NETWORK ANALYSIS USING STRUCTURAL, TEXTUAL AND TEMPORAL DATA MODALITY

Mario Karlovčec

Doctoral Dissertation
Jožef Stefan International Postgraduate School
Ljubljana, Slovenia

Supervisor: Prof. Dr. Dunja Mladenić, Jožef Stefan Institute, Jožef Stefan International Postgraduate School, Ljubljana, Slovenia

Evaluation Board:
Asst. Prof. Dr. Bernard Ženko, Chair, Jožef Stefan Institute, Ljubljana, Slovenia
Asst. Prof. Dr. Irena Nančovska Šerbec, Member, Faculty of Education, University of Ljubljana, Ljubljana, Slovenia
Prof. Dr. Jasminka Dobša, Member, Faculty of Organization and Informatics, University of Zagreb, Varaždin, Croatia

MEDNARODNA PODIPLOMSKA ŠOLA JOŽEFA STEFANA
JOŽEF STEFAN INTERNATIONAL POSTGRADUATE SCHOOL



Mario Karlovčec

NETWORK ANALYSIS USING STRUCTURAL, TEXTUAL AND TEMPORAL DATA MODALITY

Doctoral Dissertation

ANALIZA OMREŽIJ NA OSNOVI STRUKTURNE, TEKSTOVNE IN ČASOVNE VSEBINE

Doktorska disertacija

Supervisor: Prof. Dr. Dunja Mladenić

Ljubljana, Slovenia, October 2015

Baki Ani

Acknowledgments

I would like to thank my PhD supervisor, Professor Dunja Mladenić for invaluable guidance and support. I would also like to thank Marko Grobelnik and Mitja Jermol for providing inspiration for research work. Special thanks go to all of the members of the Artificial Intelligence Laboratory at the Jožef Stefan Institute, especially Primož Škraba and Jan Rupnik.

My gratitude goes to the members of my doctoral committee, Prof. Bernard Ženko, Prof. Irena Nančovska Šerbec and Prof. Jasminka Dobša, for their valuable comments and remarks.

I would also like to thank following colleagues, mentors and teachers: Borut Lužar, Matjaž Krnc, Riste Škrekovski, Blaženka Divjak, Mirko Čubrilo, Miroslav Bača, Nevenka Mlinarić, Danica Resman, Mark Jarecke, Rayid Ghani and Matthew Gee.

Děkuji Ti Terezo, že jsi vždy stála na mé straně.

Hvala majci Mariji, ocu Stanku, sestrama Sanji i Danijeli, te baki Ani.

Abstract

Network analysis is becoming increasingly important for understanding and analyzing networks of social interactions. This thesis unifies work on developing methods for analyzing networks using structural, textual and temporal data modality.

The first topic of our research is development of an algorithmic approach for identifying important subsets of nodes in the sense of cardinality of the union of neighborhoods. We describe the polynomial algorithm for finding the group with maximal degree centralization, with the best-possible approximation ratio. The constructed algorithm is tested on six real-world networks. In experiments we observe a property of unimodality of group degree centralization for group size parameter.

We also propose an approach to using the network structure as captured by the core-periphery partition, to study evolution of social networks over time. We use machine learning techniques on one side and analyze parameters of ego-networks to reveal the differences between individuals in the cores on the other. In particular, we analyzed membership in the cores of the scientific coauthoring networks, dynamics of individuals related to the network cores, and ego-networks of all individuals.

Another important aspect of our research is investigating interdisciplinarity of scientific fields based on graph of collaboration between the researchers. A new measure for network diversity is proposed that takes into account graph content and structure.

Finally, we propose an approach for conceptualizing science using collaboration and competences of the scientists. The approach is demonstrated on Slovenian national research data, in a publicly available web application named Science Atlas.

Povzetek

Analiza omrežij postaja vse bolj pomembna za razumevanje in analizo družbenih interakcij. Ta doktorska disertacija predstavlja razvoj metod za analizo omrežij s pomočjo strukturne, besedilne in časovne podatkovne komponente.

Prvi del disertacije se ukvarja z razvojem algoritemskega pristopa za določitev pomembnih vozlišč v smislu kardinalnosti z unijo vseh sosedov vozlišč v omrežjuṘazvili smo polinomski algoritem za iskanje skupine z maksimalno centralizacijo, ki ima najboljšo možno mejo aproksimacije. Zgrajeni algoritem je testiran na šestih realnih omrežjih. Pri poskusih opazujemo lastnost unimodalnosti centralizacije kot funkcije velikosti skupine.

V doktoratu predlagamo pristop za uporabo strukture omrežja, zajete s particijo na jedro in periferijo, za študij razvoja socijalnih omrežij skozi čas. Uporabljamo metode strojnega učenja za analizo ego-omrežij, s čimer odkrivamo različne vzorce obnašanja posameznikov, ki so del omrežja.

Naslednji del disertacije se ukvarja z razvojem metrike, ki uporablja strukturo omrežja in tekstualno komponento podatkov za določanje raznovrstnosti omrežja. To metriko uporabljamo za analizo interdisciplinarnosti raziskovalnega sodelovanja v Sloveniji.

Metode za vizualizacijo ter analizo teksta in omrežij se lahko uporabijo za celovit prikaz nekega družbenega sistema. V zadnjem delu disertacije se ukvarjamo z konceptualizacijo znanosti v Sloveniji s uporabo strukturalne, tekstualne in časovne komponente podatkov. Konceptualizacija je implementirana v javno dostopni spletni aplikaciji Atlas slovenske znanosti.

Contents

List of Figures **xvii**

List of Tables **xxi**

List of Algorithms **xxiii**

Symbols **xxv**

Glossary **xxvii**

1 Introduction **1**
1.1 General Introduction . . . 1
1.2 Hypothesis and Goals . . . 1
1.3 Scientific Contributions . . . 2
1.4 Thesis Structure . . . 2

2 Using Network Structure for Determining Importance of Groups **3**
2.1 Related Work . . . 3
2.2 Preliminaries . . . 5
2.3 Evaluating Degree Centralization . . . 6
2.4 The Algorithm . . . 8
2.5 Experiments . . . 11
2.5.1 Complexity analysis . . . 11
2.5.2 Datasets . . . 12
2.5.3 Experimental results . . . 12
2.6 Conclusion and Future Work . . . 14

3 Analyzing Core-Periphery Structural Property Dynamics **15**
3.1 Related Work . . . 16
3.2 Core in Collaboration Networks . . . 17
3.3 Core-Periphery Analysis Through Ego-Networks . . . 20
3.3.1 Order, size and density . . . 20
3.3.2 Selectivity and disparity . . . 21
3.3.3 Ego-betweenness . . . 21
3.3.4 Connectivity and efficiency . . . 22
3.3.5 Collaboration in cores . . . 23
3.3.6 Maximum k-cores . . . 24
3.4 Describing Core Dynamics on the Level of an Individual Using Machine Learning . . . 25
3.4.1 Feature construction . . . 25
3.4.2 Clustering and classification . . . 27
3.4.2.1 Stability . . . 27

3.4.2.2 Balance . . . 28
3.4.2.3 Strength . . . 29
3.5 Discussion . . . 30
3.6 Conclusion and Future Work . . . 32

4 Measuring Diversity of Network Elements Using Text and Structure 35
4.1 Background . . . 36
4.1.1 Basic concepts . . . 36
4.1.2 Related work . . . 37
4.1.3 Interdisciplinarity and collaboration . . . 38
4.2 Data Description . . . 39
4.3 Approach . . . 39
4.3.1 Interdisciplinarity measure . . . 39
4.3.2 Extension of Stirling's diversity for graphs . . . 40
4.3.3 Subject categories similarity matrix . . . 41
4.4 Results . . . 43
4.4.1 Interdisciplinarity of science fields and subfields based on project collaboration and co-authoring publications (1996-2013) . . . 43
4.4.2 Interdisciplinarity based on project collaboration . . . 44
4.4.3 Interdisciplinarity based on co-authoring publications . . . 47
4.4.4 Comparison of results based on project and publication collaboration 49
4.4.5 Evolution of interdisciplinarity . . . 50

5 Conceptualization of Science Using Collaboration and Competences 55
5.1 Related Work . . . 56
5.2 Approach . . . 58
5.2.1 Basic units . . . 58
5.2.2 Analysis and data models . . . 58
5.2.3 Visualization layer . . . 59
5.3 Data . . . 59
5.4 Use Cases . . . 60
5.4.1 Humanities use-case . . . 61
5.4.1.1 Bridges and cut-vertices . . . 61
5.4.1.2 Science groups collaboration diversity . . . 62
5.4.1.3 Network analysis . . . 63
5.4.1.4 Using other services . . . 63
5.4.2 Technical sciences use-case . . . 64
5.4.3 Natural sciences use-case . . . 64
5.4.4 General topic use case . . . 65
5.4.5 Competence maps of research topics and individuals . . . 66
5.4.5.1 Competence map of an individual researcher . . . 66
5.5 Results . . . 67
5.6 Conclusion and Future Work . . . 70

6 Conclusions and Future Work 73

Appendix A Implementation of GDC in C++ Using SNAP 75

Appendix B Implementationof the Diversity Measure Based on Connectivity and Text Similarity in C++ 81

Appendix C Science Atlas Application 85

C.1 Web service 85

References **89**

Bibliography **97**

Biography **99**

List of Figures

Figure 2.1: The graphic representation of Freeman centralization of group degree centrality of six networks with different sizes of groups. 13

Figure 3.1: Collaboration networks in 1980, 1990, 2000, and 2010. It is evident that in the period of ten years the number of nodes and edges strongly increases. Nodes contained in the cores of the networks are marked. . . 18
Figure 3.2: The growth of the order of cores and the whole networks in logarithmic scale. 18
Figure 3.3: Distribution of researchers within seven science fields for the core researchers (left) and overall distribution (right). 19
Figure 3.4: The number of connected components together with the proportional size of the largest component in terms of the number of vertices (left) and the diameter and maximum separation distance between core vertices during the years (right). 19
Figure 3.5: Average order (left) and size (right) of ego-networks for core vertices and overall. 21
Figure 3.6: Average values of selectivity (left) and disparity (right). 22
Figure 3.7: Average values of ego-betweennesses. 22
Figure 3.8: Average values of effective size in ego-networks (left) and average values of efficiency (right). 23
Figure 3.9: Average number of neighbors from the core (left), and the distribution of values of core neighbors in 2013 (right). 24
Figure 3.10: Average degree of maximum k-cores (left), and the distribution of maximum k-core values in 2013. 24
Figure 3.11: Architecture of the approach for studing evolution of bibliographic network structure as captured by the core-periphery analysis. Notice that a network for year m contains information from all the previous years (e.g. 1971 contains information for 1970 and 1971, as our data starts with 1970).In the network we have n researchers (R1 to Rn) and we have defined k features (f1 to fk) for describing them. 25
Figure 3.12: An illustration of a core membership information for an arbitrarily chosen researcher. The researcher in this example has periphery position in period from 1970 to 1987. In 1988 he enters the core of the network and in 2000 moves back to the periphery. 27
Figure 3.13: Stability clustering medoids. 28
Figure 3.14: Decision tree classification model for stability. 28
Figure 3.15: Balance clustering medoids. 29
Figure 3.16: Decision tree classification model for balance. 29
Figure 3.17: Strength clustering medoids. 30
Figure 3.18: Decision tree classification model for strength. 30

Figure 3.19: Core-periphery dynamics time-series data for researchers A, B and C from collaboration networks, with the year of obtaining their PhD title marked. 31

Figure 4.1: Schematic representation of the attributes of diversity, based on (Stirling, 1998). 40
Figure 4.2: Graphs with the same set of nodes from 4 groups and different set of edges. From a–h the groups in the graph are increasingly more connected. Different shades represent different scientific fields. 41
Figure 4.3: Share of research collaboration of each science with the other sciences based on research projects collaboration. 45
Figure 4.4: Probability density functions of ego-networks diversity index based on project collaboration for science fields from Table 4.5. 46
Figure 4.5: Interdisciplinarity of research fields from 1996 to 2013, based on collaboration on research projects 47
Figure 4.6: Collaboration of researchers from Medical science with researchers from other sciences over time (%). 47
Figure 4.7: Interdisciplinarity of research fields from 1996 to 2013, based on collaboration on research publications. 48
Figure 4.8: Interdisciplinarity of research fields from 1996 to 2013, based on co-authoring publications. 48
Figure 4.9: Share of research collaboration of each science with the other sciences based on research projects collaboration. 49
Figure 4.10: Probability density functions of ego networks diversity index based on co-authoring for science fields from Table 7. Diversity is on x-axis and frequency on y-axis. Intervals for x-axis are equal for each histogram (0–0.4), while intervals for y-axis differ depending on the size of the field. There are 20 classes of width 0.02 for each diagram except for Interdisciplinarity, which has 8 classes of size 0.05. Vertical lines indicate mean values. 51
Figure 4.11: Share of research collaboration of each science with the other sciences based on co-authoring publications. 52
Figure 4.12: Interdisciplinarity of science subfields based on projects and publication collaboration. Symbols in each quadrant represent the subfields in the corresponding quadrant. The subfields are ordered from more to less interdisciplinary. For example, Communication technologies represented with a rhombus in the upper right corner have high interdisciplinarity according to both projects and publications. On the other hand rhombus on the lower left corner represents Mathematics with low interdisciplinarity. 53
Figure 4.13: Average (mean) number of researchers peer project and publications from 1996 to 2013. 54

Figure 5.1: Collaboration graph for topic anthropology with manualy marked bridges (1,2 and 3), cut-vertices (A,B,..., I) and clusters (a,b,..., f). 61
Figure 5.2: Collaboration graph of an interdisciplinary architecture project. 62
Figure 5.3: Collaboration graph for topic artificial intelligence. 64
Figure 5.4: Collaboration graph for topic nanotechnology. 65
Figure 5.5: Collaboration graph for topic Slovenia. 66
Figure 5.6: Competence map of artificial intelligence topic. 67

Figure 5.7: Competence map of a researcher. Grey nodes represent projects, while blue nodes represent keywords. Projects are connected on the basis of common keywords. 69

Figure C.1: Science Atlas application architecture 85

List of Tables

Table 2.1: Some statistics and centrality results from our experiments. 13

Table 3.1: Descriptions of features derived from core-periphery time series. Each year a researcher can be in the core (position one), or in the periphery (position zero). 26
Table 3.2: Classification of researchers from the illustrative example. 31
Table 3.3: Ego-network parameters of the researchers A, B, and C in the year 2013. 32

Table 4.1: Measures of interdisciplinarity used in this study. 40
Table 4.2: Stirling's diversity versus the extension for networks, measured on illustrative graphs from Figure 4.2. 42
Table 4.3: Subject categories similarity matrix for 7 scientific fields. 42
Table 4.4: Pairs with extremely high values of similarity from the constructed scientific subfields similarity matrix. 43
Table 4.5: Interdisciplinarity of science fields measured by research project collaboration 45
Table 4.6: Percentage of research collaboration on projects between fields. The table shows how much percentage of research collaboration a field from row has with eachfield from the columns. E.g. in average 18.4%of collaborators of Medical science are from Natural science and 10.63% are from Technical science. The value on the diagonal shows the collaboration within a science, e.g., 75.14% for Humanistic. Bold values show the largest percentages of collaboration of science fields from rows with other science fields. 50
Table 4.7: Interdisciplinarity of science fields measured on the basis of co-authoring publications. Bold values indicate the largest values of mean diversity, variety, balance, disparity and Shannon's diversity. 50

Table 5.1: Attributes of main entities in SICRIS database. 60
Table 5.2: Topics and competences generated with competence map. 68
Table 5.3: Results of analysis of collaboration graphs from the use-cases. 69
Table 5.4: Examples of static keywords, keywords defined with competence and the relation between them. 70

List of Algorithms

Algorithm 2.1: Finding group with maximal degree centralization 9
Algorithm 2.2: Updating variables A, B, C and sum. 10
Algorithm 2.3: Functions *contribution* and *decreaseContribution*. The former outputs the contribution of v while the latter refreshes the dictionary *histogram* whenever the contribution of node v decreases by one. . 11

Symbols

G	...	Graph
v	...	a node
e	...	an edge
n	...	number of nodes
m	...	number of edges
$V(G)$	...	set of nodes of graph G
$E(G)$	...	set of edges of graph G
$\|V(G)\|$	...	cardinality of a set of nodes
$\|E(G)\|$	...	cardinality of a set of edges
$deg(v)$	...	node degree - number of adjacent nodes.
$N(v)$	...	set of neighbors of a node.
$\mathrm{GD}(S, G)$	...	group degree centrality of a set S in a graph G.
$\mathrm{GD}_1(S, G)$	...	group degree centralization of a set S in a graph G.

Glossary

Network is a structure that reflects the nature of interactions between entities.
Social network is a network of social interactions, with nodes often representing people.
Graph is an ordered pair $G = (V(G), E(G))$, comprising a set of nodes $V(G)$ and a set of edges $E(G)$. Graphs are mathematical constructs used for modeling networks.
*Tree*is an undirected simple graph that is connected and has no cycles.
Star is a tree with one internal node and $n - 1$ leaves.
Centrality is an indicator of node importance in terms of a real-valued function on the nodes of a graph.
Centralization is a graph-level score based on graph-level centrality, normalized by dividing by the maximum theoretical score for a graph with the same number of nodes.
Group centrality is an indicator of importance of a group of nodes in a network.
Maximum domination problem is a problem of finding a subset of nodes with size k that maximizes the cardinality of dominated nodes.
Ego-network is a network where set of nodes is determined by one node and its usualy one-step neighborhood.
Co-authorship network is a network of authors that are connected if they coauthored a publication.
Stirling's diversity is an index that combines variety, balance and disparity, to measure the diversity of a set of nodes.
Subject categories similarity matrix is a gives the pairwise similarities between subject categories, which in this work refer to seven scientific disciplines.
Cosine distance is a measure of similarity between two vectors of an inner product space that measures the cosine of the angle between them. It is a measure commonly used for comparing documents which are represented in high-dimensional spaces.

Chapter 1

Introduction

1.1 General Introduction

Network, which can be defined as an interconnected system of entities or people, recently became an extremely exposed representation for social and technical systems. The main cause for this is the development and global usage of Internet. Today, people can very easily communicate over huge distances using email and social networking applications. News, ideas, knowledge and many other kinds of information are being created, distributed and spread in different networks that connect people. Alongside the social networks, the network is becoming the underlying data model for technical systems. The developments in artificial intelligence are oriented towards connecting data in order to build intelligent systems. With connections, raw pieces of data can become concepts with relations, building network structures from which knowledge can be extracted.

This aims at developing methods for network analysis using three data modalities: structure, text and time. We use machine learning and text mining approaches together with the temporal component to develop new measures and methods for network analysis. The main interest of our research has been in understanding the position and role of individuals and groups in social networks. We have focused on developing methods that can efficiently identify groups of important nodes in large networks with millions of nodes. We have developed an approach that can characterize individual nodes according to the dynamics of some structural property, such as core-periphery position. Another focus of our research was developing a metric for measuring diversity of groups and individual network members. The measure takes into consideration both structural properties and the textual content related to the nodes. Finally, we have developed an approach for conceptualizing science using network and text.

1.2 Hypothesis and Goals

The research work of this thesis is based on the following hypothesis and goals:

- **Hypothesis 1:** The structure of network can be used to determine strength of groups in terms of cardinality to other members. This can be efficiently computed using the degree-based centrality measure for large real-world networks.
  Goals: Determine the best polynomial-time approximation algorithm for group degree centralization by reducing it to a maximum domination problem. Develop an efficient method for calculating an approximate maximum for group degree centralization of a given graph for all group sizes. Run experimental testing on large real-world networks.

- **Hypothesis 2:** Temporal evolution of structural properties of networks can be used to explain the characteristics of individuals involved in the social networks. *Goals*: Analysis of structural properties taking into account core and periphery nodes and analysis of network cores using ego-networks approach. Development of an approach for analyzing discrete core/periphery time-series using machine learning methods to describe dynamics of individual nodes.

- **Hypothesis 3:** Using textual information assigned to nodes and structure of network, we can determine the level of diversity of individual nodes and group of nodes in network. *Goals:* Development of a new network diversity metric based on Stirling's index, using text similarity.

- **Hypothesis 4:** Methods for visualization of network structure and text can be used to conceptualize a social network community.
  Goals: Analysis of text and network structure to conceptualize a scientific community.

1.3 Scientific Contributions

In this thesis we claim the following major scientific contributions:

- Polynomial time group degree centralization approximative algorithm.

- Approach for characterizing dynamics of indivudial nodes based on core/periphery structural position using clustering and decision trees.

- New interdisciplinary metric for networks based on Stirling's diversity index and text analysis based similarity matrix.

- Approach for conceptualization of science using network of collaboration and competences of individuals.

1.4 Thesis Structure

This theses is composed of six chapters. The first chapter introduces the general idea of the thesis, hypotheses and goals, and the thesis structure. The following four chapters present the main research work of the thesis. In the second chapter we introduce an algorithm for efficient computation of important groups using the structure of connections between network nodes. The third chapter introduces an approach for characterizing nodes using the time evolving structure of networks. In the fourth chapter we combine the network structure and text to develop a method for calculating network diversity. Content of the fifth chapter is conceptualization of scientific community developed with methods for analyzing text and network structure. The final sixth chapter concludes the thesis and shows directions for future work.

Chapter 2

Using Network Structure for Determining Importance of Groups

Network structure reflects the nature of interactions between subjects, or elements that form the network. Given a network G, group degree centralization is a measure of how some set of nodes stands out from other subsets of the same size in the sense of cardinality of the union of neighborhoods. In simple words, group degree centralization is determined by the number of nodes that are neighbors to the nodes in the group and are not a part of that group. Finding a group that is the strongest in terms of degree centralization can be important in a number of different real-world scenarios. For example, for diffusing some information in a friendship network such as Facebook, maximal degree centralization can identify the group of nodes to which the information should be seeded in order to be efficiently distributed across the network. For a given integer k among all k-subsets of set of nodes $V(G)$, determining the maximum group degree centralization is $NP-hard$ problem. Here we develop a polynomial algorithm with the best-possible approximation ratio that calculates all centralizations for $1 \leq k \leq n$ and altogether runs in $O\left(m + \gamma(G) \cdot \Delta(G)\right) \leq O(n^2)$ time, where n is number of nodes and m is number of edges in the graph. This makes the calculation of group degree centralization feasible for large real-world networks. The constructed algorithm is tested on six real-world networks. In experiments we observe a property of unimodality of group degree centralization for parameter k. At the end, we provide some challenges for future work.

This work was conducted in collaboration with Matjaž Krnc and Riste Škrekovski and was submitted to the Journal of Applied Mathematics and Computer Science in September, 2015.

2.1 Related Work

A social network is typically represented as a graph, where individual persons or nodes are represented as vertices, and the relationships between pairs of individuals as edges. In most networks some edges or nodes are more central than others, where central can mean variety of different things like well connected to other nodes, closer to other nodes, or between many important paths. To quantify this intuitive feeling, centrality indices were introduced. First mathematical concept of centrality of graphs was introduced 146 years ago by Jordan (1869). There are many ways to provide a measure of the relative "importance" of a node in a network, thus different motivations lead to different centrality measures that were developed in several fields. Centrality measures are used in chemistry (Jeong, Mason, Barabási, & Oltvai, 2001), psychology (Ahn, Kim, Lassaline, & Dennis, 2000), sociology (Jackson et al., 2008), geography (Lorenzen & Andersen, 2009), game

theory (Gomez et al., 2003), and many other fields. Arguably, the most common branch of centrality measures is based on the distance between the nodes of the network. Some of the standard centrality indices from this branch are betweenness, closeness and eccentricity. Classic applications of those centralities can be found in transportation-network theory (Guimera, Mossa, Turtschi, & Amaral, 2005), communication-network theory (Onnela et al., 2007) and many others. In the last years, even more widespread interest is developing in the field of chemistry (Badar, Hite, & Badir, 2013), electrical circuits (Cotilla-Sanchez, Hines, Barrows, & Blumsack, 2012) and also in the study of food webs (Bauer, Jordán, & Podani, 2010).

We will freely interchange terms vertex/node and graph/network, without any meaningful difference. Various vertex-based measures of the centrality have been proposed to determine the relative importance of a vertex within the graph. Arguably, the most basic measure is the degree of a vertex, which we study in this research work. Among other measures of centrality, some better known in network analysis are: betweenness centrality, closeness centrality, eccentricity centrality, Google PageRank, eigenvector centrality, Katz centrality, Alpha centrality, and others. For detailed definitions and discussion on various centrality indices, we refer the reader to (Brandes & Erlebach, 2005; Bavelas, 1948, 1950; Koschützki et al., 2005; Proctor & Loomis, 1951; Seeley, 1949).

In his study, Freeman (1979) realized that despite all existing vertex-centrality indices, there was a need for *graph centrality* measure based on differences in point centrality. He defined a *centralization* measure that can be used in combination with any vertex-centrality to determine to what extent some vertex in network stands out from others in terms of given centrality measure. Furthermore he used this approach to compare different graphs, depending on their highest centralization scores. In the same article Freeman remarked that the centralizations for degree centrality, betweenness centrality and closeness centrality reach its maximum if and only if G is a star. The statement was later proved in detail by Everett, Sinclair, and Dankelmann (2004). In order to compare centralization values of graphs with different sizes, Freeman in the definition of centralization used a normalized formula, dividing the expression by the theoretically largest sum of differences in any graph from the given class of graphs (Freeman, 1979).

Everett and Borgatti (1999) introduced the concept of *group centrality* which enables researchers to answer questions such as "how central is the engineering department in the informal influence network of this company?" or "among middle managers in a given organization, which are more central, the men or the women?" With these measures we can also solve the inverse problem: given the network of ties among organization members, how can we form a team that is maximally central? In Everett and Borgatti (1999), the authors introduced group centrality for measures of degree, closeness and betweenness centrality, which we use in this work. In 2006, Borgatti introduced important group centrality measure (usually called KPP) that is motivated by *key players problem* (Borgatti, 2006). In his paper he focused on finding a set of vertices for the purpose of optimally diffusing something through the network by using selected vertices as seeds, or for maximally fragmenting the network by removing the key nodes. Interestingly, Borgatti claims that previously mentioned group closeness and betweenness are not proper tools to define KPP centrality. He therefore used tools like graph fragmentation and information entropy to define KPP centrality. Miyano and Ono (2011) study the algorithmic complexity of finding the best group for so-called *k-vertex maximum domination problem* (or k-maxVD, in short). Although they claim that k-maxVD is a new variant of vertex-domination problem, it is in fact the same as group degree centrality (Everett & Borgatti, 1999), summed with a constant k. Authors determined the best possible polynomial approximation of k-maxVD is a greedy one, which we use in our work.

Several more concepts of vertex centrality with respect to some subset of vertices have been introduced throughout the last decade. S. White and Smyth (2003) introduced a measure called *personalization* that show, how central an individual is according to given subset R (group of important people) in given social network. In 2005, *subgraph centrality* has been introduced by Estrada and Rodriguez-Velazquez (2005), and characterizes the participation of each node in all subgraphs in a network, which is calculated from the spectra of the adjacency matrix of the network. In the same year, Everett and Borgatti (2005b) introduced another measure (i.e. *core centrality*), where they evaluate the extent to which a network revolves around a core group of nodes. Finally, very recently Bell (2014) introduced the concept called *subgroup centrality*, where centrality (of one vertex) is calculated only on a restricted set of vertices. Let us remark that all four mentioned centralities in principle measure importance of an individual vertex (with respect to some conditions) and are different from group centrality, proposed by Everett and Borgatti (1999). Following Freeman's approach, group centralization notion was introduced in Krnc and Škrekovski (2015).

In this chapter, we study group centralization notion for degree centrality, i.e. group degree centralization. In Section 2.2 we provide notations and definitions from Krnc and Škrekovski (2015) that we use. In Section 2.3 we mathematically analyze group degree centralization problem and determine its time-complexity. We determine its best polynomial-time approximation algorithm by reducing it to maximum domibnation problem (Miyano & Ono, 2011) and develop an efficient method for calculating an approximate maximum for group degree centralization of given graph for all group sizes. In Section 2.4 we describe the developed algorithm for studying group degree centralization, as well as the optimal value of the group size. We also analyze the complexity of the algorithm and observe the *degree-unimodality* property, which may be a new property for studying real-world networks. In the conclusion, we list some open problems and provide challenges for future work.

2.2 Preliminaries

Graphs are mathematical structures often used to model networks. First we introduce some basic definitions related to graphs, after that we give preliminaries for our method.

A graph is an ordered pair $G = (V(G), E(G))$, comprising a set of nodes $V(G)$ and a set of edges $E(G)$. An edge e is a two element subsets of nodes v. We use simbol n to denote the number of nodes and m to denote the number of edges in a graph. A graph is undirected if the set of edges contains unordered pairs of nodes, while it is directed in opposite case. Simple graph is a graph that contains no self-edges (loops) and no multiple edges. In this work we refer to undirected simple graphs when we use the term graph. Node degree $deg(v)$ is the number of nodes adjacent to a node. We denote the set of neighbors of a node with $N(v)$. A cycle in a graph consists of a sequence of nodes starting and ending at the same node, with each two consecutive nodes in the sequence adjacent to each other in the graph. A tree is an undirected simple graph that is connected and has no cycles. A star is a tree with one internal node and $n-1$ leaves.

Let G be a graph on n vertices and m edges, and let $S \subseteq V(G)$. According to Everett and Borgatti (1999), group degree centrality is defined as

$$\mathrm{GD}(S,G) = \left| \bigcup_{v \in S} N(v) \setminus S \right|.$$

Where $N(v)$ denotes all the neighboring nodes of v. For a fixed k and set of all nodes of a graph $V(G)$, let S_k^* be one of the sets from $\binom{V(G)}{k}$ that achieves the maximum value of

group degree centrality, i.e. $\mathrm{GD}_G(S_k^*) = \max_{S \in \binom{V(G)}{k}} \mathrm{GD}(S, G)$. If graph G is clear from context, we will omit the index from group degree notion.

According to Freeman (1979), Krnc and Škrekovski (2015), $\mathrm{GD}_1(S, G)$ stands for group degree centralization. Define $k := |S|$, and observe

$$\mathrm{GD}_1(S, G) \quad = \quad \frac{\sum_{S' \in \binom{V(G)}{k}} \left(\mathrm{GD}_G(S) - \mathrm{GD}_G(S')\right)}{\max_{H \in \mathcal{G}_n} \sum_{S'' \in \binom{V(H)}{k}} \mathrm{GD}_H(S^*_{|S|}) - \mathrm{GD}(S'')}. \tag{2.1}$$

According to Freeman (1979), the denominator is needed to efficiently normalize centralization to interval $[0, 1]$, for better relative comparison. Clearly $\mathrm{GD}_1(S, G)$ is maximized whenever $\mathrm{GD}(S, G)$ is maximized, and from Krnc and Škrekovski (2015) we know that the maximum value of the denominator corresponds to star graph S_n, where the optimal set S_k^* is any set containing the center of the star.

Whenever graph G is known from the context, we omit it from the notions of centrality or centralization. We denote *maximizing group size* $\mathrm{dc}(G)$ to be the positive integer such that $S^*_{\mathrm{dc}(G)}$ we achieve the maximum value of group degree centralization, i.e. $\mathrm{GD}_1(S^*_{\mathrm{dc}(G)}, G) = \max_{k \in [n]} \mathrm{GD}_1(S_k^*, G)$, and also denote $S^* := S^*_{\mathrm{dc}(G)}$. Let $\gamma(G)$ be a cardinality of a minimum set that dominates graph G (also known as *domination number*). The notation $\Delta(G)$ stands for the highest degree of any vertex in a graph, i.e. $\Delta(G) = \max_{v \in V(G)} \deg_G(v)$. A function f is said to be *unimodal* if locally there is only a single highest value in f.

2.3 Evaluating Degree Centralization

The goal of this section is to optimize the procedure of calculating group degree centralization for a given graph and a constant k. We start by calculating the denominator for group degree centralization from Equation 2.1.

Proposition 1. *Let G be a star on n vertices with center c, and let $S^* \in \binom{V(G)}{k}$ such that $c \in S^*$. Then*

$$\sum_{S' \in \binom{V(G)}{k}} \left[\mathrm{GD}(S^*) - \mathrm{GD}(S')\right] = (k+1)\binom{n-1}{k}.$$

Proof. Let us partition the sets of $\binom{V(G)}{k}$ into two parts P_1 and P_2, depending on weather or not (respectively) they include the vertex c as a member. It is easy to see that the group degree centrality of members of these parts equals $n - k$ and 1, respectively.

Consider the number of k-sets of $V(G)$ that contain the vertex c, i.e. $|P_1|$. Since all such sets contain vertex c, we have yet to choose $k-1$ set members among the remaining $n-1$ vertices, therefore we have $|P_1| = \binom{n-1}{k-1}$. With the similar argument, we get $|P_2| = \binom{n-1}{k}$. Joining those facts, we get

$$\begin{aligned}
\sum_{S' \in \binom{V(G)}{k}} \left[\mathrm{GD}(S^*, G) - \mathrm{GD}(S', G)\right] &= \binom{n}{k}(n-k) - \binom{n-1}{k-1}(n-k) - \binom{n-1}{k} \\
&= (n-k) \cdot \binom{n-1}{k} - \binom{n-1}{k} \\
&= (n-k-1)\binom{n-1}{k} \\
&= (k+1) \cdot \binom{n-1}{k+1}.
\end{aligned}$$

□

In the following proposition we use a classic graph theoretical approach of double counting to show, that the sum $\sum_{S' \in \binom{V(G)}{k}} \mathrm{GD}(S', G)$ from (2.1) can be computed efficiently.

Proposition 2. *Let G be a graph on n vertices, and let $k \leq n$ be a positive integer. It holds that*

$$\sum_{S' \in \binom{V(G)}{k}} \mathrm{GD}(S', G) = n \cdot \binom{n-1}{k} - \sum_{v \in V(G)} \binom{n - \deg(v) - 1}{k}.$$

In particular, the sum $\sum_{S' \in \binom{V(G)}{k}} \mathrm{GC}(S', G)$ can be computed in $O(n)$ steps.

Proof. For each vertex $v \in V(G)$, define its contribution $g_k(v)$ to be the number of k-sets that dominates v, i.e.

$$g_k(v) \;=\; \left| \left\{ X \in \binom{V(G) - v}{k}; X \cap N(v) \neq \emptyset \right\} \right| = \binom{n-1}{k} - \binom{n - \deg(v) - 1}{k}.$$

It follows that

$$\sum_{S' \in \binom{V(G)}{k}} \mathrm{GD}(S', G) \;=\; \sum_{v \in V(G)} g_k(v) = n \cdot \binom{n-1}{k} - \sum_{v \in V(G)} \binom{n - \deg(v) - 1}{k}, \quad (2.2)$$

which can be computed in $O(n)$ steps, traversing all vertices once. □

We join results from Propositions 1 and 2 to further develop (2.1). We can therefore claim the following.

Theorem 2.1. *For a given graph G and a group of its vertices S of size k, group degree centralization can evaluate as*

$$\mathrm{GD}_1(S, G) = \frac{\binom{n}{k} \cdot \mathrm{GD}(S, G) + \sum_{v \in V(G)} \binom{n - \deg(v) - 1}{k} - n \cdot \binom{n-1}{k}}{(k+1) \cdot \binom{n-1}{k+1}}, \quad (2.3)$$

which can be computed in $O(n)$ steps.

It is easy to see that finding the best-possible group S^* that maximizes group degree centrality (and hence group degree centralization) is $\mathcal{NP}$-hard.

Proposition 3. *Given a graph G, it is $\mathcal{NP}$-hard to find S_k^*.*

Proof. We prove the claim by reducing the problem of finding S_k^* to the existence of k-dominating set. Let us assume that there exists a polynomial algorithm for finding a k-set $S_k^* \subseteq V(G)$ such that

$$\mathrm{GD}_G(S_k^*) = \max_{S \in \binom{V(G)}{k}} \mathrm{GD}_G(S).$$

Now observe that the existence of k-dominating set is equivalent to the property

$$\mathrm{GD}_G(S_k^*) = n - k.$$

As group degree centrality of a given fixed set S_k^* can be computed in at most linear time, it is clear that the set S_k^* provides us with an answer of the existence of k-dominating set, which is a well-known $\mathcal{NP}$-hard problem. □

In Section 2.4 we present an efficient algorithm that achieves the best possible linear-time approximation for calculating group degree centrality scores for all group sizes.

2.4 The Algorithm

In this section we present the methods that can be used to calculate the estimate of group degree centralization on some big real-world networks. Also an algorithm for finding the group with maximal degree centralization is with its complexity analysis is provided. The algorithm is implemented in C++[1] and Python[2] with the help of the libraries "SNAP"[3] and "SNAP.py"[4], respectively. We run experiments on six real-world networks and discuss an interesting unimodal property of centralization for the real-world networks from the experiments.

Finding the group with the biggest degree centralization can be useful for many real-world networks which can in some cases be very large (largest network from the experiments contains 16 million edges). To calculate group degree centralization, we reduce it to *k-VertexMaximumDomination* (*k-MaxVD*, for short), which is a special case of *MaximumCoverageProblem*, that for a positive integer k finds a subset of size k that maximizes the cardinality of dominated vertices. By Proposition 3, group degree centralization cannot be calculated in polynomial time, unless $\mathcal{P} = \mathcal{NP}$. To calculate it efficiently, we use greedy algorithm for k-MaxVD from (Miyano & Ono, 2011) which is a polynomial time $(1 - 1/\epsilon)$-approximation algorithm. Furthermore, in (Miyano & Ono, 2011) it is proven that no other polynomial time constant factor approximation algorithm for k-MaxVD can have an approximation ratio better than $(1 - 1/\epsilon)$. An implementation of the procedure that calculates an approximation for the group with biggest degree centralization, for all meaningful group sizes, is given by Algorithm 2.1. Let us now describe the parts of the algorithm.

We use k-MaxVD greedy approach to efficiently find groups S, starting with $k = 1$ and increasing the group size until the graph is dominated. Every time that k increases, we add some greedily chosen vertex to the set S and remove it from G while maintaining some dictionaries that we use (*contribution* and *histogram*, in particular). In the first phase of Algorithm 2.1 (throughout lines 1–6), we pre-process our graph G and initialize the starting values of all dictionaries and other variables (conversion to digraph is assumed implicitly). In the *main loop* (lines 7–22) we first chose the vertex v to add to our group S, and accordingly update variables S and k (lines 8–10), and then calculate the group degree centrality for increased set S (in line 11). Then, throughout lines 11–19 we update the dictionaries *contribution* and *histogram*, and also remove v and some additional directed edges from the graph. Finally, in lines 21 and 22, we update some centralization variables and calculate the Freeman centralization of the centrality from line 11. We now present some details of how we maintain some values and graph properties.

For each group we efficiently calculate group degree centralization using Propositions 1 and 2. While the directed graph G that we work with is changing with each iteration, let G' be an original undirected input graph, let S be the group of vertices that we are calculating group degree centralization from, and set $k := |S|$. Note that in the algorithm treats G as a directed graph, initially with $\deg^+_G(v) = \deg^-_G(v) = \deg_{G'}(v)$ for all $v \in V(G')$. Before the beginning of each iteration of the *main loop*, the existence of a directed edge uv means, that

- in a initial graph G', we had $uv \in E(G')$,
- neither of v, u is a member of S, and

[1] https://github.com/mkarlovc/gcentralization
[2] https://github.com/mkrnc/group_degree_centralization
[3] https://snap.stanford.edu/snap/
[4] https://snap.stanford.edu/snappy/

Algorithm 2.1: Finding group with maximal degree centralization

Data: a graph G.
Result: a list *centralization* of group degree centralization scores, where *centralization*$[i]$ is an approximation of $\max_{S\in\binom{V(G)}{i}} \mathrm{GD}(S,G)$.

```
1  n ← |V(G)|, k ← 0, S ← ∅;
2  dominated ← ∅, histogram ← ∅;
3  G ← {a directed instance of graph G'};
4  for v in V(G') do
5  |  add v to histogram[deg_{G'}(v)];
6  |  dominated[v] ← False;
   end
7  while max(histogram) ≥ 0 do
8  |  v ← any vertex from histogram with the highest contribution;
9  |  S ← S ∪ v;
10 |  k ← k + 1;
11 |  centrality ← centrality+CONTRIBUTION(v);
12 |  for u ∈ N_G^+(v) do
13 |  |  DECREASECONTRIBUTION(u);
   |  end
14 |  for u ∈ N_G^-(v) do
15 |  |  DECREASECONTRIBUTION(u);
16 |  |  dominated[u] ← True;
17 |  |  for w ∈ N_G^+(u) do
18 |  |  |  DECREASECONTRIBUTION(w);
19 |  |  |  E(G) ← E(G) − uw;
   |  |  end
   |  end
20 |  G ← G − v;
21 |  UPDATECENTRALIZATIONVARIABLES();
22 |  centralization[k] ← A · centrality + B − C;
   end
```

- in a initial graph G', vertex v is not connected with any vertex from S, i.e. $v \notin \cup_{v\in S} N(v)$.

For any vertex $v \in V(G) \setminus S$, we define the *contribution* of vertex v to be the value $\mathrm{GD}(S\cup\{v\},G) - \mathrm{GD}(S,G)$ and observe that

$$\mathrm{GD}(S\cup\{v\},G) - \mathrm{GD}(S,G) = \begin{cases} \deg_G^-(v) & \text{if } v \text{ is not dominated,} \\ \deg_G^-(v) - 1 & \text{otherwise.} \end{cases}$$

The calculation of the value contribution is done by a short function *contribution* (see Algorithm 2.3). As different nodes have various contributions we define a dictionary *histogram*, initialized in line 5 of Algorithm 2.1, where keys are all possible values of contribution (for any key i, it clearly holds $-1 \le i \le \Delta(G)$), while the values are unordered sets of nodes with a given contribution. While dictionary *histogram* is initially indeed a degree-histogram, with each modification of graph G and set S we carefully update it. The goal of algorithm is, for each S, to calculate the value of (2.3). We implement this by introducing variables

$A, B, C, centrality$, each assigned for different part of the expression (2.3), i.e.

$$\mathrm{GD}_1(S,G) = \frac{\binom{n}{k}\cdot \mathrm{GD}(S,G) + \sum_{v\in V(G)}\binom{n-\deg(v)-1}{k} - n\cdot\binom{n-1}{k}}{(k+1)\cdot\binom{n-1}{k+1}},$$

which may, for algorithmic purposes, be written as $A \cdot centrality + B - C$, with

$$\begin{aligned} A &= \frac{n}{(n-k)\,(n-k-1)}, \\ centrality &= \mathrm{GD}(S,G), \\ B &= \frac{(n-k-2)!}{(n-1)!}\sum_{v\in V(G)}\frac{(n-\deg(v)-1)!}{(n-k-\deg(v)-1)!}, \\ C &= \frac{n}{n-k-1}. \end{aligned}$$

Clearly, whenever the group S increases, also all values of $A, B, C, centrality$ change. To handle variable B we also introduce dictionaries, $degDistribution$ and sum, where keys of both are all degrees of vertices in G, and the values are defined as

$$\begin{aligned} sum[i] &= \frac{(n-i-1)!}{(n-k-i-1)!}\cdot\frac{(n-k-2)!}{(n-1)!}, \\ degDistribution[i] &= \text{number of vertices in } G \text{ of degree } i. \end{aligned}$$

Algorithm 2.2: Updating variables A, B, C and sum.

```
   function updateCentralizationVariables()
1  |  A ← n/((n−k)(n−k−1)) ;
2  |  C ← n/(n−k−1);
3  |  for i in sum do
4  |  |  sum[i] ← sum[i] · (N−i−k−1)/(N−k−2)
   |  end
5  |  B ← Σ_i sum[i] · degDistribution[i]
```

Note that sum needs to be refreshed with every change of k. We first initialize both dictionaries before entering the main loop, and then we maintain their values by use of a function $updateCentralizationVariables$ (we treat these variables as global variables, therefore no parameters are needed). To avoid using big numbers, we update the value of $sum[i]$ by just multiplying it by $\frac{(n-k-i-1)}{(n-k-2)}$, whenever k increases by one. Using this, B can be calculated by a simple addition

$$B = \sum_i sum[i]\cdot degDistribution[i],$$

see line 5 of Algorithm 2.2.

Note that while the graph is changing and the group S is increasing, also contributions of the remaining vertices change. While initial contribution for some vertex v is equal its degree $\deg_G(v)$, during the main loop the contribution of v may incrementally decrease by one several times. We handle these changes by defining a function $decreaseContribution(v)$, see Algorithm 2.3.

Algorithm 2.3: Functions *contribution* and *decreaseContribution*. The former outputs the contribution of v while the latter refreshes the dictionary *histogram* whenever the contribution of node v decreases by one.

```
   function contribution()
       if v is dominated then
1          return deg⁻_G(v) − 1
       else
2          return deg⁻_G(v)
       end
   function decreaseContribution()
3      c ← CONTRIBUTION(v)
4      histogram[c] ← histogram[c] \ {v}
5      histogram[c − 1] ← histogram[c − 1] ∪ {v}
```

2.5 Experiments

2.5.1 Complexity analysis

In this section we analyze the running time complexity of Algorithm 2.1 in terms of n, m, $\Delta(G)$ and $\gamma'(G) := \frac{\epsilon}{\epsilon-1}\gamma(G)$. First notice that functions from Algorithms 2.2 and 2.3 take at most $O(\Delta(G)) \leq O(n)$ and $O(1)$, respectively. Using the greedy approach from (Miyano & Ono, 2011), in each step of the algorithm a node with biggest contribution to the group degree centrality is added to the group. We implement this by maintaining a sorted histogram of node contributions. Initially, the contributions to group centrality are equal to node degrees. After each addition of a new node to the group, the histogram is accordingly updated. Lets name the while loop from the algorithm 2.1 (line 7 to 22) the *MAIN LOOP*. The *MAIN LOOP* of the algorithm terminates when there is no node with contribution greater than -1, which occurs when all the nodes are dominated by the group. Note that the the number of iterations of the *MAIN LOOP* is bounded to $\gamma'(G)$.

It is easy to see, that throughout lines 1–10, our algorithm needs up to $O(n)$ steps to initialize all the needed variables for the *MAIN LOOP* to start. By use of the dictionary *histogram*, it takes constant time to greedily pick the vertex v with the biggest contribution, therefore line 12 altogether takes at most $O(n)$ time. Further notice that also lines 13–15 and 25 take constant time, altogether using up to additional $O(n)$ steps. Consider now the lines 16–23 and notice, that a function *decreaseContribution* is called for each removed directed edge precisely once (here we also consider edges removed in line 23. Furthermore, in line 22, deleting a vertex v from a graph means removing a vertex ($O(1)$) and removing its adjacent edges ($O(\deg(v))$). Since there are m edges in G, the lines 16–23 are altogether bounded to $O(m)$. Finally, line 24 takes constant time for updating variables A and C, while both *sum* and B takes up to $O(\Delta(G))$ time to compute, altogether summing up to $O(\gamma'(G) \cdot \Delta(G))$. Summing everything, the final complexity of finding the approximation of group degree centralization for groups of sizes $1 \leq k \leq n$ in Algorithm 2.1 equals $O(m + \gamma(G) \cdot \Delta(G)) \leq O(n^2)$. For most scale-free networks this can probably be reduced further to $O(n \log n)$. From (Miyano & Ono, 2011) we know that from approximation ratio point of view, Algorithm 2.1 is the best possible polynomial approximation for group degree centralization.

2.5.2 Datasets

Here we describe the datasets used for testing the degree centralization algorithm. We used six real-world networks ranging from several hundred thousands up to more than 16 million edges.

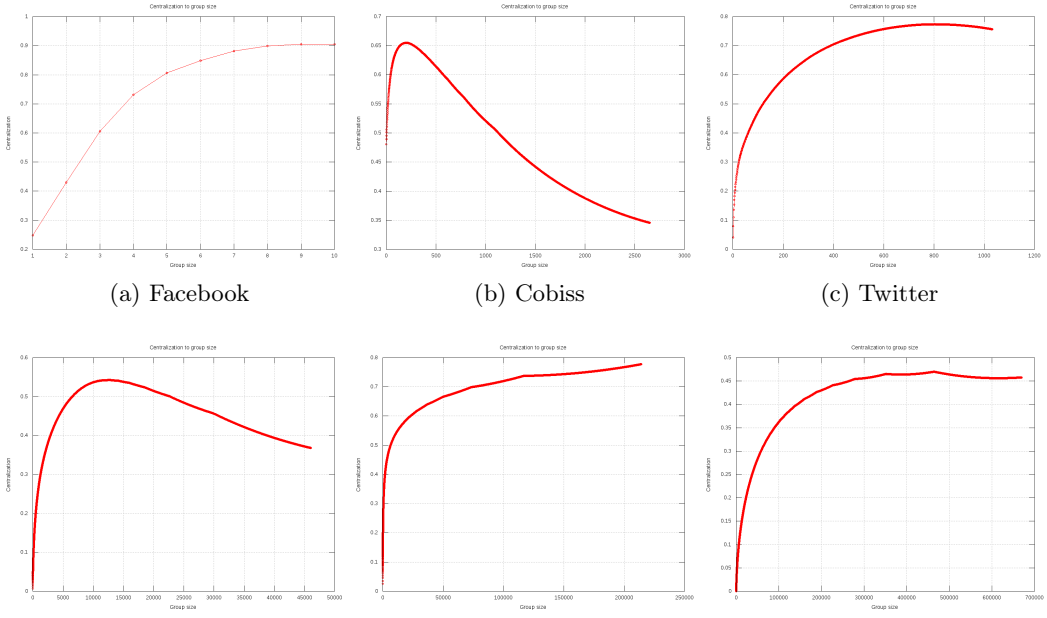
Facebook [5] is the smallest dataset in experiments containing 4039 nodes and 88234 edges (Leskovec & Krevl, 2014). It is anonymized data collected by survey participants using a Facebook application with combined ten ego-networks. The dataset was generated for study of social circles in ego-networks (Leskovec & Mcauley, 2012) and it is publicly available. Cobiss dataset is a graph of scientific co-authoring of the complete national research database in Slovenia from 1970 to 2013. Two authors are connected if they publish at least one paper together. The graph contains 25,301 nodes and 316,587 edges. The dataset was generated using database maintained by ARRS (Slovenian Research Agency) and IZUM (Institute of Information Science, Maribor, Slovenia). The data for generating this graph was obtained as part of the Slovenian Science Atlas and Co-authorship Networks of Slovenian Scholars. Twitter [5] dataset contains graph of followers, with 81,306 nodes and 1,768,149 edges (Leskovec & Krevl, 2014). The dataset was collected from public sources for study of social circles discovering study (Leskovec & Mcauley, 2012) and is publicly available. Amazon [5] dataset is a graph of frequently co-purchased products based on Amazon website in June, 2003 (Leskovec & Krevl, 2014). The graph has 403,394 nodes and 3,387,388 edges and was generated for the study of viral marketing dynamics (Leskovec, Adamic, & Huberman, 2007). The dataset is publicly available. Youtube [5] dataset contains a graph of user friendship graph (Leskovec & Krevl, 2014). The graph contains 1,134,890 nodes and 2,987,624 edges. The dataset was prepared by Mislove, Marcon, Gummadi, Druschel, and Bhattacharjee (2007). The dataset is publicly available. Patents [5] dataset is a citation graph of patents granted between 1975 and 1999 (Leskovec & Krevl, 2014). The graph contains 3,923,922 nodes (patents) and 16,522,438 edges (citations). The dataset was generated for the purpose of graph evolution study (Leskovec, Kleinberg, & Faloutsos, 2005), using the U.S. patent dataset maintained by the National Bureau of Economic Research (Hall, Jaffe, & Trajtenberg, 2001). The dataset is publicly available.

2.5.3 Experimental results

Figure 2.1 shows the value of centralization with different size of the group. In Cobiss (Figure 2.1b) network the maximal centralization is obtained with relatively small group size, and after that point adding members to the group causes drastic decreasing of centralization. Amazon (Figure 2.1d) has a similar shape, but increasing and decreasing of centralization is less intensive. In Youtube network (Figure 2.1e), relatively small group size has a high degree centralization. Further increasing of group size slowly increases centralization, which is maximal only when the group dominates all the nodes. In Facebook network (Figure 2.1a) the maximal centralization is also achieved when the group dominates all the nodes. Patents network (Figure 2.1f) is similar to Youtube, but the maximal centralization is achieved before the nodes are dominated by the group. In Twitter (Figure 2.1c), network centralization increases up to maximal point and then decreases with the same intensity.

Centralization is comparable between different networks. We can determine that the largest centralization from the networks in the experiments is obtained on the Facebook data with values of 0.9, when the group is of size 10. This is an expected result because the Facebook network was generated by combining 10 ego-networks and the centers of the ego networks are the group members identified by our algorithm. Youtube and Twitter

[5] https://snap.stanford.edu/data/



(a) Facebook (b) Cobiss (c) Twitter

(d) Amazon (e) Youtube (f) Patents

Figure 2.1: The graphic representation of Freeman centralization of group degree centrality of six networks with different sizes of groups.

have a relativity similar and high maximal centrality, which is obtained with 214,003 and 803 group members respectively. Cobiss and Amazon graphs are in the middle range of our centralization experiments, while the lowest maximal centrality has Patents network, which is also the largest and the sparsest network in our experiments.

Looking at the centralization values for all networks in the experiments, with thousands of different sizes of groups, we observed an interesting property that we believe is worth further discussion. Within a small error bound, the shape of most of plots in Figure 2.1 is unimodal, i.e., it is monotonically increasing up to a maximizing group size, while at value $\mathrm{dc}(G)$ the slope starts to have negative slope. We do not believe that such property would hold on general connected graphs. However, results of our experiments indicate that this is a property that could hold for many real-world networks.

Table 2.1: Some statistics and centrality results from our experiments.

G	$\lvert V(G)\rvert$	$\lvert E(G)\rvert$	$\mathrm{dc}(G)$	$\mathrm{GD}_1(S^*_{\mathrm{dc}(G)}, G)$	$\mathrm{GD}(S^*_{\mathrm{dc}(G)}, G)$
facebook	4,039	88,234	10	0.905393	4029
cobiss	25,301	316,587	204	0.654997	19635
twitter	81,306	1,768,149	803	0.773615	78,811
amazon	403,394	3,387,388	12,810	0.542647	320,133
youtube	1,134,890	2,987,624	214003	0.777639	920,887
patents	3,923,922	16,522,438	464,298	0.470009	3,105,485

2.6 Conclusion and Future Work

Regarding group degree centralization, various results can be found in this chapter. We developed some mathematical optimizations and also used known results of Miyano and Ono (2011) to develop a novel algorithm for estimating group degree centralization for all group sizes. In Section 2.4 we describe the developed algorithm and analyze its time complexity. We present some results on six real-world networks with various number of nodes, ranging from a few thousands to a few millions and observe an interesting property of the shape of plots in Figure 2.1, which we call the *degree-unimodality* property. We believe that this property may be a new property for studying real-world networks, in particular we believe that most of scale-free networks have the unimodality property, hence the following conjecture.

Conjecture 4. *Let G be any network on n nodes, generated by Barabassi-Albert model, and let f be a function, such that*

$$f(i) = \max_{S \in \binom{V(G)}{i}} \mathrm{GD}_1(S, G).$$

Then f is unimodal WHP (in terms of n).

It might be also interesting to investigate the parameters for which the same is true in Watts and Strogatz model, or in general $\mathcal{G}_{n,p}$ random graph model. Finally, one may study if the unimodality property is present in the real-world networks also for the group centralization of some other centrality indices, such as betweenness, closeness or eccentricity.

Chapter 3

Analyzing Core-Periphery Structural Property Dynamics

Interactions between people and many other real-world systems are dynamic. Therefore, it is important to observe the structure of networks at different points in time, in order to capture the temporal component of the observed system. In this work, we take into account both temporal and structural network properties, by analyzing a temporal dynamic of core-periphery structure in bibliographic networks. Bibliographic networks recorded over time capture publication activity and co-authorship of the actors. This means that the number of actors is typically growing over time and the social network capturing their interconnections is expected to become denser. Each year may bring a change by introducing some new actors in the network and dropping some from the previous year. As we aim at observing evolution of the network of researchers and investigate development of their career paths, we have decided to aggregate the data over years, so that each network contains all the data until the year labeling the network. For instance, as our data starts in 1970, the network for 2000 contains a union of all the data of networks from 1970 through 2000 (the aggregate for 31 years of publications).

By studying the network structure we can observe that some nodes are in the network core while others are in the periphery. Several algorithms have been proposed for calculating the core of a network. We have applied the discrete core-periphery partitioning algorithm (Lip, 2011) resulting in about 480 (out of 18 000) researchers being in the core in the final year.

The original contribution of this research work is in proposing an approach for using the network structure as captured by the core-periphery analysis to study evolution of the social network over time using machine learning techniques and studying various parameters of ego-networks. Information about a node being in a core of the network in a particular instance of time is used in generation of feature vectors describing the network nodes. As we are focusing on bibliographic networks and study of the career development of researchers over a span of over 40 years, feature vectors in our case describe individual researchers. We are proposing about 20 features to be used to describe a researcher and we have grouped the features in three groups: stability, balance, strength. For instance, stability features capture the number of transitions from periphery to core.

In our experiments, we have tested the following hypotheses:

- Researchers in the core are spread over the collaboration network;
- Ego-networks of core researchers have outstanding properties compared to the other ego-networks;
- Researchers exhibit different patterns of core-periphery position dynamics;

- There is a big overlap of the top researchers based on the publications rating and the researchers in core.

The rest of this chapter is organized as follows. Section 3.1 presents related work. Description of collaboration networks and structural properties taking into account core nodes is given in section 3.2. Core analysis using ego-networks is given in section 3.3, while section 3.4 gives discrete core-periphery time-series using machine learning methods. The proposed approach is presented in Section 3.2. In section 3.5 we discuss the results on few individuals from the dataset. Section 3.6 gives conclusions and some ideas for future work.

This research work was conducted in collaboration with Borut Lužar and Dunja Mladenić and was submitted to the Journal of Scientometrics in a paper titled "Core-Periphery Analysis of Collaboration Networks from Different Aspects".

3.1 Related Work

Labeling data points as belonging to core or periphery has been addressed in different contexts. Snyder and Kick (1979) examined data on international exchange and used the Wallerstein's model (Wallerstein, 1974) of a tripartite division of countries into core, periphery, and semi-periphery. Nemeth and Smith (1985) used the commodity exchange data and partitioned the world economy to core, strong semi-periphery, weak semi-periphery and periphery. Smith and D. R. White (1992) used the same data to analyze the structure of international economic exchange at multiple points in time. Alba and Moore (1978) developed a method for locating denser, or more cohesive parts of networks by identifying cliques which were aggregated when they were sufficiently overlapped. The method was applied to an interaction network of nearly 900 individuals, revealing aggregated subgraphs with characteristics of social circles. Doreian and Fararo (1985) analyzed journal-to-journal citation network to test position/role hypothesis stating that journals catering to the same or similar audiences, with the same type of content, have the same role in the network. They analyze 21 psychology journals in 1950 and 1960 and show that the position/role hypothesis is true. Mcauley and Leskovec (2014) studied the problem of identifying social circles within ego-networks with interpretable model that uses both the structure of the network and user profile information.

In this work we are analyzing the scientific collaboration network in Slovenia. Similar data was analyzed in Jovanovic, John, and Reschke (2010), where authors investigated cooperation in producing scientific output in successor republics of Yugoslavia. Collaboration structures in Slovenian scientific communities were analyzed by Kronegger, Mali, Ferligoj, and Doreian (2012), where co-authorship networks of four research disciplines (biotechnology, mathematics, physics and sociology) from 1986 to 2005 were compared and different co-authoring cultures were identified. Similarly, in Mali, Kronegger, and Ferligoj (2010) authors showed that the tendency to co-author in sociological community is increasing and the quality of publications in relation to structural position in the network was compared. Recently, two studies of interdisciplinarity in Slovenian science using collaboration networks have also been introduced (Karlovčec & Mladenić, 2015; Lužar, Levnajić, Povh, & Perc, 2014).

The concept of core–periphery is usually divided into discrete and continuous model. In the discrete model nodes belong either to the core which has a high density of ties, or to the periphery which has a low density of ties. In continuous model nodes obtain a degree which describes suitability of a node for core. In the rest of the chapter we consider the discrete model. Borgatti and Everett (1999) formalize intuitive notion of a core–periphery structure and recommend the model that maximizes connectivity between in-the-core nodes

and minimizes connectivity between the periphery nodes, while it ignores the connectivity between the nodes of the two partitions. They used the recommended model and developed the algorithm for detecting core–periphery structure using combinatorial optimization technique based on a genetic algorithm (Goldberg, 1989). This algorithm is slow and it can resolute to local maxima not finding an optimal partition. The algorithm was implemented in the UCINET (Borgatti, Everett, & Freeman, 2002), which is one of the most commonly used tools for social network analysis. Boyd, Fitzgerald, and Beck (2006) used the same approach of pruning the search space and developed an algorithm based on simulated annealing (Kirkpatrick, Gelatt, & Vecchi, 1983), and the Kernighan-Lin algorithm (Kernighan & Lin, 1970). Brusco (2011) developed an exact algorithm based on the branch-and-bound technique, but this algorithm is feasible only for very small networks with up to about 60 actors. Recently, Lip (2011) analyzed the most common definition of the core–periphery structure and derived that the core can in fact be determined in linear time. His algorithm is based on greedy approach and solves the problem exactly and efficiently for both symmetric and asymmetric networks. Since it is suitable for large networks and gives an optimal solution, we implemented Lip's algorithm and used it in our analysis. Using the algorithm we obtained core–periphery partitions with about 340 (out of 18 000) researchers being in the core in the final year.

3.2 Core in Collaboration Networks

Our analysis is conducted on scientific collaboration networks in Slovenia with links between two researchers whenever they co-authored a scientific publication; namely, we assign a weight of the number of publications coauthored by the two researchers to every link. The data is publicly available on SICRIS [1] and COBISS [2] web pages. We focused on original scientific articles, review articles, published scientific conference contributions, independent scientific component parts or chapters in monographs, and scientific monographs. We use data of about 18 000 researchers registered in Slovenian national scientific database (SiCRIS) in a span of 44 years (1970–2013). Note that we ignore collaboration with foreign scientists, which results in a lower level of collaboration measured for the scientists mainly performing research in narrow science fields.

Since temporal component of the data is crucial in our analysis, we constructed a collaboration network for each year in the period. After conducting initial experiments on both—separate networks with only the nodes that collaborate in the specific year and aggregated networks containing information from all the previous years—we decided to use the aggregated version of the networks. This is a standard approach for analyzing network evolution Barabási et al. (2002), Leskovec, Backstrom, Kumar, and Tomkins (2008, 2007) and proved to be good for tracking career of researchers. In Figure 3.1, evolution of the collaboration networks by decades is presented.

By the definition of discrete core (Lip, 2011), the only condition a node must fulfill in order to become a core member is to have big enough *strength*, i.e. the sum of weights on its incident edges, compared to the other nodes. Usually core members are the nodes having many collaborations with different nodes. This suggests that the researchers in the core of the network have more scientific collaboration and networking. Although core membership implicitly implies a significant number of publications, one should bear in mind that the core is not a measure of scientific productivity. For example, an extremely productive researcher is considered a peripheral node if he mostly publishes as a single author, or in collaboration with just a few other colleagues. On the other hand, a researcher with an

[1] http://sicris.izum.si/
[2] http://www.cobiss.si/

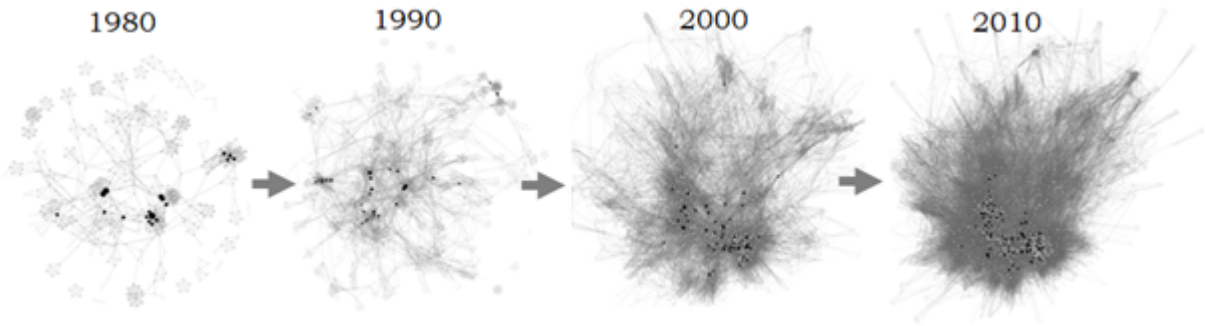


Figure 3.1: Collaboration networks in 1980, 1990, 2000, and 2010. It is evident that in the period of ten years the number of nodes and edges strongly increases. Nodes contained in the cores of the networks are marked.

average number of publications, which are coauthored by many authors, may enter the core in our model. In Subsection 3.3.1, we detect a researcher with a large number of coauthors but small number of publications, who is not a part of the core.

From the definition, it clearly follows that the number of nodes in the core grows slowly compared to the growth of entire network. In Figure 3.2, a chart confirming this fact is presented.

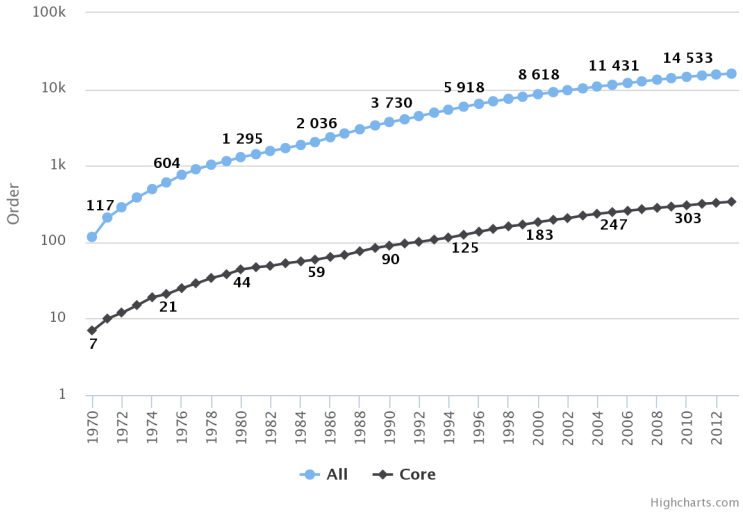


Figure 3.2: The growth of the order of cores and the whole networks in logarithmic scale.

In Figure 3.3, we compare the representation of science fields in the cores and in entire networks. In natural sciences the graphs are similar, while there is only a small number or no researchers in the cores from social sciences and humanities, although the number of researchers in those fields is relatively big. This shows that the co-authorship networking and publication activity is smaller in the social fields. On the other hand, it is evident that natural sciences are losing the leadership against engineering and technologies, and the growth of social sciences is showing in the last past years.

Obviously, the networks we study are not entirely connected. There are several small components, i.e. researchers that collaborate only with each other, but in the last years

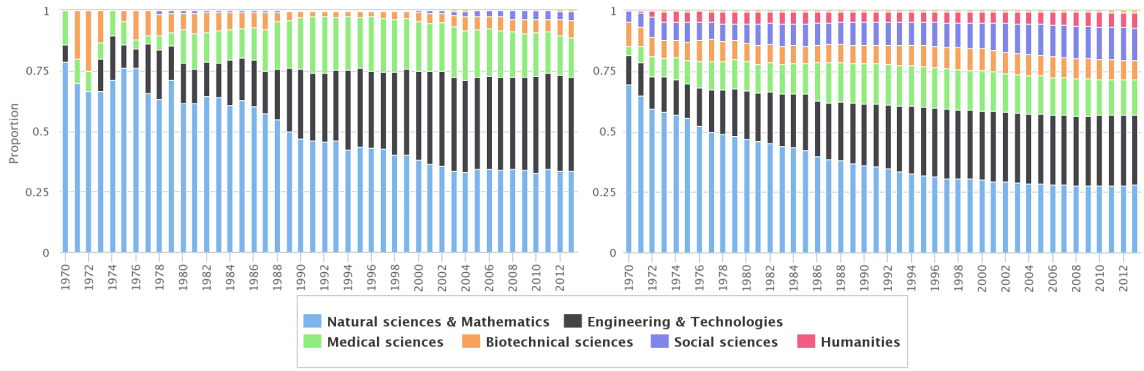


Figure 3.3: Distribution of researchers within seven science fields for the core researchers (left) and overall distribution (right).

this phenomenon is becoming more and more rare. It can be seen in Figure 3.4 that in the early years of the analysis, when the number of researchers was considerably smaller, the collaboration was at an extremely low level. There were many components in the networks and their number started to decrease in the 1990's. In 2013 there are still around 100 connected components in the network but these are mainly the artifacts from the earlier years, representatives of which already finished their scientific carrier. As our networks are aggregate, such components remained. The size of the giant component, however, has grown all the time and it comprises more than 99 percent of nodes in 2013. It is not surprising that already from 1978, all the core nodes were present in the giant component.

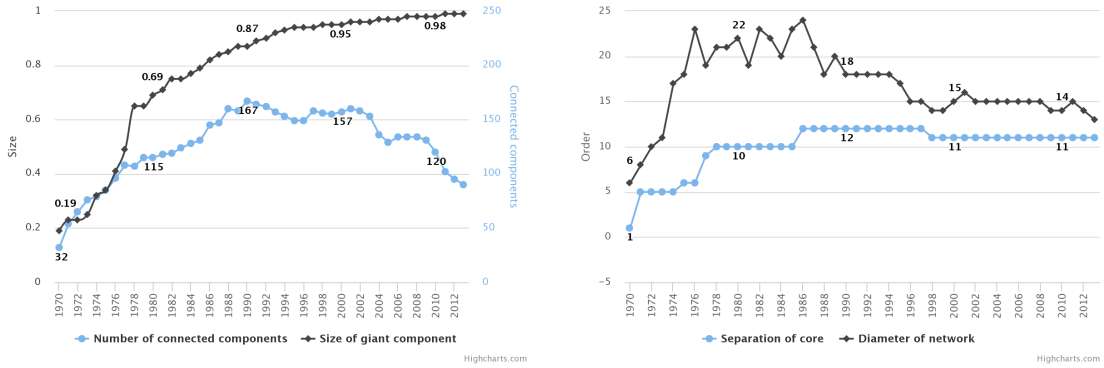


Figure 3.4: The number of connected components together with the proportional size of the largest component in terms of the number of vertices (left) and the diameter and maximum separation distance between core vertices during the years (right).

In the study of complex networks a considerable attention was given to their diameter, i.e. how far from each other the nodes can appear in the network (it is customary that when considering this parameter, only the greatest connected component is taken into account). We compare the diameter of the network to the maximum distance between two core nodes (separation) in Figure 3.4. Interestingly, the separation of the core nodes is very high, which shows that there are researchers in the core having no similar scientific interest at all (if it can be measured by the distance between the nodes). This confirms

the hypothesis that the core is in fact a backbone of the network and is not concentrated in just one part of the network.

3.3 Core-Periphery Analysis Through Ego-Networks

With datasets growing extremely fast in the past years, it is usually computationally too expensive to analyze and compute the values of global invariants of the entire networks with deterministic methods. One of the common approaches to tackle this issue is simplification by analyzing smaller networks constructed around every single node of the entire network, called *ego-networks*. This concept was first used in analysis of social networks by Bott (1957) and has received considerable attention since then.

Ego-networks may be defined in different ways, e.g. by considering connections between the neighbors of the central node or not, considering also neighbors of neighbors, etc. cf. (Hanneman & Riddle, 2005). We will focus on the most common version, where with an ego-network of a node v we mean a network comprised of v (the *central node*), all his direct neighbors and the edges between these nodes.

Several attempts to derive information about the structure of entire network from ego-networks have been reported (see e.g. Britton and Trapman (2013), Frank and Snijders (1994)), but in general it turns out that for the majority of properties only a small portion of conclusions may be deduced. Among other things, we would like to understand what kind of nodes form the core in extended, continuous meaning, which is much harder to compute compared to the discrete definition. Our analysis exposes several features that may be appropriate for such a partition.

In this section, we compute a number of network descriptors to compare their average values among all the ego-networks against the ego-networks with the central nodes taken from the cores (*core-ego-networks*), in order to expose some nontrivial differences between the core nodes and the nodes of entire networks. We compute average values for every year between 1970 and 2013 and compare the values over the years. Note that we do not exclude the core nodes when computing average values of parameters for entire networks.

3.3.1 Order, size and density

First, we compare three basic network descriptors; the order, the size and the density, i.e. respectively the number of nodes, the number of edges and the ratio between the size and the number of all possible edges in the network. As expected, on average the order and size of core-ego-networks grow much faster than the average order and size (see the charts in Figure 3.5). On one side, the average order of ego-networks is small due to a constant appearance of new researchers whose ego-networks are small for at least the first several years. On the other side, usually only the researchers with large ego-networks enter the core. It can be seen that the growth rate of ego-network sizes is about the same as the growth of the order, hence linear.

The three maximum orders of ego-networks are 198, 194 and 178 in 2013. Interestingly, there is a researcher, who is not in the core, with the order of his ego-network 159, which puts him to the eighth place in the maximum order ranking. Hence, the presence of vertices in the core does not rely entirely on the order of their ego-networks. However, the difference between distributions of the orders is in general clear (see the right chart in Figure 3.5).

Collaboration networks are not very dense in general and the density decreases by every additional node added to the ego-network. A researcher with a small number of coauthors is usually able to initiate a collaboration between his coworkers or it is likely that they have worked together before. On the other hand, core nodes have a high number

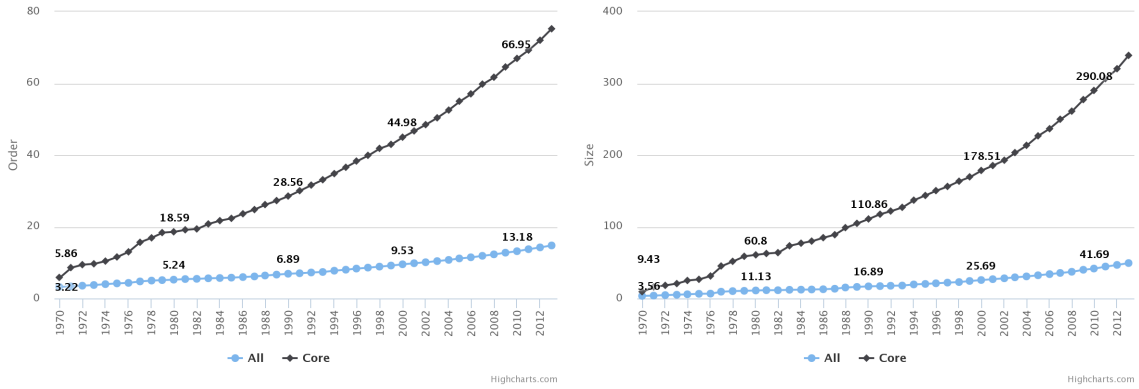


Figure 3.5: Average order (left) and size (right) of ego-networks for core vertices and overall.

of coworkers and since the density of ego-networks is relatively small, one might expect that these coworkers are weakly connected. The average density of ego-networks is therefore decreasing with the growth of ego-networks' order. The distribution of values in 2013 is close to uniform for entire networks, while the majority of values in core is below 20 percent.

3.3.2 Selectivity and disparity

The networks we consider in this work have weighted edges. Intensity of collaboration between two researchers is measured by the number of publications they coauthored. A measure that captures the proportion between the node degree and its strength is called *selectivity*, as defined in Masucci and Rodgers (2009). Selectivity hence measures the ratio between different co-authors and the sum of publications published with each of the coauthors. In Figure 3.6, it is shown that the average value of selectivity is slowly decreasing in entire networks as well as in the cores, which means that the number of publications published with each of the coauthors is increasing on average. This behavior is in a way not expected. As we have shown in the previous subsection, the order of ego-networks increases fast during the years, so the publication activity of each particular researcher is high. Surprisingly, the core nodes are still able to maintain the ratio. Again, there is an obvious distinction between the values of core-ego-networks and the ego-networks of all nodes.

A similar measure, *disparity*, is defined as the sum of squares of degree-strength ratios for every neighbor (Barthelemy, Gondran, & Guichard, 2003) (see also (Serrano, Boguñá, & Pastor-Satorras, 2006)). It was introduced in order to detect the distribution of degree-strength ratios among the neighbors of a node. In Figure 3.6, a distinction between core nodes and overall is again notable, but in general one cannot obtain any additional information complementary to the information given by the selectivity.

3.3.3 Ego-betweenness

In order to identify the most important nodes in a network various centrality measures are used. *Betweenness centrality* of a node v counts the number of shortest paths traversing v. When the central node of an ego-network is considered, we speak of the *ego-betweenness* of a node. Freeman (1982) was the first to study this parameter and later Everett and Borgatti (2005a) revisited it. Since ego-networks have diameter two, every pair of non-adjacent

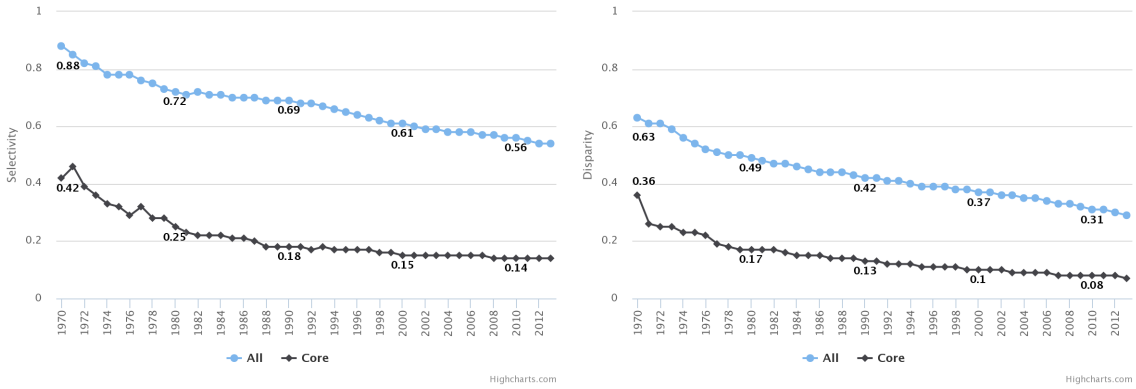


Figure 3.6: Average values of selectivity (left) and disparity (right).

nodes will have a shortest path traversing the central node (there might also be others, and each is counted proportionally) and hence the betweenness will grow with the size of the network quadratically (see Figure 3.7). Although one might expect that the nodes

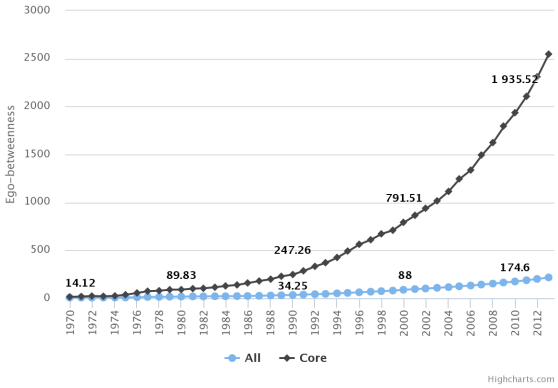


Figure 3.7: Average values of ego-betweennesses.

having the highest ego-betweenness are present in the core, there is a node ranked fifth regarding this parameter, which is not a core node. This phenomenon occurs, because we decided not to consider weights on the edges when counting the number of shortest paths. This is shown also when considering minimum and maximum values of ego-betweenness of the core nodes in 2013. The minimum value achieved is 44, while the maximum is 8325. Almost 50 percent of core nodes have value more than 1600 in 2013, while in the entire network there are only about 3 percent of such nodes.

3.3.4 Connectivity and efficiency

Recently Abbasi, Chung, and Hossain (2012) compared several measures of ego-networks to the scholarly performance of researchers. They confirmed a hypothesis that scholars with more coauthors perform better in terms of research and suggest that the scholars with higher ego-network betweenness perform better. They also considered an interesting theory due to Burt (1992) on structural holes saying, in brief, that the density of researcher's collaboration network negatively correlates to performance. With this they defined another

measure, called *efficiency*, as the proportion between the order of the network and the number of connected components, when the central node is removed (denoted the *effective size*).

First, we analyze effective sizes. Above, we already showed that smaller ego-networks typically have higher densities. One measure to confirm such a behavior is precisely the effective size. Indeed, since ego-networks are graphs of diameter two, the only relevant node to consider/remove is the central node itself. The average effective size of ego-networks is presented in the left chart of Figure 3.8. While the average number of connected components in the ego-networks of entire networks is around two, the average number grows at the same rate for core nodes, overcoming four in 2004. This measure is also an indicator of multidisciplinary work of the nodes in the core, which is somehow limited to a small number of groups. Surprisingly, there are five researchers having effective size at least 22

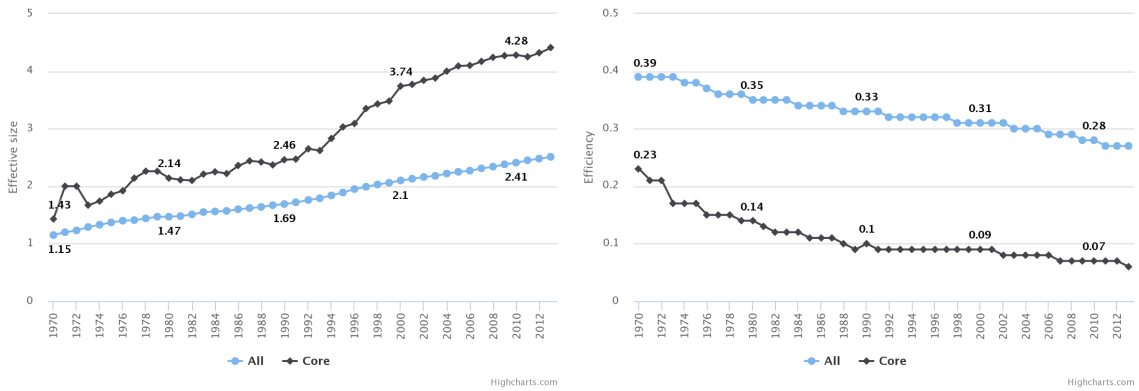


Figure 3.8: Average values of effective size in ego-networks (left) and average values of efficiency (right).

in 2013, but none of them is contained in a core; the maximum effective size in the core is 22 in 2013.

It is in a way clear that the core nodes have higher effective size due to their high degree and low density. Their efficiency however is much lower than the average efficiency overall (see Fig. 3.8). This fact is a result of a custom that the head of a research group is set to be a coauthor of publications, where he might not be involved much (if at all). In this way, the network of coworkers is still relatively dense and the efficiency measure is thus lower.

3.3.5 Collaboration in cores

In this subsection, we present an analysis of collaboration with the researchers in cores. One might expect that the core nodes form a center of their community and are not involved in areas where someone else is the leading researcher. As it can be seen in Figure 3.9, this is not the case. The core nodes are on average much more connected with other core nodes as an average node. Surprisingly, the average number of core neighbors is approximately three since 1980, whereas this number grows for core nodes and exceeds twelve in 2010. Regardless of the high degree of core nodes, it can be seen that the growth of this parameter is faster since 2002, which may show that the researchers with high publication rate connect more intense. We may speculate that they make consortia for project applications. In 2013, there are two core nodes without any neighbor from the core and additional three having

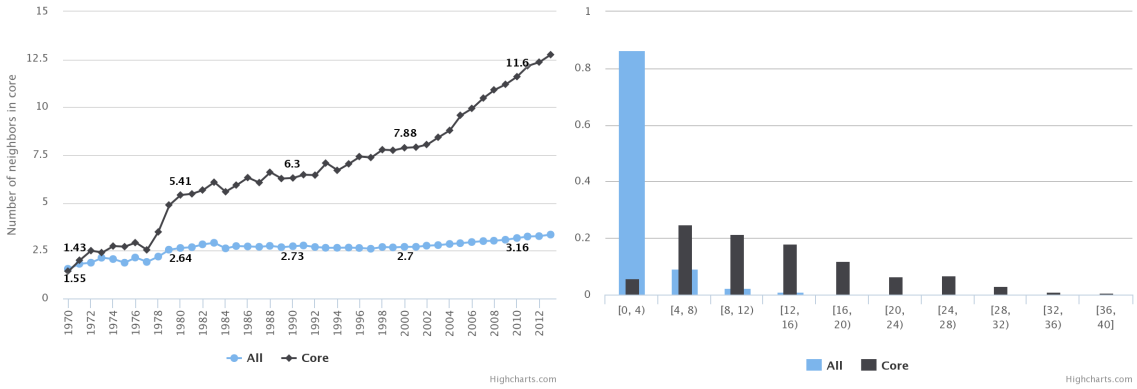


Figure 3.9: Average number of neighbors from the core (left), and the distribution of values of core neighbors in 2013 (right).

only one core neighbor. On the other hand, there are more than fifteen core nodes having at least 29 core neighbors in 2013. Additionally, there is also a non-core researcher with 30 core neighbors. The distributions of the values in the core and entire networks, however, differ importantly as it can be seen in the right chart of Figure 3.9.

3.3.6 Maximum k-cores

The *k-core* of a graph is the maximal subgraph with the minimum degree k. This concept was introduced by Seidman (1983). We have computed the maximum k for every ego-network such that its k-core is non-empty. Although the core-ego-networks are sparser, their maximum k is on average much bigger. In 2013, the maximum k for core ego-networks is almost eight, while it is just above three for ego-networks of all nodes (see the left chart in Figure 3.10).

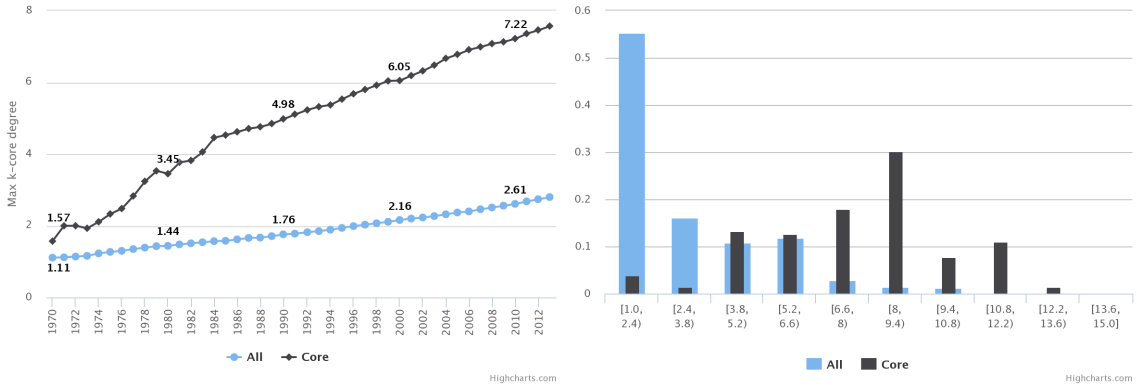


Figure 3.10: Average degree of maximum k-cores (left), and the distribution of maximum k-core values in 2013.

Notice that there is also the difference in the distribution of values. On one side the distribution is decreasing for general ego-networks, and on the other it is close to normal distribution for core ego-networks. In the right chart of Figure 3.10, the distributions for

2013 are shown.

3.4 Describing Core Dynamics on the Level of an Individual Using Machine Learning

So far we have analyzed how the structure of core ego-networks compares to perifery ego networks, by measuring different structural properties evolving in time. In this section we focus our analysis on the level of individuals in attempt to describe and model the dynamics of core position through the research career. Our hypothesis is that the core versus periphery position dynamics differs from individual to individual, but can be grouped according to stability, balance and strength patterns. In order to automatically detect the patterns and build a descriptive model we develop an approach composed of following steps: (1) constructing a feature set that quantifies discrete core-periphery time-series data, (2) group the features so that the three qualities can be revealed - stability, balance and strength, (3) perform clustering using (K-means algorithm), and (4) build descriptive model (decision tree). The proposed approach enables analysis of bibliographic networks by using the network structure as captured by the core-periphery analysis and by applying machine learning techniques. The architecture of the proposed approach is given in Figure 3.11.

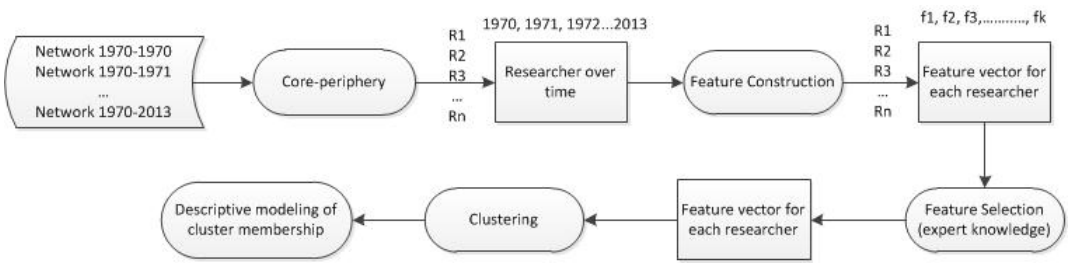


Figure 3.11: Architecture of the approach for studing evolution of bibliographic network structure as captured by the core-periphery analysis. Notice that a network for year m contains information from all the previous years (e.g. 1971 contains information for 1970 and 1971, as our data starts with 1970).In the network we have n researchers (R1 to Rn) and we have defined k features (f1 to fk) for describing them.

3.4.1 Feature construction

We experimented with 20 features listen in Table 3.1. These features were computed for core–periphery time-series for each researcher. In Figure 3.12, we present an example of a core–periphery time series for a researcher. From a set of many features that describe time-series we selected and grouped to cluster the researchers according to stability, balance and strength of retention in the core. Selection of the features for clustering was done by applying expert knowledge and analysing internal criterion, specifically maximizing intra-cluster similarity and minimizing inter-cluster similarity. For a descriptive analysis using clustering part of the approach, we provided the complete set of features at the input, so the algorithm could automatically select the most discriminative features.

Feature selection is performed based on the predefined types; for each type of features one data set is formed. On each data set clustering is performed to identify groups of

Table 3.1: Descriptions of features derived from core-periphery time series. Each year a researcher can be in the core (position one), or in the periphery (position zero).

Feature name	Description
number of ones	Number of time points in core position.
one ratio	Number of time points in core divided by the total number of time points observed.
num of zero to one	Number of transitions from periphery to core.
num of one to zero	Number of transitions from core to periphery.
num changes	Number of transitions.
number of one areas	Number of continuous sequences in core.
number of zero areas	Number of continuous sequences in periphery.
number of one area rate	Number of continuous sequences in core divided by maximum possible number of such sequences.
one to zero num areas rate	Number of continuous sequences in core divided by number of sequences.
min zero area	Size of the shortest continuous sequences in periphery.
max zero area	Size of the longest continuous sequences in periphery.
avg zero area	Average size of continuous sequences in periphery.
avg zero area rate	Average size of continuous sequences in periphery divided by number of time points.
min one area	Size of the shortest continuous sequences in core.
max one area	Size of the longest continuous sequences in core.
avg one area	Average size of continuous sequences in core.
avg one area rate	Average size of continuous sequences in periphery divided by number of time points.
one ratio X/Y	Number of core time points in Xth part of the observed period (divided on Y equal parts), divided by the size of Y.
one balance X/Y	Number of core time points in Xth part of the observed period (divided in Y equal parts), divided by the total number of core time points.
balance entropy	Shannon entropy with probability distribution given by one balance 1/6 - one balance 6/6.

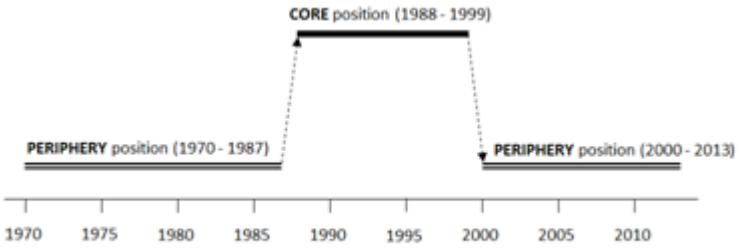


Figure 3.12: An illustration of a core membership information for an arbitrarily chosen researcher. The researcher in this example has periphery position in period from 1970 to 1987. In 1988 he enters the core of the network and in 2000 moves back to the periphery.

similar researchers. Notice that two researchers may be similar when considering stability of their research career development, but different when considering some other feature type. In order to analyze the obtained clusters we propose to apply descriptive modeling on cluster membership. In particular, we have applied decision tree algorithm to model the cluster membership, as it enables a good interpretation of the obtained model.

3.4.2 Clustering and classification

Using the identified features we perform feature engineering for stability, balance and strength, perform clustering and build descriptive models for automatic classification.

3.4.2.1 Stability

Stability describes continuity and length of retention in the core. After entering the core, highly stable individuals will stay in the core for significant amount of time, without frequent changes between the core and periphery. Experimentation and feature engineering led to choosing number of changes between core and periphery, ratio of core years in the observed period and average length of continuous core periods, as the features for clustering according to stability. K-means clustering algorithm was used to obtain clusters of individuals defined by stability. Number of clusters was determined by observing within group sum of squares (WSS) by number of clusters relation. Sum of squares declines as new clusters are added, so we are looking for the number of clusters where the reduction in sum of squares is large and where adding new clusters would result with lower reduction. In this case the chosen number of clusters was 4. The resulting 4 clusters with different degree of stability were labeled with low, medium low, medium high and high stability. The representative samples or cluster medoids are plotted in Figure 3.13.

The classification model automatically chooses `num_changes` as the single feature for modeling stability and determines thresholds for different stability classes. According to the model, if number of changes is greater than 3, samples are classified into "low stability" class. Researchers with exactly 3 changes between core and periphery are classified into class "medium low". Two changes give classification into class "medium high" and one or no changes the most stable class "high".

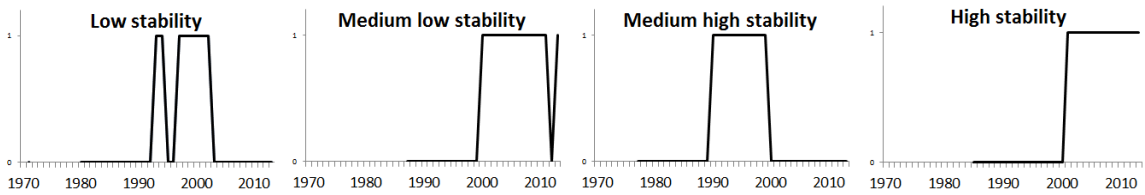


Figure 3.13: Stability clustering medoids.

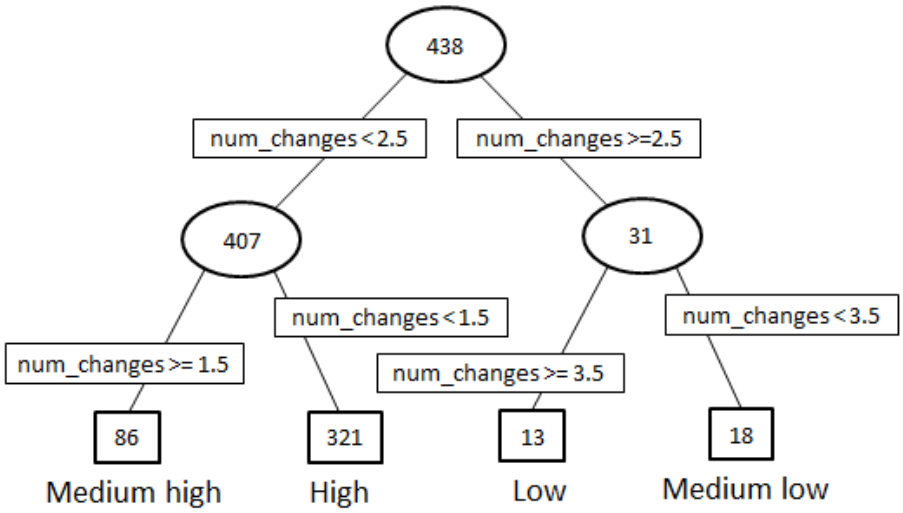


Figure 3.14: Decision tree classification model for stability.

3.4.2.2 Balance

Balance shows in which parts of the observed time period is an individual in the core. Behavior can vary from continuous core position, to core position in time chunks of different sizes distributed across the observed time period. Clusters defined by balance were identified using one ratio X/Y and one balance X/Y, where the time period was divided on six equal parts. These features describe the amount of time units in core in different parts of the observed time period. The time period is evenly divided into 2, 3, 4 and 6 parts. For example, feature one rate 2/3 has value 1.0 if an individual is in the core during the complete 2nd third of the observed time period. One balance indicates that the proportion of time spend in core during the time chunk in relation to the total number of units spent in core is measured. For example, the feature one balance 5/6 has value 0.5 if an individual was 2 time units in the core in the fifth sixth part of the period and was 4 time units in total in the core during the complete observed time period. Clusters on the basis of balance were identified using the K-means clustering algorithm. Five clusters labeled as "start", "late start", "middle", "early 1end" and "end" were identified. Medoids of the clusters are plotted in Figure 3.15. In order to generate a descriptive classification model, a decision tree was applied on instances labeled by balance clusters and described with the full set of features. Resulting decision tree is plotted in Figure 3.16. Depending on the proportion of core years in the last part of observed period, the decision tree is divided to the left branch oriented towards the end, and the right branch striving towards the beginning of the research period. At the second level of the decision tree, rates of core years in second

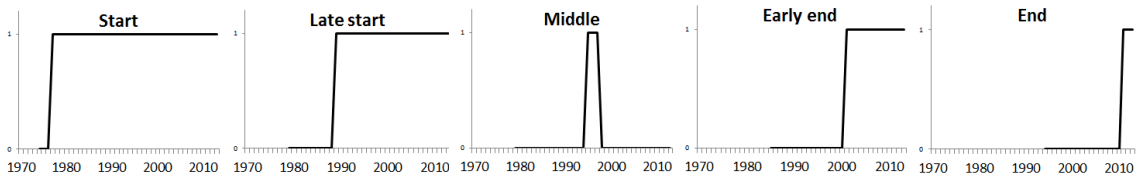


Figure 3.15: Balance clustering medoids.

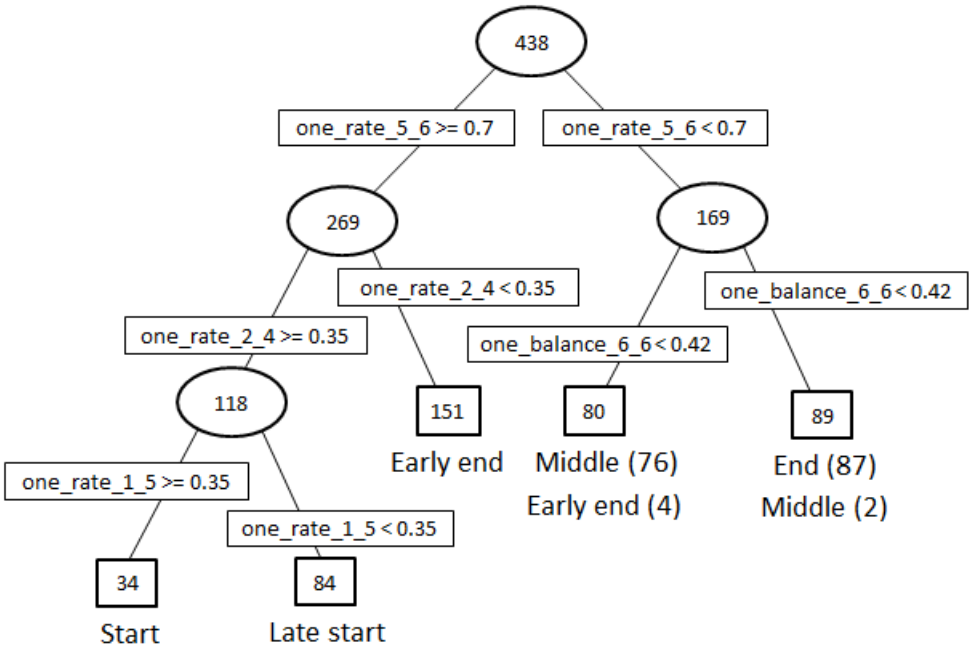


Figure 3.16: Decision tree classification model for balance.

fourth and last sixth of the period are used for splits. Rate of core years in the first fifth is used for distinguishing between "start" and "early start" class.

3.4.2.3 Strength

Strength is the ratio of the time in the core compared to the time in periphery. Individuals with high strength are in the core for the most of the observed time period. By performing cluster analysis it was determined that the strength of the core position can be appropriately defined using number of ones and core ratio features. The first feature is the number of time units a researcher was in the core position, while the second feature is core ratio divided by the total number of time units in the observed period. Clustering on the basis of strength of core position resulted with four clusters. In Figure 3.17, plots of the cluster medoids are depicted. Clusters are labeled as "low", "medium low", "medium high" and "high".

The generated clustering show that researchers can be grouped into clusters with different intensity of core stability, balance and strength. This shows that the researchers have different patterns of behavior in their research career. To further explore the clusters we conduct descriptive classification using the labels obtained with clustering.

Decision tree classification model built using the cluster labels for strength and the full

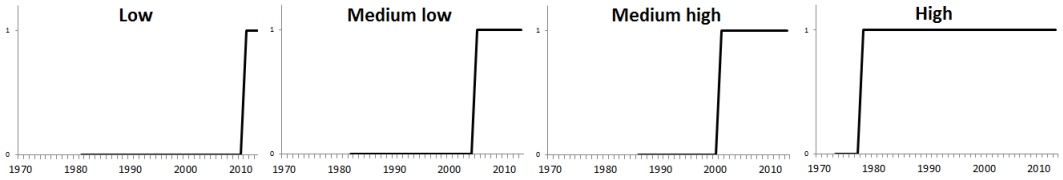


Figure 3.17: Strength clustering medoids.

set of features is plotted in Figure 3.18. Automatically automatically chooses one ratio, number of years in core and balance entropy as the features for classifying according to strength.

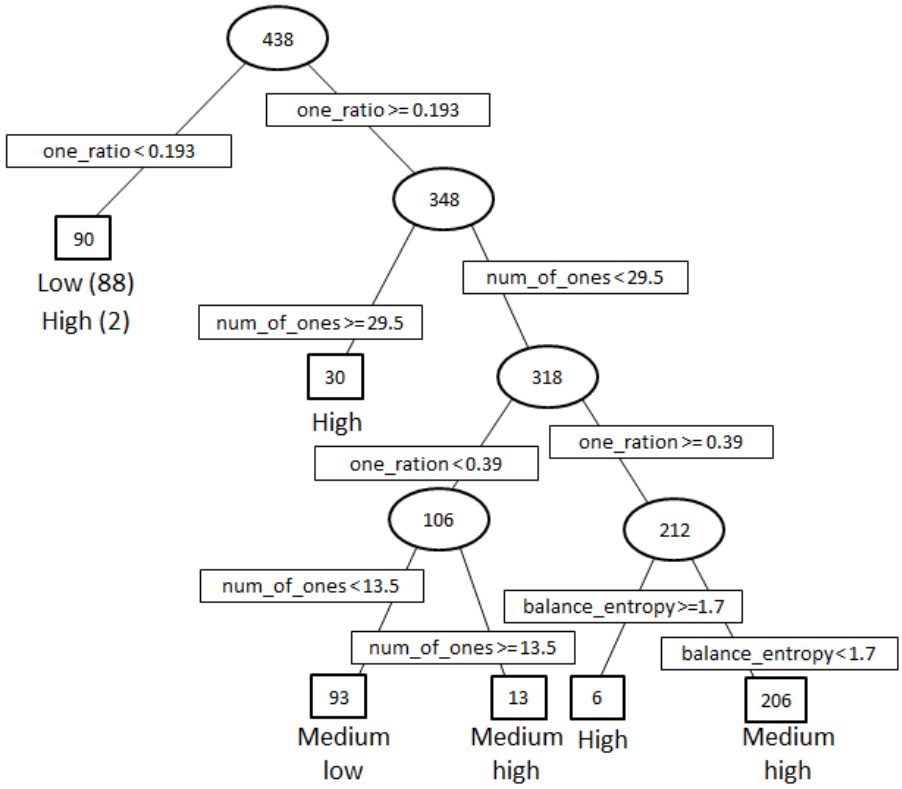


Figure 3.18: Decision tree classification model for strength.

3.5 Discussion

Researchers can have different paths in their career. Some of them may instantly engage in many collaborative activities, while others grow their collaboration network slowly in time. Some of them spend several years working in industry at different points of their career, while others concentrate only on research work and publishing scientific papers. All this influences the position researchers have in collaboration network. We elaborated a method for automatic clustering of researchers into groups according to their core–periphery collaboration network position. Here, we illustrate a usage of the classification model for

describing core–periphery dynamics on researchers from educational sciences (researcher A), pharmacy (researcher B), and computer science and informatics (researcher C). We apply the model to categorize core dynamics according to stability, balance and core strength. In Figure 3.19, the time-series data for core position with marked year of obtaining a PhD

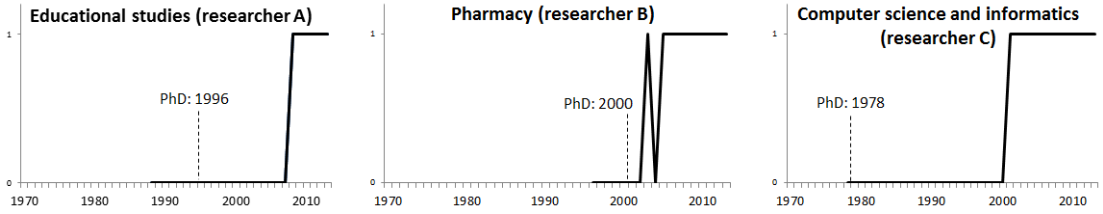


Figure 3.19: Core-periphery dynamics time-series data for researchers A, B and C from collaboration networks, with the year of obtaining their PhD title marked.

is shown. The researcher A obtained PhD in 1996, which was eight years after entering the network. The researcher enters the core in 2008 and keeps the core position until the end of the observed period. The model characterizes the researcher with medium low stability, middle end balance and medium strength. The researcher B entered the collaboration network in 1996 and obtained PhD in 2000. In 2003, he enters the core, leaves it next year, but re-enters the core in the following year and keeps the core position until the end of the observed period. Using the classification model trained on the co-authoring network core dynamics the researcher is characterized as medium-low stable, with middle-end balance and medium strength. Lower stability is caused by core-periphery alternations between 2002 and 2005, balance is middle-end based because the core position is roughly retained in the mid point since entering the network until the end, and the strength is medium because the researcher is approximately half of the time in the core position. The researcher C obtained PhD in 1978, which is the same year when he entered the collaboration network. The researcher stayed in the periphery of the networks for 22 years, entered the core in year 2000 and maintains the core position until the end of the observed period. The researcher's core dynamics is characterized by high stability, balanced at the end of the observed period (end balance) and medium strength.

Researcher	A	B	C
Stability	high	medium low	high
Balance	end	early end	end
Strength	middle low	medium	medium

Table 3.2: Classification of researchers from the illustrative example.

In Slovenia, a thorough research evaluation system is established. Every registered researcher is assigned a score regarding his publication, citation and industrial-collaboration performance. The ranking of the three analyzed researchers within their science fields is 61 for the researcher A, 16 for B, and 21 for C. Their ranking regarding the strength of the nodes within the networks is however just the opposite, the researcher A is ranked 112, B is 279, and C is ranked 196 in the year 2013 among 16012 researchers analyzed. Further analysis shows that some of the ego-network parameters of the three researchers differ substantially and the others are similar (see Table 3.3), which can be explained also with the fact that they come from different science fields.

Parameter \ Researcher	A	B	C
Order	44	100	77
Size	167	441	229
Density	0.18	0.09	0.08
Selectivity	0.49	0.26	0.17
Disparity	0.03	0.06	0.04
Ego-betweenness	531.74	3696.91	2461.77
Effective size	4	6	7
Efficiency	0.09	0.06	0.09
Neighbors in core	6	16	6
Maximum k-core degree	7	8	6

Table 3.3: Ego-network parameters of the researchers A, B, and C in the year 2013.

A comparison of overlapping between the members of cores and the researchers with the highest evaluation scores is also interesting. Surprisingly, there is only about 30 percent of core members ranked in top 100 within natural and technical sciences. The highest inclusion is achieved in biotechnical sciences, a 66 percent. This phenomenon is to be understood due to the limitations of our analyses to collaboration between Slovenian scientists only and not taking into account citations or industrial collaboration.

3.6 Conclusion and Future Work

The proposed approach combines structure analysis of a network based on core-periphery partition and machine learning methods to study evolution of collaboration networks over time. The approach is applied to collaboration network of Slovenian researchers capturing their publication activity over the last 40 years.

The analysis mainly confirmed the initial hypotheses: indeed, the core members are widely distributed over the networks; comparison and analysis of average values of various network parameters for the core-ego-networks and ego-networks overall has mainly shown as expected that the values are much different in cores. We did however obtain several results which are non-trivial, e.g. high collaboration between core members. We have shown that researchers indeed exhibit different patterns of core-periphery dynamics. Using our approach patterns of behavior can be identified and researchers can be automatically classified into categories. We used K-means clustering to identify the patterns, applied feature engineering to group the features to characterize stability, balance and strength qualities and finally build a decision-tree classification model that can automatically classify researchers. In this work we applied feature engineering for three properties of core–periphery position, but the same approach can be applied for other properties and network characteristics that can be described with time-series. Finally, we found that the core position within a limited network does not mean that the member's performance is good in general.

We will continue our analysis with analysis of core-periphery structures within networks filtered to science branches. Different publication culture in different sciences is evident and different behavior of core-periphery structure should be revealed within different sciences. Our analysis can also be combined with the analysis of communities, showing how community detection can be related to core–periphery structure, and that perhaps communities can be built around the backbone nodes. We believe that one can introduce a measure for community centers/leaders. Additionally, we will perform a similar analysis

on the networks of all co-authors of Slovenian scientists, in order to reveal the culture of international collaboration within different science fields. We already mentioned earlier that the properties of ego-network were studied in order to obtain some knowledge on entire networks. A natural question arises whether meso-scale structures can be determined through ego-networks.

Chapter 4

Measuring Diversity of Network Elements Using Text and Structure

Nodes that are connected in a network can have different levels of diversity in terms of their attributes. We developed a new measure for diversity using the structure of the network and the text that describes the nodes and used it to determine the interdisciplinarity of the network of scientists. Many believe that great advances take place at the intersections between disciplines, making the interdisciplinary research almost universally acclaimed as 'the way to go' (Porter and Rafols 2009). Interdisciplinary research is seen as more successful at making breakthroughs and generating more relevant outcomes, be it in terms of innovation for economic growth or for social needs (Rafols & Meyer, 2010). It is a complex issue as modern society increasingly demands application-oriented knowledge and the usability of scientific knowledge generally requires combination and integration of knowledge from various scientific disciplines (van den Besselaar & Heimeriks, 2001). An analysis of 17.9 million papers spanning all scientific fields shows the importance of interdisciplinarity for scientific impact (Brian Uzzi & Jones, 2013). In this study we propose a new measure for interdisciplinarity and a methodology for constructing subject categories similarity matrix based on text similarities between descriptions of categories. The methods are applied to research community in Slovenia, where innovations could stimulate the growth of the economy, strongly affected by the economic crisis. Below is the summary of key scientific contributions on methods and exploratory findings of this study. Key contributions on methods:

- Extending Stirling's diversity index for graphs, by adding a component that captures connectivity between subject categories.

- Using text similarity approach for constructing science categories similarity matrix used for interdisciplinarity calculation.

Key exploratory findings:

- The interdisciplinarity based on research collaboration in medical science increases fastest with time. The main reasons are intensive collaborations with technical, natural and social sciences in the period from 2004 to 2006 and with natural and especially technical science from 2011 to 2013.

- Biotechnology and natural sciences are the most interdisciplinary science disciplines whether we look at project collaboration or co-authorship of publications.

- In co-authoring publications social and humanistic sciences are more interdisciplinary than technical and medical sciences, while in collaborations on research projects we can see the opposite.
- The collaboration and interdisciplinarity based on projects increases more intensively with time than collaboration and interdisciplinarity based on publications.
- The annual linear growth rate of interdisciplinarity based on projects is approximately double of the rate of growth of collaboration on projects, while the interdisciplinarity based on publications is approximately half of the rate of the growth of collaboration on publications.
- Some research topics diverse by number of researchers from different science fields are less interdisciplinary than some other less diverse topics, and vice versa.
- Interdisciplinarity based on co-authoring versus project collaboration varies across topics.

This research work was conducted in collaboration with Dunja Mladenić and was published in Journal of Scientometrics (Karlovčec & Mladenić, 2015).

4.1 Background

Since there are many different definitions of interdisciplinarity, we begin with the review of the basic terms: scientific discipline, unidisciplinary, multidisciplinary, interdisciplinary and transdisciplinary science. Next, we introduce some of the related work on measuring interdisciplinarity. Finally, background on using collaboration as the basis for interdisciplinarity is given.

4.1.1 Basic concepts

Scientific disciplines are organized around the study of particular substantive phenomena (e.g., psychological, social, environmental, biological). Such organization of science is useful in that it highlights the distinctive substantive concerns (e.g., biological, psychological, social, geographical phenomena), analytic levels (e.g., cellular, cognitive, emotional, interpersonal, organizational, social), concepts, measures, and methods associated with particular fields of study (Stokols et al., 2003). The concept of a scientific discipline comprises both a body of knowledge and a social body that generates, evaluates, communicates, and teaches the corresponding knowledge, i.e. discipline is a combined cognitive and social category (Schummer, 2004). In the rest of this chapter we will use the terms science field and science subfield as a type of scientific discipline used in this research.

Unidisciplinary research relies solely on the methods, concepts, and theories associated with a single discipline, such as psychology, sociology, geography, or medicine. Multidisciplinarity refers to a process whereby researchers in different disciplines work independently or sequentially, each from his or her own discipline-specific perspective, to address a common problem. (Stokols et al., 2003)

Interdisciplinarity has a vast number of different definitions and there is no common established meaning of the term. In (Stokols et al., 2003) interdisciplinarity is defined as a process in which researchers work jointly, but from each of their respective disciplinary perspective, to address a common problem. This definition differs from others in that it does not include the integrative effect of the process, which is included in the definition of the next step - transdisciplinary. Another definition says that interdisciplinary approaches

integrate separate disciplinary data, methods, tools, concepts, and theories in order to create a holistic view or common understanding of a complex issue, question, or problem. The critical indicators of interdisciplinarity in research include evidence that the integrative synthesis is different from, and greater than, the sum of its parts (Stokols et al., 2003; Wagner et al., 2011). Guimera et al. (2005), Porter, Cohen, David Roessner, and Perreault (2007) also put emphasis on integration, saying that interdisciplinary research requires an integration of concepts, techniques, and/or data from different fields of established research. Interdisciplinarity is a process of integrating different bodies of knowledge rather than transgression of disciplinary boundaries per se. The integration is the key aspect of so-called interdisciplinary research (Rafols & Meyer, 2010; Guimera et al., 2005; Rafols & Meyer, 2010).

Transdisciplinarity is a term related to interdisciplinarity. The existing definitions emphasize the existence of frameworks and taking into account sustainability and all the possible disciplines and stakeholders. In contrast to unidisciplinary research, transdisciplinary science involves collaboration among scholars representing two or more disciplines in which the collaborative products reflect an integration of conceptual and/or methodological perspective drawn from two or more fields (Stokols et al., 2003). Transdisciplinary approaches are comprehensive frameworks that transcend the narrow scope of disciplinary worldviews. For instance, transdisciplinary science in cancer research is a form of transcendent interdisciplinary research that fosters systematic theoretical framework for defining and analyzing social, economic, political, environmental, and institutional factors in human health wellbeing. It is a new mode of knowledge production that draws on expertise from a wider range of organizations, and collaborative partnerships for sustainability that integrate research from different disciplines with the knowledge stakeholders in society. The transdisciplinary product is greater than the sum of its parts, though the scope of the overall effort is more comprehensive and the parts may be more diverse (Stokols et al., 2003). Transdisciplinary science involves the integration of theoretical and methodological perspectives drawn from different disciplines, for the purpose of generating novel conceptual and empirical analysis of a particular research topic (Rosenfield, 1993; Klein, 1996). Transdisciplinarity concerns that which is at once between the disciplines, across different disciplines, and beyond all disciplines. Its goal is (a) the understanding of the present world, of which one of the imperatives is the unity of knowledge, and (b) the solution of mega and complex problems by drawing on and seeking to integrate disciplinary and stakeholder views on the basis of some overarching theory (Klein, 1996). According to the definitions, the transdiciplinarity is a transcendent form of interdisciplinary research, relaying on the same key concept as the interdisciplinarity - integration of knowledge.

4.1.2 Related work

Porter and Rafols (2009) performed an interdisciplinary study of six research fields corresponding to the Web of Science subject categories, in the period from 1975 to 2005. They use the diversity index (Stirling, 1998) as the integration score, which was based on citations and subject categories associated with the journal that published the cited articles. The similarity matrix between subject categories was constructed by calculating the integration score on a subset of the data. The results showed that the integration score of all six fields increases over time, but significantly less than the mean number of authors per paper and the mean number of disciplines per paper. Rafols and Meyer (2010) used the Stirling's measure of diversity simplified to disparity (to avoid imposing categorization), network centrality measures and the normalized Shannon-Wiener and Herfindahl-Hirschmann (Simpson's index used in biodiversity) index to perform an interdisciplinary study on authors and publication in two case studies from bionanoscience.

Morillo, Bordons, and Gómez (2003) examined the variety and balance of scientific disciplines using pre-existing categories. Research areas and categories are described according to the quantity of their links (number of related categories) and their quality (with close or distant categories, diversity, and strength of links). Schummer (2004) investigates interdisciplinary research collaboration in nanoscience and nanotechnology in 2003 and 2004. The applied interdisciplinarity measure was based on co-authorship of authors from different scientific disciplines. Similar as work of Porter, Roessner, and Heberger (2008), this work uses co-authors' affiliation as an in interdisciplinarity indicator, which is the same method as used in our work. Adams, Jackson, and Marshall (2007) have analyzed bibliometric indicators of interdisciplinary research, with hypothesis that since it is marginal to core subjects, it would be systematically cited less often. Using the Shannon diversity index, they found no tendency for the most interdisciplinary articles to be less frequently cited on any consistent basis within or across categories, and a weak tendency for the articles with highest citation impact to be neither very mono-disciplinary nor very multidisciplinary in terms of their cited references. Work of Leydesdorff (2006) focuses on applying social network analysis measure of betweeness centrality (Freeman, 1977) as an indicator of interdisciplinarity. Betweenness centrality is shown to be an indicator of the interdisciplinarity of journals, but only in local citation environments. The same method is used in Leydesdorff (2007) for measuring the interdisciplinarity of journals that are parts of the overlap between the Science Citation Index and the Social Science Citation Index.

Other related work include (van den Besselaar & Heimeriks, 2001) where factor analysis of the journal - journal citations matrix of the core journal of a specialty is applied to describe research fields in terms of their position within scientific communication networks. The work of Zitt, Ramanana-Rahary, and Bassecoulard (2005) is a study of impact of level of aggregation (journal, specialties (ISI subject categories), sub-disciplines and disciplines) on the interdisciplinarity measure.

4.1.3 Interdisciplinarity and collaboration

It is a widely-held view that the publication is an essential part of the scientific research process (Wagner et al., 2011). It provides the function within science of correction, evaluation, and acceptance by a community. Even though research is mostly measured by publication output; there are other activities important for the research process (Porter et al., 2007). These include research projects, patents, lectures and discussions. Wagner et al. (2011) state that co-authorship of scholarly output is "a convenient, but weak indicator of interdisciplinarity". However, the only given argument is the difficulty of obtaining data about the discipline/field of authors. The inconvenience of obtaining data is not a convincing argument of weakness of a method. Moreover, in some cases (as shown in the rest of the chapter) the discipline and affiliation data is accurate and available. Research publications are recognized as the most clear-cut entities on which to gauge interdisciplinarity of the underlying research by Porter et al. (2007). The authors suggest the co-authoring of publication is the primary information for investigating the collaboration aspect of interdisciplinary research, but secondary for integration aspect—the essence of definition of interdisciplinarity. The argument is that the integration can be accomplished by a single author as well as by a team. At the same time the conducted interdisciplinarity investigation performed on publication co-referencing information was based on a premise that integration can be measured by examining the spread of a paper's references, which is not a stronger premise than the one relying on a spread of paper's authors. We argue that for investigating interdisciplinarity of research at some scope, all the information of research activity has to be considered. The availability and quality of different types of information varies in different datasets, which has to be taken into account in every inves-

tigation. We perform investigation of interdisciplinarity of research collaboration based on co-authoring publications and cooperation on research projects and, show results of both static and dynamic analysis on the case of the national research community of Slovenia. We investigate collaboration of researchers from different science fields on publications and projects, as we believe that research collaboration stimulates sharing and integration of knowledge, resulting in the interdisciplinary of research.

4.2 Data Description

The information on research activity in Slovenia was obtained from SICRIS national database containing information about researchers, research projects and organizations and COBISS national bibliographic database containing information on publications. Time period of 18 years from 1996 to 2013 was covered in the analysis. The researchers are classified by Slovenian National Research Agency classification into sciences fields and science subfields. The classification of researchers into seven science fields: (1) natural, (2) technical, (3) medical, (4) biotechnical, (5) social, (6) humanistic and (7) interdisciplinary, and 72 subfields was the basic information used for measuring interdisciplinarity. Each researcher selects the classification that best matches her research interest. The largest of the fields is technical science with 6,041 researcher, followed by natural science with 2,975, social science with 2,460, medical science with 2,050, biotechnical with 1,234, humanistic 1209 and finally interdisciplinary science with 282 researchers. The researchers have assigned keywords in their profiles that describe their research interest. On average a researcher is labeled with five keywords.

4.3 Approach

Here we introduce measures used for measuring interdisciplinarity and the extension of Stirling's index with connectivity attribute, which makes it suitable for measuring interdisciplinarity of graphs.

4.3.1 Interdisciplinarity measure

In order to quantify interdisciplinarity, we apply a general framework for analyzing diversity proposed by Stirling (2007),

$$\Delta = \sum_{ij(i \neq j)} (d_{ij})^{\alpha} (p_i p_j)^{\beta} \tag{4.1}$$

where p_i and p_j are proportional representations of elements of disciplines i and j in the system (balance) and d_{ij} is the degree of difference (disparity) attributed to elements i and j. Combinations of the values 0 and 1 of parameters α and α, yield four properties of interest: variety, balance, disparity and diversity. As shown in Equation 4.1, diversity increases with variety, balance and disparity. We apply the diversity variant of the generalized diversity framework (Equation 4.1) with values of exponents $\alpha = 1$ and $\beta = 1$, as the main measure for interdisciplinarity. By changing the parameters α and β, diversity indices can be explored for different variations of the properties of diversity-variety, balance and disparity (Figure 4.1). Note that the variant where $\alpha = 1$ and $\beta = 1$ was initially introduced by Rao (1982). Shannon's diversity is a more traditional measure based on Entropy. It does not take into account differences among the categories,

Entropy is a concept used to measure the randomness or uncertainty of a given set of elements. This formula was used by Grupp (1990) conjunction with literature statistics

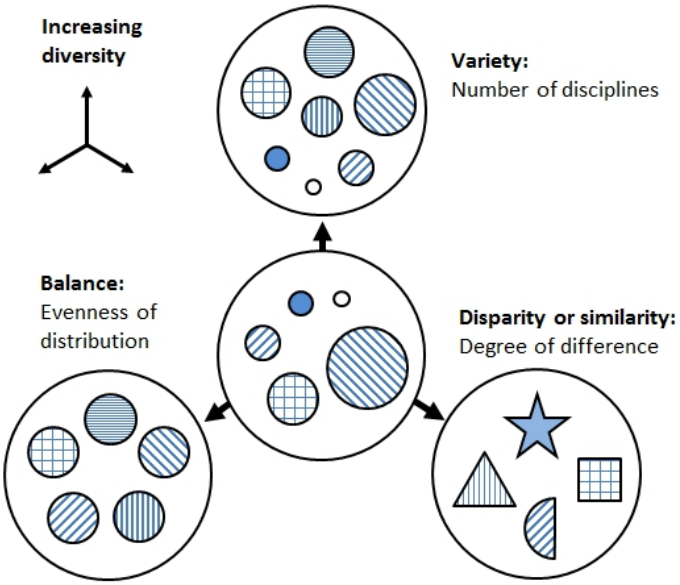


Figure 4.1: Schematic representation of the attributes of diversity, based on (Stirling, 1998).

Table 4.1: Measures of interdisciplinarity used in this study.

Property	Parameters	Equation
Diversity (balance/disparity-weighted variety)	$\alpha = 1, \beta = 1$	$\sum_{ij(i \neq j)} d_{ij} p_i p_j$
Variety (scaled variety)	$\alpha = 0, \beta = 0$	$\sum_{ij(i \neq j)} 1$
Balance (balance-weighted variety)	$\alpha = 0, \beta = 1$	$\sum_{ij(i \neq j)} p_i p_j$
Disparity (disparity-weighted variety)	$\alpha = 1, \beta = 0$	$\sum_{ij(i \neq j)} d_{ij}$
Shannon's diversity		$\sum_i p_i \log p_i$
Stirling's diversity for graphs	$\alpha, \beta, \gamma \in [0, 1]$	$\sum_{ij(i \neq j)} (d_{ij})^\alpha (p_i p_j)^\beta (e_{ij})^\gamma$

(bibliometrics) and patent statistics for studies of disorder of institutions engaged in certain R&D fields.

4.3.2 Extension of Stirling's diversity for graphs

Here we introduce the extension of the Stirling's diversity index that enables measuring diversity of networks or graphs. This is accomplished with addition of the degree of variable e_{ij}, that captures the degree of connectedness between elements i and j in the system. e_{ij} is the proportion of edges connecting elements from i and j, from all possible such edges.

$$\Delta_g = \sum_{ij(i \neq j)} (d_{ij})^\alpha (p_i p_j)^\beta (e_{ij})^\gamma \tag{4.2}$$

The parameter γ enables control of the general framework and different variations of diversity. e_{ij} is the proportions of edges connecting elements in the network i and j to all edges in the network. In the case of some elements i and j being completely disconnected $e_{ij} = 0$. In the case elements i and j are connected with all edges in the network $eij = 1$.

The measure works also on weighted graphs, in which case e_{ij} is the proportion of sum of the weights of all the edges connecting elements i and j, to sum of weight of all the edges. Figure 4.2 shows an example of eight graphs with the same set of nodes and different sets of edges. Table 4.2 contains the values of interdisciplinarity measured by Stirling's diversity and our extension of Stirling's measure for graphs. In the example in Figure 4.2 we have four categories represented with different shading. The dissimilarity between each pair of the four categories used in this example is 0.25. Notice that the absolute values of Stirling's diversity and the proposed Stirling's diversity for graphs cannot be compared, thus we are observing a relative change of values for each measure separately. The original Stirling's measure is the same for all graphs from 4.2a to 4.2h, because the set of nodes does not change for any graph. Graphs 4.2a and 4.2b have extended diversity 0, because there are no edges connecting nodes from different categories. In graphs 4.2c through 4.2h nodes from 4 different categories become increasingly connected, which increases the interdisciplinarity measured according to our extension of Stirling's diversity for graphs. The example illustrates how the extension enables measuring the interdisciplinarity where in addition to variety (number of disciplines), balance (evenness of distribution) and disparity (degree of difference), the elements have different connectivity. The measure is further tested on real networks defined by research topics in section 4.4.6 interdisciplinarity of topic subgraphs.

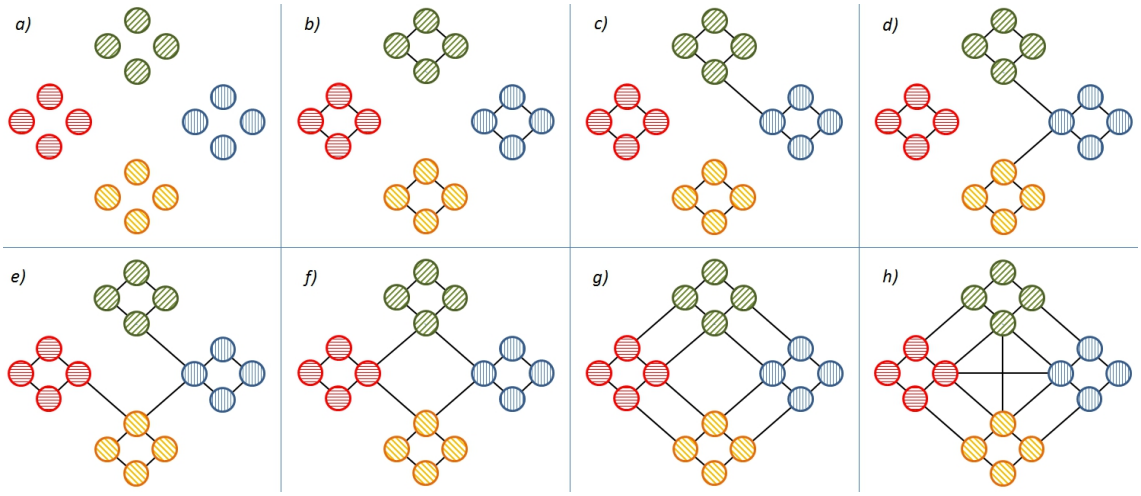


Figure 4.2: Graphs with the same set of nodes from 4 groups and different set of edges. From a–h the groups in the graph are increasingly more connected. Different shades represent different scientific fields.

4.3.3 Subject categories similarity matrix

In order to include the disparity component into interdisciplinarity, the information on similarity between subject categories is needed. Subject categories similarity matrix enables the distinction of influences that different pairs of subject categories have on interdisciplinarity.

For instance, physics and mathematics are known to be related scientific fields, thus, scientific collaboration and integration of knowledge between those two fields is less interdisciplinary than between some two less related fields like physics and literature. In the related work most of the authors constructed a similarity matrix by using a subset of the original dataset to calculate Salton's cosine similarity (Salton & McGill, 1986) between

Table 4.2: Stirling's diversity versus the extension for networks, measured on illustrative graphs from Figure 4.2.

Graph	Stirling's diversity	Extended Stirling's index
a	0.093750	0
b	0.093750	0
c	0.093750	0.000919
d	0.093750	0.001736
e	0.093750	0.002467
f	0.093750	0.003125
g	0.093750	0.005208
h	0.093750	0.006010

Table 4.3: Subject categories similarity matrix for 7 scientific fields.

Fields	Nat	Tech	Med	Bio	Soc	Hum	Int
Nat	1.000	0.323	0.193	0.239	0.046	0.016	0.103
Tech	**0.323**	1.000	0.053	0.060	0.127	0.021	0.180
Med	0.193	0.053	1.000	0.175	0.054	0.007	0.029
Bio	0.239	0.060	0.175	1.000	0.028	0.009	0.048
Soc	0.046	0.127	0.054	0.028	1.000	0.291	0.122
Hum	0.016	0.021	0.007	0.009	0.291	1.000	0.169
Int	0.103	0.180	0.029	0.048	0.122	0.169	1.000

categories, which were described with the same information (Porter & Rafols, 2009) as used for interdisciplinarity. We avoid this circularity and apply the method of measuring similarity between textual descriptions of categories, which is to our best knowledge a novel approach for constructing similarity matrix between subject categories. This approach is feasible, because researchers belonging to subject categories have manually entered keywords describing their research interest. Although interdisciplinary collaboration between two fields certainly influences the research interests of the fields, we believe that relatedness of fields causes collaboration more than collaboration causes assimilation of research interests. Also, keywords are in most cases entered when entering the base and rarely changed in the later stages. For those reasons we believe this is a valid method of constructing subject category similarity matrix. Below is the detailed approach of subject category similarity matrix construction:

1. Aggregating keywords describing researches into separate documents according to science field, so that each document contains only the keywords of researchers from one category.

2. Generate bag-of-words representation of each document, removing stop-words, applying stemming and identifying n-grams, with $n \leq 5$. This results with vector representation of documents. Elements of the vectors represent TF-IDF weighted frequencies of words in the documents.

3. Calculate cosine similarity between each pair of documents. Cosine similarity is a measures for similarity between documents. Given vectors of attributes generated in step 4, cosine similarity between two documents a and b can be represented as a dot product: $\cos\theta = \frac{ab}{||a||||b||}$

Table 4.4: Pairs with extremely high values of similarity from the constructed scientific subfields similarity matrix.

Pairs of fields		Similarity
Biotechnology	Biochemistry and molecular biology	0.58917
Urbanism	Architecture and design	0.58026
Mechanical design	Mechanical engineering	0.56665
Computer science and informatics	Telecommunications	0.54794
Ethnology	Anthropology	0.53212
Theology	Philosophy	0.51500
Urbanism	Landscape design	0.51133
Systems and cybernetics	Manufacturing technologies and systems	0.50953
Educational studies	Sport	0.50604
Culturology	Sociology	0.49513

4. Calculate disparity between fields as $1 - \cos\theta$.

Table 4.3 shows similarity matrix constructed for 7 science fields of national ARRS classification. According to the matrix the most similar pairs of science fields are: (1) natural - technical, (2) social - humanities, and (3) natural - humanities. On the other hand, the most unrelated pairs of science fields according to the matrix in Table 3 are: (1) medical - humanities, (2) biotechnical - humanities, and (3) natural - humanities. Besides similarity matrix for 7 science fields we have constructed a matrix for 72 science subfields of the national ARRS classification. Since the whole matrix would not fit in the form of a table, in Table 4.4 we are highlighting ten pairs of subfields with the highest similarity.

4.4 Results

Community of researchers can be defined based on different properties. Scientific fields and subfields are one possible way of defining research communities. Another possibility is to consider the research topic. First, we report results of interdisciplinarity investigation on the defined scientific fields and subfields. In the second part, we generate collaboration subgraphs based on different research topics and apply the extended Stirling's diversity index to calculate interdisciplinarity.

4.4.1 Interdisciplinarity of science fields and subfields based on project collaboration and co-authoring publications (1996-2013)

In order to investigate evolution of interdisciplinarity, ego-networks of researchers were constructed. Ego-network is a network containing a central node and all the nodes that collaborated with the central node in a defined time (notice that collaborations between the other nodes are not considered in the ego-network). Ego-networks were constructed for each researcher and for every year from 1996 to 2013, with addition of a network that included the complete time span. Since interdisciplinarity was calculated both based on collaboration on research projects and co-authoring scientific publications, two variations of each ego-network were constructed. Stirling's diversity index, together with the three components of the diversity (variety, balance and disparity) and the more traditional Shannon's measure were calculated for each ego-network. In order to calculate the diversity of the scientific fields, the values of the individuals (belonging to the field) were averaged with arithmetic mean, with reported standard deviation. Notice that in this way we estimate

the interdisciplinarity of the field based on collaborations without considering the content of research work.

Another way would be to construct a network for each field and determining interdisciplinarity based on a set of nodes. The set of nodes would consist of the researcher belonging to the field and the researcher connected to them. We use the ego-networks approach because it is more robust. For example, consider a scientific field a with n researchers where only one researcher collaborates with n researchers from another field, while the rest n-1 researchers from a do not collaborate. Using the second approach, the interdisciplinarity of the field would be greater than interdisciplinarity of the most interdisciplinary researcher. Using the ego-networks approach, the interdisciplinarity of the field would be influenced by the most interdisciplinary researcher, but it would be lowered by other non-interdisciplinary researchers. The downside of ego-centric approach is that the collaborators in different ego-networks can be set of the same researchers and the interdisciplinarity of the field would be the same as if this set would be different in every ego-network. However, we consider a field to be as interdisciplinary as are the researchers who belong to the field and find the ego-networks approach the most suitable.

4.4.2 Interdisciplinarity based on project collaboration

The results of interdisciplinarity of the 7 scientific fields based on project collaboration are reported in Table 4.5. Probability density functions for the diversity are plotted in Figure 4.4. The diagrams show that frequency of ego-networks with very low interdisciplinarity (from 0, to 0.02) is very high for all diagrams except for interdisciplinarity. This shows that most of the researchers for every field have un-interdisciplinary collaboration. Exception is the Interdisciplinary field which was established especially with the reason to foster interdisciplinarity, but we must bear in mind that this field is much smaller than the rest of the fields. Note that the Interdisciplinary science field is a field designed especially for the researchers who are performing highly interdisciplinary research work. It is thus expected that this research field is the most interdisciplinary. Another thing that makes this science field special is the small number of researchers belonging to the filed – 282, in opposition to 2975 (Natural), 6041 (Technical), 2050 (Medical), 1234 (Biotechnical), 2460 (Social) and 1209 (Humanities). As expected, the interdisciplinary science field has the highest value of diversity, measured by project collaboration. From the other 6 fields, the highest Stirling's diversity index is for the biotechnical sciences, followed by the natural, medical, technical, social and finally humanities. Looking at the dimensions of diversity, we can see that the technical sciences have higher variety and disparity than medical sciences, but since the balance of the medical sciences is higher, the overall diversity of the medical sciences exceeds that of the technical sciences. Comparing the Shannon's and Stirling's indexes reveals the high correlation between the two, with Pearson correlation coefficient 0.99652.

The dynamics of interdisciplinarity over scientific fields is shown in Figure 4.8. We see that over time, interdisciplinarity has an increasing trend for all science fields. The fastest growing interdisciplinarity has medical science with linear trend function $y = 0.0043x + 0.0626$ and $R^2 = 0.8449$. The biggest increases of interdisciplinarity of medical science were in year intervals 2004 – 2007 and 2011 – 2013.

In order to further analyze the interdisciplinarity of the medical science field, we have plotted the percentage of collaboration of researchers from medical sciences with researchers from other sciences (Figure 4.6). This shows that the collaboration is strongest with natural and technical sciences, with some collaboration with biotechnical and social sciences and almost no collaboration with humanities. The intervals when medical sciences recorded the fastest growth (wider lines in Figure 4.6) are caused by strong increases of collabora-

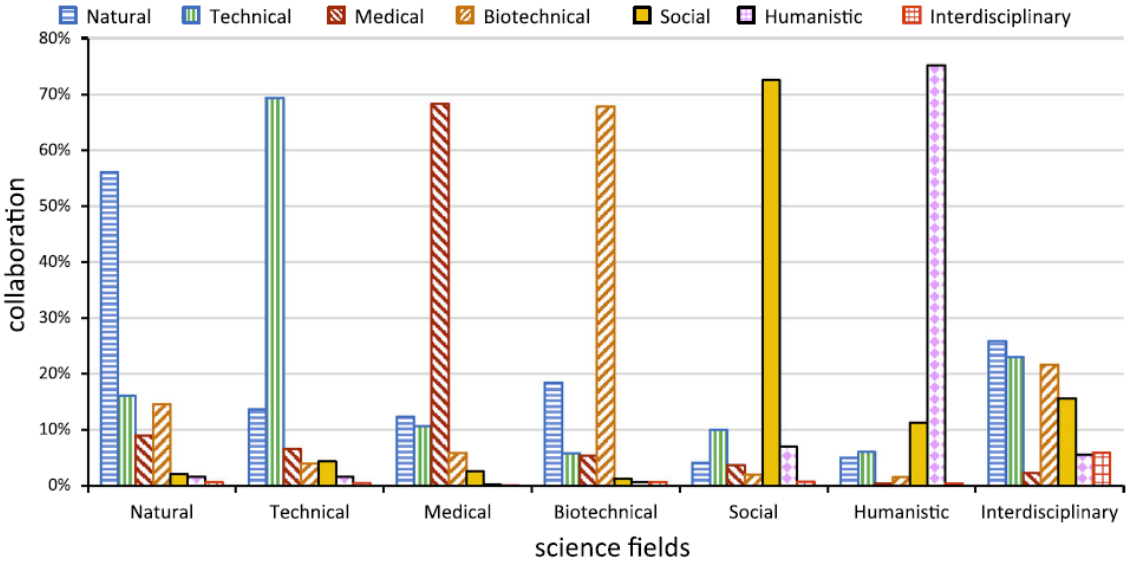


Figure 4.3: Share of research collaboration of each science with the other sciences based on research projects collaboration.

Table 4.5: Interdisciplinarity of science fields measured by research project collaboration

Field	Diversity Mean	SD	Variety[1]	Balance	Disparity	Shannon
Natural	0.1446	0.0914	0.4744	0.1854	2.7925	0.6378
Technical	0.1131	0.0980	0.3725	0.1403	2.2080	0.4907
Medical	0.1279	0.0987	0.3560	0.1486	2.1141	0.5147
Biotechnical	0.1581	0.0795	0.6610	0.1935	3.9392	0.6976
Social	0.1051	0.0964	0.3098	0.1240	1.8866	0.4308
Humanities	0.0893	0.0930	0.2574	0.1079	1.5381	0.3716
Interdisciplinary	0.2324	0.0582	0.9755	0.2756	5.9546	1.0086

tion with natural, technical and biotechnical sciences in the interval 2004-2007 and with natural and especially technical sciences in the interval 2011-2013. In the last interval the collaboration with technical science raised from 8.62% in 2011 to 14.33% in 2013, which almost equals with the collaboration with natural science (14.34%) - the most collaborative throughout the whole measured period.

We have examined in detail the percentage of research collaboration of each scientific field, results are reported in Table 6 and plotted in Figure 6. The results show that the Humanistic science has the highest percentage of within field collaboration (75.14%), which then results in the lowest interdisciplinarity of this science. From other fields, Humanities collaborate mostly with Social (11.31%), Technical (6.07%) and Natural sciences (5.06%), while the collaboration with the Biotechnology (1.58%) and Medical (0.42%) sciences is negligibly small. The strongest collaboration with Social sciences is expected, because these two sciences are related, which is confirmed with our similarity matrix constructed on the basis of keywords similarity. The similarity of humanities and social sciences is 29.1% according to our similarity matrix. This means the collaboration of Humanities with Social sciences does not contribute much to the interdisciplinarity, but the collaboration with Technical and Natural sciences even if almost twice less intense will have stronger

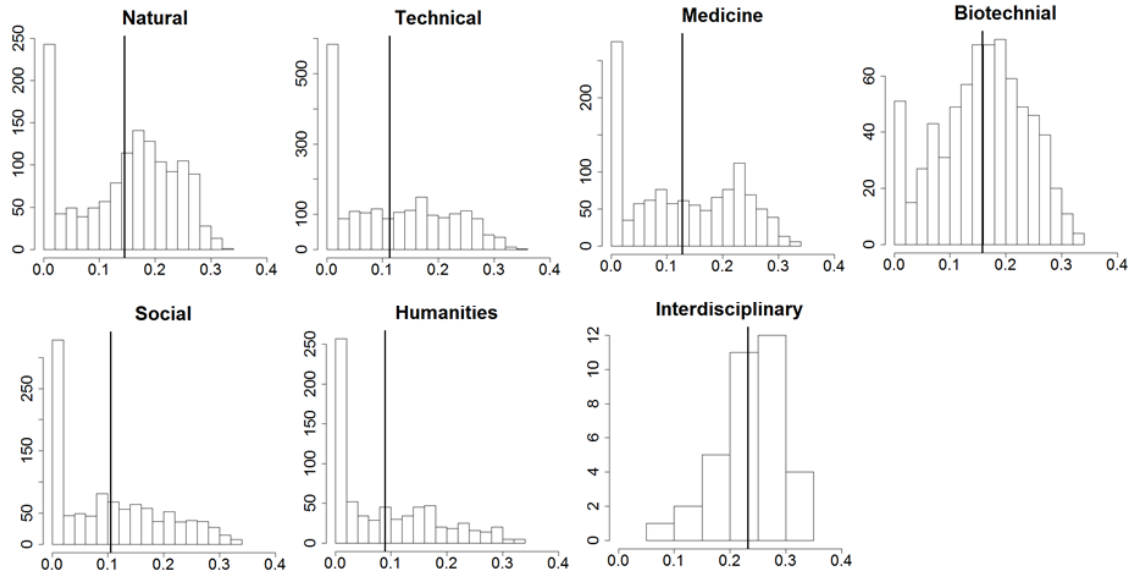


Figure 4.4: Probability density functions of ego-networks diversity index based on project collaboration for science fields from Table 4.5.

Diversity is on x-axis and frequency on y-axis. Intervals for x-axis are equal for each histogram (0–0.4), while intervals for y-axis differ depending on the size of the field. There are 20 classes of width 0.02 for each diagram except for Interdisciplinarity, which has 8 classes of size 0.05.Vertical lines indicate mean values

impact on interdisciplinarity measure because of low similarities with these fields – 2.1% and 1.6% respectively.

Social science has second strongest within the field collaboration (72.51%). From collaboration with other fields, the strongest is with Technical science (10.02%), although it is more similar to Humanities with the second largest share of collaboration (7%). Humanistic science collaborates to a lesser extent with Natural and Medical science and even less with Biotechnical and Interdisciplinary science field. Technical science has the next highest percentage within field collaboration (69.36%) and some collaboration with Natural (13.68%), Medical (6.57%), Social (4.36%), Biotechnical (3.93%) and Humanistic (1.60%) science. Interesting is the fairly high percentage of collaboration with Medical science. Since the similarity score between Technical and Medical sciences is low in our similarity matrix (0.053), this type of collaboration is an important contribution to interdisciplinarity of the Technical science. Medical science has within field collaboration 68.3%, expectedly (similarity score 0.193) strongest collaboration with the Natural science (12.31%) and surprisingly (similarity score 0.053) high collaboration with Technical science (10.63%). Biotechnical science has 67.85% of within field collaboration and a high percentage of collaboration with Natural science (18.40%), which is the strongest between fields collaboration (without taking into account the Interdisciplinary field). The percentage of within field collaboration of Natural science is by far the lowest (56.05%) (except Interdisciplinary field), the collaboration is divided mostly on Technical (16.08%), Biotechnical (14.60%) and Medical (8.95%) sciences. Interdisciplinary field is the only one with the percentage of within field collaboration (5.96%) lower than collaboration with some other fields. Since the field involves a very small number of researchers, its percentage in other fields' collaboration is very low (under 1%). The researchers from Interdisciplinary science mostly collaborate with Natural (25.90%), Technical (23.02%), Biotechnical (21.69%) and Social sciences (11.31%) and somewhat less with Humanistic (5.55%) and Medical sciences

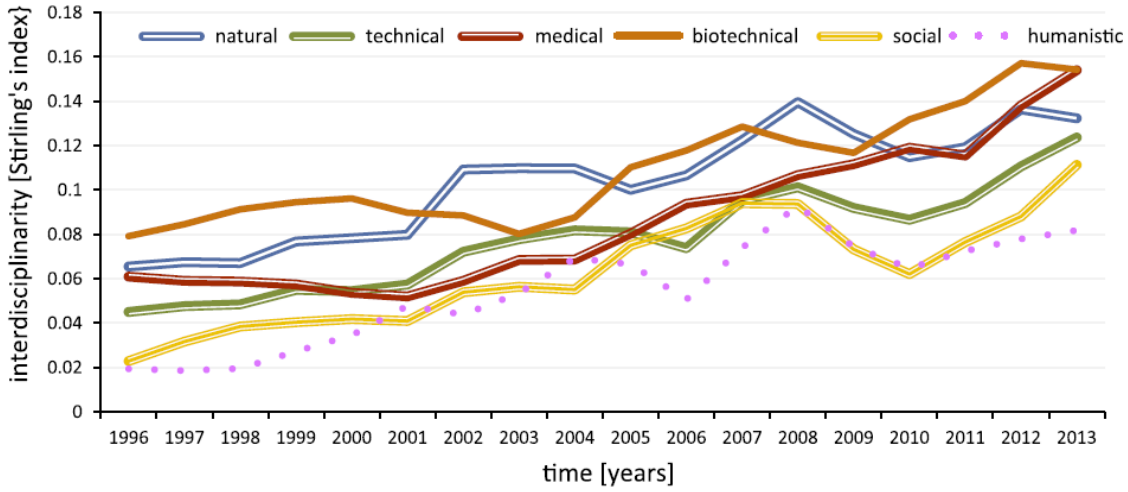


Figure 4.5: Interdisciplinarity of research fields from 1996 to 2013, based on collaboration on research projects

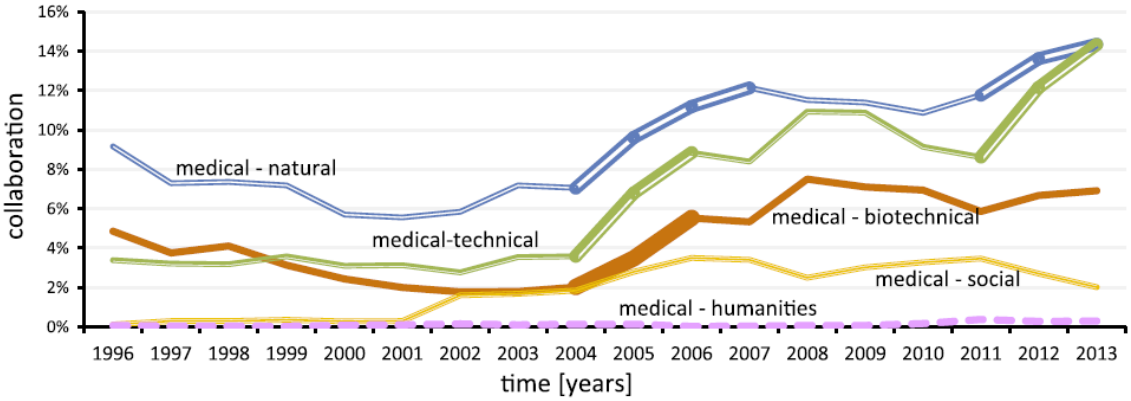


Figure 4.6: Collaboration of researchers from Medical science with researchers from other sciences over time (%).

(2.26%).

4.4.3 Interdisciplinarity based on co-authoring publications

Table 7 shows the results of interdisciplinarity, measured with Stirlings and Shannons diversity index, on the basis of co-authoring scientific publications. Probability density functions for the diversity are plotted in Figure 4.10. Like in the case of project collaboration ego-networks, we can notice that all the diagrams except the one for interdisciplinary field have class with lowest interdisciplinarity index (from 0 to 0.02) of much higher frequency than the rest of the classes. The most interdisciplinary science is Biotechnology, followed by Natural, Social, Humanistic, Medical and Technical sciences. The Interdisciplinary science field has to be considered separately, because it is a special field designed for Interdisciplinary research with a significantly smaller number of researchers. Similarly as in the case of projects, the Stirling's and Shannon's diversity indexes are highly correlated, with Pearson coefficient 0.99788. This is in accordance with the finding (Porter & Rafols, 2009) that Stirling's diversity is very closely associated with traditional diversity indicators (Shannon and Herfindahl). At this point we can notice that the two most interdisciplinary

sciences are the same as with the project collaboration and the interdisciplinarity of all sciences is growing with time (Figure 4.7). Detailed comparison of results obtained on the basis of project and publication collaboration is given in the next chapter.

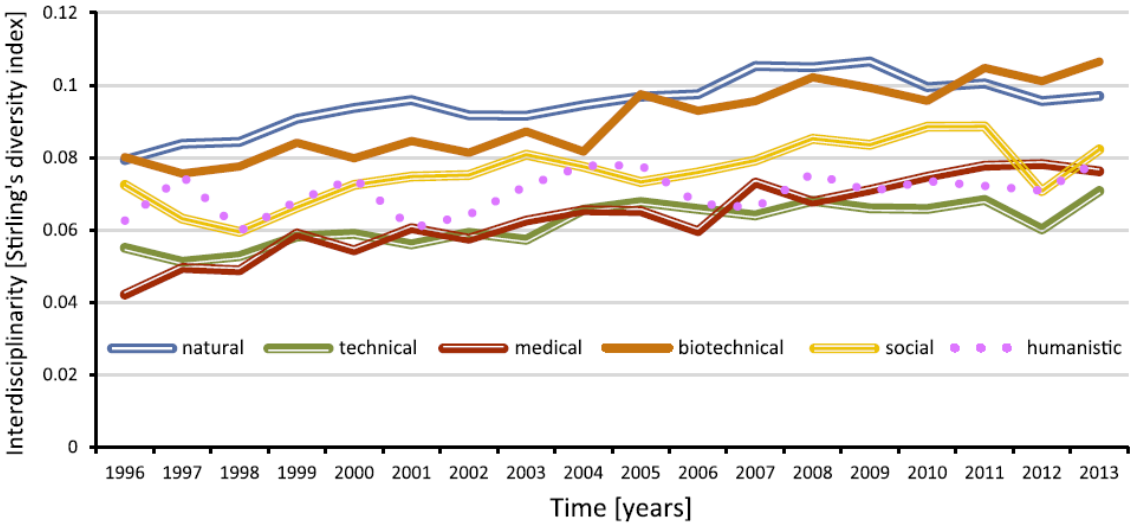


Figure 4.7: Interdisciplinarity of research fields from 1996 to 2013, based on collaboration on research publications.

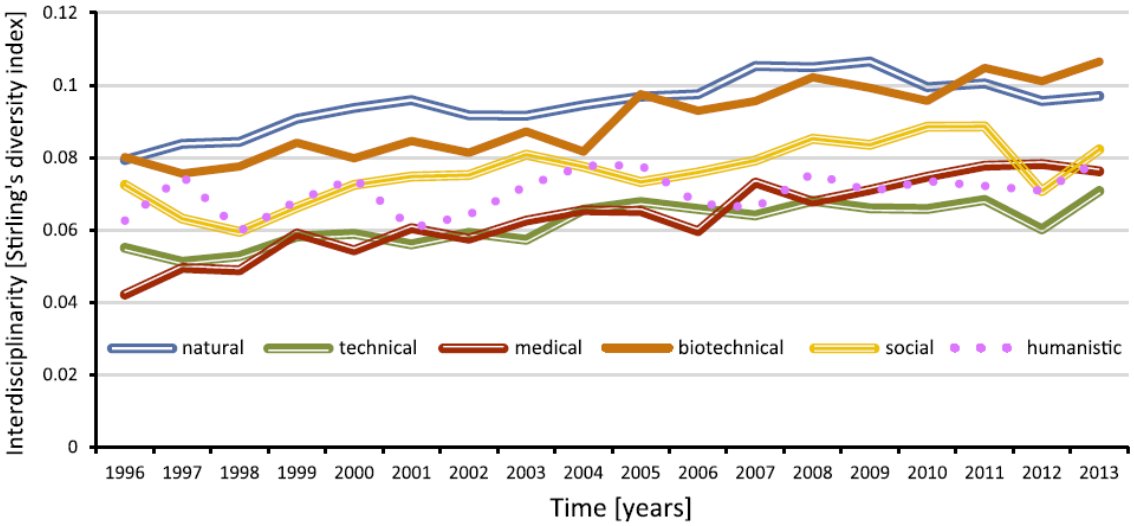


Figure 4.8: Interdisciplinarity of research fields from 1996 to 2013, based on co-authoring publications.

Table 8 shows the percentages of collaboration each science from a row has with the science from a column (plotted in Figure 9). The Medical science has the highest percent of within field collaboration (79.31%) and low collaboration with other fields, from which the most with Natural science (8.08%). Technical science has 73.02% of within collaboration, 12.72% of collaboration with the most related Natural science and some collaboration with Social (5.63%), Medical (4.06%), Biotechnical (2.42%) and Humanistic (1.66%) sciences. Humanistic science collaborates 71.32% within the field with noticeable collaboration with Social science (16.50%). Social science has slightly lower within field collaboration (69.21%)

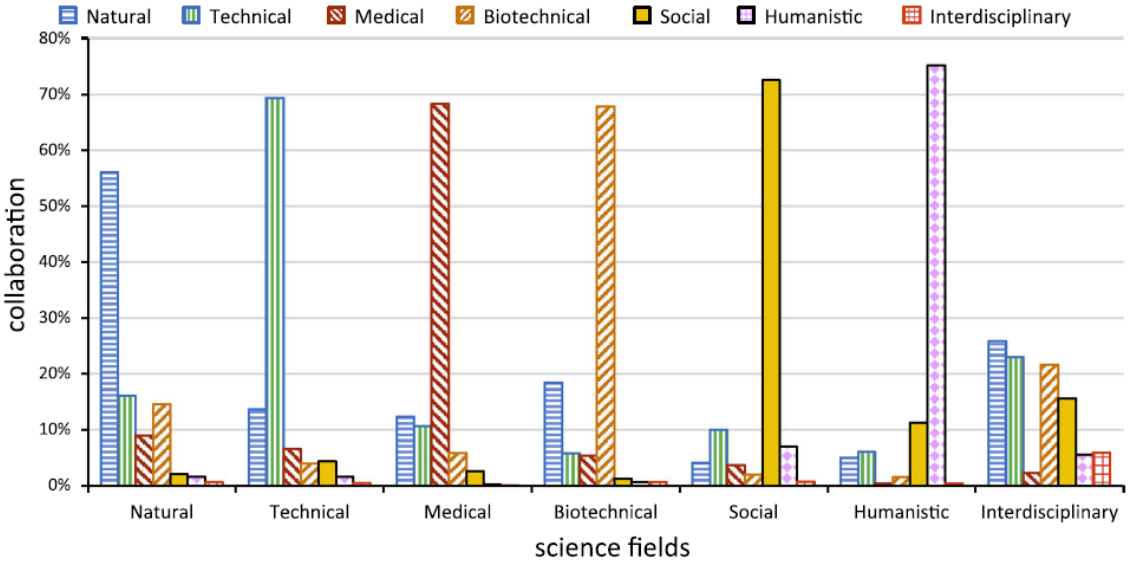


Figure 4.9: Share of research collaboration of each science with the other sciences based on research projects collaboration.

and has more balanced collaboration with other science fields: Humanistic (10.78%), Technical (8.86%), Medical (4.68%), Natural (4.71%) and Biotechnical (1.45%). Biotechnical science is the most interdisciplinary, but has the second lowest within field collaboration (65.47%), with strongest collaboration with Natural science (19.51%) and surprisingly (similarity score 0.053) high collaboration with Technical science (10.63%). Biotechnical science has 67.85% of within field collaboration and a high percentage of collaboration with Natural science.

4.4.4 Comparison of results based on project and publication collaboration

The most interdisciplinary science fields based on co-authoring publications are Biotechnologies and Natural sciences, which is the same as in the case of research collaboration on projects. However the collaboration of Social and Humanistic scientists is more interdisciplinary than the collaboration of Medical and Technical scientists, which is a different result than the one based on research projects. The Biotechnical science is the most and the Natural science is the second most interdisciplinary field in both cases – measured by projects collaboration and co-authoring. The results indicate that researchers behave differently regarding the interdisciplinarity of collaboration in the two research activities – projects and publications. However, there is some correlation between interdisciplinarity based on projects and publications. Measured on the 7 science fields the Pearson's correlation coefficient is 0.91, while measured on the 72 science subfields the Pearson's correlation coefficient is 0.49. According to interdisciplinarity based on projects and publications, we can cluster the 72 science subfields into 4 categories: 1) high projects ($>=0.2$) and publications ($>=0.2$) based interdisciplinarity, 2) high project ($>=0.2$) and low publications (<0.2) interdisciplinarity, 3) low projects (<0.2) and high publications ($>=0.2$) based interdisciplinarity and 4) low projects (<0.2) and publications (<0.2) based interdisciplinarity. Figure 10 depicts placing of the subfields on a two-dimensional plane, where one dimension is interdisciplinarity based on projects collaboration and the other one is interdisciplinarity based on co-authoring publications. We can see that if the border between high and low

Table 4.6: Percentage of research collaboration on projects between fields. The table shows how much percentage of research collaboration a field from row has with eachfield from the columns. E.g. in average 18.4%of collaborators of Medical science are from Natural science and 10.63% are from Technical science. The value on the diagonal shows the collaboration within a science, e.g., 75.14% for Humanistic. Bold values show the largest percentages of collaboration of science fields from rows with other science fields.

Fields	Nat	Tech	Med	Bio	Soc	Hum	Int
Nat	56.05	16.08	8.95	14.60	2.11	1.60	0.61
Tech	13.68	69.36	6.57	3.93	4.36	1.63	0.46
Med	12.31	10.63	68.33	5.89	2.58	0.18	0.07
Bio	18.40	5.82	5.40	67.85	1.26	0.63	0.658
Soc	4.12	10.02	3.67	1.95	72.51	7.00	0.72
Hum	5.06	6.07	0.42	1.58	11.31	75.14	0.41
Int	25.90	23.02	2.26	21.69	15.62	5.55	5.96

Table 4.7: Interdisciplinarity of science fields measured on the basis of co-authoring publications. Bold values indicate the largest values of mean diversity, variety, balance, disparity and Shannon's diversity.

Field	Diversity		Variety^2	Balance	Disparity	Shannon
	Mean	SD				
Natural	0.1404	0.0906	0.5369	0.1791	3.2059	0.6284
Technical	0.0973	0.0914	0.3558	0.1207	2.1295	0.4230
Medical	0.1100	0.0915	0.4822	0.1265	2.9243	0.4682
Biotechnical	0.1516	0.0864	0.6024	0.1851	3.6063	0.6624
Social	0.1327	0.0918	0.5130	0.1569	3.1536	0.5575
Humanities	0.1142	0.0920	0.4463	0.1395	2.7215	0.4961
Interdisciplinary	0.2357	0.0739	0.8707	0.2741	5.3290	0.9978

is set to 0.2, the clustering into four groups is quite balanced. Also, there is no obvious pattern for belonging to a group related to soft/hard, lab/office (Kronegger et al., 2012), or natural/technical science subfields types.

4.4.5 Evolution of interdisciplinarity

Evolution of interdisciplinarity observed by Mali et al. (2010) showed that the amount of coauthored publications in the field of sociology increased over the last two decades (1986-2005). We measured the mean number of authors per publication for all scientific disciplines and obtained the same result – increasing number of authors. Similarly, increasing trend is found with research projects. Both results are plotted in Figure 11. The mean number of authors per publications increases in linear trend function, which gives an average annual linear grow rate of 1.4%. The number of researchers in projects grows with liner trend function and has an annual linear growth rate of 3.5%. Measures of co-authorship were examined in Porter Rafols (2009) with findings of average growth in the period of 30 years about 75% (annual linear growth rate 2.59%), which ranged from 48% in Math, 54% in Physics to 90% in Neurosciences. We did not find any other work that measured collaboration of research projects over time, but our results indicate these main differences compared to co-authoring: the average and the variance of number of

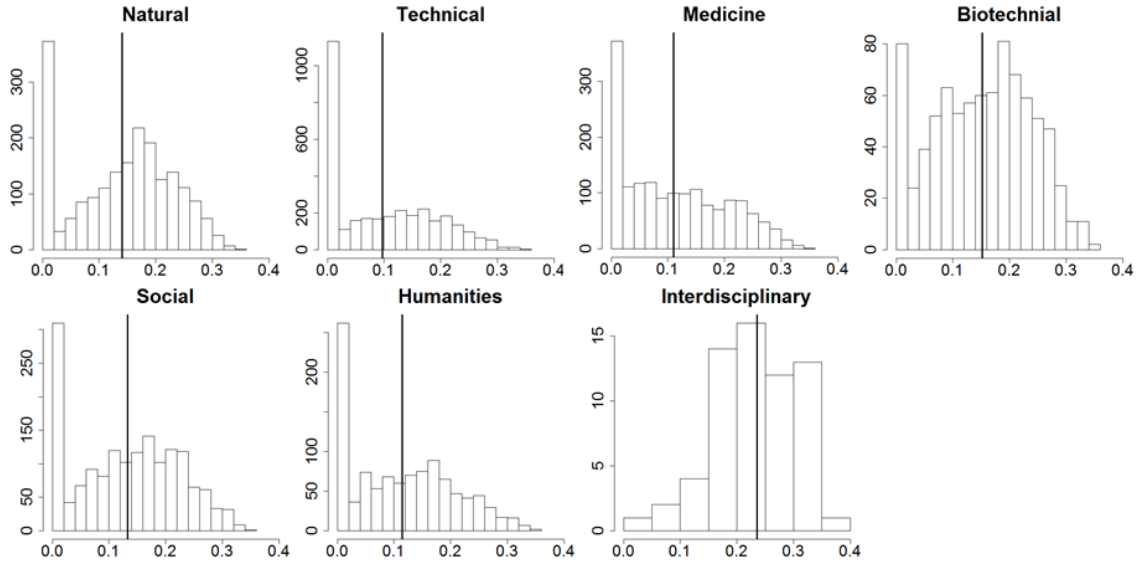


Figure 4.10: Probability density functions of ego networks diversity index based on co-authoring for science fields from Table 7. Diversity is on x-axis and frequency on y-axis. Intervals for x-axis are equal for each histogram (0–0.4), while intervals for y-axis differ depending on the size of the field. There are 20 classes of width 0.02 for each diagram except for Interdisciplinarity, which has 8 classes of size 0.05. Vertical lines indicate mean values.

researchers in projects is larger than on publications, which makes the growth much less stable than on publications. Furthermore the projects have faster growth of number of participants in time.

Both, collaboration and interdisciplinarity grow in time. The annual linear growth rate of number of collaborators in projects is 3.5%, while the annual linear growth rate of interdisciplinarity based on projects is 8.5%. The annual linear growth rate of number of authors per publication is 1.4%, while the interdisciplinarity based on publications grows with rate of 0.7%. This shows that interdisciplinarity of projects grows faster than the collaboration, while the interdisciplinarity of publications grows slower than the collaboration. Increase of collaboration does not influence the measure of interdisciplinarity (Equation 1), because only the proportional representations of elements of disciplines in the system are taken into account. With fixed distribution of field sizes, the expected value of the proportional representations of elements stays the same and does not depend on the sizes of observed ego networks.

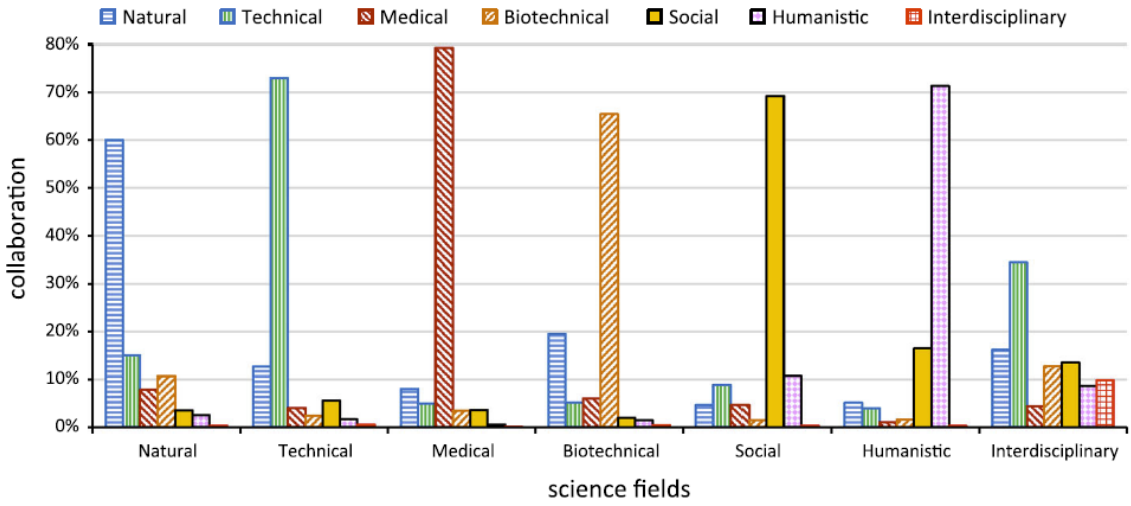


Figure 4.11: Share of research collaboration of each science with the other sciences based on co-authoring publications.

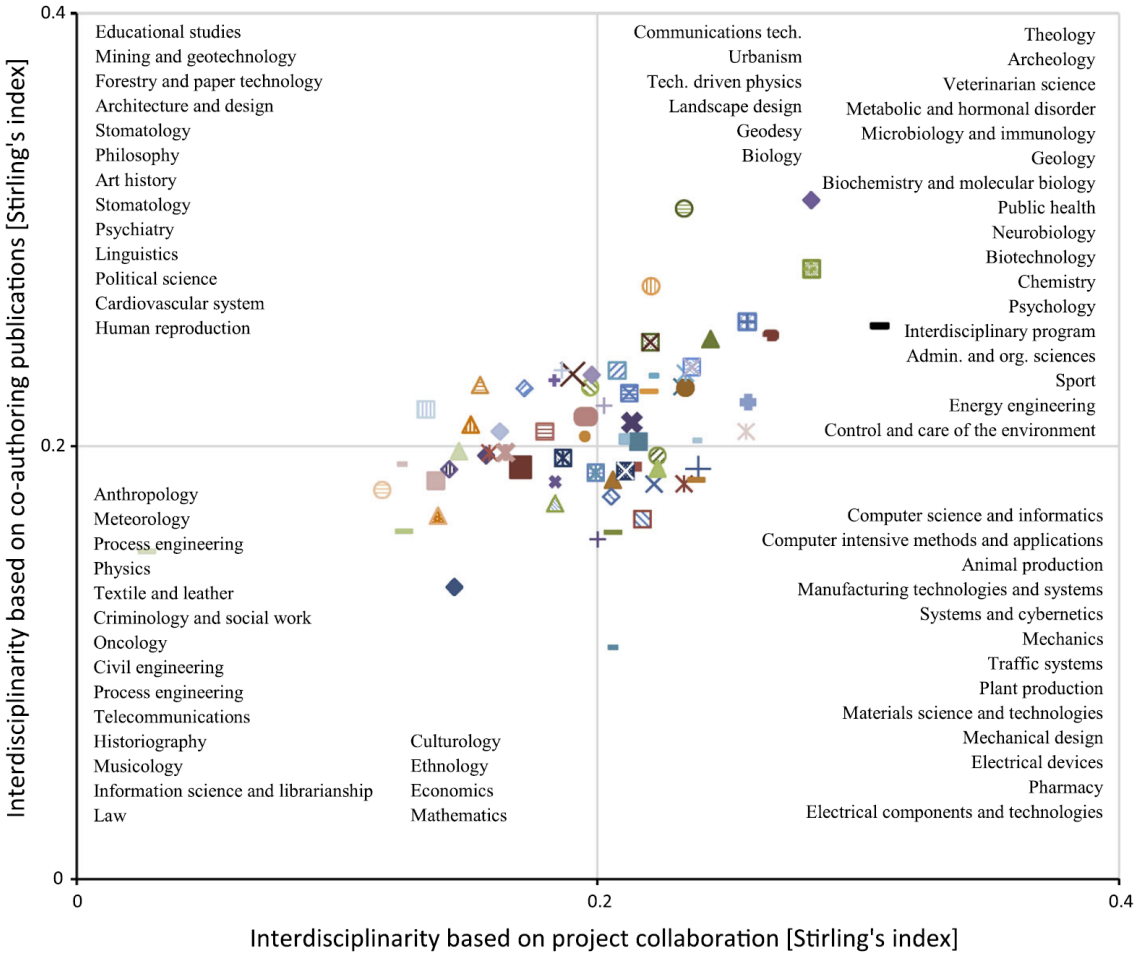


Figure 4.12: Interdisciplinarity of science subfields based on projects and publication collaboration. Symbols in each quadrant represent the subfields in the corresponding quadrant. The subfields are ordered from more to less interdisciplinary. For example, Communication technologies represented with a rhombus in the upper right corner have high interdisciplinarity according to both projects and publications. On the other hand rhombus on the lower left corner represents Mathematics with low interdisciplinarity.

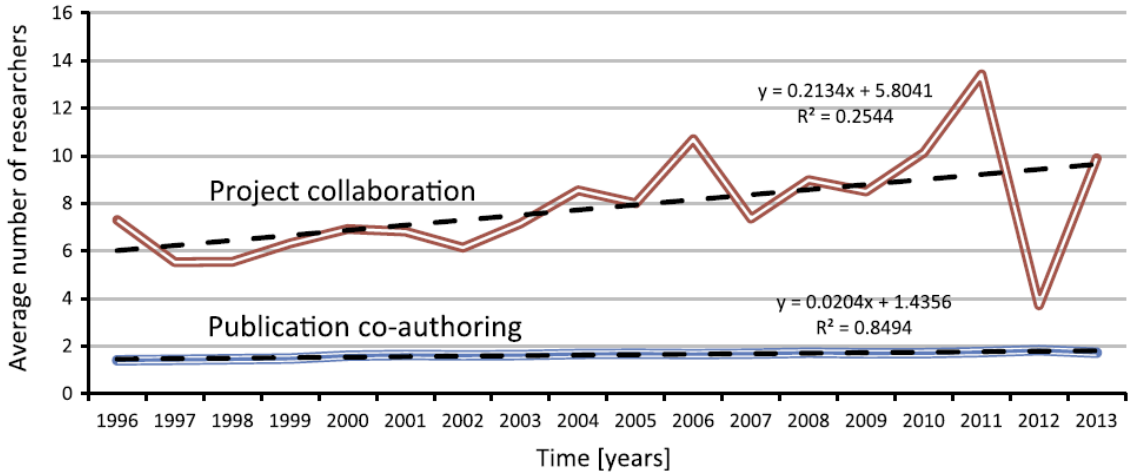


Figure 4.13: Average (mean) number of researchers peer project and publications from 1996 to 2013.

Chapter 5

Conceptualization of Science Using Collaboration and Competences

Text and network analysis methods and different visualization techniques can be used to conceptualize a social scientific community. Science, engineering, technology, and innovation are the bedrock of a successful economy, particularly as nations are moving toward knowledge-based economies (Enderby, 2004). Measured by the number of books (Lebert, 2005), papers and journals (May, 1997; Derek John, 1975; Cronin & Atkins Barsky, 2000) and patents (WIPO, 2007; eMedia Asia Ltd, 2007), science marks remarkable growth. The same trend is with the number of researchers and engineers, whose collaboration is ever more important. In these settings, it is important to be able to explore the globally connected and rapidly evolving scientific community, whose results are crucial for further development of society. Conceptualizing means forming an idea, picture, etc., of something in your mind (Merriam-Webster Online, 2009). In their discussion on Visual Conceptualizations and Models of Science Börner and Scharnhorst (2009) define conceptualization as unifying mental framework that identifies the boundaries of the system or object under study, its basic building blocks, interactions among building blocks, basic mechanisms of growth and change, and existing laws (static and dynamic).

Following the existing theories and knowledge of the basic units of science and their interrelations and dynamics, we are proposing an approach for modeling and analyzing research community based on collaboration and emerging competences. The proposed approach defines the basic unit of science and the links between them and produces collaboration and competence models, enabling network and topic analysis. We base our study around the following research questions:

1. Can conceptualization of science based on collaboration reveal important facts such as the most researchers in a topic, nodes and links that connect the community, or clusters and interdisciplinarity of the community?

2. Can conceptualization of science based on competences reveal the actual research interests of researchers? Where "actual research interest" refers to research topics derived from results of researcher's work (scientific publications and research projects).

In the rest of this chapter we show that the current research insufficiently addresses raised questions and we give contribution with the proposed approach. The chapter is organized as follows. First in section 5.1 we introduce some fundamentals and recent related work on conceptualizing, analyzing and visualizing research results. Next, in Section 5.2 we explain the approach in more detail. Section 5.3 provides a detail description of data used for the use case. By using use cases from four different categories (humanities, technical

sciences, natural sciences, general topic) we show how an analysis of scientific collaboration and competences can be performed (Section 5.4). Finally, we report results (Section 5.5) conclusions and some ideas for future work are provided in Section 5.6.

This research work was submitted and accepted for publication in Electronic Library Journal. The work was conducted in collaboration with Dunja Mladenić, Marko Grobelnik and Mitja Jermol.

5.1 Related Work

Social network theory is not a single theory, but a perspective or a frame that can be used to develop specific theories (Knoke, 1982). Social networks take prominence in sociology, network sciences, complex network analysis and other academic disciplines (Pescosolido, 2006). Topics like six degrees of separation first researched by Milgram (1967), social capital (Williams & Durrance, 2008) and link prediction (Liben-Nowell & Kleinberg, 2007) have been adopted in many different fields. The theory is interdisciplinary in nature, so the underlying terms and concepts can produce both acceptance and confusion in academic and community circles (Pescosolido, 2006). Most simply stated, a social network is a "structure of relationships linking social actors" (Marsden, 2000) or "a set of actors and ties among them" (Wasserman & Faust, 1994). Social network can be observed as the essence of social structure (Burt, 1980), or mechanisms that connect micro and macro levels of social life (Coleman, 1990).

An early work on modeling intellectual structure of scientific specialties performed by Small (1973) used co-citation networks from 1971 Science Citation Index data. It was assumed that frequently cited papers represent the key concepts, methods, or experiments in a field and that those co-citation patters can be used to map in great detail the relationship between those key ideas. Garfield's ISI Atlas of Science: Biochemistry and Molecular Biology 1979-1980 (Garfield & Shea, 1981) has 102 research specialties within the fields of biochemistry and molecular biology. The specialties were identified using the process of co-citation clustering. For each of the 102 specialties there is a short essay and a map, in which the cited documents are arranged graphically. Marshakova (1981) analyzed citations in the field of information science using 13 volumes of journal Nauchno-tekhnicheskaya Informatsiya and revealed the topic structure, the communities of authors and the names of single leading scientists. In work of Small and Greenlee (1986) authors used methodology of co-citation cluster strings which identifies coherent areas of scientific research. Recently Zhu and Guan (2013) performed complex network analysis on a specific topic - service innovation research. They analyzed 437 papers that were obtained from Web of Science database. Three types of network were generated: keyword co-occurrence network, researcher collaboration network and subject category co-occurrence network, which covered three important aspects of a scientific field: academicians, hot topics, and corresponding subject category. Similarly to the work of Zhu and Guan (2013), our approach implemented in Science Atlas generates researcher collaboration networks (later called collaboration graph) and keyword co-occurrence network (later called competence map). While Zhu and Guan (2013) manually preprocessed the data for one chosen topic, Science Atlas can automatically generate sub network for an arbitrary topic as long as there is some record in the data. The goal of the Science Atlas is to enable generation of arbitrary subgraphs and offer possibility of on-the-fly exploratory analysis by visual inspection and interaction with the graph.

There is a lot of work done in using keyword co-occurrence analysis for presenting competences in many different fields which is related to the competence map of Science Atlas. In the field of information retrieval Ding, Chowdhury, and Foo (2001) have applied

co-word analysis to trace concept or idea derivation during the period of 1987–1997. Other examples are: bibliometric analysis of keywords in the literature on ethics and dementia during the period 1980-2000 (Baldwin, Hughes, Hope, Jacoby, & Ziebland, 2003), scientific output of the field of geographic information system by using keywords (Tian, Wen, & Hong, 2008), software engineering (Coulter, Monarch, & Konda, 1998), chemistry (Callon, Courtial, & Laville, 1991), scientometrics (Courtial, 1994), neural network research (van Raan & Tijssen, 1993) and biotechnology (Rip & Courtial, 1984). Since in our approach research projects represent the main research activity of research community, we use keyword co-occurrence in research projects as the method for identifying the most important keywords of a research topic or a researcher. This method is a part of competence map which is covered in a later section. Pajek is a desktop application for analysis of large networks (Batagelj & Mrvar, 2004). It provides the user with some visualization tools and implementation of a selection of efficient algorithms for analysis of large networks. ArnetMiner (Li et al., 2008) is an online expert oriented search system. It has an advanced pipeline that includes Knowledge Acquisition and Semantic Integration for crawling data. This makes the main difference compared to the proposed approach in Science Atlas, which uses clean data from national agencies responsible for recording research data. Because of this, Science Atlas can also be used as a highly credible data source. On the other hand, ArnetMiner is broader and more flexible from the aspect of data sources, but it can be less reliable because automatic knowledge acquisition can be vulnerable to errors. In work from Zhuhadar, Yang, and Nasraoui (2012) authors evaluate three types of force directed methods for discovering communities in the network structure: (i) Fruchterman Reingold method (Fruchterman & Reingold, 1991), (ii) Force Atlas 2 (Jacomy, Venturini, Heymann, & Bastian, 2014) method and (iii) Yifan Hu method (Hu, 2005). The clusters of different layouts which could be considered as another representation of community structure were compared to a modified version of the modularity measure proposed by Blondel, Guillaume, Lambiotte, and Lefebvre (2008). In the experimental study (Blondel et al., 2008) Fore Atlas 2 obtained the best modularity measure and was used as the method for detecting communities of users in the Social Web network. The same algorithm was used for layout of the all-to-all collaboration graph of our Science Atlas. Science Atlas covers research output in Slovenia and in this respect it is related to the work from Perc (2010) on growth and structure of Slovenia's scientific collaboration network, which shows that the network matches the characteristics of small world networks. This was shown with the growth of the network that follows near linear preferential attachment and evolution of clustering coefficient and diameter of the network. Collaboration structures in Slovenian scientific communities were analyzed by Kronegger et al. (2012). Longitudinal co-authorship networks of four research disciplines (biotechnology, mathematics, physics and sociology) for 1986-2005 were compared. Different co-authoring cultures and their evolution were identified, and using blockmodeling it was shown how co-authorship structures change in all disciplines. Work on co-authorship trends and collaboration patterns in the Slovenian sociological community by Mali et al. (2010) shows that the tendency to co-author in this field is increasing and compares the quality of publications in relation to structural position in the network. The proposed approach is similar in that it enables collaboration network analysis. It does not include temporal analysis, but enables analysis of sub networks which can be generated using different criteria. In our study we tackle the insufficiency of existing research in using both collaboration and competences for conceptualizing science and use the information on research projects alongside publications. We develop a publicly available system that enables exploring the complete research community of a country.

5.2 Approach

We propose an approach for creating descriptive models of science to illustrate two major features of datasets - collaboration and competences. Our approach starts by defining basic units of science and links between them. Next, various algorithms are applied to model the data and enable network and topic analysis. Finally, data and the results of modeling are communicated in visualization layers.

5.2.1 Basic units

The atomic unit of our approach is a researcher (considered identical to author or scientist). Researcher is a person who publishes scientific papers, books and journals, produces patents, data sets, hardware, software or works on research projects (Börner, 2010). Basic unit of analysis can be any aggregation of the atomic unit, e.g. organization, project group, scientific field, country, etc. In this case the set of attributes of a basic unit is the union of attributes of all atomic units. The attributes assigned to the basic unit can be used for two tasks: linkages and networks derivation. Linkages are units that shape the connections between basic units. Linking unit can be any result of research work of the basic unit. Network derivation is a task of selecting a subset of basic units. Units can be selected if they have the same or similar enough value of an attribute.

5.2.2 Analysis and data models

The core of our approach is modeling the collaboration and competences. The inputs to collaboration model are the type of collaboration (egocentric or derived) and the selected node (in egocentric type) or selection input argument (in derived type of collaboration). Output is a set of nodes (basic units) and a set of weighted edges connecting the nodes. We name the outputs egocentric and all-to-all collaboration diagrams. Collaboration diagram is a type of sociogram that visualizes collaboration between people. Sociograms serve as simple visual illustrations in helping people to make sense of the underlying network information (Card, Mackinlay, & Shneiderman, 1999). Egocentric type of collaboration diagram is always focused on one researcher, while the others are displayed according to the relationship with the selected one. All-to-all type of collaboration diagram shows the connections between all the nodes in the diagram. In other words, if a collaboration criterion between any two pair of nodes in the diagram is satisfied, the nodes will be connected with an edge. We use the term competence map for graphs that are based on keywords and show competences of a researcher or competences related to a research field. Competence map is the implementation of competence model proposed by our approach and applied in the Science Atlas application. The graph connects research projects based on common keywords. The keywords that connect the most projects are considered the most important and are emphasized as the main keywords that show competences. Process of determining common keywords is implemented with lemmatization (Kanis & Skorkovská, 2010) and calculating Levenshtein distance (Levenshtein, 1966) between words. Lemmatization solves the problem of essentially equal keywords written in different grammatical form. For example one project can have a keyword "semantic technology", while another one can have a keyword "semantic technologies", by applying lemmatization these two words are identified as identical and the two projects will be connected. Another problem with keywords is that they can be misspelled (e.g. "artificial inteligence" compared to "artificial intelligence"), or written using different language conventions (e.g., "visualisation" compared to "visualization"). This is solved using Levenshtein distance, which is a string metric that counts the minimum number of single-character edits required to change one

word into the other. Keywords with a small arbitrary chosen edit distance were considered identical. After implementation in visualization layer the model enables network analysis - identifying important nodes and edges, observing connectivity, communities and diversity of network. The competence model similarly to collaboration model can be egocentric or derived. The output in this case is a bipartite graph where one set of nodes represent linkages (e.g. publications, patents, projects, etc.), while the second set represent keywords related to linkages. The second part of the model is a list of the most important keywords relevant to the selected subset of nodes. The selected keywords are the most co-occurring keywords from the output graph.

5.2.3 Visualization layer

Visualization layer generates highly interactive maps that support users in hypothesis generation, data exploration and decision making (MacEachren, 2004). Since output of collaboration models can be too vast to be understood at once, interaction via zooming and panning, exploration and access to details via search and selection is important. Different colors of the nodes represent different values of a typically categorical attributes, while the sizes correspond to number of connections a node has. All-to-all collaboration diagram applies Force Atlas 2 algorithm for the layout (Jacomy et al., 2014), which follows the first law of geography given by Waldo Tobler: "Everything is related to everything else, but closer things are more closely related" (Börner, 2010). The algorithm was developed with considering the transparence demand. As long as it runs, the nodes repulse each other like magnets with the same poles, while the edges attract their nodes like springs. The final configuration is expected to help the interpretation of the data by showing clusters of tightly connected researchers, central individuals and hubs. The layout used to construct the egocentric collaboration diagram is circular (Karlovčec, Mladenić, Grobelnik, & Jermol, 2011). The selected researcher is placed in the middle of the diagram. The researchers with whom the selected researcher is collaborating are placed around him in concentric circles, with the radius inversely proportional to the strength of collaboration. This results in the placement of the researchers who are more important for the collaboration to the center of the diagram.

5.3 Data

We have created a publicly available web application called Science Atlas [1] that implements the proposed approach on Slovenian scientific data. Slovenia is a smaller sized country in southern Central Europe with population around 2 million. It was chosen as the use case for the approach, because it has a well organized system of recording scientific results and collaborations. The proposed approach can be applied on other scientific datasets. Here it is applied on Slovenian dataset, demonstrating collaboration and competences analysis on a complete national science. The developed application can be useful for policy making, social science researchers, experts from different fields of industry and others. Actually, we already got some very positive feedback from Slovenian National Research Agency and a number of researchers. Science Atlas is based on SICRIS (Slovenian Current Research Information System) database of national research projects and COBISS (Co-operative Online Bibliographic System) national database of scientific publications. SICRIS and COBISS are managed by the Slovenian National Research Agency and IZUM (Institute of Information Science in Maribor). The SICRIS database contains data about national research projects registered from 1994 till present, and it is being updated on a monthly

[1] http://scienceatlas.ijs.si

basis. The main entities of the SICRIS data model are: research project, research program, researcher, organization and research group (attributes of the main entities are given in Table 5.1). The main entities are connected with the following relations: researchers – projects/program (researchers working on a project/program), researchers – organizations (researchers employed in an organization) and researchers – research group (researchers are part of a research group). SICRIS database is bilingual which enables creating all the services of Science Atlas application in both English and Slovene language. COBISS is a national database that contains data about scientific publications in Slovenia, with the main attributes: title of publication, authors, abstract and year of publication.

Table 5.1: Attributes of main entities in SICRIS database.

Entity	Attributes
research project	status data, classification, keywords, abstract, significance, report, organizations, researchers, equipment
researcher	status data, contact data, classification, keywords, video, knowledge of foreign languages, education, employment, projects, programs
organization	status data, contact data, classification, research groups, researchers, projects, programs, research equipment
research group	status, contact, classification, researchers

Science Atlas database is being continuously updated. Currently the database contains 36046 researchers, 5855 projects and 974 organizations. The textual attributes of the entities are not fully filled - only 9825 (27%) researchers and 3056 (52%) projects contains some keywords, and 3058 (52%) projects contains some abstract. Text analysis on keywords and abstracts of projects in English identified 14192 distinct terms (analysis preceded standard text preprocessing steps - stemming and stop-words removal). The 10 most frequent terms were: slovenia, research, projects, result, study, method, model, product, system, process and the most co-occurring pair of words (frequency of word above 780 and correlation above 0.40) were: project-research, project-slovenia, research-develop and research-result. The most frequent terms and associations between them give a nice summary of the complete database (Kastrin & Karlovčec, 2012). Since the foundations of Science Atlas are official data sources of national scientific activity, the developed application can be a credible atlas of research work in Slovenia. The methods used here can be applied to other bibliographic data.

5.4 Use Cases

In this section we show results of using Science Atlas for exploring scientific community in Slovenia. We perform four use cases on collaboration analysis from different science fields: anthropology (humanities), artificial intelligence (engineering), nanotechnology (natural sciences) and Slovenia (general topic). In addition, competences related to topics and competences of individuals are analyzed.

5.4.1 Humanities use-case

Search with the term anthropology returns 148 researchers. The collaboration graph constructed with this search term is shown in Figure 5.1. The graph reveals that the most of the researchers are from the field of humanities science group which is represented with pink color. The second most common science group is social science, while there are few researchers from engineering and medical sciences and one from interdisciplinary science. The structure of the graph emphasizes a large strongly connected cluster of researchers from humanities (upper right corner in Figure 5.1. Besides the large cluster, several smaller clusters are present. With exception of a smaller humanities cluster (upper left corner in Figure 5.1), it seems the clusters are connected to each other in one to six hops.

5.4.1.1 Bridges and cut-vertices

The collaboration graph reveals several brokers and bridges. Figure 3 marks some bridges (1, 2 and 3) and cut-vertices (A, B, C,..., I) which are immediately revealed thanks to the layout of the graph.

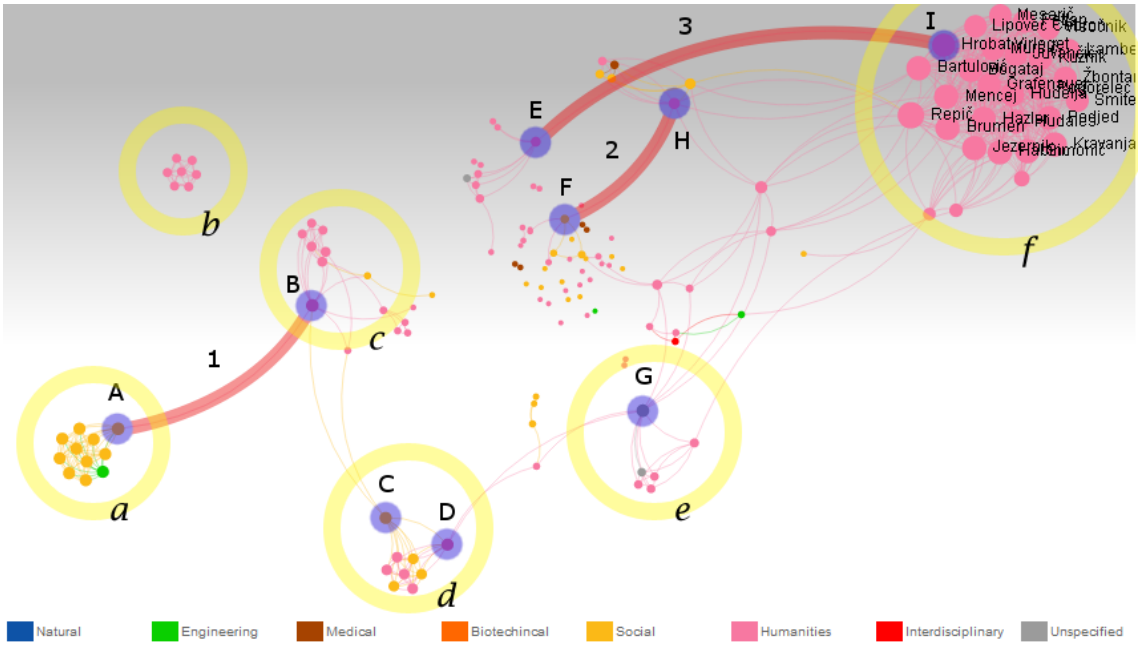


Figure 5.1: Collaboration graph for topic anthropology with manualy marked bridges (1,2 and 3), cut-vertices (A,B,..., I) and clusters (a,b,..., f).

For example, removal of the broker G from the graph would causes disconnecting of a large part (that includes brokers A, B, C and D) of the graph. Therefore the researcher represented as the broker G has an important role in connecting the research community relevant for the topic of anthropology. The edge 1 represents collaboration between two researchers – brokers A and B. From the social network analysis point of view, this is an important collaboration because it keeps a cluster on the left side (that includes broker A) connected with the rest of the graph. Furthermore this connection is even more important because it connects a cluster of researchers from the different science groups from the majority in the anthropology community – the cluster on the left is constituted of social sciences researchers and a researcher from engineering sciences.

5.4.1.2 Science groups collaboration diversity

Social sciences are the second most present research group on the graph. Very noticeable is the cluster on the lower left part of the graph (containing the broker A in Figure 5.1), constructed from social scientists and a researcher from engineering science. Using Science Atlas, we try to analyze this part of the anthropology graph in order to reveal the grounds for collaboration. Since it is interesting to see a researcher from engineering sciences in the anthropology graph, we obtain the profile page of the researcher by clicked on the node. The profile page reveals that the researcher belongs to engineering science and technologies group, field of civil engineering, with keywords: building structures and seismical engineering. Among other information, the profile page lists a project Interdisciplinary Center of Vernacular Architecture EU in Ljubljana: programme, contents, teheoretical frame of establishing. Clicking on the project reveals details that include the following keywords: architecture, heritage, culture, info system, preserving, linguistic, anthropology and human. By reading the abstract of the project, it is clear that it is an interdisciplinary project combining architecture as technical science with elements of culture and heritage. Generating collaboration graph of the projects results in the collaboration graph shown in Figure 5.2.

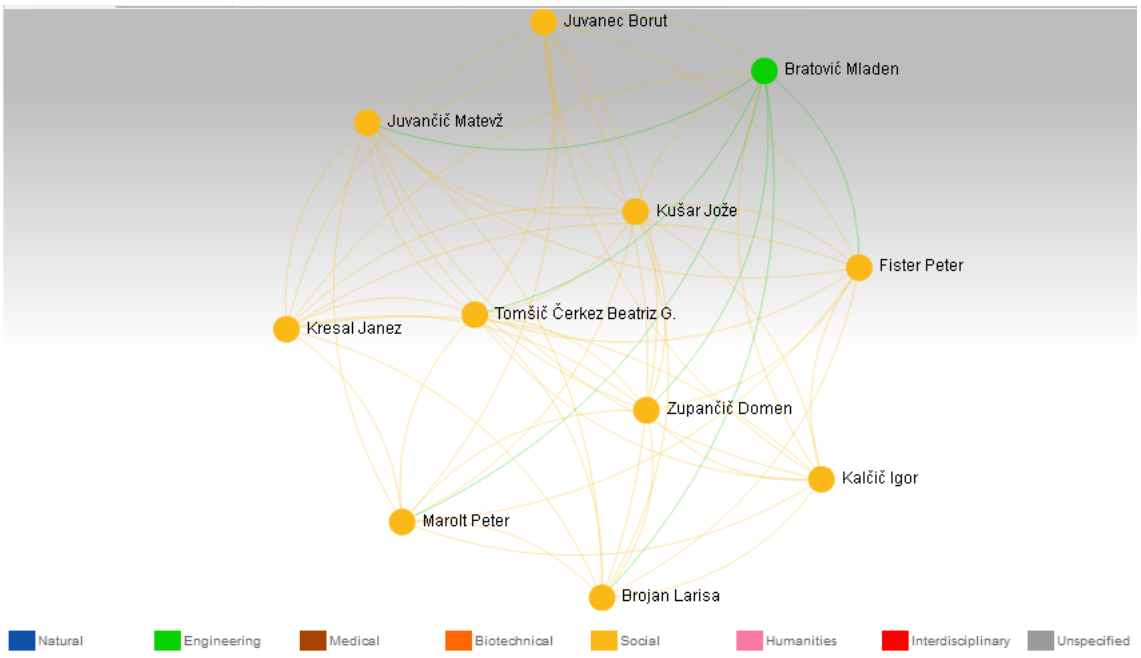


Figure 5.2: Collaboration graph of an interdisciplinary architecture project.

The graph of this project is identical to the cluster in the lower left part in Figure 2 (cluster a) representing the anthropology graph. Now it is clear that the group of social scientists and an engineer are included into the anthropology graph because of collaboration on a interdisciplinary research project that combines elements of architecture, urbanism and civil engineering. To find out what makes this cluster connected to the reset of the anthropology graph, we have to investigate the bridge 1, i.e. brokers A and B from marked in 5.2. It is enough to click on either of the nodes A or B, which will open the profile page of the researcher. Profile page lists all the collaborators and the project on which the collaboration is based on. This shows that the collaboration between brokers A and B (Figure 5.2), which connects the whole cluster of researchers with the rest of the graph is based on the project Architecture of school space in the function of hidden curriculum.

5.4.1.3 Network analysis

We have performed network analysis with Pajek in order to compare the exact results of network analysis with insight obtained with observation of visualizations in Science Atlas. The anthropology graph (Figure 5.1) suggests that most of the researchers are connected to each other. Exact analysis reveals that there are 8 connected components. Most of the researchers (103 or 70%) belong to a largest connected component. Second largest connected component contains 7 researchers (upper right part in Figure 2). There are another 6 connected components containing only 2 researchers, while 26 researchers (18%) are isolated nodes without any connection with other researchers from the graph. Existence of a large connected component containing the majority of researchers proves good connectivity observation. In Figure 5.1 9 most obvious cut-vertices are marked – researchers whose deletion would cause disconnecting of a connected component into at least two new components. Analysis with Pajek revealed another 6 cut-vertices, some of which would generate connected components of smaller size (3 or less researchers). Force Atlas 2 layout algorithm used for all-to-all type collaboration graph in Science Atlas reveals some clusters of researchers. We use the term cluster to refer to a group of nodes tightly connected to each other and weekly connected or disconnected with the rest of the graph. In work of Flake, Lawrence, Giles, and Coetzee (2002) authors use similar definition for community of webpages - collections in which each member page has more hyperlinks within the community than outside the community. Clusters defined in this way are related to the concept of bi-components from the network analysis literature. A bi-component is simply a component (a maximal connected subnetwork) without a cut-vertex (de Nooy, Mrvar, & Batagelj, 2005). In Figure 5.1 2 we can see at least 6 such clusters, which are market with circles (from a to f). Bi-components were identified using the Pajek software. Experiments with minimum sizes of bi-component set to 3, 4, 5 and 6 were made, which resulted in 10, 7, 6 and 5 identified components respectively. 6 bi-components identified with the minimum size of bi-component set to 5 correspond very closely to the clusters marked in Figure 3.

5.4.1.4 Using other services

We will try to get information on anthropology scientific community in Slovenia using other available tools and methods – SICRIS and ArnetMiner. The goal is to test how much new value and information can be obtained with Science Atlas. Science Atlas is based on publicly available data that is available on SICRIS web portal. SICRIS enables separate search for researchers, organizations, project groups, projects and programs. The search for researchers with the term “anthropology” returns 76 researchers. The reason why it returns fewer results than Science Atlas is because it takes into consideration only the keywords of researchers and not the keywords of projects when searching for researchers. SICRIS enables separate search for projects, which in the case of anthropology returns 29 projects. Although SICRIS is a stable and reliable source it does not have the ability to show the connecting relations between entities (e.g. researchers and projects). Capturing the relations between concepts and creating visualizations on top of them is the added value of Science Atlas. ArnetMiner contains Expert Finding application that helps with finding experts on a given topic. We entered “Anthropology” as the search term and retrieved 275 experts. The retrieved experts did not include the Slovenian researchers from the field of anthropology. Moreover, unlike Science Atlas, ArnetMiner does not enable generation of collaboration graph based on the researchers retrieved with some keyword.

5.4.2 Technical sciences use-case

Search with the term "artificial intelligence" returns 108 researchers. The collaboration graph constructed with this search term is shown in Figure 5.3. The graph shows that the most of the researchers are from the field of engineering science group which is represented with green color. The colors of nodes suggest that the second most common science groups are social sciences and natural sciences, while there are few researchers from humanities and medical sciences. A large connected component is dominant on the graph with two smaller isolated clusters (upper part in Figure 5.3). It is also interesting how social sciences, natural sciences and humanities occupy different edges of the largest cluster, while the engineering science holds the core. Since artificial intelligence is both an engineering and mathematics topic, while it is less obviously a humanistic science topic, we examined in more detail the node from humanities located in the middle of the largest cluster of the graph. Since the graph is interactive, this is very easy, the profile page of the researcher can be retrieved by clicking on it on the graph. This reveals that the humanities researcher is from the field of linguistics and that he is described with keywords: computational linguistics and language technologies. Since computational linguistics is a part of artificial intelligence, it is now clear why this researcher was a part of the graph. Looking at the list of projects of the researcher reveals many projects that are closely related to computer sciences, which made him strongly connected in the dominantly engineering cluster.

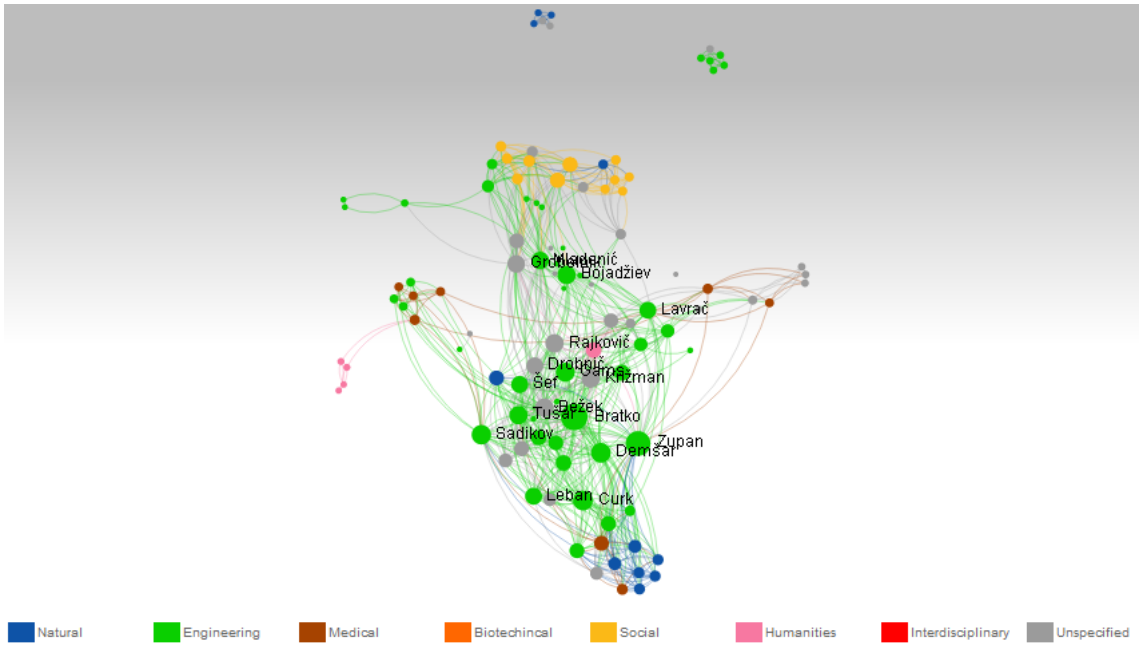


Figure 5.3: Collaboration graph for topic artificial intelligence.

5.4.3 Natural sciences use-case

Search with the term "nanotechnology" returns 53 researchers. The collaboration graph constructed with this search term is shown in Figure 5.4. The graph contains roughly the same number of researchers from natural and engineering science group. There are two connected components, one containing most of the researchers and another smaller one with 5 researchers from engineering and one with undefined category. The largest connected component contains three clusters. The largest cluster (lower middle part of

Figure 5.4) is connected with the cluster on the right side via a cut-vertex from natural sciences. The largest and the upper cluster are connected via two vertices, one from natural sciences and one with undefined category. These two vertices are not cut-vertices, because in case of removing one of them, the component stays connected with the other one. Nevertheless, both vertices are very important for the connectivity of the component, because they have almost equally many connections to both cluster which are otherwise not very connected between each other. It is interesting that although these two nodes have a very similar structure of connections, they are not directly connected with each other. We can conclude that the mentioned nodes present very important researchers from the aspect of connectivity in nanotechnology research community.

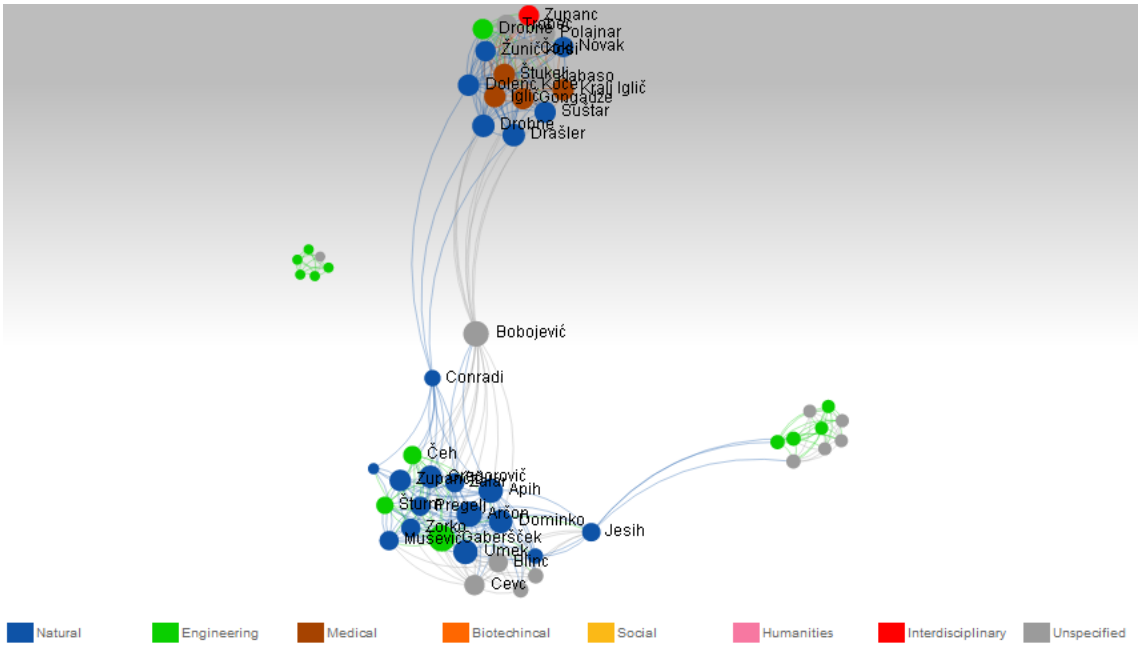


Figure 5.4: Collaboration graph for topic nanotechnology.

5.4.4 General topic use case

Search with the term "Slovenia" returns 672 researchers. The collaboration graph constructed with this search term is shown in Figure 5.5. The collaboration graph shown in Figure 5.5 is different from the rest of the collaboration graphs in two ways – the topic is general and the number of researchers is notably greater. The graph contains a single large connected component. All seven science groups are present in the graph, but the largest portions occupy humanities and natural sciences. The portion of biotechnical sciences is placed inside of natural sciences, which makes sense since these sciences are related. The similar symbiosis can be noticed with social sciences and humanities. Medical sciences occupy their own portion of the graph which is on upper right side in Figure 5.5, while engineering is interlaced with natural sciences. The only researcher from interdisciplinary sciences is connected with a large number of researchers from humanities and natural sciences and some from biotechnologies and social sciences.

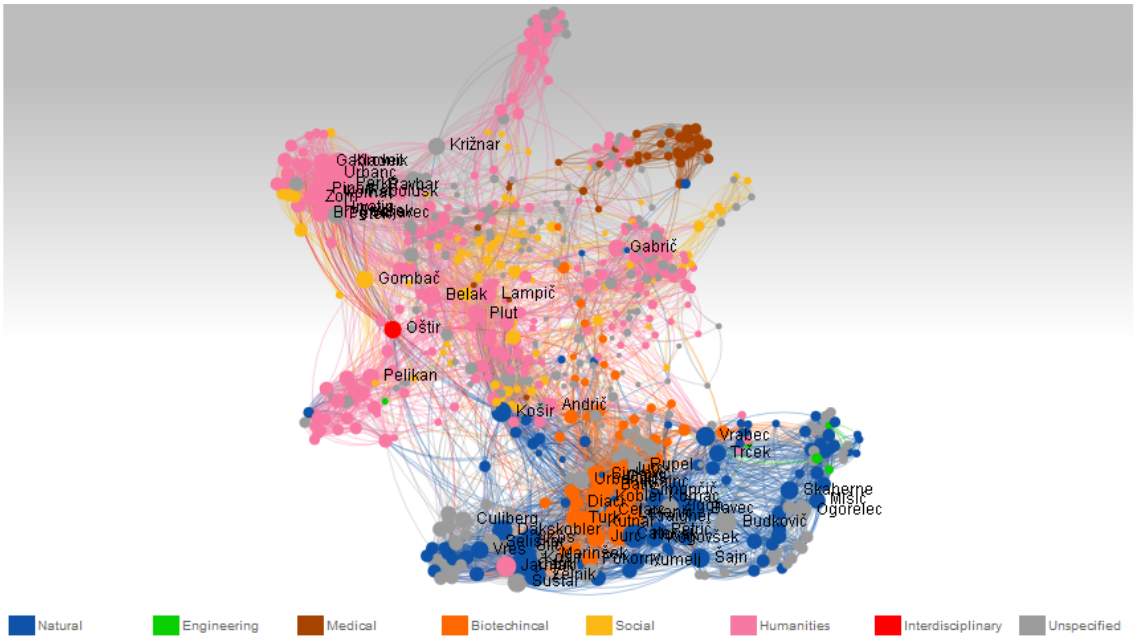


Figure 5.5: Collaboration graph for topic Slovenia.

5.4.5 Competence maps of research topics and individuals

Competence map based on a research topic shows projects and keywords related to a topic. The goal of this graph is to use keywords to show the most relevant competences related to some topic. Since the keywords are derived from the research projects, the graph shows competences derived from the actual research work. Figure 5.6 shows the competence map of the topic "artificial intelligence". The competence map is zoomed to a position close to keywords "artificial intelligence" and "machine learning". Since these two keywords appear in the most project related to the artificial intelligence, their nodes are bigger and labels are initially displayed. There are other keywords that appear on the graph, but are less important and their labels are displayed on mouse-over event – for example inductive logic programming in Figure 5.6. The list of the most important competences is explicitly given in form of list of clickable keyword list at the bottom of competence map.

For artificial intelligence the most important competences are: artificial intelligence, machine learning, statistics, mixed modes, mathematical modeling, knowledge, learning, quality assurance, data mining and neural network. In order to illustrate the appropriateness of the competence map keywords with the topics, we generate a competence map on 8 different topics (4 of which were analyzed in collaboration diagram part of the results section) and show the results in Table 5.2.

Except showing the relatedness of topics, the competence map shows the intensity of research work, number of projects and groups of similar projects.

5.4.5.1 Competence map of an individual researcher

Competence map of a researcher uses keywords of projects to show the research activity of a researcher. Figure 5.7 shows an example of competence map for a researcher with 19 projects. The most important keywords are: machine learning, language technologies, artificial intelligence, language resources, slovene language, statistics and distance learning. There are two groups of projects – group on the left side in Figure 8 connected with keywords: artificial intelligence, machine learning, language technologies, Slovene language

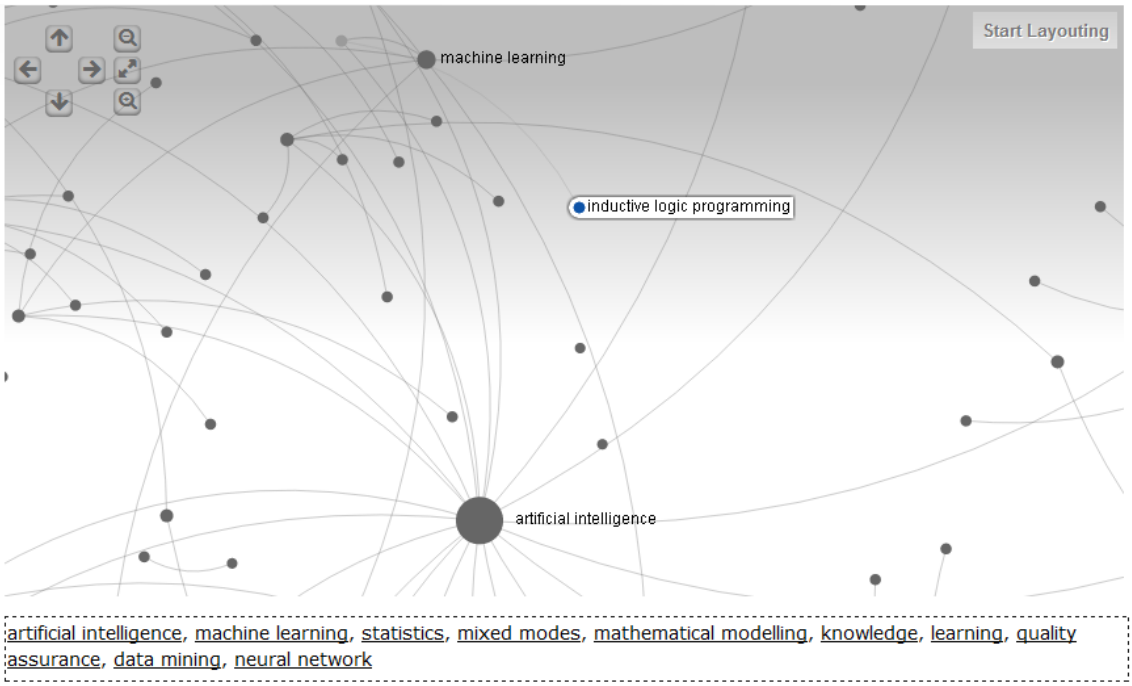


Figure 5.6: Competence map of artificial intelligence topic.

and language resources, and group on the right side connected with the keyword "statistics". The competence map suggests that the research interest of the researcher is closely related to language oriented machine learning and artificial intelligence. This can be confirmed by looking at the keywords explicitly assigned to the researcher as a separate attribute: Artificial intelligence, intelligent systems, machine learning, data-mining, text-mining, intelligent agents, learning from the Web, intelligent data analysis.

In this example the keywords derived with the competence map are very similar to the static ones. This means that the corresponding researcher from the example produces scientific results (publications and projects) that are in line with the manually entered keywords. Consider a case of a researcher who entered those same keywords to describe her research interest, but since then did not produce any research results in the form of publications or participations in research projects. In this example competence map will not produce any keywords, although the static keywords indicate expertise in machine learning and related topics. In practical usage, if someone is searching for experts in a particular field (for purpose of consulting, research collaboration, etc.), competence map will identify only those researchers with research results in exactly these topics.

5.5 Results

Here we summarize the results of testing the approach of conceptualizing science based on collaboration and competences. The analysis of collaboration of four research topics gives the answer to our first research question:

Can conceptualization of science based on collaboration reveal important facts such as: the most important researchers in a topic, nodes and links that connect the community, clusters and interdisciplinarity of the community?

The use-cases show the conceptualization of science using collaboration reveals the important facts defined with the research, including: size of the research community; most

Table 5.2: Topics and competences generated with competence map.

Topic	Science Atlas competence map keywords
anthropology	anthropology, slovenia, slovenes, historical anthropology, theology, ethics, philosophy, multiculturalism, europe, globalization
artificial intelligence	artificial intelligence, machine learning, statistics, mixed modes, mathematical modeling, knowledge, learning, quality assurance, data mining, neural network
nanotechnology	nanotechnology, adsorption, electron microscopy, nanotubes, nanofluidics, 1d nanomaterials, energy storage
Slovenia	slovenia, ecology, epidemiology, management, stratigraphy, environment, modelling, vre, culture, pcb
breast cancer	breast cancer, chemotherapy, brca1, sentinel node, hormone dependent breast cancer, polymorphism, prognosis, erythropoietin, anemia, cathepsins
sensors	sensors, thick-film technology, pressure sensors, hybrid microelectronics, electronics, actuators, processing, optical sensors, thin films, organic semiconductors
social politics	social security, privatization, social cohesion, elderly, attitudes, employment, activation
robotics	robotics, actions, manufacturing systems, medical and service robotics, humanoid robots, merkel cells, mechanoreception

important researchers – the nodes with most connections and position that connects groups are considered the most important; dominant science group – the dominant colors represent the science group with most representatives; diversity of the community – are in the community present only researchers from one science group or is the distribution uniform; connectedness of the community – is the community weekly connected with many individual nodes and disconnected components, or is the community one connected component. In order to summarize the findings we have made by visually inspecting the diagrams shown in sections 5.1 – 5.4, we have expressed the perceived size, diversity and connectivity as small (low), medium, or big (high) in Table 5.3. This evaluation is not a part of conceptualization, but a way to present the results of use-cases. For example, a user interested in inspecting research community based on topic of nanotechnology could be an entrepreneur interested in starting a company for producing nanomaterials. For him, the diagram can provide an answer if there is a sufficient scientific community he can consult for business purposes and the criteria for evaluating the properties of the collaboration diagram can be completely different in his case.

The analysis of competence map of individuals and research topics gives the answer to the second research question:

Can conceptualization of science based on competences reveal the actual research interests of researchers? Where "actual research interest" refers to research topics derived from results of researcher's work (scientific publications and research projects).

Since competence map is based on the keywords derived from the research projects and scientific collaboration, these keywords indeed indicate the research topics individuals are active and productive in. We have further analyzed the keywords generated with

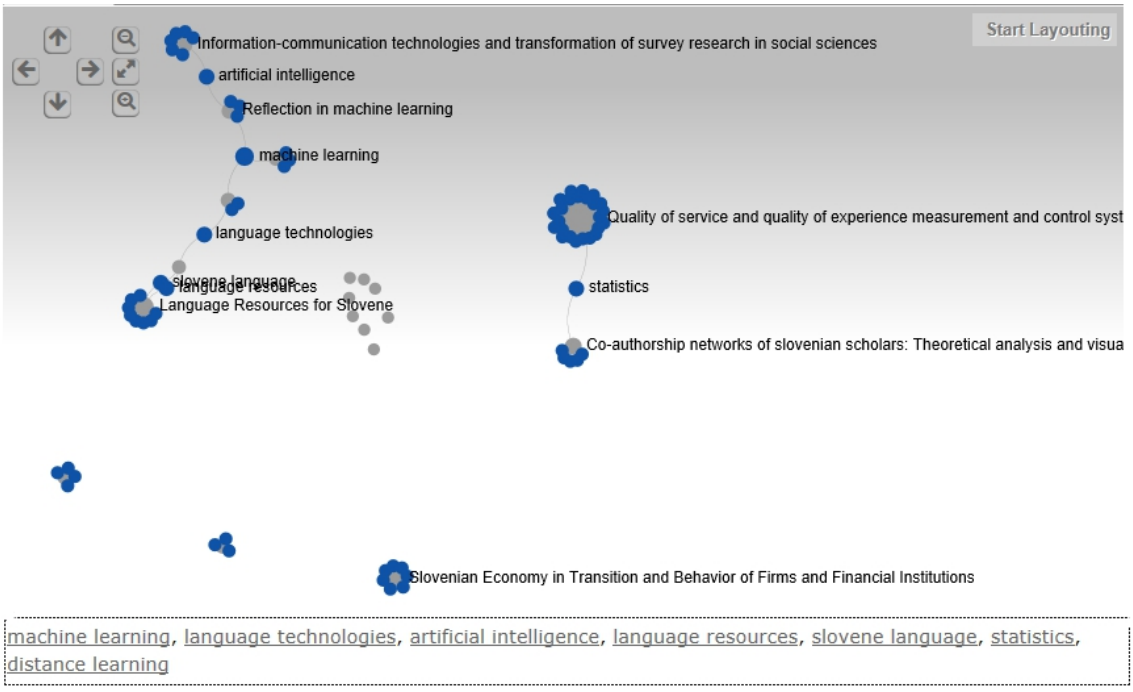


Figure 5.7: Competence map of a researcher. Grey nodes represent projects, while blue nodes represent keywords. Projects are connected on the basis of common keywords.

Table 5.3: Results of analysis of collaboration graphs from the use-cases.

Topic	Community size	Relevent researchers	Dominant science group	Diversity of the community	Connectivity of the community
anthropology	medium	Simonič, Repić, Bartulović, Baškar, Spraizer...	humanities	medium	medium low
artificial intelligence	medium	Bratko, Mladenić, Grobelnik, Erjavec...	engineering	medium	medium low
nanotechnology	medium	Bobojević, Conradi, Jesih, Blinc...	natural sciences	medium	medium
Slovenia	big	Košir, Kriznar, Oštir, Gombač, Simončič...	humanities	high	high

the competence maps and compared them with the static keywords entered usually by researcher himself, his librarian, or other person responsible for this task in a research institution ("keywords" is a typical attribute of bibliographic databases). The quantitative

analysis shows that the average cosine similarity (Salton, 1989) between the generated keywords with the static ones is 35%. The qualitative analysis is performed by manually inspecting the relatedness of the two sets of keywords. The analysis shows that the two sets are related, but the set of generated keywords is sometimes more detailed or describes a related topic. Focusing on a particular subtopic and shifting to a topic related to the initially defined one are typical types of behavior of a researcher during his career and this is well captured when comparing the two sets of keywords. This can be observed in Table 5.4.

Table 5.4: Examples of static keywords, keywords defined with competence and the relation between them.

Static keywords	Keywords generated with competence map	Relation
Human geography, urban geography, regional geography	basic research into Slovenian music history, Plečnik, Arhitecture, history of art	Topic becomes more specific
History and arts, art history, painting, sculpture and architecture	ATM, ASF, application support functions, broadband communications, broadband networks, NH, national host, economic growth, economic development...	Topic shift
Modern english drama, modern american literature, slovene immigrant literature...	presentation, critical response, mediators, translations	Topic shift

5.6 Conclusion and Future Work

Conceptualization of science should ideally state the boundaries of the modeled system or object, basic building blocks, e.g., units of analysis or key actors, interactions of building blocks (e.g., via coupled networks), basic mechanisms of growth and change, and the underlying static and dynamic laws (Börner & Scharnhorst, 2009). We have defined the basic unit of analysis (researchers), its attributes and relations, and the aggregated units. The basic units interact by collaboration on research projects and publications, continuously changing the structure of the collaboration network and competences of the basic units. On top of that we have built a collaboration diagram and a competence map that enable exploration of the proposed conceptualization of science. Specifically, conceptualization of science based on collaboration reveals the most important researchers in a topic, nodes and links that connect the community, clusters and interdisciplinarity of the community; and conceptualization of science based on competences reveal the actual research interests of researchers. This approach can be applied to existing bibliographic databases. We tested the approach on real-world data using publicly available web application Science Atlas, implemented on the national research data of Slovenia. Four use-cases from different science groups in Slovenia are described to illustrate collaboration network and topic analysis. It was shown that the approach reveals size of the research community, most important researchers, the dominant colors represent the science group with most representatives; diversity of the community and connectedness of the community. Alongside collaboration

network the approach enables topical analysis. Science Atlas implements topical analysis with competence map, a method for determining the most important keywords based on connections between research projects. Quantitative and qualitative analysis of keywords generated with competence map indicates that these keywords describe the actual research interests of researchers. In the future work we plan to expand the approach to enable temporal and geographical analysis. The practical application will be applied on other scientific datasets (such as FP7 and Horizon 2020 research projects). The proposed approach can boost many related research. We give some examples of research directions that can be inspired by this study. One such study is analysis of interdisciplinarity of different scientific topics and individual researchers using ego-centric collaboration diagrams (Karlovčec & Mladenić, 2015). This was achieved by generating collaboration diagrams and then measuring the interdisciplinarity of the researchers from the diagrams by looking at scientific fields or at their competences. Another research direction could be analyzing the structure of collaboration graphs and identifying topics in which researchers are the most densely connected. Furthermore, a research could also be directed into analyzing how different ego-networks influence the development of competences of researchers.

Chapter 6

Conclusions and Future Work

In this thesis we have developed several novel methods and approaches for analyzing networks using structural, textual and temporal data modality. Using the network structure and concept of group degree centralization, the groups of important nodes can be identified in large real world networks. Temporal evolution of structural properties such as core-periphery can be used to characterize behavior of individual network nodes. By looking at structure and textual information assigned to nodes, the level of network diversity can be determined using a diversity metric based on Stirling's index. Finally, the methods for text and network analysis can be used for conceptualizations of scientific community that can show the most important researchers, dominant science group, diversity and connectedness, the most important researchers in a topic, nodes and links that connect the community, clusters of nodes and interdisciplinarity.

The results can be summarized in the following list of scientific contributions:

- *Polynomial time group degree centralization approximative algorithm.* We have developed an algorithmic approach for efficient identification of centralization for an arbitrary group of nodes in network. We used greedy approach for maximum domination problem to develop algorithm for identifying a group with maximal degree centralization withing best possible approximation bound. We conducted experiments on several real world networks that range up to billions of edges. The algorithm was implemented in both C++ and Python and the code is publicly available online.

- *Approach for characterizing dynamics of individual nodes based on core/periphery structural position using clustering and decision trees.* We used core-periphery structural property of networks and analyzed how it evolves in time. We used the structural and temporal component to characterize the behavior of members that constitute the network. To do this, we used the k-means clustering method on time series of core-periphery position values. We also used decision tree classification algorithm to build descriptive models that can characterize individuals in terms of core stability, balance and strength.

- *New interdisciplinary metric for networks based on Stirling's diversity index and text analysis based similarity matrix.* We used textual content assigned to nodes alongside structure of network to determine diversity of nodes in networks. We used an novel approach of constructing subject categories similarity matrix using cosine distance between textual descriptions. We used this approach to extend the Stirling's index. The new metrics was used to analyze the evolution of interdisciplinarity of Slovenian research collaboration networks.

- *Approach for conceptualization of science using network of collaboration and competences of individuals.* We used the methods for analyzing network and text and created conceptualization of scientific community in Slovenia. The methods are implemented in an open interactive web application that can show the most researchers in a topic, nodes and links that connect the community, clusters and interdisciplinarity of the community, and actual research interests of researchers.

The polynomial time group degree approximative algorithm has importance for the state of the art in field of network analysis, because it enables computing group degree centralization on large networks. Measuring and comparing group degree centralization on different real world social networks can be of great importance for exploratory network analysis. The approach for characterizing nodes based on dynamics of a structural property, combines temporal component of networks and machine learning algorithms often used in data mining. Combining these approaches can give more information about the network. The developed network diversity metric uses text assigned to nodes to determine how diverse is a network. This is a new network analysis method for determining diversity, autocorrelation or homophily/heterophily, that can have many applications. The conceptualization of science using collaboration and competences illustrates the possibility of network and text analysis methods for better understanding of a social system.

In future work we will investigate the unimodality of centralization with increasing the size of the group. It is a property observed in our experiments and future investigation could lead towards additional improvements of the group degree centralization algorithm. Another interesting research question that will be a part of future work is investigating error bonds on real-world network for our approximative algorithm. Evolution of structure was extensively used for core-periphery property that produces binary time-series of structural value for individual nodes. In future work, we will expand the approach on other properties such as network community structure and other indices that produce time-series with continuous values. The results of science interdisciplinarity exploratory analysis give many possibilities for future research. The most interesting direction seems investigation of the causes why some subject topics are more interdisciplinary based on projects, while other are more interdisciplinary on publications. Another interesting phenomenon is the difference in knowledge integration across topics. Understanding and ability to direct interdisciplinarity of particular topics would of great value in the context of interdisciplinarity as an environment for exceptional scientific achievements and innovation.

Appendix A

Implementation of GDC in C++ Using SNAP

In this appendix we give C++ implemented of algorithm for finding the group with maximal group degree centralization. The algorithm is implementation using SNAP (Stanford Network Analysis Platform), which is based on a general purpose STL (Standard Template Library)-like library GLib developed at Jožef Stefan Institute.

GetGroupDegreeCentralization function calculates group degree centralization for the input group. In order to calculate group centralization with each new addition of a node into the group, we are updating variables A, B and sum. A and B are updated in constant time, while updating sum requires δ iterations in each step.

```
TFlt GetGroupDegreeCentralization(const int N, TIntH& Degs,
     const TIntH& GroupNodes, double GD, int k, TIntFltH& mem,
     TFlt& A, TFlt& B, TIntH& deg_dist, double& approx) {

  TFlt n = N;
  TFlt sum = 0;
  if (k >= 1) {
    k--;
    A *= (n - (double)k) / (n - (double)k - 2.0);
    B *= (n - (double)k - 1) / (n - (double)k - 2.0);
    // iterate sum vector to update values and get the sum
    for (THashKeyDatI<TInt, TFlt> NI = mem.BegI(); NI < mem.EndI(); NI++) {
      int i = NI.GetKey();
      // get old value
      TFlt old_val = NI.GetDat();
      TFlt new_val = old_val * ((n - i - k - 1.0) / (n - k - 2.0));
      // store new value in the sum vector
      mem.AddDat(i, new_val);
      // get the sum
      TFlt deg_val = (float)deg_dist.GetDat(i);
      sum += (deg_val * new_val);
    }
  }
  TFlt centralization = A * GD + sum - B;

  return centralization;
}
```

Decreasing contribution is the function that performs operations of decreasing group degree contribution of a node. The function takes a node from its current bucket in the histogram of contributions, decreases the contribution for 1 and moves it into the appropriate bucket. For each newly chosen group node, contributions of all incoming neighboring

nodes (in-nodes) decreased by 1. Also, contributions of all outgoing neighboring nodes (out-nodes) is decreased by 1. In addition to that the contributions are decreased by 1 for all incoming neighboring nodes of the out-nodes (out-in-nodes), and the arcs from out-in-nodes to out-nodes are removed. Finally, the chosen node is removed from the graph. This updating operations ensure that the histogram of contributions stays appropriately updated.

```
void DecreaseContribution(TIntIntHH &Histogram, TIntH& Nodes, TIntH& index,
        int id, int decrease) {
  int HistogramHashId = Nodes.GetDat(id);
  if (Histogram.IsKey(HistogramHashId)) {
    Histogram.GetI(HistogramHashId).GetDat().DelKey(id);
    index.AddDat(HistogramHashId,
    Histogram.GetI(HistogramHashId).GetDat().Len());
  }
  if (Histogram.IsKey(HistogramHashId - decrease)) {
    Histogram.GetI(HistogramHashId - decrease).GetDat().AddDat(id, id);
    index.AddDat(HistogramHashId - decrease,
    Histogram.GetI(HistogramHashId - decrease).GetDat().Len());
  }
  else {
    TIntH NewHash;
    NewHash.AddDat(id, id);
    Histogram.AddDat(HistogramHashId - decrease, NewHash);
    index.AddDat(HistogramHashId - decrease,
    Histogram.GetI(HistogramHashId - decrease).GetDat().Len());
  }
  int prev = Nodes.GetDat(id);
  Nodes.AddDat(id, prev - decrease);
}
```

MaxCPGreedy is the core function that for some inuput graph returns group with the largest group degree centralization. Using the greedy approach in each step of the algorithm a node with the biggest contribution to the group degree centrality is added to the group. We implement this by maintaining a sorted histogram of node contributions. Initially, the contributions to group centrality are equal to node degrees. After each addition of a new node to the group, the histogram is accordingly updated. In this way choosing a node with biggest contribution is a fast operation that in total takes between δ and n steps. The *main loop* of the algorithm terminates when there is no node with contribution greater than -1, which occurs when all the nodes are dominated by the group.

```
void MaxCPGreedy(const PNGraph& Graph, TIntH& GroupNodes,
        TIntFltVH& ResultVector,
        TFltV& MaxVector, bool& Unimodal, const int k) {

  // buildup container of group nodes
  // int br = 0;
  // number of nodes
  int N = Graph->GetNodes();
  // sum of contributions
  double ContributionsSum = 0;
  // Hash table of covered nodes
  TIntH Covered;
  // Indicator if it is in the group set
  TIntBoolH InGroup;
  // DEBUG: file to print out
  //FILE *F = fopen("test10.txt", "wt");
  // id of Result vector element with maximal centralization
  int MaxI = -1;
  // value of maximal centralization
  double MaxGdVal = -1;
```

```
// unimodality check
bool increasing = true;
double PrevGdc = -1;
int NumChanges = 0;

// Centralization variables
// sum vector
TIntFltH mem;
// degree distribution
TIntH deg_distribution;
// variable for updating the sum
TFlt A = 1 / ((double)N - 1.0);
// variable for updating the sum
TFlt B = (double)N / ((double)N - 1.0);

// Hash table of node contributions (initialy degrees) sorted by values
TIntH Nodes;
// Hash table of node in degrees (persuming Degs->GetDat()
// is faster then NI->GetInDeg())
TIntH Degs;
for (TNGraph::TNodeI NI = Graph->BegNI(); NI < Graph->EndNI(); NI++) {
  Nodes.AddDat(NI.GetId(), NI.GetInDeg());
  Degs.AddDat(NI.GetId(), NI.GetInDeg());
  InGroup.AddDat(NI.GetId(), false);
}
Nodes.SortByDat(false);

// u of nodes of contributions - Key() - node contribution,
// Dat() - Hash table of nodes with the distribution
TIntIntHH Histogram;
int PrevDeg = -1;
TIntH NewHash;
for (THashKeyDatI<TInt, TInt> NI = Nodes.BegI(); NI < Nodes.EndI(); NI++) {
  int CurrDeg = NI.GetDat();
  int CurrKey = NI.GetKey();
  if (CurrDeg == PrevDeg) {
    NewHash.AddDat(CurrKey, CurrKey);
  }
  else  {
    if (PrevDeg != -1) {
    Histogram.AddDat(PrevDeg, NewHash);
      }
  NewHash.Clr();
  NewHash.AddDat(CurrKey, CurrKey);
  PrevDeg = CurrDeg;
  }
}
if (NewHash.Len() > 0) {
  Histogram.AddDat(PrevDeg, NewHash);
}
Histogram.SortByKey(false);

// Index of histogram values - index is acctualy
// a histogram of contributions. Key() - contribution,
// Dat() - number of nodes with that contribution
TIntH index;
for (THashKeyDatI<TInt, TIntH> HI = Histogram.BegI();
HI < Histogram.EndI(); HI++) {
  index.AddDat(HI.GetKey(), HI.GetDat().Len());
  mem.AddDat(HI.GetKey(), 1 / ((double)N - 1));
  deg_distribution.AddDat(HI.GetKey(), HI.GetDat().Len());
```

```
}

// Choose node
int ChoosenNode = -1;
// Choose the biggest degree
int ChoosenDeg = Histogram.BegI().GetKey();
// Group nodes counter
int GroupNodesCounter = 0;

// Main loop
do {
  ChoosenNode = -1;
  // Choosing the node - this takes in sum O(maxdeg) steps, or O(n) steps
  while (ChoosenNode == -1 && ChoosenDeg >= 0) {
    if (index.IsKey(ChoosenDeg)) {
      if (index.GetDat(ChoosenDeg) > 0) {
        ChoosenNode = Histogram.GetDat(ChoosenDeg).BegI().GetDat();
      }
      else {
        ChoosenDeg--;
      }
    }
    else {
      ChoosenDeg--;
    }
  }

  if (ChoosenNode != -1) {
    ContributionsSum += ChoosenDeg;
    GroupNodes.AddDat(GroupNodesCounter, ChoosenNode);
    GroupNodesCounter++;
    // If the choosen node is in the set of covered neighbours
    // it has to be removed
    if (Covered.IsKey(ChoosenNode)) {
      Covered.DelKey(ChoosenNode);
    }
    // get all the nodes with choosen degree
    TIntH FirstHash = Histogram.GetDat(ChoosenDeg);
    // first key = choosen deg
    int FirstKey = ChoosenDeg;
    // Number of nodes with choosen degree
    int Len = FirstHash.Len();

    // Remove the index for this degree if it is the last element,
    // otherwise just decrease the size
    if (Len == 0 || Len == 1) {
      Histogram.DelKey(ChoosenDeg);
      index.DelKey(ChoosenDeg);
    }
    else {
      Histogram.GetI(ChoosenDeg).GetDat().DelKey(ChoosenNode);
      index.AddDat(FirstKey, Len - 1);
    }

    // Get the node
    TNGraph::TNodeI NI1 = Graph->GetNI(ChoosenNode);

    // A) reduce contributionto in-degree nodes by 1 (or 2)
    for (int i = 0; i < NI1.GetInDeg(); i++) {
      int nei = NI1.GetInNId(i);
      int decrease = 1;
      if (Nodes.IsKey(nei) && ChoosenNode != nei) {
```

```
      DecreaseContribution(Histogram, Nodes, index, nei, decrease);
    }
  }

  // B) for all node-outneigh neigbours
  for (int i = 0; i < NI1.GetOutDeg(); i++) {
    // Id of out neighbout
    int n_out = NI1.GetOutNId(i);
    DecreaseContribution(Histogram, Nodes, index, n_out, 1);
    if (!Covered.IsKey(n_out)) {
      Covered.AddKey(n_out);
    }
    if (ChoosenNode != n_out) {
      // Get all node-outneigh-inneigh nodes
      TNGraph::TNodeI NI1 = Graph->GetNI(n_out);
      TIntV n_out_ins;
      for (int j = 0; j < NI1.GetInDeg(); j++) {
        // Id of node-outneigh-inneigh
        int n_out_in = NI1.GetInNId(j);
        if (n_out_in != ChoosenNode) {
          n_out_ins.Add(n_out_in);
        }
      }

      if (n_out_ins.Len() > 0) {
        for (int i = 0; i < n_out_ins.Len(); i++)   {
          int n_out_in = n_out_ins[i];
          // B.1 Delete in edges
          if (Graph->IsEdge(n_out_in, n_out)) {
            Graph->DelEdge(n_out_in, n_out);
          }
          // B.2 DECREASE node-outneigh-inneigh for one
          DecreaseContribution(Histogram, Nodes, index, n_out_in, 1);
        } // end for
      } // end if
    } // end if
  } // end for B)

  // C) delete the node
  Graph->DelNode(ChoosenNode);
  Nodes.DelKey(ChoosenNode);
  Covered.Len()
  double approx = 0;
  TFlt gdc2 = TSnap::GetGroupDegreeCentralization(N, Degs,
  GroupNodes, ContributionsSum, GroupNodesCounter, mem, A, B,
  deg_distribution, approx);

  // save result vector
  TFltV vec;
  vec.Add((double)GroupNodesCounter);
  vec.Add(ChoosenNode); vec.Add(ChoosenDeg);
  vec.Add((double)Degs.GetDat(ChoosenNode));
  vec.Add(ContributionsSum); vec.Add(gdc2);
  vec.Add(approx);
  ResultVector.AddDat(GroupNodesCounter, vec);

  // check if its max
  if (gdc2 > MaxGdVal) {
    MaxI = GroupNodesCounter;
    MaxGdVal = gdc2;
    MaxVector = vec;
  }
```

```
      // count number of changes of unimodality trend
      if (PrevGdc <= gdc2 && !increasing) {
        NumChanges++;
        increasing = true;
      }
      else if (PrevGdc >= gdc2 && increasing) {
        NumChanges++;
        increasing = false;
      }
      else { }
      PrevGdc = gdc2;
    }
  } while (ChoosenNode > -1);

  if (NumChanges > 1) {
    Unimodal = false;
  }
}
```

Miain is a function that illustrates usage of max degree centralization function for a specific graph:

```
int main(int argc, char* argv[]) {
  Env = TEnv(argc, argv, TNotify::StdNotify);
  Try
    // loading input parameters
    const TStr InFNm = Env.GetIfArgPrefixStr("-i:", "./Data/cobiss.txt",
    "Input un/directed graph");
    const TStr OutFNm = Env.GetIfArgPrefixStr("-o:", "./Output/gc.out.txt",
    "Output file");

    // group nodes for output
    TIntH gk1;
    // result vectors for oputput
    TIntFltVH ResultVector;
    // max vector for output
    TFltV MaxVector;
    // unimodality variable for output
    bool Unimodal = true;

    // loading graph
    PUNGraph UGraph = TSnap::LoadEdgeList<PUNGraph>(InFNm);
    // converting to directed graph
    PNGraph DGraph = TSnap::ConvertGraph<PNGraph, PUNGraph>(UGraph);

    // max size of the group is initialy the number of nodes
    int k = UGraph->GetNodes();

    // Main part - centralization
    TSnap::MaxCPGreedy(DGraph, gk1, ResultVector, MaxVector, Unimodal, k);
  }

  Catch
    return 0;
}
```

Appendix B

Implementationof the Diversity Measure Based on Connectivity and Text Similarity in C++

CalculateEij is a function for computing pairwise probabilities:

```
TIntFltH CalculateEij(PUNGraph& Graph, TIntIntH Cats, TIntFltH P) {
  TIntFltH E;
  int e = Graph->GetEdges();
  int e_conn = 0;
  int N = P.Len();
  for(int i=0;i<P.Len();i++) {
    for(int j=i;j<P.Len();j++){
      if (i!=j) {
        for (TUNGraph::TEdgeI EI = Graph->BegEI();
        EI < Graph->EndEI(); EI++) {
          int SRC = Cats.GetDat(EI.GetSrcNId());
          int DST = Cats.GetDat(EI.GetDstNId());
          int I = P.GetKey(i);
          int J = P.GetKey(j);
          if (Cats.GetDat(EI.GetSrcNId())==P.GetKey(i) &&
          Cats.GetDat(EI.GetDstNId())==P.GetKey(j) ||
          Cats.GetDat(EI.GetSrcNId())==P.GetKey(j) &&
          Cats.GetDat(EI.GetDstNId())==P.GetKey(i))
            e_conn++;
         }
         E.AddKey((N*P.GetKey(i))+P.GetKey(j));
         E.AddDat((N*P.GetKey(i))+P.GetKey(j),(float)e_conn/(float)e);
         E.AddKey((N*P.GetKey(j))+P.GetKey(i));
         E.AddDat((N*P.GetKey(j))+P.GetKey(i),(float)e_conn/(float)e);
         e_conn = 0;
           }
    }
  }
  return E;
}
```

index is a function that computes extended Stirling's interdisciplinarity:

```
float Index(PUNGraph& Graph, TIntFltH& E, TIntIntH& Cats, TIntFltH& P,
TIntFltH& D, int Alpha, int Beta, int Gamma) {
  int N = P.Len();
  float I = 0;
  for(int i=0;i<P.Len();i++){
    for(int j=i;j<P.Len();j++){
      if (i!=j) {
```

```
        double a = D.GetDat((N*P.GetKey(i))+P.GetKey(j));
        double b = P.GetDat(P.GetKey(i));
        double c = P.GetDat(P.GetKey(j));
        double d = E.GetDat((N*P.GetKey(i))+P.GetKey(j));
        I += pow(1-D.GetDat((N*P.GetKey(i))+P.GetKey(j)),Alpha) *
        pow(P.GetDat(P.GetKey(i))*P.GetDat(P.GetKey(j)),Beta)*
        pow(E.GetDat((N*P.GetKey(i))+P.GetKey(j)),Gamma);
      }
    }
  }
  return I;
}
```

StirlingIndex is a function that combines all the pieces needed for computing interdisciplinarity based on extended Stirling's index. The function loads a list of nodes to categories associations, calculates probabilities of categories, loads a similarity matrix, calculates connectivity and calculates value of interdisciplinarity index of a graph:

```
double StirlingIndex(PUNGraph& Graph, TStr InFNmCat, TStr InFNmMat,
int Alpha, int Beta, int Gamma) {
  const int N = Graph->GetNodes();
  TIntFltH P;
  TIntIntH Cats = TSnapDetail::LoadCatList(Graph,InFNmCat,P);
  TSnapDetail::CalculatePij(Graph,P);
  TIntFltH D;
  TSnapDetail::LoadMat(Graph,InFNmMat,D);
  TIntFltH E = TSnapDetail::CalculateEij(Graph,Cats,P);
  float I = TSnapDetail::Index(Graph,E,Cats,P,D, Alpha, Beta, Gamma);
  return double(I);
}
```

This is an example of using the interdisciplinarity function:

```
int main(int argc, char* argv[]) {
  // code needed for inputing parameters
  Env = TEnv(argc, argv, TNotify::StdNotify);
  Try
  const TStr InFNmGraph = Env.GetIfArgPrefixStr("-i:",
  "artificial_intelligence_pub.txt", "Input graph (undirected graph)");
  const TStr InFNmCat = Env.GetIfArgPrefixStr("-c:",
  "artificial_intelligence_cat_pub.txt", "Categories");
  const TStr InFNmMat = Env.GetIfArgPrefixStr("-m:",
  "sciences.txt", "Matrix");
  const TStr OutFNm = Env.GetIfArgPrefixStr("-o:",
  "diversity.txt", "Output file");
  const int DivAlg = Env.GetIfArgPrefixInt("-a:", 1, "Measure: 1:Stirling");
  const int Alpha = Env.GetIfArgPrefixInt("-alp:", 1, "alpha");
  const int Beta = Env.GetIfArgPrefixInt("-bet:", 1, "beta");
  const int Gamma = Env.GetIfArgPrefixInt("-gam:", 1, "gama");

  // defining graph
  PUNGraph Graph = TSnap::LoadEdgeList<PUNGraph>(InFNmGraph, false);
  double D = 0.0;
  TStr DivendAlgStr;

  // based on input parametr -a (variable DivAlg),
  // diversity measure is choosen
  if (DivAlg == 1) {
    DivAlgStr = "Stirling";
    D = TSnap::StirlingIndex(Graph,InFNmCat,InFNmMat, Alpha, Beta, Gamma);
  }
  else { Fail; }
```

```
  Catch
  return 0;
}
```

Appendix C

Science Atlas Application

This appendix gives an overview of Slovenian Science Atlas application system architecture. The server side of the application represents the core that contains all the functionalities provided with the application. The server side of application is installed on the web server hosting machine, which is a powerful machine with big amounts of memory and processor power. The functionalities of the server are written in *C#* programing language. The architecture of the system is arranged is such a way that all the computational demanding tasks are performed on the server side. These functionalities are then accessed by the clients in the form of services which are independent on the device and software (operating system) of the users. Figure C.1 represents the architecture of the system.

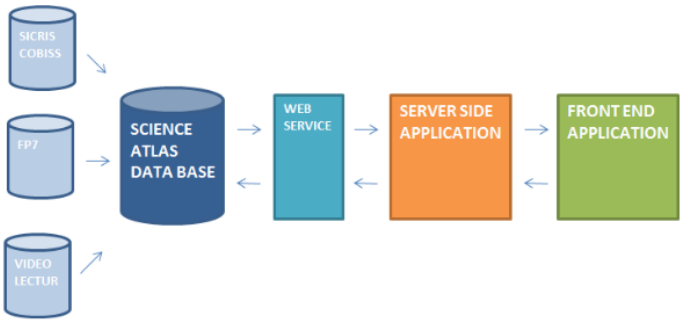


Figure C.1: Science Atlas application architecture

The data layer of the application is constructed from the database (Science Atlas Data Base) that integrates data from different sources (SICRIS, COBISS, FP& projects, VideoLectures.NET). Next, the web service is an interface between the data and the users. The request and the deployment of the data is mediated through the module that performs various functionalities on the server side module.

The last module is the front end application, with which all the functionalities are represented to user in the form of a client web application. In the rest of this report, the server side of the application will be analyzed in more detail.

C.1 Web service

Web service is the part of the application that enables the transferee of data from and to database. Using the web service, the data is transferred in a standard format (XML, JSON) and it is independent on the rest of the application which makes the provided data

easily usable in other applications. The web service is written in *C#* programing language and .NET Framework 3.5.

ProfileOfRsrJson call returns the profile information of a researcher (with the specified id input parameter) that include first and last name of the researcher, classification of the researcher (codes and names of science group, field and subfield), organizations of the researcher and contact information. In addition to this, the call returns all the projects on which a researcher has worked on. Input parameter: string id

RsrByIdsJson function returns profile information of every researcher for which the id is given in the array of strings input parameter. Input parameter: string[] ids (string of researcher ids)

RsrOnPrjJson call returns the profile information of all the researchers who worked on a project specified with the input id. Input parameter: string id (id of a project)

RsrOnPubJson call returns the profile information of all the researchers who collaborated on a publication specified with the input id. Input parameter: string id (id of a publication)

RsrInOrgJson call returns the profile information of all the researchers employed in an organization. Input parameter: string id (id of the organization)

CollaborationOfRsrOnPrj1Json call returns the profile information of the researchers that collaborated with a researcher specified with the input id, the collaboration is based on publications Input parameter: string id (id of a researcher)

CollaborationOfRsrOnPrjHeadJson call returns the profile information of the researchers that collaborated with a researcher specified with the input id, but it returns only the collaboration between the heads of the projects. The collaboration is based on projects. Input parameter: string id (id of a researcher)

CollaborationOfRsrOnPubJson call returns the profile information of the researchers that collaborated with a researcher specified with the input id, the collaboration is based on publications. Input parameter: string id (id of a researcher)

PrjOfOrgJson call returns all the projects of an organization. Input parameter: string id (id of a organization)

PrjOfRsrJson call returns all the projects of a researcher. Input parameter: string id (id of a researcher)

PrjAll call returns the ids of all the projects in the database. Input parameter: string id (id of a researcher)

PrjByRsrIdsJson call returns all the projects of a set of researchers. Set of researchers is an input parameter in the form of array of ids. This ids are putt into SQL query using a loop that combines them with OR logical statement. Input parameter: string[] id (array of researcher ids)

PrjByIdJson call returns the information about the project specified by the project id. Input parameter: string id (id of a project)

PublicationsOfRsrJson call returns all the publications of a researcher specified by the researcher id. Input parameter: string id (id of a researcher)

PubById call returns the publication specified by the id of the publication. Input parameter: string id (id of a publication)

OrgByRsrIdsJson call returns all the organization of researchers specified by the array of researcher ids, specified by the input parameter. The ids are putt into SQL query using a loop that combines them with OR logical statement. Input parameter: string[] id (ids of researchers)

OrgByIdJson call returns projects specified by the array of project ids given as the input parameter. This ids are putt into SQL query using a loop that combines them with OR logical statement. Input parameter: string[] id (ids of organizations)

LecOfRsrJson call returns all the lectures of a researcher specified by the researcher id input parameter Input parameter: string id (id of a researcher)

References

Abbasi, A., Chung, K. S. K., & Hossain, L. (2012, July). Egocentric analysis of co-authorship network structure, position and performance. *Inf. Process. Manage. 48*(4), 671–679.

Adams, J., Jackson, L., & Marshall, S. (2007). *Bibliometric analysis of interdisciplinary research*. Higher Education Funding Council for England.

Ahn, W.-k., Kim, N. S., Lassaline, M. E., & Dennis, M. J. (2000). Causal status as a determinant of feature centrality. *Cognitive Psychology*, *41*(4), 361–416.

Alba, R. D. & Moore, G. (1978). Elite social circles. *Sociological Methods and Research*, *7*(2), 167–188.

Badar, K., Hite, J. M., & Badir, Y. F. (2013). Examining the relationship of co-authorship network centrality and gender on academic research performance: the case of chemistry researchers in pakistan. *Scientometrics*, *94*(2), 755–775.

Baldwin, C., Hughes, J., Hope, T., Jacoby, R., & Ziebland, S. (2003, January). Ethics and dementia: mapping the literature by bibliometric analysis. *International journal of geriatric psychiatry*, *18*(1), 41–54. doi:10.1002/gps.770

Barabási, A. L., Jeong, H., Néda, Z., Ravasz, E., Schubert, A., & Vicsek, T. (2002). Evolution of the social network of scientific collaborations. *Physica A: Statistical Mechanics and its Applications*, *311*(3-4), 590–614. doi:DOI:10.1016/S0378-4371(02)00736-7

Barthelemy, M., Gondran, B., & Guichard, E. (2003). Spatial structure of the internet traffic. *Physica A: Statistical Mechanics and its Applications*, *319*, 633–642. doi:http://dx.doi.org/10.1016/S0378-4371(02)01382-1

Batagelj, V. & Mrvar, A. (2004). Pajek — analysis and visualization of large networks. In M. Jünger & P. Mutzel (Eds.), *Graph drawing software* (pp. 77–103). Mathematics and Visualization. Springer Berlin Heidelberg. doi:10.1007/978-3-642-18638-7_4

Bauer, B., Jordán, F., & Podani, J. (2010). Node centrality indices in food webs: rank orders versus distributions. *Ecological Complexity*, *7*(4), 471–477.

Bavelas, A. (1948). A mathematical model for group structures. *Hum. Organ. 7*(3), 16–30.

Bavelas, A. (1950). Communication patterns in task-oriented groups. *J. Acoust. Soc. Am.* 725–730.

Bell, J. R. (2014). Subgroup centrality measures. *Network Science*, *2*(02), 277–297.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, *2008*(10), P10008. Retrieved from http://stacks.iop.org/1742-5468/2008/i=10/a=P10008

Borgatti, S. P. (2006). Identifying sets of key players in a social network. *Comput. Math. Organ. Theory*, *12*(1), 21–34.

Borgatti, S. P. & Everett, M. G. (1999). Models of core/periphery structures. *Social Networks*, *21*, 375–395.

Borgatti, S. P., Everett, M. G., & Freeman, L. (2002). UCINET 6 for Windows: Software for Social Network Analysis. Harvard, MA, Analytic Technologies.

Börner, K. (2010). *Atlas of science: visualizing what we know*. MIT Press.

Börner, K. & Scharnhorst, A. (2009). Visual conceptualizations and models of science. *Journal of Informetrics*, *3*(3), 161–172. Science of Science: Conceptualizations and Models of Science. doi:10.1016/j.joi.2009.03.008

Bott, E. (1957). *Family and social network*. Tavistock, London.

Boyd, J. P., Fitzgerald, W. J., & Beck, R. J. (2006). Computing core/periphery structures and permutation tests for social relations data. *Social Networks*, *28*(2), 165–178. Retrieved from http://dblp.uni-trier.de/db/journals/socnet/socnet28.html#BoydFB06

Brandes, U. & Erlebach, T. (2005). *Network analysis: methodological foundations*. Springer Science & Business Media.

Brian Uzzi, M. S., Satyam Mukherjee & Jones, B. (2013). Atypical combinations and scientific impact. *Science*, *342*(6157), 468–472.

Britton, T. & Trapman, P. (2013). Inferring global network properties from egocentric data with applications to epidemics. *Mathematical Medicine and Biology*. doi:10.1093/imammb/dqt022. eprint: http://imammb.oxfordjournals.org/content/early/2013/10/15/imammb.dqt022.full.pdf+html

Brusco, M. (2011). An exact algorithm for a core/periphery bipartitioning problem. *Social Networks*, *33*(1), 12–19. Retrieved from http://dblp.uni-trier.de/db/journals/socnet/socnet33.html#Brusco11

Burt, R. S. (1980). Models of network structure. *Annual Review of Sociology*, *6*, 79–141.

Burt, R. S. (1992). *Structural holes: The social structure of competition*. Cambridge, MA: Harvard University Press.

Callon, M., Courtial, J. P., & Laville, F. (1991). Co-word analysis as a tool for describing the network of interactions between basic and technological research: the case of polymer chemsitry. *Scientometrics*, *22*(1), 155–205. doi:10.1007/BF02019280

Card, S. K., Mackinlay, J. D., & Shneiderman, B. (Eds.). (1999). *Readings in information visualization: using vision to think*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Coleman, J. S. (1990). *Foundations of social theory*. Harvard University Press.

Cotilla-Sanchez, E., Hines, P. D., Barrows, C., & Blumsack, S. (2012). Comparing the topological and electrical structure of the north american electric power infrastructure. *Systems Journal, IEEE*, *6*(4), 616–626.

Coulter, N. S., Monarch, I., & Konda, S. (1998). Software engineering as seen through its research literature: a study in co-word analysis. *JASIS*, *49*(13), 1206–1223. Retrieved from http://dblp.uni-trier.de/db/journals/jasis/jasis49.html#CoulterMK98

Courtial, J. P. (1994). A coword analysis of scientometrics. *Scientometrics*, *31*(3), 251–260. doi:10.1007/BF02016875

Cronin, B. & Atkins Barsky, H. (Eds.). (2000, September). *The web of knowledge: a festschrift in honor of eugene garfield*. ASIS Monograph Series.

de Nooy, W., Mrvar, A., & Batagelj, V. (2005, January). *Exploratory social network analysis with pajek (structural analysis in the social sciences)*. Cambridge University Press. Retrieved from http://www.amazon.ca/exec/obidos/redirect?tag=citeulike04-20%7B%5C&%7Dpath=ASIN/0521602629

Derek John, d. S. P. (1975). *Science since babylon*. Yale University Press.

Ding, Y., Chowdhury, G. G., & Foo, S. (2001, October). Bibliometric cartography of information retrieval research by using co-word analysis. *Inf. Process. Manage. 37*(6), 817–842. doi:10.1016/S0306-4573(00)00051-0

Doreian, P. & Fararo, T. J. (1985). Structural equivalence in a journal network. *Journal of the American Society for Information Science and Technology*, *36*, 28–37.

eMedia Asia Ltd. (2007). Bird's eye view of patent playground. EE Times Asia.

Enderby, J. (2004). Facilitating interdisciplinary research. (Chap. Knowledge, Validation, and Transfer: Science, Communication, and Economic Development). The National Acadamies Press.

Estrada, E. & Rodriguez-Velazquez, J. A. (2005). Subgraph centrality in complex networks. *Physical Review E*, *71*(5), 056103.

Everett, M. G. & Borgatti, S. P. (1999). The centrality of groups and classes. *J. Math. Sociol. 23*(3), 181–201.

Everett, M. G. & Borgatti, S. P. (2005a). Ego network betweenness. *Social Networks*, *27*(1), 31–38.

Everett, M. G. & Borgatti, S. P. (2005b). Extending centrality. *Models and methods in social network analysis*, *35*(1), 57–76.

Everett, M. G., Sinclair, P., & Dankelmann, P. (2004). Some centrality results new and old. *J. Math. Sociol. 28*(4), 215–227.

Flake, G. W., Lawrence, S., Giles, C. L., & Coetzee, F. M. (2002, March). Self-organization and identification of web communities. *Computer*, *35*(3), 66–71. doi:10.1109/2.989932

Frank, O. & Snijders, T. A. B. (1994). Estimating hidden populations using snowball sampling. *Journal of Official Statistics*, *10*, 53–67.

Freeman, L. (1977, March). A Set of Measures of Centrality Based on Betweenness. *Sociometry*, *40*(1), 35–41.

Freeman, L. (1979). Centrality in social networks conceptual clarification. *Soc. Netw. 1*(3), 215–239.

Freeman, L. (1982). Centered graphs and the structure of ego networks. *Mathematical Social Sciences*, *3*, 291–304.

Fruchterman, T. M. J. & Reingold, E. M. (1991). Graph drawing by force-directed placement. *Software: Practice and Experience*, *21*(11), 1129–1164. doi:10.1002/spe.4380211102

Garfield, E. & Shea, J. E. (Eds.). (1981). *Isi atlas of science: biotechnology and molecular genetics 1981/82*.

Goldberg, D. E. (1989). *Genetic algorithms in search, optimization and machine learning* (1st). Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc.

Gomez, D., González-Arangüena, E., Manuel, C., Owen, G., del Pozo, M., & Tejada, J. (2003). Centrality and power in social networks: a game theoretic approach. *Mathematical Social Sciences*, *46*(1), 27–54.

Grupp, H. (1990). The concept of entropy in scientometrics and innovation research. *Scientometrics*, *18*(3-4), 219–239. doi:10.1007/BF02017763

Guimera, R., Mossa, S., Turtschi, A., & Amaral, L. N. (2005). The worldwide air transportation network: anomalous centrality, community structure, and cities' global roles. *Proceedings of the National Academy of Sciences*, *102*(22), 7794–7799.

Hall, B. H., Jaffe, A. B., & Trajtenberg, M. (2001, October). *The NBER Patent Citation Data File: Lessons, Insights and Methodological Tools* (NBER Working Papers No. 8498). National Bureau of Economic Research, Inc. Retrieved from http://ideas.repec.org/p/nbr/nberwo/8498.html

Hanneman, R. A. & Riddle, M. (2005). *Introduction to social network methods*. University of California, Riverside. Riverside, CA. Retrieved from http://www.faculty.ucr.edu/~hanneman/

Hu, Y. F. (2005). Efficient and high quality force-directed graph drawing. *The Mathematica Journal*, *10*, 37–71. Retrieved from http://www.mathematica-journal.com/issue/v10i1/contents/graph_draw/graph_draw.pdf

Jackson, M. O. et al. (2008). *Social and economic networks*. Princeton University Press Princeton.

Jacomy, M., Venturini, T., Heymann, S., & Bastian, M. (2014, June). Forceatlas2, a continuous graph layout algorithm for handy network visualization designed for the gephi software. *PLoS One*.

Jeong, H., Mason, S. P., Barabási, A.-L., & Oltvai, Z. N. (2001). Lethality and centrality in protein networks. *Nature*, *411*(6833), 41–42.

Jordan, C. (1869). Sur les assemblages de lignes. *J. Reine Angew. Math*, *70*(185), 81.

Jovanovic, M. M., John, M., & Reschke, S. (2010). Effects of civil war: scientific cooperation in the republics of the former yugoslavia and the province of kosovo. *Scientometrics*, *82*(3), 627–645. Retrieved from http://dblp.uni-trier.de/db/journals/scientometrics/scientometrics82.html#JovanovicJR10

Kanis, J. & Skorkovská, L. (2010). Comparison of different lemmatization approaches through the means of information retrieval performance. In P. Sojka, A. Horák, I. Kopeček, & K. Pala (Eds.), *Text, speech and dialogue* (Vol. 6231, pp. 93–100). Lecture Notes in Computer Science. Springer Berlin Heidelberg. doi:10.1007/978-3-642-15760-8_13

Karlovčec, M. & Mladenić, D. (2015). Interdisciplinarity of scientific fields and its evolution based on graph of project collaboration and co-authoring. *Scientometrics*, *102*(1), 433–454. doi:10.1007/s11192-014-1355-y

Karlovčec, M., Mladenić, D., Grobelnik, M., & Jermol, M. (2011). Visualizations of slovenian scietific community. In *Proceedings of the 14th international multiconference information society* (pp. 129–132).

Kastrin, A. & Karlovčec, M. (2012). Collaboration networks of slovenian researchers. In *In the proceedings of 6th international conference on information technologies and information society.*

Kernighan, B. W. & Lin, S. (1970). An efficient heuristic procedure for partitioning graphs. *Bell Sys. Tech. J. 49*(2), 291–308.

Kirkpatrick, S., Gelatt, C. D., & Vecchi, M. P. (1983). Optimization by simulated annealing. *SCIENCE*, *220*(4598), 671–680.

Klein, J. T. (1996). *Crossing boundaries: knowledge, disciplinarities, and interdisciplinarities.* University of Virginia Press.

Knoke, D. (1982). *Political networks: the structural perspective.* Sage Publications.

Koschützki, D., Lehmann, K. A., Peeters, L., Richter, S., Tenfelde-Podehl, D., & Zlotowski, O. (2005). Centrality indices. In *Network analysis* (pp. 16–61). Springer.

Krnc, M. & Škrekovski, R. (2015). Group centralization of network indices. *Discr. Appl. Math. 186*, 147–157.

Kronegger, L., Mali, F., Ferligoj, A., & Doreian, P. (2012). Collaboration structures in slovenian scientific communities. *Scientometrics*, *90*(2), 631–647. doi:10.1007/s11192-011-0493-8

Lebert, M. (2005). Project gutenberg, from 1971 to 2005.

Leskovec, J., Adamic, L. A., & Huberman, B. A. (2007, May). The dynamics of viral marketing. *ACM Trans. Web*, *1*(1).

Leskovec, J., Backstrom, L., Kumar, R., & Tomkins, A. (2008). Microscopic evolution of social networks. In *Proceedings of the 14th acm sigkdd international conference on knowledge discovery and data mining* (pp. 462–470). KDD '08. Las Vegas, Nevada, USA: ACM. doi:10.1145/1401890.1401948

Leskovec, J., Kleinberg, J., & Faloutsos, C. (2005). Graphs over time: densification laws, shrinking diameters and possible explanations. In *Proceedings of the eleventh acm sigkdd international conference on knowledge discovery in data mining* (pp. 177–187). KDD '05. Chicago, Illinois, USA: ACM.

Leskovec, J., Kleinberg, J., & Faloutsos, C. (2007, March). Graph evolution: densification and shrinking diameters. *ACM Trans. Knowl. Discov. Data*, *1*(1). doi:10.1145/1217299.1217301

Leskovec, J. & Krevl, A. (2014, June). SNAP Datasets: Stanford large network dataset collection. http://snap.stanford.edu/data.

Leskovec, J. & Mcauley, J. (2012). Learning to discover social circles in ego networks. In P. Bartlett, F. Pereira, C. Burges, L. Bottou, & K. Weinberger (Eds.), *Advances in neural information processing systems 25* (pp. 548–556). Retrieved from http://books.nips.cc/papers/files/nips25/NIPS2012_0272.pdf

Levenshtein, V. (1966). Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, *10*, 707.

Leydesdorff, L. (2006). Betweenness centrality" as an indicator of the "interdisciplinarity" of scientific journals. *Journal of the American Society for Information Science and Technology*, *58*, 1303–1309.

Leydesdorff, L. (2007). Mapping interdisciplinarity at the interfaces between the science citation index and the social science citation index. *Scientometrics*, *71*(3), 391–405. doi:10.1007/s11192-007-1694-z

Li, J., Tang, J., Zhang, J., Luo, Q., Liu, Y., & Hong, M. (2008). Arnetminer: expertise oriented search using social networks. *Frontiers of Computer Science in China*, *2*(1), 94–105. doi:10.1007/s11704-008-0008-9

Liben-Nowell, D. & Kleinberg, J. (2007). The link-prediction problem for social networks. *Journal of the American society for information science and technology*, *58*(7), 1019–1031. Retrieved from http://scholar.google.com/scholar.bib?q=info:U1ggap2AFkIJ:scholar.google.com/&output=citation&hl=en&as_sdt=0,5&as_vis=1&ct=citation&cd=0

Lip, S. Z. W. (2011). A fast algorithm for the discrete core/periphery bipartitioning problem. ArXiv: 1102.5511 (physics.soc-ph)

Lorenzen, M. & Andersen, K. V. (2009). Centrality and creativity: does richard florida's creative class offer new insights into urban hierarchy? *Economic Geography*, *85*(4), 363–390.

Lužar, B., Levnajić, Z., Povh, J., & Perc, M. (2014). Community structure and the evolution of interdisciplinarity in slovenia's scientific collaboration network. *PLoS ONE*, *9*(4). doi:10.1371/journal.pone.0094429

MacEachren, A. M. (2004). *How maps work - representation, visualization, and design.* Guilford Press.

Mali, F., Kronegger, L., & Ferligoj, A. (2010). Co-authorship trends and collaboration patterns in the slovenian sociological community. *Corvinus Journal of Sociology and Social Policy*, *1*(2), 29–50.

Marsden, P. (2000). Encyclopedia of sociology. In E. F. Borgatta & R. J. V. Montgomery (Eds.), (Chap. Social Networks." pp. 2727–2735). Macmillan.

Marshakova, I. (1981). Citation networks in information science. *Scientometrics*, *3*(1), 13–25. doi:10.1007/BF02021861

Masucci, A. P. & Rodgers, G. J. (2009). Differences between normal and shuffled texts: structural properties of weighted networks. *Advances in Complex Systems*, *12*(1), 113–129.

May, R. M. (1997, February). The scientific wealth of nations. *Science 7*, *275*(5301), 793–796.

Mcauley, J. & Leskovec, J. (2014, February). Discovering social circles in ego networks. *ACM Trans. Knowl. Discov. Data*, *8*(1), 4:1–4:28. doi:10.1145/2556612

Merriam-Webster Online. (2009). Merriam-Webster Online Dictionary. Retrieved from http://www.merriam-webster.com

Milgram, S. (1967). The small world problem. *Psychology today*, *2*(1), 60–67.

Mislove, A., Marcon, M., Gummadi, K. P., Druschel, P., & Bhattacharjee, B. (2007, October). Measurement and Analysis of Online Social Networks. In *Proceedings of the 5th acm/usenix internet measurement conference (imc'07)*. San Diego, CA.

Miyano, E. & Ono, H. (2011). Maximum domination problem. In *Proceedings of the seventeenth computing: the australasian theory symposium-volume 119* (pp. 55–62). Australian Computer Society, Inc.

Morillo, F., Bordons, M., & Gómez, I. (2003). Interdisciplinarity in science: A tentative typology of disciplines and research areas. *Journal of the American Society for Information Science and Technology*, *54*(13), 1237–1249. doi:10.1002/asi.10326

Nemeth, R. J. & Smith, D. A. (1985). International trade and world-system structure: a multiple network analysis. *Review (Fernand Braudel Center)*, *8*(4), 517–560.

Onnela, J.-P., Saramäki, J., Hyvönen, J., Szabó, G., Lazer, D., Kaski, K., ... Barabási, A.-L. (2007). Structure and tie strengths in mobile communication networks. *Proceedings of the National Academy of Sciences*, *104*(18), 7332–7336.

Perc, M. (2010). Growth and structure of slovenia's scientific collaboration network. *CoRR*, *abs/1004.4824*. Retrieved from http://arxiv.org/abs/1004.4824

Pescosolido, B. A. (2006). The sociology of social networks. In C. D. Bryant & D. L. Peck (Eds.), (Chap. The Sociology of Social Networks, pp. 208–217).

Porter, A. L., Cohen, A., David Roessner, J., & Perreault, M. (2007). Measuring researcher interdisciplinarity. *Scientometrics*, *72*(1), 117–147. doi:10.1007/s11192-007-1700-5

Porter, A. L. & Rafols, I. (2009). Is science becoming more interdisciplinary? measuring and mapping six research fields over time. *Scientometrics*, *81*(3), 719–745. doi:10.1007/s11192-008-2197-2

Porter, A. L., Roessner, D. J., & Heberger, A. E. (2008). How interdisciplinary is a given body of research? *Research Evaluation*, *17*(4), 273–282.

Proctor, C. H. & Loomis, C. P. (1951). Analysis of sociometric data. *Research Methods in Social Relations*, 561–586.

Rafols, I. & Meyer, M. (2010). Diversity and network coherence as indicators of interdisciplinarity: case studies in bionanoscience. *Scientometrics*, 263–287.

Rao. (1982, February). Diversity and dissimilarity coefficients: A unified approach. *Theoretical Population Biology*, *21*(1), 24–43. doi:10.1016/0040-5809(82)90004-1

Rip, A. & Courtial, J. P. (1984). Co-word maps of biotechnology: an example of cognitive scientometrics. *Scientometrics*, *6*(6), 381–400. doi:10.1007/BF02025827

Rosenfield, P. L. (1993). The potential of transdisciplinary research for sustaining and extending linkages between the health and social sciences. *Social science and medicine*, 1343–1357.

Salton, G. & McGill, M. J. (1986). *Introduction to modern information retrieval*. New York, NY, USA: McGraw-Hill, Inc.

Schummer, J. (2004). Multidisciplinarity, interdisciplinarity, and patterns of research collaboration in nanoscience and nanotechnology. *Scientometrics*, 425–465.

Seeley, J. R. (1949). The net of reciprocal influence; a problem in treating sociometric data. *Canadian Journal of Psychology Revue Canadienne de Psychologie*, (3), 234–240.

Seidman, S. B. (1983). Network structure and minimum degree. *Social Networks*, *5*(3), 269–287.

Serrano, M. Á., Boguñá, M., & Pastor-Satorras, R. (2006, November). Correlations in weighted networks. *Phys. Rev. E*, *74*, 055101. doi:10.1103/PhysRevE.74.055101

Small, H. (1973). Co-citation in the scientific literature: a new measure of the relationship between two documents. *Journal of the American Society for Information Science*, *24*(4), 265–269. doi:10.1002/asi.4630240406

Small, H. & Greenlee, E. (1986). Collagen research in the 1970s. *Scientometrics*, *10*(1-2), 95–117. doi:10.1007/BF02016863

Smith, D. A. & White, D. R. (1992, June). Structure and dynamics of the global economy: network analysis of international trade 1965-1980. *Social Forces*, *70*(4), 857–893. Retrieved from http://www.jstor.org/stable/2580193

Snyder, D. & Kick, E. L. (1979, March). Structural position in the world system and economic growth, 1955-1970: a multiple-network analysis of transnational interactions. *American Journal of Sociology*, *84*(5), 1096–1126. doi:10.2307/2778218

Stirling, A. (1998). On the economics and analysis of diversity.

Stirling, A. (2007). A general framework for analysing diversity in science, technology and society. *Journal of The Royal Society Interface*, *4*(15), 707–719. doi:10.1098/rsif.2007.0213

Stokols, D., Fuqua, J., Gress, J., Harvey, R., Phillips, K., Baezconde-Garbanati, L., ... Trochim, W. (2003). Evaluating transdisciplinary science. *Nicotine and Tobacco Research*, *5*, 21–39.

Tian, Y., Wen, C., & Hong, S. (2008). Global scientific production on gis research by bibliometric analysis from 1997 to 2006. *Journal of Informetrics*, *2*(1), 65–74. doi:10.1016/j.joi.2007.10.001

van den Besselaar, P. & Heimeriks, G. (2001). Disciplinary, multidisciplinary, interdisciplinary - concepts and indicators. Paper for the 8th conference on Scientometrics and Informetrics – ISSI2001, Sydney. Australia, July 16-20, 2001.

van Raan, A. F. J. & Tijssen, R. J. W. (1993). The neural net of neural network research - an exercise in bibliometric mapping. *Scientometrics*, *26*(1), 169–192. Retrieved from http://dblp.uni-trier.de/db/journals/scientometrics/scientometrics26.html#RaanT93

Wagner, C. S., Roessner, J. D., Bobb, K., Klein, J. T., Boyack, K. W., Keyton, J., ... Börner, K. (2011). Approaches to understanding and measuring interdisciplinary scientific research (idr): a review of the literature. *Journal of Informetrics*, *5*(1), 14–26. doi:http://dx.doi.org/10.1016/j.joi.2010.06.004

Wallerstein, I. (1974). *Capitalist agriculture and the origins of the european world-economy in the sixteenth century*. Studies in social discontinuity. Academic Press. Retrieved from http://books.google.com.bn/books?id=ieLtAAAAMAAJ

Wasserman, S. & Faust, K. (1994). *Social network analysis: methods and applications (structural analysis in the social sciences)*. Cambridge University Press.

White, S. & Smyth, P. (2003). Algorithms for estimating relative importance in networks. In *Proceedings of the ninth acm sigkdd international conference on knowledge discovery and data mining* (pp. 266–275). ACM.

Williams, K. & Durrance, J. (2008). Social networks and social capital: rethinking theory in community informatics. *The Journal of Community Informatics*, *4*(3). Retrieved from http://ci-journal.net/index.php/ciej/article/view/465

WIPO. (2007). Wipo patent report: statistics on worldwide patent activity.

Zhu, W. & Guan, J. (2013). A bibliometric study of service innovation research: based on complex network analysis. *Scientometrics*, *94*(3), 1195–1216. doi:10.1007/s11192-012-0888-1

Zhuhadar, L., Yang, R., & Nasraoui, O. (2012, December). Toward the design of a recommender system: visual clustering and detecting community structure in a web

usage network. In *Web intelligence and intelligent agent technology (wi-iat), 2012 ieee/wic/acm international conferences on* (Vol. 1, pp. 354–361).

Zitt, M., Ramanana-Rahary, S., & Bassecoulard, E. (2005). Relativity of citation performance and excellence measures: from cross-field to cross-scale effects of field-normalisation. *Scientometrics*, *63*(2), 373–401. doi:10.1007/s11192-005-0218-y

Bibliography

Publications Related to the Thesis

Journal Articles

Karlovčec, M. & Mladenić, D. (2015). Interdisciplinarity of scientific fields and its evolution based on graph of project collaboration and co-authoring. *Scientometrics*, *102*(1), 433–454. doi:10.1007/s11192-014-1355-y

Karlovčec, M., Mladenić, D., Grobelnik, M., & Jermol, M. (in press). Conceptualizing science using collaboration and competences. *Electronic Library*.

Conference Papers

Karlovčec, M. & Mladenić. (2012). Visualizations of business and research collaboration in slovenia. In *Information technology interface*.

Karlovčec, M. & Mladenić, D. (2013). Web application for generating sub networks of slovenian research collaboration. In *In the proceedings of 6th international conference on information technologies and information society* (pp. 8–14).

Karlovčec, M., Mladenić, D., Grobelnik, M., & Jermol, M. (2011). Visualizations of slovenian scietific community. In *Proceedings of the 14th international multiconference information society* (pp. 129–132).

Kastrin, A. & Karlovčec, M. (2012). Collaboration networks of slovenian researchers. In *In the proceedings of 6th international conference on information technologies and information society*.

Other Publications

Conference Papers

Karlovčec, M. (2011). Information extraction from receipts using machine learning. In *Information technology interface*.

Karlovčec, M., Córdova-Sánchez, M., & Pardos, Z. (2012). Knowledge component suggestion for untagged content in an intelligent tutoring system. In S. Cerri, W. Clancey, G. Papadourakis, & K. Panourgia (Eds.), *Intelligent tutoring systems* (Vol. 7315, pp. 195–200). Lecture Notes in Computer Science. Springer Berlin Heidelberg. doi:10.1007/978-3-642-30950-2_25

Biography

Mario Karlovčec was born on February 11, 1987 in Čakovec, Croatia. In 2005, he started his studies at the Faculty of Organization and Informatics, University of Zagreb, Croatia. He finished the BSc Information Systems program in 2008 under the supervision of prof. Mirko Čubrilo and the MSc Software Engineering program in 2010 under the supervision of Prof. Miroslav Bača.

In the fall of 2010 he started his graduate studies at the Jožef Stefan International Postgraduate School, Ljubljana, Slovenia. He enrolled in the PhD program Information and Communication Technologies under the supervision of prof. Dunja Mladenić. During his PhD studies he was a member of the Artificial Intelligence Laboratory at the Jožef Stefan Institute, Ljubljana, Slovenia. He collaborated on EU funded projects SYMPHONY (Orchestrating Information Technologies and Global Systems Science for Policy Design and Regulation of a Resilient and Sustainable Global Economy) and SOPHOCLES (Self-Organised information PrOcessing, CriticaLity and Emergence in multilevel Systems), and nationally funded projects Slovenian Science Atlas and Co-authorship networks of Slovenian scholars: Theoretical analysis and visualization user interface development. He was an intern at FOUR32C, New York in 2013 and attended the Eric & Wendy Schmidt Data Science for Social Good Summer Fellowship at the University of Chicago in 2015.