Steve Horvath

Weighted Network Analysis

Applications in Genomics and Systems Biology

Weighted Network Analysis

Steve Horvath

Weighted Network Analysis

Applications in Genomics and Systems Biology

Springer

Steve Horvath
Professor of Human Genetics and Biostatistics
University of California, Los Angeles
Los Angeles, CA 90095-7088, USA
shorvath@mednet.ucla.edu

ISBN 978-1-4419-8818-8 e-ISBN 978-1-4419-8819-5
DOI 10.1007/978-1-4419-8819-5
Springer New York Dordrecht Heidelberg London

Library of Congress Control Number: 2011925163

To Lora, my brother Markus, my parents,
Joseph O'Brien and Joerg Zimmermann

Preface

The past decade has seen an incredible growth of network methods following publications by Laszlo Barabasi and others. Excellent text books exist on general networks and graph theory, but these books typically describe unweighted networks. This book focuses on weighted networks. In weighted networks, the pairwise connection strength between two nodes is quantified by a real number between 0 and 1. It is worth emphasizing that **most of the material also applies to unweighted networks**. Further, unweighted networks can easily be constructed from weighted networks by dichotomizing the connection strengths between nodes. While unweighted networks permit graph-theoretic visualization techniques and algorithms, weighted networks can be advantageous for many reasons including the following:

1. They preserve the continuous nature of the underlying connectivity information. For example, weighted correlation networks that are constructed on the basis of correlations between numeric variables do not require the choice of a hard threshold (Chap. 5). Dichotomizing information and (hard)-thresholding may lead to information loss.
2. They often lead to highly robust results (Zhang and Horvath 2005). In contrast, results based on unweighted networks, constructed by thresholding a pairwise association measure, often strongly depend on the threshold.
3. They can sometimes be decomposed and approximated by simpler networks. For example, networks can sometimes be approximated by "factorizable" networks (Chap. 2). Such approximations are often difficult to achieve for sparse, unweighted networks.
4. They sometimes allow for a parsimonious parametrization (in terms of modules and conformities, see Sect. 2.3).
5. They often allow one to derive simple relationships between network concepts (statistics) (Sect. 3.8 and Chap. 6). In particular, weighted *correlation* networks facilitate a geometric interpretation based on the angular interpretation of the correlation (Sect. 6.7).
6. They can be used to enhance standard data-mining methods such as cluster analysis since (dis)-similarity measures can often be transformed into weighted networks (Sect. 7.7).

Many of the applied sections in this book present analysis techniques and strategies to the wider audience of applied researchers. The book assumes little mathematical and statistical knowledge, but some sections are rather abstract. To make the book self-contained, some sections review statistical and data-mining techniques. Since several technical sections and chapters are less relevant to applied researchers, they start out with the warning that they can be skipped. I also present abstract, theoretical material since it may be useful for quantitative researchers, who carry out methodological research. In my own experience, I have found that applied researchers can be expert users of network methods and software. Of course, domain-knowledge experts have often a superior intuition about how to arrive at a meaningful analysis of their data. Many weighted network methods arose from collaborations with cancer biologists, neuroscientists, mouse geneticists, and biologists (e.g., see the acknowledgement section and references).

Although the field of weighted network analysis only began a few years ago, it is already impossible to summarize it in one book. I have tried to cite as many articles as possible, but I apologize to my colleagues for failing to cite their work. Several people are mentioned throughout the book, see the index for names and page numbers. I am acutely aware that I leave unmentioned many important ideas and techniques. My only excuse for giving too much attention to my own work is that I understand it best.

While the methods are formulated in general terms, which facilitate their application to wide variety of data, most applications involve genes, proteins, and gene expression data. It has become clear that networks have important medical and biological applications. Gene co-expression networks bridge the gap from individual genes to clinically important, emergent phenotypes. Gene networks allow one to move beyond single-gene comparisons and systematically identify biologically meaningful relationships between gene products, pathways, and phenotypes. Weighted gene co-expression network analysis (WGCNA) has been used to identify candidate disease biomarkers, to annotate genes with regard to module membership, to study the relationships between co-expression modules, and to compare the network topology of different networks. Case studies show how WGCNA can be used to screen for genes, to understand the transcriptional architecture, and to relate modules in different mouse tissues. Integrating co-expression networks with genetic marker data facilitates systems genetic applications (Sects. 11.5 and 12.3), which make use of causal testing and network edge-orienting procedures.

Freely Available R Software

This book provides an in-depth description of the `WGCNA` R package (Langfelder and Horvath 2008), which provides functions for carrying out network analysis tasks. R is a freely available, open source language and environment for statistical computing and graphics, which has become a de-facto standard in data analysis (Ihaka and Gentleman 1996; Venables and Ripley 2002; Gentleman et al. 2004,

2005; Carey et al. 2005). The R environment integrates standard data analysis and visualization techniques with packages (libraries) implementing the latest advances in data mining, statistics, and machine learning. The `WGCNA` package is available from the Comprehensive R Archive Network (CRAN), the standard repository for R add-on packages. To install it, type the following command into the R session:

```
install.packages("WGCNA")
```

Most of the R code and data presented in the book chapters can be downloaded from the following webpage:
www.genetics.ucla.edu/labs/horvath/CoexpressionNetwork/Book
Related scientific articles and presentations can be found at the following webpages:
www.genetics.ucla.edu/labs/horvath/CoexpressionNetwork/.

Other relevant R packages mentioned throughout the book are freely available on the R CRAN package resource at
www.R-project.org
and/or on the Bioconductor webpage.

Carey VJ, Gentry J, Whalen E, Gentleman R (2005) Network structures and algorithms in bioconductor. Bioinformatics 21(1):135–136

Ihaka R, Gentleman R (1996) R: A language for data analysis and graphics. J Comput Graph Stat 5(3):299–314

Gentleman RC, Carey VJ, Bates DJ, Bolstad BM, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth GK, Tierney L, Yang YH, Zhang J (2004) Bioconductor: Open software development for computational biology and bioinformatics. Genome Biol 5(10):R80

Gentleman R, Huber W, Carey V, Irizarry R, Dudoit S (2005) Bioinformatics and computational biology solutions using R and bioconductor. Springer, New York

Langfelder P, Horvath S (2008) WGCNA: An R package for weighted correlation network analysis. BMC Bioinform 9(1):559

Venables WN, Ripley BD (2002) Modern applied statistics with S, 4th edn. Springer, New York

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

Los Angeles, USA Steve Horvath

Acknowledgements

Many weighted network methods and R software code were developed in collaboration with colleagues, Postdoctoral researchers, and doctoral students. In particular, I mention **Peter Langfelder** who maintains the WGCNA R software package (Langfelder and Horvath 2008) and was the first author on several related publications (Langfelder and Horvath 2007; Langfelder et al. 2007, 2011). His contribution and those of others are mentioned throughout the book (see the index). **Bin Zhang** worked on the general framework for weighted gene coexpression network analysis (Zhang and Horvath 2005) and developed the first dynamic tree cutting algorithm Langfelder et al. (2007). **Jun Dong** worked on the relationships between network concepts (Dong and Horvath 2007; Horvath and Dong 2008).

Much of the work arose from close collaborations with applied researchers. In particular, weighted correlation networks were developed in joint discussions with cancer researchers **Paul Mischel** and **Stanley F. Nelson**, and neuroscientists **Daniel H. Geschwind** and **Michael C. Oldham** (Horvath et al. 2006; Carlson et al. 2006; Oldham et al. 2006, 2008; Miller et al. 2010). **Jake Lusis**, **Thomas Drake**, and **Eric Schadt** provided important mouse genetic applications and data (Ghazalpour et al. 2006; Chen et al. 2008). **Roel Ophoff, Giovanni Coppola**, and **Jeremy Miller** provided applications to neurological diseases (Saris et al. 2009; Miller et al. 2010). **Kenneth Lange** and **John Ranola** solved optimization problems. Many doctoral students have closely worked with me including the following: **Andy Yip** and **Ai Li** worked on extensions of the topological overlap matrix (Yip and Horvath 2007; Li and Horvath 2007). **Jason Aten** worked on causal anchors and the network edge orienting (NEO) (Aten et al. 2008). **Tova Fuller** worked on differential network analysis and causal analyses (Fuller et al. 2007). **Angela Presson** worked on integrating genetic markers with gene co-expression network analysis (Presson et al. 2008). **Chaochao Cai, Marc Carlson, Sudheer Doss, Charles Farber, Anatole Ghazalpour, Austin Hilliard, Wen Lin, Rui Luo, Michael Mason, Chris Plaisier, Lin Song, Yang Song, Atila Van Nas, Lin Wang, Kellen Winden, Wei Zhao, Yafeng Zhang, and Joerg Zimmermann** worked on WGCNA methods and applications (Ghazalpour et al. 2006; Carlson et al. 2006; Farber et al. 2009; van Nas et al. 2009; Mason et al. 2009).

Aten J, Fuller T, Lusis AJ, Horvath S (2008) Using genetic markers to orient the edges in quantitative trait networks: The NEO software. BMC Syst Biol 2(1):34

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

Chen Y, Zhu J, Lum PY, Yang X, Pinto S, MacNeil DJ, Zhang C, Lamb J, Edwards S, Sieberts SK, Leonardson A, Castellini LW, Wang S, Champy MFF, Zhang B, Emilsson V, Doss S, Ghazalpour A, Horvath S, Drake TA, Lusis AJ, Schadt EE (2008) Variations in DNA elucidate molecular networks that cause disease. Nature 452(7186):429–435

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

Farber CR, van Nas A, Ghazalpour A, Aten JE, Doss S, Sos B, Schadt EE, Ingram-Drake L, Davis RC, Horvath S, Smith DJ, Drake TA, Lusis AJ (2009) An integrative genetics approach to identify candidate genes regulating bone density: Combining linkage, gene expression and association. J Bone Miner Res 1:105–116

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

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

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

Langfelder P, Horvath S (2007) Eigengene networks for studying the relationships between co-expression modules. BMC Syst Biol 1(1):54

Langfelder P, Horvath S (2008) WGCNA: an R package for weighted correlation network analysis. BMC Bioinform 9(1):559

Langfelder P, Zhang B, Horvath S (2007) Defining clusters from a hierarchical cluster tree: The Dynamic Tree Cut library for R. Bioinformatics 24(5):719–720

Langfelder P, Luo R, Oldham MC, Horvath S (2011) Is my network module preserved and reproducible? Plos Comput Biol 7(1):e1001057

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

Mason M, Fan G, Plath K, Zhou Q, Horvath S (2009) Signed weighted gene co-expression network analysis of transcriptional regulation in murine embryonic stem cells. BMC Genomics 10(1):327

Miller JA, Horvath S, Geschwind DH (2010) Divergence of human and mouse brain transcriptome highlights Alzheimer disease pathways. Proc Natl Acad Sci USA 107(28):12698–12703

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

Presson AP, Sobel EM, Papp JC, Suarez CJ, Whistler T, Rajeevan MS, Vernon SD, Horvath S (2008) Integrated weighted gene co-expression network analysis with an application to chronic fatigue syndrome. BMC Syst Biol 2:95

Saris C, Horvath S, van Vught P, van Es M, Blauw H, Fuller TF, Langfelder P, DeYoung J, Wokke J, Veldink J, van den Berg L, Ophoff R (2009) Weighted gene co-expression network analysis of the peripheral blood from Amyotrophic Lateral Sclerosis patients. BMC Genomics 10(1):405+

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

Contents

Acronyms

AF	Adjacency function
ARACNE	Algorithm for the reconstruction of accurate cellular networks
CF	ConFormity
CPA	Common pleiotropic anchor
DPI	Data processing inequality
FDR	False discovery rate
GTOM	Generalized topological overlap measure
GS	Gene (or node) Significance
kME	Connectivity based on the module eigenvector
kIM	Connectivity, intramodular
KL	Kullback–Leibler
LEO	Local edge orienting
LRT	Likelihood ratio test
ME	Module eigenvector or eigengene
MI	Mutual information
MLE	Maximum likelihood estimation
MTOM	Multinode topological overlap measure
NEO	Network edge orienting
OCA	Orthogonal causal anchor
NCF	Network concept function
PAM	Partitioning around medoids
PPI	Protein–Protein interaction
QTL	Quantitative trait locus
SEM	Structural equation model
SFT	Scale-free topology
SNP	Single nucleotide polymorphism
SVD	Singular value decomposition
TOM	Topological overlap matrix
WGCNA	Weighted gene co-expression network analysis or weighted correlation network analysis

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).