DESCRIPTION OF THE RESEARCH PROJECT

27.1. **Scientific background, problem identification and objective of the proposed research**

Bibliographic services, including Web of Science/Knowledge, Scopus, CiteSeer, Zentralblatt Math, Google Scholar, DBLP, Math Sci, COBISS and others, provide data about scientific works (papers, books, reports, etc.). They are usually used by individual users for searching publications on selected topics, and by institutions for research evaluation and planning. They are used also in data analysis for bibliometric and scientometric research. For this purpose the data on selected topic are often transformed in the collection of bibliographic networks linking different entities (modes: works, authors, editors, journals, keywords, institutions, countries, languages, etc.).

A crucial step in construction of bibliographic networks is the entity identification/resolution (resolving synonymy / homonymy of entity names/labels). It is very important also in combining data from different sources. High precission in entity resolution task is required for obtaining high quality network data. We will develop new, high precission methods for entity resolution for specific types of entities based on interconnections (network) data. We will also create programs for conversions between bibliometric formats.

The obtained bibliographic networks are often large (thousands or even millions of entities). For their analysis efficient subquadratic algorithms should be developed, usually based on the assumption of the sparsity of networks.

An important tool for analysis of collections of networks are the derived networks obtained by combining network normalization (fractional approach) and multiplication of networks. Recently we provided a theoretical background for the fractional approach (Batagelj 2020) and showed how the temporal networks based on temporal quantities (Batagelj and Praprotnik 2016) can be applied in bibliometric analyses (Batagelj and Maltseva 2020). We intend to explore the possibilities offered by both approaches. The newly developed methods will be applied to analyze selected bibliographic data sets.

The data available in bibliographic data bases could be used also to provide higher level services for different kind of users. We will identify and explore interesting options and provide some prototype solutions.

27.2. **State-of-the-art in the proposed field of research and survey of the relevant literature**

From special bibliographies (BibTEX, EndNote) and bibliographic data bases it is possible to obtain data about works (papers, books, reports, etc.) on selected topics. A typical description of a work contains the following data: authors; title; publisher/journal; publication year and pages. In some sources, additional data are available including languages, classification of documents, keywords, authors' institution/country affiliation, lists of citations and abstracts. These data can be transformed into a collection of compatible two-mode networks on selected topics: works × authors; works × keywords; works × countries, and other pairs of characteristics describing works. Besides these networks, we can get also partitions of works by their publication years, partitions of works by journals, vector of number of pages, and, in some cases, (one-mode) citation networks.

When constructing any of these networks, the first task is to specify the nodes and which relations are linking them. In short, the network boundary problem (Marsden 1990) has to be solved. This

includes deciding whether a network is one-mode or two-mode and which node properties are important for the intended analyses. For specifying links, this amounts to answering a series of questions:
(1) Are the links directed?
(2) Are there different types of links (relations) to include?
(3) Can a pair of nodes be linked with multiple links?
(4) What are the weights on the links? (5) Is the network static, or is it changing through time?

Another problem occuring often when defining the set of nodes is the identification of nodes. The unit corresponding to a node can have different names (synonymy), or the same name can denote different units (homonymy or ambiguity). For example in the BibTEX bibliography from the Computational Geometry Database (Jones 2002) the same author appears under 7 different names: R.S. Drysdale, Robert L. Drysdale, Robert L. Scot Drysdale, R.L. Drysdale, S. Drysdale, R. Drysdale, and R.L.S. Drysdale. Insider information is needed to decide that Otfried Schwarzkopf and Otfried Cheong are the same person. At the other extreme, there are at least 57 different mathematicians with the name Wang, Li in the MathSciNet Database (TePaske-King and Richert 2001). Its editors have tried hard, from 1985, to resolve the identification of authors problem during the data entry phase. In the future, the problem could be eliminated by general adoption of initiatives such as using ResearcherID or ORCID.

Similary in the WoS work's references we find the following journal names: NUCLEIC ACIDS RES, NUCL ACIDS RES, NUCLEIC ACIDS RES S, NUCLEIC ACIDS RES S2, NUCL ACID RES, NUCL ACIDS RES S2, NUCL ACIDS S SER, NUCL ACIDS RES S, NUCL AC RES, NUCLEIC ACIDS RES S1, Nucleic Acids Res, NUCL ACIDS RES S1 or Q J R MET SOC, Q J R METEOROL SOC, Q J ROY METEOR SO S1, Q J ROY METEOR SOC, Q J ROY METEOR SOC B, QUART J ROY METEOR S, QUART J ROY METEOROL, QUART J ROY METEOROL SOC, QUART J ROYAL METEOR. The
immediate issue with all of these names is whether they denote the same journal or a small set of journals. There exists International Standard Serial Number (ISSN 2020), an international system for the identification of serial publications and other continuing resources. The problem is that the convention is not considered in WoS in the list of work's references. In resolving the journal identification problems, it is possible to use the Global serials directory (Ulrichsweb 2020) and Journal Abbreviation Sources (JAS 2020) and many other services and data sources.

The identification problem appears also when the units are extracted from plain text parts of documents. In producing keywords from the title or abstract of a work, the unimportant 'stopwords' must be eliminated first. The remaining (real) terms (words or phrases) are usually standardized by replacing them by a 'canonical' representative. For example, terms 'function', 'map', 'mapping' and 'transformation' in the mathematics literature can be considered as equivalent terms. Similar problem are equivalent terms from multi-lingual sources. To resolve this problem it is necessary to provide lists of equivalent terms or dictionaries.

Yet another source of identification problems stem from the grammar rules of the language used in a specific document. For example the action, 'go' can appear in the text in a variety of different forms including 'go', 'goes', 'gone', 'going' and 'went'. Resolving these grammar problems requires the use of stemming or lemmatization procedures from natural language processing toolkits such as NLTK (Bird et al. 2009; Perkins 2010) or MontyLingua (Liu 2004).

The general entity resolution problem is well elaborated in text mining literature (Buscaldi & Rosso, 2008; Talburt, 2011; Christen, 2012; Reitz & Hoffmann, 2013; Momeni & Mayr, 2016; Windham, 2018; Yadav & Bethard, 2019). In bibliometric analysis we need high precission solutions. We will try to develop them on the basis of specific structure (interconnections) of bibliographic data.

Public call for co-financing of research projects in 2020

From data collected from bibliographic databases we can construct different bibliographic networks. For example using the program WoS2Pajek we obtain from data collected from WoS the following two-mode networks: the authorship network WA on works × authors, the journalship network WJ on works × journals, the keywordship network WK on works × keywords, and the (one-mode) citation network Cite on works. We obtain also the following node properties: the partition year of works by publication year, the DC partition distinguishing between works with complete description (DC[w] = 1) and the cited only works (DC[w] = 0), and the vector of number of pages NP. Analyzing these networks we can get distributions of frequencies of different units (authors, journals, keywords) describing overall properties of networks. We can also identify the most important units (Cerinšek & Batagelj, 2015). An important tool in the analysis of linked (collections of) networks is the network multiplication that produces derived networks linking not directly linked sets of units – for example, the network AK = t(WA) · WK (t(A) is the transpose of matrix A) links authors to keywords (Batagelj & Cerinšek, 2013).

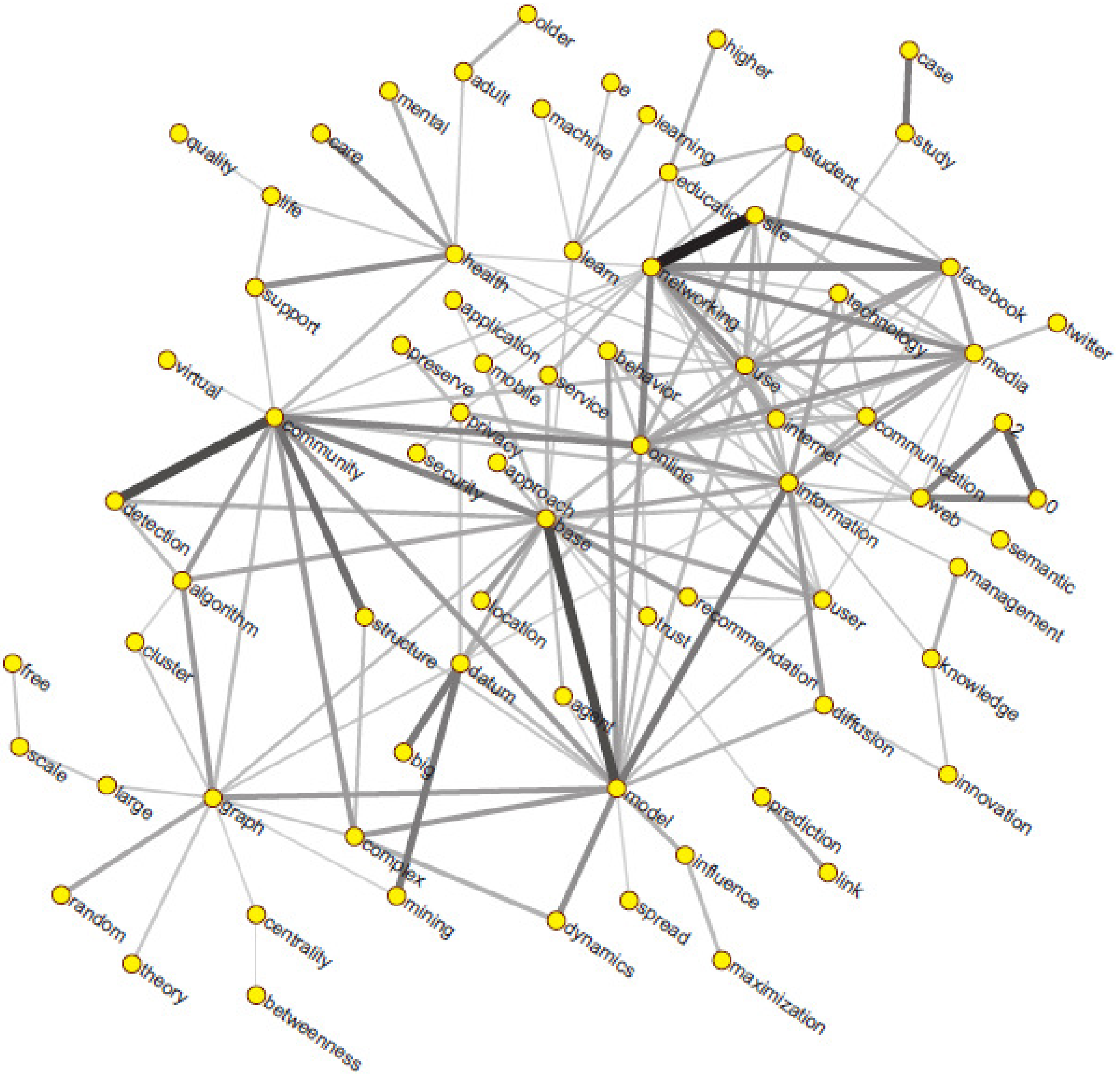


Figure 1

The fractional approach was proposed by Lindsey (1980). For example in the analysis of co-authorship the contributions of all co-authors to a work has to add to 1. Usually the contribution is then estimated as 1 divided by the number of co-authors. A normalization of network A obtained in this way is denoted n(A). An alternative rule, Newman's normalization, was given in Newman (2001) and Newman (2004) which excludes the self-collaboration. Recently several papers (Batagelj and Cerinšek, 2013; Cerinšek and Batagelj, 2015; Perianes-Rodriguez et al., 2016; Prathap and Mukherjee, 2016; Leydesdorff and Park, 2017; Gauffriau, 2017) reconsidered the background of the fractional approach. In the paper (Batagelj 2020) we proposed a theoretical framework based on the outer product decomposition to get the insight into the structure of bibliographic networks obtained with network normalization and multiplication.

In Figure 1 the main link island (the most connected nodes) in the network nKK of keywords co-appearance for the field of social network analysis (SNA) is presented. Keywords are lemmatized. nKK = t(n(WK)).n(WK), |W| = 70792, |K| = 32409, |E(nKK)| = 2799530.

A more detailed insight in the evolution of bibliographic networks is enabled by considering also the temporal information. In a temporal network, the presence and activity of its nodes and links and their values can change through time. In a description of a temporal network we have to provide information about these changes. Early applications of temporal networks were introduced in the project scheduling (CPM, Pert) in operations research (Moder & Phillips, 1970), in the transportation network analysis (Bell & Iida, 1997), and as constraints networks in artificial intelligence (Dechter, 2003). Also for data analytic tasks different approaches were proposed (Holme, 2015). Most frequently the cross-sectional approach is used in which the time consists of a finite number of time points (intervals). A time slice for a given time point t is a (static,ordinary) network for which the set of nodes/links contains all the nodes/links active in this time point. The analysis is performed on each slice separately using standard network analysis methods producing a 'time series' of results. Another interesting formalization are the time-varying graphs (Casteigts, Flocchini, & Quattrociocchi, 2012).

In the paper Batagelj and Praprotnik (2016) a longitudinal approach to analysis of temporal networks based on temporal quantities was presented. It is an alternative to the traditional cross-sectional approach. A temporal quantity is describing how the corresponding property is changing through time. The proposed approach has the following advantages: (1) it works for both discrete and continuous time (2) it internally (inside operations) adapts to the granularity of data (3) the result of a method is usually again a temporal network or a list of temporal quantities. The proposed approach can be applied to temporal bibliographic networks. It can be used also in other similar contexts. For the state-of-the-art on temporal networks see the forthcoming book Holme and Saramäki (2019). In a recent paper (Batagelj & Maltseva, 2020) we showed how the traditional bibliographic networks and information about the publication year can be transformed into the corresponding temporal (instantaneus or cumulative) network.

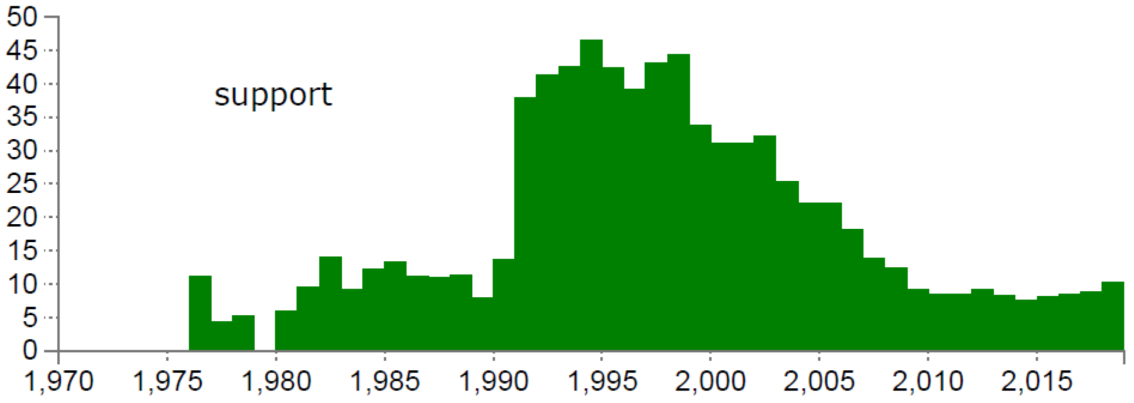


Figure 2

For example, we computed the proportion of the number of appearances of each keyword to the most frequent keyword appearance for each year based on the instantaneous temporal version of the network WK. This proportion normalizes the importance of selected keyword over time from 0 to 100%. In Figure 2 changes through years of importance in SNA literature of the keyword "support" are presented.

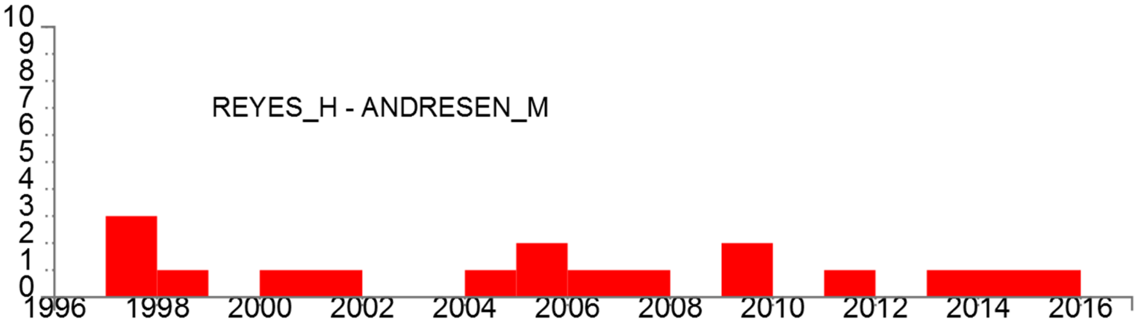


Figure 3

The instantaneous co-authorship network Coi is obtained as Coi = t(WAi) · WAi, where WAi is the instantaneous temporal version of the network WA. The weight Coi(a,b) is a temporal quantity describing the number of works authors a and b wrote together each year. Figure 3 presents the co-authorship of Reyes H. and Andersen M. in the field of peer-review (data from Batagelj, Ferligoj, Squazzoni, 2017)..

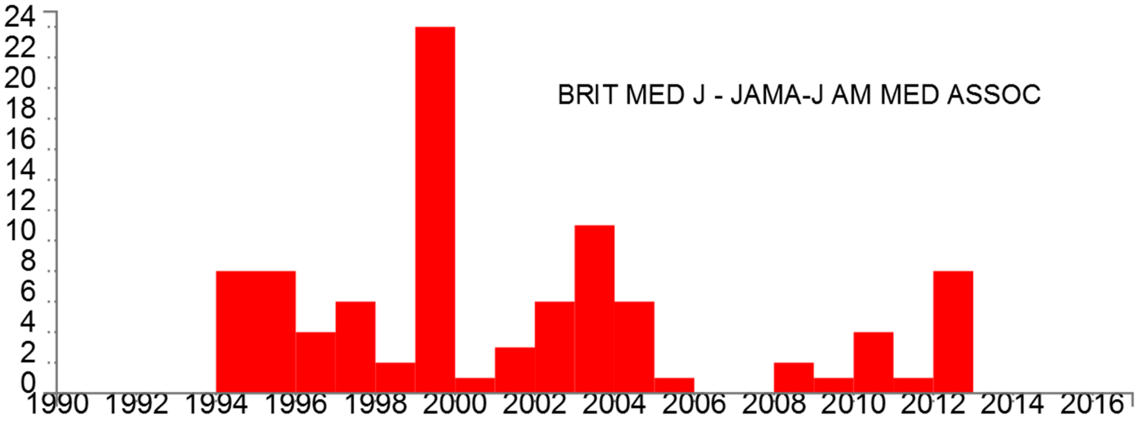
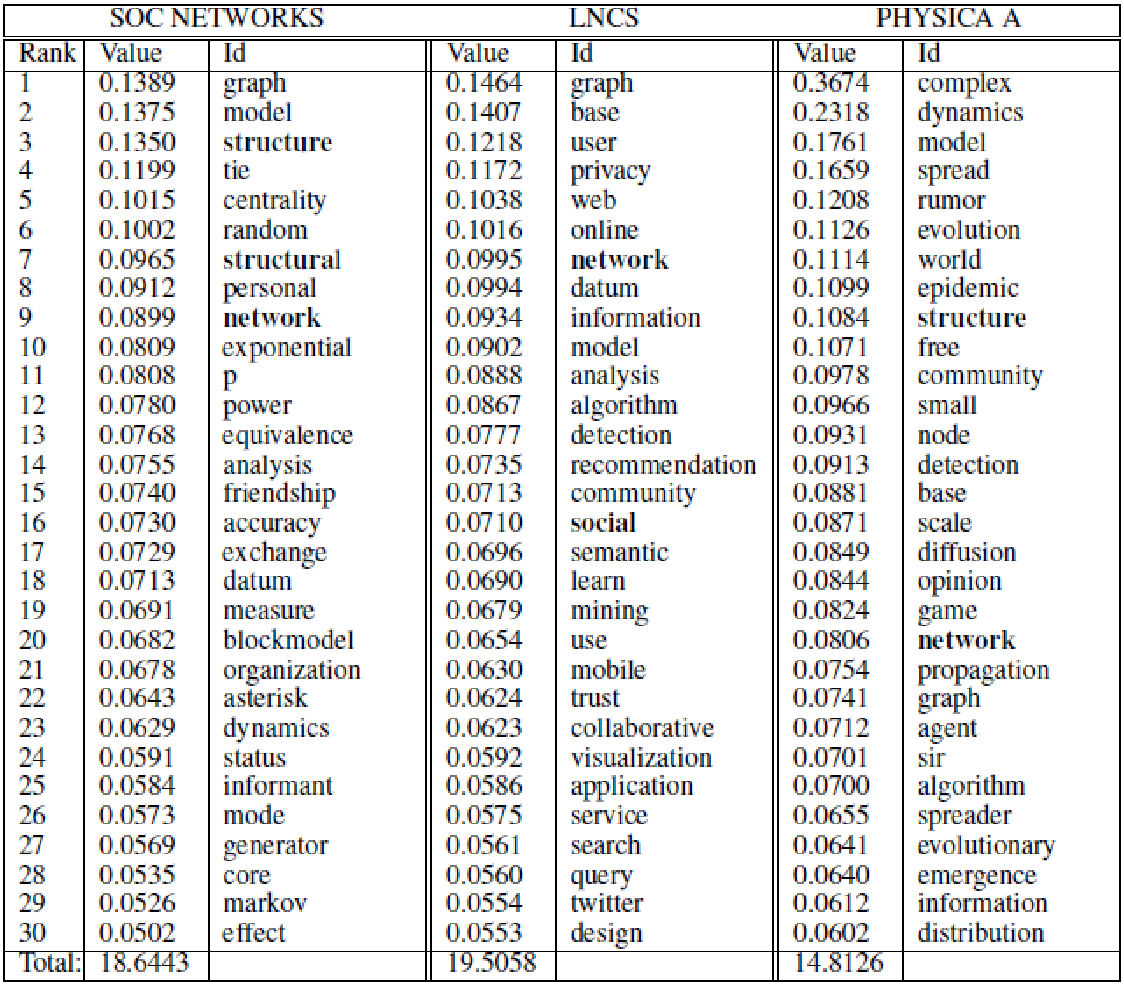


Figure 4

The derived instantaneous temporal network describing citations between journals is obtained as JCJ = t(WJi) · CiteI · WJc. Note that the first two factors are instantaneous and third network in the product is cumulative. The weight of the element JCJ(i,j) is equal to the number of citations per year from works published in journal i to works published in journal j. In a special case when i = j we get a temporal quantity describing self-citations of journal i. In Figure 4 the temporal quantity JCJ(BRIT MED J, JAMA-J AM MED ASSOC) in the field of peer-review is presented.

To identify keywords specific for a given group of authors or works we can apply the TF–IDF approach (Robertson, 2004). In Table 1 the most specific keywords for the field of SNA according to TF-IDF weights for selected three journals (groups of works) are presented.



SOC NETWORKS			LNCS		PHYSICA A	
Rank	Value	Id	Value	Id	Value	Id
1	0.1389	graph	0.1464	graph	0.3674	complex
2	0.1375	model	0.1407	base	0.2318	dynamics
3	0.1350	**structure**	0.1218	user	0.1761	model
4	0.1199	tie	0.1172	privacy	0.1659	spread
5	0.1015	centrality	0.1038	web	0.1208	rumor
6	0.1002	random	0.1016	online	0.1126	evolution
7	0.0965	**structural**	0.0995	**network**	0.1114	world
8	0.0912	personal	0.0994	datum	0.1099	epidemic
9	0.0899	**network**	0.0934	information	0.1084	**structure**
10	0.0809	exponential	0.0902	model	0.1071	free
11	0.0808	p	0.0888	analysis	0.0978	community
12	0.0780	power	0.0867	algorithm	0.0966	small
13	0.0768	equivalence	0.0777	detection	0.0931	node
14	0.0755	analysis	0.0735	recommendation	0.0913	detection
15	0.0740	friendship	0.0713	community	0.0881	base
16	0.0730	accuracy	0.0710	**social**	0.0871	scale
17	0.0729	exchange	0.0696	semantic	0.0849	diffusion
18	0.0713	datum	0.0690	learn	0.0844	opinion
19	0.0691	measure	0.0679	mining	0.0824	game
20	0.0682	blockmodel	0.0654	use	0.0806	**network**
21	0.0678	organization	0.0630	mobile	0.0754	propagation
22	0.0643	asterisk	0.0624	trust	0.0741	graph
23	0.0629	dynamics	0.0623	collaborative	0.0712	agent
24	0.0591	status	0.0592	visualization	0.0701	sir
25	0.0584	informant	0.0586	application	0.0700	algorithm
26	0.0573	mode	0.0575	service	0.0655	spreader
27	0.0569	generator	0.0561	search	0.0641	evolutionary
28	0.0535	core	0.0560	query	0.0640	emergence
29	0.0526	markov	0.0554	twitter	0.0612	information
30	0.0502	effect	0.0553	design	0.0602	distribution
Total:	18.6443		19.5058		14.8126	

Table 1

All three approaches, normalization, network multiplication, and temporal networks, can be combined and span a large, mostly unexplored space for development of new methods.

One of the tags assigned to an author is a research interest. For instance, Sicris/Cobiss database contains for each researcher a hierarchical research interest (RI). Eg. 1.01.05 represents (1) Natural Sciences, (01) Mathematics, (05) Graph Theory. However, in several cases this information is missing or it does not best represents current or main research interest of a researcher. In some cases there was an error committed while inputing the data. On the other hand the Scopus database of scientific journals contains information about the Scientific Fields (SF) that a journal is covering. For instance, the journal "Ars Mathematica Contemporanea" is classified as a Mathematical journal, covering four areas of mathematics: Algebra and Number Theory, Discrete Mathematics and Combinatorics, Geometry and Topology, and Theoretical Computer Science. Our intention is to find how RI correlate with SF and then use this bond to find outliers and errors in data. Independently we will use other independent tests to check the data, such as collaboration distance, or keywords. One expects that co-authors will tend to have similar research interests. Also a collection of keywords may be used to pinpoint a connection between research interest (RI) on the one hand and a scientific field (SF) on the other.

Another topic we intend to explore is the connection between cuts, islands and cores and research interests of authors in co-authorship networks. We intend to introduce new network invariants that will measure departure of an induced subgraph from a monochromatic one, where colors are research interests.

27.3. **Detailed description of the work programme**

WP1. Improvement of the software support for conversion of bibliographic data into networks
We already developed some programs for conversion of bibliographic data into networks (WoS2Pajek, DBLP, Zbmath, BiBTeX2Pajek). The most complex among them is WoS2Pajek. In the project we intend to extend the capabilities of WoS2Pajek with additional options (country, language, institution). Additionally we will develop a program for converting bibliographic data in other formats (BibTeX, Zbmath, DBLP, RIS) into WoS format. This will enable us to use WoS2Pajek for converting bibliographic data in any format into networks. In this way we will be also able to combine data obtained from different sources.

WP2. Methods and tools for the identification of units (entity resolution)
Construction of bibliographic networks on selected topic is an iterative process. We start with a data set of hits for a query with keywords characterizing selected topic. We create the corresponding set of networks and by an initial analysis we identify missing important units (for example, frequently cited works not included in the hits) and add them to the data set – the saturation of the data. In the following process of data cleaning we try to resolve the problem of identification of units – sometimes (for resolving ambiguity) by correcting the raw data and creating a new version of networks. The entity resolution task is well elaborated in general data mining literature. In the case of bibliographic data we can exploit (the additional knowledge about) networks structure in developing special high performance tools. We will try to develop such methods.

WP3. Theoretical research in bibliographic network analysis
An important tool in analysis of collections of linked networks (bibliographic networks are a special case) is network multiplication (Batagelj and Cerinšek, 2013) which enable us to compute derived networks. In order to consider each unit equally in the analysis of bibliographic networks, the fractional approach is used. Its theoretical background was proposed in our recent paper (Batagelj 2020). In papers Batagelj and Praprotnik (2016) and Batagelj and Maltseva (2020) we proposed a longitudinal approach to temporal network analysis and showed how it can be applied in analysis of temporal bibliographic networks. We will continue to explore the possibilities provided by these three approaches in the bibliographic network analysis.

WP4. Implementation of new methods for bibliographic network analysis with applications to real-life data
Bibliographic networks can be large (some houndred thousands or even millions of nodes). The developed software support should provide solutions that can deal also with such data efficiently – in range of some seconds or minutes. The basic support will be provided in some Python or R libraries (packages). An option is to integrate all these procedures into an user friendly tool – a kind of calculator with bibliometric data (similar to Pajek). To demonstrate the power of bibliographic (temporal) network analysis we will construct some collections of large networks on selected scientific fields and analyze them providing an insight into development and structure of the field.

WP5. Higher level bibliographic services
The bibliographic network analysis can be used also for checking the consistency of bibliographic data and services and providing automatic suggestions for missing data. It can be also used to build higher level bibliographic services for different kinds of users. For example:
Journal editors: selecting appropriate reviewers for a given paper; evaluation of reviewers.

Data base maintainers: automatic suggestion of keywords.
Authors: selecting the right journal to submit a given paper; keywords for a given paper.
Researchers: possible partners in a project.
Student: list of papers to read to enter a selected topic.

WP6. Publication of results.
We will monitor the work on the project on seminars. The obtained results will be reported on international scientific conferences and published in scientific journals. The developed software, its documentation and example data sets will be made available on Github as open-source. The annual and final reports are also part of this WP.

27.4. **Available research equipment over 5.000 €**

For most data sets a better laptop with 16 GB memory is sufficient. For very large data sets we will use the computing facilities available at the IAM.

27.5. **Project management: Detailed implementation plan and timetable**

The project is devided into six work packages (WP) with (sub)tasks:

WP1. Improvement of the software support for conversion of bibliographic data into networks
  a. Improvements in WoS2Pajek (additional options: language, country, ...), implementation in Python 3
  b. Conversions from other bibliographic formats (Scopus, DBPL, RIS, ...) into WoS format
WP2. Methods and tools for the identification of units (entity resolution)
  a. Identification of works
  b. Identification of persons
  c. Identification of journals
  d. Identification of keywords
  e. Identification of countries
WP3. Theoretical research in bibliographic network analysis
  a. New derived networks based on normalization and multiplication
  b. Temporal versions of derived networks
  c. New temporal quantities describing temporal bibliographic networks
  d. Clustering in temporal networks
WP4. Implementation of new methods for bibliographic network analysis with applications to real-life data
  a. Inclusion of new algorithms in libraries Nets, TQ and Biblio or implementation as separate programs
  b. Design and implementation of the "calculator"
  c. Creation of collections of bibliographic networks on selected topics and their analysis
WP5. Higher level bibliographic services
  a. Identification of problems and development of methods for maintainance of bibliographic data bases
  b. Identification of problems and development of methods for higher level user services
WP6. Publication (seminars, conferences, papers) of the results

The WPs are dynamically interdependent – a progress in one basic task extends the options that can be used in applications. We know the topics – we can immediately start with tasks.

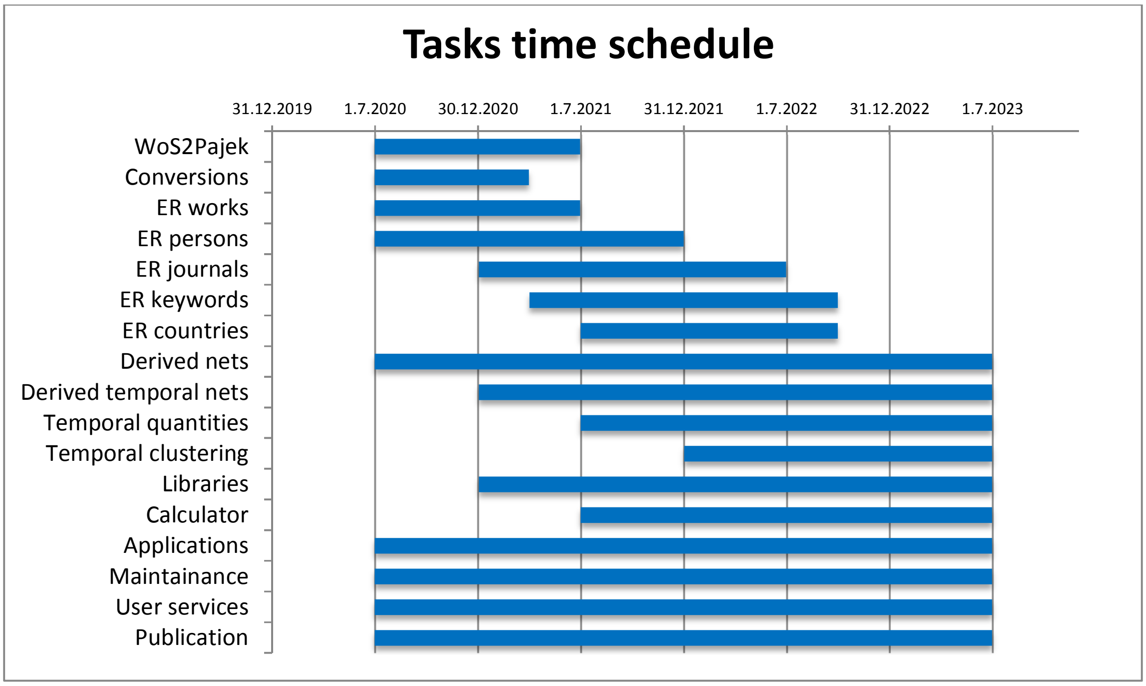


Figure 5

The time schedule of tasks is presented in Figure 5 based on the assumption that the project starts on July 1, 2020. The final time schedule will be shifted according to the real start of the project.

27.6 References

Batagelj, V, Cerinšek, M: On bibliographic networks. Scientometrics 96 (2013) 3, 845-864.

Batagelj, V., Doreian, P., Ferligoj, A., Kejžar, N.: Understanding Large Temporal Networks and Spatial Networks: Exploration, Pattern Searching, Visualization and Network Evolution. Wiley Series in Computational and Quantitative Social Science. Wiley, 2014.

Batagelj, V., Ferligoj, A., & Squazzoni, F. (2017). The emergence of a field: A network analysis of research on peer review. Scientometrics, 113(1), 503–532.

Batagelj, V., Praprotnik, S.: An algebraic approach to temporal network analysis based on temporal quantities. Social Network Analysis and Mining, 6(2016)1, 1-22

Batagelj, V.: On Fractional Approach to Analysis of Linked Networks. Scientometrics, 2020.

Batagelj, V., Maltseva, D.: Temporal Bibliographic Networks. Journal of Informetrics, 2020.

Bell, M. G. H., & Iida, Y. (1997). Transportation network analysis. Chichester: Wiley.

Bird, S., Klein, E., Loper, E.: NLTK - Natural Language Processing with Python. 2020. https://www.nltk.org/book/

Buscaldi, D., Rosso, P.: A conceptual density-based approach for the disambiguation of toponyms. Int. J. Geogr. Inf. Sci. 22, 3 (January 2008), 301–313.

Casteigts, A., Flocchini, P., Quattrociocchi, W., et al. (2012). Time-varying graphs and dynamic networks. International Journal of Parallel, Emergent andDistributed Systems, 27(5), 387–408.

Cerinšek, M., Batagelj, V.: Network analysis of Zentralblatt MATH data. Scientometrics, 102(2015)1, 977-1001.

Christen, P.: Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection. Springer-Verlag, Berlin, Heidelberg, 2012

Dechter, R. (2003). Constraint processing. San Francisco: Morgan Kaufmann.

Doreian, P., Batagelj, V., Ferligoj, A. (eds.): Advances in Network Clustering and Blockmodeling. Wiley, 2020. ISBN: 978-1-119-22470-9

Gauffriau, M.: A categorization of arguments for counting methods for publication and citation indicators. Journal of Informetrics, 11(2017)3, 672-684.

Holme, P. (2015). Modern temporal network theory: A colloquium. European Physical Journal B, 88, 234.

Holme, P., & Saramäki, J. (Eds.). (2019). Temporal network theory. Springer.

ISSN: 2020. https://www.issn.org/ , https://www.nuk.uni-lj.si/informacije/ISSN

JAS: Journal Abbreviation Sources. 2020. https://www.abbreviations.com/jas.php

Jones, B.: Computational Geometry Database. 2002. ftp://ftp.cs.usask.ca/pub/geometry/

Leydesdorff, L., Park, H.W.: Full and Fractional Counting in Bibliometric Networks. Journal of Informetrics Volume 11, Issue 1, February 2017, Pages 117–120.

Lindsey, D.: Production and Citation Measures in the Sociology of Science: The Problem of Multiple Authorship. Social Studies of Science, 10(1980)2, 145–162.

Maltseva, D., Batagelj, V.: Social network analysis as a field of invasions: bibliographic approach to study SNA development. Scientometrics, November 2019, Volume 121, Issue 2, pp 1085–1128

Maltseva, D., Batagelj, V.: Towards a Systematic Description of the Field Using Keywords Analysis: Main Topics in Social Networks. Scientometrics, 2020

Marsden, P.V.: Network Data and Measurement. Annual Review of Sociology, Vol. 16 (1990), pp. 435-463

Moder, J. J., & Phillips, C. R. (1970). Project management with CPM and PERT. Second edition. New York: Van Nostrand Reinhold Company.

Momeni, F., Mayr, P.: Evaluating Co-authorship Networks in Author Name Disambiguation for Common Names. In: Fuhr N., Kovács L., Risse T., Nejdl W. (eds) Research and Advanced Technology for Digital Libraries. TPDL 2016. Lecture Notes in Computer Science, vol 9819. Springer, (2016, p. 386-391

MontyLingua: A Free, Commonsense-Enriched Natural Language Understander for English. 2004/2020. http://alumni.media.mit.edu/~hugo/montylingua/

Newman, M.E.J.: Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. Physical Review E, 64(2001)1, 016132.

Perianes-Rodriguez, A., Waltman, L., Van Eck, N.J.: Constructing bibliometric networks: A comparison between full and fractional counting. Journal of Informetrics, 10(2016)4, 1178-1195.

Prathap, G., Mukherjee, S.: A conservation rule for constructing bibliometric network matrices. 2016. https://arxiv.org/abs/1611.08592

Reitz, F., Hoffmann, O.: Learning from the Past: An Analysis of Person Name Corrections in the DBLP Collection and Social Network Properties of Affected Entities. In Özyer, T., Rokne, J., Wagner, G., Reuser, A.H.P. (Eds.): The influence of technology on social network analysis and mining. Springer, Vienna 2013, p. 427-453

Robertson, S.: Understanding inverse document frequency: On theoretical arguments for IDF. Journal of Documentation, 60(2004)5, 503–520.

Talburt, J. (Eds.): Entity Resolution and Information Quality. Morgan Kaufmann, 2011

TePaske-King, B., Richert, N.: The Identification of Authors in the Mathematical Reviews Database. Issues in Science and Technology Librarianship No. 31 (Summer 2001), http://www.istl.org/01-summer/databases.html

Ulrichsweb: 2020. http://ulrichsweb.serialssolutions.com/

Yadav, V., Bethard, S.: A Survey on Recent Advances in Named Entity Recognition from Deep Learning models. October 2019, https://arxiv.org/abs/1910.11470

Windham, M.: Unstructured Data Analysis: Entity Resolution and Regular Expressions in SAS. SAS Institute, 2018