Institute of Informatics
University of Szeged

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

Summary of the PhD Dissertation

by

András London

Supervisor:

Dr. András Pluhár

Szeged
2017

1 Introduction

The development of *"small-world" networks* [34] 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 [19] and investigating complex systems [8]. Complex systems often can be represented by graphs (or networks), 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 [31] and Boccaletti et al. [11]). 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. Simplistically, 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: 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 focusing on the development and application of network models and graph mining tools for real-world problems.

1.1 Characteristics of Real-world Networks

The algorithms and methods developed and/or used in this study are defined on *graphs* that we use to model real-world complex systems. Formally, an undirected (directed) *graph* (or *network*) $G = (V, E)$ 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). In many cases we deal with weighted networks, means that a function $w : E \to \mathbb{R}$, that assigns a real number w_{ij} to each edge (i, j), is given.

Since Leonhard Euler invented graph theory by solving the Königsberg bridge problem in 1736, graphs have been investigated from various perspectives and applied for a wide-range of real-life problems. In the past few decades, 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 e.g. from the random graph defined by Erdős and Rényi in 1959 [18]. The seminal papers of Watts and Strogatz in 1998 [34], and Albert and Barabási in 1999 [9] gave rise a new movement of interest in the study of complex networks that are essentially graphs coming from real-life examples. Such networks appears in a wide range of areas including biology, economy and sociology, among others. These networks often have an imbalanced structure, they are dynamically evolving over time and complex in the sense that their global structural properties and functioning are not obvious from the properties of their individual parts.

In parallel with the investigation of global network properties (such as *link density*, *diameter*, *average path length*, *degree distribution*, *community* and *core/periphery structure*), the problem of rating and ranking nodes in networks has also been extensively studied. The most important contribution 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 on networks that seeks to model the user behavior of Web graph surfing [14]. PageRank is mostly used as a

network centrality measure 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, named it HITS (Hyperlink Induced Topic Search) algorithm [25]. The mathematics of PageRank and HITS, however, is general and can be applied to any graph or network in any domain [22].

2 Network Models for Some Real-life Problems

We considered various real-life problems that can be modeled by networks. The analysis of these networks proved to be quite useful for better understanding the modeled system, and we were able to extract meaningful information and answer topic specific questions. Next we briefly summarize the main results of the chapter.

2.1 Citation Networks and Scientometrics

Although in many applications PageRank scores need to be computed for all nodes of the graph, there are situations where one is interested in or capable of computing PageRank scores only for a small subset of the nodes. Chen et al. [15] developed an algorithm to approximate the PageRank score (PR) of a target node of the 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 before computing the PageRank of the target node(s). We modified their algorithm and applied it to the co-citation graph of the well-known paper by Jenő Egerváry [17] in order to give an alternative (network based) measure of its scientific impact [6]. Our algorithm can be summarized by following three steps :

1. *Subgraph building*: Starting from a set of target nodes (articles) which we are interested in, and expanding 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 can be preformed via an iteratively deepening depth-first search. Here, 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*: Using a heuristic to estimate the individual PR scores in the boundary: we add an extra term to the PR value of each boundary node that equals to the fraction of its in-coming edges to all edges in the subgraph.

3. *Calculating PR*: Running 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.

We also defined a "reaching probability" (RP) score for the same purpose. We could see that the network-based methods provided a more realistic picture of the importance of that paper than other scientometric indices (results are briefly summarized in Table 1.).

2.2 Modeling Transportation Networks

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 were especially interested in cities with a population between 100,000 and 250,000; (ii) the

Table 1: PR-score, 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 [17]	0.891	4	0.009	2	39	65
Kuhn [26]	1.189	1	0.042	1	726	1
Ford, Fulkerson [20]	0.525	8	0.004	9	39	65
Bellman [10]	0.399	11	0.003	10	18	158

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-size), but their urban morphology is different.

We were the first who performed a comprehensive network analysis (using the modern network theoretic tools) of the public transportation systems 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. Furthermore, we also assigned weights for each node and each edge by using the capacity of the vehicles and the morning peak hour schedules. 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 that validated by transportation engineers [23].

2.3 Educational Data Mining Aspects

Educational Data Mining [33] is concerned with the development, research and application of computerized methods to find patterns and features in large collections of educational data. Such features that would be hard to analyze due to the huge amount of information available and the high-level complexity of such databases.

We proposed several network representations of certain educational data and showed which are the most appropriate graph mining tools for analyzing them and what kind of additional information can be extracted by their usage [6]. Depending on the construction of the underlying graphs, we presented four families of network models and describe a case-study using one of the models. We pointed out several advantages of graph-based data mining techniques in educational systems. With the intention of evaluating the achievements of students and generating a ranking between 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 [4].

3 Network Models applied in Economics

Network models in economics are concerned with understanding of economic phenomena by using network concepts and the tools of network science [24]. 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 stock correlation-based financial networks. Next we briefly summarize the main results of the chapter.

3.1 Trade Networks

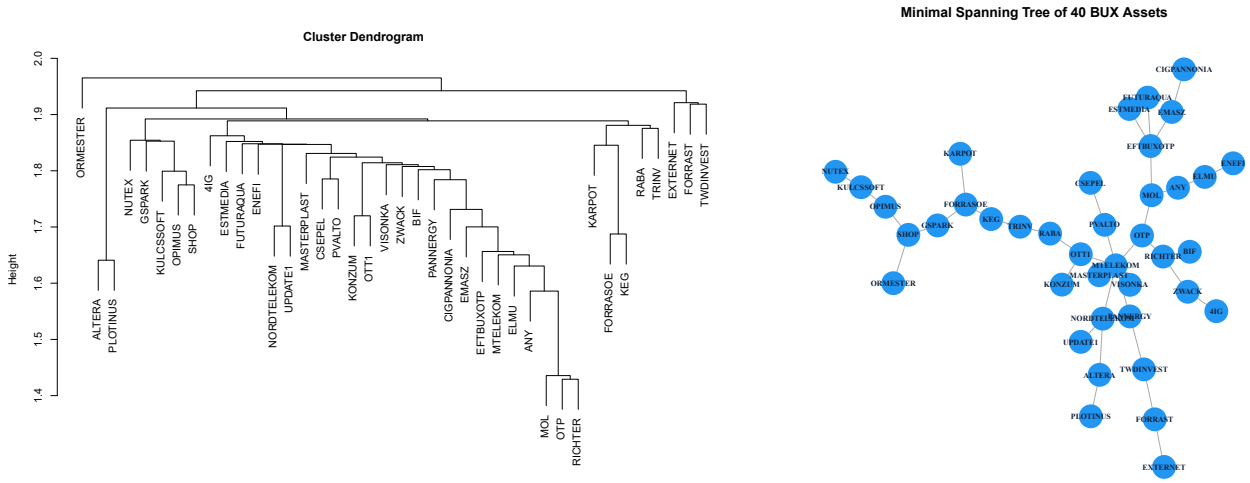
Trade networks can be studied in a simplified, but complexity preserving graph model, where the countries are represented by the nodes of the graph, while edges represent the trading relation between any two countries, often using export and import volumes in US dollars as weights. Thus, the model of 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.

We studied the trading data of the EU countries and the economic superpowers from a network perspective [30], and we found that although the export, proportional to GDP, has been growing in each European country 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 prevails, meaning that a significant percentage of the total export of the international trade is executed by just a few countries. Thereby, countries where the export volume is relatively small, but is a high proportion of the GDP, are in a strong economic dependence on the superpowers. We also pointed out that the networks have a strong core-periphery structure. There is a stable a core (with Germany, France and UK as leading economies) and a diverse periphery (contains the former Comecon countries, the countries of Balkan and South-European countries for different reasons) that changes over time. We applied a modularity optimization method [32] to identify communities and their timely evolution in the network. Then, by performing a modularity optimization based community detection method we could see, that in the trade network, peripheral countries are contained in the clusters of Russia and China, in contrast to the Western-European core countries that are in clusters where the central nodes are Germany and the USA, respectively, highlighting real economic ties among the EU countries.

3.2 Correlation Networks of Stock Markets

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.

By using tools that are common in network science we investigated the *Markowitz portfolio selection* problem [28]: 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 $\mathbf{p} = (p_1, \ldots, p_n)$ are $r_p = \sum_{i=1}^n p_i r_i = \mathbf{p}\mathbf{r}^T$ and $\sigma_p^2 = \sum_{i=1}^n \sum_{j=1}^n p_i p_j \sigma_{ij} = \mathbf{p}\Sigma\mathbf{p}^T$, respectively, where r_i is the expected return of asset i and σ_{ij} is the covariance (or correlation obtained by normalization) of the asset daily closure price



(a) (b)

Figure 1: 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

time series of i and j $(i, j = 1, \ldots, n)$ in a given time interval. The goal is to find vector $\mathbf{p}$ that minimizes the variance σ_p^2 given the minimal expected portfolio return r_p.

Since the covariance matrix, that appears in the objective function, is generally noisy and contains a large amount of information (i.e. n^2 element). Therefore, 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 which is robust against statistical uncertainty, and decrease the number of different elements in it. We used a *Random Matrix Theory* approach [29], and two versions of *hierarchical clustering* [7] (see Figure 1 for illustration) for this purpose. Moreover, to determine the expected return of the assets we used used different statistical methods.

To compare the performance of portfolio optimization in case of the different combinations of filtering procedures and return estimation techniques we analyzed the Budapest Stock Exchange data and a Yahoo Finance stock time series data. A lot of bootstrap experiments showed 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 [2, 21].

4 Network Models 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. The ranking of individuals based on the underlying graph that models their bilateral relations/comparisons 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 (e.g. road and electric) networks [27]. Next we briefly summarize the main results of the chapter.

4.1 Rating and Ranking in Sports

We defined a *time-dependent PageRank* (tdPR) algorithm and applied it for ranking players in a university table tennis competition [3]. According to our tdPR method, the ranking of a player

is not only determined by the number of his or her victories, but matters from how good players he could beat or lose against and when it happened. This means, that a good player is needed to beat for higher ranking position, but win many matches against weaker opponents does not lead anyone to the top of the ranking table. The time-dependency guarantees that the games played a long time ago do not count 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, therefore count less in the ranking. The results of our method were compared by several popular ranking techniques. We observed that our method has a good predictive power. Moreover, we think that a self-organization mechanism works in the background of the evolution of the contact graph. Obviously, players want to enter matches are expected to be exciting, and this nature of such competitions can be modeled and measured mathematically just by knowing the dates of the results. That observation gives the idea to define a special graph evolution mechanism where nodes having higher PageRank values more likely to connect to each other and this is maybe related to the emergence of an elite in sports. Further research is needed around this hypothesis. Testing our method for more sports and data sets is also another work for the future.

4.2 Probabilistic Forecasting in Sports

We presented a new model for probabilistic forecasting in sports based on linear algebraic rating methods that simply use the historical game results data of the investigated sport competition. In contrast to those techniques that use the actual respective strength, calculated using the previous results, of the two competing teams, like, for example, the celebrated *Bradley-Terry model* [13], 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 the actual relative performance of them. We consider that the smaller the rating vector, that contains the ratings of each team, changes after a certain outcome in an upcoming single game the higher the probability of that outcome occurs.

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 behind the prediction of the outcome of an upcoming match on game day r between teams i and j (outcomes are denoted by $< x : y >,\ x, y = 0, 1, \dots$) is the assumption that the more probable an outcome is the less change it causes in the rating vector $\phi^{r-1}(V)$. This changing is calculated as

$$\delta_{xy}^r = \mathrm{dist}\big(\phi^{r-1}(V), \phi^r(V) \ \big|\ \text{final result} =< x : y > \big)$$

by using some distance function $\mathrm{dist} : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$. In our experiments we used a time-dependent PageRank method and the Euclidean distance. Home-field effect was also considered such that for each team i we differentiate team home-i and team away-i resulting a $2n \times 2n$ results matrix. This matrix, in fact, describes a bipartite graph where each team appears both in the home team side and the away team side of the graph. Performing experiments on results data sets of European football championships, we observed that this model performed well (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 for the future.

5 Bipartite Network Models of Real-world Systems

A special, but rather important class of complex systems can be represented by bipartite networks, in which networks the nodes can be divided into two class, A and B, say, and links connect nodes of the different classes only. Bipartite networks naturally appears in areas ranging from social to biological systems. Next we briefly summarize the main results of the chapter.

5.1 Community Detection via Statistically Validated Projecton

We presented a method for finding the core of communities in bipartite networks using a one-mode projection method with statistical link validation [12]. In order to validate statistically each link in the projected network, we used 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 was used. The null hypothesis was 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,

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

where m is the number of nodes in B, d_i (d_j) is the degree of a_i (a_j). Then, a p-value is assigned to each pair (a_i, a_j) like so

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

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 thus the null-hypothesis is rejected and the link between a_i and a_j in the projection is validated. However, 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 was performed that associates a common level of significance to all links of the projected network. The most restrictive *Bonferroni correction* was used, that minimizes the number of false positives (i.e. type I errors), but often does not guarantee sufficient accuracy (it provides usually a large number of false negatives, i.e. type II errors).

The found cores of communities are highly informative and robust with respect to the presence of errors or missing entries in the bipartite network. We assessed the statistical robustness of cores by investigating an artificial benchmark network. We showed that this kind of filtering procedure necessarily increases the precision of the community detection, finds highly stable cores (with high precision) and suggested use, even considering the drawback that it decreases the level of accuracy in some situations. We also presented experimental results on real systems, namely an actors-movies network and a scientist-research papers network.

5.2 Rating and Ranking Nodes in Bipartite Networks

We described how a generalized version of the PageRank and HITS algorithms can be defined for bipartite networks and, as a case-study, applied on data sets of wine tasting events in order to rank tasters according to their ability and professional skill [1]. We compared the results with two

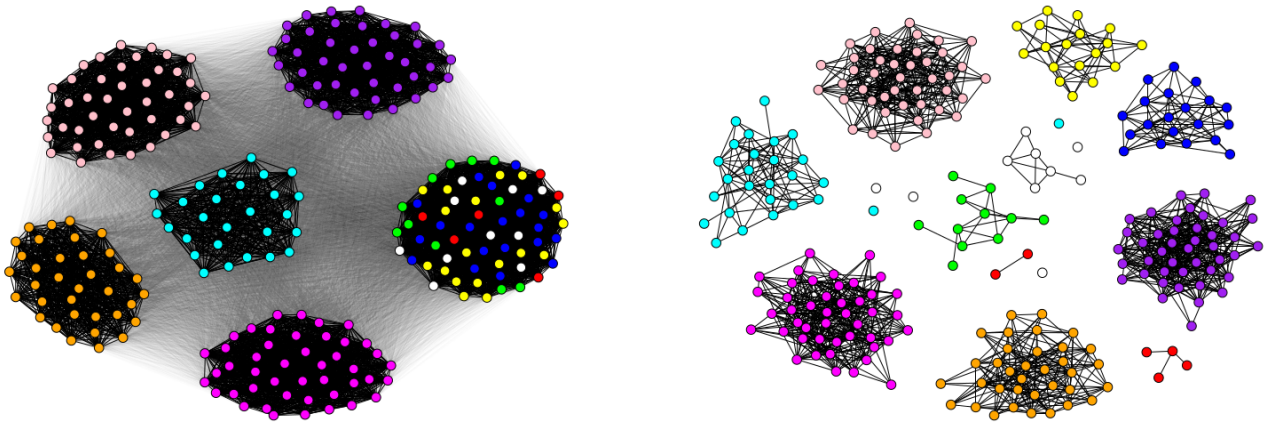


Figure 2: Communities found in the adjacency projection versus communities found in the Bonferroni projection.

simple statistical methods. Experimental results pointed out that co-HITS algorithm produced promising results confirming our a priori knowledge about the tasters involved. Furthermore it is proved to be more sophisticated than the statistical methods that produced unreasonably large differences between the tasters and ranked those tasters too high who (maybe due to their incompetency) gave the average of the scores of some of the other tasters for the wines. Another important advantage network based methods is that not each wine should be rated by each taster in order to calculate the ratings. This gives us the possibility to use such method for ranking users in a continuously evolving user-item rating database.

Summary of the Author's Contribution

The following list summarizes the key points of the dissertation. Table 2 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 to use the network approach to extract meaningful knowledge 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 performs well in general and can be applied for various real-life problems that can be modeled by networks.

II. The author addressed the question of quantifying the degree of statistical uncertainty (usually called noise) presents in real systems from different perspectives. Several methods were defined and used to filter the the part of information which is robust against statistical uncertainty (i.e. robust against errors in the data or other sources of noise). Such, 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 find communities of bipartite networks. The results show that using these techniques the classic

	[16]	[5]	[23]	[6]	[4]	[30]	[21, 2]	[3]	[12]	[1]
I.	•	•	•	•	•	•		•		•
II.							•		•	
III.									•	
IV.									•	•
Chapter	1, 2, 3	1	1	1	1	1	3	4	5	5

Table 2: Correspondence between the thesis points and publications and chapters

Markowitz solution can be outperformed, while in case of community detection the core of communities can be found with high statistical precision.

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 of communities 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 got 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 that purpose and case-study on a real data set will be presented in detail. Experimental results show that the method could be applied well in 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. This method gives a better picture and it is able to outperform several broadly used methods in terms predictive power. The author also proposes a novel rating-based forecasting framework. Against the well-known Bradley-Terry model, the main 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 outcome (i.e. final result) in an upcoming single game, the higher the probability of that outcome occurs. The results using a time-dependent PageRank method are compared to the Bradley-Terry predictions and the betting odds predictions of experts in terms of predictive accuracy. Our experiments shows that this model performs well and outperforms the advanced versions of the Bradley-Terry model in many cases, even without fine tuning parameters and optimizing the implementation.

Selected references

[1] **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.

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

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

[4] **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.

[5] **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.

[6] **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.

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

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

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

[12] 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.

[13] 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.

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

[15] 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.

[16] 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.

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

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

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

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

[21] 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.

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

[23] 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.

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

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

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

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

[28] H. Markowitz. *Portfolio selection: Efficient diversification of investments. Cowles Foundation monograph no. 16*. New York: John Wiley & Sons, Inc, 1959.

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

[30] Á. Merza, **A. London**, I. M. Kiss, A. Pelle, J. Dombi, and T. Németh. A világkereskedelem hálózatelméleti vizsgálatának lehetöségeiröl. *Közgazdasagi Szemle (Economic Review) LXIII. évfolyam*, pages 79–98, 2016.

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

[32] M. E. Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences of the USA*, 103(23):8577–8582, 2006.

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

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