Chapter 1
Networks and Fundamental Concepts

Abstract This chapter introduces basic terminology and network concepts. Subsequent chapters illustrate that many data analysis tasks can be addressed using network methods. Network concepts (also known as network statistics or network indices) can be used to describe the topological properties of a single network and for comparing two or more networks (e.g., differential network analysis). Dozens of potentially useful network concepts are known from graph theory, e.g., the connectivity, density, centralization, and topological overlap. Measures of node interconnectedness, e.g., based on generalizations of the topological overlap matrix, can be used in neighborhood analysis. We distinguish three types of *fundamental* network concepts: (1) whole network concepts are defined without reference to modules, (2) intramodular concepts describe network properties of a module, and (3) intermodular concepts describe relationships between two or more modules. Intermodular network concepts can be used to define networks whose nodes are modules.

1.1 Network Adjacency Matrix

Networks can be used to describe the pairwise relationships between n nodes (which are sometimes referred to as vertices). For example, we will use networks to describe the relationships between n genes. We consider networks that are fully specified by an $n \times n$ dimensional **adjacency matrix** $A = (A_{ij})$, where the entry A_{ij} quantifies the connection strength from node i to node j. For an *unweighted* network, A_{ij} equals 1 or 0 depending on whether a connection (also known as link or edge) exists from node i to node j.

For a *weighted network*, A_{ij} takes on a real number between 0 and 1. A_{ij} specifies the connection strength between node i and node j. For an undirected network, the connection strength (A_{ij}) from i to j equals the connection strength from j to i (A_{ji}), i.e., the adjacency matrix A is symmetric $(A_{ij}) = (A_{ji})$. For a directed network, the adjacency matrix is typically not symmetric (see Sect. 11.4). Unless we explicitly mention otherwise, we assume in the following that we are dealing with an undirected network. As a convention, we set the diagonal elements to 1, i.e., $A_{ii} = 1$.

S. Horvath, *Weighted Network Analysis: Applications in Genomics and Systems Biology*,
DOI 10.1007/978-1-4419-8819-5_1,

In summary, we study networks whose adjacencies satisfy the following conditions:

$$\begin{aligned} 0 \leq A_{ij} &\leq 1, \\ A_{ij} &= A_{ji}, \\ A_{ii} &= 1. \end{aligned} \tag{1.1}$$

Many network applications use at least one node significance measure. Abstractly speaking, we define a *node significance measure* $GS = (GS_1, \ldots, GS_n)$ as a vector with n components that correspond to the network nodes. For the ith node, GS_i quantifies the significance or importance with regard to a particular application. The only assumption is that $GS_i = 0$ means that node i is not significant with regard to the application under consideration. We should emphasize that node significance does not necessarily correspond to statistical significance. For example, GS_i can be an indicator variable that equals 1 if prior literature suggests that node i is known to be important and 0 otherwise. If a statistical significance level (p value) is available for each node, then a p value-based node significance measure can be defined as follows:

$$GS_i = -log(p\ value_i). \tag{1.2}$$

In this case, GS_i is proportional to the number of zeroes of the ith p value. In gene network applications, gene significance measures allow one to incorporate external gene information into the network analysis. In functional enrichment analysis, a gene significance measure could indicate pathway membership. In gene knockout experiments, gene significance could indicate knockout essentiality.

1.1.1 Connectivity and Related Concepts

The *connectivity* (also known as degree) of the ith node is defined by

$$k_i = \sum_{j \neq i} A_{ij}. \tag{1.3}$$

In unweighted networks, the connectivity k_i equals the number of nodes that are directly linked to node i. In weighted networks, the connectivity equals the sum of connection weights between node i and the other nodes.

1.1.2 Social Network Analogy: Affection Network

Since humans are organized into social networks, social network analogies should be intuitive to many readers. Therefore, we will refer to the following "affection network" throughout this book. Each individual is represented by a node in the affection network. We assume that the connection strength (adjacency) between two individuals reflects how much affection they feel for each other. To be specific, we

assume that the affection (adjacency) A_{ij} equals 1 if two individuals strongly like each other, it equals 0.5 if they are neutral toward each other, and it equals 0 if they strongly dislike each other. Then the scaled connectivity K_i is a measure of relative popularity: high values of K_i indicates that the ith person is well liked by many others.

1.2 Analysis Tasks Amenable to Network Methods

Networks are useful for describing the relationships between objects (interpreted as network nodes). Networks are increasingly being used to analyze high-dimensional data sets where nodes correspond to variables (e.g., gene measurements). Networks facilitate sophisticated data analysis, which can often be described in intuitive ways. As social beings we function in social networks, which is why network language and terminology are very intuitive to us. For example, a network module can be interpreted as a social clique (e.g., a club) and highly connected hub nodes as popular people. Network methods can be used to address a variety of data analysis tasks including the following:

1. *To describe direct and indirect relationships between objects.* While the network adjacency matrix encodes direct first-order relationships, higher order relationships can be measured based on shared neighbors (see, e.g., Sect. 1.3.14)
2. *To carry out a neighborhood analysis*. Roughly speaking, a neighborhood is composed of nodes that are highly connected to a given "seed" set of nodes. Thus, neighborhood analysis facilitates a guilt-by-association screening strategy for finding nodes that are close to a given seed set of interesting nodes (see Sect. 1.4).
3. *To describe network properties using network concepts (also known as network statistics)*. We describe several types of network concepts in this and subsequent chapters.
4. *To describe the module structure of a data set*. Modules (groups, clusters, cliques) of nodes can be defined in many ways. Several module detection and clustering procedures are described in Chap. 8.
5. *To define shared "consensus" modules present in multiple data sets*. By construction, consensus modules can be found in two or more networks (see Sect. 7.11.1). Consensus modules may represent fundamental preserved structural properties of the network.
6. *To identify important modules*. For example, module significance measures can be used to identify gene modules that relate to cancer survival time (Sect. 5.7). A module significance measure can be defined by averaging a node significance measure across the module genes.
7. *To measure differences in connectivity patterns between two data sets.* Differential network analysis can be used to identify changes in connectivity patterns or module structure between different conditions (Sect. 1.11). Module preservation statistics are described in Chap. 9.

8. *To find highly connected "hub" nodes*. For example, highly connected intramodular hub nodes effectively summarize or represent the module.
9. *To reduce or compress the data*. For example, focusing the analysis on modules or their representatives (e.g., intramodular hub nodes) amounts to a network-based data reduction technique. Module-based analyses greatly alleviate the multiple testing problem that plagues many statistical analyses involving large numbers of variables.
10. *To annotate objects with regard to module membership*. For example, intramodular connectivity measures can be used to annotate all network nodes with respect to how close they are to the identified modules. This can be accomplished by defining a fuzzy measure of module memberships (intramodular connectivity) that generalizes the binary module membership indicator to a quantitative measure. Fuzzy measures of module membership can be used to identify nodes that lie intermediate between (i.e., close to) two or more modules.
11. *To develop network-based or module-based node screening procedures*. For example, gene pathway-based approaches for finding biologically important genes can be defined with regard to module membership measures (intramodular connectivity). In general, node-screening criteria can be based on a variety of network concepts (e.g., based on differential network analysis).

Throughout the book, we mention additional analysis tasks that can be addressed by more specialized networks. For example, correlation networks (described in Chap. 5) are constructed on the basis of correlations between numeric variables that can be described by an $m \times n$ numeric matrix *datX*. The nodes of a correlation network correspond to the columns of the matrix *datX*. Network concepts and methods can be used to describe the correlation patterns between the variables and to reduce the data. Although other statistical techniques exist for analyzing correlation matrices, network language and concepts are particularly intuitive. Statistically speaking, networks can be used as a data exploratory techniques (similar to cluster analysis, factor analysis, or other dimensional reduction techniques), as machine learning, data mining, and variable selection techniques. While sometimes established statistical techniques can be used to address similar goals, they are often far less intuitive to applied scientists. In contrast, network methods can usually be explained using social network analogies. Often the data being analyzed correspond to network measurements, e.g., genes operate in pathways or modules. It is natural to use network methods when one tries to model pathways.

1.3 Fundamental Network Concepts

In the following, we describe existing and novel network concepts (also known as network statistics or indices) that can be used to describe local and global network properties (Dong and Horvath 2007). The prime example of a fundamental network concept is the connectivity k_i (1.3). Sometimes network concepts are defined with

regard to a node significance measure GS_i. Abstractly speaking, a **fundamental network concept** is a function of the off-diagonal elements of A and/or a node significance measure GS. Below we present several network concepts including the density, maximum adjacency ratio, centralization, hub node significance, etc.

1.3.1 Matrix and Vector Notation

If M is a matrix and β is a real number, then M^β denotes the element-wise power, i.e., the ijth element of M^β is given by M_{ij}^β. Similarly, if v is a numeric vector, then the ith component of v^β is given by v_i^β. More generally, if $f()$ is a function that maps real numbers to real numbers, then $f(v)$ denotes the vector whose ith component is given by $f(v_i)$. We define $sum(M) = \sum_i \sum_j M_{ij}$ as the sum across all matrix entries, $max(M)$ as the maximum entry of matrix M, and $max(v)$ as the maximum component of the vector v. Similarly we define the minimum function $min(\cdot)$. We define the function $S_\beta(\cdot)$ for a vector v as $S_\beta(v) = \sum_i v_i^\beta = sum(v^\beta)$. Then $mean(v) = sum(v)/n$ and $variance(v) = sum(v^2)/n - (sum(v)/n)^2$. The transpose of a matrix or vector is denoted by the superscript $^\tau$. The **Frobenius matrix norm** is denoted by

$$\|M\|_F = \sqrt{\sum_i \sum_j m_{ij}^2} = \sqrt{sum(M^2)}. \tag{1.4}$$

Further denote by I the identity matrix and by $\mathrm{diag}(v^2)$ a diagonal matrix with its ith diagonal component given by $v_i^2, i = 1, \ldots, n$.

We briefly review two types of multiplying two $n \times n$ dimensional matrices A and B. The *component-wise* product $A * B$ yields an $n \times n$ dimensional matrix whose i, jth element is given by $A_{ij} * B_{ij}$. In contrast, the *matrix multiplication* AB yields an $n \times n$ dimensional matrix whose i, jth element is given by $\sum_{l=1}^n A_{il} B_{lj}$. Note that no multiplication sign is used for the matrix multiplication. In contrast, the multiplication sign $*$ between two matrices denotes their component-wise product. The R commands for carrying out these two types of multiplication are given by `A*B` and `A\%*\%B`, respectively.

1.3.2 Scaled Connectivity

The connectivity (node degree) k_i is probably the best known fundamental network concept. Many other network concepts are functions of the connectivity. For example, the *minimum connectivity* is defined as:

$$k_{min} = min(k), \tag{1.5}$$

where $min(k)$ denotes the minimum across the n components of the vector k. The *maximum connectivity* is defined as:

$$k_{max} = max(k). \tag{1.6}$$

Consider a network concept NC_i (such as the connectivity) that depends on a node index i (where $i = 1,\ldots,n$). Denote by max(NC) the maximum observed value across the n nodes. Then the **scaled version of the network concept** is defined as follows:

$$ScaledNC = \frac{NC}{max(NC)}. \tag{1.7}$$

For example, the **scaled connectivity** K_i of the ith node is defined by

$$ScaledConnectivity_i = \frac{k_i}{k_{max}} = K_i. \tag{1.8}$$

By definition the scaled connectivity lies between 0 and 1, i.e., $0 \leq K_i \leq 1$. Note that we distinguish the scaled from the unscaled connectivity using an uppercase "K" and a lowercase 'k', respectively. By definition $k_{max} \leq n-1$. Sometimes it is convenient to define the scaled connectivity (with a capital C) as follows:

$$C_i = \frac{k_i}{n-1}. \tag{1.9}$$

To avoid confusion, we should point out that the word "scale" has different meanings in different contexts. It has no relationships to the *scale*-free topology fitting index described in the following section.

1.3.3 Scale-Free Topology Fitting Index

Many studies have explored the frequency distribution of the connectivity, which can be defined based on the discretized connectivity vector $dk = discretize(k)$. The *discretize* function takes as input a numeric vector and outputs a vector of equal length whose components indicate the bin number into which the value falls. Denote the number of equal-width bins by *no.bins*. Then the uth component $dk_u = discretize(k, no.bins)_u$ reports the bin number $r = 1,2,\ldots,no.bins$ into which k_u falls. The *discretize* function is defined in (14.10). Denote by $p(r)$ the relative frequency of the rth bin, i.e., the proportion of components of k that fall into the rth bin. The **frequency distribution** of the connectivity can be estimated with $p(dk) = (p(1),\ldots,p(no.bins))$. Using this notation, we define the **connectivity frequency** $p.Connectivity$ (sometimes denoted $p(dk)$ or $p(k)$) as follows:

$$p.Connectivity = p(dk) = p(discretize(k, no.bins)), \tag{1.10}$$

which depends on the number of bins *no.bins*. As default, we set $no.bins = 10$ when discretizing the connectivity vector *Connectivity*.

Many network theorists have studied the properties of the frequency distribution of the connectivity $p.Connectivity = p(dk)$ (Barabasi and Albert 1999; Albert and Barabasi 2000; Jeong et al. 2001; Ravasz et al. 2002; Watts 2002; Han et al. 2004; Barabasi and Oltvai 2004; Pagel et al. 2007). In many (but certainly not all) real network applications, the frequency distribution $p(dk)$ follows a power law:

$$p(r) = PositiveNumber * r^{-\gamma} \tag{1.11}$$

where $PositiveNumber = \frac{1}{\sum_{r=1}^{no.bins} r^{-\gamma}}$ and γ denote positive real numbers. In this case, the network is said to exhibit **scale-free topology** (Barabasi and Albert 1999; Barabasi and Oltvai 2004; Albert et al. 2000) with scaling parameter γ. By taking the log of both sides of (1.11), one can verify that scale-free topology implies a straight line relationship between $log(p(r))$ and $log(r)$:

$$log(p(r)) = -\gamma * log(r) + log(PositiveNumber). \tag{1.12}$$

To measure the extent of a straight line relationship between $log(p(r))$ and $log(r)$, we define the **scale-free topology fitting index**

$$ScaleFreeFit(no.bins) = cor(log(p(dk)), log(BinNo))^2 \tag{1.13}$$

as the square of the correlation coefficient (5.12) between $log(p(dk))$ and $log(BinNo)$, where $BinNo = (1,2,\ldots,no.bins)$. We often use the following abbreviation $R^2 = ScaleFreeFit$.

Networks whose scale-free topology index R^2 is close to 1 are defined to be approximately scale free. One can visually inspect whether approximate scale-free topology is satisfied by plotting $log(p(k))$ versus $log(k)$ (see Fig. 1.5). In most real networks one observes an inverse relationship between $log(p(k))$and $log(k)$, i.e., γ is positive. Scale-free networks are extremely heterogeneous, and their topology being dominated by a few highly connected nodes (hubs) that link the rest of the less connected nodes to the system. Several models have been proposed for explaining the emergence of the power-law distribution (scale-free topology). For example, it can be explained using a network growth model in which nodes are preferentially attached to already established nodes, a property that is also thought to characterize the evolution of biological systems (Albert and Barabasi 2000). Scale-free networks display a remarkable tolerance against errors (Albert et al. 2000). Many networks satisfy the scale-free property only approximately. For example, Fig. 5.7 shows that for a yeast co-expression network, the connectivity distribution $p(r)$ is better modeled using an **exponentially truncated power law** (Csanyi and Szendroi 2004)

$$p(r) = PositiveNumber * r^{-\gamma} * exp(-\alpha r)$$

where $PositiveNumber = \frac{1}{\sum_{r=1}^{no.bins} r^{-\gamma} * exp(-\alpha r)}$, γ, and α denotes positive real numbers. On a log scale, an exponentially truncated power law is given as:

$$log(p(r)) = -\gamma * log(r) - \alpha r + log(PositiveNumber) \tag{1.14}$$

Potential Uses In Sect. 4.3, we use the scale-free topology index R^2 for formulating the scale-free topology criterion for network construction.

1.3.4 Network Heterogeneity

The *network heterogeneity* measure is based on the variance of the connectivity. Authors differ on how to scale the variance (Snijders 1981). We define it as the coefficient of variation of the connectivity distribution, i.e.,

$$Heterogeneity = \frac{\sqrt{var(k)}}{mean(k)} = \sqrt{\frac{n * sum(k^2)}{sum(k)^2} - 1}. \tag{1.15}$$

This heterogeneity measure is invariant with respect to multiplying the connectivity by a scalar.

Social Network Interpretation of the Heterogeneity: The heterogeneity can be used to measure the variation of popularity (connectivity) across the individuals.

Potential Uses of the Heterogeneity: Describing the reasons for and the meaning of the heterogeneity of complex networks has been the focus of considerable research in recent years (Albert et al. 2000; Watts 2002). As mentioned before, many complex networks have been found to exhibit approximate scale-free topology, which implies that these networks are highly heterogeneous.

1.3.5 Maximum Adjacency Ratio

For weighted networks, we define the *maximum adjacency ratio* of node i as follows:

$$MAR_i = \frac{\sum_{j \neq i} (A_{ij})^2}{\sum_{j \neq i} A_{ij}}, \tag{1.16}$$

which is defined if $k_i = \sum_{j \neq i} A_{ij} > 0$. One can easily verify that $0 \leq A_{ij} \leq 1$ implies $0 \leq MAR_i \leq 1$. Note that $MAR_i = 1$ if all nonzero adjacencies take on their maximum value of 1, which justifies the name "maximum adjacency ratio". By contrast, if all nonzero adjacencies take on a small (but constant) value $A_{ij} = \varepsilon$, then $MAR_i = \varepsilon$ will be small.

Social Network Interpretation of the Maximum Adjacency Ratio: $MAR_i = 1$ suggests that the ith individual does not form neutral relationships; this individual either strongly likes or dislikes others since all A_{ij} are either 0 or 1. In contrast, $MAR_i = 0.5$ suggests the ith individual forms less intense relationships with others.

Potential Uses of the Maximum Adjacency Ratio: Since $MAR_i = 1$ for all nodes in an unweighted network, the maximum adjacency ratio is only useful for weighted networks. The MAR can be used to determine whether a hub node forms moderate relationships with a lot of nodes or very strong relationships with relatively few nodes. To illustrate this point, we show in the following simple example that the *MAR* can be used to distinguish nodes that have the same connectivity. Assume a network (labeled by *I*) for which the adjacency between node 1 and every other node equals $A^{(I)}_{1,j} = 1/(n-1)$. Then $k^{(I)}_1 = \sum_{j\neq 1} A^{(I)}_{1,j} = (n-1)/(n-1) = 1$ and $MAR^{(I)}_1 = 1/(n-1)$. For a different network (labeled by *II*) where $A^{(II)}_{1,2} = 1$ and $A^{(II)}_{1,j} = 0$ if $j \geq 3$, the connectivity $k^{(II)}_1$ still equals 1 but $MAR^{(II)}_1 = 1$.

As aside, we mention that a directed network analog of MAR_i has been used in the analysis of metabolic fluxes (Almaas et al. 2004).

1.3.6 Network Density

To simplify notation, we will make use of the function *vectorizeMatrix* which turns an $n \times n$ dimensional symmetric matrix A into a vector whose $n*(n-1)/2$ components correspond to the upper-diagonal entries of A, i.e.,

$$vectorizeMatrix(A) = (A_{12}, A_{13}, \ldots, A_{n-1,n}). \tag{1.17}$$

Using this notation, the *network density* (also known as line density (Snijders 1981)) is defined as the mean off-diagonal adjacency and is closely related to the mean connectivity.

$$\begin{aligned} Density &= mean(vectorizeMatrix(A)) \\ &= \frac{\sum_i \sum_{j>i} A_{ij}}{n(n-1)/2} \\ &= \frac{mean(k)}{n-1} \approx \frac{mean(k)}{n}, \end{aligned} \tag{1.18}$$

where $k = (k_1, \ldots, k_n)$ denotes the vector of node connectivities.

Social Network Interpretation: The density measures the overall affection among individuals. A density close to 1 indicates that all individuals strongly like each other, while a density of 0.5 suggests the presence of more ambiguous relationships.

Below, we show that many module detection (and clustering) methods aim to find subnetworks with high density.

1.3.7 Quantiles of the Adjacency Matrix

Quantiles are used to describe the distribution of a variable. The $prob = 0$ quantile of a set of numbers is the minimum, the $prob = 0.25$ quantile is the first quartile, the $prob = 0.50$ quantile is the median, and the $prob = 1.0$ quantile is the maximum. Using this terminology, we define the network concept *prob-th quantile of the adjacency* as the *prob*-th quantile of the *off-diagonal* elements of the adjacency matrix

$$quantile_{prob}(A) = quantile_{prob}(vectorizeMatrix(A)), \tag{1.19}$$

which is the quantile of the vectorized adjacency matrix (1.17). The minimum and median values across the off-diagonal elements of the adjacency matrix are denoted by $quantile_0(A) = min(A)$ and $quantile_{0.5}(A) = median(A)$, respectively. The median adjacency $quantile_{0.5}(A) = median(A)$ can be considered a robust measure of network density. In Sect. 4.5, we use general quantiles for 'calibrating' different networks.

1.3.8 Network Centralization

The *network centralization* (also known as degree centralization (Freeman 1978)) is given by

$$\begin{aligned} Centralization &= \frac{n}{n-2}\left(\frac{max(k)}{n-1} - \frac{mean(k)}{n-1}\right) \\ &= \frac{n}{n-2}\left(\frac{max(k)}{n-1} - Density\right) \\ &\approx \frac{max(k)}{n} - Density. \end{aligned} \tag{1.20}$$

The centralization is 1 for a network with star topology; by contrast, it is 0 for a network where each node has the same connectivity. Note that a regular grid network where $mean(k) = max(k)$ has centralization 0.

Social Network Interpretation of the Centralization: The centralization of the affection network is close to 1, if one individual has loving relationships with all others who in turn strongly dislike each other. In contrast, a centralization of 0 indicates that all individuals are equally popular.

Potential Uses of the Centralization: While the centralization is a widely used measure in social network studies, it has only rarely been used to describe structural differences of metabolic networks (Ma et al. 2004). We have found that the centralization can be used to describe properties of cluster trees (Dong and Horvath 2007; Horvath and Dong 2008).

1.3.9 Clustering Coefficient

The (local) *clustering coefficient* of node i is a density measure of local connections, or "cliquishness". Let us first review its definition for an unweighted network (Watts and Strogatz 1998). If the ith node is connected to a pair of other nodes (direct neighbors), these three build a triple. Thus, the number of triples equals the number of links among the neighbors of node i. If two neighbors of node i are also linked, then these three nodes build a triangle in the unweighted network. Thus, the number of triangles equals the number of links among the neighbors of node i. In an unweighted network, the ith clustering coefficient is defined as the proportion of observed triangles among all possible triangles involving node i.

$$ClusterCoef_i = \frac{\text{number of triangles involving node i}}{\text{number of triples (i.e., possible triangles)}}. \tag{1.21}$$

Algebraically, the clustering coefficient can be calculated as follows:

$$\begin{aligned} ClusterCoef_i &= \frac{\sum_{j\neq i}\sum_{k\neq i,j} A_{ij}A_{jk}A_{ki}}{\left(\sum_{j\neq i} A_{ij}\right)^2 - \sum_{j\neq i}(A_{ij})^2} \\ &= \frac{\sum_{j\neq i}\sum_{k\neq i} A_{ij}A_{jk}A_{ki} - \sum_{j\neq i} A_{ij}^2 A_{jj}}{\left(\sum_{j\neq i} A_{ij}\right)^2 - \sum_{j\neq i}(A_{ij})^2}. \end{aligned} \tag{1.22}$$

We defined the clustering coefficient for a weighted network by simply evaluating (1.22) on a weighted adjacency matrix (Zhang and Horvath 2005). One can easily prove that $0 \leq A_{ij} \leq 1$ implies that $0 \leq ClusterCoef_i \leq 1$.

Social Network Interpretation of the Clustering Coefficient: The higher the clustering coefficient of an individual, the higher is the affection among his friends. The clustering coefficient is zero if all of his friends strongly dislike each other.

Potential Uses of the Clustering Coefficient: The mean clustering coefficient has been used to measure the extent of module structure present in a network. The relationship between the clustering coefficient and the connectivity has been used to describe structural (hierarchical) properties of networks (Ravasz et al. 2002).

1.3.10 Hub Node Significance

Now we will define a network concept that makes use of a node significance measure GS_i. To measure the association between connectivity and node significance, we propose the following measure of *hub node significance*:

$$HubNodeSignif = \frac{\sum_i GS_i K_i}{\sum_i (K_i)^2}, \tag{1.23}$$

When GS_i is proportional to the scaled connectivity ($GS_i = cK_i$), the hub node significance equals the constant of proportionality: $HubNodeSignif = c$. The hub node significance equals the slope of the regression line between GS_i and the scaled connectivity K_i if the intercept term is set to 0 (see Figs. 1.5 and 5.2c).

Social Network Interpretation of the Hub Node Significance: Assume that the node significance measures the grade point average of the ith individual. Then the hub node significance can be used to assess whether there is a relationship between popularity (connectivity) and grade point average.

Potential Uses of the Hub Node Significance: Several studies have shown that the relationship between connectivity and node significance (i.e., the hub node significance) carries important biological information. For example, in the analysis of yeast networks where nodes correspond to genes, highly connected hub genes are essential for yeast survival, and hub genes tend to be preserved across species (Albert et al. 2000; Jeong et al. 2001; Albert and Barabasi 2002; Carter et al. 2004; Han et al. 2004; Oldham et al. 2006). A detailed analysis shows that the positive relationship between connectivity and knockout essentiality cannot always be observed (Carlson et al. 2006), i.e., the hub gene significance can be close to 0.

1.3.11 Network Significance Measure

We define the *network significance measure* as the average node significance of the nodes:

$$NetworkSignif = \frac{\sum_i GS_i}{n}. \tag{1.24}$$

Social Network Interpretation of the Network Significance: The network significance simply measures the average grade point average among the individuals.

1.3.12 Centroid Significance and Centroid Conformity

A **network centroid** is a suitably defined centrally located node in a network. A centroid can be defined in many different ways, e.g., based on connectivity or other centrality measures. For example, the centroid can be defined as (one of the) most highly connected node(s) in the network. If a node significance measure GS_i has been defined for the network, then the *centroid significance* is simply the node significance of the centroid:

$$CentroidSignif = GS_{i.centroid}, \tag{1.25}$$

where $i.centroid$ is the node index of the centroid.

We define the *centroid conformity* of the *i*th node as the adjacency between the centroid and the *i*th node:

$$CentroidConformity_i = A_{i,i.centroid}. \tag{1.26}$$

Social Network Interpretation of the Centroid Conformity: In our affection network, we choose the most popular individual as centroid; then his or her grade point average is the centroid significance. The centroid conformity of the *i*th individual equals his or her affection (connection strength) with the most popular individual. The mean centroid conformity equals the average amount of affection felt by the most popular individual.

Potential Uses of the Centroid Conformity: In Chap. 6, we will characterize networks where the adjacency $A_{i,j}$ can be approximated by a product of the centroid conformities:

$$A_{i,j} \approx CentroidConformity_i CentroidConformity_j.$$

In Chap. 3, we use this insight to derive relationships among seemingly disparate network concepts.

1.3.13 Topological Overlap Measure

In an unweighted network, the number of direct neighbors of nodes i and j is given by $\sum_{l\neq i,j} A_{il}A_{jl}$. Then $numerator_{ij} = \sum_{l\neq i,j} A_{il}A_{jl} + A_{ij}$ equals the number of shared neighbors plus 1 if $A_{ij} = 1$ (i.e., if a direct link exists between nodes i and j). The *topological overlap measure* $TopOverlap_{ij}$ is a normalized version of $numerator_{ij}$. Specifically, topological overlap measure between nodes i and j is given by:

$$TopOverlap_{ij} = \begin{cases} \frac{\sum_{l\neq i,j} A_{il}A_{jl} + A_{ij}}{min\{\sum_{l\neq i} A_{il} - A_{ij}, \sum_{l\neq j} A_{jl} - A_{ij}\} + 1} & \text{if } i \neq j \\ 1 & \text{if } i = j. \end{cases} \tag{1.27}$$

Note that the denominator will never be 0 due to the addition of 1. The formula for the topological overlap matrix was first suggested in the context of unweighted networks, more specifically for protein–protein interaction networks (supplementary material of (Ravasz et al. 2002)).

We adapted the definition of the topological overlap measure (1.27) to weighted network and co-expression networks (Zhang and Horvath 2005; Carlson et al. 2006; Ghazalpour et al. 2006; Oldham et al. 2006; Horvath et al. 2006; Li and Horvath 2007). In the following, we use $0 \leq A_{ij} \leq 1$ to prove that $0 \leq TopOverlap_{ij} \leq 1$. Since $\sum_{l\neq i,j} A_{il}A_{jl} \leq \sum_{l\neq i} A_{il} - A_{ij}$ and $\sum_{l\neq i,j} A_{il}A_{jl} \leq \sum_{l\neq j} A_{jl} - A_{ij}$, we find that

$$\sum_{l\neq i,j} A_{il}A_{jl} \leq min\left\{\sum_{l\neq i} A_{il} - A_{ij}, \sum_{l\neq j} A_{jl} - A_{ij}\right\}.$$

Since $A_{ij} \leq 1$, this implies that the numerator of $TopOverlap_{ij}$ is smaller than the denominator.

Social Network Interpretation of the Topological Overlap Measure: In our affection network, two individuals have a high topological overlap if they like and dislike the same people. If two individuals share the same friends, they may be part of a clique, which we refer to as a module.

Potential Uses of the Topological Overlap Measure: Erroneous adjacencies can have a strong impact on network topological inference. To counter the effects of spurious or sparse adjacencies, it can be advantageous to use node similarity measures that are based on common interaction partners or on topological metrics (Ravasz et al. 2002; Brun et al. 2003; Chen et al. 2006; Chua et al. 2006). We will use the topological overlap measure as robust measure of interconnectedness.

The topological overlap measure has two major uses: first, it can be used in conjunction with a clustering method to define network modules as described below; second, it can be used to define the neighborhood of an initial set of nodes in a network. Intuitively speaking, a neighborhood is comprised of nodes that are highly connected to a given set of nodes. Below, we describe how neighborhood analysis facilitates a guilt-by-association screening strategy for finding nodes that interact with a given set of initial nodes.

As caveat, we point out that topological overlap-based analyses will only be useful in applications that satisfy the following basic assumption: *The more neighbors are shared between two nodes, the stronger is their relationship.* Our applications and several publications provide empirical evidence that the topological overlap matrix leads to biologically meaningful results (Ravasz et al. 2002; Zhang and Horvath 2005; Carlson et al. 2006; Ghazalpour et al. 2006; Oldham et al. 2006; Horvath et al. 2006; Li and Horvath 2007; Yip and Horvath 2007). But there will undoubtedly be situations when alternative similarity measures are preferable.

1.3.14 Generalized Topological Overlap for Unweighted Networks

Here, we describe extension of the topological overlap measure (referred to as generalized topological overlap, GTOM), which considers longer ranging relationships between nodes (Yip and Horvath 2007). Importantly, this generalization only applies unweighted networks. This work was done with **Andy Yip**.

For an unweighted network, one can define a distance measure $\mathrm{d(u,i)}$ as the length of the shortest path between nodes u and i. Define

$$N_1(i,j) = \{u \neq i,j \mid \mathrm{d(u,i)} = 1 \ \& \ \mathrm{d(u,j)} = 1\}$$

as the set of common, directly linked neighbors shared by i and j, i.e., the nodes in this set are one step away from both i and j. Similarly, define

$$N_1(i,-j) = \{u \neq i,j \mid \mathrm{d(u,i)} = 1\}$$

as the set of one-step neighbors of i excluding j. The number of elements in these sets is given by

$$\begin{aligned} |(N_1(i,j))| &= \sum_{u \neq i,j} A_{iu} A_{ju} \\ |N_1(i,-j)| &= \sum_{u \neq i,j} A_{iu}, \end{aligned} \tag{1.28}$$

where $|\cdot|$ denotes the number of elements. Note that $|N_1(i,-j)|$ equals the connectivity k_i.

For an unweighted network, we can express the topological overlap measure as follows:

$$t_{ij} = \frac{|N_1(i,j)| + A_{ij}}{min\{|N_1(i,-j)|, |N_1(-i,j)|\} + 1}. \tag{1.29}$$

We use the notation t_{ij} instead of $TopOverlap_{ij}$ to remind the reader that these formulas only apply for an unweighted network. In the following, we extend the topological overlap measure to keep track of longer ranging interactions. Toward this end, we define multistep neighbors. For an unweighted network, one can define the $m = 2$ neighbors of node i as those nodes that can be reached within two steps. For an unweighted network, $t_{ij}^{[m]}$ keeps track of the number of shared neighbors that can be reached within m steps (Yip and Horvath 2007).

The mth order GTOM measure is constructed by (a) counting the number of m-step neighbors that a pair of nodes share and (b) normalizing it to take a value between 0 and 1. Specifically, denote by $N_m(i,j)$ (with $m > 0$) the set of nodes (excluding i,j) that are reachable from i and j within a path of length m, i.e.,

$$N_m(i,j) = \{l \neq i, j \mid d(l,i) \leq m \;\&\; d(l,j) \leq m\}. \tag{1.30}$$

Similarly, we define

$$N_m(i,-j) = \{l \neq i, j \mid d(l,i) \leq m\}. \tag{1.31}$$

Generalizing (1.28) to m-step neighbors, we define

$$t_{ij}^{[m]} = \frac{|N_m(i,j)| + A_{ij}}{min\{|N_m(i,-j)|, |N_m(-i,j)|\} + 1}. \tag{1.32}$$

We call the matrix $T^{[m]} = [t_{ij}^{[m]}]$ the *mth order generalized topological overlap matrix* (GTOMm). This quantity simply measures the agreement between the nodes that are reachable from i and from j within m steps. When $m = 1$, we obtain back the original TOM in formula (1.27). It is convenient and intuitive to define the zeroth order topological overlap matrix GTOM0 as $T^{[0]} = A$, which only considers the direct link between the pair of nodes in question. In an exercise, you are asked to derive a computational formula for $T^{[m]}$ (1.32).

Figure 1.1 shows the generalized topological overlap measure in simple examples.

Figures 1.2 and 1.3 present a simple network where GTOM1 and GTOM2 lead to different neighborhoods.

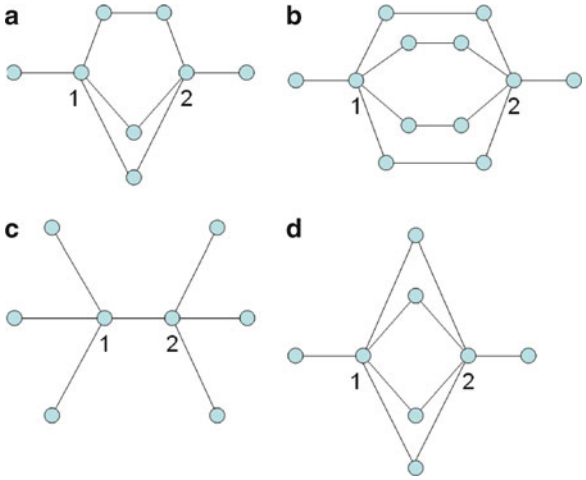

Fig. 1.1 Simple examples for illustrating the computation of the generalized topological overlap measure. (**a**) $A_{12}=0$,$t_{12}^{[1]}=0.4$, $t_{12}^{[2]}=0.67$, (**b**) $A_{12}=0$, $t_{12}^{[1]}=0$, $t_{12}^{[2]}=0.8$, (**c**) $A_{12}=1$, $t_{12}^{[1]}=0.25$, $t_{12}^{[2]}=1$, (**d**) $A_{12}=0$, $t_{ij}^{[1]}=0.67$, $t_{12}^{[2]}=0.8$

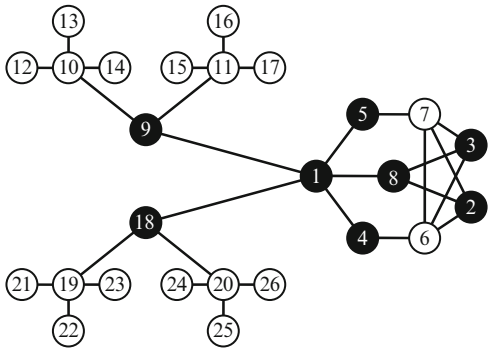

Fig. 1.2 GTOM1-based neighborhood of node 1. The eight closest neighbors of node 1 (with regard to GTOM1) are colored in black. Note that nodes 9 and 18 are part of this neighborhood

1.3.15 Multinode Topological Overlap Measure

In the following, we describe how to generalize the topological overlap measure to three or more nodes. This work was done with **Ai Li** (Li and Horvath 2007). While the standard TOM measure measures the number of neighbors shared by two nodes, the corresponding multinode measure keeps track of shared neighbors among multiple nodes. In light of formula (1.29), it is natural to define the MTOM of three different nodes i,j,k as follows:

$$t_{ijk}=\frac{|(N_1(i,j,k)|+A_{ij}+A_{ik}+A_{jk}}{min\{|N_1(i,j,-k)|,|N_1(i,-j,k)|,|N_1(-i,j,k)|\}+\binom{3}{2}}, \tag{1.33}$$

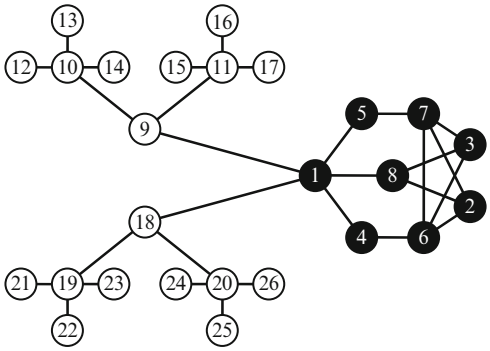

Fig. 1.3 GTOM1-based neighborhood of node 1. Note that nodes 9 and 18 are no longer among the eight closest nodes of node 1 since they do not share many of their two-step neighbors with node 1

where

$$N_1(i,j,-k) = \{u \neq i,j,k \mid \mathrm{d(u,i)} \leq 1 \;\&\; \mathrm{d(u,j)} \leq 1\}$$
$$N_1(i,j,k) = \{u \neq i,j,k \mid \mathrm{d(u,i)} \leq 1 \;\&\; \mathrm{d(u,j)} \leq 1 \;\&\; \mathrm{d(u,k)} \leq 1\}.$$

Here $N(i,j,-k)$ can be regarded as the set of the neighbors shared by i and j excluding k. The following algebraic formulas can be used to calculate these quantities:

$$|(N_1(i,j,k))| = \sum_{u \neq i,j,k} A_{iu}A_{ju}A_{ku}$$
$$|N_1(i,j,-k)| = \sum_{u \neq i,j,k} A_{iu}A_{ju}. \tag{1.34}$$

The binomial coefficient $\binom{3}{2} = 3$ in the denominator of (1.33) is an upper bound of $A_{ij}+A_{ik}+A_{jk}$ and equals the number of connections that can be formed between i, j, and k. Analogous to the proof for two nodes, one can prove that $0 \leq t_{ijk} \leq 1$. In the same way, we can define the MTOM of four nodes as follows:

$$t_{ijkl} = \frac{|(N_1(i,j,k,l)| + A_{ij} + A_{ik} + A_{il} + A_{jk} + A_{jl} + A_{kl}}{min\{|N_1(i,j,k,-l)|, |N_1(i,j,-k,l)|, |N_1(i,-j,k,l)| + |N_1(-i,j,k,l)|\} + \binom{4}{2}}, \tag{1.35}$$

where

$$N_1(i,j,k,-l) = \{u \neq i,j,k,l \mid d(u,i) \leq 1 \;\&\; d(u,j) \leq 1 \;\&\; d(u,k) \leq 1\}$$
$$N_1(i,j,k,l) = \{u \neq i,j,k \mid d(u,i) \leq 1 \;\&\; d(u,j) \leq 1 \;\&\; d(u,k) \leq 1 \;\&\; d(u,l) \leq 1\}. \tag{1.36}$$

These quantities can be computed as follows:

$$|(N_1(i,j,k,l))| = \sum_{u \neq i,j,k,l} A_{iu}A_{ju}A_{ku}A_{lu}$$
$$|N_1(i,j,k,-l)| = \sum_{u \neq i,j,k,l} A_{iu}A_{ju}A_{ku}. \quad (1.37)$$

It is straightforward to extend the definition of the topological overlap measure to more nodes. Note that the algebraic formulas for MTOM do not require that the adjacency matrix take on binary values. Since the formulas remain mathematically meaningful as long as $0 \leq A_{ij} \leq 1$, it is straightforward to use them for generalizing MTOM to **weighted networks**.

The MTOM measure is implemented in the stand-alone MTOM software, which can be downloaded from: www.genetics.ucla.edu/labs/horvath/MTOM/

1.4 Neighborhood Analysis in PPI Networks

The goal of neighborhood analysis is to find a set of nodes (the neighborhood) that is similar to an initial "seed" set of nodes. Mathematically speaking, a network interconnectedness measure is simply a node similarity measure, e.g., the adjacency matrix or the topological overlap measures could be used.

If individual network connections are susceptible to noise, it can be advantageous to use a robust interconnectedness measure, e.g., the topological overlap measure or the generalized topological overlap measure (GTOM) (Yip and Horvath 2007). A simple approach for defining a neighborhood of node i is to choose the nodes with highest adjacencies A_{ij}. In an unweighted network, this amounts to choosing the directly connected neighbors of node i. But sometimes it is advantageous to use the GTOM measures (described in Sect. 1.3.14) as we will illustrate using a protein–protein interaction network of the fly (Drosophila).

1.4.1 GTOM Analysis of Fly Protein–Protein Interaction Data

We downloaded the data from the general repository for interaction datasets (BioGRID). Protein knockout experiments in lower organisms (yeast, fly, worm) have shown that some proteins are essential for survival. In lower organisms, highly connected hub genes tend to be essential (Jeong et al. 2001, 2003; Hahn and Kern 2005; Carlson et al. 2006). Here we define neighborhoods around several seed sets of essential proteins. Specifically, we considered the 30 most highly connected essential proteins in the network and refer to them as essential hubs. The goal of the neighborhood analysis around each essential hub was to implicate other essential proteins.

The analysis assumes that a protein that is connected to an essential protein is likely to be essential as well. We used different GTOM measures to quantify whether two proteins are closely interconnected. The GTOM0 measure (i.e., the adjacency matrix) allows one to identify the directly linked neighbors of each essential hub protein. On average, each essential hub contained about 40 direct neighbors (according to GTOM0), which is why we considered a maximum neighborhood size of 40 for the GTOM1 and GTOM2 measures as well.

To assess the performance of GTOM measures, we kept track of how many essential proteins were found in the neighborhoods of each of the 30 essential hub proteins. The results were averaged over the 30 resulting neighborhood analyses. The proportions of essential proteins among the nearest neighbors of the essential hub proteins are shown in Fig. 1.4. Note that proteins that are close to essential hub proteins with respect to GTOM2 are more likely to be essential than proteins that are close with respect to GTOM1. Thus, keeping track of higher order neighbors increases the biological signal in this neighborhood analysis. But we should point out that the GTOM1 measure performs sometimes better in other neighborhood applications.

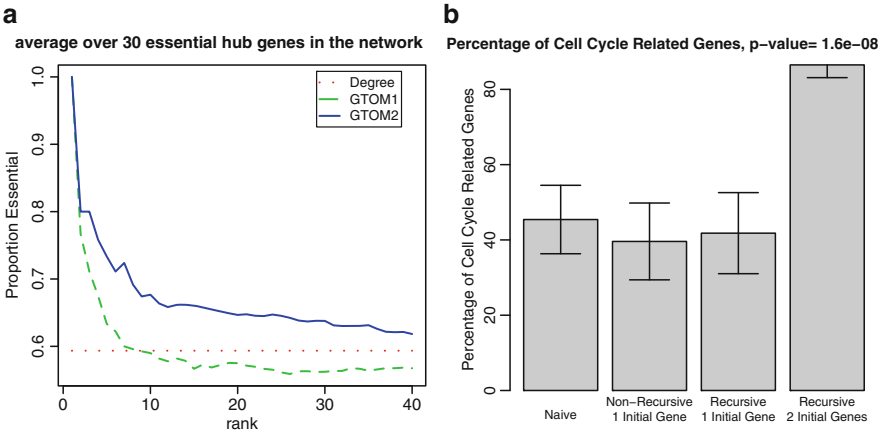

Fig. 1.4 Neighborhood analysis in protein–protein interaction networks. (**a**) Drosophila protein–protein interaction network. Proportions of essential proteins (*y*-axis) among the k nearest neighbors (*x*-axis) averaged over 30 essential hub proteins. For GTOM0 (binary adjacency matrix), the dashed horizontal line corresponds to the average proportion of essential proteins among the directly linked neighbors of the 30 essential hubs. Here GTOM2 outperforms GTOM0 or GTOM1. (**b**) Yeast Protein–Protein Physical Interaction Network (MIPS Data). Comparing the percent of cell cycle proteins R (*y*-axis) in different MTOM neighborhoods. Note that the recursive approach involving a seed of two cell cycle-related "hub" proteins performs better than approaches based on a single seed protein. Both recursive and the non-recursive MTOM neighborhood analysis involving a *single* seed protein tie with the naive approach of constructing a neighborhood based on the adjacency measure

1.4.2 MTOM Analysis of Yeast Protein–Protein Interaction Data

Here we illustrate how to use the multinode topological overlap measure (described in Sect. 1.3.15) in neighborhood analysis. This method is implemented in the MTOM software (Li and Horvath 2007). A limitation of many network similarity measures is that they measure pairwise similarity. While pairwise similarities are useful for clustering procedures, we show that it can be advantageous for neighborhood analysis to use a multi-node similarity measure such as MTOM.

Given an initial seed neighborhood, we consider two basic approaches for defining a neighborhood based on the concept of multi-node topological overlap. The default approach is to build the neighborhood recursively. The non-recursive alternative is computationally faster but produces less interconnected neighborhoods. The MTOM-based neighborhood analysis requires as input an initial seed neighborhood comprised of at least one node and the requested final size of the neighborhood $S \geq 1$.

1. **Recursive approach**
   a. For each node outside of the neighborhood set, compute its MTOM value with the nodes in the current neighborhood set.
   b. Add the node with the highest MTOM value to the neighborhood set.
   c. Repeat steps (a) and (b) until the neighborhood size S is reached.
2. **Non-recursive approach**
   a. For each node outside of the neighborhood set, compute its MTOM value with the nodes in the initial neighborhood set (the start seed).
   b. Choose the S nodes with the highest MTOM values as the neighborhood.

Since the recursive approach leads to neighborhoods with higher MTOM values, it is preferable over the computationally faster, non-recursive approach.

In the following, we describe an application of MTOM-based neighborhood analysis to predict cell cycle-related proteins in a yeast protein–protein interaction network (from the Munich Information Center for Protein Sequences (Guldener et al. 2006)). We restricted the analysis to the largest connected component comprised of 3,858 proteins with 7,196 direct pairwise interactions. To compare different neighborhood analysis approaches, we studied the neighborhoods of subsets of 101 cell cycle-related proteins found in the Kyoto Encyclopedia of Genes and Genomes (KEG). We considered a neighborhood size of 10. Within each neighborhood, we determined the number C of cell cycle-related proteins. We found that C is significantly correlated with the network connectivity k of the initial protein (Spearman correlation $r = 0.36$, p value ≤ 0.001). For this reason, we focused the neighborhood analysis on subsets of the 50 most highly connected "hub" cell cycle-related proteins. As can be seen from Fig. 1.4b, the neighborhoods of cell cycle genes tend to be enriched with other cell cycle genes as well. A major advantage of the MTOM-screening approach is the ability to input multiple initial nodes as seed set. Figure 1.4a shows that an initial seed neighborhood comprised of two cell

cycle-related hub proteins leads to far better results than using a single protein as input. But we found that this is only true for protein pairs that have high topological overlap.

1.5 Adjacency Function Based on Topological Overlap

Let us now return to the setting of a possibly weighted adjacency matrix. In practice, an original network adjacency matrix $A^{original}$ is often transformed into new network adjacency matrix denoted by A. For example, a transformation can be used to change the topological properties of a network. We define an **adjacency function** AF as a matrix valued function that maps an $n \times n$ dimensional adjacency matrix $A^{original}$ onto a new $n \times n$ dimensional network adjacency

$$A = AF(A^{original}).$$

In the following, we will describe an important adjacency function based on the topological overlap measure (1.27). The **topological overlap matrix (TOM)-based adjacency function** AF^{TOM} maps an original adjacency matrix $A^{original}$ to the corresponding topological overlap matrix, i.e.,

$$AF^{TOM}(A^{original})_{ij} = \frac{\sum_{l \neq i,j} A^{original}_{il} A^{original}_{l,j} + A^{original}_{ij}}{min\left(\sum_{l \neq i} A^{original}_{il}, \sum_{l \neq j} A^{original}_{jl}\right) - A^{original}_{ij} + 1}. \quad (1.38)$$

The TOM-based adjacency function AF^{TOM} is particularly useful when the entries of $A^{original}$ are sparse (many zeroes) or susceptible to noise. In this case it can be advantageous to replace $A^{original}$ by $AF^{TOM}(A^{original})$. This replaces the original adjacencies by a measure of interconnected that is based on shared neighbors. In Sect. 3.10, we describe how AF^{TOM} is used to transform sparse unweighted protein–protein interaction networks into weighted networks. The topological overlap measure can serve as a filter that decreases the effect of spurious or weak connections, and it can lead to more robust networks (Li and Horvath 2007; Yip and Horvath 2007; Dong and Horvath 2007).

1.6 R Functions for the Topological Overlap Matrix

There are two different `WGCNA` functions for calculating TOM based on the adjacency matrix. The first function `TOMsimilarity` takes as input a weighted (or possibly unweighted) adjacency matrix and outputs the first order ($m = 1$) GTOM measure. This function can*not* calculate higher order ($m > 1$) generalizations of the TOM measure. Often it is more convenient to calculate the TOM-based dissimilarity measure $dissTOM = 1 - TOM$, which can be accomplished with the R function `TOMdist`.

The second function `GTOMdist` calculates $distGTOM_{ij} = 1 - t_{ij}^{[m]}$, i.e., 1 minus the generalized topological overlap matrix. It is worth repeating that the input of this function is an *unweighted* network adjacency matrix which has binary entries (1 or 0). It is straightforward to use this function to calculate the generalized topological overlap measure as follows: `TOMm=1-`TOMdist(unweightedADJ,m) The following R code shows how to calculate the minimum value of the off-diagonal elements of a generalized topological overlap matrix:

```
library(WGCNA)
# number of nodes in the network
n=500
# set the seed of the random number generator
set.seed(1)
BinaryVector=sample(c(0,1),size=n*n, prob=c(.75,.25), replace=T)
BinaryMatrix=matrix(BinaryVector,nrow=n,ncol=n)
# here we define a symmetric unweighted adjacency matrix whose
unweightedADJ=BinaryMatrix * t(BinaryMatrix)
# this is the vector of values of m
mVector=0:5
minOffDiagonal=rep(NA,length(mVector))
for (i in 1:length(mVector) ){
m=mVector[i]
# Next we calculate the GTOMm matrix
# as 1 minus the corresponding dissimilarity
GTOMm=1-GTOMdist(unweightedADJ,degree=m)
# minimum off-diagonal element of GTOMm
minOffDiagonal[i]=min(GTOMm)
}
plot(mVector,minOffDiagonal,xlab="m")
abline(h=1)
minOffDiagonal
# output
0.0000000 0.0000000 0.6710098 0.9960000 0.9960000 0.9960000
```

One can show that the entries of the topological overlap matrix increase with m. In an exercise you are asked to show that the entries of GTOMm approximate 1 if m is chosen large enough.

The multinode topological overlap measure (MTOM) is implemented in a standalone software of the same name, which can be downloaded from: www.genetics.ucla.edu/labs/horvath/MTOM/.

1.7 Network Modules

Similar to the term "cluster", there is no general agreement on the meaning of the term "module". To make our results widely applicable, we provide a very general and abstract definition: a module is a set of nodes which forms a subnetwork. The subnetwork comprised of module nodes is referred to as network module. Since a particular network module may encode a pathway or a protein complex, these

special types of networks have great practical importance in biology. Assume that a module detection method (e.g., a clustering procedure) has found Q modules. Denote by $\mathscr{M}_q$ the set of node indices that correspond to the nodes in module q. We denote the adjacency matrix of the nodes inside the qth module by $A^{(q)}$. Analogously, we define $GS^{(q)}$ as the node significance measure restricted to the module nodes. Denote by $n^{(q)}$ the number of nodes inside the qth module.

Here we briefly mention an approach for defining network modules which is used in most of our applications. Typically, **we define modules as clusters of nodes that are highly interconnected** (as measured by the topological overlap measure). To simplify our notation, we introduce the **dissimilarity transformation** $D(A)$, which turns an adjacency matrix (which is a measure of node similarity) into a measure of *dissimilarity* by subtracting it from 1, i.e.,

$$D_{ij}(A) \equiv 1 - A_{ij}. \tag{1.39}$$

Note that $D(A)$ does not satisfy our definition of an adjacency matrix since its diagonal elements equal 0. Alternatively, one can use the topological overlap matrix (1.27) to define the *TOM-based dissimilarity measure*

$$\begin{aligned} dissTopOverlap_{ij} &= D_{ij}\left(AF^{TOM}(A)\right) \\ &= 1 - TopOverlap_{ij} \\ &= 1 - \frac{\sum_{u \neq i,j} A_{iu}A_{uj} + A_{ij}}{min(\sum_{u \neq i} A_{iu}, \sum_{u \neq j} A_{ju}) + 1 - A_{ij}}. \end{aligned} \tag{1.40}$$

As detailed in Sect. 8.4, we typically use the TOM-based dissimilarity as input of average linkage hierarchical clustering (Kaufman and Rousseeuw 1990), which results in a cluster tree (dendrogram) of the network. Next, network modules are defined as branches of the dendrogram. Hierarchical clustering and branch cutting are described in Sects. 8.4 and 8.6, respectively. This module detection procedure was originally proposed by Ravasz et al. 2002 for the setting of unweighted networks (Ravasz et al. 2002), but we find that it also works well for weighted networks. For example, Fig. 3.1 shows cluster trees (dendrograms) involving a gene network application. Genes or proteins of proper modules (branches) are assigned a color (e.g., turquoise, blue, etc.). Genes outside any proper module are colored gray. To date, this module detection approach has led to biologically meaningful modules in more than a hundred applications (see, e.g., Zhang and Horvath 2005; Horvath et al. 2006; Carlson et al. 2006; Ghazalpour et al. 2006; Gargalovic et al. 2006; Dong and Horvath 2007; Oldham et al. 2006, 2008).

While we often use hierarchical clustering to define modules, we emphasize that many alternative methods could be used. In the following, we assume that a module is a subnetwork inside a larger network.

1.8 Intramodular Network Concepts

*Intra*modular network concepts measure topological properties within a module.

Let us start out assuming that a module assignment Cl is available so that each node is assigned to exactly one of Q modules. For example, $Cl(i) = q$ if the ith node is in module q. The number of nodes in the qth module is denoted by $n^{(q)}$. Module assignment could be based on prior knowledge (e.g., Cl could encode groupings based on gene ontology information), or it could be the result of a clustering procedure (as described below and in Chap. 8). The following results do not depend on any particular module detection method. Instead, they are applicable for any network module, i.e., a subset of nodes which forms a subnetwork which is encoded by an $n^{(q)} \times n^{(q)}$ dimensional adjacency matrix $A^{(q)}$. Sometimes a corresponding node significance measure $GS^{(q)}$ is also available. We use the superscript (q) to denote quantities associated with the qth module. But for notational convenience, we sometimes omit superscript (q) when the context is clear.

An intramodular network concept is simply a (fundamental) network concept defined for $A^{(q)}$ and/or $GS^{(q)}$.

In the following, we present some particularly noteworthy intramodular network concepts. The **intramodular connectivity** $k_i^{(q)}$ (sometimes denoted by $\mathrm{kIM}_i^{(q)}$) is defined as the sum of connection strengths to other nodes within the same module, i.e.,

$$k_i^{(q)} = \mathrm{kIM}_i^{(q)} = \sum_{\substack{j \in \mathscr{M}_q \\ j \neq i}} a_{ij}^{(q)}, \tag{1.41}$$

where $\mathscr{M}_q$ is the set of node indices that correspond to the nodes in module q. Nodes with high intramodular connectivity are referred to as intramodular hub nodes. Intramodular connectivity has been found to be an important complementary node screening variable for finding biologically important genes (Horvath et al. 2006; Gargalovic et al. 2006; Oldham et al. 2006).

The **intramodular density** is defined as:

$$\begin{aligned} Density^{(q)} &= \sum_{i \in \mathscr{M}_q} \sum_{\substack{j \in \mathscr{M}_q \\ j \neq i}} \frac{A_{ij}^{(q)}}{n^{(q)}(n^{(q)} - 1)} \\ &= mean\left(vectorizeMatrix(A^{(q)}) \right). \end{aligned}$$

The density of nodes in a subnetwork (e.g., a module) can be used to find out whether this subnetwork is tight or cohesive. The goal of many module detection methods is to find clusters of nodes with high density (see Sect. 8.1).

We refer to the network significance (1.24) of a network module $A^{(q)}$ simply as the **module significance measure**, i.e., the module significance is the average node significance of the module nodes:

$$ModuleSignif^{(q)} = NetworkSignif(A^{(q)}) = \frac{\sum_{i \in \mathscr{M}_q} GS_i^{(q)}}{n^{(q)}}. \tag{1.42}$$

The module significance measure can be used to address a major goal of gene network analysis: the identification of biologically significant modules or pathways.

The **intramodular centroid conformity** $CentroidConformity_i^{(q)}$ with regard to the centroid (or medoid) in the qth network module $A^{(q)}$ is defined as follows:

$$CentroidConformity_i^{(q)} = a_{i,i.centroid}^{(q)}. \tag{1.43}$$

In Chap. 3.10, we will provide empirical evidence that for many network modules $a_{i,j}^{(q)}$ can be approximated by a product of conformities: $a_{i,j}^{(q)} \approx CentroidConformity_i^{(q)} * CentroidConformity_j^{(q)}$.

1.9 Networks Whose Nodes Are Modules

Here we outline how to define a network among modules, i.e., each node in the network corresponds to a module. Assume that we are studying two modules denoted by q_1 and q_2, respectively. Denote by $\mathscr{M}_{q_1}$ the set of $n^{(q_1)}$ nodes inside module q_1. The adjacencies between nodes of the two modules can be represented by an $n^{(q_1)} \times n^{(q_2)}$ dimensional sub-matrix $A^{(q_1,q_2)}$ of the full adjacency matrix A.

To define a measure of adjacency between the two modules, we summarize the matrix $A^{(q_1,q_2)}$ by a number between 0 and 1:

$$\begin{aligned} A_{q_1,q_2}^{ave} &= mean(A^{(q_1,q_2)}) = \frac{\sum_{i \in \mathscr{M}_{q_1}} \sum_{j \in \mathscr{M}_{q_2}} A_{ij}}{n^{(q_1)} n^{(q_2)}}, \\ A_{q_1,q_2}^{max} &= max(A^{(q_1,q_2)}) = max_{i \in \mathscr{M}_{q_1}, j \in \mathscr{M}_{q_2}} A_{ij}, \\ A_{q_1,q_2}^{min} &= min(A^{(q_1,q_2)}) = min_{i \in \mathscr{M}_{q_1}, j \in \mathscr{M}_{q_2}} A_{ij}. \end{aligned} \tag{1.44}$$

Since A^{ave} uses an average, it is statistically more robust than the maximum- or the minimum-based inter-adjacency measures. But in specific applications, the minimum- and maximum-based measures can be useful as well. The inter-adjacency measures (1.44) can be used to define a network between modules, e.g.,

$$A_{q_1,q_2} = \begin{cases} A_{q_1,q_2}^{ave} & \text{if } q_1 \neq q_2 \\ 1 & \text{if } q_1 = q_2 \end{cases}. \tag{1.45}$$

Denote by $A_{modules}$ the $Q \times Q$ dimensional symmetric matrix whose q_1, q_2 element is given by $A_{q_1 q_2}$ (which measures the adjacency between the two modules). The diagonal elements of $A_{modules}$ are set to 1. Note that $A_{modules}$ can be interpreted as adjacency matrix between modules, i.e., it represents a (weighted) network whose nodes are modules. Many alternative approaches exist for defining networks whose nodes are modules (see, e.g., Sects. 2.3 and 6.5). One can also define the adjacency between modules (comprised of overlapping sets of nodes) by assessing the probability that the two index sets $\mathscr{M}_{q_1}$ and $\mathscr{M}_{q_2}$ overlap. The probability that the two index sets overlap can be calculated based on the hypergeometric distribution (i.e., Fisher's exact test) (see Sect. 14.3). In gene network applications, adjacencies between modules have also been defined by measuring the probability of overlap between gene enrichment categories (Oldham et al. 2008).

1.10 Intermodular Network Concepts

*Inter*modular network concepts measure topological properties among modules. Here we use the notation from Sect. 1.9. A *fundamental intermodular network concept* $NC(q_1, q_2) = NCF(A^{(q_1,q_2)}, A^{(q_1,q_1)}, A^{(q_2,q_2)})$ is a function of $A^{(q_1,q_2)}$, $A^{(q_1,q_1)}$, and $A^{(q_2,q_2)}$. For example, the **geometric mean density of two modules** is defined as:

$$Density(q_1, q_2) = \sqrt{Density^{(q_1)} Density^{(q_2)}}. \tag{1.46}$$

Now we will describe network concepts that can be used to measure whether two modules are separated (distinct) from one another. Our **module separability statistics** contrast *inter*modular adjacencies (1.44) with *intra*modular adjacencies. We define separability statistics as 1 minus the ratio of intermodular adjacency divided by intramodular density:

$$separability.ave(q_1, q_2) = 1 - \frac{A^{ave}_{q_1,q_2}}{Density(q_1, q_2)}, \tag{1.47}$$

$$separability.max(q_1, q_2) = 1 - \frac{A^{max}_{q_1,q_2}}{Density(q_1, q_2)}, \tag{1.48}$$

$$separability.min(q_1, q_2) = 1 - \frac{A^{min}_{q_1,q_2}}{Density(q_1, q_2)}. \tag{1.49}$$

The separability statistics take on (possibly negative) values smaller than 1. The closer a separability statistic value is to 1, the more separated (distinct) are the two modules.

1.11 Network Concepts for Comparing Two Networks

Network concepts can be used to describe differences between two networks. Assume that two $n \times n$ dimensional adjacency matrices $A^{[\text{ref}]}$ and $A^{[\text{test}]}$ are available for a set of n nodes. $A^{[\text{ref}]}$ and $A^{[\text{test}]}$ may encode the connectivity patterns among genes before and after a biological perturbation experiment. $A^{[\text{ref}]}$ may encode the gene connectivity pattern in brain tissue, while $A^{[\text{test}]}$ reports the corresponding connectivity pattern in blood tissue.

It is natural to use network concepts to describe the differences between the reference and the test network. For example, if $NC^{[\text{ref}]}$ and $NC^{[\text{test}]}$ denote the values of a network concept in the reference and test network, then one can define a **differential network concept** as follows:

$$Diff.NC = NC^{[\text{ref}]} - NC^{[\text{test}]}. \tag{1.50}$$

Based on the above-mentioned fundamental (single) network concepts, one can define the following differential network concepts:

$$\begin{aligned} Diff.K &= Diff.ScaledConnectivity_i = K_i^{[\text{ref}]} - K_i^{[\text{test}]} \\ Diff.Density &= \frac{2}{n*(n-1)} \sum_i \sum_{i \neq j} \left(A_{ij}^{[\text{ref}]} - A_{ij}^{[\text{test}]} \right) \\ Diff.ClusterCoef_i &= ClusterCoef_i^{[\text{ref}]} - ClusterCoef_i^{[\text{test}]} \\ Diff.TopOverlap_{ij} &= TopOverlap_{ij}^{[\text{ref}]} - TopOverlap_{ij}^{[\text{ref}]}. \end{aligned} \tag{1.51}$$

If the ith node is highly connected in the first network but has a low connectivity in the second network, then $Diff.ScaledConnectivity$ takes on a large positive value.

By ranking the nodes according to a suitably defined differential network concept (1.50), one can find nodes that have different connectivity patterns across two networks. Sometimes it is useful to consider two differential network concepts to screen for interesting nodes.

For example, in correlation network applications, it can be useful to plot a node significance measure GS_i (e.g., based on a Student t-test of differential expression 10.6) on the y-axis and a measure of differential connectivity $Diff.K_i$ (1.51) on the x-axis. By thresholding both GS_i and $Diff.K_i$, one can define sectors that contain nodes with high node significance and/or high differential connectivity. To find significance thresholds, one can use permutation tests as described in Fuller et al. (2007). We refer to the scatter plot along with horizontal and vertical lines (corresponding to the threshold values) as **sector plot** for differential network analysis. Differential network analysis was used to identify highly significant sectors of obesity-related genes (Fuller et al. 2007) and gender-specific genes (van Nas et al. 2009) based on mouse gene expression data. Differential network analysis has also been used to study how network connections change with age (Swindell 2007) and between primate brains (Oldham et al. 2006).

Let us now define network concepts for measuring the similarity between two networks. Consider a (single) network concept NC_i with a node index i (e.g., the scaled connectivity K_i). To measure whether $NC^{[\text{ref}]}$ is correlated with $NC^{[\text{test}]}$ across nodes, one can define the **network concept correlation**:

$$cor.NC = cor(NC^{[\text{ref}]}, NC^{[\text{test}]}). \tag{1.52}$$

The correlation coefficient (5.10) is defined in Sect. 5.1.1. For example, the **connectivity correlation** is defined as the correlation coefficient between the connectivity vectors $K[[1]]$ and $K[[2]]$:

$$cor.K = cor(K^{[\text{ref}]}, K^{[\text{test}]}). \tag{1.53}$$

Since the correlation coefficient is scale invariant, the connectivity correlation also equals $cor.k = cor(k^{[\text{ref}]}, k^{[\text{test}]})$, where k denotes the (unscaled) connectivity vector. The connectivity correlation can be used to determine whether the connectivity patterns are preserved. High values of $cor.k$ indicate that the two networks have a similar module structure.

A network concept NC_{ij} with two node indices (e.g., the adjacency matrix or the topological overlap matrix) can be vectorized and correlated between the two networks:

$$cor.NC = cor(vectorizeMatrix(NC^{[\text{ref}]}), vectorizeMatrix(NC^{[\text{test}]})). \tag{1.54}$$

For example, the **adjacency correlation** is defined by

$$cor.Adj = cor\left(vectorizeMatrix(A^{[\text{ref}]}), vectorizeMatrix(A^{[\text{test}]})\right). \tag{1.55}$$

The above-mentioned network concepts can also be applied to a subnetwork formed by the nodes of a module. For example, connectivity preservation statistics quantify how similar connectivity of a given module is between a reference and a test network. For example, module connectivity preservation can mean that, within a given module q, nodes with a high connection strength in the reference network also exhibit a high connection strength in the test network. This property can be quantified by the correlation of intramodular adjacencies in reference and test networks. Specifically, if the entries of the first adjacency matrix $A^{[\text{ref}](q)}$ are correlated with those of the second adjacency matrix $A^{[\text{test}](q)}$, then the adjacency pattern of the module is preserved in the second network. Therefore, we define the *adjacency correlation* of the module q network as:

$$cor.Adj^{(q)} = cor\left(vectorizeMatrix(A^{[\text{ref}](q)}), vectorizeMatrix(A^{[\text{test}](q)})\right). \tag{1.56}$$

High $cor.Adj^{(q)}$ indicates that adjacencies within the module q in the reference and test networks exhibit similar patterns. If module q is preserved in the second network, the highly connected hub nodes in the reference network will often be

highly connected hub nodes in the test network. In other words, the intramodular connectivity $\text{kIM}^{[\text{ref}](q)}$ (1.41) in the reference network should be highly correlated with the corresponding intramodular connectivity $\text{kIM}^{[\text{test}](q)}$ in the test network. Thus, we define the correlation of intramodular connectivities as:

$$cor.kIM^{(q)} = cor\left(\text{kIM}^{[\text{ref}](q)}, \text{kIM}^{[\text{test}](q)}\right), \tag{1.57}$$

where $\text{kIM}^{[\text{ref}](q)}$ and $\text{kIM}^{[\text{test}](q)}$ are the vectors of intramodular connectivities of all nodes in module q in the reference and test networks, respectively. These and other measures will be used in Chap. 9 to define network-based measures of module preservation.

1.12 R Code for Computing Network Concepts

The function `fundamentalNetworkConcepts` in the `WGCNA` R package computes many of the above-mentioned (fundamental) network concepts based on an adjacency matrix and optionally a node significance measure. These network concepts are defined for any symmetric adjacency matrix (weighted and unweighted).

The following R code shows how to calculate network concepts for a randomly generated adjacency matrix.

```
library(WGCNA)
 n=1000
# set the seed for the random number generator
 set.seed(1)
 # now we simulate a random n*n dimensional matrix
# whose entries lie between 0 and 1
CF=runif(n)^2
# now we define A[i,j]=CF[i]*CF[j]
A=outer(CF,CF)
# diagonal elements are set to 1
diag(A)=1
# network connectivity equals the row sum -1
Connectivity = apply(A,2,sum,na.rm=T)-1
# partition the graphics window into 2 panels
par(mfrow=c(1,2))
# now we evaluate the scale free topology fit of the network
scaleFreePlot(Connectivity,main="Scale Free Topology")
# calculate several scale free topology fitting indices
scaleFreeFitIndex(Connectivity)
# define the scaled connectivity
K=Connectivity/max(Connectivity)
# now we simulate a node significance measure
trueHubNodeSignif=0.5
GS=trueHubNodeSignif*K+rnorm(n,sd=.02)
# Fundamental network concepts
NC=fundamentalNetworkConcepts(A,GS)
```

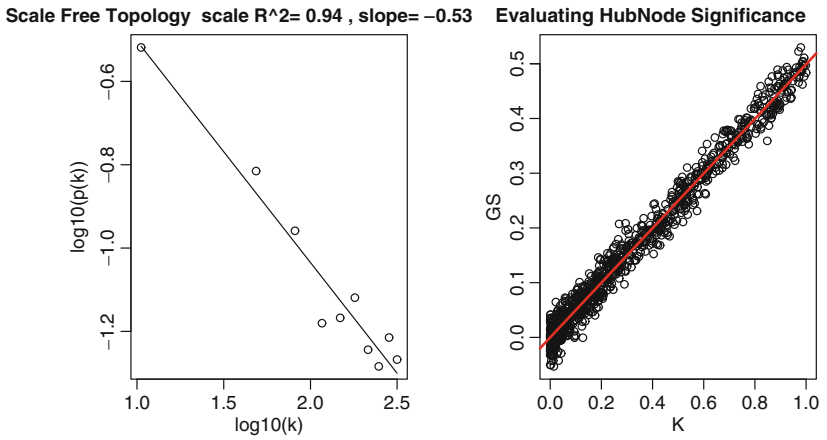

Fig. 1.5 Simulated network for illustrating scale-free fit (*left panel*) and the hub node significance (*right panel*). The R code used for generating this plot can be found in the text. Scale-free topology is approximately satisfied with $R^2 = 0.93$. The (observed) hub node significance is the slope of the line in the right figure, which results from using a linear model (without intercept term) to regress GS on K

```
NC
# scatterplot of GS versus K
plot(K,GS,main="Evaluating HubNodeSignificance")
 # fit a regression line without intercept term
lm1=lm(GS~K-1)
abline(lm1,col="red",lwd=2)
# The following output shows that the slope of the regression line
# equals the observed hub node significance (0.4990)
summary(lm1)
```

The graphical results from this R code are presented in Fig. 1.5.

1.13 Exercises

1. Exercise regarding network concepts. Consider an unweighted block diagonal adjacency matrix with two blocks. The first and second blocks contain $n^{(1)}$ and $n^{(2)}$ nodes, respectively. Each nonzero element of the first and second blocks equals b_1 and b_2, respectively. Calculate the following network concepts: connectivity k_i, scaled connectivity K_i, density, centralization, MAR, clustering coefficient, and topological overlap. Hint: Sect. 3.11. If i is in the first block, then $k_i = (n^{(1)} - 1)b_1$.

2. Exercise regarding alternative definitions of the topological overlap. Show that the following matrix satisfies the conditions of an adjacency matrix:

$$TopOverlap_{ij}^{average} = \begin{cases} \frac{\sum_{l\neq i,j} A_{il}A_{jl}+A_{ij}}{(\sum_{l\neq i} A_{il}+\sum_{l\neq j} A_{jl})/2-A_{ij}+1} & \text{if } i \neq j \\ 1 & \text{if } i = j. \end{cases} \tag{1.58}$$

   To calculate this alternative TOM measure, use `optionTOMDenom="mean"` in function `TOMsimilarity` (implemented by P. Langfelder).
3. Exercise regarding a computational formula for the mth order generalized topological overlap matrix GTOMm (based on (Yip and Horvath 2007)). Recall the definition of the GTOMm matrix $T^{[m]} = [t_{ij}^{[m]}]$ (1.32):

$$t_{ij}^{[m]} = \frac{|N_m(i,j)| + A_{ij}}{min\{|N_m(i,-j)|, |N_m(-i,j)|\} + 1}.$$

   Define the matrix $\tilde{A} = A - I$ whose diagonal elements equal 0. Let $\tilde{A}^m = \tilde{A} \ldots \tilde{A}$ denotes the matrix power based on the matrix multiplication described in Sect. 1.3.1.

   (i) Argue that the i,jth element of $\tilde{A}^m$ counts the number of paths of length m connecting nodes i and j. Note that the connecting paths may contain cycles.

   (ii) Argue that the matrix

$$S^{[m]} \equiv [s_{ij}^{[m]}] = \tilde{A} + \tilde{A}^2 + \ldots + \tilde{A}^m$$

   (where powers denote matrix powers) gives precisely the number of distinct paths with length smaller than or equal to m connecting each pair of nodes.

   (iii) Denote by $N_m(i)$ (with $m > 0$) the set of nodes (excluding i itself) that are reachable from i within a path of length m, i.e.,

$$N_m(i) = \{l \neq i \mid d(i,l) \leq m\} \tag{1.59}$$

   Note that $N_m(i) \equiv \{l \neq i \mid s_{il}^{[m]} > 0\}$. Define the binary matrix $B^{[m]}$ as follows:

$$b_{il}^{[m]} = \begin{cases} 1 \text{ if } s_{il}^{[m]} > 0 \text{ and } i \neq l, \\ 0 \text{ otherwise,} \end{cases}$$

   Show that $N_m(i) = \{l \neq i \mid b_{il}^{[m]} = 1\}$.

   (iv) Show that the number of shared m-step neighbors, $|N_m(i,j)| = |N_m(i) \cap N_m(j)|$ can be calculated as the inner product of the ith and the jth columns of $B^{[m]}$ which equals the i,jth element of $\left(B^{[m]}\right)^2$. Hint: $B^{[m]}$ is a symmetric matrix.

(v) Show that $|N_m(i)|$ is given by the i diagonal entry of $\left(B^{[m]}\right)^2$.

(vi) Show how the above results can be used to compute $T^{[m]}$ (1.32).

(vii) Why is it computationally advantageous to compute $S^{[m]}$ recursively by the matrix product between $\tilde{A}$ and $(S^{[m-1]} + I)$?

4. Exercise regarding the asymptotic behavior of GTOMm for large m. Recall the definition of the GTOMm matrix $T^{[m]} = [t_{ij}^{[m]}]$ (1.32).

(i) Use the R code in Sect. 1.6 to show empirically that the minimum value of the GTOMm measure increases with m.

(ii) Consider the situation where each pair of nodes is connected by a path of length $\leq m$. In this case, show that $|N_m(i)| = n - 1$ and $|N_m(i,j)| = |N_m(i) \cap N_m(j)| = n - 2$ and

$$t_{ij}^{[m]} = \frac{n - 2 + A_{ij} + 2I_{i=j}}{n - A_{ij}},$$

where n is the number of nodes in the network.

(iii) Show that large n implies that $T_{ij}^{[m]} \approx 1$.

(iv) Argue that for large enough m, the generalized topological overlap measure between each pair of nodes is 1.

(v) Argue that the GTOMm measure becomes uninformative if m is chosen too large. Comment: As default value, we choose $m = 1$. But sometimes $m = 0$ or $m = 2$ leads to more meaningful measures of interconnectedness. We expect that $m > 2$ is useful only when the unweighted input adjacency matrix contains many zeroes.

References

Albert R, Barabasi AL (2002) Statistical mechanics of complex networks. Rev Mod Phys 74:47–97

Albert R, Barabasi AL (2000) Topology of evolving networks: Local events and universality. Phys Rev Lett 85(24):5234–5237

Albert R, Jeong H, Barabasi AL (2000) Error and attack tolerance of complex networks. Nature 406(6794):378–382

Almaas E, Kovacs B, Vicsek T, Oltvai ZN, Barabasi AL (2004) Global organization of metabolic fluxes in the bacterium Escherichia coli. Nature 427:839–843

Barabasi AL, Albert R (1999) Emergence of scaling in random networks. Science 286(5439):509–512

Barabasi AL, Oltvai ZN (2004) Network biology: Understanding the cell's functional organization. Nat Rev Genet 5(2):101–113

Brun C, Chevenet F, Martin D, Wojcik J, Guenoche A, Jacq B (2003) Functional classification of proteins for the prediction of cellular function from a protein-protein interaction network. Genome Biol 5(1):R6

Carlson M, Zhang B, Fang Z, Mischel P, Horvath S, Nelson SF (2006) Gene connectivity, function, and sequence conservation: Predictions from modular yeast co-expression networks. BMC Genomics 7(7):40

Carter SL, Brechbuler CM, Griffin M, Bond AT (2004) Gene co-expression network topology provides a framework for molecular characterization of cellular state. Bioinformatics 20(14):2242–2250

Chen J, Hsu W, Lee ML, Ng S (2006) Increasing confidence of protein interactomes using network topological metrics. Bioinformatics 22:1998–2004

Chua NH, Sung W, Wong L (2006) Exploiting indirect neighbours and topological weight to predict protein function from proteinprotein interactions. Bioinformatics 22:1623–1630

Csanyi G, Szendroi B (2004) Structure of a large social network. Phys Rev 69:1–5

Dong J, Horvath S (2007) Understanding network concepts in modules. BMC Syst Biol 1(1):24

Freeman L (1978) Centrality in social networks: Conceptual clarification. Soc Networks 1:215–239

Fuller TF, Ghazalpour A, Aten JE, Drake T, Lusis AJ, Horvath S (2007) Weighted gene coexpression network analysis strategies applied to mouse weight. Mamm Genome 18(6–7):463–472

Gargalovic PS, Imura M, Zhang B, Gharavi NM, Clark MJ, Pagnon J, Yang WP, He A, Truong A, Patel S, Nelson SF, Horvath S, Berliner JA, Kirchgessner TG, Lusis AJ (2006) Identification of inflammatory gene modules based on variations of human endothelial cell responses to oxidized lipids. Proc Natl Acad Sci USA 103(34):12741–12746

Ghazalpour A, Doss S, Zhang B, Plaisier C, Wang S, Schadt EE, Thomas A, Drake TA, Lusis AJ, Horvath S (2006) Integrating genetics and network analysis to characterize genes related to mouse weight. PloS Genet 2(2):8

Guldener U, Munsterkotter M, Oesterheld M, Pagel P, Ruepp A, Mewes HW, Stumpflen V (2006) MPact: The MIPS protein interaction resource on yeast. Nucleic Acids Res 34:436–441

Hahn MW, Kern AD (2005) Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. Mol Biol Evol 22(4):803–806

Han JD, Bertin N, Hao T, Goldberg DS, Berriz GF, Zhang LV, Dupuy D, Walhout AJ, Cusick ME, Roth FP, Vidal M (2004) Evidence for dynamically organized modularity in the yeast protein-protein interaction network. Nature 430(6995):88–93

Horvath S, Dong J (2008) Geometric interpretation of gene co-expression network analysis. PLoS Comput Biol 4(8):e1000117

Horvath S, Zhang B, Carlson M, Lu KV, Zhu S, Felciano RM, Laurance MF, Zhao W, Shu Q, Lee Y, Scheck AC, Liau LM, Wu H, Geschwind DH, Febbo PG, Kornblum HI, Cloughesy TF, Nelson SF, Mischel PS (2006) Analysis of oncogenic signaling networks in glioblastoma identifies ASPM as a novel molecular target. Proc Natl Acad Sci USA 103(46):17402–17407

Jeong H, Mason SP, Barabasi AL, Oltvai ZN (2001) Lethality and centrality in protein networks. Nature 411:41

Jeong H, Oltvai Z, Barabasi A (2003) Prediction of protein essentiality based on genome data. ComPlexUs 1:19–28

Kaufman L, Rousseeuw PJ (1990) Finding groups in data: An introduction to cluster analysis. Wiley, New York

Li A, Horvath S (2007) Network neighborhood analysis with the multi-node topological overlap measure. Bioinformatics 23(2):222–231

Ma HW, Buer J, Zeng AP (2004) Hierarchical structure and modules in the Escherichia coli transcriptional regulatory network revealed by a new top-down approach. BMC Bioinform 5(1):199

van Nas A, GuhaThakurta D, Wang SS, Yehya N, Horvath S, Zhang B, Ingram-Drake L, Chaudhuri G, Schadt EE, Drake TA, Arnold AP, Lusis AJ (2009) Elucidating the role of gonadal hormones in sexually dimorphic gene coexpression networks. Endocrinology 150(3):1235–1249

Oldham MC, Horvath S, Geschwind DH (2006) Conservation and evolution of gene coexpression networks in human and chimpanzee brains. Proc Natl Acad Sci USA 103(47):17973–17978

Oldham MC, Konopka G, Iwamoto K, Langfelder P, Kato T, Horvath S, Geschwind DH (2008) Functional organization of the transcriptome in human brain. Nat Neurosci 11(11):1271–1282

Pagel M, Meade A, Scott D (2007) Assembly rules for protein networks derived from phylogenetic-statistical analysis of whole genomes. BMC Evol Biol 7(Suppl 1):S16

Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabasi AL (2002) Hierarchical organization of modularity in metabolic networks. Science 297(5586):1551–1555

Snijders TA (1981) The degree variance: An index of graph heterogeneity. Soc Networks 3:163–174

Swindell W (2007) Gene expression profiling of long-lived dwarf mice: Longevity-associated genes and relationships with diet, gender and aging. BMC Genomics 8(1):353

Watts DJ (2002) A simple model of global cascades on random networks. Proc Natl Acad Sci USA 99(9):5766–5771

Watts DJ, Strogatz SH (1998) Collective dynamics of 'small-world' networks. Nature 393 (6684):440–442

Yip A, Horvath S (2007) Gene network interconnectedness and the generalized topological overlap measure. BMC Bioinform 8(8):22

Zhang B, Horvath S (2005) General framework for weighted gene coexpression analysis. Stat Appl Genet Mol Biol 4:17