Scientometrics (2020) 123:357–382
https://doi.org/10.1007/s11192-020-03365-0

Towards a systematic description of the field using keywords analysis: main topics in social networks

Daria Maltseva[1] · **Vladimir Batagelj**[1,2,3]

Received: 17 September 2019 / Published online: 25 January 2020

Abstract
This paper presents the results of the analysis of keywords used in Social Network Analysis (SNA) articles included in the WoS database and main SNA journals, from 1970 to 2018. 32,409 keywords were obtained from 70,792 works with complete descriptions. We provide a list of the most used keywords and show subgroups of keywords which are connected to each other. To go deeper, we place the keywords into the contexts of selected groups of authors and journals. We use temporal analysis to get an insight into some keyword usage. The distributions of the number of keyword types and tokens over time show fast growth starting from 2010s, which is the result of the growth in the number of articles on SNA topics and applications of SNA in various scientific fields. Even though the most frequently used keywords are trivial or general, the approaches used for the normalization of network link weights allow us to extract keywords representing substantive topics and methodological issues in SNA.

Keywords Social network analysis · Bibliographic networks · Temporal network analysis · Keyword co-occurrence networks · Fractional approach · TF–IDF index

Introduction

Social network analysis (SNA) is a rapidly developing scientific field that has appeared and grown significantly over the past 50 years, in the number of scientific publications and in the different disciplines involved (Borgatti and Foster 2003; Otte and Rousseau 2002). Until the 2000s the field was mostly developed inside different branches of social sciences, it then received significant attention from researchers in the natural sciences,

✉ Daria Maltseva
dmaltseva@hse.ru

Vladimir Batagelj
vladimir.batagelj@fmf.uni-lj.si

1 National Research University Higher School of Economics, 11 Pokrovsky Boulevard, Moscow, Russia 101000

2 Institute of Mathematics, Physics and Mechanics, Jadranska 19, 1000 Ljubljana, Slovenia

3 University of Primorska, Andrej Marušič Institute, 6000 Koper, Slovenia

which led to the so-called *"invasion of the physicists"* (Bonachich 2004) and resulted in the development of Network Science (Freeman 2004, 2011). To a large extent, this increase in interest in the topic was also due to the emergence of the World Wide Web in the 1990s and online social networks in the 2000s. This inevitably led to the extension of thematic areas where the methodology of network analysis is applied.

The usual way to study thematic areas and get important information on the topics developed in different scientific areas is to analyze the keywords used in their publications. In today's academic world, keywords have become an important part of the information about publications, as it is usually obligatory to provide them with an article or book. However, when keywords are not provided by the author, they can be assigned to the paper by the journal or database, or automatically extracted from the title. Thus, the topical identity of any field can easily be constructed based on the metadata of the academic works.

Although the development of SNA has attracted the attention of a number of researchers, this attention has mostly been given to explorations of co-authorship structures (Batagelj et al. 2014; Leydesdorff et al. 2008; Otte and Rousseau 2002), citation structures of works or journals (Batagelj et al. 2014; Hummon and Carley 1993; Leydesdorff et al. 2008) and bibliographic coupling (Batagelj et al. 2014; Brandes and Pich 2011) between works, authors, or journals in the whole field. Different subfields (Batagelj et al. 2014, 2020; Hummon et al. 1990; Kejžar et al. 2010) and subdisciplines within the field (Borgatti and Foster 2003; Lazer et al. 2009; Otte and Rousseau 2002; Varga and Nemeslaki 2012) have also been studied. However, there are few examples of analysis of the main topics in SNA, where the data comes from one journal (Leydesdorff et al. 2008) or special subtopics (Batagelj et al. 2014, 2020).

Study on the development of SNA (Maltseva and Batagelj 2019), based on the analysis of networks of articles from the *Web of Science (WoS)* database matching the query "social network*", influential works, and those published in the main journals in the field (up to 2018), has shown that the number of publications on SNA topics has grown significantly, and on average it doubles every 3 years. This is due to the huge interest to SNA from other disciplines, such as physics, economics, computer science, media studies, and—surprisingly—from behavioral biology. We assume that the growth in the number of articles and disciplines involved should be reflected in the topics observed in the field, and with the analysis of keywords, we observe the scope of contexts in which SNA is applied. In this article, we present the core concepts which unify the field, and, vice versa, the concepts showing disciplinary differences. The extraction of such information is used to compare different units—authors, groups of authors, or journals. This analysis reveals important information for the systematic description of the current development of SNA.

This paper is organized as follows. "Literature review" section presents several previous studies on keywords used in SNA. "Data" section describes the dataset and some issues of the network construction from the original two-mode network connecting works with keywords. "Basic network properties" section presents some statistical properties of this network and a list of the most-used keywords. In "Keywords co-occurrence" section we provide the analysis of the network of keyword co-occurrences. In "Keywords and authors" and "Keywords and journals" sections, we show the possibility of checking the keywords associated with authors, groups of authors or journals, using two-mode networks connecting works with authors and works with journals. We used a temporal version of the original network to get an insight into the dynamics of keyword usage. Our approach to bibliometric network analysis has already shown its productivity in a number of studies of different scientific fields and topics (Batagelj et al. 2014, 2017, 2020; Kejžar et al. 2010). The approach to temporal network analysis in Batagelj and Maltseva (2020) and Batagelj and Praprotnik (2016) is applied to large bibliographic networks for the first time.

Literature review

There are few studies describing the field of SNA using keyword analysis and the datasets used do not cover the whole field—they describe either the keywords associated with one journal (Leydesdorff et al. 2008), or literature on specific topics in SNA (Batagelj et al. 2014, 2019).

Leydesdorf et al. (2008) present an analysis of the topical development of SNA through the analysis of networks of title word co-occurrences of articles published in the journal *Social Networks*. During the period 1988–2006, 165 title words occurred more than once in a single year, and were included in the analysis. The authors find that over time, *"particular issues reappear, notably centrality, measurement and measure, and concepts relating to data collection"*, and *"less frequently, concepts related to balance, blockmodels or equivalence appear"*, which shows the methodological identity of the journal. For different years, the set of words in the network was different, meaning that the title words in the publications of each year provide a specific selection from the larger vocabulary of a discourse shaped and reproduced at the level of the specialty.

To provide more stable results, the semantic domain was enlarged: 6071 titles containing the keywords *social network* or *social networks*, were harvested from Google Scholar, and 172 words (occurring 8 or more times in any single year) were used for additional analysis. This resulted in a more stable structure; however, the titles of most publications focused on substantive issues rather than methodological ones (*social capital*, concepts referring to less-privileged social groups, such as *minorities, women, patients*, and *the elderly*). Authors also showed that some words changed their network position over time, such as the theoretical concepts of *capital* and *community*, which became central for the application of SNA across the social sciences. The words *method* and *model* also moved to the center over time, suggesting the rise of methodological reflection among scholars investigating social networks.

In a bibliometric analysis of the literature on centrality (Batagelj et al. 2014) and the literature on network clustering and blockmodeling (Batagelj et al. 2019), the authors presented lists of the most frequent keywords, constructed from the titles and keywords provided in the full descriptions of articles in the WoS. Most of the top keywords were either expected, or trivial (*social, network*), or generic with limited value (*model, graph, structure*, etc.). According to authors, as a tool of explanation, the keywords should be examined with great care in clearly defined contexts—in some groups of closely related works or authors. Kejžar et al. (2010) presented a method to construct such subnetworks of the topics of selected groups of authors from the two-mode networks of works with authors and works with keywords; this method was used in this article.

Previous studies show the importance of the analysis of relatively large datasets, which not only validate the results, but also make the analysis more systematic. We can expect the appearance (and reappearance) of the expected, or trivial and generic, words associated with SNA—which might be regarded as its core concepts. There should also be words associated with substantive issues, devoted to the topics which are being studied in different subfields of SNA. This paper aims to uncover these topics.

Data

Data collection

The data collection, cleaning and network construction were presented in detail by Maltseva and Batagelj (2019). Our dataset consists of publications from the *WoS Core Collection* database matching the query "social network*", and other works highly cited in the SNA field, and published in main SNA journals, up to 2018. The first part of the dataset is based on the SN5 data collected for the Viszards session at the Sunbelt 2008 (Batagelj et al. 2014), and contains all the records obtained for the query `"social network*"` and articles from the journal *Social Networks*, until 2007. Obtained descriptions of the works can be of two types: with full descriptions (*hits*), and cited only (*terminal*, listed only in the CR field of a work description in WoS). We additionally searched for the terminal works without full descriptions which were most frequently cited, and papers on SNA of around 100 social networkers. The final version of SN5 contained 7950 works with a full description (hits), and 193,376 works (hits and cited only). The SN5 data were extended in June 2018 using the same search scheme. Starting from 2007, 576 articles from *Social Networks* journal were added. Additionally, in 2018, all the articles from the network—related journals contained in the WoS were included—such as *Network Science*, *Social Network Analysis and Mining*, *Journal of Complex Networks* (in total, 431 article). Again, we additionally collected full descriptions for terminal works with high (at least 150) citation frequencies. We also included manual descriptions of important terminal works from the dataset **BM** on blockmodeling (Batagelj et al. 2019). Finally, our dataset included 70,792 WoS records with complete descriptions (hits).

Using **WoS2Pajek 1.5** (Batagelj 2017), we transformed our data into a collection of networks: a one-mode citation network **Cite** on works (from the field CR of the WoS file description) and two-mode networks—the authorship network **WA** on works × authors (from the field AU), the journal network **WJ** on works × journals (from the field CR or J9), and the keyword network **WK** on works × keywords (from the fields ID, DE or TI). The keywords are single words—phrases were split to components. They were lemmatized and stopwords were removed. After data cleaning, from 70,792 hits we produced networks with sets of the following sizes: works $|W| = 1{,}297{,}133$, authors $|A| = 395{,}971$, journals $|J| = 69{,}146$, key words $|K| = 32{,}409$. We removed multiple links and loops and obtained *basic* networks **CiteN**, **WAn**, **WJn**, and **WKn**. For the terminal works only partial information is provided: the name of the *first* author, journal, publication year, journal issue, and the first page number. That is why it is not correct to use these networks for the analysis of keywords and authors. We constructed *reduced* networks containing only works with complete descriptions **CiteR**, **WAr**, **WJr**, and **WKr**, where the sizes of sets are as follows: works $|W| = 70{,}792$, authors $|A| = 93{,}011$, journals $|J| = 8943$, key words $|K| = 32{,}409$. The total number of keywords is lower than the total number of documents, which means that the same keywords reappear in papers several times. In this paper, we use these three two-mode networks for the analysis.

Even though the initial search was oriented towards *social* networks, an additional 'saturation' search of the papers which were cited a lot by the field's representatives, as well as inclusion of the works from journals important for the field, and the most prominent authors allowed us to improve the dataset in sense of the broader inclusion of the publications related to network analysis in general. Thus, the dataset covers not only the works of social scientists, but also influential papers published by physicists, biological scientists, information and computer scientists, etc. This additional search allowed us also to include influential papers, usually published earlier, that could have been overlooked by our search queries because they do not use the contemporary terminology.

Derived networks

A two-mode network can be represented as a two-mode matrix. A pair of two-mode networks can be multiplied, if the second set of nodes in the first network is equal to the first set of nodes in the second network. If all weights in two-mode networks are equal to 1, then the product of the weights will also be equal to 1 and therefore a [*u*, *v*] element of the product matrix counts the number of ways we can move from node *u* using the first network through the second set and afterwards using the second network to node *v* (Batagelj and Cerinšek 2013; Batagelj et al. 2014). In our case, this shared set is the set of works (papers, reports, books, etc.), which *links* bibliographic networks to each other. Using the multiplication of two-mode networks, we constructed *derived* networks.

Multiplying a network **WK** with its transpose, we obtain the network of keyword co-occurrences $\mathbf{KK} = \mathbf{WK}^T * \mathbf{WK}$. The weight of an edge between two nodes $w[k_1, k_2]$ in the keyword co-occurrence network **KK** tells us in how many works the keywords k_1 and k_2 were used together. Multiplying different compatible two-mode networks, we construct the network of authors and keywords $\mathbf{AK} = \mathbf{WA}^T * \mathbf{WK}$, counting in how many works author *u* used the keyword *k*, and journals and keywords $\mathbf{JK} = \mathbf{WJ}^T * \mathbf{WK}$ counting how many times journal *j* used the keyword *k*.

Normalization in derived networks

Derived networks can have some deficiencies, such as overrating the contribution of bibliographic entities with many ties (works with many authors or keywords, journals with many works). To deal with such cases, the *fractional approach* (Batagelj and Cerinšek 2013; Batagelj 2019; Gauffriau et al. 2007) was used. This takes into account the contribution of bibliographic entities (works, authors, or journals), normalizing their weights so that their input to the resulting network is equal to 1.

Let us provide an example of a two-mode network of works × keywords **WK**. In a regular network, the outdegree is equal to the number of keywords of the work, and the indegree is equal to the number of works in which the same keywords are used. The normalization creates the network **nWK** where the weight of each arc is divided by the sum of weights of all arcs having the same initial node as this arc (the outdegree of a node):

$$n(\mathbf{WK})[w, k] = \frac{\mathbf{WK}[w, k]}{\max(1, \mathrm{outdeg}(w))}$$

where *w* is a work and *k* is a keyword. The contribution of each paper *w* is equal to 1, and we assume that each keyword takes an equal place among others. The proposed normalization is applied to different two-mode networks **WK**, **WA**, and **WJ**, and thus the product networks **nKK**, **nAK**, and **nJK** are also normalized.

For **JK**, we also applied the *TF–IDF approach* (term frequency—inverse document frequency) to the normalization (Robertson 2004), which allows us to evaluate the importance of a word to a document in a corpus of documents. A detailed description of each derived network construction and normalization is presented in the corresponding sections below.

Temporal networks

Applying the *temporal quantities* approach (Batagelj and Maltseva 2020; Batagelj and Praprotnik 2016) to the **WKr** network, we constructed temporal networks, using Python libraries Nets and TQ (Batagelj 2014). These networks can be of two types—instantaneous (with values given per year) **WKins**, and cumulative **WKcum**. They are stored in the `json` format. Using the multiplication and normalization of temporal networks, different derived temporal networks can be constructed. The construction of these networks is described in the corresponding sections below.

Basic network properties

Statistical distribution

For the works with full descriptions ($DC = 1$), the keywords are supposed to be presented in special fields `DE` (Author Keywords) and `ID` (Keywords Plus). However, for some publications this information is not provided. In such cases the keywords are constructed by **WoS2Pajek** from the titles of works. All composite keywords were split into single words, and lemmatization was used to deal with the *word-equivalence problem*. However, the works which are cited only ($DC = 0$) do not have keywords—in our case, 95% of the works in the **WKn** network.

In **WKr**, the network constructed from works with complete descriptions, the number of keywords in 70,792 works varies from 1 to 84 (Fig. 1, top). The distribution of the number of keywords used in all works (Fig. 1, bottom) shows that large numbers of keywords are mentioned only once (16,164), twice (3919), or three times (1970). The usage of these keywords is episodic, and it shows the wide scope of the contexts where SNA is applied. There are also keywords which are used intensively, constructing the core concepts of the field.

Figure 2 presents the temporal distributions of the number of *all* keywords (top) and *unique (different)* keywords (bottom) used in SNA publications. The observed rise of the number of keywords used is due to the fast growth in the number of articles on SNA topics starting from 2007, which was shown in Batagelj and Maltseva (2019). In 2007 the number of keywords used was around 30,000; in 2017 it was 160,000. The number of *different* keywords also shows the growth in the range of scientific fields and disciplines where SNA is applied—in 2005 it was around 3000; in 2017 it was four times larger.

The most used keywords

The most frequent keywords are presented in Table 1. Not surprisingly, the words *social* and *network* are mentioned in the largest number of works, followed by *analysis*, which is trivial, but also shows the relevance of the data to the topic being studied. Some other frequently used words—*graph, structure, relationship, role, tie* (marked in boldface)—are related to network analysis, while others—*datum, base, information, research, theory, model, algorithm, approach, pattern, effect*—to scientific research in general (they are generic with limited value). General graph theoretic terms such as *node, edge, arc, link, path, connection*, do not appear among the top terms. There

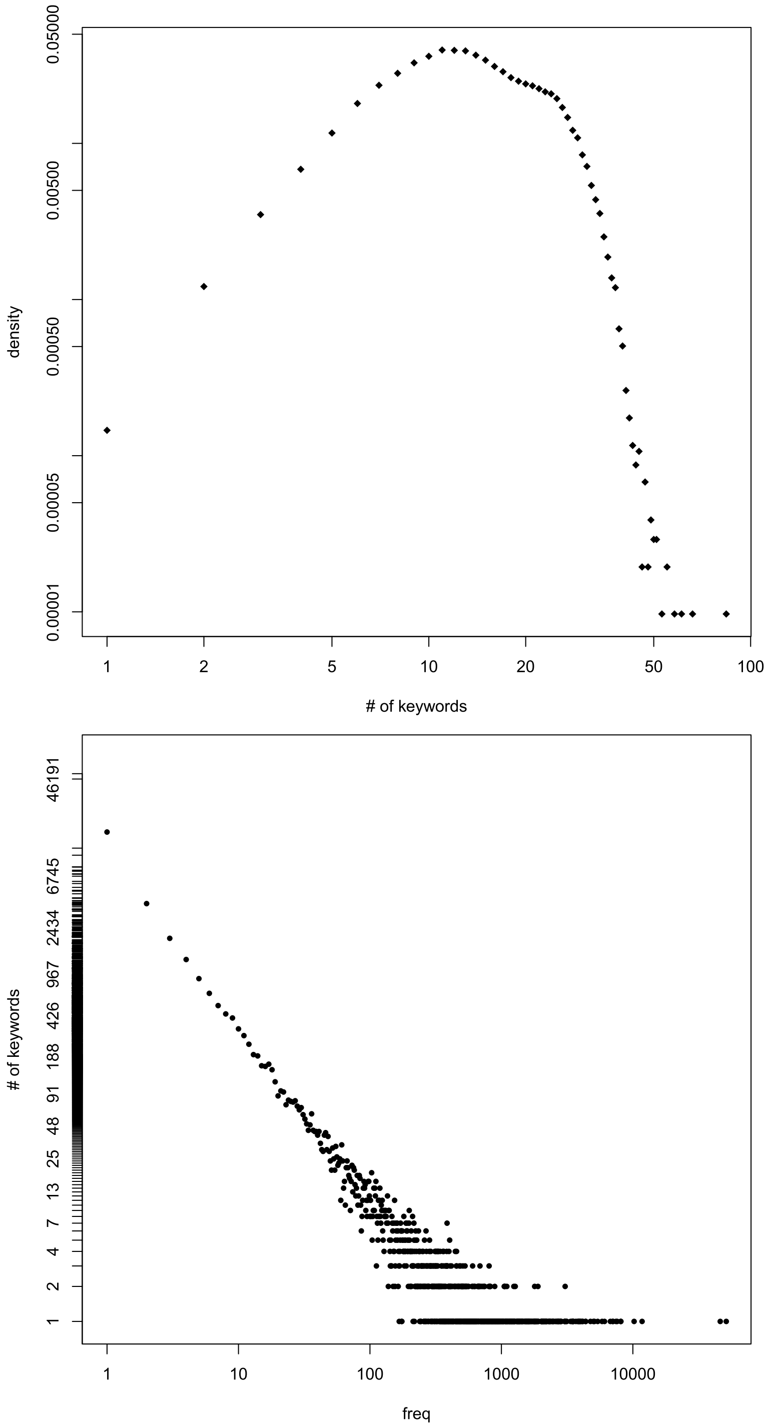

Fig. 1 Logarithmic plots with distributions of the number of keywords per paper (top) and number of keywords used in all works (bottom)

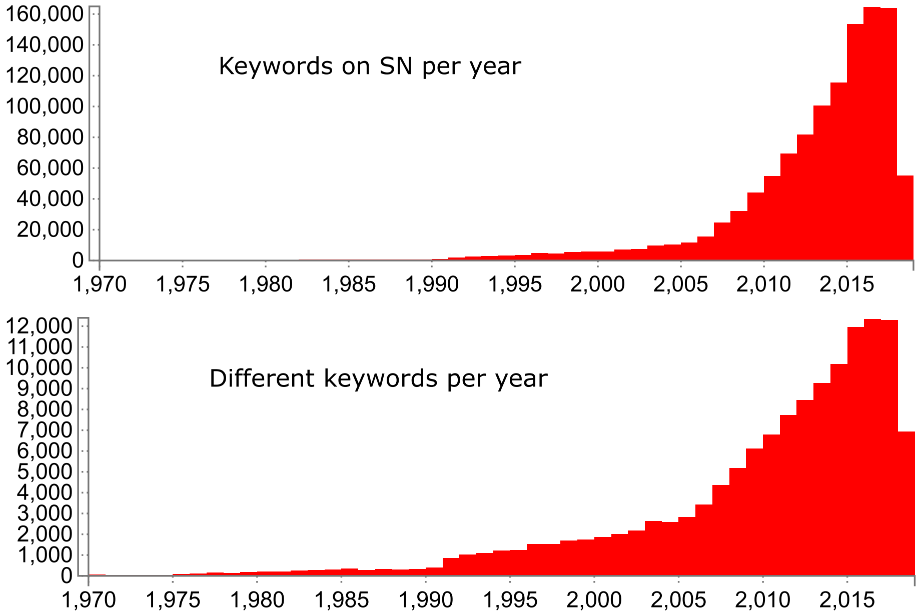

Fig. 2 WKins: distributions based on keywords and works

are also words related to exact substantive topics being studied in network analysis—*online, networking, facebook, internet, site, web; health; behavior; education; support; communication; influence; innovation; trust; risk; family; community*. We note that keywords can have different meanings in different contexts, therefore their identification in different subgroups (of authors or journals) can give us better understanding of the topic structure of SNA.

We counted the proportion of the number of appearances of *each keyword* to the *most frequent keyword* appearance for each year based on the **WKins** network. This proportion normalizes the importance of certain keyword over time from 0 to 100%. The proportions for the most used keywords (Table 1) over time are presented in the figures below: the most frequent keywords up to 100% and 50% (Fig. 3), and up to 30%, 14%, and 10% (Fig. 4). It is expected that the keywords *social* and *network* get the maximum levels of usage in almost all the years starting from the 1970s. Other keywords presented in Fig. 3 have maximum usage in the 1970s due to the small number of works published in this period. However, it shows that these words—*structure, theory, graph, relationship, role, innovation*—are used for all of the recent history of SNA. It is interesting that these keywords were very frequently used in the early years (up to 1970s).

Some of the most used keywords, presented in Figs. 3 and 4, have been in use for a long time—these are *community, support, health, algorithm, behavior, tie*). Some of the words appeared later—in the 1980s and 1990s (*trust, technology, service, web, risk*), or the 2000s (*internet, media, online, facebook*). Mentioned since the 1990s, the word *detection* grew in the 2010s, presumably due to studies of community detection. The word *animal*, which "surprisingly" appeared at the analysis of the citation network (Maltseva and Batagelj 2019), is presented in the field from 1990s.

Table 1 WKn net indegree: the most used keywords

Rank	Value	Id	Rank	Value	Id
1	51,332	**social**	31	3485	**structure**
2	46,191	**network**	32	3479	life
3	11,751	**analysis**	33	3444	risk
4	10,219	model	34	3358	research
5	8104	community	35	3143	learn
6	8090	use	36	3116	influence
7	7596	base	37	3054	student
8	7439	information	38	3054	impact
9	7061	health	39	3049	perspective
10	7023	behavior	40	3042	complex
11	6745	online	41	3024	theory
12	6087	networking	42	2859	organization
13	5833	media	43	2828	**relationship**
14	5404	support	44	2802	algorithm
15	5101	communication	45	2776	education
16	5013	study	46	2714	group
17	4759	datum	47	2704	mobile
18	4376	management	48	2698	**tie**
19	4372	internet	49	2695	adult
20	4164	knowledge	50	2633	approach
21	4126	user	51	2608	care
22	4023	facebook	52	2551	adolescent
23	3984	technology	53	2479	**role**
24	3907	site	54	2472	state
25	3888	web	55	2467	innovation
26	3855	self	56	2434	pattern
27	3784	**graph**	57	2385	effect
28	3676	performance	58	2339	people
29	3534	service	59	2333	trust
30	3512	dynamics	60	2332	family

Keywords co-occurrence

Network construction

We applied the column projection to the normalized reduced **WKr** network to construct the normalized one-mode network **nKK**:

$$\mathbf{nKK} = n(\mathbf{WK})^T * n(\mathbf{WK})$$

In this network, the loops were deleted and bidirected arcs were transformed to edges (with the summation of the arc weights). The obtained network **nKK** consists of 32,409 nodes and 2,799,530 edges. The weight $\mathbf{nKK}[i,j]$ on the edges between the nodes (keywords) is equal to the *fractional* co-occurrence of keywords i and j in the same works. It holds that

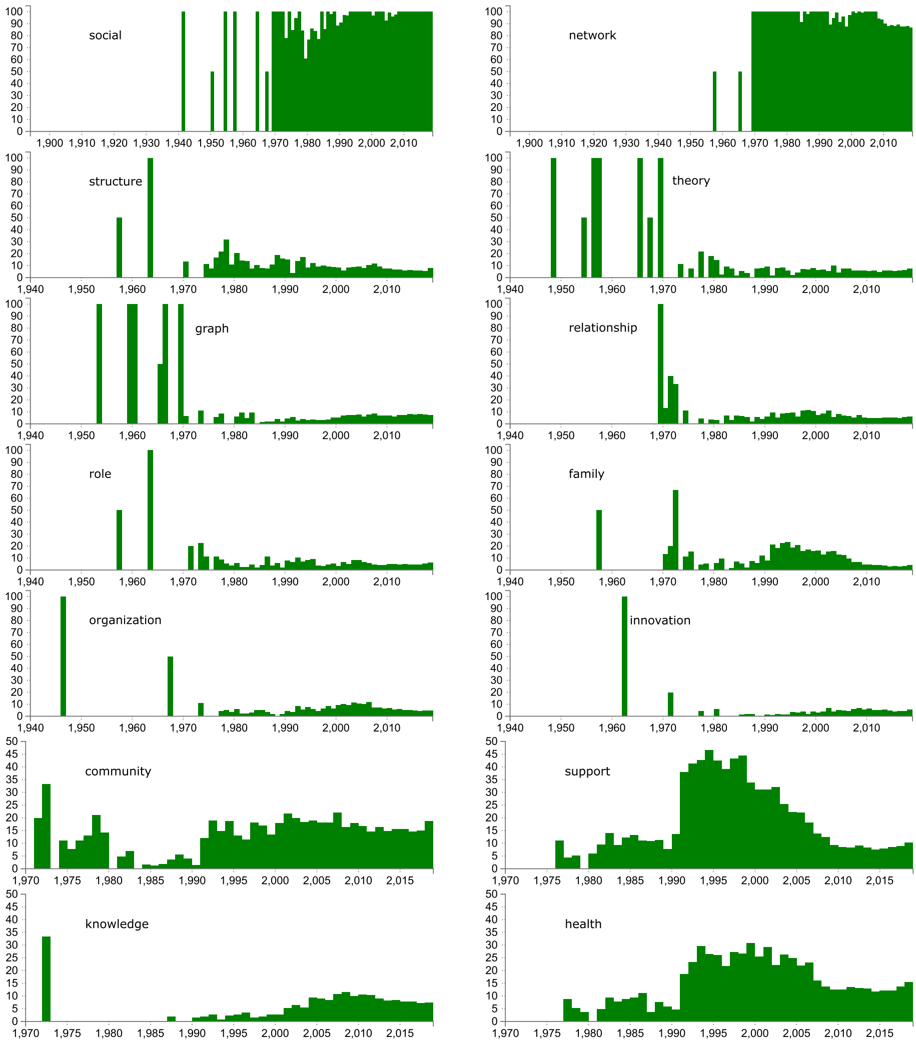

Fig. 3 Distribution of proportions of keywords: scales of 100% and 50%

$\mathbf{nKK}[i,j] = \mathbf{nKK}[j,i]$ and $\sum_{i,j} \mathbf{nKK}[i,j] = |W|$—each work has value 1 that is redistributed over keywords (Batagelj et al. 2019).

Keyword co-occurrence network analysis

An exploratory analysis showed that in the **nKK** network the most frequent words *social*, *network*, and *analysis* connected most of the other keywords, which is why we excluded these three nodes from the network. Using the *Link Islands approach* (Batagelj et al. 2014), we searched for subnetworks sized from 2 to 75 nodes. A large number of islands (342) was obtained, where the majority of islands (301) represented only pairs of keywords. The main island includes 75 nodes; there are also some islands of smaller sizes.

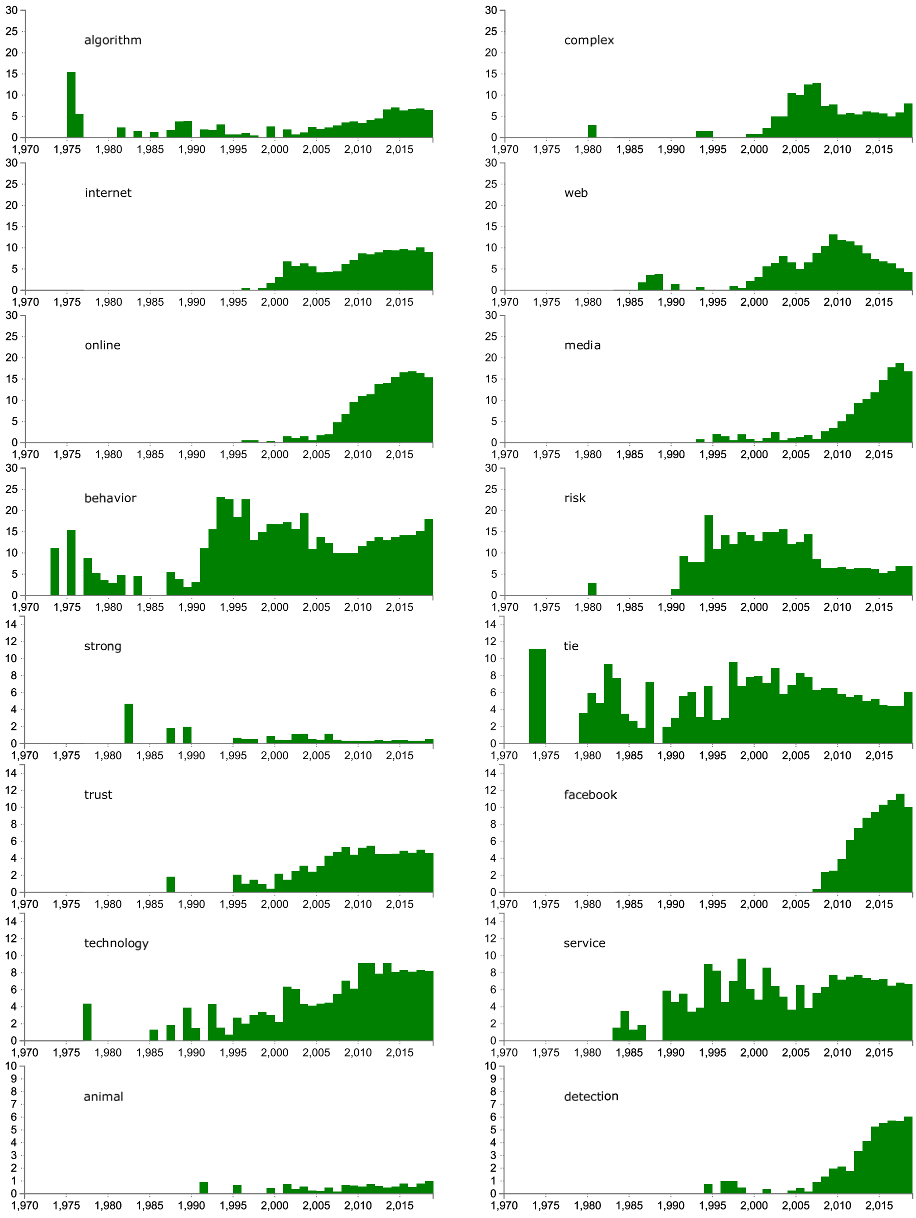

Fig. 4 Distribution of proportions of keywords: scales of 30%, 14%, and 10%

A large part of the main island (Fig. 5) consists of keywords on the topic of networking sites and social media (*networking, online, site, service, internet, web 2.0, semantic, technology, media, facebook, twitter, technology*). Other words connected to this group are *information*, *use*, *user*, *privacy*, and *security*, presumably raising the issues of networking service usage. *Information* is also connected to the words *diffusion, innovation, knowledge*, and *management*.

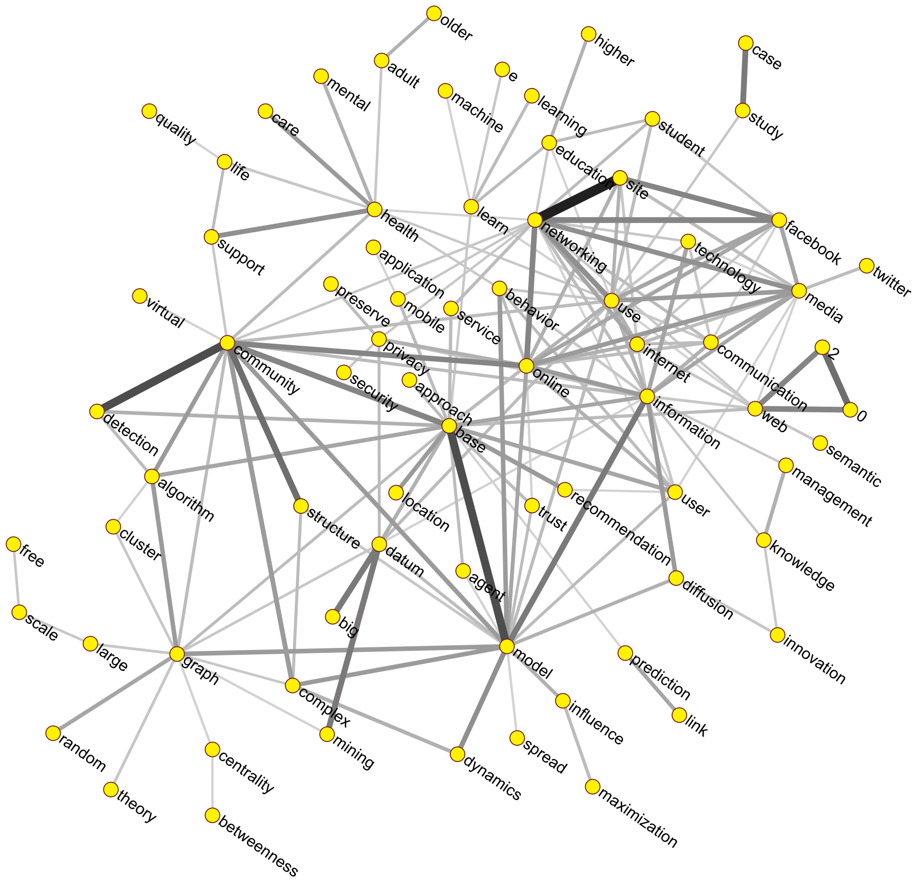

Fig. 5 The main island of the **nKK** network

Other central keywords are *base*, connected to the words *model* and *community* (also connected to each other). *Model* is connected to the words *dynamics, complex, spread, influence* (with latter connected to *maximization*), and *community*—by *detection, structure, complex, algorithm*, and *virtual*. Another group of words connected *graph* is *algorithm, model, random, theory, centrality* (connected to *betweenness*), *large* (connected to *scale* linked to *free*). Other locally highly connected groups are formed by the words *datum, big*, and *mining*, *prediction* and *link*. These nodes, which are the largest part of the main component, form a group of keywords on the methodological issues of SNA.

Some words appearing in this subnetwork are associated with substantial topics in SNA: on health (*health, support, life, care, mental, adult, behavior*) and education (*education, higher, student, learn, e–, learning*). *Learn* is also connected to the word *machine*, a developing topic in computer science.

Other islands identify some expressions from topics being developed in SNA (*strength, weak, tie; corporate - interlock - directorate; triadic - closure; small - world*, or some broad topics from substantive studies (*organ - donor - donation; persecutory - delusion - paranoia; trade - international - migration*), and some stable phrases with limited value (*special, issue, introduction*).

To go deeper into the meaning of the keywords, we looked at them in different contexts—the contexts associated with selected groups of authors and journals which were found to be important during our previous analysis of co-authorship, citation and bibliographic coupling structures among authors and journals.

Keywords and authors

Network construction

To construct the network of authors and keywords **AK**, we used the normalized reduced networks **WAr** and **WKr**. The first network was transposed and then multiplied by the second in the following way:

$$\mathbf{nAK} = n(\mathbf{WA})^T * n(\mathbf{WK})$$

The obtained network is normalized. In this network, the weight $\mathbf{nAK}[a, k]$ of the edge between the nodes a and k is equal to the fractional use of author a of keyword k. It can be extended to a group of authors C, for a given keyword k:

$$\mathbf{nAK}[C, k] = \sum_{a \in C} \mathbf{nAK}[a, k]$$

In this section, we used the results of the analysis of co-authorship networks between the authors in the field of SNA. From the network **WAr**, which consisted of 70,792 works and 93,011 authors, we created collaboration network $\mathbf{Ct}'$ (Batagelj and Cerinšek 2013; Batagelj et al. 2014). We used normalization proposed by Newman (2001), who interpreted collaboration in a "strict" way—as a collaboration only with others (excluding single authored papers). In this case, for the initial **WAr** network the weight of each arc is divided by the sum of the weights of all arcs having the same initial node (its outdegree) subtracting the initial author (which is 1). Then the network $\mathbf{Ct}'$ is constructed by the transposition of the regularly normalized $\mathbf{n}(\mathbf{WA})$ network and multiplying it by the Newman normalized $\mathbf{n}'(\mathbf{WA})$ network.

$$n'(\mathbf{WA})[w, a] = \frac{\mathbf{WA}[w, a]}{\max(1, \text{outdeg}(w) - 1)}$$

then

$$\mathbf{Ct}' = \mathbf{n}(\mathbf{WA})^T * \mathbf{n}'(\mathbf{WA})$$

The obtained $\mathbf{Ct}'$ is undirected and does not have loops. The contribution of a complete subgraph corresponding to each work is 1. The weights of the edges between the nodes (authors) are equal to the total contribution of the "strict collaboration" of authors i and j to works they wrote together. The total contribution for an author is counted by line weights—it is equal to the sum of the weights of all the works he or she co-authored.

Keywords used by selected groups of authors

To extract the groups of authors collaborating with each other from the $\mathbf{Ct}'$ network, we used the *Islands approach* (Batagelj et al. 2014). We generated 14,222 simple islands of between 2 and 50 nodes (in sum, 45,524 nodes, or 45% of all nodes in the network). The sizes and number of islands show that there are many groups of collaborating authors that can be extracted out of the $\mathbf{Ct}'$ network. There are different ways to identify the islands for the further inspection, based on the size of islands, largest values of line weights, or specific names. To get islands with really strong ties, we removed all the lines lower the threshold of 7.5 from the $\mathbf{Ct}'$ network and extracted the network of 32 nodes. Then we manually searched for the islands to which these 32 nodes belong, and extracted them. Another approach used was to search for the structures for some well-known authors.

For presenting the keywords associated with groups of authors, we have chosen simple islands represented by `BARABASI_A` (8 authors), `BORGATTI_S`, `SNIJDERS_T` (4 authors each), `CHRISTAKIS_K`, `SKVORETZ_J` (3 authors each), `WASSERMAN_S`, `PATTISON_P`, `VALENTE_T`, `DOREIAN_P` (2 authors each) for the extraction of keywords. The selected islands with the members of each group are presented in Fig. 6. The top-20 keywords for each group are presented in Tables 2, 3 and 4. The top keywords for these clusters are the trivial keywords *network* and *social*. Other keywords can provide some description of the topics being studied by selected groups of authors, oriented either on methodological or substantive issues.

The island of Borgatti, Everett, Boyd, and Halgin can be attributed to the methodological group, having the keywords *graph, centrality, role, regular, equivalence, semigroup, structure, clique*, and *homomorphism*, as can the pair of Doreian and Conti with the words *equivalence, evolution, journal, balance, blockmodeling, generalized, regular, ranking*. For Robins and Pattison the words are *model, graph, random, Markov, logit, logistic, regression, exponential, p, semigroup, asterisk, multirelational*. The pair of Wasserman and Faust can be represented with the words *correction, model, exchange, stochastic, structure, statistical, blockmodel, equivalence, logit, triad* (there are also *logistic* and *regression* in 23th and 24th places). The group of 4 authors connected to Snijders has the keywords *Markov, random, friendship, behavior, peer, inference, influence, stochastic, actor, longitudinal, orient* which reflect their work in stochastic actor-oriented models. The island represented

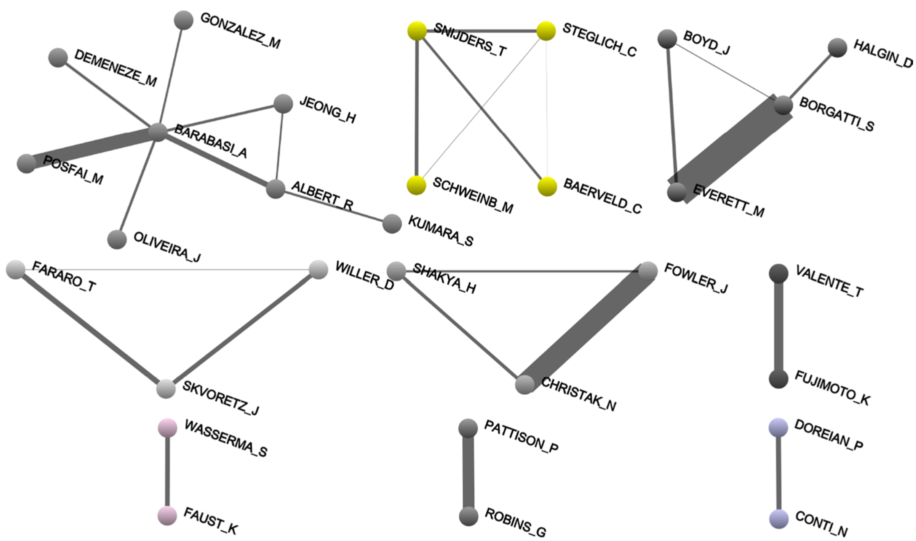

Fig. 6 Collaboration network: selected simple islands

Table 2 Keywords used in the clusters of authors from Fig. 6 (1)

	BORGATTI_S		BARABASI_A		CHRISTAKIS_K	
Rank	Value	Id	Value	Id	Value	Id
1	4.9303	**network**	7.0709	**network**	3.1788	**network**
2	2.5918	**social**	2.0782	**social**	2.9358	**social**
3	2.0858	graph	1.7068	dynamics	1.0204	spread
4	1.4210	centrality	1.6670	complex	1.0192	behavior
5	1.4202	analysis	1.6362	scale	0.7261	health
6	1.3399	role	1.5946	web	0.5512	large
7	1.2780	regular	1.5516	community	0.5169	model
8	1.2424	equivalence	1.4709	world	0.4778	smoking
9	1.0530	semigroup	1.3622	internet	0.4522	human
10	1.0000	correction	1.1906	model	0.4479	cooperation
11	0.9891	**structure**	1.1858	free	0.4313	obesity
12	0.7755	clique	1.0210	evolve	0.4125	influence
13	0.7576	homomorphism	1.0087	science	0.3973	life
14	0.7241	relation	0.9808	random	0.3728	dynamics
15	0.6346	power	0.9476	wide	0.3715	evolution
16	0.6301	betweenness	0.8178	human	0.3463	analysis
17	0.6287	exchange	0.8076	theory	0.3286	cosponsorship
18	0.6232	algorithm	0.7561	small	0.3044	norm
19	0.6167	similarity	0.7536	graph	0.3036	trial
20	0.5595	ebloc	0.6603	phenomenon	0.2985	study
Total	63.0810		76.6373		46.8865	

by Skvoretz has the keywords *power, exchange, bias, model, correction, theorem, approximation, simulation, dynamic*.

The network science representatives—the group of 8 authors with Barabási, Posfai, Albert, and others—can also be attributed to the methodological stream, having the words *dynamics, complex, scale, web, community, world, internet, model, free, evolve*, and *random*.

The top keywords for other selected groups cover some substantive issues. The group of Fowler, Christakis, and Shakya have keywords *spread, behavior, health, smoking, human, cooperation, obesity, influence, evolution, dynamics*. The group of Valente is represented by the words *health, diffusion, behavior, innovation, peer, adolescent, influence, smoking, prevention, cigarette, leader*. As an example, it is interesting to compare the latter with the description on the official home page of Thomas Valente, who is working on the topics of *social networks, behavior change, and program evaluation* and *uses social network analysis, health communication, and mathematical models to implement and evaluate health promotion programs designed to prevent tobacco and substance abuse, unintended fertility, and STD/HIV infections*, and is *also engaged in mapping community coalitions and collaborations to improve health care delivery and reduce healthcare disparities*.

Some simple islands form larger general islands. The general island formed by the groups of `SNIJDERS_T`, `SKVORETZ_J`, `WASSERMAN_S`, `PATTISON_P`, and `DOREIAN_P` is presented in Fig. 7. The keywords for this island are presented in Table 5.

Table 3 Keywords used in the clusters of authors from Fig. 6 (2)

PATTISON_P			SNIJDERS_T		VALENTE_T	
Rank	Value	Id	Value	Id	Value	Id
1	2.2196	**network**	2.6375	**network**	2.5536	**network**
2	2.0729	**social**	2.0902	**social**	1.9553	**social**
3	1.7567	model	1.6702	model	1.0000	untitled
4	1.3084	graph	1.0692	graph	0.9419	health
5	0.8939	random	0.8857	dynamics	0.8737	diffusion
6	0.8583	markov	0.7390	markov	0.7802	behavior
7	0.8531	logit	0.6903	random	0.7402	innovation
8	0.8220	logistic	0.6734	friendship	0.6974	model
9	0.8220	regression	0.6228	datum	0.6521	use
10	0.8012	exponential	0.5932	statistical	0.6349	peer
11	0.7055	analysis	0.5780	behavior	0.6216	adolescent
12	0.6752	p	0.5547	analysis	0.5717	influence
13	0.5530	statistical	0.5423	peer	0.5610	smoking
14	0.5038	**structure**	0.5383	inference	0.5371	analysis
15	0.3561	semigroup	0.5346	influence	0.5247	prevention
16	0.3522	asterisk	0.4623	stochastic	0.4987	cigarette
17	0.3368	process	0.4612	actor	0.4979	opinion
18	0.3333	multirelational	0.4480	selection	0.4860	leader
19	0.3249	family	0.4372	longitudinal	0.4545	risk
20	0.3031	dynamics	0.3785	orient	0.4491	intervention
Total	38.6110		46.6732		44.8812	

We can see that the keywords with largest values are more commonly used words, such as *network, social, model, analysis, graph, structure, datum, structural, theory, group, method*. However, there are also special words on methodological issues, mentioned in the islands above, such as *correction, exchange, equivalence, random, power, markov, evolution, statistical, dynamics, generalized, regression, exponential, blockmodel, logit, p, cluster, logistic, dynamic, blockmodeling*. This is a group of authors dealing with methodological issues in SNA.

Keywords and journals

Network construction

To construct the derived network of journals and keywords, **JK**, we used the normalized reduced networks **WJr** and **WKr**. The first network was transposed and then multiplied by the second in the following way:

$$\mathbf{nJK} = n(\mathbf{WJ})^T * n(\mathbf{WK})$$

Table 4 Keywords used in the clusters of authors from Fig. 6 (3)

	SKVORETZ_J		WASSERMAN_S		DOREIAN_P	
Rank	Value	Id	Value	Id	Value	Id
1	3.8058	**network**	2.4529	**network**	6.0097	**network**
2	1.6586	power	1.6875	**social**	3.7088	**social**
3	1.6277	exchange	1.0000	correction	1.5308	equivalence
4	1.5218	**social**	0.9414	analysis	1.4972	evolution
5	1.2301	bias	0.7270	model	1.4917	journal
6	1.0751	model	0.5509	graph	1.2177	**structural**
7	1.0000	correction	0.4818	datum	1.0395	measure
8	0.9204	**structure**	0.4595	method	0.9402	**structure**
9	0.7765	theory	0.4457	exchange	0.8107	group
10	0.6341	theorem	0.4319	stochastic	0.7987	balance
11	0.5001	tie	0.4282	**structure**	0.6923	analysis
12	0.4119	**structural**	0.3554	statistical	0.5395	actor
13	0.3972	weak	0.3501	blockmodel	0.5067	blockmodeling
14	0.3905	approximation	0.3438	kinship	0.4917	utility
15	0.3905	simulation	0.3308	equivalence	0.4870	model
16	0.3883	dynamic	0.3118	**structural**	0.4711	generalized
17	0.3436	theoretical	0.3079	logit	0.4667	stand
18	0.3371	strength	0.2666	relation	0.4339	connectivity
19	0.3108	analysis	0.2611	triad	0.4333	ranking
20	0.3105	sociology	0.2611	census	0.4238	regular
Total	33.5190		29.1417		48.1875	

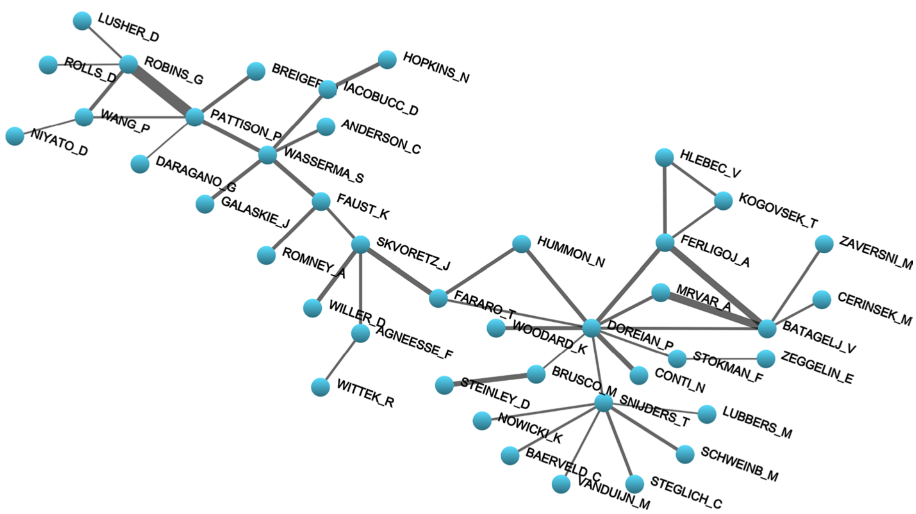

Fig. 7 Collaboration network: a general island formed out of several simple islands

Table 5 Keywords used in the general island of authors (Fig. 7)

Rank	Value	Id	Rank	Value	Id
1	30.0225	**network**	21	1.8844	generalized
2	20.1127	**social**	22	1.8226	journal
3	8.6241	model	23	1.8012	regression
4	7.3574	analysis	24	1.7816	exponential
5	6.0054	graph	25	1.7772	blockmodel
6	5.5047	**structure**	26	1.7639	logit
7	3.1894	datum	27	1.7326	balance
8	3.0265	**structural**	28	1.7253	p
9	3.0000	correction	29	1.6844	measure
10	2.9594	exchange	30	1.6639	algorithm
11	2.7971	equivalence	31	1.6584	cluster
12	2.6809	random	32	1.6381	approach
13	2.5432	theory	33	1.6222	actor
14	2.5255	power	34	1.5873	logistic
15	2.5081	markov	35	1.5509	relation
16	2.4107	evolution	36	1.5398	introduction
17	2.2839	group	37	1.5356	bias
18	2.2531	statistical	38	1.5144	dynamic
19	2.1939	method	39	1.4467	blockmodeling
20	2.1816	dynamics	40	1.4391	friendship

The network is normalized. In this network, the weight, **nJK**[j, k], on the edges between the nodes j and k is equal to the *fractional contribution* of journal j for given keyword k; or for a group of journals C:

$$\mathbf{nJK}[\mathrm{C,k}] = \sum_{j \in C} \mathbf{nJK}[j, k]$$

We used the *TF–IDF approach* to line weighting (Robertson 2004), which allows us to evaluate the importance of a word to a document in a corpus of documents. The importance increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus. In our case, **TF** shows the number of times a keyword appears in a selected journal, divided by the total number of keywords in the journal, and **IDF** is the logarithm of the number of journals in the corpus divided by the number of journals where the specific keyword appears. We used the reduced networks **WJr** and **WKr** for the **JKr** network construction, and calculated **TF–IDF** indexes for the keywords in the following way:

$$\mathbf{TF-IDF}(\text{keyword, JOUR}) = \mathbf{TF}(\text{keyword, JOUR}) * \mathbf{IDF}(\text{keyword})$$

$$\mathbf{TF}(\text{keyword, JOUR}) = \frac{\text{\# of times keyword appeared in JOUR}}{\text{Total \# of keywords in JOUR}}$$

$$\mathbf{IDF}(\text{keyword}) = \log \frac{\text{\# of JOURs}}{\text{\# of JOURs with keyword}}$$

Keywords in selected journals

In our analysis, we identified journals intensively used in SNA. To present the analysis of the keywords associated with these journals, we have chosen journals *Social Networks* (SOC NETWORKS), *Lecture Notes in Computer Science* (LECT NOTES COMPUT SC), *Physica A* (PHYSICA A), *PLOS ONE* (PLOS ONE), *American Journal of Sociology* (AM J SOCIOL), and *Animal Behaviour* (ANIM BEHAV).

Using the *fractional approach* to network normalization, we extracted the top keywords associated with the selected journals (Tables 6, 7). As shown above, the most used keywords are trivial, and many frequently used words are generic, giving limited

Table 6 Selected journals and keywords: fractional approach (1)

SOC NETWORKS			LNCS		PHYSICA A	
Rank	Value	Id	Value	Id	Value	Id
1	80.4616	**network**	133.7777	**social**	31.3976	**network**
2	48.9783	**social**	127.8566	**network**	26.2949	**social**
3	19.4413	model	30.2293	base	16.0265	complex
4	16.7508	**structure**	26.2529	analysis	14.2695	model
5	16.3657	analysis	23.1342	graph	10.9905	dynamics
6	12.5557	graph	22.3405	model	7.8218	community
7	11.5428	centrality	19.9326	information	5.6892	**structure**
8	9.6468	tie	19.4907	online	5.6236	spread
9	8.5549	datum	18.2891	user	5.4035	base
10	8.3412	**structural**	18.0402	web	5.0177	world
11	6.7409	personal	17.5844	community	4.2681	information
12	6.7131	power	16.669	datum	4.1653	evolution
13	6.3861	measure	14.9789	use	3.9429	scale
14	5.5022	community	14.9197	privacy	3.8668	small
15	5.4449	organization	11.8789	algorithm	3.8318	online
16	5.3927	group	9.761	learn	3.5402	detection
17	5.3734	random	9.682	influence	3.4912	analysis
18	5.1904	theory	9.5126	service	3.3722	free
19	5.126	exchange	9.4945	networking	3.2738	graph
20	5.0863	communication	9.0722	detection	3.2399	epidemic
21	5.053	equivalence	8.9043	trust	3.2016	diffusion
22	5.0385	correction	8.7554	recommendation	3.0552	opinion
23	5.0248	dynamics	8.2151	mobile	3.0089	behavior
24	4.9305	support	8.0529	search	2.7643	centrality
25	4.6783	friendship	7.9805	approach	2.7129	game
26	4.5928	relation	7.8424	media	2.6598	rumor
27	4.289	effect	7.7627	semantic	2.6145	algorithm
28	3.9814	role	7.5789	mining	2.3185	node
29	3.9413	note	7.5489	twitter	2.2702	effect
30	3.8269	use	7.3363	application	2.2489	propagation
Total	1132.333		1991.5		470	

Table 7 Selected journals and keywords: fractional approach (2)

PLOS ONE			AM J SOCIOL		ANIM BEHAV	
Rank	Value	Id	Value	Id	Value	Id
1	33.6687	**network**	8.5444	**network**	3.9838	**social**
2	28.8837	**social**	7.7785	**social**	3.6231	**network**
3	7.7329	dynamics	3.0574	model	2.0308	behavior
4	7.355	behavior	2.4432	**structure**	1.4271	group
5	6.9541	complex	1.8205	tie	1.4138	**structure**
6	6.205	model	1.5329	market	1.3927	dynamics
7	6.0393	community	1.2604	dynamics	1.2574	association
8	5.0104	analysis	1.1914	organization	1.235	pattern
9	4.5741	health	1.1189	power	1.1695	analysis
10	4.0371	population	1.0609	theory	1.0874	population
11	3.7547	pattern	1.0159	friendship	1.0206	evolution
12	3.7441	use	0.9794	family	0.9795	animal
13	3.7143	evolution	0.9614	exchange	0.9346	dominance
14	3.6934	**structure**	0.9296	action	0.9012	size
15	3.6774	information	0.8719	behavior	0.8691	behaviour
16	3.5965	scale	0.8073	weak	0.8675	female
17	3.5392	human	0.8024	collective	0.8178	individual
18	3.5196	online	0.8013	class	0.8022	organization
19	3.2497	base	0.7803	strength	0.7789	wild
20	3.239	spread	0.7695	community	0.7159	selection
21	3.1964	risk	0.7671	world	0.6431	reproductive
22	3.1864	communication	0.7567	analysis	0.6024	primate
23	3.1044	disease	0.716	culture	0.5953	dolphin
24	2.7954	study	0.7036	state	0.5862	fission
25	2.7086	cooperation	0.6914	**structural**	0.5387	interaction
26	2.4118	world	0.6909	diffusion	0.536	transmission
27	2.3582	emergence	0.6746	industry	0.5213	fusion
28	2.2015	twitter	0.648	embeddedness	0.495	success
29	2.075	influence	0.5805	small	0.4761	macaque
30	2.0716	impact	0.5781	unit	0.4729	male
Total	667		132		107	

value. However, other words represent the features of the discourse provided by each of the selected journals.

For *Social Networks*, such keywords are *centrality, measure, random, equivalence, role*, describing methodological issues, and *community, organization, group, exchange, communication, support, friendship*, focusing on substantive ones. Comparing this set with the keywords associated with the *American Journal of Sociology*, we can support the observation of Leydesdorff et al. (2008) that in the social sciences SNA is used in studies on substantive, not methodological issues. Keywords for `AM J SOCIOL` are *market, organization, power, friendship, family, exchange, action, collective, behavior,* *class, community, culture, state, industry*; however, there are also keywords reflecting the traditional terms of SNA (*strength, weak, embeddedness, diffusion, small, world*).

For *Lecture Notes in Computer Science*, the special keywords are those describing computer networks and services (*online, web, privacy, service, networking, recommendation, mobile, media, twitter*) and those representing the computer science issues being studied (*detection, semantic, mining*). For *Physica A*, the most used keywords identify the methodological and substantive issues which network scientists are working on (*complex, dynamics, evolution, community, detection, spread, small, world, free, scale, epidemic, diffusion, propagation, opinion, behavior, rumor, online*); the "traditional" SNA term *centrality* also appears in the list. The most frequently used keywords for the general scientific journal *PLOS ONE* are similar to those in *Phisica A* (*dynamics, complex, behavior, community, evolution, online*), but health research has a bigger focus (*health, population, human, risk, disease, spread, influence*). In *Animal Behaviour*, attention is given to the studies of animals and nature, which are associated with the keywords *animal, individual, population, female, male, wild, selection, reproductive, primate, dolphin, fission, macaque*, being studied in sense of the *behavior* (and *behaviour*), *group, structure, dynamics, association, evolution, dominance, organization, interaction, transmission*, and *success*.

The results obtained by the *TF–IDF approach* (Tables 8, 9) are similar to the results of fractional normalization. However, they even more clearly show the special features of the discourses provided in the selected journals. For all the journals, besides *LNCS*, the keyword *social* moved away, and *network* is far from the first place. In the list of the top keywords in *Animal Behaviour*, the trivial and generic words with limited value are replaced by the terms from biology. The words *structure* and *structural* remain in the lists of the top keywords in all the journals. We can conclude that this approach better identifies the keywords associated with some substantial topics developing in SNA.

Discussion and conclusion

This paper provides an insight into the topics developed in SNA and reveals important information for its systematic description. As previous studies have shown, the identification of the keywords used in publications can provide important information on the discourse developed in the field and its main streams and topics, either methodological or substantial. However, it was also shown that the results of such an analysis should be examined with a great care. Small samples mean the networks for separate years can be significantly different, both in the set of words and their peripheral or central positions. The most used keywords can be trivial and anticipated before the analysis, or generic with limited value. Last but not least, the results are inevitably connected to the data, which are in turn dependent on the databases used for data collection and the queries used for identifying works: depending on these, the results can reflect certain disciplines, fields or subfields, and can be oriented to methodological or substantive issues.

In this study, we used the keywords obtained from the works published in the WoS database matching the search query "social network*", influential works, and those published in the main journals indexed in the WoS. The time coverage is from the very first articles published in 1970s, up to 2018. 32,409 keywords were obtained from 70,792 works with complete descriptions, from the fields Author Keywords, Keywords Plus, and titles. The distributions of the numbers of all keywords used and the unique (different) keywords over time show accelerated growth starting from 2007, which was the

Table 8 Selected journals and keywords: TF–IDF index (1)

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	

result of the increasing number of works on SNA in various scientific fields, and disciplines applying SNA in their studies, as shown by Maltseva and Batagelj (2019). The wide scope of the contexts where SNA is applied is also shown by the large number of keywords which are used episodically.

In our analysis, we looked at the distribution of the most frequently used keywords in a two-mode network of works × keywords and the islands obtained from the normalized one-mode network of keyword co-occurrence. To go deeper, we placed the keywords into the clearly defined contexts: selected groups of authors closely connected to each other according to their co-authorship, and selected journals representing different disciplines. These results support the conclusions made in previous studies: the most used

Table 9 Selected journals and keywords: TF–IDF index (2)

PLOS ONE			AM J SOCIOL		ANIM BEHAV	
Rank	Value	Id	Value	Id	Value	Id
1	0.0841	dynamics	0.0739	model	0.1059	wild
2	0.0732	complex	0.0659	**structure**	0.1033	dominance
3	0.0670	behavior	0.0588	friendship	0.1012	dolphin
4	0.0667	population	0.0554	dynamics	0.0935	animal
5	0.0547	spread	0.0552	tie	0.0920	fission
6	0.0538	disease	0.0540	segregation	0.0883	association
7	0.0510	evolution	0.0534	interracial	0.0865	reproductive
8	0.0490	health	0.0497	organization	0.0848	bottle
9	0.0488	human	0.0470	action	0.0824	nose
10	0.0482	pattern	0.0445	market	0.0806	female
11	0.0472	risk	0.0435	racial	0.0774	dynamics
12	0.0471	**network**	0.0420	exchange	0.0755	**structure**
13	0.0469	scale	0.0419	industry	0.0742	behavior
14	0.0468	model	0.0380	**network**	0.0720	pattern
15	0.0461	cooperation	0.0374	state	0.0717	group
16	0.0445	transmission	0.0366	collective	0.0707	size
17	0.0414	hiv	0.0364	unit	0.0692	primate
18	0.0402	emergence	0.0357	logit	0.0688	population
19	0.0397	epidemic	0.0354	world	0.0675	evolution
20	0.0390	**structure**	0.0346	small	0.0666	fusion
21	0.0359	community	0.0338	embeddedness	0.0597	baboon
22	0.0356	infection	0.0336	race	0.0597	macaque
23	0.0348	size	0.0335	power	0.0563	individual
24	0.0342	sex	0.0333	diffusion	0.0562	behaviour
25	0.0331	influenza	0.0320	job	0.0531	selection
26	0.0330	adult	0.0314	class	0.0527	tit
27	0.0311	infectious	0.0309	culture	0.0524	male
28	0.0308	individual	0.0309	intergroup	0.0521	bottlenose
29	0.0307	analysis	0.0308	**structural**	0.0517	tursiop
30	0.0304	game	0.0304	theory	0.0501	reticula
Total	13.7681		7.7723		12.0405	

keywords are trivial—*social* and *network*,—and many other frequently used keywords have limited value: they express terms commonly used in research in general (such as *datum, base, information, research, theory, algorithm, approach*), or in SNA (such as *graph, structure, relationship, role, tie*). Temporal analysis showed the constant presence and usage of these words (counted as the proportion of the number of the appearances of *each keyword* to the *most frequent keyword* appearance for each year).

Another group of topics identified in SNA can be assigned to the methodological stream. These topics appeared in the analysis of co-occurrence networks and (mostly) in the lists of frequent keywords used by selected groups of authors and by selected journals. In the main island of the **KKn** network, we identify the topics of graph theory, dynamic and complex network models, models of spread and influence maximization, agent based models, random graph models, large scale-free networks, community detection algorithm, link prediction, graph centrality, innovation diffusion, semantic web, machine learning, big data, and data mining. Other islands identify some topics traditionally developed in SNA, such as the strength of weak ties, triadic closure, interlock directorates, and small world. With the previous group of trivial and general keywords in SNA, these words can be seen as the core concepts of the field.

Besides these keywords, the list of the most used keywords largely provided the keywords representing some substantive topics developed in SNA: networking sites and social media; community, family, health, education studies; trust and support; innovation and influence. A temporal analysis of the keywords associated with these topics showed that while some of them (*community, support, health, behavior*) are present in the field from 1970s, others appeared later, in the 1980s and the 1990s (*technology, service, web*) and in the 2000s (words connected to internet and media studies). Some of the words could change their topical origin over time—for example, the word *community*, which could be associated with the studies of offline communities in 1970s and 1980s, online communities in the 1990s and 2000s, and the algorithms of community detection from 2010s (as the usage of the word *detection* also increased from this time). The analysis of the co-occurrence network **nKK** adds other substantive issues connected with networking sites and social media—the topics of information privacy, security and information management.

Methodological and substantive streams are also found in the selected groups of authors and journals representing different disciplines. We identified a set of social network analysts (the group's representatives are Borgatti, Pattison, Wasserman, Doreian, Snijders, and Skvoretz) working mostly on the the methodological issues of SNA, and several other groups (representatives are Valente and Christakis) who work on substantial issues. The group of network scientists (represented by Barabási) was also attributed to the methodological stream. The analysis of keywords for journals showed the disciplinary differences between the selected sources. The comparison of *Social Networks* with the *American Journal of Sociology* showed that the former is mostly methodologically oriented while the later applies the tools of SNA for substantive studies, supporting the previous observation of Leydesdorff et al. (2008). *Lecture Notes in Computer Science* is devoted to the topics of internet networks and services, developed by the computer scientists. *Physica A* is in a way similar to the general scientific journal *PLOS ONE*—both focus on issues developed in network science; however, the latter also focuses on health studies. *Animal Behavior* publishes works on the social networks of animals. The proposed *fractional* and *TF–IDF* approaches showed their strengths in the identification of the keywords for selected subgroups, and the latter was better at identifying keywords associated with substantial topics. We suppose that these approaches can be further used for the extraction of the unit (author or journal) identities and their clustering according to similarity, and this could be a direction for future research.

There are some limitations in the current study. First of all, the initial search was oriented towards ***social*** networks, and thus some works related to a broader field of network analysis in general could have been overlooked. At the same time, the search query for "network analysis" would be too broad (beyond the data analysis), including the works related to computer networks, optimization problems on networks, etc. That is why, on the first step, our search was somehow limited. However, on the second step, we extended the results of the original query and added works initially not included in the search. This additional 'saturation' search of the papers which were cited a lot by the field's representatives, as well as inclusion of the works from journals important for the field, and the most prominent authors allowed us to improve the dataset in the sense of broader inclusion of the publications related to network analysis in general. Thus, the obtained dataset covers not only the works of social scientists, but also influential papers published by physicists, biological scientists, information and computer scientists, etc. This search allowed also to include additional influential papers, usually published earlier, that could have been overlooked by our search queries because they do not use the contemporary terminology. Second, our dataset is based on the information available in the WoS. Adding publications from the journals not indexed in the WoS, or the analysis of some smaller datasets (e.g., articles from specific journals) could provide extra results. For the further analysis, the obtained dataset can be extended through the additional search queries, such as "complex network*" and "network science", and usage of other bibliographic databases, which will make the view of the whole landscape of network analysis more complete and conclusive. Although we do not expect substantial changes in the top level results. Third, even though the choice of authors and journals is motivated by the previous analysis of co-authorship, citation and bibliographic coupling structures among authors and journals, the choice of the authors' groups and journals is partially subjective. That is why it should be seen as an illustration of a methodological approach. Finally, the approach of temporal network analysis, which is applied to large bibliographic networks for the first time, needs further developments in the reading and visualization of the results. This is one of the tasks for the future.

Acknowledgements We would like to express our special gratitude to our colleague, professor Anuška Ferligoj (University of Ljubljana and the International Laboratory for Applied Network Research, Moscow) for her advice and comments which greatly improved the manuscript. We appreciate the help of David Connolly (Academic Writing Center, Higher School of Economics, Moscow) with the proofreading of the article. This work is supported in part by the Slovenian Research Agency (Research Program P1-0294 and Research Projects J1-9187 and J7-8279), project COSTNET (COST Action CA15109), and by Russian Academic Excellence Project '5-100'. The funding sources had no involvement in the study and article.

References

Batagelj, V. (2014). Nets—Python package for network analysis. Available at: https://github.com/bavla/Nets/tree/master/source.

Batagelj, V. (2017). WoS2Pajek. Networks from Web of Science. Version 1.5 (2017). Available at: http://vladowiki.fmf.uni-lj.si/doku.php?id=pajek:wos2pajek.

Batagelj, V. (2019). *On fractional approach to analysis of linked networks*. Available at: arxiv:1903.00605.

Batagelj, V., & Maltseva, D. (2020). Temporal bibliographic networks. *Journal of Informetrics*. https://doi.org/10.1016/j.joi.2020.101006

Batagelj, V., & Cerinšek, M. (2013). On bibliographic networks. *Scientometrics*, *96*(3), 845–864.

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.

Batagelj, V., Ferligoj, A., & Doreian, P. (2020). Bibliometric analysis of the network clustering literature. In P. Doreian, V. Batagelj, & A. Ferligoj (Eds.), *Advances in network clustering and blockmodeling* (pp. 63–102). Hoboken, NJ: Wiley.

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

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

Bonacich, P. (2004). The invasion of the physicists. *Social Networks*, *26*, 285–288.

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

Brandes, U., & Pich, C. (2011). Explorative visualization of citation patterns in social network research. *Journal of Social Structure*, *12*(8), 1–19.

Freeman, L. C. (2004). *The development of social network analysis. A study in the sociology of science*. Vancouver, BC: Empirical Press.
Freeman, L. C. (2011). The development of social network analysis-with an emphasis on recent events. *The SAGE Handbook of Social Network Analysis*, *21*(3), 26–39.
Gauffriau, M., Larsen, P., Maye, I., Roulin-Perriard, A., & von Ins, M. (2007). Publication, cooperation and productivity measures in scientific research. *Scientometrics*, *73*(2), 175–214.
Hummon, N. P., & Carley, K. (1993). Social networks as normal science. *Social Networks*, *15*(1), 71–106.
Hummon, N. P., Doreian, P., & Freeman, L. C. (1990). Analyzing the structure of the centrality-productivity literature created between 1948 and 1979. *Science Communication*, *11*(4), 459–480.
Kejžar, N., Černe, S. K., & Batagelj, V. (2010). Network analysis of works on clustering and classification from web of science. *Classification as a tool for research* (pp. 525–536). Berlin, Heidelberg: Springer.
Lazer, D., Mergel, I., & Friedman, A. (2009). Co-citation of prominent social network articles in sociology journals: The evolving canon. *Connections*, *29*(1), 43–64.
Leydesdorff, L., Schank, T., Scharnhorst, A., & De Nooy, W. (2008). Animating the development of Social networks over time using a dynamic extension of multidimensional scaling. *El Profesional de Informacion*. https://doi.org/10.3145/epi.2008.nov.04.
Maltseva, D., & Batagelj, V. (2019). Social network analysis as a field of invasions: Bibliographic approach to study SNA development. *Scientometrics*, *121*(2), 1085–1128. https://doi.org/10.1007/s11192-019-03193-x.
Newman, M. E. (2001). Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. *Physical Review E*, *64*(1), 016132.
Otte, E., & Rousseau, R. (2002). Social network analysis: A powerful strategy, also for the information sciences. *Journal of Information Science*, *28*(6), 441–453.
Robertson, S. (2004). Understanding inverse document frequency: On theoretical arguments for IDF. *Journal of Documentation*, *60*(5), 503–520.
Varga, A. V., Nemeslaki, A. (2012). Do organizational network studies constitute a cohesive communicative field? Mapping the citation context of organizational network research. *Journal of Sociology and Social Anthropology* 5(64), XV: 349–364.