

Core-periphery dynamics in collaboration networks: the case study of Slovenia

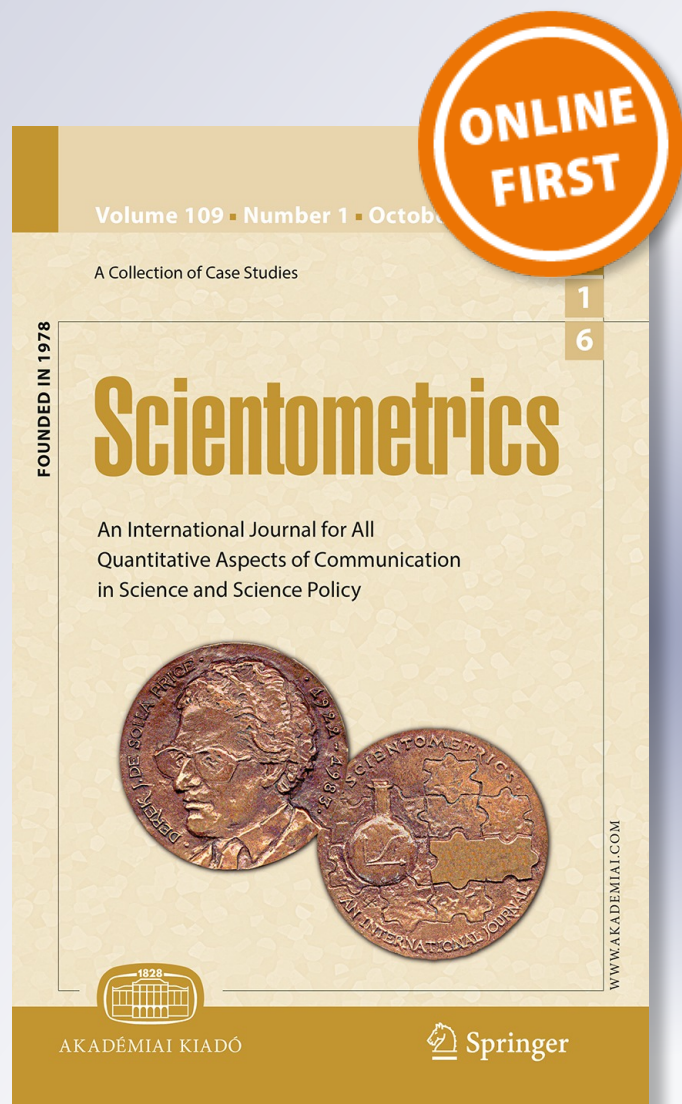
Mario Karlovčec, Borut Lužar & Dunja Mladenić

Scientometrics

An International Journal for all
Quantitative Aspects of the Science of
Science, Communication in Science and
Science Policy

ISSN 0138-9130

Scientometrics
DOI 10.1007/s11192-016-2154-4



Your article is protected by copyright and all rights are held exclusively by Akadémiai Kiadó, Budapest, Hungary. This e-offprint is for personal use only and shall not be self-archived in electronic repositories. If you wish to self-archive your article, please use the accepted manuscript version for posting on your own website. You may further deposit the accepted manuscript version in any repository, provided it is only made publicly available 12 months after official publication or later and provided acknowledgement is given to the original source of publication and a link is inserted to the published article on Springer's website. The link must be accompanied by the following text: "The final publication is available at link.springer.com".

Core-periphery dynamics in collaboration networks: the case study of Slovenia

Mario Karlovčec¹  · Borut Lužar² · Dunja Mladenić¹

Received: 12 February 2016
© Akadémiai Kiadó, Budapest, Hungary 2016

Abstract The paper presents analysis of core-periphery structure and transition dynamics of individuals between the core and periphery in collaboration networks of Slovenian researchers over 44 years. We observe the dynamics of individuals from three different aspects: regarding the length of the presence in the core-strength, the length of intervals of permanent presence in the core-stability, and the presence in different time periods-balance. We use clustering and classification machine learning techniques in order to automatically group individuals with similar dynamics of behaviour into common classes. We study collaboration networks of Slovenian researchers based on their publication records. The data we used comprises about 18,000 researchers registered in Slovenian national databases of researchers together with their publications from 1970 to 2013.

Keywords Core-periphery structure · Collaboration network · Machine learning · Clustering

Introduction

Analysis of collaboration between scientists gained a new dimension when tools for analyzing large complex networks have become more powerful and available to a bigger spectrum of researchers. One of the first such analyses was performed by Newman (2001),

✉ Mario Karlovčec
mario.karlovceec@ijs.si

Borut Lužar
borut.luzar@gmail.com

Dunja Mladenić
dunja.mladenic@ijs.si

¹ Jožef Stefan Institute, Ljubljana, Slovenia

² Faculty of Information Studies, Novo Mesto, Slovenia

who confirmed that scientists form a small world when they are connected in terms of publication co-authorship. Publication activity of scientists is one of the basic means by which scientific achievements are distributed and revealing patterns appearing in scientific collaboration is an important tool in discovering how to increase research efficiency. For instance, recently Ioannidis et al. (2014) confirmed that continuity of scientist's publishing activity results in findings with higher impact.

In the paper, we analyze collaboration between Slovenian scientists in the period from 1970 to 2013 based on the core-periphery partitions of the networks observing evolution of the networks and development of scientists' career paths. Partitioning of networks into local, global or intermediate substructures has proved to help understand the properties of networks in more detail. In recent years, studies of meso-scale structures are receiving considerable attention, with community detection being the leading topic. *Communities* in networks are subgraphs, which are "more dense as if some nodes from their neighborhood would be adjoined to them". As there is no strict definition of a community, various types of network partitioning into communities have been proposed (see e.g. Harenberg et al. (2014) for a survey).

In our analysis, we use another meso-scale network partitioning, called *core-periphery*. The nodes belonging to the *core* should be densely interconnected, while the nodes in *periphery* are connected sparsely. We study the dynamics of individuals migrating between core and periphery from three aspects: regarding the length of the presence in the core-strength, the length of intervals of permanent presence in the core-stability, and the presence in the core in different time periods-balance. We use clustering and classification machine learning techniques in order to automatically group individuals with the same dynamics behaviour into common classes.

The rest of this paper is organized as follows. In Sect. 2, we present related work, in Sect. 3 we describe the data, methodology and overview some basic properties of the analyzed networks. Core-periphery time-series analysis using machine learning methods is given in Sect. 4, while in Sect. 5, we discuss the results on few individuals from the dataset. Section 6 gives conclusions and some ideas for future work.

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 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 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 analyze scientific collaboration network in Slovenia. Studying collaboration between researchers and the structure of its relations has been intense in the last two decades: we refer e.g. to (Glänzel 2001a, b; Glänzel and Schubert 2005; Perc 2014). One of the first studies on Slovenian collaboration networks have been presented by Perc (2010a), where the author studied the evolution of the network, and Mali et al. (2010), where it has been 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 has been compared. Perc provided further results on slovenian publication data in Perc (2010b). Collaboration structures in Slovenian scientific communities were analyzed by Kronegger et al. (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. Recently, two studies about interdisciplinarity in Slovenian science using collaboration networks have also been introduced Karlovčec and Mladenić (2015); Lužar et al. (2014). Similar data was analyzed by Jovanović et al. (2010), where the authors investigated collaboration in producing scientific output in successor republics of Yugoslavia, and by Hâncean et al. (2014) investigating productivity of Romanian sociologists.

There are two types of the core-periphery concept: a discrete and a continuous model. In the discrete model nodes belong either to the core which has a high density of edges, or to the periphery which has a low density of edges. In continuous models each node obtains a value describing its "coreness". Computing these values is usually time consuming and the current methods can be applied to networks of moderate sizes only. Thus, as our networks are very large, we consider the discrete model in the rest of the paper.

Borgatti and Everett (1999) formalize an intuitive notion of a core-periphery structure and recommend the model that maximizes connectivity between the nodes in core and minimizes connectivity between the nodes in periphery, while it ignores the connectivity between the nodes of the two partitions. They used the recommended model and developed an 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 et al. (2002), which is one of the most commonly used tools for social network analysis. Boyd et al. (2006) used the same approach of pruning the search space and developed an algorithm based on simulated annealing Kirkpatrick et al. (1983), and the Kernighan-Lin algorithm Kernighan and 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.

Lip (2011) analyzed the most common definition of the core-periphery structure and deduced that the core can in fact be determined in time linear in the order of the network. His algorithm is based on greedy approach and solves the problem exactly and efficiently for both symmetric and asymmetric networks. Basically, it is shown that the optimal core-periphery partition of a network (in the discrete sense) is obtained by putting the first k nodes with the degree at least k into the core, and leave the others in the periphery. Due to its time efficiency it is suitable for arbitrarily large networks and gives an optimal solution. Therefore, we implemented Lip's algorithm and used its output partitions in our analysis.

Data description

Our analysis is conducted on scientific collaboration networks in Slovenia with two researchers being connected whenever they co-author a scientific publication; namely, we assign a weight of the number of publications coauthored by the two researchers to every edge. We focused on journal and conference publications. In particular, we consider 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, what results in a lower level of collaboration measured for the scientists mainly performing research in narrow science fields. We also want to emphasize that the data before the year 1990 may be incomplete, as a systematic recording of publications has been initiated later. However, in our opinion, the records for earlier years have been added in such an amount that starting an analysis in 1970 still gives a sample representative enough.

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 et al. 2007, 2008) and proved to be better for tracking career of researchers.

By the definition of the 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, in our model an extremely productive researcher is likely to be classified as a peripheral node if she mostly publishes as a single author or in collaboration with just a few other colleagues. On the other hand, a researcher with an average number of publications, which are coauthored by many authors, may enter the core in our model.

From the definition, one immediately infers that the number of nodes in the core grows slowly (from 7 in 1970 to 303 in 2010) compared to the growth of the entire network. In Fig. 1, a chart in the logarithmic scale confirming this fact is presented.

In addition to their bibliographic records, the data about Slovenian researchers also contain classification in which science branch an individual researcher is active. The researchers are classified according to Slovenian Research Agency¹ into seven branches: Natural sciences and mathematics, Engineering sciences and technologies, Medical sciences, Biotechnical sciences, Social sciences, Humanities, and Interdisciplinary studies. In Fig. 2, we compare representation of the science fields in the cores and in entire networks. Observe that the proportion of researchers in the cores and in entire networks are very similar during the observed time period, which is an indicator that the cores are well

¹ <http://www.arrs.gov.si/en/gradivo/sifranti/sif-vpp.asp>.

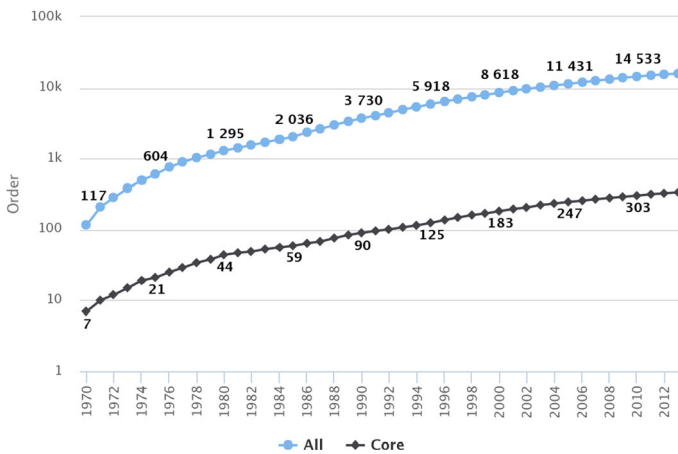


Fig. 1 The growth of the order of cores and the entire network in the logarithmic scale. The number of core nodes grows steadily, slightly slower than the network itself. In 1970 about 6 % of the nodes of the entire network were in the core, while in 2010 only 2 % of all the nodes are in the core

distributed within entire networks. There are however small deviations in the social sciences and humanities, as the publishing culture in these fields is undoubtedly different.

Clearly, 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 this phenomenon is becoming more and more rare. It can be seen in Fig. 3 that in the early years of 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 1990s. 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 career. 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 % of nodes in 2013. It is not surprising that already from 1978, all the core nodes were present in the giant component.

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 Fig. 4. As already the analysis of science fields suggested, the separation of the core nodes is high, which confirms that there are researchers in the core having no similar scientific interest at all. 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 networks.

Analysis of core-periphery dynamics

In this section, we describe an approach to automatically group individuals according to the dynamics of their transitions between the core and periphery throughout their research careers. We group transition patterns according to three aspects: *strength*, *stability*, and *balance*. The information about a node being in the core of a network in a particular

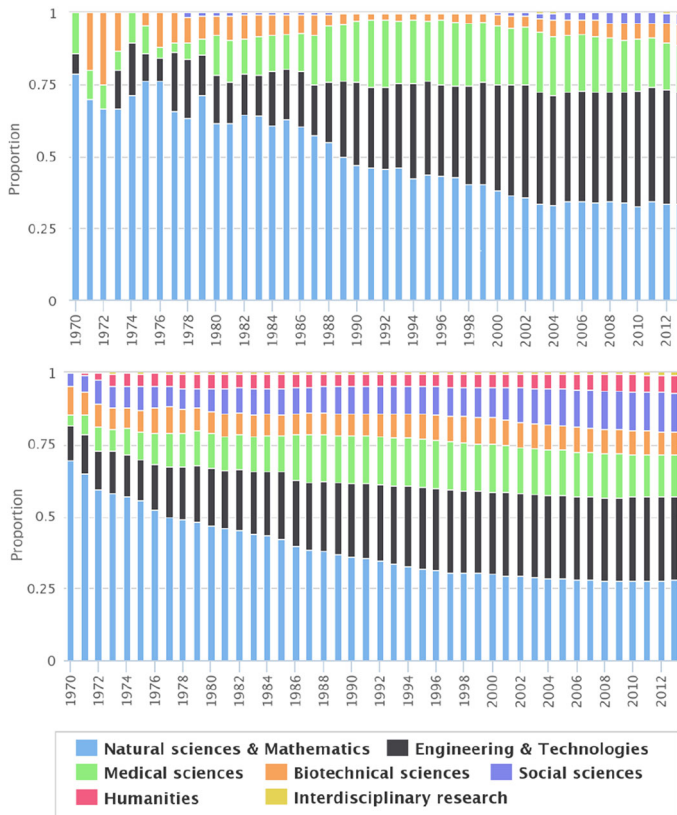


Fig. 2 Distribution of researchers within seven science fields for the core researchers (*top*) and the overall distribution (*bottom*). One can notice that the proportion of core researchers from social sciences and humanities is considerably smaller than in the entire network. Biotechnical sciences exhibit a drop in the number of core researchers in the decade of 1990–2000

instance of time is used to build a set of transition features. The analysis is performed through the following four steps (see Fig. 5):

1. computing the core-periphery partition of the collaboration networks for each year. In this step, we obtain 438 researchers that were at least in one year a part of the core, which are used in next steps of the approach;
2. constructing a feature set that quantifies discrete core-periphery time-series data based on researcher's participation in the core or periphery in different time points;
3. clustering to obtain groupings of individuals according to different degrees of core stability, balance and strength;
4. building classification models describing the clusters.

Computing the core-periphery partitions

In the first step, we apply Lip's algorithm (described in Sect. 2) to the 44 collaboration networks obtaining a core-periphery partition for each. Then, we use the partition data to determine a 44-dimensional binary vector $\mathcal{C}(r)$ for each researcher r appearing in any of the

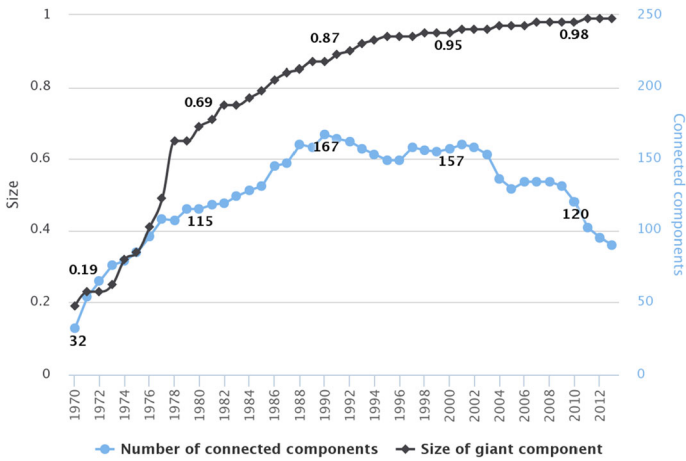


Fig. 3 The number of connected components together with the proportional size of the largest component in terms of the number of nodes. The size of the giant component has grown from 19 to 99 % during the observed four decades

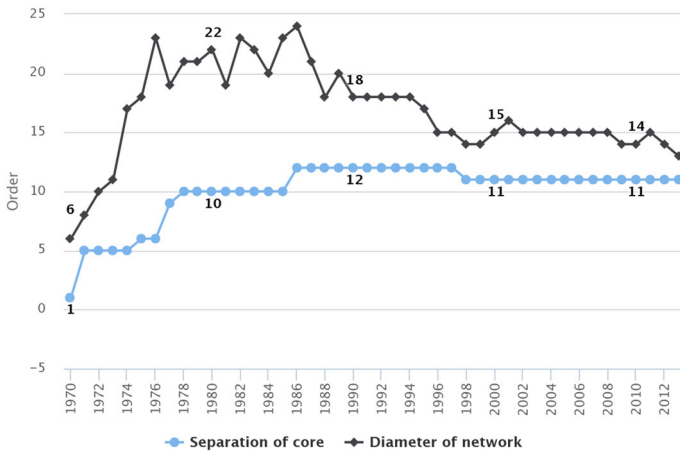


Fig. 4 The diameter and the maximum separation distance between the core nodes during the observed period. As the separation of core remains between 10 and 12 over the last three decades, we conclude that the core is a backbone of the network rather than its center

networks, where the i -th component of $\mathcal{C}(r) = 1$ if the researcher r is a member of the core in the i -th network, and $\mathcal{C}(r) = 0$ otherwise.

Feature construction

We experimented with about 20 features and finally selected the features listed in Table 1. The features were computed for core-periphery time-series for each researcher (see Fig. 6 for an example of a core-periphery time series for a researcher).

Using the selected features, we selected and grouped to clusters the researchers according to stability, balance, and strength of retention in the core. Selection of the

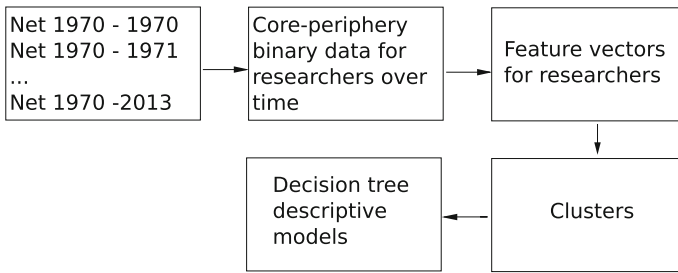


Fig. 5 Architecture of the proposed approach using the collaboration network structure as captured by the core-periphery analysis to study evolution of the network using machine learning techniques. The idea is to observe researchers based on their track record as seen from the core-periphery analysis. We capture the dynamics of individual researchers in feature vectors constructed from core-periphery data and use decision trees to describe typical behavioral patterns

Table 1 Descriptions of features derived from core-periphery time series

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 changes	Number of transitions
One ratio X/Y	Number of core time points in X th 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 X th 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

The focus is on dynamics of entering and leaving core of the network and the proportion of years in the core

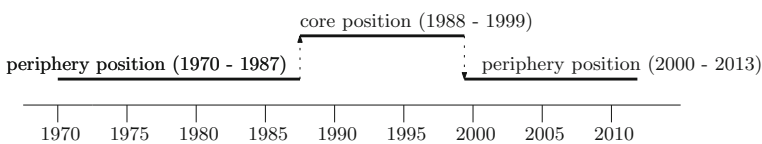


Fig. 6 An illustration of a core membership information for an arbitrarily chosen researcher. The researcher in this example has periphery position in the period from 1970 to 1987. In 1988, she enters the core and in 2000, she moves back to the periphery. This kind of behavior may be noticed for a number of researchers, with the period of being in the core varying in duration

features for clustering was done by applying expert knowledge and analyzing internal criteria, specifically maximizing intra-cluster similarity and minimizing inter-cluster similarity. We experimented with different sets of features and observed the composition of obtained clusters. The goal was to obtain clusters that clearly divide researchers according to the three sociological qualities we are observing. For a descriptive analysis using the 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.

Clustering according to stability, balance, and strength

The goal of this approach is to describe individuals in the networks according to dynamics of transitions between the core and periphery. To achieve it, we apply K-means clustering of the nodes described with the features listed in Table 1. As a result of clustering we obtain groups of nodes, which are similar according to some interpretable characteristics. Namely, we group the features in such a way that the resulting clusters represent groups of researchers with different core retention stability, core retention balance over time, and strength of the core position.

Stability

Stability describes continuity and length of retention in the core. After entering the core, highly stable individuals will retain in the core for significant amount of time, without frequent transitions between the core and periphery. Experimentation and feature engineering led to choosing the number of transitions between the 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. The 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 four clusters have been chosen, namely *low*, *medium low*, *medium high*, and *high* stability. The representative samples or cluster medoids are depicted in Fig. 7.

Figure 8 represents distribution into clusters for six science branches into which the researchers from the core are classified. The distribution in the branches contributing most of the researchers to the core mainly follows the overall distribution presented in the last column.

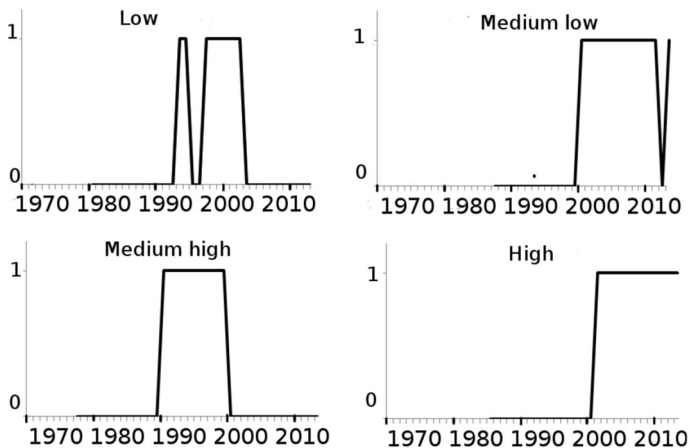


Fig. 7 Stability clustering medoids, where we observe four kinds of typical behavior. While some researchers are frequently entering and leaving the core, the others are entering only a couple of times or just once and staying in the core for a longer period

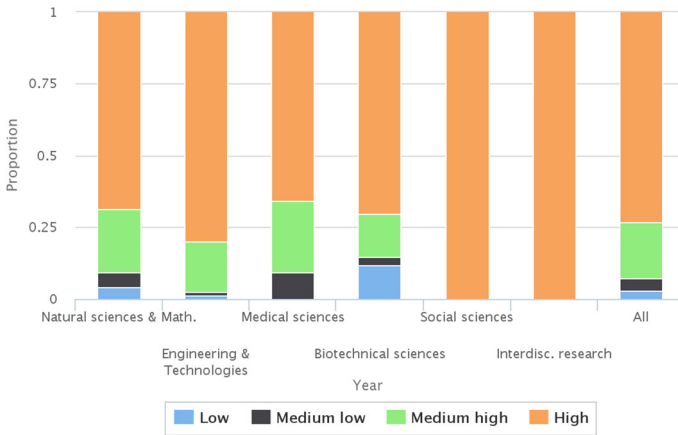


Fig. 8 The distribution of core researchers into clusters per science branch. The majority of researchers (73.3 %) that enter the core have high stability, with the highest proportion in Engineering and Technologies (notice that the statistics for Social sciences and Interdisciplinary research is based on a small sample)

Balance

Balance shows in which parts of the observed time period an individual is present 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 has been divided into 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 ratio $2/3$ has value 1.0 if an individual is present in the core during the complete second 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 has been present in the core for two time units in the fifth sixth part of the period and has been four time units in total in the core during the complete observed time period. Clusters based on the balance have been identified using the K-means clustering algorithm. Five clusters have been identified and labeled *start*, *start middle*, *middle*, *end middle* and *end*. The representative samples or cluster medoids are depicted in Fig. 9.

Figure 10 represents distribution into clusters for six science branches. Again, the distribution in the branches follows the overall distribution, but most of the researchers classified into the start cluster come from natural sciences and medicine, which is not surprising as other branches have developed later.

Strength

Strength represents the ratio of the time in the core compared to the time in periphery. Individuals with high strength are present 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 the core ratio divided by the total number of time units in the observed period. Clustering on the basis

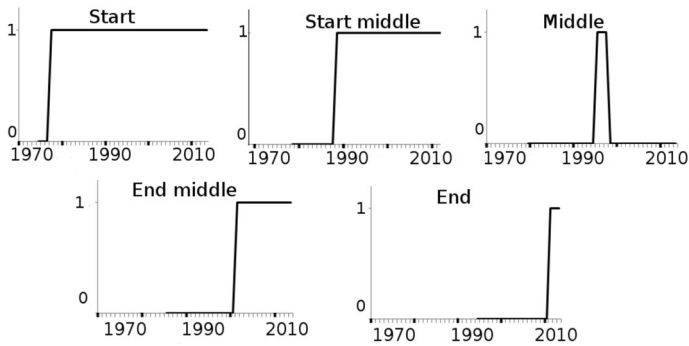


Fig. 9 Balance clustering medoids, where we observe five typical behaviors relative to the time of the observed period when a researcher was in the core

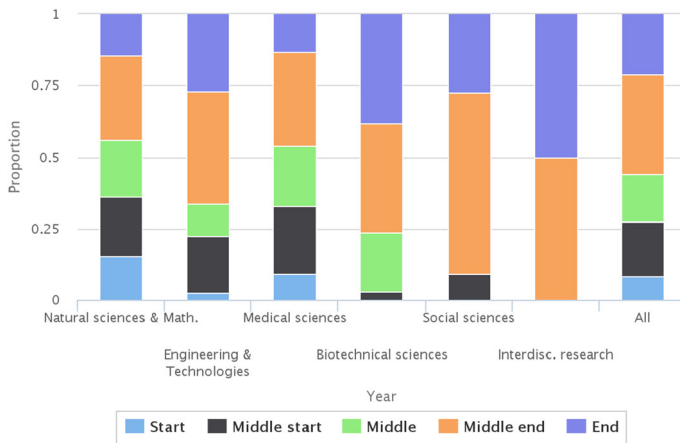


Fig. 10 The distribution of core researchers into clusters per science branch. Most of the researchers are in the core in the middle or towards the end of the observed period. The proportion of researchers in the core at the start is the largest for Natural sciences, about 18 %, and the lowest for Biotechnical sciences, where there is a large proportion of researches in the core at the end (notice that the statistics for Social sciences and Interdisciplinary research is based on a small sample)

of strength of core position resulted in four clusters, labeled *low*, *medium low*, *medium high* and *high*. In Fig. 11, plots of the cluster medoids are depicted.

Figure 12 represents distribution into clusters for six science branches. Similarly as above, the distribution in the branches follows the overall distribution. It is somehow surprising that most of the nodes are classified into the clusters medium high and high. This in particular means that it is very likely that a researcher which enters a core, will be present there for most of his career.

Descriptive classification modeling

In order to build models that produce descriptions of samples with respect to stability, balance, and strength, the obtained clusters are used to build classification models. In the previous steps, we used only the selected features, but here we are using all the features as

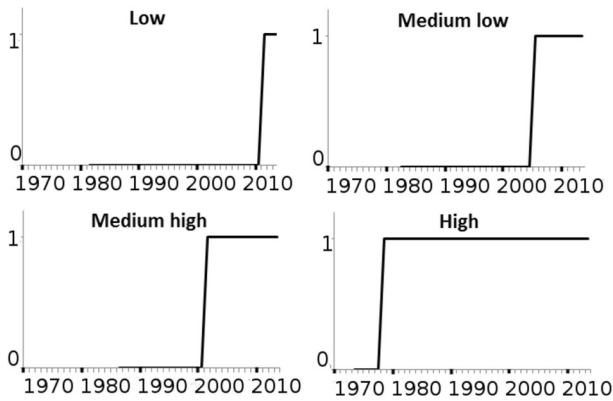


Fig. 11 Strength clustering medoids, where we observe four groups of typical behavior. While some researchers are in the core for rather short time or long time, the majority is in core for medium to high time

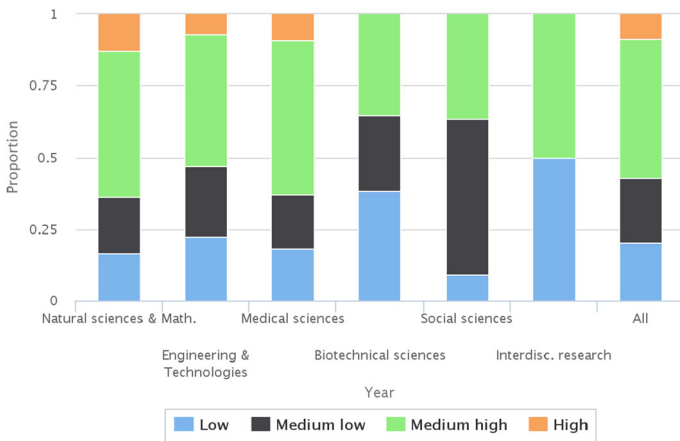


Fig. 12 The distribution of core researchers into strength clusters per science branch. Majority of the researchers from the core are more than medium of their active research time in the core. Being for high time in the core is observed for Natural sciences, Engineering and Technologies and Medical sciences. The proportion of researchers with low strength is the highest in Social sciences

the input for the classification algorithm. In this way, we see if the algorithm selects the features that are related to stability, balance, and strength, respectively. The samples were labeled according to the cluster membership and described with the complete set of features from Table 1. The labeled samples were used as the training data to build decision trees, which are descriptive classification models. The leafs of the trees give the class, while the paths from the root to the leafs define the rules for each class. To keep the models simple, we applied pruning and limited the depth to three layers, what slightly generalizes the models. The resulting models for balance and strength have classification accuracy below 100%, while the stability model classifies each sample to the same class as determined by the clustering.

Figures 13, 14 and 15 show decision trees for stability, balance and strength. Numbers in nodes represent the number of researchers in each step of decision tree. Decision tree has

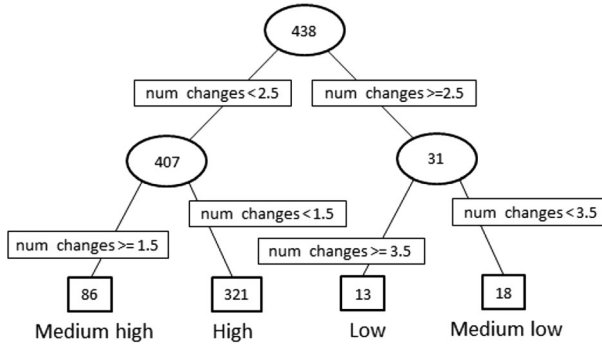


Fig. 13 The decision tree classification model for stability. In this tree ‘num changes’ was the only feature used for classifying the researchers into four groups with different stability

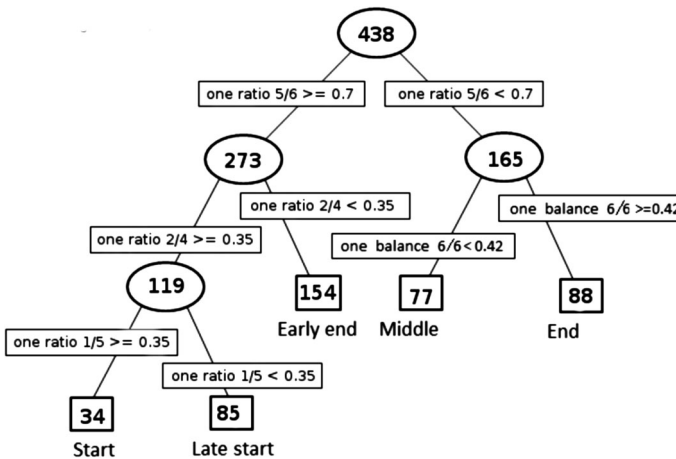


Fig. 14 The decision tree classification model for balance showing that the proportion of years in core in the last part of the observed period is the most informative for determining the balance

root in the top most node. It contains all researchers that were members of core at least once in their career. Every node of the tree has two successors, together containing all the researchers from their parent, divided to the left or the right successor regarding the value of the selected feature (the feature and the value interval is determined on the edge connecting the parent with its successor). Descriptions of features used in the decision trees are given in Table 1. For example, the root of the tree in Fig. 13 contains 438 researchers. All the researchers with the value of ‘num changes’ smaller than 2.5 are moved to the left child node, while all the others are moved into the right child node. Nodes that have no successors are called leafs. Leafs are the final step of decision tree and they determine the class of researchers.

The decision tree based on stability is plotted in Fig. 13. The classification model automatically chooses the ‘num changes’ as the single feature for modeling stability and determines thresholds for different stability classes. For example if the number of transitions is greater than three, the samples are classified into low stability class.

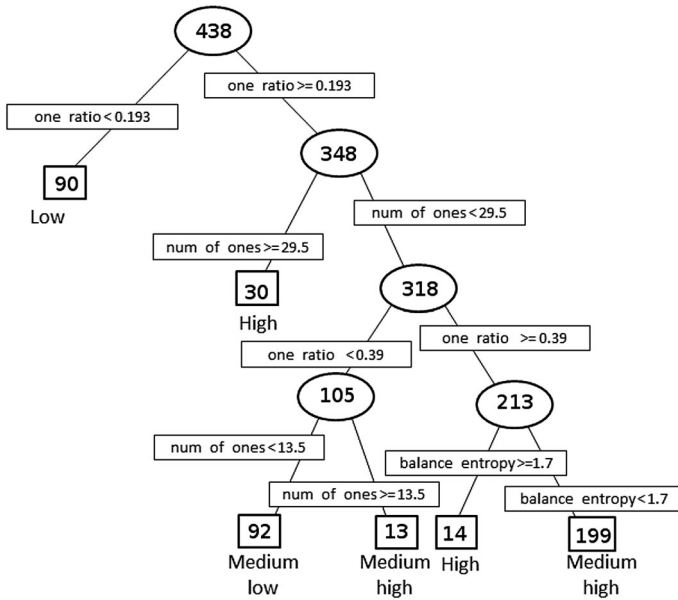


Fig. 15 The decision tree classification model for strength. This model uses ‘one ratio’, number of years in core and ‘balance entropy’ variables (described in 1), the researchers are classified into five different strength groups

In order to generate a descriptive classification model for balance, the decision tree was applied on instances labeled by balance clusters and described with the full set of features. The resulting decision tree is plotted in Fig. 14. Depending on the proportion of core years in the last part of observed period, the decision tree is divided into 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, ratios of core years in second fourth and last sixth of the period are used for splits. Ratio of core years in the first fifth is used for distinguishing between start and middle start class.

The decision tree classification model build using the cluster labels for strength and the full set of features is plotted in Fig. 15.

For classifying according to strength decision tree classification model automatically chooses one ratio and the number of years in core as the most important properties. In the first level is the core ration feature which immediately classifies researchers into low strength class if the ratio of core years is less than 19.3 % of total years in the collaboration network. If the ratio is above this value, absolute number of years in the core and balance entropy are used for determining the class.

Discussion

Researchers can have different career paths. 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 a collaboration network. We elaborated a method for automatic clustering of researchers into groups according to their core-periphery position in collaboration networks over time. Here, we illustrate 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 their core dynamics according to stability, balance and core strength. In Fig. 16, the time-series data for their core positions is presented. We looked up for the years in which the exemplary researchers obtained their PhD's to put results in the context of researcher's academic career. Note that this variable was not used in the analysis, but rather as an additional information for interpretation of exemplary results.

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 high stability, middle end balance and medium strength.

The researcher *B* entered the collaboration network in 1996 and obtained PhD in 2000. In 2003, she 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 as she 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 the observed period (end balance) and medium strength. Classification for researcher *A*, *B* and *C* is summarized in Table 2.

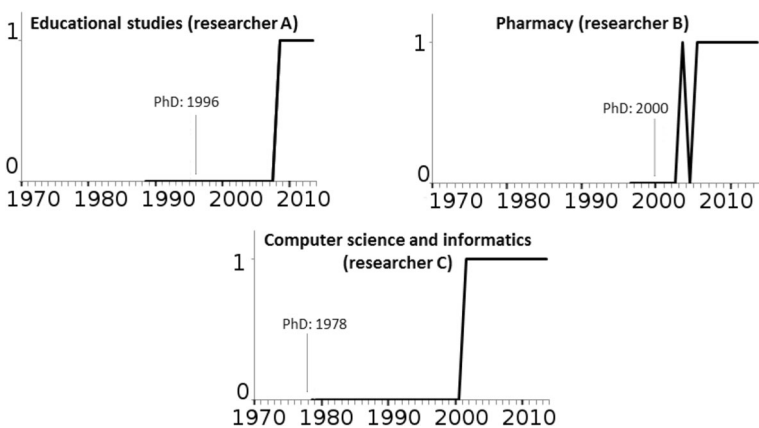


Fig. 16 Core-periphery dynamics time-series data for researchers *A*, *B* and *C* from collaboration networks, with the year of obtaining their PhD title marked

Table 2 Classification of researchers from the illustrative example

Researcher	A	B	C
Stability	High	Medium low	High
Balance	End	Middle-end	End
Strength	Middle low	Medium	Medium

Conclusion and future work

The proposed approach combines structural 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 44 years.

We have shown that researchers exhibit different patterns of core-periphery dynamics. Using our approach patterns of behavior can be identified and researchers can be automatically classified into categories. This can be used to interpret behavior of individual in a network that changes in time, as illustrated in the discussion section.

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 characteristic that can be described with time-series.

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. 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. Additionally we want to expand this line of research beyond descriptive modeling and try to predict the structural position of individuals.

Acknowledgments The authors thank two anonymous referees for their remarks which helped to improve the manuscript. This research was initiated in the framework of the project L7-4119 financed by Slovenian Research Agency ARRS. The second author acknowledges a partial support from ARRS Program P1-0383, Creative Core-FISNM-3330-13-500033, and the National Scholarship Programme of the Slovak Republic.

References

- Alba, R. D., & Moore, G. (1978). Elite social circles. *Sociological Methods and Research*, 7(2), 167–188.
- Barabási, A. L., Jeong, H., Nédá, 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.
- Borgatti, S., & Everett, M. (1999). Models of core/periphery structures. *Social Networks*, 21, 375–395.

- Borgatti, S., Everett, M. G., & Freeman, L. C. (2002). *UCINET 6 for Windows: Software for social network analysis*. Harvard: Analytic Technologies.
- 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.
- Brusco, M. (2011). An exact algorithm for a core/periphery bipartitioning problem. *Social Networks*, 33(1), 12–19.
- 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.
- Glänzel, W. (2001a). Coauthorship patterns and trends in the sciences (1980–1998): A bibliometric study with implications for database indexing and search strategies. *Library Trends*, 50(3), 461–473.
- Glänzel, W. (2001b). National characteristics in international scientific co-authorship relations. *Scientometrics*, 51(1), 69–115.
- Glänzel, W., & Schubert, A. (2005). The use of publication and patent statistics in studies of S&T systems. In H. F. Moed, W. Glänzel, & U. Schmoch (Eds.), *Handbook of quantitative science and technology research* (pp. 257–276). Dordrecht: Springer Netherlands.
- Goldberg, D. E. (1989). *Genetic algorithms in search, optimization, and machine learning* (1st ed.). Boston: Addison-Wesley Longman Publishing Co., Inc.
- Hâncean, M. G., Perc, M., & Vlăsceanu, L. (2014). Fragmented romanian sociology: Growth and structure of the collaboration network. *PLoS One*, 9(11), e113271.
- Harenberg, S., Bello, G., Gjeltema, L., Ranshous, S., Harlalka, J., Seay, R., et al. (2014). Community detection in large-scale networks: A survey and empirical evaluation. *Wiley Interdisciplinary Reviews: Computational Statistics*, 6(6), 426–439.
- Ioannidis, J. P. A., Boyack, K. W., & Klavans, R. (2014). Estimates of the continuously publishing core in the scientific workforce. *PLoS One*, 9(7), e101698.
- Jovanović, 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.
- Karlović, M., & Mladenović, D. (2015). Interdisciplinarity of scientific fields and its evolution based on graph of project collaboration and co-authoring. *Scientometrics*, 102, 433–454.
- Kernighan, B. W., & Lin, S. (1970). An efficient heuristic procedure for partitioning graphs. *Bell System Technical Journal*, 49(2), 291–307.
- Kirkpatrick, S., Gelatt, C. D., & Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, 220(4598), 671–680.
- Kronegger, L., Mali, F., Ferligoj, A., & Doreian, P. (2012). Collaboration structures in Slovenian scientific communities. *Scientometrics*, 90(2), 631–647.
- Leskovec, J., Kleinberg, J., & Faloutsos, C. (2007). Graph evolution: Densification and shrinking diameters. *ACM Transactions on Knowledge Discovery from Data*, 1(1), 2.
- 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, New York, NY, USA, KDD '08*, (pp. 462–470).
- Lip, S. Z. W. (2011). A fast algorithm for the discrete core/periphery bipartitioning problem. ArXiv preprint, article no. 1102.5511.
- 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), e94429.
- 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*, 2(1), 29–50.
- Mcauley, J., & Leskovec, J. (2014). Discovering social circles in ego networks. *ACM Transactions on Knowledge Discovery from Data*, 8(1), 4.
- Nemeth, R. J., & Smith, D. A. (1985). International trade and world-system structure: A multiple network analysis. *Review*, 8(4), 517–560.
- Newman, M. E. J. (2001). The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences*, 98(2), 404–409.
- Perc, M. (2010a). Growth and structure of Slovenia's scientific collaboration network. *Journal of Informetrics*, 4(4), 475–482.
- Perc, M. (2010b). Zipf's law and log-normal distributions in measures of scientific output across fields and institutions: 40 years of Slovenia's research as an example. *Journal of Informetrics*, 4(3), 358–364.
- Perc, M. (2014). The matthew effect in empirical data. *Journal of the Royal Society Interface*, 11(98), 20140378.
- Smith, D. A., & White, D. R. (1992). Structure and dynamics of the global economy: Network analysis of international trade 1965–1980. *Social Forces*, 70(4), 857–893.

- Snyder, D., & Kick, E. L. (1979). 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.
- Wallerstein, I. M. (1974). *The modern world-system I: Capitalist agriculture and the origins of the European world-economy in the sixteenth century (Studies in Social Discontinuity)*. New York: Academic Press.