iteration, the size of the largest component is stored to generate an empirical null distribution of maximal component sizes, shown here. The red line indicates the cut-off for declaring a component size as statistically significant ($\alpha = 0.05$). The black line shows the observed size of the component illustrated in **(c)**. In this case, $p = 0.037$. **(e)** Projecting the network into anatomical space, we see that functional connectivity differences between patients and controls involve a distributed network of connections (yellow edges), largely centered on frontal and temporal areas. In this plot, node sizes correspond to the number of edges attached to each node in this network. Repeating the analysis using the FDR instead of the NBS found only one significant connection showing a difference between patients and controls (Zalesky et al., 2010a). The NBS thus offers a considerable gain in power, but cannot reject the null hypothesis at the level of individual edges.

matrix is then treated as a pseudo-network—a collection of edges showing an effect of interest at a nominal level of statistical confidence. This pseudo-network of test statistic values is then subjected to a breadth-first search (Box 6.3) to identify the number and size of connected components. These components can be interpreted as subnetworks of edges showing a common statistical effect of interest (Figure 11.4c). The size of each component is stored. The size of a connected component can be measured as the number of edges it comprises or the sum of the test statistic values associated with those edges. The latter approach is referred to as *cluster mass* in traditional image-based cluster statistics and can account for variation in effect sizes (Bullmore et al., 1999). Just counting the number of edges comprising a component ignores the effect size associated with each edge. If we sum the test statistic values, two components comprising the same number of edges can now have different sizes depending on the magnitude of the effect at each connection. In this way, the size of a cluster reflects a combination of both its topological extent and its effect size.

The observed component sizes are used by the NBS for statistical inference, so selecting an appropriate component-forming threshold is of critical importance. In principle, the threshold specifies the minimum test statistic value for which the null hypothesis can be potentially rejected and should be set according to the type of effects that are expected in the network. If we expect a strong effect restricted to a small subset of edges, the threshold should be conservative (e.g., convert the test statistics to p-values and remove all connections with $p < 0.001$, uncorrected), as it will result in smaller components comprising edges that respond strongly to the experimental manipulation. On the other hand, we might expect an effect to be weak at any individual edge, but to have a broad distribution across a large number of edges. In this case, we should use a less conservative threshold (e.g., $p < 0.05$, uncorrected). It is often difficult to know what type of effects should be expected, so it is useful to repeat analyses across a range of component-forming thresholds. Such an analysis allows us to understand whether effects are indeed strong but limited to a small number of edges, or weaker with a broader distribution across the network. In any case, it is important to remember that the FWER is controlled, at the level of connected components, irrespective of the primary threshold.

After the sizes of the observed components are computed, permutation testing (Box 11.1) is used to estimate a corrected p-value for the size of each observed component. In the example depicted in Figure 11.4, this involves randomly shuffling the labels assigned to each network so that the new "patient" and "control" groups comprise a mixture of actual patients and controls. The analysis is then repeated and the size of the *largest* component is stored. We repeat this procedure many times to generate an empirical null distribution of maximal component size. The corrected p-value for a component of size m is then

given by the proportion of permutations for which the largest component is equal to or greater in size than m (Figure 11.4d).

Using the null distribution of maximal component size ensures control of the FWER. To understand why, recall that the FWER is the probability of one or more false positives at any connection. Therefore, as shown by Nichols and Hayasaka (2003), if we use s_i to denote the size of the ith component and α the desired FWER, it follows that,

$$\begin{aligned} FWER &= P\left(\geq 1 \text{ component declared significant} \mid H_0\right) \\ &= 1 - P\left(0 \text{ components declared significant} \mid H_0\right) \\ &= 1 - P\left(\bigcup\nolimits_i \{s_i \leq t_\alpha\} \mid H_0\right) \\ &= 1 - P\left(\max\{s_i\} \leq t_\alpha\right) = \alpha, \end{aligned} \tag{11.8}$$

where H_0 denotes the global null hypothesis. By storing the size of the largest component, the null distribution $F(t_\alpha) = P(\max\{s_i\} \leq t_\alpha)$ is estimated empirically. Each permutation contributes a new data point to this sample of the null distribution. The component size threshold, t_α, for an α-level significance is then determined such that $t_\alpha = F^{-1}(1-\alpha)$. The crucial detail here is that ensuring the largest component is less than t_α guarantees that all components are less than this critical threshold. In particular, the above derivation shows that for all of the components to be of size less than t_α, we require that the largest component, $\max\{s_i\}$, is less than t_α.

With the NBS, the null hypothesis is always rejected at the level of components, not at the level of individual connections. Strictly speaking, it is therefore invalid to make inference about a specific connection embedded within a substantially larger component for which the null hypothesis is rejected. In other words, it is only valid to make inference about the component (subnetwork) as a whole. This is a consequence of weak control of the FWER. Thus, while the NBS can offer a considerable gain in statistical power (Zalesky et al., 2010a), this gain comes at the expense of not being able to localize effects to individual edges (e.g., Figure 11.4e).

11.2.7 Spatial Pairwise Clustering

Spatial pairwise clustering (SPC) is closely related to the NBS (Zalesky et al., 2012a). It is also a permutation-based approach that controls the FWER in the weak sense when performing mass univariate testing on all connections in a network. SPC differs from the NBS in the way that "clusters" are defined among connections in a network. Whereas the NBS defines clusters in terms of connected components, SPC uses a more stringent pairwise clustering approach that takes into account the network's spatial embedding.

To understand the difference between these clustering approaches, consider two connections (u_1, v_1) and (u_2, v_2). With the NBS, these two connections form

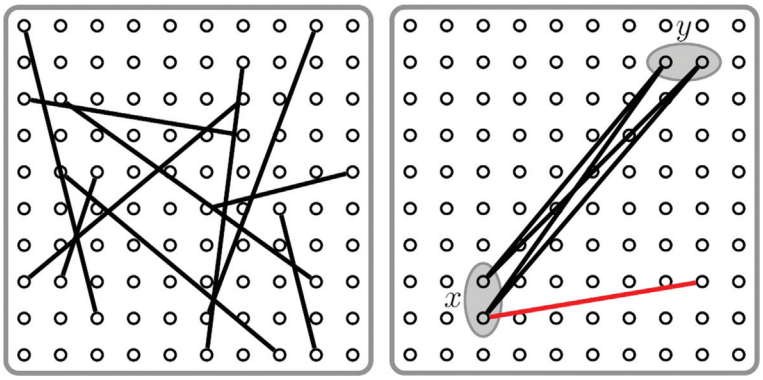

FIGURE 11.5 Connections forming a spatial pairwise cluster provide localised evidence for an effect. Connections distributed randomly (left) versus connections forming a spatial pairwise cluster (right). Each open circle is a node, while lines represent connections exceeding a predefined test statistic threshold. The pairwise cluster comprises four connections. The red connection is not part of the cluster because only one of its nodes is a neighbor. The presence of a pairwise cluster provides evidence of an effect between the two groups of nodes shaded gray. *Figure reproduced from Zalesky et al. (2012a) with permission.*

a cluster if $u_1 = u_2$ or $v_1 = v_2$, or equivalently, $u_1 = v_2$ or $v_1 = u_2$. A cluster defined as such is a connected component of size two. SPC defines clusters by taking into account the network's spatial embedding. In particular, (u_1, v_1) and (u_2, v_2) form a spatial pairwise cluster if the pair of nodes u_1 and u_2 as well as the pair of nodes v_1 and v_2 are spatial neighbors, or the same node. The concept of neighbors here is flexible. A pair of nodes can be considered spatial neighbors if they share a common border, or if the Euclidean distance between their centers does not exceed a given threshold. Figure 11.5 shows an example of a spatial pairwise cluster in a network that is embedded on a two-dimensional grid. Nodes are considered neighbors in this example if they are spatially **adjacent** relative to the grid.

The main advantage of SPC compared to the NBS is that effects can in some cases be localized with greater specificity to individual pairs of regions. The null hypothesis can therefore be rejected individually for each pair of regions. In contrast, the NBS is more likely to assimilate connectivity effects evident across distinct regions into a single, all-encompassing network. Figure 11.6 provides an example of this phenomenon in a study aiming to localize changes in human brain functional connectivity measured with **EEG** during a working memory task performed at different memory loads. SPC identified three separate spatial pairwise clusters, each corresponding to a different group of brain regions between which connectivity changed as a function of the working memory load. The null hypothesis can therefore be evaluated for each cluster individually. In contrast, the NBS identified a single subnetwork that

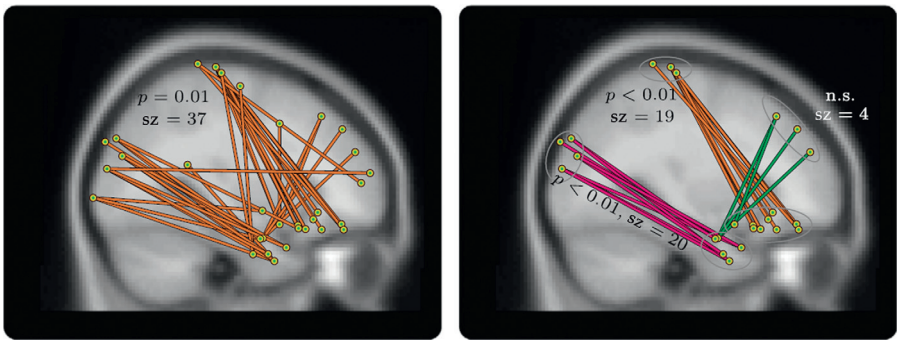

FIGURE 11.6 **Comparison of connections identified by the NBS and by spatial pairwise clustering (SPC).** Changes in functional brain connectivity across different working memory loads in healthy individuals were localized with the NBS (left) and SPC (right). Functional connectivity was measured with source analysis of EEG data. With the NBS, a single subnetwork can be declared significant, whereas SPC enables finer localization of this subnetwork, identifying three distinct spatial pairwise clusters, of which two can be declared significant ($p < 0.05$). The connections comprising a common cluster are colored identically. sz, size of the cluster; n.s., not significant. *Figure reproduced from Zalesky et al. (2012a) with permission.*

assimilated all three of these distinct spatial pairwise clusters identified with SPC. With the NBS, the null hypothesis was therefore only rejected at the level of the entire subnetwork.

SPC has a higher computational burden than the NBS. Moreover, the finer localization provided by SPC may be to the detriment of statistical power. In particular, a spatial pairwise cluster in itself may be too small in size to reach statistical significance; however, if it is assimilated into a connected component comprising other pairwise clusters, the connected component may be large enough for the null hypothesis to be rejected with the NBS.

SPC is most advantageous when performing inference on networks with high spatial resolution, such as voxel-based networks constructed with MRI or high-density source analysis of M/EEG. For coarser networks, it is more likely that a difference in connectivity between two regions is only sampled by a single pair of nodes, and thus a pairwise cluster cannot be formed. In these cases, the NBS may provide greater power. Variants of the SPC have also been developed for the analysis of dynamic functional connectivity networks (Hipp et al., 2011).

Clusters among connections in a network can in principle be defined in many different ways. Connected components and spatial pairwise clusters are perhaps the most obvious and intuitive definitions, hence their use in the NBS and SPC, respectively. Since there is no need to derive a theoretical null distribution when using permutation tests, the NBS and SPC methodologies can be

trivially adapted to suit any sensible definition of a cluster. For example, in their analysis of functional connectivity in hypercapnic states, Ing and Schwarzbauer (2014) define clusters as spatially contiguous groups of nodes/voxels, where each node/voxel is associated with at least one suprathreshold connection. In this way, we can identify groups of voxels that are associated with a connectivity effect. Clusters in the context of the NBS could also be defined in terms of network modules (Chapter 9).

11.2.8 The Screening-Filtering Method

Meskaldji and colleagues (2014) propose a screening-filtering method to control the FWER across a family of J tests. A closely related method was originally called subnetwork-based analysis (Meskaldji et al., 2011). The approach is generic and not specific to connectomic analysis.

The J tests in a family are first partitioned into m subsets, where each subset comprises $s = J/m$ tests. A summary statistic is determined for each subset by summing the s test statistics associated with each test and normalizing this sum by $\sqrt{s}$, thus yielding a normally distributed summary statistic with zero mean and unity variance under the null hypothesis. The subsets are then divided into two classes according to whether or not their summary statistic exceeds a prescribed threshold. As part of this screening step, the null hypothesis is accepted for all tests comprising subsets with a summary statistic below the prescribed threshold. The remaining *positive* subsets are given further consideration in a filtering step. The rationale here is that if the prescribed threshold is chosen to control the FWER across the m subsets, then weak control is guaranteed across the family of J tests.

The filtering step involves dividing the original p-values corresponding to all the tests comprising the positive subsets with a relaxation coefficient, $r > 0$. The classic Bonferonni correction is then applied to the relaxed p-values. The computation of r is achieved with an approximation algorithm. For a given r, the expected number of false positives can be computed if the proportion of true null hypotheses per each positive subset is known. These proportions are unknown in practice, and thus the algorithm loops over all feasible proportions, returning the worst-case expected number of false positives. The algorithm continues incrementing r by a small quantity until the worst-case expected number of false positives exceeds a desired α.

The success of the screening-filtering method is contingent on a judicious partitioning of the J tests into m subsets. In particular, a partitioning is required in which the tests corresponding to true positives are all allocated to as few of the m subsets as possible. Such a partitioning requires prior knowledge about the location of true positives, which is of course unknown. Meskaldji and

colleagues (2014) evaluate the power of the screening-filtering method using simulated datasets in which all the true positives are assigned to $m_1 = 2$, 5, or 10 *affected* subsets out of a total of $m = 20$ or 50 subsets. Each of the affected subsets comprises a proportion $\pi = 0.25$, 0.5, or 0.75 of true positives. Therefore, the total number of true positives is given by $J\pi m_1/m$. These parameter choices may provide a somewhat unrealistic representation of the partition quality that can be achieved in practice when we have no prior knowledge about the location of true positives. For example, when $J = 2000$, $m_1 = 5$, $m = 20$, and $\pi = 0.25$, in the absence of any prior knowledge, the probability of actually finding a given subset with $\pi \geq 0.25$ is in the order of 10^{-9}.

11.2.9 Sum of Powered Score

Kim et al. (2014) propose a sum of powered score (SPU) test, which was originally devised to identify rare variants in genome-wide association studies. Strictly speaking, the approach provides only a global test, and thus between-group differences cannot be localized to specific connections or nodes. If the null hypothesis is rejected, we can only claim that an effect is present somewhere in the network.

In the absence of any nuisance covariates, the simplest SPU test boils down to computing a score, U_j, for each hypothesis $j = 1, \ldots, J$, according to,

$$U_j = \sum_{i=1}^{n} X_{ij}(Y_i - \bar{Y}), \tag{11.9}$$

where X_{ij} is the value of the observation for the jth hypothesis of the ith individual, $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ is a binary vector partitioning the sample into two groups and $\bar{Y}$ is its mean. For example, X_{ij} might contain the edge weight for the jth connection of the ith individual. Intuitively, U_j measures the covariance between the observed data values $(X_{1j}, \ldots, X_{nj})$ and the demeaned grouping vector. The rationale here is that evidence against the null hypothesis grows as a function of the covariance between the observed data values and the grouping vector.

An omnibus test statistic is then computed according to $T_{\mathrm{SPU}} = \sum_{j=1}^{J} U_j^{\gamma}$, where $\gamma \geq 1$ is a parameter that can be chosen to give more weight to larger elements of U. Care must be taken when setting γ, since odd values may render the test statistic insensitive to bidirectional effects. To perform statistical inference, a null distribution is computed for T_{SPU} using a permutation test (Box 11.1), thereby yielding a corrected p-value. Kim and colleagues (2014) also consider an adaptive test statistic $T_{\mathrm{aSPU}} = \min_{\gamma \in \Gamma} P_{\mathrm{SPU}}(\gamma)$, where $\Gamma = \{1, 2, \ldots, \infty\}$ and $P_{\mathrm{SPU}}(\gamma)$ is the p-value computed for T_{SPU} with γ. This accounts for the arbitrariness in selecting γ. Once again, a permutation test is used to compute a p-value for T_{aSPU}.

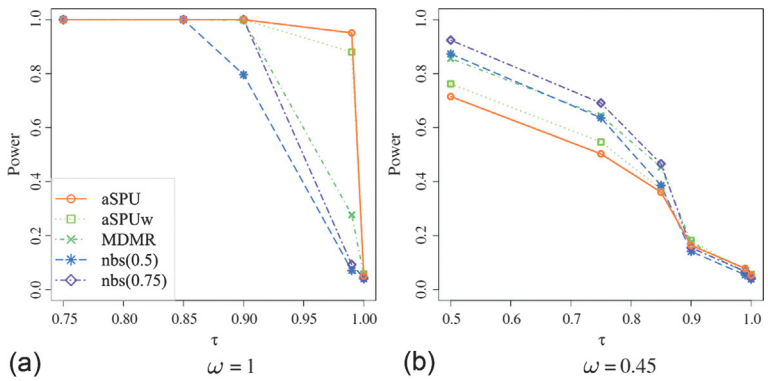

FIGURE 11.7 Comparison of the statistical power of the sum of powered score (SPU) and the network-based statistic (NBS). Between-group differences were simulated in a network comprising 2701 connections. The vertical axis (power) shows the probability of rejecting the global null hypothesis of no between-group difference anywhere in the network. The horizontal axis shows the proportion of connections at which the null hypothesis is true. The null hypothesis is true everywhere when the value of the horizontal axis is one, and thus this point represents the type 1 error rate. The parameter ω controls the effect size at each connection. In terms of the global null hypothesis, the NBS outperforms SPU when the effect size is small **(b)**, but SPU may be advantageous when the effect size is large and confined to a limited number of connections **(a)**. Note however that SPU does not localize rejection of the null hypothesis to particular connections or subnetworks. MDMR (multivariate matrix distance regression) is an alternative approach evaluated by Kim et al. (2014; see Section 11.3.2). Note that nbs(x) denotes the NBS with a primary threshold of x, and aSPU and aSPUw denote adaptive and weighted variants of the SPU, respectively. *Figure reproduced from Kim et al. (2014) with permission.*

To localize any effects to specific nodes or connections, Kim and colleagues suggest ranking the J tests based on their U score and plotting the true positive rate for the top-m ranked connections as a function of m. However, for any given value of m, we have no estimate of the FWER. Kim and colleagues used simulated data to compare the statistical power of the SPU test relative to the NBS. We can see in Figure 11.7 that the NBS outperforms SPU when the effect size is small, but SPU is advantageous when the effect size is large and confined to a limited number of connections.

11.3 MULTIVARIATE APPROACHES

As we have already seen, mass univariate hypothesis testing allows inference at the level of individual connectome elements and subnetworks. Although this offers tremendous localizing power, there may be fundamental limits to what we can learn about a connectome by studying its elements in isolation. This viewpoint follows from the notion of emergent phenomena—a fundamental tenet of systems biology—which contends that complex biological functions

emerge from complex interactions that cannot be trivially reduced to isolated elements and processes (Bassett and Gazzaniga, 2011).

Multivariate approaches seek to recognize and learn complex patterns among multiple connectome elements and utilize these patterns for inferential classification or prediction. This is in contrast to the methods described earlier in the chapter, which focus on fitting a model to minimize the sum of squared errors. We may wish to classify a group of patients into putative disease subtypes, or predict a patient's diagnosis based on connectivity patterns. Unlike mass univariate hypothesis testing, multivariate approaches usually avoid the multiple comparisons problem because the significance of an entire connectivity pattern is evaluated using a single test (Varoquaux and Craddock, 2013).

Multivariate approaches encompass a broad range of methods, including pattern recognition algorithms, principal component and discriminant analysis, and machine and deep learning algorithms. Many of these methods have been around for decades, but their application to the field of connectomics is in its early stages. Some readers may be familiar with the application of these methods to functional MRI data, under the guise of multivoxel pattern analysis (MVPA; Norman et al., 2006). MVPA utilizes pattern classification and machine learning methods to discover patterns of activity across multiple voxels in functional MRI data that can be used to distinguish between distinct cognitive states. In this section, we briefly overview some of the main approaches that have been used in the analysis of brain connectivity data.

11.3.1 Support Vector Machines

Support vector machines (SVMs) are the most common multivariate approach used in the connectomics literature. SVMs in their simplest form are used to classify an individual into one of two groups based on the presence or absence of complex combinations of connectome features. The most discriminatory feature combinations are learned during a training phase in which the SVM is presented with individuals from both groups and told the particular group to which each individual belongs. An optimization problem is then solved to generate a model that optimally maps each individual to a high-dimensional feature space in such a way that the two groups are maximally separated. To classify new individuals, they are mapped to the high-dimensional feature space and assigned to one of the two groups based on which side of the gap they fall. The reader is referred to the classic tutorial by Burges (1998) for an in depth mathematical treatment of SVMs. The topic is also covered in numerous books on machine learning and artificial intelligence (e.g., Mohri et al., 2012).

To evaluate the classification accuracy of an SVM, we usually apportion the data into training and validation subsets using methods such as *k*-fold cross-validation. The training subset is used to train the SVM, after which the SVM is then asked to classify each individual comprising the validation subset. Ideally, the training and validation subsets should comprise datasets acquired independently using a range of different techniques to ensure generalizability across different settings, although this is not always possible. In any case, it is crucial that the individuals allocated to the validation subset are kept hidden during the training phase. Feature selection is used to identify network elements that provide the greatest power in predicting an individual's membership of a particular group. The *t*-statistic is an example of a simple feature selector, in the sense that networks elements with a high *t*-statistic are likely to be more predictive. While feature selection is an important step to prevent model over-fitting, it should never be performed on the validation subset, since doing so can substantially inflate the estimated classification accuracies.

SVMs can also be extended to analyze continuous outcome variables, using a variant called support vector regression. Following the work of Dosenbach et al. (2010) in which an individual's age was predicted on the basis of his/her resting-state functional connectivity (Figure 11.8a), a number of SVM-based studies have reported the ability to classify various outcomes, such as cognitive or disease state, based on connectivity measures (e.g., Cole et al., 2013; Ekman et al., 2012; Zeng et al., 2012; Figure 11.8b). A particular advantage of these methods is that they allow inferences about individual cases. For example, it is in principle possible with SVMs to measure some aspect of brain structure or function in a new patient, compare the measure to an existing data set (the training set) and determine the probability that the new person belongs in one group or the other (i.e., is this person a patient or control?).

In view of this potential, SVMs and machine learning more generally have been proposed as methods that can potentially assist in providing a more objective, personalized and biologically based diagnosis of brain disorders. The ultimate aim of this work is to be able to determine the probability that a given individual carries a diagnosis of a particular disease given their pattern of brain connectivity. However, much of the work to date has focused on simply classifying patients and controls based on measures of brain structure or function. These distinctions are often trivial (and less expensive) for most clinicians to make without recourse to multivariate analysis of network parameters. Predicting illness characteristics that are more difficult to discern based on clinical examination, such as treatment response and illness course, is a more challenging task. SVMs trained to detect connectome pathology that is predictive of a particular response to a drug, for example, will most likely offer the greatest potential to influence clinical decision-making.

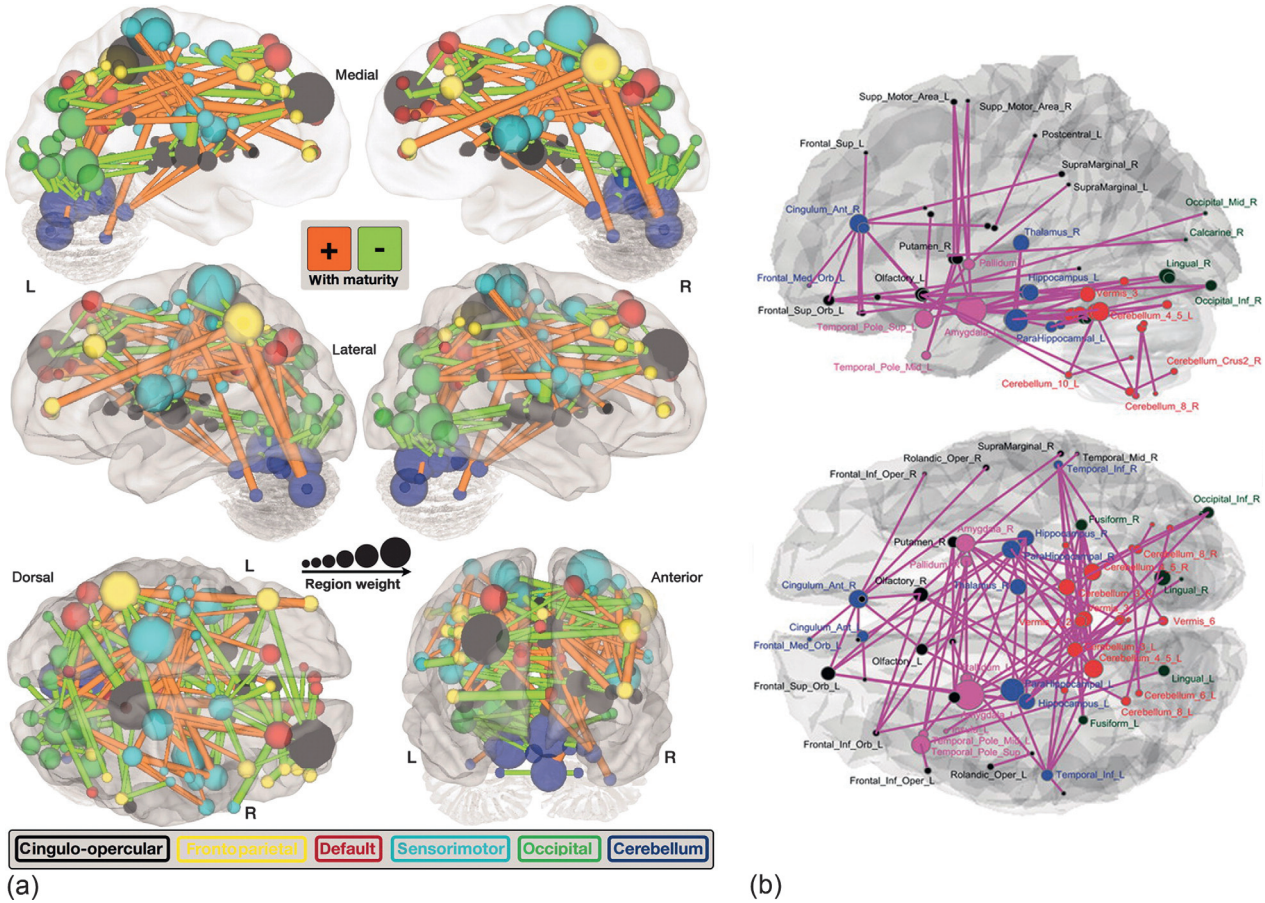

FIGURE 11.8 Combining connectomics and machine learning to investigate the demographic and clinical variability of human brain networks. **(a)** Functional links, as measured with human resting-state functional MRI, that contribute to predicting participant age in a support vector machine. Edge thickness scales with weight; node size with regional strength. Links positively correlated with age are in orange, links negatively correlated with age are in green. Nodes are colored according to the broad functional system to which they belong (see key at bottom of figure). L and R correspond to left and right hemispheres. **(b)** Functional links, measured with human resting-state functional MRI, that contribute to the discrimination between patients with major depression and healthy controls. In this analysis, patients were classified with 94% accuracy based on functional connectivity measures alone. Nodes are colored according to the functional system to which they belong (blue is default mode network; purple is an affective system; green is visual cortex; red is cerebellum; and gray denotes other areas). ***(a)*** *Reproduced from Dosenbach et al. (2010) and* ***(b)*** *from Zeng et al. (2012) with permission.*

11.3.2 Other Multivariate Approaches

Shehzad et al. (2014) develop a novel multivariate approach to analyze brain connectivity data. Unlike SVMs, their method does not allow for inference at the level of individuals. It is rather a multivariate statistic that combines evidence across multiple connections to maximize statistical power, although this can be to the detriment of highly focal effects. First used in the analysis of functional MRI data, the method performs inference on connections, but is implemented in such a way that the null hypothesis is rejected at the level of spatially contiguous clusters of voxels, not connections, yielding outputs that are similar in presentation to classic functional MRI activation studies.

In the analysis of functional MRI data, the method proceeds as follows. For each individual, we compute the functional connectivity between a given gray-matter voxel and all other gray-matter voxels, and store all these values as a single vector, with one vector for each of n individuals. Using these vectors, we then form a symmetric $n \times n$ matrix expressing the "distance" between each pair of individuals in terms of their connectivity patterns. In particular, element (i, j) of the distance matrix is populated with the distance measure $\sqrt{2(1-r_{ij})}$, where r_{ij} denotes the Pearson correlation coefficient between the vector of connectivity values for the ith and jth individual. This distance measure is zero for individuals with perfectly correlated connectivity patterns, one for uncorrelated patterns and two for perfectly negatively correlated patterns. Next, we use an established approach called multivariate distance matrix regression (MDMR) to test if the distances between individuals within the same group are shorter than the distances between individuals across different groups. MDMR is a multivariate test yielding a pseudo-F-statistic that assesses between-group differences or some other desired effect. This entire procedure is then repeated for all gray-matter voxels. To control the FWER across all voxels, a cluster-based permutation test is performed (Nichols and Holmes, 2001), which yields clusters of voxels for which the null hypothesis can be rejected. It is important to note that this approach does not test specific connections for an effect, but rather tests whether all the connections originating from a single node can cumulatively provide sufficient evidence to reject the null hypothesis at that node.

Deep learning and artificial neural networks are another group of promising multivariate approaches that are only beginning to be used to identify informative connectivity patterns in the connectome and other neuroimaging data (Plis et al., 2014). Partial least squares (PLS) is a well-established multivariate approach that can also be used to perform multivariate inference on brain networks (Krishnan et al., 2011; McIntosh et al., 1996). The basic idea of PLS is to

identify latent variables that express the largest amount of covariation between a set of response variables (e.g., clinical diagnosis or symptom scores), and a set of predictor variables (e.g., brain network connectivity or topological measures). Permutation testing can be used to assess the significance of each latent variable. Canonical correlation analysis (CCA) is related to PLS and has been used to identify multivariate relations linking demographic and psychometric measures with patterns of human brain functional connectivity (Smith et al, 2015). CCA is essentially a doubly multivariate approach that involves analyzing "correlations of correlations", and which does not readily allow for the statistical identification of specific connections, subnetworks and regions that underpin these multivariate relations. While these multivariate approaches have not gained substantial traction in the connectomics literature, it is likely that they will play a greater role in future studies using multivariate inference to link components of network topology to clinical, cognitive and neurobiological data.

11.4 SUMMARY

We began this chapter with a look at methods for thresholding connectivity matrices. Thresholding is intended to suppress spurious connections that may arise from measurement noise and imperfect connectome reconstruction techniques. While thresholding is not a mandatory step before performing statistical inference, it can potentially improve statistical power and interpretability. Density-based thresholding treats differences in connection density across populations as a confound, whereas weight-based methods assume that differences in density reflect meaningful variation in the distribution of connectivity weights. Methods for local thresholding can additionally be used to address the problems posed by network fragmentation that can occur with global thresholding. However, the process of local thresholding can in and of itself introduce nontrivial topological structure. Integration over a range of thresholds avoids arbitrariness in the choice of threshold.

The second part of the chapter focused on methods for performing statistical inference on brain networks. Inference can be performed at the level of network-wide measures, which is referred to as omnibus or global testing. We can also test hypotheses at the level of individual network elements. This allows us to identify the specific network elements responsible for an effect. In other words, we can reject the null hypothesis at the level of individual nodes, connections or subnetworks. For example, we may want to pinpoint specific connections that substantially differ in their connectivity weight between two populations of individuals. This is referred to as mass univariate hypothesis testing. We saw that the multiple comparisons problem is particularly sizeable when mass univariate testing is performed on all connections.

Network-specific methods for dealing with the multiple comparisons problem have been developed that provide greater power than generic multiple comparisons such as the FDR. These network-specific methods exploit the fact that effects in a brain network, as with many other biological and engineered networks, are more likely to form interconnected clusters, rather than appearing across many disconnected network elements.

Finally, we briefly discussed multivariate approaches for statistical inference. The advantage of methods such as support vector machines is that inference can be made at the level of individuals. Multivariate methods are likely to play a greater role in the future of statistical connectomics, as datasets continue to increase in size and complexity.