Complex network models, graph mining and information extraction from real-world systems

Doctoral Thesis

by

András London

Supervisor

András Pluhár

PhD School in Computer Science

Department of Computational Optimization

University of Szeged, Szeged, Hungary

2018

Acknowledgment

First of all, I would like to thank to my supervisor András Pluhár, who has taught me mathematical (and skeptic) thinking, for his guidance and great patience. I am also very thankful to Péter Csermely for his encouragement to enter to the world of scientific research.

I am grateful to all my co-authors, particularly József Németh, Tamás Németh, Andor Háznagy, Anita Pelle, Christian Bongiornio and Rosario N. Mantegna who I can also call friends. Special thanks goes to my colleagues Tibor Csendes, Tamás Vinkó and Balázs Bánhelyi for their support.

I would also like to thank David P. Curley for scrutinizing and correcting this thesis from a linguistic point of view.

I would like to thank to all of my friends. I am a lucky man since it would be hard to list all of them. Without them, my life would not be as beautiful as it is.

Special thanks to Kornél, Norbi and Pircsi, who have always been there with me, for their constant support. Last, but not least, I wish to thank my parents who always supported and encouraged my education and my sister Mári and brother Gábor. I would like to dedicate this thesis to them as a way of expressing my gratitude and appreciation.

András London, January 2018.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

> "*Something else an academic education will do for you. If you go along with it any considerable distance, it'll begin to give you an idea what size mind you have. What it'll fit and, maybe, what it won't. After a while, you'll have an idea what kind of thoughts your particular size mind should be wearing. For one thing, it may save you an extraordinary amount of time trying on ideas that don't suit you, aren't becoming to you. You'll begin to know your true measurements and dress your mind accordingly*"
>
> J.D. Salinger, *The catcher in the rye*

Data-driven science is a rapidly growing area with the main goal being to extend the use of computers, from data analysis to making hypotheses. New knowledge simply emerges as plausible patterns found by mining data and related to these observed patterns, a range of questions can be addressed and hopefully answered. In essence, the aim is to generate knowledge from data. An analysis of the massive quantities of data produced by and about people, machines, and their interactions have received enormous attention by computer scientists, physicists, mathematicians, economists, political scientists, sociologists and bio-informaticians, among others, for the past few years. The huge amount of available data allows us to study *complex systems* that appear in such fields as biology, economics and the social sciences. Therefore, data mining, or knowledge discovery in large databases, has become one of the most important challenges in scientific fields and in industry, including for instance the pharmaceutical industry and the online social media organizations.

The development of *"small-world" networks* [183] has significantly changed and extended the research directions of graph theory, a part of mathematics which provides the theoretical toolkit for the study of complex systems. Alongside this, research on mining graph and network data has been increasingly growing over the past few years, and it has become the most promising approach for extracting knowledge from relational data [64]) and investigating complex systems [8]. Complex systems can often be represented by networks (or graphs), where nodes (also called vertices) stand for individuals or entities of the system, while links (also called edges) represent the interaction between pairs of these individuals (for some excellent reviews, see e.g. Newman, 2003 [145] and Boccaletti et al, 2006 [18]). The network approach is not only useful for simplifying and visualizing

enormous amounts of data, but it is also effective in identifying the most important elements and finding their key interactions. In essence, the aim of data mining is to generate knowledge from data by discovering common patterns and features in different data sets, while graph-based data mining, usually known simply as *graph mining*, is the extraction of knowledge from a graph (i.e. a network) representation of the data.

Complex network modeling and analysis and data mining have similar goals; namely, given the data representing a complex system, the goal is to extract (or synthesize) information from it, by creating a model (either a complex network representation, or a data mining model) on which successive steps of the analysis can be performed. The goal of this dissertation is to present the author's work which focuses on the development and application of network models and data mining tools for real-world problems.

In this dissertation, we commence with a brief introduction to the basics of graph theory, the main concepts of network science and data mining tools that are needed to understand later chapters.

Chapter 2 presents examples of real-life problems where the network approach is a natural way of mathematical modeling. Firstly, the author proposes a local PageRank algorithm, whose motivation comes from the area of "scientometics", to measure the influence of scientific papers using their local citation network. Then, a comprehensive analysis of public transportation systems will presented using various network models. The study provides a first step small-scale study of complex transportation systems of Hungarian cities by comparing their global and local characteristics. Lastly in this chapter the author will introduce potential network representations of a real social system based on educational data. Results of network analysis will also be presented.

Chapter 3 is concerned with economic networks, namely the international trade network and stock correlation based financial networks. In the first part of the chapter network analysis of the timely evolution of the trade network of the European Union is discussed. Afterwards, correlation-based financial networks will be defined and applied to the portfolio selection problem.

In Chapter 4, the task of rating nodes in networks is addressed and applied specifically in ranking sport teams and players. A novel, time-dependent PageRank method to rank players based on the results graph of a sport competition will be presented. Afterwards, a new network-based probabilistic model is introduced for forecasting in sports and it will be compared with the fundamental Bradley-Terry model and with experts' betting odds based on measures of accuracy and predictive power.

In Chapter 5, we focus on complex systems that can be modeled by bipartite networks. Firstly, a community detection methodology is presented using a statistically validated one-mode projection approach. It will be shown how the link validation-based filtering procedure necessarily increases the precision of the community detection and it is able to find the core of the communities even in the case of very noisy data. Then a generalized version of the PageRank and HITS algorithms will be adapted to bipartite networks in order to rank nodes in them. A case study for wine tasting events will then be discussed.

In Chapter 6, we summarize the dissertation both in English and in Hungarian.

1.1 Brief (Hi)story of Network Science

Historically, the study of networks has been in the domain of *graph theory*, a branch of discrete mathematics. Since 1736, when Leonhard Euler invented graph theory by solving the Königsberg bridge problem (to find a round trip that crosses each bridge of the city exactly once), graphs have been investigated from various perspectives and applied to a wide-range of real-life problems. Graph theory has proved to be one of the most powerful tools in mathematical modeling and the graph theoretical framework has provided solutions to many difficult practical questions. Such questions as to what the maximum flow is from the source to the sink in a network of pipes, how to assign n people to n jobs with maximum utility, and how many colors are enough to color the regions of a map without coloring two neighboring regions with the same color, etc. The first book on graph theory was published in 1936, written by a Hungarian mathematician named Dénes Kőnig. Also in the first part of the 20th century, remarkable achievements were made using graphs in some special context. For instance, in the social sciences in the early 1920s, where studies focused on relationships among people, such as friendships or communication between members of a group, and in economics, where trade and other economic transactions among nations or firms were investigated. Probability theory became widely used for investigating graphs after the seminal contributions of Erdős and Rényi [58, 59]. Their eight papers on the topic gave rise to *random graph theory*, while the probabilistic method became one of the most effective techniques in problem solving in graph theory and combinatorics. Another direction of research concentrated on graphs with very strict structures; these are the so-called *perfect graphs* [128]. These are far from being random, and pop up in several applications and beautiful theorems.

However, later it turned out empirically that the "typical" structure of graphs that model real relational data, i.e. the structure of *real-world networks*, is very different from the random graph defined by Erdős and Rényi. In the last two decades, some seminal papers gave rise to a new movement of direction, namely the study of *complex networks*. They are networks with a highly irregular structure, dynamically evolving over time and complex in the sense that their global properties and functioning are not obvious from the properties of their individual parts. Namely, these are the works of Watts and Strogatz (1998) [183] attempting to describe *small-world networks*[1] mathematically, and Albert and Barabási (1999) [9] describing a *"preferential attachment"* algorithm that generates *"scale-free" networks*[2] characterized by a *power-law degree distribution*. The preferential attachment model has been rigorously analyzed by Bollobás and Riordan [20] who cleared up and confirmed the heuristics associated with the model. It should also be added what Jackson points out ([93], Ch. 3), that many network degree distributions exhibit "fat-tails", like a power law, when compared to a Poisson random graph, but it is not clear that these distributions really are power laws.

Besides to the power law degree distribution or small-world properties, the *community*

[1] Small-world networks are often characterized by a small average path length and high clustering coefficient (see Sec. 1.2.2)

[2] A scale-free network is a graph whose degree distribution (see Sec. 1.2.2) follows a power law; i.e., $\Pr(d_i = k) = ck^{-\gamma}$, where d_i is the degree of a node i, c is a normalizing constant and $\gamma > 1$.

structure turned out to be a significant and common feature of complex networks; see e.g. [71] and [70] for a good introduction and survey, respectively. Practically speaking, *community detection* in a graph is a partition of the nodes into disjoint sets (often called communities, or clusters), such that nodes in the same community are more densely connected to each other than to the rest of the graph. Sometimes the so-called overlapping communities, where any node may be a member of more than one community, are in the focus of interest [24, 148]. In general, the communities in networks reflect the similarities and common features of the nodes that they contain. Newman and Girvan introduced the modularity optimization method to find communities in real-world networks [146]. Since then, a myriad of papers has appeared on the topic which became one of the most important topics in network science. Similar structural studies uncovered important *core/periphery network* structures [23], where the concept of the network core usually refers to a central and densely connected set of network nodes, while the periphery of the network denotes a sparsely connected, typically non-central set of nodes, which are linked to the core. A Core/periphery structure has been detected in many complex systems including biological networks, animal and human social networks and related networks, such as the World Wide Web and Wikipedia; engineered networks (such as the Internet, power-grids or transportation networks), as well as networks of the world economy. A survey in the topic by the author of this thesis and co-authors can be found in [44].

In parallel with the investigation of global network properties, the problem of rating and ranking nodes (representing real actors) in networks has also been extensively studied. One of the most important contributions from our perspective are those that at the end of 1990s, Sergey Brin and Larry Page, founders of Google Inc., developed a special random walk algorithm in networks that seeks to model the user behavior of Web graph surfing [28]. PageRank is mostly used as a network centrality measure (see Sec. 1.2.3) and utilizing PageRank can help us understand the complex network better by focusing on what PageRank reveals as important. Independent of Brin and Page, Kleinberg proposed a different approach to measure the importance of a web page [103]. While PageRank computes the pagerank scores on the entire graph, the Kleinberg's HITS algorithm (Hyperlink Induced Topic Search) tries to distinguish between hubs (nodes that link to many authorities) and authorities (nodes that have in-coming links from hubs) within a subgraph of relevant pages. The mathematics of PageRank and HITS, however, is general and can be applied to any graph or network in any domain, and it is successfully utilized in social and information network analysis as well as in biology, chemistry, neuroscience, and physics [79]. Modified versions of them with various applications will be discussed in different parts of the thesis.

Very recently, complex network theory and *data mining* have been used together in a variety of problems. Methods for extracting patterns from data have a long history and providing merely a brief history and introduction to data mining is beyond the scope of this study. It transpired that out that differences between network theory and data mining may, in some situations, provide an added value when both of them are used in combination [189]. The term data mining now mostly refers to the process of using

methodologies and techniques taken from different areas like statistics, probability theory, database technology, machine learning and data visualization. Viewing data mining as a tool for knowledge discovery/information extraction in a step-by step approach, from problem understanding via (mathematical) modeling to evaluation and deployment, was proposed in [185].

The road map of this thesis conceptually proceeds in the following way. Given a real-world system and associated problems, a complex network model is created to represent the system. Firstly, an analysis of the network (representing the system) can provide an initial picture of the nature and fundamental features of it and difficulties of the problem are addressed. However trying to solve special problems often leads to the development of new network models, new tools and techniques that may be applicable to a broader range of problems. This thesis seeks to present the author's work on the topic, but before going into the details, some basic definitions and concepts need to be presented.

1.2 Characteristics of Real-world Networks

Now, we will give a brief introduction to graph theory and an overview of the main definitions that characterize the structural properties of complex networks. Particular attention will be paid to the community structure and core/periphery structure as global characteristics of complex networks and stochastic graph algorithms like PageRank and HITS as they are widely-used graph-based data mining tools.

1.2.1 Basic Definitions

Formally, an undirected (directed) *network* or *graph* $G = (V, E)$ consists of two sets V and E, where $V \neq \emptyset$, while E is a set of unordered (ordered) pairs of elements of V. The elements of $V = \{1, 2, \ldots, n\}$ are called nodes (or vertices) and the elements of E are called links (or edges). A network is mathematically represented by its *adjacency matrix* $A = [a_{ij}]_{i,j=1,\ldots,n}$, which is an $n \times n$ matrix with entries $a_{ij} = 1$ if there is an edge (directed edge) between i and j and $a_{ij} = 0$ otherwise. For an undirected network if the (i, j) edge exists, then $a_{ij} = a_{ji} = 1$, i.e. A is symmetric. If a function $w : E \to \mathbb{R}$ that assigns a real number w_{ij} to each edge (i, j) is given, then we say that the network is *weighted*.

For a network G of n nodes the number of links lies between 0 (*empty graph*) and $n(n-1)/2$ (*complete graph*). G is said to be *sparse*, if $|E| \cong cn$ and *dense* if $|E| \cong cn^2$ where c is a positive-valued constant.

The *degree* d_i of node i is the number links that are connected to i. If the network is directed, we can define the *in-degree* d_i^+ and *out-degree* d_i^- of a node i, these being the number of incoming links to i and the number of outgoing links from i, respectively. The weighted degree of a node can be calculated in a similar way using $w_i = \sum_i w_{ij}$ $(i = 1, \ldots, n)$, which is sometimes called the *strength* of i.

A *subgraph* $G' = (V', E')$ of $G = (V, E)$ is a graph where $V' \subseteq V$ and $E' \subseteq E$. If it contains all links of G that connects two nodes in V', it is said to be the *induced*

subgraph by V'. A *clique* is a maximal complete subgraph of three or more nodes.

A *walk* $(i, k_1), (k_1, k_2), \ldots, (k_m, j)$ between two nodes i and j is an alternating sequence of nodes and edges, starting and ending at i and j, resp., in which each edge in the sequence is adjacent to its two endpoints. The length of the walk is the number of edges on it. If all the nodes along the walk are distinct, then the walk is a *path*. The *shortest path* between i and j is a path between them where the length of the path is minimized. The (sub)graph is (strongly) *connected* if, for every pair of nodes i and j of the subgraph, there is a (directed) path from i to j.

1.2.2 Global Characteristics

The *number of links* in a network, the *average degree* and the *link density* are computed using the following formulas:

$$m = \frac{1}{2}\sum_{i=1}^{n} d_i = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}, \quad \overline{d} = \frac{1}{n}\sum_{i=1}^{n} d_i = \frac{2m}{n}, \quad \rho = \frac{m}{\binom{n}{2}} = \frac{\overline{d}}{n-1}, \tag{1.1}$$

where the bar denotes the average.

Next, let ℓ_{ij} be the shortest path between nodes i and j. The *diameter* of the network is defined as the maximum of the shortest paths among all pairs of nodes. That is,

$$D(G) = \max_i \max_{j \neq i} \ell_{ij}, \tag{1.2}$$

The *average path length* is defined as

$$\overline{\ell} = \frac{2}{n(n-1)} \sum_i \sum_{j \neq i} \ell_{ij}, \tag{1.3}$$

which exists only if there are no unconnected nodes in the network and Eq. (1.3) is usually restricted to this case. In real (especially social) networks, the average path length is usually small, typically less than $\log n$.

The list of the node degrees is called the *degree sequence* of the network. The *degree distribution* $\mathcal{P}(d)$, a key characteristic of real-world networks, is defined as the fraction of nodes having degree d; or, equivalently, it is the probability that a uniformly randomly chosen node has degree d. In the case of directed networks, we can distinguish the in-degree and out-degree distributions. It has turned out that many real networks have a "fat-tailed" or "heavy-tailed" degree distribution. More precisely, power-law distributions, given in the form $\mathcal{P}(d) \sim cd^{-\gamma}$, have been observed many times [10]. Networks with such degree distribution are called *scale-free*[3].

The *clustering coefficient* is defined as

$$C = \frac{3 \times \#\{\text{triangles}\}}{\#\{\text{connected triples of nodes}\}} \tag{1.4}$$

[3]The term scale-free originated from a branch of statistical physics called the theory of phase transitions. The higher order moments of scale-free probability distributions are infinite and hence fluctuations around the average may be arbitrarily large. i.e. there is no meaningful internal scale.

and measures how the connected triples of a network tend to form triangles, which is very common e.g. in social networks emphasizing the paradigm that "two individuals with a common friend are likely to know each other".

Community structure

Another key property of complex networks is called the *community structure*. Finding communities (also called clusters) in a network informally means finding a way to partition the nodes into disjoint sets (subgraphs) such that nodes in the same set are more densely connected to each other than to the rest of the network. Typically, a community in a network means the similarity and common features of the nodes that it contains.

Numerous different algorithms have been developed to find communities in networks (for a comprehensive work of the topic, see e.g. [70]). We should mention here a widely used one called the "Leuven" method by Blondel et al. [17]. This is based on the *modularity* maximization method developed by Girvan and Newman [78]. This is a heuristic based on the idea that a null-model random graph is not expected to have a cluster structure like the original one. Given the network G, the modularity function which needs to be maximized, is defined as

$$Q(G) = \frac{1}{2m} \sum_{i,j} (w_{ij} - \frac{w_i w_j}{2m}) \delta(C_i, C_j), \tag{1.5}$$

which is a scalar-valued function that takes values between $-1/2$ and 1; w_{ij} represents the weight (or just presence, in the case of unweighted networks) of edge (i, j), w_i is the strength of node i (or just the degree), C_i is the community to which node i is assigned. Here, $\delta(C_i, C_j) = 1$ if $C_i = C_j$ and $\delta(C_i, C_j) = 0$ otherwise while m is the sum of the weights over all edges (or simply the total number of links in the unweighted case).

It has been shown that modularity maximization is NP-complete [26]. For this reason, several methods, ranging from simulated annealing to spectral optimization and greedy methods have been developed providing that the partitioning of a graph which gains the highest modularity value. In the case of the greedy Leuven method, initially each node of the network forms a community. The first step consists of a sequential sweep over all nodes. Given a node i, the gain in weighted modularity is computed. This gain comes from putting i in the community of its neighbor node j and picks the community of the neighbor that yields the largest increase of modularity, as long as it is positive. At the end of the sweep, the first level partition is obtained. For the next step, communities are condensed into single nodes, and two condensed communities ("supernodes") are connected if there is at least an edge between nodes of the corresponding communities. In this case, the weight of the edge between the supernodes is the sum of the weights of the edges between the given communities at a lower level. The two steps of the algorithm are repeated, yielding new hierarchical levels and "supergraphs".

Core and periphery of a network

Informally, the concept of a network core usually means a central and densely connected set of network nodes, while the periphery of the network represents a sparsely connected, typically non-central set of nodes, which are linked to the core. The "and" is important in the above informal definition, since all nodes of the core are mostly central, but certainly not every set of central nodes forms a network core. The concept of a network core may be approached from many directions (including various core defining algorithms; rich-clubs referring to an interconnected set of network hubs; network nestedness; the bow-tie structure of directed networks, as well as the highly robust onion network structures; for a detailed survey on core-periphery network, see [44]), and hence there are many possible definitions for it. In this thesis, we only describe the first formal approach to deal with core-periphery structure by Borgatti and Everett [23]. Their discrete approach was based on a comparison of the adjacency matrix A of the network with an ideal core/periphery network model consisting of a fully-linked core and a periphery that is fully connected to the core, but there are no links between any two nodes in the periphery. If δ denotes the (row) vector of length n with entries equal to one or zero, and if the corresponding node belongs to the core or the periphery, respectively, then $\Delta = \delta^t\delta$ is the adjacency matrix of the ideal core-periphery network of n nodes, where $\Delta_{ij} = 1$ if $\delta_i = 1$ and $\delta_j = 1$, and $\Delta_{ij} = 0$ otherwise. Determining how a core-periphery structure a network has is an optimization problem that attempts to find the vector δ such that the expression

$$\rho = \sum_{i,j} a_{ij}\Delta_{ij} \tag{1.6}$$

achieves its maximum value. The coefficient ρ is maximal when the adjacency matrix A and the matrix Δ are identical. Eq. 1.6 is essentially an unnormalized Pearson correlation coefficient applied to matrices rather than vectors. A network exhibits a core-periphery structure if the correlation between the ideal structure and the data is large. Of course, a statistical test for the significance of the core/periphery structures found by the algorithm is needed. For weighted networks, the optimal core/periphery subdivision is a partition obtained in a way that maximizes the weight of within core-group edges, and minimizes the weight of within periphery-group edges. A detailed description of calculations and implementations can be found in [23] and `https://sites.google.com/site/bctnet/`, respectively.

1.2.3 Local Characteristics

In complex networks, *centrality* generally refers to the class of measures that represent the most important and "central" nodes of the network from some given perspective. Here, we shall mention only a few that turned out to be interesting for several reasons in different fields.

The *degree centrality* is simply refers to the degree d_i of node i (in the case of directed networks, the in- and out-degrees are distinct) and it tells us how large the neighborhood

of i is. For example, it simply says that how many friends one has in her friendship network.

The *closeness centrality* [164] of a node i is defined as

$$C(i) = \frac{1}{\sum_{i \neq j} \ell_{ij}}. \tag{1.7}$$

Here, in general, the greater the value, the smaller the length of the shortest paths to all other nodes from i. The concept is important, for example, in the investigation of road and transportation networks, and in the analysis of information diffusion in social networks [22].

The *eccentricity* e of a node i is the longest distance between i and any other node in the network. That is,

$$e(i) = \max_{j \neq i} \ell_{ij}. \tag{1.8}$$

Let σ_{jk} be the number of shortest paths between nodes j and k and let $\sigma_{jk}(i)$ be the number of shortest paths between them that pass through node i. The *betweenness centrality* [73] of node i is defined as

$$BC(i) = \sum_{j \neq i \neq k} \frac{\sigma_{jk}(i)}{\sigma_{jk}}. \tag{1.9}$$

In complex networks, the larger the number of paths that pass through a certain node (or edge), the greater the betweenness of this node (or edge) and more central it is in this viewpoint. Betweenness has similar importance as closeness in the investigation of social (e.g. friendship), technological (e.g. transportation) and biological (e.g. protein-protein interaction) networks.

PageRank

The PageRank algorithm [28] was originally developed to measure and provide a good approximation of the importance of Web pages by considering their position in the Web graph. The PageRank score of a node $i \in V$ of graph G is defined as the recursion

$$PR(i) = \frac{\lambda}{n} + (1 - \lambda) \sum_{j \in N^+(i)} \frac{PR(j)}{d^-(j)}, \tag{1.10}$$

where $N^+(i) = \{j \in N : j \to i \text{ exists}\}$, which is the set of nodes having an edge to node i, while $\lambda \in [0, 1]$ is a free parameter (usually given a value between 0.1 and 0.2). The PageRank recursion formula defined by Eq. 1.10 can be written in vector equation form like so

$$\mathbf{PR} = [\frac{\lambda}{n} \mathbb{1}\mathbb{1}^T - (1 - \lambda) A D^{-1}] \mathbf{PR}, \tag{1.11}$$

where A is the adjacency matrix of G, D is a diagonal matrix such that $d_{ii} = \sum_{\ell=1}^{n} a_{i\ell}$ and $d_{ij} = 0$, if $i \neq j$, I is the $n \times n$ identity matrix and finally $\mathbb{1}$ is the n-dimensional vector that has each component equal to 1. Eq. 1.11 shows that $\mathbf{PR}$ is the eigenvector of

the matrix $(\lambda/n)\mathbb{1}\mathbb{1}^T - (1-\lambda)AD^{-1}$. Due to the fact that the corresponding eigenvalue of 1 is the largest eigenvalue of this matrix, which is a consequence of the *Frobenius-Perron* theorem for row-stochastic matrices [147], $\mathbf{PR}$ is in fact the *steady-state* solution a random walk on the nodes of the graph that can be described as follows. Starting from a node i, a random surfer selects one of the node's outgoing edges uniformly at random, moves to the end node j of that edge, and repeats this process from j, etc. The parameter λ can be understood as a "damping factor" which guarantees that the random walk restarts in some node of the graph, chosen uniformly at random, in every $1/\lambda$-th step, almost surely (i.e. with probability 1). This should guarantee that the process would not stop by reaching a node with zero out-degree. If the surfer reaches a node, the number of visits of that node increases by one. The damping factor ensures that each node receives a contribution λ/N for each step.

Assuming that $\mathbb{1}\mathbf{PR} = 1$, means that PageRank is a discrete probability distribution over the nodes of the graph, and Eq. (5.14) implies that the PageRank vector $\mathbf{PR}$ can be calculated as

$$\mathbf{PR} = \frac{\lambda}{N}[I - (1-\lambda)AD^{-1}]^{-1}\mathbb{1}, \tag{1.12}$$

and we can write

$$\mathbf{PR} = \frac{\lambda}{N}\mathbb{1}\sum_{k=1}^{\infty}((1-\lambda)AD^{-1})^k, \tag{1.13}$$

whose form gives us a useful power method for PageRank calculation (Alg. 1).

HITS

Independent of Brin and Page, Kleinberg [18] proposed a different approach to measure the importance of a web page. While PageRank computes the PageRank scores on the entire graph, the Kleinberg's HITS algorithm (Hyperlink Induced Topic Search) tries to distinguish between hubs and authorities within a subgraph of relevant pages, where hub scores and authority scores of the nodes are recursively calculated from each other. A good hub is a node that is connected to many authorities, while a good authority is a node that has in-coming links from good hubs. Mathematically speaking, the hub and authority scores can be calculated recursively as

$$h(i) = \sum_{j:i\to j} a(j) \quad \text{and} \quad a(i) = \sum_{j:j\to i} h(j), \tag{1.14}$$

where $a(i)$ and $h(i)$ are the authority and hub scores of node i, respectively (Alg. 2). The scores converge starting from any initial scores of the nodes. Writing this in matrix form, authority scores are calculated as $\mathbf{a} = A^T\mathbf{h}$, while hub scores are calculated as $\mathbf{h} = A^T\mathbf{a}$. Combining them, we get $\mathbf{a} = A^TA\mathbf{a}$ and $\mathbf{h} = AA^T\mathbf{h}$; hence it is apparent that they are principal eigenvectors of matrices A^TA and AA^T, respectively.

Algorithm 1: Power method for PageRank computation

Input : G directed graph
Output: PageRank vector $\mathbf{PR}$

Initialize $\mathbf{PR}_0 = \frac{\lambda}{N}\mathbb{1}$
$k = 1$
repeat
  $\mathbf{PR}_{k+1} := \frac{\lambda}{N}\mathbb{1} + \lambda A D^{-1}\mathbf{PR}_k$
  $k = k + 1$
until $||\mathbf{PR}_{k+1} - \mathbf{PR}_k||_1$
return $\mathbf{PR}_{k+1}$

Algorithm 2: HITS algorithm

Input : G directed graph
Output: Hub and authority scores of the nodes

Initialize all (node) weights to 1
repeat
  for all hub $i \in H$ **do**
    $h_i = \sum_{j \in F(i)} a_j$
  end for
  for all authority $i \in A$ **do**
    $a_i = \sum_{j \in B(i)} h_j$
  end for
until the weights converge
normalize

1.2.4 Brief Summary of the Author's Contribution

	[44]	[125]	[87]	[126]	[124]	[140]	[75, 127]	[123]	[21]	[122]
I.	•	•	•	•	•	•		•		•
II.							•		•	
III.									•	
IV.									•	•
Chapter	1, 2, 3	1	1	1	1	1	3	4	5	5

Table 1.1. Correspondence between the thesis points and publications/chapters

The following list summarizes the key points of the dissertation. Table 1.1 shows the connection between the thesis points and the publications of the author.

I. The author points out that many real-world systems can be modeled by networks and suggests using graph-based data mining and network analysis as a first step of investigating such systems. Each case study explains that, after collecting appropriate data, how the network approach, especially network analysis and applying rating methods, can be used to extract meaningful information from the system being modeled. New methods are also developed by slightly modifying some widely-used stochastic graph algorithms. In particular, a local PageRank approximation method and a new version of the generalized co-HITS are constructed to rate nodes in a network. The methods perform well in general and can be used for various

real-life problems modeled by networks.

II. The question of quantifying the degree of statistical uncertainty (usually called "noise") presents in real systems is addressed from different perspectives. Several methods were defined and used to filter the part of information which is robust against statistical uncertainty (i.e. robust against errors in the data or other sources of noise). In particular, network-based, random matrix theory-based and statistics based, methods were applied to correlation networks used for portfolio optimization and also used to detect cores of communities of bipartite networks. The results tells us that using these techniques, the classical Markowitz solution can be outperformed on the the one hand, and community cores can be found with high precision on the other.

III. The author demonstrates that information present in a bipartite network could be used to detect cores of communities of each set of bipartite system. Using Monte-Carlo simulations, the results indicate that the cores found are very stable and detecting them is very precise although the methodology may be not very accurate in some cases. The key concept is to consider statistically validated networks obtained by starting from the original bipartite network. The information carried by the statistically validated network can highly informative and could be used to detect communities of a given set that are robust with respect to the algorithm of detection and to the presence of errors or missing entries in the given database. Experimental results on real data are also presented. Staying with bipartite networks, the question of rating nodes of a bipartite network is also addressed. A general framework of a HITS type algorithm is presented for this purpose and a case study on a real data set is presented in detail. Our experimental results confirm that the method could be readily applied for many real-life situations.

IV. The problem of rating and ranking sport players and teams is addressed from a network analysis perspective. A time-dependent PageRank method is defined to rate players using the graph of game results data. The method gives a better picture than several broadly used methods and it is able to outperform them in terms predictive power. The author also proposes a novel rating-based forecasting framework. Against the widely known Bradley-Terry model, the key idea behind the model is that if a rating correctly reflects the actual performance of teams considered, then the smaller the changes in the rating vector, contains the ratings of the teams, after a certain event (final result) in an upcoming single game, the higher the probability of that event occurs. The results using several rating methods were compared to the Bradley-Terry predictions and the betting odds predictions of experts in terms of predictive accuracy. The authors showed that the new model outperforms the advanced versions of the Bradley-Terry model in many cases, even without fine tuning parameters and optimizing the implementation.

Chapter 2

Network Models for Some Real-life Problems

In this chapter we present some examples of real-life problems that can be modeled by complex networks. The analysis of these networks proved to be quite useful for gaining a better understanding the system being modeled, extracting meaningful information and answering certain specific questions.

First, following the network approach the main goal is to measure the influence of a single article regardless of the characteristics of the academic subject. Based on the previous results of [43] and by applying the experimental results of [37] that later mathematically proved to be applicable for many classes of graphs in [7], we use a local PageRank approximation for this purpose. It should be mentioned that we do not wish to attempt to determine the scientific worth of the articles, which will probably be judged in the future; rather we want to measure "the impact" of the papers in their field. We describe how a local PageRank method can be applied to determine the influence of a research paper. As a case study, we apply it to the co-citation graph of the well-known paper by Jenő Egerváry [53] and highlight the main advantages of our approach in Scientometrics.

Afterward, we will study engineered networks; namely we will analyze public transportation networks. We perform a comprehensive network analysis with the main goal of identifying the similarities of, and differences between the transportation networks of five Hungarian cities. In particular, we compare the global and local characteristics of the networks to get a detailed picture of the differences in the organization of public transport, which may have arisen for historical, geographical and economic reasons. As a result, we will highlight inconsistencies, organizational problems and identify which are the most sensitive routes and stations of the network.

Lastly in this chapter, we will introduce a novel example of a real social system, taken from the world of public education, which is suitable for network representation. We propose several network representations of certain educational data and show which are the most appropriate graph mining tools for analyzing them and what kind of additional information can be extracted by their usage. Depending on the construction of the underlying graphs, we present four families of network models and describe a case study using one of the models. We point out several advantages of graph-based data mining

techniques in educational systems.

2.1 Citation Networks and Scientometrics

The relevance of Scientometrics – which seeks to measure the productivity and quality of scientific research – has long been discussed in the academic domain. The most popular measures are the scientific citation indices due to their easy accessibility. Several of these indices have been introduced such as the h-index (or Hirsch-index) proposed by Hirsch [89], the g-index proposed by Egghe [54], the w-index and the maximum index both proposed by Woeginger [186]. Each of them is based on the citation records of the researchers. These indices have been extensively criticized since they are highly dependent on the scientific field (like the number of active researchers and available journals, popularity of the area and gender ratio etc.; see, e.g. [2, 107, 184]). Another drawback of their usage is that they do not give a clear picture of the influence and quality of any given paper.

Several studies have sought to address this problem using the network approach. Co-citation networks – in which nodes represent single articles and a directed edge represents a citation from a citing article to a cited article, describes the relation between citations of different papers – were widely studied previously [34, 97, 116]. Chen et al. [35] applied the PageRank algorithm to co-citation networks. Later Raddichi et al. [156] defined an iterative ranking method similarity to different ranking algorithms such as PageRank, CiteRank [180] and HITS in order to evaluate the influence of single articles by using co-authorship networks. In these networks networks nodes represent publications and weighted edges represent the number of common authors among them. Several modifications and variants of network models have been introduced in the context of Scientometrics (see, e.g. [65, 121, 171, 188]).

More recently, the Eigenfactor Score and the Article Influence Score [16] have been developed to estimate the relative influence of single articles based on citation networks as well. Besides this, we should mention that the underlying algorithms can also be applied to journals, authors, and institutions.

2.1.1 A Local PageRank Approximation

Although in many applications PageRank scores need to be computed for all nodes of the graph, there are situations where one is interested in computing PageRank scores only for a small subset of the nodes. Chen et al. [37] developed an algorithm to approximate the PageRank scores of target nodes of a graph with high precision. Their algorithm crawls a small subgraph around the target node(s) and applies various heuristics to calculate the PageRank scores of the nodes at the boundary of this subgraph. Then it computes the PageRank of the target node(s) by just using the crawled subgraph and the estimates for the boundary nodes. With simulations, they showed, on the one hand, that this algorithm gives a good approximation on average. On the other hand, they also pointed out that high in-degree nodes could make the algorithm very expensive and imprecise.

From now on, we will use the same notions as in [7]. An algorithm is said to be an *ε-approximation* of the PageRank, if for a graph $G = (V, E)$, a target node $i \in V$ and a given error parameter $\varepsilon > 0$, the algorithm outputs a value $PR'(i)$ satisfying

$$(1 - \varepsilon)PR_G(i) \leq PR'(i) \leq (1 + \varepsilon)PR_G(i), \tag{2.1}$$

where $PR_G(i)$ is the PageRank value of node i in the original graph. For a directed path $p = (k_1, \ldots, k_t)$ from node k_1 to k_t, we define $w(p) = \prod_{i=1}^{t-1} 1/d_i^-$, that is the reaching probability of k_t from k_1 in a given path, where the transition probability values are proportional to the number of outgoing edges. Let $p_t(i, j)$ be the set of all directed path of length t from i to j. Then the *influence* of node i on the PageRank of node j at radius t is defined as

$$I_t(i, j) = \sum_{p \in p_t(i,j)} w(p), \tag{2.2}$$

and hence, the total influence of i on j is

$$I(i, j) = \sum_{t=0}^{\infty} I_t(i, j). \tag{2.3}$$

Using the definition of influence, the PageRank of node j at radius r can be defined as

$$PR_G^r(j) = \frac{\lambda}{n} \sum_{t=0}^{r} \sum_{i \in V(G)} (1 - \lambda)^t I_t(i, j). \tag{2.4}$$

It can be proved that for every node $j \in G$, $PR_G(j) = \lim_{r \to \infty} PR_G^r(j)$ holds (whose proof can be found in [7], say). An interesting question is that how small the radius r can be such that the PageRank approximation would even be appropriate.

In [7], it was proved that the hardness and inappropriate nature of local approximation of PageRank on certain graphs (constructed examples) is caused by two factors; namely, the existence of high in-degree nodes and the slow convergence of the PageRank iteration algorithm. We shall see that in our case (and in most of the co-citation graphs, along with most real-world networks) these properties do not hold. It was also shown that the several variants of the approximation algorithms proposed by Chen et al. are still efficient on graphs that have bounded in-degrees and admit fast PageRank convergence.

We are given a graph $G = (V, E)$, a node $j \in V$ and the approximation parameter ε. The *point-wise influence mixing time* of j is defined as

$$T_G^{\varepsilon}(j) = \min\{r \geq 0 : \frac{PR_G(j) - PR_G^r(j)}{PR_G(j)} < \varepsilon\}. \tag{2.5}$$

The algorithm we use computes $PR_G^r(j)$ for a given node j and it follows from the definitions that it runs with $r = T_G^{\varepsilon}(j)$ and gives an ε-approximation of PR. To complete the description of the theoretical background, we should examine the upper bound on $T_G^{\varepsilon}(j)$ (i.e. on radius r).

For graph $G = (V, E)$ with $j \in V$ and $r \geq 0$ the *crawl size* at radius r is defined as

$$C_G^r(j) = \#\{i \in G : \exists p_t(i, j) \text{ with } t \leq r\}, \tag{2.6}$$

which is the number of nodes within a distance r from j. It follows from the definition that if the local PageRank algorithm runs for r iterations, its cost is $C_G^r(u)$. A trivial upper bound for the crawl size is that $C_G^r(u) < d^r$, where d is the maximum in-degree of G. It suggests that if both r and the maximum in-degree are low, the brute force algorithm, which uses Eq. 2.4 by recursively calculating the influences, is efficient. It can be also proved that for any directed graph G, for the number of iterations that the local PageRank algorithm needs to run, $r = \mathcal{O}(\log(1/PR_G(j)))$ is always sufficient (while in practice, such as in our case, a much lower radius could be enough).

2.1.2 Reaching Probabilities

A possible simplification of the PageRank method is just consider the reaching probability values of the nodes in the network. We would like to know the probability of reaching a node j starting from an arbitrarily chosen node $i \neq j$ of the network. The reaching probability, RP of a node j can be defined as the recursion

$$RP(j) = \sum_{i \in N^+(j)} p_{ij} RP(i), \tag{2.7}$$

where p_{ij} is the reaching probability of node j from a neighbor node i. Now it is natural to assume that reaching any neighboring node from i has the same probability, so we can use $p_{ij} = 1/d^-(i)$ in Eq. 2.7. With this choice, Eq. 2.7 is the PageRank equation without the damping factor. However, in contrast to the calculation of PageRank, we do not wish to evaluate the vector $\mathbf{RP}$ in the steady state. Instead, we will only determine the reaching probability of a given node j, which can be calculated as

$$RP(j) = \frac{1}{n} \sum_{i \in V} I(i, j) \tag{2.8}$$

where $I(i, j)$ is as defined in Eq. 2.3. For published articles, RP can be interpreted as the probability that a given article found by someone (e.g. a scientist) starts to read an article and "jumps" to another randomly chosen article cited by the current one.

2.1.3 A Case-study

A co-citation network is defined as a directed graph $G = (V, E)$ of n nodes, where each node $i \in V$ refers to an article and there is a directed edge $(i, j) \in E$ from node i to node j if article j is cited in article i. Our method, which seeks to measure the "influence" of a scientific article, is based on the following three steps:

1. *Subgraph building*: Start from a set of target nodes (articles) we are interested in their scientific impact, expand backward in the reverse direction the nodes having

out-going links to the target nodes. The procedure halts after a fixed number of levels. This may be performed via an iteratively deepening depth-first search. In this task, the graphs contain all nodes, from which the target nodes can be reached in at most three steps and we consider the induced subgraph of these nodes.

2. *Estimating PR of the boundary*: We use a heuristic to estimate the individual PR scores in the boundary. An extra term to the PR value of each boundary node is added that represents the fraction of its in-coming edges to all edges in the subgraph.

3. *Calculating PR and RP*: We run the PageRank algorithm on the subgraph. In each step for the boundary nodes, the estimated PR value is used and the PageRank damping factor value is added to each node. In addition, we also calculate the reaching probability (RP) of the target node(s) in the subgraph (using the same values for the boundary nodes as in the case of PageRank).

Although the PageRank values cannot be calculated exactly without having to run the algorithm on the full graph, the estimation heuristic we defined gives an acceptable approximation of them in the constructed subgraph. We also note that the convergence of the PageRank is guaranteed by this method, unlike that defined by Csendes and Antal for the same purpose [43]. Here, we set the radius size $r = 3$ from the target nodes for two reasons. The first is that the number of nodes in the fourth layer is $\mathcal{O}(n)$ (here, n is the number of nodes of the crawled graph G) and the in-degrees are bounded by a constant, hence, it is sufficient to consider the number of in-coming links to the boundary nodes from the fourth layer, and ignore the linking structure between them to get a good approximation of the PR-scores. The second reason is we assume that for the articles at a distance more than four, the target articles do not have much impact in any scientific sense (which may be a reasonable assumption in Scientometrics).

Algorithm 3: The local PageRank approximation algorithm

Input : Seed article
Output: The PR-score of the article from its local co-citation network

Build the article's local co-citation network G with radius r
$PR^0_G(j) = \frac{\lambda}{n}$
$\text{layer}_0 = j$
$I_0(j, j) = 1$
for node i in layer_r **do**
  $PR(i) = |N^+(v)|/|E(G)|$
end for
for t=1, . . . , r **do**
  for i in layer_t **do**
    $I_t(i, j) = \frac{1}{d_i^-} \sum_{k:i\to k} I_{t-1}(k, j)$
  end for
  $PR^t_G(j) = PR^{t-1}_G(j) + \frac{\lambda}{n} \sum_{i\in\text{layer}_t} (1-\lambda)^t I_t(i, j)$
end for
return $PR^r_G(j)$

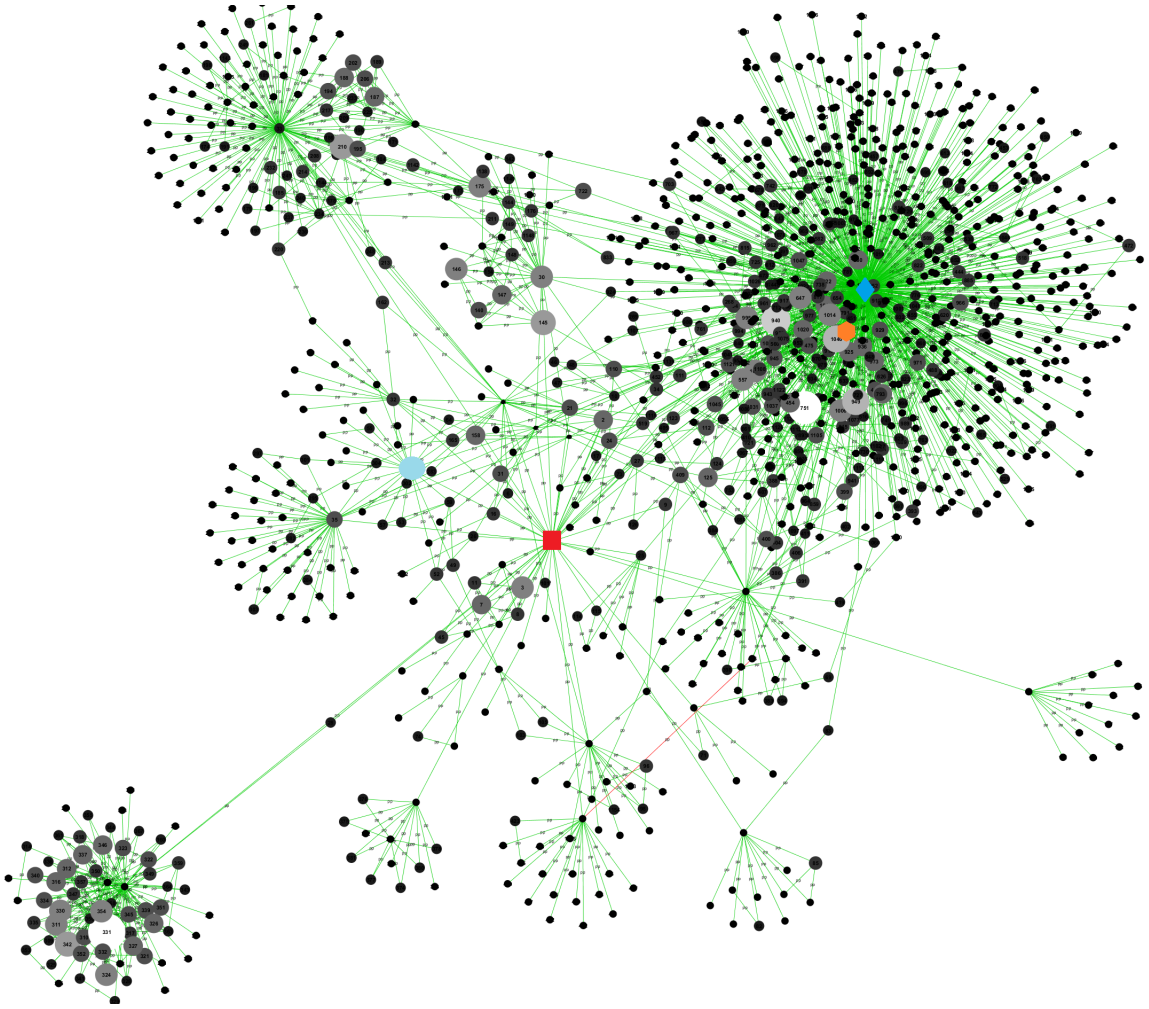


Figure 2.1. Local co-citation network containing the famous paper of Egerváry marked by red square. The blue diamond, the orange hexagon and the light blue circle represent Kuhn's paper, Ford and Fulkerson's paper and Bellman's paper, respectively. The size of the nodes refers to the number of citations.

Egerváry's Paper and its Citation Network

As is known, Harold Kuhn developed an algorithm for solving the assignment problem [106] and he called it the Hungarian method, acknowledging the contribution of Jenő Egerváry and Dénes Kőnig [53, 104]. The paper by Egerváry received just a few citations (probably because it was written in Hungarian), while some of its citing papers received many more: for Egerváry's paper 38 citations can be found in the ISI Web of Knowledge database, while the article by Kőnig and Kuhn received 215 and 726, respectively. In contrast to classic scientometrics that only takes into account the direct number of citations, we shall see that the network-based methods provide a more realistic picture of the importance of Egerváry's paper.

We constructed a network which contains the following articles as nodes: the famous paper by Jenő Egerváry: *On combinatorial properties of matrices* (published in Hungarian, 1931), the three articles which are referred in Egerváry's paper, and citing articles of these papers within the radius $r = 3$ in the network. We will now examine the network that is induced by these nodes, as described in the first phase of our method; it contains

Table 2.1. PR-score (with $\lambda = 0.2$), reaching probabilities and number of citations of the well-known publications in the Egerváry co-citation graph. The PR value has been multiplied by 100.

Publication	PR-Score	PR-rank	RP-score	RP-rank	#Cites	Cite rank
Egervári [53]	0.891	4	0.009	2	39	65
Kuhn [106]	1.189	1	0.042	1	726	1
Ford, Fulkerson [67]	0.525	8	0.004	9	39	65
Bellman [14]	0.399	11	0.003	10	18	158

$n = 1155$ nodes and 1923 edges. Figure 5.7 shows the network, where the paper by Egerváry is marked by red square. We applied the modified local PageRank algorithm (with $\lambda = 0.1, 0.15, 0.2, 0.25$) to this network and also calculated the reaching probabilities of the nodes. We observed that the PageRank method is robust against the choice of λ. The results (with $\lambda = 0.2$) are summarized in Table 2.1 for four notable publications in the co-citation network.

Observations

First, we observe that the choice of the damping factor λ does not influence the final ranking of the first ten publications; only small changes can be seen in the rest of the rank list. The ranks and the relative values of the papers provide a more realistic picture of their importance than the number of their citations. It is not surprising that Kuhn's paper PR value is the highest by far, the 726 citations for this paper being extraordinary in the field. The second and third articles in the PR ranking became D. Kőnig: *Graphs and their applications for the theory of determinants and sets* (1916, in Hungarian, 215 citations) and G. Frobenius: *Über zerlegbare Determinanten* (1917, 11 citation), respectively. Both articles were cited in Egerváry's paper, which became the fourth highest ranked paper although it received only 39 citations and it comes 65th in the citation ranking. The very high position of Frobenius's paper in the ranking is definitely due the reputation it receives from Egerváry's article. It is worth stressing that Ford and Fulkerson's article, which received the same number of citations as Egerváry's article, was ranked lower but it is still in the top ten. These two facts also tells us about advantages of the network-based evaluation, since this paper was also quite important in the development of Operations Research. We should also mention, that the important paper of Bellman was ranked 11th (although it got just 18 citations), which offers a much clearer picture of its impact (unlike its citation rank). It is also interesting to see that Egerváry's article comes second in RP-ranking, which means that someone who comes across the article at random and checks the articles in the field will find the Egerváry's paper with the second highest probability.

2.2 Modeling Transportation Networks

A great deal of effort has been concentrated on investigating transportation systems for many decades because of its practical importance. In the past decade, partly due to the development of small-world networks and modern network theory, several studies

have treated public transportation systems as complex networks, and several statistical properties have been discovered, like the small-world property and scale-free distribution of various graph parameters [50, 115, 163, 179]. In most of these studies, the public transportation network (PTN) model represents nodes as stations and stops of a public transportation system, and edges that connect consecutive stations along a route.

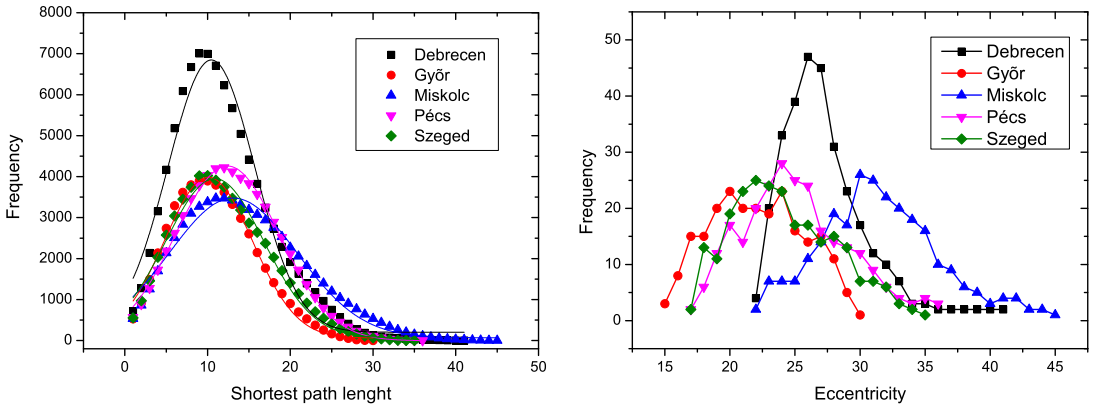
2.2.1 Data Collection and Modeling

We selected 5 Hungarian cities (Debrecen, Győr, Miskolc, Pécs, Szeged) to study their urban public transportation systems. The choice of the cities was based on the following criteria: (i) we are especially interested in cities with a population between 100,000 and 250,000; (ii) the characteristics (like land use and economic role) and the organization of the public transportation of these cities are similar; but (iii) the geographical conditions (landscape, hydrography, size of the area) are different. The areas lie between 162 and 462 km^2 (so these are medium-sized), but their urban morphology is different. In Miskolc and Pécs the land undulates, while in Győr and Szeged a river which crosses the city is the main factor that determines the shape of the city. In Debrecen, there are no restricting factors on the morphology. We should also mention that railway tracks may have a similar role to that of the rivers on the morphology. This phenomenon appeared in all the cities investigated. The above-mentioned characteristics have had a high impact on the development of the cities and also on the organization of the public transportation systems. Table 2.2 summarizes the basics characteristics of the cities and their PTN network models. Here, "links-simple" refers to the number of links in the simplified graphs (no multiple edges), while "links-multiple" refers to the number of links in the model where each line between two stations is represented by a link. In order to perform a comprehensive network analysis of the public transportation network of these cities, the first step was to generate the transportation networks (i.e the representing graphs). This was done by modeling stations/stops as nodes and lines that connect them as directed links. If a line runs between two stops in both directions, as is usually the case, we can decompose the link that represents this line into two directed links due to the orientation. Furthermore, we can also assign weights for each node and each edge by using the capacity of the vehicles. This can be performed as follows:

1. Assign the lines to the stations where they stop by using the transport schedules.
2. Classify the stations that belong together.
3. Determine the *morning peak hour capacity* of each vehicle using the types of the vehicles (the data provided by the public transport companies of the cities).

Table 2.2. The codes of the vehicle types are as follows: B: bus, E: electric trolleybus, T: tram

City	Area (km^2)	Pop. ($\times$ 1000)	Density (inhab./km^2)	Nodes	Links–simple	Links–multiple	Lines	Diameter	Avg. path length	Vehicle types
Debrecen	461	204	442.5	306	711	1772	53	41	11.7	BET
Győr	174	129	741.4	230	529	1391	43	30	10.8	B
Miskolc	236	161	682.2	257	535	977	35	45	14.5	BT
Pécs	163	147	901.8	256	569	1960	55	36	13	B
Szeged	281	162	576.5	242	558	1192	40	35	11.8	BET



(a) (b)

Figure 2.2. **a**: Shortest path length distribution; solid lines show a fit to the function. **b**: Eccentricity distribution.

Merging the stations into a single one was necessary for the following reasons. It frequently occurs that stops belonging to the same node have different names. In a special case, it can happen that there are four different names of the same stop in a 4-way crossroads. On the one hand these stops can be viewed as just one stop, while on the other hand this classification allows us to unambiguously cover the road network of the city with the PTN. In a big public transportation interchange or terminal where a high number of lines intersect, usually the lines have different stops. These stops were also merged. In the case where a line makes two stops in two stations that were merged, we will treat it as just one stop of the line for this node. In the case where the route of the line is a one-way instead of a two-way between two consecutive stations, the stops were not merged.

A calculation of the maximal capacities of the different lines was performed based on the evaluation of the vehicle capacities[1] in the morning peak hours (6-8 am). For each single line, we collected the follow-up interval of it and multiplied it by the capacities of the vehicles belonging to this line between 6-6:59 am and 7-7:59 am. By averaging the two values, we obtained the *average morning peak hour capacity* (AMPHC) of the line. For each node and link, we assigned the sum of AMPHCs of the lines that stop at that node or pass through that link between two consecutive stations at least once. By considering the morning peak hours, it can be seen that number of passengers that go from the outer districts to the inner city area is significantly higher than the number passengers that go in the opposite direction. Based on this observation, we are able to identify traffic source and traffic sink districts. We should also mention that all of this data is available on `http://www.epito.bme.hu/uvt/dolgozok/feltoltesek/haznagy/ptncomplexanalysis.zip`.

[1]The following types of vehicles are considered: mini bus: 30 persons; normal bus/trolleybus: 60 persons; articulated bus/trolleybus: 100 persons. In the case of trams, the situation is more complicated. The types of trams are different for every city; moreover the passenger capacities are calculated in different ways by the different manufacturers. To calculate the tram capacities, we used the formula $3 \times (\ell w - sr^2)/10^6$, where ℓ is the length of the vehicle, w is the width of the vehicle (both in millimeters), s is the number of seats and r^2 is the are taken up by a single seat with $r = 500$ millimeter. Here, we got the following results: Debrecen: Ganz KCSV6: 142 (persons), CAF Urbos 3: 227; Miskolc: Skoda 26T: 214, Tatra KT8D: 187; Szeged: Tatra T6A2: 81, Tatra KT4: 101, Pesa 120Nb: 187.

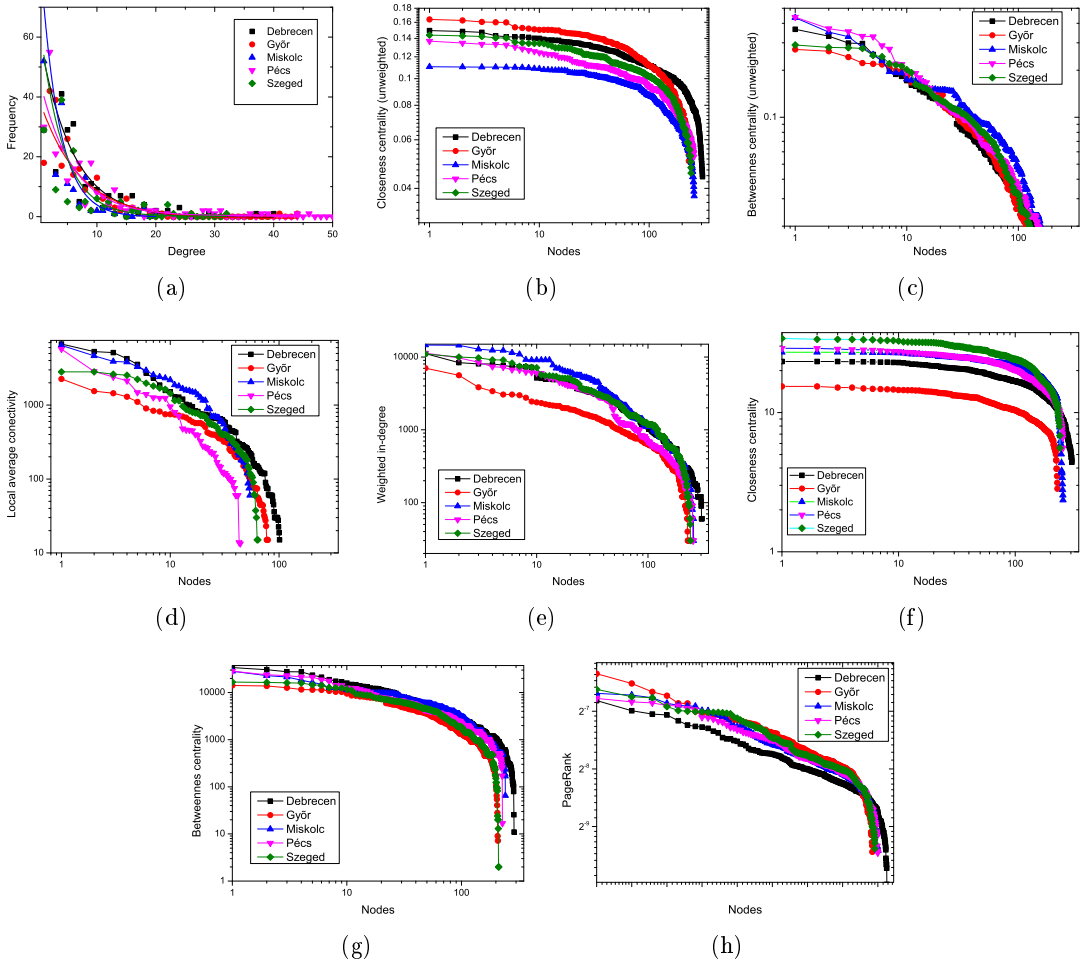


Figure 2.3. Local network properties of five Hungarian PTNs

2.2.2 A Comprehensive Network Analysis

Global Network Characteristics

We performed an analysis of the networks both in the weighted and unweighted case[2].

In practice, the diameter presents the longest route (i.e. number of stations along the longest route) in the network if a passenger uses the optimal routes, which means that she uses the shortest route between any two stations. In Table 2.2 we list the diameters for the PTNs. It is interesting to note that the diameter does not correlate with the area of the city.

The average path length corresponds to how many stations there are between two stations on the shortest route on average, if we choose these stations randomly. We can see in Table 2.2 that the the PTNs reveal a small-world feature from the average path lengths point of view, since $\bar{\ell} \sim \log N$, i.e. the average distance between the nodes is proportional to the logarithm of the number of nodes. The number of shortest paths

[2]An extension of the definition of the centrality measures to weighted networks can be performed using the w_{ij} edge weights in the following way, say. The weighted degree of a node i is simply defined as $w_i = \sum_j w_{ij}$. In the case of PageRank, w_{ij}/w_i is used d_j^-. The weighted closeness and betweenness can be defined by using $c_{ij} = 1/w_{ij}$ and $dist(i,j) = \sum_{(u,v)\in P} c_{uv}$, where P is a path between i and j and the weighted shortest path ℓ_{ij}^w is defined as the minimum of $dist(i,j)$.

from a node i is defined as $\ell_i = \sum_{i \neq j} \ell_{ij}$. Fig. 2.2(a) tells us that the distribution of the shortest paths is close to a normal distribution with mean that varies between 10.8 and 14.5 (Table 2.2) and variance between 5.2 and 7.7.

The eccentricity tells us how far a stop/station is from the most distant stop in the PTN. In Fig. 2.2(b) we plotted the eccentricity distribution of the five PTNs. The shape of the function is quite different in the case of Debrecen, due to its extensive area and Miskolc, where many peripheral areas increase the distances between certain stops/stations.

Fig. 2.3(a) shows the degree distributions in the unweighted case, where multiple links are allowed, which has an exponential decay $\mathcal{P}(d) \sim \exp(-d/\hat{d})$, where $\hat{d}$ is of the order of the average node degree. In contrast, the weighted degree distribution (Fig. 2.3(e)) of the (weighted) networks has a power-law decay $\mathcal{P}(d) \sim d^{-\gamma}$, where γ varies between 1.05 to 1.2.

In order to find communities, we use the Leuven *modularity optimization method* described in Sec. 1.2.2. The communities of the PTNs are shown in figures 2.4(b), 2.4(d), 2.4(f), 2.5(b) and 2.5(d). The results indicate the following common features of the networks. On the one hand, for each city, the center of it contains one or two communities and most of the peripheral lines have different community classes. On the other, we observed that if the city lies in a valley (Miskolc) or is bounded by mountains on one side (Pécs) and hence the arrangement of the city is asymmetric, then it has some special characteristics. The central core of the networks have been extended (figs. 2.4(e) and 2.5(a)) and this part of the transportation network can be partitioned into three or four communities.

Local Network Characteristics

The degree d_i of node i (in the case of directed networks, the in- and out-degrees are used) tells us how big the neighborhood of i is. The weighted in-degree centralities of the five PTNs can be seen in Fig. 2.3(e). The distributions have a power-law decay, as we noticed earlier.

Let N_i be the set of neighbors of u and $G[N_i]$ be the subnetwork *induced* by the nodes in N_i. The degree of a node j in the subnetwork $G[N_i]$ is denoted by $d^{G[N_i]}(j)$. The *local average connectivity* [118] of node i is defined as

$$LAC(i) = \frac{1}{d_i} \sum_{j \in N_i} d^{G[N_i]}(j) \tag{2.9}$$

and it describes how close its neighbors are. In a public transportation system it basically means that if a stop/station cannot be used for some reason, the neighboring stops become disconnected from each other. Nodes with high LAC values are the locally central nodes. Fig. 2.3(d) shows the distribution of LAC for the 5 PTNs. We observed that the distributions fit a power-law decay with degree exponent between 1.2 to 1.4. The closeness centrality values for the unweighted case and weighted case can be seen in figures 2.3(b) and 2.3(f), respectively. The distributions display similar shapes for each city,

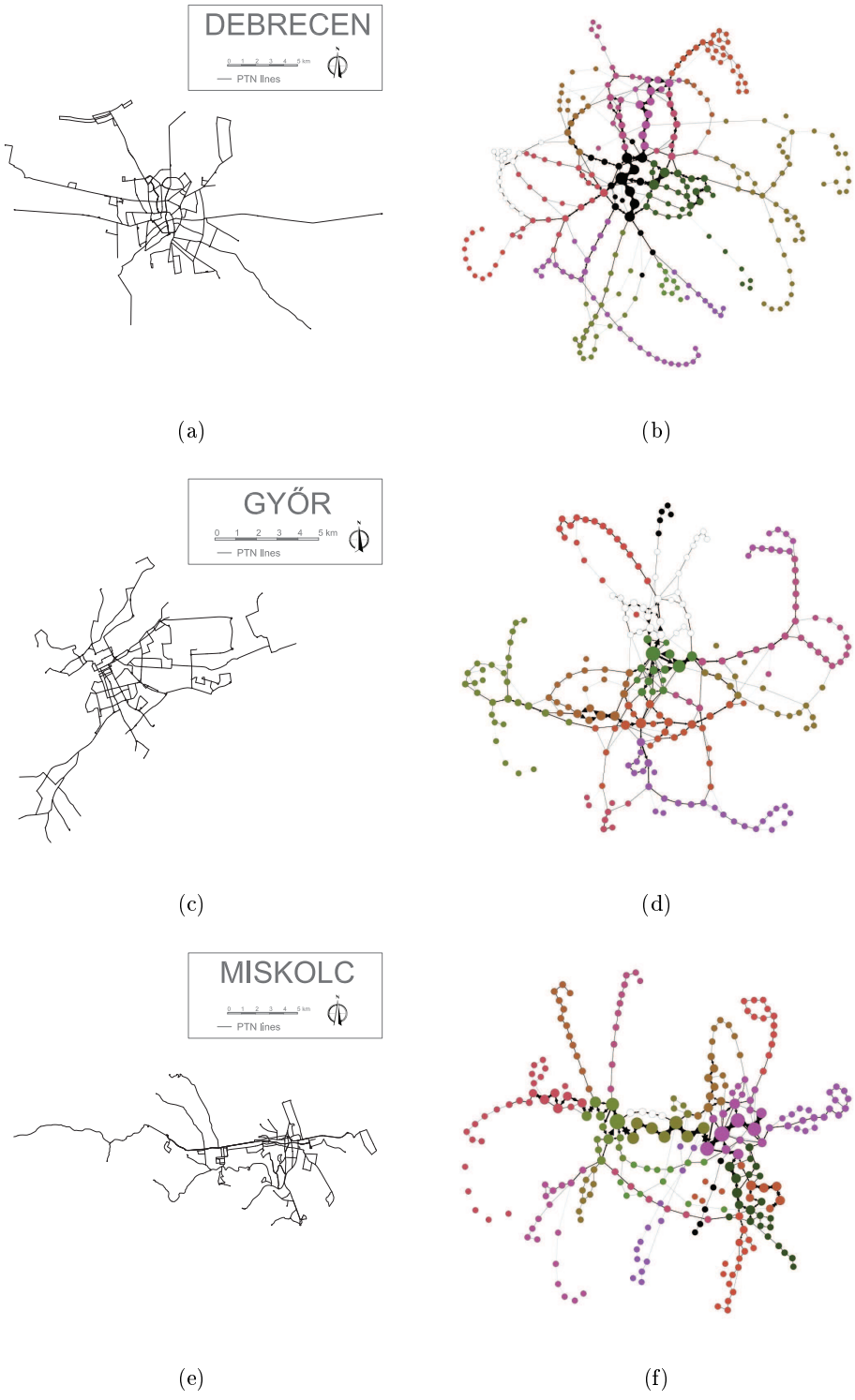
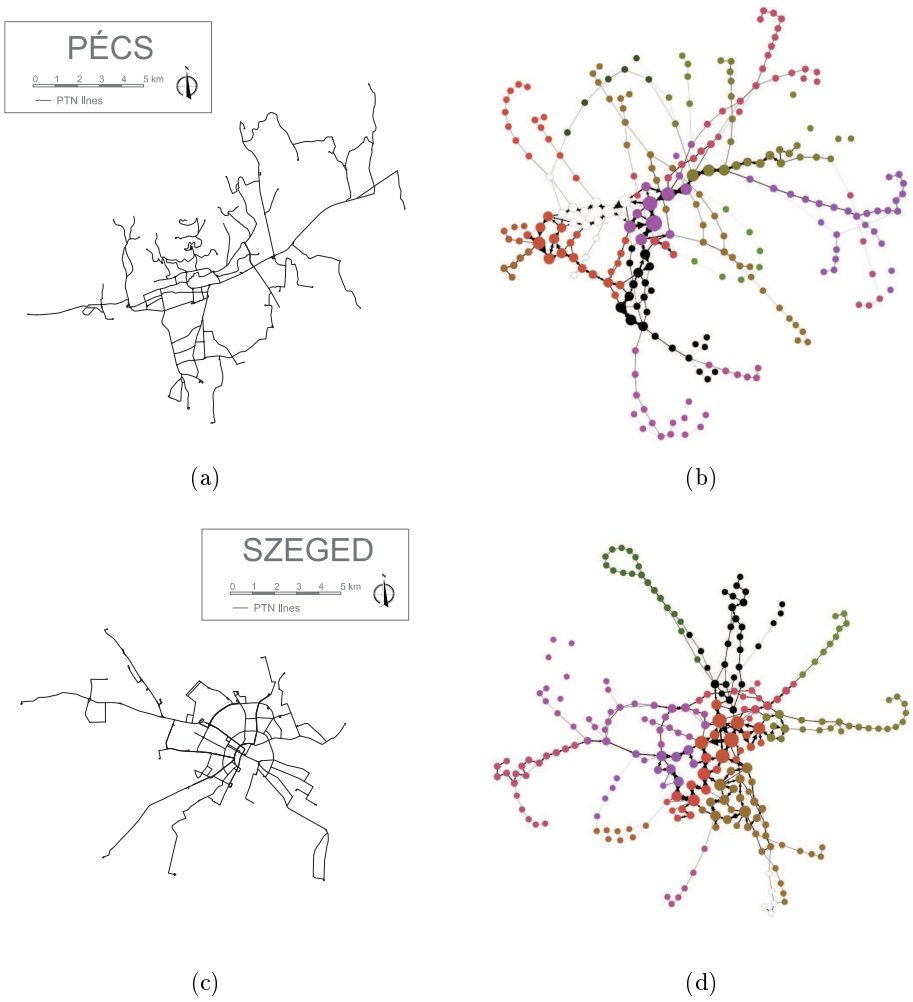


Figure 2.4. Simple maps of the lines of the transportation system of the cities. The partition of the PTNs into communities using the modularity optimization method. Nodes having the same color belong to the same cluster; the bigger a node, the higher its in-degree is; and the thicker an edge, the greater its capacity is.

but interesting observations can be made by comparing the the unweighted and weighted closeness values for one city. The centrality values in the unweighted networks tell us how central and important the nodes are according to the structure of the network. By



(a) (b)

(c) (d)

Figure 2.5. Simple maps of the lines of the transportation system of Pécs and Szeged.

considering the schedules and capacities of lines in the PTN we need to assign weights to the links, the nodes get closer to or farther from each other from a transportation point of view. The unweighted and weighted C values for each city can be seen (plotted on the same scale) in figures 2.6(a)–2.6(e). In the case where the centrality value in the unweighted network of a node is bigger than the value in the weighted case tells us that although the node has central position in the network, the stop that represented by this node may not be well exploited in the transportation sense. However, if the relation between the unweighted and weighted case is the opposite, the stop is overloaded according to the network traffic.

The betweenness centralities for the unweighted and weighted case can be seen in figures 2.3(b) and 2.3(f) and display similar shapes as in the case of closeness. The unweighted and weighted BC values for each city can be seen (plotted on the same scale) in figures 2.6(f)–2.6(j). Similar to closeness, if the BC value of a node in the weighted network is greater than its value in the unweighted case, the given stop may be overloaded in the PTN. The opposite relation refers to a stop with spare capacity.

In [119], PageRank was used to identify the key nodes in a transportation system and also for traffic simulations [143] to find important nodes that have a high impact on

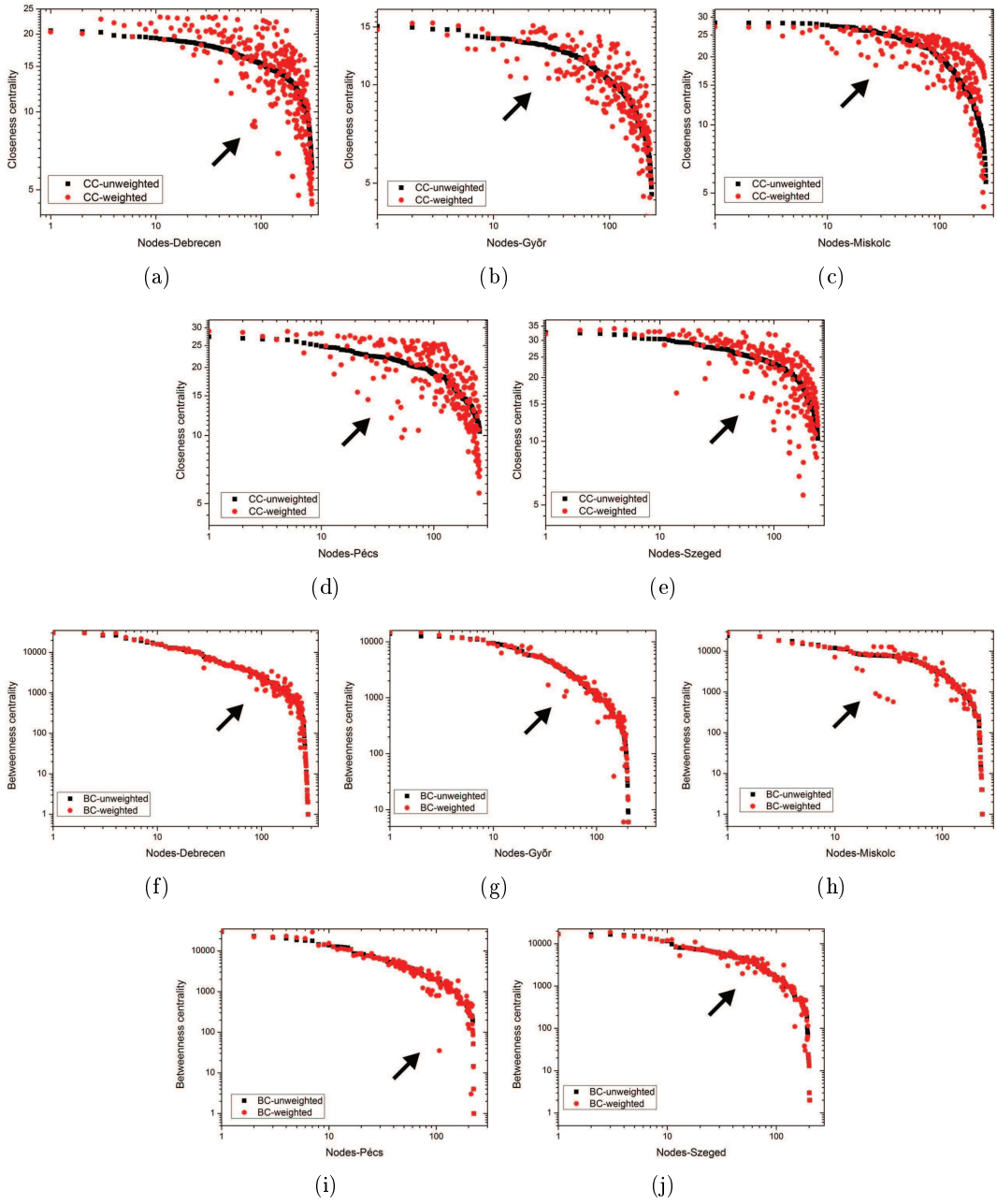
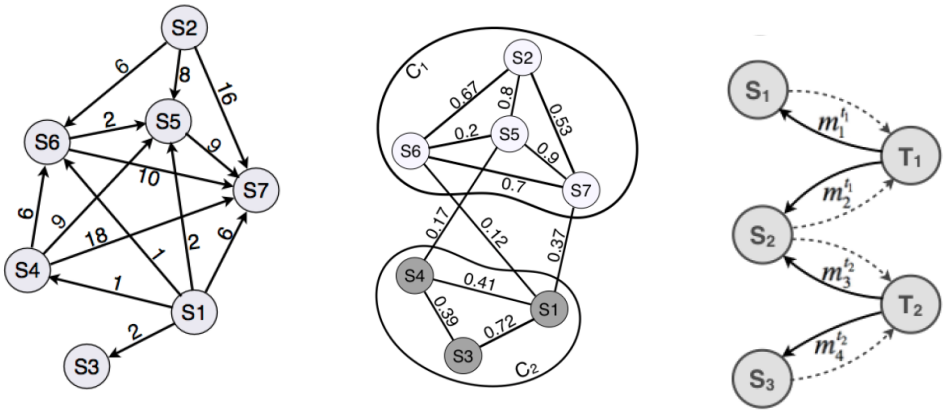


Figure 2.6. The unweighted and weighted betweenness and closeness centrality measures for each city. The values are in decreasing order of the centrality values for the unweighted networks.

transportation efficiency. It is interesting to observe that the PageRank distributions are similar for all the five weighted PTNs (Fig. 2.3(h)), which is probably due to the fact that organizational rules of the schedules are similar.

2.3 Educational Data Mining Aspects

Educational Data Mining [160] is concerned with the development, research and application of computerized methods to find patterns and features in large collections of educational data. Such features are hard to analyze due to the huge amount of information available and the high-level complexity of such databases. Data of interest is not restricted



(a) (b) (c)

Figure 2.7. Toy examples for the network models. (a): a directed weighted graph of the students. (b): a similarity-based weighted graph of the students with communities (c): a bipartite graph of students and teachers

to interactions of individuals in an educational system (e.g., navigation behavior, input to quizzes and interactive exercises), but they might also include data from collaborating students (e.g. text chat), administrative data (like school, school district, teacher), and demographic data (like gender, age, school grades). Some discussions on educational data mining can be found in [88, 159, 160, 166]. Databases of educational institutes, where the data is produced by complex administration systems, contain the administration of the daily work of teachers and students, like descriptions of the lessons including the equipment and educational methods that were used, the areas of competence that have been developed, the students who participated and their marks and level, among other things. Since a large amount of detailed data has become available via administration activities and there is an opportunity to obtain more information about the participants of the educational system than e.g. using classical questionnaire methods. Such relevant issues, which have long been of interest, like measuring the progress and achievements of the students, the efficiency of the teacher's work, level of difficulty, data visualization and the detection of incidental problems of the students (like drug or alcohol abuse, crisis in the family) may be investigated and addressed using different kinds of data mining techniques.

Here, we discuss the possible application of the ubiquitous complex network approach for information extraction from educational data. We define several suitable network representations of data available in such administration systems and present some possible ways of how graph mining techniques can be used to get detailed information about them.

2.3.1 Graph-based Concepts on the Educational Sphere

Directed Graphs based on the Marks of the Students

The first network model of the students is a generalization of the one defined in [124]. In this model, each node represents a student and a link between two students is defined in the following way. We will assume that two students can be compared directly if they

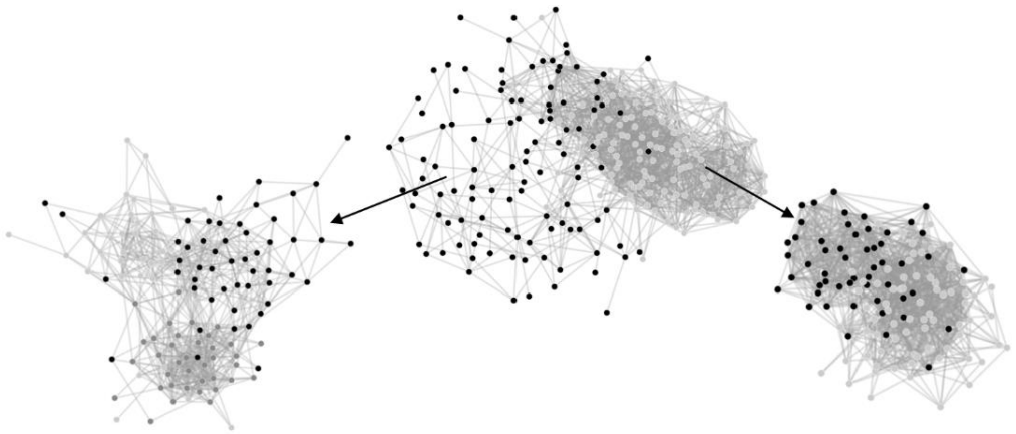


Figure 2.8. Community structure of a network of students (middle). The two subgraphs (left and right) induced by the two communities were re-clustered.

received an end-of-year mark in at least one common subject. If the end-of-year marks of the students i and j are $(m_1^i, \ldots, m_t^i)$ and $(m_1^j, \ldots, m_t^j)$, respectively, then we can calculate the weight

$$w_{ij} = \sum_{k=1}^{t} c_k (m_k^i - m_k^j), \tag{2.10}$$

and add a directed link with weight $|w_{ij}|$ between nodes i and j. The link goes from j to i, if $w_{ij} > 0$, and it goes in the opposite direction if $w_{ij} < 0$. The constant term c_i refers to the level of difficulty of a subject, which can also be measured by a network-based approach (see below) or by applying some statistical methods. As a short concrete example, suppose Anne and Bob received the end-year marks $(4, 5, 5, 5, 5)$ and $(5, 3, 3, 3, 4)$ for Mathematics, Literature, History, English and Art, respectively. Then $w_{AB} = 6$ with $c_k = 1, \forall k$, means that Anne is 6 points better than Bob, if all the subjects have the same difficulty. Fig 2.7(a) shows a toy example for the model. One possible way of determining the subject difficulty values is to use the average of the end-of-year marks of each subject and assume that the higher the average, the less difficult the subject is. By using the cumulative distribution of the marks, one can define an alternative way for calculating the c_k values by comparing these distributions. It is also possible to find out how difficult it is to a get a certain mark from a teacher and incorporate this parameter into the formula that used to compute the edge weights.

Undirected Graphs based on Similarities of the Marks of the Students

The second network model is a family of undirected and weighted networks. As before, the nodes represent students, while a weighted edge between two students is defined by a *similarity measure* S of the lists containing the end-of-year marks of their common subjects (that were not necessarily taught by the same teachers). For example, the *Jaccard similarity* measure [92] is defined as the fraction of the marks that are the same as all the marks in common for two students (a toy example can be seen in Fig. 2.7(b)). One may use several similarity functions to define the weight of similarity of two students.

Figure 2.8 shows the community structure of the network of 255 students in their tenth year in a Hungarian secondary school. The weights were defined by the Jaccard similarity measure. We observed in our preliminary studies that the network contained two main communities. The community of students who performed well in the school (Fig. 2.8, middle, grey community) and community of students with a weaker academic performance in school (Fig 2.8, middle, black community), respectively. We also found that the network had a more refined structure by re-clustering the two main communities, and we identified clusters of students who were better in the natural sciences and students who were better in the arts, respectively. We should also mention that while these studies were not too detailed, such investigations might be the subject of a future study.

Bipartite Graphs of Students and Teachers

In order to evaluate how difficult it is to get a good mark from a certain teacher, we propose a family of *bipartite graphs* (see Sec. 5.1 for more details on bipartite graphs) as network models based on the earlier results of [49] and [122]. We consider a bipartite graph, $G = (A, B, E)$. In the model, the elements of A are students from the same school, while the set B stands for their teachers. We can define a directed edge from a node $b \in B$ to a node $a \in A$ with weight m_a^b, if the teacher who is represented by node b gave an end-of-year mark m_a^b to student who is represented by node a. However, we also define a directed edge from a to b, based on the assumption that it is more difficult to get a good mark from this teacher if the mark he or she gave is lower than the average of the student's marks (a toy example can be seen in Fig. 2.7(c)). Next, we can easily construct a weighted directed graph of the teachers using the same technique as that described in [122]. With this projection, a network of the teachers can be constructed where a directed and weighted link from a teacher b_i to another teacher b_j shows how much more "consistent" a teacher is than the other. The consistency is measured via the average difference of the marks that the teacher gave to each of his or her students and the average of the students' marks. Once this network is given, we can apply the PageRank method, say, on it in order to assign scores to the teachers. These scores may provide a realistic evaluation of the consistency of their marking habits; moreover, these scores can be used to compare students by normalizing their marks using this evaluation of the teachers.

Bipartite Graphs of Students and Subjects

Similar to the evaluation of the teachers, we can also evaluate how difficult it is to a get a good mark in a certain subject. For this purpose, we consider a bipartite graph of students and subjects, i.e. we simply replace the set of teachers (defined above) by the set of subjects. A directed and weighted link from a subject (say Maths) to a node $a \in A$ (which represents the student a) is defined with the weight m_a^b if the student a got the end-of-year mark m_a^b. Then, from the student a weighted link to the subject Maths is defined, where the weight represents, for instance, the difference between mark m_a^b and the average of the student's marks. A network of the subjects can be defined and by using

some evaluation technique (like PageRank), a ranking of the subjects according to their level of difficulty can be obtained. These scores can be used as weights for the calculation of the students' performance and also for the evaluation of the teachers.

2.3.2 Student Evaluation based on Networks

With the intention of evaluating the achievements of students and generating a ranking between them, we defined a modified PageRank algorithm as a data mining technique. The Simple Network Workflow for Schools system (SNW) which we investigated is a complex administration software package of more than sixty institutes of public education (including elementary schools, technical colleges, secondary schools and educational institutes of arts) with electronic diaries, quality management, measurement, and evaluation systems.

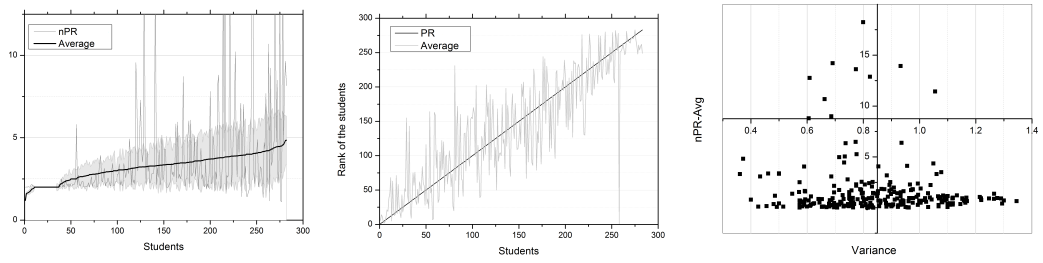
Data set, Mathematical model and Experimental results

As a first step, we collected data from the database of the SNW-system. We used a dataset of 283 students in the same (secondary) school in the same year and examined all their end-year school reports. We defined a weighted and directed network of these students, as described in the first model in Sec. 2.3.1. We used $c_k \equiv 1$ in Eq. 2.10 and normalized the weights as $p_{ij} = w_{ij} / \sum_{j:i\to j} w_{ij}$ in order to get a row stochastic adjacency matrix with the transition probability values and applied the PageRank method to this network.

First, we checked the sensitivity of our method for different values of the damping factor. We observed that the method is robust against the choice of λ, which was confirmed by the high correlation of the resulting PageRank vectors that contain the PR scores of the students (the Pearson correlation was over 0.9 for each pair of result vectors). In further studies, we chose $\lambda = 0.2$, which is usual in PageRank computations.

We used *Kendall's τ correlation* method [101] to quantify the rank correlation between the rankings obtained by the PageRank and the average method, which is simply gives a ranking of students by comparing the average of their end year marks. Although the correlation coefficient was 0.68, which displays a positive correlation, many differences can be seen between the two methods (see e.g. Fig. 2.9(a)). We normalized the PR values, such that the values obtained from using the two methods are on the same scale. We observed in general that if the end-year average of a student is high, but the PR value of hers is relatively small, then the student has only a few subjects, where it is "easy" to get a five mark. This assertion was justified by checking all the marks of those subjects. However, if a student i has a low average, but her PR value is high compared to the others, it is normally true that most of the students taught by the same teachers also have low marks such as i, but a high PR value of i means that she exceeded the performance of their schoolmates.

We checked the relationship between the PR values (and ranks resulting from this; the relation between the PR ranks and average ranks is shown in Fig. 2.9(b)) along with the variance of the end-year marks. It can be seen in Fig. 2.9(a) that outstanding PR values occur when the variance is high. Generally speaking, it was noticed that if the variance of the marks of a student was high and their PR value was also high, then the



(a) (b) (c)

Figure 2.9. (a) Relation between normalized PR scores and the average of the end-year marks; the light grey area shows the variance domain of the marks. (b) Ranks of the students obtained by the different methods, ordered by the PR scores. (c) Relation between the variance of the end-year marks and difference of the normalized PR scores and average values

student is talented in at least one subject. A high variance of the marks is due to the large variety in the marks, but a large PR value must be caused by just one (or only a few) subject (of the same teacher), where the student was significantly better than her classmates.

It is also interesting to examine the relationship between the difference of the normalized PR scores and average values, and the variance of the end-year marks (see Fig. 2.9(c)). It can be seen that the difference between the PR score and the average of the marks is small in general regardless of whether the variance is high or low. It suggests that we should be pay more attention to those students where this difference is high. In such cases, we should also discover, what causes this big difference.

Applying this network-based method, the talents, the problematic students, the strict or overly lenient teachers can be filtered out. Examining students in the same class, uniformly high or uniformly low PR scores (for instance) can also provide a fair picture of the difficulty of each subject and/or the personality of the teacher of a certain subject as well as the achievements of a class of students from a global point of view. The PageRank method is also very effective in finding the best students in the same year. After filtering out the "outliers" (e.g. students who have just a very few marks because, for example, they moved to another school), PR scores provide a fairly good relative order of the students with respect to their achievements. Such rankings can also be useful in deciding which students need to be rewarded at the end of the year, and it is also useful for the teachers and parents to follow the educational progress of the students and children, respectively.

2.4 Summary

In this chapter, we considered three real-world systems and presented some possible complex network models for them. In each case, after data collection, a network of the "actors" of the system was defined and analyzed using standard and, in some case, new graph mining techniques.

Firstly, a new local PageRank approximation algorithm was applied to a co-citation network on the citation environment of the seminal paper by Jenő Egerváry. It follows

from the implementation of the PageRank algorithm that citations received from more important papers contribute more to the ranking of the cited paper than those coming from less important ones. Furthermore, simplicity and fast computability are advantages of our method. However, co-citation networks provide more detailed contextual information (compared to the number of citations) for evaluating the impact of an article. We hope that network-based ranking methods will gain more space in Scientometrics since they offer a more objective picture of the impact of scientific publications. Now it seems that one of the most challenging tasks in citation network based scientometics is to generate the local co-citation network of any article automatically, but hopefully these data sets will be provided by the owners of such databases in the future.

Secondly, a comprehensive network analysis was performed on the public transportation network of five Hungarian cities. Although previous studies often used unweighted networks, one novelty of our study was to consider directed and weighted edges, where the weight of a link referred to the morning peak hour capacity of that link obtained by using the capacities of the vehicles (bus, tram, trolleybus) and schedules of the lines that go though that link. We should add that the modal split (that is, the percentage of travelers using a particular type of transportation) and the real number of passengers in the PT vehicles are the key descriptors of public transport systems from an optimization point of view. However, we presented an alternative approach which requires a smaller amount of data, but gives a "first glance" global picture of the PTNs. In the future, we would like to analyze bigger cities and also cities in different countries with similar layouts (medium-sized, similar urban structure and land use) with network theoretic tools using more detailed data (where in addition to the schedules and capacities, the geographical distances are also given between the consecutive stations). We would also like to address the question of transfers between routes. The results of this study accord well with the earlier studies in the area of classical PTN modeling. We think that the kinds of methods applied here could assist experts in the planning of urban public transportation systems and they could be integrated into the classical PT organization methodology.

Thirdly, we proposed four different suitable network representations of students, teachers and subjects in public education and presented some possible ways of how graph mining techniques could be used to provide more detailed information about them. Analyzing these networks using real data sets might be a fruitful direction in the future. Then, we defined a PageRank-based graph algorithm and applied it to a network of students in a secondary school. By applying our method, the achievements and ranking of the students are not only analyzed and determined by simple statistical techniques, but the use of pairwise comparisons of the students to obtain a complex network representation of this system was also considered. We observed that our method gave a better picture of the students' relative performance, and it can also identify outstanding and relatively weak students. In our experiments, the PageRank method gave an especially good picture of the students in the case where we want to investigate whether the student is outstandingly better than her schoolmates.

Chapter 3

Network Models applied in Economics

Having presented complex network models of several real systems in the previous chapter, we will now focus on networks that arise in the field of economics. Providing only a brief introduction and survey on recent findings of the topic is out of the scope of this thesis, instead we refer to [93] and [94] as a good textbook and a very recent survey of the topic, respectively.

Firstly, we show how graphs can be used to model trade networks of countries. We give a brief overview of possible network representations of the international trade and present some approaches to extract information from the system using network analysis. We present a brief case study that investigates the timely evolution of the trade network of the European Union, focusing especially on the evolution of communities and different rankings of the countries and paying special attention to the former member countries of the Council for Mutual Economic Assistance (Comecon).

Secondly, we discuss the concept of correlation-based financial networks. We show how different "noise" filtering techniques can be applied on the correlation matrix (i.e. correlation network) containing the pairwise correlations of stock time series. Then, we examine the performance of the fundamental Markowitz portfolio optimization model using stock time series data of various stock exchanges and investment period intervals. The performance of the methods is compared using the estimated and realized returns and risks, respectively. The results indicate, in accordance with previous studies, that the estimated risk, in general, is closer to the realized risk using filtering methods. We also draw some conclusions according to the the expected return estimation, namely, our results tells us that the use of the James-Stein "shrinkage" estimator the reliability of the portfolio can be improved.

3.1 Trade Networks

Investigating trade systems and the world-trade is an important area of modern economics. One of the most important indicators is the world-trade ranking of the countries, which not only indicates the wealth of the countries, but implicitly contains information

about the efficiency of their economic relations with other countries. However, the ranking is usually done according to their export/import volumes in US dollars. In this approach the most developed and rich countries are at the top of the ranking, but not necessary due to the fact that their trade network is efficient, broad and competitive [105]. The usual statistical indicators give a relatively objective picture of the countries' economy, but they do not give much information either about the international trade as a continuously evolving economic system or the relations each country has with other countries.

Complex networks analysis provides a detailed picture of complex trade systems and their evolutionary dynamics. Trade networks can be studied in a simplified, but complexity preserving graph model, where the countries are represented by the nodes of the graph, while edges represent the trading relation between any two countries, often using export and import volumes in US dollars as weights. Thus, the model of the system is a directed and weighted graph, where the direction and weight of an edge refer to the direction and volume of the cash flow, respectively.

It should be mentioned that from another aspect, international trade can be modeled by a bipartite graph, where nodes represent countries and products, and a weighted edge between a country and a product represents the ratio to the total amount of the product imported or exported. The world trade web, if defined in this way is highly *nested* [60], which informally means that small-degree nodes tend to be connected only with high-degree nodes, resulting in a core-periphery like bipartite network. Such nestedness remained constant between 1985 and 2009 [29], and most probably greatly contributes to the stability of the world trade. (The stability of the international trade network was recently confirmed by applying a different method [170].) Interestingly, distances of countries did not play a crucial role in shaping the world trade network [152].

3.1.1 Structure and Evolution of Trade Networks

The earliest studies of trade networks dealt with undirected and static model graphs (see, e.g. [168]), but recent results appeared based on the analysis of evolving directed networks (see e.g. [5, 60, 63, 74, 151]). Analyzing the structure and the the temporal dynamics of these networks has recently been used to confirm the globalization of the world economy [5, 91]. Although in [91] the authors claim that there is strong evidence for the globalization of the world-trade, many studies highlight the co-existence of processes opposite to globalization, sometimes referred as "regionalization" [5, 151]. Recently, by investigating the International Trade Network, it has been found that the global changes of the trade network are closely related to changes at the regional level [192].

3.1.2 Studies on Trade Network of the EU

Now we will we briefly present, through an example, how the network approach can be used to investigate trade networks; in particular, we will analyze the timely evolution of the trade network of the European Union (for more details, see the author's paper [140].) We used the import/export volumes data available in the United Nations Comtrade Database [177]. The EU 28 countries together with non-EU countries, like the

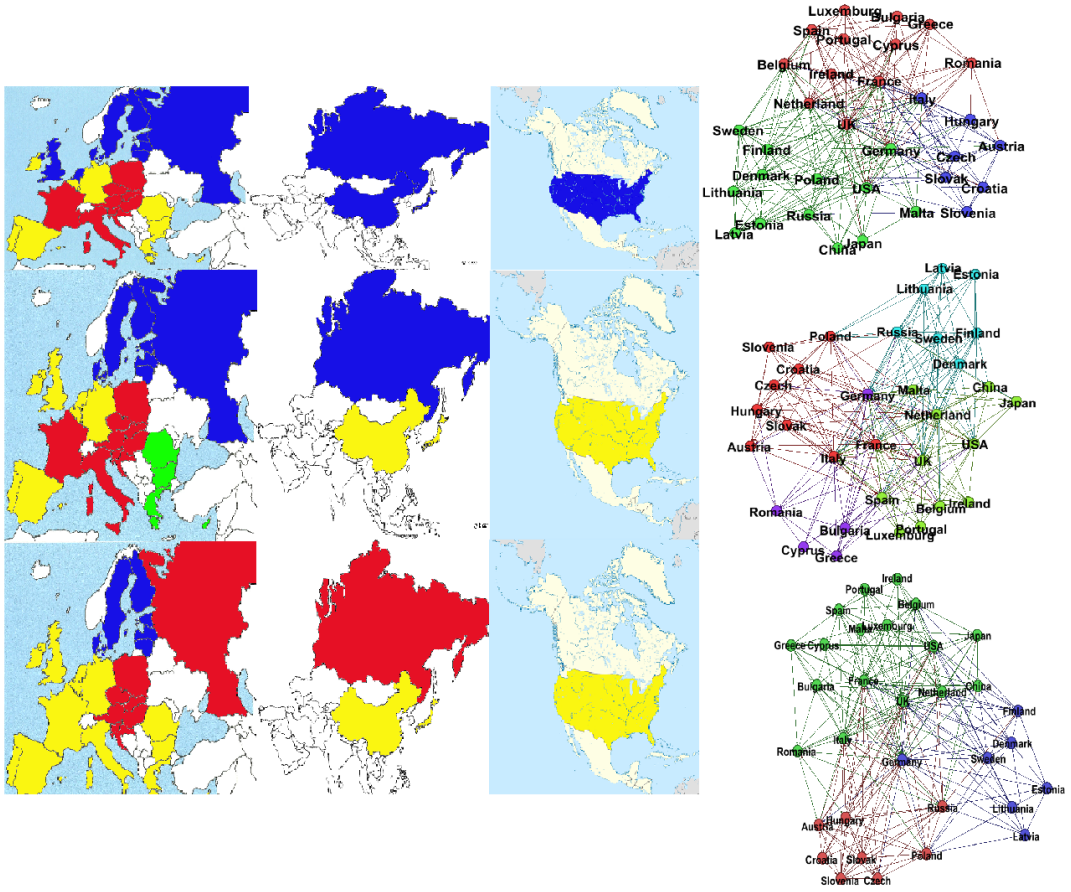


Figure 3.1. Communities of the trade network in 2004, 2007 and 2013 (top to bottom).

USA, Russia, China and Japan were included in the study in the period 1995-2013. We especially focused on the membership expansion years. That is

1. 1995: the EU had fifteen members, namely Austria, Belgium, Finland, Netherlands, Luxemburg, Germany, France, Italy, Denmark, Ireland, the United Kingdom, Greece, Spain, Sweden and Portugal
2. 2000: no enlargement
3. 2004: ten new members joined, namely: Cyprus, the Czech Republic, Estonia, Poland, Latvia, Hungary, Malta, Slovakia and Slovenia
4. 2007: Bulgaria and Romania joined the EU
5. 2013: Croatia joined the EU

Communities in the trade network

For each country we considered the import/export trade volumes relative to the country's GDP. It can be seen that each country's export volume (relative to the GDP) increased after joining the EU. The fastest rate of growth was produced by the Central European countries (the Czech Republic, Hungary, Poland and Slovakia), but this growth was

caused by the increasing trade between each other and the also with Russia and China. During the period examined (1995–2013), the trade network was mainly characterized by four communities, namely West Europe, East-Central Europe, North Europe and South Europe (or Balkan) communities, respectively (Fig. 3.1). In these communities most of the countries were stable members, but there were countries that belonged to different communities in different periods and some of the communities merged for a short period before splitting up again. The most stable community consisted of the Scandinavian and Baltic countries. The West-Europe community displayed a higher fluctuation, but the core countries, namely Benelux countries, France, Ireland, Spain and Portugal, were stable. In the 2004 enlargement, Germany and Italy became the core members of the East-Central European community by the increasing trade with the new countries that had just joined to the EU.

It can also be seen that the former Comecon countries, and Cyprus and Greece did not became stable members of any community perhaps due to their historical legacies. While the Central-East European community contained the countries of the Habsburg-, later Austrian-Hungarian, Monarchy, the Balkan community consisted of the countries of the Ottoman Empire until the end of the 19th century [56]. Interestingly, in the latter case Greece already joined the EU in 1981, it actually was in the community contained Bulgaria, Cyprus and Romania. Temporary merging with and later separation from the western clusters points out the integration challenges of these regions and confirms importance of regional effects – related to the historical and geographical conditions – which can still not be altered by the concept of a customs union and aspirations for EU integration.

In another aspect, we may define the trade network of the same countries using the total import/export volumes, instead of the volumes relative to GDP. In these networks a central core appeared for each given year (see Fig. 3.2). The core-periphery classification was calculated using the weighted version of the Borgatti-Everett algorithm. In each examined year the network core was formed by the leading West European countries (France, Germany, Italy, Netherlands and UK) and the USA. Investigating the role of China, Japan and Russia, we observe that in 1995 Japan was in the core but later it dropped out. In contrast, China became a core member over time, and this is consistent with the results in [192] which says that China took over the leading position of the Far East region from Japan. The role of Russia is similar to that of China's, it moved to the network core by 2013, which was probably caused by the growing trading relations with the former Comecon countries that joined the EU.

Ranking countries

Usually the international trade ranking of countries is performed according to their export and/or import volumes counted in US Dollar. Using this approach the rich and well-developed countries lie at the top of the listing, not necessary due to the fact that their trade network is efficient, broad and competitive [61]. Since the PageRank and HITS algorithms often work well in ranking nodes according to their network position, we used

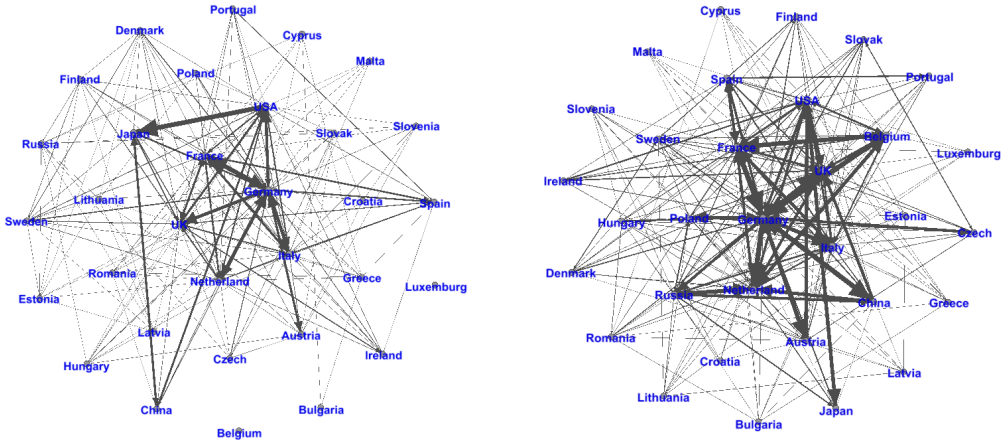


Figure 3.2. Total export network of the examined countries in 1995 and 2013. The thicker the link, the higher the export volume between the countries

them to get a trade ranking of EU countries in the period examined. The results obtained by the two algorithms are in general agreement with each other. The PageRank values of the countries display a Pareto-like distribution (i.e. a power-law), which means that the total export of the EU is transacted by a small proportion of the member countries (France, Germany, Italy, Netherlands, UK). By comparing the PageRank scores with the GDP, we can get a more detailed picture: countries with a lower export/GDP rank than PageRank rank are more important in the export network than would simply be expected based on their export volumes (like the Czech Republic and Hungary, Fig. 3.3, left). Applying the HITS algorithm to the networks, the authority scores of the countries reveal how big importing countries the export goes into, while hub scores show how big exporting countries the import comes from. By comparing the in-degree (i.e. total import relative to GDP) with the hub score, and the out-degree (i.e. total export relative to GDP) with the authority score for each country we can draw the following simple conclusions. Countries with higher rank according the hub scores than in-degree rank are the leading economies of the EU (France, Germany, UK) and the Baltic and Central-East Europe countries (including Hungary, see Fig. 3.3, middle). In the case of the leading economies this is caused by the active and significant trading between each other (in the network core) since these countries have a high hub and authority score at the same time. The smaller countries with higher hub scores than export relative to GDP, trade with the big importing countries and this may result in significant advantages during periods of economic growth. Smaller out-degree rank than hub score rank implies that the exported products utilized in countries with smaller authority scores (Fig. 3.3 right). For bigger economies the large diversity of the trading partners is naturally better; however for smaller economies the combination of small export volumes (relative to GDP) with low-prestige trading partners may be risky especially in periods of economic crisis.

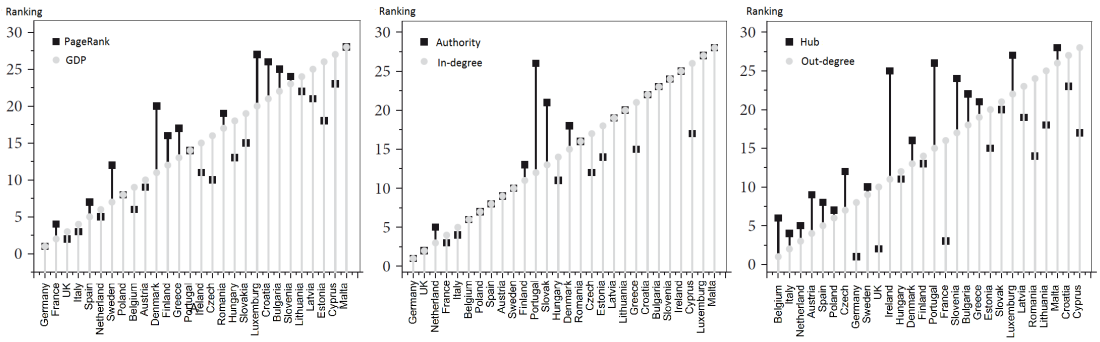


Figure 3.3. Comparing different rankings of the countries in 2013.

3.2 Networks based on Stock Correlations

After gaining certain insights from our analysis of trade networks, now we will introduce the key concepts of using networks in finance. In a financial market the performance of a company is judged by the company's stock price, while the value of a company is determined by the stock price multiplied by the number of shares outstanding (that is, the company's stock currently held by all its shareholders). Though the exact nature of the interactions among companies is not known in general, it is natural to think that these interactions are reflected in the equal-time correlations of their stock prices. These correlations play a central role in investment theory and risk management, including the classic *Markowitz portfolio theory*.

The interactions of companies, measured by stock price correlations, can be viewed as an evolving complex system of stocks (as units of the system), and hence applying network theory, which provides an approach to investigate complex systems, may be useful here. Mantenga was the first who defined networks based on correlations [132] and many articles have appeared on the topic since then (see, e.g. [174] for a good survey and for more references).

Now let us consider the price time series of n given assets and let us denote the closure price of asset i at time t (here it is a day) $(t = 1, \ldots, T)$ by $P_i(t)$. The daily logarithmic return[1] of i is defined as

$$r_i(t) = \log \frac{P_i(t)}{P_i(t-1)} = \log P_i(t) - \log P_i(t-1). \tag{3.1}$$

The correlation coefficient between stock i and j is defined as

$$C_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_i \sigma_j}}, \tag{3.2}$$

where σ_{ij} is the covariance between stock i and j and σ_i is the standard deviation of

[1] This is common mainly because of the following reasons. (i) if we assume that prices are log-normally distributed (which, may or may not be true for a given price series), then $r_i(t)$ is normally distributed; (ii) when returns are very small (common for trades with short holding durations), the log-returns are close in value to raw returns.

stock i, calculated as

$$\sigma_{ij} = \overline{r_i r_j} - \overline{r_i}\,\overline{r_j} \text{ and } \sigma_i = \sigma_{ii} = \overline{r_i^2} - \overline{r}_i^2. \tag{3.3}$$

Above, the bar denotes the temporal average. That is,

$$\overline{r_i} = \frac{1}{T}\sum_{t=0}^{T} r_i(t), \quad \overline{r_i^2} = \frac{1}{T}\sum_{t=0}^{T} r_i^2(t) \text{ and } \quad \overline{r_i r_j} = \frac{1}{T}\sum_{t=0}^{T} r_i(t) r_j(t). \tag{3.4}$$

We should note that, however, C_{ij}, σ_{ij} and σ_i are, in theory, calculated using the (joint) probability distributions of $\{r_i(t)\}_{t=0,\ldots,T}$ and $\{r_j(t)\}_{t=0,\ldots,T}$, which were defined as sample quantities (i.e. they are estimated using the realized values of the given time series). Lastly, the correlation matrix is denoted by $\mathbf{C} = (C_{ij})_{i,j=1,\ldots,n}$ and the covariance matrix is denoted by $\mathbf{\Sigma} = (\sigma_{ij})_{i,j=1,\ldots,n}$.

3.2.1 Correlation Networks and Statistical Uncertainty

Recently, the analysis of the correlation coefficient matrix of stock time series has become the focus of interest [39, 55, 108, 109, 161, 172]. Many attempts have been made in order to quantify the degree of statistical uncertainty present in the correlation matrix and filter information that is robust against this uncertainty [39, 84, 108, 109, 131]. The filtered correlation matrices have been successfully used in portfolio optimization in terms of risk reduction [109, 161, 172]. Below we describe two approaches used for the correlation matrix filtering, namely the *random matrix theory* approach and the *hierarchical clustering* approach.

Random Matrix Theory

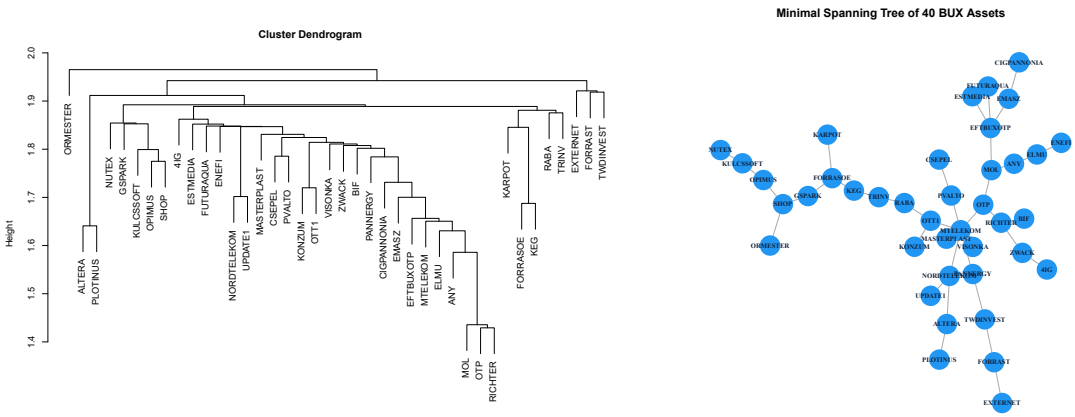
A simple random matrix is a matrix whose elements are random numbers from a given distribution [137]. In the context of stock portfolios, random matrix theory (RMT) can be useful to investigate the effect of statistical uncertainty in the estimation of the correlation matrix [172]. Given the time series of length T of the returns of n assets and assuming that the returns are independent Gaussian random variables with zero mean and unit variance, in the limit $n \to \infty$, $T \to \infty$ such that $Q = T/n$ is fixed, the distribution $\mathcal{P}_{rm}(\lambda)$ of the eigenvalues of the random correlation matrix ($\mathbf{C}_{rm}$) is given by

$$\mathcal{P}_{rm}(\lambda) = \frac{Q}{2\pi\sigma^2}\frac{\sqrt{(\lambda_{\max} - \lambda)(\lambda - \lambda_{\min})}}{\lambda}, \tag{3.5}$$

where $\lambda_{\min}$ and $\lambda_{\max}$ are the minimum and maximum eigenvalues, respectively [167], and they have the form

$$\lambda_{\max,\min} = \sigma^2(1 + \frac{1}{Q} \pm 2\sqrt{\frac{1}{Q}}). \tag{3.6}$$

Previous studies have pointed out that the largest eigenvalue of correlation matrices from returns of financial assets is completely inconsistent with Eq. 3.5 and it refers to the common behavior of the stocks in the portfolio [108, 153]. Since Eq. 3.5 is strictly valid



(a) (b)

Figure 3.4. Indexed hierarchical tree - obtained by the single linkage procedure - and the associated MST of the correlation matrix of 40 assets of the Budapest Stock Exchange.

only for $n \to \infty$, $T \to \infty$, one can construct random matrices for the given n and T values of the data sets used and compare the largest eigenvalues and the spectrum $\mathbf{C}$ and $\mathbf{C}_{rm}$ (i.e. compare the spectrum of the matrix constructed from real data and a random matrix of the same size). Since $\mathrm{Trace}(\mathbf{C}) = n$, the variance of the part not explained by the largest eigenvalue can be quantified as $\sigma^2 = 1 - \lambda_{\mathrm{largest}}/n$. Using this, we can recalculate λ_{min} and λ_{max} in Eq. 3.6 and construct a filtered diagonal matrix, using the singular value decomposition, got by setting to zero all eigenvalues of $\mathbf{C}$ smaller than λ_{max} and transforming it to the basis of $\mathbf{C}$ by setting the diagonal elements to one.

Hierarchical clustering

The correlation matrix $\mathbf{C}$ has $n(n-1)/2 \sim n^2$ elements, hence it contains a huge amount of information even for a small number of assets considered in the portfolio selection problem. As shown by Mantegna and others [132], that the *single linkage* hierarchical clustering algorithm (closely related to minimal spanning trees (MST) of graphs) provides economically meaningful information using just $n-1$ elements of the correlation matrix. To construct the MST, the correlation matrix $\mathbf{C}$ is converted into a distance matrix $\mathbf{D}$, e.g. following [132, 133], using the $d_{ij} = \sqrt{2(1-\rho_{ij})}$ ultrametric distance. Ultrametric distances are the class of distances that satisfy the inequality $d_{ij} \leq \max\{d_{ik}, d_{kj}\}$, which is a stronger assumption that the standard triangular inequality. The distance matrix $\mathbf{D}$ may be viewed as representing a fully connected graph of the assets with edge weights d_{ij} representing a similarity between their time series. For this graph (matrix) we can use, for example, the *Kruskal algorithm* in order to obtain the MST of $n-1$ elements and then construct the filtered correlation matrix $\mathbf{C}_{um}$ using the $n-1$ correlation coefficients derived from the $n-1$ distances in the MST.

Another widespread hierarchical clustering procedure is the *average linkage* algorithm. While the single linkage clustering procedure basically follows the greedy Kruskal MST method, the average linkage algorithm, for each iteration step, defines the distance between an element and a cluster as the average distance between the element and each

element in the cluster. For a detailed description, see e.g. [174].

3.2.2 Application to Portfolio Optimization

Portfolio optimization is one of the fundamental problems in asset management that seeks to reduce the risk of an investment by diversifying it into assets expected to fluctuate independently [57]. In his seminal work [134], Markowitz formulated the problem as a quadratic programming task. Namely, given the expected return of the portfolio, the risk, a quadratic function that is measured via the covariances of the asset time series, has to be minimized. Formally, given n risky assets, a portfolio composition is determined by the weights p_i $(i = 1, \ldots, n)$, such that $\sum_i^n p_i = 1$, indicating the fraction of wealth invested in asset i. The expected return and the variance of the portfolio $\mathbf{p}$ are

$$r_p = \sum_{i=1}^{n} p_i r_i = \mathbf{p}\mathbf{r}^T \tag{3.7}$$

and

$$\sigma_p^2 = \sum_{i=1}^{n}\sum_{j=1}^{n} p_i p_j \sigma_{ij} = \mathbf{p}\Sigma\mathbf{p}^T, \tag{3.8}$$

respectively, where r_i is the expected return of asset i and $\mathbf{\Sigma}$ is the covariance matrix contains the pairwise covariances of the asset time series in a given time interval. Vectors here are now treated as row vectors.

In the classic Markowitz model [134] risk is measured by the variance providing a quadratic optimization problem that consists in finding vector a $\mathbf{p}$ which minimizes σ_p for a given "minimal expected return" value of r_p. Here, we will assume that short selling is allowed and therefore p_i can be negative. The solution of this problem, found by Markowitz, is

$$\mathbf{p}^* = \lambda\mathbf{\Sigma}^{-1}\mathbf{1}^T + \gamma\mathbf{\Sigma}^{-1}\mathbf{r}^T, \tag{3.9}$$

with $\mathbf{1} = (1, \ldots, 1)$; and the other parameters are

$$\lambda = (C - r_p B)/D \text{ and } \gamma = (r_p A - B)/D,$$

where

$$A = \mathbf{1}\mathbf{\Sigma}^{-1}\mathbf{1}^T, B = \mathbf{1}\mathbf{\Sigma}^{-1}\mathbf{r}^T, C = \mathbf{r}\mathbf{\Sigma}^{-1}\mathbf{r}^T, D = AC - B^2.$$

A possible RMT approach for portfolio optimization is to use $\mathbf{\Sigma}_{rm}$ (which can be readily calculated from $\mathbf{C}_{rm}$) instead of $\mathbf{\Sigma}$ in the Markowitz model. Similarly, we can use $\mathbf{\Sigma}_{sl}$ and $\mathbf{\Sigma}_{al}$, instead of the empirical covariance matrix $\mathbf{\Sigma}$, got by applying the single- and average linkage procedures, respectively.

3.2.3 Results

Estimators of the expected returns

In the case of stationary independent normal returns, the *maximum likelihood estimator* is the sample mean of the past observations of $r_i(t)$ and it was defined as $\overline{r}_i$ in Eq. 3.3. Thus, for the portfolio we can define

$$\overline{\mathbf{r}}_{ML} = (\overline{r}_1, \ldots, \overline{r}_n), \tag{3.10}$$

The maximum likelihood return estimation can be highly inefficient since assets with high past returns are likely to contain more positive estimation errors than others. The positive part trimming could further reduce the risk, and the *James-Stein estimator* [95] provides a constructive shrinkage estimator to do this. The James-Stein estimation for the expected return for asset i is

$$\overline{\mathbf{r}}_{JS} = (1-w)\overline{\mathbf{r}}_{ML} + w\overline{r}_0\mathbf{1}, \tag{3.11}$$

where

$$\overline{r}_0 = \frac{\mathbf{1}\boldsymbol{\Sigma}^{-1}\overline{\mathbf{r}}_{ML}^T}{\mathbf{1}\boldsymbol{\Sigma}^{-1}\mathbf{1}^T}, w = \frac{\lambda}{\lambda + T} \text{ and } \lambda = \frac{(n+2)(T-1)}{(\overline{\mathbf{r}}_{ML} - \overline{r}_0\mathbf{1})\boldsymbol{\Sigma}^{-1}(\overline{\mathbf{r}}_{ML} - \overline{r}_0\mathbf{1})^T}$$

In this calculation, each sample mean is shrunk toward the average return of the minimum variance portfolio $\overline{r}_0$.

For a small sample size, usually below 50, it was observed that there is no evidence that common asset expected returns are different. If all expected returns are assumed to be equal, the *minimum-variance* portfolio is efficient and

$$\overline{\mathbf{r}}_{MV} = \overline{r}_0\mathbf{1}. \tag{3.12}$$

Data

To compare the performance of the methods, we decided to analyze the data set of $n = 40$ stocks traded in the Budapest Stock Exchange (BSE) in the period 1995-2016, using 5145 records of daily returns per stock. The second data set contained the stock time series of $n = 48$ companies of the Information Technology sector (Hardware + Software), which are available on Yahoo Finance (YF) (https://finance.yahoo.com/), in the same period as the BSE data with 5395 records of daily returns of each stock.

We considered $t = t_0$ as the time when the optimization is performed. Since the covariance matrix has $\sim n^2$ elements while the number of records used in the estimation is nT, the length of the time series needs to be $T >> n$ in order to get small errors on the covariance. However, for large T the non-stationarity of the time series appears likely. This problem is known as the *curse of dimensionality*. Because of this, we computed the covariance matrix and expected returns using the $[-T, 0]$ interval, i.e. letting $T = 50 \approx n$, $T = 100 > n$ and $T = 500 >> n$ days preceding the $t = 0$.

Furthermore, applying filtering techniques we tried to filter the part of the covariance matrix which is less affected by statistical uncertainty. To quantify and compare the different methods applied here, we will use the measures described below.

We should also mention here, that treating the Markowitz portfolio selection method as a quadratic programming problem is particularly simple when $\mathbf{\Sigma}$ (in Eq. 3.8) is positive semi-definite and the constraints are equalities (as in Eq. 3.7). It is not difficult to see that the positive semi-definiteness is valid for the original covariance matrix and also for the filtered matrix got by using the RMT method. In [4] it was proved that the filtered correlation matrix obtained by the single linkage clustering procedure is always positive definite if all the elements of the obtained filtered correlation matrix are positive. This is usually the case for correlations of stock time series and it was observed for all the matrices we used. Moreover it was also proved there, that the filtered correlation matrix obtained by using the average linkage clustering method is also positive definite under the same conditions as in the case of the single linkage procedure.

Performance evaluation

To measure the performance of the portfolios determined by the different models, we use the following quantities for the estimated return and risk at the time of investment and the realized risk and returns after the investment period. For portfolio p, the *ex-ante Sharpe ratio* measures the excess return per unit of risk:

$$S_p = \frac{\hat{r}_p - r_f}{\sigma_p}, \tag{3.13}$$

while the *ex-post Sharpe ratio* is defined by a similar equation, but with the realized return r_p. Here, r_f is the risk-free rate of return. The portfolio risk, based on the estimation of the correlation matrix, is calculated as

$$R_p = \frac{|\sigma_r^2 - \hat{\sigma}_p^2|}{\hat{\sigma}_p^2}, \tag{3.14}$$

where $\hat{\sigma}_p^2$ is the predicted risk, and σ_r^2 is the realized risk of the portfolio.

Simulation setup and results

We implemented our simulation environment in R. We are given a data set of stock time series and the input parameters `timeInverval` T , the vector of `startingTimes` $\mathbf{t_0} = (t_0^1, \ldots, t_0^k)$ and the $\mathbf{r_p} = (r_p^1, \ldots, r_p^\ell)$ vector of `expectedReturns` (equal steps between the average return and the maximal return over all asset by default). The simulation procedure is performed via the following steps:

1. For each starting time t_0^j, the `asset.solve.Complete.R()` subroutine checks whether the portfolio optimization can be performed for the given starting time on the interval $[-T, t_0^j]$:

   - if yes, it calculates the optimal portfolio;

- if not, it goes to the next starting time t_0^{j+1};

2. The subroutine stores portfolio weights and the data required for performance evaluation.

The subroutine `asset.solve.Complete.R()` works as follows:

1. It determines the expected returns using maximum likelihood, James-Stein and minimum variance estimations.

2. It determines the covariance matrix of the stock time series.

3. It calculates the filtered covariance matrices using the RMT, the single linkage and average linkage procedures.

4. Portfolio optimization is performed for each return estimation.

   - Using the Lagrange multipliers method of the 'Rsolnp' package [76], it calculates the optimal weights for each covariance matrix
   - It calculates the portfolio risk according to the optimal weights.
   - It determines the realized risk and Sharpe-ratio.

In order to improve the running times, the 'doParallel' R package [31] was used (here we will not go into the details of parallelization).

To check the robustness of the methods, a standard bootstrap experiment was performed. We chose 50 starting times randomly and solved the optimization problem using the time series on the intervals $[-T, t_0^j]$ $(T = 50, 100, 500, j = 1, \ldots, 50)$. For each portfolio, the predicted risk was calculated using Eq. 3.8 for fixed expected returns from the average $\sum_{i=1}^{n} r_i/n$ to the maximum expected return $\max\{r_i : i = 1, \ldots, n\}$ with equal spans. The Lagrange multiplier method, which is available in the 'Rsolnp' R package, was used for the optimization. In each case, the portfolios with realized returns in the top and bottom 10% were dropped. The realized risk using the determined stock weights at time t_0^j, the realized covariance matrix and realized returns were calculated on $[t_0^j, T]$.

Figures 3.5 and 3.6 show the ratio of the realized risk σ_p^2 (continuous line) and the predicted risk $\hat{\sigma}_p^2$ (dashed line) as the function of the expected return r_p obtained by using the different procedures for the BSE data set and Yahoo data set, respectively. For each T, the time of the investment t_0^j $(j = 1, \ldots, 50)$ and the set of stocks were the same.

For the BSE data set, the classic method and the RMT method provide similar realized returns that are always higher using hierarchical clustering (single and average linkage). In spite of this, the risk ratio R_p (i.e. the reliability of the portfolio) is also significantly decreased (see Fig. 3.5, and Tab. 3.1 "Risk Ratio" column), but the deviation of the realized returns increased. The Sharpe ratio of the hierarchical clustering methods was smaller than those got using the other methods, since the estimated risk was often higher than that using the classic and the RMT methods. It can be seen that each method

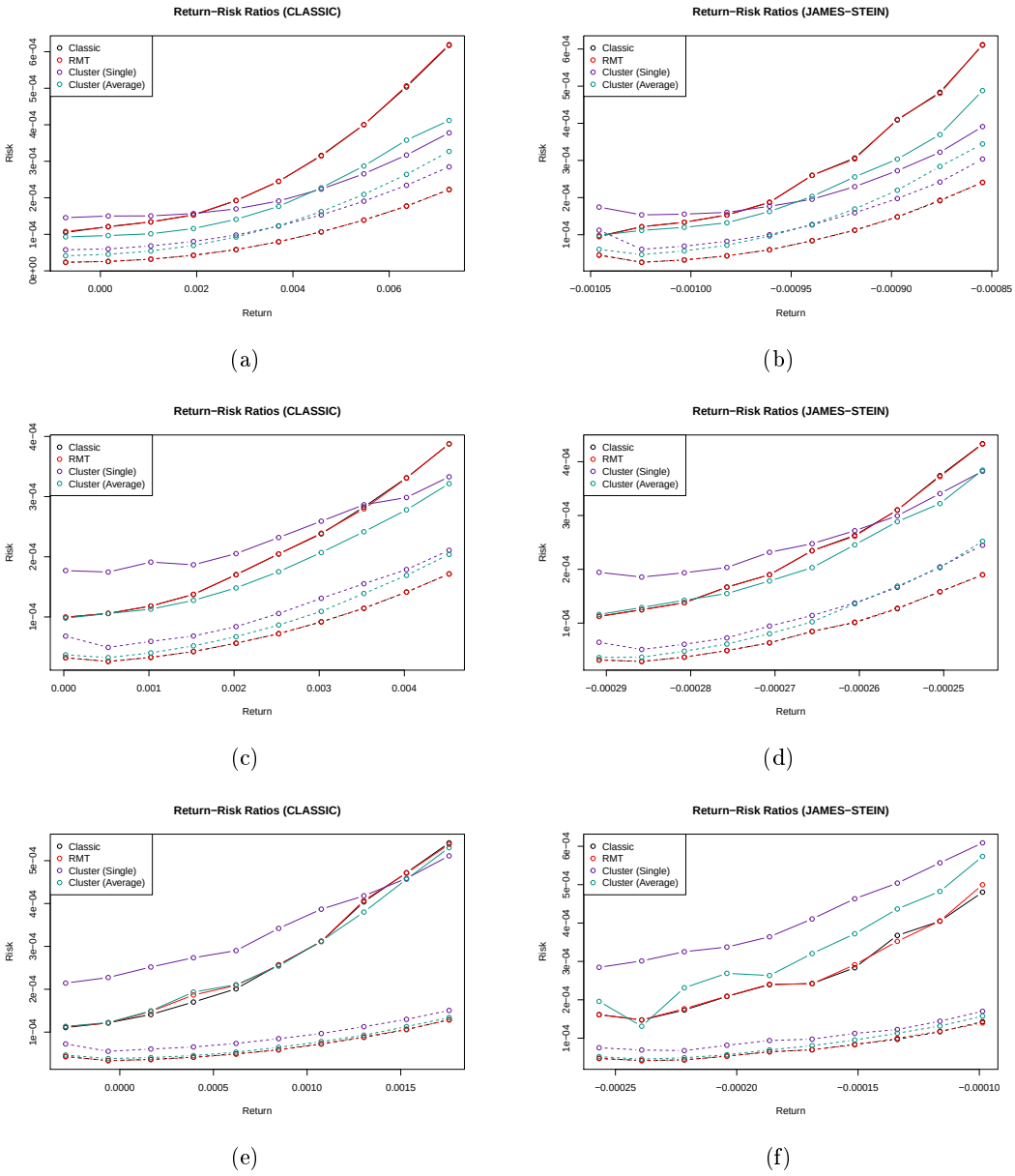


Figure 3.5. The ratio of the realized risk σ_p^2 and the predicted risk $\hat{\sigma}_p^2$ as the function of expected portfolio return (continuous line) and realized return (dashed line) for the different procedures for $T = 50, 100, 500$ (top-down) using the maximum likelihood estimator (left panels) and the James-Stein estimator (right panels). The data set contains 40 BSE stocks for the period 1995-2016.

provided better expected returns and a smaller risk ratio (i.e. higher reliability) for the smaller values of T ($T = 50, 100$, see Tab. 3.1). The results tells us that the James-Stein return estimation, although it increases the deviation of the realized returns, provides a smaller risk ratio and an improvement on the Sharpe ratio. The Sharpe ratio of the minimum variance portfolio (see Tab. 3.1 last four columns) was the highest due to the very small expected risk that the method estimated, while its reliability is significantly smaller than those got using the other return estimators.

For the Yahoo data set, the same is true for the realized returns as in the case of BSE data set. Here, the smallest risk ratio was obtained when $T = 100$ days (Fig. 3.6(c) and Fig. 3.6(d)). It can also be seen that using the James-Stein return estimator pro-

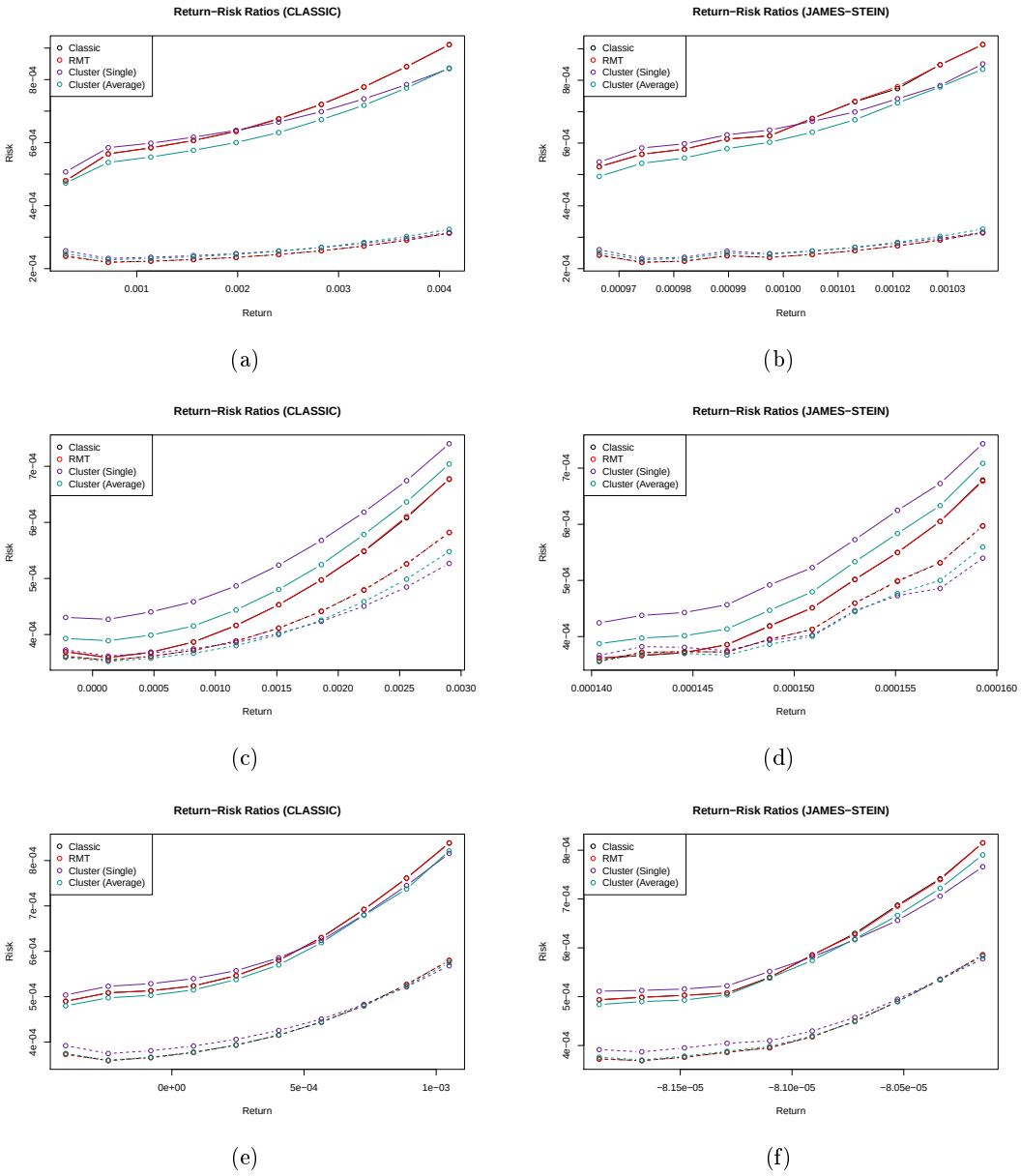


Figure 3.6. The ratio of the realized risk σ_p^2 and the predicted risk $\hat{\sigma}_p^2$ as the function of expected portfolio return (continuous line) and realized return (dashed line) for the different procedures as $T = 50, 100, 500$ (top-down) using the maximum likelihood estimator (left panels) and the James-Stein estimator (right panel). The data set contains 48 IT sector companies with available historical time series data in the Yahoo finance page in the period 1995-2016.

vided better results (realized returns, Sharpe ratio), while the usage minimum variance estimator decreased the risk ratio in some cases.

3.3 Summary

In this chapter, we provided a brief insight into the network modeling in economic systems. Firstly, we showed how network analysis could be applied to trade networks of countries. We studied the trading data of the EU countries and the economic superpowers from a network perspective, and we found that although the export, proportional to GDP,

has been growing in each European countries since joining the EU, the former Comecon countries have not significantly increased their GDP proportional exports to other EU countries, but have increased in the direction of Russia and China. By applying different ranking algorithms (out-degree, PageRank, HITS) to the network, we learned that the Pareto-principle (or Zipf-law, or "80-20"-law) prevails, meaning that a significant percentage of the total export of the world is executed by just a few countries. Thereby, countries where the export volume is relatively small, but have a high proportion of the GDP, are in a strong economic dependence on the superpowers. We showed that such networks have a strong core-periphery structure. We applied a modularity optimization method to identify those communities in the network that change over time. We found that the European countries in the periphery are contained in the clusters of Russia and China, in contrast to the Western-European countries that are in clusters where the central nodes are Germany and the USA, respectively, highlighting real economic dependencies among the EU countries.

Next, we investigated the Markowitz portfolio selection problem using filtered correlation matrices (networks) got by using different filtering procedures, namely a random matrix theory approach and hierarchical clustering approach. We used several estimators to determine the expected return of a portfolio. A lot of experiments have shown that, using filtered covariance matrices, the classic Markowitz solution can be outperformed in terms of realized returns and reliability, meaning that the realized risk and the estimated risk are closer to each other in that case. Our simulations revealed that the different filtering procedures provide different portfolio optimization results. Namely, the most useful methods may be different depending on the risk level of the portfolio, the investment period size and reliability of the risk and return estimation. We think that other filtering procedures combined with different return estimators could also provide interesting or better results with different parameter settings (e.g. expected returns, portfolio size, investment period length) of the optimization problem.

BSE data set		Average return estimator				James-Stein estimator				Min variance estimator			
	Filtering	Realized Return	Realized Return (sd)	Sharpe ratio	Risk Ratio	Realized Return	Realized Return (sd)	Sharpe ratio	Risk Ratio	Realized Return	Realized Return (sd)	Sharpe ratio	Risk Ratio
$T = 50$	Classic	0.00123	0.00465	8.18922	2.30546	0.00117	0.00466	6.85066	2.14136	0.00084	0.00461	298776	3.32719
	RMT	0.00124	0.00465	7.92646	2.31340	0.00118	0.00466	7.14867	2.16424	0.00084	0.00461	214031	3.32716
	Single linkage	0.00121	0.00304	4.77064	0.73042	0.00192	0.00502	11.82008	0.76601	0.00107	0.00452	155939	1.17548
	Average linkage	0.00073	0.00189	0.51529	0.52383	0.00185	0.00475	10.52615	0.68557	0.00100	0.00447	266844	1.34028
$T = 100$	Classic	0.00013	0.00153	7.66392	2.01878	0.00101	0.00328	12.57921	1.99415	0.00070	0.00280	520868	3.52232
	RMT	0.00015	0.00154	7.87615	2.02137	0.00099	0.00325	12.34801	2.00979	0.00069	0.00280	508169	3.52259
	Single linkage	0.00119	0.00294	12.90469	1.44612	0.00164	0.00352	13.56478	1.45191	0.00080	0.00282	406972	2.65070
	Average linkage	0.00017	0.00114	8.95616	1.20554	0.00133	0.00339	13.76679	1.22242	0.00072	0.00280	422477	3.28602
$T = 500$	Classic	-0.00047	0.00121	-11.41798	3.06961	-0.00015	0.00115	-1.61974	2.70563	0.00026	0.00080	23937	4.75457
	RMT	-0.00052	0.00127	-14.16135	3.29680	-0.00016	0.00115	-1.86695	2.72909	0.00026	0.00080	24031	4.75513
	Single linkage	-0.00011	0.00121	-2.19925	2.94493	0.00013	0.00132	1.26614	2.88015	0.00034	0.00077	20362	3.17707
	Average linkage	-0.00053	0.00127	-13.73743	2.96084	-0.00009	0.00108	-0.25504	2.79321	0.00028	0.00075	22544	4.24691
Yahoo data set													
$T = 50$	Classic	-0.00058	0.00335	-3.76019	1.69646	-0.00055	0.00331	-3.46238	1.64100	-0.00075	0.00318	-9.20659	1.44308
	RMT	-0.00057	0.00334	-3.62685	1.69695	-0.00055	0.00331	-3.45527	1.64236	-0.00075	0.00317	-9.19923	1.44350
	Single linkage	-0.00048	0.00335	-0.55588	1.57894	-0.00049	0.00335	-0.55640	1.58649	-0.00074	0.00331	-4.67897	1.45398
	Average linkage	-0.00045	0.00328	0.37105	1.44763	-0.00045	0.00328	0.53901	1.45527	-0.00068	0.00318	-3.80231	1.28830
$T = 100$	Classic	0.00030	0.00223	1.58302	0.07236	0.00033	0.00226	1.64966	0.06094	0.00024	0.00199	0.10471	0.01050
	RMT	0.00030	0.00223	1.59425	0.07074	0.00034	0.00226	1.66608	0.05837	0.00024	0.00199	0.09566	0.01217
	Single linkage	0.00012	0.00213	0.62113	0.26152	0.00014	0.00219	0.55639	0.25039	0.00009	0.00187	-0.46180	0.15097
	Average linkage	0.00022	0.00218	1.08339	0.16704	0.00024	0.00222	1.11380	0.15679	0.00017	0.00194	-0.33963	0.08565
$T = 500$	Classic	-0.00030	0.00070	-1.11800	0.38939	-0.00029	0.00070	-1.05358	0.36550	-0.00024	0.00062	-0.69456	0.38257
	RMT	-0.00030	0.00070	-1.11424	0.38804	-0.00029	0.00070	-1.05042	0.36223	-0.00024	0.00061	-0.69082	0.38035
	Single linkage	-0.00032	0.00069	-1.05861	0.37295	-0.00031	0.00070	-1.03993	0.34634	-0.00027	0.00061	-0.67158	0.36608
	Average linkage	-0.00030	0.00068	-1.08412	0.36562	-0.00029	0.00069	-1.03587	0.35142	-0.00024	0.00059	-0.69585	0.35530

Table 3.1. Bootstrap experiments using 50 random samples for each value of T when the return is the mean of the average expected return of the portfolio and the maximal expected return over all stocks.

Chapter 4

Network Models and Linear Algebra for Rating and Prediction

The problem of assigning scores to a set of individuals based on their pairwise comparisons appears in many areas and activities. For example in sports, players or teams are ranked according to the outcomes of games that they played; the impact of scientific publications can be measured using the relations among their citations. Web search engines rank websites based on their hyperlink structure. The centrality of individuals in social systems can also be evaluated according to their social relations. As we saw earlier, the ranking of individuals based on the underlying graph that models their bilateral relations has become the central ingredient of Google's search engine and later it appeared in many areas from social network analysis to optimization in technical networks (e.g. road and electric networks) [110].

In the previous chapters we presented models and examples for rating and ranking based on network structure: in Chapter 2 we defined a network algorithm for rating scientific papers; then we used a similar procedure to measure the performance of students in public education based on their pairwise comparisons. In Chapter 3 countries were ranked using the trade network of them. Now in this chapter we introduce see rating and ranking methods of nodes of bipartite networks.

In this chapter we introduce more research and results of the author in the topic. Firstly, we discuss how network models and related linear algebraic methods can be used to rate the actors (players, agents) of the modeled system and we make predictions for the future events, based on pairwise comparison graphs (or matrices, as we also refer to them here). The final goal of this chapter is to present a new model for probabilistic forecasting in sports based on linear algebraic rating methods which simply use the historical game results data of the investigated sport competitions. In contrast to those techniques that use the actual respective strength of the two competing teams, we provide a (complex) system level approach. The assumption of our model is that if a rating of the teams after a game day correctly reflects the actual relative performance, i.e. the performance of the system of teams in the competition, then the smaller the performance of the system's changes after a certain event occurs in an upcoming single game the higher the probability that that event will occur. We discuss several prediction methods including the widely-

used Bradley-Terry model, the betting odds predictions and our proposed method in detail. We present our initial empirical results obtained by measuring the accuracy and the predictive power of the methods presented.

4.1 Rating and Ranking in Sports

Ranking in sports is important for those who are interested in the various professional or amateur leagues as fans, managers, financial investors and for the growing number of gamblers who bet on offline or online platforms [178]. Ranking and, in fact, performance rating of athletes and sport teams play a crucial role in sports betting from both the better's and the betting agency's point of view.

In many sports, only the win/loss ratio is considered (see e.g. the most popular sports in the U.S.) for ranking the teams or players, i.e. a higher value indicates a higher position in the ranking. In the case of equal win/loss rates, the result(s) of the head-to-head matches between the players/teams in question and other simple statistics are considered to determine the ranking positions. In many sports, instead of the round-robin system, the type of the most relevant competitions is a single-elimination tournament (also called knock-out or cup) maybe with a preceding group stage. Thus the players play just few matches against only a small subset of their competitors. The official ranking of the players is usually determined by a sport specific rating system (e.g. see tennis, table tennis, combat sports, etc.). In fact, in a tournament, in a regular season or in a given period each player/team plays with only a subset of the others and a player/team who plays against weaker opponents have a considerable advantage compared with those who play against stronger ones.

Many approaches trace back the ranking problem to the solution of a system of linear equations, where the entries of the coefficient matrix refer in some way to the results of the games played. From the study of this pairwise comparison scheme (for early studies see e.g. [25, 46, 102]), several matrix-based ranking algorithm have appeared in sports (see e.g. [42] for chess teams, [45, 155] for tennis players, and [19, 38, 83, 136, 142] for American football teams). For a good mathematical guide to ranking in sports, see e.g. [96], while some useful comprehensive studies are e.g. [82] and [136].

4.1.1 Some Linear Algebraic Rating Methods

Next, we give a short description of the ranking methods we will use. Hundreds of ranking methods have been appeared in the long history of ranking in sports: for a more detailed introduction on ranking methods, we refer to [12] and [111]. The selection of the methods we used satisfy the following criterion: (1) each method is based on linear algebra, (2) each method has been proved to be successful in real applications, and (3) each method has a simple formulation with, in most cases, a closed solution. Before going into more detail, some definitions and notations, that are consistent with the network terminology are introduced.

Let $V = (1, \ldots, n)$ be the set of n teams (or players) and let R be the number of

game days in a competition among the teams in V. A *rating* is a function $\phi^r : V \to \mathbb{R}^n$ that assigns a score to each team after each game day r ($r = 1, \ldots, R$). This is considered as the quantitative "strength" of the teams. A *ranking* $\sigma^r : V \to V$, after game day r, is an ordering of the teams that is simply obtained by sorting the teams according to the rating ϕ^r. For rating and ranking the teams we consider only the game result information, i.e. win and loss or final result information. We note that the methods can be easily extended to the case when ties are allowed. Furthermore the matrix contains the final scores of the games can also be considered; for more details we refer to [12].

Let W be the $n \times n$ matrix with entries $W_{ij} = \#\{i \text{ won against } j\}$. The elements of the $n \times 1$ vectors $\mathbf{w} = W\mathbb{1}$, $\mathbf{l} = W^T\mathbb{1}$ and $\mathbf{t} = (W + W^T)\mathbb{1}$ are the number of wins, losses and the total number of games played by team i ($i = 1 \ldots, n$), respectively, where $\mathbb{1}$ is the $n \times 1$ with all entries equal to 1. Since each game considered here is either a win or a loss, thus $\mathbf{t} = \mathbf{w} + \mathbf{l}$. We define $T = \mathrm{diag}(t_i)$, namely the diagonal matrix with entries $T_{ii} = t_i$, $(i = 1 \ldots, n)$ and $T_{ij} = 0$ if $i \neq j$. By using these notations, we can describe some widely-used linear algebraic rating methods within unified framework.

Winning percentage (WP)

The Winning Percentage of a team i after game day r is simply defined as $\phi^r_{WP,i} = w_i/t_i$. The vector of winning percentages of the teams after game day r can be computed as

$$\phi^r_{WP} = T^{-1}\mathbf{w}. \tag{4.1}$$

The advantage of the method is that it can be easily calculated and interpreted. The main disadvantage is that it do not take into account the strength of the opponent teams, only the outcomes of the single games.

Massey's least squares method (M)

Kenneth Massey in his bachelor's thesis (1997) applied the least squares method for ranking sports and assumed that the rating difference between two teams was proportional to the score difference of the game between them (if they played) [136]. Let $Y_{r,i,j}$ be a random variable that denotes the score difference between i and j in an upcoming game r. Then

$$\mathbb{E}(Y_{r,i,j}) = \phi^r_{M,i} - \phi^r_{M,j}. \tag{4.2}$$

If $Y_{r,i,j} = y_{r,i,j}$ after game r and X is the $m \times n$ (m is the number of all game played) matrix with entries $x_{r,i} = 1$ if i won in game r, $x_{r,i} = -1$ if i defeated in game r and $x_{r,i} = 0$ otherwise, then the rating of the teams is given by the solution of the linear system

$$X\phi^r_M = \mathbf{y}, \tag{4.3}$$

where $\mathbf{y}$ is the $m \times 1$ vector of the realized score differences. Multiplying 4.3 by X^T from the left we get $X^TX\phi^r_M = X^T\mathbf{y}$ and denoting $X^TX = M$ and $X^T\mathbf{y} = \mathbf{p}$, the rating of

the teams after game r can be obtained solving the linear system

$$M\phi_M^r = \mathbf{p}, \tag{4.4}$$

where $M = T - W - W^T$ contains the total number of games played by the teams in the diagonal; if $i \neq j$ then $M_{ij} = -W_{ij} - W_{ji}$; that is, the number of games played between teams i and j with a negative sign, while $\mathbf{p}$ contains the total score differences of each team. By using $A = W - W^T$ the system 4.5 is equivalent to $T\phi_M^r - A\phi_M^r = \mathbf{p}$, and hence $\mathbf{r} = T^{-1}A\phi_M^r + T^{-1}\phi_M^r$. It follows that

$$\phi_{M,i}^r = \frac{1}{T_{ii}} \sum_j A_{ij} r_j + \frac{p_i}{T_{ii}}, \tag{4.5}$$

where the first term is the average rating of teams against i has played, while the second term is the average point spread of team i. In the case of competitions where there are more than one game is played between some pair of teams we may use $\mathbf{p} = \mathbf{w} - \mathbf{l}$. Since $\text{rank}(M) < n$, the linear system Eq. 4.5 does not have a unique solution. To overcome this problem, one possible solution is to replace any row in M with a row with all entry equals to 1 and replace the corresponding entry of vector $\mathbf{w} - \mathbf{l}$ with zero.

Colley's least squares method (C)

Colley's method [38] is also a modification of the least squares method by using an observation called Laplace's rule of succession (see [162], page 148) which states that if one observed k successes out of r attempts, then $(k+1)/(r+1)$ is a better estimation for the next event to be a success than k/r. Since $t_i = w_i + l_i$, we have

$$w_i = \frac{w_i + n_i - l_i}{2} = \frac{w_i - l_i}{2} + \frac{t_i}{2}. \tag{4.6}$$

Colley observed that the second term is the summation of all terms equal to $1/2$, corresponding to the default rating of a team with zero played games. Generalizing this using the opponent's strength

$$\frac{t_i}{2} = \sum_{j=1}^{t_i} \phi_{C,j}^r, \tag{4.7}$$

it follows that

$$\phi_{C,i}^r = \frac{w_i + 1}{t_i + 2} = \frac{(w_i - l_i)/2 + \sum_{j=1}^{t_i} \phi_{C,j}^r + 1}{t_i + 2}. \tag{4.8}$$

Rearranging and writing it in linear system form, we get

$$C\phi_C^r = \mathbf{b}. \tag{4.9}$$

The rating vector ϕ_C^r of the teams is the solution of the linear system Eq. 4.9, where $C = M + 2I$ (here, I is the identity matrix) and $\mathbf{b} = \mathbf{1} + (\mathbf{w} - \mathbf{l})/2$. It can be easily seen that the linear system Eq. 4.9 always has a unique solution.

Keener method (K)

Keener's method [100] is a so-called spectral rating method which uses the Perron-Forbenius eigenvector for the rating and (after round r) it is given by the solution of the eigenvalue equation

$$T^{-1}W\phi_K^r = \lambda\phi_K^r \tag{4.10}$$

In Eq. 4.11, λ is the dominant eigenvalue of the matrix $T^{-1}W$. This exists for a matrix with non-negative entries, and any other eigenvalue is smaller in absolute value. The corresponding eigenvector (which is called the Perron-Frobenius eigenvector) has non-negative entries and it gives the rating of the teams. Originally, the method was defined for the case where the final scores of the games are considered. The Keener matrix, also based on the Laplace's rule of succession, is defined as $K_{ij} = (W_{ij}+1)/(W_{ij}+W_{ji}+2)$. Then, the Keener rating vector of the teams is given by the solution of the equation

$$K\phi_K^r = \lambda\phi_K^r. \tag{4.11}$$

PageRank method (PR)

In our result matrix representation of sports game outcomes, using the result matrix W instead of adjacency matrix A and using $\mathbf{l}\mathbb{1}^t$ instead of D, the PageRank rating vector of the teams defined as

$$\phi_{PR} = \mathbf{\Pi} = \frac{\lambda}{N}[I - (1-\lambda)W^t(\mathbf{l}\mathbb{1}^t)^{-1}]^{-1}\mathbb{1}. \tag{4.12}$$

Assuming that $\mathbb{1}\mathbf{PR} = 1$ Eq. 4.12 implies that

$$\mathbf{\Pi} = [\frac{\lambda}{N}\mathbb{1}\mathbb{1}^t - (1-\lambda)W^t(\mathbf{l}\mathbb{1}^t)^{-1}]\mathbf{\Pi}, \tag{4.13}$$

which shows that $\mathbf{\Pi}$, is the eigenvector of the matrix $\frac{\lambda}{N}\mathbb{1}\mathbb{1}^t - (1-\lambda)W^t(\mathbf{l}\mathbb{1}^t)^{-1}$, belongs to the eigenvalue 1, which is the largest (dominant) eigenvalue of this matrix by a consequence of the Frobenius-Perron theorem for row-stochastic matrices (see e.g. [141], Ch. 8.) as we mentioned earlier.

Time-dependent PageRank method (tdPR)

We modified the PageRank algorithm such that the weight (i.e the transition probability) of each edge decreases whenever a new edge appears in the graph. Formally, the new approach is that after the kth match was played in a given period, the weight of the latest edge becomes 1, the second latest becomes $1/2$, the ith latest becomes $1/i$, the oldest one becomes $1/k$. We normalize the weights such that the matrix obtained become row-stochastic (i.e. each row summing to 1) and we recalculate the ranking every time a new result is registered in the database by solving the equation

$$\phi_{tdPR} = \mathbf{\Pi} = \frac{\lambda}{N}[I - (1-\lambda)W^t_{\text{mod}}(\mathbf{l}\mathbb{1}^t)^{-1}]^{-1}\mathbb{1}, \tag{4.14}$$

where the entries of W_{mod} are then the new transition probability values, calculated as we described.

Network Representation of the Methods

We would like to emphasize, that several of the above-defined methods have an interpretation on a graph. Using the game results data set, one can define a directed multigraph[1], where nodes represent players/teams, while links between them represent outcomes of games they played. The links are directed and each of them is going from the loser team to the winning team. If ties are also considered they can be represented by two directed links with opposite directions and half weight. In this case, matrix W is the adjacency matrix of the directed multigraph representation of the results, $\mathbf{w}$ and $\mathbf{l}$ contain the in- and out-degrees of nodes, respectively. From a network science perspective, Massey's M matrix is the graph Laplacian if the result matrix is treated as the matrix of a symmetric undirected graph. The rating vector ϕ_M defined in Eq. 4.5 is then equivalent to the potential vector over a resistor network defined W with supply vector $\mathbf{w} - \mathbf{l}$ [72]. The PR and td-PR methods are the modifications of the classic PageRank algorithms performed on the results graph.

4.1.2 Experimental Results

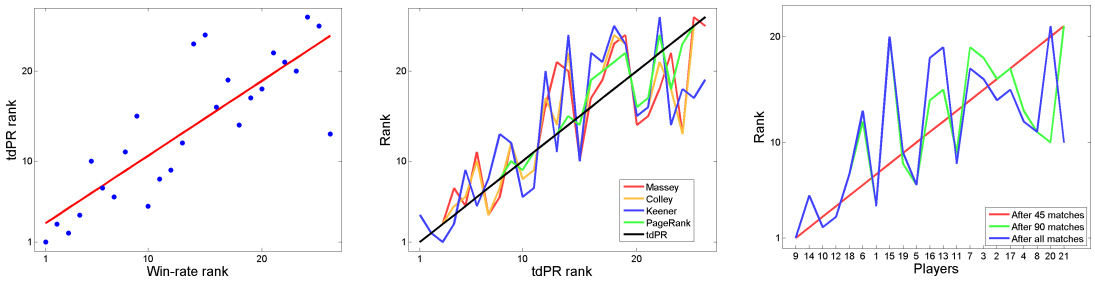
We applied the methods described above to the table tennis competition of the Institute of Informatics at the University of Szeged (the data set we used can be found in the website `http://www.inf.u-szeged.hu/~london/TableTennisResults.txt`). In that competition, there is no any rule for the selection of the opponents or the date of the match. The only restriction is that 7 days must have elapsed between two matches of the same players. Without considering the organizational rules and by considering the results only in a given period, it can be seen that these features occurred in many sports where the competition is not a round-robin.

In Table 4.1, we report the scores of the players obtained by the different ranking methods. In the case of the PR and the tdPR algorithms, we used $\lambda = 0.1, 0.2, 0.3, 0.4$, respectively. We found that the td-PR score is robust against these variations of λ (the Pearson correlation was more than 0.95 for each pair). Furthermore, the td-PR method was proved to be very effective in finding the top players of the competition that could be justified *a posteriori* by knowing the players skills.

We used *Kendall's τ rank correlation* [101] to quantify the correlation between the different methods. The rank correlation coefficient is defined as $\tau = (n_c - n_d)/\binom{n}{2}$, where n_c (n_d) is the number of such pairs that have the same (opposite) order in both ranking list. However, the td-PR score is positively correlated with the winning percentage, differences can be seen by comparing the two methods. The relation between the td-PR and the WP is shown in Fig. 4.1(a).

A relevant outlier on the list is player 14, having a win ratio 50%, who precedes players 5, 23, 19 and 21, but the latter has a better WP than himself. He is placed at

[1] That is a graph where multiple links are allowed.



(a) (b) (c)

Figure 4.1. (a) The scatter plot of the tdPR rank vs. the WP rank. (b) The results obtained by the different ranking methods. (c) The tdPR ranks of the players after 45, 90 and 180 played games.

position 4 and this is consistent with the fact that he was just defeated by players (player 10, player 12) who are ranked higher.

Fig. 4.1(b) shows the relation between td-PR and the other ranking methods. Despite the high correlation between td-PR and the other methods, we observed that the time-dependent method has a better predictive power. We considered the first half of the total number matches that had been played since the start of the competition and calculated the td-PR values of that period. Then we checked the results of the upcoming matches and the changes in the ranking. It can be observed that the players with much a higher td-PR score after the first half of the total matches played won a high proportion of their matches against players with smaller td-PR values in the later part of the competition. Our observations suggest that the difference between the td-PR values of the players can provide a reliable prediction for the upcoming matches. Fig. 4.1(c) shows the td-PR ranks of the players after 45, 90 and 180 played games. We should mention that Fig. 4.1(c) only contains those players, who had already played at least one played match after the first 45 matches of the competition. Obviously, at that time we could not predict the results of those players who joined later in the competition.

Table 4.1. Ratings obtained by the different methods; the ordering of the players is given by the decreasing order of the td-PR values

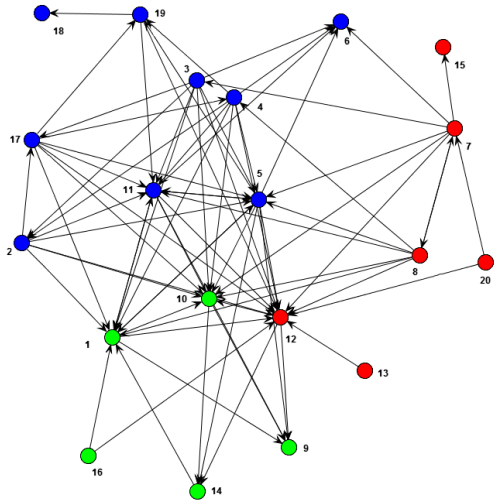
Player	#Plays	#Wins	Win ratio	Massey	Colley	Keener	PR	td-PR
9	13	13	1.000	1.418	1.074	0.229	0.113	0.138
10	29	25	0.862	0.972	0.923	0.238	0.089	0.093
12	30	26	0.867	0.859	0.882	0.245	0.083	0.085
1	63	44	0.698	0.497	0.722	0.233	0.071	0.075
14	6	3	0.500	0.658	0.717	0.198	0.064	0.070
5	38	22	0.579	0.266	0.604	0.200	0.050	0.052
23	5	3	0.600	0.779	0.736	0.199	0.047	0.047
18	16	8	0.500	0.555	0.700	0.192	0.046	0.045
11	24	11	0.458	0.209	0.564	0.193	0.039	0.040
19	10	6	0.600	0.454	0.664	0.200	0.042	0.039
21	13	7	0.538	0.325	0.615	0.199	0.035	0.032
8	19	6	0.316	-0.338	0.354	0.181	0.031	0.032
26	1	0	0.000	-0.503	0.407	0.194	0.031	0.029
4	19	3	0.158	-0.474	0.265	0.172	0.025	0.026
6	10	5	0.500	0.269	0.586	0.194	0.030	0.025
2	17	3	0.176	-0.380	0.307	0.177	0.022	0.024
17	13	2	0.154	-0.437	0.286	0.178	0.019	0.020
3	13	1	0.077	-0.615	0.213	0.171	0.019	0.020
7	12	2	0.167	-0.650	0.219	0.176	0.018	0.018
16	2	0	0.000	-0.322	0.401	0.191	0.024	0.018
13	2	0	0.000	-0.322	0.401	0.191	0.024	0.018
22	14	1	0.071	-0.433	0.277	0.169	0.016	0.016
24	4	1	0.250	-0.507	0.349	0.191	0.023	0.016
15	5	1	0.200	-0.174	0.416	0.188	0.017	0.010
25	3	0	0.000	-1.060	0.186	0.191	0.015	0.007
20	5	0	0.000	-1.047	0.136	0.184	0.010	0.004

Table 4.2. Kendall's τ rank correlation between the different methods.

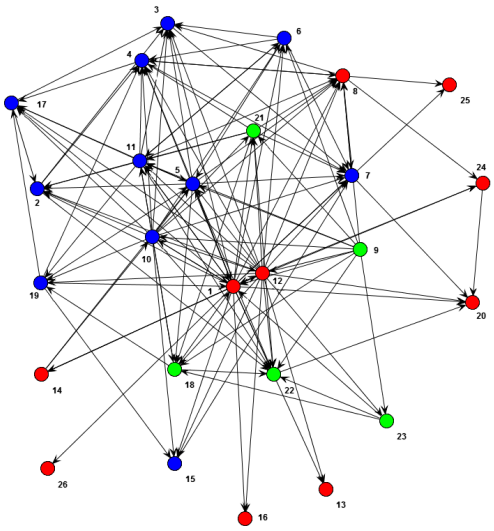
	Win/loss	Lsm	Colley	Keener	PR	tdPR
Win/loss	1.000					
MASSEY	0.705	1.000				
COLLEY	0.748	0.895	1.000			
KEENER	0.655	0.606	0.711	1.000		
PR	0.723	0.735	0.803	0.662	1.000	
tdPR	0.723	0.674	0.705	0.563	0.902	1.000

Further Ideas

We also ran a clustering algorithm (Leuven method) to see whether there exists a deeper organizational pattern behind the evolution of the result network. Fig. 4.2 shows the network with clusters that are shown in various colors. It depicts the contact graph of the players after 90 played matches (up) and the state of the championship after more than 180 matches (down). It is interesting to see the changes of the clusters in the two graphs. First, we can observe that most of the newcomer players want to play against



(a)



(b)

Figure 4.2. The contact graph of the players after 90 matches played (a) and the state of the championship after more than 180 matches (b). Nodes having same color belong to the same cluster.

the current best players (in td-PR rank/a priori) and expect to jump to the top of the ranking table. Second, it seems that players having closer td-PR values are more likely to play with each other than players having a smaller td-PR value and lower ranking position. Thus, we conjecture that the td-PR scores have a good explanatory power for a self-organizing mechanism of free-time (and perhaps professional) sports as well. This could explain the appearance of different strength classes in most of the sports, where it is more difficult to predict results within a class than results between different classes. Furthermore, from a graph theoretical point of view, a new type of "regularity" (for details, see [41]) can be defined on directed graphs, where the fraction of in/out edges of a node is around $1/2$ in the same class, and tends to 1 (or 0) between different classes.

4.2 Probabilistic Forecasting in Sports

In general, making predictions in sports is a difficult task. Traditionally, predictions have been made by experts like sports commentators, sports journalists, former players and coaches based on their experience and intuition [69]. The predictions generally appear in the form of betting odds, which, in the case of "fixed odds", provide a fairly acceptable source of expert's predictions regarding for the outcomes of sport games [154]. In the age of information and high-performance computers, as a multimillion dollar market, the sports betting market has been pervaded by a huge amount of statistics, produced after every single game, aim to evaluate the performance of teams and players [129, 169]. Thanks to the increasing quantity of available data the statistical ranking, rating and prediction methods have became more dominant in sports in the last decade. A key question is how accurate these evaluations are; more concretely, how accurately the outcomes of the upcoming games can be predicted based on the statistics, ratings and forecasting models in hand. In recent years, several statistics-based and machine learning methods have been applied to the historical results data of sport competitions.

Statistics-based forecasting models are used to predict the outcome of games based on some relevant information of the competing teams and/or players of the teams. As a detailed survey of the scientific literature of rating and forecasting methods in sports is beyond the scope of this dissertation, we will refer to only some important and recent results in the topic. The celebrated *Bradley-Terry model* [25] (with several extensions [47]) for data from paired-comparisons was developed to estimate the probability that one object will be preferred to another. Applications of the model include sport competitions as well, where the teams are the objects and the comparisons are the games between them with preferences corresponding to wins and losses (and also ties, in many sports). For some papers with a detailed literature overview and sport applications, see e.g. [30, 33, 182]). In Sec. 4.2.2 we will give a detailed description of the Bradley-Terry model. Other popular approaches are the Poisson goal distribution-based analysis (with extensions of home-field effect and tie-effect), where the game results are predicted by the number of points scored by the competing teams that are considered to be independent Poissonian random variables with means determined by the respective offense and defense abilities of the teams. For some references, see for instance [51, 99, 130]. A large family of prediction models only consider the game results win, loss (and tie) and they usually apply some probit regression model. For instance [68] and [80] consider team quality, actual performance and match significance and compare the statistical methods to expert's views represented by the published betting odds. More recently, well-known data mining techniques, like artificial neural networks, decision trees and support vector machines, have also become very popular, and some references, without being exhaustive include [40, 48, 98, 117]. A notable part of prediction models that use only the historical data of game results contains the ranking and rating-based prediction methods. Some recent articles on the topic are [12, 32, 77, 113, 173].

4.2.1 Betting Odds

Bookmakers determine *betting odds* for the games according to their expectations of outcome probabilities. Here we deal with fixed odds, means that they do not vary over time depending on the betting volumes. These "fixed-odds" represent the predictions of bookmakers [154]. Recent studies have pointed out that calculating probabilities from betting odds is an appropriate forecasting method with increasing efficiency [68, 169]. However, the efficiency of betting markets has frequently been questioned and formerly outperformed by statistical methods in some cases, see e.g. [51, 80].

From the technical point of view, if the betting odds for an upcoming game between team i and team j are $\text{odds}(i)$ and $\text{odds}(j)$, respectively, it means that if one bets \$1 to i's win and it comes out, he wins $\text{odds}(i)$ dollars, while if j wins, then the bettor loses his \$1 (similarly, one can bet to team j's win). We can calculate the probabilities of the respective events as

$$\Pr_{(\text{odds})}(i \text{ beats } j) = \frac{1/\text{odds}(i)}{1/\text{odds}(i) + 1/\text{odds}(j)} \tag{4.15}$$

and

$$\Pr_{(\text{odds})}(j \text{ beats } i) = \frac{1/\text{odds}(j)}{1/\text{odds}(i) + 1/\text{odds}(j)}. \tag{4.16}$$

We should note here that odds provided by betting agencies do not represent the true chances (as imagined by the bookmaker) that the event will or will not occur, but are the amount that the bookmaker will pay out on a winning bet. The odds include a profit margin which effectively means that the payout to a successful bettor is less than that represented by the true chance of the event occurring. This means mathematically that $1/\text{odds}(i) + 1/\text{odds}(j)$ is more than one. This profit expected by the agency is known as the "over-round on the book".

4.2.2 The Bradley-Terry Model

The *Bradley-Terry model* [25] is a widely-used method to assign probabilities for the possible outcomes when a set of n individuals are repeatedly compared with each other in pairs. For two elements i and j, according to the model, the probability that i beats j defined as

$$\Pr(i \text{ beats } j) = \frac{\pi_i}{\pi_i + \pi_j}, \tag{4.17}$$

where $\pi_i > 0$ is a parameter associated to each individual $i = 1, \ldots, n$, representing the overall skill, or "intrinsic strength" of it. Equivalently, π_i/π_j represents the odds in favor i beats j, therefore this is a "proportional-odds model". Suppose that i and j played N_{ij} games against each other with i winning W_{ij} of them, and all games are considered to be independent. The likelihood is given by

$$L(\pi_i, \ldots, \pi_n) = \prod_{i<j} \left[\frac{\pi_i}{\pi_i + \pi_j}\right]^{W_{ij}} \left[\frac{\pi_j}{\pi_i + \pi_j}\right]^{N_{ij} - W_{ij}}. \tag{4.18}$$

Then the log-likelihood is

$$\ell(\pi_i, \ldots, \pi_n) = \sum_{1 \le i \ne j \le n} \left[W_{ij} \log \pi_i - W_{ij} \log(\pi_i - \pi_j) \right]$$
$$= \sum_{i=1}^{n} W_{ij} \log \pi_i - \sum_{1 \le i < j \le n} N_{ij} \log(\pi_i + \pi_j) \tag{4.19}$$

which need to be maximized.

One possible derivation of the model assumes team i produces an unobserved score S_i, no matter which is the opposing team, with the cumulative distribution function

$$S_i \sim F_i(s) = \exp[-e^{-(s - \log \pi_i)}]. \tag{4.20}$$

It follows that distribution of the difference $S_i - S_j$ follows a logistic distribution function

$$S_i - S_j \sim F_{ij}(s) = \frac{1}{1 + e^{-(s - (\log \pi_i - \log \pi_j))}}, \tag{4.21}$$

which implies that

$$\Pr(S_i > S_j) = \Pr(S_i - S_j > 0) = 1 - \frac{1}{1 + e^{\log \pi_i - \log \pi_j}} = \frac{\pi_i}{\pi_i + \pi_j}. \tag{4.22}$$

Extension with Home advantage and Tie

A natural extension of the Bradley-Terry model with "home-field advantage", according to [1], say, is to calculate the probabilities as

$$\Pr(i \text{ beats } j) = \begin{cases} \frac{\theta \pi_i}{\theta \pi_i + \pi_j}, & \text{if } i \text{ is at home} \\ \frac{\pi_i}{\pi_i + \theta \pi_j}, & \text{if } j \text{ is at home} \end{cases} \tag{4.23}$$

where $\theta > 0$ measures the strength of the home-field advantage (or disadvantage).

Considering also a tie as a possible final result of a game, the following calculations, proposed in [158], can be used :

$$\Pr(i \text{ beats } j) = \frac{\pi_i}{\pi_i + \alpha \pi_j}, \tag{4.24}$$

$$\Pr(i \text{ ties } j) = \frac{(\alpha^2 - 1)\pi_i \pi_j}{(\pi_i + \alpha \pi_j)(\alpha \pi_i + \pi_j)} \tag{4.25}$$

where $\alpha > 1$. Combining them is straightforward. In our experiments, we used the Matlab implementations found at `http://www.stats.ox.ac.uk/~caron/code/bayesbt/` using the *expectation maximization* algorithm described in detail in [33].

4.2.3 A Rating-based Model: a general framework

Now we will describe our model, which is a rating-based method, where the rating used only deals with the win-lose or the final score statistics of the teams of the given com-

petition. The crucial assumption of the model, unlike e.g. the Bradley-Terry model, is that the rating of the teams evaluated after a given game day correctly reflects the actual performance, relative to each other, of the teams. Suppose that before game day r ($r = k, \dots, R-1$), for some k, the rating vector of the teams $V = (1, \dots, n)$ in a competition is $\phi^{r-1}(V) = (\phi_1^{r-1}, \dots, \phi_n^{r-1})$. We assume, that this rating is a good approximate of the performance the teams. The key idea for the predicting the outcome of an upcoming match on game day r between teams i and j is the assumption that the more probable an outcome is the less change it will cause in the rating vector $\phi^{r-1}(V)$. Mathematically, let us define the distances

$$\delta^r_{\{i \text{ beats } j\}} = \operatorname{dist}\big(\phi^{r-1}(V), \phi^r(V) \;\big|\; \{i \text{ beats } j\}\big) \tag{4.26}$$

and

$$\delta^r_{\{j \text{ beats } i\}} = \operatorname{dist}\big(\phi^{r-1}(V), \phi^r(V) \;\big|\; \{j \text{ beats } i\}\big), \tag{4.27}$$

by using some distance function $\operatorname{dist} : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$. Practically speaking, $\delta^r_{\{\}}$ measures how the rating vector changes after a certain game outcome on game day r. Then we can simply assign probabilities for the events $\{i \text{ beats } j\}$ and $\{j \text{ beats } i\}$, such that

$$\Pr(\{i \text{ beats } j\}) = \frac{f(\delta^r_{\{i \text{ beats } j\}})}{f(\delta^r_{\{i \text{ beats } j\}}) + f(\delta^r_{\{j \text{ beats } i\}})} \tag{4.28}$$

and

$$\Pr(\{j \text{ beats } i\}) = \frac{f(\delta^r_{\{j \text{ beats } i\}})}{f(\delta^r_{\{ \text{ beats } j\}}) + f(\delta^r_{\{j \text{ beats } i\}})}, \tag{4.29}$$

respectively, by using some $f : \mathbb{R} \to \mathbb{R}$ non-increasing function. Within this framework the rating function ϕ, the distance function δ and the non-increasing function f can be chosen independently.

Considering ties as well in our rating based-model the probabilities can be calculated as

$$\frac{f(\delta^r_{\{i \text{ beats } j\}})}{f(\delta^r_{\{i \text{ beats } j\}}) + f(\delta^r_{\{i \text{ ties } j\}}) + f(\delta^r_{\{j \text{ beats } i\}})}, \tag{4.30}$$

$$\frac{f(\delta^r_{\{i \text{ ties } j\}})}{f(\delta^r_{\{i \text{ beats} j\}}) + f(\delta^r_{\{i \text{ ties } j\}}) + f(\delta^r_{\{j \text{ beats } i\}})}. \tag{4.31}$$

The home-field advantage may be defined in various ways. Since in our experiments we used a time-dependent PageRank method, we will describe a possible way of defining home-field advantage. Furthermore, we give a more detailed description of the model in that case.

Table 4.3. Accuracy results on football data sets. The values where the difference between the Bradley-Terry method and the PageRank method was higher than 0.01 are shown in bold.

League	Season	Betting odds error	Bradley-Terry error	PageRank method error
Premier League	2011/12	0.58934	0.60864	**0.59653**
	2012/13	0.56461	0.59744	**0.58166**
	2013/14	0.54191	**0.55572**	0.59406
	2014/15	0.55740	0.60126	0.60966
Bundesliga	2011/12	0.58945	0.59994	**0.59097**
	2012/13	0.57448	0.59794	**0.58622**
	2013/14	0.55724	**0.57803**	0.60125
	2014/15	0.57268	0.60349	0.60604
La Liga	2011/12	0.54598	**0.57837**	0.58736
	2012/13	0.56417	**0.58916**	0.60205
	2013/14	0.57908	**0.58016**	0.60473
	2014/15	0.52317	0.55888	0.56172

4.2.4 Experimental Results

Forecasting Accuracy

To measure the accuracy of the forecasting we calculate the mean squared error, which is often called *Brier scoring rule* in the forecasting literature [27], described as follows. The Brier score measures the mean squared difference between the predicted probability assigned to the possible outcomes for event E and the actual outcome o_E. Suppose that for a single game g, between i and j, the forecast is $\mathbf{p}^g = (p_w^g, p_t^g, p_l^g)$ containing the probabilities of i wins, the game is a tie and i loses, respectively. Let the actual outcome of the game be $\mathbf{o}^g = (o_w^g, o_t^g, o_l^g)$ where exactly one element is 1, the other two elements are 0. Noting that the number of games played (and predicted) is N, BS is defined as

$$BS = \frac{1}{N}\sum_{g=1}^{N} ||\mathbf{p}^g - \mathbf{o}^g||_2^2 = \frac{1}{N}\sum_{g=1}^{N}[(p_w^g - o_w^g)^2 + (p_t^g - o_t^g)^2 + (p_l^g - o_l^g)^2]. \quad (4.32)$$

The best score achievable is 0. In the case of three possible outcomes (win, lost, tie) we can easily see that the forecast $\mathbf{p}^g = (1/3, 1/3, 1/3)$ (for each game g and any N) gives accuracy $BS = 2/3 = 0.666$. We consider this value as a worst-case benchmark. One question of our investigation is that how better BS values can be achieved using our method, and how close we can get to the good betting agencies' predictions.

Results on football data sets

We test our model using the following setup. For rating the teams, a time-dependent PageRank method is used. The damping factor is $\lambda = 0.1$, while we use an exponential function 0.98^{α} for time-dependency, where α denotes the number of game days elapsed between the last and the first game day that we consider for calculations. We define each day as a game day in which on a day at least one match is played. The construction of the modified PageRank matrix used in Eq. 4.14 is carried out as follows. For any game day in which we make a forecast, we consider the results matrix that contains all the results of the previous $T = 40$ game days. To take into account the home-field effect, for

each team i we distinguish team home-i and team away-i. We define a $2n \times 2n$ results matrix S, which, in fact, describes a bipartite graph where each team appears both in the home team side and the away team side of the graph. Thus, a home team and an away team PageRank values are calculated for each team. We would like to establish a connection between team home-i and team away-i using the assumption that home-i is not weaker than away-i. In our implementation we assume that home-i had a win 2-1 against away-i to give a positive bias for home-i at the beginning. In our experiments this setup performed well, but it was not optimized precisely. For the 40 game days time window, the entries of the results matrix S are defined as $S_{ij} = \#\{\text{scores team home-i achieved against team away-j}\}$. Each entry is multiplied by the time-dependency function, then the row stochastic PageRank matrix is constructed and PageRank scores are calculated according Eq. 4.14.

Using the above-defined results matrix S and the PageRank rating vector ϕ, we assign probabilities to the outcomes {home team win, tie, away team win} of an upcoming game in game day r between team home-i and team away-j as follows. Before the game day in which we make the forecast, let the calculated PageRank rating vector be $\phi_{40}^{r-1}(V)$. Since now we use the results matrix S, we should consider final scores instead of win-tie-loss outcomes considered in the model description above, to calculate the $\delta_{\{\}}^{r}$ values defined in Eq. 4.26 and Eq. 4.27. We use δ_{xy}^{r} to measure how the rating vector of the teams changes if the result of an upcoming game between teams i and j, denoted as $x : y$, where $x, y = 0, 1, \ldots$ are the scores achieved by team i and team j, respectively[2]. We define δ_{xy}^{r} as the Euclidean distance between $\phi_{40}^{r-1}(V)$ and $\phi_{40}^{r}(V)$ that is the rating vector for the new results matrix obtained by adding x to S_{ij} and y to $S_{n+j,i}$. In the results graph interpretation this simply means that an edge from node away-j to node home-i with weight x and an edge from node home-i to node away-j with weight y are added to the graph, respectively. Our assumption is that if an outcome $x : y$ has a high probability and it occurs, then it causes a small change in the PageRank vector; hence δ_{xy} will be small. To simplify the notations let $\{\delta_1, \ldots, \delta_m\}$ be the distance values obtained by considering different results $\{E_1, \ldots, E_m\}$ of the upcoming game between team i and team j. The goal now is to calculate the probability that a certain result occurs conditioned to $\{\delta_1, \ldots, \delta_m\}$. To do this, we use the following simple statistics-based machine learning method. Let $f^{+}()$ be the probability density function of δ_i random variable where the event (game result) E_i occurred. In our implementation $E_i \in \{0:0, 1:0, 1:1, \ldots, 5:5\}$, assuming that the probability of other results equals 0. Similarly, let $f^{-}()$ be the probability density function of δ_i random variable in which case the event (game result) E_i did not occur. To approximate the $f^{+}()$ and $f^{-}()$ functions, for each game we use the training data set contains all results and related δ_i $(i = 1, \ldots, m)$ values of the preceding $K = 40$ game days of that game. In our experiments, the gamma distribution (and its density function) turned out to be a fairly good approximate for $f^{+}(\delta)$ and $f^{-}(\delta)$.

Assuming that $\delta_1, \ldots, \delta_m$ are independent, using the Bayes theorem and the law of

[2]We should note here that if the result is $0:0$, then $x = y = 1/2$ is used.

total probability, we get

$$\begin{aligned}\Pr(E_i|\{\delta_1,\ldots,\delta_m\}) &= \frac{\Pr(\delta_1,\ldots,\delta_m|E_k)\Pr(E_i)}{\Pr(\delta_1,\ldots,\delta_m)} = \frac{\prod_k \Pr(\delta_k|E_i)\Pr(E_i)}{\sum_\ell \Pr(\delta_1,\ldots,\delta_m|E_\ell)\Pr(E_\ell)} = \\ &= \frac{\prod_k \Pr(\delta_k|E_i)\Pr(E_i)}{\sum_\ell \prod_k \Pr(\delta_k|E_\ell)\Pr(E_\ell)} = \frac{f^+(\delta_i)\prod_{k\neq i} f^-(\delta_k)\frac{1}{m}}{\sum_\ell f^+(\delta_\ell)\prod_{k\neq l} f^-(\delta_\ell)\frac{1}{m}} = \\ &= \frac{f^+(\delta_i)\prod_{k\neq i} f^-(\delta_k)}{\sum_\ell f^+(\delta_\ell)\prod_{k\neq l} f^-(\delta_\ell)}. \qquad (4.33)\end{aligned}$$

We should note here that using Eq. 4.33 we assign probabilities to concrete game final results, which is another novelty of our model. Then, for the upcoming game between i and j, based on Eq. 4.33, the outcome probability of the event $\{i \text{ beats } j\}$ is calculated as

$$\Pr(i \text{ beats } j) = \sum_{\substack{k:\ E_k \text{ encodes a result} \\ \text{of team-i win}}} \Pr(E_k|\{\delta_1,\ldots,\delta_m\}), \qquad (4.34)$$

where we sum over those E_k results for which team-i beats team-j (i.e. 1:0, 2:0, 2:1, 3:0, 3:1, etc.). The probabilities $\Pr(i \text{ ties } j)$ and $\Pr(j \text{ beats } i)$ can be calculated in a similar way.

Our initial results are summarized in Table 4.3, which contain the accuracy scores (i.e. Brier scores, using Eq. 4.32) of the different forecasting methods applied in different seasons of various European football championships. To calculate the betting odds probabilities we used the betting odds provided by bet365 bookmaker available at `http://www.football-data.co.uk/`. We could see that these predictions gave the best accuracy score (BS) in each case. We highlighted the values where the difference between the Bradley-Terry method and the PageRank method was higher than 0.01. Although we can see that slightly more than half of the cases the Bradley-Terry model gives a better accuracy, the results are still promising considering the fact that the parameters of the method and our implementation are far from being optimized.

4.3 Summary

In this study, we defined a time-dependent PageRank-based algorithm and applied it for ranking players in a university table tennis competition. According to our tdPR method, the ranking of a player is not only determined by the number of his or her victories, but it depends on how good the players are he could beat or lose against. It means that a good player is needed to beat for higher ranking position, but winning many matches against weaker opponents does not lead anyone to the first position in the ranking table. The time-dependency of weights of the matches guarantee that the matches played a long time ago do not count as much in the ranking. Another aim of the time-dependency is to pressure the players to play regularly or else their results would be out of date; then they would count much less in the ranking. We also observed that our method has good predictive power. This may be interesting in other aspects of sports, like estimating the betting odds for games. We think that a self-organization pattern operates

in the background of the evolution of the contact graph. Obviously, players who want to enter matches are expected to be exciting, but the nature of such competitions can be modeled and measured mathematically just by knowing the time-series of the results. This observation suggest we should define a special preferential attachment mechanism where players having higher PageRank values are more likely to play (contact) with each other and this is may be related to the emergence of an elite group in sports. Further research is needed to evaluate this hypothesis, and testing our method for different sports and data sets is also another plan for the future.

Next, we presented a new model for probabilistic forecasting in sports based on rating methods that simply use the historical game results data of the investigated sport competition. In contrast to those techniques that use the current respective strength, calculated using the previous results of the two competing teams, like, the celebrated Bradley-Terry model, we provided a "forward-looking" type network based approach. The assumption of our model is that the rating of the teams after a game day is correctly reflects their current relative performance. We consider that the smaller the rating vector, which contains the ratings of each team, and it changes after a certain event occurs in an upcoming single game, the higher the probability that this event will occur. Performing experiments on results data sets of European football championships, we observed that this model performed well (it outperformed the advanced versions of the Bradley-Terry model in some cases) in terms of predictive accuracy. However, we should note here, that parameter fine tuning and optimizing certain parts of our implementation are tasks that need to be examined in the future.

Chapter 5

Bipartite Network Models of Real-world Systems

In the previous chapters we saw examples of complex network models and their applications to real-life systems, from scientometrics through educational data mining to economic modeling. A special, but rather important class of complex systems can be represented by bipartite networks, in which the nodes of the network can be divided into two classes, A and B, say, and links only connect nodes of the different classes. In this chapter, after introducing the main definitions, concepts and tools for analyzing bipartite networks the author's results will be presented.

Firstly, a method for finding the core of communities (in other words clusters) is presented for bipartite networks using a one-mode projection method with statistical validation. Cores of communities are highly informative and robust with respect to the presence of errors and/or missing entries in the bipartite network. We assess the statistical robustness of cores by investigating an artificial benchmark network. We will show that this kind of filtering procedure necessarily increases the precision of the community detection, finds highly stable cores (with high precision) and suggested uses, even with the drawback that it decreases the level of accuracy in some situations. We also present experimental results on real systems that can be modeled via bipartite networks.

Secondly, we describe how a generalized version of the PageRank and HITS algorithms can be defined for bipartite networks and, as a case study, when applied on data sets of wine tasting events in order to rank tasters according to their ability and professional skill. In general, we will show that our ranking performs well due to our apriori knowledge about the tasters, and it is able filter out incompetent tasters, who, for example, gave the average score of some other tasters (i.e. cheating in some way) for the wines tasted. Furthermore, we point out that our method gives a clearer picture about the competence of wine tasters than other statistical methods that can be readily applied.

5.1 Bipartite Networks

In this chapter, we will deal with the type of complex systems that can be modeled by bipartite networks. Bipartite networks naturally appear in areas ranging from social to

biological systems and examples include, among many others, the actors–movies network (where the two classes of nodes represent actors and movies, and there is a link between an actor and the movie if the actor played in that movie) [150], scientists–research papers cooperation networks (a link exist if a scientist is one of the authors of the paper) [144], diseases–genes networks (links represent gene-disease associations) [36], plants–pollinators mutualistic networks (a link exists if a plant species is pollinated by a pollinator species) [13], banks–firms money transfer networks (links represent loan relations between banks and firms) [135] and words co-occurrence networks (where the two classes of nodes are words and sentences/texts, and there is a link between two nodes if a given word occurs in the represented sentence) [149]. Although these types of networks contain a large amount of information about the system, retrieving this information is generally hard.

Two fundamental approaches have been used to analyze bipartite networks. The first is the so-called "direct" approach, where the bipartite network is analyzed directly by jointly analyzing the two sets A and B via the linking structure between them. The second approach is called the "projection" method, in which the network is converted into two one-mode projections (i.e. two "unipartite" networks of set A and set B, respectively) and then they are analyzed separately. There are several reasons for thinking that the direct approach is better. A key idea is that important structural information may be lost by using one-mode projections [114]. However, recent studies have pointed out that data is not necessarily lost by using projections [62, 138]. A real advantage of using the second approach is the availability of the arsenal of well-refined techniques present in the literature for analyzing "unipartite" complex networks that usually cannot be used for bipartite networks directly.

Formally, a *bipartite network* $G = (A, B, E)$ is a triple where $A = \{a_1, \dots a_n\}$ and $B = \{b_1, \dots b_m\}$ represent the set of nodes of the two parts, respectively, while $E \subseteq A \times B$ denotes the set of edges that only connect nodes of the different parts. Let M be the $n \times m$ bipartite adjacency matrix of G, where $M_{ij} = 1$ if $(a_i, b_j) \in E$ and $M_{ij} = 0$, otherwise.

Here, without being exhaustive, we mention two characterizations of bipartite networks. They are: (i) a network is bipartite if and only if it does not contain an odd cycle[1]; (ii) a network is bipartite if and only if it is 2-colorable[2] (see e.g. in [6], ps. 7-8.).

5.1.1 Communities in Bipartite Networks

The community structure in a bipartite network can be revealed in various ways depending on questions of interest [11, 85, 139, 191]. Often, the communities of only one side is analyzed by using a one-mode projection method. In recent years, according to different definitions of communities in bipartite networks, many methods have been proposed to find them using both the direct [3, 11, 85, 187] and the projection approach [62, 112, 138], but still many problems and questions arise in the topic. One of the main questions that

[1] A cycle is odd, if the number of its edges is odd.

[2] It means that the nodes of the graph can be colored by two colors such that no adjacent nodes have the same color.

we investigate here is the reliability of the adjacency projection based on the community structure of the projected network.

5.1.2 One-mode Projections

Most of the existing approaches simply construct a one-mode projection by assigning a weight to each pair of nodes in A (or B, respectively) based on the number of their neighbors in B (or in A). We call these projections "adjacency projections". The *co-occurrence matrix* C_A is defined as $C_A = MM^T$, where $(C_A)_{ij}$ counts the number of common neighbors of $a_i \in A$ and $a_j \in A$. The simplest adjacency projection is the undirected projected network G_A defined by the weighted adjacency matrix C_A, with the weight $n_{ij} = (C_A)_{ij}$ or, equivalently,

$$n_{ij} = \sum_{k=1}^{m} M_{ik} M_{jk}. \tag{5.1}$$

The co-occurrence matrix C_B and the projected network of the B side can be defined similarly. Defining other types of weights (e.g. by using similarity measures, correlation coefficients [66]) leads to different types of adjacency projections. Just to mention a few, the *Jaccard similarity* [92] is defined as the fraction of the number of common neighbors of a_i and a_j and the number of nodes in their common neighborhood. That is,

$$s_{ij}^{J} = \frac{n_{ij}}{d_i + d_j - n_{ij}}. \tag{5.2}$$

Another frequently used similarity measure is one of the *collaborative filtering* methods, which is defined as

$$s_{ij}^{CF} = \frac{n_{ij}}{\min\{d_i, d_j\}}, \tag{5.3}$$

and it also referred as pairwise nestedness in the literature. The *Pearson correlation coefficient* can be also regarded as a similarity measure [175], which is defined by the formula

$$s_{ij}^{P} = \frac{n_{ij} - d_i d_j / m}{\sqrt{d_i(1 - d_i/m) d_j (1 - d_j/m)}}. \tag{5.4}$$

5.2 Statistically Validated Projections

When one constructs a projected network from the original bipartite network, the network heterogeneity (i.e. the heterogeneous degree distribution) of the original network makes it difficult to distinguish between (i) links whose presence in the projected network cannot be explained in terms of random co-occurrence of neighbors in the original network and (ii) links that are consistent with a random null hypothesis taking into account the heterogeneity of the bipartite network [176]. Roughly speaking, when one works with just a sample of a data set, the smaller the sample size, the higher the chance that it is not a good representative of the real data set and it has a random nature. Nevertheless, many real-world systems (viz. the data) are very noisy and/or the presence of many links in

the system can statistically be regarded as random and the adjacency projection methods may produce a significant distortion. To avoid this, some projection methods which use a filtering procedure via link validation have been developed [120, 165, 176]. The main idea is to verify whether a given (a possible) link in the projected network is consistent (or not) with a null hypothesis of random connectivity between its nodes and their neighborhood of the original bipartite network. If the answer is yes, then the link in the projection is not validated, and hence not drawn in the projection. If the null hypotheses is rejected, then the link is validated and drawn in the projection between the pair in question.

5.2.1 Hypotheses Testing

In order to validate statistically each link in the projected network, we can use the projection where two nodes a_i and a_j in A (and b_k and b_ℓ in B) are connected only if the number of neighbors that they share is not consistent with the null hypothesis of random co-occurrence of the common neighbors. To test this hypothesis, the one-side *hypergeometric test* is used. The null hypothesis is that nodes a_i and a_j are randomly connected to the elements of set B; namely, the probability that nodes a_i and a_j share exactly x neighbors in set B is given by the hypergeometric distribution,

$$H(x|m, d_i, d_j) = \frac{\binom{d_i}{x}\binom{m-d_i}{d_j-x}}{\binom{m}{d_j}}. \tag{5.5}$$

Then a p-value is assigned to each pair (a_i, a_j) like so

$$p_{ij} = 1 - \sum_{x=0}^{n_{ij}-1} H(x|m, d_i, d_j). \tag{5.6}$$

To reject the null hypothesis, usually a fixed level of significance is used; often it has a value of $\alpha = 0.01$ or $\alpha = 0.05$. If the p-value is less than or equal to the significance level α, it suggests that the observed data is inconsistent with the assumption that the null hypothesis is true and hence in this case the null-hypothesis is rejected. However, the hypothesis tests that incorrectly reject the null hypothesis (i.e make type I error(s)) are more likely to occur when one considers a set of statistical tests simultaneously. To try to avoid this, a multi-comparison test can be performed that associates a common level of significance to all links of the projected network. The most restrictive one is the *Bonferroni correction* [52], that is, to set $\alpha_B = \alpha/N_t = 0.01/N_t$, where N_t is the number of tests performed. Now, N_t could be the number of all possible pairs of nodes of the set A, i.e. $n(n-1)/2$, or the number of links of the adjacency projection. The Bonferroni correction minimizes the number of false positives (i.e. type I errors), but it often does not guarantee sufficient accuracy (it usually provides a large number of false negatives, i.e. type II errors). The FDR correction [15] reduces the number of false negatives by controlling the expected proportion of rejected null hypothesis without significantly expanding the number of false positives. The control of the FDR is calculated as follows: p-values from all the different N_t tests are first arranged in increasing order

$(p_1 < p_2 < ... < p_k < ... < p_{N_t})$, and then the null hypothesis is rejected for links until the p-value of rank k_{max} is such that $p_{k_{max}} < k_{max}\ \alpha_B$. Here, we will mostly use the statistical validation with FDR correction but results with the Bonferroni correction will also be applied. It should be mentioned here that when the Bonferroni correction does not provide any rejection, this is also the case for the FDR correction.

Comparing the different partitions

For our comparison we shall apply two widely used indicators. The first is the adjusted *Rand index* [157] and the second is an adjusted version of the *Wallace index* [181]. In other words, the comparison is made by considering adjusted versions of the *accuracy* and *precision* of the pairs of nodes observed in the given partition with respect to the reference partition.

Let X and Y be two partitions (into communities) of the same projected network. Here, we will use the following simple notations:

1. TP: True positives are the node pairs that are in the same community under X and Y.
2. FP: False positives are the pairs that are in different communities under X, but in the same cluster under Y.
3. TN: True negatives are the pairs that are in different communities under X and Y.
4. FN: False negatives are the pairs that are in the same communities under X and in different ones under Y.

As for accuracy, it is usually referred to as the Rand index in the case of graph clustering. This is the fraction of true results among the total number of cases examined. Namely,

$$\mathrm{RI} = \frac{TP + TN}{TP + FP + TN + FN}. \tag{5.7}$$

The Rand index varies between zero (absence of any accuracy in the given partition) and one (total accuracy in the partitioning). However, also in the presence of random partitioning a certain amount of accuracy may arise by chance. To take into account this possibility an adjusted version of the Rand index has been introduced [90]. It is defined as

$$\mathrm{ARI} = \frac{(TP + TN) - \mathbb{E}(TP + TN)}{(TP + FP + TN + FN) - \mathbb{E}(TP + TN)}, \tag{5.8}$$

where $\mathbb{E}(TP + TN)$ is the expected value of the true assessment between a random partition and the reference partition. For a random partition compared with another partition the value of ARI is on average close to zero. Negative values of the index describe cases where the membership of the two partitions is very different from that in a random case. By considering a set of N elements, and two partitions of these elements $X = \{X_1, X_2, \ldots, X_r\}$ and $Y = \{Y_1, Y_2, \ldots, Y_s\}$ and by defining n_{ij} as the number of

elements in common between partition X_i and Y_j, the Adjusted Rand index can also be written as

$$ARI = \frac{\sum_{i,j}\binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right]/\binom{N}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right]/\binom{N}{2}}, \tag{5.9}$$

where $a_i = \sum_j^s n_{ij}$ and $b_j = \sum_i^r n_{ij}$.

The precision of the pairwise classification is defined as

$$P = \frac{TP}{TP + FP}. \tag{5.10}$$

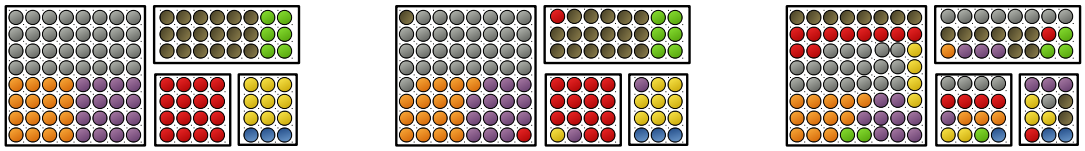
When two memberships are compared pairwise, the precision is usually referred to as one of the Wallace indices. Also for the case of the Wallace index, one can consider an adjusted version of it. Hereafter we provide the definition of an adjusted version of the Wallace index that we call the *Adjusted Wallace Index* (AWI)

$$AWI = \frac{TP - \mathbb{E}[TP]}{TP + FP - \mathbb{E}[TP]}, \tag{5.11}$$

where

$$\mathbb{E}[TP] = \frac{(TP + FP)(TP + FN)}{TP + FP + TN + FN}. \tag{5.12}$$

We note that $\mathbb{E}(TN)$ can similarly be defined to calculate Eq. 5.8. It is also worth mentioning that AWI varies between $-\infty$ and one. A value of one indicates that the partition obtained for a certain number of pairs is fully included in the reference partition. In Fig. 5.1, we provide an illustrative example. The correct partition is indicated by the different boxes, i.e. the system has four communities of different size. In each panel, different colors indicate an alternative partition relative to the reference one. In the example, the alternative partition has eight communities. In panel a) of the figure the communities of the partition are always contained in the communities of the reference partition and hence AWI is equal to one. In panel b) the communities shown by the color attributes are only partially contained in the reference partition. For example the red nodes are mainly in one box but two of them are associated with the largest and the second largest community, respectively. In this example the AWI is equal to 0.88, indicating a high but not perfect precision of the membership of pairs of nodes in the given partition. In panel c) the identified partition is quite different from the reference partition and almost all the boxes contain nodes of all colors. In this last case, AWI is close to zero, i.e. the value of the Wallace index is close to the one expected by a random null hypothesis.



(a) (b) (c)

Figure 5.1. (a): AWI = 1.0 (b): AWI = 0.88 (c): AWI = 0.03

5.2.2 Performance Evaluation on Benchmark Networks

Synthetic Benchmark Networks

We will explain how our approach works by considering certain synthetic networks. To be exact, we will generate bipartite networks with a well-defined community structure as follows. Let q be a fixed integer and $\{s_1^A, \ldots, s_q^A\}$ and $\{s_1^B, \ldots, s_q^B\}$ be the partitions of set A and B, having n and m nodes, respectively. For each set, the partitions are all of the same size (namely S_A and S_B), thus $n = qS_A$ and $m = qS_B$, respectively. We will evaluate the effectiveness of modularity optimization by considering the effect of missing or misclassified links on the artificial benchmark network.

Our synthetic network is first obtained by connecting nodes of set s_i^A to corresponding nodes (s_i^B) of set B with probability p_c. In this way with the parameter p_c we control the density of the links. This starting procedure leads to q disjoint bipartite components of the bipartite network (see panels a) and b) of Fig. 5.2) as an example with $q = 5$, $n = 25$, $m = 16$, and $p_c = 1$).

With the aim of modeling possible sources of randomness, or errors present in the original databases describing real systems, a second step in the generation of the artificial benchmark is to perturb the network by using the following procedure. Let us call p_r the probability that a link is misplaced due to some randomness or error. For each node i of set A with d_i links, a fraction $p_r d_i$ of links is selected and these links are randomly distributed to all possible nodes of set B, avoiding multiple links. The probability p_r therefore quantifyies the uncertainty added to the generated artificial benchmark. In the limiting case where $p_r = 0$ we go back to the original network, while in the opposite limit of $p_r = 1$ we get a completely random bipartite network that destroys the original community structure. In panel c) of Fig. 5.2, we show the artificial benchmark network characterized by $q = 5$, $S_A = 5$, $S_B = 16$, $p_c = 1$, and $p_r = 0.2$.

Results on the synthetic networks

We investigate the artificial network benchmark described above by performing community detection on the projected networks of it (typically on set A). Specifically, the community detection is performed on three benchmark networks. The first is the weighted projected network, referred to as the FULL network, connoting the fact that in this case we use all information available for all the actual links and their weights, obtained by

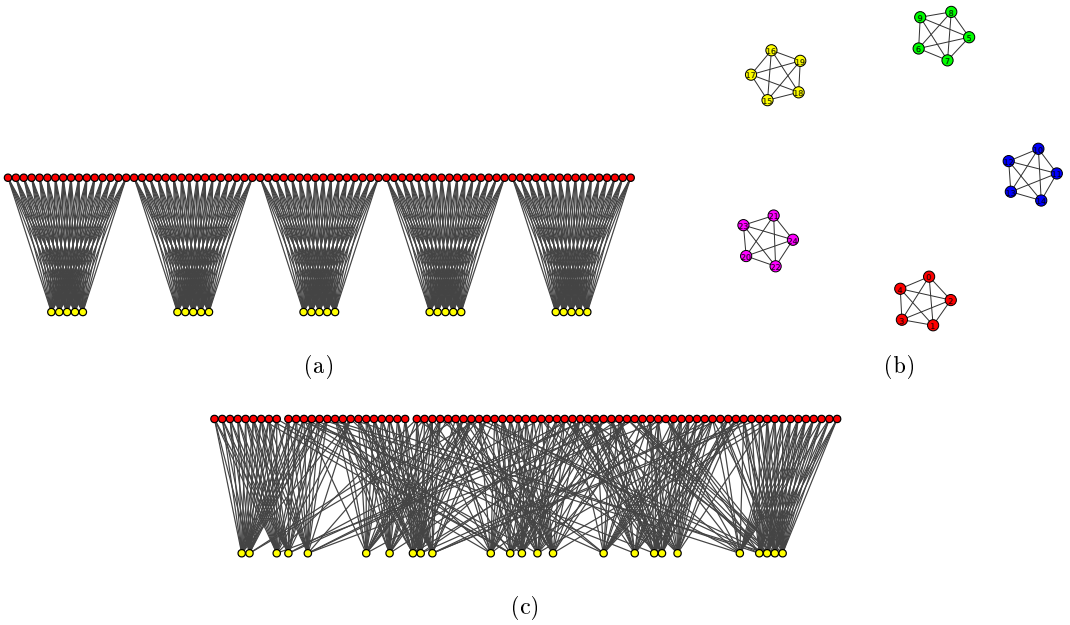
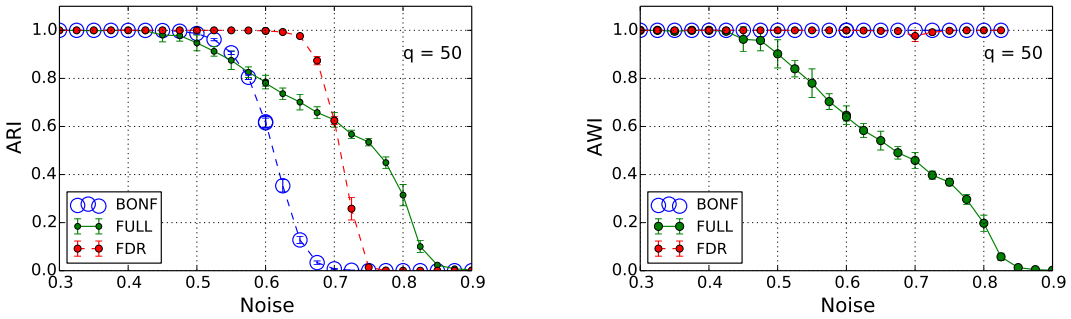


Figure 5.2. (a) network. (b) Adjacency projection of the benchmark. (c) Benchmark bipartite with $p_r = 0.2$.

starting from the bipartite network. The second one is a statistically validated version of the weighted network got with the procedure described in Section 5.2.1 when the multiple hypothesis test correction is the Bonferroni correction. We will call this network the Bonferroni network. The third one is a statistically validated version of the weighted network obtained with the control of the False Discovery Rate (FDR) correction. We will call this last type of network the FDR network.

For all three types of networks we performed a community detection by using modularity optimization. To be precise, we used the Louvain algorithm. To investigate the robustness of the partition obtained with this algorithm we repeated the community detection by using different starting sequences. With this approach the output of the Louvain algorithm is stochastic and different partitions can be obtained for close values of the modularity.

In Fig. 5.3, we have plotted the ARI and AWI values measured between the partition obtained by performing community detection of the three types of projected networks and the reference partition. The different settings of the benchmark were decided by choosing $S_A = 50$, $S_B = 50$, $p_c = 0.8$, $q = 50$ and several values of p_r ranging from 0.3 to 0.9 in steps of 0.025. In the top panel of Fig. 5.3, we plotted the ARI as a function of the probability of misplacement p_r of a link in the bipartite network. For the FULL network (green symbols), ARI is close to one for low values of p_r and starts to decrease for values of p_r greater than 0.4. ARI has values close to zero when p_r is greater than 0.9. The failure of the community detection procedure in detecting the correct membership is due to the fact that because of the misplacement of links, the algorithm is unable to detect all the communities of the reference partition and it merges some of them. A similar pattern of success is observed for the partitions obtained by SVNs. In fact, for the FDR network (red symbols) we can observe a value of ARI close to one for low values



(a) (b)

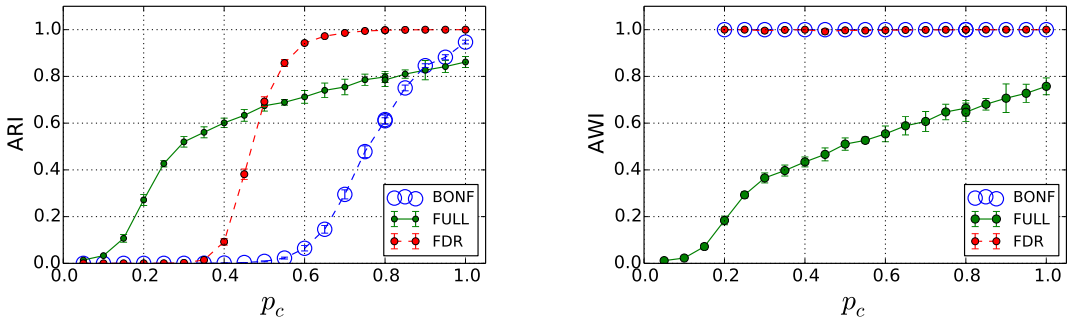
Figure 5.3. ARI and AWI measured between the partition obtained by performing community detection of the three type of projected networks (FULL (green symbols), FDR (red symbols) and Bonferroni (blue symbols) and the reference partition of the artificial benchmark. The benchmark was set by choosing $S_A = 50$, $S_B = 50$, $p_c = 0.8$, $q = 50$. Simulations and community detection are performed for several values of p_r ranging from 0.3 to 0.9 in steps of 0.025. Error bars were obtained by performing 10 realizations of the artificial benchmark

of p_r and it is close to zero for high values of p_r. It is worth noting that for the specific parameters of the benchmark there is an interval of p_r ($0.5 \leq p \leq 0.7$) where ARI of the FDR network is higher than the corresponding ARI value of the FULL network. The Bonferroni network has a similar pattern, but a decrease of ARI is seen for smaller values of p_r ($p_r \approx 0.5$). The reason for the decrease of the ARI for the FDR and the Bonferroni network is completely different from that for the full network. In fact for the partitions of these SVNs, ARI decreases because the statistical test loses power and the number of nodes present in them decreases as a function of p_r. This implies that the number of disconnected subgraphs (present in the SVNs and/or detected by the Louvain algorithm) increases, while the number of connected nodes decreases.

In the bottom panel of Fig. 5.3, AWI values for the three types of networks have been plotted. For the FULL network, the pattern of AWI is similar to the pattern of ARI. It starts very close to one and decreases to zero starting from $p_r \approx 0.4$. The behavior of AWI of the SVNs is quite different. In fact it remains very close to 1 until it abruptly reaches zero when the SVNs become empty, i.e. all the nodes are isolated. In other words, the precision of the classification of pairs of nodes is always high for SVNs and the problem they have in providing informative partitions is not precision but rather accuracy. All the partitions provided by them are statistically verified, but the level of accuracy progressively decreases in the presence of high levels of link misplacement.

So far we have investigated the role of the link misplacement in the detection of communities of the artificial benchmark. Another source of difficulty in community detection in real systems may originate from an insufficient coverage of the data. For this reason we evaluated the performance of our approach on artificial benchmarks characterized by a different level of link coverage. In Fig. 5.4, we have plotted ARI and AWI for simulations got by setting the same parameters used previously for $p_r = 0.6$ and different values of p_c ranging from 0 to 1 in steps of 0.05.

Panels (a) of Fig. 5.4 indicate that the ability of the community detection algorithm to detect the underlying benchmark decreases with decreasing p_c both for the FULL network



(a) (b)

Figure 5.4. Homogeneous Set. All simulations were obtained by setting $p_r = 0.6$ and $q = 20$. ARI values between the obtained partition and the underlying benchmark for (a) different values of p_c ranging from 0 to 1 in steps of 0.05, $S_A = S_B = 50$, (b) S_A ranging from 5 to 100 in step of 5. $S_B = 50$ and $p_c = 0.8$, and (c) S_B ranging from 0 to 1 in steps of 0.05, $S_A = 50$ and $p_c = 0.8$. In panels (d), (e) and (f) we have AWI values obtained using the same parameters as those for the corresponding ARI. The average value is obtained by performing ten different realizations. The error bar indicates one standard deviation.

and also for the SVNs. However, in this case the reason for this failure is also different for the two approaches. In the case of the FULL network the algorithm fails to detect the correct partition because it progressively merges several communities progressively when p_c decreases. Despite this, the major problem observed for the partitions got from SVNs is due to the fact that the accuracy of the statistical validation decreases for values of p_c lower than 0.7. Again panel (b) of Fig. 5.4 tells us that the problem is, however, not so much a problem of precision, as previously observed in our investigations, as a function of p_r.

In summary, both as a function of p_r and as a function of p_c the partitions observed with the approach of SVNs are partitions which are very precise in classifying the membership of pairs of nodes, although they might present a poor accuracy in the presence of high values of p_r or low values of p_c. The membership obtained by investigating the SVNs can therefore be viewed as statistically validated cores of the communities present in a given network.

5.2.3 A Case-study on Real Data

We will also investigate two widely studied real bipartite networks. The first is a the bipartite network of scholars and papers posted in the cond-mat archive [144]. The second is a classic bipartite network of actors and movies obtained by using information present in the International Movie Data Base (IMDB).

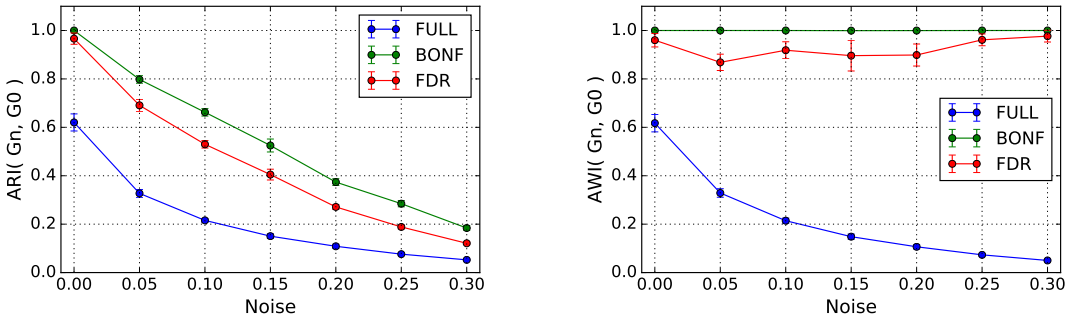
Co-authorship network

We will first investigate a co-authorship bipartite network. This bipartite network was constructed by Mark Newman based on preprints posted to the Condensed Matter section of arXiv E-Print Archive between 1995 and 1999. The dataset is available on the web page `https://toreopsahl.com/datasets/` and it consists in $16,726$ authors and $22,015$ papers. Our analysis was limited to the largest connected component of $13,861$ authors

and $19,466$ papers. We projected the bipartite network for the above set of authors. We also evaluated the FDR projected SVN. The FULL network has $44,619$ links and the FDR network has $7,768$ links. We performed community detection on them with the Louvain algorithm. For each network, the community detection was performed by applying the algorithm 1000 times with different initial conditions.

The 1000 partitions obtained for the FULL network have modularity ranging from 0.864 to 0.867. To investigate the degree of similarity among partitions of top values of modularity we selected partitions with modularity higher than the one of the 99 percentile of the 1000 outputs of Louvain algorithm. In particular, we selected 10 out 1000 partitions of highest modularity. We then estimated the ARI between all distinct pairs of these 10 partitions. These 45 pairs have an average mutual ARI of 0.65 with values ranging between a value of 0.59 (minimum) and 0.71 (maximum). As already noted in previous studies [81, 190], these partitions are quite different from each other in spite of the fact that the modularity of the partitions is almost identical (bounded within the interval $0.8666, 0.8670$). We got a quite different result when we considered the top 10 partitions obtained by performing community detection in the FDR SVN. In fact these 10 partitions are the same and the ARI among all of them is one. It worth noting that the FDR partition is not fully contained in any partition obtained from the FULL network. In fact, the interval of the AWI index is quite different from one and it covers a relatively limited interval of values $(0.57, 0.66)$.

By investigating the links and the communities that are obtained with SVNs, we can extract “cores" of the communities that are statistically robust. These “cores" are also quite stable with respect to errors that might be present in the database. To make this point explicit, we put some noise in the database by modifying it in a similar way to what we did with our artificial benchmark networks when we used values of p_r different from zero. In panel (a) of Fig. 5.5, we have plotted ARI values between the best partition of the FULL, that we label as $G0$, and the 100 best partitions obtained for values of p_r ranging from 0.05 to 0.3. In the same panel we also show the results of an analog investigation performed for the FDR SVN. The partitions obtained from FDR SVNs are always significantly more robust to noise that the ones obtained by performing community detection in the FULL network. In panel (b) of Fig. 5.5 we show the AWI for the same investigations. It is worth noting that similar to what we observed for the artificial benchmark networks, the cores of communities detected by investigating the FDR SVN show a decreasing similarity (i.e. ARI values) with the uncorrupted partition $G0$, not due to decrease of precision but rather due to decrease of accuracy. In fact, the AWI value of FDR does not go below 0.85 for all values of p_r, whereas we observed values of the AWI as low as 0.1 of the partitions obtained from the FULL network when $p_r = 0.3$. In other words, the informativeness of the detected cores of communities is robust with respect to noise added to the database. This behavior is similar to that observed for the artificial benchmark.



(a) (b)

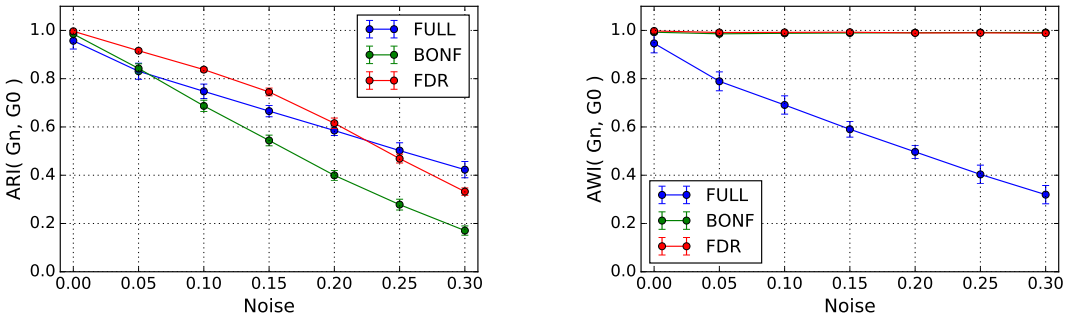
Figure 5.5. Co-authorship database. (a) The average ARI value between 100 partitions of the FULL network (blue symbols), the Bonferroni SVN (Green symbols), and FDR SVN (red symbols) obtained for different values of p_r and the uncorrupted best partition $G0$. Different partitions of high modularity are obtained with the Louvain algorithm using different initial conditions. (b) The average AWI of the same partitions.

IMDB

The second dataset we investigated was the classic bipartite system of actors and movies. We downloaded data about this system from the International Movie Data Base (IMDB) (`http://www.imdb.com/interfaces`). From the information given in the database we constructed several bipartite networks. A link between an actor and a movie is considered if the actor played in that movie, during a selected period of time. For our study we chose all movies present in the database during the time period from 1950 to 2015, with the exception of TV series, talk shows, animation films, short and adult movies.

An analysis for different periods of time was defined by a time-window of 5 years starting from 1950. Within each selected time interval, we constructed the bipartite network that lists movies released in that period and all the actors that played in these movies. As for the previous system, an analysis was performed on the largest connected component of the period in question. The bipartite networks were projected onto the movie side. The results of our investigations are summarized in Table 5.1 later on. Each row of the table refers to a different time period of investigation labeled by the first year of the chosen time period. The size of the investigated projected networks varied over time from the lowest value of 9,143 nodes and 686,398 links to the highest value 127,911 nodes and 1,487,598 links for the periods 1950-1954 and 2010-2014, respectively. The link density for the FULL projected network of movies varied from $1.82\ 10^{-4}$ (for 2010-2014) to $1.64\ 10^{-2}$ (for 1950-1954), i.e. in all cases the projected networks are quite sparse. The Bonferroni and FDR SVNs are significantly sparser than the FULL network. Actually, the percentage of the all links observed in them never exceeds 13.5 % for FDR and 2.6 % for Bonferroni SVNs (see the third and fourth columns of Table 5.1).

For each period of time and for the FULL, the Bonferroni, and the FDR SVNs we obtained 1000 output partitions using the Louvain algorithm with different initial conditions.To evaluate the differences observed between pairs of partitions obtained we computed the ARI among the 10 partitions of the 99 percentile of the 1000 best outputs. The average value of ARI is reported in the sixth, seventh, and eight column of Table 5.1 for



(a) (b)

Figure 5.6. IMDB database. Time period 1990-1994. (a) The average ARI value between 100 partitions of the FULL network (blue symbols), the Bonferroni SVN (green symbols), and FDR SVN (red symbols) obtained for different values of p_r and the uncorrupted best partition $G0$. Different partitions of high modularity are obtained with the Louvain algorithm using different initial conditions. (b) The average AWI of the same partitions.

the FULL, the Bonferroni, and the FDR networks, respectively. The values of ARI are always above 0.9 for all types of networks, suggesting that for this database the modularity optimization of the FULL network provides quite reliable results in most cases. In fact, values of the ARI lower than 0.97 are observed only for the last three time periods, suggesting that the reliability of the modularity optimization is very high for several time periods except the last three. The partitions obtained with the SVNs networks are rather stable for all time periods including the last three indicating that, for this database as well, SVNs detect cores of communities. This conclusion is also supported by the observed AWI values between the Bonferroni and the FULL network (ninth column of Table 5.1), and between the FDR and the FULL network (tenth column of Table 5.1). In both cases the AWI is very close to one for all time periods except the last three, when the modularity optimization of the FULL network becomes less reliable.

As for the IMDB bipartite networks of the period 1990-1994 we included noise in the bipartite network by modifying it in a similar way to that with our artificial benchmark networks and with the co-authorship database. In panel (a) of Fig. 5.6 we have plotted the average value of ARI between 100 partitions of the FULL network obtained for values of p_r ranging from 0.05 to 0.3 and the best partition $G0$ observed in the absence of noise. In the same panel we also show the results of an analogous investigation performed for the Bonferroni and FDR SVNs. The partitions obtained from FDR SVNs are for a large interval of p_r significantly more similar and therefore more robust to noise than those obtained by performing community detection in the FULL network. In panel (b) of Fig. 5.6, we have plotted the AWI values for the same investigations. Again the AWI value is close to one for the partitions of the SVNs, confirming once again the high degree of precision of the method in the detection of cores of communities. As for the previous cases, by combining the two measurements, we find that the decreasing values of the ARI with the uncorrupted partition $G0$ for the Bonferroni and the FDR SVNs are not due to a decrease in precision, but are rather it is due to a decrease in accuracy of the SVN method.

Table 5.1. Summary of IMDB investigations.

Time period	Nodes	Links	Bonf % of links	FDR % of links	AVG(ARI) Full	AVG(ARI) Bonf	AVG(ARI) FDR	AWI (Bonf,Full)	AWI (FDR,Full)
1950-54	9143	686398	1.4	8.2	0.992 (0.985,0.999)	0.994 (0.987,1.0)	0.92 (0.857,0.987)	1.00	0.98
1955-59	11253	519240	1.8	9.1	0.994 (0.984,1.0)	1.0 (1.0,1.0)	1.0 (1.0,1.0)	1.00	0.97
1960-64	12392	506639	1.9	10.7	0.997 (0.994,1.0)	1.0 (1.0,1.0)	0.998 (0.991,1.0)	1.00	0.97
1965-69	14782	633135	2.1	10.7	0.979 (0.958,0.995)	0.995 (0.989,1.0)	0.992 (0.979,1.0)	1.00	0.98
1970-74	15958	620634	2.2	11.1	0.982 (0.963,0.996)	0.993 (0.982,1.0)	0.978 (0.944,1.0)	0.99	0.97
1975-79	14996	522389	2.6	13.3	0.981 (0.971,0.995)	0.998 (0.997,1.0)	0.988 (0.971,0.998)	0.99	0.95
1980-84	15401	485082	2.5	13.5	0.991 (0.978,0.998)	1.0 (1.0,1.0)	0.984 (0.964,0.999)	1.00	0.95
1985-89	16846	569253	2.1	13.2	0.99 (0.983,0.997)	1.0 (1.0,1.0)	0.989 (0.974,1.0)	1.00	0.94
1990-94	17001	458604	1.9	10.2	0.977 (0.94,0.996)	0.998 (0.997,1.0)	0.984 (0.969,0.998)	0.99	0.98
1995-99	20311	402736	1.4	7.1	0.982 (0.974,0.989)	1.0 (1.0,1.0)	0.996 (0.982,1.0)	1.00	0.97
2000-04	31231	470828	1.4	7.2	0.961 (0.934,0.975)	1.0 (1.0,1.0)	0.96 (0.838,1.0)	0.98	0.93
2005-09	62496	788713	1.5	5.7	0.956 (0.937,0.969)	1.0 (1.0,1.0)	0.995 (0.986,1.0)	0.93	0.72
2010-14	127911	1487598	1.1	4.4	0.908 (0.859,0.96)	0.991 (0.984,1.0)	0.972 (0.933,0.997)	0.88	0.70

5.3 Rating and Ranking Nodes in Bipartite Networks

In the previous section we dealt with community detection in bipartite network. Now we will continue with the analysis of bipartite networks and in this section we will define a generalized version of the HITS algorithm that can be applied to weighted bipartite networks for rating and raking purposes. Although, it is a case-study, here we use this HITS based algorithm only for evaluating the quality of wine tasters, such rating and ranking methods are well suited for the investigation of user-item type rating databases that form the basis of recommendation systems, say.

5.3.1 A Generalized co-HITS Algorithm

Consider a bipartite graph $G = (A, B, E)$ where $A = \{a_1, a_2, \ldots, a_n\}$ and $B = \{b_1, b_2, \ldots, b_m\}$ are the two independent sets of n and m nodes and E is the set of edges. Now G is a weighted directed graph. Given $a_i \in A$ and $b_j \in B$, let $w(a_ib_j) > 0$ and $w(b_ja_i) > 0$ denote the weights of the directed edges (a_i, b_j) and (b_j, a_i), respectively; otherwise let $w(a_ib_j) = w(b_ja_i) = 0$. We assume, that the weights are normalized such that $\sum_{b_j \in Y} w(a_ib_j) = 1$ and $\sum_{a_i \in A} w(b_ja_i) = 1$ (this can be assumed without loss of generality, e.g let $w(a_ib_j) = w'(a_ib_j)/\sum_{j \in Y} w'(a_ib_j)$, where w' was the original weight of the link without normalization). The weight w can be viewed as the transition probability from a node in A (or in B) to a node in B (in A) of a random walk process. On the nodes of this bipartite graph a random walk can be naturally defined, where $\overrightarrow{W} = W(AB) = (w(a_ib_j))_{ij} \in \mathbb{R}^{n \times m}$ denotes the transition matrix from A to B and $\overleftarrow{W} = W(BA) = (w(b_ja_i))_{ji} \in \mathbb{R}^{m \times n}$ denotes the transition matrix from B to A. For the nodes on one side, a "hidden" transition probability $w(a_ia_k)$ from a_i to a_k can be defined as

$$w(a_ia_k) = \sum_{b_j \in B} w(a_ib_j)w(b_ja_k), \tag{5.13}$$

and using this definition $\sum_{a_k \in A} w(a_ia_k) = 1$ will also hold.
Note that $W_A = W(AA) = \overrightarrow{W}\overleftarrow{W} = w(a_ia_k)_{ik} \in \mathbb{R}^{n \times n}$ is a hidden transition probability matrix over A; the W_B matrix over B can be obtained in a similar way.

The *generalized co-HITS algorithm* can be applied on such directed weighted bipar-

tite graphs like that defined above. The algorithm assigns scores to the nodes of the graph via an iterative procedure as follows. Let p_i^0 and q_j^0 be the initial scores of the nodes a_i and b_j, respectively. The algorithm is described by the following recursion equations:

$$p_i = (1 - \lambda_A)p_i^0 + \lambda_A \sum_{b_j \in B} w(b_j a_i) q_j, \tag{5.14}$$

and

$$q_j = (1 - \lambda_B)q_j^0 + \lambda_B \sum_{a_i \in A} w(a_i b_j) p_i, \tag{5.15}$$

where $\lambda_A \in [0, 1]$ and $\lambda_B \in [0, 1]$ are real-valued parameters. By substituting Eq. 5.15 for q_j in Eq. 5.14 we see that

$$\begin{aligned} p_i &= (1 - \lambda_A)p_i^0 + \lambda_A(1 - \lambda_B) \sum_{b_j \in B} w(b_j a_i) q_j^0 + \\ &+ \lambda_A \lambda_B \sum_{a_k \in A} w(a_k a_i) p_k. \end{aligned} \tag{5.16}$$

It can be easily seen that the HITS algorithm, and the personalized PageRank algorithm [86] are just special cases of the Co-HITS algorithm. If $\lambda_A = \lambda_B = 1$, then Eq. 5.16 becomes

$$p_i = \sum_{a_k \in A} w(a_k a_i) p_k, \tag{5.17}$$

which is one part (e.g. for hubs) of the original HITS recursion. It is worth noting here, that this is the stationary state of the Markov chain defined by a random walk on the weighted graph defined above [147]. And if $\lambda_B = 1$, then

$$p_i = (1 - \lambda_A)p_i^0 + \lambda_A \sum_{a_k \in A} w(a_k a_i) p_k, \tag{5.18}$$

which is the recursion formula of the personalized PageRank algorithm.

5.3.2 A Case-study: Wines and Tasters

We investigated how the generalized co-HITS algorithm can be used to determine the quality of wine tasters. However, there are several methods available for evaluating the quality of wines, often by using the scores that a wine received in a wine tasting event, but it is still an open question that how to evaluate the competence and professional skills of the tasters, also mentioned in the article of Csendes and Antal [43]. Here we applied the generalized co-HITS algorithm for the datasets of two wine tasting events and compared the results with two simple statistical methods. The experimental results show that co-HITS algorithm produced promising results, and they seem to confirm our apriori knowledge about the tasters involved. Furthermore it proved to be more sophisticated than the statistical methods: both of them produced unreasonably large differences between the tasters and ranked those tasters too high who (perhaps due to their incom-

petence) gave the average of the scores of some other tasters for the wines.

Usually, wine tasting is a personal and subjective procedure for determining the quality of wines. Different wines are scored in an anonymous way called blind tasting (i.e. the tasters do not know which wine is being tasted). Each taster scores the wines she or he tasted and the wines are ranked according to these scores. Before we apply the co-HITS algorithm to provide a ranking of the tasters according to their competencies, we shall consider the following natural assumptions:

1. First of all, the wines are sorted by the points they received (i.e. there is no reference value for them).
2. Tasters are sorted by only considering the scores that the wines received from the tasters.
3. There is no cheater among the tasters (i.e. they score more or less on the "same scale").

Now, we will describe how the Co-HITS algorithm can be applied on the wine tasting data. Let A and B (defined previously) be the set of wine tasters and wines, respectively. We start from the same p_i^0 value for each $a_i \in A$ taster. Let $w'(a_ib_j)$ be the score that wine b_j obtained from taster a_i and let $w(a_ib_j) = w'(a_ib_j)/\sum_{b_j \in B} w'(a_ib_j)$ be its normalization. To be consistent to our first assumption, we define the q_j^0 value (for wine b_j) as the average of the scores that the wine received. Then, we define the weight $w(b_ja_i)$ in the following way. Let us suppose that wine b_j was tasted by ℓ different tasters and let us define

$$D(b_j) = \sum_{a_i \in A} \left| q_j^0 - w'(a_ib_j) \right|, \tag{5.19}$$

which is the sum of differences from the average score received by wine b_j. Finally, let

$$w(b_ja_i) = \frac{\left| D(b_j) - \left| q_j^0 - w'(a_ib_j) \right| \right|}{(\ell - 1)D(b_j)}. \tag{5.20}$$

Note that $\sum_{a_i \in A} w(b_ja_i) = 1$, hence each weight $w(b_ja_i)$ can be regarded as a transition probability from b_j to a_i. Fig. 5.7 shows an example for the calculation of the weights. The weight between two tasters a_i and a_k can be defined as the hidden transition probability defined by Eq. 5.13. Then the solution $p = (p_1, p_2, \ldots, p_m)$ of the HITS equation $p = W_A p$ provides the evaluation and ranking of the tasters.

We tested our model in the selected data of two wine tasting events. The first event was the Szeged Wine Fest in 2009, where 104 wines were blind tasted by four groups of five tasters. Each group tasted 33-34 different wines. The second dataset is a bit more specific: only red wines from the wine region of Villány were blind tasted by seven groups, each containing six tasters. Each group tasted 40-48 different wines. In both events, each wine was scored in accordance with the widely used and accepted international 100 point rating system.

We compared the results obtained by using the Co-HITS algorithm with two simple statistical methods which seems natural to use for our purpose. The first statistics-based

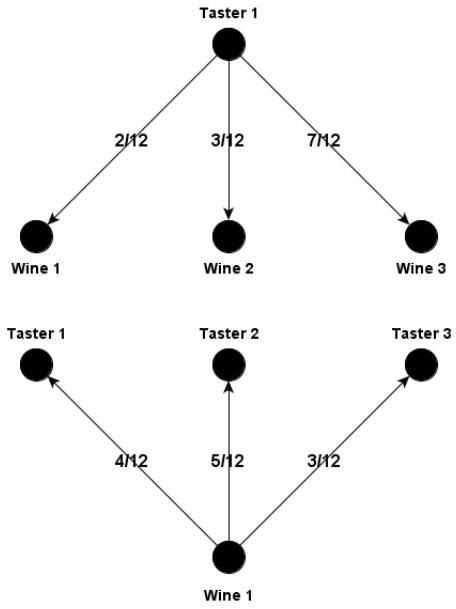


Figure 5.7. Weights of the graph when taster 1 assigns the scores 20, 30 and 70 for wine1, wine2, and wine3, respectively (up) and when wine 1 received the scores 20, 30, and 70 from taster 1, taster 2 and taster 3, respectively (down).

ranking method (SM1) evaluates the sum of differences, S_i, from the average score that each wine received for each taster a_i. Then, the tasters are ranked according to the increasing order of the S_i values. Formally,

$$S_i = \sum_{b_j \in B} |q_j^0 - w'(a_i b_j)|. \tag{5.21}$$

We consider the normalized points $(1 + \min_{a_i \in A} S_i)/(1 + S_i)$ for all a_i, (thus, the score of the taster with minimal S_i value will be 1).

The second statistical method (SM2) we used was the Pearson correlation coefficient between the scores that a taster asssigned to a wine and the average score that wine received. In other words, we are interested in how the scores of a taster correlate with the average scores of the wines received. The calculated values are normalized such that tasters with the highest correlation get 1. Table 5.2 and Table 5.3 show the detailed results obtained by applying the three different methods on the Szeged Wine Fest data and wine tasting data from the wine region of Villány, respectively. The calculated values can be interpreted as normalized merit values where the larger is the better. For each method the best taster of each group is shown in bold.

For better illustration, Fig. 5.8 shows the summarized results on the Szeged Wine Fest data. For each taster, the three different colored bars from the left to the right refers to the methods used for calculations, namely Co-HITS, SM1, and SM2, respectively. The results show that the Co-HITS algorithm produces more sophisticated results than SM1 and SM2. The stochastic process calculates closer values between the tasters. Consistent with this fact, much larger differences that the statistical methods produced can hardly be justified based on the concrete dataset. It should be mentioned that all three methods

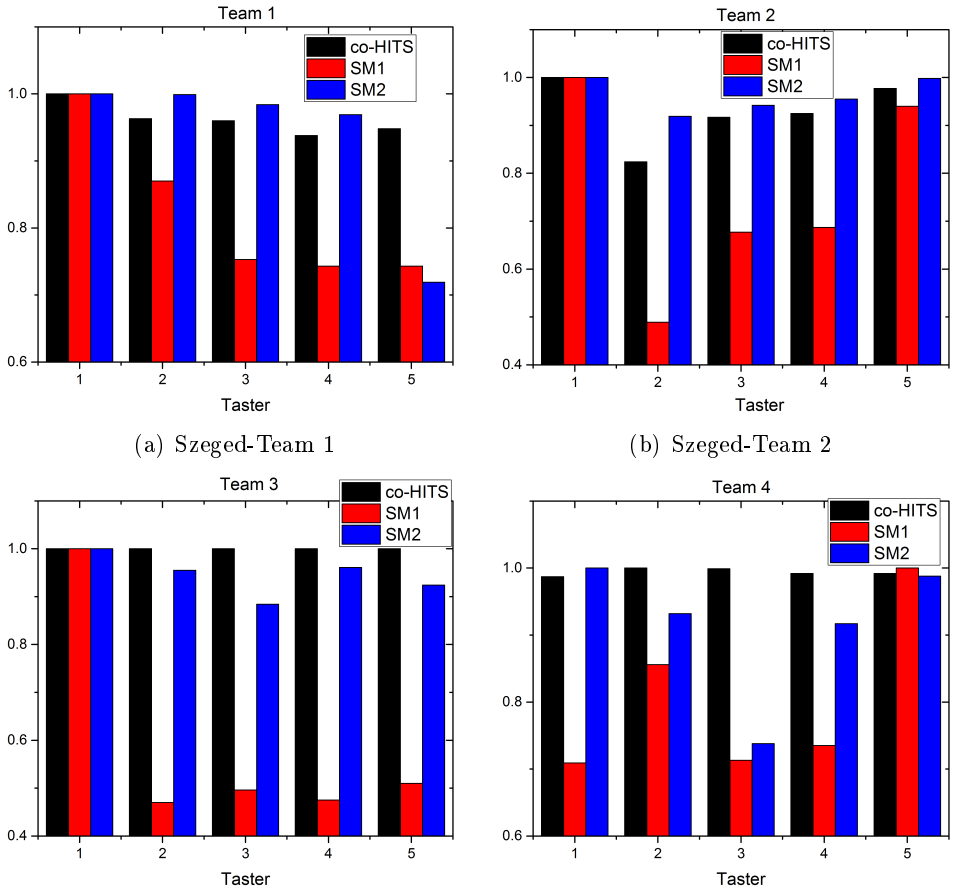
Table 5.2. Test results on the 2009 Szeged Wine Fest data

	Team 1			Team 2		
Taster	co-HITS	SM1	SM2	co-HITS	SM1	SM2
1	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
2	0.963	0.870	0.999	0.824	0.489	0.919
3	0.960	0.753	0.984	0.917	0.677	0.942
4	0.938	0.743	0.969	0.925	0.687	0.955
5	0.948	0.743	0.719	0.977	0.940	0.998
	Team 3			Team 4		
Taster	co-HITS	SM1	SM2	co-HITS	SM1	SM2
1	1.000	**1.000**	**1.000**	0.987	0.709	**1.000**
2	1.000	0.470	0.955	**1.000**	0.856	0.932
3	1.000	0.496	0.884	0.999	0.713	0.738
4	1.000	0.475	0.961	0.992	0.735	0.917
5	1.000	0.510	0.924	0.992	**1.000**	0.988

Table 5.3. Test results on the Villány data

	Team – co-HITS						
Taster	1	2	3	4	5	6	7
1	0.843	0.908	0.890	**1.000**	**1.000**	0.969	0.958
2	0.970	0.941	0.980	0.985	0.994	0.984	0.961
3	0.894	0.986	0.941	0.967	0.955	**1.000**	0.933
4	0.957	**1.000**	0.977	0.946	0.944	0.982	**1.000**
5	0.966	0.899	**1.000**	0.978	0.938	0.944	0.932
6	**1.000**	0.901	0.870	0.950	0.966	0.945	0.944
	Team – SM1						
Taster	1	2	3	4	5	6	7
1	0.491	0.556	0.495	0.991	0.901	0.891	0.746
2	0.779	0.672	0.932	**1.000**	**1.000**	0.932	0.760
3	0.478	0.794	0.625	0.872	0.839	**1.000**	0.870
4	0.638	**1.000**	0.892	0.665	0.644	0.919	**1.000**
5	0.781	0.613	**1.000**	0.781	0.651	0.822	0.705
6	**1.000**	0.492	0.505	0.856	0.829	0.739	0.678
	Team – SM2						
Taster	1	2	3	4	5	6	7
1	0.939	0.954	0.965	0.982	**1.000**	0.983	0.898
2	0.971	0.933	0.943	0.998	0.999	**1.000**	**1.000**
3	0.829	0.961	0.804	**1.000**	0.997	0.947	0.963
4	0.949	**1.000**	0.966	0.913	0.934	0.952	0.981
5	0.958	0.951	**1.000**	0.921	0.964	0.935	0.928
6	**1.000**	0.917	0.850	0.999	0.997	0.923	0.962

produced the same results for the best taster in many cases and the differences mostly appeared in the rest of the ranking lists. It can be observed that SM1 prefers the "closeness to the average" (due to its definition) and SM2 is better if the scores co-movement with the average is higher. It follows from these observations that both statistical methods can



(a) Szeged-Team 1 (b) Szeged-Team 2

(c) Szeged-Team 3 (d) Szeged-Team 4

Figure 5.8. Evaluation of the tasters by the different methods on the Szeged Wine Fest data

offer an opportunity for cheating, while the stochastic nature and iterative calculation of co-HITS scores is able the detect the outliers. The network-based algorithm considers the wine tasting data not only as a database that contains the scores of individual tasters, but also as a complex network that shows each tasters' relationship to one another. The relation between the tasters can be defined well for the purpose of this investigation. Therefore, the Co-HITS algorithm may give a better picture about the quality of tasters and as a byproduct it may give a better picture about the wines as well. Moreover, one of the main advantages of the graph-based method is that it also works on incomplete datasets, where not all the wines are tasted by a taster, or a taster tastes just a portion of all wines.

5.4 Summary

In this chapter, after introducing bipartite networks and some concepts for analyzing them, we showed that information present in a bipartite network can be used to detect cores of communities of each set defining the bipartite system. Simulation results revealed that the detected cores are highly stable and their detection is very precise although the methodology may, in same cases, be not so accurate. The cores of communities are found by considering statistically validated networks obtained by starting from the original

bipartite network. The information carried by these statistically validated network is highly informative and could be used to detect the membership of the investigated sets that are robust with respect to the algorithm of detection and to the presence of errors or missing entries in the database. The usefulness of the statistical validation approach can be assessed by a measure of similarity between pairs of partitions that are obtained by a stochastic community detection algorithm and that differ between them only for a tiny value of the function of the quality of a clustering. Here, we used the Adjusted Rand Index (ARI) and an adjusted version of the Wallace Index (AWI). In the presence of partitions characterized by very similar values of the quality function and presenting low values of ARI between them, one should consider it informative only on subsets of those partitions that are statistically stable. We suggest that in such cases investigations should focus on cores of the partitions obtained by performing community detection on SVNs. In this study, we considered an algorithm based on modularity optimization, but we think that our results are general and not strictly related to the chosen algorithm. They should be valid for any algorithm based on the maximization of a quality function.

Next, we defined a generalized version of the HITS algorithm that can be applied to weighted bipartite networks that, for instance, were obtained from user-item rating databases. However, as a case study, we used the HITS based algorithm to evaluate the quality of wine tasters, which may also be applicable in areas where people evaluate someone or something, such as sports that include figure skating, diving and synchronized swimming; social events that include singing contests and other tasting events such as a cooking competition or beer tasting. We observed that our ranking method performed well. It was able to filter out incompetent users, who, for example, gave the average score of the others for the items. Furthermore, our method can provide a clearer picture about the competence of users. In future work, we plan to refine the HITS algorithm for various applications: it would be interesting to use other modifications of HITS, and different rules for the weights of the network. We could analyze suitable null models and artificially generated data sets, and discuss the advantages and drawbacks of applying rating algorithms like this.

Chapter 6

Summary

Research on mining graph and network data has been continuously growing over the past few years, and it has become the most promising approach for extracting knowledge from relational data and investigating complex systems. It has become natural to represent such data and systems by means of graphs, where nodes stand for individuals or entities of the system, while edges represent the interaction or some relationship between pairs of these individuals or entities. Network theory, often combined with data mining tools, attempts to understand the origins and characteristics of networks that unify the components in various complex systems. This dissertation provides a summary of the author's work and results in the area of complex networks modeling and analysis. The main focus of this dissertation was to present general concepts of modeling with networks, network analysis and also present the author's results concentrating on the aim of extracting meaningful information from the modeled systems.

6.1 Characteristics of Real-world networks

The algorithms and methods developed and described in this thesis are defined on graphs that seek to model real-world complex systems. The first chapter of the dissertation provided a brief introduction to graph theory and an overview of the main definitions that characterize the structural properties of complex networks. We paid special attention to the community structure and core/periphery structure as global characteristics of complex networks and stochastic graph algorithms, namely PageRank and HITS since they are widely-used graph-based data mining tools.

6.2 Network Models for Some Real-life Problems

In Chapter 2 we presented several examples of real-world systems that can be modeled by networks. We highlighted the use graph-based data mining and network analysis as a first step to investigating such systems. Each case study explains, that after collecting appropriate data, how the network approach, especially concerning community detection and rating algorithms, can be used to extract meaningful information from the system we modeled. New methods are developed by slightly modifying some widely-used stochastic

graph algorithms. The results were published in a journals [44, 125] and conference proceedings [87, 124, 126]. The following paragraphs briefly summarize the main results of the chapter.

A local PageRank approximation with a case study

Although in many applications PageRank values need to be computed for all nodes of the graph, there are situations where we are interested in or capable of computing PageRank scores only for a small subset of the nodes. A local PageRank approximation method was developed based on the one proposed in [37] to assign "scientometrical" scores to research publications based on their local co-citation networks. We defined a "reaching probability" score for the same reason. As a case study, the local co-citation network of Egerváry's famous paper was examined and we saw that the network-based methods provided a more realistic picture of the importance of that paper than other scientometric indices.

Analyzing public transportation networks

Several network models were defined for a public transportation network (PTN) and a comprehensive analysis involving the PTN of five Hungarian cities was carried out. We were the first who performed a comprehensive network analysis (using modern network theoretic tools) of the public transportation systems of these cities. Our study examined directed and weighted edges, where the weight of a link referred to the morning peak hour capacity of the represented line, got by using the capacities of the vehicles (bus, tram, trolleybus) and schedules of the lines that go though that link. We compared the global and local characteristics of the networks and showed that they reveal a small-world feature (in terms of diameter and average path lengths) and scale-free distribution of various node centrality measures. We got a detailed picture of the differences in the organization of public transport, which may have arisen for historical, geographical and economic reasons. As a result, we highlighted some inconsistencies, organizational problems and identified which are the most sensitive routes and stations of the network justified by transportation engineers.

Introduce networks for educational data mining

We introduced a novel example of a real social system taken from the world of public education that can be modeled by networks. We proposed different network representations of relational educational data and mentioned several appropriate graph mining tools that could be used to analyze them. We discussed what kind of information could be extracted by their usage. Depending on the construction of the underlying graphs, we introduced four families of network models and performed a case study using one of them. With the intention of evaluating the achievements of students and generating a ranking among them, we defined a modified PageRank algorithm. We observed that the PageRank scores provide a fairly good relative order of the students with respect

to their achievements. Moreover, their progression can be monitored continuously using this method. Lastly, we pointed out several advantages of using graph-based data mining techniques in educational systems.

6.3 Network Models in Economics

In Chapter 3, we discussed various network models applied in economics and presented case studies including the analysis of the timely evolution of an international trade network and portfolio optimization using correlation-based financial networks. The results were published in a journal [140] and a conference proceeding [75] and another paper submitted to a journal [127]. The following paragraphs briefly summarize the main results of the chapter.

Case-study on a trade network

We demonstrated how network analysis could be applied to the trade networks of countries. As a case study, we investigated the timely evolution of the trade network of the European Union, focusing in particular on the evolution of communities and different trade rankings of the countries. We found in the EU that there is a core (with Germany, France and UK as leading economies) and a periphery (containing e.g. the former Comecon countries and the Balkans). In the trade network, peripheral countries are contained in the clusters of Russia and China, in contrast with the Western-European core countries that lie in clusters where the central nodes are Germany and the USA, respectively, highlighting real economic ties among he EU countries.

Financial networks and portfolio optimization

The question of quantifying the degree of statistical uncertainty (usually called "noise") presents in correlation-based financial systems was addressed. We applied different filtering techniques on the covariance matrix (in fact, on the correlation matrix obtained by normalization) to filter out the part of information that is robust against statistical uncertainty, and decrease the number of different elements in it. We used a Random Matrix Theory approach, and two versions of hierarchical clustering methods. Moreover, to determine the expected return of the assets we applied different statistical estimations. The methods were first applied to correlation matirces (networks) and then used for portfolio optimization. A large set of experiments revealed that using filtered correlation matrices, the classic Markowitz solution can be outperformed in terms of realized returns and reliability, which means that the realized risk and the estimated risk are closer to each other using filtering procedures.

6.4 Network Models and Linear Algebra for Rating and Prediction

In Chapter 4, the problem of rating and ranking sport players and teams was addressed from a network analysis perspective. A time-dependent PageRank method was designed to rate players with the graph defined using the game results data. Our algorithm was compared to several widely-used rating methods and it transpired that it provided a better ranking and predictive power in some situations. We also proposed a novel rating-based forecasting framework. The results were published in part in a journal [123] and some of them will be published. The following paragraphs briefly summarize the main results of the chapter.

Rating and Ranking in Sports

A novel ranking method which may be useful in sports like tennis, table tennis and American football, especially where players or teams play only a subset of opponents, was introduced and analyzed. In order to rank the players or teams, a time-dependent PageRank method was developed and applied on the directed and weighted graph representing players and game results in a sport competition. The method was tested on the results dataset of the table tennis competition of the researchers of the Institute of Informatics at the University of Szeged. The results obtained using our method were compared with several popular ranking techniques. We found that our approach worked well in general and that it had a good predictive power.

Forecasting in sports

We also proposed a novel rating-based forecasting framework. Against e.g. the well-known Bradley-Terry model, the main idea behind the model is that if a rating correctly reflects the actual relative performance of the teams in question, then the smaller the change in the rating vector, containing the rating of the teams, and after a certain event (e.g. win/loss) in an upcoming single game, the higher the probability will be that that event will occur. The results using a time-dependent PageRank rating method were compared to the Bradley-Terry predictions and the predictions of experts' betting odds based on their accuracy and predictive power. We found that our method outperforms the Bradley-Terry model in some cases, but in the future we would like to carry out a more systematic analysis of this.

6.5 Bipartite Network Models of Real-world Systems

In Chapter 5, we dealt with bipartite network models of complex systems. Such networks include, for instance, diseases-genes networks, plants-pollinators mutualistic networks, scientists-research papers cooperation networks and actors-movies networks. First of all, a methodology was presented in order to find the core of communities in bipartite networks and it was tested on synthetic benchmark networks and real bipartite systems. Then, we

discussed how a generalized version of PageRank and HITS algorithms could be defined for bipartite networks and, in a case study, we applied it on wine tasting datasets in order to rank tasters based on their ability and professional skills. The results were published in a journal [21] and appeared in conference proceeding [122]. The following paragraphs briefly summarize the main results of the chapter.

Statistical validation and the core of communities in bipartite networks

We demonstrated that information present in a bipartite network could be used to detect cores of communities of each set of the bipartite system being modeled. Using Monte-Carlo simulations, the results indicated that the cores found are very stable and detecting them is very precise although the methodology may not always be very accurate in a statistical sense. The key concept was to consider statistically validated networks obtained by starting from the original bipartite network. The identified communities of a given set are robust against the algorithm of detection and to the presence of errors or missing entries in the given database. Case studies on real data sets were also presented.

Rating nodes in bipartite networks

The question of rating nodes of a bipartite network was also addressed. A general framework of a HITS-type algorithm was presented for that purpose and a case study on a real data set was elaborated. We demonstrated that our method gives a clearer picture about the competence of wine tasters than other available statistical methods that can be readily applied here. Another important advantage of network-based methods is that not each wine should be rated by each taster to calculate the ratings. This allows us to use such methods for ranking users in a continuously evolving user-item rating database.

6.6. Összefoglaló

A hálózattudomány valamint a gráf alapú adatbányászat valós komplex rendszerek tanulmányozásának, illetve relációs adatokból való információ kinyerés fő eszközeivé váltak. Jelentőségük rendkívüli mértékben megnőtt az utóbbi két évtizedben, köszönhetően a rendelkezésre álló adatok robbanásszerű megnövekedésének, továbbá annak, hogy a gráf a matematikai modellezés egyik leghasznosabb eszköze. Komplex rendszerek és relációs adatok gráffal való modellezése - mely gráf csúcsai a rendszer entitásai, míg élei az entitás párok közti valamilyen kapcsolatot, illetve hasonlóságot reprezentálnak - kézenfekvővé vált. A hálózattudomány, adatbányászati eszközökkel kombinálva, valós rendszerek gráf modelljei szerkezetének és fejlődési dinamikájának tanulmányozását célozza. Ezen disszertáció a szerző munkájának és eredményeinek összefoglalása a hálózatos modellezés és hálózatkutatás területén. A fő hangsúlyt általános módszerek alkalmazására és új ötletek bemutatására helyeztük, melyek célja mindig a modellezett valós rendszer vizsgálata és információ kinyerési lehetőségek feltárása.

Valós hálózatok jellemzői

A dolgozatban bemutatott algoritmusok és módszerek gráfokon, illetve gráfokat leíró mátrixokon értelmezettek. Az első fejezet egy rövid bevezető, mely tárgyalja az gráfelméleti alapfogalmakat és áttekintést ad komplex hálózatok strukturális tulajdonságainak vizsgálatáról. Itt tárgyaltuk többek közt a hálózatok közösség szerkezete és mag-periféria szerkezete fogalmakat, mint globális tulajdonságok, illetve a PageRank és a HITS sztochasztikus gráf algoritmusokat, mint széles körben használt gráfos adatbányászati eszközöket.

Valós rendszerek hálózatos modelljei

A disszertáció második fejezetében különböző valós rendszerek hálózatos modelljeire láttunk példákat. Általános konklúziónk, hogy gráfos adatbányászati módszerek alkalmazása javasolt mintegy első lépés a gráfokkal modellezhető komplex rendszerek vizsgálatában, majd az elemzés eredményeinek segítségével megfogalmazható hipotézisek tesztelése mély statisztikai eszköztárral egy következő lépcsőfok lehet. Mindhárom bemutatott esettanulmány (hivatkozási hálózat, tömegközlekedési hálózatok, oktatási adatok vizsgálata) jól szemlélteti a gráf-bányászati elemzés legfontosabb lépéseit: releváns adatok összegyűjtése, a gráf modell(ek) megalkotása, globális tulajdonságok vizsgálata (mint például fokszám eloszlás és közösség szerkezet), illetve a gráf pontjainak és/vagy éleinek értékelése-rangsorolása. A elemzés segítségével információt kapunk a modellezett rendszerről, illetve hipotéziseket fogalmazhatunk meg működéséről, növekedéséről és egyes részeinek, entitásainak rendszerbeli szerepéről. Ismert gráf algoritmusok módosításával új módszereket fejlesztettünk és teszteltünk különböző problémákra. Az eredmények nemzetközi folyóiratokban [44, 125] és konferencia kiadványokban [87, 124, 126] jelentek meg. A következő pontokban összegezzük a fejezet fő eredményeit.

Lokális PageRank közelítő algoritmus bemutatása és egy esettanulmány

Számos alkalmazás esetén szükséges a gráfpontok PageRank értékének kiszámítása, ugyanakkor vannak olyan szituációk, amikor csak egy részgráf estén tudjuk és/vagy szeretnék kiszámítani azokat. Egy lokális PageRank közelítő algoritmust adtunk meg a [37]-ban javasolt módszer egy változataként, továbbá definiáltuk az „elérési valószínűség" értéket is, elsősorban tudományos publikációk értékelésének céljából. Esettanulmányunkban Egervári Jenő híres cikkének [53] hivatkozási környezetét vizsgáltuk a tárgyalt gráfos adatbányászati módszerekkel. Rámutattunk, hogy a hálózatos megközelítés objektívebb képet ad a mű fontosságáról, mint más tudománymetriai indexek.

Tömegközlekedési hálózatok elemzése

Több hálózatos modellt mutattunk be tömegközlekedési hálózatok vizsgálatára. Esettanulmányunkban öt magyar város tömegközlekedésének átfogó hálózatelemzését adtuk meg. Elsőként végeztünk átfogó összehasonlító hálózatelemzést (a modern hálózatkutatás eszközeivel) magyar városok tömegközlekedési hálózatain. A modellünkben gráf csúcsai a megállókat reprezentálják, élei irányítottak és súlyozottak: az irány a két csúcs közti járat irányát mutatja. A súlyokat pedig a közlekedő járatok reggeli csúcsidei kapacitásait és menetrendjét felhasználva definiáltuk. A hálózatok globális és lokális tulajdonságait hasonlítottuk össze, és láttuk, hogy a kisvilág tulajdonság megjelenik mind az úthossz, mind pedig egyes csúcs centralitási értékek eloszlása esetén. A hálózatok topológiája és egyes lokális tulajdonságai közti különbségek főleg történeti, földrajzi és gazdasági okokból fakadnak. Rá tudtunk világítani néhány inkonzisztenciára és szervezésbeli problematikára, továbbá közlekedés mérnökök által is megerősített érzékenynek tűnő csomópontokat és útvonalakat határoztunk meg.

Hálózatos modellek oktatási adatok vizsgálatához

Új példát mutattunk hálózattal modellezhető társadalmi rendszerre oktatási adatokat vizsgálva. Oktatási adminisztrációs rendszerekből kinyerhető adatok különböző gráfos reprezentációit mutattuk be és sorra vettük a legkézenfekvőbb hálózatelemzési és gráfbányászati lehetőségeket. Tárgyaltuk, hogy milyen típusú és mélységű információ nyerhető ki ezen módszerek segítségével. Négy modellcsaládot mutattunk be, az egyiket részletes esettanulmányban is vizsgáltuk rangsorolási céllal. A tanulók fejlődésének vizsgálatához és rangsorolásukhoz egy módosított PageRank algoritmust adtunk meg. Azt kaptuk, hogy a tanulók páronkénti összehasonlításával nyert hálózaton a a tanulókhoz rendelt PageRank értékek alapján egészen jó sorrendet tudunk felállítani a tanulók között a tanulmányi teljesítményükre vonatkozóan és folyamatosan követni tudjuk a tanulmányi fejlődésüket. Végül rámutattunk a gráfos adatbányászat használatának további előnyeire és lehetőségeire oktatási rendszerekben.

Hálózatos modellek a közgazdaságtanban

A harmadik fejezetben a közgazdaságtan területén alkalmazható hálózatos modelleket ismertettünk. Két esettanulmányt mutattunk be. Elsőként egy nemzetközi kereskedelmi hálózat időbeli fejlődését vizsgáltuk, majd korreláció alapú pénzügyi hálózatok alkalmazását néztük meg optimális részvényportfólió összeállítása céljából. Az eredmények hazai folyóiratban [140] és nemzetközi konferencia kiadványban jelentek meg [75], továbbá nemzetközi folyóiratban [127] kerülnek publikálásra. A következő pontokban összegezzük a fejezet fő eredményeit.

Az európai kereskedelmi hálózat vizsgálata

Bemutattuk hogyan alkalmazható a gráfos adatbányászat országok kereskedelmi hálózatának vizsgálatára. Esettanulmányunkban az Európai Unió országai és a gazdasági nagyhatalmak időben változó kereskedelmi hálózatát tanulmányozzuk, fókuszálva a közösségszerkezet változásaira és a kereskedelmi (import/export) rangsorok kialakulására. Megmutattuk, hogy a vizsgált hálózatok erős mag-periféria szerkezetet mutatnak. Közösség kereső eljárást alkalmazva láttuk, hogy a periférián lévő országok jellemzően az Oroszország, illetve Kína által fémjelzett klaszterekbe esnek. Ezzel szemben a magban lévő országok a német és amerikai központú klaszterekben helyezkednek el. A közösségszerkezet és gráfalgoritmusok által kereskedelmi rangsorok együttesen objektív képet adnak az EU országai közti gazdasági (függőségi) viszonyokra.

Pénzügyi hálózatok és portfólió optimalizálás

A fejezet második részében korreláció alapú pénzügyi hálózatokkal foglalkoztunk. Ezen hálózatok pontjai részvényeket reprezentálnak, két részvény között pedig az árfolyam idősoraik közti Pearson korrelációs együttható teremt kapcsolatot. Az így modellezett rendszerben jelen levő statisztikai bizonytalanság (melyet gyakran zajnak is hívnak) mérésének és szűrésének lehetőségeit tárgyaltuk. Különböző technikákat alkalmaztunk, hogy leválasszuk a mátrix azon részét, mely robusztus a statisztikai bizonytalansággal (véletlenszerűséggel) szemben, illetve, hogy csökkentsük a benne lévő elemek számát. A használt módszerek a véletlen mátrixok elméletén, illetve hierarchikus klaszterezési eljárásokon alapulnak. A szűrési eljárásokon túl a várható hozamok számításához is több statisztikát kipróbáltunk. A módszereket a Markowitz portfólió optimalizálási problémára alkalmaztuk, melynek célfüggvényében implicit módón jelenik meg a részvények közti korreláció alapú hálózat. Bootstrap szimulációs eredményeink azt mutatják, összhangban korábbi teszteredményekkel, hogy a klasszikus Markowitz megoldás javítható az elért hozamok, illetve a portfólió megbízhatósága, azaz a becsült és realizált kockázat eltérése tekintetében.

Értékelés és rangsorolás hálózatokban

A negyedik fejezetben sportolók és sportcsapatok értékelésén-rangsorolásán keresztül mutattunk be további hálózat alapú értékelési modelleket. Egy új, időfüggő PageRank modellt definiáltunk és alkalmaztunk meccsvégeredmény adatok által definiált, úgynevezett eredmény gráfokra. Szintén ebben a fejezetben egy új előrejelzésre használható modellt mutattunk be. Módszereinket több, széles-körben elterjedt eljárással hasonlítottuk össze. Az eredmények nemzetközi folyóiratban [123] jelentek meg, tovább egy részük később kerül publikálásra. A következő pontokban összegezzük a fejezet fő eredményeit.

Sportolók és sportcsapatok értékelése és rangsorolása

Egy új, elsősorban sportokra kifejlesztett használható rangsoroló módszert mutattunk be és vizsgáltunk valós sporteredmény adatokon. A játékosok, illetve csapatok rangsorolásához egy időfüggő PageRank módszert adunk meg és alkalmazunk a sportverseny eredményeit reprezentáló irányított és súlyozott gráfon. Az eljárást az SZTE Informatika Intézet belső asztalitenisz bajnokságának adatain teszteltük összehasonlítva számos elterjedt rangsoroló módszerrel Eredményeink azt mutatják, hogy módszerünk általánosan jól működik rangsorolási célra és jó predikciós erővel rendelkezik.

Sporteredmények előrejelzése

Bemutattunk egy új, hálózat alapú modellt sporteredmények előrejelzése céljából. Szemben a széles körben elterjedt Bradley-Terry féle páronkénti összehasonlításokon alapuló modellel, a mi módszerünk alapötlete az, hogy ha egy értékelő módszer pontosan tükrözi a csapatok közti aktuális erőviszonyokat, akkor egy következő mérkőzés egy adott kimenetele annál valószínűbb, minél kevésbé változtatja meg ezt a relatív erősorrendet. Egy időfüggő PageRank értékelőt használva hasonlítottuk össze eredményeinket a Bradley-Terry valószínűségekkel és a fogadási irodák által adott oddsok alapján számolt valószínűségekkel. Megmutattuk, hogy több esetben a módszerünk pontosabb és jobb predikciós erővel rendelkezik a Bradley-Terry modellnél, ugyanakkor megjegyezzük, hogy a modell alapos tanulmányozása jövőbeni kutatások tárgyát képezi.

Valós rendszerek páros gráf modelljei

Az ötödik fejezetben komplex rendszerek páros (szerencsésebb, de kevésbé elterjedt elnevezésben kétrészes) gráf modelljeivel foglalkoztunk. Ilyen hálózatok például a betegség-gén, növény-beporzó, kutató-publikáció, vagy a színész-film hálózatok. Elsőként egy módszertant mutattunk be kétrészes hálózatok közösségeinek, illetve közösségei magjának meghatározására. Ezután a PageRank és HITS algoritmusok egy általánosítását tárgyaltuk, majd bemutattuk egy lehetséges új alkalmazását borkóstolási adatokra. Az eredmények egy nemzetközi folyóiratban [21] kerülnek publikálására, illetve egy nemzetközi konferencia kiadványban [122] jelentek meg. A következő pontokban összegezzük a fejezet fő eredményeit.

Statisztikai validáció és közösségek magja páros gráfokban

Megmutattuk, hogy egy kétrészes hálózat szerkezetében rejlő információ használható közösségek magjának meghatározására, mindkét színosztályban. Monte-Carlo szimulációkkal kapott eredményeink mutatják, hogy ezen magok erősen stabilak és precízen megtalálhatók (kevés elsőfajú hiba), bár néhány esetben a módszer nem túl pontos (másodfajú hiba fellép). A kulcsötlet az, hogy az eredeti hálózatból kapott statisztikailag validált hálózatokat hozunk létre, majd azon végzünk közösségkeresést. A detektált közösségek robusztusak abban az értelemben, hogy nem függenek közösségkereső algoritmustól, illetve az hiányzó vagy hibás adatokra sem érzékenyek. Végül bemutattuk valós adatokon való vizsgálódásaink eredményeit is.

Csúcsok értékelése páros gráfokban

A dolgozat utolsó részében kétrészes hálózat pontjaink értékelési lehetőségeit vizsgáltuk. A HITS módszer páros gráfokra vett általánosítását mutattuk be és alkalmaztuk borversenyek adain borkostolók szakértelmének és értékeléseik konzisztenciájának meghatározása céljából. Megmutattuk, hogy, a priori tudásunk szerint, a módszer objektív képet mutat a kostolók hozzáértéséről szemben más természetes módon alkalmazható statisztikai elemzésekkel. Fontos előnye a hálózat alapú módszernek, hogy akkor is jól működik, amikor a kóstolók nem minden bort kóstolnak meg, így egy folyamatosan változó online kóstolási adatbázis esetén is képes lehet objektív rangsorolást adni a kóstolókról és a borokról is.

Bibliography

[1] A. Agresti. *Categorical data analysis*. John Wiley & Sons, New York, 1996.

[2] S. Alonso, F. J. Cabrerizo, E. Herrera-Viedma, and F. Herrera. H-index: A review focused in its variants, computation and standardization for different scientific fields. *Journal of Informetrics*, 3(4):273–289, 2009.

[3] T. Alzahrani, K. J. Horadam, and S. Boztas. Community detection in bipartite networks using random walks. In *Complex Networks V*, pages 157–165. Springer, 2014.

[4] M. R. Anderberg. *Cluster analysis for applications: probability and mathematical statistics: a series of monographs and textbooks*, volume 19. Academic press, 2014.

[5] I. Arribas, F. Perez, and E. Tortosa-Ausina. Measuring globalization of international trade: theory and evidence. *World Development*, 37(1):127–145, 2009.

[6] A. S. Asratian, T. M. Denley, and R. Häggkvist. *Bipartite graphs and their applications*. Cambridge University Press, 1998.

[7] Z. Bar-Yossef and L.-T. Mashiach. Local approximation of PageRank and reverse PageRank. In *Proceedings of the 17th conference on Information and knowledge management*, pages 279–288. ACM, 2008.

[8] A.-L. Barabási. The network takeover. *Nature Physics*, 8(1):14–16, 2012.

[9] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

[10] A.-L. Barabási and E. Bonabeau. Scale-free networks. *Scientific American*, 288(5):50–59, 2003.

[11] M. J. Barber. Modularity and community detection in bipartite networks. *Physical Review E*, 76(6):066102, 2007.

[12] D. Barrow, I. Drayer, P. Elliott, G. Gaut, and B. Osting. Ranking rankings: an empirical comparison of the predictive power of sports ranking methods. *Journal of Quantitative Analysis in Sports*, 9(2):187–202, 2013.

[13] J. Bascompte and P. Jordano. Plant-animal mutualistic networks: the architecture of biodiversity. *Annual Review of Ecology, Evolution, and Systematics*, pages 567–593, 2007.

[14] R. Bellman. Mathematical aspects of scheduling theory. *Journal of the Society for Industrial & Applied Mathematics*, 4(3):168–205, 1956.

[15] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, pages 289–300, 1995.

[16] C. T. Bergstrom, J. D. West, and M. A. Wiseman. The eigenfactor^{TM} metrics. *The Journal of Neuroscience*, 28(45):11433–11434, 2008.

[17] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008.

[18] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang. Complex networks: Structure and dynamics. *Physics Reports*, 424(4):175–308, 2006.

[19] V. Boginski, S. Butenko, and P. M. Pardalos. Matrix-based methods for college football rankings. *Economics, Management and Optimization in Sports*, pages 1–13, 2004.

[20] B. Bollobás and O. M. Riordan. Mathematical results on scale-free random graphs. *Handbook of graphs and networks: from the genome to the internet*, pages 1–34, 2003.

[21] C. Bongiorno, A. London, S. Miccichè, and R. N. Mantegna. Core of communities in bipartite networks (submitted to Physical Review E). *arXiv preprint arXiv:1704.01524*, 2017.

[22] S. P. Borgatti. Centrality and network flow. *Social networks*, 27(1):55–71, 2005.

[23] S. P. Borgatti and M. G. Everett. Models of core/periphery structures. *Social Networks*, 21(4):375–395, 2000.

[24] A. Bóta, M. Krész, and A. Pluhár. Dynamic communities and their detection. *Acta Cybernetica*, 20(1):35–52, 2011.

[25] R. A. Bradley and M. E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3-4):324–345, 1952.

[26] U. Brandes, D. Delling, M. Gaertler, R. Görke, M. Hoefer, Z. Nikoloski, and D. Wagner. On finding graph clusterings with maximum modularity. In *International Workshop on Graph-Theoretic Concepts in Computer Science*, pages 121–132. Springer, 2007.

[27] G. W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950.

[28] S. Brin and L. Page. Reprint of: The anatomy of a large-scale hypertextual web search engine. *Computer Networks*, 56(18):3825–3833, 2012.

[29] S. Bustos, C. Gomez, R. Hausmann, and C. A. Hidalgo. The dynamics of nestedness predicts the evolution of industrial ecosystems. *PloS One*, 7(11):e49393, 2012.

[30] K. Butler and J. T. Whelan. The existence of maximum likelihood estimates in the Bradley-Terry model and its extensions. *arXiv preprint math/0412232*, 2004.

[31] W.-S. Calaway, Rich and D. Tenenbaum. *Foreach Parallel Adaptor for the 'parallel' Package*, 2015. R package version 2.14.

[32] T. Callaghan, P. J. Mucha, and M. A. Porter. Random walker ranking for NCAA division IA football. *American Mathematical Monthly*, 114(9):761–777, 2007.

[33] F. Caron and A. Doucet. Efficient bayesian inference for generalized Bradley–Terry models. *Journal of Computational and Graphical Statistics*, 21(1):174–196, 2012.

[34] C. Chen. Visualising semantic spaces and author co-citation networks in digital libraries. *Information Processing & Management*, 35(3):401–420, 1999.

[35] P. Chen, H. Xie, S. Maslov, and S. Redner. Finding scientific gems with Google's PageRank algorithm. *Journal of Informetrics*, 1(1):8–15, 2007.

[36] W. Chen, J. Lu, and J. Liang. Research in disease-gene network based on bipartite network projection. *Complex System and Complexity Science*, 6(1):13–19, 2009.

[37] Y.-Y. Chen, Q. Gan, and T. Suel. Local methods for estimating PageRank values. In *Proceedings of the 13th International conference on Information and knowledge management*, pages 381–389. ACM, 2004.

[38] W. N. Colley. Colley's bias free college football ranking method: the Colley matrix explained. `http://www.colleyrankings.com/matrate.pdf`, 2002.

[39] T. Conlon, H. J. Ruskin, and M. Crane. Random matrix theory and fund of funds portfolio optimisation. *Physica A: Statistical Mechanics and its Applications*, 382(2):565–576, 2007.

[40] A. C. Constantinou, N. E. Fenton, and M. Neil. Pi-football: A bayesian network model for forecasting association football match outcomes. *Knowledge-Based Systems*, 36:322–339, 2012.

[41] B. Csaba and A. Pluhár. A weighted regularity lemma with applications. *International Journal of Combinatorics*, 2014.

[42] L. Csató. Ranking by pairwise comparisons for Swiss-system tournaments. *Central European Journal of Operations Research*, 21(4):783–803, 2013.

[43] T. Csendes and E. Antal. Pagerank based network algorithms for weighted graphs with applications to wine tasting and scientometrics. In *Proceedings of the 8th International Conference on Applied Informatics*, pages 209–216, 2010.

[44] P. Csermely, A. London, L.-Y. Wu, and B. Uzzi. Structure and dynamics of core/periphery networks. *Journal of Complex Networks*, 1(2):93–123, 2013.

[45] G. Dahl. A matrix-based ranking method with application to tennis. *Linear Algebra and its Applications*, 437(1):26–36, 2012.

[46] H. A. David. Ranking from unbalanced paired-comparison data. *Biometrika*, 74(2):432–436, 1987.

[47] R. R. Davidson and P. H. Farquhar. Bibliography on method of paired comparisons. *Biometrics*, 32(2):241–252, 1976.

[48] D. Delen, D. Cogdell, and N. Kasap. A comparative analysis of data mining methods in predicting NCAA bowl outcomes. *International Journal of Forecasting*, 28(2):543–552, 2012.

[49] H. Deng, M. Lyu, and I. King. A generalized co-hits algorithm and its application to bipartite graphs. In *Proceedings of the 15th SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 239–248. ACM, 2009.

[50] S. Derrible. Network centrality of metro systems. *PloS One*, 7(7):e40575, 2012.

[51] M. J. Dixon and P. F. Pope. The value of statistical forecasts in the UK association football betting market. *International Journal of Forecasting*, 20(4):697–711, 2004.

[52] O. J. Dunn. Multiple comparisons among means. *Journal of the American Statistical Association*, 56(293):52–64, 1961.

[53] J. Egerváry. Mátrixok kombinatorikus tulajdonságairól. *Matematikai és Fizikai Lapok*, 38:16–28, 1931.

[54] L. Egghe. An improvement of the h-index: The g-index. *ISSI Newsletter*, 2(1):8–9, 2006.

[55] M. El Alaoui. Random matrix theory and portfolio optimization in Moroccan stock exchange. *Physica A: Statistical Mechanics and its Applications*, 433:92–99, 2015.

[56] A. Éltető. Versenyképesség a közép-kelet-európai külkereskedelemben. *Közgazdasági Szemle (Economic Review), L évfolyam*, pages 269–281, 2003.

[57] E. J. Elton, M. J. Gruber, S. J. Brown, and W. N. Goetzmann. *Modern portfolio theory and investment analysis*. John Wiley & Sons, 2009.

[58] P. Erdős and A. Rényi. On random graphs I. *Publicationes Mathematicae (Debrecen)*, 6:290–297, 1959.

[59] P. Erdős and A. Rényi. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, 5(1):17–60, 1960.

[60] L. Ermann and D. L. Shepelyansky. Ecological analysis of world trade. *Physics Letters A*, 377(3):250–256, 2013.

[61] L. Ermann and D. L. Shepelyansky. Google matrix analysis of the multiproduct world trade network. *The European Physical Journal B*, 88(4):84, 2015.

[62] M. G. Everett and S. P. Borgatti. The dual-projection approach for two-mode networks. *Social Networks*, 35(2):204–210, 2013.

[63] G. Fagiolo, J. Reyes, and S. Schiavo. World-trade web: Topological properties, dynamics, and evolution. *Physical Review E*, 79(3):036115, 2009.

[64] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From data mining to knowledge discovery in databases. *AI Magazine*, 17(3):37, 1996.

[65] D. Fiala, F. Rousselot, and K. Ježek. Pagerank for bibliographic networks. *Scientometrics*, 76(1):135–158, 2008.

[66] A. Fiasconaro, M. Tumminello, V. Nicosia, V. Latora, and R. Mantegna. Hybrid recommendation methods in complex networks. *arXiv preprint arXiv:1412.3697*, 2014.

[67] L. R. Ford and D. Fulkerson. Solving the transportation problem. *Management Science*, 3(1):24–32, 1956.

[68] D. Forrest, J. Goddard, and R. Simmons. Odds-setters as forecasters: The case of English football. *International Journal of Forecasting*, 21(3):551–564, 2005.

[69] D. Forrest and R. Simmons. Forecasting sport: the behaviour and performance of football tipsters. *International Journal of Forecasting*, 16(3):317–331, 2000.

[70] S. Fortunato. Community detection in graphs. *Physics Reports*, 486(3):75–174, 2010.

[71] S. Fortunato and C. Castellano. Community structure in graphs. In *Computational Complexity*, pages 490–512. Springer, 2012.

[72] M. Franceschet and E. Bozzo. The Massey's method for sport rating: a network science perspective. *arXiv preprint arXiv:1701.03363*, 2017.

[73] L. C. Freeman. Centrality in social networks conceptual clarification. *Social Networks*, 1(3):215–239, 1979.

[74] D. Garlaschelli and M. I. Loffredo. Structure and evolution of the world trade network. *Physica A: Statistical Mechanics and its Applications*, 355(1):138–144, 2005.

[75] I. Gera, B. Bánhelyi, and A. London. Testing the Markowitz portfolio optimization method with filtered correlation matrices. In *Proceedings of the Middle-European Conference on Applied Theoretical Computer Science*, pages 44–47, 2016.

[76] A. Ghalanos and S. Theussl. *'Rsolnp': General Non-linear Optimization Using Augmented Lagrange Multiplier Method*, 2015. R package version 1.16.

[77] R. Gill and J. Keating. Assessing methods for college football rankings. *Journal of Quantitative Analysis in Sports*, 5(2), 2009.

[78] M. Girvan and M. E. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826, 2002.

[79] D. F. Gleich. Pagerank beyond the web. *SIAM Review*, 57(3):321–363, 2015.

[80] J. Goddard and I. Asimakopoulos. Forecasting football results and the efficiency of fixed-odds betting. *Journal of Forecasting*, 23(1):51–66, 2004.

[81] B. H. Good, Y.-A. de Montjoye, and A. Clauset. Performance of modularity maximization in practical contexts. *Physical Review E*, 81(4):046106, 2010.

[82] A. Y. Govan. Ranking theory with application to popular sports. *PhD dissertation, North Carolina State University, Raleigh, North Carolina*, 2008.

[83] A. Y. Govan, A. N. Langville, and C. D. Meyer. Offense-defense approach to ranking team sports. *Journal of Quantitative Analysis in Sports*, 5(1):1–19, 2009.

[84] T. Guhr and B. Kälber. A new method to estimate the noise in financial correlation matrices. *Journal of Physics A: Mathematical and General*, 36(12):3009, 2003.

[85] R. Guimerà, M. Sales-Pardo, and L. A. N. Amaral. Module identification in bipartite and directed networks. *Physical Review E*, 76(3):036102, 2007.

[86] T. Haveliwala, S. Kamvar, and G. Jeh. An analytical comparison of approaches to personalizing pagerank. Technical report, Stanford University, 2003.

[87] A. Háznagy, I. Fi, A. London, and T. Németh. Complex network analysis of public transportation networks: a comprehensive study. In *4th International Conference on Models and Technologies for Intelligent Transportation Systems*, pages 371–378. IEEE, 2015.

[88] C. Heiner, N. Heffernan, and T. Barnes. Educational data mining. In *Supplementary Proceedings of the 12th International Conference of Artificial Intelligence in Education*, 2007.

[89] J. E. Hirsch. An index to quantify an individual's scientific research output. *Proceedings of the National academy of Sciences*, pages 16569–16572, 2005.

[90] L. Hubert and P. Arabie. Comparing partitions. *Journal of Classification*, 2(1):193–218, 1985.

[91] D. Hummels. Transportation costs and international trade in the second era of globalization. *The Journal of Economic Perspectives*, 21(3):131–154, 2007.

[92] P. Jaccard. A comparative study of the floral distribution in alps and jura. *Bull. Walden Soc. Nat. Sci*, 37:547–579, 1901.

[93] M. O. Jackson. *Social and economic networks*. Princeton University Press, 2010.

[94] M. O. Jackson, B. Rogers, and Y. Zenou. Networks: An economic perspective. *Oxford Handbook of Social Network Analysis, R. Light and J. Moody (Eds.)*, 2016.

[95] W. James and C. Stein. Estimation with quadratic loss. In *Proceedings of the 4th Berkeley symposium on mathematical statistics and probability*, volume 1, pages 361–379, 1961.

[96] T. Jech. The ranking of incomplete tournaments: A mathematician's guide to popular sports. *The American Mathematical Monthly*, 90(4):pp. 246–264+265–266, 1983.

[97] H. Jeong, Z. Néda, and A.-L. Barabási. Measuring preferential attachment in evolving networks. *Europhysics Letters*, 61(4):567, 2003.

[98] A. Joseph, N. E. Fenton, and M. Neil. Predicting football results using bayesian nets and other machine learning techniques. *Knowledge-Based Systems*, 19(7):544–553, 2006.

[99] D. Karlis and I. Ntzoufras. Analysis of sports data by using bivariate Poisson models. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 52(3):381–393, 2003.

[100] J. P. Keener. The Perron-Frobenius theorem and the ranking of football teams. *SIAM Review*, 35(1):80–93, 1993.

[101] M. G. Kendall. A new measure of rank correlation. *Biometrika*, 30:81–93, 1938.

[102] M. G. Kendall and B. Babington Smith. On the method of paired comparisons. *Biometrika*, 31(3/4):324–345, 1940.

[103] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.

[104] D. König. Über graphen und ihre anwendung auf determinantentheorie und mengenlehre. *Mathematische Annalen*, 77(4):453–465, 1916.

[105] P. Krugman, M. Obstfeld, and M. Melitz. *International Economics: Theory and Policy*. Addison-Wesley, 2011.

[106] H. W. Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97, 1955.

[107] M. J. Kumar. Editorial: Evaluating scientists: Citations, impact factor, h-index, online page hits and what else? *IETE Technical Review*, 26(3):165–168, 2009.

[108] L. Laloux, P. Cizeau, J.-P. Bouchaud, and M. Potters. Noise dressing of financial correlation matrices. *Physical Review Letters*, 83(7):1467, 1999.

[109] L. Laloux, P. Cizeau, M. Potters, and J.-P. Bouchaud. Random matrix theory and financial correlations. *International Journal of Theoretical and Applied Finance*, 3(03):391–397, 2000.

[110] A. N. Langville and C. D. Meyer. *Google's PageRank and beyond: The science of search engine rankings*. Princeton University Press, 2011.

[111] A. N. Langville and C. D. Meyer. *Who's #1?: the science of rating and ranking*. Princeton University Press, 2012.

[112] D. B. Larremore, A. Clauset, and A. Z. Jacobs. Efficiently inferring community structure in bipartite networks. *Physical Review E*, 90(1):012805, 2014.

[113] J. Lasek, Z. Szlávik, and S. Bhulai. The predictive power of ranking systems in association football. *International Journal of Applied Pattern Recognition*, 1(1):27–46, 2013.

[114] M. Latapy, C. Magnien, and N. Del Vecchio. Basic notions for the analysis of large two-mode networks. *Social Networks*, 30(1):31–48, 2008.

[115] V. Latora and M. Marchiori. Is the Boston subway a small-world network? *Physica A: Statistical Mechanics and its Applications*, 314(1):109–113, 2002.

[116] S. Lehmann, B. Lautrup, and A. Jackson. Citation networks in high energy physics. *Physical Review E*, 68(2):026113, 2003.

[117] C. K. Leung and K. W. Joseph. Sports data mining: Predicting results for the college football games. *Procedia Computer Science*, 35:710–719, 2014.

[118] M. Li, J. Wang, X. Chen, H. Wang, and Y. Pan. A local average connectivity-based method for identifying essential proteins from the network level. *Computational Biology and Chemistry*, 35(3):143–150, 2011.

[119] G. Lianxiong, W. Jianping, and R. Liu. Key nodes mining in transport networks based in pagerank algorithm. In *Control and Decision Conference, 2009. CCDC'09. Chinese*, pages 4413–4416. IEEE, 2009.

[120] G. Linden, B. Smith, and J. York. Amazon. com recommendations: Item-to-item collaborative filtering. *IEEE Internet Computing*, 7(1):76–80, 2003.

[121] X. Liu, J. Bollen, M. L. Nelson, and H. Van de Sompel. Co-authorship networks in the digital library research community. *Information Processing & Management*, 41(6):1462–1480, 2005.

[122] A. London and T. Csendes. Hits based network algorithm for evaluating the professional skills of wine tasters. In *8th International Symposium on Applied Computational Intelligence and Informatics*, pages 197–200. IEEE, 2013.

[123] A. London, J. Németh, and T. Németh. Time-dependent network algorithm for ranking in sports. *Acta Cybernetica*, 21(3):495–506, 2014.

[124] A. London and T. Németh. Student evaluation by graph based data mining of administrational systems of education. In *Proceedings of the 15th International Conference on Computer Systems and Technologies*, pages 363–369. ACM, 2014.

[125] A. London, T. Németh, A. Pluhár, and T. Csendes. A local pagerank algorithm for evaluating the importance of scientific articles. *Annales Mathematicae et Informaticae*, 44:131–141, 2015.

[126] A. London, Á. Pelyhe, C. Holló, and T. Németh. Applying graph-based data mining concepts to the educational sphere. In *Proceedings of the 16th International Conference on Computer Systems and Technologies*, pages 358–365. ACM, 2015.

[127] I. London, András and and B. Bánhelyi. Testing portfolio selection models using various estimators of expected returns and filtering techniques for correlation matrices (submitted). 2017.

[128] L. Lovász. Perfect graphs. *Selected topics in graph theory*, 2:55–87, 1983.

[129] S. Luckner, J. Schröder, and C. Slamka. On the forecast accuracy of sports prediction markets. In *Negotiation, Auctions, and Market Engineering*, pages 227–234. Springer, 2008.

[130] M. J. Maher. Modelling association football scores. *Statistica Neerlandica*, 36(3):109–118, 1982.

[131] Y. Malevergne and D. Sornette. Collective origin of the coexistence of apparent random matrix theory noise and of factors in large sample correlation matrices. *Physica A: Statistical Mechanics and its Applications*, 331(3):660–668, 2004.

[132] R. N. Mantegna. Hierarchical structure in financial markets. *The European Physical Journal B-Condensed Matter and Complex Systems*, 11(1):193–197, 1999.

[133] R. N. Mantegna and H. E. Stanley. *Introduction to econophysics: correlations and complexity in finance*. Cambridge University Press, 1999.

[134] H. Markowitz. Portfolio selection: Efficient diversification of investments. Cowles foundation monograph no. 16, 1959.

[135] L. Marotta, S. Miccichè, Y. Fujiwara, H. Iyetomi, H. Aoyama, M. Gallegati, and R. N. Mantegna. Bank-firm credit network in Japan: An analysis of a bipartite network. *PLoS One*, 10(5):e0123079, 2015.

[136] K. Massey. Statistical models applied to the rating of sports teams. *Bluefield College (Master thesis)*, 1997.

[137] M. L. Mehta. *Random matrices*. Academic Press, 2004.

[138] D. Melamed. Community structures in bipartite networks: A dual-projection approach. *PLoS One*, 9(5):e97823, 05 2014.

[139] D. Melamed, R. L. Breiger, and A. J. West. Community structure in multi-mode networks: Applying an eigenspectrum approach. *Official Journal of the International Network for Social Network Analysts*, 33:18–23, 2013.

[140] Á. Merza, A. London, I. M. Kiss, A. Pelle, J. Dombi, and T. Németh. On the possible use of network science in the analysis of world trade (in Hungarian). *Közgazdasagi Szemle (Economic Review) LXIII. évfolyam*, pages 79–98, 2016.

[141] C. D. Meyer. *Matrix analysis and applied linear algebra*, volume 2. SIAM, 2000.

[142] S. Motegi and N. Masuda. A network-based dynamical ranking system for competitive sports. *Scientific Reports*, 2:904, 2012.

[143] N. Mukai. Pagerank-based traffic simulation using taxi probe data. *Procedia Computer Science*, 22:1156–1163, 2013.

[144] M. E. Newman. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences*, 98(2):404–409, 2001.

[145] M. E. Newman. The structure and function of complex networks. *SIAM review*, 45(2):167–256, 2003.

[146] M. E. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical Review E*, 69(2):026113, 2004.

[147] J. R. Norris. *Markov chains*. Cambridge University Press, 1998.

[148] G. Palla, I. Derényi, I. Farkas, and T. Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043):814–818, 2005.

[149] M. Patel, J. A. Bullinaria, and J. P. Levy. Extracting semantic representations from large text corpora. In *4th Neural Computation and Psychology Workshop, London, 9–11 April 1997*, pages 199–212. Springer, 1998.

[150] T. P. Peixoto. Parsimonious module inference in large networks. *Physical Review Letters*, 110(14):148701, 2013.

[151] C. Piccardi and L. Tajoli. Existence and significance of communities in the world trade web. *Physical Review E*, 85(6):066119, 2012.

[152] F. Picciolo, T. Squartini, F. Ruzzenenti, R. Basosi, and D. Garlaschelli. The role of distances in the world trade web. In *8th International Conference on Signal Image Technology and Internet Based Systems (SITIS)*, pages 784–792. IEEE, 2012.

[153] V. Plerou, P. Gopikrishnan, B. Rosenow, L. A. N. Amaral, and H. E. Stanley. Universal and nonuniversal properties of cross correlations in financial time series. *Physical Review Letters*, 83(7):1471, 1999.

[154] P. F. Pope and D. A. Peel. Information, prices and efficiency in a fixed-odds betting market. *Economica*, pages 323–341, 1989.

[155] F. Radicchi. Who is the best player ever? A complex network analysis of the history of professional tennis. *PloS One*, 6(2):e17249, 2011.

[156] F. Radicchi, S. Fortunato, B. Markines, and A. Vespignani. Diffusion of scientific credits and the ranking of scientists. *Physical Review E*, 80(5):056103, 2009.

[157] W. M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850, 1971.

[158] P. Rao and L. L. Kupper. Ties in paired-comparison experiments: A generalization of the Bradley-Terry model. *Journal of the American Statistical Association*, 62(317):194–204, 1967.

[159] C. Romero and S. Ventura. Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications*, 33(1):135–146, 2007.

[160] C. Romero, S. Ventura, M. Pechenizkiy, and R. S. Baker. *Handbook of educational data mining*. CRC Press, 2010.

[161] B. Rosenow, V. Plerou, P. Gopikrishnan, and H. E. Stanley. Portfolio optimization and the random magnet problem. *Europhysics Letters*, 59(4):500, 2002.

[162] S. M. Ross. *Introduction to probability models*. Academic Press, Ninth edition, 2007.

[163] C. Roth, S. M. Kang, M. Batty, and M. Barthelemy. A long-time limit for world subway networks. *Journal of The Royal Society Interface*, page rsif20120259, 2012.

[164] G. Sabidussi. The centrality index of a graph. *Psychometrika*, 31(4):581–603, 1966.

[165] F. Saracco, R. Di Clemente, A. Gabrielli, and T. Squartini. Grandcanonical projection of bipartite networks. *arXiv preprint arXiv:1607.02481*, 2016.

[166] O. Scheuer and B. M. McLaren. Educational data mining. In *Encyclopedia of the Sciences of Learning*, pages 1075–1079. Springer, 2012.

[167] A. M. Sengupta and P. P. Mitra. Distributions of singular values for some random matrices. *Physical Review E*, 60(3):3389, 1999.

[168] M. A. Serrano and M. Boguná. Topology of the world trade web. *Physical Review E*, 68(1):015101, 2003.

[169] M. Spann and B. Skiera. Sports forecasting: a comparison of the forecast accuracy of prediction markets, betting odds and tipsters. *Journal of Forecasting*, 28(1):55, 2009.

[170] T. Squartini and D. Garlaschelli. Stationarity, non-stationarity and early warning signals in economic networks. *Journal of Complex Networks*, 3(1):1, 2015.

[171] C. Su, Y. Pan, Y. Zhen, Z. Ma, J. Yuan, H. Guo, Z. Yu, C. Ma, and Y. Wu. Prestigerank: A new evaluation method for papers and journals. *Journal of Informetrics*, 5(1):1–13, 2011.

[172] V. Tola, F. Lillo, M. Gallegati, and R. N. Mantegna. Cluster analysis for portfolio optimization. *Journal of Economic Dynamics and Control*, 32(1):235–258, 2008.

[173] J. A. Trono. Rating/ranking systems, post-season bowl games, and 'the spread'. *Journal of Quantitative Analysis in Sports*, 6(3), 2010.

[174] M. Tumminello, F. Lillo, and R. N. Mantegna. Correlation, hierarchies, and networks in financial markets. *Journal of Economic Behavior & Organization*, 75(1):40–58, 2010.

[175] M. Tumminello, S. Miccichè, L. J. Dominguez, G. Lamura, M. G. Melchiorre, M. Barbagallo, and R. N. Mantegna. Happy aged people are all alike, while every unhappy aged person is unhappy in its own way. *PloS One*, 6(9):e23377, 2011.

[176] M. Tumminello, S. Miccichè, F. Lillo, J. Piilo, and R. N. Mantegna. Statistically validated networks in bipartite complex systems. *PLoS One*, 6(3):e17994, 03 2011.

[177] UNComtrade. United Nations commodity trade statistics database. http://comtrade.un.org, 2010.

[178] N. Vlastakis, G. Dotsis, and R. N. Markellos. How efficient is the european football betting market? Evidence from arbitrage and trading strategies. *Journal of Forecasting*, 28(5):426–444, 2009.

[179] C. Von Ferber, T. Holovatch, Y. Holovatch, and V. Palchykov. Public transport networks: empirical analysis and modeling. *The European Physical Journal B-Condensed Matter and Complex Systems*, 68(2):261–275, 2009.

[180] D. Walker, H. Xie, K.-K. Yan, and S. Maslov. Ranking scientific publications using a model of network traffic. *Journal of Statistical Mechanics: Theory and Experiment*, 2007(06):P06010, 2007.

[181] D. L. Wallace. Comment. *Journal of the American Statistical Association*, 78(383):569–576, 1983.

[182] C. Wang and M. L. Vandebroek. A model based ranking system for soccer teams. *Research report, available at SSRN 2273471*, 2013.

[183] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, 1998.

[184] M. C. Wendl. H-index: however ranked, citations need context. *Nature*, 449(7161):403–403, 2007.

[185] R. Wirth and J. Hipp. Crisp-dm: Towards a standard process model for data mining. In *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining*, pages 29–39. Citeseer, 2000.

[186] G. J. Woeginger. An axiomatic characterization of the Hirsch-index. *Mathematical Social Sciences*, 56(2):224–232, 2008.

[187] Y. Xu, L. Chen, B. Li, et al. Density-based modularity for evaluating community structure in bipartite networks. *Information Sciences*, 317:278–294, 2015.

[188] E. Yan and Y. Ding. Discovering author impact: A pagerank perspective. *Information Processing & Management*, 47(1):125–134, 2011.

[189] M. Zanin, D. Papo, P. A. Sousa, E. Menasalvas, A. Nicchi, E. Kubik, and S. Boccaletti. Combining complex networks and data mining: why and how. *Physics Reports*, 635:1–44, 2016.

[190] P. Zhang and C. Moore. Scalable detection of statistically significant communities and hierarchies, using message passing for modularity. *Proceedings of the National Academy of Sciences*, 111(51):18144–18149, 2014.

[191] P. Zhang, J. Wang, X. Li, M. Li, Z. Di, and Y. Fan. Clustering coefficient and community structure of bipartite networks. *Physica A: Statistical Mechanics and its Applications*, 387(27):6869–6875, 2008.

[192] Z. Zhu, F. Cerina, A. Chessa, G. Caldarelli, and M. Riccaboni. The rise of China in the international trade network: a community core detection approach. *PloS One*, 9(8):e105496, 2014.