DESCRIPTION OF THE RESEARCH PROJECT

Higher-level bibliographic services

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

Bibliographic services, including Web of Science/Knowledge, Scopus, CiteSeer^X, zbMATH (formerly known as Zentralblatt MATH), Google Scholar, DBLP, MathSciNet, COBISS, arXiv 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 a selected topic are often transformed in the collection of bibliographic networks linking different entities (modes: works, authors, editors, journals, keywords, institutions, countries, languages, etc.) (Batagelj, Ferligoj, & Squazzoni, 2017).

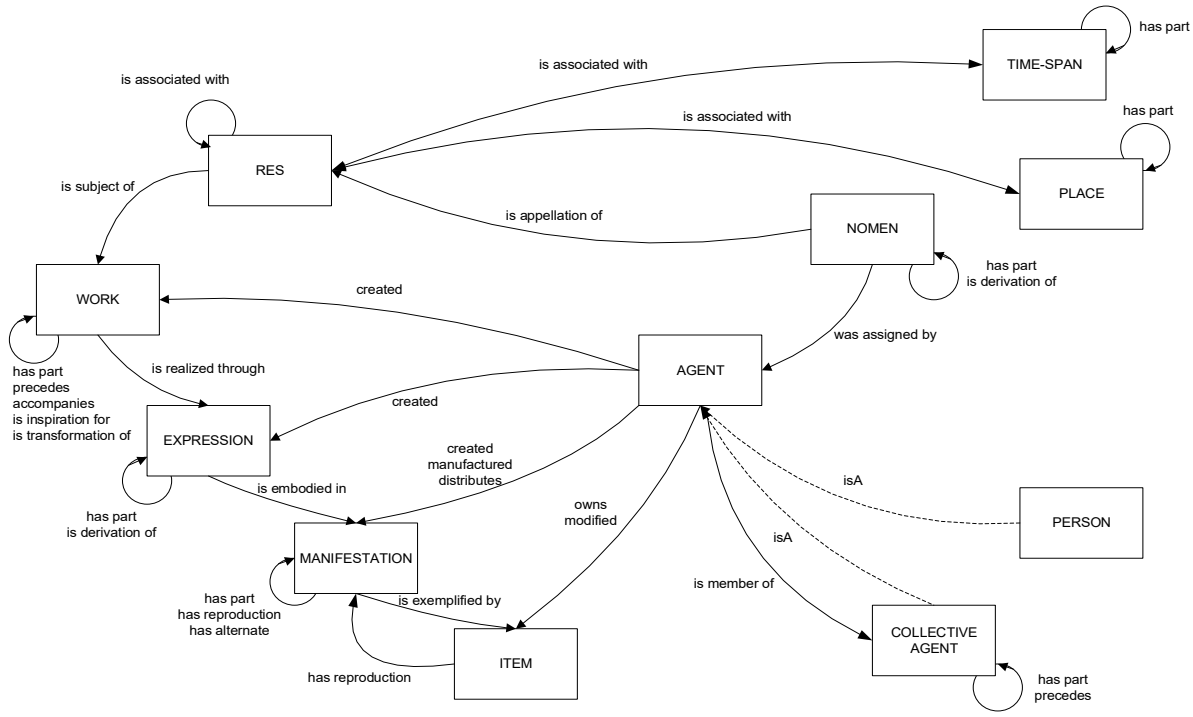

Figure 1: A scheme of basic types of entities in IFLA-LRM (Žumer, 2018).

According to the IFLA Library Reference Model (Žumer, 2018), the bibliographic universe is also represented as a multi-mode multi-relational network, enabling clustering on different levels, for example versions of papers (preprints, published) or editions of a monograph (see Figure 1).

The data collected in different bibliographic databases can be used to provide higher order bibliographic and bibliometric services (such as: what to read (contact/visit)? – a list of important articles/books (authors, institutions) on selected topic; where to publish? – a list of journals suitable for the publication of an article; reviewer selection – a list of reviewers suitable for a submitted article; career application – a candidate's activity report draft; etc.) for different types of users (students, researchers, teachers, decision makers, funding agencies, research institutions, database managers, etc.). The main goal of the project is the identification of potential higher order services and development of some prototype solutions. To support this goal we

have to provide high quality data often obtained by combining data from different databases. We also have to develop new algorithms for some analytical problems.

Most of the higher level services will be based on domain analysis – a field that is based on theoretical work by Birger Hjørland (e.g., Hjørland & Albrechtsen, 1995; Hjørland, 2002). Smiraglia (2013) sees domain analysis in the context of knowledge organization as set of techniques for extraction and analysis of semantic intellectual content of a coherent group. As pointed out by Gutierres Castanha and Wolfram (2018), domain analysis thus enables identification of various aspects of a research area, e.g. trends, patterns, main ideas and authors. As such, it is concerned with the characteristics of communication within a domain (particularly, scholarly communication).

The data obtained from the established bibliographic services and databases stated above deal primarily with peer-reviewed and published scientific works, and are usually well structured, standardised and of high quality. Due to peer review and publication processes such data reflect the "verified edge of the science" with a lag of 3 months or usually much more. In rapidly developing research fields (e.g. artificial intelligence, COVID-19 related research, etc.) these data resemble rather "old news".

On the other hand, the primary sources of the newest research results are preprints. There are several established public preprint archives (HAL, arXiv, bioRxiv, etc.). The data authors usually provide for preprints include title, list of authors, abstract and a PDF file of a paper. Sometimes preprints are published in several repositories and in different versions. It is quite a challenge to collect (by web scraping and PDF parsing) and match (advanced clustering) all these publications and provide at least basic standardization. Fortunately, this work is already being done and current datasets have been on disposal for research purposes. Most notable examples include Semantic Scholar Open Research Corpus (SSORC) and its cleaner derivative S2ORC (updated monthly) by Allen Institute of AI (Lo, Wang, Neumann, Kinney, & Weld, 2020; L. L. Wang et al., 2020). Recently, the extract of the latter, CORD-19 database specialized on COVID-19 related papers and preprints, sparked huge interest in AI and text mining research communities.

Clearly, the data in S2ORC and CORD-19 databases are significantly less standardised and structurally consistent than the data from established bibliographic services of published papers. This opens new research challenges in bibliographic network data enhancement and data analysis. In the future, such methods could find applications in various search and analytic engines and in a wider context than scientific literature (blogs, news articles, etc.)

A crucial step in construction of bibliographic networks is the entity (works, authors, etc.) identification/resolution (resolving synonymy/homonymy of entity names/labels). It is a necessary step in combining data from different sources. High precision in entity resolution is required for obtaining high quality network data. One of the goals of the project is to start developing robust and high precision methods for entity resolution for specific types of entities, based on their interconnections (network) data. Our methods will aim at entity identification across all major bibliographic formats and types of data from services and databases stated above and will build on previous research, such as Freire, Borbinha, and Calado (2012); Freire (2014).

An important tool for analysis of collections of networks are the derived networks obtained by combining network normalisation (fractional approach) and multiplication of networks (Maltseva & Batagelj, 2019, 2020). Recently we provided a theoretical background for the fractional approach (Batagelj, 2020) and showed how the temporal networks based on temporal quantities (Batagelj & Praprotnik, 2016) can be applied in bibliometric analyses (Batagelj & 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.

In particular we plan to carry out various bibliographic analysis on open data provided by the highest ranking Slovenian international scientific journals (Ars Mathematica Contemporanea, Acta Chimica Slovenica, etc.). This will be carried out in collaboration with their editors.

Other prototype analyses (methodologies) will include tools and analyses of current trends in selected research (sub-)fields (most probably information science, mathematics, network analysis, artificial intelligence, COVID-19 research, etc.)

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. In the past we have developed some smaller specialized libraries (Nets, TQ and Biblio) in Python and an established tool Pajek. Due to the availability of various algorithms and tools in different programming languages, typical bibliographic network analysis consists of mixing R (statistics, visualisations), Python (text processing, specific algorithms) and Pajek (large network analysis, closed source, based on Pascal/Delphi). The nature of development of bibliographic network analytic methods involves a lot of interactive work with data and many trials and errors. With big datasets containing tens or hundreds of gigabytes of data, the speed and efficiency of execution is crucial. Hence the core algorithms have to be implemented in a fast C-like programming language which is also easy to use and easy to maintain. The obvious candidate is the Julia programming language which is steadily gaining its reputation as a fast data analytics language (Bezanson, Edelman, Karpinski, & Shah, 2017). Our goal is to start developing an open source Julia library for bibliographic network analysis that can primarily be used directly, but can also be integrated with packages in R and Python. In particular, integration of code/packages/modules from other languages like C/C++, Python and R is significantly easier compared to other languages and often quite straightforward.

We believe that our methods, algorithms, tools and the derived high-quality data will provide the foundation of higher level services for different kinds of users (researchers, planners, publishers, etc.).

In summary, the objectives of the project include:

- Identify potential higher level services and carry out several prototype bibliographic analyses and develop related tools, motivated by needs of selected end-users (editors, scientists on specific fields, funding agencies, research institutions, etc.).
- Development of methods and algorithms for high quality bibliographic entity resolution based on bibliographic network analysis. This will enable us to implement processing pipelines that can be applied on bibliographic databases in order to obtain periodically refreshed high quality and up to date bibliographic data (including preprint data).
- Further development of methodologies and algorithms for analysis of bibliographic networks, based on our past research (2-mode networks, fractional approach, temporal networks and temporal quantities) motivated by specific types of analyses with emphasis on how the science is developing in "real-time" (based on preprints).

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

Although bibliometric approach is only one of 11 proposed by identified by Hjørland (2002) for domain analysis, it is the most commonly used approach (e.g., Smiraglia, 2013, 2015; Chen & Xiao, 2016; S. Wang, 2019; Gutierres Castanha & Wolfram, 2018; Agrahari, Chaudhary, & Singh, 2018). However, there are different ways of limiting the coherence of a group in domain analysis, e.g. with the starting point being a seminal volume (Smiraglia, 2015), most prominent output (Smiraglia, 2013) or articles in a particular journal (Gutierres Castanha & Wolfram, 2018; Agrahari et al., 2018). As the limits of a domain are determined by the examined frame of publications, it would be benefitial to study publications in different contexts/databases, taking into account both citation indexes (e.g. Scopus/Web of Science), as well as broader collections, e.g. COBISS.

Despite various bibliometric analyses of scholarly publication in Slovenia, they were in large part made at macro-level (e.g., Kronegger, Mali, Ferligoj, & Doreian, 2012; Karlovčec & Mladenić, 2015; Karlovčec, Lužar, & Mladenić, 2016; Lužar, Levnajić, Povh, & Perc, 2014; Bartol, Budimir, Dekleva-Smrekar, Pusnik, & Juznic, 2014). Therefore, true domain analyses in Slovenian context are lacking. We propose such an analysis, taking into account all of the researchers that have a particular field of research assigned to them. As interdisciplinary fields can be of greater interest, one possibility would be e.g. analysis of publications by authors to whom the field 5.13 – Social Sciences/Information and Library Science is assigned. Such an analysis would e.g. study the coherence of the group, its research topics, patterns and trends of publication, collaboration and citation.

From special bibliographies (BibTEX, EndNote) and bibliographic databases 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. This 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 also get 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 BibTEXbibliography 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 & Richert, 2001). Its editors have tried hard, from 1985, to resolve the identification of the author's problem during the data-entry phase. Significant growth of contributions by Chinese scientists and their full name similarity in roman transcriptions adds additional complexity to the problem. In the future, the problem could be eliminated by general adoption of initiatives such as using ORCID.

The author name resolution becomes even more challenging when data from preprints are considered, as the process of entering authors is even less controlled. A preliminary analysis of the CORD-19 database presented us with several additional challenges. In a small percentage of papers the authors are wrongly parsed, often due to special characters, as not all web pages use UTF-8 encoding and some use HTML special character codes (like `&Aacute;`). Sometimes

this can be fixed by certain replacements. Sometimes web pages that are being scraped contain badly entered data. We have found that about 3% of article records of the World Health Organization repository have long lists of author names and multi word surnames which are just separated by commas and thus full author's names are not deterministically separable. There are challenges with long multi-word family surnames (e.g. Portuguese, Spanish ones), positioning of "from family or location" adjectives in different languages (von, van der, de, der, della, etc.) and generational name additions (like Jr., Sr., 2nd, III). Sometimes papers are obtained from several sources and with different versions, even with different numbers of authors. Some problems can be addressed by improved parsing, but sometimes the parsed results are unusable and have to be omitted from further analyses.

Similarly 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. ISSN is an international identifier, but unfortunately rarely used in databases. In resolving the journal identification problems, it is possible to use the Global serials directory *Ulrichsweb* (2022) and *Journal Abbreviation Sources* (2022) 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 'stop words' 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 problems 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, Klein, & Loper, 2019; 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, 2019; Yadav & Bethard, 2019). In bibliometric analysis we need high precision solutions. We will try to develop them on the basis of specific structure (interconnections) of bibliographic data. From data collected from bibliographic databases we can construct different bibliographic networks (Batagelj, Doreian, Ferligoj, & Kejžar, 2014). For example using the program `WoS2Pajek` we obtain from data collected from WoS (Web of Science) the following two-mode networks: the authorship network WA on works $\times$ authors, the journalship network WJ on works $\times$ journals, the keywordship network WK on works $\times$ keywords, and the (one-mode) citation network $Cite$ on works. We also obtain 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 $AK = t(WA) \cdot WK$

($t(A)$ is the transpose of matrix A) links authors to keywords (Batagelj & Cerinšek, 2013).

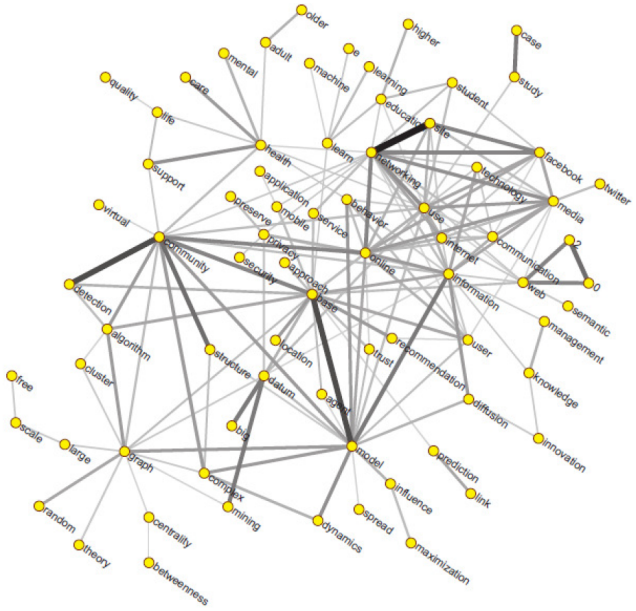

Figure 2: The main link island in the normalized keywords network for SNA.

The fractional approach was proposed by de Solla Price and Beaver (1966) and later 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, which excludes the self-collaboration was proposed by Newman (2001). Recently several papers (Batagelj & Cerinšek, 2013; Cerinšek & Batagelj, 2015; Perianes-Rodríguez, Waltman, & van Eck, 2016; Prathap & Mukherjee, 2016; Leydesdorff & 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 2 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 often 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, & Santoro, 2012).

In the paper Batagelj and Praprotnik (2016) a longitudinal approach to analysis of temporal networks based on temporal quantities was proposed. It is an alternative to the traditional cross-sectional approach. A temporal quantity is describing how the corresponding property is

changing through time. This 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 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 (instantaneous or cumulative) network.

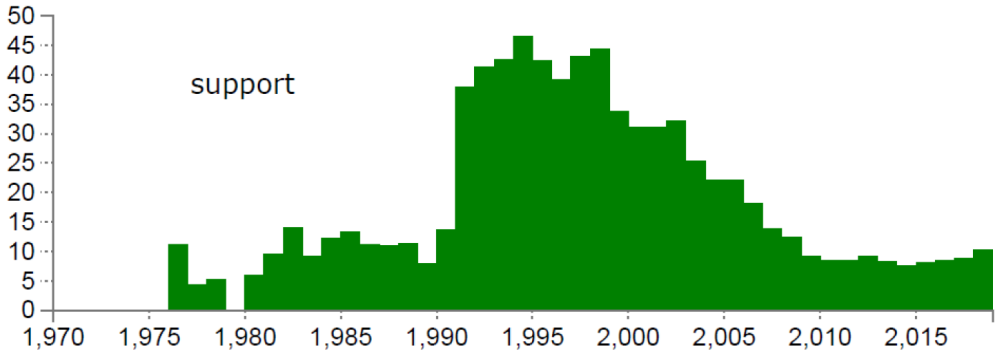

Figure 3: Frequency distribution of appearances per year of the keyword "support" in SNA literature.

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 a selected keyword over time from 0 to 100%. In Figure 3, changes through years of importance in SNA literature of the keyword "support" are presented.

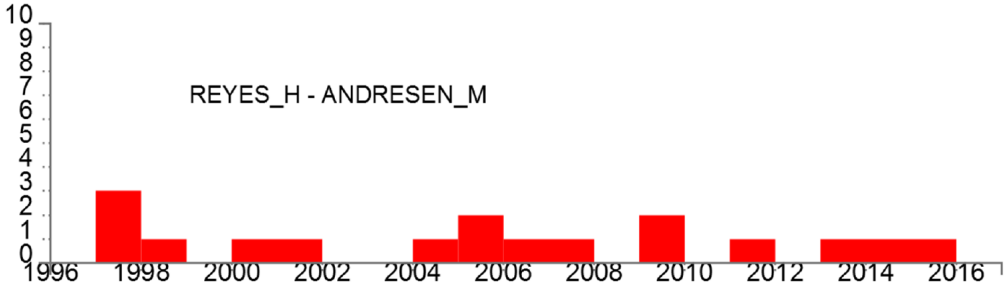

Figure 4: Distribution of number of works on peer-review that authors Reyes and Andersen wrote together each year.

The instantaneous co-authorship network Coi is obtained as $Coi = t(WAi) \cdot WAi$, 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 4 presents the co-authorship of Reyes, H. and Andersen, M. in the field of peer-review (data from Batagelj et al., 2017).

The derived instantaneous temporal network describing citations between journals is obtained as $JCJ = t(WJi) \cdot CiteI \cdot WJc$. Note that the first two factors are instantaneous and the third network in the product is cumulative. The weight of the element $JCJ(i, j)$ is equal to the

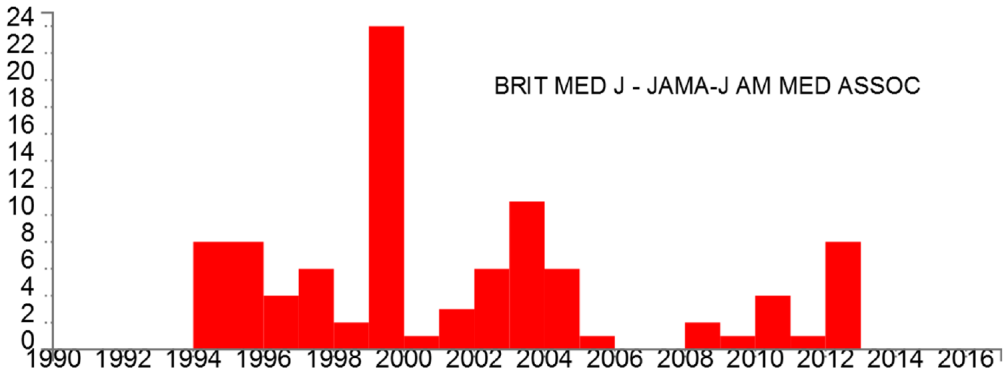

Figure 5: Distribution of number of citations from the journal BRIT MED J to JAMA-J AM MED ASSOC in the peer-review literature.

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 5 the temporal quantity $JCJ(\texttt{BRIT MED J}, \texttt{JAMA-J AM MED ASSOC})$ 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.

Table 1: The most specific keywords on SNA for selected journals according to TF–IDF weights.

	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	

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 his research interest. The topic of determining research fields in Slovenia and elsewhere is briefly discussed by Bartol et al. (2014). Other authors, e.g. Önder, Schweitzer, and Yilmazkuday (2021) have worked on determining research fields from bibliographic data. For instance, the SICRIS/COBISS database contains for each researcher a hierarchical research interest (RI). E.g. 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 represent the current or main research interest of a researcher. In some cases there was an error committed while inputting 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 RIs correlate with SFs 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.

Recently, in two distinct occasions, we have used collaboration network and collaboration distance to solve practical problems (J. Pisanski & Pisanski, 2019; T. Pisanski, Pisanski, & Pisanski, 2020). If this project is fully financed we will make a detailed analysis of various ways to construct such networks, mainly based on fractional and temporal approach (Batagelj, 2020; Batagelj & Maltseva, 2020). Large networks only become accessible if we prune them via procedures such as the MST-based algorithm for Pathfinder networks (Quirin, Cordón, Guerrero-Bote, Vargas-Quesada, & Moya-Anegón, 2008).

We also want to explore structural virality (Goel, Anderson, Hofman, & Watts, 2016) computed for trees following (Mohar & Pisanski, 1988) applied for instance to trees arising from word suffixes (or prefixes) in a text. We also plan to extend the structural virality algorithm from trees to Pathfinder networks.

Our plan is to be able to work on relatively big datasets. For instance metadata of CORD-19 contains more than 420k papers including titles, abstracts and lists of authors. The size of the imported CSV is more than 0.5 GB. Data cleaning and general vectorized table operations (filters, joins, group-by, summarize) are performed relatively fast using the optimized library `data.table` in R on an average 4-core/16 GB RAM machine, the slowest ones within a few seconds. Network operations carried out with our Python libraries (Nets, TQ, Biblio) are much slower, in minutes, and need to be implemented in a faster programming language. We plan to start developing a specialized open-source bibliographic network analysis package in Julia, to support the developed analytic methods and algorithms. Julia is ideal, as it offers C-like speed while providing much better maintainability and ease of use (McNicholas & Tait, 2019; Sengupta & Edelman, 2019; Perla, Sargent, & Stachurski, 2020).

23.3. Detailed description of the work programme

WP1. Project management, coordination and dissemination

This work package runs throughout the project and consists of three main tasks. The detailed description of the work in this package and overall project management is available in the section 23.5. Project management.

T1.1 – Coordination. There are 3 partners in the project which have already established long-term cooperation. We will monitor the work on the project on monthly seminars.

T1.2 – Reporting. Done on a yearly basis, as required by the financer (SRA/ARRS). Principal investigator will assign a member of the project to coordinate the collection of achievements in the reporting period, to prepare and submit the annual report. Financial reporting and funds monitoring will be performed by the accounting departments of the partners.

T1.3 – Dissemination. 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.

WP2. Identification of higher order services and implementation of prototype solutions

The main goal of the project is the identification of potential higher order services and development of some prototype solutions, based on investigation of actual needs and contexts of different interest groups. To support this goal we have to provide high quality data often obtained by combining data from different databases. We also have to develop new algorithms for some analytical problems.

WP3. Methods and tools for the identification of units (entity resolution)

In the 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) network structure in developing special high performance tools. We will try to develop such methods. In such a way a raw data transformation pipeline is built incrementally. Such a pipeline is used for processing fresh updates of input data. The outputs of the pipeline are cleaned elementary bibliographic networks that can be used for analysis. The work in this WP will involve problems of identification of works, authors, journals, countries, and determining keywords, building on existing solutions such as the ones described in Smiraglia and Cai (2017). Results of each of the tasks will include methods for detection of anomalies, resolution proposals and algorithmic updates implementing the resolution proposal into the transformation pipeline. The results will be published as open-source code leveraging our Julia library and will be described in periodic reports.

WP4. 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 enables 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. The main tasks in this WP include:

- **T4.1 – New derived networks based on normalization and multiplication; extension to weighted networks**

- **T4.2 – Temporal versions of derived networks**
- **T4.3 – New temporal quantities describing temporal bibliographic networks**
- **T4.4 – Clustering in temporal networks**

The result of tasks will be new methods with demonstrations. Methods will be implemented in WP5. Progress will be described in periodic reports.

WP5. Development of new methods for bibliographic network analysis in a new Julia package

Bibliographic networks can be large (some hundred thousands or even millions of nodes). The developed software support should provide solutions that can deal also with such data efficiently – in a range of some seconds or minutes. The core of this WP will be implementation of the new library in the Julia programming language, that is interoperable with Python, R and Pajek. The library will be based on experiences gained from Python libraries Nets, TQ and Biblio. We will also develop direct data imports/exports from JSON based formats and direct and rich network visualization methods leveraging modern Javascript/HTML/CSS based libraries (vis.js, vis-network, d3.js, NetworkD3). An important part of the library will be data structures and analytic algorithms to support data cleaning in WP3. The tasks in this WP include:

- **T5.1 – Implementation and optimization of basic data structures and algorithms from Nets and TQ in Julia.** This includes implementation of reading/writing data in selected formats and basic integration with Python, R and Pajek, when specific packages or functionalities are needed (e.g. text processing, statistical analyses, advanced network analytic algorithms).
- **T5.2 – Implementation of advanced algorithms.** Advanced algorithms will be implemented based on methods developed in WP4.
- **T5.3 – Visualization methods.** Integrations with selected Javascript/HTML/CSS visualization libraries.
- **T5.4 – Testing, optimization and documenting the library.**

WP6. Demonstration of applications

The prototype demonstrations will be developed with selected end users, considering their needs and use cases. We will work with editors of selected journals (e.g. Ars Mathematica Contemporanea, Acta Chimica Slovenica). The use cases we may consider include: selecting appropriate reviewers, evaluation of reviewers, quality of data evaluation, automatic suggestion of keywords, etc. We will also consider use cases from the "consumer" side, namely authors, researchers and students. This may include keyword suggestions, journal suggestions, possible partners for research collaboration, papers to read for selected topics. Demonstrations will be focused on selected research fields (e.g. mathematics, social network analysis, etc.). The work package will be divided into the following tasks.

- **T6.1 – Bibliometric analysis in selected research fields with demonstration of the newly developed methods**
- **T6.2 – Applications for journal managers**
- **T6.3 – Applications for authors, researchers and students**

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.

23.4. Available research equipment over 5.000 €

For most data sets a better laptop with at least 32 GB memory is sufficient. For very large data sets we will use the computing facilities available at the UP.

23.5. Project management: Detailed implementation plan and timetable

The project is spanned over three organizations: University of Primorska, Faculty of Mathematics, Natural Sciences and Information Technology (UP FAMNIT); Institute of Mathematics and Physics, Ljubljana (IMFM); and University of Ljubljana, Faculty of Arts (UL FF). Roughly, UL FF will be focused mainly on information science aspects, while UP FAMNIT and IMFM will be responsible primarily for algorithm design, data processing and prototyping.

Project management. The principal investigator (PI) will be responsible for the achievement of milestones (reports) and deliverables. He will also ensure that the project is up and running within the set deadlines. Members of work package groups will have weekly meetings to assess progress on the project and identify possible risks. Results and work in progress will be presented in seminars. The communication on the project will be carried out via electronic media and meetings. Documents and computer code will be stored in GIT repositories.

General methodology. The development of methods in data analysis consists of analyzing data, setting approximate conjectures, developing and implementing analytic algorithms supporting the method, testing and at the end evaluation of results and the method itself. Software development of a library consists of cycles that include planning, implementation, testing, optimization and documentation. Initial cycles can be simplified and consist only of planning, prototyping and evaluation for the purpose of development of methods which lead consolidation cycles (planned code rewrites). All the above stated methods are iterative processes.

Project timeline. The expected project duration is 3 years. Most of the work packages will be active throughout the project. The only exception is WP6 which will be active during the final year of the project.

References

Agrahari, A., Chaudhary, C. P., & Singh, S. N. (2018). Domain analysis of D-Lib magazine: A bibliometric study. *Webology*, *15*(1). Retrieved from `http://www.webology.org/2018/v15n1/a165.pdf`

Bartol, T., Budimir, G., Dekleva-Smrekar, D., Pusnik, M., & Juznic, P. (2014). Assessment of research fields in Scopus and Web of Science in the view of national research evaluation in Slovenia. *Scientometrics*, *98*(2), 1491–1504. doi: 10.1007/s11192-013-1148-8

Batagelj, V. (2020). On fractional approach to analysis of linked networks. *Scientometrics*, *123*(2), 621–633. doi: 10.1007/s11192-020-03383-y

Batagelj, V., & Cerinšek, M. (2013). On bibliographic networks. *Scientometrics*, *96*(3), 845–864. doi: 10.1007/s11192-012-0940-1

Batagelj, V., Doreian, P., Ferligoj, A., & Kejžar, N. (2014). *Understanding large temporal networks and spatial networks: Exploration, pattern searching, visualization and network evolution.* Chichester: Wiley. doi: 10.1002/9781118915370

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. doi: 10.1007/s11192-017-2522-8

Batagelj, V., & Maltseva, D. (2020). Temporal bibliographic networks. *J. Informetr.*, *14*(1), Article No. 101006. doi: {10.1016/j.joi.2020.101006}

Batagelj, V., & Praprotnik, S. (2016). An algebraic approach to temporal network analysis based on temporal quantities. *Soc. Netw. Anal. Min.*, *6*(1), 28:1–28:22. doi: 10.1007/s13278-016-0330-4

Bell, M. G. H., & Iida, Y. (1997). *Transportation network analysis.* Chichester: Wiley. doi: 10.1002/9781118903032

Bezanson, J., Edelman, A., Karpinski, S., & Shah, V. B. (2017). Julia: A fresh approach to numerical computing. *SIAM Rev.*, *59*(1), 65–98. doi: 10.1137/141000671

Bird, S., Klein, E., & Loper, E. (2019). *Natural language processing with python: Analyzing text with the natural language toolkit.* Retrieved from `https://www.nltk.org/book/`

Buscaldi, D., & Rosso, P. (2008). A conceptual density-based approach for the disambiguation of toponyms. *Int. J. Geogr. Inf. Sci.*, *22*(3), 301–313. doi: 10.1080/13658810701626251

Casteigts, A., Flocchini, P., Quattrociocchi, W., & Santoro, N. (2012). Time-varying graphs and dynamic networks. *Int. J. Parallel Emergent Distrib. Syst.*, *27*(5), 387–408. doi: 10.1080/17445760.2012.668546

Cerinšek, M., & Batagelj, V. (2015). Network analysis of Zentralblatt MATH data. *Scientometrics*, *102*(1), 977–1001. doi: 10.1007/s11192-014-1419-z

Chen, G., & Xiao, L. (2016). Selecting publication keywords for domain analysis in bibliometrics: A comparison of three methods. *J. Informetrics*, *10*(1), 212–223. doi: 10.1016/j.joi.2016.01.006

Christen, P. (2012). *Data matching: Concepts and techniques for record linkage, entity resolution, and duplicate detection.* Berlin, Heidelberg: Springer. doi: 10.1007/978-3-642-31164-2

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

Freire, N. (2014). Word occurrence based extraction of work contributors from statements of responsibility. *Int. J. Digit. Libr.*, *14*(3-4), 141–148. doi: 10.1007/s00799-014-0113-3

Freire, N., Borbinha, J., & Calado, P. (2012). An approach for named entity recognition in poorly structured data. In E. Simperl, P. Cimiano, A. Polleres, Ó. Corcho, & V. Presutti (Eds.), *The semantic web: Research and applications* (Vol. 7295, pp. 718–732). Springer. doi: 10.1007/978-3-642-30284-8_55

Gauffriau, M. (2017). A categorization of arguments for counting methods for publication and citation indicators. *J. Informetr.*, *11*(3), 672–684. doi: 10.1016/j.joi.2017.05.009

Goel, S., Anderson, A., Hofman, J. M., & Watts, D. J. (2016). The structural virality of online diffusion. *Manag. Sci.*, *62*(1), 180–196. doi: 10.1287/mnsc.2015.2158

Gutierres Castanha, R. C., & Wolfram, D. (2018). The domain of knowledge organization: A bibliometric analysis of prolific authors and their intellectual space. *Knowl. Organ.*, *45*(1), 13–22. doi: 10.5771/0943-7444-2018-1-13

Hjørland, B. (2002). Domain analysis in information science: Eleven approaches – traditional as well as innovative. *J. Documentation*, *58*(4), 422–462. doi: 10.1108/00220410210431136

Hjørland, B., & Albrechtsen, H. (1995). Toward a new horizon in information science: Domain-analysis. *J. Am. Soc. Inf. Sci.*, *46*(6), 400–425. doi: 10.1002/(sici)1097-4571(199507)46:6\<400::aid-asi2\>3.0.co;2-y

Holme, P. (2015). Modern temporal network theory: A colloquium. *Eur. Phys. J. B*, *88*(9), Article No. 234. doi: 10.1140/epjb/e2015-60657-4

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

Jones, B. (2002). *Computational geometry database.* Retrieved from `ftp://ftp.cs.usask.ca/pub/geometry/`

Journal abbreviation sources. (2022). Retrieved from `https://www.abbreviations.com/jas.php`

Karlovčec, M., Lužar, B., & Mladenić, D. (2016). Core-periphery dynamics in collaboration networks: the case study of Slovenia. *Scientometrics*, *109*(3), 1561–1578. doi: 10.1007/s11192-016-2154-4

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

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

Leydesdorff, L., & Park, H. W. (2017). Full and fractional counting in bibliometric networks. *J. Informetr.*, *11*(1), 117–120. doi: 10.1016/j.joi.2016.11.007

Lindsey, D. (1980). Production and citation measures in the sociology of science: The problem of multiple authorship. *Soc. Stud. Sci.*, *10*(2), 145–162. doi: 10.1177/030631278001000202

Liu, H. (2004). *Montylingua: A free, commonsense-enriched natural language understander for english (version 2.1).* Retrieved from `http://alumni.media.mit.edu/~hugo/montylingua/`

Lo, K., Wang, L. L., Neumann, M., Kinney, R., & Weld, D. S. (2020). S2ORC: The semantic scholar open research corpus. In D. Jurafsky, J. Chai, N. Schluter, & J. R. Tetreault (Eds.), *Proceedings of the 58th annual meeting of the association for computational linguistics, ACL 2020* (pp. 4969–4983). Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.447

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. doi: 10.1371/journal.pone.0094429

Maltseva, D., & Batagelj, V. (2019). Social network analysis as a field of invasions: Bibliographic approach to study SNA development. *Scientometrics*, *121*(2), 1085–1128. doi: 10.1007/s11192-019-03193-x

Maltseva, D., & Batagelj, V. (2020). Towards a systematic description of the field using keywords analysis: Main topics in social networks. *Scientometrics*, *123*(1), 357–382. doi: 10.1007/s11192-020-03365-0

Marsden, P. V. (1990). Network data and measurement. *Annu. Rev. Sociol.*, *16*, 435–463. doi: 10.1146/annurev.so.16.080190.002251

McNicholas, P. D., & Tait, P. A. (2019). *Data science with julia.* New York: Chapman and Hall/CRC. doi: 10.1201/9781351013673

Moder, J. J., & Phillips, C. R. (1970). *Project Management with CPM and PERT* (2nd ed.). New York: Van Nostrand Reinhold.

Mohar, B., & Pisanski, T. (1988). How to compute the Wiener index of a graph. *J. Math. Chem.*, *2*(3), 267–277. doi: 10.1007/bf01167206

Momeni, F., & Mayr, P. (2016). Evaluating co-authorship networks in author name disambiguation for common names. In N. Fuhr, L. Kovács, T. Risse, & W. Nejdl (Eds.), *Research and advanced technology for digital libraries, TPDL 2016* (Vol. 9819, pp. 386–391). Springer. doi: 10.1007/978-3-319-43997-6_31

Newman, M. E. J. (2001). Scientific collaboration networks. ii. shortest paths, weighted networks, and centrality. *Phys. Rev. E*, *64*(1), Article No. 016132. doi: 10.1103/physreve.64.016132

de Solla Price, D. J., & Beaver, D. (1966). Collaboration in an invisible college. *Am. Psychol.*, *21*(11), 1011–1018. doi: 10.1037/h0024051

Önder, A. S., Schweitzer, S., & Yilmazkuday, H. (2021). Specialization, field distance, and quality in economists' collaborations. *J. Informetrics*, *15*(4), 101222. doi: 10.1016/j.joi.2021.101222

Perianes-Rodríguez, A., Waltman, L., & van Eck, N. J. (2016). Constructing bibliometric networks: A comparison between full and fractional counting. *J. Informetrics*, *10*(4), 1178–1195. doi: 10.1016/j.joi.2016.10.006

Perkins, J. (2010). *Python text processing with NLTK 2.0 cookbook.* Packt.

Perla, J., Sargent, T. J., & Stachurski, J. (2020). *Quantitative economics with julia.* Retrieved from `https://julia.quantecon.org/`

Pisanski, J., & Pisanski, T. (2019). The use of collaboration distance in scheduling conference talks. *Informatica*, *43*(4), 461–466. doi: 10.31449/inf.v43i4.2832

Pisanski, T., Pisanski, M., & Pisanski, J. (2020). A novel method for determining research groups from co-authorship network and scientific fields of authors. *Informatica*, *44*(2), 139–145. doi: 10.31449/inf.v44i2.3079

Prathap, G., & Mukherjee, S. (2016). A conservation rule for constructing bibliometric network matrices. *arXiv*. Retrieved from `https://arxiv.org/abs/1611.08592`

Quirin, A., Cordón, O., Guerrero-Bote, V. P., Vargas-Quesada, B., & Moya-Anegón, F. (2008). A quick MST-based algorithm to obtain Pathfinder networks $(\infty, n-1)$. *J. Assoc. Inf. Sci. Technol.*, *59*(12), 1912–1924.

Reitz, F., & Hoffmann, O. (2013). Learning from the past: An analysis of person name corrections in the DBLP collection and social network properties of affected entities. In T. Özyer, J. G. Rokne, G. Wagner, & A. H. P. Reuser (Eds.), *The influence of technology on social network analysis and mining* (Vol. 6, pp. 427–453). Springer. doi: 10.1007/978-3-7091-1346-2_19

Robertson, S. (2004). Understanding inverse document frequency: On theoretical arguments for IDF. *J. Documentation*, *60*(5), 503–520. doi: 10.1108/00220410410560582

Sengupta, A., & Edelman, A. (2019). *Julia high performance: Optimizations, distributed computing, multithreading, and gpu programming with julia 1.0 and beyond* (2nd ed.). Packt.

Smiraglia, R. P. (2013). Is FRBR a domain? Domain analysis applied to the literature of the FRBR family of conceptual models. *Knowl. Org.*, *40*(4), 273–284. doi: 10.5771/0943-7444-2013-4-273

Smiraglia, R. P. (2015). Domain analysis of domain analysis for knowledge organization: Observations on an emergent methodological cluster. *Knowl. Org.*, *42*(8), 602–614. doi: 10.5771/0943-7444-2015-8-602

Smiraglia, R. P., & Cai, X. (2017). Tracking the evolution of clustering, machine learning, automatic indexing and automatic classification in knowledge organization. *Knowl. Org.*, *44*(3), 215–233. doi: 10.5771/0943-7444-2017-3-215

Talburt, J. R. (2011). *Entity resolution and information quality.* Morgan Kaufmann. doi: 10.1016/

c2009-0-63396-1

TePaske-King, B., & Richert, N. (2001). The identification of authors in the mathematical reviews database. *Issues Sci. Technol. Librariansh.*(31). doi: 10.5062/f4kh0k9m

Ulrichsweb. (2022). Retrieved from `http://ulrichsweb.serialssolutions.com/`

Wang, L. L., Lo, K., Chandrasekhar, Y., Reas, R., Yang, J., Eide, D., ... others (2020). CORD-19: The COVID-19 open research dataset. *ArXiv*. Retrieved from `https://arxiv.org/abs/2004.10706`

Wang, S. (2019). The intellectual landscape of the domain of culture and ethics in knowledge organization: An analysis of influential authors and works. *Cat. Classif. Q.*, *57*(4), 227–243. doi: 10.1080/01639374.2019.1614710

Windham, M. (2019). *Unstructured data analysis: Entity resolution and regular expressions in sas*. SAS Institute.

Yadav, V., & Bethard, S. (2019). A survey on recent advances in named entity recognition from deep learning models. *arXiv*. Retrieved from `https://arxiv.org/abs/1910.11470`

Žumer, M. (2018). Ifla library reference model (ifla lrm): Harmonisation of the frbr family. *Knowl. Org.*, *45*(4), 310–318. doi: 10.5771/0943-7444-2018-4-310