A Positional Approach
for Network Centrality

Dissertation submitted for the degree of
Doctor of Natural Sciences

Presented by

David Schoch

at the

Universität
Konstanz

Faculty of Sciences

Department of Computer and Information Science

Date of the oral examination: 19.11.2015
1st referee: Prof. Dr. Ulrik Brandes
2nd referee: Prof. Dr. Nils Weidmann
3rd referee: Prof. Dr. David Krackhardt
4th referee: apl. Prof. Dr. Sven Kosub

Dedicated to the central people of my life

Deutsche Zusammenfassung

Die Netzwerkforschung ist ein aufstrebendes und sehr aktuelles Forschungsgebiet welches unterschiedliche Disziplinen wie Mathematik, Physik, und Soziologie miteinander verknüpft. Konzepte der Netzwerktheorie haben Eingang in die Analyse vieler Forschungsfragen gefunden sodass sich Methoden stetig weiterverbreiten und neue entwickelt werden. Trotz des hohen Bekanntheitsgrades ist das Erforschen von Netzwerken noch eine junge Disziplin sodass es noch an vielen grundlegenden Theorien fehlt. Dieser Mangel führt dazu, dass einige Konzepte nicht klar definiert sind und die Anzahl an existierenden Methoden zur Analyse von Netzwerken unüberschaubar zu sein scheint.

In dieser Arbeit wird konkret das Konzept der Zentralitätsmessung in Netzwerken behandelt. Bei der Analyse eines Netzwerks ist in vielen Fällen die (strukturelle) Wichtigkeit von Knoten oder Akteuren von Interesse. Neben naiven Ansätzen wie dem Knotengrad wurden im Laufe der Zeit viele komplexe Maße konzipiert, welche auf unterschiedlichen graphentheoretischen Modellen begründet sind. Ein bekanntes Beispiel ist der PageRank, welcher die Basis für Suchmaschinen bildet. Die Vielfalt stellt eine große Herausforderung dar, ein angemessenes Maß für spezifische Forschungsfragen auszuwählen. Die Auswahl begründet sich daher meistens auf dem 'Trial-and-Error-Prinzip', d.h. es werden verschiedene Maße getestet und verglichen um ein zufriedenstellendes Resultat zu erreichen. Obwohl in der Literatur bereits viel zur Abgrenzung des Konzepts beigetragen wurde, fehlt es weiterhin an theoretischen Grundlagen und Richtlinien für die Zentralitätsmessung.

Die vorliegende Arbeit beginnt mit der Einführung des Zentralitätsbegriffs sowie der Darlegung der genannten Mängel. Speziell wird auf einen Anwendungsfall in der Biologie eingegangen. In einer bekannten Studie über Protein Interaktionsnetzwerke wurde gezeigt, dass die Sterblichkeit von Proteinen mit Zentralitätsmaßen nachgewiesen werden kann. Was zunächst mit der Anwendung des Knotengrads begann, kulminierte in einer Jagd nach dem Maß welches am stärksten mit der Sterblichkeit von Proteine korelliert. Im Gegensatz zur Literatur wird in dieser Arbeit die Schlüssigkeit des proklamierten Zusammenhangs hinterfragt. Eine Reanalyse der ursprünglichen Studie deckt statistische Unstimmigkeiten auf, welche den beschriebene Zusammenhang stärker erscheinen lassen als tatsächlich beobachtet werden konnte. Eine breiter angelegte Studie mit verschiedenen Datensätzen zeigt, dass der Zusammenhang zwischen Zentralitätsmaßen und Sterblichkeit stark Daten abhängig ist und generell schwankende Resultate aufweist.

Im weiteren Verlauf wird eine neue Konzeptualisierung der Zentralitätsmessung vorgeschlagen. Es wird argumentiert, dass Zentralität nicht als Auffinden von Mustern in Daten durch maschinelles Lernen, sondern als Konzept der Messtheorie behandelt werden sollte. Eine zentrale Rolle spielt dabei der Begriff der Dominanz in Netzwerken. Ein Knoten der zu den selben Knoten und möglicherweise zusätzlichen benachbart ist als ein anderer, dominiert diesen Knoten. Es wird gezeigt das die geläufigsten Zentralitätsmaße diese sogenan-

nte Nachbarschaftsinklusionsordnung erhalten, d.h. der dominierende Knoten immer mindestens den gleichen Rang in der induzierten Rangfolge hat. Im Zuge der Beweisführung werden einige relevante Schritte der Zentralitätsmessung vereinheitlicht. Die Herleitung indirekter Beziehungen, welche auf graphentheoretischen Modellen beruhen, wird anhand von algebraischen Strukturen, den Halbringen, geführt. Diese Vereinheitlichung ermöglicht es, Bedingungen aufzustellen welche indirekte Beziehungen erfüllen müssen um die Nachbarschaftsinklusionsordnung zu erhalten. Des Weiteren wird ein neuer allgemeiner Ansatz zur Netzwerk Analyse, der Positionsansatz, angewandt um Zentralität als Positionsvergleich von Akteuren zu charakterisieren.
Der restliche Verlauf ist den Auswirkungen dieser Rekonzeptualisierung gewidmet. Eine wichtige Rolle spielt dabei die Klasse der eindeutig geordneten Graphen. Ist die Nachbarschaftsinklusionsordnung vollständig, induzieren alle Zentralitätsmaße die selbe Rangfolge und würden daher in empirischen Situationen widersprüchliche Erklärungen liefern. Im Umkehrschluss bedeutet dies, je weiter entfernt ein Netzwerk von einer vollständigen Ordnung ist, desto unterschiedlicher können die Rangfolgen von Zentralitätsmaßen sein. Diese Beobachtung wird verwendet um die Korrelation zwischen unterschiedlichen Zentralitätsmaßen zu untersuchen. Im Gegensatz zum allgemeinen Konsens in der Literatur, dass Korrelation abhängig von der Definition der Maße ist, wird gezeigt, dass die Korrelation abhängig vom Abstand zum nächstliegenden vollständig geordneten Graphen ist. Je geringer der Abstand, desto höher die Korrelation.
Den Abschluss der Arbeit bildet eine Diskussion über die Anwendbarkeit des neuen Zentralitätskonzepts in empirischen Studien und der Verallgemeinerung der Dominanz in Netzwerken. Anhand kleiner synthetischer Netzwerke werden neue Formen der Dominanz erarbeitet und das Beispiel aus der Biologie wird wieder aufgegriffen um Indizien gegen den beschriebenen Zusammenhang zu sammeln.

Table of Contents

List of Figures iii

List of Tables v

1 Introduction 1

2 Preliminaries 5

I Network Centrality based on Indices 9

3 The Concept of Network Centrality 11
3.1 Centrality Indices 11
3.2 Conceptualization of Centrality Indices 21
3.3 Illustrative and Motivational Examples 29
3.4 Discussion 31

4 Centrality in Empirical Research 33
4.1 Early Developments and Applications 33
4.2 Recent Developments and Applications 36
4.3 Discussion 37

5 Centrality in Protein Interaction Networks 41
5.1 Introduction 42
5.2 Material and Methods 43
5.3 Re-Analysis of Original Study 48
5.4 Results for S.cerevisiae 56
5.5 Results for Multiple Organisms 58
5.6 Summary 61

II Network Centrality based on Positions 63

6 Centrality, Measurement and Positions 65

6.1 Introduction 65
6.2 Measurement Theory 65
6.3 Characterizing Centrality via Positions 68

7 Re-Conceptualizing Network Centrality 77
7.1 Formal Definitions and Terminology 77
7.2 Semirings for Indirect Relations 81
7.3 Neighborhood-Inclusion and Indirect Relations 85
7.4 Centrality Indices and Neighborhood-Inclusion 89
7.5 Summary and Discussion 91

8 Uniquely Ranked Graphs 93
8.1 Definitions and Properties of Threshold Graphs 93
8.2 Threshold Distance Measures 95

9 Correlation and Threshold Distance 99
9.1 Introduction 99
9.2 Material and Methods 100
9.3 Results for Random Graph Models 104
9.4 Results for Real Networks 106
9.5 Summary 110

10 Generalization and Application of the Dominance Concept 111
10.1 Further Notions of Dominance 111
10.2 The Scope of Rankings and its Implications 116
10.3 Centrality-Lethality Hypothesis revisited 118
10.4 Summary and Discussion 119

11 Conclusion 121

Bibliography 125

List of Figures

2.1 Examples for induced subgraphs 6

3.1 Periodic table of centrality indices 12
3.2 Radial and medial walk positions 25
3.3 Weights of eigenvectors for three indices 27
3.4 Krackhardt's Kite and graph with 4 centers 30
3.5 Minimal graphs with different centers 31
3.6 Graph with extreme discordant rankings 31
3.7 Graph with unique ranking 32

4.1 Redrawn from Moreno's group structures 34
4.2 Redrawn from Leavitt's communication patterns 34
4.3 Redrawn from russian trading routes 35
4.4 Redrawn from Krackhardt's hierarchy and advice seeking network 35

5.1 PPI of *S. cerevisiae* 41
5.2 Overlapping proteins in PINs of *S. cerevisiae* 44
5.3 Example for separation plots 47
5.4 Degree distribution of PIN of *S. cerevisiae* 48
5.5 Attack tolerance of *S. cerevisiae* 50
5.6 Correlation and MCC of *S. cerevisiae* 51
5.7 Less equal, greater than test of *S. cerevisiae* 52
5.8 Degree distributions of lethal and non-lethal proteins 53
5.9 Alternative degree distributions of PIN of *S. cerevisiae* 53
5.10 ROC and PR performance of degree 55
5.11 Rank sized degree distr. for instances of *S. cerevisiae* 56
5.12 Attack tolerance test for instances of *S. cerevisiae* 57
5.13 Barplot of AUC values on PINs of *S. cerevisiae* 57
5.14 Separation plots for *S. cerevisiae* 58
5.15 Evaluation on 20 organisms 59
5.16 AUC organisms 60

6.1 Example for positions in networks. 71

6.2 Example for dominance in networks. 72
6.3 Example for dominance under total homogeneity in networks. . . . 74

7.1 Example for neighborhood-inclusion 78

8.1 Examples for threshold graphs 94
8.2 Illustration of $t_1(G)$ and $t_e(G)$ 97
8.3 Illustration of the ambiguity of t_m 98

9.1 Correlation vs Majorization Gap on ER 104
9.2 Contradictory results of correlation coefficients. 105
9.3 Correlation vs Majorization Gap on BA 106
9.4 Correlation vs Majorization Gap on PINs 107
9.5 Correlation vs spectral gap on PINs 107
9.6 Illustrating numerical issues of centrality scores 108
9.7 Correlation vs Majorization Gap on Valente's dataset 109
9.8 Illustrating numerical issues of scores 109

10.1 Threshold Graphs with five vertices 112
10.2 Totally ordered graphs with five vertices 112
10.3 Example graph for automorphic dominance 112
10.4 Graphs with five vertices and newly derived forms of dominance . 114
10.5 Graphs with five vertices where discordance may occur 114
10.6 Rankings of discordant graphs with five nodes 115
10.7 Example graphs with nine vertices 117

List of Tables

2.1 Properties of binary relations 8

3.1 Categorized measures of betweenness 16
3.2 Axiomatic systems of the literature 23
3.3 Flow processes and centrality indices 24
3.4 Walk involvement of centrality indices 25

5.1 Network statistics for PINs of *S. cerevisiae* 43
5.2 Network statistics for PINs of 20 organisms 45
5.3 Structure of a contingency table 46
5.4 Results for logistic regression for degree and lethality 54
5.5 Statistical test for power law in PINs of *Saccharomyces cerevisiae* . . . 56
5.6 Mean correlation of indices on PINs of *S. cerevisiae* 58

6.1 Decision making Example 67
6.2 Comparing positions with additivity and homogeneity 74

7.1 One dimensional semirings for indirect relations 82

9.1 Summary statistics for protein dataset 101
9.2 Summary statistics for Valente's dataset 101

10.1 Statistics for neighborhood-inclusion in eight PINs of yeast. 119

CHAPTER 1

Introduction

"The role of mathematics in empirical science is puzzling, mysterious, and in my opinion has defied rational explanation."

– Narens, 1981

One of the founding fathers of *sociology*, Auguste Comte, envisioned it to be the science unifying all scientific disciplines [49]. Although this ambitious aim was never achieved, some of his ideas lived on and helped forming sociology as is. One of his main ideas was to divide the study of social systems into the study of *social statics* and *social dynamics*. In the words of Comte,

> *"The statical study of sociology consists in the investigation of the laws of action and reaction of the different parts of the social system."* [49, p. 457]

When dealing with the components of a social system, Comte refused to see individuals as atomic elements, since "the scientific spirit forbids us to regard society as composes of individuals." [49, p. 502]. In his view, social systems can only be studied by looking at society in terms of relations among individuals. He thus offered a complementary view on social sciences to the classical atomistic approach of Thomas Hobbes [99]. Early sociologists adopted Comte's *structural perspective* to study the patterning of social connections trying to specify different kinds of social ties in different social systems [76].

The most explicit use of a structural perspective can be found in the work of Georg Simmel, who argues that "society exists where a number of individuals enter into interaction." [176, p. 23]. The structural perspective in sociology prevailed and with Moreno's development of *sociometry* [139], the groundwork was laid for a new subfield of sociology, called *social network analysis* [76].

Contrasted to an atomic view of entities in a social system, the *network perspective* draws attention to the dyadic domain, the relations among entities or actors. More importantly, these relations are not disjoint but intertwined and most certainly dependent on each other. A simple example is given by a set of individuals, the friendship relations among them and the associated phenomenon of 'the friend of my friend is also my friend'. Network analysts seek to uncover the patterning of ties (i.e. relations) in which actors are embedded and try to explain why those patterns arise and interpret their consequences. Networks offer a new perspective to tackle complex problems beyond the scope of traditional methodology such that social network analysis has now advanced from a subfield of sociology to *network science*, a maturing field of its own [40]. Since networks arise in many different contexts ('Networks are everywhere'), the network paradigm has become scientifically relevant across disciplinary boundaries. This inherent interdisciplinary and the prospect of significant advances are the compelling forces that draw more and more scientific areas towards network science.

Network scientific studies with promising results exist in many areas. Individuals are connected through at most five acquaintances, known as the *six degrees of separations* or the *small-world effect* [137,196], obesity spreads via direct and indirect social ties [43], your friends have more friends than you do [70] accompanied with *preferential attachment* or 'the rich get richer' phenomenon of networks [9], and the *strength of weak ties* [90] to name a few studies that made their way to popular science. Networks, as it seems, are taking over the scientific landscape.

One of the key concepts in network science is *network centrality*. Centrality seeks to provide the answer to the question of who (or what) is important in a network depending on the underlying process forming the network and the empirical phenomenon in question. In a nutshell, an actor in a network is more central if he or she has better relations, where the definition of *better relations* depends on the conceptualization of *structural importance*.

Applications of centrality range from simple problems like 'who is the most popular individual?' in a friendship network to more complex tasks as 'which infected individual should be targeted to prevent the spread of a disease most efficiently?'. Early applications of centrality date back to the work of Moreno in the 1930s and although many researchers have contributed since then to a better understanding of centrality, its theoretical foundations mostly remain nebulous. Network analysts have differing interpretations of what constitutes a central position in a network and the number of methods to determine structural importance has drastically increased. The vast amount of methods poses various difficulties in empirical research, such as identifying a suitable approach, which most often culminates in trial-and-error efforts probing different methods until a satisfactory result is obtained.

In this thesis, we aim for a re-conceptualization of network centrality with a more solid theoretical basis which is additionally more accessible in empirical research.

The first part of this thesis deals with the current conception of centrality based on real valued functions. In Chapter 3, commonly used centrality indices are presented and several attempts to conceptualize network centrality are discussed. We also conjecture about a novel framework for centrality indices by means of spectral decompositions. We end the chapter by some illustrative examples to motivate later theoretical considerations. Chapter 4 is then dedicated to applications of centrality in empirical research. Working through historical advances, we discuss potential weak points of the current state which are examined extensively in Chapter 5 with a well known application of centrality in systems biology.

The second part comprises the main contributions of this thesis. We discuss centrality in the context of the theory of measurement in Chapter 6 and motivate an alternative approach based on positions in networks. The considerations in Chapter 6 lead to a re-conceptualization of centrality by a form of dominance in networks. This approach is formally introduced in Chapter 7. In the course, we introduce a general framework to derive indirect relations in graphs with the algebraic structure of semirings. In Chapter 8, we present a class of graphs which are uniquely ordered, i.e. all centrality indices induce the same ranking of nodes. This class is used to explain correlations among centrality indices in Chapter 9. Chapter 10 is devoted to applications of our theoretical results. We derive further forms of dominance in graphs and illustrate how the concepts of dominance can be used in empirical work. We investigate a set of small graphs and also reconsider the application of centrality in systems biology with newly developed methods.

In Chapter 11 we give final concluding remarks and discuss the provided contributions.

Parts of the thesis have already been published or presented at various conferences.

- Results of Section 5.3 are submitted [171].
- Results of Section 5.5 are published in [170].
- Results of Chapter 7 are submitted [173].
- Partial Results of Chapters 7 to 9 are published in [172].
- Results of Chapter 9 have been presented at the Sunbelt Conference 2014.

Chapter 2

Preliminaries

Mathematical Conventions

We use common notations for the set of natural and real numbers, where $\mathbb{N} = \{1, 2, \ldots\}$ and $\mathbb{N}_0 = \{0, 1, 2, \ldots\}$. The set of all positive real numbers including zero is denoted as $\mathbb{R}_0^+$.

Matrices are generally denoted with capital letters and an entry (i, j) of a matrix A is addressed by A_{ij}. The *identity matrix* I is defined as

$$I_{uv} = \begin{cases} 1 & u = v \\ 0 & \text{otherwise.} \end{cases}$$

For general diagonal matrices $D \in \mathbb{R}^{n,n}$ we use the shorthand notation

$$diag(D_{11}, D_{22}, \ldots, D_{nn}).$$

Vectors are expressed in bold lowercase letters where the ith entry of a vector $\boldsymbol{x}$ is either addressed with x_i or $\boldsymbol{x}(i)$. The all ones vector of length n is denoted as $\mathbf{1}_n$. We define $J = \mathbf{1}_n \mathbf{1}_n^T$ to be the all ones matrix.

Graph Theory

Throughout the thesis we will make use of standard graph-theoretic notations [27,195].

Graphs. A graph is an ordered tuple $G = (V(G), E(G))$ consisting of a set of *nodes* or *vertices* $V(G) = \{v_1, v_2, \ldots, v_n\}$ with $n = |V|$ and a set $E(G) = \{e_1, e_2, \ldots, e_m\}$ with $m = |E|$ of *edges* or *ties*. In empirical settings, we also refer to nodes as *actors* and *entities* and edges as *relations* or *ties* among actors. Vertices are denoted in subscript notation v_i where $i \in \{1, \ldots, n\}$ if they are treated as sequences. Otherwise we use i, j, u, v, s and t to denote specific vertices. Edges are commonly referred to in set notation providing the participating vertices. If the graph is clear from the context, we use V and E for simplicity.

A graph is called *undirected* if the edge set consists of unordered pairs, i.e.

$$E = \{\{u, v\} : u, v \in V\} \subseteq \binom{V}{2}.$$

If $\{u, u\} \notin E$ for all $u \in V$ and multiple edges between two vertices are forbidden, we speak of a *simple* graph. The *density* of an undirected simple graph is given by

$$\rho(G) = \frac{m}{\binom{n}{2}} \, .$$

Other notions of graphs stem from different conceptualizations of the edge set. A graph is *directed* if edges are ordered tuples, i.e.

$$E = \{(u, v) : u, v \in V\} \subseteq V \times V.$$

A graph is weighted if there is a function $\omega : E \to \mathbb{R}$ assigning weights to the edge set. The tuple $G = (V, E, \omega)$ then describes a *weighted* graph.

In the main part of this thesis, we will only deal with simple undirected and unweighted graphs which we simply refer to as graphs for brevity.

Two vertices u and v are said to be *adjacent* if $\{u, v\} \in E$. The *neighborhood* of a vertex u is the set of all adjacent vertices, i.e. $N(u) = \{v : \{u, v\} \in E\}$ and $N[u] = N(u) \cup \{u\}$ its closed neighborhood. The *degree* d_u of a vertex u is defined as the cardinality of its neighborhood. The *degree sequence* of a graph is then defined as $d(G) = [d_1, d_2, \ldots, d_n]$. Henceforth, we assume that this sequence is ordered non-increasingly, i.e. $d_1 \geq \cdots \geq d_n$.

GRAPH STRUCTURES. A *subgraph* $H = (V(H), E(H))$ of $G = (V(G), E(G))$ is a graph where $V(H) \subseteq V(G)$ and $E(H) \subseteq E(G)$. A subgraph H is called *induced* if for all $u, v \in V(H)$ it holds that $\{u, v\} \in E(H) \iff \{u, v\} \in E(G)$.

An induced subgraph H is a *clique* if it is complete, i.e. $\{u, v\} \in E(H)$ for all $u, v \in V(H)$. Cliques with n vertices are denoted by K_n. An *independent set* $I \subseteq V(G)$ induces a subgraph $H = (I, E(H))$, where $E(H) = \emptyset$.

A graph is called a *split graph* if its vertex set V can be partitioned into $V = C \cup I$, such that C induces a clique and I induces an independent set.

A graph is called *bipartite,* if its node set can be divided into two disjoint independent sets V_1 and V_2, i.e. $V = V_1 \cup V_2$ and edges only connect vertices in V_1 to vertices in V_2. Other notable (sub)graph structures and their denotation used in this thesis are shown in Figure 2.1.

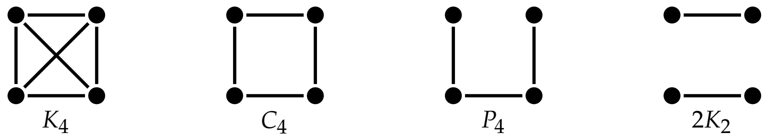


FIGURE 2.1: Examples for simple graph structures.

MATRICES. The connectivity of a graph can be represented by an *adjacency matrix A*, where

$$A_{uv} = \begin{cases} 1 & \{u,v\} \in E \\ 0 & \text{otherwise.} \end{cases}$$

The adjacency matrix of a graph G is symmetric and irreducible if G is undirected and connected and therefore has real eigenvalues with a set of orthonormal eigenvectors. The *spectral decomposition* of A is given by

$$A = X \Lambda X^T \ ,$$

where $\Lambda = \text{diag}(\lambda_1,\ \lambda_2, \dots \lambda_n)$ is the diagonal matrix of eigenvalues with $\lambda_1 > \lambda_2 \geq \ldots \geq \lambda_n$, and $X = [\boldsymbol{x}_1\ \boldsymbol{x}_2\ \ldots \boldsymbol{x}_n]$ are the corresponding eigenvectors.

The *spectral gap* is commonly defined as the difference between the principal and the second largest eigenvalue, i.e. $|\lambda_1 - \lambda_2|$. In the course of this thesis we make use of the fraction λ_2/λ_1 to limit the spectral gap on the interval $[0,1]$. The *fundamental weight* w_i of an eigenvector $\boldsymbol{x}_i$ is defined as

$$w_i = \sum_{j=1}^{n} \boldsymbol{x}_i(j)$$

and $\boldsymbol{w}$ is the vector of all fundamental weights [189].

Another matrix associated with graphs is given by the *Laplacian matrix L*. Its entries are defined as

$$L_{uv} = \begin{cases} d_u & u = v \\ -1 & \{u,v\} \in E \\ 0 & \text{otherwise.} \end{cases}$$

TRAJECTORIES AND DISTANCES. A *walk* of length $k \in \mathbb{N}$ in a graph G is defined as an alternating sequence

$$v_0,\ \{v_0,v_1\}\ ,\ v_1,\ \{v_1,v_2\}\ , \ldots,\ v_{k-1},\ \{v_{k-1},v_k\}\ ,\ v_k$$

of $k+1$ nodes and k edges. A walk is called a *trail* if all edges are pairwise distinct. A walk with $v_0 = s$ and $v_k = t$ is called (s,t)-walk. A walk is *closed* if $v_0 = v_k$.

The number of (s,t)-walks of certain lengths can be calculated by powers of the adjacency matrix, i.e. the entry $[A^k]_{st}$ gives the number of (s,t)-walks of length k [87]. Any (s,t)-walk can also be thought of as a *random walk*. Starting at vertex s, one of its neighbors is chosen uniformly at random and the random process is continued with the chosen vertex until the destination t is reached.

A *path* is a walk where all vertices are pairwise distinct. A path that connects s and t with a minimal number of edges is called a *shortest path* or *geodesic* and the length of a shortest (s,t)-path is defined as the *distance* between s and t, denoted by $dist(s,t)$.

Two (s,t)-paths are *vertex disjoint* if the set of intermediary vertices does not overlap. Similarly, two (s,t)-paths are *edge disjoint* of the edge sets of the paths do not overlap.

A graph is *connected* if there exists a path between all pairs of vertices. Otherwise, the graph is disconnected and is composed of several *connected components*.

Order Relations

A *binary relation* $\succcurlyeq$ on a set $\mathcal{N}$ is defined as a subset of the Cartesian product of $\mathcal{N}$, i.e. $\succcurlyeq \subseteq \mathcal{N} \times \mathcal{N}$. If $(a, b) \in \succcurlyeq$ we say that a and b are related by $\succcurlyeq$. Instead of $(a, b) \in \succcurlyeq$ and $(a, b) \notin \succcurlyeq$ we commonly use infix notation $a \succcurlyeq b$ and $a \not\succcurlyeq b$. Two elements $a, b \in \mathcal{N}$ are called *comparable* if $a \succcurlyeq b$ or $b \succcurlyeq a$ (or both) holds. Otherwise they are *incomparable*. Some important properties of a binary relation $\succcurlyeq$ over a set $\mathcal{N}$ are given in Table 2.1.

property	definition
complete	all pairs comparable
reflexive	$a \succcurlyeq a$ holds for all $a \in \mathcal{N}$
irreflexive	$a \succcurlyeq a$ holds for no $a \in \mathcal{N}$
symmetric	$a \succcurlyeq b \iff b \succcurlyeq a$ for all $a, b \in \mathcal{N}$
anti-symmetric	$a \succcurlyeq b \implies b \not\succcurlyeq a$ for all $a, b \in \mathcal{N}$
transitive	$a \succcurlyeq b \ \wedge\ b \succcurlyeq c \implies a \succcurlyeq c$ for all $a,\ b,\ c \in \mathcal{N}$

Table 2.1: Properties of binary relations.

A binary relation is called a *preorder* if it is reflexive and transitive. A preorder is also referred to as a *partial ranking*. A complete preorder is called a *weak order* or *ranking*.

An *equivalence* relation is a symmetric preorder. We usually denote an equivalence relation by $\sim$. The *equivalence class* of an element $a \in \mathcal{N}$ is defined as $[a] = \{b \in \mathcal{N}\ :\ a \sim b\}$. The set of all equivalence classes in $\mathcal{N}$ w.r.t. $\sim$ is denoted by $\mathcal{N}/_{\sim}$. Two important equivalence relations on graphs are described in the following.

Structural Equivalence. Two vertices $u,\ v \in V$ are *structurally equivalent*, if their neighborhoods are identical, i.e. $N(u) = N(v)$, denoted by $u \sim v$ [30]. A similar form of equivalence is, e.g. given by $N[u] = N[v]$.

Automorphic Equivalence. A more general notion of equivalence is given by *automorphic equivalence*. Two vertices $u,\ v \in V$ are automorphic equivalent if there exists an automorphism $\pi : V \to V$ with $\pi(u) = v$, denoted by $u \sim_\pi v$. It holds that structural equivalence implies automorphic equivalence.

The notation $\succcurlyeq$ is reserved for a special binary relation among vertices of a graph. Commonly we denote a binary relation with $\geq$. Its definition is either given explicitly, or it is obvious from the context.

Part I

Network Centrality

based on Indices

CHAPTER 3

The Concept of Network Centrality

> *"There is certainly no unanimity on exactly what centrality is or its conceptual foundations, and there is very little agreement on the proper procedure for its measurement."*
>
> – Freeman,1979

3.1 CENTRALITY INDICES

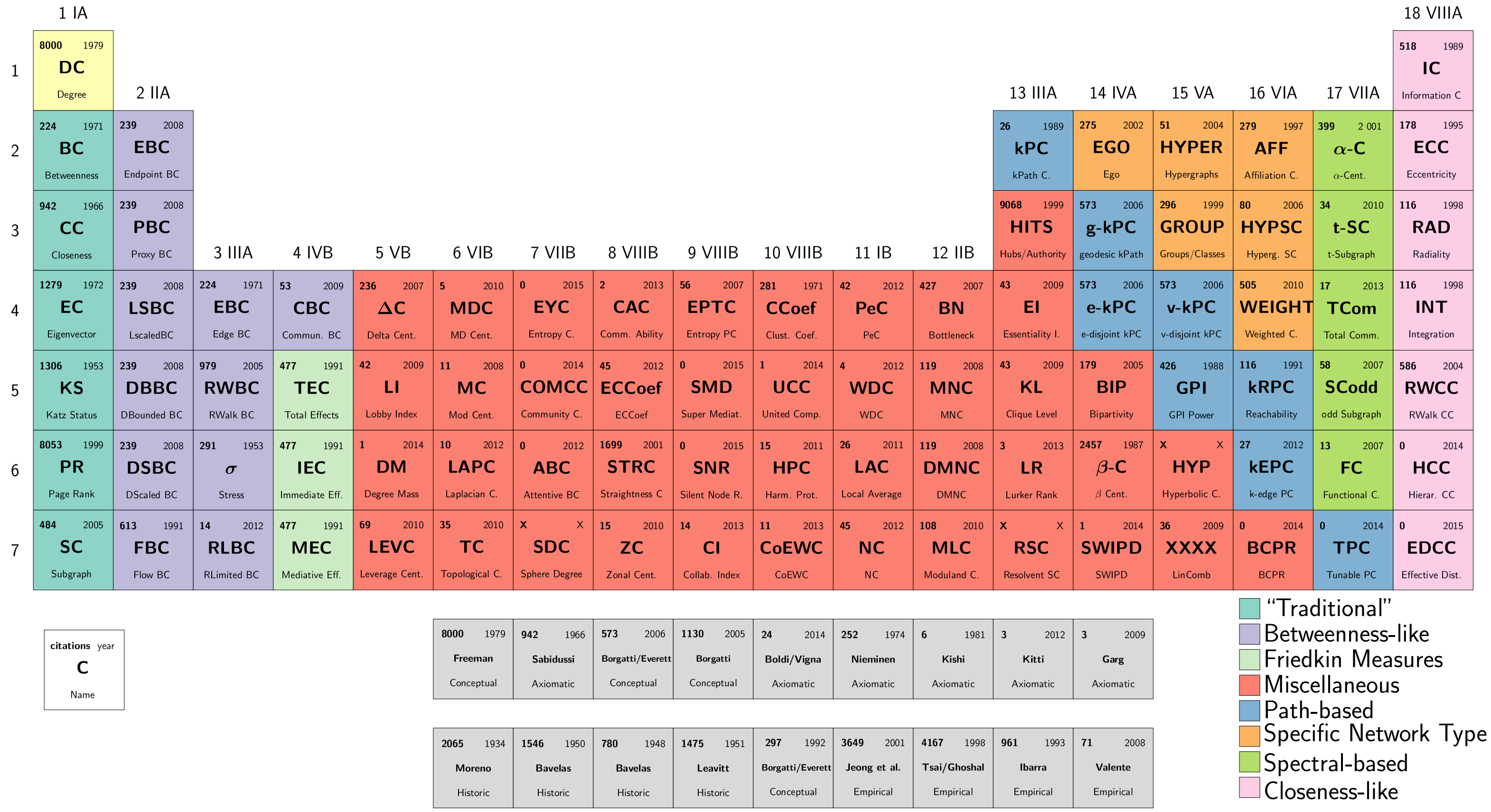
The purpose of network centrality is to identify important actors or a general importance ranking in a network. Importance by means of network structures gives rise to the term *structural importance*, contrasted to a perceived *individual importance*. Structural importance arises from network topological properties alone, whereas individual importance can potentially be any external attribute of actors comprising a network. Structural importance is determined by so called *measures of centrality*, commonly defined in terms of indices $c : V \to \mathbb{R}$ interpreted as

$$c(u) > c(v) \iff u \text{ is more central than } v \, .$$

Throughout this thesis, we use measure of centrality, centrality measure and centrality index interchangeably for mappings $c : V \to \mathbb{R}$ which determine structural importance.

Since the meaning of structural importance is by no means unambiguous, a vast amount of different indices exist (cf. Figure 3.1). In addition, any mapping $c : V \to \mathbb{R}$ induces a ranking of the vertices, but not every such ranking might represent a plausible concept of structural importance. Several possibilities to narrow down the number of feasible indices are presented in Section 3.2.

In the following, we introduce some of the standard, or most commonly used measures of the literature, their variants and formal connections among them. At this point, we restrict ourselves to the graph-theoretic notations. The intuition behind indices and their general applicability are discussed in more detail in Section 3.2. Additionally, we only present the definitions for undirected graphs. However, most of the indices can be transferred to directed



	1 IA	2 IIA	3 IIIA	4 IVB	5 VB	6 VIB	7 VIIB	8 VIIIB	9 VIIIB	10 VIIIB	11 IB	12 IIB	13 IIIA	14 IVA	15 VA	16 VIA	17 VIIA	18 VIIIA
1	8000 1979 DC Degree																	518 1989 IC Information C
2	224 1971 BC Betweenness	239 2008 EBC Endpoint BC											26 1989 kPC kPath C.	275 2002 EGO Ego	51 2004 HYPER Hypergraphs	279 1997 AFF Affiliation C.	399 2 001 α-C α-Cent.	178 1995 ECC Eccentricity
3	942 1966 CC Closeness	239 2008 PBC Proxy BC											9068 1999 HITS Hubs/Authority	573 2006 g-kPC geodesic kPath	296 1999 GROUP Groups/Classes	80 2006 HYPSC Hyperg. SC	34 2010 t-SC t-Subgraph	116 1998 RAD Radiality
4	1279 1972 EC Eigenvector	239 2008 LSBC LscaledBC	224 1971 EBC Edge BC	53 2009 CBC Commun. BC	236 2007 ΔC Delta Cent.	5 2010 MDC MD Cent.	0 2015 EYC Entropy C.	2 2013 CAC Comm. Ability	56 2007 EPTC Entropy PC	281 1971 CCoef Clust. Coef.	42 2012 PeC PeC	427 2007 BN Bottleneck	43 2009 EI Essentiality I.	573 2006 e-kPC e-disjoint kPC	573 2006 v-kPC v-disjoint kPC	505 2010 WEIGHT Weighted C.	17 2013 TCom Total Comm.	116 1998 INT Integration
5	1306 1953 KS Katz Status	239 2008 DBBC DBounded BC	979 2005 RWBC RWalk BC	477 1991 TEC Total Effects	42 2009 LI Lobby Index	11 2008 MC Mod Cent.	0 2014 COMCC Community C.	45 2012 ECCoef ECCoef	0 2015 SMD Super Mediat.	1 2014 UCC United Comp.	4 2012 WDC WDC	119 2008 MNC MNC	43 2009 KL Clique Level	179 2005 BIP Bipartivity	426 1988 GPI GPI Power	116 1991 kRPC Reachability	58 2007 SCodd odd Subgraph	586 2004 RWCC RWalk CC
6	8053 1999 PR Page Rank	239 2008 DSBC DScaled BC	291 1953 σ Stress	477 1991 IEC Immediate Eff.	1 2014 DM Degree Mass	10 2012 LAPC Laplacian C.	0 2012 ABC Attentive BC	1699 2001 STRC Straightness C	0 2015 SNR Silent Node R.	15 2011 HPC Harm. Prot.	26 2011 LAC Local Average	119 2008 DMNC DMNC	3 2013 LR Lurker Rank	2457 1987 β-C β Cent.	X X HYP Hyperbolic C.	27 2012 kEPC k-edge PC	13 2007 FC Functional C.	0 2014 HCC Hierar. CC
7	484 2005 SC Subgraph	613 1991 FBC Flow BC	14 2012 RLBC RLimited BC	477 1991 MEC Mediative Eff.	69 2010 LEVC Leverage Cent.	35 2010 TC Topological C.	X X SDC Sphere Degree	15 2010 ZC Zonal Cent.	14 2013 CI Collab. Index	11 2013 CoEWC CoEWC	45 2012 NC NC	108 2010 MLC Moduland C.	X X RSC Resolvent SC	1 2014 SWIPD SWIPD	36 2009 XXXX LinComb	0 2014 BCPR BCPR	0 2014 TPC Tunable PC	0 2015 EDCC Effective Dist.

citations year C Name

8000 1979 Freeman Conceptual	942 1966 Sabidussi Axiomatic	573 2006 Borgatti/Everett Conceptual	1130 2005 Borgatti Conceptual	24 2014 Boldi/Vigna Axiomatic	252 1974 Nieminen Axiomatic	6 1981 Kishi Axiomatic	3 2012 Kitti Axiomatic	3 2009 Garg Axiomatic

2065 1934 Moreno Historic	1546 1950 Bavelas Historic	780 1948 Bavelas Historic	1475 1951 Leavitt Historic	297 1992 Borgatti/Everett Conceptual	3649 2001 Jeong et al. Empirical	4167 1998 Tsai/Ghoshal Empirical	961 1993 Ibarra Empirical	71 2008 Valente Empirical

- 🟩 "Traditional"
- 🟪 Betweenness-like
- 🟩 Friedkin Measures
- 🟥 Miscellaneous
- 🟦 Path-based
- 🟧 Specific Network Type
- 🟩 Spectral-based
- 🟪 Closeness-like

FIGURE 3.1: Stylized table of centrality indices from the literature.

networks and other graph classes like *ego networks* [132], *hypergraphs* [26, 67], *affiliation networks* [69], and weighted graphs [152] or to quantify centrality of edges [110]. Further, there exist notions of *group centrality* which quantifies the centrality of subsets of nodes [68]. We also do not discuss methods to standardize or normalize scores as well as the concept of graph *centralization* [75, 195]. For details of any of the mentioned concepts, please refer to the corresponding literature.

REMARK. *Although centrality is defined by real valued mappings, we effectively only deal with the induced rankings in the majority of applications and the actual scores play a secondary role. The scale of measurement is discussed in depth in the second part of the thesis. For now it suffices to note that we can alter certain definitions of indices, e.g. by dropping scaling factors or constants, without altering the induced rankings.*

DEGREE CENTRALITY is the most simple form of a centrality index. It is defined as

$$c_d(u) = d_u \; .$$

Degree centrality is a purely *local* measure since it only depends on the direct neighborhood of a vertex. A simple application example is popularity in friendship networks, i.e. 'who has the most friends?'. It is among the only indices that can be directly applied, since it does not require any form of transformation, e.g. calculating shortest paths.

BETWEENNESS CENTRALITY was introduced by Freeman [74] and Anthonisse [6], based on Shimbel's *stress centrality* [175]. Shimbel assumes that the number of shortest paths containing a node u is an estimate for the amount of 'stress' the node has to sustain in a network. In this sense, the more shortest paths run through a vertex the more central it is. Formally, stress centrality is defined as

$$c_{stress}(u) = \sum_{s,t \in V} \sigma(s,t|u) \; ,$$

where $\sigma(s,t|u)$ is the number of shortest paths from s to t passing through u. By convention, we set $\sigma(s,t|u) = 0$ if $u \in \{s,t\}$. Instead of the absolute number of shortest paths, betweenness centrality quantifies the relative number of shortest paths passing through a vertex u. This relative number is given by

$$\delta(s,t|u) = \frac{\sigma(s,t|u)}{\sigma(s,t)} \; ,$$

where $\sigma(s,t)$ is the total number of shortest paths connecting s and t. If $s = t$ we set $\sigma(s,t) = 1$. The expression $\delta(s,t|u)$ can be interpreted as the extent to which u controls the communication between s and t and is also referred to as *shortest path dependency* of s and t on u [34]. Betweenness can thus be defined as

$$c_b(u) = \sum_{s,t \in V} \delta(s,t|u) \; .$$

We can further break down the definition by defining a *dyadic dependency* of a *sender s* on a *broker u* as

$$\delta(s|u) = \sum_{t \in V} \delta(s,t|u) \; . \tag{3.1}$$

The betweenness of a broker u is then given as the sum over all possible senders s, i.e.

$$c_b(u) = \sum_{s \in V} \delta(s|u) \; .$$

The interpretation of betweenness is not only restricted to communication. More generally, betweenness quantifies the influence of vertices on the transfer of items or information through the network with the assumption that it follows shortest paths. Many different variants of shortest path betweenness have been proposed to incorporate additional assumptions, e.g. the specific location of a vertex u on a shortest (s,t)-path or its length. Some of these variants are given in the following.

$$\begin{aligned}
&(i) \text{ proximal source:} && c_{bs}(u) = \sum_{s,t \in V} \delta(s,t|u) \cdot A_{us} \\
&(ii) \text{ proximal target:} && c_{bt}(u) = \sum_{s,t \in V} \delta(s,t|u) \cdot A_{ut} \\
&(iii) \text{ } k\text{-bounded distance:} && c_{bk}(u) = \sum_{s,t \in V} \delta(s,t|u) \cdot \mathbb{1}_{dist(s,t) \leq k} \\
&(iv) \text{ length-scaled:} && c_{bd}(u) = \sum_{s,t \in V} \frac{\delta(s,t|u)}{dist(s,t)} \\
&(v) \text{ linearly-scaled:} && c_{bl}(u) = \sum_{s,t \in V} \delta(s,t|u) \cdot \frac{dist(s,u)}{dist(s,t)}
\end{aligned}$$

where

$$\mathbb{1}_{dist(s,t) \leq k} = \begin{cases} 1 & dist(s,t) \leq k \\ 0 & \text{otherwise.} \end{cases}$$

Details of these variants can be found in [35]. Other variants of the general betweenness concept rely on different assumptions of transfer in networks besides shortest paths.

A measure by means of network flow was defined by Freeman et al. [79]. The authors assume information as flow and assign to each edge a non-negative value representing the maximum amount of information that can be passed between the endpoints. The aim is then to measure the extent to which the maximum flow between two vertices s and t depends on a vertex u. Denote by $f(s,t)$ the maximum (s,t)-flow w.r.t. constraints imposed by edge capacities and the amount of flow which must go through u by $f(s,t|u)$. Similarly to shortest path betweenness, *flow betweenness* is then defined as

$$c_f(u) = \sum_{s,t \in V} \frac{f(s,t|u)}{f(s,t)} \; .$$

The value of $f(s,t|u)$ can be determined by erasing u from the graph G. Denoting with $\tilde{f}(s,t)$ the maximum (s,t)-flow in the resulting graph $G-u$, it holds that $f(s,t|u) = f(s,t) - \tilde{f}(s,t)$.

The index was introduced as a betweenness variant for weighted networks but can be readily applied to unweighted networks. In the case of simple undirected and unweighted networks, the maximum (s,t)-flow is equivalent to the number of edge disjoint (s,t)- paths and $f(s,t|u)$ is the minimum number of such paths u lies on [72].

Yet another variant was proposed by Newman [144]. His *random walk betweenness* calculates the expected number of times a random (s,t)-walk passes through a vertex u, averaged over all s and t. Newman shows, that his variant of betweenness can also be calculated with a current-flow analogy by viewing a graph as an electrical network. Random walk betweenness is then equivalent to the amount of current that flows through u averaged over all s and t. Thus, his measure is also known as *current flow betweenness*. Details and formal definitions of his versions can be found in [38, 144].

All variants of betweenness can be described in a more general form considering a flow of information analogy. Depending on the assumption of how information is 'flowing' between a sender s and a target t, the set $P(s,t)$ contains all possible information channels to transmit the piece of information. This set might contain all shortest (s,t)-paths if the information has to be transmitted as fast as possible or all random (s,t)-walks when the delivery time does not play any role. Basically any kind of trajectory on a graph can be thought of as an information channel. The set $P(s,t|u)$ contains all information channels where the vertex u is in a position to control the information flow. For shortest path betweenness, u is in a controlling position if he is part of an information channel and for proximal target betweenness if it presents the information to the target t. In the former case $P(s,t|u)$ comprises all elements of $P(s,t)$ that contain u as an intermediary and in the latter all elements that contain the edge $\{u,t\}$. Again, the position of control could be defined as any location on a trajectory. A measure of relative betweenness is then defined with aggregation rules over the two specified sets, commonly the fraction of their cardinalities. This fraction can also be weighted according to specified rules, e.g. as in length scaled betweenness. Aggregating over all possible sources and targets, we can define a generic betweenness index as

$$c_{bg}(u) = \sum_{s,t \in V} \frac{|P(s,t|u)|}{|P(s,t)|} \cdot \omega(s,t) ,$$

where $\omega(\cdot)$ is the weighting function. Table 3.1 shows the presented betweenness variants categorized by information channels and position of control.

Many other variants are possible, e.g. k-betweenness mentioned in [32], where $P(s,t)$ is the set of all (s,t)-paths of length at most k.

Closeness centrality was first mentioned in the work of Bavelas [14] and later formally defined by i.a. Sabidussi [169]. It is defined as the reciprocal

Information channels $P(s,t)$	Position of control $P(s,t\|u)$	Weighting ω	Index
shortest paths	any intermediary	1	$c_b(u)$
	any intermediary	$\frac{1}{dist(s,t)}$	$c_{bd}(u)$
	any intermediary	$\frac{dist(s,u)}{dist(s,t)}$	$c_{bl}(u)$
	any intermediary	$\mathbb{1}_{dist(s,t)\leq k}$	$c_{bk}(u)$
	first intermediary	1	$c_{bs}(u)$
	last intermediary	1	$c_{bt}(u)$
edge disjoint paths	any intermediary	1	$c_f(u)$
random walks	any intermediary	1	$c_{rwb}(u)$

TABLE 3.1: Categorized measures of existing betweenness measures.

of the sum of the distances of a vertex to all other nodes in the network, i.e.

$$c_c(u) = \frac{1}{\sum\limits_{t\in V} dist(u,t)} \ .$$

Vertices in a network are thus considered more central if they have a small total distance to all other vertices in the network. By definition of graph-theoretic distances, closeness is ill-defined on unconnected graphs. A close variant applicable to both connected and unconnected graphs is given by

$$c_{hc}(u) = \sum_{t\in V} \frac{1}{dist(u,t)} \ .$$

This variant was proposed by various researcher. Among the first are Gil-Mendieta & Schmidt who refer to it as *power index* [85]. Rochat later introduced it as *harmonic closeness* [165].

It has long been suggested, that closeness and betweenness are dual to each other. Brandes et al. [37] show that

$$\sum_{u\in V} c_b(u) = \sum_{s,t\in V} (dist(s,t)-1) = \sum_{s\in V} (c_c(s)^{-1}-1) \ .$$

Therefore, betweenness is a redistribution of aggregated closeness values (or vice versa). Additionally, it holds that

$$\sum_{u\in V} \delta(s,t|u) = dist(s,t) - 1$$

and thus

$$(c_c(s))^{-1} - (n-1) = \sum_{u\in V} \delta(s|u) \ .$$

The proofs and further theoretical considerations can be found in [37]. Defining a matrix M with entries $M_{su} = \delta(s|u)$, we observe that betweenness is defined as column sums and closeness as row sums, illustrating a form of duality between the two measures.

As in the case of betweenness, many different variations of closeness have been proposed, mostly to correct for the fact that the 'classical' closeness is not properly defined on unconnected networks. Valente & Foreman [187] introduce *integration* as an index which measures how well a vertex is integrated in a network. It is defined as

$$c_{int}(u) = \frac{\sum_{t \in V}(\mathrm{diam}(G) + 1 - dist(u,t))}{n-1} ,$$

where $\mathrm{diam}(G) = \max_{s,t \in V} dist(s,t)$ is the *diameter* of the graph. Since the diameter is a constant, we can omit it from the calculations without altering the induced ranking. The same holds for the denominator. We can thus redefine integration as

$$c_{int}(u) = -\sum_{t \in V} dist(u,t) ,$$

which yields a ranking close to harmonic closeness and on connected graphs also as the classical closeness. Although integration purports to measure the integration of a node in a graph, it effectively measures the same as closeness in terms of rankings.

Many other variants of closeness exist, e.g. *random walk closeness* [151], which can also be found in [110]. However, these variants are not vital for the remainder of the thesis.

A variant of particular importance in the upcoming parts of the thesis was defined by Stephenson & Zelen [180]. Their *information centrality* is based on counting all paths between two vertices and the edge overlap among these paths. Having all paths at hand, a matrix is formed that contains the lengths of all paths on the diagonal and the overlap on the off diagonal entries. This matrix is inverted and a harmonic mean of each row is formed.

The authors interpret this procedure from an information-theoretic point view. They argue that the information content of a path is inversely proportional to the length of a path and the edge overlap represents a covariance among paths.

Stephenson & Zelen show, that the calculations actually do not have to be performed explicitly but can be derived by inverting a matrix $C = (L + J)^{-1}$, where L is the Laplacian matrix and J the matrix of all ones. With the matrix C, information centrality equates to

$$c_{inf}(u) = \left(C_{uu} + \frac{T - 2R}{n}\right)^{-1} ,$$

where

$$T = \sum_{v=1}^{n} C_{vv} \quad \text{and} \quad R = \sum_{v=1}^{n} C_{uv} .$$

By definition of the Laplacian matrix, we observe for the matrix $B = L + J$ that

$$\sum_{v=1}^{n} B_{uv} = n \qquad \forall u \in V$$

holds, implying for its inverse $C = B^{-1}$ that

$$\sum_{v=1}^{n} C_{uv} = \frac{1}{n} \qquad \forall u \in V \, .$$

The term $\frac{T-2R}{n}$ is thus a constant for all vertices in the network and we can define information centrality to be

$$c_{inf}(u) = \frac{1}{C_{uu}} \, ,$$

without altering the induced ranking.

The interpretation of the index is not straightforward and its underlying intuition is commonly not understood. Brandes & Fleischer offer an explanation with a current-flow analogy [38]. The index can be interpreted as the harmonic mean of the effective resistance towards a vertex u. Hence, information centrality can also be referred to as *current-flow closeness*.

EIGENVECTOR CENTRALITY was introduced by Bonacich [24] and is part of the category of *feedback centralities*. Measures in this class assume that the centrality of a node is conditional on the centrality of its neighbors. Nodes are highly central if they are connected to other highly central nodes. If we define the centrality of a vertex as the sum of the centrality scores of its adjacent vertices, we obtain

$$c_e(u) = \sum_{v \in V} A_{uv} c_e(v) \, .$$

The centrality scores can be calculated by solving the system of equations $\boldsymbol{y} = A\boldsymbol{y}$, which, however, only has a solution if $\det(A - I) = 0$. Instead, we solve the eigenvalue problem $\lambda_1 \boldsymbol{y} = A\boldsymbol{y}$, where the principal eigenvalue is chosen since the entries of its associated eigenvector have the same sign. The solution is therefore given by the eigenvector $\boldsymbol{x}_1$ and eigenvector centrality is thus defined as

$$c_e(u) = \boldsymbol{x}_1(u).$$

The principal eigenvector can also be computed with the power iteration by repeatedly multiplying A to an arbitrary vector $\boldsymbol{b}_0$ until convergence, i.e.

$$\frac{A^k \boldsymbol{b}_0}{\|A^k \boldsymbol{b}_0\|} \xrightarrow{k \to \infty} \boldsymbol{x}_1 \, .$$

Since an entry $[A^k]_{uv}$ is the number of (u, v)-walks of length k, eigenvector centrality of a node u can also be seen as the limit proportion of walks of the same length starting at u.

Results of eigenvector centrality on graphs with poorly connected dense clusters are difficult to interpret. In this case, the eigenvector will draw most of the weight to one cluster and conceal the actual importance of nodes. The index is thus best applied to graphs with a *core-periphery structure*. Ideally, the vertex set of a graph can be partitioned into two subsets, one inducing a clique (the core) and the other an independent set (the periphery). The eigenvector

x_1 then puts the majority of weight as anticipated to the core. To measure how close a graph is to this idealized structure, we can use Everett & Borgatti's ρ-measure [31], defined as

$$\rho = \sum_{u,v \in V} A_{uv} c_u c_v,$$

where $c_u \in [0, 1]$ is the *coreness* of a node. The larger ρ is the more concentrated is the graph and so, closer to an ideal core-periphery structure. When dealing with unconnected graphs, eigenvector centrality should be calculated for each component separately and then scaled by the size of the component.

Google's *PageRank* is undoubtedly one of the most famous adaptations of eigenvector centrality for directed graphs [153]. It is defined as

$$c_{pr}(u) = \sum_{v \in N^-(u)} \alpha \frac{c_{pr}(v)}{d_v^+} + (1 - \alpha),$$

where $N^-(u)$ is the incoming neighborhood of u, d_v^+ the out-degree of v and α a damping factor, commonly set to 0.85. Although PageRank is attributed to the work of Page & Brin, an equivalent index was already introduced in 1990 by Friedkin & Johnson [82, 83].

A feedback centrality dating back to 1953 was introduced by Katz [103]. Similarly to eigenvector centrality, all walks emanating from a node u are summed up but longer walks are penalized by an attenuation factor α. Formally, *Katz status* is defined as

$$c_{katz}(u) = \sum_{k=1}^{\infty} \sum_{v \in V} \alpha^k [A^k]_{uv}.$$

In order for the series to converge, α has to be chosen such that it is smaller than the reciprocal of the largest eigenvalue of A. In this case, Katz status can be calculated with the closed form

$$c_{katz}(u) = \left[\left(I - \alpha A \right)^{-1} \cdot \mathbf{1}_n \right) \Big]_u .$$

A close variant is Bonacich's *β-centrality*, whose definition also allows for a negative attenuation factor β [24]. It is given by

$$c_{\alpha,\beta}(u) = \alpha \sum_{k=1}^{\infty} \sum_{v \in V} \beta^k [A^{k-1}]_{uv} \ ,$$

where α is merely a scaling parameter, such that it can be omitted without altering the induced ranking. With $|\beta| \leq \frac{1}{\lambda_1}$, a closed form is given by

$$c_{\alpha,\beta}(u) = \left[\left(I - \beta A \right)^{-1} A \cdot \mathbf{1}_n \right) \Big]_u \ .$$

Katz status and eigenvector centrality can be considered as positive feedback centralities, since the centrality of a vertex is higher if it is connected to other vertices with a high centrality score. In contrast, Bonacich's β-centrality with a negative β is a negative feedback centrality, since vertices are considered central, if they are connected to vertices with low centrality score. This kind of centrality is particularly of interest in bargaining situations since bargaining power comes from being in a better position than negotiating partners.

SUBGRAPH CENTRALITY was recently introduced by Estrada & Rodriguez-Velazquez [66]. It is closely related to eigenvector centrality and Katz status, since it also involves counting walks. The difference is that only closed walks are considered and longer walks are inversely weighted by the factorial of their length, i.e.

$$c_s(u) = \sum_{k=0}^{\infty} \frac{\left[A^k\right]_{uu}}{k!}.$$

The weighting by factorials is a convenient choice since it guarantees convergence of the series. Its closed form is given by the matrix exponential, such that

$$c_s(u) = [e^A]_{uu}.$$

Estrada & Rodriguez-Velazquez also consider variants, where only walks of even or odd length are summed up, giving rise to *odd subgraph centrality* and *even subgraph centrality* defined as

$$c_{se} = \sum_{k=0}^{\infty} \frac{\left[A^{2k}\right]_{uu}}{(2k)!} \quad \text{and} \quad c_{so} = \sum_{k=0}^{\infty} \frac{\left[A^{2k+1}\right]_{uu}}{(2k+1)!}.$$

All three measures can also be expressed with the spectral decomposition of A. With the closed forms

$$\cosh(x) = \sum_{k=0}^{\infty} \frac{x^{2k}}{(2k)!} \quad \text{and} \quad \sinh(x) = \sum_{k=0}^{\infty} \frac{x^{2k+1}}{(2k+1)!},$$

we obtain the following spectral forms of all three subgraph variants:

$$\begin{aligned} c_s(u) &= \sum_{j=1}^{n} e^{\lambda_j} \boldsymbol{x}_j^2(u) \\ c_{se}(u) &= \sum_{j=1}^{n} \cosh(\lambda_j) \boldsymbol{x}_j^2(u) \\ c_{so}(u) &= \sum_{j=1}^{n} \sinh(\lambda_j) \boldsymbol{x}_j^2(u), \end{aligned}$$

where $\boldsymbol{x}_j(u)$ is the uth entry of the eigenvector $\boldsymbol{x}_j$ associated with the eigenvalue λ_j. The proof can be found in [66].

A close variant called *bipartivity* was introduced by the same set of authors [65]. In bipartite graphs, all closed walks have even length, such that a measure of the 'bipartiteness' of a vertex can be defined as

$$c_{bip}(u) = \frac{c_{se}(u)}{c_s(u)} = \frac{\sum\limits_{j=1}^{n} \cosh(\lambda_j) \boldsymbol{x}_j^2(u)}{\sum\limits_{j=1}^{n} e^{\lambda_j} \boldsymbol{x}_j^2(u)}.$$

In a bipartite graph, all vertices have a bipartivity score of 1.

Instead of only considering closed walks, Benzi & Klymko's *total communicability* quantifies all walks starting at a vertex u, with the same weighting scheme as subgraph centrality [15]. The index can equivalently be defined as a series, its closed form and in spectral form as

$$\begin{aligned} c_{tc}(u) &= \sum_{k=0}^{\infty} \sum_{v \in V} \frac{\left[A^k\right]_{uv}}{k!} \\ c_{tc}(u) &= \sum_{v \in V} [e^A]_{uv} \\ c_{tc}(u) &= \sum_{j=1}^{n} e^{\lambda_j} \left(\sum_{v=1}^{n} \boldsymbol{x}_j(v) \right) \boldsymbol{x}_j(u) \, . \end{aligned}$$

Both, subgraph centrality and total communicability also exist in parameterized form [16], i.e.

$$\begin{aligned} c_s^{\beta}(u) &= \sum_{k=0}^{\infty} \frac{\left[\beta A^k\right]_{uu}}{k!} = [e^{\beta A}]_{uu} \\ c_{tc}^{\beta}(u) &= \sum_{k=0}^{\infty} \sum_{v \in V} \frac{\left[\beta A^k\right]_{uv}}{k!} = \sum_{v \in V} [e^{\beta A}]_{uv} \, , \end{aligned}$$

where $\beta \in \mathbb{R}^+$. It was shown that the parameterized forms, in the limit cases of β, interpolate between degree and eigenvector centrality, that is

$$\begin{aligned} c_d(u) \xleftarrow{\beta \to 0+} \quad & c_{sc}^{\beta}(u) \quad \xrightarrow{\beta \to \infty} c_e(u) \\ c_d(u) \xleftarrow{\beta \to 0+} \quad & c_{tc}^{\beta}(u) \quad \xrightarrow{\beta \to \infty} c_e(u) \end{aligned}$$

holds for all $u \in V$. The proofs can be found in [16].

The concept of indices based on the matrix exponential has also been used to define a betweenness measure called *communicability betweenness* [64].

Interpretations for indices based on the matrix exponential for social settings are not straightforward and mainly stem from analogies drawn to physical processes. According to the authors of total communicability, the weighting scheme with factorials allows for a physical interpretation by continuous-time quantum walks. The communicability between nodes s and t represents the probability that a particle starting from s ends up at t after wandering on a graph "due to the thermal fluctuation" [64, p. 6]. This thermal fluctuation can be seen as a form of random noise and thus the particle as an information carrier in a network.

3.2 Conceptualization of Centrality Indices

An immediate restriction imposed on centrality indices from a graph-theoretic point of view is that they should only depend on the structure of the network.

This implies that *structural* or, more general, *automorphically* equivalent vertices should always be equally central [110]. Little to none, however, is known about other shared or defining properties of centrality indices such that measures could be defined arbitrarily. Several attempts have been made to delineate and break down the space of indices, which are discussed in this section.

Freeman's Conceptual Clarification

Freeman's seminal work of 1979 constitutes a first successful effort to establish a clear concept of network centrality [75]. He recognized the inherent ambiguity and tried to resolve conceptual issues. Freeman reduced the already existing abundance of measures to the following three competing concepts of centrality using communication in human groups as analogy.

DEGREE as an indicator of communication *activity*. An actor with high degree is "in the thick of things" [75, p. 219]. Actors with low degree are seen as peripheral, isolated from any ongoing communication processes.

BETWEENNESS as indicator for *control* of communication. A person with a high betweenness score can influence a group by distorting or withholding information that is being transmitted via shortest paths. A low betweenness score, on the other hand, limits the potential of being influential.

CLOSENESS as an indicator of *efficiency* or *independence*. An actor with a high closeness score has a low distance to all other actors, so that he or she does not depend on others as intermediaries for information. Being at long distance to others makes actors more dependent on intermediaries to obtain information.

For each concept, Freeman states several alternative of the literature, pointing out that those are "often unnecessarily complicated" [75, p. 220], "absolutely unintelligible from any theoretical perspective whatever" [75, p. 220] or "tend to add unnecessary and confusing complications that make them difficult to interpret" [75, p. 225]. Having these statements in mind and looking at the presented indices which mostly emerged after his work, it becomes apparent that his efforts were not thoroughly acknowledged. His work is mainly perceived as the introduction of degree, betweenness and closeness as centrality indices.

Axiomatization of Centrality Indices

The objective of an *axiomatization* should be to understand and to describe as completely as possible the implications of a list of properties, i.e. the axioms. Historically, many disciplines profited and substantially advanced by introducing axiomatic systems to a common problem. The *Von Neumann-Morgenstern utility theorem* in decision making [140], *Arrows impossibility theorem* in social choice theory [8] and axiomatic systems for bargaining theory [166], to name just a few outside the scope of modern and ancient mathematics.

Numerous attempts have been made to formalize the concept of centrality by introducing a system of axioms. Ideally, a combination of axioms describes the behavior of centrality indices to an extent that facilitates interpretative statements about centrality rankings and aids in the selection of indices. Sabidussi's [169] seminal work appears to be the first along these lines, and many others have followed [20, 109, 116, 146, 147, 168, 188].

The different approaches mostly follow similar guidelines. Axioms are chosen to be desirable, or intuitively plausible properties of indices under graph transformations such as adding or switching edges, e.g. adding an edge to a vertex should never decrease its centrality score [169]. Axiomatic approaches under this premise are typically restrictive, i.e. the investigation of whether indices are valid according to the axioms often only leaves a few possibilities (cf. Table 3.2), demonstrated with simple counterexamples [108].

	Sabidussi [169]	Ruhnau [168]	Landherr [116]	Boldi [20]
degree	yes	no	no	no
betweenness	no	yes	no	no
closeness	no	no	no	no
eigenvector	no	yes	no	no
Katz status	no	?	yes	no
harm. closeness	no	no	no	yes

TABLE 3.2: Axiomatic systems and whether indices fulfill all axioms.

Although axiomatic systems seem to be well-defined, it does not suffice to justify the exclusive focus on indices fulfilling those. As the literature shows, there are many ideas about *intuitive plausible* that favor different indices, yet none are general enough to encompass the concept as a whole. This impedes general theorems about network centrality and sometimes only shifts the focus from the definition of indices to the definition of axioms.

Centrality indices generally behave very inconsistent under edge transformations so that finding a common ground of how scores change is virtually impossible. As we have pointed out before, however, actual scores do not play a role in applications. We could allow for score variations as long as it does not alter the ranking. Therefore, if we seek for an axiomatic system it should focus on the induced rankings instead of the function values.

After all, some axiomatizations may well serve as *representational theorems* of certain groups of indices [3, 109].

Classification of Indices

The classification of indices is more conceptually oriented. It provides terminology and intuition to reason about the features embodied in centrality indices and relates formal definitions with substantive motivations. Mentionable work in this line are the frameworks provided by Borgatti [28] and Borgatti & Everett [32].

BORGATTI'S FRAMEWORK is based on network flows. He argues that centrality indices measure different kinds of traffic flowing through a network which can be categorized along two key dimensions. The kind of trajectory the traffic follows (shortest paths, walks, etc.) and its dyadic diffusion mechanism (parallel duplication, serial duplication, or transfer). Simple examples are described in the following.

TRANSFER OF MONEY. A banknote is passed around randomly within an economy. It does not follow prescribed routes and can easily move several times between two individuals. In graph-theoretic terms, the banknote traverses the network via walks. Since it is additionally indivisible, the dyadic diffusion process is a transfer.

GOSSIP. Typically, gossip is told to various individuals but one at a time confidentially. The story can thus be familiar to many people at the same time. Unlike a banknote, it usually does not traverse the same edge twice but can reach an individual several times via different edges. It thus follows trails in the network and it diffuses with a serial duplication.

PACKAGE DELIVERY. A package should be delivered in the fastest way possible. In a network of roads as edges and intersections as nodes, the driver selects the shortest rout possible to its destination. Therefore, the package is transferred via shortest paths.

Borgatti uses additional examples and simulates these processes to determine which centrality index models it the best. His resulting categorization of indices can be found in Table 3.3. Borgatti notes that although indices are distinct, they quantify similar outcomes in terms of network flow and many of the considered flow process are not covered by any index. In particular the ones he considered the most important, gossip and infection, are not summarized by any existing index.

	Parallel duplication	Serial duplication	Transfer
Geodesics		Closeness	Closeness Betweenness
Paths	Closeness Degree		
Trails	Closeness Degree		
Walks	Closeness Degree Eigenvector Katz status		

TABLE 3.3: Flow processes and centrality by Borgatti [28].

BORGATTI & EVERETT'S FRAMEWORK is purely based on graph-theoretic notions. They show that the assessment of a nodes involvement in the walk structure of a graph is a unifying property of several indices. They identified four dimensions which can be used to distinguish between centrality indices: the considered walk type (e.g. geodesic or edge disjoint), the walk property (volume or length), the walk position (radial or medial, see also Figure 3.2) and the summary type (e.g. aggregation or mean). The classification according to walk position and property is given in Table 3.4.

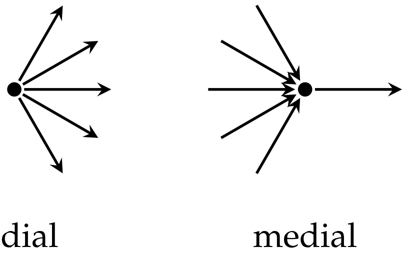


FIGURE 3.2: Illustration of walk positions.

	Radial	Medial
Volume	degree eigenvector Katz status	betweenness
Length	closeness information	

TABLE 3.4: Walk involvement and centrality by Borgatti and Everett [32].

Both frameworks provide certain guidelines for empirical research and prove helpful to answer the question 'which index to choose?' in applications. Measures that fall in the same category can potentially be used interchangeably and one can reasonably ask which performs best. However, the frameworks do not allow for sharp distinctions and provable statements.

A Spectral Framework for Centrality Indices

In Section 3.1, we saw that many indices can be defined in terms of the spectral decomposition of the adjacency matrix of a graph. From a more general point of view, any vector in a n-dimensional space can be written as a linear combination of orthogonal vectors that span the space. In the case of graphs, an orthogonal basis is formed by the eigenvectors X. Each vector $\boldsymbol{y}$ in the space spanned by X can be written as

$$\boldsymbol{y} = \sum_{j=1}^{n} r_j \boldsymbol{x}_j \, ,$$

where r_j are real-valued scalars. Let $\boldsymbol{c} = (c(u_1),\ c(u_2), \ldots,\ c(u_n))$ be the vector of centrality scores of an arbitrary index $c : V \to \mathbb{R}_0^+$. Since centrality indices are defined with the adjacency matrix A, we conjecture that it should be possible to express $\boldsymbol{c}$ as a linear combination of eigenvectors of A, i.e.

$$\boldsymbol{c} = \sum_{j=1}^{n} r_j \boldsymbol{x}_j .$$

It remains to determine the coefficients r_j. A trivial case is eigenvector centrality, where $r_1 = 1$ and $r_j = 0$ for all $1 < j \leq n$.

A spectral form for degree is derived in the following. Since degree can be defined as the row sums of the adjacency matrix, its vector can be written as

$$\boldsymbol{c}_d = A\mathbf{1}_n = X\Lambda X^T \mathbf{1}_n = X\Lambda \boldsymbol{w} \,.$$

In summation form we have

$$\boldsymbol{c}_d = \sum_{j=1}^{n} \lambda_j w_j \boldsymbol{x}_j ,$$

i.e. $r_j = \lambda_j w_j$.

A more general approach is given in the following theorem.

Theorem 3.1. *Let $f : \mathbb{R}^{n,n} \to \mathbb{R}^{n,n}$ be a matrix function defined as a power series*

$$f(X) = \sum_{k=0}^{\infty} \alpha_k X^k \,.$$

Further, let $c : V \to \mathbb{R}$ be an arbitrary centrality index. Then, the following statements hold true.

(i) $\boldsymbol{c} = f(A)\mathbf{1}_n \iff c(u) = \sum_{j=1}^{n} f(\lambda_j) w_j \boldsymbol{x}_j(u) \qquad \forall u \in V$

(ii) $\boldsymbol{c} = diag(f(A)) \iff c(u) = \sum_{j=1}^{n} f(\lambda_j) \boldsymbol{x}_j^2(u) \qquad \forall u \in V$

Proof. The proof for both statements is straightforward due to the following known equality [136, 150]:

$$f(A) = f(X\Lambda X^T) = Xf(\Lambda)X^T \,.$$

Therefore,

$$\boldsymbol{c} = f(A)\mathbf{1}_n = Xf(A)X^T \mathbf{1}_n = Xf(\Lambda)\boldsymbol{w} \,.$$

The centrality score for a vertex u is then the uth entry of $\boldsymbol{c}$ and we can write

$$c(u) = \sum_{j=1}^{n} f(\lambda_j) w_j \boldsymbol{x}_j(u)$$

for all $u \in V$.

For indices based on the diagonal entries we have

$$\mathrm{diag}(f(A)) = \mathrm{diag}(Xf(\Lambda)X^T)$$

and thus for the centralities scores it holds that

$$c(u) = \sum_{j=1}^{n} f(\lambda_j) \boldsymbol{x}_j^2(u)$$

for all $u \in V$. □

The theorem implies, that if we know the function f for centrality indices, we can represent them in one of the two described forms. Besides the already known spectral forms of subgraph centrality and total communicability, we can also derive a spectral form for Katz status. It holds that

$$f_{katz}(A) = (I - \alpha A)^{-1} .$$

The vector of scores is defined as the row sums of the resulting matrix. Therefore, Theorem 3.1(i) applies and we can write Katz status as

$$c_{katz}(u) = \sum_{j=1}^{n} \frac{1}{1 - \alpha \lambda_j} w_j \boldsymbol{x}_j(u) .$$

The benefit of the spectral representation is that we can compare indices analytically by how much emphasis is put on individual eigenvectors. For indices defined as row sums, the weight of the ith eigenvector is given by $r_i = f(\lambda_i) w_i$. Since the fundamental weight is independent of the function f, we can solely focus on $f(\lambda_j)$. Figure 3.3 illustrates the functions for degree, Katz status and total communicability.

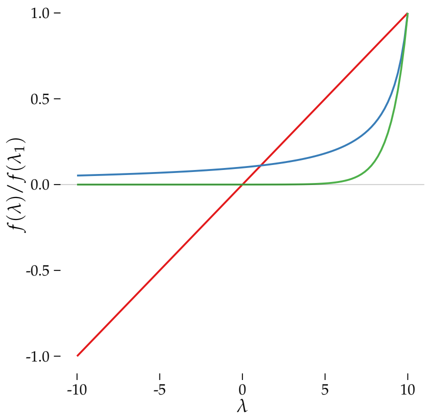


FIGURE 3.3: Weighting functions of eigenvectors for spectral forms of degree(red), total communicability(green) and Katz status(blue). Weights are normalized with the function value of the principal eigenvalue.

Of course, the actual weighting depends on the actual distribution of eigenvalues and is generally not expected to be equidistant as in the considered example. Yet, we can see that the weighting of Katz status and total communicability is similar so that we expect them to produce similar results in terms of vertex rankings. Note that eigenvectors with associated negative eigenvalues do not play a significant role for these two measures. In contrast, degree puts negative weights on these vectors.

The example shows the analytic advantages of the spectral forms, however, we could only derive spectral forms for a small subset of indices. Defining a function with power series representation for most indices, e.g. betweenness and closeness, is non-trivial and potentially not possible at all. But we can establish some analytic connections among indices by means of different matrices associated with a graph.

The Laplacian matrix with spectral decomposition $L = Y\Lambda^{(L)}Y^T$ can be used to derive a spectral representation of degree in diagonal form, i.e.

$$\boldsymbol{c}_d = \operatorname{diag}(L) = \sum_{j=1}^{n} \lambda_j^{(L)} \boldsymbol{y}_j^2 = \sum_{j=1}^{n-1} \lambda_j^{(L)} \boldsymbol{y}_j^2 \ ,$$

where the third equality holds since $\lambda_n^{(L)} = 0$ for all graphs [135].

Another index that can be expressed with the spectrum of the Laplacian is information centrality. Its vector can be written as

$$\boldsymbol{c}_{inf}^{-1} = \operatorname{diag}((L + J)^{-1}) \ ,$$

where the inversions must be understood component-wise. Let $Z\Lambda^{(B)}Z^T$ be the spectral decomposition of $B = (L + J)$. We can thus write

$$\boldsymbol{c}_{inf}^{-1} = \sum_{j=1}^{n} \frac{1}{\lambda_j^{(B)}} \boldsymbol{z}_j^2 \ .$$

The eigenvectors and eigenvalues of L and B are closely related by

$$\lambda_j^{(B)} = \lambda_{j-1}^{(L)} \text{ and } \boldsymbol{z}_j = \boldsymbol{y}_{j-1} \qquad \text{for } j = 2, \ldots, n \ .$$

Also $\lambda_1^{(B)} = n$ and $\boldsymbol{z}_1 = r\mathbf{1}_n$ where $r \in \mathbb{R}^+$ holds. Since the entries of $\boldsymbol{z}_1$ are all equal, we can omit it from the summation, since it does not alter the rankings. Thus we obtain

$$\boldsymbol{c}_{inf}^{-1} = \sum_{j=2}^{n} \frac{1}{\lambda_j^{(B)}} \boldsymbol{z}_j^2 = \sum_{j=1}^{n-1} \frac{1}{\lambda_j^{(L)}} \boldsymbol{y}_j^2 \ .$$

Observe, that we can now formally compare degree and information centrality quite easily. We have $f_d(\lambda_j) = \lambda_j$ for degree and $f_{inf}(\lambda_j) = 1/\lambda_j$ for the inverse values of information centrality. We thus expect a very high rank correlation between the two measures, which we show to be true in Chapter 9.

Unfortunately, the spectrum of the Laplacian and the adjacency matrix are seemingly unrelated. The same holds true for the distance matrix which can be used for a spectral form of closeness. If a connection would exist, e.g. there is a function $g : \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$ such that $g(A) = L$, we could express all centrality indices by means of the eigenvalues and eigenvectors of the adjacency matrix. This would greatly facilitate the formal comparison of indices since we would have a general framework for all indices and can establish analytic connections with the respective functions.

Justification of Indices

In the previous subsections, we introduced several existing approaches to conceptualize network centrality. None of them, however, are generally accepted ways to define the concept of network centrality. In the absence of a formal basis, researchers mainly rely on two different methods to justify new indices, which we briefly discuss here and relate them to the topics covered in this thesis.

STAR PROPERTY. It appears that the only requirement that is both formally established and substantively accepted is the *star property*. In the words of Freeman,

> *"A person located in the center of a star is universally assumed to be structurally more central than any other person in any other position in any other network of similar size."* [75, p. 218]

This statement is frequently invoked as a justification for newly defined indices. If an index attains the highest value for the center of a star, it can be considered as a measure of centrality. The argument is certainly intuitively understandable, yet there is no analytic justification that it should universally hold true. In Chapter 7, we provide a formal substantiation for the star property but also show that it is not decisive enough to distinguish between well-defined and contrived indices. We introduce a class of networks in Chapter 8, which serves the purpose as a benchmark for centrality indices unambiguously.

CORRELATION. When new indices are introduced, most often a correlation analysis with existing indices is performed. Its motivation is given by a general consensus in the literature, described by Valente et al.:

> *"If centralities are not highly correlated, they indicate distinctive measures, associated with different outcomes."* [186, p. 1]

That is, a weak correlation with existing indices justifies the adoption of the new index since it measures structural importance on a different, or even new level. The topic of correlation among centrality indices is revisited in Chapter 9, where we show that the correlation is only weakly, if at all, related with the definition of indices and strongly depends on the underlying network structure.

3.3 ILLUSTRATIVE AND MOTIVATIONAL EXAMPLES

We now have a basic understanding of how centrality is measured and we have seen that there exists a myriad of different ways for its quantification. We now go through some small examples to point out differences and similarities among indices and motivate later theoretical explorations. Concentrating on the four most widely used measures degree, betweenness, closeness, and eigenvector centrality, we start by investigating induced *centers*, i.e. the vertex with the highest score on different graphs.

Four different conceptualizations of structural importance (number of neighbors, distances, dyadic dependencies and limit proportion of walks) may generate an expectation of obtaining different centers for each index. The graph on the right in Figure 3.4 shows that this can indeed be the case.

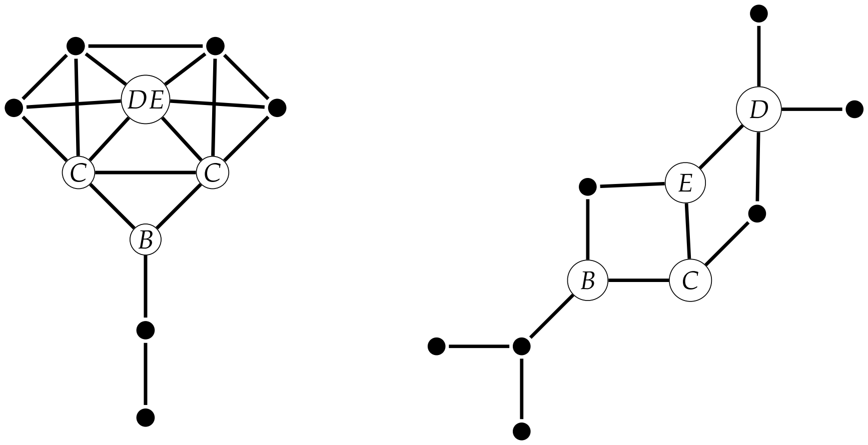


FIGURE 3.4: Krackhardt's Kite and a graph with four different centers according to degree (D), betweenness (B), closeness (C) and eigenvector centrality (E)

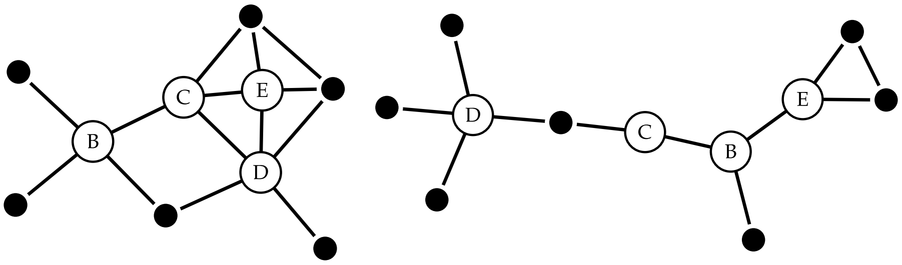
However, centers can also coincide. The graph in Figure 3.4 on the left, known as *Krackhardt's Kite*, was introduced by David Krackhardt to illustrate the differences between degree, betweenness and closeness [112]. The three measures all induce different centers, yet if we add eigenvector centrality to the analysis, its center coincides with the degree center.

The already mentioned star shaped graphs provide the extreme cases for coinciding centers since all indices should induce the same center. Extreme examples for disjoint centers are depicted in Figure 3.5. The left graph is the vertex minimal graph where all four indices induce a different center whilst the right graph is edge minimal [39].

The presented small graphs with a restricted set of indices already give a feeling of what is possible on larger graphs. If we can have four centers with four indices in a graph with ten vertices, it might well be possible to define indices in a way that any vertex in a graph, reasonably or not, can be considered as the most central. The situation gets even more complicated by taking the whole ranking into account.

Figure 3.6 shows a graph with nine vertices, where the ranking of eleven selected centrality indices is extremely discordant. On the other hand, Figure 3.7 shows a nine vertices graph, with no discordance for all eleven cases.

The former example creates the impression, that any ranking is possible with the proper index, whereas the latter conveys that there is only one concept of centrality necessary. Admittedly, both graphs are on the extreme ends and graphs will generally be 'somewhere in between'. However, they give rise to



(A) $n = 10, m = 14$

(B) $n = 11, m = 11$

FIGURE 3.5: Redrawn from [39].(a) vertex and (b) edge minimal graphs with different centers according to degree (D), betweenness (B), closeness (C) and eigenvector centrality (E).

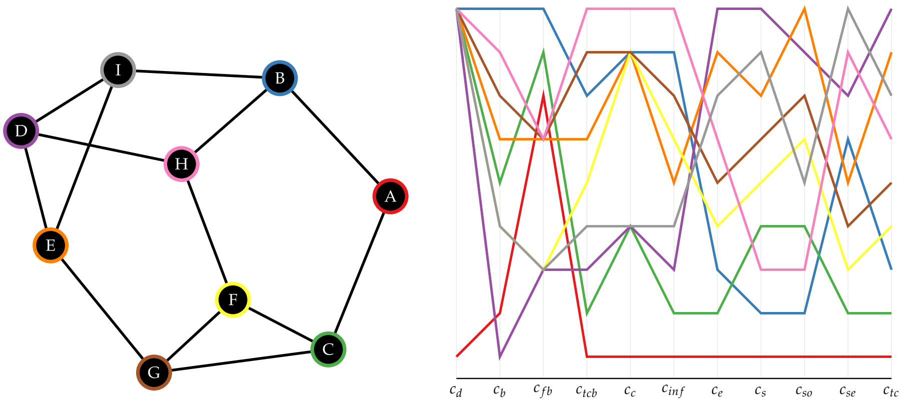


FIGURE 3.6: Graph with extreme discordant rankings for eleven different centrality indices, depicted as parallel coordinates on the right (ranking from top to bottom).

several questions. What structural properties of graphs lead to one or the other extreme? How can we incorporate these properties in the concept of centrality? These and more theoretical questions are tackled later in the second part of the thesis. In the following two chapters we first deal with the application of centrality indices in empirical research.

3.4 Discussion

The literature is flooded with centrality indices and new ones are introduced on a regular basis. This poses severe problems in empirical research, since possibilities are endless to pick a suitable measure. Ideally, the substantive nature of the relations forming the network in question should determine the appropriate definition of structural importance and so the measure to be used. That is, none of the introduced indices is superior to others and every measure can be appropriate for some yet not all questions related to structural

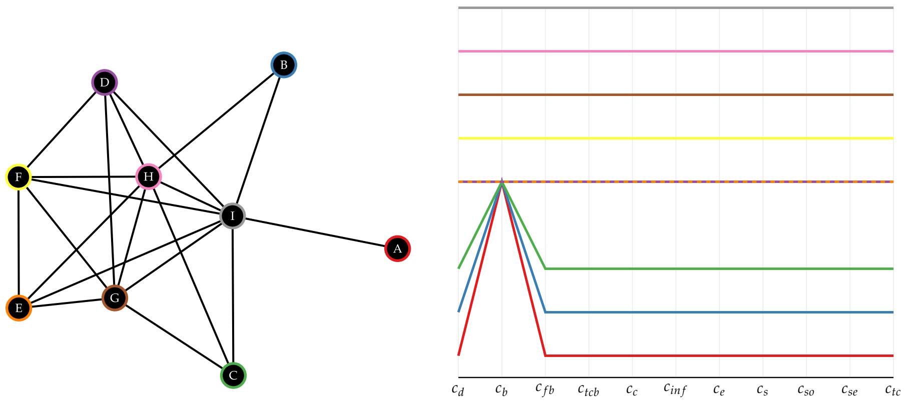


FIGURE 3.7: Graph with almost unique ranking for eleven different centrality indices, illustrated with parallel coordinates on the right (ranking from top to bottom).

importance. The application of indices should thus be driven by the nature of the network and the empirical phenomenon in mind. This suggests that the use of centrality should be seen as a procedure of *measurement*. Therefore, it is inevitable to understand what indices are actually measuring. The classifications presented in Section 3.2 provide guidelines for the most common indices and confidence for their application. But, as already noted by Freeman in 1979, many others are mathematically too complex to apply in a non-physical context. Admittedly, many phenomena in nature can be explained by physical models, e.g. the Navier-Stokes equations describing the motion of fluids, and even social science make use of methods drawn from physics [179]. In order to be meaningful, however, the connection of a physical model and an empirical phenomenon should be evident. It is not immediately obvious, e.g. that particles wandering on a graph by laws of thermal fluctuation describe a needle passed around in a drug-user network [16].

In the next chapter, we turn to the actual application of centrality in empirical research. We discuss several important examples from the early stage of social network analysis and proceed to how centrality is used more recently.

CHAPTER 4

Centrality in Empirical Research

> *"Triggered by recently available data on large real networks [. . .] combined with fast computer power on the scientist's desktop, an avalanche of quantitative research on network structure currently stimulates diverse scientific fields."*
> – Bornholdt & Schuster, 2003

4.1 EARLY DEVELOPMENTS AND APPLICATIONS

The foundations for the concept of network centrality were laid in the 1930s by Moreno's development of *sociometry* [139]. In the words of Moreno

> *"Sociometric explorations reveal the hidden structures that give a group its form: the alliances, the subgroups, the hidden beliefs, the forbidden agendas, the ideological agreements, the 'stars' of the show."* [139]

Moreno studied all kinds of sentiments among individuals within groups and defined several structures that may arise in this context. Notable structures for centrality are given in Figure 4.1. His perception of central individuals is tied to the number of choices, i.e. edges, an individual receives. He defines a dominating individual, the *star*, to be an individual who receives at least five choices.

Although he did not explicitly introduce the term centrality, Moreno was among the first who made significant contributions for social network analysis and thus also for centrality.

The original idea of structural centrality was later developed by Bavelas [13, 14] and Leavitt [118] in experiments on communication patterns. Five persons played a game in which they had to solve a puzzle, where each person is given a unique piece of information. The participants could transmit information through predefined communication channels (cf. Figure 4.2) until every person obtained the solution. The hypothesis was that decentralized communication given in a circle should be the most efficient structure for solving the problem and that the centralized wheel should be the least efficient.

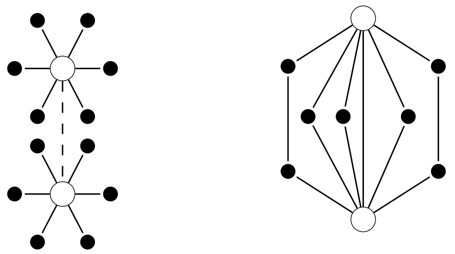


FIGURE 4.1: Redrawn from Moreno's group structures involving stars [139]. (Left) Two centralized sub-groups with two dominating individuals (white nodes). (Right) A group in which two dominating individuals are strongly related, both directly and indirectly.

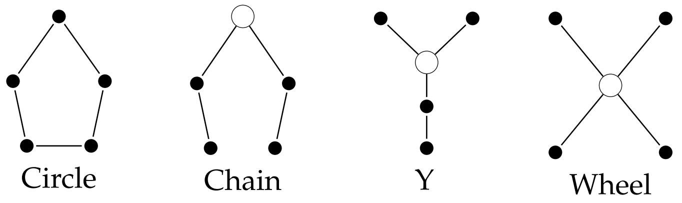


FIGURE 4.2: Redrawn from Leavitt's communication patterns [118]. While the circle does not exhibit any centrality feature, the centrality of the white node increases from the chain towards the wheel.

The results of the experiment, however, were exactly the opposite such that Bavelas and Leavitt concluded that centralization was the most important aspect for organizational communication. The more centralized an organization is, the better it communicates and performs, at least theoretically. Their conceptualization of centrality was based on the distance of each node to all others in the graph, later formalized as closeness.

Over time, the concept of centrality was applied to research questions beyond the scope of communication networks. Pitts investigated the advantageous geographical position of Moscow from a network perspective and its political fortunes as a consequence thereof [156]. He examined the Russian trading routes network of the 12th - 13th century with graph-theoretic distances used as a measure of *system effort* and Shimbel's stress centrality. He found that Moscow can be considered the most central in both measures as depicted in Figure 4.3. Pitts concluded that this central position may have led to social and economic advantages associated with the growth of Moscow.

A remarkable centrality study not involving any index was done by Krackhardt [113]. Manuel, a new manager in a business unit, implemented some organizational changes which lead to backlogs in the workflow of the unit. Looking at the advice seeking network drawn by Krackhardt in Figure 4.4, he realized the (informal) central position of Nancy and reconstituted some of previous workflow in collaboration with Nancy. The notable part of the study is the absence of a quantitative assessment of centrality. One could obviously argue that Nancy has the highest indegree, or use feedback centrality measures.

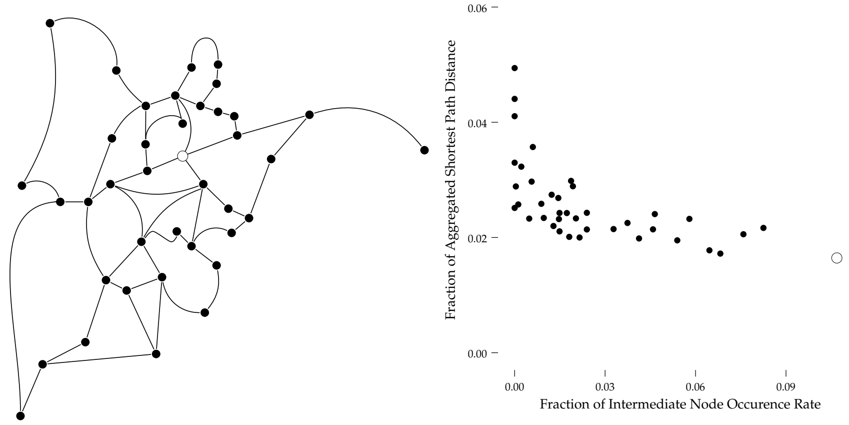


FIGURE 4.3: Redrawn from Russian trading routes in the 12th–13th centuries [156]. (Left) Graph of the trading routes among 39 Russian cities and (Right) scatterplot of fractions of aggregated shortest path distances and intermediate node occurrence rate, i.e. stress centrality. The white vertex and circle represent Moscow.

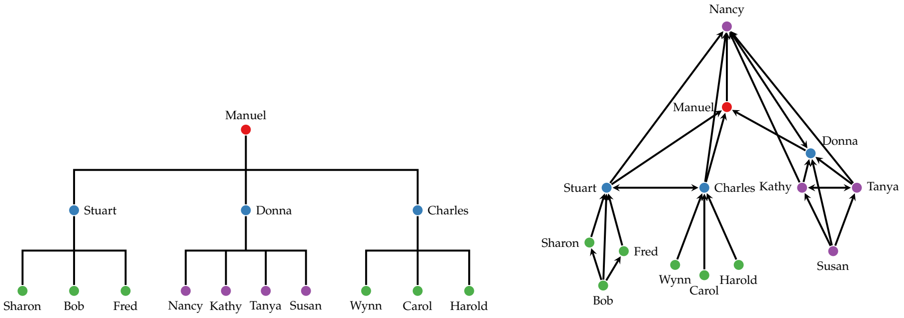


FIGURE 4.4: Redrawn from Krackhardt's hierarchy and advice seeking network [113]. (Left) Formal status hierarchy in a business unit. (Right) Advise seeking network as an informal status hierarchy. Directed edges according to 'who do you turn to for work-related questions?'. The network is drawn so that as many edges as possible point upwards.

But, as Krackhardt explains, "these pictures communicate much more than any number or statistical results" [113, p. 165]. Additionally, the case study nicely shows how different structural importance and perceived individual importance can be.

Over time, centrality indices were more often used as indicators or explanatory variables for certain actor attributes. Tsai & Ghoshal hypothesized in a seminal work that

> *"the centrality [betweenness centrality] of a business unit in interunit social interaction will be positively associated with the level of its perceived trustworthiness"* [185, p. 466].

The association between trustworthiness and betweenness was evaluated by the correlation of these two actor variables.

Note the different uses of centrality in the study of Pitts and in the work of Tsai & Goshal. Pitts outlines specific reasons why stress and closeness are appropriate in his context and argues about their connection to empirical phenomena. Tsai & Ghoshal, on the other hand, purely rely on correlation based evidence. This correlation approach became prevalent in later studies.

4.2 Recent Developments and Applications

The early work on network centrality was reserved for studies in social sciences and related fields. Towards the end of the 20th century, however, "there was a revolutionary change in the field" [78, p. 4]. An article about the *small-world effect* by Watts and Strogatz [196] and one about *scale-free* networks by Barabási and Albert [9] drew the natural science, specifically physics and biology to network research and the application of centrality indices. Networks are no longer exclusively social but rather *complex* [143].

One of the earliest and also most notable application of centrality in biology was a study by Jeong et al. of a protein interaction network [101]. They found that protein lethality can be explained in part by the number of interactions a protein is involved in. This study led to an exploding number of follow up studies, where increasingly sophisticated centrality indices were found to correlate even stronger with protein lethality.

Other applications in biology include metabolic networks [128, 192] and protein folding [190]. The advances in these two research efforts are very much in line with the aforementioned case. Vendruscolo et al. are among the first to hypothesize that betweenness centrality is able to pinpoint critical residues in the folding of proteins [190]. Later, several studies argue that closeness is more effective in this task [5, 42, 55].

Ambedkar et al. investigated human protein interactions data to uncover genes associated with *diabetes mellitus* [4]. Using 14 different centrality indices, they report the top 10 genes of all indices as potential drug targets for diabetes mellitus. Many others enqueue in this line of research with similar approaches [111, 126, 193, 205].

In the course of the years, more and more research areas joined the trend of applying centrality indices to respective research question. In the aftermath of the financial crises of 2007-08, researchers were concerned with the question to which extent it is possible to predict systemic risk, i.e. the risk of a default of the financial system. Battiston et al. were among the first to design an index, called *DebtRank*, to identify systemically important financial institutions within financial networks [12]. Banks are now not only 'too-big-to-fail' but also 'too-central-to-fail'. Others follow with similar indices, e.g. *SinkRank* of Soramäki et al. [177] or the *liquidity spreading index* of León et al. [120]. Additional applications in the financial sector include the building of well-diversified portfolios to reduce investment risk [159].

Very recently, also cosmic networks "foray into this new arena" of network science [100, p. 2000]. Centrality indices were applied to an observed galaxy distribution and Hong & Dey explored whether these tools can be useful for cosmological and astrophysical studies [100].

4.3 Discussion

The number of applications of centrality in different contexts is ever-expanding due to the prospect of new insights and hope for the possibility of facilitating substantial progress. Considering the field of biology, a majority of research questions can only be answered with time consuming and expensive experiments, which can potentially be simplified through computational efforts. One of the key applications in biology is the identification of important proteins in protein interaction networks. The aim is to find potential drug targets to fight diseases, e.g. the mentioned diabetes mellitus and "if a protein can be considered in advance as a drug target, the process of drug discovery can be greatly improved and the cost of experiment can be dramatically reduced." [126, p. 19]. Applying centrality indices to a protein interaction network can be done without any effort and allows for results in no time.

What is all too often not understood is that *network science* is not just a rag-bag of methods but a scientific field by itself [40]. A lot of effort has been made to formalize concepts for network analysis, yet they are all too often neglected by researchers outside the field [77] and too much attention is given to previous anecdotal evidence, although results are promising on first sight.

Many issues are attached to recent studies and the application of centrality indices in general. Three crucial points in this context are outlined in the following.

Availability of Data. The formation of protein interaction databases like String [182] Dip [199] and Biogrid [178] simplified the access to massive amounts of data and potential investigation of protein interaction networks. This availability, however, involves the risk that we end up with *data-driven hypotheses* and/or *post hoc theorizing*. In the case of protein interaction data, we have to take several things into account to not end up with the mentioned fallacies. Interactions among databases are not consistent since they have different sources, the data contain false positives, nomenclature issues exist (i.e. different naming schemes among databases) and the data is incomplete. For these reasons alone, it is not sufficient to derive results from single datasets without validating them on others.

On the other extreme, we might be faced with a total lack of data as, e.g. in the case of banking networks. Financial data is mostly restricted or highly confidential, such that researchers rely on techniques to produce networks that are supposed to be similar as the real network [177]. Several models exist to produce random networks, e.g. the Erdős-Rényi model [61,86] for uniform random graphs or the Barabási-Albert model [9] for scale-free networks. Although these models possess some features of real-world networks, they are generally to prescribed to be a decent substitute of real data and results on these generated data has to be handled with care.

APPLICABILITY OF CENTRALITY. Network centrality is commonly used as explanatory, independent, as well as intermediate variable in empirical research. Research hypotheses typically state that the level of some variable of the network, i.e. vertex attributes like protein lethality or the risk of default for a bank, is either positively or negatively associated with some centrality index, constituting a *centrality effect*. The selection of a centrality index is usually the weakest part of a research design, as little reliable knowledge exists that places one index over another. Moreover, if associations cannot be confirmed empirically, one is at loss. The consequence is that the application of centrality indices all too often culminates in trial-and-error efforts. A new study uncovers a connection between a centrality index and an empirical phenomenon and follow up work engages in a hunt for more effective indices, where choices are purely made by performance and not with substantiated arguments. Although the frameworks of Section 3.2 would give some guidelines to reason about specific choices, the justification is mostly done with an appeal to common practice by adopting "an 'agnostic' perspective by looking at some of the common centrality/peripherality measures" [159, p. 2] or an appeal to authority: "We chose this index because, as Freeman argued, it is the most suitable centrality measure [. . .]" [185, p. 469]. If no suitable index can be found new ones are designed to fit the data or, in extreme cases, hybrid indices are formed with linear combinations "based on the observation that by using it we constantly obtain better performing portfolios" [159, p. 7]. This methodology creates the impression, that applications involving centrality indices are drifting towards the field of data mining. However, centrality should not be viewed as a data mining task but rather as a procedure of measurement.

AVAILABILITY OF TOOLS. The effortless application of centrality indices is due to the vast amount of tools available to perform network analysis. UCINET[1], Gephi[2], visone[3], Pajek[4] and CentiBiN, a software specifically for biological networks [102], all offer the possibility to apply various indices on networks with a single click. Many indices are defined for connected and unweighted (or otherwise limited classes of) networks, but implementations often output results for networks outside of this scope. Studies need to check carefully whether results obtained from such analyses are meaningful.

[1] https://sites.google.com/site/ucinetsoftware/home

[2] https://gephi.org/

[3] http://visone.info/

[4] http://mrvar.fdv.uni-lj.si/pajek/

The preceding points should not be seen as a taunt of the current state of research connected to centrality but rather address some deficiencies inherent in the concept. Due to the non-existence of a prevailing definition for centrality, the degrees of freedom are high enough to permit any kind of explorations under the pretext of the concept.

We substantiate our allegations by examining the role of centrality in protein interaction network in its entirety in Chapter 5. We use this application as a prime example to emphasize current problems since it includes instances of all three categories.

Chapter 5

Centrality in Protein Interaction Networks

"The most highly connected proteins in the cell are the most important for its survival"

– Jeong et al., 2001

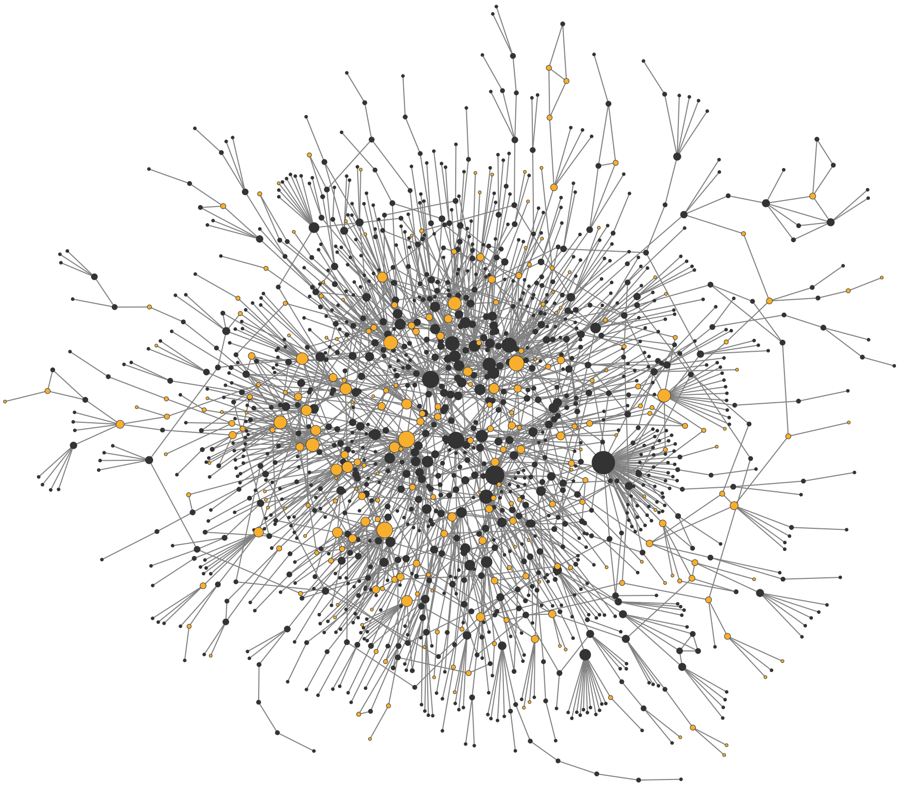


FIGURE 5.1: Protein interactions in a cell of *Saccharomyces cerevisiae*.

5.1 Introduction

A remarkable example for the seemingly successful application of centrality indices can be found in the area of systems biology in discriminating proteins by means of their importance for a cell's survival. Identifying proteins which are essential for the survival of a cell with experiments is a time consuming and expensive process. However, it is a necessity for drug design against diseases since the knock-off of essential proteins (henceforth also referred to as lethal proteins) causes death for a cell [46]. Thus, it truly sounds appealing to support this procedure by applying network analytic methods on the protein interaction network (PIN) to determine structural features shared by lethal proteins. In a seminal work by Jeong et al., it was indeed found that the lethality status of a protein can be partially explained by the number of interactions with other proteins [101]. That is, the higher the degree of a protein, the higher the probability it causes death of the cell if knocked-off. The study was performed on a PIN of *Saccharomyces cerevisiae* (a form of yeast), which is often used as a model organism, since it is easy to cultivate. The findings led to an exploding number of follow up studies claiming an even stronger associations between lethality and different centrality indices and terms like a *centrality-lethality hypothesis* and sometimes even a *centrality-lethality rule* are proclaimed [97].

Comparative studies of existing indices were conducted [62, 63] or new ones were introduced and specifically designed to correlate with lethality [123, 194]. A different line of research even tried to combine indices in various ways. Del Rio et al. show, that the correlation can be significantly improved by combining two centrality indices [54]. Chua et al. introduce a probabilistic approach combining the results of several indices [44]. They use a supervised technique, i.e. they train and determin the relative importance of each method according to a set of known lethal proteins. Others make use of a random forest approach [154].

Although it was already conjectured in [101], that similar outcomes can be expected for PINs of various organisms, only a few studies dealt with other organisms like *Drosophila melanogaster* (fruitfly) and *Caenorhabditis elegans* (a worm) [93]. In a recent study, Raman et al. review the centrality-lethality hypothesis across PINs of 20 different organisms [162]. Using a bootstrapping approach, they show that degree and betweenness centrality of lethal proteins are significantly higher than the network average. In contrast, closeness centrality was found to be less indicative of lethality.

There exist some debates from a biological point of view concerning the data consistency, i.a. false positive lethal proteins impairing the reliability of the results [97, 206]. The high correlation between degree and lethality could well be due to a sampling bias towards lethal proteins. That is, if a specific protein is found to be lethal, it is more likely that this protein is analyzed for interactions rather than a non-lethal protein [52]. Others argue that it is not possible to identify lethal proteins with the network structure alone [125, 184], making methods that combine centrality indices with biological information (e.g. gene expression data [124, 183, 204], orthologous information [155]) the most promising approaches.

Despite these objections, the validity of the centrality-lethality hypothesis is widely accepted and considered as a successful application of centrality measurement. So far there has been little discussion about the reliability of results and the actual applicability of centrality indices in this domain. This chapter is devoted to a general discussion about the application of centrality measurement in PINs. We start with a simple re-analysis of the original study conducted by Jeong et al. in Section 5.3, pointing towards several shortcomings and misleading results in the analysis. Due to the importance of *S. cerevisiae* as a model organism, we perform a more extensive analysis in Section 5.4 to scrutinize the plausibility of the hypothesis. In Section 5.5, we examine the conjecture that the results are portable to other organisms and further demonstrate data dependencies of the results. Our results are summed up in Section 5.6 with some general remarks on empirical studies performed with centrality indices.

5.2 Material and Methods

Data for S. cerevisiae

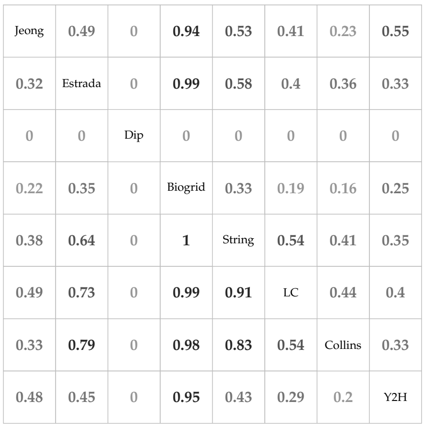
Many previous studies used a single instance of the PIN of *S. cerevisiae* to evaluate an association of lethality with newly defined indices [62,63,204]. However, there are various instances available through different protein databases and other resources, offering the opportunity to test the validity of the hypothesis on a larger sample. We therefore use several versions from different available sources: two previously used networks by Jeong et al. and Estrada et al. , three self compiled versions from different protein interaction databases and three instances from the literature. In all cases, we only consider the biggest component of the network. Detailed network statistics are provided in Table 5.1.

Name	Proteins	Interactions	Lethal proteins	Source (Version)
Jeong	1870	2277	22%	[101]
Estrada	2224	6609	26%	[63]
Dip	2131	4813	31%	[199] (01/01/15)
Biogrid	6238	223575	18%	[178] (3.2.111)
String	2032	9497	38%	[182] (9.0)
LC	1213	2556	44%	[164]
Collins	1004	8319	41%	[48]
Y2H	1647	2518	23%	[201]

Table 5.1: Network statistics of the used PINs of *S. cerevisiae*.

Remark. *For the re-analysis of [101], we rebuild the network used in the study. The raw data contain self-interactions, multiple interactions, several components and isolates, and proteins with an unknown lethality status. In order to replicate the results of Jeong et al. , we consider the version without loops and exclude proteins with unknown lethality status from later analysis. In Section 5.4, we then only use the biggest component.*

Figure 5.2 shows the fraction of overlapping proteins among the eight instances. The network derived from Biogrid includes almost all proteins from the other networks, yet the overlap is generally very low such that the networks provide a broad sample of the complete PIN of *S. cerevisiae*.



Jeong	0.49	0	0.94	0.53	0.41	0.23	0.55
0.32	Estrada	0	0.99	0.58	0.4	0.36	0.33
0	0	Dip	0	0	0	0	0
0.22	0.35	0	Biogrid	0.33	0.19	0.16	0.25
0.38	0.64	0	1	String	0.54	0.41	0.35
0.49	0.73	0	0.99	0.91	LC	0.44	0.4
0.33	0.79	0	0.98	0.83	0.54	Collins	0.33
0.48	0.45	0	0.95	0.43	0.29	0.2	Y2H

FIGURE 5.2: Fraction of overlapping proteins among the eight PIN instances of *S. cerevisiae*. Due to a differing naming convention, the overlap between the Dip instance and others could not be calculated.

The lethality status for the proteins were obtained from the *database of essential genes* [203] (v10.0).

Data for Multiple Organisms

The protein interactions of 20 organisms were obtained from the String Database (version 9.0). The database contains experimentally identified interactions from published literature as well as computationally predicted interactions. Each interaction is given a confidence score indicating the probability of an actual interaction. In contrast to Raman et al. who only used 700 as a lower bound, we construct eight networks using $\{600, 650, 700, 750, 800, 850, 900, 950\}$ as lower bounds for the interaction scores for each organism. Again, only the biggest component of each network is considered. The lethality status for the proteins were obtained from the *database of essential genes* [203] (v10.0). Summary statistics of the networks with confidence score 700 are shown in Table 5.2.

Organism	Proteins	Interactions	Lethal proteins
Abayli	2080	12498	22%
Athaliana	6726	69388	1%
Bsubtilis	2899	20404	7%
Celegans	4744	46877	2%
Dmelanogaster	5251	93660	4%
Ecoli	3686	25843	17%
Fnovicida	1230	7476	27%
Hinfluenzae	1307	8670	40%
Hpylori	1149	7773	23%
Mgenitalium	413	3354	83%
Mpulmonis	474	2955	56%
Mtuberculosis	2845	18206	9%
Paeruginosa	3658	20983	7%
saNCTC	1662	9009	19%
sasaN31	1576	8800	18%
Scerevisiae	5461	105893	20%
Spneumoniae	1425	8387	6%
Ssanguinis	1366	7846	15%
Styphimurium	3249	20635	5%
Vcholerae	2611	15398	16%

TABLE 5.2: Statistics for PINs of 20 organisms with confidence score 700.

Centrality Indices

A vast amount of indices was applied in the context of the centrality-lethality hypothesis. Since our intention is to draw general conclusions and not to find the best performing index, we just focus on a subset of indices. Our set contains degree, betweenness, closeness and eigenvector centrality as well as information centrality, subgraph centrality and bipartivity. The last two have been reported to be the best performing indices on the Estrada instance of *S. cerevisiae* [62, 63]. Additionally we apply a new measure, the *hyperbolic index*, introduced in the following.

Let u be a vertex. On the subgraph induced by $N[u]$, we sum up all closed walks of even length and weight them decreasingly with their length and additionally with its density, i.e. the local *clustering coefficient*

$$ccoef(u) = \frac{2\left|\left\{v_i, v_j\right\} \in E : v_i, v_j \in N(u)\right|}{d_u(d_u - 1)},$$

which describes the density of the subgraph induced by $N(u)$. Thus, c_{hyp} is defined as

$$c_{hyp}(u) = ccoef(u)\left[\sum_{v \in N[u]} \sum_{k=0}^{\infty} \frac{(A^{[u]})^{2k}_{vv}}{(2k)!}\right],$$

where $A^{[u]}$ is the adjacency matrix of the subgraph induced by $N[u]$. This formula can be simplified in the following way:

$$\begin{aligned} c_{hyp}(u) &= ccoef(u)\left[\sum_{v\in N[u]}\sum_{k=0}^{\infty}\frac{(A^{[u]})^{2k}_{vv}}{(2k)!}\right] \\ &= ccoef(u)\left[\sum_{v\in N[u]}\frac{(e^{A^{[u]}})_{vv}+(e^{-A^{[u]}})_{vv}}{2}\right] \\ &= ccoef(u)\left[\sum_{v\in N[u]}\cosh(A^{[u]})_{vv}\right] \end{aligned}$$

The hyperbolic index is mainly used for illustrative purposes and obtaining surprising results.

Evaluation Methods

Many different evaluation methods have been used to ascertain an association between centrality and lethality, where the choice is also a matter of interpreting the hypothesis.

The most common approach in the literature is to rank the proteins according to a centrality index and calculate the fraction of lethal proteins in the top x ranked proteins and compare the value with random sampling (henceforth referred to as top rank approach) [62,63,123,124,125,194,204]. Although a high fraction within the top ranked proteins suggests that lethal proteins occupy central positions, there might still be plenty of low ranked lethal proteins which are neglected. Additionally, we run the risk of cherry picking results by fixing x at appealing values. Yet this approach is sufficient if we are only concerned with strategies to sample lethal proteins. Other approaches like comparing mean values of indices [93] or bootstrap sampling [162] fall in the same class.

A second category comprises methods used in binary classification problems, i.e. centrality indices are treated as predictors for the lethality status of a protein. Metrics in this field rely on *contingency tables*, as depicted in Table 5.3. These tables are constructed by setting a threshold for accurate predictions and determining the values of each cell. Used metrics relying on a contingency table are mainly *accuracy* [123,194] and the *receiver operating characteristic* (ROC) [93,183,184]. Accuracy is defined as

		actually	
		lethal	non-lethal
predicted	lethal	true positive	false positive
	non-lethal	false negative	true negative

TABLE 5.3: Structure of a contingency table.

$$ACC = \frac{TP+TN}{TP+FP+FN+TN}\,.$$

Since accuracy is only determined for one chosen threshold, it suffers from similar drawbacks as the top rank approach [160]. For illustrative purposes, we apply *Matthew's correlation coefficient* as an alternative to accuracy [134]. It is defined as

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \,.$$

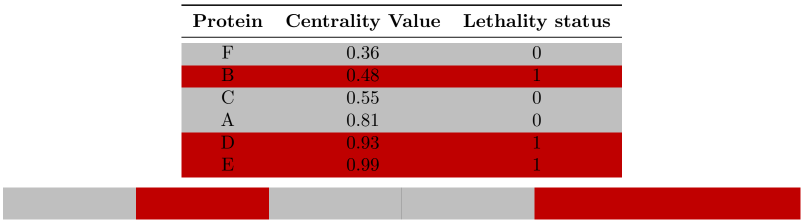
In contrast to accuracy and Matthew's correlation coefficient, ROC curves avoid the supposed subjectivity in the threshold selection process by considering all possible cutoffs. The curve is created by plotting the true-positive rate (TPR, also known as recall) against the false-positive rate (FPR, also defined as 1-specificity), which are defined as

$$FPR = \frac{FP}{FP+TN} \quad \text{and} \quad TPR = \frac{TP}{TP+FN} \,.$$

To facilitate the interpretation of the curve, the area under the curve (AUC) is used as a summary statistic. AUC values are bounded between 0 and 1, where a model that yields a value of 0.5 is no better than randomness and higher (lower) scores indicate a better (worse) prediction than one could expect by chance. It is argued that ROC curves are less informative and tend to overestimate the performance if the class distributions are highly skewed [53]. Since the fraction of lethal proteins in each considered PIN varies strongly and mostly is quite small, we here use *Precision Recall* (PR), as suggested in [53], which is mainly used in information retrieval [130, 161]. In PR space, Recall is plotted against Precision, which are defined as

$$Recall = \frac{TP}{TP+FN}, \qquad Precision = \frac{TP}{TP+FP} \,.$$

Again, AUC can be used as a summary statistic with similar interpretation as in ROC space, the difference being that random performance depends on the total number of lethal proteins. We also make use of *separation plots* which is a purely visual method to not solely rely on single value statistics [91]. An example of its functionality is given in Figure 5.3.



Protein	Centrality Value	Lethality status
F	0.36	0
B	0.48	1
C	0.55	0
A	0.81	0
D	0.93	1
E	0.99	1

FIGURE 5.3: Example for separation plots. Proteins are ordered according to centrality scores and color coded according to their lethality status (lethal=1, non-lethal=0). The ranking is visualized as a colored rectangle, indicating the position of the lethal proteins from left to right.

5.3 Re-Analysis of Original Study

The scale-free property

The deductive reasoning of Jeong et al. leading to the results are based on frequently invoked paradigms about *real-world* networks. Many man-made networks, or those that occur naturally in *complex systems* are said to be *scale-free* [9]. Although their is no precise definition of what 'scale-freeness' is [22,23], it is commonly associated with a *power-law* degree distribution [9]. That is, the fraction $p(x)$ of vertices in a network having a degree x is approximately given by

$$p(x) \propto x^{-\alpha} ,$$

where typically $2 < \alpha < 3$ can be observed. Empirical evidence, however, suggests that the scaling is better described by $p(x) \propto (x + x_{min})^{-\alpha}$, since networks do not follow the power law for small x. Additionally, it is observed that the scaling for large x is better described with an exponential function after a cutoff x_c. Thus, we can write

$$P(x) \propto (x + x_{min})^{-\alpha} \exp(-\frac{x + x_{min}}{x_c}).$$

The analysis of the PIN of *S. cerevisiae* by Jeong et al. naturally starts with the claim that its degree distribution follows a power law with an exponential cutoff. This was shown by a partial straight line fit of degree frequencies on a doubly logarithmic scale with parameters $x_{min} = 1$, $x_c = 20$ and $\alpha = 2.4$ (cf. Figure 5.4 (left)). Beforehand, the degree distribution was condensed into six data bins to calculate relative frequencies.

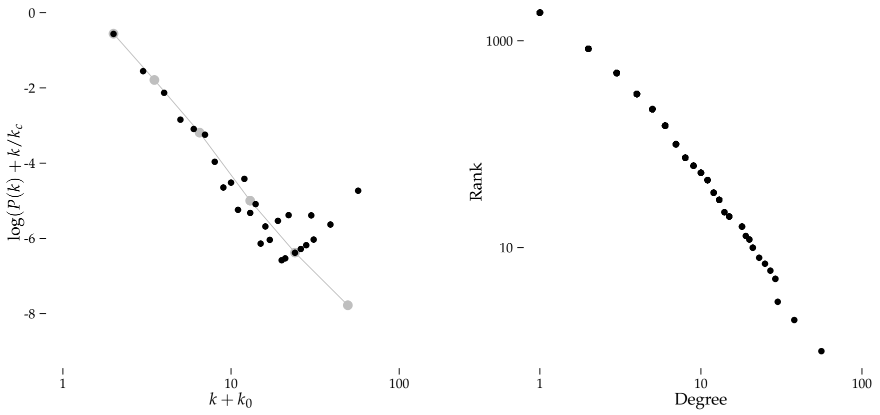


FIGURE 5.4: (left) Chosen illustration for degree distribution by Jeong et al. (cf. Figure 1b in [101]). Black points show actual frequencies without binning (gray line). (right) size-rank plot of the degree distribution on a doubly logarithmic scale.

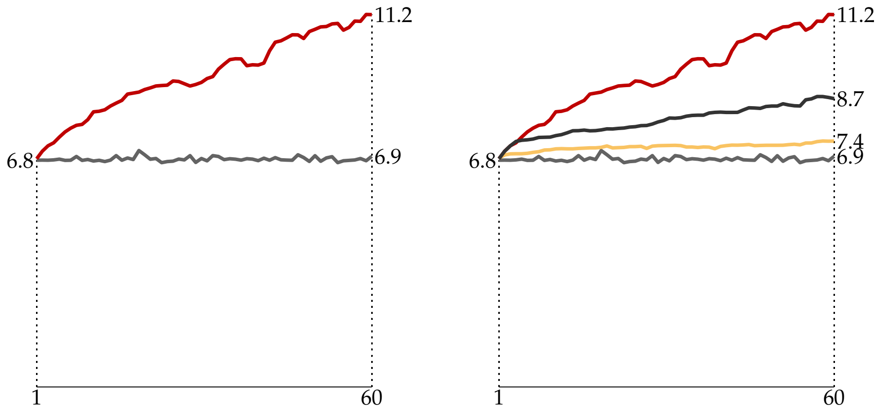
A more robust way for a visual assessment of the degree distribution is the size-rank plot as suggested by Li et al. [122]. A straight line fit on a doubly logarithmic scale can be a better indicator for a power law degree distribution. Figure 5.4 shows the size-rank plot of the degree distribution on the right. Its 'straightness' actually suggests a power law distribution.

Although log-log plots are an accepted justification for scale-freeness, it is certainly not a statistically sound argument. Especially since various ways of data binning can be used to overcome low frequency regions in the dataset. Size and boundary values for data bins are left to the researcher and yield a high level of arbitrariness, altering the results drastically [122]. Granted, the analysis was done at the beginning of the 'power law age' and more sophisticated methods were not yet at hand.

With the development of new tools, many claims about apparent scale-free networks could be refuted or are at least highly questionable, e.g. with a method introduced by Clauset et al. [47]. First, the parameters α and x_{min} are estimated by a *method of maximum likelihood*. Then, the goodness-of-fit is determined with the *Kolmogorov-Smirnov statistic*(KS) and a p-value is calculated as the fraction of the KS for synthetic data whose value exceed the KS for the real data. The power law distribution can be ruled out if the p-value is sufficiently small $(p \leq 0.1)$. Performing this test, we obtain $\alpha = 3.02$, $x_{min} = 5$ and $p = 0.44$. That is, we can not rule out a power law distribution, neither visually nor statistically. However, we also do not have enough evidence for a power law, since we i.a. did not test for other distribution.

Fragility under targeted attack

A prominent inferred feature of scale-free networks is its ‘robust yet fragile’ nature, prompted by Albert et al. [2]. This property arises from the fact that the removal of random nodes does not alter the network topology (*error tolerance*), yet the removal of highly connected so called *hubs* disrupts the path structure significantly leading to an increase in the average shortest path length (*attack vulnerability*). The authors go as far as to say that scale-free networks are the *only* class of networks that display this error tolerance. Interestingly they already state that “such decreased attack survivability is useful for drug design” [2, p. 381]. Jeong et al. use this reasoning on the PIN of *S. cerevisiae* and show that the network diameter increases *rapidly* when highly connected proteins are removed (cf. Figure 5.5a). Therefore, they deduce that the hub proteins are most likely lethal. Yet, if we target highly connected lethal and nonlethal proteins separately, we observe in Figure 5.5b that the network diameter increases more if nonlethal proteins are removed. That is, the network is actually more fragile to targeted attacks against highly connected nonlethal proteins.



(A) Illustration by Jeong et al.

(B) Targeting lethal/nonlethal proteins.

FIGURE 5.5: Attack tolerance of the network. Increase of the mean shortest path distance when up to 60 proteins ($\sim$ 4%) are removed randomly (grey), high degree proteins are targeted (red) and high degree lethal (orange) and nonlethal (black) are targeted.

Highly connected, more likely to be lethal

From the asserted error tolerance, the authors form an analogy for the biological process and hypothesize that “[...] on average less connected proteins should prove to be less essential than highly connected ones” [101, p. 1]. *On average* and *less essential* leave an ample scope on how to interpret and analyze this hypothesis. The authors claim a correlation of 0.75 between degree and lethality and state that

> *"although proteins with five or fewer links constitute about* 93% *of the total number of proteins we find that only about* 21% *of them are essential. By contrast, only some* 0.7% *of the yeast proteins with known phenotypic profiles have more than* 15 *links, but single deletion of* 62% *or so of these proves lethal"* [101, p. 1].

It must be noted that these results were obtained by binning all proteins with a degree greater than 15 "to produce a more reliable result"(Supplementary material of [101]). The results itself sound convincing and tempting enough to constitute a strong association between lethality and degree. However, as stated before, the binning process can drastically change the nature of the results. As can be seen on the left in Figure 5.6, the correlation strongly depends on how we bin high degree proteins. The same holds true for Matthew's correlation coefficient, which is shown in Figure 5.6 on the right.

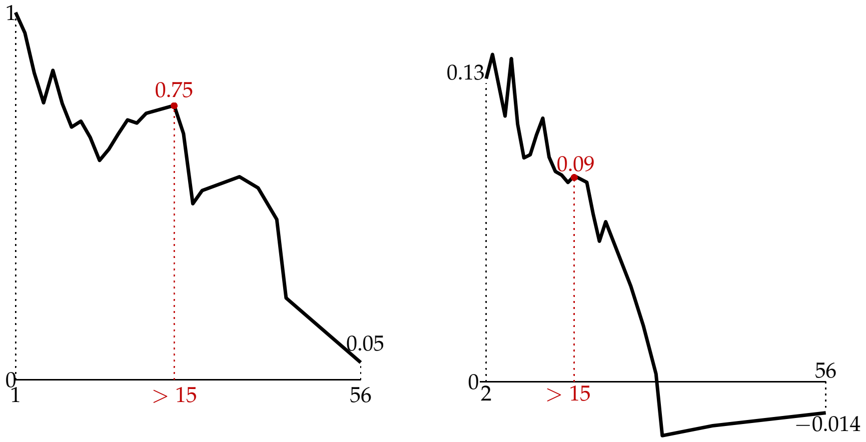
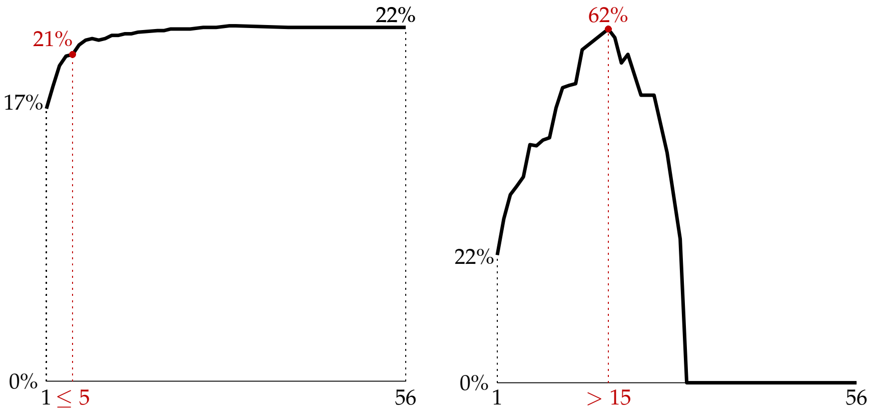


FIGURE 5.6: (left) Correlation between lethality and degree when different data binning for high degree proteins are used. Red value indicates reported value in [101]. (right) Matthew's correlation coefficient for the same data binning procedure. Red value indicates the cutoff used in the original work.

The same issue can be observed in the second test. The share of 21% of lethal proteins is actually quite close the total percentage of 22% in the total dataset (cf. Figure 5.7a) and the reported 62% are the highest possible portion of lethal proteins with *high* degree (cf. Figure 5.7b). Changing the upper bound of 15 to higher values would not yield any appealing results.



(A) % lethal ≤ degree

(B) % lethal > degree

FIGURE 5.7: Percentage of lethal proteins in (A) low degree and (B) high degree proteins with different data binnings for low/high degree.

Statistical Testing

After the last subsection, we constitute that the impression of an effect is mainly due to the chosen data binning and that other data contracting techniques would have allowed for different outcomes. The question is, if there is any plausible statistical evidence for the claimed effect. The possibilities are manifold to (a) interpret the statement 'the higher the degree the more likely to be lethal' and (b) choose an appropriate statistical test to analyze a derived hypothesis. For many statistical tests, we have to assume that the data is normally distributed. The extreme skewness of the dependent variable (22% lethal vs. 78% non-lethal) and the independent variable (skewed degree distribution) thus prevents any test were this assumption has to be made. Also, we have to assume that the lethality status of the proteins are independent. This assumption might be reasonable from a statistical but not from a biological point of view. From now, we tacitly assume independence and try to asses the centrality-lethality hypothesis with statistical tools.

We first examine if there is any significant difference between the degree distribution of lethal and non-lethal proteins with a two-sided *Wilcoxon rank-sum test*. The resulting test statistic is significant ($p = 8.984^{-10}$) so that we can reject the hypothesis of an equal distribution. Figure 5.8 shows the two distributions separately.

The Figure displays the skewness of the data quite well. The great majority of both lethal and non-lethal proteins have a degree less or equal five. As Jeong et al. already pointed out, these proteins comprise around 93% of the whole dataset. Figure 5.9 shows the degree distributions with an alternative representations.

For proteins with a degree less than five we can actually see, that the fraction of lethal proteins gradually increases. However, for higher degree proteins,

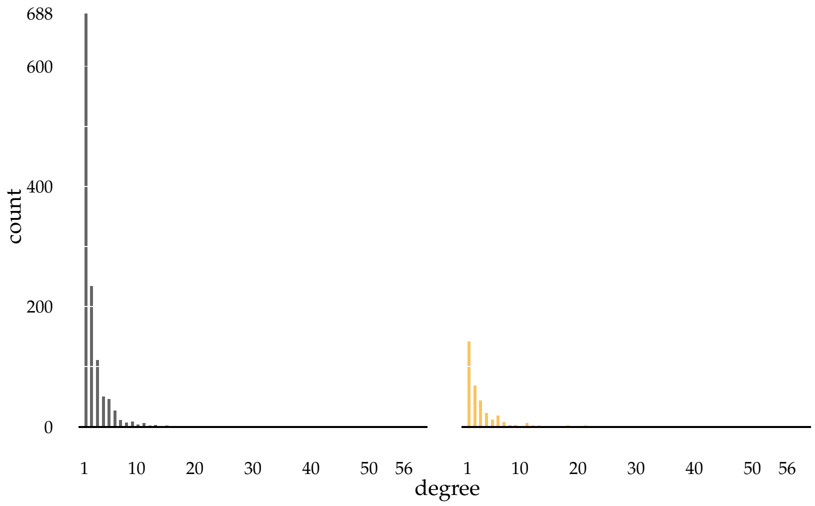


FIGURE 5.8: Degree frequencies of non-lethal (right) and lethal (left) proteins.

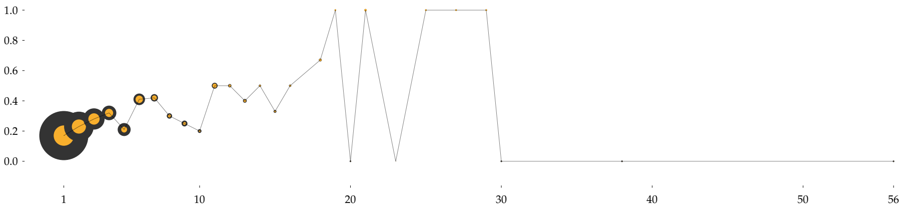


FIGURE 5.9: Alternative representation of the degree distribution together with the fraction of lethal proteins. The area of the circles describe the number of proteins for given degrees. The orange circles are the number of lethal proteins per degree.

we do not observe any striking pattern. Since the low degree proteins are the great majority in the dataset, we can expect that any statistical analysis is biased towards them. Nevertheless, we assess the posed claim 'the higher the degree, the more likely to be lethal' by means of a regression. We can model the lethality status of a protein as a *binary response variable* y where $y_i = 1$ if protein i is lethal and $y_i = 0$ otherwise. The degree is used as a continuous *explanatory variable* d. With a *logistic regression*, we can estimate the probability of y based on the variable d, i.e. $P(y_i|d = d_i)$. The first column in Table 5.4 shows the result of the regression for the whole dataset.

	$d_i \geq 1$	$d_i \geq 2$	$d_i \geq 3$	$d_i \geq 4$	$d_i \geq 5$
(Intercept)	-1.49^{***}	-1.18^{***}	-0.99^{***}	-0.92^{***}	-0.92^{***}
	(0.08)	(0.12)	(0.16)	(0.22)	(0.27)
degree	0.090^{***}	0.056^{**}	0.040	0.036	0.036
	(0.019)	(0.020)	(0.021)	(0.023)	(0.025)
AIC	1623.24	853.34	525.52	340.37	249.40
BIC	1633.93	862.50	533.59	347.52	255.90
Log Likelihood	-809.62	-424.67	-260.76	-168.19	-122.70
Deviance	1619.24	849.34	521.52	336.37	245.40
Num. obs.	1552	722	419	264	191

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

TABLE 5.4: Results for logistic regression with degree as explanatory and lethality status as response variable. The name of the models indicate which proteins were included.

The log-odds for degree are 0.09 and significant $(p < 0.001)$, translating to an increase in probability of 1.09 that a protein is lethal when degree is incremented by 1. This actually suggests that the higher the degree, the higher the probability for a protein to be lethal. However, the increase in probability is very low and not as convincing the results of Jeong et al. would suggest. Additionally, if we gradually truncate the data by removing low degree nodes from the dataset, we observe in columns two to five of Table 5.4 that the effect gets weaker and even non-significant. As we argued before, this hints at the mentioned statistical bias. The regression is mainly driven by the first 5 data points signifying an inflated effect.

Binary Classification

Besides statistical testing we can also treat the hypothesis as a classification problem and use ROC and PR to evaluate the performance of degree to detect lethal proteins. Figure 5.10 shows the ROC and PR curves for the classification. With an AUC= 0.6 for ROC and AUC= 0.31 for PR (random performance 0.22), degree is not significantly better than a random classifier. The separation plot additionally reveals how scattered the lethal proteins in ranking induced by degree are.

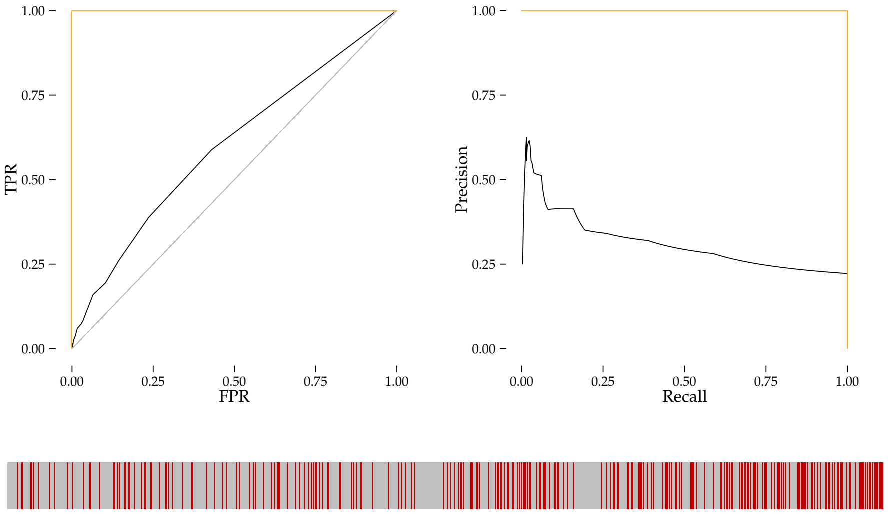


FIGURE 5.10: (left) ROC and (right) PR curve of the performance of degree to detect lethal proteins. The gray line indicates random performance and the orange lines the respective best performance possible. The separation plot shows the position of lethal proteins (red) in the degree ranking.

5.4 Results for S.cerevisiae

Degree Distribution and Attack Tolerance

Figure 5.11 show the degree distribution as a rank sized plot for all eight instances. It can be seen that the PINs are quite distinct from each other. The distributions are mostly far from being 'a straight line', i.e. scale-free.

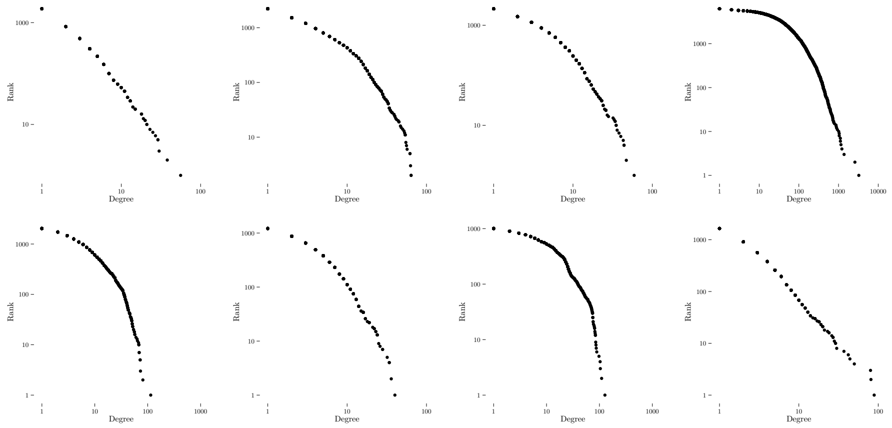


Figure 5.11: Rank sized plot of degree distribution on a doubly logarithmic scale for eight PINs of *S. cerevisiae*. From top left to bottom right the instances are: Jeong, Estrada, Dip, Biogrid, String, LC, Collins and Y2H.

Table 5.5 additionally shows the parameters for the fitted power law degree distributions as well as the Kolmogorov-Smirnov statistic with the associated p-value. The results shows that we can rule out a power law distribution only for the Collins data set. As before, the higher p-values do not indicate that the other datasets contain a power law degree distribution, but only tell that we can not rule it out.

	x_{min}	α	KS	p-value
Jeong	5	3.03	0.033	0.44
Estrada	14	3.26	0.047	0.22
Dip	12	3.75	0.030	0.71
Biogrid	345	3.94	0.035	0.73
String	34	4.81	0.046	0.83
LC	11	3.72	0.044	0.37
Collins	20	2.86	0.059	0.02
Y2H	7	2.78	0.021	0.91

Table 5.5: Statistical test for power law in PINs of *S. cerevisiae*.

Figure 5.12 shows the results of the attack tolerance test conducted in Section 5.3. Similar to the degree distributions, the results vary among the datasets. Only the LC dataset shows the anticipated behaviour, i.e. the removal of lethal

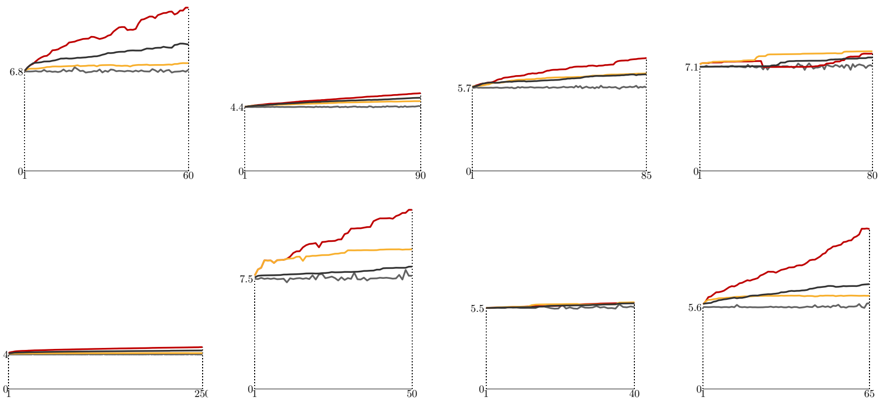


FIGURE 5.12: Attack tolerance test for eight PINs of *S. cerevisiae* as done in Figure 5.5. From top left to bottom right the instances are: Jeong, Estrada, Dip, Biogrid, String, LC, Collins and Y2H.

proteins with a high degree disrupts the network more than the non-lethal proteins.

AUC and Separationplots

The predictive powers of the used centrality indices are summarized in Figure 5.13. We can see that the overall performance differs between the eight instances and generally does not deviate too much from randomness. Differ-

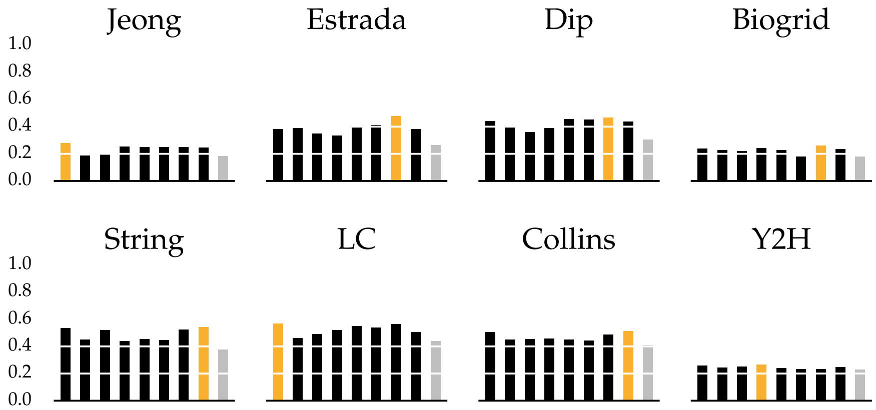


FIGURE 5.13: AUC values for degree, eigenvector centrality, closeness, betweenness, subgraph centrality, bipartivity, hyperbolic index, information centrality for eight instances of the PIN of *S. cerevisiae*. Orange bars indicate best performing index and gray bars random classifiers.

ent indices perform best on each network with the hyperbolic index outperforming others on three instances. However, the AUC values are quite similar for all indices on each network. To determine the similarity in the rankings of lethal proteins, we calculated the average correlation of indices on each network using Kendall's τ. The scores, shown in Table 5.6, indicate only a weak correlation on all networks. The separation plots of the best performing indices per network are shown in Figure 5.14. Observe that there is a dense region of lethal proteins at the top of the ranking, however, most lethal proteins seem to be scattered across the ranking.

Jeong	Estrada	Dip	Biogrid
0.45 (0.18)	0.59 (0.16)	0.52 (0.15)	0.71 (0.14)
String	LC	Collins	Y2H
0.48 (0.23)	0.42 (0.20)	0.47 (0.25)	0.51 (0.19)

TABLE 5.6: Mean pairwise Kendall's τ of centrality rankings of lethal proteins. Standard deviation in brackets.

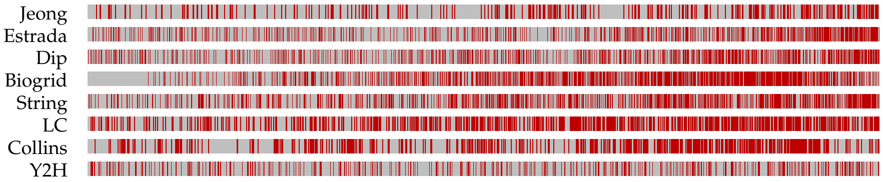


FIGURE 5.14: Separation plots of the best performing index per network.

5.5 Results for Multiple Organisms

Figure 5.15 summarizes the performances for the 20 networks with 700 as interaction threshold. In contrast to the results reported in [162], we see that closeness actually performs better than degree and betweenness in most of the networks. It must be noted though, that Raman et al. considered disconnected components as well. However, our reexamination of their results showed, that they used the ill-defined version of closeness [170]. Similar to the results of *S. cerevisiae*, different indices perform best and generally the performance varies among organisms. Note that eigenvector centrality outperforms the more sophisticated walk-based measures subgraph centrality and bipartivity on most of the networks.

Figure 5.16 illustrates that prediction accuracy depends heavily on the chosen interaction threshold. That is, depending on how we construct the networks, different indices yield the best performance.

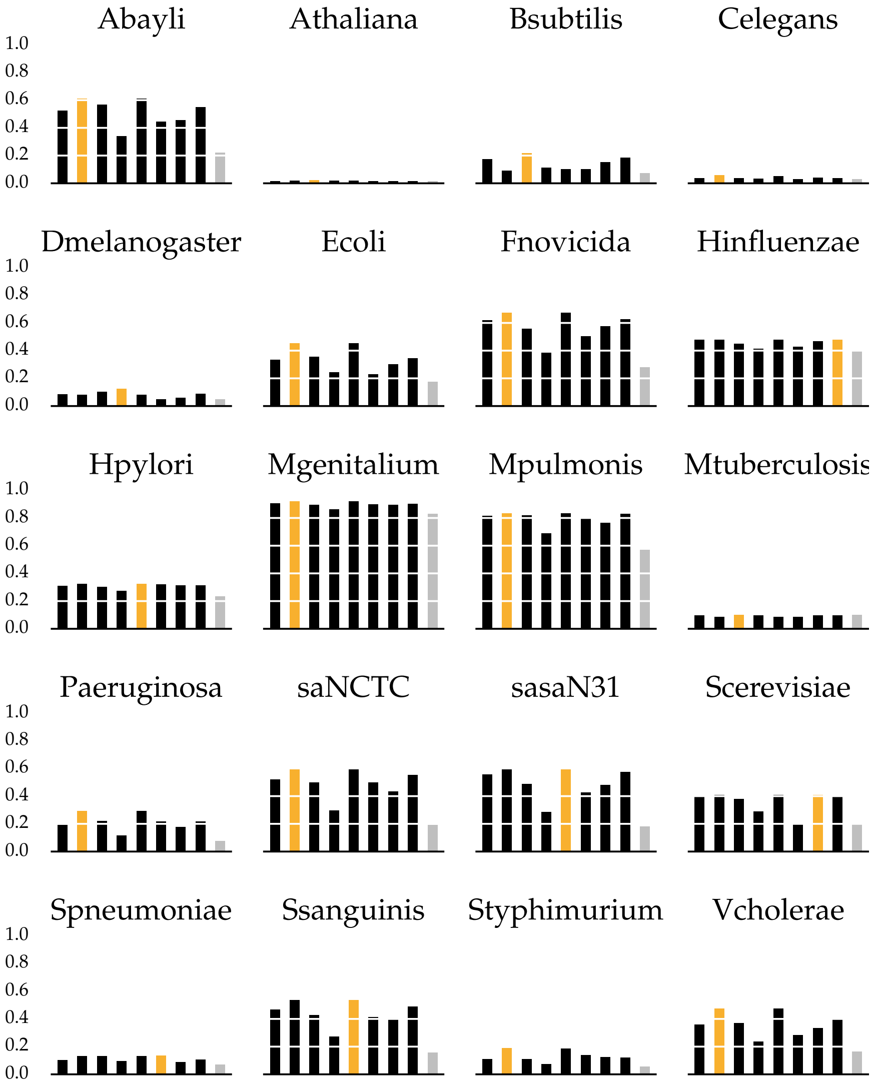


FIGURE 5.15: AUC values for degree, eigenvector centrality, closeness, betweenness, subgraph centrality, bipartivity, hyperbolic index, information centrality on 20 different organisms with interaction threshold 700. Orange bars indicate best performing index and gray bars random classifiers.

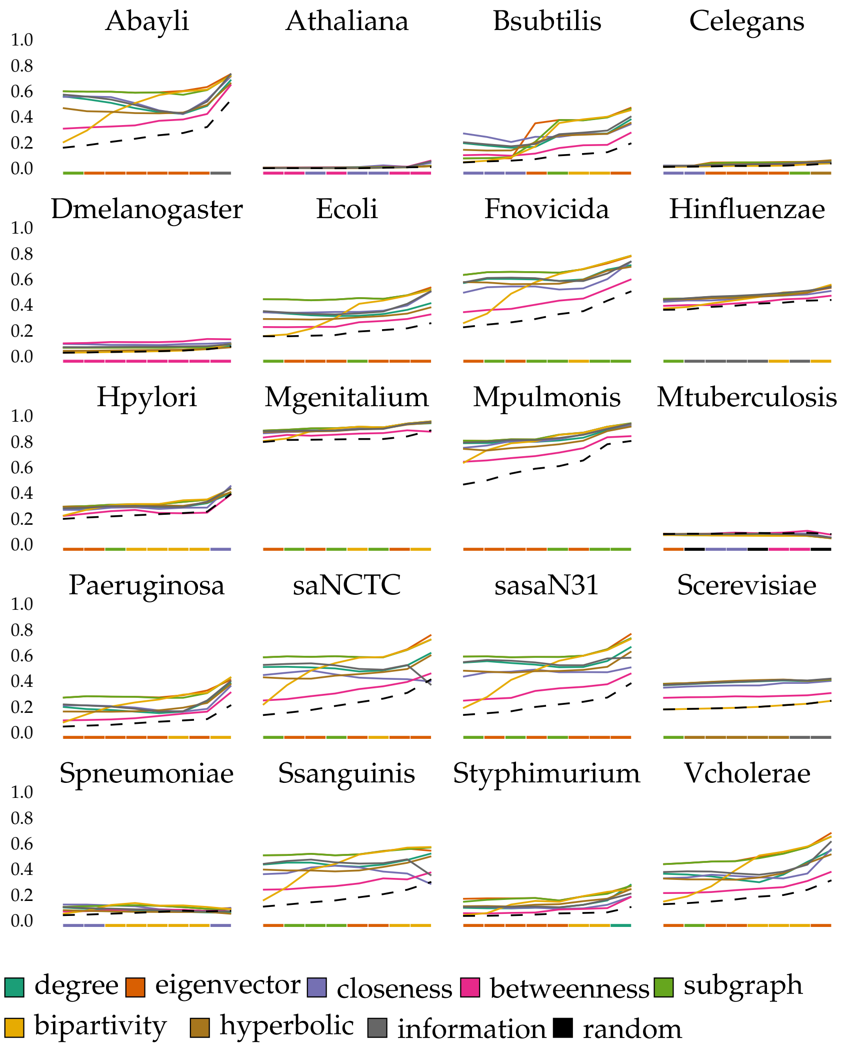


FIGURE 5.16: AUC values when threshold for interaction data is altered from 600 to 950. Colored rectangles indicate best performing index.

5.6 Summary

In this chapter, we assessed the general plausibility of the proclaimed centrality effect in PINs of *S. cerevisiae* and other organisms. Section 5.3 has shown, that the original reasoning why high degree proteins are more likely to be lethal does not withstand closer examinations. The chain of arguments is based on analogies from *known facts* about networks applied to protein interaction which are themselves questioned by the literature [22, 23, 47, 122]. Even if we accept a scale-free and 'robust yet fragile' nature of the network, it is still more vulnerable to targeted attacks on non-lethal proteins than lethal ones, making the initial argument unfounded. Apart from that, our re-analysis partially confirmed the results in [101], yet they are mostly due to chosen data bins, i.e. reported data points are the ones that make the case for a centrality effect most strongly. Our statistical analysis revealed that the seeming effect stems from the proteins with degree less equal 5. In this subset of the data, we can indeed observe an increase in lethal proteins, yet the rest of the dataset does not allow for generalization. The skeweness of the data is the main reason why statistical tests find an overall positive effect.

In Section 5.4, we reviewed elicited follow up work, which was mainly concerned with strengthening the association between lethality and centrality for PINs of *S. cerevisiae*. In contrast to others, we used a larger sample of PINs and discussed the reliability of results obtained by centrality indices. First, we have seen that the network topologies differ strongly among the considered instances such that no general conclusion about its nature, e.g. scale-free or not, can be drawn. Further, no index has a consistent high association with lethality across networks. The performance is generally comparable among indices, yet the best performing index depends on the considered network. Since the induced rankings of lethal proteins are additionally only weakly correlated, we can conclude that lethal proteins do not share common structural features, but instead hold diverse positions within the network. Additionally, the separation plots of the best performing indices showed how scattered lethal proteins are across the rankings.

Transferring results to a greater variety of organisms has shown to be futile in Section 5.5. Reexamination of [162] showed that the results of Raman et al. are skewed for two main reasons: inappropriate use of a version closeness that is ill-defined on disconnected networks, and restriction to a single threshold for interactions. Our analyses showed yet again that the performance of centrality indices are inconsistent with differing best performing index on each organism. The robustness of results was tested by altering the threshold of considered interactions for the PINs. Results depended heavily on how the networks are constructed, such that the best performing index varied even within organisms.

One could certainly argue that our considered set of indices was not suitable to draw any considerable conclusions. So far, however, there is no justification for the appropriateness of any centrality index in the first place. Several have been offered [97, 206] but neither of them attempts to explain the underlying processes that make proteins in certain network positions more likely to be lethal than others. In the absence of a substantive theory that justifies a corresponding centrality concept, choosing an existing or crafting a new index is an exercise in data fitting. Some measure will necessarily turn out to fit best, yet they do not offer any substantial explanations. We demonstrated these issues with the newly defined hyperbolic index. Its definition lacks any biological appeal and outperformed other measures on three instances of *S. cerevisiae*.

Without a testable theoretical explanation at hand, hypothesized centrality effects need to be tested on a large number of networks to gain confidence in the results. *Data-driven hypothesis* by means of a single instance can be a useful way to proceed (cf. [105]), but conclusions have to be drawn with care and are only informative if underlying data is representative of a sufficient wide class of cases. For the presented problem, results from a single PIN of an organism can not be generalized for several reasons; Topologies differ, the set of included proteins have a small overlap and interactions are not unambiguous. Therefore, we have to rely on a bigger sample of networks to obtain convincing arguments for or against conjectured effects.

Although we gathered evidence against the plausibility of the centrality-lethality hypothesis, we can not refute its validity for the aforementioned reasons. The inconsistent findings suggest that either the effects differ, or the data are of varying accuracy. We return to the centrality-lethality hypothesis in Chapter 10 and examine it with newly developed methods.

Part II

Network Centrality

based on Positions

CHAPTER 6

Centrality, Measurement and Positions

"Measurement juxtaposes science and philosophy, because only through measurement does science approach real life."

– Krebs, 1987

6.1 INTRODUCTION

The first part of this thesis was concerned with the current conception of centrality. We have seen that there is a lack of conceptual clarity and little knowledge about theoretical fundamentals of indices. Especially in recent years, applications of centrality drifted towards data mining tasks. That is, indices are used to find patterns in data that potentially can explain certain empirically observed phenomena. In Chapter 5 we thoroughly investigated one such application, the centrality-lethality hypothesis, and showed that this data-driven approach to network centrality has its drawbacks and limitations.

In this first chapter of the second part, we initiate to view centrality as a proper procedure of measurement. We present some common tools offered by theories about measurement and discuss how they can be adapted for network centrality. We also briefly introduce a novel positional approach to network analysis recently proposed by Brandes [36]. Comparing position under varying premises will offer new theoretical insight for centrality indices, which in turn suggests a new characterization of centrality concepts. The mathematical evaluation of these considerations in this chapter is done in Chapter 7.

6.2 MEASUREMENT THEORY

What we have learned thus far is that there is not *one* concept of centrality, but different competing *operational definitions*, i.e. indices translating the concept into measurement of some kind. Evidently, being central in a network is ambiguous with many different interpretations. This is not necessarily bad to begin with if we compare the concept to measurement procedures in general.

Considering two individuals A and B which can be measured according to different physiological quantities, e.g. height and weight. Person A may be heavier than B but B might well be taller. There is no intrinsic physiological ordering of A and B, yet they are comparable by means of some observable empirical structure, depending on *what* we want to measure. As trivial this example sounds, as tedious was laying the basics of measurement ultimately leading to the understanding we have today [57,58]. According to the *representational theory of measurement* (RTM), measurement may be regarded as

> *" [. . .] the construction of homomorphisms (scales) from empirical relational structures of interest into numerical relational structures that are useful."* [114, p. 9]

In the received interpretation of RTM, it is assumed that we are given an *empirical relational structure* (ERS) together with an *numerical relational structure* (NRS). The ERS is defined as a tuple $(\mathcal{X}, R, \circ)$, where $\mathcal{X}$ is a set of objects, R is a set of relations among the objects in $\mathcal{X}$ and there exists an order relation $\succcurlyeq$ among those. A closed operation (mostly a concatenation) on $\mathcal{X}$ is given by $\circ$. The NRS is defined as the tuple $(\mathbb{R}, \geq, +)$, with the conventional definition of $+$ and $\geq$ on the real numbers. The base set $\mathcal{X}$ can be measured if we can formulate a *representation theorem*.

Theorem 6.1 (Representation Theorem)**.** *Let $(\mathcal{X}, R, \circ)$ be an ERS with an order relation $\succcurlyeq \in R$ and $(\mathbb{R}, \geq, +)$ a NRS. A mapping $\phi : \mathcal{X} \to \mathbb{R}$ is called a <u>representation</u> of $\mathcal{X}$ in $\mathbb{R}$ if it is a homomorphism, i.e. for all $x, y \in \mathcal{X}$*

$$x \succcurlyeq y \implies \phi(x) \geq \phi(y)$$

holds.

Note that the existence of such a homomorphism is i.a. guaranteed if $\succcurlyeq$ is a weak order [33]. The exact characterization of the type of scale a measurement procedure yields is then given by a *uniqueness theorem*.

Definition 6.2 (Uniqueness)**.** *Let $(\mathcal{X}, R, \circ)$ be an ERS with an order relation $\succcurlyeq \in R$, $(\mathbb{R}, \geq, +)$ a NRS and $\phi : \mathcal{X} \to \mathbb{R}$ a homomorphism. A transformation $\phi \mapsto \phi'$ is <u>permissible</u> if and only if ϕ and ϕ' are both homomorphisms into the same NRS.*

A distinction is commonly made between nominal, ordinal, interval and ratio scales [181]. Representation and uniqueness are two of the key concepts for RTM. Other more philosophical concerns deal with *meaningfulness* [141], e.g. 'which assertions about measurement make sense?', and *validity* [1], e.g. 'does a homomorphism measure what it is supposed to measure?'. RTM thus provides a well-founded theoretical basis for formulating hypotheses and performing tests under the premise of measurement.

Although it is well established, the concept is occasionally debated and extended [98,145]. One such extension is *conjoint measurement* (CM), where a concatenation operation on $\mathcal{X}$ is not defined and objects are evaluated according to product sets [33]. That is, the base set of the ERS is given by a n-ary Cartesian product $\mathcal{X} = X_1 \times X_2 \times \cdots \times X_n$ and $\succcurlyeq$ is a binary relation on this product set. Similar to RTM, the intention is to build a numerical representations of $\succcurlyeq$ and study its uniqueness. A family of representations of CM are *additive value functions* defined as

$$x \succcurlyeq y \iff \sum_{i=1}^{n} v_i(x_i) \geq \sum_{i=1}^{n} v_i(y_i)\,, \tag{6.1}$$

where $x, y \in \mathcal{X}$ and $v_i : X_i \to \mathbb{R}$ are *partial value functions*. Many situations in decision theory involve the study of binary relations on product sets, e.g. in *multiple criteria decision making* [202] and *decision under uncertainty* [71].

We here consider a small decision problem taken from Hammond et al. to illustrate tools of CM for later references [95].

PROBLEM. *A consultant is faced with the task to rent new office space. There are five locations to choose from, all meeting a number of requirements. For the final decision, he compares the location according to five distinct characteristics (or attributes): commute time (X_1, in minutes), ease of access (X_2, percentage of clients in the close area), level of service offered (X_3, scale with three levels:* 3*(everything available),* 2 *(telephone and fax) and* 1 *(no facilities)), size of office (X_4, in square feet) and monthly cost (X_5, in dollars). The preferences for each attribute are independent from others and well-defined. The consultant prefers lower values for X_1 and X_5 and higher values for X_2, X_3 and X_4. His task is now to find the best location among the five alternatives by evaluating them with the given values on the five attributes in Table 6.1.*

	X_1	X_2	X_3	X_4	X_5
a	45	50	3	800	1850
b	25	80	2	700	1700
c	20	70	1	500	1500
d	25	85	3	950	1900
e	30	75	1	700	1750

TABLE 6.1: Evaluation of the five locations on the five considered attributes.

For simplicity, we assume that X_1 and X_5 are scaled such that higher values are preferred. Further, we can describe Table 6.1 as a two dimensional variable $x \in \mathbb{R}^{5,5}$, where x_{ij} is the value of attribute j for alternative i.

Before applying any advanced tools of CM, we can compare the alternatives by a concept of *dominance* [33].

Definition 6.3. *Let $x \in \mathbb{R}^{n,m}$ describe a decision problem with n alternatives and m attributes equipped with a binary relation $\geq$. Further, let u and v be two alternatives. If $x_{uj} \geq x_{vj}$ holds for all $j \in \{1, \ldots, m\}$, then u dominates v denoted by $u \succcurlyeq v$.*

It is pretty clear that dominated alternatives are no option for a final choice in any decision problem and can thus be omitted. Generally, only a small fraction of alternatives will meet the condition of Definition 6.3, e.g. the only case of dominance in our example is $b \succcurlyeq e$ eliminating e. To further narrow down the choices we can make use of the *even swaps* technique [104]. The goal of even swaps is to create pseudo dominated alternatives by assessing trade offs, i.e. what gain in a attribute i can compensate for a loss in j. Say, for alternative d, we would forgo 100 square meters if it reduces the cost by 100 Dollar each month. Alternative d dominates a under this circumstances and a can thus be eliminated. By repeatedly applying this strategy we will ultimately end up with a single alternative as our preferred choice.

Although the even swaps technique is simple and seems obvious, it comes with several disadvantages. First, trade offs are defined from a purely subjective perspective and are not further substantiated. Second, the technique becomes infeasible for decision problems with a multitude of alternatives and attributes and the appearance of a new alternative or attribute would require to reset the process again. Last, the technique only provides us with the most preferred option and does not yield any form of preference ranking.

Therefore, we generally rely on the additive value function model in order to solve decision problems as above. By defining reasonable partial value functions, we can quantify all choices and obtain a preference ranking that can be used for the final decision. An important property for well defined functions is that the dominance relations should be preserved. That is, a dominated alternative should not become more preferable than its dominator. Defining appropriate partial value functions in general is not a trivial task. The process of building these functions is explained in greater detail in [33].

6.3 Characterizing Centrality via Positions

In this section, we discuss centrality in the context of measurement. We relate it to concepts introduced in the last section and point towards differences and similarities. Afterwards, we briefly introduce positions in networks and how they can be used to assess centrality.

Centrality and Measurement

Compared to the thorough investigation of measurement of RTM and CM, very little attention has been drawn to develop a well-founded procedure of measurement in networks such that centrality almost seems like an ad hoc concept. We briefly discuss four key issues of measurement in the context of centrality.

Representation. We could define an ERS as the graph $G = (V, E)$, i.e. $\mathcal{X} = V$ and $E \subseteq R$. However, a concatenation operation on V is missing since we can not combine vertices directly in a rational way. Further, we do not evaluate centrality according to attributes but on relations among actors. In a broader sense, the set of alternatives and attributes coincide. Concatenations on E are well-defined by extending direct relations to trajectories like walks and paths to obtain new relations. This derived relations can be subsumed in R together with E.

Uniqueness. The scale of measurement of centrality indices is an interval scale. We can not say that a vertex is twice as central as another but more importantly, the centrality scores themselves are more or less meaningless and do usually not play any role in empirical studies. Even if scores are normalized it is not justifiable to say a vertex with a betweenness of 1 is twice as 'between' as a vertex with betweenness 0.5. We thus do not lose any information if we weaken the scale to ordinal particularly because we are only concerned about ranking the vertices.

MEANINGFULNESS. The term meaningfulness describes concepts that are "relevant to the underlying measurement situation" [141, p. 31]. Relevance is very loosely defined, since it is not obvious what characteristics describe a satisfactory meaningfulness concept. If a procedure of measurement can be considered meaningful, independent of its conception, we can draw inferences based on the measurement results leading to valid conclusions. Applying indices haphazardly to networks does certainly not lead to valid conclusions, no matter what meaningfulness corresponds to. Meaningful results can only be obtained by a certain amount of preparatory work. Some phenomena might be directly transferable into measurement in networks. With the help of e.g. the classification of indices in Section 3.2, a suitable index can be chosen and we are able to draw valid conclusions afterwards. Other phenomena, however, might not be directly measurable. Taking the example of perceived trustworthiness of individuals in a communication network [185]. No centrality indices can measure the level of trustworthiness, but indirect the social process that leads to a high level of trust. In order to do so, we have to identify this process and translate it into graph-theoretic notion. Say, a person can be considered trustworthy if many others seek their advice or communicate frequently with them. This process would translate into degree centrality. If the underlying process can not be identified, we are unable to define an appropriate measurement procedure and are at risk to fall for anecdotal evidences.

VALIDITY. It is safe to say that validity is the major concern in measurement. One of few who recognizes the problem of validity in the context of centrality is Friedkin. In one of his seminal works, he states that

> *" [...] measures that have been derived from a social process can only be meaningfully applied to situations in which the social process occurs."* [81, p. 1480]

Meaningfully equates to validly in this context. He defined three measures which are specifically designed to fit three different social processes and are only valid in the respective context [81]. We have seen in Chapters 4 and 5 that validity of measures is often tested by correlation with empirically observed phenomena. As we have already argued, this is not enough to justify the appropriateness of indices. Going back to our introductory example, we might observe a correlation between body weight and height, yet this does not justify the use of a weighting scale to determine body height. The question is, if the assumed order or effect we intend to measure is actually inherent in a network or just inferred by the application of centrality indices.

REMARK. *Validity is not the same as reliability. A measure that is always off by a constant factor is reliable but not valid. Reliability is thus the extent to which a measure yields similar results consistently. We will see in Chapter 9 that some centrality indices can not even be considered reliable due to numerical inconveniences.*

Especially the considerations about representation show that the methods described in the last section can not readily be applied to networks since the prerequisites are different. However, we can observe several similarities which will be explored in the following subsection. Our goal is not to develop a profound theory of measurement for network centrality, but use analogies to motivate a new characterization of centrality using network positions.

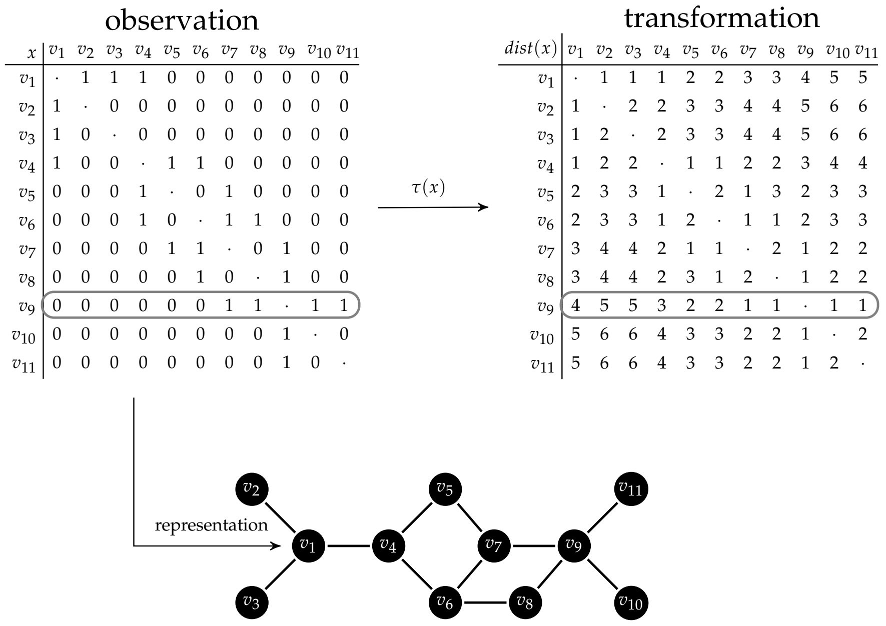
Networks and Positions

We can define a *dyadic variable* holding direct relations among n actors with

$$x_{ij} = \begin{cases} 1 & i \text{ and } j \text{ related} \\ 0 & \text{otherwise,} \end{cases}$$

and thus containing the actual *network data*. A graph can then be used as a visual *representation* of x. Note that we now make a distinction between networks and graphs instead of using the terms interchangeably as done before. Additionally, we use $\mathcal{N}$ as the set of actors and denote $\mathcal{D} = \mathcal{N} \times \mathcal{N}$ as the *dyadic domain*, i.e. we say $x \in \mathcal{R}^{\mathcal{D}}$, where $\mathcal{R}$ is an ordered value range.

The variable x constitutes our *observation* obtained from an experiment, study or other empirical efforts. We are, however, usually not merely interested in these direct relations among actors, but rather in indirect ones. These relations can be manifold, ranging from distances to dyadic dependencies and others that are e.g. used in the context of centrality indices. All kinds of indirect relations can be derived from x by a yet unspecified function $\tau(x)$. The obtained network $\tau(x)$ is a *transformation* of our observed network data with a different set of relations. Figure 6.1 illustrates the points made with an example network and distances as derived relations.



observation

x	v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8	v_9	v_{10}	v_{11}
v_1	·	1	1	1	0	0	0	0	0	0	0
v_2	1	·	0	0	0	0	0	0	0	0	0
v_3	1	0	·	0	0	0	0	0	0	0	0
v_4	1	0	0	·	1	1	0	0	0	0	0
v_5	0	0	0	1	·	0	1	0	0	0	0
v_6	0	0	0	1	0	·	1	1	0	0	0
v_7	0	0	0	0	1	1	·	0	1	0	0
v_8	0	0	0	0	0	1	0	·	1	0	0
v_9	0	0	0	0	0	0	1	1	·	1	1
v_{10}	0	0	0	0	0	0	0	0	1	·	0
v_{11}	0	0	0	0	0	0	0	0	1	0	·

$\xrightarrow{\tau(x)}$

transformation

$dist(x)$	v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8	v_9	v_{10}	v_{11}
v_1	·	1	1	1	2	2	3	3	4	5	5
v_2	1	·	2	2	3	3	4	4	5	6	6
v_3	1	2	·	2	3	3	4	4	5	6	6
v_4	1	2	2	·	1	1	2	2	3	4	4
v_5	2	3	3	1	·	2	1	3	2	3	3
v_6	2	3	3	1	2	·	1	1	2	3	3
v_7	3	4	4	2	1	1	·	2	1	2	2
v_8	3	4	4	2	3	1	2	·	1	2	2
v_9	4	5	5	3	2	2	1	1	·	1	1
v_{10}	5	6	6	4	3	3	2	2	1	·	2
v_{11}	5	6	6	4	3	3	2	2	1	2	·

FIGURE 6.1: Illustration of different dyadic variables. Distances can be derived from the direct relations x by a function $\tau(x)$. Dots on the diagonal indicate, that 'self' relations are not allowed. The gray row marks the position of v_9 for both relations. A graph is used to represent the direct relations.

The value range of x and $dist(x)$ are weakly ordered, i.e. a present relation is better than an absent one and lower distances are preferred.

The ith row of these networks, can be seen as the *position* of actor i within the network. It describes how he or she relates to all other actors. It is a generalization of positions in social space, i.e. *Blau space* [17, 18], adding a relational dimension.

This definition of a network position leaves out a lot of formalities and preliminary considerations of the newly introduced *positional approach to network analysis* [36]. Positions serve as a unifying approach to several network related questions, like

- centrality via ordering of positions,
- roles via equivalences of positions,
- cohesion via similarity of positions, and
- micro and macro structure via aggregations of positions.

The positional approach not only unifies methods but also suggests new ones from alternative instantiations. Giving a more detailed description is out of scope for this thesis but can be found in [36]. In the following, we focus on the comparison of positions in order to obtain (partial) centrality rankings. In the following, the position of an actor i in a network x is denoted by $pos(i|x)$.

Comparing Positions

Recall that we considered an actor to be more central than another if he or she has better relationships. In the context of positions, we thus want to assess whether an actor has a preferable position relative to others. Similar to decision making tasks we can argue about preferable positions by pairwise comparisons with a notion of *positional dominance* similar to Definition 6.3.

Definition 6.4. *Let $x \in \mathcal{R}^{\mathcal{D}}$ describe a network where the value range $\mathcal{R}$ is ordered by $\geq$. Further, let $u, v \in \mathcal{N}$. If $x_{ut} \geq x_{vt}$ holds for all $t \in \mathcal{N}$, then the position of u dominates v's position denoted by $pos(u|x) \succcurlyeq pos(v|x)$.*

Note that we commonly exclude $\{u, v\}$ from the comparison of u and v. This can be done without altering results, since $x_{uv} = x_{vu}$ and x_{uu} as well as x_{vv} are undefined in symmetric and simple networks. If $pos(u|x) \succcurlyeq pos(v|x)$ and $pos(v|x) \succcurlyeq pos(u|x)$ holds, then actor u and v occupy equivalent positions, denoted by $pos(u|x) \sim pos(v|x)$.

The positional dominance relations of the example in Figure 6.1 for both networks are given in Figure 6.2. Observe, that the dominance relations in x and $dist(x)$ are exactly the same. That is, $dist(x)$ *preserves* the dominance of the observed relations. This observation is of particular interest for theoretical considerations in the next chapter.

x	v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8	v_9	v_{10}	v_{11}
v_1		$\succcurlyeq$	$\succcurlyeq$								
v_2	$\preccurlyeq$		$\succcurlyeq$	$\preccurlyeq$							
v_3	$\preccurlyeq$	$\succcurlyeq$		$\preccurlyeq$							
v_4		$\succcurlyeq$	$\succcurlyeq$								
v_5						$\preccurlyeq$					
v_6					$\succcurlyeq$						
v_7								$\succcurlyeq$		$\succcurlyeq$	$\succcurlyeq$
v_8							$\preccurlyeq$			$\succcurlyeq$	$\succcurlyeq$
v_9										$\succcurlyeq$	$\succcurlyeq$
v_{10}							$\preccurlyeq$	$\preccurlyeq$	$\preccurlyeq$		$\succcurlyeq$
v_{11}							$\preccurlyeq$	$\preccurlyeq$	$\preccurlyeq$	$\succcurlyeq$	

$dist(x)$	v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8	v_9	v_{10}	v_{11}
v_1		$\succcurlyeq$	$\succcurlyeq$								
v_2	$\preccurlyeq$		$\succcurlyeq$	$\preccurlyeq$							
v_3	$\preccurlyeq$	$\succcurlyeq$		$\preccurlyeq$							
v_4		$\succcurlyeq$	$\succcurlyeq$								
v_5						$\preccurlyeq$					
v_6					$\succcurlyeq$						
v_7								$\succcurlyeq$		$\succcurlyeq$	$\succcurlyeq$
v_8							$\preccurlyeq$			$\succcurlyeq$	$\succcurlyeq$
v_9										$\succcurlyeq$	$\succcurlyeq$
v_{10}							$\preccurlyeq$	$\preccurlyeq$	$\preccurlyeq$		$\succcurlyeq$
v_{11}							$\preccurlyeq$	$\preccurlyeq$	$\preccurlyeq$	$\succcurlyeq$	

FIGURE 6.2: Dominance relations of the networks given in Figure 6.1. An entry '$\succcurlyeq$' in cell (i, j) indicates that actor i dominates actor j.

As for the decision making problem, we are, however, still left with many incomparable pairs of actors since positional dominance only yields a partial ranking. We therefore need further considerations to obtain a complete ranking by *quantification* of positions.

The usual approach to centrality is to aggregate the position of an actor with indices, i.e.

$$c_\tau(u) = \sum_{\{u,t\} \in \mathcal{D}} \tau(x)_{ut} ,$$

and we say

$$c_\tau(u) > c_\tau(v) \implies pos(u|\tau(x)) \text{ is better than } pos(v|\tau(x)).$$

Tacitly, however, we then assume that relations are *additive*. Although this seems natural it is not a straightforward assumption for any relation. Establishing additivity took much effort in other fields, e.g. psychology [197] and social sciences [163, 198], but has not yet been thoroughly discussed for relations in networks and it is mostly seen as given due to its simplicity.

A further assumption is the *homogeneity* among all actors. When relations are summed up, all relations are treated as equal and are not further differentiated. In our example network, the actors 4 and 7 both have three direct relations and would be considered equally central according to degree centrality. Imagine, however, the network represents the result of a survey in a company, where employees were asked to name the best performing staff members[1]. It surely makes a difference if an employee is named by a manager instead of a janitor. If actor 1 represents a manager and actor 9 a janitor, we would conclude that 4 is performing better than 7. The same reasoning can be applied to any other derived relation, e.g. when the network represents locations and connections among those. Being at short distance to a supermarket is generally more beneficial than being close to a prison. Therefore, a present ordinal actor attribute should not be neglected and provides additional useful information for centrality measurement.

Depending on the presence or absence of additivity and homogeneity, we have different possibilities to compare positions. If additivity and homogeneity can not be assumed, we are left with positional dominance for pairwise comparisons. On the bright side, we do not necessarily have to consider indirect relations $\tau(x)$ at all if

$$pos(u|x) \succcurlyeq pos(v|x) \implies pos(u|\tau(x)) \succcurlyeq pos(v|\tau(x))$$

holds true.

If relations are non-additive but actors are homogeneous, we can define a second form of dominance.

Definition 6.5. *Let $x \in \mathcal{R}^{\mathcal{D}}$ be a network where the value range $\mathcal{R}$ is ordered by $\geq$. Further, let $u, v \in \mathcal{N}$. If there exists an automorphism $\pi : \mathcal{N} \rightarrow \mathcal{N}$ such that $x_{u\pi(t)} \geq x_{vt}$ holds for all $t \in \mathcal{N}$, then the position of u dominates v's position under the total homogeneity assumption denoted by $pos(u|x) \succcurlyeq_h pos(v|x)$.*

Obviously, positional dominance implies dominance under homogeneity,

$$pos(u|x) \succcurlyeq pos(v|x) \implies pos(u|x) \succcurlyeq_h pos(v|x)$$

but not vice versa.

Dominance under total homogeneity can be checked by sorting the relationship values non-increasingly and comparing them front to back. The result for our example is shown in Figure 6.3. In contrast to positional dominance, we do not observe a connection between dominance in x and $dist(x)$, which also holds true for any $\tau(x)$. For our observed network x we have a complete

[1] This will most certainly not result in a symmetric network, since answers are unlikely to be reciprocated. The example just serves as an illustration.

x	v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8	v_9	v_{10}	v_{11}
v_1		$\succcurlyeq_h$	$\succcurlyeq_h$	$\succcurlyeq_h$	$\succcurlyeq_h$	$\succcurlyeq_h$	$\succcurlyeq_h$	$\succcurlyeq_h$	$\preccurlyeq_h$	$\succcurlyeq_h$	$\succcurlyeq_h$
v_2	$\preccurlyeq_h$		$\succcurlyeq_h$	$\preccurlyeq_h$	$\preccurlyeq_h$	$\preccurlyeq_h$	$\preccurlyeq_h$	$\preccurlyeq_h$	$\preccurlyeq_h$	$\succcurlyeq_h$	$\succcurlyeq_h$
v_3	$\preccurlyeq_h$	$\succcurlyeq_h$	$\preccurlyeq_h$	$\preccurlyeq_h$	$\preccurlyeq_h$	$\preccurlyeq_h$	$\preccurlyeq_h$	$\preccurlyeq_h$	$\preccurlyeq_h$	$\succcurlyeq_h$	$\succcurlyeq_h$
v_4	$\succcurlyeq_h$	$\succcurlyeq_h$	$\succcurlyeq_h$		$\succcurlyeq_h$	$\succcurlyeq_h$	$\succcurlyeq_h$	$\succcurlyeq_h$	$\preccurlyeq_h$	$\succcurlyeq_h$	$\succcurlyeq_h$
v_5	$\preccurlyeq_h$	$\succcurlyeq_h$	$\succcurlyeq_h$	$\preccurlyeq_h$		$\preccurlyeq_h$	$\preccurlyeq_h$	$\succcurlyeq_h$	$\preccurlyeq_h$	$\succcurlyeq_h$	$\succcurlyeq_h$
v_6	$\succcurlyeq_h$	$\succcurlyeq_h$	$\succcurlyeq_h$	$\succcurlyeq_h$	$\succcurlyeq_h$		$\succcurlyeq_h$	$\succcurlyeq_h$	$\preccurlyeq_h$	$\succcurlyeq_h$	$\succcurlyeq_h$
v_7	$\succcurlyeq_h$	$\succcurlyeq_h$	$\succcurlyeq_h$	$\succcurlyeq_h$	$\succcurlyeq_h$	$\succcurlyeq_h$		$\succcurlyeq_h$	$\preccurlyeq_h$	$\succcurlyeq_h$	$\succcurlyeq_h$
v_8	$\preccurlyeq_h$	$\succcurlyeq_h$	$\succcurlyeq_h$	$\preccurlyeq_h$	$\succcurlyeq_h$	$\preccurlyeq_h$	$\preccurlyeq_h$		$\preccurlyeq_h$	$\succcurlyeq_h$	$\succcurlyeq_h$
v_9	$\succcurlyeq_h$	$\succcurlyeq_h$	$\succcurlyeq_h$	$\succcurlyeq_h$	$\succcurlyeq_h$	$\succcurlyeq_h$	$\succcurlyeq_h$	$\succcurlyeq_h$		$\succcurlyeq_h$	$\succcurlyeq_h$
v_{10}	$\preccurlyeq_h$	$\succcurlyeq_h$	$\succcurlyeq_h$	$\preccurlyeq_h$	$\preccurlyeq_h$	$\preccurlyeq_h$	$\preccurlyeq_h$	$\preccurlyeq_h$	$\preccurlyeq_h$		$\succcurlyeq_h$
v_{11}	$\preccurlyeq_h$	$\succcurlyeq_h$	$\succcurlyeq_h$	$\preccurlyeq_h$	$\preccurlyeq_h$	$\preccurlyeq_h$	$\preccurlyeq_h$	$\preccurlyeq_h$	$\preccurlyeq_h$	$\succcurlyeq_h$	

$dist(x)$	v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8	v_9	v_{10}	v_{11}
v_1		$\succcurlyeq_h$	$\succcurlyeq_h$	$\preccurlyeq_h$		$\preccurlyeq_h$	$\preccurlyeq_h$			$\succcurlyeq_h$	$\succcurlyeq_h$
v_2	$\preccurlyeq_h$		$\succcurlyeq_h$	$\preccurlyeq_h$	$\preccurlyeq_h$	$\preccurlyeq_h$	$\preccurlyeq_h$	$\preccurlyeq_h$	$\preccurlyeq_h$	$\preccurlyeq_h$	$\preccurlyeq_h$
v_3	$\preccurlyeq_h$	$\succcurlyeq_h$		$\preccurlyeq_h$	$\preccurlyeq_h$	$\preccurlyeq_h$	$\preccurlyeq_h$	$\preccurlyeq_h$	$\preccurlyeq_h$	$\preccurlyeq_h$	$\preccurlyeq_h$
v_4	$\succcurlyeq_h$	$\succcurlyeq_h$	$\succcurlyeq_h$				$\succcurlyeq_h$	$\succcurlyeq_h$		$\succcurlyeq_h$	$\succcurlyeq_h$
v_5		$\succcurlyeq_h$	$\succcurlyeq_h$			$\preccurlyeq_h$				$\succcurlyeq_h$	$\succcurlyeq_h$
v_6	$\succcurlyeq_h$	$\succcurlyeq_h$	$\succcurlyeq_h$		$\succcurlyeq_h$			$\succcurlyeq_h$		$\succcurlyeq_h$	$\succcurlyeq_h$
v_7	$\succcurlyeq_h$	$\succcurlyeq_h$	$\succcurlyeq_h$	$\succcurlyeq_h$				$\succcurlyeq_h$		$\succcurlyeq_h$	$\succcurlyeq_h$
v_8		$\succcurlyeq_h$	$\succcurlyeq_h$	$\preccurlyeq_h$		$\preccurlyeq_h$	$\preccurlyeq_h$			$\succcurlyeq_h$	$\succcurlyeq_h$
v_9	$\succcurlyeq_h$	$\succcurlyeq_h$	$\succcurlyeq_h$							$\succcurlyeq_h$	$\succcurlyeq_h$
v_{10}	$\preccurlyeq_h$	$\succcurlyeq_h$	$\succcurlyeq_h$	$\preccurlyeq_h$	$\preccurlyeq_h$	$\preccurlyeq_h$	$\preccurlyeq_h$	$\preccurlyeq_h$	$\preccurlyeq_h$		$\succcurlyeq_h$
v_{11}	$\preccurlyeq_h$	$\succcurlyeq_h$	$\succcurlyeq_h$	$\preccurlyeq_h$	$\preccurlyeq_h$	$\preccurlyeq_h$	$\preccurlyeq_h$	$\preccurlyeq_h$	$\preccurlyeq_h$	$\succcurlyeq_h$	

FIGURE 6.3: Dominance relations under total homogeneity of the networks given in Figure 6.1. An entry '$\succcurlyeq_h$' in cell (i, j) indicates that actor i dominates actor j under the total homogeneity assumption.

ranking of positions,

$$\begin{aligned} pos(v_9|x) &\succcurlyeq pos(v_1|x) \sim pos(v_4|x) \sim pos(v_6|x) \sim pos(v_7|x) \\ &\succcurlyeq pos(v_5|x) \sim pos(v_8|x) \\ &\succcurlyeq pos(v_2|x) \sim pos(v_3|x) \sim pos(v_{10}|x) \sim pos(v_{11}|x). \end{aligned}$$

The ranking naturally coincides with degree centrality, since the value range of x is dichotomous. For derived networks $\tau(x)$ with a larger value range, we do not necessarily expect a complete ranking such that we still need further methods if we need a complete ranking.

In cases where additivity can be assumed but not homogeneity, we generally need more specifications about inhomogeneities to compare positions.

The considerations about additivity and homogeneity are summarized in Table 6.2.

relations	*actors* homogeneous	 inhomogeneous
additive	$c : \mathcal{N} \to \mathbb{R}_0^+$	?
non-additive	$\succcurlyeq_h$	$\succcurlyeq$

TABLE 6.2: Comparing positions with different assumptions about additivity and homogeneity.

Note that when additivity is assumed, it is always possible to rank actors completely. Constructing the ranking, however, must be done with caution. Figure 6.1, showed that there are incomparable pairs of actors even under total homogeneity according to distance. Yet by applying closeness centrality, we obtain a complete ranking. We might thus run the risk of inferring a ranking which is not intrinsic to a network.

In the case of non-additivity we are mostly left with a partial ranking. However, there are special cases where we can obtain a complete ranking. One such case is discussed in depth in Chapter 8 and others in Chapter 10.

Assuming additivity and homogeneity involve a high amount of arbitrariness when it comes to the quantification of centrality in terms of indices, yet in the absence of theories about both properties we have to keep relying on them. However, this chapter indirectly provided us with some new theoretical insights about centrality indices. By comparing positions with positional dominance, we obtained

$$pos(u|x) \succcurlyeq pos(v|x) \implies pos(u|\,dist(x))) \succcurlyeq pos(v|\,dist(x))) \implies c_c(u) \geq c_c(v).$$

In the upcoming part we generalize this statement to a larger class of indirect relations $\tau(x)$, i.e.

$$pos(u|x) \succcurlyeq pos(v|x) \implies pos(u|\tau(x))) \succcurlyeq pos(v|\tau(x))) \implies c_\tau(u) \geq c_\tau(v).$$

We could now proceed by showing that the implications hold for individual relations separately, yet this would prevent us from drawing general conclusion. In contrast, we show that the statement holds for a large class of transformations τ obtained from indirect relations defined by specific algebraic structures, such that we can derive conditions for the preservation of dominance. The upcoming chapter will deal with these formalities.

Chapter 7

Re-Conceptualizing Network Centrality

"Ideally, measures should grow out of advanced theoretical efforts; they should be defined in the context of explicit process models. Before such models can be developed, however, a certain amount of conceptual specification is necessary;"
– Freeman, 1979

7.1 Formal Definitions and Terminology

In this section, we define the basic concepts and set up some standard notation and terminology outlined in the previous chapter.

Networks

So far we have addressed network centrality from a graph-theoretic perspective, i.e. undirected graphs $G = (V, E)$. However, graphs are merely a *representation* of underlying network data.

Definition 7.1. *A network is a mapping $x : \mathcal{D} \to \mathcal{R}$ (or vector $x \in \mathcal{R}^{\mathcal{D}}$) assigning values in a range $\mathcal{R}$ to dyads from an interaction domain $\mathcal{D} \subseteq \mathcal{N} \times \mathcal{N}$ comprised of ordered pairs of nodes $\mathcal{N}$.*

Undirected graphs are thus representations of a dichotomous network on the interaction domain $\mathcal{D} = (\mathcal{N} \times \mathcal{N}) \setminus \{(i,i) \ : \ i \in \mathcal{N}\}$ with value range $\mathcal{R} = \{0, 1\}$, where the values represent the presence or absence of relationships. In general, however, there are no restrictions put on the value range such that it can be any set of numbers (e.g. $\mathbb{R}$), intervals (e.g. $[0, 1]$), or any other kind of objects (e.g. timestamps).

To facilitate distinction and comparison we assume the existence of a preorder $\leq$ on $\mathcal{R}$. This preorder is either bounded from below by a special element $0 \in \mathcal{R}$ with $0 \leq a$ for all $a \in \mathcal{R}$ or from above by $\infty \in \mathcal{R}$ for all $a \in \mathcal{R}$. The distinction is made such that larger values represent more beneficial relationships or smaller values represent less costly relationships.

In the upcoming parts, we focus on dichotomous networks where self relations are forbidden, i.e. $x \in \mathcal{B}^{\mathcal{D}}$ where $\mathcal{B} = \{0,1\}$, $\mathcal{D} = (\mathcal{N} \times \mathcal{N}) \setminus \{(i,i) : i \in \mathcal{N}\}$ and $x_{ij} = x_{ji}$ for all $i \neq j \in \mathcal{N}$.

Neighborhood-Inclusion Preorder

Definition 6.4 introduced positional dominance as a binary relation in networks x. If actor u's position dominates the position of v in x, we say

$$pos(u|x) \succcurlyeq pos(v|x).$$

Domination was also used for indirect relations $\tau(x)$, however, we here focus on the domination in x and under which circumstances it is preserved by $\tau(x)$. We thus simplify our notation by

$$pos(u|x) \succcurlyeq pos(v|x) \iff u \succcurlyeq v.$$

The induced preorder of domination is the core element of the upcoming parts. In graph-theoretic terms, it translates to the well-studied *neighborhood-inclusion* or *vicinal preorder* for undirected and unweighted graphs [19, 129].

Corollary 7.2. *Let $G = (V, E)$ be the simple undirected graph representing a network x. Then*

$$u \succcurlyeq v \iff N[u] \supseteq N(v).$$

An example of neighborhood-inclusion is depicted in Figure 7.1. In the following, we use positional dominance and neighborhood-inclusion interchangeably, keeping in mind that the former is actually more general.

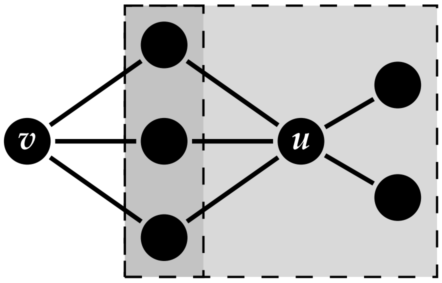


FIGURE 7.1: Illustration of neighborhood-inclusion. The neighborhood of vertex v is completely contained in the closed neighborhood of u.

Semirings

Most of the definitions in this section are adapted from Gondran & Minoux [88], Batagelj [10] and Mohri [138].

Definition 7.3. *An algebraic structure $(\mathcal{R}, \oplus, \odot, \overline{0}, \overline{1})$ on a set $\mathcal{R}$ is a semiring if and only if*

(i) $(\mathcal{R}, \oplus, \overline{0})$ is a commutative monoid with a neutral element $\overline{0}$.

1. $\forall a, b \in \mathcal{R}: \quad a \oplus b \in S$
2. $\forall a, b \in \mathcal{R}: \quad a \oplus b = b \oplus a$
3. $\forall a \in \mathcal{R}: \quad \overline{0} \oplus a = a \oplus \overline{0} = a$
4. $\forall a, b, c \in \mathcal{R}: \quad a \oplus (b \oplus c) = (a \oplus b) \oplus c$

(ii) $(\mathcal{R}, \odot, \overline{1})$ *is a monoid with neutral element* $\overline{1}$.

1. $\forall a, b \in \mathcal{R}: \quad a \odot b \in \mathcal{R}$
2. $\forall a \in \mathcal{R}: \quad \overline{1} \odot a = a \odot \overline{1} = a$
3. $\forall a, b, c \in \mathcal{R}: \quad a \odot (b \odot c) = (a \odot b) \odot c$

(iii) $\odot$ *distributes over* $\oplus$

(iv) $\overline{0}$ *is an annihilator for* $\odot$*:* $\forall a \in \mathcal{R}: \quad a \odot \overline{0} = \overline{0} \odot a = \overline{0}$

The binary relation $\oplus$ is called *aggregation* and $\odot$ *concatenation*. If both binary operations and neutral elements are understood, we denote the semiring simply by $\mathcal{R}$. A concatenation $a \odot b$ can conveniently be written as a juxtaposition ab and $a \odot a$ as a^2.

Definition 7.4. *Let* $(\mathcal{R}, \oplus, \odot, \overline{0}, \overline{1})$ *be a semiring and*

$$a^{(k)} = \overline{0} \oplus a \oplus \ldots \oplus a^k \quad (\forall a \in \mathcal{R}).$$

An element a is <u>*k-closed*</u> *if* $a^{(k)} = a^{(k+1)}$.

It is easy to prove by induction that

$$a \in \mathcal{R} \text{ } k\text{-closed} \implies a^{(k)} = a^{(k+l)} \quad (\forall l \geq 1)\,.$$

holds.

Definition 7.5. *Let* $(\mathcal{R}, \oplus, \odot, \overline{0}, \overline{1})$ *be a semiring. For k-closed elements* $a \in \mathcal{R}$*, the* <u>*closure operation*</u> a^* *is defined as*

$$a^* = \bigoplus_{l=0}^{\infty} a^l = a^{(k)}$$

and the <u>*strict closure operation*</u> *as*

$$\bar{a} = a \odot a^*.$$

An important property of semirings for our context is *monotonicity*.

Definition 7.6. *Let* $(\mathcal{R}, \oplus, \odot, \overline{0}, \overline{1})$ *be a semiring and* $\geq$ *a preorder over* $\mathcal{R}$*. The semiring is* <u>*monotonic*</u> *if the following statements hold for all* $a, b, c \in \mathcal{R}$*:*

(i) $(a \geq b) \implies (a \oplus c \geq b \oplus c)$

(ii) $(a \geq b) \implies (a \odot c \geq b \odot c)$

(iii) $(a \geq b) \implies (c \odot a \geq c \odot b)$

A semiring is strictly monotonic if

$$\begin{aligned}
&\textit{(i)}\ \ (a > b) \implies (a \oplus c > b \oplus c)\\
&\textit{(ii)}\ \ (a > b) \implies (a \odot c > b \odot c)\\
&\textit{(iii)}\ \ (a > b) \implies (c \odot a > c \odot b)
\end{aligned}$$

Monotonicity requires that the set $\mathcal{R}$ is preordered, which is guaranteed if $\mathcal{R} \subseteq \mathbb{R}$ assuming the natural order. However, it is not necessarily unambiguous when $\mathcal{R}$ is multidimensional. Monotonicity could be defined componentwise or after any form of aggregation.

Formal Power Series and Centrality

A *formal power series* can be loosely defined as a power series in which questions of convergence are ignored by using indeterminates as variables [149]. We can write a formal power series as

$$\sum_{k=0}^{\infty} \gamma_k X^k \,,$$

where X is an indeterminate and $\gamma = (\gamma_k)_{k=0}^{\infty}$ is a sequence of real numbers indexed by the natural numbers. An example of formal power series are *generating functions*. The main purpose of formal power series is to study properties of the sequences γ.

Centralities relying on counting walks can be described as a formal power series of the weighting factors $(\gamma_k)_{k=0}^{\infty}$. Assuming we are given the sequence of (s,t)-walk counts of lengths $k > 0$

$$\omega_{st} = [\omega_{st}^0,\ \omega_{st}^1,\ \omega_{st}^2, \dots,\ \omega_{st}^k, \dots],$$

we can calculate the respective indirect relations by an infinite scalar product

$$\langle \gamma, \omega_{st} \rangle = \sum_{k=0}^{\infty} \gamma_k \omega_{st}^k \,.$$

where $(\gamma_k)_{k=0}^{\infty}$ are the weighting factors of a walk based centrality index. Examples for such sequences are

$$\begin{aligned}
&(i)\ \text{eigenvector:} && \gamma = [0, 1, 1, 1, \dots]\\
&(ii)\ \text{subgraph:} && \gamma = [1, 1, 1/2, \dots, 1/k!, \dots]\\
&(iii)\ \text{Katz:} && \gamma = [0, \alpha, \alpha^2, \dots, \alpha^k, \dots]\\
&(iv)\ \text{Bonacich:} && \gamma = [0, 1, \beta, \beta^2, \dots, \beta^{k-1}, \dots]
\end{aligned}$$

7.2 Semirings for Indirect Relations

Walks and Paths

Although indirect relations in networks rely on different graph-theoretic concepts, they can all be derived in a similar fashion with the notion of walks and paths. A *walk* of length $k \in \mathbb{N}$ is defined as an alternating sequence

$$i_0,\ \{i_0, i_1\},\ i_1,\ \{i_1, i_2\}, \dots,\ i_{k-1},\ \{i_{k-1}, i_k\},\ i_k$$

of $k+1$ nodes and k dyads. For brevity, we only consider the sequence of dyads from now on. Dyads are unordered, since we only consider undirected graphs, i.e. symmetric networks. Note the slight notational adjustment from edges to dyads. A walk can be *any* sequence of dyads of the above kind and does not necessarily have to be realizable in the graph, in contrast to the definition given in Chapter 2. Other trajectories like (s,t)-walks, closed walks and paths are adjusted in the same way. We define the set of all possible (s,t)-walks of length k to be $\mathcal{D}^k_{st}$ and $\mathcal{D}^k = \bigcup_{s,t \in \mathcal{N}} \mathcal{D}^k_{st}$. Correspondingly, we define $\mathcal{D}^{(k)}_{st}$ and $\mathcal{D}^{(k)}$ for walks of length at most k, and $\mathcal{D}^*_{st}$ and $\mathcal{D}^*$ for walks of arbitrary length.

Indirect Relations from Path Algebras

The value range of symmetric dichotomous networks $x \in \mathcal{B}^{\mathcal{D}}$ permits a semiring structure $(\mathcal{B}, \max, \min, 0, 1)$ which can only be used to derive reachability as an indirect relation. In order to obtain further relations, we therefore have to map the network into a suitable value range and additionally have to include the set $\{(i,i) : i \in \mathcal{N}\}$ for numerical conveniences. These two steps can be described in terms of a monomorphism.

Definition 7.7. *Let $x \in \mathcal{B}^{\mathcal{D}}$ be a symmetric network on the interaction domain $\mathcal{D} = (\mathcal{N} \times \mathcal{N}) \setminus \{(i,i) : i \in \mathcal{N}\}$ and $(\mathcal{R}, \oplus, \odot, \bar{0}, \bar{1})$ be a semiring. Let $g_1 : \mathcal{B} \to \mathcal{B} \cup \{\cdot\}$ be an injective function with*

$$x'_{st} = g_1(x_{st}) = \begin{cases} 1 & (s,t) \in \mathcal{D} \wedge x_{st} = 1 \\ 0 & (s,t) \in \mathcal{D} \wedge x_{st} = 0 \end{cases}$$

and $g_2 : \mathcal{B} \cup \{\cdot\} \to \mathcal{R}$ be a function with

$$g_2(x'_{st}) = \begin{cases} y_{st} & x'_{st} = 1 \\ \bar{0} & x'_{st} = 0 \\ \bar{1} & x'_{st} = \cdot. \end{cases}$$

The network $y = g(x) \in \mathcal{R}^{\mathcal{D}}$ with $g = g_2 \circ g_1$ is the <u>*monomorphic transformation*</u> *of x in $\mathcal{R}$.*

The function g_1 is necessary in order to distinguish absent and undefined relations. Note that the network $y = g(x)$ does not yet hold a new relation but rather transfered the observed network into an extended interaction domain with differing value range.

TABLE 7.1: One dimensional semirings for indirect relations

Relation	$\mathcal{R}$	$\oplus$	$\odot$	$\overline{0}$	$\overline{1}$	order
reachability	$\mathcal{B}$	$\vee$	$\wedge$	0	1	$\geq$
shortest path	$\mathbb{R}_0^+$	min	$+$	∞	0	$\leq$
longest path	$\mathbb{R}_0^+$	max	$+$	∞	0	$\geq$
max. reliability	$[0,1]$	max	$\times$	0	1	$\geq$
max. capacity	$\mathbb{R}_0^+$	max	min	0	∞	$\geq$

Since the monomorphically transformed network $y = g(x)$ permits a semiring structure $(\mathcal{R}, \oplus, \odot, \overline{0}, \overline{1})$, we can first extend $y \in \mathcal{R}^{\mathcal{D}}$ to walks $P \in \mathcal{D}^*$ by concatenation

$$y(P) = \bigodot_{(i,j)\in P} y_{ij} .$$

Afterwards, we aggregate over the set of walks applying the closure operator

$$y^*_{st} = \bigoplus_{P\in\mathcal{D}^*_{st}} y(P) .$$

If $\mathcal{R} \subseteq \mathbb{R}_0^+$ holds, we directly obtain a new relation $\tau(x)_{st} = y^*_{st}$. Some one-dimensional semirings are given in Table 7.1 together with the respective order relation.

However, many relations of interest in the context of centrality can not be derived from a one-dimensional semiring. The following definition gives a suitable semiring in order to derive dyadic dependencies and can be found in [10].

Definition 7.8. *Let $(\mathbb{R}_0^+ \times \mathbb{N}_0, \geq)$ be a preordered set where $\geq$ is an adapted lexicographical order defined as*

$$(a,i) \geq (b,j) \iff a < b \vee (a = b \wedge i \geq j). \tag{7.1}$$

Then, $\left(\mathbb{R}_0^+ \times \mathbb{N}_0, \oplus, \odot, (\infty, 0), (0,1)\right)$ is called the geodetic semiring with concatenation

$$(a,i) \odot (b,j) = (a+b, i\cdot j)$$

and aggregation

$$(a,i) \oplus (b,j) = \left(\min(a,b), \begin{Bmatrix} i & a<b \\ i+j & a=b \\ j & a>b \end{Bmatrix} \right) .$$

The monomorphic transformation of a network x in the geodetic semiring is given by

$$y_{st} = \begin{cases} (1,1) & (s,t) \in \mathcal{D} \wedge x_{st} = 1 \\ (\infty,0) & (s,t) \in \mathcal{D} \wedge x_{st} = 0 \\ (0,1) & (s,t) \notin \mathcal{D}. \end{cases}$$

Then, an element

$$y^*_{st} = (dist(y)_{st}, \sigma(y)_{st}),$$

of the strict closure comprises the distance and the number of shortest paths between s and t. The geodetic semiring is monotonic with the lexicographic order given in Equation (7.1), however, it is not strictly monotonic in $\oplus$. Consider $c < a < b \in \mathbb{R}^+_0$ and $j < i < k \in \mathbb{N}_0$. Then $(a,i) > (b,j)$ but $(a,i) \oplus (c,k) = (c,k) = (b,j) \oplus (c,k)$.

In order to obtain dyadic dependencies from the strict closure y^*, we have to define a mapping into the real numbers. To do so, we need the following Lemma.

Lemma 7.9. *Let $y^* \in (\mathbb{R}^+_0, \mathbb{N}_0)^{\mathcal{D}}$, where $y^*_{st} = (dist(y)_{st}, \sigma(y)_{st})$. It holds that*

$$\sigma(s,t|i) = \begin{cases} \sigma(y)_{si} \cdot \sigma(y)_{it} & dist(y)_{si} + dist(y)_{it} = dist(y)_{st} \\ 0 & \textit{otherwise.} \end{cases}$$

With Lemma 7.9, we can define a function

$$\delta(y^*)_{si} = \begin{cases} \sigma(y)_{si} \cdot \sum\limits_{(i,t)\in\mathcal{D}} \left(\frac{\sigma(y)_{it}}{\sigma(y)_{st}} \cdot \mathbf{1}_{st}(i) \right) & s \neq i \\ 0 & s = i. \end{cases}$$

where $\mathbf{1}_{st}(i)$ is a Kronecker delta-esque function defined as

$$\mathbf{1}_{st}(i) = \begin{cases} 1 & dist(y)_{si} + dist(y)_{it} = dist(y)_{st} \\ 0 & \text{otherwise,} \end{cases}$$

such that $\tau(x) = \delta(y^*)$ yields the desired relation.

Semirings for indirect relations do not necessarily have to be of finite dimension. The following definition introduces a non finite semiring which can be used to derive walk counts of arbitrary length.

Definition 7.10. *Let $(\Omega, \geq)$ be the preordered set of all sequences $\alpha = (\alpha_k)_{k=0}^{\infty}$ with $\alpha_k \in \mathbb{R}$ for all $k \in \mathbb{N}$ and binary relation $\geq$ defined as*

$$\alpha \geq \beta \iff \sum_{k=0}^{k_0} \alpha_k \geq \sum_{k=0}^{k_0} \beta_k \qquad \forall k_0 \in \mathbb{N}_0. \tag{7.2}$$

Then $(\Omega, \oplus, \odot, [0,0,0,\ldots], [1,0,0,\ldots])$ is called the semiring of infinite sequences with concatenation

$$\alpha \odot \beta = [\alpha_0\beta_0,\ \alpha_1\beta_0 + \alpha_0\beta_1,\ \alpha_2\beta_0 + \alpha_1\beta_1 + \alpha_0\beta_2, \ldots,\ \sum_{j=0}^{k} \alpha_j \beta_{k-j}, \ldots]$$

and aggregation

$$\alpha \oplus \beta = [\alpha_0 + \beta_0,\ \alpha_1 + \beta_1,\ \alpha_2 + \beta_2, \ldots] \,.$$

The monomorphic transformation of a network x in Ω is given by

$$y_{st} = \begin{cases} [0,1,0,\dots] & (s,t) \in \mathcal{D} \ \wedge \ x_{st} = 1 \\ [0,0,0,\dots] & (s,t) \in \mathcal{D} \ \wedge \ x_{st} = 0 \\ [1,0,0,\dots] & (s,t) \notin \mathcal{D}. \end{cases}$$

and the strict closure by

$$y^*_{st} = [w^0_{st}, w^1_{st}, \dots, w^k_{st}, \dots]$$

where w^k_{st} is the number of $(s,t)-$walks of length k. To derive a relation used for walk based indices, we have to project the walk sequence into the real numbers with the previously defined infinite scalar product, i.e.

$$\tau(x)_{st} = \langle \gamma, y^*_{st} \rangle,$$

where γ is a weighting sequence of a walk based index. Additional specification might be necessary if i.a. only closed walks are considered as for subgraph centrality.

We conclude this section with an important semiring to enumerate specific paths or walks in network [88].

Definition 7.11. *Let $Q \in \mathcal{D}^k_{it}$ and $P \in \mathcal{D}^k_{jt}$. P and Q are k-similar, denoted by $P \bowtie_k Q$, if they only differ in the first dyad.*

Let $Q \in \mathcal{D}^k_{it}$ with $P = (i,i_1),(i_1,i_2),\dots,(i,i_l),\dots,(i_{k-1},t)$. A walk $P \in \mathcal{D}^{k-l}_{jt}$ with $P = (i,i_l),(i_l,i_{l+1}),\dots,(i_{k-1},t)$ is called an l-contraction of P, denoted by $P \rhd_l Q$.

Abusing set theoretic notation, we will address the overlap of two k-similar walks P and Q by $P \cap Q$.

Definition 7.12. *Let $(\mathcal{P}(\mathcal{D}^*), \geq)$ be a preordered set with binary relation $\geq$ defined as*

$$\mathcal{W}_i \geq \mathcal{W}_j \implies \forall Q \in \mathcal{W}_i \ \exists P \in \mathcal{W}_j \ : \ (Q \bowtie P \vee Q \rhd_1 P).$$

Then $\left(\mathcal{P}(\mathcal{D}^*), \bigcup, \circ_\psi, \emptyset, \mathcal{D}^0\right)$ is called the semiring of enumerated $\psi-$trajectories, where for

$$\begin{aligned} P &= \{i_0,i_1\}, \ \{i_1,i_2\}, \dots \{i_{k-1},i_k\} \in \mathcal{D}^k \quad \textit{and} \\ Q &= \{j_0,j_1\}, \ \{j_1,j_2\}, \dots \{j_{l-1},j_l\} \in \mathcal{D}^l, \end{aligned}$$

concatenation is defined as

$$P \circ_\psi Q = \begin{cases} \{i_0,i_1\}, \ \{i_1,i_2\}, \dots \{i_{k-1},i_k\}, \ \{i_k,j_1\}, \ \{j_1,j_2\}, \dots \{j_{l-1},j_l\} & i_k = j_0 \textit{ and} \\ & \textit{has property } \psi \\ \emptyset & \textit{otherwise.} \end{cases}$$

The property ψ is used to constrain the considered trajectories. For instance, to enumerate walks or paths up to length k, walks or paths with length k. The monomorphic transformation of a network x in $\mathcal{P}(\mathcal{D}^*)$ is given by

$$y_{st} = \begin{cases} \{(s,t)\} & (s,t) \in \mathcal{D} \wedge x_{st} = 1 \\ \emptyset & (s,t) \in \mathcal{D} \wedge x_{st} = 0 \\ \{(\cdot,\cdot)\} & (s,t) \notin \mathcal{D}. \end{cases}$$

and the strict closure by

$$y^*_{st} = \{P_1, P_2, \dots\}$$

where $P_i \in \mathcal{D}^*_{st}$ are trajectories that fulfill property ψ. Indirect relations derived from these sets of trajectories can be of various kind. Commonly, however, the cardinality ,i.e.

$$\tau(x)_{st} = |y^*_{st}|$$

is used.

7.3 Neighborhood-Inclusion and Indirect Relations

We now have the basic terminology and notations to consider the preservation of dominance or, equivalently, the neighborhood-inclusion preorder in undirected unweighted graphs for relations $\tau(x)$ and indices of the form

$$c_\tau(u) = \sum_{\{u,t\} \in \mathcal{D}} \tau(x)_{uv} \, .$$

Definition 7.13. *Let $x \in \mathcal{R}^{\mathcal{D}}_{\epsilon}$ be a network. An index $c_\tau : \mathcal{N} \to \mathbb{R}^+_0$ preserves domination if*

$$u \succeq v \implies c_\tau(u) \geq c_\tau(v) \, .$$

Domination is strictly preserved if

$$u \succ v \implies c_\tau(u) > c_\tau(v) \, .$$

We can relate the preservation to properties of $\tau(x)$.

Corollary 7.14. *Let $x \in \mathcal{R}^{\mathcal{D}}_{\epsilon}$ be a network and $c_\tau : \mathcal{N} \to \mathbb{R}^+_0$ be an index with*

$$c_\tau(u) = \sum_{\{u,t\} \in \mathcal{D}} \tau(x)_{ut} \, .$$

Further, let $u, v \in \mathcal{N}$ and $u \succeq v$. Domination is preserved if one of the following statements holds true.

(i) $\forall t \in \mathcal{N} \, : \, \tau(x)_{ut} \geq \tau(x)_{vt}$

(ii) $\forall \, t \in \mathcal{N} \; \exists \, \pi : \mathcal{N} \to \mathcal{N} \, : \, \tau(x)_{u\pi(t)} \geq \tau(x)_{vt}$

The second condition in Corollary 7.14 is equivalent to domination under total homogeneity in Definition 6.5. Thus, $(i) \implies (ii)$ but not vice versa. In the forthcoming we therefore develop sufficient conditions for Corollary 7.14(i) to be fulfilled.

We start by giving sufficient conditions for the preservation of dominance in the strict closure of a semiring.

Theorem 7.15. *Let y be the monomorphic transformation of $x \in \mathcal{B}^{\mathcal{D}}$ in the semiring $(\mathcal{R}, \oplus, \odot, \overline{0}, \overline{1})$, where $(\mathcal{R}, \geq)$ is a preordered set. Then, for u, $v \in \mathcal{N}$,*

$$u \curlyeqsucc v \implies \forall t \in \mathcal{N} \setminus \{u,v\} \ \forall k \in \mathbb{N} : \quad \bigoplus_{P \in \mathcal{D}_{ut}^{(k)}} y(P) \geq \bigoplus_{Q \in \mathcal{D}_{vt}^{(k)}} y(Q).$$

holds true if the following statements hold true.

(i) $(\mathcal{R}, \geq)$ *is monotonic*

(ii) $\forall (s,t) \in \mathcal{D} \, : \, \overline{1} \geq y_{st} \geq \overline{0}$

(iii) $\forall k \in \mathbb{N} \ \forall P \in \mathcal{D}^k \ \exists Q \in \mathcal{D}^k \ : \ P \bowtie_k Q$

(iv) $\forall k \in \mathbb{N} \ \forall P \in \mathcal{D}^k \ \exists Q \in \mathcal{D}^{k-1} \ : \ P \rhd_1 Q$

Proof. We proof the theorem by showing that if (i)-(iv) are fulfilled, then

$$u \curlyeqsucc v \implies \forall Q \in \mathcal{D}_{vt}^{(k)} \ \exists P \in \mathcal{D}_{ut}^{(k)} \ : \ (Q \bowtie P \vee Q \rhd_1 P) \wedge (y(P) \geq y(Q))$$

holds true for all $t \in \mathcal{N} \setminus \{u,v\}$ and $k \in \mathbb{N}$.
W.l.o.g let $t \in \mathcal{N} \setminus \{u,v\}$ and $k \in \mathbb{N}$ be arbitrarily chosen.

We define two sets

$$\left[\mathcal{D}_{vt}^{l}\right]_{-} = \left\{Q \in \mathcal{D}_{vt}^{l} \ : \ \{v,u\} \notin Q \ \vee \ \{i_0, i_1\} \neq \{v,u\}\right\} \quad \text{and}$$
$$\left[\mathcal{D}_{vt}^{l}\right]_{+} = \left\{Q \in \mathcal{D}_{vt}^{l} \ : \ \{i_0, i_1\} = \{v,u\}\right\},$$

with $l \leq k$ such that

$$\left[\mathcal{D}_{vt}^{l}\right]_{-} \cup \left[\mathcal{D}_{vt}^{l}\right]_{+} = \mathcal{D}_{vt}^{l}.$$

Consider an arbitrary $Q \in \left[\mathcal{D}_{vt}^{l}\right]_{-}$. With (iii) we can define $P \in \mathcal{D}_{ut}^{l}$ such that $P \bowtie_l Q$. Let $R = P \cap Q$. If $(\mathcal{R}, \geq)$ is monotonic, it follows that

$$y(P) = y_{ui_1} \odot y(R) \geq y_{vi_1} \odot y(R) = y(Q).$$

Since $\left[\mathcal{D}_{vt}^{l}\right]_{-} \subseteq \left[\mathcal{D}_{vt}^{(k)}\right]_{-}$, we have

$$\forall Q \in \left[\mathcal{D}_{vt}^{(k)}\right]_{-} \ \exists P \in \mathcal{D}_{ut}^{(k)} \ : \ P \bowtie_{(k)} Q \wedge y(P) \geq y(Q).$$

Consider now an arbitrary $Q \in \left[\mathcal{D}_{vt}^{l}\right]_{+}$. With (iv) we can define $P \in \mathcal{D}_{ut}^{l-1}$ such that $Q \rhd_{l-1} P$. With (ii), it follows that

$$y(P) \geq y_{vu} \odot y(P) = y(Q),$$

since $\overline{1} \geq y_{vu}$ implies $y(P) \geq y_{vu} \odot y(P)$. We thus have

$$\forall Q \in \left[\mathcal{D}_{vt}^{l}\right]_{+} \ \exists P \in \mathcal{D}_{ut}^{l-1} \ : \ Q \rhd_1 P \wedge y(P) \geq y(Q).$$

In summary, we have shown that

$$\forall Q \in \mathcal{D}_{vt}^{(k)} \; \exists P \in \mathcal{D}_{ut}^{(k)} \; : \; (Q \bowtie_{(k)} P \vee Q \rhd_1 P) \wedge (y(P) \geq y(Q))$$

which directly implies that

$$\bigoplus_{P \in \mathcal{D}_{ut}^{(k)}} y(P) \geq \bigoplus_{Q \in \mathcal{D}_{vt}^{(k)}} y(Q) \qquad \forall k \in \mathbb{N}.$$

holds. □

The geodetic semiring and the semiring of infinite sequences as well as all one dimensional semirings in Table 7.1 fulfill the requirements, except the longest path semiring which violates (ii).

Corollary 7.16.

$$u \succcurlyeq v \implies y_{ut}^* \geq y_{vt}^* \qquad \forall t \in \mathcal{N} \setminus \{u, v\}$$

if the requirements of Theorem 7.15 are fulfilled.

Theorem 7.15 and Corollary 7.16 excluded the direct comparison of $\{u, v\}$, $\{v, u\}$ and the diagonal elements. These special cases are treated in the following corollary.

Corollary 7.17. *Let $x \in \mathcal{B}^{\mathcal{D}}$ and y its monomorphic transformation in $(\mathcal{R}, \oplus, \odot, \overline{0}, \overline{1})$ with $\mathcal{R} \subseteq \mathbb{R}_0^+$. Then*

(i) $\forall u \neq v \in \mathcal{N} \; : \; y_{uv}^* = y_{vu}^*$.

(ii) If Theorem 7.15 (i) and (ii) hold then $y_{uu}^* = \overline{1}$ for all $u \in \mathcal{N}$.

While the first statement is trivial for all symmetric networks, the second is fulfilled when a self relation is considered the most beneficial or least costly relation, e.g. $dist(x)_{uu} = 0$, which is ensured by Theorem 7.15 (i) and (ii). We can thus state a general result for indirect relations derived from a one dimensional semiring.

Corollary 7.18. *Let $x \in \mathcal{B}^{\mathcal{D}}$ and y its monomorphic transformation in $(\mathcal{R}, \oplus, \odot, \overline{0}, \overline{1})$ with $\mathcal{R} \subseteq \mathbb{R}_0^+$. If the semiring fulfills (i)-(iv) in Theorem 7.15, then*

$$pos(u|x) \succcurlyeq pos(v|x) \implies pos(u|\tau(x))) \succcurlyeq pos(v|\tau(x)))$$

Corollary 7.19. *Closeness centrality preserves positional dominance.*

The preservation is strict, since the underlying semiring is strictly monotonic. The following theorem deals with the geodetic semiring.

Theorem 7.20. *Let $x \in \mathcal{B}^{\mathcal{D}}$ and y its monomorphic transformation in the geodetic semiring. If $u \succcurlyeq v$ then the following statements hold.*

(i) $\delta(y^*)_{su} \geq \delta(y^*)_{sv} \quad \forall s \in \mathcal{N} \setminus \{u, v\}$

(ii) $\delta(y^*)_{uv} = 0$

(iii) $\delta(y^*)_{vu} \geq 0$

Proof. (i) is fulfilled due to the definition of the adapted lexicographic order.

(ii) Since $dist(y)_{uv} + dist(y)_{vt} \geq dist(y)_{uv} + dist(y)_{ut} > dist(y)_{ut}$, it holds that $\mathbf{1}_{ut}(v) = 0$ for all $t \in \mathcal{N}$ and thus $\delta(y^*)_{uv} = 0$

(iii) If $x_{uv} = 1$ and there exists a $t \in \mathcal{N}$ such that $x_{ut} = 1$ and $x_{vt} = 0$ then $dist(y)_{vu} + dist(y)_{ut} = dist(y)_{vt}$ holds and therefore $\delta(y^*)_{vu} > 0$. In general, we thus have $\delta(y^*)_{vu} \geq 0$. □

Corollary 7.21. *Betweenness centrality and its variants preserves positional dominance.*

The preservation is non-strict, since the underlying semiring is not strictly monotonic.

For the semiring of infinite sequences, we note that Corollary 7.17(i) also holds since walk counts are symmetric. The following lemma deals with the diagonal elements.

Lemma 7.22. *Let $x \in \mathcal{B}^{\mathcal{D}}$ and y its monomorphic transformation in the semiring of infinite sequences. Then*

$$u \succcurlyeq v \implies w_{uu}^k \geq w_{vv}^k \qquad \forall k \in \mathbb{N}$$

Proof. Let $Q \in \mathcal{D}_{vv}^k$ be a closed walk. Suppose that $(v, i_1) \neq (v, u)$. According to Theorem 7.15, there exists a $P' \in \mathcal{D}_{vu}^k$ such that $P' \bowtie Q$ and $y(P') \geq y(Q)$. Since x is symmetric, it holds that $P' \in \mathcal{D}_{uv}^k$. Thus, there exists $P \in \mathcal{D}_{uu}^k$ with $P \bowtie P'$ and $y(P) \geq y(P')$. It follows that $y(P) \geq y(Q)$.

Now suppose that $(v, i_1) = (v, u)$. We then just need the fact that for any i_j on a closed (v, v)-walk we can define a closed (i_j, i_j)-walk with the same set of dyads. Therefore, we can define a $P \in \mathcal{D}_{uu}^k$, such that $y(P) = y(Q)$.

Together, we obtain that

$$\forall Q \in \mathcal{D}_{vv}^k \;\; \exists P \in \mathcal{D}_{uu}^k \; : \; y(P) \geq y(Q)$$

and therefore $w_{uu}^k \geq w_{vv}^k$ □

The following theorem deals with properties of the weighting sequences γ such that dominance is preserved.

Theorem 7.23. *Let $x \in \mathcal{B}^{\mathcal{D}}$ and y its monomorphic transformation in the semiring of infinite sequences and $u \succcurlyeq v$. Further, let $\gamma = (\gamma_k)_{k=0}^{\infty}$. Then*

(i) $\gamma_0 \geq \gamma_1 \geq \ldots \geq 0 \implies \langle \gamma, w_{ut} \rangle \geq \langle \gamma, w_{vt} \rangle \qquad \forall t \in \mathcal{N} \setminus \{u, v\}$

(ii) $\gamma_k \geq 0 \quad \forall k \in \mathbb{N} \implies \langle \gamma, w_{uu} \rangle \geq \langle \gamma, w_{vv} \rangle$

The proof is omitted since (i) is a generalization of order preserving function for an infinite dimensional vector space [96, 133] and (ii) is a direct consequence from Lemma 7.22.

Corollary 7.24. *(i) Indices based on walk counts preserve positional dominance if γ is positive and non-increasing.*

(ii) Indices based on closed walk counts preserve positional dominance if $\gamma_k \geq 0$ for all $k \in \mathbb{N}$.

With the semiring of enumerated $\psi-$trajectories, we can broaden the class of indices preserving dominance even more.

Corollary 7.25. *Let $x \in \mathcal{B}^{\mathcal{D}}$ and y its monomorphic transformation in the semiring of enumerated ψ-trajectories. The indirect relation $\tau(x) = |y^*|$ preserves dominance if*

(i) ψ conform trajectories are walks or path with length at most $k \in \mathbb{N}$

(ii) ψ conform trajectories are closed walks or path with length at most or exact $k \in \mathbb{N}$

The proofs for both statements are essentially the same as for the semiring of infinite sequences.

7.4 Centrality Indices and Neighborhood-Inclusion

In the last section we presented sufficient conditions for the neighborhood-inclusion preorder to be preserved given a great variety of indirect relations. The implication is that a great deal of centrality indices, e.g. betweenness (and its variants), closeness (and its variants), eigenvector centrality and subgraph centrality (and its variants) to name a few all preserve the neighborhood-inclusion preorder. Besides the relations used in indices, we also showed that dominance is preserved for a great variety of other relations, such as common path optimization problems and indices relying on counts of trajectories up to a length k. Based on these findings, we propose a new characterization for centrality indices.

Proposition 7.26. *Let $c : \mathcal{N} \rightarrow \mathbb{R}_0^+$ be an index. Then, c is a measure of centrality, if and only if it preserves the neighborhood-inclusion preorder.*

The preservation of dominance by neighborhood-inclusion is in line with the preservation of dominance for additive value functions in CM. Alternatives that are preferable in each dimension to another one should not be ranked lower in any preference ranking. In the context of centrality, an actor that has better relations to all other actors than another should never be less central.

Based on Proposition 7.26, we can state the following result about the whole set of potential centrality indices.

Corollary 7.27. *The set of feasible centrality indices according to Proposition 7.26 are the completions of the neighborhood-inclusion preorder.*

The devastating implication of Corollary 7.27 is that the set of centrality indices could potentially be much greater then it already is. Since the problem of counting the completions is #P-complete [41], the actual number can not be determined. However, if the neighborhood-inclusion preorder is complete, the problem is reduced to two cases, stated in the following theorem.

Theorem 7.28. *Let c_1 and c_2 be two centrality indices according to Proposition 7.26. Let $G = (V, E)$ be a graph where the neighborhood-inclusion preorder is complete. Then the following statements hold true.*

(i) If c_1 and c_2 strictly preserve the neighborhood-inclusion preorder, then

$$\begin{aligned} c_1(u) > c_1(v) &\iff c_2(u) > c_2(v) \qquad \forall u, v \in V \quad \textit{and} \\ c_1(u) = c_1(v) &\iff c_2(u) = c_2(v) \qquad \forall u, v \in V. \end{aligned}$$

(ii) If either c_1 or c_2 preserves the neighborhood inclusion non-strictly, then

$$c_1(u) \geq c_1(v) \iff c_2(u) \geq c_2(v) \qquad \forall u, v \in V.$$

Note, that Theorem 7.28 also substantiates the star property in a formal way. Star shaped graphs are completely ordered since the center dominates all leafs which in turn mutually dominate each other, i.e. they are structurally equivalent.

Incompatible Measures and the Star Property

We have shown that a vast amount of existing centrality measures preserve the neighborhood-inclusion preorder. The question is, if there are indices that do not meet this condition and if so, why it is reasonable to exclude them from the set of centrality measures.

A simple index which generally does not fulfill Proposition 7.26 is the number of 2-paths starting at a vertex. As a simple illustrative example consider a star graph, where the number of 2-paths for the center and leaf nodes are given by

$$c_{2path}(u) = \begin{cases} 0 & u = \text{center} \\ n-2 & \text{otherwise.} \end{cases}$$

Obviously, this measure also does not fulfill the star property. The next example, however, shows that preserving dominance is a tighter requirement than the star property. The scores of the hyperbolic index, defined in Section 5.2,

$$c_{hyp}(u) = ccoef(u) \left[\sum_{v \in N[u]} \cosh(A^{[u]})_{vv} \right]$$

on a star shaped network can be calculated as a function of the number of vertices n with

$$c_{hyp}(u) = \begin{cases} \frac{2}{n}\left[2\cosh(\sqrt{n-1}) + (n-2)\right] & u = \text{center} \\ 2\cosh(1) & \text{otherwise.} \end{cases}$$

Since $\cosh(x)$ is monotonically increasing faster than $2/n$ is decreasing, the center of the star attains the highest value for all $n > 2$. Although the practical relevance of this measure was already questioned, we could argue that it is a centrality index according to the star property.

However, it can be shown that the hyperbolic index generally does not preserve the neighborhood-inclusion preorder. This is i.a. due to the scaling with the clustering coefficient. It is easy to see that the clustering coefficient itself does not preserve neighborhood-inclusion. Admittedly, it is not intended to be a measure of centrality, yet it is used for several indices as a scaling factor [123].

A prominent index that does not necessarily preserve dominance is given by Bonacich's β-centrality. Its representational sequence is defined

$$\gamma = [0, 1, \beta, \beta^2, \ldots, \beta^{k-1}, \ldots],$$

where $|\beta| < \frac{1}{\lambda_1}$ to ensure convergence. If β is chosen to be greater than zero, Bonacich points out that

> *"[...], $c_{\alpha,\beta}$ is a conventional centrality in which each unit's status is a positive function of the statuses of those with which it is in contact." [25, p. 1170]*

Indeed, for $\beta > 0$ the requirements of Proposition 7.26 are fulfilled, since γ is monotonically decreasing and $\gamma_k \geq 0$ for all $k \in \mathbb{N}_0$ so that the requirements of Corollary 7.24 are fulfilled. However, this does no longer hold true if β is negative. In this case, γ is an alternating sequence such that the preservation of the neighborhood-inclusion preorder can not be guaranteed. Bonacich describes the meaning of a negative β in the context of bargaining situations.

> *"In bargaining situations, it is advantageous to be connected to those who have few options; power comes from being connected to those who are powerless. Being connected to powerful others who have many potential trading partners reduces one's bargaining power." [25, p. 1171]*

His work was based on the findings of Cook et al., who showed that power does not equal centrality in exchange network [50]. Having better relationships in this situation is not an indicator for better bargaining positions. Therefore, β-centrality with a negative β is no admissible centrality measure but rather a measure of power in exchange networks.

7.5 Summary and Discussion

This chapter dealt with the theoretical principles for the preservation of dominance by centrality indices. We now have sufficient conditions at hand to show that

$$\begin{aligned} u \succcurlyeq v &\iff pos(u|x) \succcurlyeq pos(v|x) \\ &\implies pos(u|\tau(x))) \succcurlyeq pos(v|\tau(x))) \\ &\implies c_\tau(u) \geq c_\tau(v). \end{aligned}$$

holds for a great variety of indirect relations $\tau(x)$. This lead to a new characterization of centrality concepts by enforcing preservation on indices. From the point of MCDM, the requirenment is a natural extension for the preservation of dominance in the additive value function model.

Compared to the conceptualizations of centrality given in Section 3.2, our characterization can be positioned between the weak star property and the restrictive axiomatic systems. We have shown that fulfilling the star property is not enough to exclude contrived indices. We later introduce a super class of star shaped networks which provides a stronger requirement for centrality indices.

Our characterization encompasses all prominent indices and is not restricted to selected groups of indices as for the axiomatic approaches. Further, it focuses solely on induced rankings, which we have argued to be more important than the actual centrality scores. Obviously, Proposition 7.26 could also be formulated as an axiom for centrality. The resulting axiomatic definition would, however, be of a different nature then previous systems. The discussed axiomatic systems in Chapter 3 are *normative*. As in other contexts, e.g. *social choice theory*, a normative axioms does not imply that every index behaves according to them. Instead, they are a basis for suggesting a behavior that we would like indices to follow. In other words, enforcing properties on indices. Proposition 7.26 could be seen as a *descriptive* axiom. That is, we observed a certain behavior all indices seem to follow, i.e. the preservation of the neighborhood-inclusion preorder, and formulate it as an axiom.

Whether or not a centrality index preserves the neighborhood-inclusion preorder can be verified without much mathematical effort. However, it prevents us from deriving any general statements about theoretic conditions that indices have to fulfill. The generalization of indirect relations by means of semirings provides a holistic view of relations on networks and allows for deducing sufficient conditions when the neighborhood-inclusion preorder is preserved. Their application is by no means a novel approach. Semirings are commonly used in order to produce generic algorithms for path finding problems and called *path algebras* [88, 138]. Here, we are not interested in designing algorithms but rather defining algebraic structures as a theoretic basis for indirect relations used in the context of network centrality. Further, they facilitate theoretical investigations, neglecting concerns about computational complexity.

The given sufficient conditions under which neighborhood-inclusion is preserved show that we are faced with an infeasible number of possibilities to craft new indices. From a theoretical perspective, any index based on walk counts with a monotonic representational sequence γ can be termed a measure of centrality. Even worse, if γ comprises a free parameter we can tune indices ad infinitum to obtain desired outcomes. This is, however, not in agreement with our arguments given in the previous chapter. Centrality should be seen as a procedure of measurement and not as a tool to uncover patterns in data.

CHAPTER 8

Uniquely Ranked Graphs

> *"Threshold graphs have a beautiful structure and possess many important mathematical properties such as being the extreme cases of certain graph properties."*
>
> – Mahadev & Peled, 1995

8.1 DEFINITIONS AND PROPERTIES OF THRESHOLD GRAPHS

In Chapter 7, we introduced the preservation of the neighborhood-inclusion preorder as a new characterization for centrality indices. An important implication was given in Theorem 7.28, i.e. there is only one possible ranking of vertices if the preorder is complete. We discuss a class of graphs which fulfill this property in the following.

Definition 8.1. *A graph $G = (V, E)$ is called a threshold graph if the neighborhood-inclusion preorder is complete. The set of all threshold graphs with n vertices is denoted by $\mathcal{T}_n$.*

Definition 8.1 ensures that all centrality indices in the sense of Proposition 7.26 induce the same ranking according to Theorem 7.28 on a threshold graph. Threshold graphs and their applications have been studied extensively in the literature [45,56,94]. However, they were never considered in the context of network centrality.

Giving a detailed introduction on this graph class is out of the scope of this thesis. An extensive review on topics related to threshold graphs can be found in [129]. Here, we focus on properties with implications for network centrality. Star graphs, e.g. are a proper subclass of threshold graphs such that we can strengthen the star property by requiring agreement of centrality rankings on threshold graphs.

Alternative characterizations for threshold graphs are given in the following theorem.

Theorem 8.2. *Let $G = (V, E)$ be a simple undirected graph. The following statements are equivalent*

(i) G is a threshold graph.

(ii) There exist vertex weights $\omega : V \to \mathbb{R}_0^+$ and a threshold $t' \geq 0$ such that $\{u, v\} \in E \iff \omega(u) + \omega(v) > t'$.

(iii) G can be constructed from the one-vertex graph by repeatedly adding an isolated vertex or a dominating vertex which is connected to every other vertex that has been added before.

(iv) G does not contain an induced P_4, C_4 or $2K_2$.

(v) G is a split graph and the neighborhood of the independent set is nested.

The proof can be found in [129].

The degree sequence of a threshold graph (henceforth *threshold sequence*) is *unigraphic*, i.e. the structure of a threshold graph is uniquely determined by its degree sequence up to node relabeling. Theorem 8.2(v) implies that threshold graphs have a perfect *core-periphery structure*, i.e. $V = K \cup I$, where the vertices in K form a clique and I is an independent set. Theorem 8.2(iii) can be exploited to store threshold graphs very efficiently.

Definition 8.3. *Let $G = (V, E) \in \mathcal{T}_n$. The binary creation sequence $B_n = b_1 b_2 \dots b_n$ of G is defined as*

$$b_i = \begin{cases} 1 & i \text{ is a dominating vertex} \\ 0 & \text{otherwise} \end{cases} \qquad 1 \leq i \leq n$$

The value of b_1 can be set arbitrarily, yet we choose to set it to 1.

Some simple examples for threshold graphs and their binary creation sequences are shown in Figure 8.1.

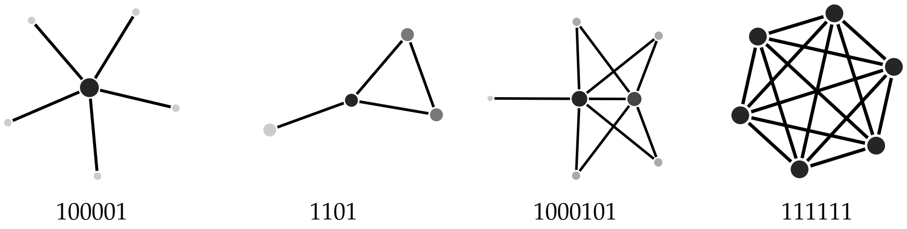


FIGURE 8.1: Examples of threshold graphs and their binary sequences.

Besides its compact form, the binary creation sequences posses another advantageous property.

Corollary 8.4. *Let B_n be a binary creation sequence of a threshold graph $G = (V, E)$. Then the following holds true*

$$b_i = b_{i+1} \iff i \text{ and } i+1 \text{ are structurally equivalent} \qquad \forall\, 2 \leq i \leq n-1.$$

Taking advantage of Corollary 8.4, we can compress the representation even more by solely giving the sizes of the equivalence classes in a *run-length encoding* [92]. This representation does not come with any loss of information. In fact, the sequence is enough to compute nearly all structural features and centrality scores in linear time [92].

8.2 Threshold Distance Measures

Although threshold graphs play important roles in a variety of disciplines, it is rather unrealistic to encounter them in real-world social networks. Yet, we can determine if an arbitrary graph has a similar structure as a threshold graph by several means. Similarity or distance measures are of particular importance for our definition of centrality indices. Graphs that are structurally close to a threshold graph exhibit a close to complete neighborhood-inclusion preorder, such that the number of possible vertex rankings is reduced significantly. We examine this observation in more detail in Chapter 9. Here, we give a brief overview of potential measures to quantify distances between an arbitrary graph and the class of threshold graphs.

Edit Distance

Distances between graphs are commonly calculated with the *graph edit distance* [84].

Definition 8.5. *Let $G = (V, E(G))$ and $H = (V, E(H))$ be two graphs with the same vertex set V. The edit distance between G and H is defined as*

$$e(G, H) = |E(G) \triangle E(H)| \ ,$$

where $\triangle$ is the symmetric difference.

Informally, the edit distance is the number of edges that have to be added and deleted to turn one graph into another. Special cases of graph editing for our context are given in the following definition.

Definition 8.6. *Let $G = (V, E)$ be a graph. and $F \subseteq V \times V$. Three instances of threshold distances are*

(i) threshold editing:

$$t_e(G) = \min\{|F| : H = (V, E\triangle F) \text{ is a threshold graph}\}$$

(ii) threshold completion:

$$t_+(G) = \min\{|F| : E \cap F = \emptyset \ \wedge \ H = (V, E\triangle F) \text{ is a threshold graph}\}$$

(iii) threshold deletion:

$$t_-(G) = \min\{|F| : F \subseteq E \ \wedge \ H = (V, E\triangle F) \text{ is a threshold graph}\}$$

The computational complexity of threshold editing has long been stated as open [142]. Very recently, it was shown that determining $t_e(G),\ t_+(G)$ and $t_-(G)$ is $\mathcal{NP}-$ complete even if G is a split graph [59].

It is ,however, possible to obtain a specially designed threshold graph from any graph with minimal edit distance based on rank-1 approximations.

Definition 8.7. *Let $G = (V, E)$ be a graph and let $A = X\Lambda X^T$ be the spectral decomposition of its adjacency matrix. The t-binarized rank-1 approximation is defined as*

$$A_{uv}^{[t]} = \begin{cases} 1 & x_1(u)\lambda_1 x_1(v) > t \\ 0 & \textit{otherwise} \end{cases}$$

with $t \in \mathbb{R}_0^+$.

The associated graph of a t-binarized rank-1 approximation $A^{[t]}$ is given by $G_t = (V, E_t)$. The main result about t-binarized rank-1 approximations is given in the following Theorem.

Theorem 8.8. *Let A be the adjacency matrix of a graph $G = (V, E)$. A t-binarized rank-1 approximation of A is the adjacency matrix of a threshold graph for all $t \in \mathbb{R}^+$.*

Proof. Let $t \in \mathbb{R}^+$. Definition 8.7 implies that

$$\{u, v\} \in E_{[t]} \iff x_1(u)\lambda_1 x_1(v) > t\ .$$

Let $\alpha \in \mathbb{R}_0^+$, such that $\alpha > \max\left\{1/\min_{u\in V} x_1(u), \sqrt{\lambda_1/t}\right\}$. The following transformations are equivalent according to the Theorem of Perron Frobenius:

$$\begin{aligned} & x_1(u)\lambda_1 x_1(v) > t \\ \iff & x_1(u)x_1(v) > \frac{t}{\lambda_1} \\ \iff & \alpha^2 x_1(u)x_1(v) > \alpha^2 \frac{t}{\lambda_1} \\ \iff & \log(\alpha x_1(u)) + \log(\alpha x_1(v)) > 2\log(\alpha) + \log(t) - \log(\lambda_1) \end{aligned}$$

Setting $\omega(u) = \log(\alpha x_1(u))$ and $t' = 2\log(\alpha) + \log(t) - \log(\lambda_1)$ we obtain

$$\{u, v\} \in E \iff \omega(u) + \omega(v) > t'\ .$$

The choice of α ensures, that $\omega(u) > 0$ for all $u \in V$ and $t' > 0$ for all $t \in \mathbb{R}_0^+$. Hence, condition (ii) of Theorem 8.2 is fulfilled and $G_{[t]} = (V, E_{[t]})$ is a threshold graph for every $t \in \mathbb{R}_0^+$. □

Definition 8.9. *Let $G = (V, E)$ be a graph. The rank-1 threshold distance is defined as*

$$t_1(G) = \min_{t\in\mathbb{R}_0^+} \left\{E \triangle E_{[t]}\right\}$$

where $E_{[t]}$ is the edge set of a t-binarized rank-1 approximation graph.

Corollary 8.10. *Let $G = (V, E)$. It holds that $t_1(G) \geq t_e(G)$.*

Although $t_1(G)$ is easy to obtain, Corollary 8.10 signifies that it is not optimal in the sense of minimum number of edits. An example for this case is depicted in Figure 8.2.

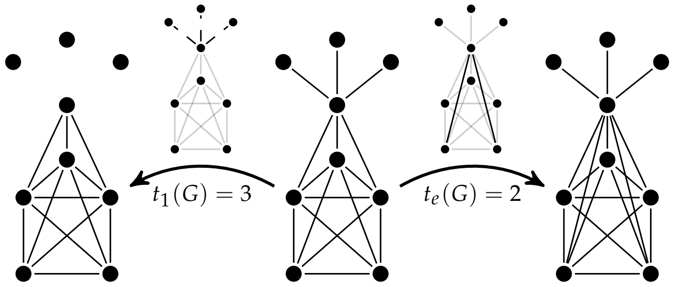


FIGURE 8.2: Illustration of the difference between $t_1(G)$ and $t_e(G)$. From the original graph (middle) three edges have to be deleted according to $t_1(G)$ whereas two edges have to be added according to $t_e(G)$ in order to obtain a threshold graph with the respective minimal distance.

Edge Rotation Distance

Since the problem of threshold editing is $\mathcal{NP}-$ complete, we have to rely on measures that quantify the distance to a threshold graph by different means. One conceivable alternative is given in the following definition.

Definition 8.11. *Let $G = (V, E)$ be a graph. The edge rotation distance $t_r(G)$ is the minimum number of edge rotations, i.e. changing one endpoint of an edge, which are necessary to turn G into a threshold graph.*

Since an edge rotation is equivalent to two edits, we have the following corollary.

Corollary 8.12. *For all graphs $G = (V, E)$,*

$$t_e(G) \leq 2t_r(G)$$

holds.

The edge rotation distance is closely related to two existing measures of non-thresholdness of degree sequences. The first is due to Hammer et al. [94].

Definition 8.13. *Let $d(G)$ be the degree sequence of a graph $G = (V, E)$. The threshold gap is defined as*

$$t_{tg}(d(G)) = \min_{G^* \in \mathcal{T}_n} \frac{1}{2} \|d(G) - d(G^*)\|_1 \,,$$

where $\|\cdot\|_1$ denotes the L_1-norm.

Further theoretical details of this measure can be found in [94]. The second measure operating on degree sequences, called *majorization gap*, was introduced by Arikati and Peled [7]. We first need the definition of a sequence closely related to degree sequences.

Definition 8.14. *The corrected conjugate sequence $d'(G)$ of a degree sequence $d(G)$ is given by the formula*

$$d'_k = |\{i : i < k \,\wedge\, d_i \geq k-1\}| + |\{i : i > k \,\wedge\, d_i \geq k\}| \qquad 1 \leq k \leq n.$$

Definition 8.15. *For a degree sequence $d(G)$ and its corrected conjugated sequence $d'(G)$,*

$$t_{mg}(d(G)) = \frac{1}{2}\sum_{k=1}^{n} \max\left\{d'_k - d_k, 0\right\}$$

is the majorization gap.

The majorization gap counts the number of *reverse unit transformations* in order to turn a degree sequence into a threshold sequence [129]. A reverse unit transformation is equivalent to changing two entries in the degree sequence. Note, that although this operation is equivalent to rotating an edge, it is not equivalent to the edge rotation distance. Degree sequences do not uniquely determine a graph, such that an edge rotation from a reverse unit transformation is not unambiguously defined. The following theorem is due to Mahadev & Peled [129]

Theorem 8.16. *For every degree sequence $d(G)$,*

$$t_{tg}(d(G)) = t_{mg}(d(G))$$

The proof can be found in [129].

Corollary 8.17. *Let $d(G)$ be a degree sequence of a graph $G = (V, E)$. Then*

$$t_r(G) \geq t_{mg}(d(G))$$

An example that $t_{mg}(d(G))$ is just a lower bound for the edge rotation distance is given in Figure 8.3.

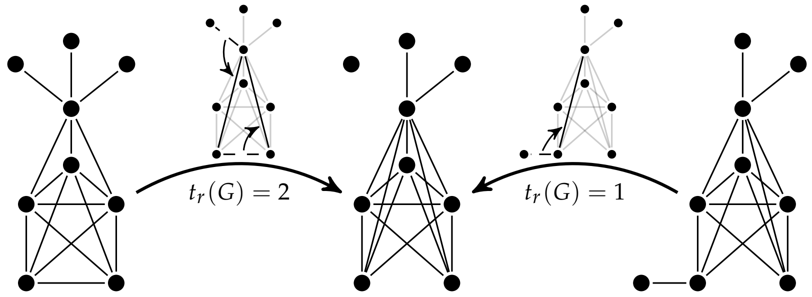


FIGURE 8.3: Illustration of the ambiguity of t_{mg}. The left and right graphs have the same degree sequence [6, 5, 5, 5, 4, 4, 1, 1, 1] and the same majorization gap of 1. However, two rotations are necessary for the left graph to obtain the threshold graph in the middle and only one for the right graph.

Since we do not have a reliable method at hand to calculate the rotational distance exactly, we apply the majorization gap in the forthcoming chapter. We have to keep in mind, however, that it is just an approximation for the rotational distance.

CHAPTER 9

Correlation and Threshold Distance

"Correlation is a minefeld for the unwary."

– Embrechts et al., 2001

9.1 INTRODUCTION

The prevalent opinion about correlation among centrality indices is best described in the words of Valente et al.:

> *"If centralities are not highly correlated, they indicate distinctive measures, associated with different outcomes."* [186, p. 1]

A weak correlation between two indices thus implies that they measure importance on different structural levels. Therefore, it seems convenient to employ a correlation analysis to justify new indices [15,144]. A weak correlation with existing measures presumably signifies that the new index evaluates structural importance on a different level proving its novelty.

A second line of research deals with the questions of how correlated measures of centrality in general are [11,21,119,121,127,167,186]. These studies intend to unveil similarities of indices with respect to the anticipated outcome, hence investigating their redundancy. Indices that measure the same structural importance are alleged to be used interchangeably and preferably the computationally less expensive one should be chosen [121]. Most of these studies tested the correlation on a small set of networks such that an influence of the network structure is ruled out beforehand. As a consequence thereof, results often contradict each other. Lee finds that "... the degree and the betweenness are correlated much strongly [sic] than other centrality measures" [119, p. 6], while Lozares et al. find that the correspondence "are weaker or partial among Degree and Betweenness" [127, p. 222].

A third type of studies deals with the robustness of indices towards perturbation of a network [29,51,73,148,157]. The goal is to examine the stability of results when dealing with missing data or sampled networks. Work in this category mostly use networks from random graph models and it was found that the network topology actually has a non-negligible effect on the reliability of centrality.

Dependencies of centralities can also be proven analytically. Benzi & Klymko examine a class of parametrized walk-based measures and showed that parameters can be tuned such that i.a. subgraph centrality and total communicability interpolate between degree and eigenvector centrality in the limit cases of the parameters [16]. An important role for the convergence plays the spectral gap. The bigger it is, the stronger should the correlation between walk based measures be.

In all mentioned cases, correlations are almost exclusively derived with the Pearson correlation coefficient, undoubtedly the most commonly used correlation coefficient across various disciplines. This measure, however, makes strong assumptions on the underlying data such that it is unreasonable choice in many cases [60].

In contrast to the received meaning of correlation in the literature, we show in this chapter that correlation is contingent on the network structure. In particular, the distance to the closest threshold graph serves as an explanation for correlations and is thus not associated with the definition of indices. In Section 9.3, we provide a correlation analysis on two random graph models, the Erdős-Renyi model and the Barábasi-Albert model for preferential attachment. This section is also used to demonstrate the sensitivity of results when different correlation measures are chosen and to illustrate why testing on generated data is not advisable.

Section 9.4 is devoted to an analysis on real-world networks. We use a set of 400 self compiled protein interaction networks as well as 60 social networks. The latter were used by Valente et al. [186]. Additionally we point to several numerical issues that may have an influence on the correlation when indices have an enormous value range, questioning their reliability.

9.2 Material and Methods

Random Graph Models

The Erdős-Rényi, or $\mathcal{G}(n, p)$ model is the simplest model for random graphs. Each dyad is realized as an edge with probability p, independently of each other [61, 86].

The Barabási-Albert model is an algorithm to generate random networks using a preferential attachment mechanism [9]. Starting from an initial network with n_0 vertices, new nodes are added iteratively and connected to $m_1 \leq n_0$ existing vertices, where the probability to connect to a specific vertex is proportional to its degree. This mechanism eventually leads to what is referred to as a scale-free network with a power law degree distribution as compared to a Poisson distribution for $\mathcal{G}(n, p)$.

Both models suffer from several shortcomings when compared to heterogeneous real-world networks. Nevertheless, they are commonly used to test new algorithms or hypotheses connected to certain network structures due to their simplicity.

For our analysis, we sample the parameters for both models. After choosing the number of vertices n uniformly at random from the interval $[100, 1000]$, we sample p or m_1 such that we obtain a network with density ≤ 0.3.

Real-world Networks

We use two sets of networks derived from real-world phenomena. The first comprises PINs of 400 organisms (henceforth protein dataset) taken from the String Database (version 9.0) [182]. The networks are constructed with interactions that have a confidence score greater than 950. We only consider the biggest component of each network. Table 9.1 summarizes some basic statistics of the dataset.

	No. of vertices	No. of edges	density
mean	388	1303	0.37
min	28	70	0.001
max	3598	8890	0.69

TABLE 9.1: Network statistics of the protein dataset.

The second set consists of 60 social networks (henceforth Valente's dataset) obtained from a correlation study by Valente et al. [186]. Table 9.2 summarizes some basic statistics of the dataset.

	No. of vertices	No. of edges	density
mean	60	525	0.19
min	32	51	0.04
max	159	7009	0.94

TABLE 9.2: Network statistics of Valente's dataset.

Centrality Indices

As in Chapter 5, we choose a small subset of existing indices for our analysis (Abbreviation used in figures in parentheses). We consider degree (DC), betweenness (BC), closeness (CC), eigenvector (EC) and subgraph centrality (SC), as well as information centrality (INF). Additionally, we use total communicability (TC) for a comparison of walk based indices.

Threshold Distance

We discussed several possibilities to quantify the distance of an arbitrary graph to its closest threshold graph in Chapter 8. We have seen, that most measures are either computationally too complex or only give a rough approximation of the actual distance. We here choose the majorization gap as an approximation of the rotational distance. To make results comparable, we normalize it by the number of edges in each graph. A maximum majorization gap of 1 thus means that we have to rotate all edges in the graph and the minimum of 0 is only reached for actual threshold graphs.

Correlation Coefficients

A vast amount of correlation coefficients can be found in the literature, derived from different assumptions about underlying data [131]. The most commonly used coefficient is *Pearson product-moment correlation coefficient* (henceforth Pearson's ρ), which measures the strength of a linear association between two variables x and y. It is defined as

$$\rho(x,y) = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \overline{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \overline{y})^2}},$$

where $\overline{x}$ and $\overline{y}$ are the mean of x and y respectively. Due to its popularity, it was almost exclusively used in correlation studies for centrality indices. However, the assumption of a linear dependence between the scores of indices is too strong and nonlinear dependencies are not adequately captured with Pearson's ρ. A common workaround, not only limited to network analytic studies is to use the logarithmic scores, although ρ is not invariant under nonlinear transformations [60]. That is, drawn conclusion from correlations obtained by logarithmic scores are potentially fallacious. Further, it is only fully applicable if two variables have a joined normal distribution [60], which can not be universally assumed for centrality indices. For these reasons, we employ Pearson's ρ on the scores and the logarithmic scores (denoted as $\rho_{\log}$) only for illustrative purposes.

Since we are solely interested in induced rankings of indices, it is more convenient to use rank based correlation measures. Five different configurations of pairs (i,j) can be observed when comparing two rankings induced by indices c_1 and c_2:

(i) *concordance* if $c_1(i) > c_1(j)$ and $c_2(i) > c_2(j)$

(ii) *discordance* if $c_1(i) > c_1(j)$ and $c_2(i) < c_2(j)$

(iii) *tie* if $c_1(i) = c_1(j)$ and $c_2(i) = c_2(j)$

(iv) *right tie* if $c_1(i) \neq c_1(j)$ and $c_2(i) = c_2(j)$

(v) *left tie* if $c_1(i) = c_1(j)$ and $c_2(i) \neq c_2(j)$

Measures of rank correlation rely on aggregating the appearance of subsets of these configuration, depending if ties are assumed to be present or not and normalizing them [89,106,158]. For ease of exposition, we assume that r and s are variables containing the scores of two different centrality indices. Further, we define an inner product

$$\langle r,s \rangle = \sum_{i<j} \operatorname{sign}(r_i - r_j) \operatorname{sign}(s_i - s_j),$$

where

$$\operatorname{sign}(x) = \begin{cases} 1 & x > 0 \\ 0 & x = 0 \\ -1 & x < 0\,. \end{cases}$$

A simple measure based on the counts of (i) and (ii) is *Goodman and Kruskal's* γ. It can be defined as

$$\gamma(r,s) = \frac{\langle r,s \rangle}{\sum\limits_{i<j} \left|\operatorname{sign}(r_i - r_j)\operatorname{sign}(s_i - s_j)\right|}.$$

The downside of this measure is, that it neglects potentially present ties in the ranking. Yet, the measure is useful to compare fine grained and coarse grained indices, i.e. when many left or right ties are present. A measure that accounts for ties is version *b* of *Kendall's rank correlation coefficient* (henceforth Kendall's τ). Together with an induced norm

$$\|r\| = \sqrt{\langle r,r \rangle}\,,$$

it can then be defined as

$$\tau(r,s) = \frac{\langle r,s \rangle}{\|r\| \cdot \|s\|}.$$

This definition is equivalent to Kendall's original work [107, 191].

However, Kendall's τ is also not unconditionally guaranteed to be free from defects. Particularly, problems may arise when dealing with huge networks. The scores of central vertices tend to be highly correlated in many reasonable rankings, yet most of the peripheral vertices are ranked in slightly different ways, introducing a large amount of noise and leading to a low value for τ. This phenomenon motivates the use of weighted correlation measures, which correct for this issue. Numerous weighted approaches have been proposed in the literature [115, 174, 200], where we use a recently introduced version by Vigna [191], due to its scaling properties for large networks. The main idea is to define a weighted inner product

$$\langle r,s \rangle_w = \sum_{i<j} \operatorname{sign}(r_i - r_j)\operatorname{sign}(s_i - s_j) w(i,j)$$

and accordingly a weighted correlation coefficient

$$\tau_w(r,s) = \frac{\langle r,s \rangle_w}{\|r\|_w \cdot \|s\|_w},$$

There are several mathematical adjustments to be made to this formulation, which can be found in the original work of Vigna. We here use an additive hyperbolic weight function, i.e.

$$w(rank(i), rank(j)) = \frac{1}{rank(i) + rank(j) + 1}.$$

9.3 Results for Random Graph Models

Erdős-Rényi Model

Figure 9.1 shows the results on the 1000 $\mathcal{G}(n,p)$ graphs. We can observe that all graphs have a large distance from threshold graphs. However, all graphs still have a high correlation for all indices. Although we do not have a comprehensive explanation for this phenomenon, there is a good case to believe that the reasons are given by the fairly homogeneous structure of the networks. Further, we do not observe any differences between Kendall's τ and its weighted version. The results of Goodman and Kruskal's γ are not shown since they are not distinctive of the results with Kendalls' measures.

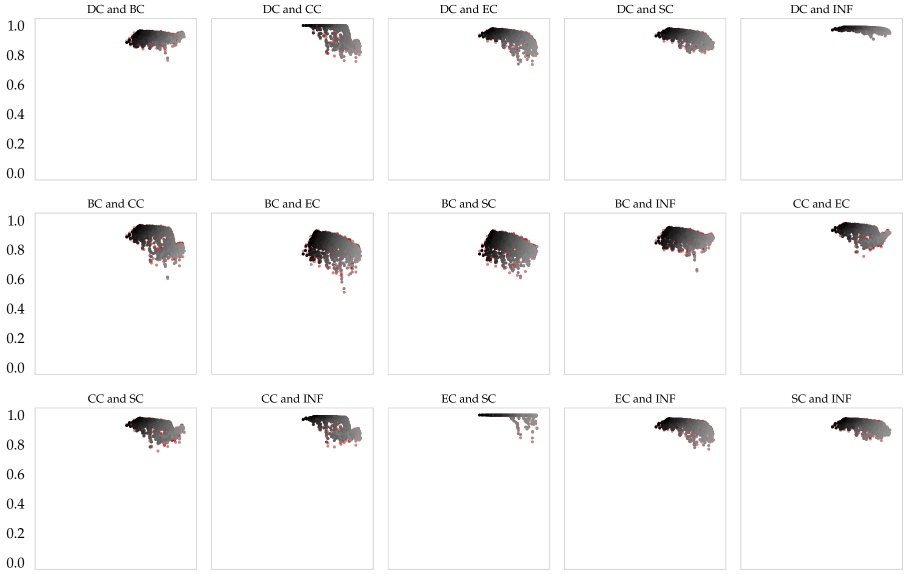
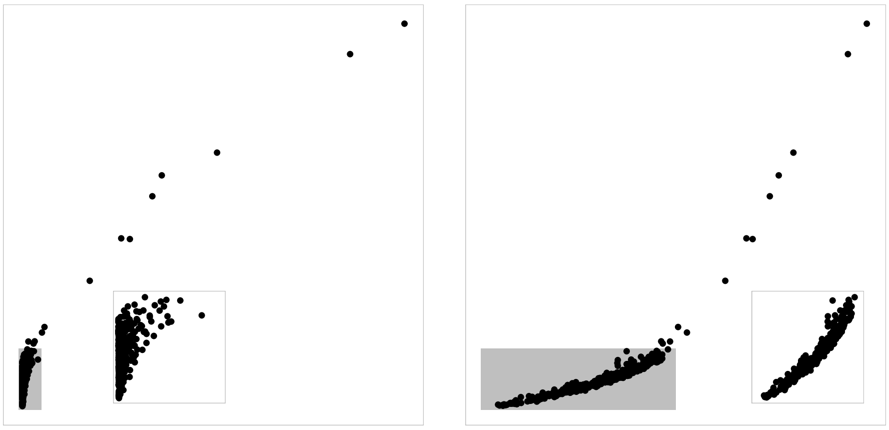


FIGURE 9.1: Correlation of several indices (y-axis) and normalized majorization gap (x-axis) for 1000 different $\mathcal{G}(n,p)$ graphs. Correlation is measured with τ (shades of grey, representing the density of the network) and τ_ω (red).

Barábasi-Albert Model

Figure 9.2 illustrates the potential contradictory results that may arise for different coefficients. While Pearson's ρ suggests a higher association between betweenness and eigenvector centrality, its logarithmic version and the two versions of Kendall's τ give a higher value for closeness and eigenvector centrality. Additionally, the scores have a high variability for both comparisons.



(A) Betweenness and eigenvector

(B) Closeness and eigenvector

	ρ	$\rho_{\log}$	τ	τ_ω
betweenness and eigenvector	0.91	0.59	0.26	0.61
closeness and eigenvector	0.81	0.86	0.89	0.92

FIGURE 9.2: Illustration for contradictory results obtained by correlation coefficients. The figure shows the scatter plot between (a) betweenness and eigenvector centrality and (b) closeness and eigenvector centrality on a network created with the Barábasi-Albert model ($n = 500$ and $m_0 = 12$).

Figure 9.3 shows the results on the 1000 randomly generated preferential attachment graphs. We can observe a significant difference in the results of Kendall's τ and its weighted version. It becomes apparent that the weak correlation of betweenness with other indices is mostly due to the ranking noise in the lower ranks. Also, subgraph centrality is perfectly correlated with eigenvector centrality in all considered networks.

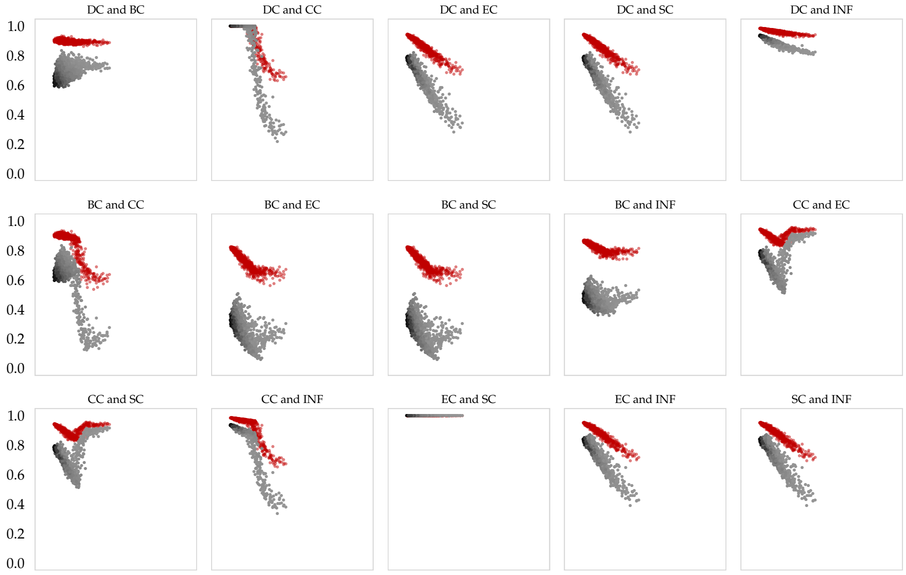


FIGURE 9.3: Correlation of several indices (y-axis) and normalized majorization gap (x-axis) for 1000 different preferential attachment graphs. Correlation is measured with τ (shades of grey, representing the density of the network) and τ_ω (red).

An interesting behavior can be observed for closeness correlated with eigenvector and subgraph centrality. The correlation seems to decay linearly at first, yet it increases again for a greater distance. Goodman and Kruskal's γ is identical with Kendall's τ for all but degree and information centrality where γ is one for all networks.

9.4 Results for Real Networks

Results for the protein dataset are shown in Figure 9.4. Since the weighted version of Kendall's τ produced correlation scores close to the unweighted version, we omit the results. We can observe a strong decay of the correlation with an increasing normalized majorization gap for most pairs of indices, except for correlations including betweenness.

In [15,16] it was analytically shown, that the correlation between subgraph centrality and total communicability with eigenvector centrality depends on the size of the *spectral gap*, i.e. the difference of the principal and the second largest eigenvalue. Figure 9.5 indicates that there is no strong coherence between correlation and the spectral gap on the protein dataset.[1].

[1] As a reminder, we use λ_2/λ_1 as spectral gap to keep values in the interval $[0, 1]$ The closer to one the fraction is, the smaller is the spectral gap.

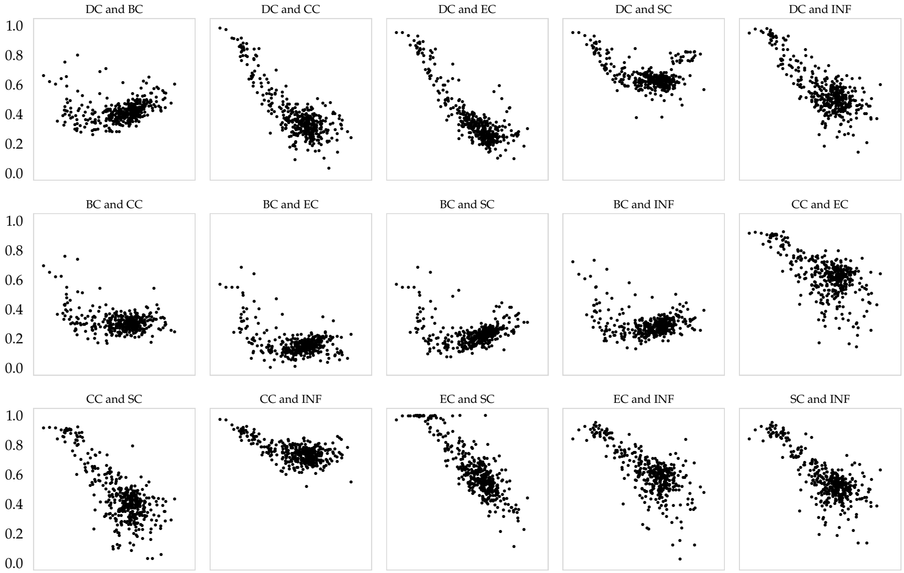


FIGURE 9.4: Correlation of several indices (y-axis) and normalized majorization gap (x-axis) for the protein dataset. Correlation is measured with Kendall's τ.

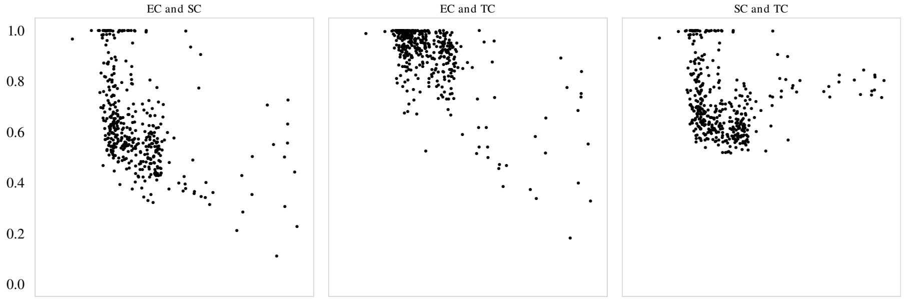


FIGURE 9.5: Correlation of several indices (y-axis) and spectral gap (x-axis) for the protein dataset. Correlation is measured with Kendall's τ.

We also checked for numerical stability of the results. In particular subgraph centrality has an enormous value range due to the exponential function and we might expect numerical problems when calculated. Figure 9.6 indeed shows that we obtain differing results for subgraph centrality when centrality scores are rounded to 8 digits thus indicating stability issues.

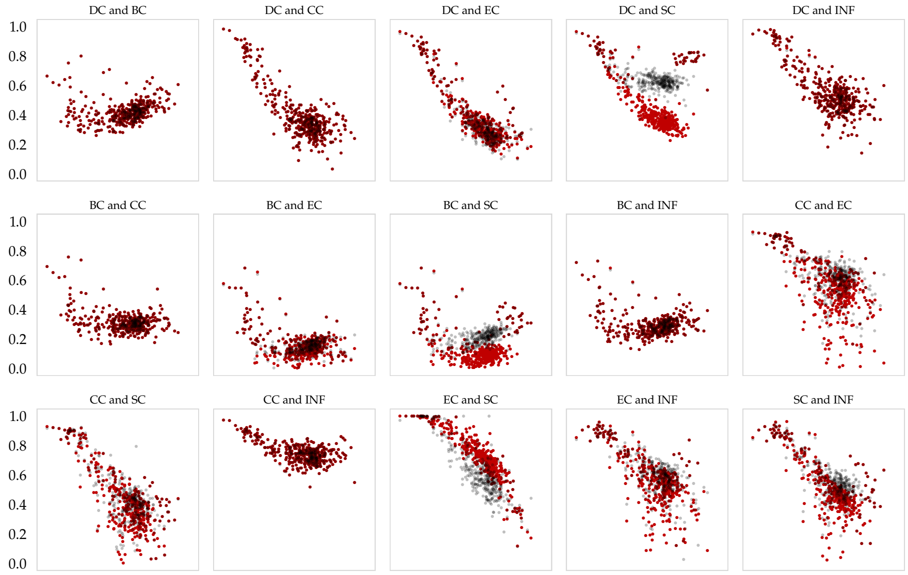


FIGURE 9.6: Correlation of several indices (y-axis) and normalized majorization gap (x-axis) for the protein dataset. Correlation is measured with Kendall's τ on unaltered scores (black) and rounded to 8 digits (red).

The results for Valente's dataset shown in Figure 9.7 are comparable to those from the protein dataset. Notable is the apparent decaying correlation between subgraph and eigenvector centrality for a low majorization gap. However, this is again due to numerical stability issues depicted in Figure 9.8.

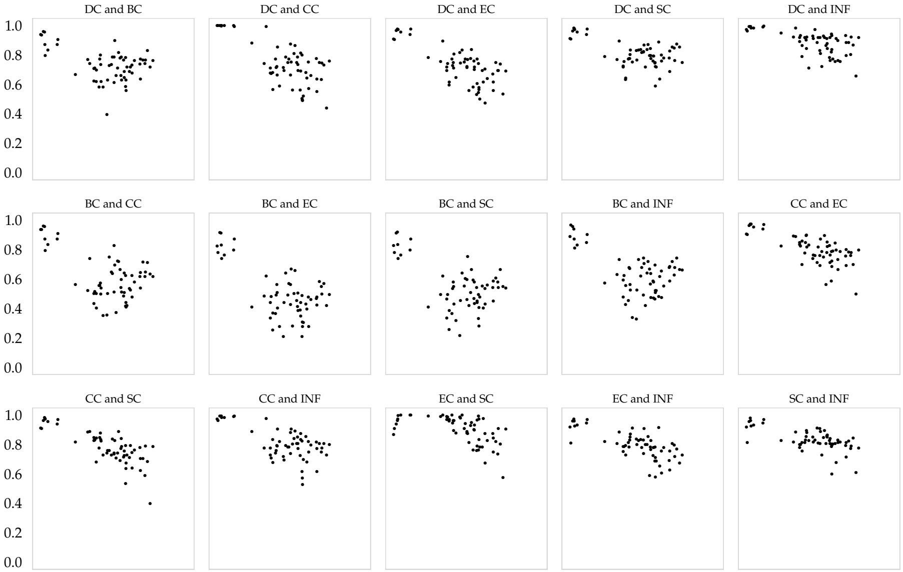


FIGURE 9.7: Correlation of several indices (y-axis) and normalized majorization gap (x-axis) for Valente's dataset. Correlation is measured with Kendall's τ.

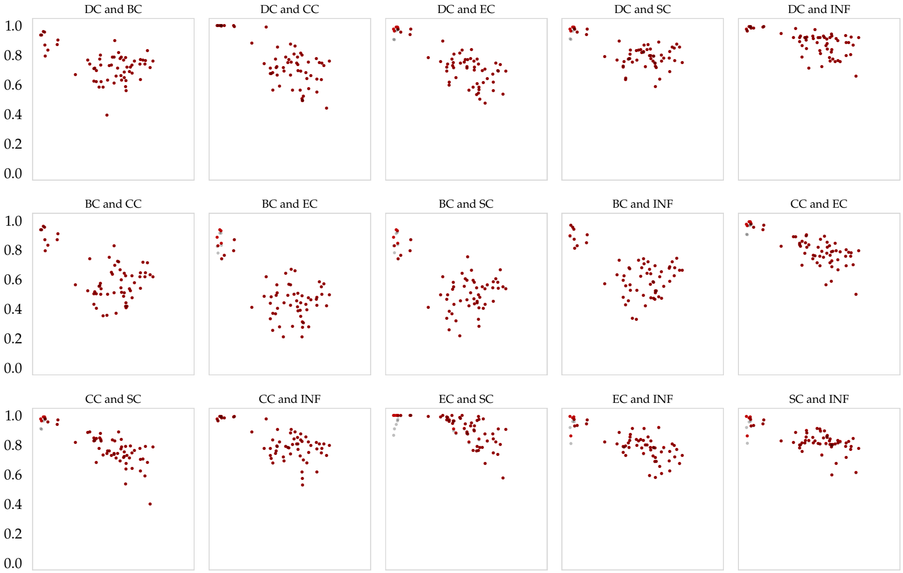


FIGURE 9.8: Correlation of several indices (y-axis) and normalized majorization gap (x-axis) for Valente's dataset. Correlation is measured with Kendall's τ on unaltered scores (black) and rounded to 8 digits (red).

9.5 Summary

In this chapter, we showed that correlations among different centrality indices mainly depend on the network structure, specifically on the distance to its closest threshold graph. This stands in stark contrast with the literature where correlations are assumed to be contingent on the definitions of indices.

The results on random graphs in Section 9.3 have shown that there exists a strong dependence of correlations on the distance to threshold graphs for networks from the Barábasi-Albert model. However, this does not hold true for the $\mathcal{G}(n,p)$ model where strong correlations are generally observed. The results suggest that one has to use caution when probing new indices or testing algorithms on generated data. For networks sampled from the $\mathcal{G}(n,p)$ model, we can expect similar results for any set of indices no matter how parameters are chosen. For the Barábasi-Albert model we might observe differing results depending on how parameters are chosen. If tests on random graph models have to be conducted, it must be ensured that a big range of parameters is chosen to obtain a broad sample of graphs from the respective model.

Section 9.4 has shown that similar outcomes can be expected for real-world networks, although with a higher amount of variability. Real-world networks inherit many different non trivial structural properties, such that it would be premature to constitute a strong or even causal dependence between the distance to a threshold graph and an observed correlation. Nevertheless, we could observe a linear decay in correlation for most pairs of indices when the majorization gap increases. An aggravating phenomenon is the numerical issue observed for subgraph centrality. Since its values can be far beyond any reasonable value range, statistical software might run into their numerical boundaries yielding wrong results due to cancellation. In terms of measurement theory, subgraph centrality can thus not be considered as a reliable measure.

The choice of an appropriate correlation coefficient has shown to be nontrivial and that wrong choices can lead to contradictory results. Especially the use of Pearson's ρ, logarithmic or not, is not convenient to capture dependence among indices, since it assumes indices to be on an interval scale. Thus, rank correlation measures are generally the preferred choice.

As it was already stated in Chapter 8, the majorization gap is just a rough estimate for the distance to the closest threshold graph. It is also not to be expected that edge rotations have a uniform impact on correlations. Certain rotations have a greater effect while others may have no effect at all. Besides the neighborhood-inclusion preorder there are certainly more structural properties of networks that drive all measures of centrality we miss by only focusing on the distance to threshold graphs. In the upcoming chapter, we introduce further concepts of dominance relations in networks which have an additional influence on the correlation of indices.

The predetermination of rankings by the neighborhood-inclusion preorder is not the universal reason for observed correlations. Pairs of indices can be perfectly correlated independent thereof, e.g. a network diameter of two is sufficient for degree and closeness to induce the same ranking.

CHAPTER 10

Generalization and Application of the Dominance Concept

"Problems cannot be solved by the level of awareness that created them."
– Albert Einstein

10.1 Further Notions of Dominance

In his seminal work of 1979, Freeman considered all 21 connected and non-isomorphic graphs with five vertices (henceforth Freeman's dataset) to illustrate differences of degree, closeness and betweenness, focusing on the attained scores for the respective indices [75].

We here reconsider Freeman's dataset, albeit focusing on the neighborhood-inclusion and the predetermination of rankings in the dataset. In this course, we develop further notions of dominance which tighten the set of feasible rankings for centrality indices even more.

In Chapter 8, we presented threshold graphs as a class of uniquely ordered graphs. Freeman's dataset contains eight threshold graphs, which are shown in Figure 10.1.

Another group of graphs which are completely ordered is shown in Figure 10.2. The five depicted graphs are no threshold graphs but contain automorphic equivalent vertices.

An obvious generalization of dominance w.r.t. automorphic equivalence is given in the following corollary.

Corollary 10.1. *If $u \succcurlyeq v$ and $v \sim_\pi w$ then $u \succcurlyeq w$.*

The corollary is a result of the transitivity of $\succcurlyeq$. In order to symbolically distinguish neighborhood-inclusion from this indirect form of dominance, we denote it by $\succcurlyeq_\pi$.

Before proceeding with Freeman's dataset, consider the graph shown in Figure 10.3. Although u and v are not comparable according to neighborhood-inclusion, u should *intuitively* be more central than v. We can formalize this intuition with an automorphic counterpart to neighborhood-inclusion.

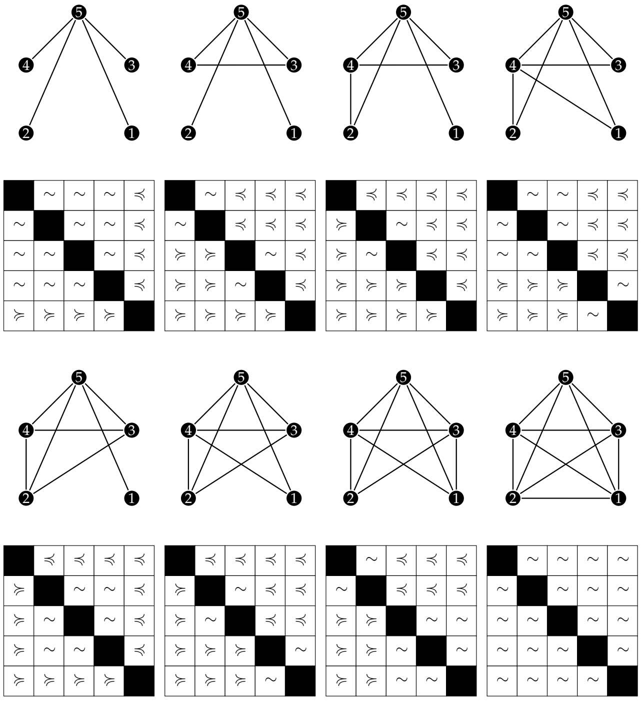


FIGURE 10.1: Threshold graphs with five vertices. The matrices show the order relations among vertices.

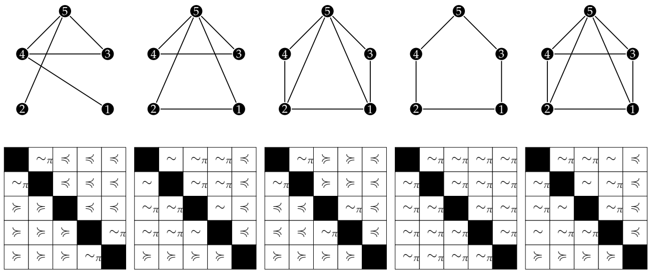


FIGURE 10.2: Graphs with five vertices where the ranking is completely predetermined due to neighborhood-inclusion and automorphic equivalence. The matrices show the order relations among vertices.

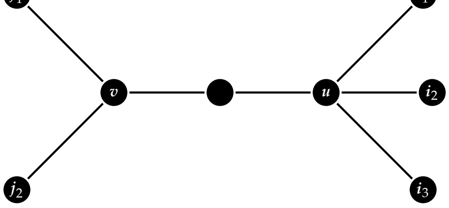


FIGURE 10.3: Example graph motivating two new definitions of dominance.

Definition 10.2. *Let $G = (V, E)$ be a simple undirected graph and $u, v \in V$ and $N_u \subseteq N(u)$. If there exists a function $\varphi : V \to V$ with the following properties*

(i) *φ restricted to $V \setminus N_u$ is an automorphism*

(ii) *$\varphi(v) = u$*

(iii) *φ restricted to N_u is the identical function*

we say that u dominates v automorphically, denoted by $u \succcurlyeq_{\varphi} v$.

It is easy to see that $u \succcurlyeq v \implies u \succcurlyeq_{\varphi}$ by setting $N_u = N(u) \setminus N(v)$ and letting φ be the identity function. The reverse direction does not necessarily hold. With the notion of automorphic dominance, we thus have a coarser form of dominance, as in the case of structural and automorphic equivalence.

Going back to Figure 10.3, we can also argue about the ordering of the structural equivalence classes $[i] := \{i_1, i_2, i_3\}$ and $[j] := \{j_1, j_2\}$. Although they have the same degree, it seems that $[i]$ has an advantage over $[j]$ considering indirect relations.

A corresponding form of dominance is described in the following definition

Definition 10.3. *Let $G = (V, E)$ be a simple undirected graph and $u, v \in V$ with $|N(u)| = |N(v)|$. If for all $j \in N(v)$, there exists an $i \in N(u)$ such that either $i \succcurlyeq j$ or $i \succcurlyeq_{\varphi} j$ holds true, then we say that u dominates v in terms of indirect relations, denoted by $u \succcurlyeq_{\tau} v$.*

Theorem 10.4. *Let $c : V \to \mathbb{R}_0^+$ be a centrality index according to Proposition 7.26. Then*

(i) $u \succcurlyeq_{\varphi} v \implies c(u) \geq c(v)$

(ii) $u \succcurlyeq_{\tau} v \implies c(u) \geq c(v)$

hold true for all $u, v \in V$.

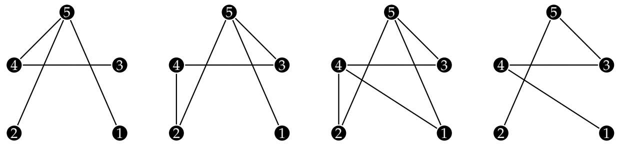
With the semiring framework of Chapter 7 we can apply the same reasoning as for the neighborhood-inclusion and the proofs of all central theorems in Section 7.3 can easily be transfered to the new notions of dominance and are thus omitted.

The new forms of dominance can be found in two graphs of Freeman's dataset and are shown in Figure 10.4.

With the newly derived forms of dominance, we expect more graph classes except the threshold graphs to be totally ordered. Two classes we expect to be totally ordered are given in the following.

Conjecture. *Paths P_n and complete bipartite graphs $K_{n,m}$ are uniquely ranked for all $n, m \in \mathbb{N}$.*

In Freeman's dataset, $K_{2,3}$ and P_5 are present. The two graphs are the two rightmost in Figure 10.4.



■	$\sim$	$\succcurlyeq_\tau$	$\preccurlyeq$	$\preccurlyeq$
$\sim$	■	$\succcurlyeq_\tau$	$\preccurlyeq$	$\preccurlyeq$
$\preccurlyeq_\tau$	$\preccurlyeq_\tau$	■	$\preccurlyeq$	$\preccurlyeq$
$\succcurlyeq$	$\succcurlyeq$	$\succcurlyeq$	■	$\preccurlyeq_\varphi$
$\succcurlyeq$	$\succcurlyeq$	$\succcurlyeq$	$\succcurlyeq_\varphi$	■

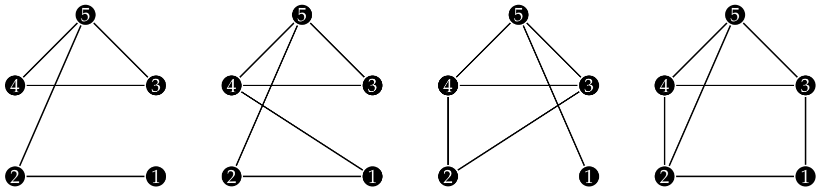
■	$\preccurlyeq$	$\preccurlyeq$	$\preccurlyeq_\varphi$	$\preccurlyeq$
$\succcurlyeq$	■	$\sim$	$\succcurlyeq_\tau$	$\preccurlyeq_\varphi$
$\succcurlyeq$	$\sim$	■	$\succcurlyeq_\tau$	$\preccurlyeq_\varphi$
$\succcurlyeq_\varphi$	$\preccurlyeq_\tau$	$\preccurlyeq_\tau$	■	$\preccurlyeq$
$\succcurlyeq$	$\succcurlyeq_\varphi$	$\succcurlyeq_\varphi$	$\succcurlyeq$	■

■	$\sim$	$\sim$	$\preccurlyeq_\pi$	$\preccurlyeq_\pi$
$\sim$	■	$\sim$	$\preccurlyeq_\pi$	$\preccurlyeq_\pi$
$\sim$	$\sim$	■	$\preccurlyeq_\pi$	$\preccurlyeq_\pi$
$\succcurlyeq_\pi$	$\succcurlyeq_\pi$	$\succcurlyeq_\pi$	■	$\sim$
$\succcurlyeq_\pi$	$\succcurlyeq_\pi$	$\succcurlyeq_\pi$	$\sim$	■

■	$\sim_\pi$	$\preccurlyeq$	$\preccurlyeq$	$\preccurlyeq_\pi$
$\sim_\pi$	■	$\preccurlyeq$	$\preccurlyeq_\pi$	$\preccurlyeq$
$\succcurlyeq$	$\succcurlyeq$	■	$\succcurlyeq_\tau$	$\succcurlyeq_\tau$
$\succcurlyeq$	$\succcurlyeq_\pi$	$\preccurlyeq_\tau$	■	$\sim_\pi$
$\succcurlyeq_\pi$	$\succcurlyeq$	$\succcurlyeq_\tau$	$\sim_\pi$	■

FIGURE 10.4: Graphs with five vertices where the ranking is completely predetermined due to newly derived forms of dominance. The matrices show the order relations among vertices.

Till now, all graphs could be uniquely ordered without the application of a centrality indices by applying several forms of dominance. Figure 10.5 shows the only graphs, where distinct indices may rank pairs of nodes differently.



■	$\preccurlyeq$	$\preccurlyeq_\varphi$	$\preccurlyeq_\varphi$	$\preccurlyeq$
$\succcurlyeq$	■	$\lesseqgtr$	$\lesseqgtr$	$\preccurlyeq_\varphi$
$\succcurlyeq_\varphi$	$\lesseqgtr$	■	$\sim$	$\preccurlyeq$
$\succcurlyeq_\varphi$	$\lesseqgtr$	$\sim$	■	$\preccurlyeq$
$\succcurlyeq$	$\succcurlyeq_\varphi$	$\succcurlyeq$	$\succcurlyeq$	■

■	$\sim_\pi$	$\lesseqgtr$	$\preccurlyeq_\varphi$	$\preccurlyeq$
$\sim_\pi$	■	$\lesseqgtr$	$\preccurlyeq$	$\preccurlyeq_\varphi$
$\lesseqgtr$	$\lesseqgtr$	■	$\preccurlyeq$	$\preccurlyeq$
$\succcurlyeq_\varphi$	$\succcurlyeq$	$\succcurlyeq$	■	$\sim_\pi$
$\succcurlyeq$	$\succcurlyeq_\varphi$	$\succcurlyeq$	$\sim_\pi$	■

■	$\preccurlyeq_\varphi$	$\preccurlyeq$	$\preccurlyeq$	$\preccurlyeq$
$\succcurlyeq_\varphi$	■	$\preccurlyeq$	$\preccurlyeq$	$\preccurlyeq$
$\succcurlyeq$	$\succcurlyeq$	■	$\sim$	$\lesseqgtr$
$\succcurlyeq$	$\succcurlyeq$	$\sim$	■	$\lesseqgtr$
$\succcurlyeq$	$\succcurlyeq$	$\lesseqgtr$	$\lesseqgtr$	■

■	$\preccurlyeq_\varphi$	$\preccurlyeq_\varphi$	$\preccurlyeq$	$\preccurlyeq$
$\succcurlyeq_\varphi$	■	$\sim$	$\lesseqgtr$	$\lesseqgtr$
$\succcurlyeq_\varphi$	$\sim$	■	$\lesseqgtr$	$\lesseqgtr$
$\succcurlyeq$	$\lesseqgtr$	$\lesseqgtr$	■	$\sim$
$\succcurlyeq$	$\lesseqgtr$	$\lesseqgtr$	$\sim$	■

FIGURE 10.5: Graphs with five vertices where discordance may occur, shown in the gray cells of the matrices.

All feasible rankings can be achieved with the four traditional measures as shown in Figure 10.6.

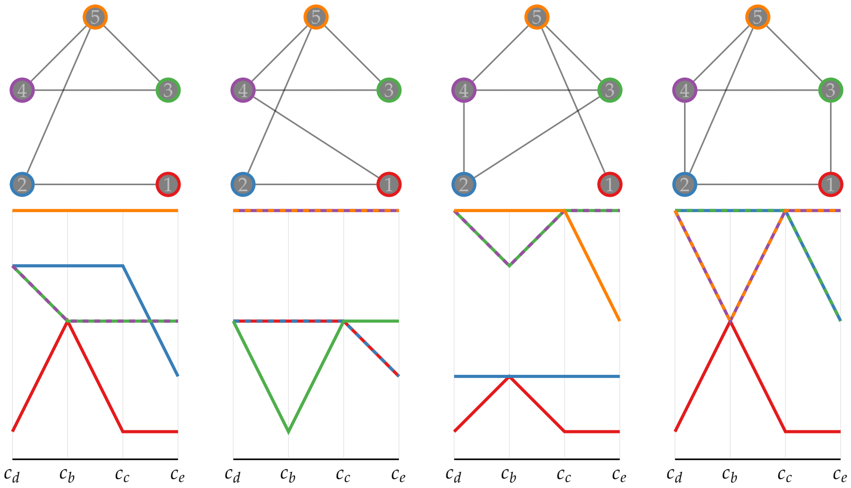


FIGURE 10.6: Induced rankings of degree, betweenness, closeness and eigenvector centrality as parallel coordinates for the four graphs shown in Figure 10.5.

10.2 The Scope of Rankings and its Implications

With positional dominance, dominance induced by automorphic equivalence, automorphic dominance and dominance by indirect relations, we now have identified four forms of dominance in networks, which are preserved by centrality indices in line with Proposition 7.26. Together with structural and automorphic equivalence, we thus have six structural properties of networks that drive all measures of centrality. Depending on the considered network, they can tighten the set of feasible centrality rankings significantly. This six properties were, e.g. enough to rank 17 out of 21 graphs without applying any index. For the remaining 4 graphs, effectively only 4 pairs were incomparable.

In empirical studies we rather expect to encounter graphs with far more vertices and a visual analysis, as partially done for Freeman's dataset, is not feasible. Furthermore, we do not yet have an algorithm to determine comparable pairs of vertices according to the new dominance relations. We therefore have to rely on the neighborhood-inclusion preorder for larger graphs, keeping in mind that it only provides a lower bound of the predetermination of a centrality ranking.

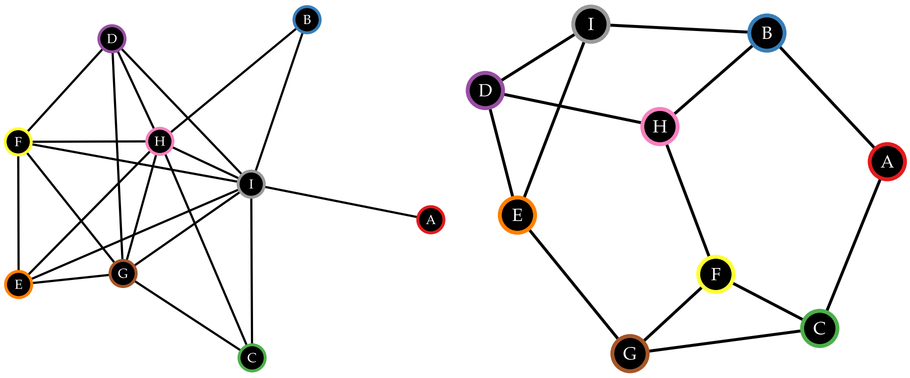
In the following, we show that our new characterization of centrality concepts proves helpful for empirical studies. We illustrate how our methods can be applied to support empirical studies in hypothesis testing and to reduce the need for trial-and-error approaches.

The density of the neighborhood-inclusion preorder can be used as an a priori test before any centrality related hypothesis is posed. If the preorder is sparse, i.e. only a few pairs of vertices are comparable, then the network structure permits a high degree of freedom for centrality rankings. We can quite easily define indices which yield desired outcomes, yet we have to be careful with the interpretation of results. There might well be an index holding a different rationale for an empirical phenomenon, which might either describe it even better, or in the worst case give rise to a contrary explanation.

In cases where the neighborhood-inclusion preorder is (nearly) complete we have to be equally careful with drawing conclusions. All centrality indices induce a similar ranking since the majority of the ranking is predetermined. Therefore, we can not attribute results to a specific index since the network structure does not permit any other outcome. We are thus also confronted with competing explanations for observed phenomena, e.g. dual measures like betweenness and closeness explain an outcome equally well.

Recall Leavitt's communication experiment from Chapter 4. Already Freeman noticed that degree, betwenness and closeness induce the same ranking on the wheel, chain, circle and Y used in the experiment [80]. An explanation can now be given with the new notions of dominance. The star graph is a threshold graph, all vertices in the circle are automoprhic equivalent, chain graphs (paths) are uniquely ordered and the Y is uniquely ordered due to automorphic dominance. With these graphs we can therefore not make any substantial argument for the use of a specific index since all indices induce the same ranking. Freeman reconsidered Leavitt's experiment with yet a different set of graphs. He used the first two graphs in Figure 10.5 and the third of Figure 10.2. In the first two graphs, we observed pairs of nodes that can be ranked differently. Since all possible rankings are achievable with degree, betweenness and closeness, we can observe differences in the outcomes and certain indices could explain varying performances in the experiments. Freeman conducted the experiments and noted that degree and betweenness are more indicative than closeness for solving tasks within groups.

Besides Leavitt's and Freeman's experiments we can now also explain the extremely different outcomes when various indices are applied to our introductory example graphs with nine vertices (cf. Figure 10.7).



(A) uniquely ranked graph

(B) discordant graph

FIGURE 10.7: Example graphs with nine vertices discussed in Section 3.3

The graph in Figure 10.7(a) with an observed unique ranking is in fact a threshold graph with the complete ranking

$$I \succcurlyeq H \succcurlyeq G \succcurlyeq F \succcurlyeq E \sim D \succcurlyeq C \succcurlyeq B \succcurlyeq A$$

and no other ranking is possible for indices according to Theorem 7.28. If this graph would have been encountered in an empirical setting, we have to have the aforementioned points in mind to not attribute outcomes to a specific index.

The graph in Figure 10.7(b) on the other hand, does not hold any form of dominance or equivalence relation, i.e. their is no predetermined ranking. That is, we have the stated huge degree of freedom to determine a centrality ranking. If we do not allow for ties in the ordering, we theoretically have up to $9! = 362880$ possibilities to rank the vertices. Ties in a ranking are, however, not only a result of equivalence relations but can occur for arbitrary subsets of vertices due to the aggregation of relations. The number of weak orderings for n objects is given by the *ordered Bell number* defined as

$$a(n) = \sum_{k=0}^{n} k! S(n,k) = \sum_{k=0}^{n} \sum_{j=0}^{k} (-1)^{k-j} \binom{k}{j} j^n ,$$

where $S(n,k)$ is the *Stirling number of the second kind*. In our case with $n = 9$ we have 7087261 possibilities to weakly order the vertices. Most of these weak orderings are out of scope with reasonable graph-theoretic approaches, yet taking a data-driven approach we could quite easily define indices to obtain any of these orderings. Recall that we can interpret walk based indices as a scalar product

$$c(u) = \sum_{t \in V} \langle \gamma, \omega_{ut} \rangle \,,$$

such that we can choose γ (e.g. with a multidimensional linear program) to be a monotonically decreasing series yielding a desired outcome. This procedure is of course not in line with our definition of centrality designated as a form of measurement. It should just serve as an illustration of the potential of a deceitful index-driven approach.

10.3 Centrality-Lethality Hypothesis revisited

In Chapter 5, we examined in depth the centrality-lethality hypothesis on several instances of the PIN of yeast. We saw that different indices perform best on each instance and we demonstrated with the hyperbolic index that indices producing reasonable results can be tailored with apparent ease. In this section, we investigate the centrality-lethality hypothesis with our newly developed tools for structural evidence to justify the assumption of a centrality effect and offer reasons for our observed outcomes.

The low density of the neighborhood-inclusion preorder for all eight PINs of yeast (cf. first column of Table 10.1) offers an explanation for several aspects observed in Chapter 5. This first includes the comparable performances yet weak correlations of several indices and second, the reasonable performance of the crafted hyperbolic index. The network structure in all eight instances gives ample scope for centrality rankings such that different indices might rank different lethal proteins on top. This prevents any general statement about a connection between a specific index and the lethality status of a protein. We can, however, argue about a general possibility of a centrality effect without appealing to indices by testing if the neighborhood-inclusion preorder is in accordance with the hypothesis.

Recall that the hypothesis we are working with states that centrality is positively associated with the lethality status of a protein, a dichotomous categorical variable. A general plausibility test we can perform is therefore twofold. First we check for the possibility to structurally discriminate the two kinds of proteins and second if lethal proteins dominate non-lethal ones such that they can be ranked on top. For the structural discriminability, the automorphic and structural equivalence classes should be homogeneous. We find, however, that the compositions are quite heterogeneous (cf. second column of Table 10.1), that is a high fraction of lethal proteins is not discriminable from one or more non-lethal proteins by any structural measure and thus also by any centrality index. Additionally, the third column of Table 10.1 shows, that a great deal of lethal proteins are dominated by at least one non-lethal protein. Hence, an optimal performance is not achievable with any centrality index, since several non-lethal proteins are always ranked higher than the lethal proteins they dominate.

It must be pointed out that an optimal performance is anyways quite a strong demand. We can estimate an upper bound for the performance of centrality indices, by minimizing the number of inversions of non-lethal/lethal proteins in a completion of the neighborhood-inclusion preorder. Finding the completion with minimum inversions is $\mathcal{NP}-$ hard [117], such that we have to rely on heuristics to estimate the upper bound. The fourth column of Table 10.1 shows estimated upper bounds for the AUC values of PR and the highest achieved AUC in Chapter 5 in parenthesis. The results show, that we can potentially come very close to an optimal performance, yet current methods are far away from this upper bound. It is unlikely to attain this accuracy with the network structure alone, suggesting that approaches incorporating further information or attributes, e.g. biological properties, are more promising.

Dataset	Comparable pairs	Indistinguishable from non-lethals	Dominated by non-lethals	Upper bound for AUC
Jeong	0.88%	25%	37%	0.98(0.28)
Estrada	0.64%	13%	30%	0.97(0.48)
DIP	0.55%	11%	29%	0.94(0.47)
BIOGRID	0.89%	5%	2%	0.99(0.26)
STRING	0.37%	8%	26%	0.95(0.54)
LC	0.72%	14%	28%	0.97(0.57)
Collins	1.06%	15%	34%	0.96(0.51)
Y2H	0.87%	12%	44%	0.95(0.27)

TABLE 10.1: Statistics for neighborhood-inclusion in eight PINs of *S. cerevisiae*. AUC values in parenthesis denote the best performance of centrality indices.

10.4 SUMMARY AND DISCUSSION

The main purpose of this chapter was to illustrate how the theoretical frameworks developed in Chapter 7 can be used to derive further dominance relations in graphs and how the neighborhood-inclusion preorder can be applied in empirical settings. The new dominance relations tighten the set of feasible centrality rankings and also suggest that there are further graph classes that are uniquely ranked besides threshold graphs, e.g. paths and complete bipartite networks. However, we currently do not have an algorithm to determine dominance besides neighborhood-inclusion, such that examinations as done with Freeman's dataset are only feasible for small graphs. In times of the big data conundrum this poses of course non negligible problems. We thus have to rely on the neighborhood-inclusion alone, keeping in mind that it only yields the lower bound for the predetermination of any centrality ranking.

Our new insights were used to revisit the centrality-lethality hypothesis, offering explanations for certain observed phenomena in Chapter 5. We gathered some evidence *against* a centrality effect by showing that many lethal proteins are either indistinguishable from or dominated by non-lethal proteins.

The sparsity of the neighborhood-inclusion preorder prevents us from completely refuting the hypothesis. There might well be an index which describes the structural position of lethal proteins in a PIN in a nearly optimal way. We could find this index with optimization techniques and terminate the hunt for the best index once and for all. Still, even in a perfect setting where PINs are error-free and complete, this purely data-driven approach does not allow for deriving substantive conclusions. Once determined, it will (a) be hard to interpret what the index actually measures and (b) even with an interpretation not at all be clear what the biological connection is.

The presented results might give the impression that using our new characterization of centrality only reveals adverse consequences in empirical research since a sparse as well as a dense neighborhood-inclusion preorder seem to impede a proper reasoning about centrality effects. This certainly holds true for the index-driven approach to centrality. Indices quantify high-dimensional relations into a one-dimensional ranking structure, such that information gets lost, is neglected or leveled off.

Comparing network positions, as described in Chapter 6, offers the possibility to gradually build a ranking by incorporating additional information to tighten the order step by step. We actually do not even have to go as far as to define a complete ranking. A partial ranking already bears information about the relationship of positions.

Chapter 11

Conclusion

> *"The time has come, it would seem, to stop, take stock and try to make some sense of the concept of centrality and the range and limits of its potential for application."*
>
> – Freeman, 1979

The main goal of this thesis was to provide a better theoretical foundation for network centrality to reduce its conceptual ambiguity. We did so by reconceptualizing centrality based on positional dominance. The preservation of dominance is the first structural property besides structural and automorphic equivalence that drive all measures of centrality. With our conceptualization, we were able to explain several empirically observed phenomena of centrality indices and we could offer new insights in discussed topics of the literature. In summary, our theoretical results are the following.

SPECTRAL FORMS OF INDICES. We explored the possibility of representing all centrality indices with the spectral decomposition of the adjacency matrix and conjectured about two general representations. We could observe a close relation of information centrality and degree which was not that obvious before. A general spectral framework might, however, fail to come into use, since measures like betweenness and closeness are most likely not representable in a non-trivial way.

INDIRECT RELATIONS VIA SEMIRINGS. We used the algebraic structure of semirings as a general framework to define indirect relations on networks. Semirings were already used in the context of generic path algorithms, however, not in the context of centrality. Additionally, we introduced semirings for relations based on walk counts which were not established hitherto. Semirings proved helpful for all our considerations and should remain useful in future theoretical research about centrality.

PRESERVATION OF DOMINANCE. We derived several sufficient conditions for the preservation of dominance. Indices based on paths preserve dominance if they are defined on a monotonic idempotent semiring. For walk based indices we found that dominance is preserved if the representational sequence is positive and monotonically decreasing for arbitrary walks and positive for closed walk counts. Both cases incorporate indirect relations which are not yet considered as measures of centrality.

UNIQUELY RANKED GRAPHS. The notions of dominance revealed, that there exist different graph classes which are uniquely ordered, i.e. every index yields the same ranking. Besides threshold graphs, we found that paths and complete bipartite graphs are potentially completely predetermined by dominance. It is to be expected that there are several such classes. The requirement of agreement of all indices on these classes generalized the star property which was the only known uniquely ordered class so far.

CORRELATION AMONG INDICES. The existence of uniquely ranked graphs gave rise to a new explanation for correlation among indices. In contrast to the literature, where correlations are believed to be related with the definition of indices, we showed that it mainly depends on the network structure, specifically on the distance to the closest threshold graph. The relation between threshold distance and correlation is, however, not the only driving force. First, more general forms of dominance were neglected and other network properties like the diameter were not consider. We thus expect the dependence of correlation on network structure to be even stronger than illustrated in this thesis.

The second goal of the thesis was to investigate centrality in empirical research and offering a more versatile approach for its assessment.

We reviewed the literature and discussed obtained results for the centrality-lethality hypothesis in detail. The provided results show that centrality as a kind of data mining task involves an elevated risk of producing fallacious conclusions. Results from the literature do not withstand closer examinations and proclaimed effects in PINs are weak at best and strongly data dependent. Applying the dominance principle, we were able to provide some evidence against the hypothesis, yet we could not entirely refute it.

The centrality-lethality hypothesis is not a unique case and just exemplifies the limitations and issues of the index driven assessment of centrality. An index producing desired results can always be found among the plethora of measures, if not, new ones can be crafted until satisfaction. It is due to this opportunistic approach that negative results are rarely seen in centrality studies. There is no doubt that phenomena exist where an assessment of centrality is not possible because of the complexity of relationships and/or the processes involved. Inferring an effect in such cases by applying indices in a haphazard way defies the nature of things.

The positional approach paired with tools from measurement theory offer a chance of restructuring centrality at the root of the concept. Viewing an actor not just as a vertex in a graph, but rather as an entity with multi-dimensional relations of various kinds give us the chance to gradually assess the position of an actor, relative to others depending on varying conditions. It thus breaks down the analytic process into steps where in each step domain-specific requirements can be adapted leading to well-defined intermediate results. Positional dominance, or neighborhood-inclusion for binary symmetric relations, always serves as a starting point since it puts the highest requirements on relations. If afterwards additivity and/or homogeneity can be assumed on various levels, we obtain a huge space of options to proceed, which always depend on the specific context. Especially the case where homogeneity can not be assumed poses some challenges, where no universal solution exists. Thus, more empirical studies have to be conducted in order to validate the potential advantages or uncover weaknesses and especially to define proper methodological procedures.

If we do not accept the positional approach, we at least have to acknowledge the preservation of the neighborhood-inclusion preorder as a shared feature of all centrality indices. Enforcing such features on indices has shown to be futile and only shifts the focus from indices towards defining axiomatic systems. Uncovering shared properties of measures, on the other hand, yield insight in the driving forces of indices which can later be used to explain empirically observed phenomena from a theoretical point of view. This, in turn, suggest the shift of defining indices towards examining theoretical foundations of measures giving a better understanding of what indices actually do. A solid theoretical foundation is of utmost importance for empirical research in order to prevent fallacious courses of action.

In empirical research, we will, however, most certainly not be dealing with the simple form of symmetric and binary networks. Binarizing and/or symmetrizing those kinds of network is not an uncommon procedure, yet it comes with loss of information. Centrality indices have to be adapted to incorporate potential weights and asymmetric links in order to provide distinctive outcomes. In contrast, positional dominance is readily applicable to any kind of network. Most of the theoretical considerations evolving around it are also transferable, such that it is independent of the network in question. A major challenge pose dynamics on networks and networks evolving over time. Understanding these dynamics is already a complex task for itself such that defining any reasonable concept of centrality for such cases seems out of scope, at least for the moment.

The thesis only dealt with the concept of network centrality but many aspects tackled are transferable to other network analytic methods. The concept of clustering (or community detection) is, besides centrality, the most widely used network analytic procedure in empirical research. Corresponding to centrality, a proper foundation is missing and the number of methods to find cohesive groups is as extensive as for centrality.

Admittedly, network science is still in its infancies and many fundamental theories are naturally not yet established. Its popularity across scientific boundaries, however, lead to the spread and innovation of methods. This, paired with the ever-expanding availability of data, ensures the growth of fields of application but shifts the focus away from reflecting about theory and proper methodology. Network science should not be seen as an attempt to unify scientific areas or savior for all research problems, but as a scientific field by itself with proper theory and well-elaborated methods to investigate networks of different origins with enough confidence.

Bibliography

[1] Robert Adcock. Measurement validity: A shared standard for qualitative and quantitative research. In *American Political Science Association*, volume 95, pages 529–546, 2001.

[2] Réka Albert, Hawoong Jeong, and Albert-László Barabási. Error and attack tolerance of complex networks. *Nature*, 406(6794):378–382, 2000.

[3] Alon Altman and Moshe Tennenholtz. Ranking systems: the PageRank axioms. In *Proceedings of the 6th ACM conference on Electronic commerce*, pages 1–8. ACM, 2005.

[4] Chintagunta Ambedkar, Kiran Kumar Reddi, Naresh Babu Muppalaneni, and Duggineni Kalyani. Application of centrality measures in the identification of critical genes in diabetes mellitus. *Bioinformation*, 11(2):90, 2015.

[5] Gil Amitai, Arye Shemesh, Einat Sitbon, Maxim Shklar, Dvir Netanely, Ilya Venger, and Shmuel Pietrokovski. Network analysis of protein structures identifies functional residues. *Journal of Molecular Biology*, 344(4):1135–1146, 2004.

[6] Jac M. Anthonisse. The rush in a directed graph. *Stichting Mathematisch Centrum. Mathematische Besliskunde*, (BN 9/71):1–10, 1971.

[7] Srinivasa R. Arikati and Uri N. Peled. Degree sequences and majorization. *Linear Algebra and its Applications*, 199:179–211, 1994.

[8] Kenneth J Arrow. A difficulty in the concept of social welfare. *The Journal of Political Economy*, pages 328–346, 1950.

[9] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.

[10] Vladimir Batagelj. Semirings for social networks analysis. *Journal of Mathematical Sociology*, 19(1):53–68, 1994.

[11] Komal Batool and Muaz A Niazi. Towards a methodology for validation of centrality measures in complex networks. *PloS ONE*, 9(4):e90283, 2014.

[12] Stefano Battiston, Michelangelo Puliga, Rahul Kaushik, Paolo Tasca, and Guido Caldarelli. Debtrank: Too central to fail? financial networks, the fed and systemic risk. *Scientific Reports*, 2, 2012.

[13] Alex Bavelas. A mathematical model for group structures. *Human Organizations*, 7(3):16–30, 1948.

[14] Alex Bavelas. Communication patterns in task-oriented groups. *Journal of the Acoustical Society of America*, 22(6):725–730, 1950.

[15] Michele Benzi and Christine Klymko. Total communicability as a centrality measure. *Journal of Complex Networks*, 1(2):124–149, 2013.

[16] Michele Benzi and Christine Klymko. On the limiting behavior of parameter-dependent network centrality measures. *SIAM Journal on Matrix Analysis and Applications*, 36(2):686–706, 2015.

[17] Peter M. Blau. *Inequality and Heterogeneity: A Primitive Theory of Social Structure*. The Free Press, New York, NY, 1977.

[18] Peter M. Blau. A macrosociological theory of social structure. *American Journal of Sociology*, 83(1):26–54, 1977.

[19] Frank Boesch, Charles Suffel, Ralph Tindell, and Frank Harary. The neighborhood inclusion structure of a graph. *Mathematical and Computer Modelling*, 17(11):25–28, 1993.

[20] Paolo Boldi and Sebastiano Vigna. Axioms for centrality. *Internet Mathematics*, 10(3-4):222–262, 2014.

[21] John M Bolland. Sorting out centrality: An analysis of the performance of four centrality models in real and simulated networks. *Social Networks*, 10(3):233–253, 1988.

[22] Béla Bollobás and Oliver Riordan. The diameter of a scale-free random graph. *Combinatorica*, 24(1):5–34, 2004.

[23] Béla Bollobás and Oliver Riordan. Robustness and vulnerability of scale-free random graphs. *Internet Mathematics*, 1(1):1–35, 2004.

[24] Phillip Bonacich. Factoring and weighting approaches to status scores and clique identification. *Journal of Mathematical Sociology*, 2:113–120, 1972.

[25] Phillip Bonacich. Power and centrality: A family of measures. *American Journal of Sociology*, 92(5):1170–1182, 1987.

[26] Phillip Bonacich, Annie Cody Holdren, and Michael Johnston. Hyperedges and multidimensional centrality. *Social Networks*, 26(3):189–203, 2004.

[27] John Adrian Bondy and Uppaluri Siva Ramachandra Murty. *Graph Theory with Applications*, volume 290. Macmillan London, 1976.

[28] Stephen P. Borgatti. Centrality and network flow. *Social Networks*, 27(1):55–71, 2005.

[29] Stephen P. Borgatti, Kathleen M. Carley, and David Krackhardt. On the robustness of centrality measures under conditions of imperfect data. *Social Networks*, 28(2):124–136, 2006.

[30] Stephen P. Borgatti and Martin G. Everett. Notions of position in social network analysis. *Sociological Methodology*, 22:1–35, 1992.

[31] Stephen P. Borgatti and Martin G. Everett. Models of core/periphery structures. *Social Networks*, 21(4):375–395, 1999.

[32] Stephen P. Borgatti and Martin G. Everett. A graph-theoretic framework for classifying centrality measures. *Social Networks*, 28(4):466–484, 2006.

[33] Denis Bouyssou and Marc Pirlot. Conjoint measurement tools for MCDM. In *Multiple Criteria Decision Analysis: State of the Art Surveys*, pages 73–112. Springer, 2005.

[34] Ulrik Brandes. A faster algorithm for betweenness centrality. *Journal of Mathematical Sociology*, 25(2):163–177, 2001.

[35] Ulrik Brandes. On variants of shortest-path betweenness centrality and their generic computation. *Social Networks*, 30(2):136–145, 2008.

[36] Ulrik Brandes. Network positions. *submitted*, 2015.

[37] Ulrik Brandes, Stephen P. Borgatti, and Linton C. Freeman. Maintaining the duality of closeness and betweenness centrality. *Social Networks*, 44:153–159, 2016.

[38] Ulrik Brandes and Daniel Fleischer. Centrality measures based on current flow. In Volker Diekert and Bernard Durand, editors, *Proceedings of the 22nd International Symposium on Theoretical Aspects of Computer Science (STACS'05)*, volume 3404 of *Lecture Notes in Computer Science*, pages 533–544. Springer-Verlag, 2005.

[39] Ulrik Brandes and Jan Hildenbrand. Smallest graphs with distinct singleton centers. *Network Science*, 2(3):416–418, 2014.

[40] Ulrik Brandes, Garry Robins, Ann McCranie, and Stanley Wasserman. What is network science? *Network Science*, 1(1):1–15, 2013.

[41] Graham Brightwell and Peter Winkler. Counting linear extensions. *Order*, 8(3):225–242, 1991.

[42] Eric Chea and Dennis R Livesay. How accurate and statistically robust are catalytic site predictions based on closeness centrality? *BMC Bioinformatics*, 8(1):153, 2007.

[43] Nicholas A. Christakis and James H. Fowler. The spread of obesity in a large social network over 32 years. *New England Journal of Medicine*, 357:370–379, 2007.

[44] Hon Nian Chua, Kar Leong Tew, Xiao-Li Li, and See-Kiong Ng. A unified scoring scheme for detecting essential proteins in protein interaction networks. In *Tools with Artificial Intelligence, 2008. ICTAI'08. 20th IEEE International Conference on*, volume 2, pages 66–73, 2008.

[45] Vašek Chvátal and Peter L. Hammer. Aggregation of inequalities in integer programming. *Annals of Discrete Mathematics*, 1:145–162, 1977.

[46] Anne E. Clatworthy, Emily Pierson, and Deborah T Hung. Targeting virulence: a new paradigm for antimicrobial therapy. *Nature Chemical Biology*, 3(9):541–548, 2007.

[47] Aaron Clauset, Cosma Rohilla Shalizi, and Mark E. J. Newman. Power-law distributions in empirical data. *SIAM Review*, 51(4):661–703, 2009.

[48] Sean R. Collins, Patrick Kemmeren, Xue-Chu Zhao, Jack F. Greenblatt, Forrest Spencer, Frank C.P. Holstege, Jonathan S. Weissman, and Nevan J. Krogan. Toward a comprehensive atlas of the physical interactome of saccharomyces cerevisiae. *Molecular & Cellular Proteomics*, 6(3):439–450, 2007.

[49] Auguste Comte. *The Positive Philosophy of Auguste Comte*. W. Gowans, 1868.

[50] Karen S. Cook, Richard M Emerson, and Mary R. Gillmore. The distribution of power in exchange networks: Theory and experimental results. *American Journal of Sociology*, 89(2):275–305, 1983.

[51] Elizabeth Costenbader and Thomas W. Valente. The stability of centrality measures when networks are sampled. *Social Networks*, 25(4):283–307, 2003.

[52] Stéphane Coulomb, Michel Bauer, Denis Bernard, and Marie-Claude Marsolier-Kergoat. Gene essentiality and the topology of protein interaction networks. *Proceedings of the Royal Society B: Biological Sciences*, 272(1573):1721–1725, 2005.

[53] Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 233–240. ACM, 2006.

[54] Gabriel del Rio, Dirk Koschützki, and Gerardo Coello. How to identify essential genes from molecular networks? *BMC Systems Biology*, 3(1):102, 2009.

[55] Antonio del Sol, Hirotomo Fujihashi, Dolors Amoros, and Ruth Nussinov. Residue centrality, functionally important residues, and active site shape: Analysis of enzyme and non-enzyme families. *Protein Science*, 15(9):2120–2128, 2006.

[56] Persi Diaconis, Susan Holmes, and Svante Janson. Threshold graph limits and random threshold graphs. *Internet Mathematics*, 5(3):267–320, 2008.

[57] José A. Díez. A hundred years of numbers. an historical introduction to measurement theory 1887–1990: Part I: The formation period. two lines of research: Axiomatics and real morphisms, scales and invariance. *Studies In History and Philosophy of Science Part A*, 28(1):167–185, 1997.

[58] José A. Díez. A hundred years of numbers. an historical introduction to measurement theory 1887–1990: Part II: Suppes and the mature theory. representation and uniqueness. *Studies In History and Philosophy of Science Part A*, 28(2):237–265, 1997.

[59] Pål Grønås Drange, Markus Sortland Dregi, Daniel Lokshtanov, and Blair D Sullivan. On the threshold of intractability. In *Algorithms - ESA 2015*, volume 9294, pages 411–423. Springer Berlin Heidelberg, 2015.

[60] Paul Embrechts, Alexander McNeil, and Daniel Straumann. Correlation and dependence in risk management: properties and pitfalls. *Risk Management: Value at Risk and Beyond*, pages 176–223, 2002.

[61] Paul Erdős and Alfred Rényi. On random graphs I. *Publicationes Mathematicae Debrecen*, 6:290–297, 1959.

[62] Ernesto Estrada. Protein bipartivity and essentiality in the yeast protein-protein interaction network. *Journal of Proteome Research*, 5(9):2177–2184, 2006.

[63] Ernesto Estrada. Virtual identification of essential proteins within the protein interaction network of yeast. *Proteomics*, 6(1):35–40, 2006.

[64] Ernesto Estrada, Desmond J. Higham, and Naomichi Hatano. Communicability betweenness in complex networks. *Physica A: Statistical Mechanics and its Applications*, 388(5):764–774, 2009.

[65] Ernesto Estrada and Juan A. Rodríguez-Velázquez. Spectral measures of bipartivity in complex networks. *Physical Review E*, 72(4):046105, 2005.

[66] Ernesto Estrada and Juan A. Rodríguez-Velázquez. Subgraph centrality in complex networks. *Physical Review E*, 71(5):056103, 2005.

[67] Ernesto Estrada and Juan A Rodríguez-Velázquez. Subgraph centrality and clustering in complex hyper-networks. *Physica A: Statistical Mechanics and its Applications*, 364:581–594, 2006.

[68] Martin G Everett and Stephen P Borgatti. The centrality of groups and classes. *The Journal of Mathematical Sociology*, 23(3):181–201, 1999.

[69] Katherine Faust. Centrality in affiliation networks. *Social Networks*, 19(2):157–191, 1997.

[70] Scott L Feld. Why your friends have more friends than you do. *American Journal of Sociology*, pages 1464–1477, 1991.

[71] Peter C. Fishburn. Normative theories of decision making under risk and under uncertainty. In *Non-Conventional Preference Relations in Decision Making*, volume 301, pages 1–21. Springer, 1988.

[72] Lester R. Ford and Delbert R. Fulkerson. Maximal flow through a network. *Canadian Journal of Mathematics*, 8(3):399–404, 1956.

[73] Terrill L. Frantz, Marcelo Cataldo, and Kathleen M. Carley. Robustness of centrality measures under uncertainty: Examining the role of network topology. *Computational and Mathematical Organization Theory*, 15(4):303–328, 2009.

[74] Linton C. Freeman. A set of measures of centrality based on betweenness. *Sociometry*, 40(1):35–41, 1977.

[75] Linton C. Freeman. Centrality in social networks: Conceptual clarification. *Social Networks*, 1(3):215–239, 1978/79.

[76] Linton C. Freeman. *The Development of Social Network Analysis: A Study in the Sociology of Science*. Empirical Press, Vancouver, BC, 2004.

[77] Linton C. Freeman. Going the wrong way on a one-way street: Centrality in physics and biology. *Journal of Social Structure*, 9(2), 2008.

[78] Linton C. Freeman. The development of social network analysis–with an emphasis on recent events. *The SAGE Handbook of Social Network Analysis*, pages 26–54, 2011.

[79] Linton C. Freeman, Stephen P. Borgatti, and Douglas R. White. Centrality in valued graphs: A measure of betweenness based on network flow. *Social Networks*, 13(2):141–154, 1991.

[80] Linton C. Freeman, Douglas Roeder, and Robert R. Mulholland. Centrality in social networks: II. experimental results. *Social Networks*, 2(2):119–141, 1980.

[81] Noah E. Friedkin. Theoretical foundations for centrality measures. *American Journal of Sociology*, 96(6):1478–1504, 1991.

[82] Noah E. Friedkin and Eugene C. Johnsen. Social influence and opinions. *Journal of Mathematical Sociology*, 15(3-4):193–206, 1990.

[83] Noah E. Friedkin and Eugene C. Johnsen. Two steps to obfuscation. *Social Networks*, 39:12–13, 2014.

[84] Xinbo Gao, Bing Xiao, Dacheng Tao, and Xuelong Li. A survey of graph edit distance. *Pattern Analysis and Applications*, 13(1):113–129, 2010.

[85] Jorge Gil-Mendieta and Samuel Schmidt. The political network in Mexico. *Social Networks*, 18(4):355–381, 1996.

[86] Edgar N. Gilbert. Random graphs. *The Annals of Mathematical Statistics*, 30(4):1141–1144, 1959.

[87] Chris Godsil and Gordon F. Royle. *Algebraic Graph Theory*, volume 207. Springer Science & Business Media, 2013.

[88] Michel Gondran and Michel Minoux. *Graphs and Algorithms*. Wiley Chichester, 1984.

[89] Leo A. Goodman and William H. Kruskal. Measures of association for cross classifications*. *Journal of the American Statistical Association*, 49(268):732–764, 1954.

[90] Mark S. Granovetter. The strength of weak ties. *American Journal of Sociology*, 78(6):1360–1380, 1973.

[91] Brian Greenhill, Michael Ward, and Audrey Sacks. The'separation plot': A new visual method for evaluating the predictive power of logit/probit models. In *APSA 2009 Toronto Meeting Paper*, 2009.

[92] Aric Hagberg, Pieter J. Swart, and Daniel A. Schult. Designing threshold networks with given structural and dynamical properties. *Physical Review E*, 74(5):056116, 2006.

[93] Matthew W. Hahn and Andrew D. Kern. Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. *Molecular Biology and Evolution*, 22(4):803–806, 2005.

[94] Peter L. Hammer, Toshihide Ibaraki, and B. Simeone. Threshold sequences. *SIAM Journal on Algebraic Discrete Methods*, 2(1):39–49, 1981.

[95] John S. Hammond, Ralph L. Keeney, and Howard Raiffa. *Smart Choices: A Practical Guide to Making Better Decisions*, volume 226. Harvard Business Press, 1999.

[96] Godfrey Harold Hardy, John Edensor Littlewood, and George Pólya. *Inequalities*. Cambridge university press, 1952.

[97] Xionglei He and Jianzhi Zhang. Why do hubs tend to be essential in protein networks? *PLoS genetics*, 2(6):e88, 2006.

[98] Conrad Heilmann. A new interpretation of the representational theory of measurement. *Philosophy of Science*, page to appear, 2014.

[99] Thomas Hobbes and Edwin Curley. *Leviathan: with selected variants from the Latin edition of 1668*, volume 8348. Hackett Publishing, 1994.

[100] Sungryong Hong and Arjun Dey. Network analysis of cosmic structures: network centrality and topological environment. *Monthly Notices of the Royal Astronomical Society*, 450(2):1999–2015, 2015.

[101] Hawoong Jeong, Sean P. Mason, Albert-László Barabási, and Zoltan N. Oltvai. Lethality and centrality in protein networks. *Nature*, 411:41–42, 2001. Brief communications.

[102] Björn H. Junker, Dirk Koschützki, and Falk Schreiber. Exploration of biological network centralities with centibin. *BMC Bioinformatics*, 7(1):219, 2006.

[103] Leo Katz. A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39–43, 1953.

[104] Ralph L. Keeney and Howard Raiffa. *Decisions with Multiple Objectives: Preferences and Value Trade-offs*. Cambridge university press, 1993.

[105] Douglas B. Kell and Stephen G. Oliver. Here is the evidence, now what is the hypothesis? the complementary roles of inductive and hypothesis-driven science in the post-genomic era. *BioEssays*, 26(1):99–105, 2004.

[106] Maurice G. Kendall. A new measure of rank correlation. *Biometrika*, pages 81–93, 1938.

[107] Maurice G. Kendall. The treatment of ties in ranking problems. *Biometrika*, pages 239–251, 1945.

[108] G. Kishi. On centrality functions of a graph. In *Graph Theory and Algorithms*, pages 45–52. 1981.

[109] Mitri Kitti. Axioms for centrality scoring with principal eigenvectors. *Social Choice and Welfare*, pages 1–15, 2015.

[110] Dirk Koschützki, Katharina Anna Lehmann, Leon Peeters, Stefan Richter, Dagmar Tenfelde-Podehl, and Oliver Zlotowski. Centrality indices. In *Network Analysis*, pages 16–61. Springer, 2005.

[111] Max Kotlyar, Kristen Fortney, and Igor Jurisica. Network-based characterization of drug-regulated genes, drug targets, and toxicity. *Methods*, 57(4):499–507, 2012.

[112] David Krackhardt. Assessing the political landscape: Structure, cognition, and power in organizations. *Administrative Science Quarterly*, 35(2):324–369, 1990.

[113] David Krackhardt. Social networks and the liability of newness for managers. In Cary L. Cooper and Denise M. Rousseau, editors, *Trends in Organizational Behavior*, volume 3, pages 159–173. Wiley, New York, NY, 1996.

[114] David H. Krantz, R. Duncan Luce, Patrick Suppes, and Amos Tversky. *Foundations of Measurement (Additive and Polynomial Representations), vol. 1*. Academic Press, New York, 1971.

[115] Ravi Kumar and Sergei Vassilvitskii. Generalized distances between rankings. In *Proceedings of the 19th International Conference on World Wide Web*, pages 571–580. ACM, 2010.

[116] Andrea Landherr, Bettina Friedl, and Julia Heidemann. A critical review of centrality measures in social networks. *Business & Information Systems Engineering*, 2(6):371–385, 2010.

[117] Eugene L. Lawler. Sequencing jobs to minimize total weighted completion time subject to precedence constraints. *Annals of Discrete Mathematics*, pages 75–90, 1978.

[118] Harold J. Leavitt. Some effects of certain communication patterns on group performance. *The Journal of Abnormal and Social Psychology*, 46(1):38, 1951.

[119] Chang-Yong Lee. Correlations among centrality measures in complex networks. *arXiv preprint physics/0605220*, 2006.

[120] Carlos León, Clara Lía Machado, and Miguel Sarmiento. Identifying central bank liquidity super-spreaders in interbank funds networks. *Available at SSRN 2413056*, 2014.

[121] Cong Li, Qian Li, Piet Van Mieghem, H. Eugene Stanley, and Huijuan Wang. Correlation between centrality metrics and their application to the opinion model. *The European Physical Journal B*, 88(3):1–13, 2015.

[122] Lun Li, David Alderson, John C. Doyle, and Walter Willinger. Towards a theory of scale-free graphs: Definition, properties, and implications. *Internet Mathematics*, 2(4):431–523, 2005.

[123] Min Li, Jianxin Wang, Xiang Chen, Huan Wang, and Yi Pan. A local average connectivity-based method for identifying essential proteins from the network level. *Computational Biology and Chemistry*, 35(3):143–150, 2011.

[124] Min Li, Jianxin Wang, and Yi Pan. A new essential protein discovery method based on the integration of protein-protein interaction and gene expression data. *BMC Systems Biology*, 6(1):15, 2012.

[125] Min Li, Jianxin Wang, Huan Wang, and Yi Pan. Essential proteins discovery from weighted protein interaction networks. In *Bioinformatics Research and Applications*, pages 89–100. 2010.

[126] Zhan-Chao Li, Wen-Qian Zhong, Zhi-Qing Liu, Meng-Hua Huang, Yun Xie, Zong Dai, and Xiao-Yong Zou. Large-scale identification of potential drug targets based on the topological features of human protein–protein interaction network. *Analytica Chimica Acta*, 871:18–27, 2015.

[127] Carlos Lozares, Pedro López-Roldán, Mireia Bolibar, and Dafne Muntanyola. The structure of global centrality measures. *International Journal of Social Research Methodology*, 18(2), 2015.

[128] Hong-Wu Ma and An-Ping Zeng. The connectivity structure, giant strong component and centrality of metabolic networks. *Bioinformatics*, 19(11):1423–1430, 2003.

[129] Nadimpalli V.R. Mahadev and Uri N. Peled. *Threshold graphs and related topics*, volume 56. 1995.

[130] Christopher D. Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. MIT press, 1999.

[131] Johannes S. Maritz. *Distribution-free Statistical Methods*, volume 17. CRC Press, 1995.

[132] Peter V. Marsden. Egocentric and sociocentric measures of network centrality. *Social Networks*, 24(4):407–422, 2002.

[133] Albert Marshall, David Walkup, and Roger Wets. Order-preserving functions: applications to majorization and order statistics. *Pacific Journal of Mathematics*, 23(3):569–584, 1967.

[134] Brian W. Matthews. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405(2):442–451, 1975.

[135] Russell Merris. Laplacian matrices of graphs: a survey. *Linear Algebra and its Applications*, 197:143–176, 1994.

[136] Carl D Meyer. *Matrix Analysis and Applied Linear Algebra*. SIAM, 2000.

[137] Stanley Milgram. The small-world problem. *Psychology Today*, 1(1):61–67, 1967.

[138] Mehryar Mohri. Semiring frameworks and algorithms for shortest-distance problems. *Journal of Automata, Languages and Combinatorics*, 7(3):321–350, 2002.

[139] Jakob L. Moreno. *Who Shall Survive? A New Approach to the Problem of Human Interrelations*. Nervous and Mental Disease Publishing Co, Washington, DC, 1934.

[140] Oskar Morgenstern and John Von Neumann. *Theory of Games and Economic Behavior*. Princeton University Press, 1953.

[141] Louis Narens. A general theory of ratio scalability with remarks about the measurement-theoretic concept of meaningfulness. *Theory and Decision*, 13(1):1–70, 1981.

[142] Assaf Natanzon, Ron Shamir, and Roded Sharan. Complexity classification of some edge modification problems. *Discrete Applied Mathematics*, 113(1):109–128, 2001.

[143] Mark E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45(2):167–256, 2003.

[144] Mark E. J. Newman. A measure of betweenness centrality based on random walks. *Social Networks*, 27(1):39–54, 2005.

[145] Reinhard Niederée. What do numbers measure?: A new approach to fundamental measurement. *Mathematical Social Sciences*, 24(2):237–276, 1992.

[146] Juhani Nieminen. On the centrality in a directed graph. *Social Science Research*, 2(4):371–378, 1973.

[147] Juhani Nieminen. On the centrality in a graph. *Scandinavian Journal of Psychology*, 15:332–336, 1974.

[148] Qikai Niu, An Zeng, Ying Fan, and Zengru Di. Robustness of centrality measures against network manipulation. *Physica A: Statistical Mechanics and its Applications*, 438:124–131, 2015.

[149] Ivan Niven. Formal power series. *American Mathematical Monthly*, 76:871–889, 1969.

[150] Ben Noble and James W Daniel. *Applied Linear Algebra*, volume 3. Prentice-Hall New Jersey, 1988.

[151] Jae Dong Noh and Heiko Rieger. Random walks on complex networks. *Physical Review Letters*, 92(11):118701, 2004.

[152] Tore Opsahl, Filip Agneessens, and John Skvoretz. Node centrality in weighted networks: Generalizing degree and shortest paths. *Social Networks*, 32(3):245–251, 2010.

[153] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank citation ranking: bringing order to the web. Technical Report 1999-66, 1999.

[154] Keunwan Park and Dongsup Kim. Localized network centrality and essentiality in the yeast–protein interaction network. *Proteomics*, 9(22):5143–5154, 2009.

[155] Wei Peng, Jianxin Wang, Weiping Wang, Qing Liu, Fang-Xiang Wu, and Yi Pan. Iteration method for predicting essential proteins based on orthology and protein-protein interaction networks. *BMC Systems Biology*, 6(1):87, 2012.

[156] Forrest R. Pitts. The medieval river trade network of russia revisited. *Social Networks*, 1(3):285–292, 1979.

[157] John Platig, Edward Ott, and Michelle Girvan. Robustness of network measures to link errors. *Physical Review E*, 88(6):062812, 2013.

[158] János Podani. A measure of discordance for partially ranked data when presence/absence is also meaningful. *Coenoses*, 12:127–130, 1997.

[159] Francesco Pozzi, Tiziana Di Matteo, and Tomaso Aste. Spread of risk across financial markets: better to invest in the peripheries. *Scientific Reports*, 3, 2013.

[160] Foster J. Provost, Tom Fawcett, and Ron Kohavi. The case against accuracy estimation for comparing induction algorithms. In *ICML*, volume 98, pages 445–453, 1998.

[161] Vijay Raghavan, Peter Bollmann, and Gwang S. Jung. A critical investigation of recall and precision as measures of retrieval system performance. *ACM Transactions on Information Systems (TOIS)*, 7(3):205–229, 1989.

[162] Karthik Raman, Nandita Damaraju, and Govind Krishna Joshi. The organisational structure of protein networks: revisiting the centrality–lethality hypothesis. *Systems and Synthetic Biology*, pages 1–9, 2013.

[163] Georg Rasch. *Probabilistic Models for some Intelligence and Attainment Tests.* Chicago:MESA, 1960.

[164] Teresa Reguly, Ashton Breitkreutz, Lorrie Boucher, Bobby-Joe Breitkreutz, Gary C. Hon, Chad L. Myers, Ainslie Parsons, Helena Friesen, Rose Oughtred, Amy Tong, et al. Comprehensive curation and analysis of global interaction networks in saccharomyces cerevisiae. *Journal of Biology*, 5(4):11, 2006.

[165] Yannick Rochat. Closeness centrality extended to unconnected graphs: The harmonic centrality index. In *ASNA*, number EPFL-CONF-200525, 2009.

[166] John E. Roemer. Axiomatic bargaining theory on economic environments. *Journal of Economic Theory*, 45(1):1–31, 1988.

[167] Richard B. Rothenberg, John J. Potterat, Donald E. Woodhouse, William W. Darrow, Stephen Q. Muth, and Alden S. Klovdahl. Choosing a centrality measure: epidemiologic correlates in the colorado springs study of social networks. *Social Networks*, 17(3):273–297, 1995.

[168] Britta Ruhnau. Eigenvector-centrality—a node-centrality? *Social Networks*, 22(4):357–365, 2000.

[169] Gert Sabidussi. The centrality index of a graph. *Psychometrika*, 31(4):581–603, 1966.

[170] David Schoch and Ulrik Brandes. Centrality as a predictor of lethal proteins: Performance and robustness. *MMB & DFT 2014*, page 11, 2014.

[171] David Schoch and Ulrik Brandes. Centrality-lethality hypothesis lacks support. *submitted*, 2015.

[172] David Schoch and Ulrik Brandes. Stars, neighborhood inclusion, and network centrality. In *SIAM Workshop on Network Science*, 2015.

[173] David Schoch and Ulrik Brandes. Re-conceptualizing centrality in social networks. 2016. To appear.

[174] Grace S. Shieh. A weighted Kendall's tau statistic. *Statistics & Probability Letters*, 39(1):17–24, 1998.

[175] Alfonso Shimbel. Structural parameters of communication networks. *The Bulletin of Mathematical Biophysics*, 15(4):501–507, 1953.

[176] Georg Simmel. *Soziologie. Untersuchungen über die Formen der Vergesellschaftung*. Duncker & Humblot, Berlin, 1908.

[177] Kimmo Soramäki and Samantha Cook. SinkRank: An algorithm for identifying systemically important banks in payment systems. *Economics: The Open-Access, Open-Assessment E-Journal*, 7(2013-28):1–27, 2013.

[178] Chris Stark, Bobby-Joe Breitkreutz, Teresa Reguly, Lorrie Boucher, Ashton Breitkreutz, and Mike Tyers. Biogrid: a general repository for interaction datasets. *Nucleic Acids Research*, 34(suppl 1):D535–D539, 2006.

[179] Dietrich Stauffer and Sorin Solomon. Physics and mathematics applications in social science. In *Encyclopedia of Complexity and Systems Science*, pages 6804–6810. Springer, 2009.

[180] Karen A. Stephenson and Marvin Zelen. Rethinking centrality: Methods and examples. *Social Networks*, 11:1–37, 1989.

[181] Stanley Smith Stevens. On the theory of scales of measurement. *Science*, 103(2684):677–680, 1946.

[182] Damian Szklarczyk, Andrea Franceschini, Michael Kuhn, Milan Simonovic, Alexander Roth, Pablo Minguez, Tobias Doerks, Manuel Stark, Jean Muller, Peer Bork, et al. The string database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Research*, 39(suppl 1):D561–D568, 2011.

[183] Xiwei Tang, Jianxin Wang, and Yi Pan. Identifying essential proteins via integration of protein interaction and gene expression data. In *Bioinformatics and Biomedicine (BIBM) 2012 IEEE International Conference on*, pages 1–4, 2012.

[184] Kar Leong Tew, Xiao-Li Li, and Soon-Heng Tan. Functional centrality: detecting lethality of proteins in protein interaction networks. In *Genome Inform*, volume 19, pages 166–177, 2007.

[185] Wenpin Tsai and Sumantra Ghoshal. Social capital and value creation: The role of intrafirm networks. *Academy of Management Journal*, 41(4):464–476, 1998.

[186] Thomas W Valente, Kathryn Coronges, Cynthia Lakon, and Elizabeth Costenbader. How correlated are network centrality measures? *Connections*, 28(1):16, 2008.

[187] Thomas W. Valente and Robert K. Foreman. Integration and radiality: Measuring the extent of an individual's connectedness and reachability in a network. *Social Networks*, 20(1):89–105, 1998.

[188] René van den Brink and Robert P. Gilles. Measuring domination in directed networks. *Social Networks*, 22(2):141–157, 2000.

[189] Piet Van Mieghem. Graph eigenvectors, fundamental weights and centrality metrics for nodes in networks. *arXiv preprint arXiv:1401.4580*, 2014.

[190] Michele Vendruscolo, Nikoley V. Dokholyan, Emanuele Paci, and Martin Karplus. Small-world view of the amino acids that play a key role in protein folding. *Physical Review E*, 65(6):061910, 2002.

[191] Sebastiano Vigna. A weighted correlation index for rankings with ties. In *Proceedings of the 24th International Conference on World Wide Web*, pages 1166–1176, 2015.

[192] Andreas Wagner and David A Fell. The small world inside large metabolic networks. *Proceedings of the Royal Society of London B: Biological Sciences*, 268(1478):1803–1810, 2001.

[193] Chenqu Wang, Wei Jiang, Wei Li, Baofeng Lian, Xiaowen Chen, Lin Hua, Hui Lin, Dongguo Li, Xia Li, and Zhicheng Liu. Topological properties of the drug targets regulated by microrna in human protein–protein interaction network. *Journal of Drug Targeting*, 19(5):354–364, 2011.

[194] Huan Wang, Min Li, Jianxin Wang, and Yi Pan. A new method for identifying essential proteins based on edge clustering coefficient. In *Bioinformatics Research and Applications*, pages 87–98. Springer-Verlag, 2011.

[195] Stanley Wasserman and Katherine Faust. *Social Network Analysis. Methods and Applications*. Cambridge University Press, Cambridge, UK, 1994.

[196] Duncan J. Watts and Steven H. Strogatz. Collective dynamics of 'small world' networks. *Nature*, 393:440–442, 1998.

[197] Benjamin D. Wright. Additivity in psychological measurement. *Measurement and Personality Assessment*, pages 101–112, 1985.

[198] Benjamin D. Wright. A history of social science measurement. *Educational Measurement: Issues and Practice*, 16(4):33–45, 1997.

[199] Ioannis Xenarios, Danny W. Rice, Lukasz Salwinski, Marisa K. Baron, Edward M. Marcotte, and David Eisenberg. Dip: the database of interacting proteins. *Nucleic Acids Research*, 28(1):289–291, 2000.

[200] Emine Yilmaz, Javed A Aslam, and Stephen Robertson. A new rank correlation coefficient for information retrieval. In *Proceedings of the 31st annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 587–594. ACM, 2008.

[201] Haiyuan Yu, Pascal Braun, Muhammed A Yıldırım, Irma Lemmens, Kavitha Venkatesan, Julie Sahalie, Tomoko Hirozane-Kishikawa, Fana Gebreab, Na Li, Nicolas Simonis, et al. High-quality binary protein interaction map of the yeast interactome network. *Science*, 322(5898):104–110, 2008.

[202] Milan Zeleny and James L. Cochrane. *Multiple Criteria Decision Making*, volume 25. McGraw-Hill New York, 1982.

[203] Ren Zhang, Hong-Yu Ou, and Chun-Ting Zhang. Deg: a database of essential genes. *Nucleic Acids Research*, 32(suppl 1):D271–D272, 2004.

[204] Xue Zhang, Jin Xu, and Wangxin Xiao. A new method for the discovery of essential proteins. *PLoS ONE*, 8(3):e58763, 2013.

[205] Mingzhu Zhu, Lei Gao, Xia Li, Zhicheng Liu, Chun Xu, Yuqing Yan, Erin Walker, Wei Jiang, Bin Su, Xiujie Chen, et al. The analysis of the drug–targets based on the topological properties in the human protein–protein interaction network. *Journal of Drug Targeting*, 17(7):524–532, 2009.

[206] Elena Zotenko, Julian Mestre, Dianne P. O'Leary, and Teresa M. Przytycka. Why do hubs in the yeast protein interaction network tend to be essential: reexamining the connection between the network topology and essentiality. *PLoS Computational Biology*, 4(8):e1000140, 2008.