Network science-based analysis of human factors in systems engineering

DOI:10.18136/PE.2020.758

PhD Thesis

by

László Gadár
environmental engineer

Supervisor

Dr. János Abonyi, DSc

The University of Pannonia
Doctoral School of Chemical Engineering and Material Sciences

July 2020

Network science-based analysis of human factors in systems engineering

Thesis for obtaining a PhD degree in the
Doctoral School of Chemical Engineering and Material Sciences
of the University of Pannonia

Written by:

László Gadár

Supervisor: János Abonyi, DSc

Propose acceptance yes / no ....
János Abonyi, DSc

The PhD-candidate has achieved .... % in the comprehensive exam.

Veszprém, ...
(Chairman of the Eximination Committee)

As reviewer, I propose acceptance of the thesis:

Name of Reviewer: Dr. Balázs Lengyel yes / no ....
(reviewer)

Name of Reviewer: Dr. Péter Pollner yes / no ....
(reviewer)

The PhD-candidate has achieved ....% at the public discussion.

Veszprém, ...
(Chairman of the Committee)

The grade of the PhD Diploma .... (……….%)

Veszprém, ...
(Chairman of UDHC)

Kivonat

Termelő rendszerek emberi tényezőinek vizsgálata hálózattudományi eszközökkel

A dolgozat célja a negyedik ipari forradalom emberi tényezőkre történő hatásainak vizsgálata, cél-orientált hálózattudományi módszerek alkalmazásával és fejlesztésével. A technológiai változások bevezetésének sikeressége az alkalmazók kompetenciáján, képességein, alkalmazkodásán is múlik. A szükséges kompetenciák egyik forrása az oktatási rendszer, valamint munkahelyi környezetben a szakmai vezetés, mentor rendszer. A fejlesztési beruházások számára elengedhetetlen a befektetési tőke, amelynek eloszlása földrajzilag nem egyenlő, és egyes régiók vonzóképessége nagyobb. Három komplex rendszer vizsgálatára, a rendszerek jobb megismerésére alkalmas hálózattudományi eszközök fejlesztése a kutatás célja.

A dolgozatban bemutatott egyik újdonság, hogy az összes diplomást anonim módon tartalmazó, hazai diplomás pályakövetési rendszer adatain alapulva a felsőfokú végzettek diplomái és az általuk betöltött munkakörök egy kétoldalú hálózatba rendezhetők. A hálózat moduljainak feltárásával megállapítható a szakok és a munkakörök horizontális illeszkedése, munkakörök kompetencia szükséglete.

A befektetési tőke számára vonzó régiók meghatározását a hazai cégtulajdonosi hálózatból levezetett települési hálózat alapján mutatja be a dolgozat. A cégtulajdonosok lakóhelyének és a telephelyének hálózatában az élek kialakulásának valószínűsége távolságfüggő. A dolgozat bemutatja, hogy az élek becslése különböző null-modellekkel lehetővé teszi távolságfüggetlen modulok meghatározását, így a távolságtól független vonzóképességi tényezők feltárását.

A termelő rendszerekben dolgozók tanulási képessége egyenlőtlen, ezért a technológia váltáskor ajánlott a szakmai mentor rendszer, betanulást támogató vezetés. Jelen dolgozat a termelő rendszerek munkahelyi szociális hálózatából határozza meg az Ipar 4.0 fejlesztések sikeres bevezetéséhez a kulcsembereket. A munkahelyi hálózatot többdimenziós kapcsolatokkal írja le az újonnan fejlesztett módszer. A kapcsolatokban rendszeresen együtt jelen levő dimenziók feltárásával lehetővé teszi a potenciális mentorok kijelölését, vezetők erősségének és gyengeségének meghatározását, szervezetfejlesztési vagy egyéni fejlesztési területek meghatározását.

Abstract

Network science-based analysis of human factors in systems engineering

The dissertation aims to investigate the effects of the Fourth Industrial Revolution on human factors by applying and developing goal-oriented network science methods. The success of the introduction of technological changes depends, among others, on the competences, abilities and adaptation of the users. One of the sources of the necessary competencies is the education system, as well as the professional leadership and mentoring system in the work environment. Investment capital, which is essential for development investment, geographically differing and some regions has a higher attractiveness. The goal was the development of network science based tools suitable for a better understanding of these three complex systems.

One of the novelties presented in the dissertation is that the degrees of higher education programs and the jobs they hold can be arranged in a bipartite network. The network based on the data of the Hungarian graduate career tracking system, which includes all graduates anonymously. The horizontal match of the higher education degrees and occupations, thus the competence requirement of the jobs can be determined by exploring the modules of the network.

In case of investment capital, the definition of attractive regions derived from the Hungarian company ownership network. The likelihood of the formation of edges decreases with increasing geographical distance. The dissertation shows that the estimation of edges with different null models increasingly allows the determination of distance-independent modules, thus uncovering attractiveness factors.

The learning ability of those working in production systems is unequal, so when technology is changing, it is recommended to build a supporting leadership and mentoring system. This dissertation provides an efficient method to identify the key people from the social network of production systems for the successful implementation of Industry 4.0 developments. It describes the community of workplaces as a multidimensional network and exploring the dimensions that often occur together in the edges. The newly developed method allows the selection of suitable mentors, the determination of the strengths and weaknesses of leaders, as well as organizational development or individual development areas.

Abstrakt

Untersuchung der menschlichen Faktoren in Produktionssystemen mit netzwerkwissenschaftlichen Methoden

Ziel der Dissertation ist die Auswirkungen der vierten industriellen Revolution auf menschliche Faktoren durch Anwendung und Entwicklung zielorientierte netzwerkwissenschaftlicher Methoden zu untersuchen. Der Erfolg der Einführung technologischer Veränderungen hängt auch von der Kompetenz, den Fähigkeiten und der Anpassung der Benutzer ab. Eine der Quellen für die erforderlichen Kompetenzen ist das Bildungssystem sowie das professionelle Management- und Mentorensystem im Arbeitsumfeld. Investitionskapital, das geografisch ungleich ist und in einigen Regionen eine höhere Attraktivität aufweist, ist für Entwicklungsinvestitionen von wesentlicher Bedeutung. Ziel der Forschung ist ein netzwerkwissenschaftliches Werkzeug zu entwickeln, die zur Untersuchung von drei komplexen Systemen und zum besseren Verständnis der Systeme geeignet sind.

Eine der in der Dissertation vorgestellten Neuheiten ist, dass die Hochschulabschlüsse und durch Sie besetzte Positionen in einem wechselseitigen Netzwerk erscheinen. Die verwendeten Daten stammen aus der Datenbank des Karriere-Tracking-Systems für Absolventen. Die Datenbank enthält anonym die Daten aller Absolventen. Module wurden aus dem Netzwerk untersucht. Jedes Modul zeigt gut die horizontale Passform der Hauptfächer und Berufe der Hochschulbildung und damit den Bedarf an beruflichen Kompetenzen.

Durch Erkundung der Module des Netzwerks, der horizontalen Anpassung der Kurse und der Arbeitskreis kann der Bedarf an der Kompetenz der Jobs bestimmt werden. Die Definition von für Investitionskapital attraktiv Regionen wird in der Dissertation auf der Grundlage des aus dem ungarischen Geschäftsinhabernetzwerk abgeleiteten Abwicklungsnetzwerks dargestellt. Die Wahrscheinlichkeit der Bildung von Kanten im Wohn- und Geschäftsnetz von Geschäftsinhabern nimmt mit zunehmender geografischer Entfernung ab. Die Dissertation zeigt, dass die Schätzung von Kanten mit unterschiedlichen Nullmodellen zunehmend die Bestimmung entfernungsunabhängiger Module ermöglicht und damit entfernungsunabhängige Attraktivitätsfaktoren untersucht.

Die Lernfähigkeit in Produktionssystemen arbeitenden ist ungleich, deswegen in Fall Technologie Änderung wird ein professionelles Mentoren System und eine Führung empfohlen, die das Lernen unterstützen. Geeignete Methoden sind erforderlich, um geeignete Mentoren und Führungskräfte zu identifizieren. Diese Dissertation identifiziert die Schlüsselpersonen aus dem sozialen Netzwerk von Produktionssystemen am Arbeitsplatz für die erfolgreiche Umsetzung von Industrie 4.0-Entwicklungen. Das Arbeitsplatznetzwerk wird in der neu entwickelten Methode durch mehrdimensionale Verbindungen beschrieben. Indem Sie die Dimensionen untersuchen, die regelmäßig in Beziehungen zusammen vorhanden sind, können Sie potenzielle Mentoren identifizieren, Stärken und Schwächen von Führungskräften festlegen und Bereiche für die Organisationsentwicklung oder die individuelle Entwicklung identifizieren.

Acknowledgements

I am grateful to my supervisor, Prof. Dr János Abonyi, who gave me professional support, encouragement, motivation, enthusiasm and adjustable deadlines. Without them, the work would not have been completed. It is an honour to work with him on a matter that we hope will get a better understanding of the world around us.

I would like to thank Tamás Szonda, the CEO and owner of Innopod Solutions Kft., for encouraging me to write this thesis. He was always curious and eagerly awaiting the next results, and he was proud of the developments.

Last but not least, I thank my family for allowing me to write and create separately, and they tolerated I spend time without them. I hope that when my children grow up, they will understand what I did in the closed room. I want to thank my wife, who patiently organized the family without me and encouraged me to write my thesis.

List of abbreviations

Table 1 Abbreviations in Chapter 2

Abbreviation	Explanation
$\mathbf{A}$	adjacency matrix of the bipartite graph of educational programs and occupations
A_{ij}	element of the weighted adjacency matrix represents how many graduates of the i-th bachelor program are working on the j-th profession
$\tilde{A}_{ij}$	element of the purified weighted adjacency matrix considering an α threshold parameter, defined by Eq. 2.6
$\mathcal{U}$	represents the set of educational programs in the bipartite graph
$\mathcal{V}$	represents the set of occupations in the bipartite graph
k_i	represents the number of students graduated in the i-th program $k_i = \sum_j A_{ij}$
k_j	represents the total number of employees having a given (j-th) profession $k_j = \sum_i A_{ij}$
k_{sat}	the empirical value of degree of saturation
k_{cut}	the empirical value of degree of cut; the distribution of degrees can be considered linear in log-log scale between the range k_{sat} and k_{cut}
L	represents the total number of links in the network, $L = \sum_{i,j} A_{ij}$
n_c	The number of communities in the network
C_c	The c-th community
N_c	Set of nodes in c-th community
L_c	Set of links connects nodes in c-th community
M_c	The modularity value of c-th community, defined by Eq. 2.3
A_{C_c}	the sum of actual links in the c-th community

Abbreviation	Explanation
P_{C_c}	the sum of expected links in the c-th community, where expectation based on configuration model
LR_{C_c}	the ratio of actual and expected links in the c-th community, the so called Louvain Ratio of the c-th community defined by Eq. 2.8
BSc/BA	Bachelor of Science and Bachelor of Arts, the bachelor level of graduation
HEd	Higher Education degree
ISCO	International Standard Classification of Occupations
STEM	Science, technology, engineering, and mathematics

Table 2 Abbreviations in Chapter 3

Abbreviation	Explanation
p	person/investor who is equivalent to the owner of a company
co	company
$[l]$	level of the settlement hierarchy (see Eq. 3.2)
$entity^{[l]}$	aggregation of an *entity* at level l of the settlement hierarchy
$\mathbf{A}^{[p,co]}$	bi-adjacency matrix of person-company ownership network
$a_{i,j}^{[p,co]}$	an element (edge weight) of the $\mathbf{A}^{[p,co]}$ bi-adjacency matrix of person-company ownership network
$\mathbf{A}^{[p,l]}$, $\mathbf{A}^{[co,l]}$	incidence matrices of person-location and company-location bipartite networks at the level l of the settlement hierarchy
$\mathbf{A}^{[l]}$	simpler notation of an adjacency matrix of location network at l level of settlement hierarchy (see Eq. 3.3)
$k_j^{[l,in]}$	in-degree of the j-th node (geographic region) at level l of the settlement hierarchy
$k_i^{[l,out]}$	out-degree of the i-th node (geographic region) at level l of the settlement hierarchy
$n_j^{[l,co]}$, $n_j^{[l,p]}$	numbers of companies and people in the j-th region at level l of the settlement hierarchy
$N^{[co]}$, $N^{[p]}$	number of companies and people/owners/investors in the network
L	number of links in the network
C	set of communities (each node is member of exactly one community)
$C^{[l]}$	set of communities at level l of the settlement hierarchy (C^1 denotes the set of towns)
$n_c^{[l]}$	number of communities at level l of the settlement hierarchy
$f(C)$	generally a metric as a function of community structure that indicates the goodness-of-fit of the community on the bases of the connectivity of nodes in it
$f(C^{[l]})$	metric of the goodness-of-fit of the community structure which is the level l of the settlement hierarchy

Abbreviation	Explanation
M	a special $f(C)$ defined by Eq. 3.18 called modularity of network
M_c	modularity of community c (sum of the modularity of each community yields the modularity M of the network)
$D_i^{[l,in]}$, $D_i^{[l,ex]}$	internal and external densities of the i-th community at level l of the settlement hierarchy, defined by Eq. 3.11 and Eq. 3.12
$O_i^{[l]}$	openness of the i-th community at level l of the settlement hierarchy, defined by Eq. 3.13
$E_i^{[l]}$	expansion of the i-th community at level l of the settlement hierarchy, defined by Eq. 3.14
$LCA_i^{[l]}$	link-collection ability of i-th community at level l of the settlement hierarchy, defined by Eq. 3.15
$CR_i^{[l]}$	cut ratio of the i-th community at level l of the settlement hierarchy, defined by Eq. 3.16

Table 3 Abbreviations in Chapter 4

Abbreviation	Explanation
$\mathcal{G}$	labelled and directed multidimensional network $\mathcal{G} = (V, E, D)$
V	set of nodes
E	set of edges, and $E = \{(u, v, d); u, v \in V, d \in D\}$; in directed network the edges (u, v, d) and (v, u, d) are distinct
D	set of labels on edges, defines the dimensions of edges, $D = \{d_1, d_2, ..., d_M\}$, where d_i is a dimension
$\mathcal{M}$	a multilayer network which is a pair $\mathcal{M} = (\mathcal{G}, \mathcal{C})$, where $\mathcal{G} = \{G_\alpha; \alpha \in \{1, ..., M\}\}$ is a family of graphs $G_\alpha = (V_\alpha, E_\alpha)$ (called layers of $\mathcal{M}$)
M	the number of layers of $\mathcal{M}$
α	a layer of $\mathcal{M}$, $\alpha \in \{1, ..., M\}$
$\mathcal{C}$	the set of edges between nodes of the different layers G_α and G_β with $\alpha \neq \beta$; mathematically $\mathcal{C} = \{E_{\alpha\beta} \subseteq V_\alpha \times V_\beta; \alpha, \beta \in \{1, ..., M\}, \alpha \neq \beta\}$, E_α are called intralayer and $E_{\alpha\beta}(\alpha \neq \beta)$ are referred to as interlayer connections
I	a set of items in frequent pattern mining, $I = \{I_1, I_2, ..., I_M\}$ (in market basket analysis, I_i represents a given product)
T	the set of transactions, $T = \{t_1, t_2, ..., t_m\}$, where $t_i \subseteq I$
$\mathcal{D}$	the database of all transactions, $\mathcal{D} = \{T_1, T_2, ..., T_{max}\}$
$\mathcal{E}$	the database of all multidimensional edges, $\mathcal{E} = \{E_1, E_2, ..., E_{max}\}$, where $E_k = \{(u, v, d); u, v \in V, d \in D\}$ is a multidimensional edge, which is a set of dimensions, $E_k \subseteq D$
$\mathcal{E}_{u^{out}}$	represent the outgoing set of multidimensional edges of a node $u \in V$, $\mathcal{E}_{u^{out}} = \{E_1, E_2, ..., E_{max}\}$, $E_k = \{(u^{out}, v^{in}, d); u, v \in V, d \in D\}$
$\mathcal{E}_{u^{in}}$	represent the incoming set of multidimensional edges of $u \in V$, $\mathcal{E}_{u^{in}} = \{E_1, E_2, ..., E_{max}\}$, $E_l = \{(v^{out}, u^{in}, d); u, v \in V, d \in D\}$
C	is a multidimensional edge $C \subseteq E$
s_{min}	a user-specified minimum support
$s_T(C)$	represents the probability of multidimensional edge C, and if $s_T(C) \geq s_{min}$ then C referred to as frequent

Abbreviation	Explanation
A	antecedent of an $A \Rightarrow B$ association rule ($A \cap B = \varnothing$), $A \subset D$
B	consequent of an $A \Rightarrow B$ association rule ($A \cap B = \varnothing$), $B \subset D$
$c_T(A \Rightarrow B)$	the probability of finding B under the condition that multidimensional edges also contain A, the confidence of an $A \Rightarrow B$ association rule ($A \cap B = \varnothing$), $c_T(A \Rightarrow B) = P(B\|A) = \frac{s_T(A \cup B)}{s_T(A)}$
l	this is the so called lift ($A \Rightarrow B$), which means that how much B increases (lift) the likelihood of A
λ	this is so called leverage ($A \Rightarrow B$), which means that how much more often A and B occur together, than expected under independence

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Fourth Industrial Revolution (Industry 4.0) has started, and it is having an impact on more and more areas of our lives. The concept of Industry 4.0 was first introduced by the German government in 2013. Industry 4.0 is the trend towards automation and data exchange in manufacturing technologies and processes which include cyber-physical systems (CPS), the internet of things (IoT), industrial internet of things (IIOT), cloud computing, cognitive computing and artificial intelligence. Industry 4.0 initially took place in engineered cyber-physical systems, and now continually spreading and growing in our culture. The cyber system of the digital age, with its hyper computing power, communication infrastructure, algorithms and information processing, is changing many other systems, such as social systems, enterprise systems, biological systems.

In my thesis, I have attempted to develop network science-based problem-oriented methods, which I will apply to study the human factors in systems engineering at macro and micro level of economy related to Industry 4.0. While working as a data analyst in a personal and organisational development company, I encounter problems with human factors when examining production systems. Working as a researcher in a research group supported by the Hungarian Academy of Sciences, my goal is to study processes behind the various phenomena affecting higher education and/or affected by higher education. New types of problems present new types of challenges, and the solutions require advanced data analysis methods.

Before introducing the developed methodologies, I place my work in a broader context in the Introduction chapter. I do not wish to analyse the technical definition, achievements, results and future directions of the Fourth Industrial Revolution, because it goes beyond the subject of the thesis. Nor do I want to deal with the microeconomic and macroeconomic issues of production systems.

The focus of Industry 4.0 is on increasing the efficiency in production. The result is a digital transformation that is changing not only production systems but also the work of other systems, from public administration to health care to education.

Digitalisation, the appearance of industrial robots, the large amount of available data, the analytical capabilities and algorithms, and the emergence of new tools pose new challenges for company staff. We are in the age of the growing demand for digital skills in the workplaces, in communication, in everyday life.

The changes have an impact on the labour market, education and organizations. There is a shortage in the labour market for skilled employees, moreover, jobs are being lost and/or transformed. The rapid rearrangement of employee competencies is indicated by estimates that say, 65% of the positions that Generation Z will occupy do not yet exist[1]. The role of the man is changing in the digitalized and robotic work environment, and employees need to prepare for transformation. Human factors will always be present in places where people are, and this will be the case in production systems as well.

The impact of digital change on people and employees should be continuously monitored and measured at different measurement levels. The results of measurements help to prepare for future challenges, to forecast processes more accurately, to determine intervention points, and to make the right decisions.

In my thesis, I developed three measurement methods related to human factors using network science toolkit, narrowing the wide broad topic area: labour market matching of higher education programs, spatial characteristics of investment decisions, and identification of key people in workplaces. The common baseline of these issues is whether methodological tools of networks science can be used to better understand some human behaviour in production systems at different aggregation levels.

In the first section of the Introduction chapter, I focus on the macroeconomic measurement points that researchers take into account to predict the future, to determine the readiness of different levels of geographical regions to Industry 4.0. I believe that researchers in this field determine the best proxy indicators of areas related to human factors, and I investigate the relationships with my work.

In the second section, I consider the human factors in preparing for the changes taking place in the production systems, that are on the front lines of the 4th Industrial Revolution. The decline in human-human relationships and the increase in human-machine relationships raise several issues.

In the final part of the Introduction chapter, I review my goals and motivations for what tasks and problems I seek answers to and why I selected these research areas. I survey the network analysis methodologies that can support to achieve my goals and gives the direction how I would like to answers my research questions. Network science tools open new research possibilities, so I consider it important to review the field in this chapter.

1.1 Measure the effect of digital transformation - a macroeconomic perspective

The digitalisation and robotisation of production systems involve challenges for companies, for politics as well as the whole economics. It is interesting to survey the current state of science, what processes are brought to the researchers' attention, and what measurement points have been selected to monitor changes. Developments and interventions at production site, settlement, region as well as country level are related to results of these measurements.

The impact of Industry 4.0 on increased investments in equipment and network infrastructure, on the personnel and material costs, on the patterns of demand according to occupations and skills, and on the increasing demand for goods are in the focus of scenario researchers [2]. Their results show that Industry 4.0 will accelerate the structural transformation towards more services and we have to be prepared to a significant labour force movement between sectors and occupations which will be higher than the change of the number of employees overall (considering entering new ones and leaving old ones).

Measurement points of change caused by Industry 4.0 can be well tracked through readiness research. Researchers of Industry 4.0 readiness carry out studies at different levels. A case study on Industry 4.0 readiness of Hungarian production plants investigates the data collection and utilisation processes in terms of manufacturing, the strengths and weakness of implementation strategy, existing equipment, IT perspective, innovation of products, training need of employees and issues related to national economic policy with a comprehensive questionnaire [3]. A research pointed out that the small and medium-sized enterprises sector showed the need to support in the aspect of preparations for revolution 4.0 [4], which indicate the fast-changing at a micro-level.

At city level Ref. [5] proposed a particular purposed measurement system called Smart Collaboration Index to assess innovation capabilities of cities which measures not only the current performance and the potential of the individual players of the ecosystem but also their collaboration capability and potential according to quadruple helix model.

Investigation of NUTS regions in terms of readiness is a gap in the literature.However, it has a significant effect both on companies and countries. Ref. [6] has attempted to identify open access indicators at the NUTS2 region level. Indicators which explicitly able to measure the readiness of regions related to higher education, labour market, innovation activities, investment, technological readiness. Their results show that the employment factors and the innovation activities are the main variables in the rankings of regions in terms of the Industry 4.0 readiness.

Authors analysed the Industry 4.0 readiness of Central and Eastern European countries at the country level. They examined the technological, entrepreneurial and governance competencies [7]. The "Invest east, export west" policy of West European countries formed thanks to the lower labour costs, good skills, and improving local business conditions. Among all variables of so-called I-Com Industry 4.0 Index include indicators of the physical and human capital to support manufacturing. The use of technologies, such as the degree of 4G coverage, the percentage of STEM graduates, the share of ICT specialists in total employment, the extent to which firms provide ICT training to their staff and the share of data workers in total employment.

The Networked Readiness Index of World Economic Forum also measures assessing countries' preparedness to reap the benefits of emerging technologies and capitalize on the opportunities presented by the digital transformation and beyond with 53 individual but related indicators (that is why called networked index) [8]. Their readiness subindex measures skills among others and impact subindex infer the economic and social impact of new technologies. Whenever relevant, the Index looks at what the different actors in society, both private and public, can do to contribute to the country's networked readiness.

All indicator system suggest that the digitalisation and robotics brought by Industry 4.0 have a powerful on changing the skills of employees and members of society as a whole. Digital technologies are disrupting career paths creating the need for new skills. The number of graduates and the number of people working in the job are used to classify regions, but these two factors are related. However, it is complicated to determine the relationship between degrees and jobs.

1.2 Measure human factors at workplaces - a microeconomic perspective

Many workers think that machines and automation steal people's work, and this revolution is, therefore, dangerous [9]. Industry 4.0 does not endanger people's work more than Ford's moving assembly lines created for enhanced efficiency. The role of man in production is changing and shifting towards control and supervision rather than the specific physical work. In fact, workers need to be open to the evolving challenges and tasks, especially openness for digital competences, is required.

Human systems engineering (HSE) is the field and commonly used intended as a structured approach to influence the intangible reality in organisations in a desirable direction. HSE combines engineering and psychology to design systems consistent with human capabilities and limitations. In other words, making technology that works for humans. HSE provides a circle of the following steps: planning, analysis, design, test and evaluation. During the planning phase, the missions and scenarios are analysed. The analysis phase contains function analysis, function allocation, task analysis. After the planning and analysis step, the developed elements will be applied in practice in system design and in the test and evaluation steps. In the thesis, I focus on the analysis phase because developed methods in the thesis related to this step.

Man-man collaboration has been complemented with man-machine (e.g. engine, computer) and machine-machine cooperation to integrate each other's strengths and to improve the efficiency of the production system. It is worth considering the differences, strengths and weaknesses of the machine and human competencies according to the given technical level when adjusting the level of automation and allocates tasks and function to human and/or machine.

Deciding which functions (tasks, jobs) of a human-machine system should be allocated to the human and which to the machine is one of the most essential activities within human factors research [10, 11, 12]. In 1951, the Fitts list [13] was the beginning of function allocation research and still the most widely used function allocation technique despite the severe criticisms [14, 15]. The original Fitts list is a list of 11 statements about whether a human or a machine performs a specific function better. Those functions that are better performed by machines should be automated, while the other tasks should be assigned to the human operator. Although, not all of its 11 statements valid today because machines have improved significantly in the past time, but still an essential approximation that describes the most important regularities of automation [12].

In addition to taking into account the competences of the human and machine in the workflow, the desired level of automation needs to be examined in the function allocation step. Ref. [16] provides an intuitive flowchart of what should be automated. For each type of automation (acquisition, analysis, decision, and action), a level of automation between low (manual) and high (full automation) is chosen according to automation criteria. The level is then evaluated by applying the primary evaluative principles of human performance consequence, and adjusted if necessary, in an iterative manner [16]. Sheridan and Verplank [17] introduced the list of 10 levels of automation which is based on the extent of decision and action done by man or machine in a task. Recent research emphasises the fact that automation introduces various problems such as behavioural adaptation, mistrust and complacency, skill degradation, degraded situation awareness, issues when reclaiming control and disruption to mental workload [12].

From the observations of how past function allocation methodologies have failed, some specific lessons learned. Additional techniques are needed to analyse human cognitive requirements [10]. Nowadays, the process is called cognitive engineering, whose goal is to provide a better fit between the human operator and the system so that the operator can more effectively perform tasks [18]. If hardware, software, and human interaction requirements are not integrated during design, it will fall on the human user/operator to do that integration in addition to the work demands of the job at hand. System design deficiencies become operations problems and require highly skilled users (or mentors) to overcome these deficiencies. These skill requirements drive increased training demands.

At the design phase, the cognitive task analysis (CTA) is used to capture people's tasks and goals within their work domain. It aimed at understanding tasks that require a lot of cognitive activity (e.g. decision making, problem-solving, memory, attention, judgement) from the user and is still an important technique to uncover system or operator level intervention points in a production workflow [19]. CTA is a structured framework specifically developed for considering the development and analysis of these complex socio-technical systems. These complex cognitive systems often involve people interacting with computers and also interacting with each other via computers in intricate networks of humans and technology. CTA can show what makes the workplace work and what keeps it from working as well as it might. [19] CTA focuses on constraints, it develops a model of how work can be conducted within a given work domain, without explicitly identifying specific sequences of actions. [20]

Some example how Industry 4.0, the automation and robotics change the production nowadays with the involvement of operators. Smart factories increasing the automation and enhance the interaction between operators and machines, which is generated a vast amount of data via different sensors and carrying the potential for further improvement. The focus of the Factory2Fit project supported by the EU is a knowledge-sharing platform called "Solution" [21]. The aim of this system is to increase the worker's motivation, satisfaction and productivity with becoming knowledge workers in a smart factory with fulfilling careers. The system collects data from operators' work and shares best practices with others. Ref. [22] reviewed the recent trends on Human-Cyber-Physical Systems (H-CPS) that is integrate the operators into a flexible multi-purpose production system creating the Operator 4.0 paradigm [23]. Authors highlighted that smart sensors, Internet of Things infrastructure wearable devices and data-driven analytic and monitoring provide a significant added value and cost reduction solution to operators in a concept of the smart factory where human and machine cooperate with each other. The last example is the sequence-mining based analysis of sensor-generated alarm data from an automated process system highlighted the benefits of the application of temporal alarm suppression rules because related faults and root cause can be uncovered [24].

The lesson learnt is that people are an integral part of the technical transformation. With the introduction of digitalization and robotics, it is necessary to develop new competencies in production systems, but learning and adoption of new knowledge are not the same for everyone. Production systems need good leadership, mentors who can support their colleagues to make technological change as smooth as possible. Fast, cheap, efficient, intelligent information discovery solutions are needed to find the right people and formulate organizational development proposals.

1.3 Uncertainties and research questions

It is still uncertain how Industry 4.0 will impact work, organisations, leadership, skills, education. Due to the rapid transformation of the labour market, it is questionable how education accommodates to the situation and how degrees match occupations at the labour market. In addition, the Industry 4.0 indicator systems presented above regularly use the number of graduates (e.g. STEM) and/or the number of people working in a given job. Questions arise, is there a reliable methodology for monitoring the relationship between education and the labour market? How well does education programs match with occupations? What types of professions match with the degree and which degree can be converted extensively? Is it a good readiness indicator which measures only the number of graduates in a given field? Or only the number of people working in a given job? Presumably, the relationship between degrees and jobs needs to be better examined. It is required an advanced methodology which can provide information about the matching of the labour market and education.

Developments require investment capital, as robotics and efficiency-enhancing solutions are money-intensive. Industry 4.0 readiness indicators provide information about an area or subregion for investors to make their decisions. Of course, it is questionable how much data is examined and whether it is obtained at all by them. However, how does investor capital move spatially? What are the driving forces behind investors who live in a locality investing in another? If we can draw a spatial network, how can we interpret connections? How much influence a city or capital its region? Which regions are attractive and to whom?

The impact of Industry 4.0 is most felt in workplaces. People's digital footprint grows, organisations become more data-rich, and the need for data analysis increases. Employees will be more demanded of flexibility, openness to innovation, adaptation, cooperation, training, problem-solving and communication. However, less monotony tolerance, memory, and precision will be required as machines and automation replace them. Managers increasingly need to act as leaders rather than managers. Their main tasks will be motivation, inspiration, knowledge management, performance analysis, and creating a trusted climate. As the automation process intensifies, change management needs to become increasingly employee-focused for sustainability. [25] Implementing changes and learning how human-machine relationships work properly, requires disseminating of competencies that need good leaders, good mentors, competent professionals, and retention of key people. However, who are they? Of course, managers know their employees to some extent, based on everyday practice. Much valuable information can be lost as the size of the organisation increases. We have seen from the literature above that interview surveys exist to assess cognitive needs, and key persons but these are costly and time-consuming techniques, besides, it is difficult to summarise the information received in case of a large number of interviews.

Can it be determined with a cheaper and simpler methodology that who is an appropriate mentor to develop his/her colleagues by transferring and/or sharing knowledge? Is the manager capable of creating a trusting climate and can motivate her/his employees? Who needs to be improved by a personal developer? Who is a key person in the organisation whose retention needs to be paid more attention? Who is influential and accepted among employees and why?

1.4 Aims and objectives

My goal is to find empirical answers to research questions with the usage and development new network science methodologies in a goal-oriented manner.

Specifically, my aim is to investigate the human factors of production systems listed below, using network science methods, considering the available reliable databases:

- the transition of knowledge acquired in education into work, that is to say, the competencies needed to fill a position or the qualifications need to jobs, and find horizontal similarity between educations and occupations,
- the spacial indicators of the decision situation in the business ownership network and to describe attractive regions,
- the nature of multidimensional relationships between the different types of connections in the social network of employees, and characterize actors with connection types and define similar (key) actors with similar multidimensional relationships.

Examining the matching of education and occupations requires exploring the frequent career paths. Graduates with a specific competency package get the jobs that best suit them [26]. The Administration Database of Hungarian Authorities contains career paths of a grade cohort, which also summarises occupations they work with a specific graduate degree. A career path connects education with the job, where the links are the graduates. Thus, the problem of training-occupation matching is transformed to the uncover of modules of a bipartite network where one set of nodes are educations, and the other set of nodes are professions. My goal is to examine the career path data as a bipartite network to explore modules to determine the similarity of elements.

A network with specific properties emerge when the formation of relationships between spatially embedded nodes effected by the costs and risks. An investor living in a settlement wish to minimize the risk [27]. The network of settlements in which the edges point to the direction of investment can be modelled with a spatial network. My goal in this decision-based network is to determine the attractiveness of settlements.

My methodological aim is to investigate the effects of different null models that approximate the edge formation on the determination of attractiveness. My purpose is to use the 'regions' (modules) where edges are more likely to appear than the null model predicts, to determine the factors of attractiveness.

Human relationships are multidimensional because feelings, evaluations, perceptions, intentions, interests emerge between interacting people. One way to analyse a network with complicated connection types is to separate relationships by dimensions (layers). However, at the dyad level, the appearance of dimensions occur together is not likely to exist independently [28]. The presence of overlaps is the result of a combination of several phenomena. My goal is to develop a survey and analysing method to uncover the properties of multidimensional relationships of employees. My further aim is to develop a goal-oriented approach at the dyad level for analysing frequently together occur dimensions. Different actors contribute to varying extent to overlapping dimensions in him or her incoming or outgoing edges. My goal is to develop a method which finds similar entities by overlapping dimensions. With the help of this new method I want to examine motivating leaders, mentors, areas of required personal development.

1.5 Methodological aspects

I set my goals of applying and developing network science methodologies to answer my research questions, so in the Introduction chapter, I consider it important to provide an overview of the field of science. Two of the empirical chapters focus on searching modularity to find similar elements and one chapter pay attention to multilayer networks, so I open a discussion on them in the Introduction chapter.

Researchers have distinguished between complicated and complex systems. The main idea behind complex systems is that the ensemble behaves in a way not predicted by the components. The interactions matter more than the nature and the performance of the units. [29] If complex systems can be understood through connections, then network models should be used to study them. Network science provides a broad analytic tool for understanding multilevel, multi-label, multilayer networks. To better understand some selected phenomena related to Industry 4.0, I interpreted databases as a network, developed methods, divide elements to clusters and analysed them.

1.5.1 Methodological opportunities with networks

A network is a great model to represent connected entities, which is indicated with the revolutionary growth of new methodologies and articles since Erdős and Rényi [30] through Watts and Strogatz [31], as well as Barabási [32] and Newman [33] to nowadays. The dynamics and structure of the system of interconnected elements are being knowledgeable by methods of network science. Among others, the following options are opened when analysing a system as a network. I would like to demonstrate that with the development of network science, a huge number of methodological possibilities open, but not all of them are applied in my empirical research, but they may emerge in my later research.

- defining the properties of nodes in a network
  - determination the embeddedness of nodes in the network
  - centralities [34, 35]
  - influential entities [36, 37]
  - the role in a multilayer network [38]
  - structurally and regularly equivalence (similarity) [33]
- defining densely connected subnetworks
  - modules, communities of nodes [39, 40, 41, 42, 43]
  - and steps for solving the resolution limit problem [44, 45]
  - modules in a multilayer network [46]
- defining structural properties at the dyad level
  - reciprocity [47, 48, 49]
  - transitivity, local and global clustering coefficient [33]
  - overlaps of edges in a multirelational network [50]
- defining structural properties of the network
  - degree distribution of nodes [51, 52]
  - components [33]
  - paths and small world effect [31]
  - homophily or assortative mixing [53]
- processes on networks
  - spreading phenomena [54]
  - percolation, resilience, robustness [55, 56, 57]
  - dynamical systems [58]

There is a tremendous amount of analytical potential, which results from thinking in networks. It is impossible to use all available analytical method, and therefore goal-oriented selection is necessary, to get a better understanding of the system under investigation.

1.5.2 Finding modules

The analysis of vertical matching of the educations and occupations is equal to find densely connected elements in a bipartite network. Thus, the problem of the applicability of competences acquired in training can be transformed into finding modules in a bipartite network where the one set of nodes are educations and the other set of nodes are the occupations.

The likelihood of emerging links between spatially embedded vertices is usually distance dependent. Thus, evidently, the number of connections between nearby vertices is higher, so the random configuration model based modules will be geographic regions [41]. However, my goal is to find other attractiveness factors besides geographical distance, so different null models should be used when exploring the modules.

Finding modules is one of the main analytical methods used in this work. Therefore some of its properties need to be discussed. A module is a unit whose structural elements are densely connected among themselves and relatively weakly connected to items in other community. A complex system can be managed by dividing it up into smaller pieces and looking at each one separately [59]. The presence of modules and the degree of modularity is one of the most important structural characteristics of the network. Network modularity, by definition, is a difference that compares the number of connections within a module to the expected number of links compared to the null model [60]. Community structure algorithms are maximizing the modularity and thus uncovering densely connected units of the network.

Define community structure is performed in two consecutive steps: first, detection of meaningful community structure, and the second, evaluation of the appropriateness of the detected communities. One of the main directions of community detection algorithms is greedy algorithms [61, 62]. Another leading trend in the defining community structure based on random walking like infomap method [63]. But there are several other methods developed by researchers [64].

Modularity based community detection has a resolution limit, and small communities remain undetected. These algorithms fail to detect modules which contain less than $\sqrt{L}$ edges, where L is the total number of edges in the network [65]. RB [66] and AFG [45] methods can handle this resolution limit problem by modifying the modularity function with adjusting the contribution of the null model and adding self-loops to the nodes, respectively.

In addition to the resolution limit, another limitation of the community detection is that a node is only included in one module. Structurally, it may be possible for one or more vertices to belong to multiple modules. Identifying these a priori unknown building blocks is crucial to the understanding of the structural and functional properties of networks [43]. Palla et al. introduced an approach to uncover overlapping communities to understand the modular structure of complex systems better.

Since then, there have been developed many other methods of exploring overlapping communities. [67, 68, 69, 70]

The community detection in complex systems with spatially embedded nodes caused another challenge for researchers. The distance-dependent edge formation proved by the deterrence function shows that when the configuration model or Newman-Girvan modularity is previously applied as a null model, the communities overlook the spatial nature of the system and modules reflect geographical regions [41]. The selection of the reference network or null model determines the factors that the researcher considers when finding modules as mesoscale structural elements of the network [71, 72]. If the null model better approximates the edge weights of the studied network, than the value of modularity decreases, however, the forces of formation modules less effected by geographical distance. If the reference network contains economic factors or gravity-like driving forces, the methodology may also be suitable for defining attractiveness factors.

1.5.3 Application of multilayer networks

It is easy to realize that treating all the network's links on an equivalent footing is a too big constraint, and may occasionally result in not fully capturing the details present in some real-life problems, leading even to incorrect descriptions of some phenomena that are taking place on real-world networks [73]. A set of people in a social network interact with different patterns, different levels, people have different aims to contact others and connections are not equal. Strong and weak ties [74], multiple relationships are around us. A multilayer network is an intuitive model to describe complex systems.

The decomposition of a complex system into layers providing new insights into the structure and function. The multilayer modelling of human brain networks obtained new achievements based on magnetic resonance imaging and resulting in better understand the functional connectivity of neurons [75]. Detect communities in a network with multiple connections by layers helps to define similar entities which frequently being in the same community [76]. The interlayer connected transport network model of a city where layers represent different modes of public transport (bus, tram, subway etc.) helps to find the intervention point to reach better diffusion of users and categorizing zones [77]. The degree of centrality of nodes distributed in different layers helps to characterize them by function [38, 78].

The coexistence of several types of interactions among the entities of a complex system is responsible for substantial differences in the kind and variety of behaviours. Analysis of multilayer networks become a hot topic in the complexity science. However, it has a various challenge in the future. [79] One of these is to find meaningful correlations between layers which is reflected in this thesis.

1.6 Outline

I have shown that Industry 4.0 is implemented in a socio-technological complex system where machine/computer/IoT cooperate. Complex systems are best known through relationships because some property of elements expresses in an interaction. Systems need to analyse as network, separate into components to make conclusions. In my dissertation, I examine three (two macro and one micro-level) aspects of a socio-technological complex system related to human components with developing new methods:

- the relationship between employees skills and university degrees
- the formation and characteristics of a network influenced by geographical distance in the business owner network,
- multidimensional relationships of co-workers, leaders.

Graphical abstract of my thesis shown in Figure 1.1. I would like to represent the related examined elements with connections and separate the individual chapters that appear in the dissertation with dashed lines. Although each of the separated parts is a chapter of the dissertation, and they are also related. An employee with a specific higher educational degree and skills has a multidimensional relationship with her/his colleagues. Her/his work is influenced by the investor, who expects results and performance from her/him. I would also like to demonstrate with this figure that I do not deal with human-machine and human-IoT relation because the focus of my thesis is on human factors.

Chapter 2 (Modularity based node similarity in a bipartite network) provides a methodological innovation for the relationship between university degrees and occupations by establishing a bipartite network. I studied similar degrees and occupations with uncovering the modules in the network. I also analyse which education and occupation have a focused or diffused relationship with the other set of nodes in the bipartite network.

In Chapter 3 (Modularity based attractivity in a spatial network) I examine the network of settlements based on business ownerships. The network can represent the attractiveness of settlements for investors also in Industry 4.0 investment projects. Methodological development and difficulties in understanding the system are related to the spatial characteristics of the network.

Chapter 4 (Evaluation of network, clusters and node characteristics with overlapping dimensions of multidimensional edges) at the micro-level explores the multidimensional relationships of employees of companies. As a methodological development, I examine the multidimensional relationships between employees and use overlaps of several layers to qualify and cluster nodes.

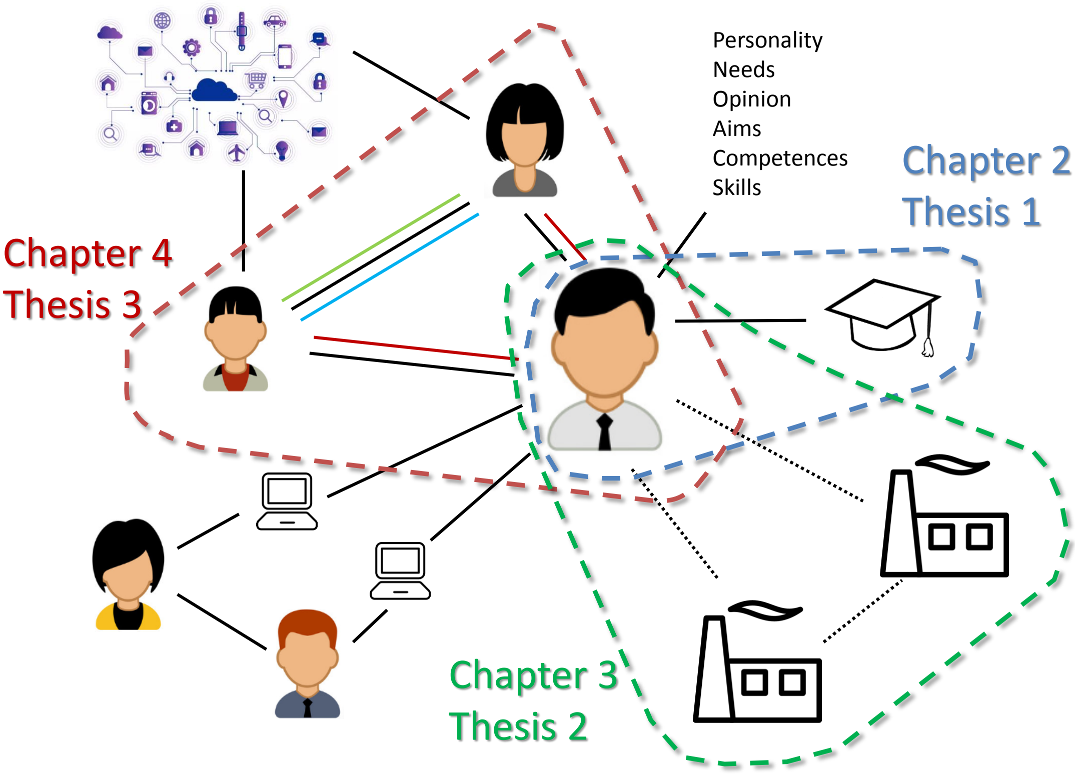


Figure 1.1 Schematic representation of a complex sociotechnological system pointed out the contents of thesis with dashed lines.

Chapter 2

Modularity based node similarity in a bipartite network

Abstract To study education – occupation matchings we developed a bipartite network model of education to work transition and a graph configuration model based metric. The career paths of more than seven-thousand Hungarian students based on the integrated database of the National Tax Administration, the National Health Insurance Fund, and the higher education information system of the Hungarian Government were studied. A brief analysis of gender pay gap and the spatial distribution of overeducation is presented to demonstrate the background of the research and the resulted open dataset. We highlighted the hierarchical and clustered structure of the career paths based on the multi-resolution analysis of the graph modularity. The results of the cluster analysis can support policymakers to fine-tune the fragmented program structure of higher education.

2.1 Introduction

Policymakers need solid information on how labour market evaluates higher education graduates. Institutions also should collect and analyse relevant information about their graduates for the management of their programs [80]. Since the salary and the chance of finding a job are important decision factors at the college attendance [81], university and program level public information about the career paths are also important to candidates of higher education [82].

Although self-reported data can have validity problems, questionnaire based databases are useful to study education-occupation matches. Among these, the Reflex database is the most comprehensive information source in Europe. The analysis of this database showed that graduates working in the field of their study have higher income and satisfaction, so they are a happier members of the society [83].

Administrative data can replace traditional questionnaires to offer much more objective information for evidence-based educational policy in decision-making [84]. In Hungary, the 2007/CI law prescribes that governmental organisations should review their decisions by using administrative data. As a new element, under the Government Decree No. 389/2016, the basic financial support for Hungarian higher education institutions changed based on the overeducation data calculated from the administrative databases. In Austria database of the whole state insurance system is accessible in anonymized form, which is also ready to career path analysis [85]. With administrative data, we can also measure the added value of higher education institutes by combining information about persistence rates, graduation rates, and post-college earnings [86]. The use of administrative data has a long tradition in Northern Europe. Finland recently connected administrative and survey data sources [87]. Based on the register of Statistics of Finland some employers were suggested to be interviewed to study unemployment of young graduates and transition from higher education to work [88]. The Swedish Ladok database was used to determine the influence of higher education institutions on labour market by regression analysis. The availability of extensive, longitudinal data made it possible to the evaluate the matching of the occupation and the level of the degree among engineering, teaching, nursing, business specialisations [26].

In this work, a new method was developed to dig deeper by focusing a goal oriented network mining tool to evaluate the matching of programs and occupations on the more detailed, at program level.

In recent years, network-type models have been proven to be useful in understanding complex systems in different subject areas (e.g. sociology, economy, industry, and biology [89]). Real life entities (e.g. people, universities, educational programs) can be characterised by numerous categorical properties (e.g. education can characterise people). Relationships between entities and values of a selected property can be modelled with a two-mode network (also known as a bipartite graph) [90].

The proposed network model is based on the integration of the databases of the National Tax Administration, the National Health Insurance Fund, and the data warehouse of the Hungarian higher education. This administrative dataset covers 15 thousand people graduated in 2009/2010 academic year and worked in 2012 May. Based on the data of 7402 Bachelor students we defined a bipartite graph of 110 bachelor programs and 113 occupations encoded by the third level of International Standard Classification of Occupations (ISCO) code system. The nodes of the resulted network are connected by 7402 links that represent the employees who received their bachelor level in a given program and work in a given profession. To demonstrate the power of administrative database, we present a brief analysis of gender pay gap and the spatial distribution of overeducation.

The analysis of the bipartite network shows that both the programs and the occupations follow a power law distribution which reflects there is a structure in the carrier paths. The key idea is that the weights of the edges with the expected number of edges of a random graph that has the same strengths as the studied network was compared. This configuration model seems the most sophisticated reference because it takes into account the expected number of links by weights of given program and occupation [91].

To search patterns in education-occupation transition in different levels of details, we cluster the graph by looking for subgraphs whose vertices are more likely to be connected to one another than to the vertices outside the subgraph [40]. To evaluate the consistency of the detected clusters we use a graph modularity based measure which assesses the quality of the clusters based on the number of edges of the configuration model [39]. A multi-resolution type analysis of the network by the step by step removal of the weak connections was elaborated. The results highlight that the educational programs have a hierarchical structure.

A large number of higher education programs can lead to a fragmented and inefficient education system. The results confirm that the extracted clusters can support decisions related to the monitoring and (re-)design of the program structure.

2.2 Methods

2.2.1 Bipartite graph model of the education to work transition

The vertices of the bipartite graph model of the education to work transition are divided into two disjoint sets, $\mathcal{U}, \mathcal{V}$. The $\mathcal{U}$ represents the educational programs, and the $\mathcal{V}$ represents the sets of occupations. Edges connect a program and an occupation. The edges are weighted, and the weights are representing the number of graduated students in a given program connected to a specific profession. The graph can be represented by an $\mathbf{A}$ adjacency matrix, where the A_{ij} element of the matrix represents how many graduates of the i-th bachelor program are working on the j-th profession.

By following this arrangement, the sum of the i-th row, represents the number of students graduated in the i-th program, while the sum of the j-th column represents the total number of employees having a given (j-th) profession. These sums can be considered as the strength of the nodes, calculated as $k_i = \sum_j A_{ij}$ and $k_j = \sum_i A_{ij}$, respectively.

Not all nodes in a network have the same number of edges (same node strength). The probability that a node has $< k >$ edges can be described by a distribution function $P(k)$. The analysis of the strength distribution can show how the graduates are distributed among the programs and the occupations.

2.2.2 Evaluation of the education-occupation match

To decide which education programs and occupation pairs are relevant and which can be considered as a "noisy" individual case, we propose a measurement to evaluate the strengths of the connections.

The core idea is that we can compare the A_{ij} weight of the edge with the expected edge weight of a degree preserving random graph that has the same degrees as the studied network. This configuration model, which is often referred as a random network with a pre-defined degree sequence [92], seems the most sophisticated application because it takes into account the expected number of links by degrees of given program and occupation.

If the edges were randomly distributed, $\frac{k_i k_j}{L}$ would be the expected number of links between the i-th program and j-th occupation, where L represents the total number of links in the network, $L = \sum_{i,j} A_{ij}$, while k_i and k_j are the strengths of the program and occupation nodes, respectively [40].

Since in the case of random matching $\frac{k_i k_j}{L}$ graduates of the i-th program would choose the j-th occupation, the difference between the actual and the expected number of graduates in the case of random arrangement can be calculated as:

$$A_{ij} - \frac{k_i k_j}{L} \tag{2.1}$$

which difference can be used as a measure of the strength of the education - occupation matchings.

2.2.3 Simultaneous clustering the programs and the occupations

In the previous session, the connection of individual educational programs and occupations was evaluated. To provide information about the whole structure of the network, the edges to obtain groups of similar programs and professions were clustered.

To formalise this clustering problem, we utilise the modularity measure introduced by Newman [39] and improved for bipartite graphs by Barber [40]. A module of the network is a subgraph whose vertices are more likely to be connected to one another than to the vertices outside the subgraph. Modularity reflects the extent, relative to a random configuration network, to which edges are formed within modules instead of between modules:

$$Q = \frac{1}{L} \sum_{ij} (A_{ij} - \frac{k_i k_j}{L}) \delta(C_i, C_j) \tag{2.2}$$

where the Kronecker delta function δ is equal to one when nodes i and j are classified as being in the same module (i.e. they have the same label value) or zero otherwise.

The modularity can be determined for each community of a network. A network with n_c communities, the following modularity value is used to determine M_c community modularity value. Each C_c community with N_c nodes are connected with by L_c links, $c = 1, \dots n_c$.

$$M_c = \frac{1}{L} \sum_{(i,j) \in C_c} (A_{ij} - \frac{k_i k_j}{L}) \tag{2.3}$$

The M_c modularity value of a c cluster can be either positive, negative or zero. In the case of zero, the community has as many links as a random subgraph. If it is a positive value, then the C_c subgraph tend to be a community, while a negative M_c means it is not.

The commonly used multi-level modularity optimisation algorithm (so called Louvain algorithm) to find clusters in the programs-occupation bipartite graph was used. This algorithm uses an iterative procedure to assign each node to a module by maximising the modularity [62]. Although the Louvain algorithm is stochastic, in modularity optimization, edges with the most different number of connections than random will be placed in a module. In the case of large networks, the iteration process can lead to modules with different nodes, and there are several ways to decrease this problem [93, 94]. In the presented case, the number of network vertices and edges is small, and the modules explored show the most characteristic differences, especially for education programs and jobs with high strength, which mostly indicate the match of a particular educational area and job group.

The rows and the columns of the adjacency matrix of the bipartite graph can be reordered to visualise the similarities of the programs and relationships (see later in chapter Clustering and visualisation).

2.2.4 Multi-resolution cluster analysis

2.2.4.1 Improvement of the resolution

The modularity always increases when small communities are assigned to one group [95]. Modularity optimisation with the null model p^{NG} has a resolution threshold which means, it fails to identify small communities in large networks and communities consisting of less than ($\sqrt{L/2}$-1) internal links [96]. Reichardt and Bornholdt (RB) generalized the modularity function by introducing an adjustable γ_r parameter [97, 66] to handle this problem, which for our directed and weighted networks is:

$$M_{RB}^{\text{dir}} = \frac{1}{L} \sum_i \sum_j \left(a_{ij} - \gamma_r \frac{k_i^{out} k_j^{in}}{L} \right) \delta(C_i, C_j) \tag{2.4}$$

Arenas, Fernandez and Gomez (AFG) also proposed a multi-resolution method by adding r self-loops to each node [45]. This algorithm increases the strength of a node without altering the topological characteristics of the original network, as: $\mathbf{A}_r = \mathbf{A} + r\ \mathbf{I}$, where $\mathbf{I}$ denotes the identity matrix and r the weight of the self-loops of each node.

$$M_{AFG} = \frac{1}{L'} \sum_i \sum_j \left(a'_{i,j} - \frac{k_i^{out\prime} k_j^{in\prime}}{L'} \right) \delta(C_i, C_j) \tag{2.5}$$

where $L' = L + Nr$; $L = \sum_{i,j} a_{i,j}$; $k_i^{out\prime} = k_i^{out} + r$; $k_j^{in\prime} = k_j^{in} + r$;

$$a'_{i,j} = \begin{cases} a_{i,j}, & \text{if } i \neq j, \\ a_{i,j} + r, & \text{if } i = j. \end{cases}$$

These methods still have the intrinsic limitation, so large communities may have been split before small communities became visible. The theoretical results indicated that this limitation depends on the degree of interconnectedness of small communities and the difference between the sizes of the communities, while independent of the size of the whole network[95].

It should be noted that the modularity decreases when $p_{i,j}$ more closely approximates the real $a_{i,j}$ values which is equivalent to finding the null model that most closely fits.

2.2.4.2 Modified multi-resolution method

Since we would like to determine how the significant matchings are structured, we applied a method for cleaning the network by step by step removing of the weak connections. Stronger and weaker connections are also present in the network structure compared to the random configuration model. The main structural elements of the network can get by increase the threshold parameter shown in Equation 2.6. The method contributes to the decomposition of modules through the most typical connections. Thus we can conclude the hierarchical structure of the network.

$$\tilde{A}_{ij} = \begin{cases} A_{ij}, & if\, A_{ij} \geq \frac{k_i k_j}{L} \cdot \alpha \\ 0, & if\, A_{ij} < \frac{k_i k_j}{L} \cdot \alpha \end{cases} \tag{2.6}$$

As the equation 2.6 describes the cleaning procedure has a $0 \leq \alpha$ threshold parameter. When $\alpha = 0$, none of the edges are removed. It should be noted, that α can be considered as a minimum relative edge strength. After the pruning of network, all connections will have α times larger weight than weight would be expected based on the random configuration model:

$$\frac{\tilde{A}_{ij}}{\frac{k_i k_j}{L}} \geq \alpha\, \forall i, j \tag{2.7}$$

It should be noted that 2.7 equation measures how the given edge contributes to the Louvain ratio used to measure the compactness of a module/cluster:

$$LR_{C_c} = \frac{A_{C_c}}{P_{C_c}} \, , \, A_{C_c} = \sum_{(i,j) \in C_c} A_{ij} \, , \, P_{C_c} = \sum_{(i,j) \in C_c} \frac{k_i k_j}{L} \, . \tag{2.8}$$

As can be seen later in the subsection 2.2.2, it is interesting to analyse how the step-by step increase of this threshold α parameter decreases the network density, what is the ratio of the non-significant edges, how characteristically structured the network.

Since after this pruning modularity based clustering was also applied, and the resulted method can be considered as a special multi-resolution analysis technique.

Modularity optimisation based community detection has a resolution limit, failing to detect communities smaller than a scale that depends on the total number of edges in the network and degree of interconnectedness of the communities [65]. To handle this problem multi-resolution methods were introduced by adjusting the resolution of the algorithms by modifying the modularity function, weighting the contribution of the null model [44] or adding self-loops to the nodes [45]. These methods still have the intrinsic limitations that large communities may have been split before small communities become visible [98].

Since the problem is that modularity based community detection algorithms join small fully connected subgraphs that connected only by weak edges into larger groups [99], our methodology which gradually removes the less important connections by the increase of α can also be considered a graph-modification based multi-resolution approach that handles this problem. It should be noted, this algorithm was developed with the aim to find how the statistically significant education-occupation matchings are structured.

2.3 Results and discussion

2.3.1 Administrative data of the hungarian career path tracking system

The studied administrative government data were collected as the integration the databases of the Hungarian tax office and National Health Insurance Fund in 2014. The database was designed by using individual hash codes, so although it does not contain personal information. It allows the micro data level analysis of student career paths.

The integrated database contains 70 variables about

- personal data (date of birth, county of address, citizenship, gender)

- occupational data in May 2012 (employment relationship, occupation, gross wage, etc.)
- employer data (county of company headquarters, company size, company activity, etc.)
- if the graduate runs her/his own enterprise the primary data of the employer is the own enterprise
- educational data (institute, faculty, program where the graduate graduated)

The dataset contains 29873 individual records. Among these only 15253 people have occupational data [100]. It must be noted, that this integration was the first made by the related governmental organisations in Hungary, so probably this is the reason why only the half of the persons were correctly merged. Table 2.1 shows contents of published database. The correctly identified 15253 people graduated in 398 education programs delivered by 52 institutions and they worked in 402 occupations encoded by the fourth level (unit groups) of the International Standard Classification of Occupations (ISCO) code system. In this work, we focus on just bachelor degree graduates who worked in a known workplace. Among 15253 people 7402 has a bachelor degree in 45 institutions, 110 programs, and works in 113 third level occupation groups.

Table 2.1 Variables of the dataset

Column name	Description
ID	ID of the graduate
Gender	Gender
Inst_HUN	Institution name in Hungarian
Pr_area_HUN	Training program areas in Hungarian
Pr_HUN	Training programs in Hungarian
Pr_ENG_bachelor	Training programs of bachelors in English
Grad_level	Degree levels
FEOR4	4th level of Hungarian ISCO
FEOR3	3rd level of Hungarian ISCO
ISCO1	1st level of ISCO
ISCO3	3rd level of ISCO
Req_HEd	Jobs that does require HE degree
Head_couty_HUN	Couty of company headquarter
Weekly_hrs	Weekly working hours
Mountly_wage_HUF	Mountly gross wage in HUF

2.3.2 Measuring overeducation

Interesting point of the dataset that according to the main group of ISCO code we can determine that a given occupation requires higher education degree or not. Table 2.2

shows how much percentage of the graduates has jobs that require higher education degree.

Table 2.2 Distribution of graduates working in occupation category that requires higher education degree (HEd)

Education level	Require HEd	Not require HEd	Number of graduates
Higher vocational trainings	38.9%	61.1%	921
Bachelor	68.8%	31.2%	7402
Collage (equal with Bachelor)	66.9%	33.1%	1889
Master	90.5%	9.5%	1334
University (equal with Master)	84.8%	15.2%	3252
Special teacher programs	67.5%	32.5%	409

The Shankey diagram of the BSc/BA graduates (see Figure 2.1) shows that who graduated in computer science and information technology, health science, engineering science works more likely in an occupation that requires higher education degree compared with graduates in sports science, arts and humanities, natural sciences, agricultural science.

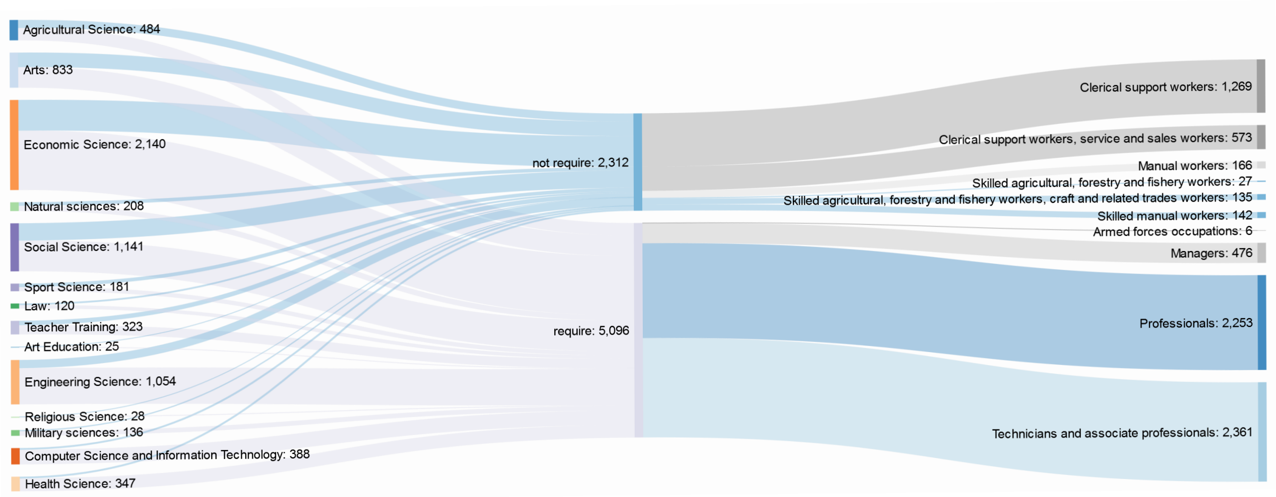


Figure 2.1 Distribution of bachelor graduates working in an occupation that requires higher education degree.

Working in a field that matches to the education has a positive effect on job performance and satisfaction [83]. Results of Iriondo and Pérez-Amaral indicates that overeducated workers suffer a wage penalty since earnings depend mainly on the educational requirements of jobs [101]. Three primary measures of education - job mismatch can be distinguished based on how the required education level is determined. The first method relies on the self-assessment [102], the second approach evaluates the "realised matches" [103], while the third "job analysis method" refers to a systematic evaluation of the "professional job analysts" who specify the required level of education for the job titles in an occupational classification [104] [105]. These last two methods require large scale and up to date administrative data-based studies, similar that we would like to deliver in our research.

Groot and Maassen van den Brink conducted a meta-analysis of 25 studies on overeducation and found that the matching based methods show 13.1% of overeducation, while self-assessment based studies estimate the much higher percentage of overeducated employees, 28.6% [106]. McGuiness and Sloane used the REFLEX dataset to study overeducation in the UK. When both education and skill mismatch variables were included in the model, overskilling reduced job satisfaction consistently for both sexes. In the UK 36% of the respondents felt as overeducated, which is quite high, compared to the 14% measured elsewhere in Europe [107], and 17% in Taiwan in 2008 [108].

A much more objective study has been performed in Spain where the Spanish Wage Structure Survey (WSS) dataset was used to examine the effects of educational mismatch on wages. Based on this employer-worker microdata 32-37% overeducation rates were calculated [109].

Our database allows a more detailed analysis. Similarly to other countries, this dataset also shows 26-39% of overeducated employees. The spatial distribution of occupations of the graduates was also investigated. Figure 2.2 shows how much percentage of the bachelor graduates are working in jobs that require higher education. This figure well illustrates that the problem of overeducation is differentiated spatially and find reasons need deeper research which is far from this work.

2.3.3 Evaluation of the degree distributions

The cumulative degree distributions of the bipartite network are shown in Figure 2.3 and 2.4. The k_i weighted degrees of the five biggest programs are: Business Management: 874, Communication and Media Science: 469, Andragogy: 367, Tourism: 364, Finance and Accounting: 310. The k_j degrees of the top five occupations are: Business services agents: 614, financial and mathematical associate professionals: 419, Engineering professionals (excluding electrotechnology): 410, Legal professionals: 372, Sales and purchasing agents and brokers: 371.

To evaluate the whole structure of the network power-law, exponential, Poisson,

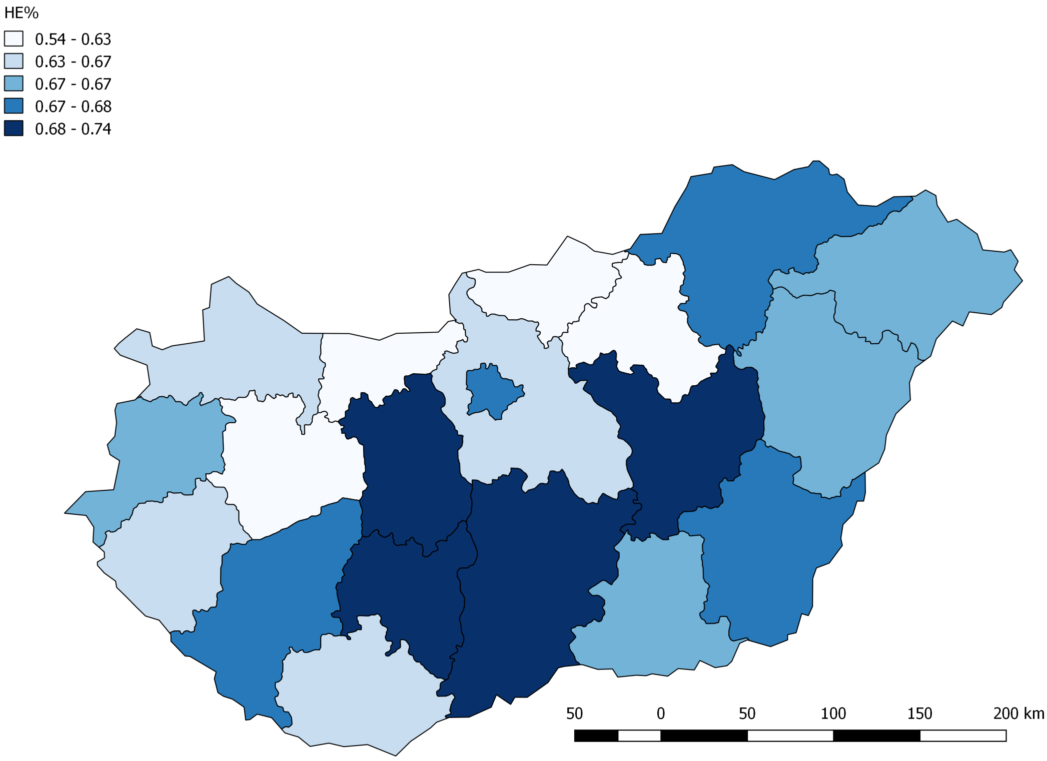


Figure 2.2 Distribution of graduates that work on occupation which requiring higher education degree by counties in Hungary

log-normal distributions was fitted to strengths of nodes in R with the help of the *poweRlaw package* [110].

As can be seen, there is a well defined linear region between k_{sat} and k_{cut}. The slope of this linear trend gives γ which is 2.00. There are less number of small strength nodes (occupations) then power-law fit would require therefore in $k < k_{sat}$ region data point are below the extrapolation of fitted line. Similarly, there are less number of high strength nodes or hubs then power-law fit would require thus in $k_{cut} < k$ region data point are also below the extrapolation of fitted line.

The strength distribution of these scale free networks can be described by a power-law tail function, $P(k) = k^{-\gamma}$ [52], and the γ parameter is one of the most important property of a graph.

The question is that which model is closer to the empirical distribution. The p-values shown in Table 2.3 suggest that we cannot reject the null hypothesis that the data follows power-law distribution.

Power-law and log-normal distributions were compared using Vuong's test statistic [111].The two sided p-value of comparison test shows that both distributions are equally close to the empirical distribution.

When the nodes of a network are randomly connected, γ is bigger than three. $2 < \gamma < 3$ relates to scale-free networks [92]. If a scale-free network has $2 < \gamma < 3$, than $< k^2 >$ diverges as $N \rightarrow \infty$, making the network ultra small. This is a consequence

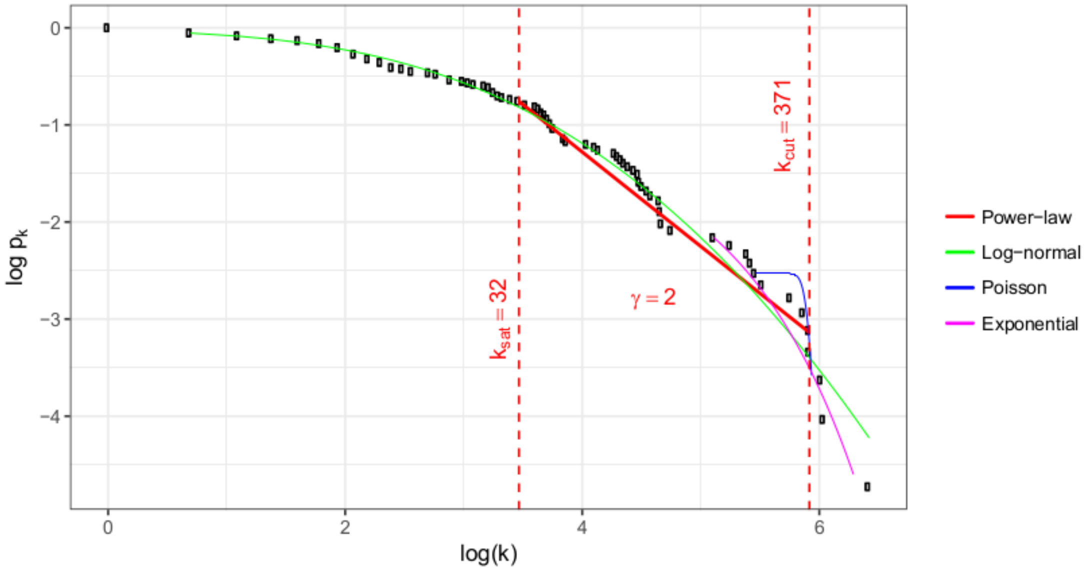


Figure 2.3 Distribution of the weighted degrees of the occupations

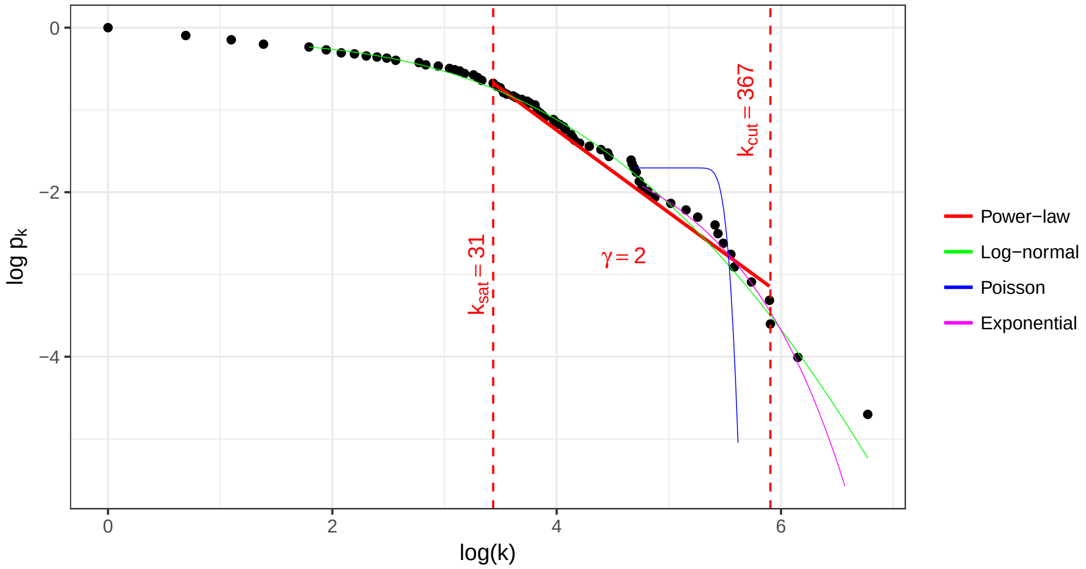


Figure 2.4 Distribution of the weighted degrees of the bachelor programs

Table 2.3 Results of fitting power-law to bipartite graph

Graph set	K_{min}	K_{max}	k_{sat}	k_{cut}	D	p	γ
Programs	1	875	31	367	0.084	0.93	2.00
Occupations	1	614	32	371	0.101	0.87	2.00

of the hubs, that act as bridges between many small degree nodes. Price's model suggests, that the growing network models primary target is to explain the origination of networks with strongly skewed degree distributions[112, 113]. His idea was that old nodes get new edges proportionally to the number of existing edges, which is also known as preferential attachment. The reason behind its origin, whether it is based on randomness or optimization, is not clear yet[114].

Our analysis shows that the studied bipartite network has scale-free structure, which assumes that graduates prefer occupations with high strength, but their preferences or optimization strategy need to be studied more detailed.

2.3.4 Evaluation of education - occupation matching

The proposed graph configuration model based measure can be used for ranking the education - occupation matchings. Tables 2.4 and 2.5 show the strongest and the weakest educational program - occupation pairs. Tables give examples how works the comparison of the real and expected edge strengths.

Table 2.4 Top 10 strongest connection

Programs	Professions	# links	# links (conf. model)
Business Management	Financial and mathematical profess.	153	49
Mechanical Engineer	Engineering professionals	114	14
Engineering IT Specialist	Software and applications developers	102	3
Tourism and Hospitality	Client information workers	83	17
Finance and Accounting	Financial and mathematical associate	80	18
Electrical engineering	Electrotechnology engineers	65	3
Nursing and Patient Care	Other health professionals	55	2
Business Management	Numerical and material recording clerks	58	10
Business Management	Business services agents	84	36
Nursing and Patient Care	Nurses and midwives	47	2

Although the tables show only a tiny slice of the most typical and least typical career path, they are very close to preliminary expectations. The most typical career paths include matching jobs for graduates in economics, engineering and health sciences. Among the least typical relationships, mainly economics graduates, were in engineering and IT-type jobs. In the case of such relationships, it is not typical that a previous degree caused the emerge the connection. Only one case was that a previous bachelor degree was engineering informatics.

2.3.5 Degree correlation and centrality measures

Programs with a relatively small number of connections are not "spread". In some of these programs, there are a lot of graduates, but they work in a few kind of occupations.

Table 2.5 Top 10 weakest connection

Programs	Professions	# links	# links (conf. model)
Tourism and hospitality	Engineering professionals	1	20
Electrical engineering	Business services agents	3	22
Communication and Media Science	Engineering professionals	6	26
Finance and Accounting	Client information workers	1	14
Electrical engineering	Financial and mathematical associate	1	14
Andragogy	Engineering professionals	4	20
Finance and Accounting	Engineering professionals	3	18
Commerce and Marketing	Engineering professionals	1	14
Business Management	Software and applications developers	9	28
Communication and Media Science	Software and applications developers	2	14

These programs are the following: computer science, engineering (electrical engineer, mechanical engineer, civil engineer, chemical engineer, mechatronic engineer), special need teacher, nursing and patient care, laboratory and diagnostic imaging analyst, and economic science (finance and accounting, international management, management organisation). Conversely, only small number of agricultural, teacher training, liberal arts, light industrial engineering graduates work in several kinds of occupation.

To get more information about the structure of our network, we calculated the degree correlation as the correlation of the node degree with the node degree of the neighbour nodes. The results show that our graph is disassortative, which means high-degree components (hubs) tend to connect to low-degree nodes, while low-degree nodes are connected to hubs [92]. In practice, this means that graduates of programs on that few people graduated work in popular occupations.

Computer science, nursing, and engineering programs are interesting in the context of eigenvector centrality as well. These programs are mainly connected to occupations that have few kind of "supplier", so they are not so embedded in the graph. This group of programs have relatively high betweenness and low closeness centrality which means that they include many shortest path, but they are far from the centre of the graph. This phenomenon predicts that the related programs - occupations connections are strong and form subgraphs that have fewer connections to other parts of the graph.

2.3.6 Clustering and visualisation

During clustering each program and occupation were assigned to one module, so each module contains a set of programs and professions. The result of clustering can be seen in Figure 2.5. This figure shows the adjacency matrix with columns and rows ordered according to the result of the clustering. The resulting five clusters are marked by A-E letters are the diagonal blocks of this matrix. Table 2.6 shows the Louvain ratio (see Eq 2.8) for every module.

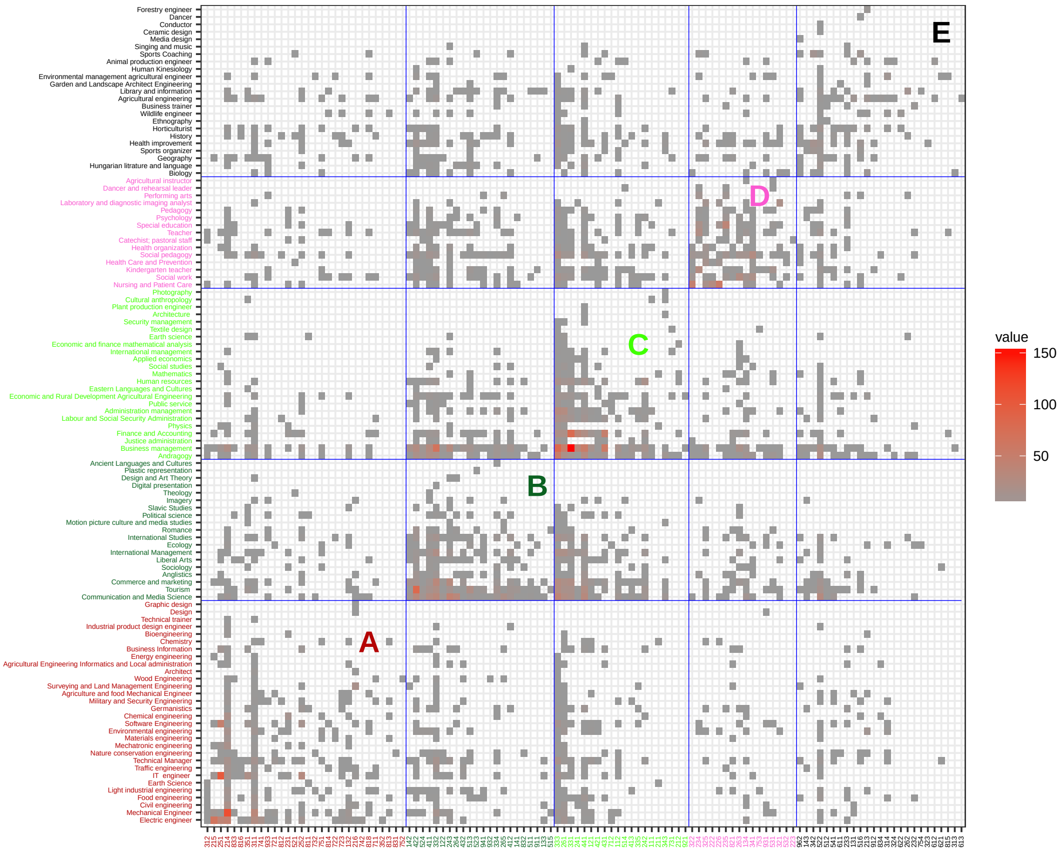


Figure 2.5 The modules obtained by the Louvain algorithm of purified program/occupation bipartite graph.

Table 2.6 Louvain ratios of the pairs of program-occupation clusters

	A	B	C	D	E
E	0.614	0.860	0.584	0.666	**4.152**
D	0.246	0.664	0.498	**4.051**	0.694
C	0.209	0.829	**1.826**	0.429	0.572
B	0.225	**2.010**	0.962	0.348	0.797
A	**3.200**	0.430	0.370	0.208	0.556

Module 'A' consists of engineering programs supplemented with design, technical trainer and germanistics programs. This cluster highlights that 25% of graduates of germanistics work in manufacturing and IT sector, indicating the strong presence of the German industry requiring advanced knowledge of German language.

Module 'B' consists of economy, social, political, and language teacher programs weakly connected to sales, office workers, client information, journalist, brokers, marketing and PR professionals.

Module 'C' contains management, financial economic, and agricultural programs, along with small programs such as physics, earth science, cultural anthropology. With them, financial, business, clerk, trade workers, keyboard operators, and service worker occupations are associated. It should be noted that there is a relatively strong connection between the 'B' and 'C' modules.

Module 'D' connects medical, pedagogy, teacher, social work, dancer and arts type programs with health, teaching, child care workers, medical technicians, and personal care workers. In this module, there are the fewest number of occupations which do not require higher education diploma.

Module 'E' collects programs with a small number of graduates. This module shows how agricultural, natural science, teacher, sport, art type of programs are connected to operators, technicians, workers, vocational education teacher, animal producer, crop grower, cooks, salesperson, services manager, and life science professionals.

In the case of the A, D and E modules this ratio shows a larger difference from the null model compared to the B and C models, which indicates the stronger connection of programs and occupation in the A, D, and E modules.

During our work, we tested several clustering algorithms, including the BRIM algorithm (bipartite, recursively induced modules) developed specifically for bipartite graphs [40].

As Figure 2.6 shows, the first module contains mainly teaching, humanity and art programs. The second module exhibit business, economic, finance, HR, social work, nursing, medical programs. In this module, almost half of the occupations do not require higher education diploma, like cooks, hairdresser, personal services, cashier, personal care, food preparation assistant, elementary worker.

The third module represents natural and technical sciences programs, like engineering, IT, physics, ecology, earth science connected to production, manufacturing, information managers, life science, engineering professionals.

2.3.7 Application of multi-resolution cluster analysis

Louvain modularity optimisation algorithm was performed with different α which is a kind of a multi-resolution method to study modules. The aim of this application is to study the hierarchical system of modules shown in Figure 2.5. If the modules are hierarchically structured, then removing weak links are broken modules up into further elements.

Firstly the number of remained and removed links was investigated. Figure 2.7 shows the distribution of the strong edges that remaining with the adjustment the α threshold parameter. It shows that $\alpha = 2.5$, half of the connections remain, which

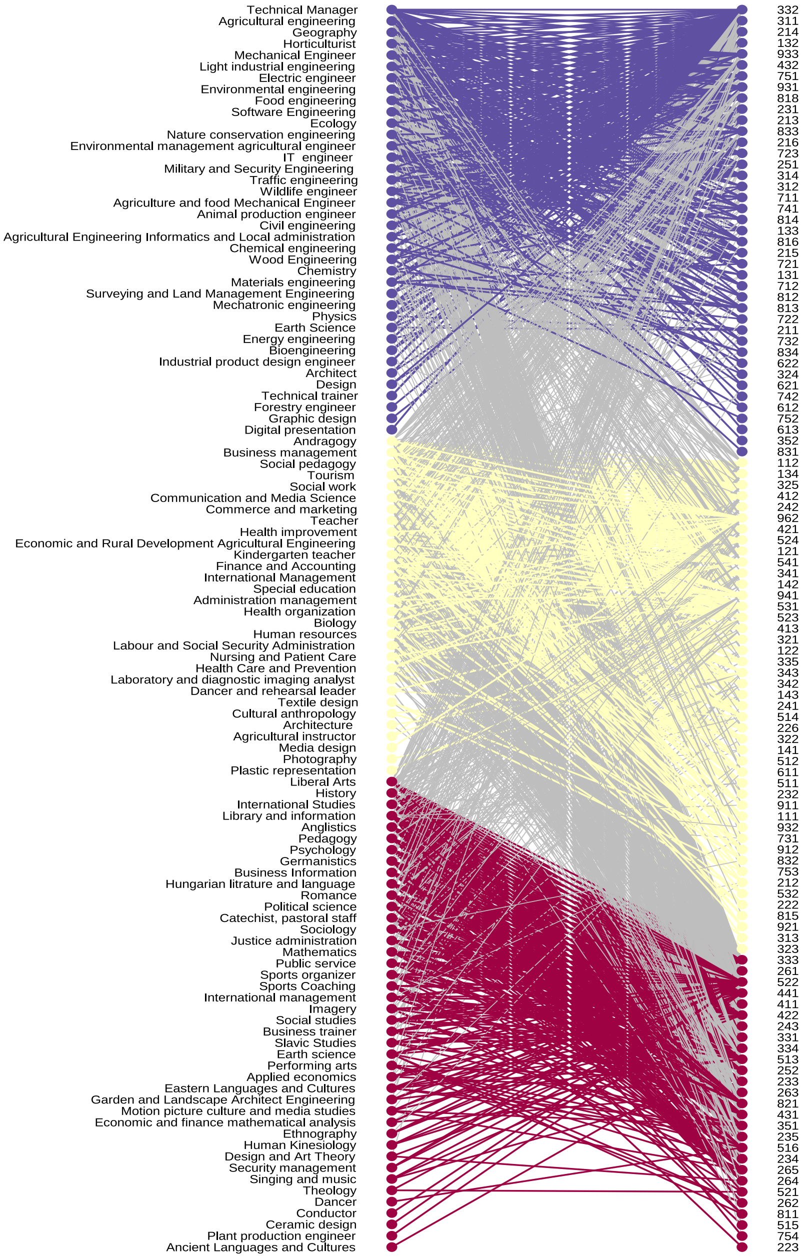


Figure 2.6 Clustering and reordering of the bipartite graph with Barber algorithm (grey edges indicate connections between modules; blue, yellow, red edges are in the three modules)

appear to be the strongest ones.

Secondly the number of clusters was investigated with the adjustment of the α parameter. The Figure 2.8. shows the number of clusters resulted from Louvain algorithm in the function of α.

A hierarchical structure was detected since all communities found at a value of $\alpha_2 > \alpha_1$ are sub-communities of the communities found at α_1. The similarities of the nodes was measured based on whether their share the same cluster in different resolutions. A hierarchical splitting occurs when the cluster of health type programs

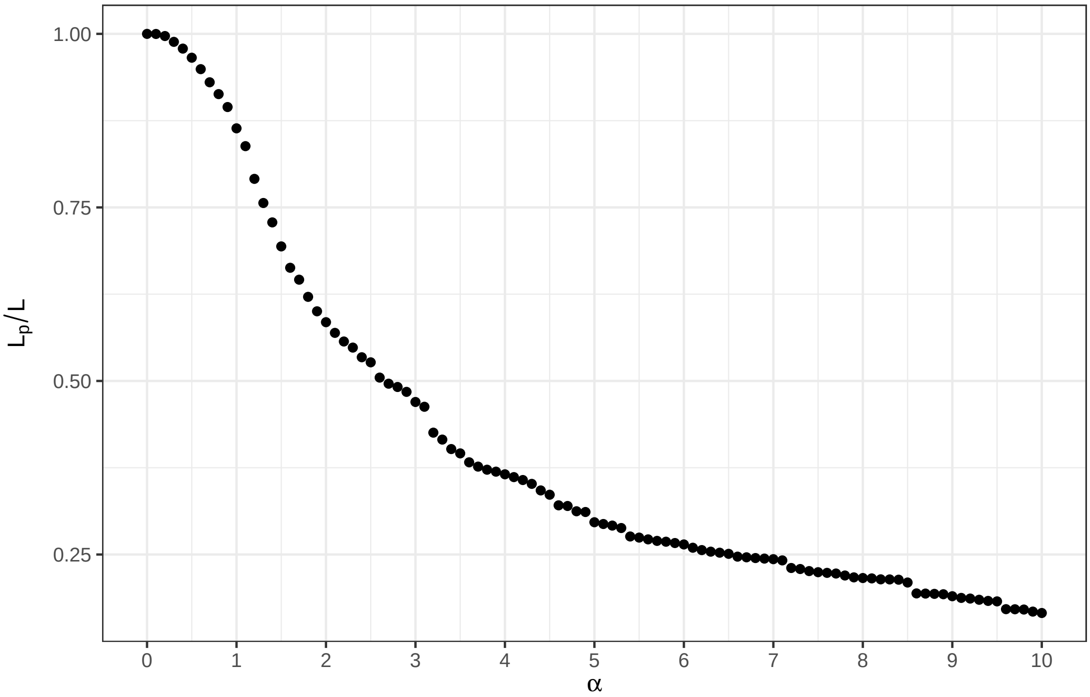


Figure 2.7 Distribution of the education - occupation significance values as the ratio of the remaining edges after pruning with different α

splits into nursing and medical professionals. Similarly, the cluster of pedagogy and social programs is divided into teaching and social professionals and the group of child care workers.

The relationships of the clusterings resulted in different resolution level is visualised in Figure 2.9. Each dot in the rows shows the contents of the modules for α_1, and similarly each dot in the columns shows the contents of the modules for α_2. If a module in a row represented with two or more points then it means that this module divided into different parts by the increasing of α parameter.

As can be seen, somewhat hierarchical splitting occurred in the first cluster. By increasing α IT engineer, business information, software engineering, germanistics programs with software developer, database professionals, information technology operations technicians occupations separated (see Figure 2.9 $x = 4, y = 1$)

2.4 Conclusion

Administrative data based career path analysis can of support governmental policy making and program development of higher education institutes. To support the extraction of useful information from these databases we developed a graph-based data structure to represent the career path of higher education graduates. Education - occupation mismatch can be analysed based on the bipartite graph of bachelor programs and occupations encoded by International Standard Classification of Occupa-

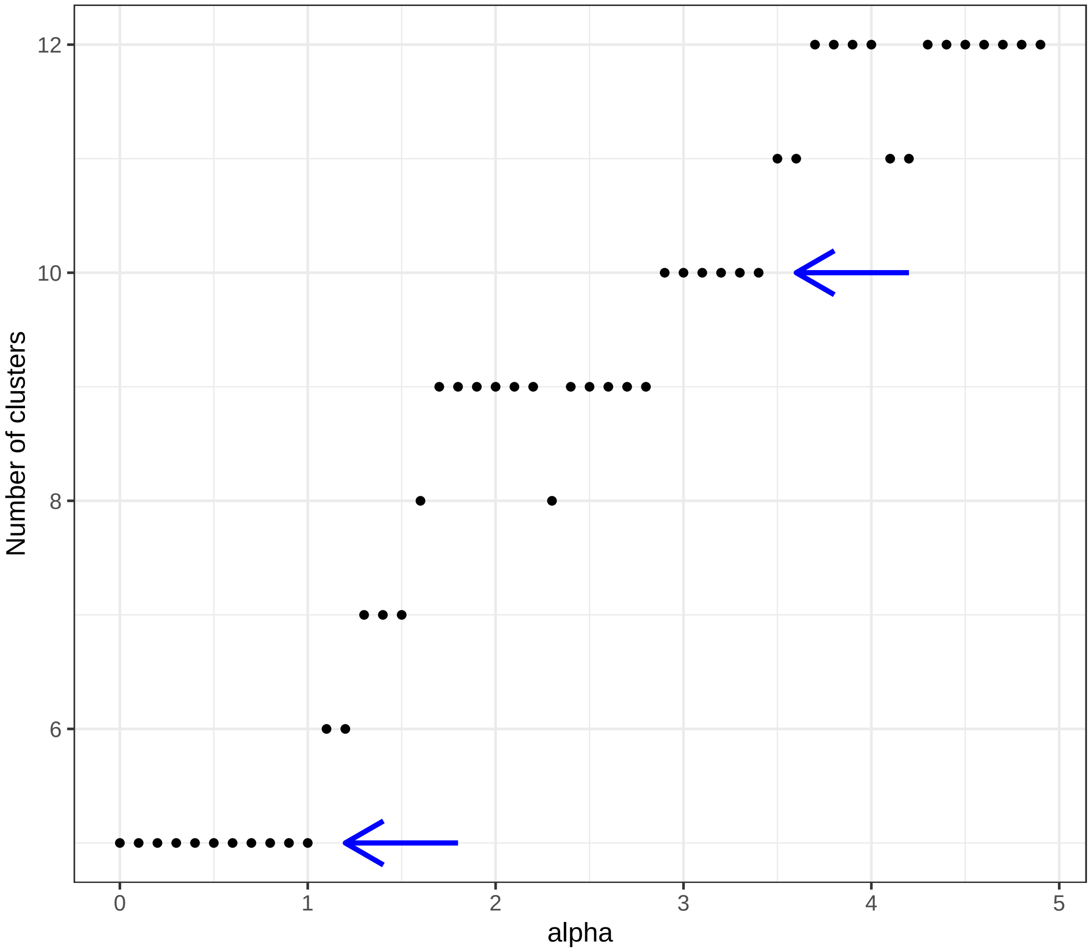


Figure 2.8 Number of clusters in case of different α

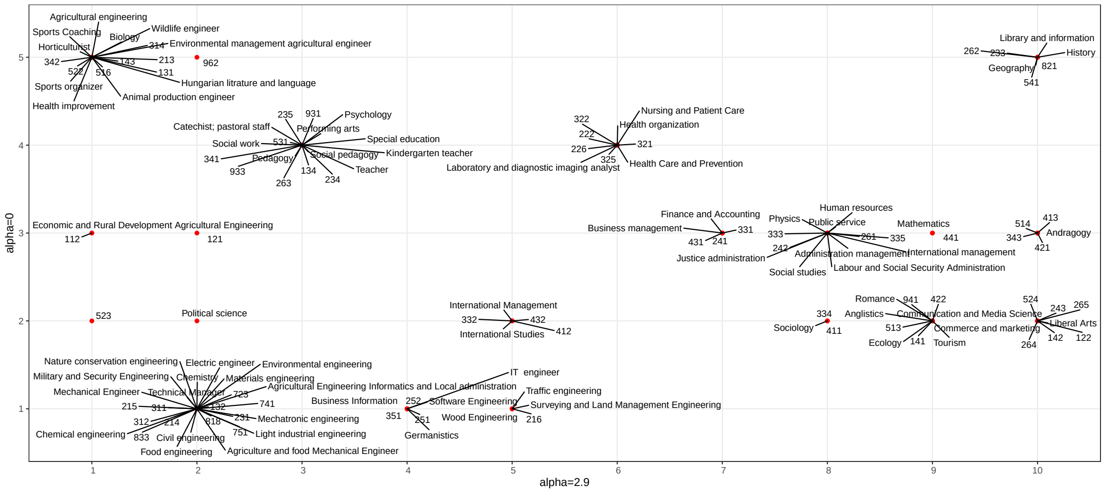


Figure 2.9 Relationship of clusters generated in step one and step three of the multi-resolution analysis

tions (ISCO) code system. We modified the Newman modularity measure to evaluate the matching of the programs and the professions. Based on this measure the hidden structure of career paths can also be clustered and visualised.

The proposed network model is applied on the integrated databases of the National

Tax Administration, the National Health Insurance Fund, and the data warehouse of the Hungarian higher education. To demonstrate the information content of this administrative database, we presented a brief analysis of the gender pay gap and the spatial distribution of the overeducation. Similarly to other countries, we showed 26-39% of overeducated employees.

The transition of graduates from higher education to employment is affected by individual characteristics. However, graduates with well-defined qualification start working in a somewhat similar profession. Our graph model gives the opportunity to cluster the typical career paths and find outliers whose education and occupation does not match. The results illustrate that the proposed multi-resolution type community finding approach provides useful results, as it highlights the groups of programs that are strongly connected to groups of bachelor programs.

The analysis of the clusters allows us the more sophisticated analysis of the performances of the programs in the labour market. For example, our method showed that significant proportion of graduates of Germanistics work as a system administrator indicating the strong presence of German origin industry in Hungary.

Such results can be useful for education policy experts and decision makers who can see the structure of the Bachelor programs from the objective viewpoint of the labour market. The resulted orderings and matching measures can support the policymakers to fine-tune the fragmented program structure of the Hungarian higher education. We found bachelor programs that are almost identical in their content, but they are different from the view of the labour market because graduates work in the different occupation. For example, pedagogy and andragogy use similar methods, but the graduates of andragogy work profession that is more related to communication and media science. (Probably this was one of the reasons why the andragogy Bachelor program has been closed in Hungary in 2017.)

The results are also informative to students and applicants of the higher education who want to be prepared for a finding job with good expectations.

2.5 Contributions to Industry 4.0 issues

As an impact of Industry 4.0, the labour market is changing drastically, because the change in technology requires the obtain new skills and competences. One source of skills is education, which needs to prepare for changes in the future.

My dissertation contributes to defining the relationship between education and the labour market. Examining the horizontal match of education programs and jobs is a hard problem. The method has developed in thesis contributes to a more accurate examination of the early career path of graduates based on an administrative database of the whole cohort of graduates.

The number of STEM graduates and/or the number of people working in STEM jobs are often used indicators to specify the readiness of regions for Industry 4.0. Still, none of the indicators examines the correlation between them in depth. Based on the results, it can be concluded that the two indicators show similar phenomena. The matching of engineering programs and jobs is high. However, not all STEM graduates work in STEM jobs. Also, not only STEM graduates work in STEM jobs.

Chapter 3

Modularity based attractivity in a spatial network

Abstract How are ownership relationships distributed in the geographical space? Is physical proximity a significant factor in investment decisions? What is the impact of the capital city? How can the structure of investment patterns characterize the attractiveness and development of economic regions? To explore these issues, we analyse the network of company ownership in Hungary and determine how are connections are distributed in geographical space. Based on the calculation of the internal and external linking probabilities, we propose several measures to evaluate the attractiveness of towns and geographic regions. Community detection based on several null models indicates that modules of the network coincide with administrative regions, in which Budapest is the absolute centre, and where county centres function as hubs. Gravity model-based modularity analysis highlights that besides the strong attraction of Budapest, geographical distance has a significant influence over the frequency of connections and the target nodes play the most significant role in link formation, which confirms that the analysis of the directed company-ownership network gives a good indication of regional attractiveness.

3.1 Introduction

Mining valuable information from social networks is a hard problem due to its dynamic nature [115, 116], complex structure [117, 118] and multidimensionality [119]. This chapter deals with the structural issues as it tries to evaluate regional attractiveness based on a set of goal-oriented null models identified to describe the geographical distributions of company ownership relations.

Complex multivariate socio-economic data is widely used to monitor regional policy [120, 121]. As the usage of a different set of variables results in various rankings, the definition and selection of socio-economic variables is the key issue in these applications. The drawback of these indicator-based approaches is that although economic behaviour is socially constructed and embedded in networks of interpersonal relations [122] and strong related to location [123], the network structure of the economy is neglected.

This chapter adds a viewpoint to regional studies based on the analysis of how the network of personal investments and the founding of companies relate to the settlement hierarchy. We assume that the socially embedded economy must have a network-based imprint in the company-ownership network which is a good indication of regional attractiveness.

Attractiveness is meaningful in preferential attachment networks, where the likelihood of a new connection is proportional with degree [51] and fitness [124] of the node. These models were generalized to handle initial attractiveness [125] and latecomer nodes with a higher degree of fitness [32, 124]. It is important to note that these models generate power-law (degree) distributions that are similar to the distribution of socio-economic variables of settlements indicating that preferential attachment is a process that can be used to describe city grow [126, 127, 128, 129, 130]. In the case of geographically distributed networks, the likelihood of link formation is dependent on distance due to the cost of establishing connections and spatial constraints [41]. Connection costs also favour the formation of cliques and thus increase the clustering coefficient [131]. Space is important in social networks as most individuals connect with their spatial neighbours [131] to minimize their effort and maintain social ties [132], e.g. the majority of our friends are in our spatial neighbourhood [133]. The probability $P(d)$ that distance d separates two connected individuals is found to behave as $P(d) \sim d^{-2}$ in terms of Belgian mobile phone data [134], or generally $P(d) \sim d^{-\delta}$, as has been shown in the case of the social network of more than one million bloggers in the USA [135], in friendship network of Facebook users, and in email communication networks [136, 137].

The attractiveness of airports [138], countries for foreign investments [139] and tourist destinations [140] is evaluated based on socio-economic variables. As many origins and destinations are present in these applications, the theory of bilateral trade flows accounts for the relative attractiveness of origin-destination pairs. The gravity model is one of the most successful empirical models implemented in economics to describe such interactions across the space [141]. Originally developed by Newton, the law of gravity can also be used in economics to describe the extent of interactions between two points of mass in networks. Almost 40 years ago, before the emergence of network science, Anderson suggested that as a force between two mass points, the number of trips from location i to location j follows the (economic version) of the 'Gravity' law, $F(d) \sim P(d) \sim I_1^{\alpha} I_2^{\alpha} d^{-\delta}$ [142]. Nowadays, many complex networks embedded in space and spatial constraints may have an effect on their connectivity patterns such as trade markets [143], migration [144], traffic flow [145] and mobile communication [134] that can be successfully modelled by a gravity model, which was also successfully applied in link prediction [146].

It is assumed that regions that heavily rely on local resources consist of more internal connections that form modules in networks, so the modularity of the networks which reflect socio-economic relationships can be used to measure regional attractiveness. The goal of modularity analysis is to separate the network into groups of vertices that have fewer connections between them than inside the communities [33]. In social network analysis, community detection is a basic step in understanding the structure, function and semantics of networks [118]. Community analysis is performed in two separate phases: first, detection of meaningful community structure from a network, and second, evaluation of the appropriateness of the detected community structure [72]. Systematic deviations from a random configuration allow us to define a quantity called modularity, that is a measure of the quality of partitions. Newman-Girvan modularity considers only the degree of nodes as a null model which is equivalent to rewiring the network whilst preserving the degree sequence [147, 60]. This random model overlooks the spatial nature of the network thus modules are blind to spatial anomalies and fails to uncover modules determined by factors other than mere physical proximity [41], which is the reason why several distance-dependent null models have been proposed recently [148, 72, 41, 149].

The goal is to use the tools of network community detection to evaluate the attractiveness of the elements of settlement hierarchies (towns, statistical sub-regions, counties, regions) based on their modularities as well as internal and external connection densities. The internal connections of the ownership network through the point of view of Newman-Girvan, spatial and gravity based null models was studied. As the modularity is based on the difference between the actual and evaluated values of weight of edges, the more accurately describes the null model the real spatial network, the total modularity tends to be zero, so the modules highlight the hidden structural similarities. A visualization technique was developed to analyse these unknown effects on community structure which can explain the attractiveness of a settlement/region.

Besides measuring the attractiveness, the Louvain community detection algorithm [150, 151] was utilized to identify closely related regions. The complete investment network of Hungarian companies was examined to explore how the ownership connections are geographically distributed, what is the structure of the network, what are the common connection directions as well as how the extracted information is correlated to the settlement hierarchy. The studied database contains information about the owners and addresses of the companies.

The results highlight that distance dependence of the investment connections is more significant than was found in online social networks [137, 152, 133]. The analysis shows that the network is hierarchical and modular as well as shaped according to the settlement hierarchy, in which Budapest is the absolute centre, and the centres of counties function as hubs.

The outline of this chapter is as follows: Section 3.2.1 presents the company ownership network and the metrics related to attractiveness. Section 3.2.3 describes the null models designed in this work to measure modularity as well as handle physical proximity and presents how closely related regions can be explored based on the modularity-related merging of towns and sub-regions. The results and discussion are provided in Section 3.3.

3.2 Problem formulation: settlement hierarchy and community structure in personal investment patterns

3.2.1 Network representation of personal investment patterns

The proposed methodology is based on the analysis of a directed investment network represented by an asymmetric bi-adjacency matrix $\mathbf{A}^{[p,co]}$, which elements are defined as:

$$a_{i,j}^{[p,co]} = \begin{cases} 1 & \text{if the } i\text{-th person owns the } j\text{-th company} \\ 0 & \text{otherwise} \, . \end{cases} \tag{3.1}$$

As the addresses of the owners and their companies are known, connections between companies and their owners define ties between geographic locations. Investments are evaluated in this model through ownership relationships. The owner is interested in the effective operation of her/his company and therefore invests in it.

According to the levels of the settlement hierarchy, a four-level study can be defined to describe how towns, regions or counties are connected through company ownerships (see Fig. 3.1). Although companies also own shares in other companies, as we intended to study the attractiveness of economic regions based on personal investment decisions, we examined only companies that belong to individuals.

The levels of the settlement hierarchy $[l]$ are defined based on the nomenclature of territorial units for statistics classification (NUTS) and the two levels of local administrative units (LAUs):

$$l = \begin{cases} 1 & \text{town/settlement - LAU 2, formally NUTS 5 level} \\ 2 & \text{statistical sub-region - LAU 1, formally NUTS 4 level} \\ 3 & \text{small regions / counties, NUTS 3 level} \\ 4 & \text{regions of regional policies, NUTS 2 level} \end{cases} \tag{3.2}$$

(Please note, for simplicity, the term “town” is used for all cities and villages.)

People and their companies are assigned to geographic regions by the $\mathbf{A}^{[co,l]}$ and $\mathbf{A}^{[p,l]}$ incidence matrices, which elements are defined as:

- $a_{i,j}^{[co,l]}$ with element one if the headquarter of the i-th company is situated in the j-th geographic region at the level l of the settlement hierarchy,
- $a_{i,j}^{[p,l]}$ with element one if the i-th person is situated in the j-th geographic region at the level l of the settlement hierarchy,

so the directed weighted network that defines the number of investment connections between the regions can be defined as:

$$\mathbf{A}^{[l]} = \left(\mathbf{A}^{[p,l]}\right)^T \times \mathbf{A}^{[p,co]} \times \mathbf{A}^{[co,l]} \,. \tag{3.3}$$

Although companies may have many local divisions, the links between the towns are defined only by connecting the permanent addresses of the owners and the location of the headquarter. This arrangement results in a transparent and easily interpretable network as people and companies are assigned to only one location. The resultant network describes how investments unite the locations, e.g. the adjacency matrix $\mathbf{A}^{[1]}$ defines the number of links between the towns, and the degrees of the nodes represent the number of incoming and outgoing investments to the j-th and from the i-th town, respectively:

$$k_j^{[l,in]} = \sum_i a_{i,j}^{[l]} \tag{3.4}$$

$$k_i^{[l,out]} = \sum_j a_{i,j}^{[l]} \,. \tag{3.5}$$

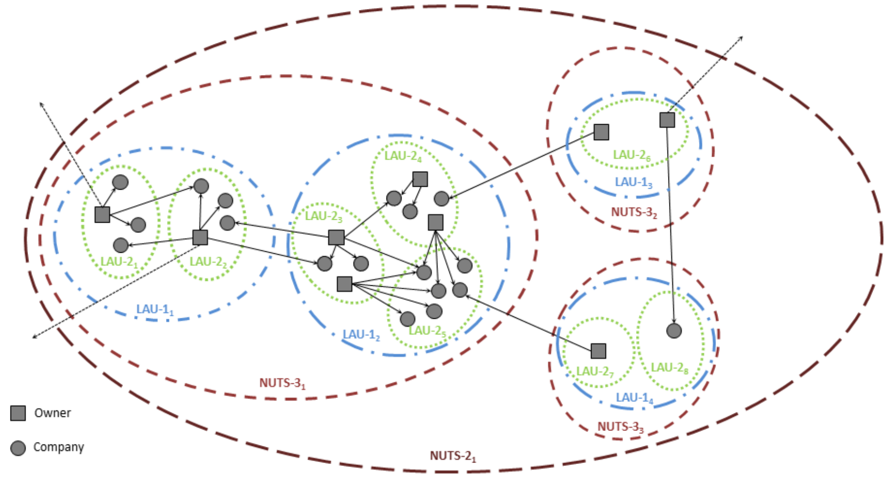


Figure 3.1 Company-ownership relations connect the elements of the settlement hierarchy (Settlement (LAU 2), statistical sub-region (LAU 1), small-region (NUTS 3), region (NUTS 2)).

The total number of investment relationships is equal to the sum of the edge weights of the networks:

$$L = \sum_i \sum_j a_{i,j}^{[l]}\,, \forall l \tag{3.6}$$

where i and j represents the indices of the geographic regions at the level l of the settlement hierarchy.

It should be noted that as L represents the total number of connections; its value is independent of at which hierarchy level the edge weights are summarised.

Similarly, the total number of companies and investors can be calculated by summing the number of companies and people at any hierarchy level, respectively:

$$N^{[co]} = \sum_{j=1} n_j^{[l,co]}\ , N^{[p]} = \sum_{j=1} n_j^{[l,p]}\ , \forall l\,, \tag{3.7}$$

where j represents the index of the geographic regions at the level l of the settlement hierarchy.

As people and companies are assigned only to one geographical regions with the $\mathbf{A}^{[co,l]}$ and $\mathbf{A}^{[p,l]}$ incidence matrices, the number of people and companies at the j-th region of the $[l]$-th level of the settlement hierarchy can be calculated as:

$$n_j^{[l,co]} = \sum_i a_{i,j}^{[co,l]} \tag{3.8}$$

$$n_j^{[l,p]} = \sum_i a_{i,j}^{[p,l]}\,. \tag{3.9}$$

The number of internal and external links of the network and the analysis of the local densities can be used to measure the attractiveness of the regions (see the Subsection 3.2.2). Then the following Subsection 3.2.3 of the work focuses on models that can be used to explore the communities in the network.

3.2.2 Internal and external connection-based evaluation

Finding community structure means the assignment of the nodes into groups, where within the nodes are highly connected and across the nodes of the communities they are much loosely connected to each other [153].

The density of the whole network can be calculated as:

$$D = \frac{L}{N^{[p]} N^{[co]}}\,. \tag{3.10}$$

while the internal density of the region is calculated as:

$$D_i^{[l,in]} = \frac{a_{i,i}^{[l]}}{n_i^{[l,p]} n_i^{[l,co]}} \,. \tag{3.11}$$

$D_i^{[l,in]}/D$ compares internal complexity of the regions to the whole network.

The probability of an external tie, in other words the external density, can be calculated a similar fashion,

$$D_i^{[l,ex]} = \frac{\sum_{i \neq j} a_{i,j}^{[l]}}{N^{[l,p]} \left(N^{[l,co]} - n_i^{[l,co]} \right)} \,. \tag{3.12}$$

where $N^{[l,co]} - n_i^{[l,co]}$ represents the number of companies that are outside of the i-th region at the $[l]$-th level of the settlement hierarchy.

To evaluate the openness as a measure of the attractiveness of the region, the ratio of the external to internal probabilities can be defined as:

$$O_i^{[l]} = \frac{D_i^{[l,ex]}}{D_i^{[l,in]}}. \tag{3.13}$$

Apart from taking into account internal and external links, the direction of the connections can be considered. Expansion computes for the number of edges pointing outside the community [72]:

$$E_i^{[l]} = \frac{\sum_i a_{i,j}^{[l]} - \sum_i a_{i,i}^{[l]}}{n_i^{[l,p]}} \tag{3.14}$$

Similarly, the ability of a community to collects links can be determined by the normalized number of links that point inside the community:

$$LCA_i^{[l]} = \frac{\sum_j a_{i,j}^{[l]} - a_{i,i}^{[l]}}{n_i^{[l,co]}} \tag{3.15}$$

Cut ratio is similar to the internal density as it computes the fraction of edges pointing out and the number of possible edges that are pointing outside the community:

$$CR_i^{[l]} = \frac{\sum_j a_{i,j}^{[l]} - a_{i,i}^{[l]}}{n_i^{[l,p]} \left(N^{[l,co]} - n_i^{[l,co]} \right)} \tag{3.16}$$

3.2.3 Evaluation of the community structure in the settlement hierarchy

The key idea of the methodology is that geographical regions can be interpreted as non-overlapping communities of investors and companies as they belong to exactly one region among the set of these regions on the l-th level of the hierarchy, $C^{[l]} =$

$\{C_1^{[l]}, C_2^{[l]}, \cdots, C_l^{[l]}, \ldots, C_{n_{c,nk}}^{[l]}\}$.

From the view of a community, the external degree is the number of links that connect the i-th community to the rest of the network, while the internal degree is the number of links between companies and owners in the same community, in other words, at the same location at the l-th level of the hierarchy. Recently a wide variety of $f(C)$ metrics have been proposed to evaluate the quality of communities on the basis of the connectivity of their nodes [72]. The following subsections will demonstrate how these metrics can be interpreted to evaluate the attractiveness of geographical regions.

3.2.3.1 Modularity of a region and level of a settlement hierarchy

Classical modularity optimization-based community detection methods utilize $f(C)$ metrics that are based on the difference between the internal number of edges and their expected number. [60, 154]

$$f(C) = (\text{fraction of edges within communities}) - (\text{expected fraction of such edges})\,. \tag{3.17}$$

In the case of the proposed directed network this difference can be formulated as:

$$f(C^{[l]}) = \frac{1}{L}\sum_{i,j}\left(a_{i,j}^{[1]} - p_{i,j}^{[1]}\right)\delta\left(C_i^{[l]}, C_j^{[l]}\right) \tag{3.18}$$

where $p_{i,j}^{[1]}$ represents the number of estimated investments proceeding from the i-th to the j-th town and $\delta\left(C_i^{[l]}, C_j^{[l]}\right)$ is the Kronecker delta function that is equal to one, if the i-th and j-th towns are assigned to the same region on the l-th level of the hierarchy (e.g., $\delta\left(C_A^{[2]}, C_B^{[2]}\right) = 1$ when towns A and B are situated in the same statistical sub-region).

The modularity of the partition $C^{[l]}$ can be calculated as the sum of the modularities of the $C_c^{[l]}$, $c = 1, \ldots, n_c^{[l]}$ communities:

$$M_c^{[l]} = \frac{1}{L}\sum_{(i,j)\in C_c^{[l]}} (a_{i,j}^{[1]} - p_{i,j}^{[1]})\,. \tag{3.19}$$

The value of the modularity $M_c^{[l]}$ of a cluster/region $C_c^{[l]}$ can be positive, negative or zero. If it is equal to zero, the community has as many links as the null model predicts. When the modularity is positive, then the $C_c^{[l]}$ subgraph tends to be a community that exhibits a stronger degree of internal cohesion than the model predicts.

Using the proposed matrix representation, the calculation of the internal links at a given level of the hierarchy is straightforward, so the modularity can be easily calculated based on the diagonal elements of the adjacency matrices of the network and its null model:

$$f(C^{[l]}) = \sum_{c=1}^{n_c^{[l]}} M_c^{[l]} = \frac{1}{L}\sum_c a_{c,c}^{[l]} - \frac{1}{L}\sum_c p_{c,c}^{[l]} \tag{3.20}$$

where $a_{c,c}^{[l]}$ represents the number of internal links in the c-th community/region on the l-th hierarchy level while $p_{c,c}^{[l]}$ is the expected number of these internal links calculated by the null model.

3.2.3.2 Null models for representing regional attractiveness

The critical element of the methodology is how the $p_{i,j}^{[1]}$ connection probabilities of the towns are calculated. The most widely applied *null model* is the random configuration model which calculates the edge probabilities assuming a random graph conditioned to preserve the degree sequence of the original network:

$$p_{i,j}^{[1]} = \frac{k_i^{[1,out]} k_j^{[1,in]}}{L} \tag{3.21}$$

This randomized null model is inaccurate in most real-world networks [149].

As we measure the attractiveness of the regions based on the probability of link formation, it is beneficial to utilize attractiveness-related variables in the model as well as take the distance-dependent link structure into account. Firstly, we generalize the model by defining the node importance measures I_i^{out} and I_j^{in}:

$$p_{i,j}^{[1]} = \gamma \; I_i^{out} I_j^{in} \, . \tag{3.22}$$

As is expected from the null model, to fulfil the following equality:

$$\sum_{i,j} p_{i,j}^{[1]} = \sum_{i,j} a_{i,j}^{[1]} = L \, , \tag{3.23}$$

the importance measures are normalized as $\sum_i I_i^{out} = 1$ and $\sum_j I_j^{in} = 1$:

$$I_i^{out} = \frac{x_i^{\alpha}}{\sum_j x_j^{\alpha}} \; , I_j^{in} = \frac{x_j^{\beta}}{\sum_i x_i^{\beta}} \tag{3.24}$$

where the parameters $\alpha, \beta > 0$ reflect the importance of the x_i and x_j variables used to express the probability of forming an edge from the i-th to the j-th node. Please note, when $\alpha = 1$ and $\beta = 1$, $x_i = k_i^{[1,out]}$, $x_j = k_j^{[1,in]}$, and $\gamma = L$, the model is identical to the random configuration model of a weighted directed graph.

To model the probability of distance-dependent link formation the model defined by Eq. 3.22 is extended by a deterrence function $f(d_{i,j})$ which describes the effect of space [131]:

$$p_{i,j}^{[1]} = \gamma \; I_i^{out} I_j^{in} f(d_{i,j}) \tag{3.25}$$

The function $f(d_{i,j})$ can be directly measured from the data by a binning procedure similarly to that used in [41]:

$$f(d) = \frac{\sum_{i,j|d_{i,j}=d} a_{i,j}^{[1]}}{\sum_{i,j|d_{i,j}=d} I_i^{out} I_j^{in}} \tag{3.26}$$

whose function is proportional to the weighted average of probability $(1/\gamma)\ a_{i,j}^{[1]}/\left(I_i^{out} I_j^{in}\right)$ of a link existing at distance d.

When the distance dependence of the connection probability is handled by an explicit function, various modifications of the gravity law-based configuration model can be defined: $f(d) = 1/d_{i,j}^{\delta}$ [145, 155], $f(d) = \exp(-d_{i,j}/\delta)$ [156], or $f(d) = d_{i,j}^{-\delta} \exp(-d_{i,j}/\kappa)$ [157].

To ensure that the sum of the expected number of links is equal to L (see Eq. 3.23), in this distance-dependent model γ should be normalized as:

$$\gamma = \frac{L}{\sum_{i,j} I_i^{out} I_j^{in} f(d_{i,j})} \tag{3.27}$$

Several models can be defined based on what kind of indicators are selected in the model. When the nodes are considered to be equally important, in other words, $I_i = I_j = 1$, only the distance determine the link formation probability, $f(d_{i,j})$. The importance of the nodes can be interpreted as the number of investors and companies, so $I_i = \left(n_i^{[l,p]}\right)^{\alpha}$ and $I_j = \left(n_j^{[l,co]}\right)^{\beta}$. The null model can be defined based on the random configuration model, which results in the selection of the variables as $I_i = \left(k_i^{[l,out]}\right)^{\alpha}$ and $I_j = \left(k_j^{[l,in]}\right)^{\beta}$. Finally, socio-economic indicators, like the number of inhabitants, or their complex combinations can be utilized.

When $f(d) = 1/d_{i,j}^{\delta}$, the parameters α, β, δ can be estimated as a regression problem. The identified parameters indicate the sensitivity, i.e. importance, of the variables that can be sorted by their importance as suggested in classical gravity-low based studies, like in [131].

3.2.3.3 Economic relations of the regions

Connections that interlink communities indicate their relationships and possibilities to merge modules/regions that are strong connected. We combine regions and determine the gain of the merged modularity similar way to the Louvain community detection algorithm [150]. The $\Delta M_{i,j}$ modularity change obtained by merging the i-th and j-th communities can be calculated as the difference between the actual and predicted number of interlinking nodes:

$$\Delta M_{i,j}^{[l]} = \frac{1}{L}\left(a_{i,j}^{[l]} - p_{i,j}^{[l]}\right) + \frac{1}{L}\left(a_{j,i}^{[l]} - p_{j,i}^{[l]}\right) \tag{3.28}$$

The resultant symmetric modularity gain matrix can be calculated as:

$$\Delta \mathbf{M}^{[l]} = \left(\mathbf{B}^{[l]}\right)^{T} + \mathbf{B}^{[l]} \tag{3.29}$$

where $\mathbf{B}^{[l]} = \mathbf{A}^{[l]} - \mathbf{P}^{[l]}$ is the so-called modularity matrix [147].

The Louvain community detection algorithm moves a node i in the community for which the gain in modularity is the largest. If no positive gain occurs, i remains in its original community. After merging the nodes/regions, a new network is constructed whose nodes are in the communities identified earlier. This method can be used to explore regions (modules) formed by the elements of the l-th settlement hierarchy with different null models. Although model-based communities can be identified by this approach and compared to regions of a larger hierarchy level as modules of ground truth, the main goal of the analysis of $\mathbf{M}^{[l]}$ is to measure the strength of relationships between the regions.

The following section demonstrates the applicability of the previously presented toolset in the analysis of the ownership network of Hungarian companies.

3.3 Results and discussion

3.3.1 Description of the studied dataset

The studied dataset represents $L = 1,077,090$ ownership relations between $N^{[p]} = 531,249$ people and $N^{[co]} = 868,591$ Hungarian companies in 2013. It should be noted only less than 10% of the ownership connections are defined based on how companies possess shares in other companies, so, although only personal investments are studied, the results reflect the attractiveness of the towns and regions as the generated network covers more than the 90% of the investment type connections.

The owners and companies were assigned to settlements, and the related settlement hierarchy covers $3,155$ towns (level LAU 2, former level NUTS 5), 175 statistical subregions (level LAU 1, former level NUTS 4), 20 small regions/counties in level NUTS 3, and 7 regions in level NUTS 2.

74% of the connections remain within the borders of the towns, which also reflects the high degree of modularity of the network (for more details see Table 3.1). $302,781$ connections are within Budapest and $45,559$ point out of the city, while $89,944$ connections point into the capital. The map of the regional connections between the people and companies can be generated using the obtained connectivity matrix and the latitudes and longitudes of the towns (see Fig. 3.2). It can be seen that the network reveals a hierarchical and modular structure reflecting that the Hungarian economy is concentrated around the capitals of the counties and Budapest, the capital of the country. The majority of the companies are situated in these locations; consequently, the network follows the structure of online social networks [152], in other words is also structured according to the settlement hierarchy, in which Budapest is the absolute centre of the network and the centres of counties also function as hubs.

Table 3.1 Number of edges inside the settlement hierarchies

	Town-level (LAU 2)	sub-Region-level (LAU 1)	County-level (NUTS 3)	Region-level (NUTS 2)
Number of nodes, N	3,111	175	20	7
Number of internal ties	797,492	846,309	893,559	969,995
Number of external ties	279,598	230,781	183,531	107,095

3.3.2 Network topology analysis

The degree distribution was determined in all levels of the settlement hierarchy by following the methodology presented in Ref. [32]. Fig. 3.3 shows that the distribution shows small degree saturation and high-degree cutoff. Several distribution functions were fitted. The two-sided Voung's test statistic [111] showed that exponential and Poisson distributions which reflect the randomness of connections could be rejected. According to this test, the power-law and log-normal distribution cannot be rejected. The estimated parameters are shown in Table 3.2. The power-law distribution of the incoming and outgoing connections reflect the preferential attachment-type structure of the network.

The Poisson distribution is characteristic of the theoretical random networks presented by Erdős-Rényi [30], in which the degrees of the vertices are around a value. The probability that two vertices are connected is in each case p, and such networks

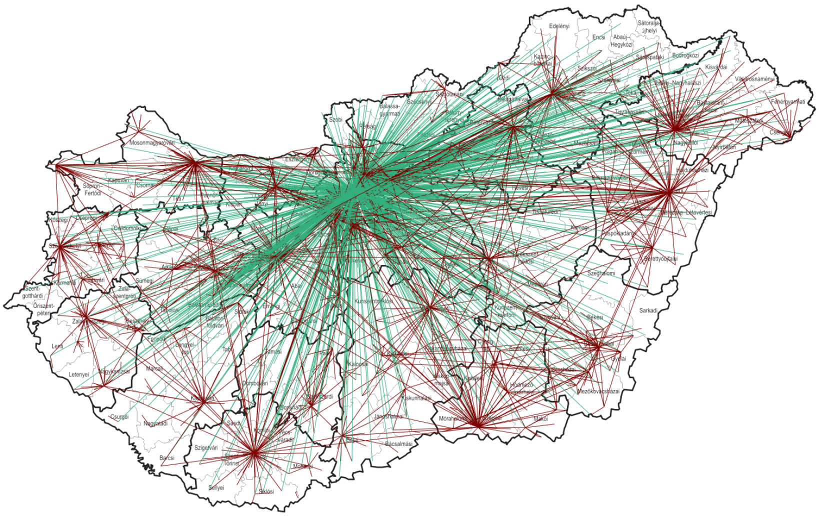


Figure 3.2 Map of the town-level company ownership network. Edges with more than 10 ownership connections are shown. Edges connected to the capital (Budapest) are denoted by green lines.

are truly random. Edge formation does not take into account any properties of the vertices. In most real networks new nodes prefer to link to the more connected nodes, the process called preferential attachment. The probability that two vertices are connected is highly depend on the degree of nodes and hubs are more attractive than other nodes.[92] This means in the context of the network under study that connections with larger cities are preferred.

Table 3.2 Parameters of the power-law distributions fitted to networks at different settlement hierarchy levels.

Distribution	k_{sat}	k_{cut}	γ
$k_j^{[1,in]}$ (LAU 2)	120	15061	1.85
$k_j^{[1,out]}$ (LAU 2)	138	19709	1.87
$k_j^{[2,in]}$ (LAU 1)	1974	392724	2.04
$k_j^{[2,out]}$ (LAU 1)	2070	348339	2.04
$k_j^{[3,in]}$ (NUTS 3)	19693	392724	2.54
$k_j^{[3,out]}$ (NUTS 3)	20401	348339	2.49
$k_j^{[4,in]}$ (NUTS 2)	74161	557112	3.31
$k_j^{[4,out]}$ (NUTS 2)	77042	519967	3.35

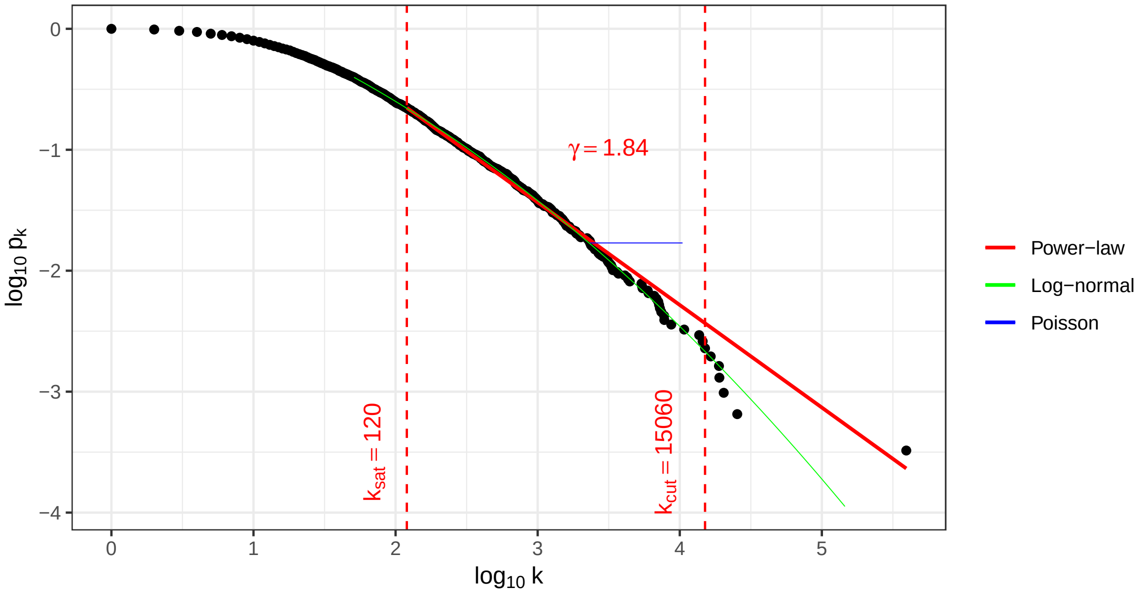


Figure 3.3 Distribution of the $k_j^{[1,in]}$ edges at the LAU-2 settlement hierarchy level.

In hierarchical networks, nodes with high degree tend to connect to nodes that are less connected to others [158]. Therefore, the hierarchical structure of the network is reflected by the dependence of the local clustering coefficient $C(k)$ on the degree of the nodes. As Fig. 3.4 shows, $C(k)$ decreases with increasing k with $C(k) \approx k^{-0.3}$ which indicates the hierarchical structure of the network [158, 159]. The assortativity of network (see Figure 3.5) also strength the hierarchical structure because slightly disassortative structure shows that high degree nodes tend to connect with low degree nodes and vica versa.

Based on the mutually reinforcing results, the network shows a structure in which low-degree vertices tend to connect to regional centres, and the relationships of hubs are somewhat less likely than the formation of small-to-large type connections. The structure of the subregional, county and national centres is also visible from the indicators of the network topology, where regional hubs tend to connect with smaller settlements instead of each other.

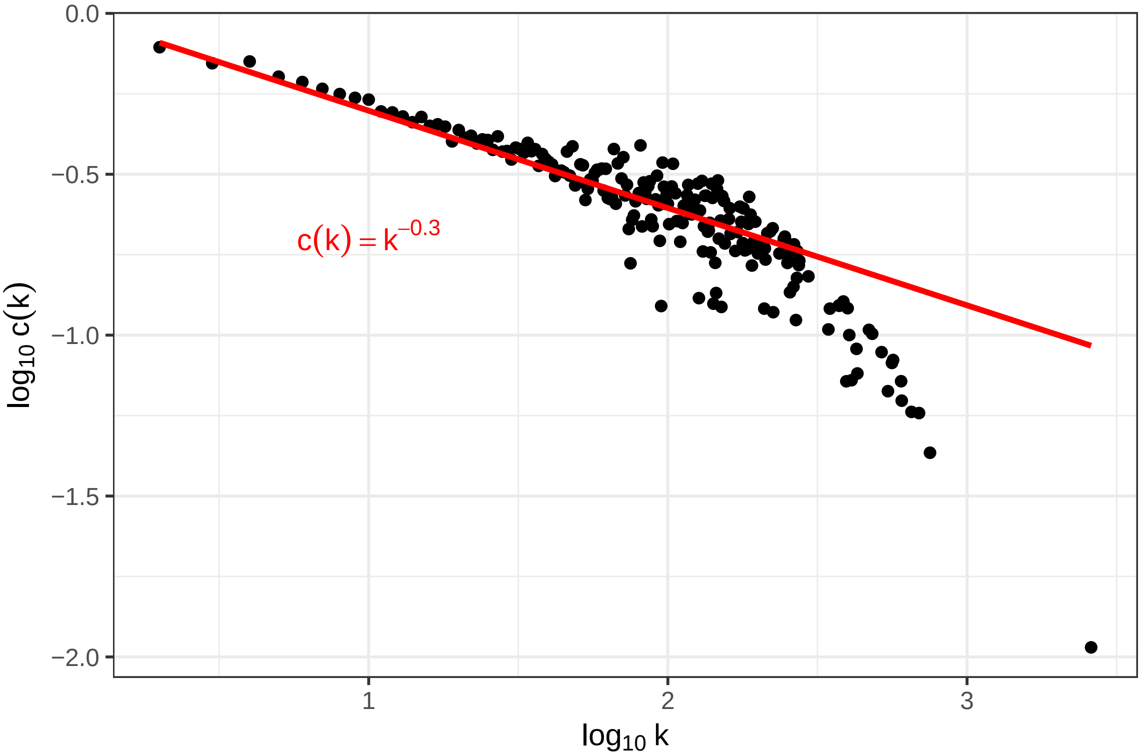


Figure 3.4 Local clustering coefficient as a function of the $k_j^{[1,in]}$ node degrees

3.3.3 Measuring attractiveness

The densities inside towns and regions can highlight the modular structure of the company-ownership network. As shown in Figure 3.6, these densities are significantly higher in most sub-regions and a negative correlation exits between the size of the regions and the number of their inner connections ($r = 0.298$, $p < 10^{-4}$). The result shows that in the case of smaller settlements, the enterprises are typically owned by a local owner. In contrast, in the case of larger towns, the proportion of entrepreneurs outside the city is significantly higher.

As illustrated by the results, smaller locations are much more isolated than larger ones, like Budapest. The same result is obtained by the analysis of the external density-based opennes measure which we consider as a main measure of attractiveness (see

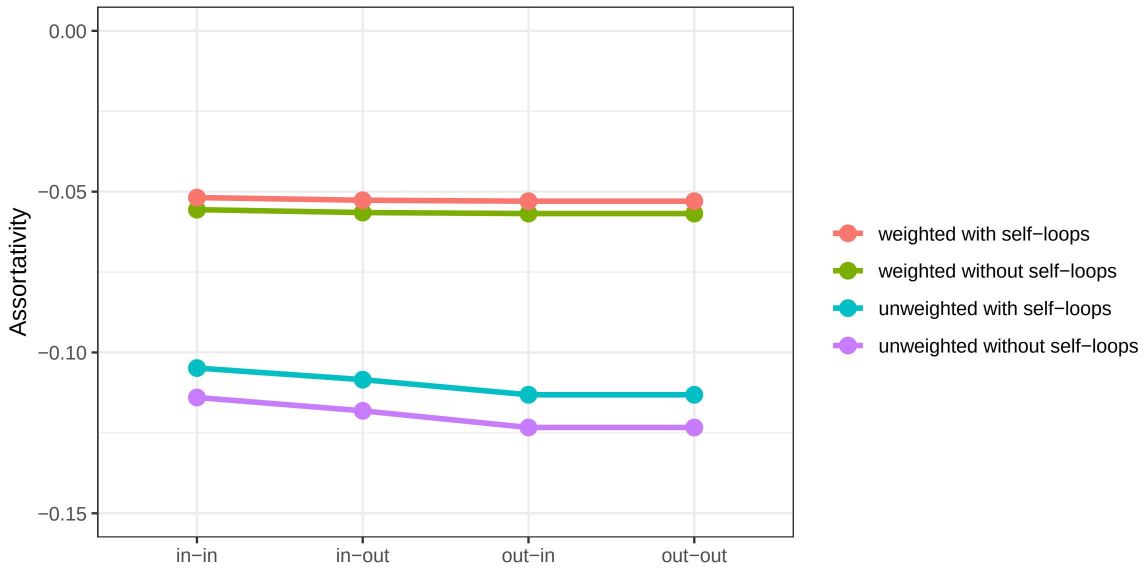


Figure 3.5 Degree correlations of ownership network in LAU-2 settlement hierarchy

Subsection 3.2.2 for more details). As is shown in Fig. 3.7, bigger regions exhibit lager openness values reflecting their higher degree of attractiveness ($r = 0.94$, $p < 10^{-10}$).

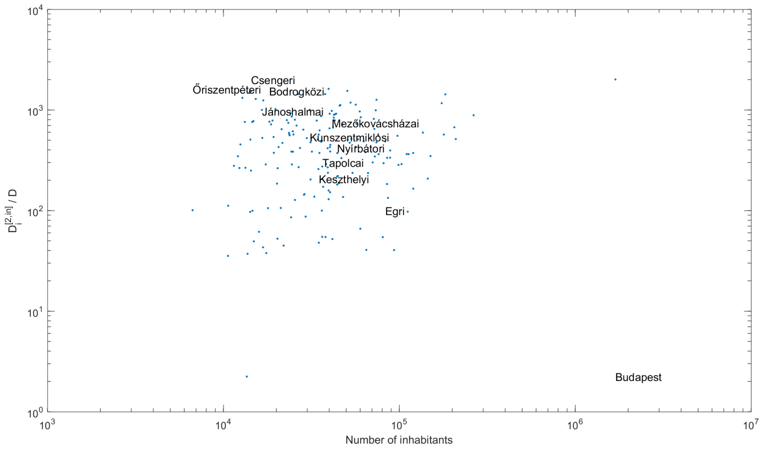


Figure 3.6 Network density as a function of the number of inhabitants on the level LAU 1.

3.3.4 The effect of geographical distance

To address the effect of distance decay on link formation, the observed ties between the towns were compared with their expected number calculated from a probabilistic model.

A resolution of 10 km was used for binning the distance distribution (see Fig. 3.8). The exponent of distance decay according to our data is -1.1057. It should be noted that the effect of the capital city is so high, that the probability of forming connections

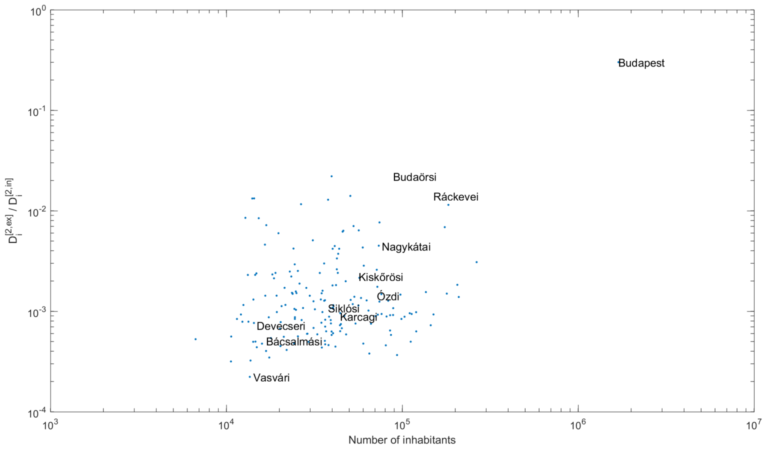


Figure 3.7 Openness of small regions (LAU 1 level) as a function of the number of their inhabitants.

with Budapest is slightly less distance-dependent, the exponent of distance decay with regard to these connections is only -0.6385.

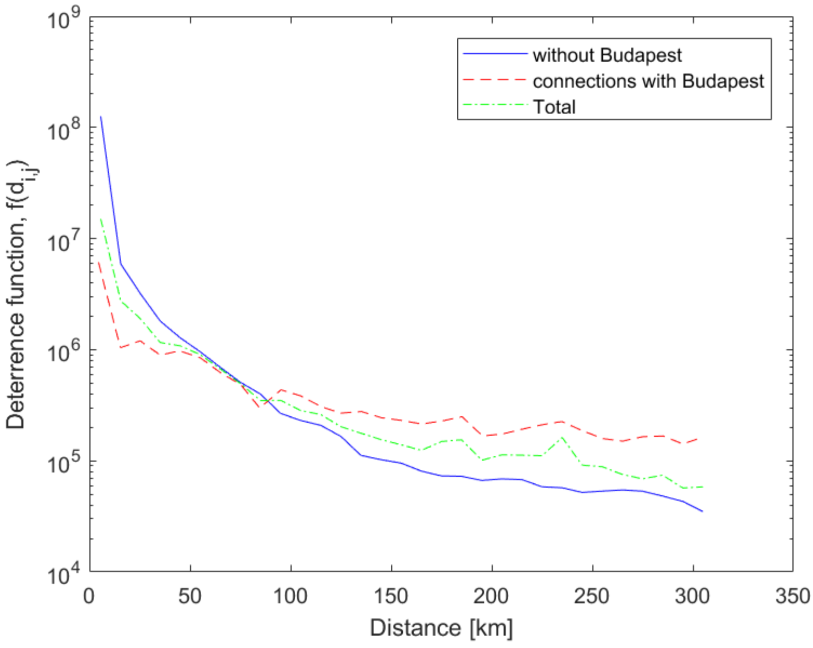


Figure 3.8 Empirically derived deterrence function determined by Eq. (3.26), where $I_i^{[in]} = n_i^{[1,p]}, I_j^{[in]} = n_j^{[1,co]}$

The distance-dependent link formation probability can be explained by the costs of establishing and maintaining the connections are also distance-dependent. This assumption can be confirmed by that the distance has a much stronger effect on investment ties than on online social networks in Hungary (where the exponent of distance decay is -0.6) [152], probably since the cost of keeping connections is less dependent on distance than the management of a company far from the permanent address of the owner.

3.3.5 Comparison of the null models

Based on the utilized distance function three different types of models can be defined. When $f(d)$ is a deterrence function defined by Eq. 3.26, the models are denoted as $p^{spa} = \gamma I_i^{out} I_j^{in} f(d)$. $p^{\alpha,\beta} = \gamma (I_i^{out})^{\alpha} (I_j^{in})^{\beta} f(d)$ represents the parametric version of this model, when the exponents α,and β are optimized to achieve a more accurate approximation of connections between towns. The objective function of optimization described with 3.30. $p_{i,j}^{grav} = \gamma (I_i^{out})^{\alpha} (I_j^{in})^{\beta} / d^{\delta}$ represents the gravity-type models.

$$\min_{\alpha,\beta,\gamma} E_m \left(\alpha, \beta, \gamma\right) = \frac{1}{L} \| \mathbf{A}^{[1]} - \mathbf{P}^{[1]} \|_2 \tag{3.30}$$

Five sets of I_i^{out}, I_j^{in} variables were defined, including simple metrics like the numbers of nodes and edges [115] in addition to socio-economic variables, like the number of inhabitants and Total Domestic Income (total income received by all sectors of the economy including the sum of all wages, profits and taxes, minus subsidies). Based on the combination of different variables and distance functions, 15 different models were identified.

As summarized in Table 3.3, by taking the distance into account the accuracy of the model is significantly improved. Among distance-dependent models the gravity model perform best. (In comparison, the accuracy of the distance independent random configuration model is 0.16494).

The Total Domestic Income (TDI) is one of the best indicators in terms of the minimal difference between the null model and the real network. The identified α, β and δ parameters reflect the importance of the I_i^{out}, I_j^{in} and d variables in the models (e.g. in the case where $I_j^{in} = TDI_j$ and $I_i^{out} = TDI_i$ the resultant nonlinear regression model is: $p_{i,j} = 0.12 \cdot \frac{(I_i^{out})^{0.37} \cdot (I_j^{in})^{0.81}}{d^{1.58}}$ (see Table 3.4), which can be interpreted as the number of connections between location i and location j is increased by 0.37% as a result of 1.0% growth of TDI in location i. Similarly, the number of connections between location i and location j is increased by 0.81% as a result of 1.0% growth of TDI in location j. According to the gravity-type models, the importance of the target/destination locations (β) is greater than the importance of the sources (α) regardless of how the strengths of the nodes are interpreted.

3.3.6 Evaluation of the modularities

As modularity-based community detection evaluates the set of $a_{i,j}^{[1]} > p_{i,j}^{[1]}$ edges (and the related nodes) whose weights are underestimated by the null model (see Eq.3.18), we designed a plot that compares $a_{i,j}^{[1]}$ with $p_{i,j}^{[1]}$ to highlight the set of potential edges that can be used to form communities.

Table 3.3 Performances of distance-dependent null models

Nodes/Null models	p^{spa}	$p^{\alpha,\beta}$	p^{grav}
$I_i^{out} = I_j^{in} = 1$	0.28100	0.28113	0.28093
$I_i^{out} = N^{[p]}, I_j^{in} = N^{[co]}$	0.08915	0.01359	0.00651
$I_i^{out} = k_i^{[1,out]}$, $I_j^{in} = k_j^{[1,in]}$	0.05759	0.01389	0.00642
I_i^{out} =Inhabitants$_i$, I_j^{in} =Inhabitants$_j$	0.12106	0.01456	0.00650
I_i^{out} =TDI$_i$, I_j^{in} =TDI$_j$	0.07142	0.01482	0.00644

Table 3.4 Coefficients of the parametric models that reflect the importance of the variables

	$p^{\alpha,\beta} = \gamma(I_i^{out})^{\alpha}(I_j^{in})^{\beta} f(d)$		$p_{i,j}^{grav} = \gamma(I_i^{out})^{\alpha}(I_j^{in})^{\beta}/d^{\delta}$		
Nodes/Parameters	α	β	α	β	δ
$I_i^{out} = N^{[p]}, I_j^{in} = N^{[co]}$	1.08373	0.91787	0.34984	0.67191	1.63711
$I_i^{out} = k_i^{[1,out]}$, $I_j^{in} = k_j^{[1,in]}$	1.05439	0.94455	0.35652	0.69045	1.59439
I_i^{out} =Inhabitants$_i$, I_j^{in} =Inhabitants$_j$	0.99347	1.15642	0.40654	0.88313	1.52391
$I_i^{out} = TDI_i, I_j^{in} = TDI_j$	0.98571	1.03669	0.37367	0.81425	1.58060

Four null models based on the $I_1 := k_i^{[1,out]}$ and $I_2 := k_j^{[1,in]}$ Newman and Girvan model are compared in Figure 3.9. In all models, the inner connections (represented by +) form a separate cluster which confirms that 74% of the connections remain within the borders of the towns. The first model (p^{NG}) shows that more inner connections exits than would be expected based on the random configuration network. The spatial models p^{Spat} and $p^{\alpha,\beta}$ handle the dependence on distance of the connections, so a slightly smaller difference is shown in the number of the experienced and expected inner connections. It is reflected in Fig. 3.10 that during the aggregation procedure the qualitative behaviour of the models does not change.

The difference between the expected number of interconnections is higher in the case of smaller settlements which indicates that small regions are not as attractive as would be expected from their number of nodes. The gravity model p^{Grav} estimates well the inner connections thanks to the exponents $\alpha = 0.35652$ and $\beta = 0.69045$ whose parameters effectively represent that the increase in the number of connections affects the attractiveness in a nonlinear fashion. This phenomenon is much more interesting when the utilized variables can be interpreted as economic potentials. When TDI is applied in the gravity model, $\alpha = 0.37367$ and $\beta = 0.81425$. These values and Fig. 3.11 confirm that gravity-based models behave similarly and, therefore, reflect the same mechanism of attractiveness.

Figure 3.9 at LAU2 level and Figure 3.10 at LAU1 level compares the goodness of fit as models estimate the connections of the towns and subregions. Modularity matrix, as the input of calculations of modularities, takes into account the differences between the real (A_{ij}) and expected (P_{ij}) number of connections of node i and j.

Different models evaluate the number of real connections which is shown in Figure 3.9. In this figure, Red line indicate the case when $A_{ij} = P_{ij}$, and points are the edges. Point above the red line are edges that overestimated by models ($A_{ij} < P_{ij}$, therefore $A_{ij} - P_{ij} < 0$). Points under the red line are underestimated by models ($A_{ij} > P_{ij}$, therefore $A_{ij} - P_{ij} > 0$). The underestimated edges hold more information for modularity optimisation algorithm which wants to find densely connected nodes.

The random configuration model (indicated with p^{NG} in Figure 3.9) underestimates the weights of edges of close nodes because of geographical distant play a significant role in the formation of edges. The non-parametric (p^{Spa}) and parametric ($p^{\alpha,\beta}$) version of random model which consider the empirical distance deterrence function have better accuracy. The weights estimated by the gravity model (p^{Grav}) are closest to real weights. The errors of models are summarized in Table 3.3. The estimation accuracy also differs significantly on the inner connections represented with + symbols.

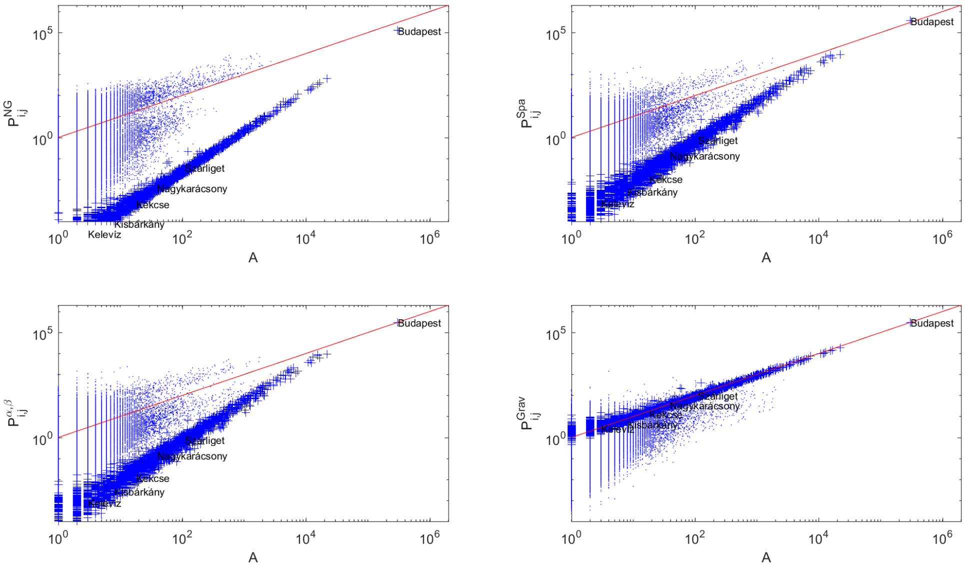


Figure 3.9 Comparison between the number of the edge weights $a^{[1]}_{i,j}$ and their estimated values $p^{[1]}_{i,j}$ generated by different null models on the town level (LAU 2) settlement hierarchy when $I^{out}_i = k^{[1,out]}_i$ and $I^{in}_j = k^{[1,in]}_j$. The + symbols represent the inner connections that form a separate cluster. This plot directly reflects the goodness of fit as the model estimates the connections of the towns.

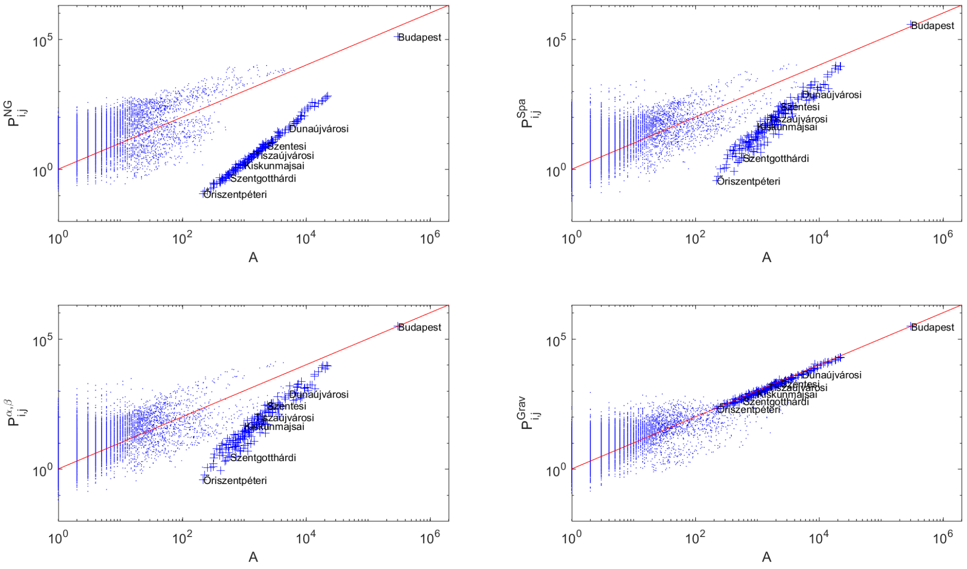


Figure 3.10 Comparison between the number of the edge weights $a^{[2]}_{i,j}$ and their estimated values $p^{[2]}_{i,j}$ generated by different null models at level LAU 1 of the settlement hierarchy when $I^{out}_i = k^{[1,out]}_i$ and $I^{in}_j = k^{[1,in]}_j$. The + symbols represent the inner-connections that form a separate cluster. This plot reflects that during the aggregation procedure, the qualitative behavior of the models does not change, furthermore, the same phenomena can be observed as in Fig. 3.9.

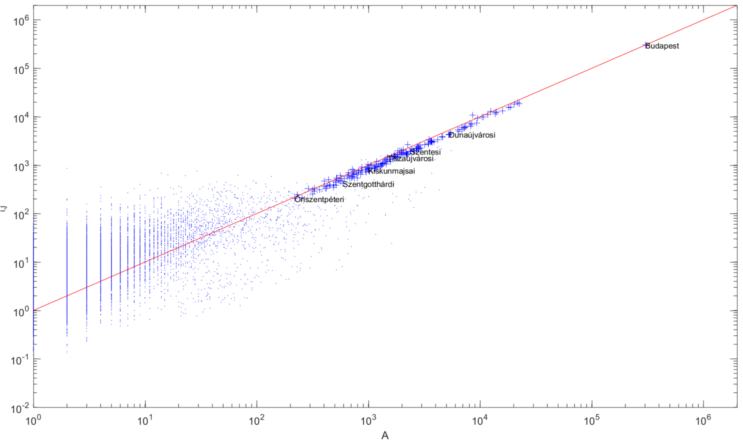
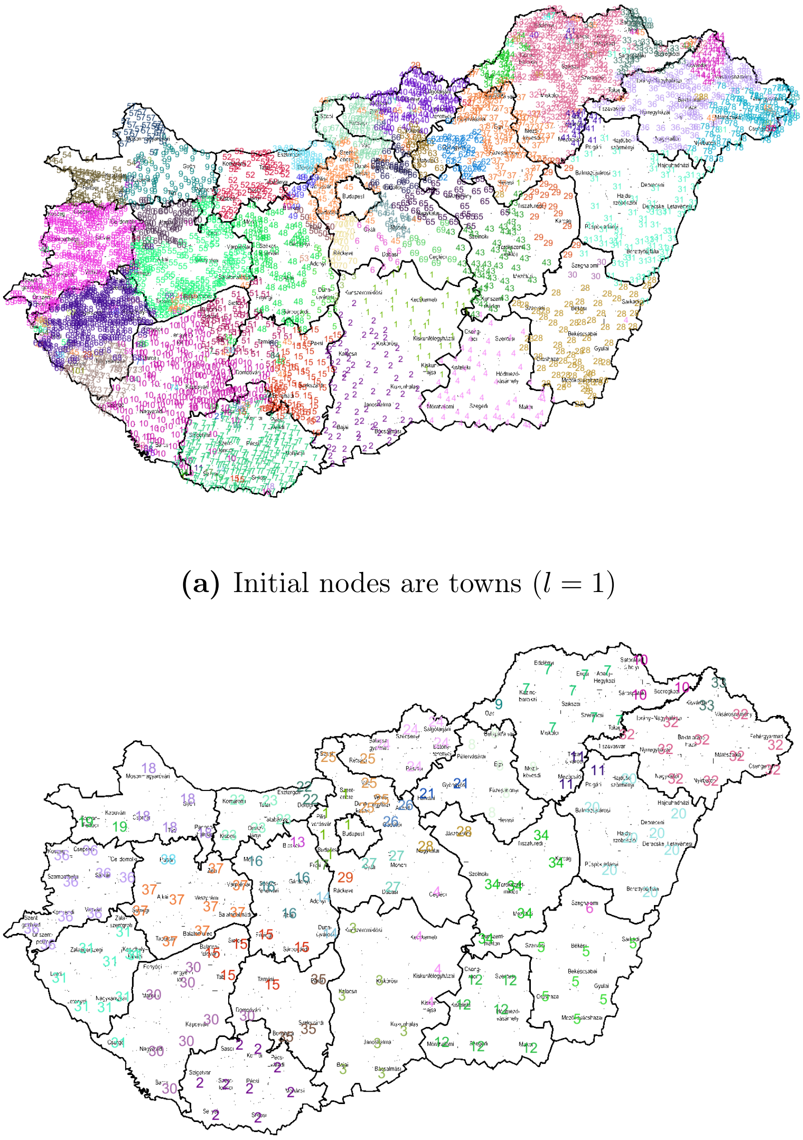


Figure 3.11 Comparison between the number of the edge weights $a^{[2]}_{i,j}$ and their estimated values $p^{[2]}_{i,j}$ generated by the gravity null model at level LAU 1 of the settlement hierarchy when $I^{out}_i = TDI_i$ and $I^{in}_j = TDI_j$. The + symbols represent the inner-connections that form a separate cluster.

3.3.7 Forming communities

Connections that interlink communities are indicative of their relationships. The effect of these interlinks can be studied by the change in modularity (see Eq. 3.28) expressed as $\Delta\mathbf{M}^{[l]} = \left(\mathbf{B}^{[l]}\right)^T + \mathbf{B}^{[l]}$.

To determine the community structure the MATLAB implementation [160] of the greedy Louvain algorithm [161] was used. Towns and sub-regions were used as an initial partitions $\mathbf{B}^{[l]}$. As is shown in Figure 3.12, the community structure formed based on the null model p^{NG} almost perfectly reconstructs the counties confirming that the settlement structure is reflected in terms of the personal investments.



(a) Initial nodes are towns ($l = 1$)

(b) Initial nodes are sub-regions (l =2)

Figure 3.12 Communities formed by the Louvain method and Newman-Girvan (NG) null model (I_i=k_i^{out} and I_j=k_j^{in}) reflect the settlement hierarchy as the resultant communities are almost identical to the counties.

Different null models provide different viewpoints with regards to community detection. The NG null model does not handle the distance dependence of the connections so the matrix $\mathbf{B}^{[l]} = \mathbf{A}^{[l]} - \mathbf{P}^{[l]}$ of the modelling errors reflects the distance dependence of the connections. Therefore, the resulted communities form spatial clusters. On the contrary, communities formed by the gravitational models reflect distance-dependent differences less. According to the resultant maps, the attractiveness of Budapest is highlighted as only small since closed regions were not assigned to the module of the capital (see Figure 3.13(a)). It is interesting to note that all the centres of counties were assigned to the community of Budapest in gravitational model which also confirms the hierarchical structure of the network. To highlight the hierarchical structure and increase the sensitivity of the model, a resolution parameter was introduced into the model that can be adapted to detect similar region-pairs as is shown in Figure 3.13(b).

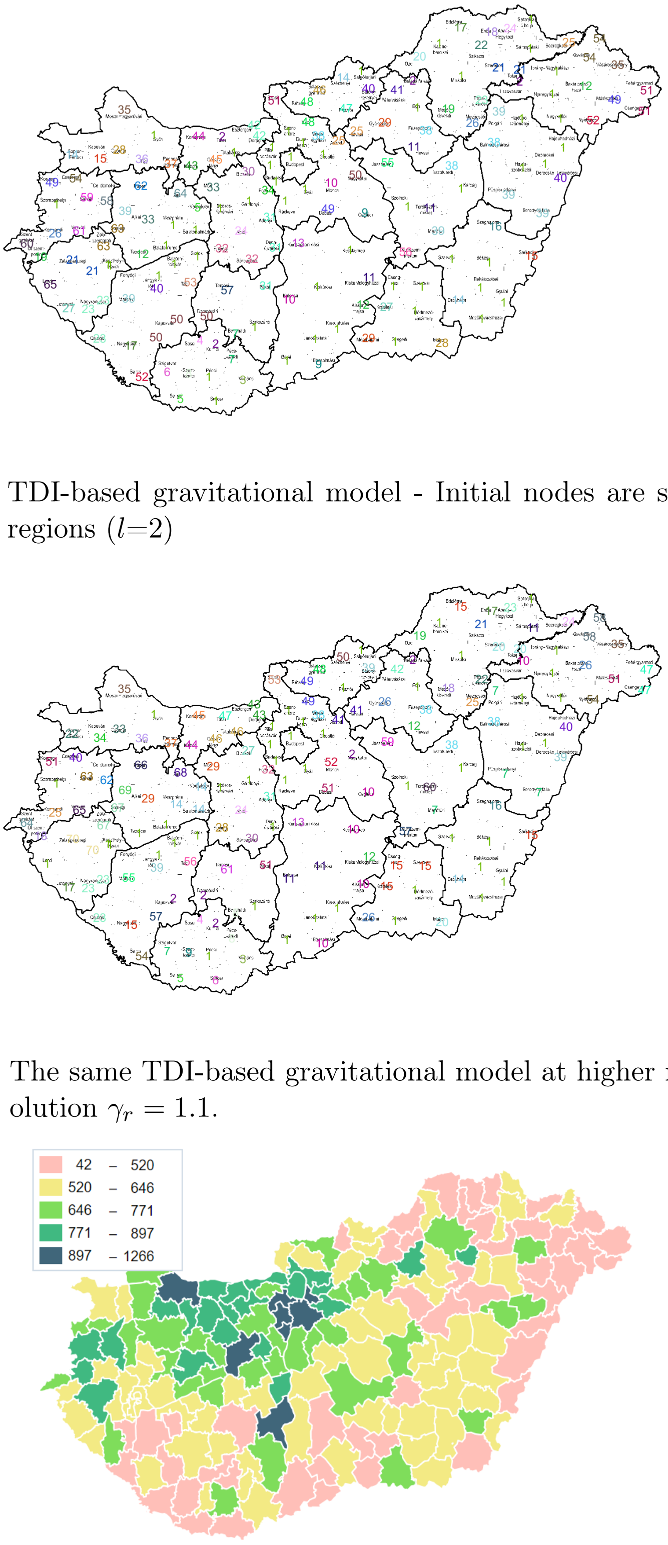
Communities formed with the NG null model (see Fig. 3.12) and the TDI-based gravity models (see Figure 3.13) significantly differ. The interpretation of the communities and these differences should rely on the understanding of the concept of the modularity. The utilised modularity detection algorithm generates partitions in which the links are more abundant within communities than would be expected from the employed model.

As the NG null model only uses the basic structural information encoded in the adjacency matrix, when the probabilities of the connections are dependent on distance, the resulted communities will represent closer geographical regions. As Table 3.1 and Figures 3.9 and 3.10 show, most of the connections remain within the county borders, so it is natural that the resultant 30 communities are almost identical to the counties.

Since the Hungarian road network reflects the administrative regions, it can be shown that the distance strongly affects the probability of the connections. This distance dependence of the connection probability can be incorporated into the null model by the proposed gravity model.

In this case, the resultant communities will reflect another unmodelled surplus in the number of connections. When the attractiveness and the distances are considered in the null model, the communities will reflect the additional economic attractiveness/similarity of the regions.

As Figure 3.13 shows, the algorithm generates a huge cluster of a well developed regions with Budapest, the larger cities and county seats with high TDIs; and several small communities related to isolated and less developed subregions.



(a) TDI-based gravitational model - Initial nodes are subregions (l=2)

(b) The same TDI-based gravitational model at higher resolution $\gamma_r = 1.1$.

(c) Spatial distribution of the TDI per capita (in 1000 HUF)

Figure 3.13 Communities formed by the Louvain method and gravitational null models reflect the attractiveness of Budapest as only less developed closed regions were not assigned to the module of the capital.

3.4 Conclusions

Regional policy-making and monitoring are firm-centered, incentive-based and state-driven. Personal investments define ties between geographical locations. We analyzed the structure of this ownership network and proposed a methodology to characterize regional attractiveness based on a set of null models identified to approximate the probabilities of link formation. According to the levels of the settlement hierarchy, a four-level study was conducted.

Based on the calculation of the internal and external network densities, several measures were proposed to evaluate the attractiveness and development of towns and geographical regions. The results indicate that small and less competitive regions have less internal connections, while larger cities are much more open.

To provide a more in-depth insight into the network, the dependence of link formation on distance was studied. The probability of connections between owners and their companies shows a much more rapid degree of distance decay than experienced in social networks. The attractiveness of the capital is so high that its connections are much less dependent on distance than other cities.

Based on the combination of three deterrence models and five sets of indicators, 15 different null models were identified besides the classical Newman-Girvan random configuration model. Communities statistically have more significant edge weights that would be wired according to the null model. As it was highlighted that underestimated link probabilities are the sources of modularity, a scatter plot was designed to visualize how the null model approximates the real structure of the network.

The identification of gravity-type models highlighted that link formation is non-linearly dependent on the studied variables. Furthermore, the target nodes are much more important when determining the probability of link formation than the source nodes which also confirms why the structural analysis of company ownership networks can be used to measure regional attractiveness.

The Louvain community detection algorithm was applied to form clusters of cities and sub-regions and compared the resultant communities to administrative regions. When the null model more closely approximates the real structure of the network, then the modularity is expected to be lower. As community detection forms modules which internal link densities are significantly higher than would be expected from the applied null models, spatial clusters that were highlighted by the distance independent random configuration model are almost identical to the counties. Communities generated based on the gravitational models - which correctly estimate the number of internal nodes and the dependence of link formation on distance - exploited the attractiveness of the capital, as they form a massive cluster that includes most of the centres of each county, bigger cities and the competitive tourist regions, while the remaining small clusters reflect isolated regions that are less developed and less attractive.

3.5 Contributions to Industry 4.0 issues

Industry 4.0 solutions require a lot of investment capital. My results contribute to the evaluation of the spatial movement of investment capital, and the exploration of attractive regions. The proposed methodology is appropriate to characterize regional attractiveness based on a set of null models that avoids the noisy effect of geographical distance.

The results show that the movement of investment capital is strongly distance-dependent. In small settlements, investors are more local, while as the size of the settlement increases, the economy becomes more open. The attractiveness of the capital is mostly related to similarly developed regions.

Chapter 4

Evaluation of network, clusters and node characteristics with overlapping dimensions of multidimensional edges

Abstract Network analysis can be applied to understand organisations based on patterns of communication, knowledge flows, trust, and the proximity of employees. A multidimensional organisational network was designed, and association rule mining of the edge labels applied to reveal how relationships, motivations, and perceptions determine each other in different scopes of activities and types of organisations. Frequent itemset-based similarity analysis of the nodes provides the opportunity to characterize typical roles in organisations and clusters of co-workers. A survey was designed to define 15 layers of the organisational network and demonstrate the applicability of the method in three companies.

The novelty of our approach resides in the evaluation of people in organisations as frequent multidimensional patterns of multilayer networks. The results illustrate that the overlapping edges of the proposed multilayer network can be used to highlight the motivation and managerial capabilities of the leaders and to find similarly perceived key persons.

4.1 Introduction

In the early 1980's Tichy suggested that organisational research should incorporate a network perspective [162, 163]. In the 1990's six themes (turnover and absenteeism, power, work attitudes, job design, leadership, motivation) dominated the research of micro-organisational behaviour [164]. Researchers have highlighted that centrality in advice networks may differ from that of friendship networks. Advice network centrality is a better indicator of the "real" hierarchy [165], because individuals may seek out advice from others who they would not consider leaders, and may perceive leaders whom they would not necessarily consider going to for advice [166]. Social influence derived from friendship networks has stronger effects on job-satisfaction [167] because social network relationships are likely to affect the performance and receipt of organisational citizenship behaviour (OCBs), which includes attitudes like job satisfaction. The spread of OCBs in organisations may be facilitated or hindered by social relationships. [168].

Social Network Analysis (SNA) is widely used to support these studies. As the attitude of the members of social networks attitudes might influence each other, predicting their behaviour requires an advanced model of the connections [169]. The analysis of network reciprocity can also provide useful information about cooperation-promoting mechanisms [47]. The modelling forming connections is crucially important when we would like to understand how social networks evolve [170]. Recently it has been proven that combining the methods in game theory, agent-based modelling, machine learning, and computational sociology is a useful approach to understand the mechanisms of network formation [171].

In organisational networks, the nodes are co-workers and ties are defined based on the researchers are focusing on. Edges can be defined based on communication [172], advice [173], friendship [174] relationships, or all of these [175]. For getting a better understanding of what factors affect the formation of communication networks, connection types defined by the theory of structuration [176] were shown to be useful [177].

The importance of the multilayer nature of intra-organisational networks was realized more than thirty years ago [178, 179]. The informal structure of an organisation is complex and multilayered as people are involved in multiple, dynamic and overlapping webs of relationships [165]. Ref. [180] was the first work to argued that multiplex models should significantly improve the analysis of organisations. In the early studies Multi-Theoretical Multi-Level models [181, 182] and multilayer networks [183, 184, 185] were used to provide a deeper insight into organisations. The theory of multilayer networks [186, 73] is a rapidly growing field in network science. Nowadays multilayer networks are widely used in SNA [187, 188, 189, 50, 28]. Multiplex networks are a special case of multilayer networks where nodes are the same everywhere, and different edges lie on different layers [190]. In these models, the nodes can be characterized by their activities on different layers, which provides a better understanding of their roles [191, 38, 192].

As can be seen, organisational networks have been considered to be multilayer networks since the early 1990s, but no method could handle the multidimensional aspect of the problem. Finding informative correlations between layers of multilayer networks is still considered as one of the primary goals of network science [73].

The research aims to develop a methodology for the complex assessment of organisations that handles the multidimensional nature of the relationships.

Their are several network science and organisational science based contributions of this work.

- Based on our organisational development experience, requirements of business partners, and the literature of organisational network analysis a multilayer organisation network with 15 layers representing interaction-, rating-/perception-, and friendship-type connections were defined.
- The key idea behind the development of this model is that the connection types of the proposed multidimensional organisational network can be analysed by association rule mining algorithm to reveal how relationships, motivations, and perceptions as layers of an organisation influence each other in different scopes of activities.
- The novelty of this approach resides in the evaluation of people in organisations as frequent multidimensional patterns of incoming/outgoing multidimensional edges.
- It is demonstrated that the overlapping edges of the proposed multilayer network can be used to characterize clusters of co-workers and to evaluate skills of the leaders e.g. the motivation and managerial capabilities.
- The similarity analysis of the nodes enables the organisation to be clustered from different aspects e.g. to find similarly perceived key persons.

This chapter is organized as follows. In the first part of the Methods section, the multidimensional organisational network model is introduced. The second part of this section presents how frequent pattern mining can be used to extract information from multidimensional networks. It is believed that the proposed approach can be widely applied to find significant correlations between layers of multilayer networks. The Results and Conclusions section demonstrates how the proposed approach can be used in the development of three organisations.

4.2 Methods

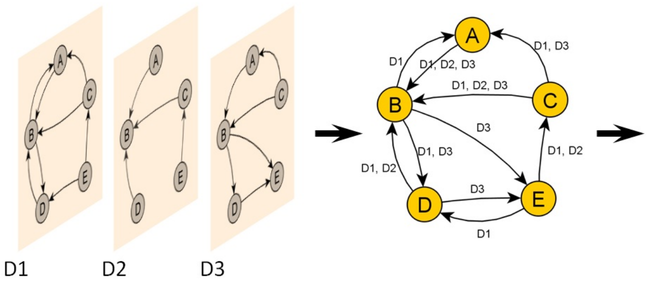
4.2.1 Multidimensional representation of organisational networks

In the proposed multidimensional organisational network the nodes represent the employees and labelled edges reflect how the members of the organisation communicate, work together, rate and motivate each other, and their personal relationships. Labelled and directed connections define multiple edges form a multidimensional network

$\mathcal{G} = (V, E, D)$, where V represents the node set, D the set of edge labels defines the dimensions of edges, and E denotes the edge set, $E = \{(u, v, d); u, v \in V, d \in D\}$ as can be seen in Fig. 4.1. In directed graph the edges (u, v, d) and (v, u, d) are distinct.

As each label can be mapped into an independent network, the model can be interpreted as a multilayer network. A multilayer network is a pair $\mathcal{M} = (\mathcal{G}, \mathcal{C})$, where $\mathcal{G} = \{G_\alpha; \alpha \in \{1, ..., M\}\}$ is a family of graphs $G_\alpha = (X_\alpha, E_\alpha)$ (called layers of $\mathcal{M}$) and $\mathcal{C} = \{E_{\alpha\beta} \subseteq X_\alpha \times X_\beta; \alpha, \beta \in \{1, ..., M\}, \alpha \neq \beta\}$ is the edge set between nodes of different layers G_α and G_β with $\alpha \neq \beta$ [73]. E_α are called intralayers and $E_{\alpha\beta}(\alpha \neq \beta)$ are referred to as interlayer-connections.

The studied intra-organisational networks can be considered to be directed multiplex networks which are a special type of multilayer networks. Multiplex networks consist of a fixed set of nodes connected by different types of links. In our case the $G = (V, E, D)$ multidimensional network is associated with a multiplex network with layers $\{G_1, ..., G_{|D|}\}$ where $\alpha \in D$, $G_\alpha = (X_\alpha, E_\alpha)$, $X_\alpha = V$, $E_\alpha = \{(u, v) \in V \times V; (u, v, d) \in E \, \text{and} \, d = \alpha\}$.



from	to	D1	D2	D3
A	B	1	1	1
B	A	1	0	0
B	D	1	0	1
B	E	0	0	1
C	A	1	0	1
C	B	1	1	1
D	B	1	1	0
D	E	0	0	1
E	C	1	1	0
E	D	1	0	0

Figure 4.1 Representations of a multidimensional network

Based on our organisational development experience, requirements of our business partners, and the literature of organisational network analysis connection-/interaction-, rating-/perception-, and friendship-type layers were defined in this model:

1. Connection-type layers
   L1: get advice from
   L2: get priorities from
   L3: get feedback from
   L4: communication with
   L5: working together with

2. Rating-type layers
   L6: he/she helps to find information
   L7: he/she provides the best working relationship
   L8: he/she has great professional knowledge
   L9: he/she motivates me

L10: he/she is capable of solving complex tasks
L11: he/she is capable of managing colleagues
L12: he/she is a key person in the organisation

3. Friendship-network layers
   L13: he/she gets along easily with me
   L14: I would like to have dinner with him/her
   L15: I would like to work together with him/her as a part of a problem-solving team

An online survey was designed to identify the connections. In the survey, there were as many questions as layers. Respondents were asked to mark the names of co-workers that fit the question and were not restricted to a fixed number of answers to minimize measurement error [193].

The combination of layers is believed to capture the essence of an organisation, making it possible to extract information about working connections, trust, employee's perceptions of each other, and leadership.

4.2.2 Frequent pattern mining of edge labels in multidimensional networks

Discovering statistically significant correlations between layers of multilayer networks is one of the major goals of network science over the next years [73]. A recently developed edge-overlap measure evaluates the conditional probability of finding a directed link on a layer given the presence of a directed link between the same nodes on another layer [38, 194] which can handle with pairs of dimension. The method is feasible for examining the overlap of a small number of dimensions. As the coexistence of links with different labels between any nodes i and j forms frequent patterns of any number of dimensions, it was found that frequent pattern mining provides a new opportunity to describe correlations between layers.

Frequent itemset mining was initially developed for market basket analysis, and it is used nowadays for almost any task that requires the discovery of regularities between (nominal) variables [195]. This concept has been extended to frequent graph-based substructure pattern mining [196].

This work differs from methods developed for frequent subgraph mining in unilayered (labeled) networks [197]. Labelling network motifs in protein-protein interaction (PPI) networks [198] and text networks [199] is also a similar problem. While in these tasks the labels are attached to the nodes, in intra-organisational network case the problem requires the identification and characterisation of the frequent multidimensional edges.

As this is the first attempt to introduce frequent itemset mining into the analysis of multidimensional networks, the technique is summarized in Table 4.1. The dimensions $D = \{d_1, d_2, ..., d_M\}$ of the network are considered to be a set of items $I = \{I_1, I_2, ..., I_M\}$ (in market basket analysis, I_i represents a given product). The set of transactions of the items $T = \{t_1, t_2, ..., t_m\}$ are defined as a set such that $t_i \subseteq I$ is identical to a given edge $E_i = \{(u_i, v_j, d); u, v \in V, d \in D\}$ in a multigraph between nodes u_i and v_j.

The aim is to identify frequently occurring subsets of edge dimensions and mine valuable information concerning multidimensional networks based on the analysis of these itemsets. The occurrence of an itemset C is measured as number of transactions (multidimensional edges) that the itemset contains. When this frequency is divided by the size of the transaction set $|\mathcal{D}|$ which is identical to the number of edges $|E|$, the calculated support of $s_T(C)$ represents the probability of multidimensional edge C. The $C \subseteq T$ is referred to as frequent when $s_T(C) \geq s_{min}$ exceeds a user-specified minimum s_{min}. The goal of frequent itemset mining is to find all frequent itemsets $C \subseteq I$ in database $\mathcal{D}$ [195].

The resultant frequent itemsets can be used to form $A \Rightarrow B$ association rules where A and B are disjoint subsets of C, as $A \subset C$, $B \subset C$ and $A \cap B = \varnothing$ [200].

A often called as antecedent and B as consequent. The rule $A \Rightarrow B$ holds in the transaction set $\mathcal{D}$/edge set E with support s, where s is the percentage of transactions/multidimensional edge in $\mathcal{D}/E$ that contain $A \cup B$, or say, both A and B. Other words the probability that a transaction/multidimensional edge contains the union of set A and set B or occurrence frequency of item set $(A \cup B)$ is $P(A \cup B)$

$$s = support(A \Rightarrow B) = P(A \cup B) \tag{4.1}$$

The rule $A \Rightarrow B$ has a confidence c in the transaction set $\mathcal{D}$/edge set $\mathcal{E}$, where c is the percentage of transactions in $\mathcal{D}/\mathcal{E}$ containing A that also contain B. The confidence of the rule represents the $P(B|A)$ conditional probability:

$$c_T(A \Rightarrow B) = P(B|A) = \frac{P(A \cup B)}{P(A)} = \frac{s_T(A \cup B)}{s_T(A)} = \frac{\text{count}(A \cup B)}{\text{count}(A)} \tag{4.2}$$

when A is independent of B, $P(A \cup B) = P(A)P(B)$. The lift l is a correlation measure that is based on the ratio of these probabilities:

$$l = lift(A \Rightarrow B) = \frac{P(A \cup B)}{P(A)P(B)} = \frac{s_T(A \cup B)}{s_T(A)s_T(B)} \tag{4.3}$$

when $l < 1$ A is negatively correlated with B, meaning that the occurrence of

A leads to the absence of B. When $l > 1$, then A and B are positively correlated, meaning that the occurrence of A implies the occurrence of B [201]. Rules with high level of lift usually exhibit relatively low degree of support [202]. An alternative to lift is leverage that states how much more often A and B occur together than as independent random variables [203].

$$\lambda = leverage(A \Rightarrow B) = s_T(A \cup B) - s_T(A)s_T(B) \tag{4.4}$$

<table>
<tr><th></th><th>Frequent itemset mining</th><th>Multidimensional network</th></tr>
<tr><td>Item base</td><td>$I = \{I_1, I_2, ..., I_M\}$
I_i, for example, represents a product</td><td>$D = \{d_1, d_2, ..., d_M\}$
d_i is a dimension</td></tr>
<tr><td>Transaction</td><td>$T = \{t_1, t_2, ..., t_m\}$
is a set of items
$T \subseteq I$</td><td>$E_k = \{(u, v, d); u, v \in V, d \in D\}$
is a multidimensional edge, which is a set of dimensions
$E_k \subseteq D$</td></tr>
<tr><td>Database</td><td>$\mathcal{D} = \{T_1, T_2, ..., T_{max}\}$
all transactions</td><td>$\mathcal{E} = \{E_1, E_2, ..., E_{max}\}$
all multidimensional edges</td></tr>
<tr><td>Frequent itemset</td><td>$C \subseteq T$ is referred to as the frequent itemset
$s_T(C) \geq s_{min}$</td><td>$C \subseteq E$ is referred to as the frequent dimension set
$s_T(C) \geq s_{min}$</td></tr>
<tr><td rowspan="2">Association rule</td><td colspan="2">$A \Rightarrow B$, where A and B are disjoint sets of items;
A: antecedent, B: consequent</td></tr>
<tr><td>$A \subset I$, $B \subset I$ are sets of items, with $A \cap B = \varnothing$</td><td>$A \subset D$, $B \subset D$ are sets of dimensions, with $A \cap B = \varnothing$</td></tr>
<tr><td rowspan="2">Support of a rule</td><td colspan="2">$s_T(A \Rightarrow B) = P(A \cup B)$</td></tr>
<tr><td>probability that a transaction contains $A \cup B$</td><td>probability that a multidimensional edge contains $A \cup B$</td></tr>
<tr><td rowspan="2">Confidence</td><td colspan="2">$c_T(A \Rightarrow B) = P(B|A) = \frac{P(A \cup B)}{P(A)} = \frac{s_T(A \cup B)}{s_T(A)}$</td></tr>
<tr><td>probability of finding B under the condition
that transactions also contain A</td><td>probability of finding B under the condition
that multidimensional edges also contain A</td></tr>
<tr><td>Lift</td><td colspan="2">$l = lift(A \Rightarrow B) = \frac{P(A \cup B)}{P(A)P(B)}$
B increases (lift) the likelihood of A
if $l < 1$ negative correlation; $l = 1$ independent;
$l > 1$ positive correlation exists between A and B</td></tr>
<tr><td>Leverage</td><td colspan="2">$\lambda = leverage(A \Rightarrow B) = s_T(A \cup B) - s_T(A)s_T(B)$
how much more often A and B occur together
than expected under independence</td></tr>
</table>

Table 4.1 Corresponding nomenclature of frequent itemset mining and multidimensional networks

The computational complexity of the proposed methodology is determined by the utilized frequent itemset mining algorithm. The complexity of the most widespread Apriori algorithm is $\mathcal{O}\left(M^2m\right)$ [204], where M represents the number of items and m the number of data records, thus finding the frequent connection types has quadratic dependence on the M connection types and linear scalability in the $m = |E_k|$ number of connection. As $M = 15$, it can be concluded that the calculation of the proposed measures can be computed very quickly even for large networks.

In this subsection, an analogy between the measures of network science and frequent pattern mining was presented. In the following subsection, how frequent itemsets and association rule mining can be used to understand the formation of connections is demonstrated.

4.2.3 Node characterisation based on incoming multidimensional edges

In (organisational) network research, three levels (dyadic, actors/nodes, networks) of the analysis can be distinguished [205]. At the dyadic level the frequent occurrence of the edge dimensions can be analysed. Analysis at the level of the actors requires information to be aggregated with regard to the types of edges to characterize the nodes. For example, to measure the degree of innovation and problem-solving abilities of the employees, the centrality of the actors in the communication network of the organisation can be studied [206]. The selection of suitable dimensions plays an important role in these ratings, e.g. as information exchange is reflected in the advice network, the perception of information access is mostly determined by the advice centrality [165]. In multilayer networks, nodes can be characterized based on their activities at different layers [73]. The distribution of degrees of nodes among layers can be described by its entropy of the multiplex degree which is similar to the multiplex participation coefficient published in Ref. [38].

In the following a novel method for the characterisation of nodes is introduced by calculating the frequent patterns of the incoming/outgoing multidimensional edges of ego-networks. The directed edge set $\mathcal{E}_{u^{out}} = \{E_1, E_2, ..., E_{max}\}$, $E_k = \{(u^{out}, v^{in}, d); u, v \in V, d \in D\}$ consist of outgoing edges of a node $u \in V$; and $\mathcal{E}_{u^{in}} = \{E_1, E_2, ..., E_{max}\}$, $E_l = \{(v^{out}, u^{in}, d); u, v \in V, d \in D\}$ represents the incoming edges of a node $u \in V$. Frequent dimensions of outgoing and incoming edges are specific to the nodes. The outgoing edges are related to the perceptions, ratings and connections to others, while the incoming patterns reflect how an actor is rated. Association rules $A \Rightarrow B$ valid for $\mathcal{E}_{u^{out}}$, $\mathcal{E}_{u^{in}}$ or $\mathcal{E}_u = \mathcal{E}_{u^{out}} \cup \mathcal{E}_{u^{in}}$ provide the specifications of the node.

As a node can support or weaken association rules with its incoming/outgoing multidimensional edges, the measures of the association rules can be utilized as fingerprints of the organisational network. The similarity between the nodes can be evaluated based on the incoming and outgoing patterns. Based on clustering of the nodes, similar key persons and leaders can also be identified which approach is similar to frequent pattern mining-based community detection [76].

As modularity is based on the difference between the actual and expected number of edges [207], the analysis of this difference can reflect attractiveness and talent in individual and organisational levels. Community detection algorithms explore densely linked groups of nodes, so these algorithms can highlight central nodes [208], leaders of communities [209] and hierarchical structure [210, 211].

The following section demonstrates that based on the similarities of the multidimensional incoming and outgoing connections the clusters of co-workers can be determined and use extracted knowledge can be used to characterize typical roles in the organisations.

4.3 Results and discussions

To demonstrate the applicability of the proposed methodology, leaders and key persons are identified based on the incoming edges and the determination of the effects with regard to the advice network based on frequent patterns containing the advice (L1) dimension are sought.

4.3.1 The studied organisational networks

Connections of 83 (response rate (RR): 75%), 57 (RR: 93%) and 203 (RR: 94%) employees from A: a not-for-profit arts organisation, B: a multinational manufacturing company, and C: a cultural institute, respectively were studied. The complexity of Company A is illustrated in Figure 4.2. The number of nodes and edges with their support is shown in Table 4.2. The high support of L13 in Company A indicates a friendly atmosphere. The reciprocity in the L13 layer is 43-44 % for all companies, which correlates well with other studies [165].

The number of two or more dimensional edges is shown in Table 4.3 which indicates that the majority of edges are multidimensional, only 26-33 % of the edges are one-dimensional. The dimensions L4, L8 and L13 (55 % of the one-dimensional in Company A), as well as the L14 and L15 tend to appear alone.

4.3.2 Analysis of the extracted association rules

Finding meaningful association rules is one of the biggest challenges in data mining. Filtering the rules based on confidence and support is an obvious approach, but in some cases, the grouping of the rules based on variables is necessary [212], e.g., the

	Company A	Company B	Company C
Number of nodes	83	57	203
Number of multidimensional edges	1709	925	5766
P(L1)	0.200	0.310	0.280
P(L2)	0.063	0.163	0.132
P(L3)	0.199	0.166	0.187
P(L4)	0.304	0.403	0.362
P(L5)	0.274	0.356	0.349
P(L6)	0.253	0.308	0.284
P(L7)	0.415	0.345	0.341
P(L8)	0.315	0.384	0.349
P(L9)	0.103	0.171	0.146
P(L10)	0.150	0.225	0.215
P(L11)	0.088	0.251	0.162
P(L12)	0.221	0.329	0.268
P(L13)	0.714	0.484	0.363
P(L14)	-	0.428	0.339
P(L15)	0.301	0.471	0.323

Table 4.2 Support values of the edge labels in the studied organisations

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Company A	0.33	0.19	0.12	0.08	0.06	0.06	0.04	0.04	0.02	0.01	0.02	0.02	0.01	0.01	-
Company B	0.26	0.15	0.11	0.09	0.06	0.06	0.04	0.04	0.04	0.03	0.03	0.02	0.03	0.02	0.03
Company C	0.30	0.18	0.12	0.10	0.06	0.06	0.04	0.04	0.03	0.03	0.02	0.02	0.01	0.01	0.02

Table 4.3 Proportion of the number of dimensions in multi-edges

setting of a high-threshold support would exclude rare dimensions from the rules (like L2 and L9).

A positive correlation is indicated between the antecedent and consequent sets of all rules with a lift greater than one. Only two rules exist in Company A that possess

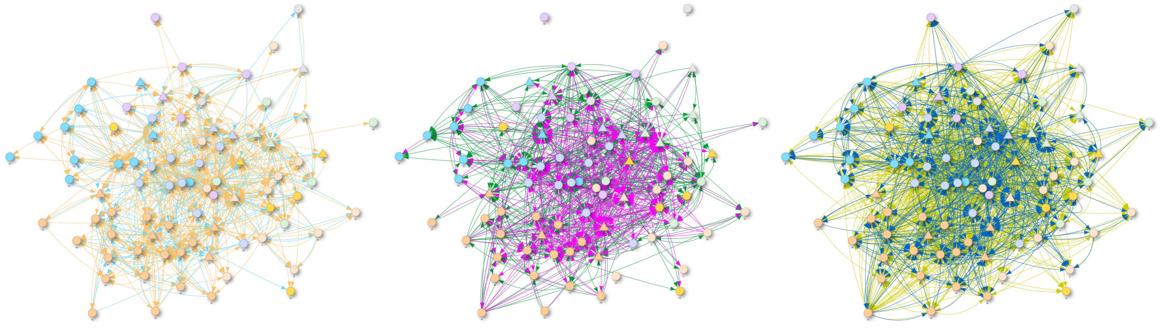
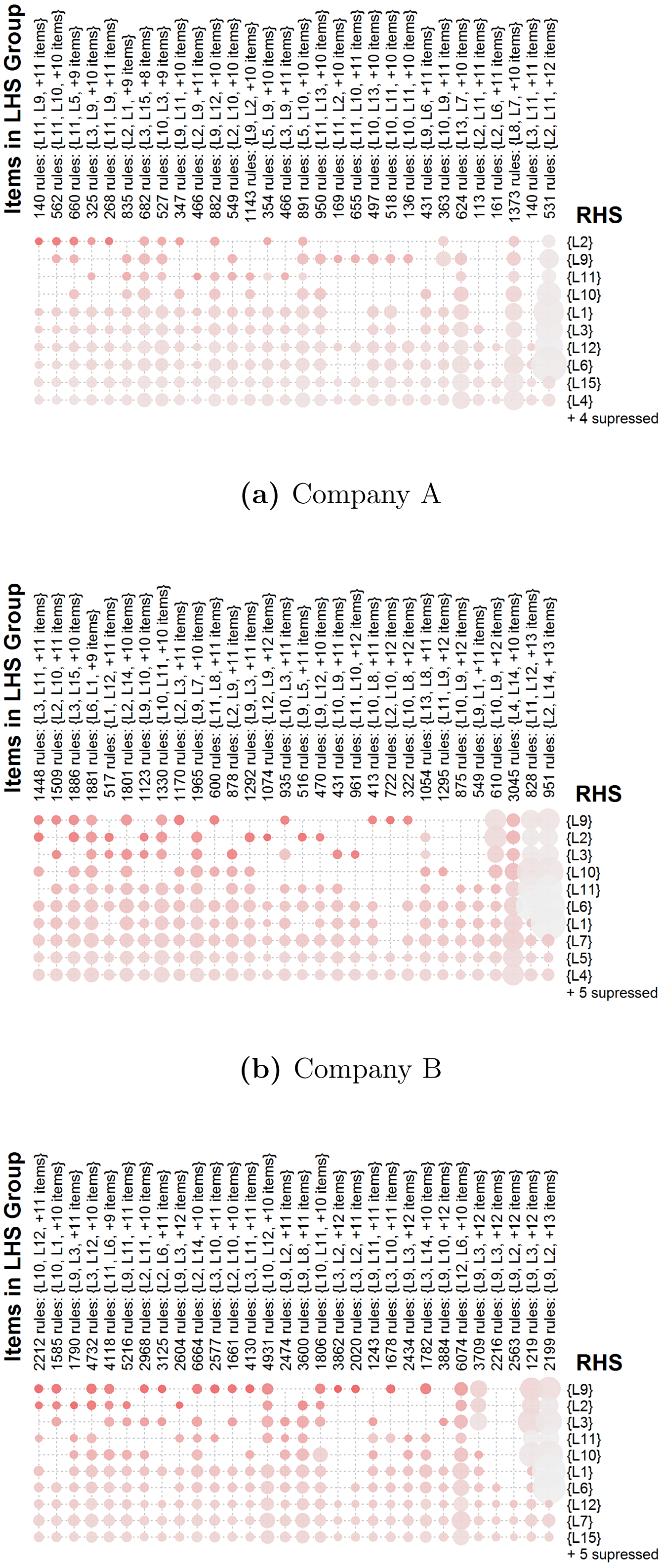


Figure 4.2 Six layers of the organisational network of Company A (left: light blue is L1, orange is L4; middle: dark green is L8, magenta is L12; right: dark yellow is L13, dark blue is L15. The nodes are coloured according to the departments they belong to. The shape of nodes corresponds to the positions as triangles represent leaders and circles stand for the employees.)

negative leverages. The L8 ⇒ L13 rule can be found on 370 edges that is less than 380 edges expected under independent conditions, which indicates on average that it is hard to get along well with people who possess a high degree of professional knowledge.

The extracted rules are summarized as grouped matrices [202] in Figure 4.3, where the antecedents that are statistically dependent on the same consequents are grouped and shown in columns with their two most frequent dimensions written on the axes. Consequents are arranged in the rows. The bubbles are coloured according to the median lift of the rules in the groups, while the sizes of the bubbles represent the medians of the supports. The resultant plots highlight that important consequents are very similar in all companies, namely L2, L9, L11, L3, L10 and L1 which refer to leadership, motivation, managerial capability, giving feedback, solving complex tasks and giving advice respectively.

The confidence values of the rules can serve as layer-overlap measures. In Figure 4.4 the columns are antecedents and the rows are consequents of rules $Column \Rightarrow Row$, furthermore, the values of the matrix show the confidences of the rules. As expected, layers L2 and L9 exhibit a strong correlation between almost all other layers, while it seems that the precedences of the edges L9, L10 and L11 increase the probabilities of connection types L7, L8 and L12.

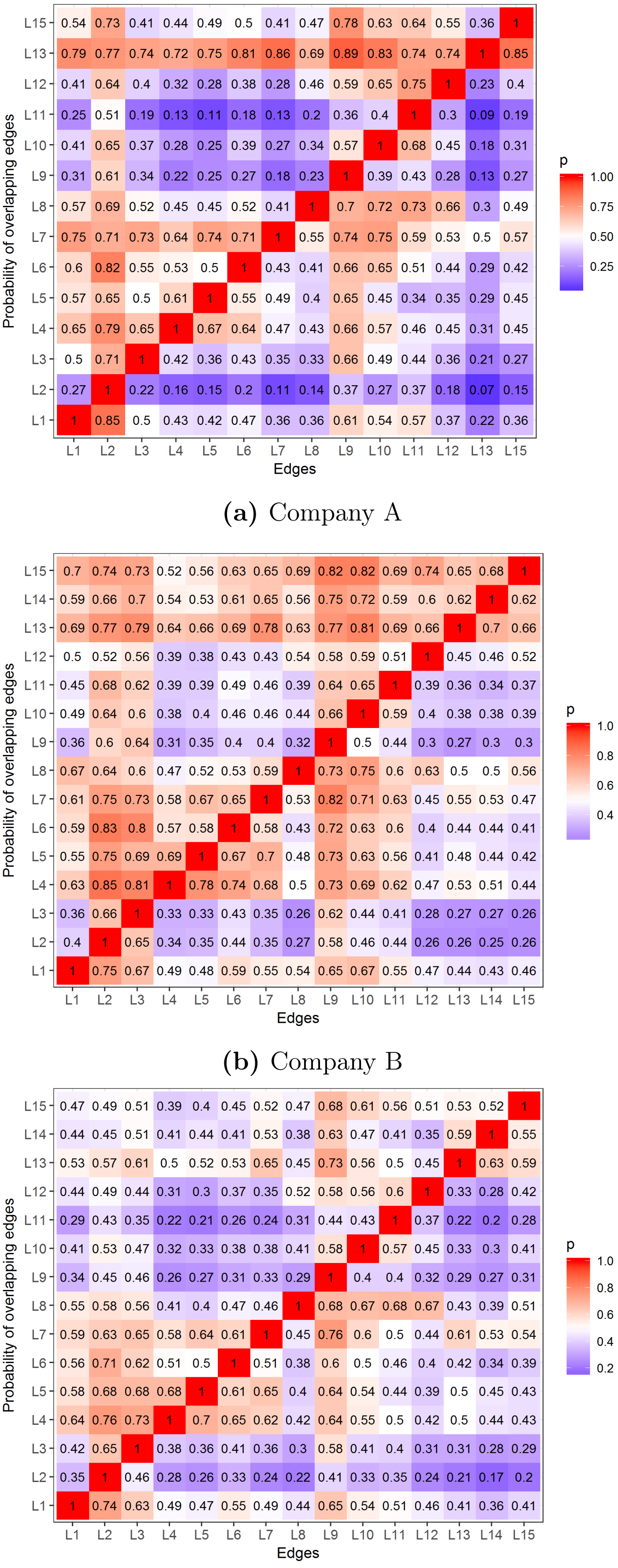


(a) Company A

(b) Company B

(c) Company C

Figure 4.3 Summary of rules (size is proportional to support, colour is proportional to lift)



(a) Company A

(b) Company B

(c) Company C

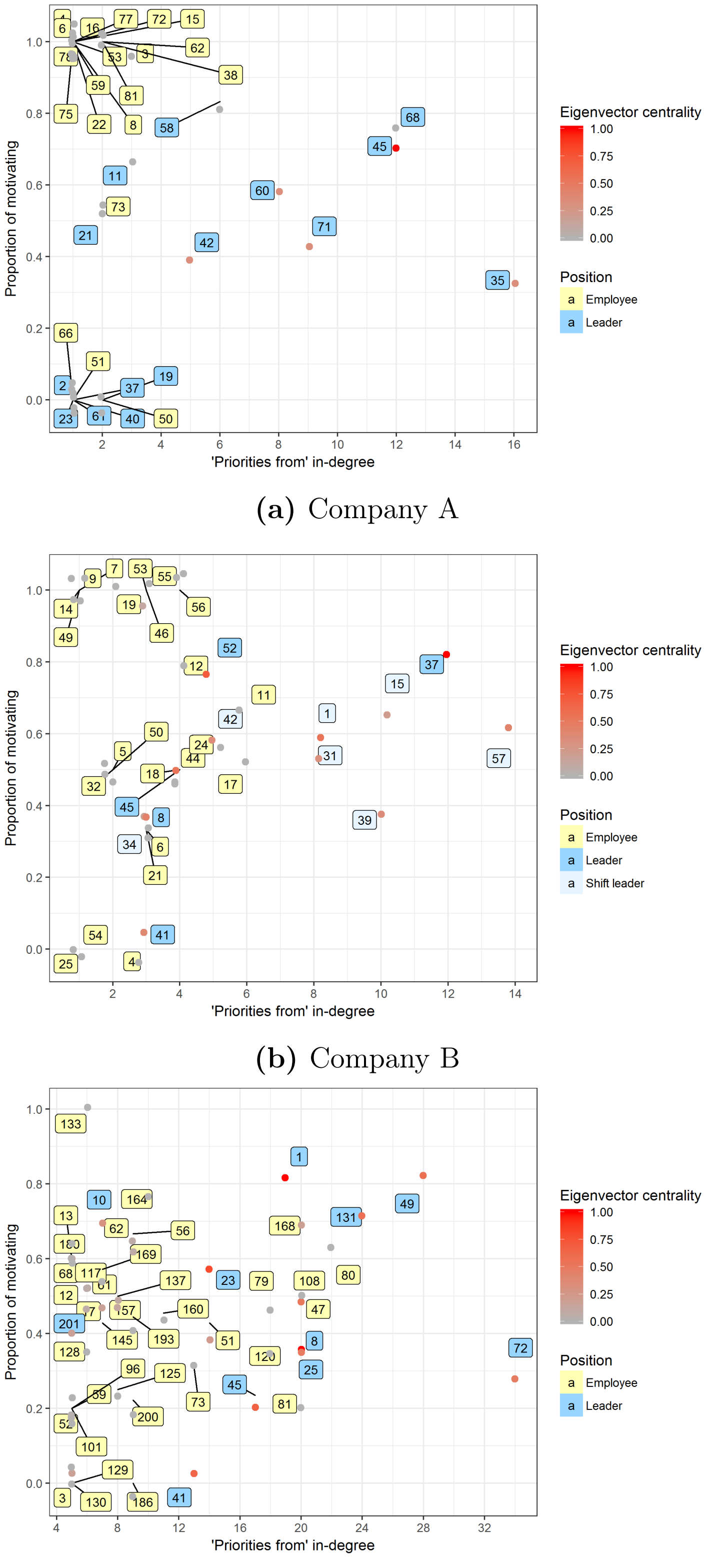
Figure 4.4 The probability of dimensions in rows given the dimensions in the columns in the case of all three companies

4.3.3 Characterisation of the leaders

The appearance of dimension L2 in a multidimensional edge shows who is considered to be a leader in an organisation because he/she provides instruction in a workflow. The confidences of the leader-related rules are shown in the second columns of the matrices in Figure 4.4. The $c(L2 \Rightarrow L9) = P(L9|L2)$ confidence of the $L2 \Rightarrow L9$ rule is a good measure of how a co-worker perceives the motivation of his/her leader.

The two-dimensional evaluation of actors is represented in Figure 4.5. The x-axis is the in-degree on the "priorities from" (L2) layer, and the y-axis shows the conditional probability of the presence of "motivates me" (L9) dimension along the same L2 edge. The in-degree centrality does not correlate with the motivating capability (Pearson's ρ between the in-degrees and the motivating capability of the nodes is 0.38 at Company A; -0.09 at Company B; and 0.21 at Company C), so the two dimensions provide additional information about actors. However, high and low social capital correlate with the in-degree centrality which reflects the eigenvector centrality captures the importance of the actors [213]. Eigenvector centralities of actors on the L2 layer are also well correlated with in-degrees (Pearson's ρ between the in-degree and the eigenvector centrality of the nodes is 0.71 at Company A; 0.68 at Company B; and 0.67 at Company C). The differences in the eigenvector centrality among actors with the same in-degree can be studied in Figure 4.5. The leader numbered as '45' in Company A has much higher eigenvector centrality than leader numbered as '68', but they have the same motivating capability that indicates that leader '45' motivates more important people than leader '68' which increases his/her overall importance.

The fact that there is no correlation between the numbers of motivation type connections and eigenvector centrality (Company A: 0.11; Company B: -0.06; Company C:0.17) shows that the capability of motivating may a personal trait. The plots can be utilized to evaluate the performance of the leaders and support decisions related to organisational development.



(a) Company A

(b) Company B

(c) Company C

Figure 4.5 Motivating leaders. For the sake of interpretability of the figures, a small amount of random variation is added to the location of each point to avoid overlapping and persons with more than four in-degree are plotted at Company C.

4.3.4 Clustering-based identification of the key persons

Finding influential employees in organisations should differ from the analysis of formal organisational charts. Research questions like "who is considered to be a key person?" require detailed analysis. A clustering-based algorithm to answer such questions was developed. Similarly evaluated people can be clustered based on how similarly their incoming edges support the association rules. The Partitioning Around Medoids (PAM) algorithm [214] was applied to identify the clusters (see Figure 4.6) as it lends itself to clustering based on the specified distance matrix [215], it has the robustness to noise [216] and performs better for large datasets than the also popular k-means algorithm [217].

Clusters 1 (blue) and 2 (yellow) include the top managers of Company A. People in Clusters 3 (grey) and 4 (red) are evaluated as key persons to the same extent. Members of Cluster 3 (grey) advise more co-workers, while members of Cluster 4 (red) are evaluated as possessing greater professional knowledge. 38 % of Cluster 3 (grey) are middle managers. There is no leader in Cluster 4 (red), which suggests that members select advisor based on their status [173].

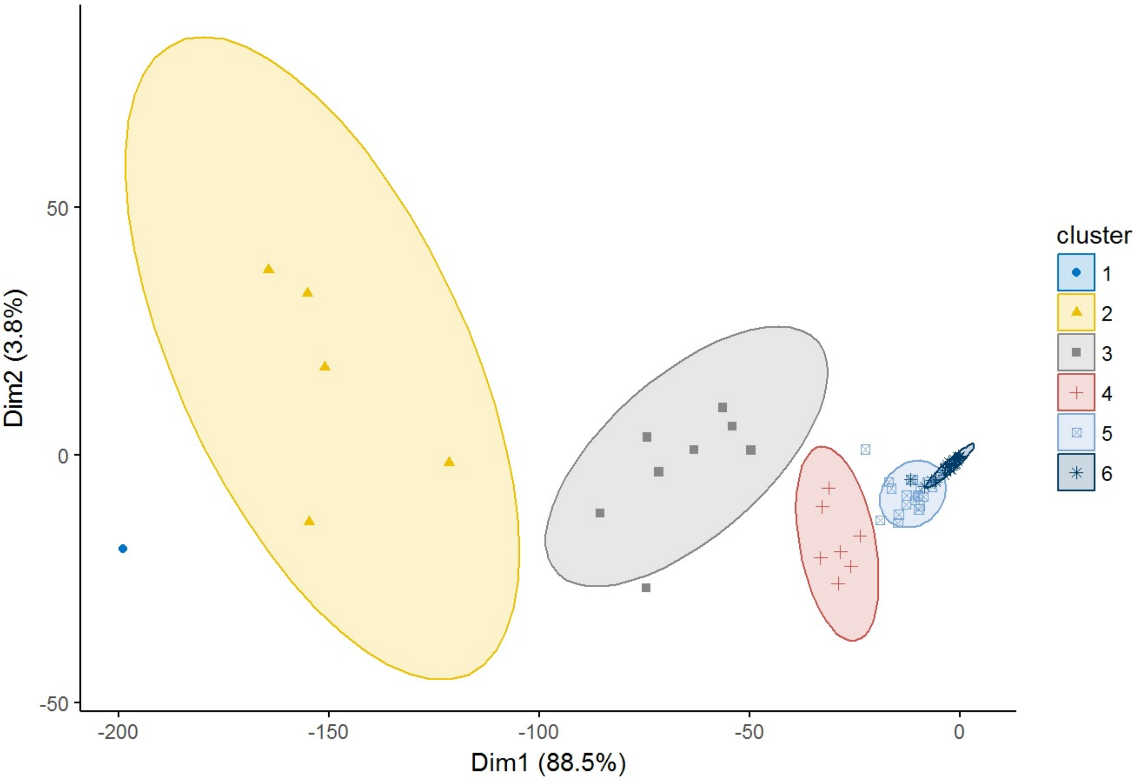


Figure 4.6 Clusters of key persons in Company A visualized by principal component analysis. The numbers at the axes labels show the percentage of the variance represented by the principal component.

4.3.5 Effects of the advice network

According to the literature, at least two kinds of processes drive how sources of advice are selected: namely status recognition and homophily [173]. The extraction of valuable information concerning these effects based on the analysis of the Lift($L1 \Rightarrow B$) values was attempted.

Table 4.4 shows that the edge types of leadership (L2 and L3), motivating behaviour (L9), information resources (L10) and cognitive ability (L6) increase the likelihood of advice (L1). Although there are some specificities in term of the networks of different companies, e.g. lift($L1 \Rightarrow L2$) is much greater in the case of Company A, but its trends are very similar. The high confidence values in the L1 columns of Figure 4.4 indicate that this connection type has positive effects on contact type, trusted relationships and the judgement of professional knowledge.

	Company A	Company B	Company C
L2	4.25	2.43	2.63
L3	2.51	2.16	2.25
L4	2.14	1.56	1.76
L5	2.08	1.55	1.67
L6	2.36	1.91	1.96
L7	1.82	1.78	1.74
L8	1.80	1.73	1.57
L9	3.04	2.10	2.34
L10	2.72	2.17	1.93
L11	2.83	1.78	1.81
L12	1.84	1.52	1.62
L13	1.11	1.42	1.47
L14	-	1.38	1.29
L15	1.80	1.48	1.46

Table 4.4 Lift($L1 \Rightarrow B$) values at the studied companies

Dimensions that predict the occurrence of L1 are described by the $A \Rightarrow L1$ family of rules. The confidence($A \Rightarrow L1$) of these rules represents the probability of the occurrence of L1 given the existence of the A dimensions. As is shown in Table 4.5, professional knowledge (L8) leads to a far more significant increase in the probability of get advice dimension (L1) than communication (L4) in the case of the existence of a leader (L2), working relationships (L5) and best working relationship (L7), as well as information sources (L6). However, L4 significantly increases the probability of the advice-type connection when motivation (L9), the capability of solving complex tasks (L10), the ability to manage colleagues (L11), and key person (L12) exist. In other words, the confidences of {L2 or L5 or L6 or L7} $\cup$ L8 $\Rightarrow$ L1 are greater than the confidences of {L9 or L10 or L11 or L12} $\cup$ L8 $\Rightarrow$ L1, and the confidences of {L2 or

L5 or L6 or L7} ∪ L4 ⇒ L1 are less than the confidences of {L9 or L10 or L11 or L12} ∪ L4 ⇒ L1.

	Company A				Company B				Company C			
Antecedents (A)		∪ L4	∪ L8	∪ L8 ∪ L9		∪ L4	∪ L8	∪ L8 ∪ L9		∪ L4	∪ L8	∪ L8 ∪ L9
L2	0.850	0.882	0.919	0.943	0.755	0.797	0.907	0.915	0.735	0.816	0.873	0.950
L3	0.502	0.683	0.655	0.761	0.668	0.733	0.847	0.901	0.630	0.736	0.813	0.883
L4	0.428		0.637	0.806	0.485		0.761	0.839	0.492		0.747	0.868
L5	0.415	0.503	0.582	0.752	0.480	0.558	0.703	0.784	0.466	0.567	0.745	0.869
L6	0.472	0.594	0.627	0.806	0.592	0.674	0.807	0.840	0.550	0.660	0.769	0.864
L7	0.363	0.549	0.554	0.780	0.551	0.649	0.734	0.818	0.488	0.617	0.713	0.850
L8	0.360	0.637			0.538	0.761			0.440	0.747		
L9	0.607	0.752	0.693		0.651	0.730	0.756		0.654	0.812	0.772	
L10	0.544	0.746	0.616	0.764	0.673	0.797	0.722	0.827	0.540	0.763	0.649	0.832
L11	0.566	0.913	0.651	0.854	0.551	0.717	0.685	0.804	0.505	0.791	0.599	0.822
L12	0.368	0.621	0.473	0.722	0.470	0.715	0.637	0.766	0.455	0.744	0.567	0.819
L13	0.221	0.486	0.462	0.734	0.441	0.635	0.651	0.795	0.411	0.626	0.635	0.812
L14					0.426	0.625	0.653	0.806	0.360	0.628	0.622	0.834
L15	0.361	0.571	0.549	0.754	0.458	0.716	0.648	0.805	0.409	0.672	0.605	0.811

Table 4.5 Confidences of the rule $A \Rightarrow L1$ of the companies

Most leaders (L2) give advice, especially in Company A. The probabilities are increased as the dimensions of the rules increase. Almost the same trends are shown in Tables 4.4 and 4.5 where the types of connections that are related to the advice network are presented. This result correlates well with the findings of Ref. [173] which show that advice is more likely to be sought from colleagues of higher statuses.

4.4 Conclusions

Organisational networks have been considered to be multilayer networks since the early 1990s, but so far no feasible method of handling their multidimensionality has been found. It has been demonstrated that frequent pattern mining can be applied to reveal statistically significant correlations between the layers and that the method is applicable regarding edge, actor and organisational level analyses. Frequently occurring outgoing edges have been shown to be related to perceptions and ratings, while incoming patterns reflect how the actor is rated. It was also highlighted that measures of the association rules could be used to define the fingerprints of organisational networks. The applicability of the methodology was demonstrated by the characterisation of leaders and key persons in three organisations. In the future, the utilisation of an extracted rule-base for the design of personal development programs, and the determination of a property-preserving multidimensional edge reordering algorithm to support goal-oriented organisational development is desired. The method can be applied to other multilayer networks where layers can represent dimensions and appropriate to make rankings.

4.5 Contributions to Industry 4.0 issues

The success of Industry 4.0 developments and the achievement of real efficiency depend on how employees in production systems adapt to new technologies. The technological change reduces man-to-man and increases the number of man-to-machine connections. The new situation requires employees to obtain new skills. The change also entails a change in leadership, mentoring network, learning structures, knowledge management.

My dissertation contributes to an organisational analysis methodology to identify key people and define areas for organisational development. It ultimately contributes the more efficient dissemination of new competencies and faster adaptation of Industry 4.0 developments.

Conclusion

The Fourth Industrial Revolution (Industry 4.0) has become an integral part of our culture. The ongoing technological change is affecting, among others, the competence needs of those working in production systems. As the new technological efficiency solutions spread in the everyday life of employees, the intensity of human-human relations decreases during human-machine relations increase. This phenomenon requires new competencies for staff members. Technological change has an impact on the education system, employee leadership, supporting mentor system, and investor capital movements which are the subject of observation in this thesis.

Technological change affects the education system because some of the competencies learned by the employees in schools. The application of new technologies can be successful if employees learn and adapt the new knowledge. However, not everyone learns with equal efficiency, good leaders, and professional mentors needed in productions systems to support colleagues.

Technological developments require investment capital, as robotics and efficiency-enhancing solutions are money-intensive. An investor wishes to minimize the risk, which affects how far capital moves geographically and relates to attractive regions.

Complex systems can be understood through connections, and network models are better to study them. In this work, the effects of digital transformation investigated with goal-oriented network science methods in order to better understand the selected human factors in production systems. The development of network science methods resulted new types of empirical results emerged on the three human factors of production systems, and novelties for network science.

In this thesis, the skill requirement of jobs is evaluated through education and occupation matching, what kind of graduation have people in a particular job on average. It was found that methodologically the relationships between higher education and the labour market can be examined with a bipartite network, and the horizontal matching of jobs and higher education degrees can be identified by exploring modules. One of the most interesting field for Industry 4.0 is the engineers who are typically graduated in engineering programs.

Results were showing the flow of the investor capital based on the estimation of the strengths in the spatial network of ownerships. The best evaluation using the gravity model showed that an investor living in developed regions prefers to invest in production systems in his/her settlement or other developed regions. The method is capable of identifying distance independent factors of attractivity.

In order to identify key persons of Industry 4.0, the relationships between employees were modelled with a multidimensional network to describe the complex system of an organization. The frequently overlapping dimensions at dyads provide information on key players or, in contrary, personal development demand. It was realized that employees are more likely to seek professional advice from their managers than those with high professional knowledge.

In conclusion, a better understanding of the human factors is greatly supported by the network science developments presented in this dissertation.

Contributions

This dissertation contributes to the area of analysis human factors related to Fourth Industrial Revolution. Specifically, it introduces novel thinking and techniques to the fields of analyse the matching of higher education programs and occupations, spatial movement of investment capital, identify key persons in systems engineering.

The interactions matter more than the performance of the units in complex systems, therefore, thinking in networks offers good opportunities to explore background processes. This dissertation offers innovative network science based analytical and methodological approaches in

1. education programs and occupations matching by uncovering modules from education-occupation bipartite network for better understand the relationship of education and labour market and skill demand induced by Industry 4.0,

2. determining attractive regions by spatial network of investment capital with comparison real weights of edges and different null models for characterizing geographical movement of investments required by developments to Industry 4.0,

3. identify key persons and personal development areas of employees by frequently together occurring dimensions in edges of co-worker social network for effective technological change and knowledge management in systems engineering.

Theses

Thesis 1 I demonstrated that transitions of graduates to the labour market can be modelled with a bipartite graph where the nodes are educations and occupations, which can be analysed with network science methods to find education-job matches.[218]

Background *Hungarian authorities provided anonymized linked database of graduates, which allowed the microdata level analysis of student career paths. One of the challenges is to find the right person for the right job, whose qualifications are related to the task he or she is performing.*

Results *From the information content of this rich database I extracted and combined the educations and occupations to form a bipartite graph where weights of edges are representing the number of graduates in a given program connected to a specific profession. The bipartite network model is suitable for determining, among others, the horizontal and vertical matching of educations and occupations with finding modules.*

Application *The method has been used to uncover hidden patterns in matching of professions and related educations. It is shown that network science methods are an efficient way to fine-tune the performance of the education-occupation matching methods.*

Thesis 2 I worked out a method to identify the attractiveness of economic regions based on the analysis of the distance-dependent owner-company networks and similarity of regions resulted by different null models.[219]

Background *Interorganization relationships between owners and leaders of businesses determine the operation of production systems, which, moreover, geographically distributed.*

Results *Personal decisions on investments define directed ties between geographical locations from the entrepreneur to his or her company. The probability of tie formation is distance dependent, but the proposed methodology is appropriate to characterize regional attractiveness based on a set of null models that avoids the noisy effect of geographical distance. Based on the calculation of the internal and external network densities, several measures were proposed to evaluate the attractiveness and development of towns and geographical regions.*

Application *It is shown, that the increasing geographical distance decrease the* *willingness to make investment decisions. It is shown that the personal investment network can be applicable in identifying attractive regions. Regions that attractive for capital are also attractive for employment; thus, investment network is also reflecting factors of jobs.*

Thesis 3 I designed multilayer network in which the members of production systems are connected with multidimensional edges. I demonstrated that frequent pattern mining can be applied to reveal statistically significant correlations (overlaps) between the layers.[220]

Background *In production units, staff, leaders and production tools are in a complicated relationship with each other, which correct functioning affects productivity, satisfaction, etc.*

Results *A survey was designed to evaluate the complex relationships of leaders and followers. The communication, knowledge flows, advice, trust, friendship (15 topics) dimension of relationships surveyed in three companies. I demonstrated that at the dyadic level, the frequently together occurred dimensions could be analysed with association rule mining. The method provides a robust evaluation of the correlation between proximities, ratings and friendship dimensions. It is proved that frequent pattern mining is an effective method to find overlaps between layers in a multilayer network.*

Application *The organization, leaders, can be characterized by the assessment of coexisted dimensions on edges. Frequently occurring dimension on outgoing edges are related to perceptions and ratings, while incoming patterns reflect how the actor is rated (e.g. motivating leader). Similarly evaluated people can be clustered based on how similarly their incoming edges support the association rules.*

Tézisek

1. tézis. Bebizonyítottam, hogy a diplomások munkaerőpiacra történő átmenetét kétoldalú gráffal lehet modellezni, amelyben a csúcsok a végzett-ségek illetve a munkakörök, és hálózattudományi módszerekkel elemezni a végzettségek és munkakörök horizontális illeszkedését modulok feltárásával. [218]

Háttér *Több magyar hatóság anonimizált, összekapcsolt adatbázisa alapján mikroadat szinten rendelkezésre áll a végzettek korai karrierjének adatállománya. A munkaerőpiacon kihívást jelent a megfelelő ember megfelelő munkakörben történő alkalmazása, a munkakörhöz leginkább kapcsolódó végzettséggel rendelkezők megtalálása.*

Eredmények *Az információban gazdag adatállományból a végzettségeket és a munkaköröket kapcsoltam össze, és rendeztem az adatokat kétoldalú hálózatba. Az élek súlyai az adott diplomával rendelkezők adott munkakörben dolgozók számát mutatja. A két-oldalú hálózatban a modulok feltárása alkalmas a végzettségek és munkakörök horizontális és vertikális illeszkedésének vizsgálatára.*

Alkalmazás *A módszer a végzettek és a munkakörök horizontális illeszkedésének rejtett mintázainak feltárására került alkalmazásra. Megmutattam, hogy a hálózattudományi módszerek alkalmazása hatékony módja a végzettség-munkakör illeszkedésé-nek megállapításához.*

2. tézis. Kidolgoztam egy módszert, amely különböző null modellek alkalmazásával a távolságfüggő tulajdonosi hálózatból képes meghatározni vonzóképes és hasonló régiókat.[219]

Háttér *A tulajdonosok és vezetők szervezetközi hálózata, amely földrajzilag tagolt, befolyásolja a termelő rendszerek működését. Elemezhető adatbázisként a magyarországi cégtulajdonosi hálózat áll rendelkezésre, amelyet települések hálózatára konvertáltam a földrajzi viszonyok vizsgálata érdekében.*

Eredmények *A befektetési döntések egy földrajzilag meghatározott hálózatban is leképezhetők, amelyben a befektető lakóhelye és a befektetésének székhelye van összekötve irányított módon. Az él kialakulás valószínűsége csökken a távolság növekedésével. Különböző null-modellekkel a távolság zavaró hatása kiküszöbölhető, és régiók távolságfüggetlen vonzóképességi tényezői meghatározhatók. Különböző külső és belső kap-csolati sűrűséget leíró hálózati mutatók alapján becsülhető a régiók, városok vonzóképes-sége.*

Alkalmazás *A távolság növekedésével csökken a befektetési hajlandóság. Be lett bizonyítva, hogy a befektetési döntések hálózata alkalmas vonzóképes régiók feltárására.*

3. tézis. Termelő rendszerek munkahelyi szociális kapcsolatait multilayer há-lózatba rendeztem, amelyben a munkatársak többdimenziós kapcsolatokkal vannak összekötve. Bebizonyítottam, hogy a gyakori elemhalmazok keresése alkalmazható a szignifikánsan együttesen megjelenő dimenziók bányászatára, és jelenségek feltárására.[220]

Háttér *A termelő rendszerekben a munkatársak, vezetők és a termelő eszközök bonyolult kapcsolatban vannak egymással, és a helyes működés befolyásolja a termelékenységet, elégetettséget, hatékonyságot, ezért a kapcsolatok jellemzésére módszerek szükségesek.*

Eredmények *A komplex kapcsolati viszonyok feltárására kérdőíves módszert fejlesztettem. Három vállalat kommunikációs, tudásáramlási, szakmai tanácsadási, bizalmi, baráti (15 tématerület) dimenzióit mértem fel. Bemutattam, hogy az élekben a gyakran együtt megjelenő dimenziók elemezhetők az asszociációs szabályok keresésével. A módszer robosztus eredményeket szolgáltat a kapcsolatok, értékelések, barátságok korrelációjáról. Be lett bizonyítva, hogy a gyakori elemhalmazok keresése egy hatékony módja a multilayer hálózatokban a rétegek átlapolásának vizsgálatához.*

Alkalmazás *A termelő rendszerek vezetőinek gyengeségei és erősségei meghatározhatóak az együttes dimenziók elemzésével. A gyakran együtt megjelenő dimenziók a kimenő élekben a percepciók és az érzékelésekkel függenek össze, míg a bejövő élekben a csúcspontok értékelésével (pl. motiváló vezető). A hasonlóképpen érzékelt szereplőket klaszterekbe lehet rendezni a bejövő élekben megjelenő együttes dimenziók alapján, amelynek elemzését az asszociációs szabályok feltárása nagymértékben támogat.*

Publications

Scientific journal articles

In an international journal

1. *GADAR L.* AND ABONYI J. (2018). Graph configuration model based evaluation of the education-occupation match. PLOS ONE 13, 3., DOI: 10.1371/journal.pone.0192427 (Q1) Related database and program codes: `https://data.mendeley.com/datasets/wkb7s93y42/1`

2. *GADAR, L.*, KOSZTYAN, Zs.T., AND ABONYI, J. (2018) The Settlement Structure Is Reflected in Personal Investments: Distance-Dependent Network Modularity-Based Measurement of Regional Attractiveness. COMPLEXITY, DOI: 10.1155/2018/1306704 (Q1)

3. *GADAR L.* AND ABONYI J. (2019). Frequent pattern mining in multidimensional organizational networks. SCIENTIFIC REPORTS 9 (1), DOI: 10.1038/s41598-019-39705-1 (D1)

4. *GADAR L.*, KOSZTYAN ZS.T., TELCS A., ABONYI J. (2020): A multilayer and spatial description of the Erasmus mobility network, SCIENTIFIC DATA 7 (41), DOI: 10.1038/s41597-020-0382-1 (D1)
   Related database and program codes: `https://data.mendeley.com/datasets/vnxdvh6998/3`

In a foreign language specialized journal in Hungary

1. FÖLDÉNYI, R. AND *GADÁR, L.* (2005) Relationships between the structure, transport, and toxicity of chloroacetanilide type herbicides. CENTRAL EUROPEAN JOURNAL OF OCCUPATIONAL AND ENVIRONMENTAL MEDICINE 11, 4, 293–300.

In a Hungarian edition of a specialized journal in Hungarian language

1. ÁLLÓ, A., *GADÁR, L.*, AND FÖLDÉNYI, R. 2008. Analitikai módszer fejlesztése patak üledék jellemző szerves szennyezőinek mennyiségi meghatározására. MAGYAR KÉMIAI FOLYÓIRAT - KÉMIAI KÖZLEMÉNYEK (1997-) 2, 63–67.

Book chapter

1. DOMOKOS, E., FEJES, L.U.A., FÜLÖP, T., ET AL. 2011. Földünk állapota. Institute of Environmental Engineering, University of Pannonia,, Veszprém. Hungary

Conference publication in journal or proceeding

Foreign language

1. ZSOLT, T.K., *LÁSZLÓ, G.*, AND ANDRÁS, T. 2018. Student and Teaching Mobility & Knowledge Transfer: a social network analysis study. In: Articles of Third University Mission International Conference. 1–5.
2. *GADÁR, L.*, KOSZTYÁN, Z.T., AND ABONYI, J. 2018. Measurement of regional attractiveness based on company-ownership networks. A Magyar Regionális Tudományi Társaság (MRTT) XVI. vándorgyűlése, Nemzetközi konferencia, Helyszín: Kecskemét, Neumann János Egyetem Gazdaságtudományi Kar
3. *LASZLO GADAR* AND JANOS ABONYI. 2018. Application of multilayer networks in organizational mining. NETSCI INTERNATIONAL SCHOOL AND CONFERENCE ON NETWORK SCIENCE, 11-15 June 2018, Paris
4. *LÁSZLÓ, G.*, ANDRÁS, T., VIVIEN, V.C., MARCELL, T.K., AND ZSOLT, T.K. 2018. Student mobility analysis (ERASMUS). Poster to IREG-9 Conference on Ranking and Accreditation - two roads to the same goal?, Belgium / Hasselt, 23-25 May,
5. *GADÁR, L.*, TISZA, Á., AND FÖLDÉNYI, R. 2006. The Role of Humic Substances and Different Metal Ions in the Retention of Acetochlor in Soil. In: Humic Substances – Linking Structure to Functions. Proceedings. 865–868., Karlsruhe, Germany, 2006. 07. 30 – 2006. 08. 04.
6. MÓD, R., FÖLDÉNYI, R., AND *GADÁR, L.* 2004. Investigation on adsorption of chloroacetanilide herbicides on Hungarian soils. In: Sixth International Sym-

posium and Exhibition on Environmental Contamination in Central and Eastern Europe and the Commonwealth of Independent States.

Hungarian language

1. TELCS, A., BANÁSZ, Z., CSÁNYI, V., *GADÁR, L.*, AND KOSZTYÁN, Z.T. 2018. Rangsorok, ligák, mobilitás, kollaboráció vizsgálata. Poszter az MTA-PE Budapest Rangsor Kutatócsoport kutatási eredményeiről, a Pannon Egyetem Multidiszciplináris Kiválósági Központ bemutatkozó konferenciáján (szimpóziumán), Veszprém, április 11.

2. ÉRSEK, C., *GADÁR, L.*, AND FÖLDÉNYI, R. 2007. Szulfonil-karbamid típusú herbicidek mint talajszennyezők. In: XXX. Kémiai Előadói Napok. 120–125.

3. *GADÁR, L.*, ÁLLÓ, A., ÉRSEK, C., AND FÖLDÉNYI, R. 2007. Ipari szennyvíztisztító fémszennyezésének hatása szerves szennyezők sorsára felszíni folyóvíz üledékében. In: Országos környezetvédelmi konferencia kiadványa. 53–60.

4. *GADÁR, L.*, IVÁDY, L., AND FÖLDÉNYI, R. 2006. Ipari szennyvíztisztító fémszennyezésének hatása felszíni folyóvíz üledékére. In: Országos Környezetvédelmi Konferencia. 277–284.

5. *GADÁR, L.*, TISZA, Á., AND FÖLDÉNYI, R. 2006. Acetoklór szorpciója talajon fémionok jelenlétében. In: The 13th Symposium on Analytical and Environmental Problems. 72–75.

6. ÉRSEK, C., *GADÁR, L.*, AND FÖLDÉNYI, R. 2005. A tribenuron-metil adszorpciója talajokon. In: Proceedings of the 12th Symposium on Analytical and Environmental Problems. 134–138.

7. *GADÁR, L.* AND FÖLDÉNYI, R. 2005. Ipari eredetű szennyvíz hatása felszíni folyóvíz üledékére. In: Proceedings of the 12th Symposium on Analytical and Environmental Problems. 129–133., Szeged, 2005. 09. 26.,

8. SERFŐZŐ, N., *GADÁR, L.*, AND FÖLDÉNYI, R. 2005. A talaj szemcseméretének hatása két herbicid adszorpciójára. In: Proceedings of the 12th Symposium on Analytical and Environmental Problems. 139–143.

9. *GADÁR, L.*, FÖLDÉNYI, R., AND MÓD, R. 2003. Klór-acetanilid típusú gyomirtószerek mozgékonyságának vizsgálata talajokon. In: Proceedings of the 10th Symposium on Analytical and Environmental Problems. 175–179., Szeged, 2003. 09. 29.

Bibliography

[1] World Economic Forum. The future of jobs: employment, skills and workforce strategy for the fourth industrial revolution, 2016.

[2] Wolter M.I. Zika G. Helmrich R., Weber E. The consequences of industry 4.0 for the labour market and education. In Campbell D. (eds) Bast G., Carayannis E., editor, *The Future of Education and Labor, Arts, Research, Innovation and Society*.

[3] Gábor Nick, Ádám Szaller, Júlia Bergmann, and Tamás Várgedő. Industry 4.0 readiness in hungary: model, and the first results in connection to data application. *IFAC-PapersOnLine*, 52(13):289–294, 2019.

[4] Robert Ulewicz and Kanchana Sethanan. Quality of educational services–industry 4.0 requirements. *Kvaliteta Obrazovnih Usluga-Zahtjevi*, 4:137–149, 2019.

[5] G Nick and F Pongrácz. How to measure industry 4.0 readiness of cities. *Int. Sci. J. Ind*, 4:2, 2016.

[6] Timea Czvetko. Assessing regional aspects of industry 4.0 readiness - development of the industry 4.0+ indicator system. Master's thesis, University of Pannonia, 2019.

[7] Wim Naudé, Aleksander Surdej, and Martin Cameron. The past and future of manufacturing in central and eastern europe: Ready for industry 4.0? 2019.

[8] SOUMITRA Dutta. The networked readiness index 2016. In *World Economic Forum*, 2016.

[9] Daniela Wurhofer, Thomas Meneweger, Verena Fuchsberger, and Manfred Tscheligi. Reflections on operators' and maintenance engineers' experiences of smart factories. In *Proceedings of the 2018 ACM Conference on Supporting Groupwork*, pages 284–296, 2018.

[10] Harold E Price. The allocation of functions in systems. *Human factors*, 27(1):33–45, 1985.

[11] PA Hancock and SF Scallen. The future of function allocation. *Ergonomics in design*, 4(4):24–29, 1996.

[12] Joost CF de Winter and Dimitra Dodou. Why the fitts list has persisted throughout the history of function allocation. *Cognition, Technology & Work*, 16(1):1–11, 2014.

[13] Paul M Fitts. Human engineering for an effective air-navigation and traffic-control system. 1951.

[14] Andreas Bye, Erik Hollnagel, and Tor Steinar Brendeford. Human–machine function allocation: a functional modelling approach. *Reliability Engineering & System Safety*, 64(2):291–300, 1999.

[15] Robert R Hoffman, Paul J Feltovich, Kenneth M Ford, and David D Woods. A rose by any other name... would probably be given an acronym [cognitive systems engineering]. *IEEE Intelligent Systems*, 17(4):72–80, 2002.

[16] Raja Parasuraman, Thomas B Sheridan, and Christopher D Wickens. A model for types and levels of human interaction with automation. *IEEE Transactions on systems, man, and cybernetics-Part A: Systems and Humans*, 30(3):286–297, 2000.

[17] Thomas B Sheridan and William L Verplank. Human and computer control of undersea teleoperators. Technical report, Massachusetts Inst of Tech Cambridge Man-Machine Systems Lab, 1978.

[18] John R Gersh, Jennifer A McKneely, and Roger W Remington. Cognitive engineering: Understanding human interaction with complex systems. *Johns Hopkins APL technical digest*, 26(4):377–382, 2005.

[19] Beth Crandall, Gary Klein, Gary A Klein, and Robert R Hoffman. *Working minds: A practitioner's guide to cognitive task analysis.* Mit Press, 2006.

[20] Daniel P Jenkins, Neville A Stanton, and Guy H Walker. *Cognitive work analysis: coping with complexity.* CRC Press, 2017.

[21] `https://factory2fit.eu/`. [Online; accessed 07/02/2020].

[22] Tamás Ruppert, Szilárd Jaskó, Tibor Holczinger, and János Abonyi. Enabling technologies for operator 4.0: A survey. *Applied Sciences*, 8(9):1650, 2018.

[23] David Romero, Peter Bernus, Ovidiu Noran, Johan Stahre, and Åsa Fast-Berglund. The operator 4.0: human cyber-physical systems & adaptive automation towards human-automation symbiosis work systems. In *IFIP interna-*

tional conference on advances in production management systems, pages 677–686. Springer, 2016.

[24] Gyula Dorgo and Janos Abonyi. Sequence mining based alarm suppression. *IEEE Access*, 6:15365–15379, 2018.

[25] Brainstorming with experienced organization developers at innopod ltd.

[26] Caroline Berggren. The influence of higher education institutions on labor market outcomes. *European Education*, 42(1):61–75, 2010.

[27] Sevilay Uslu Divanoğlu, Dr BAĞCI, et al. Determining the factors affecting individual investors' behaviours. *International Journal of Organizational Leadership*, 7:284–299, 2018.

[28] Yohsuke Murase, János Török, Hang-Hyun Jo, Kimmo Kaski, and János Kertész. Multilayer weighted social network model. *Physical Review E*, 90(5), Nov 2014.

[29] Rick Nason. *It's Not Complicated: The Art and Science of Complexity in Business.* University of Toronto Press, 2017.

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

[31] Duncan J Watts and Steven H Strogatz. Collective dynamics of 'small-world'networks. *Nature*, 393(6684):440, 1998.

[32] Albert-László Barabási. Network science book. *Boston, MA: Center for Complex Network, Northeastern University. Available online at: http://barabasi.com/networksciencebook*, 2014.

[33] Mark Newman. *Networks: An Introduction.* OUP Oxford, Mar 2010. Google-Books-ID: q7HVtpYVfC0C.

[34] Stephen P. Borgatti. Centrality and network flow. *Social Networks*, 27(1):55–71, Jan 2005.

[35] Stephen P. Borgatti and Martin G. Everett. A graph-theoretic perspective on centrality. *Social Networks*, 28(4):466–484, Oct 2006.

[36] Qian Li, Tao Zhou, Linyuan Lü, and Duanbing Chen. Identifying influential spreaders by weighted leaderrank. *Physica A: Statistical Mechanics and its Applications*, 404:47–55, Jun 2014.

[37] Jian-Guo Liu, Zhuo-Ming Ren, and Qiang Guo. Ranking the spreading influence in complex networks. *Physica A: Statistical Mechanics and its Applications*, 392(18):4154–4159, Sep 2013.

[38] Federico Battiston, Vincenzo Nicosia, and Vito Latora. Structural measures for multiplex networks. *Physical Review E*, 89(3), 2014.

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

[40] Michael J. Barber. Modularity and community detection in bipartite networks. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 76(6):1–11, 2007.

[41] P. Expert, T. S. Evans, V. D. Blondel, and R. Lambiotte. Uncovering space-independent communities in spatial networks. *Proceedings of the National Academy of Sciences*, 108(19):7663—-7668, May 2011.

[42] Tanmoy Chakraborty, Sriram Srinivasan, Niloy Ganguly, Animesh Mukherjee, and Sanjukta Bhowmick. Permanence and community structure in complex networks. *arXiv:1606.01543 [physics]*, Jun 2016.

[43] Gergely Palla, Imre Derényi, Illés Farkas, and Tamás Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043):814–818, Jun 2005.

[44] Jörg Reichardt and Stefan Bornholdt. Statistical mechanics of community detection. *Physical Review E*, 74(1):016110, 2006.

[45] Alex Arenas, Alberto Fernandez, and Sergio Gomez. Analysis of the structure of complex networks at different resolution levels. *New Journal of Physics*, 10(5):053039, 2008.

[46] Chuan Wen Loe and Henrik Jeldtoft Jensen. Comparison of communities detection algorithms for multiplex. *Physica A: Statistical Mechanics and its Applications*, 431:29–45, Aug 2015.

[47] Xuelong Li, Marko Jusup, Zhen Wang, Huijia Li, Lei Shi, Boris Podobnik, H Eugene Stanley, Shlomo Havlin, and Stefano Boccaletti. Punishment diminishes the benefits of network reciprocity in social dilemma experiments. *Proceedings of the national academy of sciences*, 115(1):30–35, 2018.

[48] Tiziano Squartini, Francesco Picciolo, Franco Ruzzenenti, and Diego Garlaschelli. Reciprocity of weighted networks. *Scientific Reports*, 3(1), Dec 2013.

[49] Per Block. Reciprocity, transitivity, and the mysterious three-cycle. *Social Networks*, 40:163–173, Jan 2015.

[50] Mark E Dickison, Matteo Magnani, and Luca Rossi. *Multilayer social networks*. Cambridge University Press, 2016.

[51] Albert-Lászlo Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, Oct 1999.

[52] Réka Albert and Albert László Barabási. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74(1):47–97, 2002.

[53] Miller McPherson, Lynn Smith-Lovin, and James M. Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27(1):415–444, 2001.

[54] Sergio Gómez, Alberto Fernández, Sandro Meloni, and Alex Arenas. Impact of origin-destination information in epidemic spreading. *arXiv:1804.02581 [cond-mat, physics:physics]*, Apr 2018.

[55] Duncan S Callaway, Mark EJ Newman, Steven H Strogatz, and Duncan J Watts. Network robustness and fragility: Percolation on random graphs. *Physical review letters*, 85(25):5468, 2000.

[56] Davide Cellai, Eduardo López, Jie Zhou, James P. Gleeson, and Ginestra Bianconi. Percolation in multiplex networks with overlap. *Physical Review E*, 88(5), Nov 2013.

[57] Ginestra Bianconi and Sergey N. Dorogovtsev. Multiple percolation transitions in a configuration model of a network of networks. *Physical Review E*, 89(6), Jun 2014.

[58] Mason A Porter and James P Gleeson. Dynamical systems on networks. *Frontiers in Applied Dynamical Systems: Reviews and Tutorials*, 4, 2016.

[59] Carliss Young Baldwin and Kim B Clark. *Design rules: The power of modularity*, volume 1. MIT press, 2000.

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

[61] Aaron Clauset, Mark EJ Newman, and Cristopher Moore. Finding community structure in very large networks. *Physical review E*, 70(6):066111, 2004.

[62] Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 10008(10):6, 2008.

[63] Martin Rosvall and Carl T Bergstrom. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, 105(4):1118–1123, 2008.

[64] Santo Fortunato. Community detection in graphs. *Physics reports*, 486(3):75–174, 2010.

[65] Santo Fortunato and Marc Barthélemy. Resolution limit in community detection. *Proceedings of the National Academy of Sciences*, 104(1):36–41, 2007.

[66] Jörg Reichardt and Stefan Bornholdt. Statistical mechanics of community detection. *Physical Review E*, 74(1):016110, 2006.

[67] Andrea Lancichinetti, Santo Fortunato, and János Kertész. Detecting the overlapping and hierarchical community structure in complex networks. *New Journal of Physics*, 11(3), Mar 2009.

[68] Huawei Shen, Xueqi Cheng, Kai Cai, and Mao-Bin Hu. Detect overlapping and hierarchical community structure in networks. *Physica A: Statistical Mechanics and its Applications*, 388(8):1706—-1712, Apr 2009.

[69] Zhihao Wu, Youfang Lin, Huaiyu Wan, Shengfeng Tian, and Keyun Hu. Efficient overlapping community detection in huge real-world networks. *Physica A: Statistical Mechanics and its Applications*, 391(7):2475—-2490, Apr 2012.

[70] Belfin R.V., Grace Mary Kanaga E., and Piotr Bródka. Overlapping community detection using superior seed set selection in social networks. *Computers and Electrical Engineering*, 70:1074—-1083, Aug 2018.

[71] Marta Sarzynska, Elizabeth A. Leicht, Gerardo Chowell, and Mason A. Porter. Null models for community detection in spatially embedded, temporal networks. *Journal of Complex Networks*, 4(3):363–406, Sep 2016.

[72] Tanmoy Chakraborty, Ayushi Dalmia, Animesh Mukherjee, and Niloy Ganguly. Metrics for community analysis: A survey. *ACM Comput. Surv.*, 50(4):54:1—-54:37, Aug 2017.

[73] S. Boccaletti, G. Bianconi, R. Criado, C.I. del Genio, J. Gómez-Gardeñes, M. Romance, I. Sendiña-Nadal, Z. Wang, and M. Zanin. The structure and dynamics of multilayer networks. *Physics Reports*, 544(1):1–122, Nov 2014.

[74] M.S. Granovetter. The Strength of Weak Ties. *The American Journal of Sociology*, 78(6):1360–1380, 1973.

[75] Multilayer modeling and analysis of human brain networks. 6.

[76] Michele Berlingerio, Fabio Pinelli, and Francesco Calabrese. Abacus: frequent pattern mining-based community discovery in multidimensional networks. *Data Mining and Knowledge Discovery*, 27(3):294–320, Nov 2013.

[77] Tanuja Shanmukhappa, Ivan WH Ho, K Tse Chi, Xingtang Wu, and Hairong Dong. Multi-layer public transport network analysis. In *2018 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1–5. IEEE, 2018.

[78] Arda Halu, Raúl J. Mondragón, Pietro Panzarasa, and Ginestra Bianconi. Multiplex pagerank. *PLoS ONE*, 8(10):e78293, Oct 2013.

[79] Federico Battiston, Vincenzo Nicosia, and Vito Latora. The new challenges of multiplex networks: Measures and models. *The European Physical Journal Special Topics*, 226(3):401–416, Feb 2017.

[80] ENQA. Standards and Guidelines for Quality Assurance in the European Higher Education Area (ESG). page 32, 2015.

[81] Skuza Agnieszka Sojkin Bogdan, Bartkowiak Paweł. Determinants of higher education choices and student satisfaction: The case of Poland. *Higher Education*, 63(5):565–581, 2012.

[82] Leon Cremonini, Don Westerheijden, and Jürgen Enders. Disseminating the right information to the right audience: Cultural determinants in the use (and misuse) of rankings. *Higher Education*, 55(3):373–385, 2008.

[83] Adela García-Aracil. Effects of college programme characteristics on graduates' performance. *Higher Education*, 69(5):735–757, 2015.

[84] Anders Wallgren and Britt Wallgren. To understand the Possibilities of Administrative Data you must change your Statistical Paradigm! In *Joint Statistical Meetings*, number Section on Survey Research Methods, pages 357–365, 2011.

[85] Attila Gál. Adminisztratív adatok pályakövetési célú felhasználásának nemzetközi gyakorlata (in hungarian). In Orsolya Garai and Zsuzsanna Veroszta, editors, *Államigazgatási adatbázisok a diplomás pályakövetésben*, pages 5–46. 2013.

[86] Jesse M Cunha and Trey Miller. Measuring value-added in higher education: Possibilities and limitations in the use of administrative data. *Economics of Education Review*, 42:64–77, 2014.

[87] I Garam. Study on the relevance of international student mobility to work and employment. *Finnish Employers' View on Benefits of Studying and Work Placement Abroad. Helsinki: Centre for International Mobility*, 2005.

[88] Ellu Saar, Marge Unt, and Irena Kogan. Transition from educational system to labour market in the european union a comparison between new and old members. *International journal of comparative sociology*, 49(1):31–59, 2008.

[89] David Young, Ron Borland, and Ken Coghill. An actor-network theory analysis of policy innovation for smoke-free places: Understanding change in complex systems. *Am J Public Health.*, 100:1208–1207, 2010.

[90] Steven Morris, Gary G Yen, et al. Construction of bipartite and unipartite weighted networks from collections of journal papers. *arXiv preprint physics/0503061*, 2005.

[91] Albert-László Barabási. *Network Science Graph Theory*, chapter 2.7. Creative Commons: CC BY-NC-SA 2.0, 2014.

[92] Albert-László Barabási. *Network science.* 2015.

[93] Ludo Waltman and Nees Jan Van Eck. A smart local moving algorithm for large-scale modularity-based community detection. *The European physical journal B*, 86(11):471, 2013.

[94] Vincent A Traag, Ludo Waltman, and Nees Jan van Eck. From louvain to leiden: guaranteeing well-connected communities. *Scientific reports*, 9(1):1–12, 2019.

[95] Ju Xiang and Ke Hu. Limitation of multi-resolution methods in community detection. *Physica A: Statistical Mechanics and its Applications*, 391(20):4995–5003, 2012.

[96] Santo Fortunato and Marc Barthélemy. Resolution limit in community detection. *Proceedings of the National Academy of Sciences*, 104(1):36–41, 2007.

[97] Jörg Reichardt and Stefan Bornholdt. Detecting fuzzy community structures in complex networks with a Potts model. *Physical Review Letters*, 93(21):218701, 2004.

[98] Jussi M Kumpula, Jari Saramäki, Kimmo Kaski, and János Kertész. Limited resolution and multiresolution methods in complex network community detection. *Fluctuation and Noise Letters*, 7(03):L209–L214, 2007.

[99] Ju Xiang and Ke Hu. Limitation of multi-resolution methods in community detection. *Physica A: Statistical Mechanics and its Applications*, 391(20):4995–5003, 2012.

[100] Szilvia Nyüsti and Zsuzsanna Veroszta. *Diplomás pályakövetési adatok 2013 Adminisztratív adatbázisok integrációja (in Hungarian).* 2013.

[101] Iñaki Iriondo and Teodosio Pérez-Amaral. The effect of educational mismatch on wages in Europe. *Journal of Policy Modeling*, 38(2):304–323, 2016.

[102] Greg J Duncan and Saul D Hoffman. The incidence and wage effects of overeducation. *Economics of Education Review*, 1(1):75–86, 1981.

[103] Richard R Verdugo and Naomi Turner Verdugo. The impact of surplus schooling on earnings: Some additional findings. *Journal of Human Resources*, pages 629–643, 1989.

[104] Joop Hartog. Over-education and earnings: where are we, where should we go? *Economics of education review*, 19(2):131–147, 2000.

[105] Parvinder Kler. Graduate overeducation in australia: A comparison of the mean and objective methods. *Education Economics*, 13(1):47–72, 2005.

[106] Wim Groot and Henriette Maassen Van Den Brink. Overeducation in the labor market: a meta-analysis. *Economics of education review*, 19(2):149–158, 2000.

[107] Seamus McGuinness and Peter J. Sloane. Labour market mismatch among UK graduates: An analysis using REFLEX data. *Economics of Education Review*, 30(1):130–145, 2011.

[108] Chia Yu Hung. Overeducation and undereducation in Taiwan. *Journal of Asian Economics*, 19(2):125–137, 2008.

[109] Inés P. Murillo, Marta Rahona-López, and Maria del Mar Salinas-Jiménez. Effects of educational mismatch on private returns to education: An analysis of the Spanish case (1995-2006). *Journal of Policy Modeling*, 34(5):646–659, 2012.

[110] Colin S. Gillespie. Fitting heavy tailed distributions: The poweRlaw package. *Journal of Statistical Software*, 64(2):1–16, 2015.

[111] Quang H Vuong. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica: Journal of the Econometric Society*, pages 307–333, 1989.

[112] Derek J De Solla Price. Networks of scientific papers. *Science*, pages 510–515, 1965.

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

[114] Albert-László Barabási. Luck or reason. *Nature*, 489(7417):507–508, 2012.

[115] Lucy Kendrick, Katarzyna Musial, and Bogdan Gabrys. Change point detection in social networks – Critical review with experiments. *Computer Science Review*, 29:1–13, 2018.

[116] Katarzyna Musial, Marcin Budka, and Krzysztof Juszczyszyn. Creation and growth of online social network. *World Wide Web*, 16(4):421–447, 2013.

[117] Piotr Bródka, Katarzyna Musial, and Przemysław Kazienko. A method for group extraction in complex social networks. In *World Summit on Knowledge Society*, pages 238–247. Springer, 2010.

[118] Meng Qin, Di Jin, Dongxiao He, Bogdan Gabrys, and Katarzyna Musial. Adaptive community detection incorporating topology and content in social networks. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, pages 675–682. ACM, 2017.

[119] Przemysław Kazienko, Katarzyna Musial, Elżbieta Kukla, Tomasz Kajdanowicz, and Piotr Bródka. Multidimensional social network: model and analysis. In *International Conference on Computational Collective Intelligence*, pages 378–387. Springer, 2011.

[120] S Tamer Cavusgil, Tunga Kiyak, and Sengun Yeniyurt. Complementary approaches to preliminary foreign market opportunity assessment: Country clustering and country ranking. *Industrial Marketing Management*, 33(7):607–617, 2004.

[121] Cristina Del Campo, Carlos MF Monteiro, and Joao Oliveira Soares. The European regional policy and the socio-economic diversity of European regions: A multivariate analysis. *European Journal of Operational Research*, 187(2):600–612, 2008.

[122] Ash Amin. An institutionalist perspective on regional economic development. *International Journal of Urban and Regional Research*, 23(2):365–378, 1999.

[123] David R Bell. *Location is (still) everything: The surprising influence of the real world on how we search, shop, and sell in the virtual one.* Houghton Mifflin Harcourt, 2014.

[124] Ginestra Bianconi and A-L Barabási. Competition and multiscaling in evolving networks. *EPL (Europhysics Letters)*, 54(4):436–442, 2001.

[125] Sergey N Dorogovtsev, José Fernando F Mendes, and Alexander N Samukhin. Structure of growing networks with preferential linking. *Physical Review Letters*, 85(21):4633–4636, 2000.

[126] Aharon Blank and Sorin Solomon. Power laws in cities population, financial markets and internet sites (scaling in systems with a variable number of components). *Physica A: Statistical Mechanics and its Applications*, 287(1-2):279–288, 2000.

[127] Gilles Duranton and Diego Puga. The growth of cities. In *Handbook of Economic Growth*, volume 2, pages 781–853. Elsevier, 2014.

[128] Xavier Gabaix. Zipf's law for cities: an explanation. *The Quarterly Journal of Economics*, 114(3):739–767, 1999.

[129] Matthieu Cristelli, Michael Batty, and Luciano Pietronero. There is more than a power law in Zipf. *Scientific Reports*, 2(812):1–7, 2012.

[130] Meredith Reba, Femke Reitsma, and Karen C Seto. Spatializing 6,000 years of global urbanization from 3700 BC to AD 2000. *Scientific Data*, 3:160034, 2016.

[131] Marc Barthélemy. Spatial networks. *Physics Reports*, 499(1–3):1—-101, Feb 2011.

[132] George Kingsley Zipf. *Human behavior and the principle of least effort.* Addison-Wesley Press, 1949.

[133] Salvatore Scellato, Cecilia Mascolo, Mirco Musolesi, and Vito Latora. Distance matters: Geo-social metrics for online social networks. In *Proceedings of the 3rd Conference on Online Social Networks*, WOSN'10. USENIX Association, 2010.

[134] Renaud Lambiotte, Vincent D. Blondel, Cristobald de Kerchove, Etienne Huens, Christophe Prieur, Zbigniew Smoreda, and Paul Van Dooren. Geographical dispersal of mobile communication networks. *Physica A: Statistical Mechanics and its Applications*, 387(21):5317—-5325, Sep 2008.

[135] David Liben-Nowell, Jasmine Novak, Ravi Kumar, Prabhakar Raghavan, and Andrew Tomkins. Geographic routing in social networks. *Proceedings of the National Academy of Sciences*, 102(33):11623—11628, Aug 2005.

[136] Lars Backstrom, Eric Sun, and Cameron Marlow. Find me if you can: Improving geographical prediction with social and spatial proximity. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 61—70. ACM, 2010.

[137] Jacob Goldenberg and Moshe Levy. Distance is not dead: Social interaction and geographical distance in the internet era. *arXiv:0906.3202 [physics]*, Jun 2009.

[138] Aisling Reynolds-Feighan and Peter McLay. Accessibility and attractiveness of European airports: A simple small community perspective. *Journal of Air Transport Management*, 12(6):313–323, 2006.

[139] Alexander Peter Groh and Matthias Wich. A composite measure to determine a host country's attractiveness for foreign direct investment. *SSRN Electronic Journal*, 833(11):1–30, 2009.

[140] Charles E Gearing, William W Swart, and Turgut Var. Establishing a measure of touristic attractiveness. *Journal of Travel Research*, 12(4):1–8, 1974.

[141] James E Anderson. The gravity model. *Annual Review of Economics*, 3(1):133–160, 2011.

[142] James E. Anderson. A theoretical foundation for the gravity equation. *The American Economic Review*, 69(1):106–116, 1979.

[143] Kunal Bhattacharya, Gautam Mukherjee, Jari Saramäki, Kimmo Kaski, and Subhrangshu S Manna. The international trade network: weighted network analysis and modelling. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(2):P02002, 2008.

[144] Moshe Levy. Scale-free human migration and the geography of social networks. *Physica A: Statistical Mechanics and its Applications*, 389(21):4913–4917, 2010.

[145] Woo-Sung Jung, Fengzhong Wang, and H. Eugene Stanley. Gravity model in the Korean highway. *EPL (Europhysics Letters)*, 81(4):48005, 2008.

[146] Akanda Wahid-Ul-Ashraf, Marcin Budka, and Katarzyna Musial-Gabrys. Newton's gravitational law for link prediction in social networks. In *International Workshop on Complex Networks and their Applications*, pages 93–104. Springer, 2017.

[147] E. A. Leicht and M. E. J. Newman. Community structure in directed networks. *Physical Review Letters*, 100(11):118703, Mar 2008.

[148] Remy Cazabet, Pierre Borgnat, and Pablo Jensen. Using degree constrained gravity null-models to understand the structure of journeys' networks in bicycle sharing systems. In *ESANN 2017 - European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, Apr 2017.

[149] Xin Liu, Tsuyoshi Murata, and Ken Wakita. Extending modularity by incorporating distance functions in the null model. *CoRR*, abs/1210.4007, 2012.

[150] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008.

[151] Philipp Schuetz and Amedeo Caflisch. Efficient modularity optimization by multistep greedy algorithm and vertex mover refinement. *Physical Review E*, 77(4):046112, 2008.

[152] Balázs Lengyel, Attila Varga, Bence Ságvári, Ákos Jakobi, and János Kertész. Geographies of an online social network. *PLoS ONE*, 10(9):1–13, 2015.

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

[154] Jaewon Yang and Jure Leskovec. Defining and evaluating network communities based on ground-truth. *Knowledge and Information Systems*, 42(1):181–213, 2015.

[155] Gautier Krings, Francesco Calabrese, Carlo Ratti, and Vincent D Blondel. Urban gravity: a model for inter-city telecommunication flows. *Journal of Statistical Mechanics: Theory and Experiment*, 2009(7):L07003, 2009.

[156] Duygu Balcan, Vittoria Colizza, Bruno Gonçalves, Hao Hu, José J. Ramasco, and Alessandro Vespignani. Multiscale mobility networks and the spatial spreading of infectious diseases. *Proceedings of the National Academy of Sciences*, 106(51), 2009.

[157] Pablo Kaluza, Andrea Kölzsch, Michael T. Gastner, and Bernd Blasius. The complex network of global cargo ship movements. *Journal of The Royal Society Interface*, 7(48):1093–1103, 2010.

[158] Erzsébet Ravasz and Albert-László Barabási. Hierarchical organization in complex networks. *Physical Review E*, 67(2):026112, 2003.

[159] Sergey N Dorogovtsev, Alexander V Goltsev, and José Ferreira F Mendes. Pseudofractal scale-free web. *Physical Review E*, 65(6):66–122, 2002.

[160] Inderjit S Jutla, Lucas GS Jeub, and Peter J Mucha. A generalized Louvain method for community detection implemented in MATLAB. *URL http://netwiki. amath. unc. edu/GenLouvain*, 2011.

[161] Peter J Mucha, Thomas Richardson, Kevin Macon, Mason A Porter, and Jukka-Pekka Onnela. Community structure in time-dependent, multiscale, and multiplex networks. *Science*, 328(5980):876–878, 2010.

[162] Noel M Tichy. Networks in organizations. *Handbook of organizational design*, (2), 1981.

[163] Noel M. Tichy, Michael L. Tushman, and Charles Fombrun. Social network analysis for organizations. *The Academy of Management Review*, 4(4):507–519, Oct 1979.

[164] David Krackhardt and Daniel J Brass. *Intraorganizational networks: The micro side*. Sage Publications, Inc, 1994.

[165] Herminia Ibarra and Steven B. Andrews. Power, social influence, and sense making: Effects of network centrality and proximity on employee perceptions. *Administrative Science Quarterly*, 38(2):277–303, Jun 1993.

[166] Dorothy R Carter, Leslie A DeChurch, Michael T Braun, and Noshir S Contractor. Social network approaches to leadership: An integrative conceptual review. *Journal of Applied Psychology*, 100(3):597, 2015.

[167] David Krackhardt and Martin Kilduff. Friendship patterns and culture: The control of organizational diversity. *American Anthropologist*, 92(1):142–154, Mar 1990.

[168] Daniel J Brass. A social network perspective on organizational citizenship behavior. In *The Oxford Handbook of Organizational Citizenship Behavior*, page 317. Oxford University Press, 2018.

[169] Zhan Bu, Huijia Li, Jie Cao, Zhiang Wu, and Lu Zhang. Game theory based emotional evolution analysis for chinese online reviews. *Knowledge-Based Systems*, 103:60–72, 2016.

[170] Yezheng Liu, Lingfei Li, Hai Wang, Chunhua Sun, Xiayu Chen, Jianmin He, and Yuanchun Jiang. The competition of homophily and popularity in growing and evolving social networks. *Scientific reports*, 8, 2018.

[171] Yuan Yuan, Ahmad Alabdulkareem, et al. An interpretable approach for social network formation among heterogeneous agents. *Nature communications*, 9(1):4704, 2018.

[172] Peter J Carrington, John Scott, and Stanley Wasserman. *Models and methods in social network analysis*, volume 28. Cambridge university press, 2005.

[173] Emmanuel Lazega, Lise Mounier, Tom Snijders, and Paola Tubaro. Norms, status and the dynamics of advice networks: A case study. *Social Networks*, 34(3):323–332, 2012.

[174] Stephen P Borgatti and Daniel S Halgin. On network theory. *Organization science*, 22(5):1168–1181, 2011.

[175] Thomas J. Zagenczyk, Russell L. Purvis, Mindy K. Shoss, Kristin L. Scott, and Kevin S. Cruz. Social influence and leader perceptions: Multiplex social network ties and similarity in leader–member exchange. *Journal of Business and Psychology*, 30(1):105–117, Mar 2015.

[176] Anthony Giddens. *The constitution of society: Outline of the theory of structuration*. Univ of California Press, 1984.

[177] Robert Whitbred, Fabio Fonti, Christian Steglich, and Noshir Contractor. From microactions to macrostructure and back: A structurational approach to the evolution of organizational networks. *Human Communication Research*, 37(3):404–433, 2011.

[178] Mark Granovetter. Economic action and social structure: The problem of embeddedness. *American journal of sociology*, 91(3):481–510, 1985.

[179] Joseph Galaskiewicz and Wolfgang Bielefeld. *Nonprofit organizations in an age of uncertainty: A study of organizational change*. Transaction Publishers, 1998.

[180] Stanley Wasserman and Katherine Faust. *Social network analysis: Methods and applications*, volume 8. Cambridge university press, 1994.

[181] Peter R Monge and Noshir S Contractor. *Theories of communication networks*. Oxford University Press, USA, 2003.

[182] Noshir Contractor, Peter Monge, and Paul M. Leonardi. Multidimensional networks and the dynamics of sociomateriality: bringing technology inside the network. *International Journal of Communication*, 5:682–720, 2011.

[183] Radosław Michalski and Przemysław Kazienko. *Social network analysis in organizational structures evaluation*, page 1832–1844. Springer, 2014.

[184] Meng Cai, Wei Wang, Ying Cui, and H Eugene Stanley. Multiplex network analysis of employee performance and employee social relationships. *Physica A: Statistical Mechanics and its Applications*, 490:1–12, 2018.

[185] Wen Zhou, Weidong Bao, Xiaomin Zhu, Ji Wang, and Chao Chen. Integrating relationships and attributes: A model of multilayer networks. In *Data Science in Cyberspace (DSC), IEEE International Conference on*, pages 127–136. IEEE, 2016.

[186] M. Kivela, A. Arenas, M. Barthelemy, J. P. Gleeson, Y. Moreno, and M. A. Porter. Multilayer networks. *Journal of Complex Networks*, 2(3):203–271, Sep 2014.

[187] M. Magnani and L. Rossi. *The ML-Model for Multi-layer Social Networks*, page 5–12. Jul 2011.

[188] Annalisa Socievole, Floriano De Rango, and Antonio Caputo. *Wireless contacts, Facebook friendships and interests: Analysis of a multi-layer social network in an academic environment*, page 1–7. IEEE, 2014.

[189] Liu Weiyi, Chen Lingli, and Hu Guangmin. Mining essential relationships under multiplex networks. *arXiv preprint arXiv:1511.09134*, 2015.

[190] Kyu-Min Lee, Byungjoon Min, and Kwang-Il Goh. Towards real-world complexity: an introduction to multiplex networks. *The European Physical Journal B*, 88(2), Feb 2015.

[191] Anders Mollgaard, Ingo Zettler, Jesper Dammeyer, Mogens H. Jensen, Sune Lehmann, and Joachim Mathiesen. Measure of node similarity in multilayer networks. *PLOS ONE*, 11(6), Jun 2016.

[192] Alves Iván and Inaki Aldasoro. Multiplex interbank networks and systemic importance: An application to european data. *Journal of Financial Stability*, 35:17–37, 2018.

[193] Paul W Holland and Samuel Leinhardt. The structural implications of measurement error in sociometry. *Journal of Mathematical Sociology*, 3(1):85–111, 1973.

[194] Federico Battiston, Vincenzo Nicosia, and Vito Latora. The new challenges of multiplex networks: Measures and models. *The European Physical Journal Special Topics*, 226(3):401–416, Feb 2017.

[195] Christian Borgelt. Frequent item set mining. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2(6):437–456, Nov 2012.

[196] Xifeng Yan and Jiawei Han. gspan: Graph-based substructure pattern mining. In *2002 IEEE International Conference on Data Mining, 2002. Proceedings.*, pages 721–724. IEEE, 2002.

[197] Lise Getoor and Christopher P Diehl. Link mining: a survey. *Acm Sigkdd Explorations Newsletter*, 7(2):3–12, 2005.

[198] J. Chen, W. Hsu, M. L. Lee, and S. K. Ng. *Labeling network motifs in protein interactomes for protein function prediction*, page 546–555. Apr 2007.

[199] Vanessa Queiroz Marinho, Graeme Hirst, and Diego Raphael Amancio. Labelled network subgraphs reveal stylistic subtleties in written texts. *Journal of Complex Networks*, 2017.

[200] T Imielinski, Arun Swami, and R Agarwal. Mining association rules between sets of items in large databases. page 207–216. ACM Press, 1993.

[201] Jiawei Han, Jian Pei, and Micheline Kamber. *Data Mining: Concepts and Techniques.* Elsevier, Jun 2011.

[202] Michael Hahsler and Sudheer Chelluboina. Visualizing association rules: Introduction to the r-extension package arulesviz. *R project module*, page 223–238, 2011.

[203] G. Piatetsky-Shapiro. Discovery, analysis and presentation of strong rules. In G. Piatetsky-Shapiro and W. J. Frawley, editors, *Knowledge Discovery in Databases*, pages 229–248. AAAI Press, 1991.

[204] Markus Hegland. The apriori algorithm–a tutorial. In *Mathematics and computation in imaging science and information processing*, pages 209–262. World Scientific, 2007.

[205] S. P. Borgatti and P. C. Foster. The network paradigm in organizational research: A review and typology. *Journal of Management*, 29(6):991–1013, Dec 2003.

[206] Jill E Perry-Smith and Christina E Shalley. The social side of creativity: A static and dynamic social network perspective. *Academy of management review*, 28(1):89–106, 2003.

[207] Mark EJ Newman. Modularity and community structure in networks. *Proceedings of the national academy of sciences*, 103(23):8577–8582, 2006.

[208] Hui-Jia Li, Zhan Bu, Aihua Li, Zhidong Liu, and Yong Shi. Fast and accurate mining the community structure: Integrating center locating and membership optimization. *IEEE Transactions on Knowledge and Data Engineering*, 28(9):2349–2362, Sep 2016.

[209] Hui-Jia Li and Jasmine J. Daniels. Social significance of community structure: Statistical view. *Physical Review E*, 91(1), Jan 2015.

[210] Joerg Reichardt and Stefan Bornholdt. Statistical mechanics of community detection. *Physical Review E*, 74(1), Jul 2006.

[211] Hui-Jia Li, Hao Wang, and Luonan Chen. Measuring robustness of community structure in complex networks. *EPL (Europhysics Letters)*, 108(6):68009, Dec 2014.

[212] Antony Unwin, Heike Hofmann, and Klaus Bernt. The twokey plot for multiple association rules control. In *Principles of Data Mining and Knowledge Discovery*, page 472–483. Springer, Berlin, Heidelberg, Sep 2001.

[213] Dawn Iacobucci, Rebecca McBride, Deidre L Popovich, and Maria Rouziou. In social network analysis, which centrality index should i use?: Theoretical differences and empirical similarities among top centralities. *Journal of Methods and Measurement in the Social Sciences*, 8(2):72–99, 2017.

[214] Alan P Reynolds, Graeme Richards, Beatriz de la Iglesia, and Victor J Rayward-Smith. Clustering rules: a comparison of partitioning and hierarchical clustering algorithms. *Journal of Mathematical Modelling and Algorithms*, 5(4):475–504, 2006.

[215] Mark Van der Laan, Katherine Pollard, and Jennifer Bryan. A new partitioning around medoids algorithm. *Journal of Statistical Computation and Simulation*, 73(8):575–584, 2003.

[216] Aruna Bhat. K-medoids clustering using partitioning around medoids for performing face recognition. *International Journal of Soft Computing, Mathematics and Control*, 3(3):1–12, 2014.

[217] T Velmurugan and T Santhanam. Computational complexity between k-means and k-medoids clustering algorithms for normal and uniform distributions of data points. *Journal of computer science*, 6(3):363, 2010.

[218] Laszlo Gadar and Janos Abonyi. Graph configuration model based evaluation of the education-occupation match. *PloS one*, 13(3):e0192427, 2018.

[219] Laszlo Gadar, Zsolt T Kosztyan, and Janos Abonyi. The settlement structure is reflected in personal investments: distance-dependent network modularity-based measurement of regional attractiveness. *Complexity*, 2018, 2018.

[220] László Gadár and János Abonyi. Frequent pattern mining in multidimensional organizational networks. *Scientific reports*, 9(1):1–12, 2019.