Authors in the field of SNA: Derived networks and temporal analysis

Vladimir Batagelj
IMFM Ljubljana, IAM UP Koper and NRU HSE Moscow

Daria Maltseva
NRU HSE Moscow

XXXIX Sunbelt Social Networks Conference
Montreal, June 18, 2019

Outline

Data

Statistics

Coauthorship

References

1 Data
2 Statistics
3 Coauthorship
4 References

Vladimir Batagelj: vladimir.batagelj@fmf.uni-lj.si
Daria Maltseva: d_malceva@mail.ru

Current version of slides (June 18, 2019 at 04 : 19): slides PDF
https://github.com/bavla/SocNet/blob/master/Temp/SNAauthors.pdf

Data

Our dataset consists of articles from the WoS database (*Core Collection*), found with the query "social network*", and those published in the main SNA journals indexed in the WoS.

Using **WoS2Pajek 1.5**, we transformed our data into a collection of linked networks:

- one-mode citation network **Cite** on works (field CR)

and two-mode networks

- the authorship network **WA** on works × authors (field AU),
- the journal network **WJ** on works × journals (fields CR or J9),
- the keyword network **WK** on works × keywords (fields ID, DE or TI).

We obtained also a vector:

- **year** with work's publication year (field PY)

Data

Works can be of two types: with full descriptions (*hits*), and cited only (*terminal*, listed in CR field). For the terminal works only partial information is provided: the name of the *first* author, journal, publication year, journal issue and the number of the first page.

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. The obtained *basic* networks are labeled **CiteN**, **WAn**, **WJn**, and **WKn**.

As terminal works contain information on the first author only, it is not correct to use full networks for the analysis of connections between authors. For the further analysis, we constructed *reduced* networks 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 $|J| = 8,943$, keywords $|K| = 32,409$.

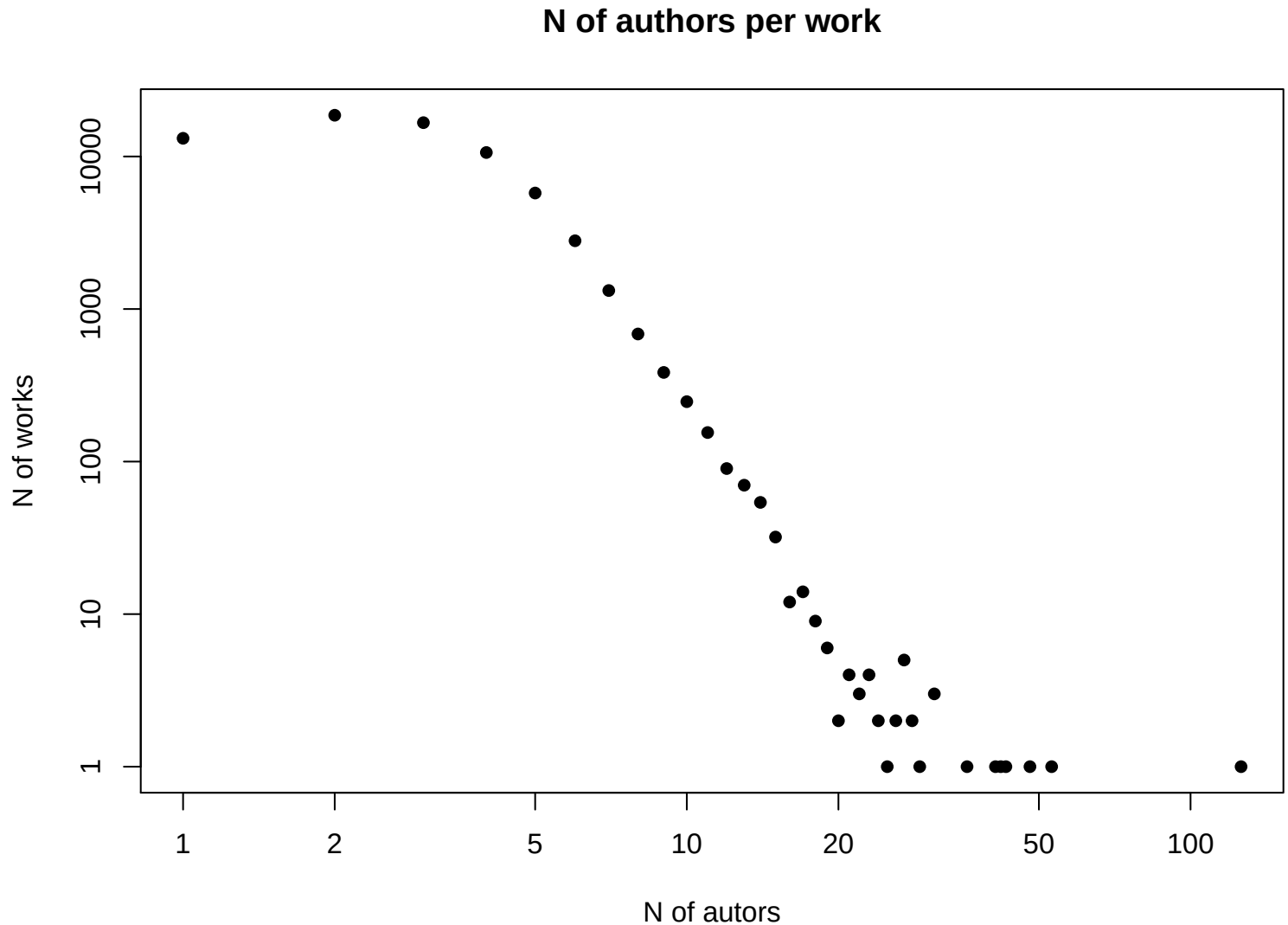
Number of authors per work

The distribution of the number of authors per work in **WAr** network (its outdegree) presented in figure on the following slide.

It shows that one fifth part (19%) of all works are written by a single author, while for a half of all works there are 2 (26%) or 3 (24%) authors.

For some works, however, the number of authors is very high. The extreme case is the work *Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking* published in *Nature Biotechnology* in 2016, which has 126 authors. Almost all works with a large number of co-authors belong to the field of natural sciences (medical, health, epidemiological, and behavioral studies), where the inclusion of all researchers involved in a project to the list of authors is a frequent practice. However, the third rated article *Discussion on the paper by Handcock, Raftery and Tantrum* published in *Royal Statistical Society Journa Series A: Statistics in Society* is written by 48 social networks scientists.

Number of authors per work



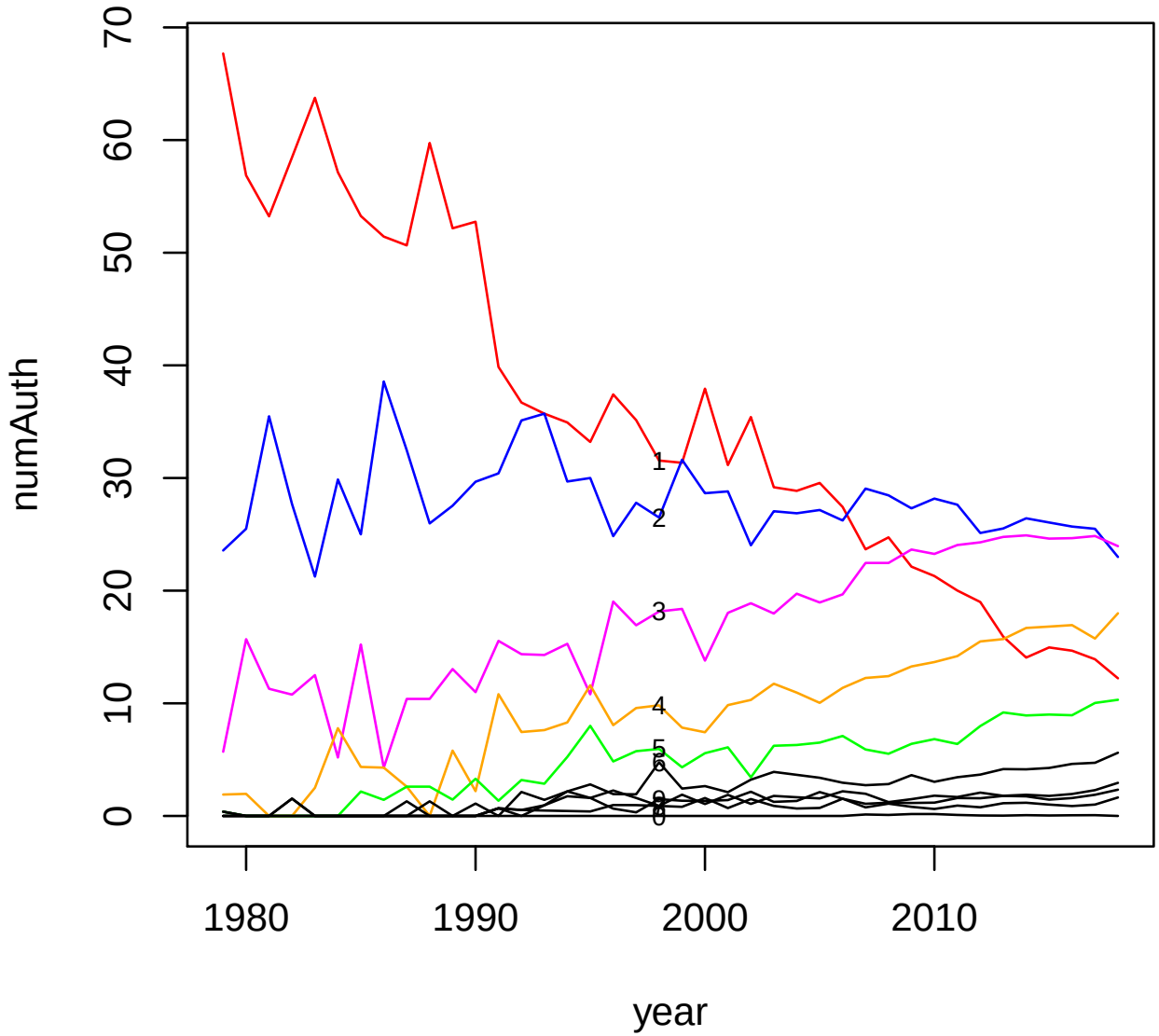
Number of authors: temporal distribution

Combining the number of authors with the publication year we get their temporal distribution, describing how the number of authors is changing through years. See next slide.

The results show that since 1980s, the number of single author papers droped from 70% to almost 10%. The number of papers authored by a pair of authors is relatively constant – around 25%. The numbers of papers authored by 3 and more authors are increasing (3: from 6% to 25%, 4: from 2% to 17%, 5: from 0% to 10%, etc.).

Besides the general trend to higher collaboration the reason could be also the expansion of SNA to other disciplines (physics, computer science, neuro science, biology, chemistry, etc.) with different writing cultures.

Number of authors: temporal distribution

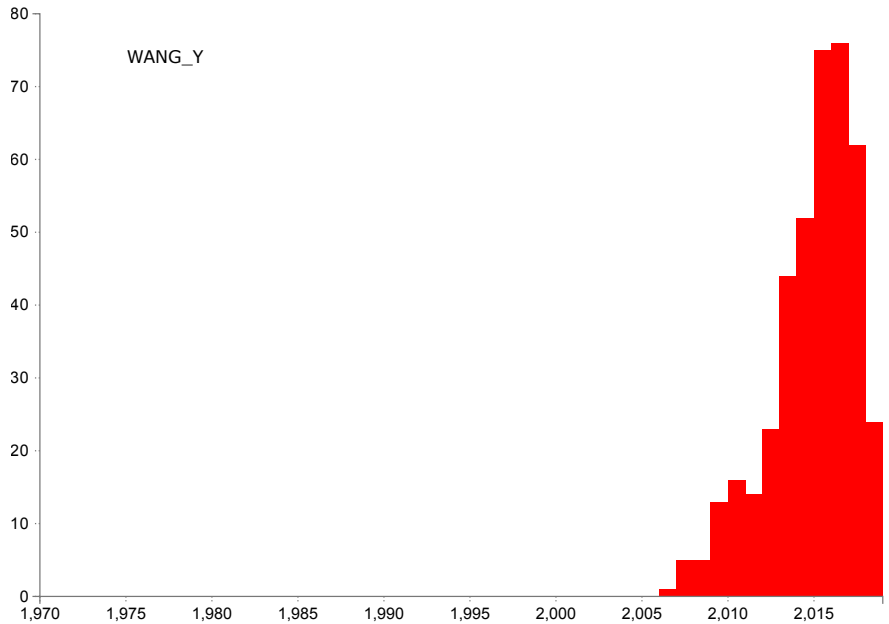


Works per author

The distribution of works per author in **WAr** network (its indegree) shows that almost all of the authors with the largest number of papers have Chinese or Korean surnames. The authors with the largest numbers of works are the following (number of articles in parentheses): WANG_Y (410), WANG_X (339), ZHANG_Y (332), LIU_Y (321), CHEN_Y (317), ZHANG_J (310), LI_J (305), LI_Y (304), LI_X (287).

The issue of the super-productivity of these groups of authors was discussed by Harzing (2015). The well-known "three Zhang, four Li" effect is that 85% of people in China have one of only around 100 surnames. Thus, there is a high chance that different authors, having the same surname and first letter of the name, merge during the analysis, creating "multiple personalities".

Works per author



As an example, the above figure shows the distribution of the number of works per year for a selected author – Y. WANG. It is hard to believe that someone can have 76 works published in one year.

The problem of "multiple personalities" could be overcame if we would use a special ID (such as ORCID) for each scientist; unfortunately, this information is not provided in the WoS (yet).

The most prolific authors (**WAr** indegree)

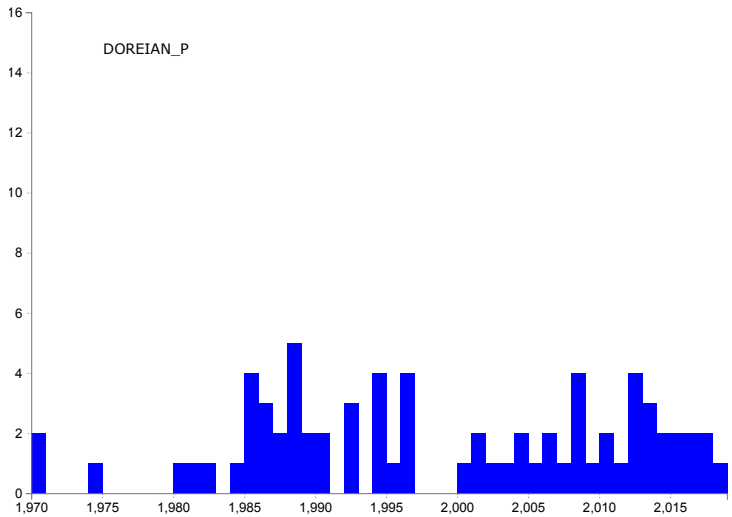
Chinese and Koreans excluded

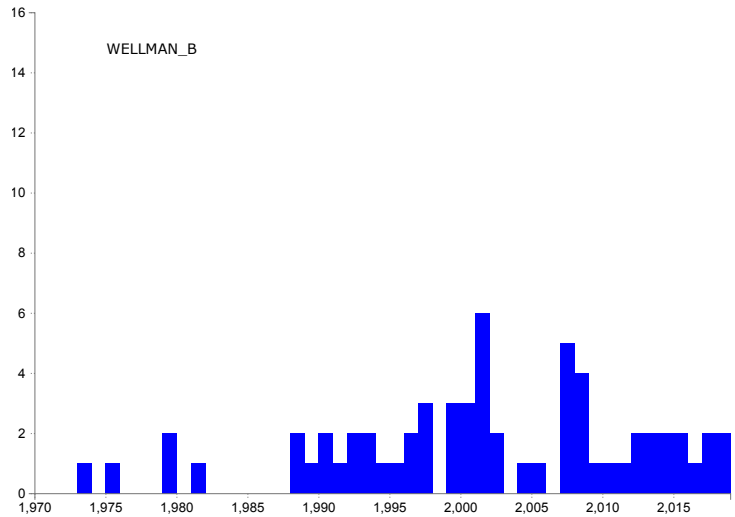
Rank	Orig rank	Author	Indegree	Rank	Orig rank	Author	Indegree
1	45	LATKIN_C	130	26	211	SCHNEIDE_J	52
2	72	VALENTE_T	97	27	212	LEYDESDO_L	51
3	84	DUNBAR_R	91	28	217	LITWIN_H	50
4	102	NEWMAN_M	81	29	228	RICE_E	48
5	121	CHRISTAK_N	74	30	232	KAWACHI_I	47
6	126	DOREIAN_P	72	31	233	BONACICH_P	46
7	127	CARLEY_K	72	32	234	PARK_Y	46
8	129	BURT_R	71	33	237	RODRIGUE_M	46
9	130	BORGATTI_S	71	34	238	NGUYEN_H	46
10	139	SNIJDERS_T	67	35	239	CROFT_D	46
11	140	BARABASI_A	67	36	249	EVERETT_M	44
12	146	FOWLER_J	65	37	252	FERNANDE_M	44
13	149	KAZIENKO_P	64	38	255	CONTI_M	44
14	150	ROBINS_G	64	39	256	MORRIS_M	43
15	152	WELLMAN_B	63	40	259	CONTRACT_N	43
16	163	FALOUTSO_C	60	41	266	WHITE_H	42
17	167	RAHMAN_M	59	42	267	SKVORETZ_J	42
18	172	PATTISON_P	58	43	275	PENTLAND_A	41
19	176	TUCKER_J	58	44	276	WILLIAMS_M	41
20	181	HOSSAIN_L	56	45	280	MOODY_J	40
21	187	JOHNSON_J	54	46	289	FRIEDMAN_S	40
22	194	NGUYEN_T	54	47	290	MARSDEN_P	39
23	196	MARTINEZ_M	53	48	292	BERKMAN_L	39
24	207	GONZALEZ_M	52	49	301	KRACKHAR_D	38
25	209	RODRIGUE_J	52	50	306	MORENO_M	38

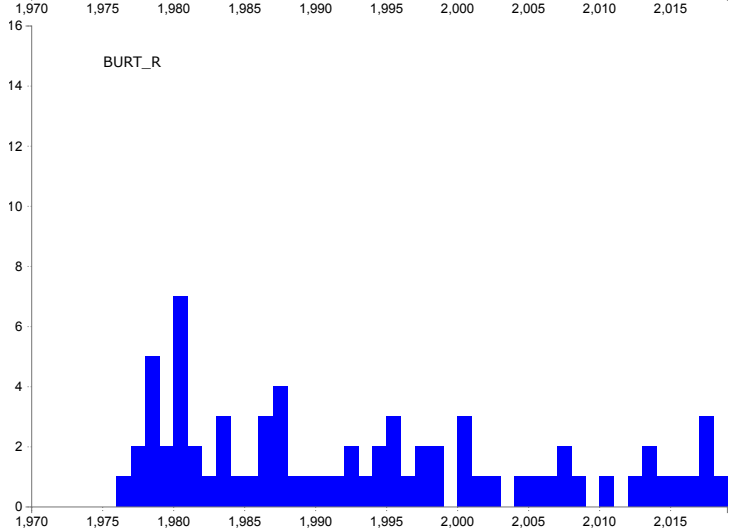
Temporal networks

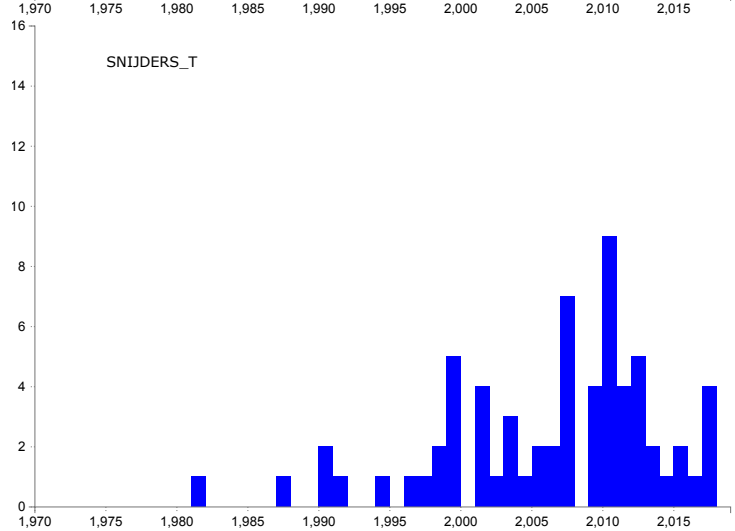
Batagelj, V., Praprotnik, S. (2016) An algebraic approach to temporal network analysis based on temporal quantities. Social Network Analysis and Mining, 6(1), 1-22.
Batagelj, V., Maltseva, D. (2019). Temporal Bibliographic Networks. https://arxiv.org/abs/1903.00600

Number of works per year for selected authors
WAins

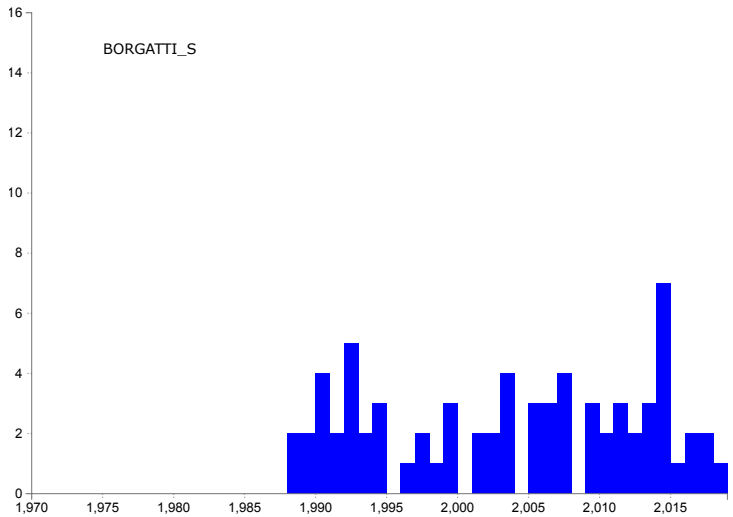


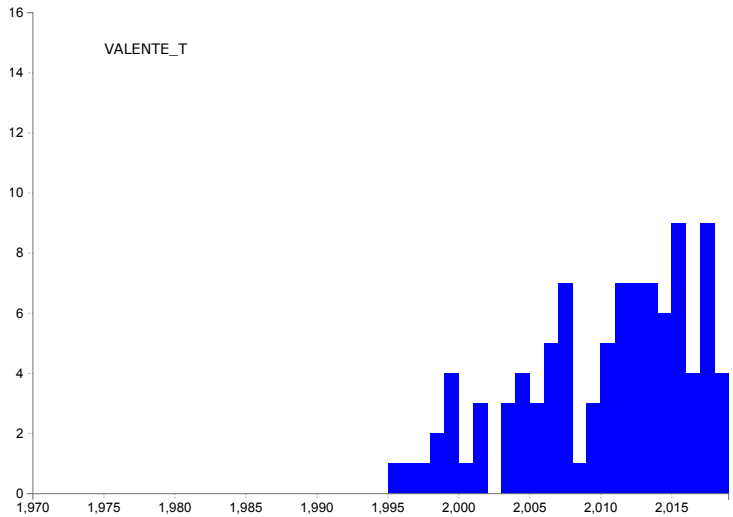


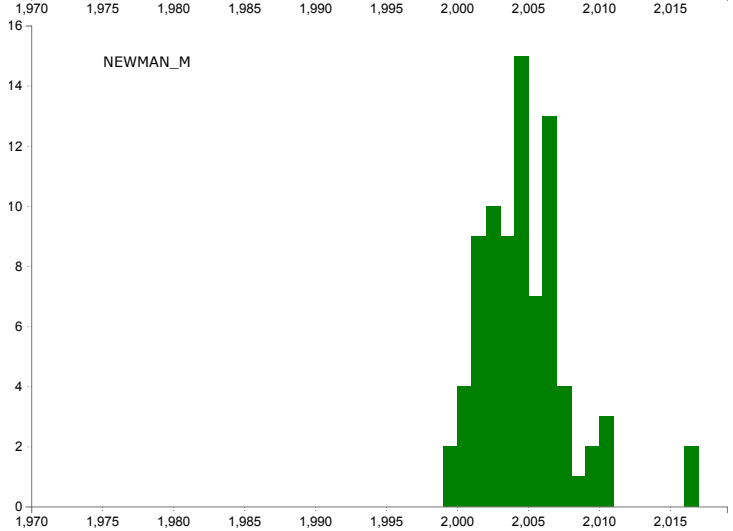


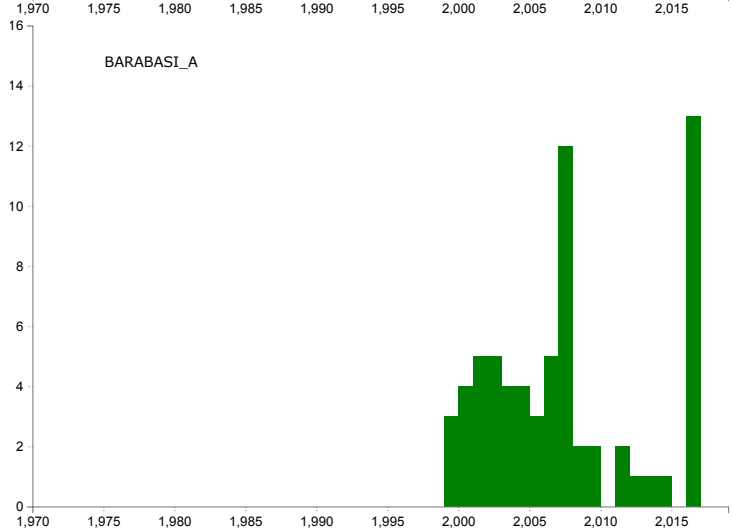


Number of works per year for selected authors
WAins

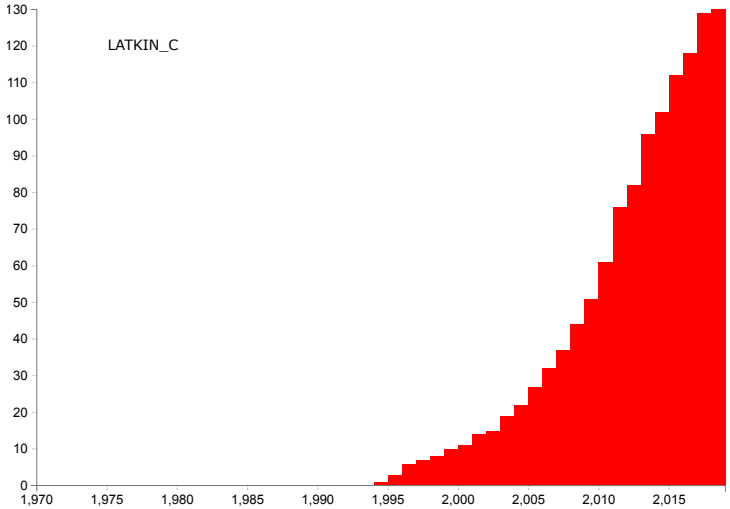


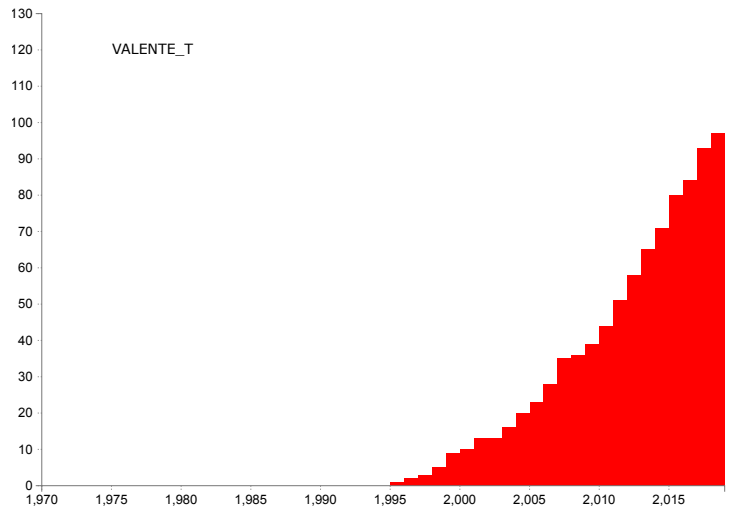


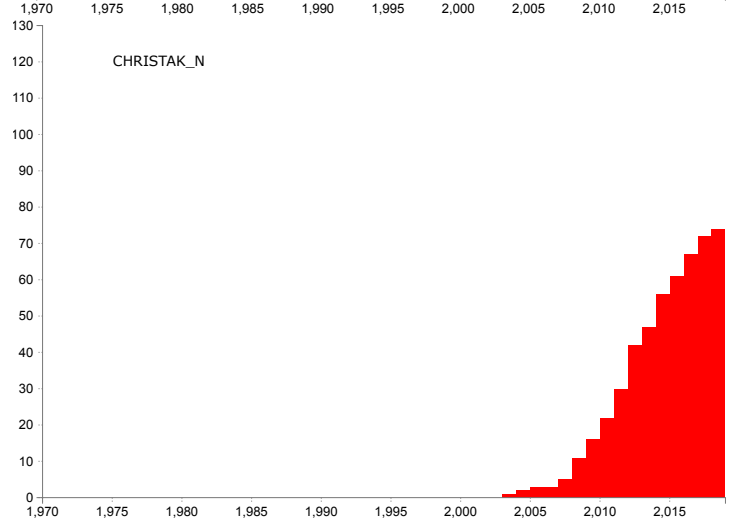


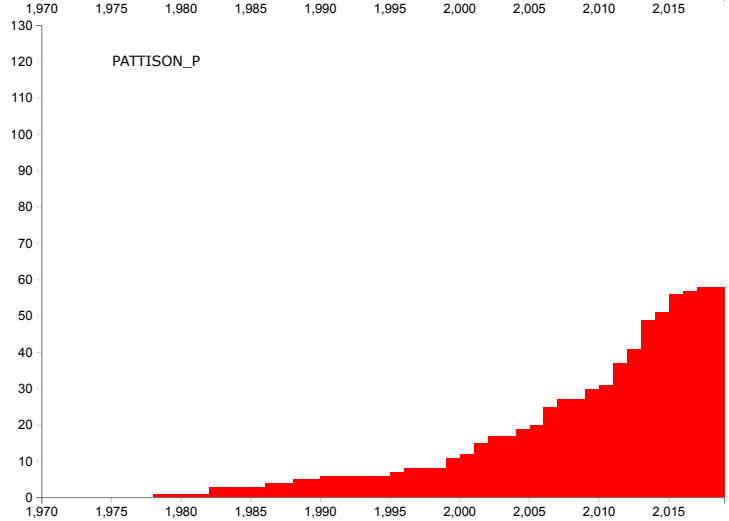


Number of works per year for selected authors
WAcum









Collaboration between authors – coauthorship

traditional

There are different ways to create one-mode networks of collaboration between authors **AA** from two-mode network **WA** (Batagelj and Cerinšek, 2013). Using **WAr** network, which consists of 70,792 works and 93,011 authors, and different types of normalization, we created three coauthorship networks – **Co**, **Cn**, and **Ct**$'$.

The traditional way to obtain the coauthorship network **Co** is

$$\mathbf{Co} = \mathbf{WAr}^T * \mathbf{WAr}$$

The weight $\mathbf{Co}[i,j]$ of the edge between the nodes i and j is equal to the total number of works author i and j wrote together. The loops in **Co** are equal to the total number of works that each author has (which is also equal to the indegree values of the **WAr** network).

Works with many coauthors are overrepresented in **Co**.

Collaboration between authors – coauthorship

fractional

Coauthorship network **Cn** is based on the *fractional approach*

$$\mathbf{n}(\mathbf{WAr})[w,a] = \frac{\mathbf{WAr}[w,a]}{\max(1, \mathrm{oudeg}[w])}$$

then $\mathbf{Cn} = \mathbf{WAr}^T * \mathbf{n}(\mathbf{WAr})$.

The weight $\mathbf{Cn}[i,j]$ of the edge between the nodes (authors) i and j is equal to the contribution of author i to works that he or she wrote together with author j (which can be asymmetric). The *author's total contribution* to all his/her works is the corresponding diagonal value (loop) in **Cn**.

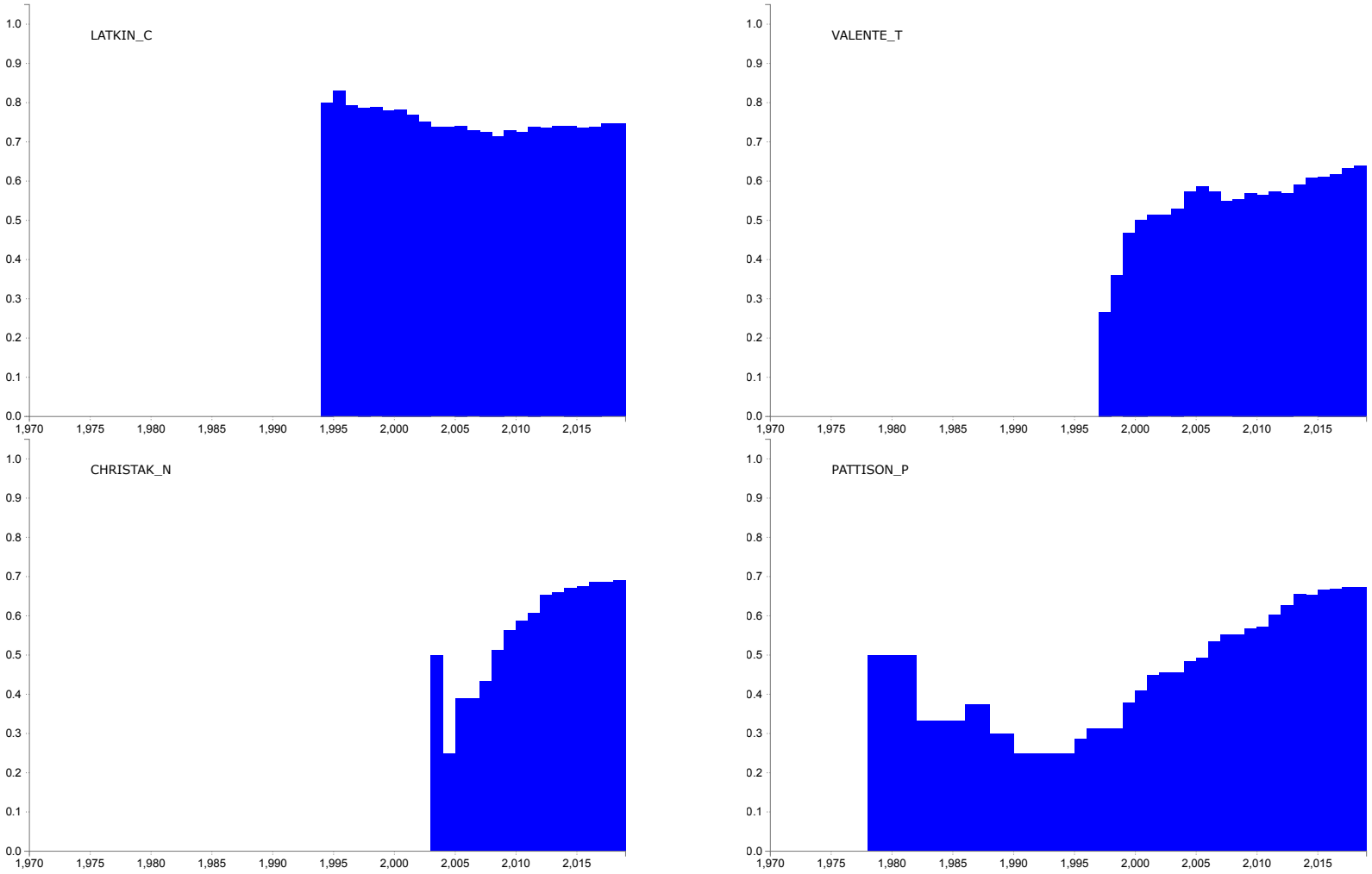
Based on this, the *self-sufficiency index Si* and the *collaborativeness index Ki* were proposed:

$$S_i = \frac{cn_{ii}}{\mathrm{indeg}_{\mathbf{WAr}}(i)} \qquad \text{and} \qquad K_i = 1 - S_i$$

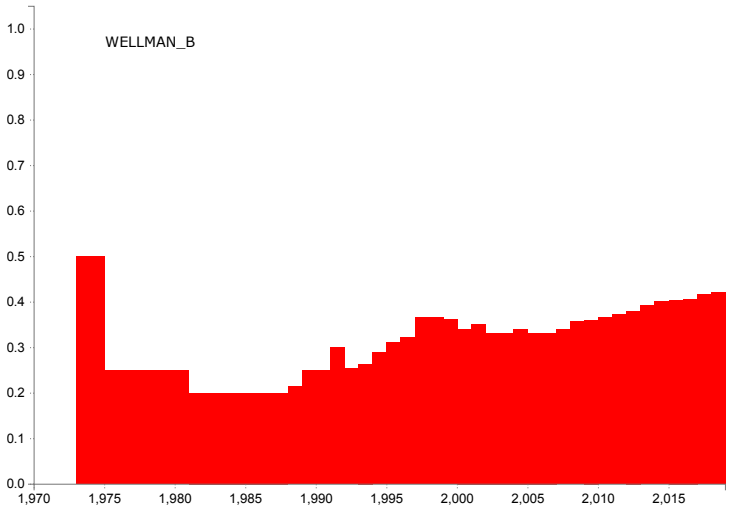
Collaborativeness

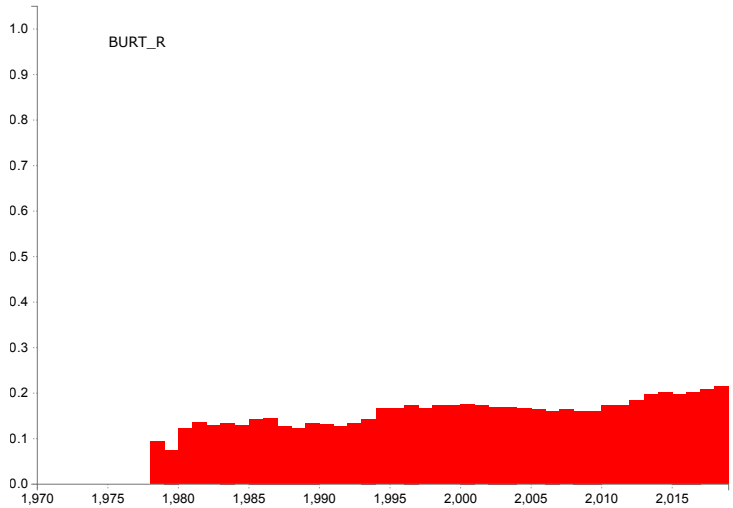
i	Author	Total	cn_{ij}	K	i	Author	Total	cn_{ij}	K
1	LATKIN_C	130	32,99	**0,75**	31	EVERETT_M	44	22,58	0,49
2	VALENTE_T	97	34,96	**0,64**	32	MORRIS_M	43	17,22	**0,60**
3	DUNBAR_R	91	40,02	**0,56**	33	CONTRACT_N	43	14,15	**0,67**
4	NEWMAN_M	81	**50,02**	0,38	34	WHITE_H	42	27,28	0,35
5	CHRISTAK_N	74	22,89	**0,69**	35	SKVORETZ_J	42	20,07	**0,52**
6	DOREIAN_P	72	**46,19**	0,36	36	PENTLAND_A	41	14,12	**0,66**
7	CARLEY_K	72	28,11	**0,61**	37	MOODY_J	40	17,7	**0,56**
8	BURT_R	71	**55,73**	0,22	38	SMITH_A	40	14,2	**0,65**
9	BORGATTI_S	71	29,72	**0,58**	39	MARSDEN_P	39	30,17	0,23
10	SNIJDERS_T	67	29,63	**0,56**	40	BERKMAN_L	39	14,3	**0,63**
11	BARABASI_A	67	27,61	**0,59**	41	SMITH_M	39	13,29	**0,66**
12	FOWLER_J	65	20,14	**0,69**	42	KRACKHAR_D	38	18,24	**0,52**
13	KAZIENKO_P	64	21,97	**0,66**	43	JACKSON_M	38	17,78	**0,53**
14	ROBINS_G	64	19,67	**0,69**	44	THELWALL_M	37	18,41	**0,50**
15	WELLMAN_B	63	36,43	0,42	45	FRIEDKIN_N	36	28,17	0,22
16	FALOUTSO_C	60	17,86	**0,70**	46	SINGH_A	36	14,5	**0,60**
17	RAHMAN_M	59	19,18	**0,67**	47	WASSERMA_S	35	15,64	**0,55**
18	PATTISON_P	58	18,94	**0,67**	48	BRANDES_U	35	14,39	**0,59**
19	JOHNSON_J	54	21,19	**0,61**	49	GONZALEZ_A	35	14,13	**0,60**
20	MARTINEZ_M	53	21,9	**0,59**	50	KLEINBER_J	34	15,05	**0,56**
21	GONZALEZ_M	52	17,76	**0,66**	51	FARINE_D	34	14,04	**0,59**
22	RODRIGUE_J	52	15,9	**0,69**	52	BATAGELJ_V	33	14,64	**0,56**
23	SCHNEIDE_J	52	13,89	**0,73**	53	BREIGER_R	31	19,73	0,36
24	LEYDESDO_L	51	33,28	0,35	54	WILLIAMS_A	31	14,5	**0,53**
25	LITWIN_H	50	32,42	0,35	55	SCOTT_J	28	17,54	0,37
26	RICE_E	48	13,09	**0,73**	56	MASUDA_N	28	14,26	0,49
27	BONACICH_P	46	34	0,26	57	FREEMAN_L	27	20,03	0,26
28	RODRIGUE_M	46	13,21	**0,71**	58	WATTS_D	27	13,67	0,49
29	BONACICH_P	46	34	0,26	59	LAZEGA_E	26	14,17	0,46
30	CROFT_D	46	11,6	**0,75**	60	FAUST_K	25	13,5	0,46

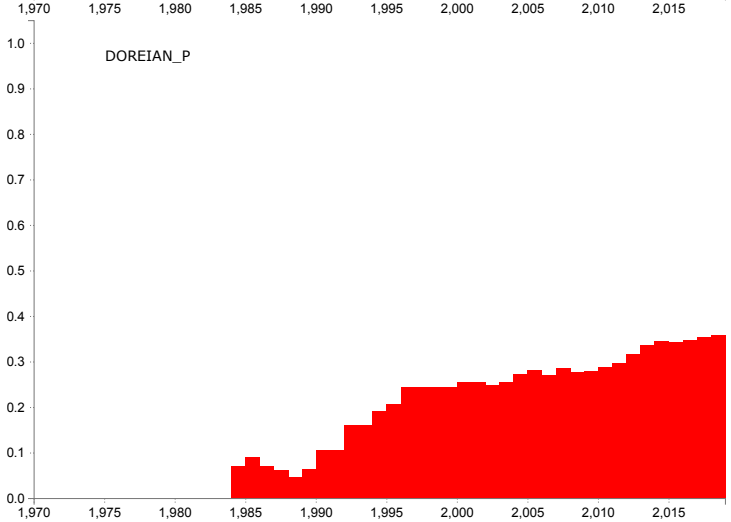
Temporal collaborativness

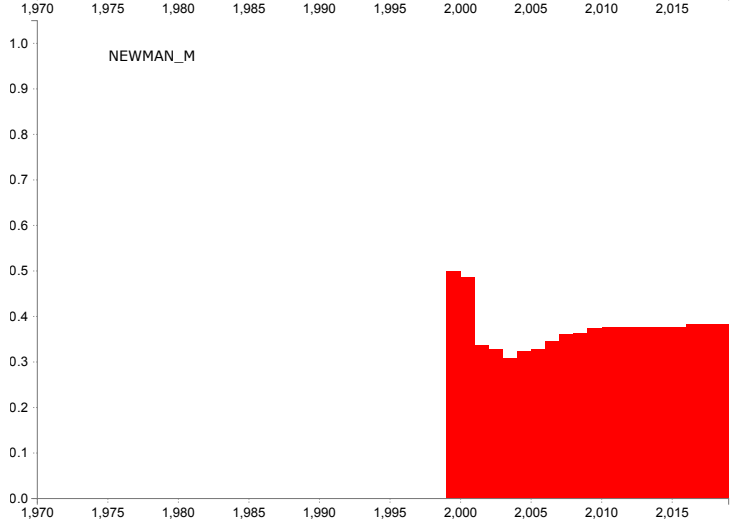


Temporal collaborativness









Collaboration between authors – coauthorship

Newman's normalization

Newman interpreted collaboration in a "strict" way – as a collaboration only with others (not with the author himself or herself, excluding single authored papers). This gives another normalized coauthorship network $\mathbf{Ct}'$

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

and $\mathbf{Ct}' = \mathbf{n}(\mathbf{WA})^T * \mathbf{n}'(\mathbf{WA})$.
The obtained $\mathbf{Ct}'$ is undirected and does not have loops.

The contribution of a complete subgraph corresponding to each work is 1. The weight $\mathbf{Ct}'[i,j]$ of the edge between the nodes (authors) i and j is equal to the total contribution of the "strict collaboration" of authors i and j to works they wrote together. The total contribution of an author is equal to the row sum of weights.

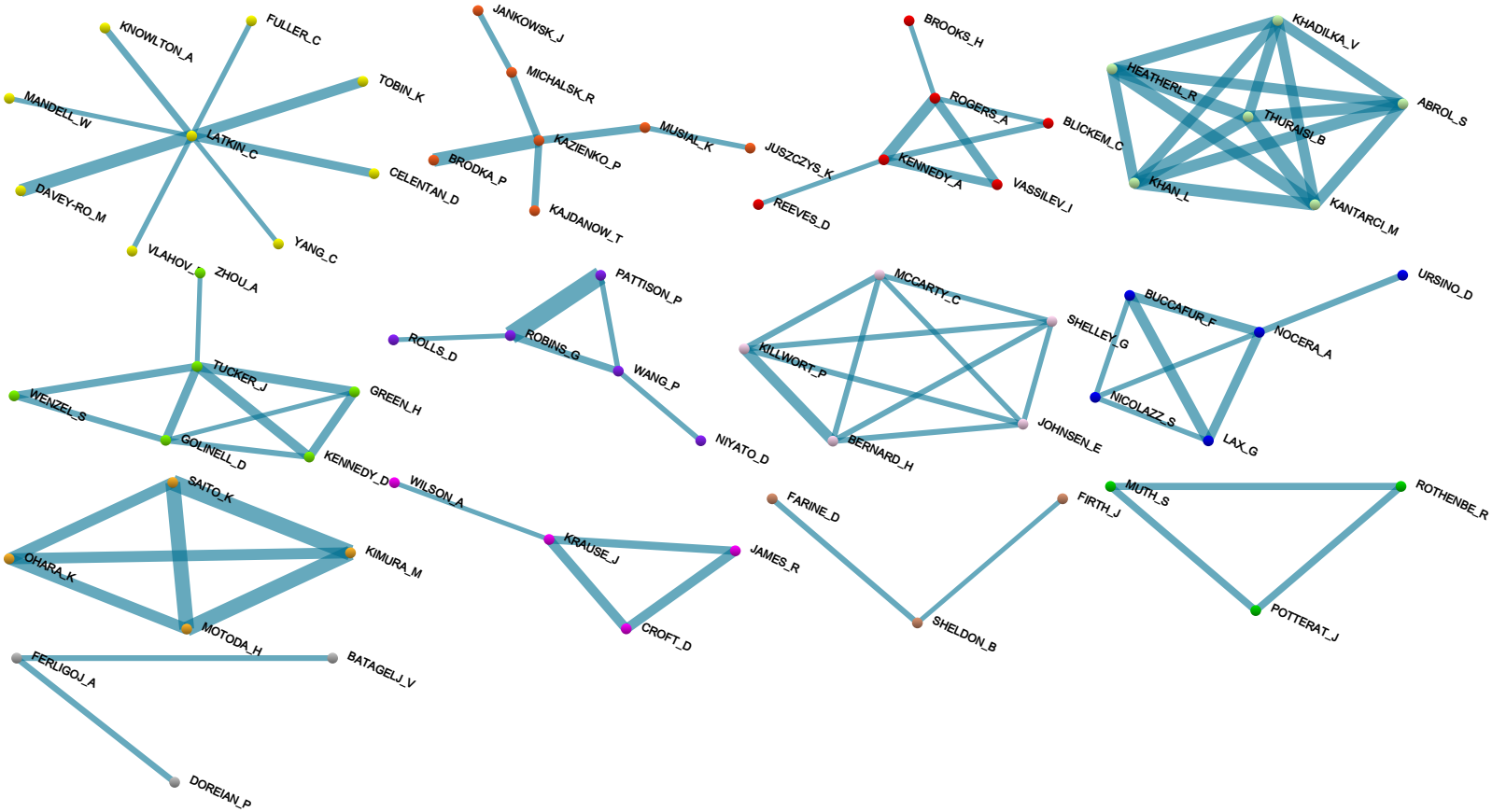
Groups of the most collaborating authors

We made in **Co** a link cut at the level of 10 (at least 10 works written together) and got a subnetwork of 420 nodes, which includes a lergest component of 58 nodes, a component of 9 nodes, 2 components of 7 nodes, 4 components of 6 nodes, 4 components of 5 nodes, 9 components of 4 nodes, and 23 components of 3 nodes and 95 components of 2 nodes.

Almost half of the nodes (45%) belong to the latter. Thus, there are not so many authors who have 10 works written in collaboration, and even for them it is more common to work in pairs.

The main component is formed of Chinese and Koreans.

Selected components in **Co**



Groups of collaborating authors in **Ct′**

Islands

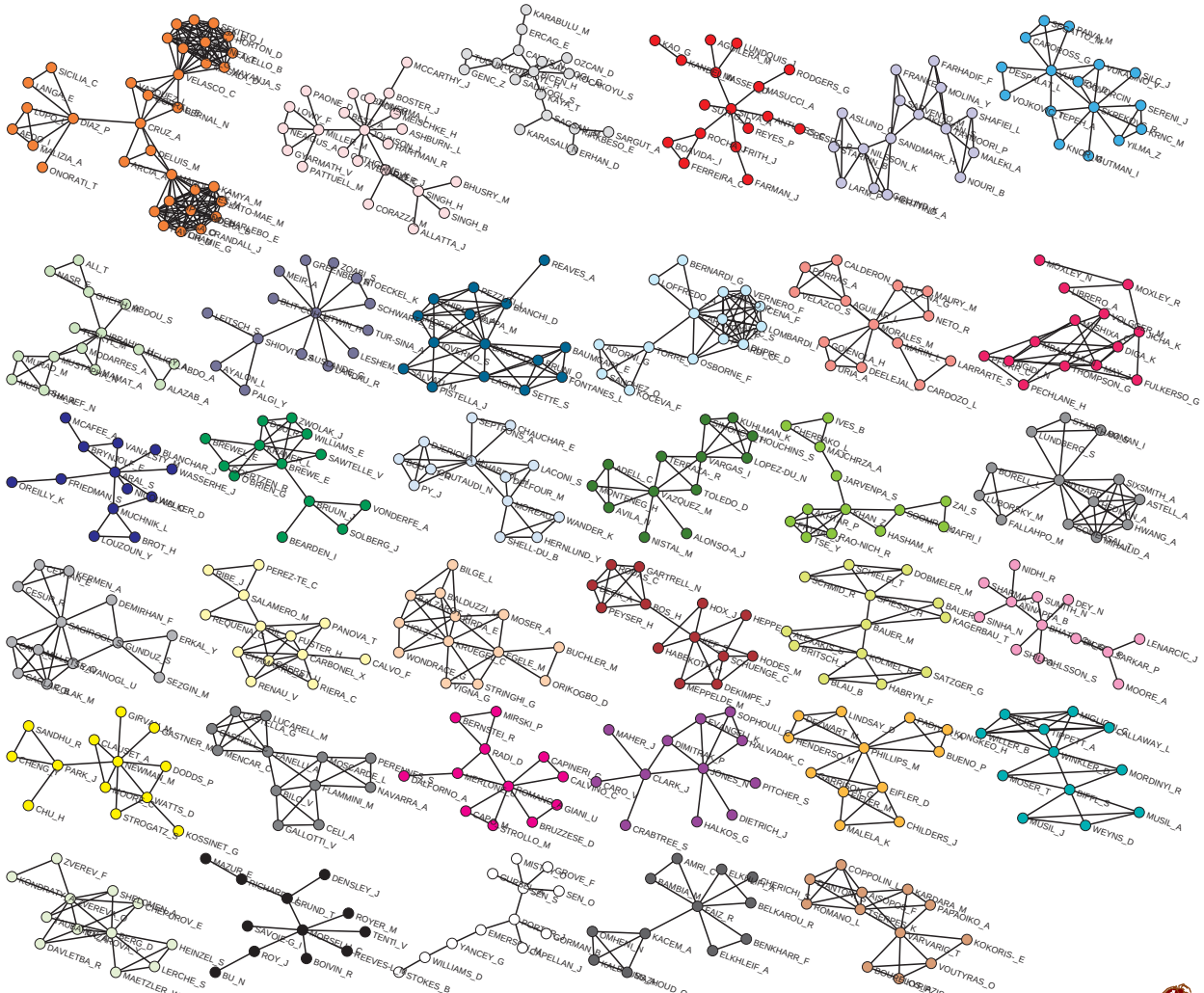
To extract the groups of authors collaborating with each other from the **Ct′** network, we used the *islands approach* (simple and general islands).

We obtained 14,222 *simple islands* of size between 2 and 50 nodes (which are 45,524 nodes, or 45% of all nodes in the network).

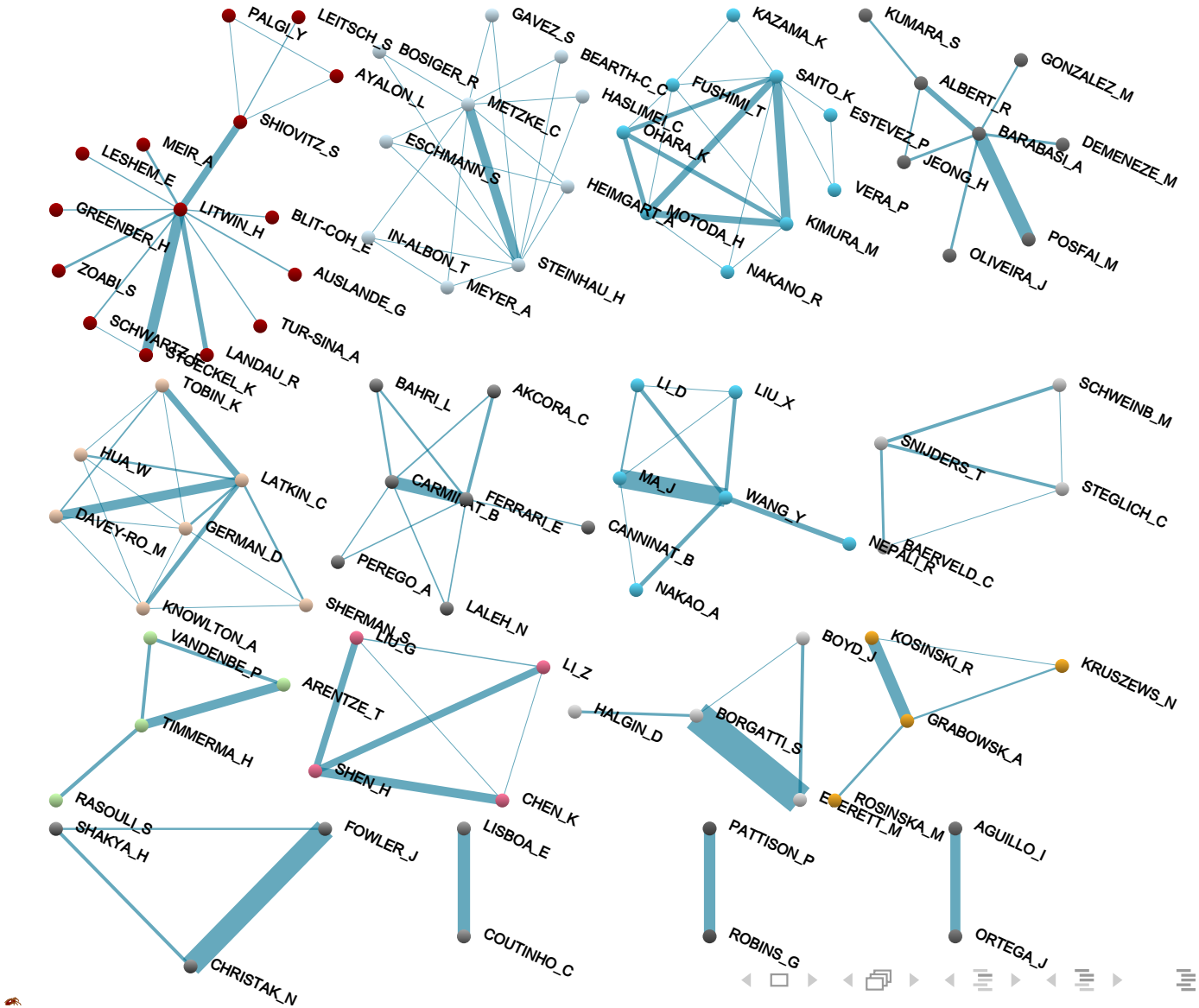
The four largest islands consists of, respectively, 35, 23, 21, and 19 nodes; another 69 islands have between 12 and 18 nodes. The largest part of the network (78%) consists of clusters of relatively small sizes: 2 (28%), 3 (24%), 4 (15%), and 5 (10%) nodes.

The variation of the upper and lower thresholds changes the situation: with the threshold [20, 100] we get just 3 islands composed of 79 nodes (0.1% of all nodes), which sizes are 35, 23 and 21 nodes. We decided to use the threshold [2, 50] for further analysis.

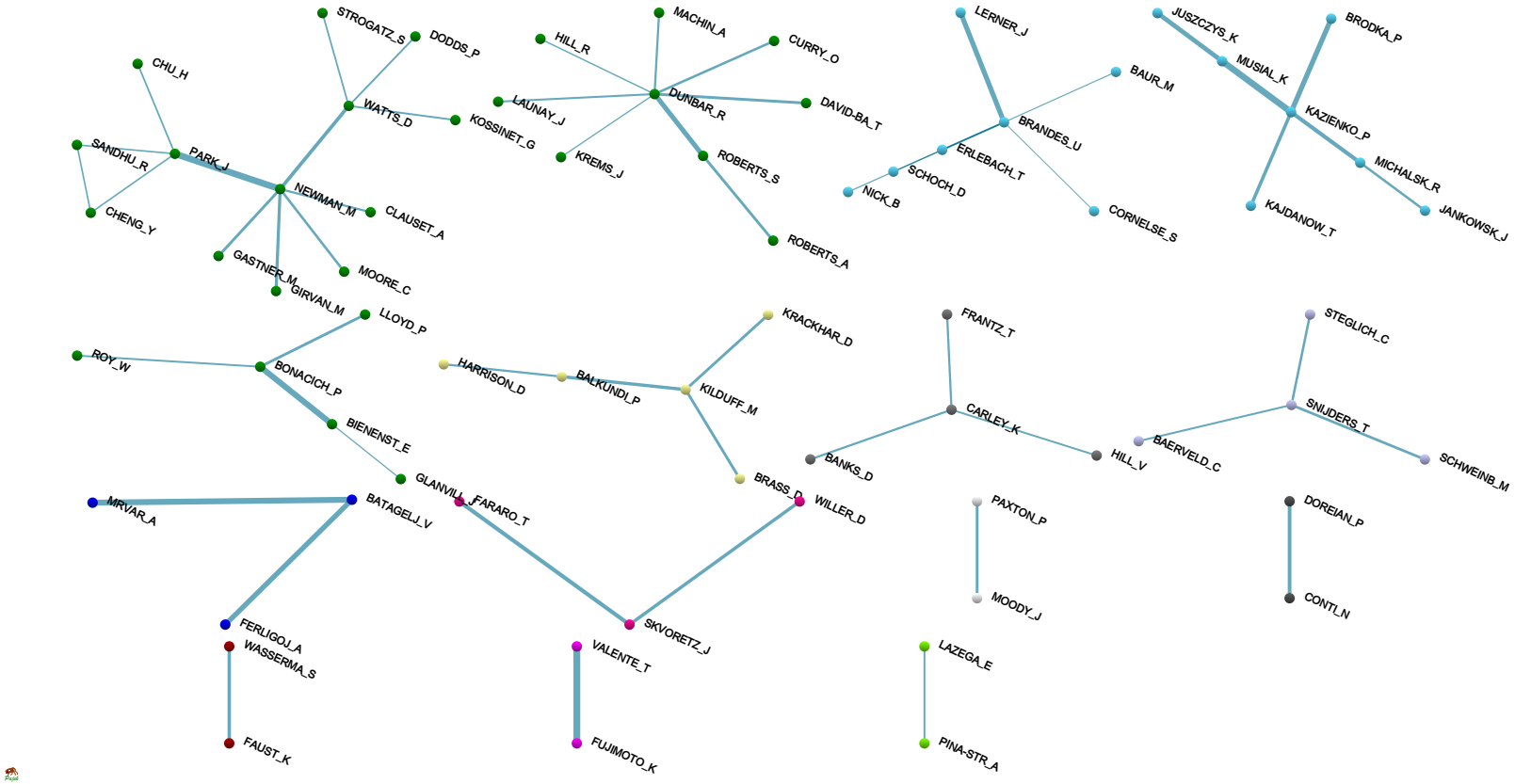
Selected simple islands in $\mathbf{Ct}'$



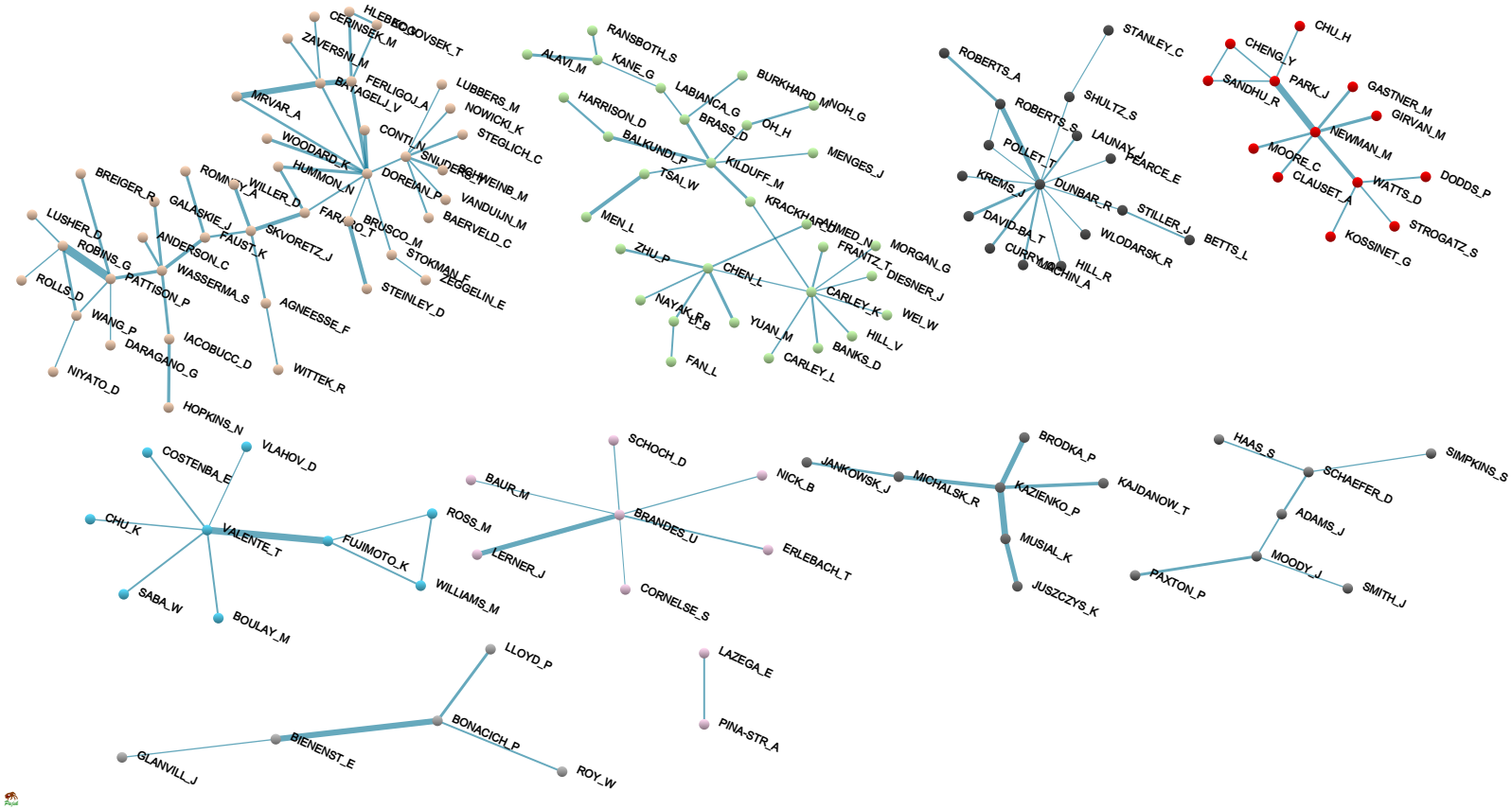
Simple islands for authors with the largest weights in $\mathbf{Ct'}$



Simple islands for selected authors in $\mathbf{Ct}'$



General islands for selected authors in $\mathbf{Ct}'$



Conclusions

Other questions about authors can be addressed:

- temporal collaboration
- combining authorship network with citation network

$$\mathbf{CiteAn} = (\mathbf{WAr})^T * \mathbf{n}(\mathbf{CiteR}) * \mathbf{WAr}$$

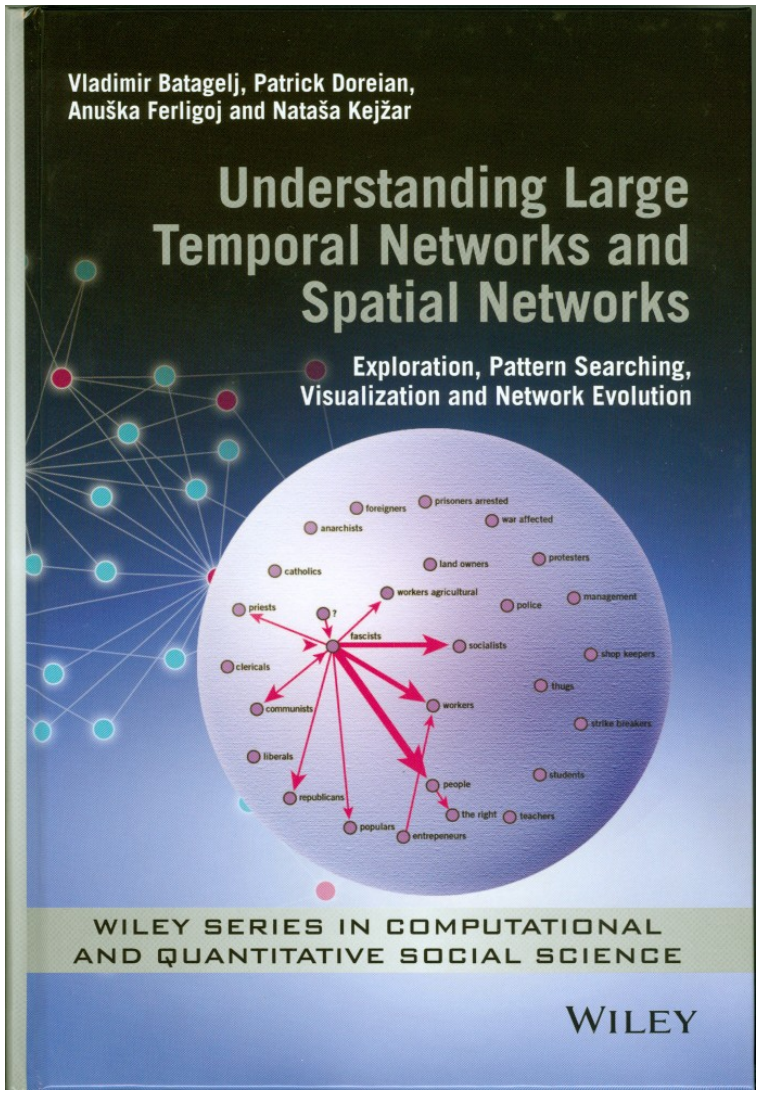
$$\mathbf{ACAn} = \mathbf{WAins}^{\mathbf{T}} * \mathbf{n}(\mathbf{CiteIns}) * \mathbf{WAcum}$$

  self-citation, being cited, citation islands
- bibliographic coupling among authors

$$\mathbf{biCo} = \mathbf{CiteR} * (\mathbf{CiteR})^T$$

$$\mathbf{ACoj} = \mathbf{n}(\mathbf{WAr})^T * \mathbf{biCo} * \mathbf{n}(\mathbf{WAr})$$

Understanding large networks



The detailed description of methods is available in chapters 2 and 3 in the book:

Vladimir Batagelj, Patrick Doreian, Anuška Ferligoj and Nataša Kejžar: Understanding Large Temporal Networks and Spatial Networks: Exploration, Pattern Searching, Visualization and Network Evolution. Wiley Series in Computational and Quantitative Social Science. Wiley, October 2014.

References I

Barabasi, A.L., Jeong, H., Neda, Z., Ravasz, E., Schubert, A., Vicsek, T.: Evolution of the social network of scientic collaborations. Physica **311** (2002) 590–614

Batagelj, V.: Wos2pajek – networks from web of science (2007). http://vladowiki.fmf.uni-lj.si/doku.php?id=pajek:wos2pajek

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

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

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

De Nooy, W., Mrvar, A., Batagelj, V.: Exploratory Social Network Analysis with Pajek; Third Edition. Structural Analysis in the Social Sciences, Cambridge University Press, September 2018.

References II

Newman, M.E.: The structure of scientific collaboration communities. Proceedings of the National Academy of Science (PNAS) **98** (2001) 404–409

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

Zaveršnik, M., Batagelj, V.: Islands. In: XXIV International Sunbelt Social Network Conference, Portorož, Slovenia (2004)

Pajek's wiki. http://pajek.imfm.si

Vladimir Batagelj, Andrej Mrvar: Pajek manual.