Network Analysis of Works on Clustering and Classification from Web of Science

Nataša Kejžar[1], Simona Korenjak Černe[2], and Vladimir Batagelj[3]

[1] University of Ljubljana, Faculty of Social Sciences, Kardeljeva pl. 5, 1000 Ljubljana, Slovenia natasa.kejzar@fdv.uni-lj.si
[2] University of Ljubljana, Faculty of Economics, Kardeljeva pl. 17, 1000 Ljubljana, Slovenia simona.cerne@ef.uni-lj.si
[3] University of Ljubljana, Faculty of Mathematics and Physics, Jadranska 19, 1000 Ljubljana, Slovenia vladimir.batagelj@fmf.uni-lj.si

Summary. Web of Science (WoS) is a database that provides information about current and past articles published in over 10,000 of the most prestigious, high impact research journals in the world from year 1970 on. A file with full information – records about selected articles – can be downloaded and further analyzed. We collected from WoS complete records on articles from Journal of Classification, articles citing these articles, and articles in WoS cited by them at least 10 times. A special program WoS2Pajek was developed for converting such data into Pajek network files. The citation network between articles, networks of articles × authors, articles × keywords, articles × journals, and the partition according to publication year were obtained from the data. These networks were analyzed in order to identify the most important authors, works and topics that have been involved in the field in the last decades.

Key words: Network Analysis, Classification Society, Journal of Classification, bibliographic data, Web of Science.

1 Introduction

Web of Science (WoS) [19] is an online academic service provided by Thomson Reuters. It provides access to seven world's leading citation databases: Science Citation Index, Social Sciences Citation Index, Arts & Humanities Citation Index, Index Chemicus, Current Chemical Reactions, Conference Proceedings Citation Index: Science, and Conference Proceedings Citation Index: Social Science and Humanities. It covers data on over 10,000 of the highest impact journals of science, technology, social sciences, arts, and humanities, and over 110,000 books and conference proceedings.

WoS allows one to get full information, a record, about an article, a book or other work: its title, authors, abstract, keywords, publication properties

(keywords, journal, volume, pages, publication year, etc.) and its references. From such bibliographic data many analyses can be done.

Network analysis has been used by Newman (2001) [16] who observed some scientific collaboration networks or, as he called them, acquaintance networks. He analyzed the networks from four different databases of published papers in the 5-year period of 1995–99 inclusive (MEDLINE, Los Alamos ArXiv, SPIRES and NCSTRL). He discovered some significant statistical differences between prespecified different scientific communities. Collaboration networks were analyzed also for some other data such as Erdős network [5], boards of directors [17], movies database (IMDB) [1], etc. A wide research of the dynamics of collaboration networks for the fields of mathematics and neuro-science was done by Barabási et al. in 2002 [2]. They looked at the network from 1991–98 and investigated the intensity of collaboration, the average separation (in terms of shortest paths) of authors, the clustering coefficients between the fields, as well as through time. They proposed models for the evolution of collaboration networks. The ones predicting connectivity distribution are based on continuum theory, however those that deal with other quantities are studied by Monte Carlo simulations.

Many different network analyses of bibliographic data sets were published in the field of scientometrics. The primary interest of the field is to study science using the scientific methods of science. Citations between scientific works can be studied directly or on agglomerated level as citations between journals or authors. Very informative visualizations of the structure for the whole science (natural and social sciences) from WoS data was constructed by Börner [9] and Boyack et al. [8, 10]. Visualizations of scientific networks through time and development of various methodological concepts were done by Leydesdorff (see e.g. [14, 15]).

Research was done mainly on 1-mode networks (of collaboration between authors or citations between works or journals). However, as Dorogovtsev and Mendes [11] pointed out, networks obtained from bibliographic databases are inherently bipartite (2-mode).

In this paper we look at records from WoS database for the field of clustering and classification. We further limit to records from and related to the Journal of Classification (as one of the most important journals in classification) in order to reveal the relevant (groups of) works, authors and topics.

2 Networks from WoS

Initially we intended to analyze the entire field of clustering and classification.

Searches from WoS were done for all years (from 1970–2008) and topics (a) `"cluster analy*"` (67,962 records), (b) `"clustering*"` (49,216), and (c) `"classificat*"` (220,190). Additional search was done for all years and publication name (d) `"Journal of Classification"` and extended with related

works. The results were converted into networks in Pajek [6] format using program WoS2Pajek [4].

The usual ISI name of a work (field CR) has the following structure

```
GRANOVET.MS, 1973, AM J SOCIOL, V78, P1360
GRANOVETTER M, 1983, SOCIOLOGICAL THEORY, V1, P203
```

which allows for many inconsistencies. Program WoS2Pajek supports also shorter names (similar to the names used in HistCite [12] output) in the format:

LastNm[:8] + '_' + FirstNm[0] + '(' + PY + ')' + VL + ':' + BP

that eliminate most of the inconsistencies. For example: `GRANOVET_M(1973)78:1360`.

WoS2Pajek produces the following networks:

- citation network **Ci** (stored in file `Cite.net`) of works only
- 2-mode network works × authors **WA** (`WA.net`)
- 2-mode network works × journals **WJ** (`WJ.net`)
- 2-mode network works × keywords **WK** (`WK.net`)
  As keywords are considered regular keywords and also all words from title and abstract without stopwords.

Preliminary network analyses of networks from `"cluster analy*"` showed that the hard core clustering community – members of IFCS (with the exception of some really fundamental works like Ward's *Hierarchical grouping to optimize an objective function* from 1963, or Sneath & Sokal's *Numerical Taxonomy*, 1973) don't play a prominent role in the broad field. Most of the important authors/works, however, belong to the field of biology. This could be due to different publishing cultures in the involved scientific communities (number of coauthors, number of references) or due to the use of the terms cluster, clustering and classification for different meanings.

This was the reason to limit our further analyses in this paper to *JoC data set* which consists of the WoS records on: (a) articles from Journal of Classification (JoC), (b) articles citing these articles, and (c) articles cited at least 10 times from (a or b) articles and having descriptions in WoS.

3 Analyses of records from JoC

There are 81,581 different works in the JoC data set of which 4,188 have full description records – 599 from JoC. The works come from 9,448 different journals and there are 37,690 authors in the data. Note that for references only the first author is known.

In the original data there was 1 loop (selfreference) in the citation network **Ci**. The inspection of the original paper showed that the error was in the WoS data. We removed the loop from the network and also transformed multiple arcs into single arcs.

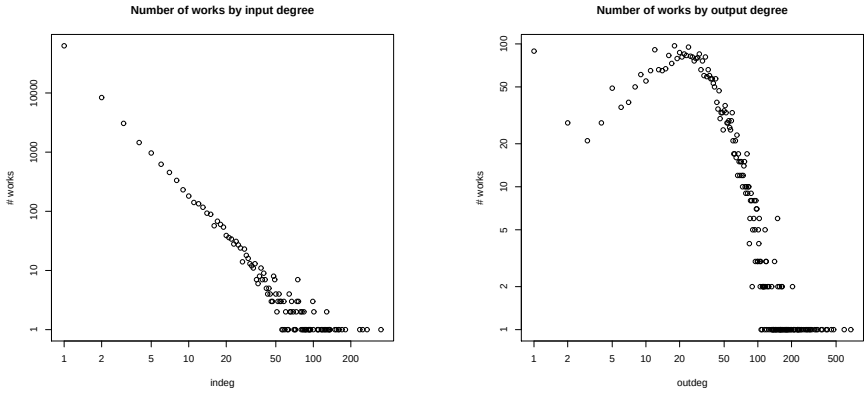

Fig. 1. Number of works by degree.

Figure 1 shows the input and output degree distributions for the citation network in log-log scale. The number of works with input degree (number of citations received) decreases very rapidly (power-law). The 15 most cited works with the number of received citations at their beginning are: 349 – HUBERT_L(1985)2:193, 270 – HARTIGAN_J(1975):, 249 – DEMPSTER_A(1977)39:1, 236 – SNEATH_P(1973):, 181 – SCHWARZ_G(1978)6:461, 170 – GOWER_J(1986)3:5, 161 – WARD_J(1963)58:236, 159 – RAND_W(1971)66:846, 153 – JOHNSON_S(1967)32:241, 149 – KAUFMAN_L(1990):, 136 – SAITOU_N(1987)4:406, 134 – JAIN_A(1988):, 134 – MCLACHLA_G(1988):, 132 – KRUSKAL_J(1964)29:1, 129 – ROHLF_F(1999)16:197.

The output degree (number of citations made), however, has a more peculiar shape. It starts high, steps down for 2 and 3, then increases till around 40 and then rapidly decreases. The largest outdegree have works that are either books or overview articles. Note that only works with a full description are considered since only referenced works (without full description) have output degree 0.

Boundary

For further analyses we limit the size of the network (boundary problem) to the works with full descriptions and referenced only works that are referenced often enough – at least k times. We delete vertices for which it holds $(0 < \text{indeg}(v) < k) \wedge (\text{outdeg}(v) = 0)$. In our case we selected $k = 3$.

Frequencies of publications in journals

Let us look at the largest indegrees in the **WJ** network. The journal names in WoS are not unified (normalized) – the same journal can appear under different names. For example: J Roy Stat Soc B, J R Stat Soc B, J Royal Stat Soc B, J Roy Stat Soc B 4, J Roy Stat Soc B Met, J Roy Stat Soc Ser B-Stat Met, J Roy Statist Soc Ser B Metho; P National Academy S, Proc Nat Acad Sci USA, P Natl Acad Sci USA; Multivar Behav

Res, Multivariate Behav R, Multivariate Behav Res, Multivariate Behavio; J Am Stat Assoc, J Amer Statist Assn,. . .

The list of journals in the bounded network with at least 50 published articles (first number is the number of published articles from the journal) contains: 1009 – J Classif, 425 – Psychometrika, 248 – Syst Biol, 215 – Mol Biol Evol, 207 – Syst Zool, 197 – J Am Stat Assoc, 136 – Comput Stat Data Anal, 120 – Evolution, 117 – Lect Note Comput Sci, 108 – P Natl Acad Sci USA, 104 – Pattern Recogn, 101 – Biometrics, 99 – Bioinformatics, 96 – Multivar Behav Res, 96 – J Mol Evol, 89 – Brit J Math Stat Psy, 88 – IEEE T Pattern Anal, 85 – Cladistics, 82 – Biometrika, 76 – J Roy Stat Soc B, 72 – Science, 71 – J Math Psychol, 70 – Nature, 68 – Math Biosci, 60 – J Marketing Res, 58 – Mol Phylogenet Evol, 56 – Ann Stat, 54 – Genetics, 54 – Discrete Appl Math, 52 – J Theor Biol, 52 – Soc Networks, 51 – Ecology, 51 – Annu Rev Ecol Syst, 51 – Pattern Recogn Lett.

Distribution of articles by the number of authors

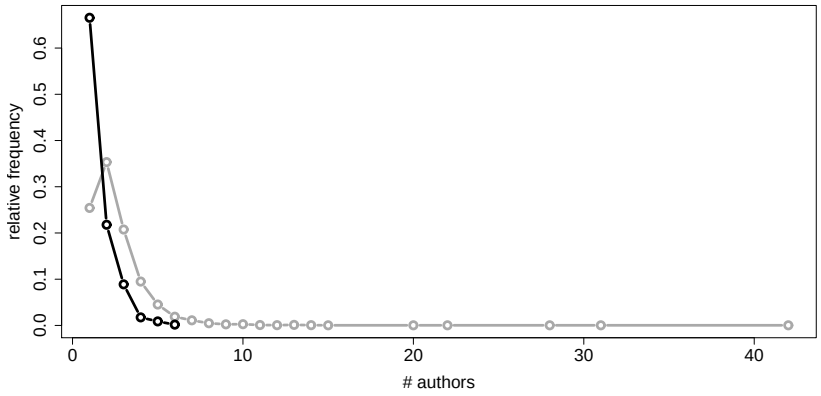

Fig. 2. Distribution of articles by the number of authors (black – JoC, gray – other).

The largest number of (co)authors in the articles from JoC is 6 (see Figure 2), while in other works the number of (co)authors is much larger. Most of the articles from JoC (almost 70 %) have only one author, while in other works 2 authors are more common. This confirms the conjecture that the JoC community has a different publishing 'culture' than the others.

3.1 Collaboration network

The collaboration network **Co** can be obtained from the 2-mode network **WA** by network multiplication $\mathbf{Co} = \mathbf{WA}^T * \mathbf{WA}$.

In larger collaboration networks we usually try to identify their dense parts using (generalized) cores (Seidman, 1983 [18]; Batagelj, Zaveršnik, 2002

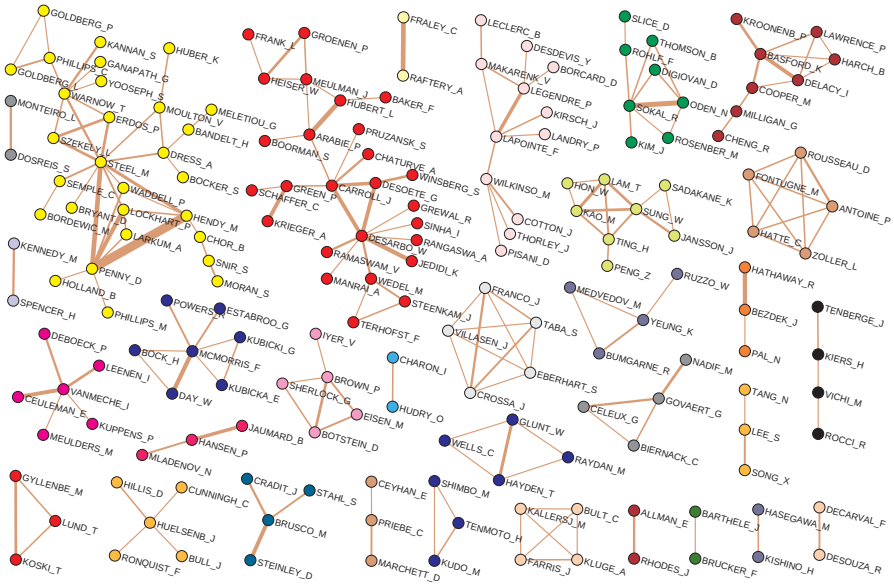

Fig. 3. Main part of line cut at level 3 of collaboration network. Colors represent cuts (connected subnetworks).

[7]). The collaboration network is a sum of cliques determined by each set of authors. For our network the results obtained by standard cores consist mainly of cliques corresponding to the papers with many coauthors. More interesting results are obtained by applying cores to the subnetwork of lines with weight at least 2 or using pS-cores. Even better view on the collaboration structure is obtained by the line cut at level 3 – we preserve in a network only lines with weight at least 3. This network has 298 vertices, 276 lines and 84 components. In Figure 3 its main part is presented.

3.2 Citation network analysis

Main path and CPM path in the citation network

To measure the importance (weight) of arcs in acyclic networks we use the methods proposed by Hummon and Doreian (1989) [13]. An efficient algoritm for computing these weights in large networks was developed by Batagelj in 1991 [3] and implemented in Pajek by Batagelj and Mrvar. The SPC (Search Path Count) method counts for each arc (u, v) the number of different paths from source (initial vertex) to sink (terminal vertex) passing through it. Therefore the higher the number, more paths pass the arc – more important is the arc. Citation networks are (almost) acyclic. The problem emerges if there are cycles (nontrivial strong components) in the network. Bounded JoC network

has 7 such strong components of size 2. All except one are citations between two works from the same publication. We shrink each of them into one vertex.

After weights are computed the main path and CPM path can be determined. Main path is a path from the source vertex to the sink, starting with the arc with the largest weight and selecting at each step the arc to the neighbors with the largest weight. CPM (Critical Path Method) determines the source-sink path(s) with the largest total sum of weights.

Main path: HOLDER_M(2008)57:814, STEEL_M(2008)57:243, COTTON_J(2007)56: 445, WILKINSO_M(2007)56:330, WILKINSO_M(2005)54:419, EULENSTE_D(2004)53:299, PISANI_D(2002)269:915, PISANI_D(2002)51:151, SEMPLE_C(2000)105:147, SANDERSO_M(1998)13:105, LAPOINTE_F(1997)46:306, PURVIS_A(1995)44:251, BULL_J(1993)42:384, DEQUEIRO_A(1993)42:368, BARRETT_M(1991)40:486, DONOGHUE_M(1989)20:431, KLUGE_A(1989)38:7, SOKAL_R(1986)17:423, DESOETE_G(1985)2:173, DESOETE_G(1984)1:235, CARROLL_J(1984)1:25, CARROLL_J(1983)48:157, PRUZANSK_S(1982)47:3, #TVERSKY_A(1982)89:123, SHEPARD_R(1980)210:390, CARROLL_J(1980)31:607, SHEPARD_R(1979)86:87, ARABIE_P(1978)17:21, WHITE_H(1976)81:730, BREIGER_R(1975)12:328, SHEPARD_R(1974)39:373, ARABIE_P(1973)10:148, CARROLL_J(1970)35:283, (HORAN_C(1969)34:139, BLOXOM_B(1968):, CLIFF_N(1968)33:225, MCGEE_V(1968)3:233, YOUNG_F(1967)12:498, ROSS_J(1966)31:27, SHEPARD_R(1966)3:287, TUCKER_L(1966)31:279, WOLD_H(1966):391, TUCKER_L(1964):109, TUCKER_L(1963)28:333, TORGERSO_W(1958):, ECKART_C(1936)1:211). The articles in brackets are all linked to the previous node.

Main topics of the works on the main path are *supertree methods in the consensus setting* in the latest works (mainly published in Systematic Biology), and *multidimensional scaling* in earlier works, published mainly in Journal of Mathematical Psychology and Psychometrika.

CPM path: GOKER_M(2008)8:86, AUCH_A(2006)7:350, THINES_M(2006)110:646, HUSON_D(2006)23:254, DELSUC_F(2005)6:361, GUINDON_S(2003)52:696, CHOR_B(2000)17:1529, STEEL_M(2000)17:839, MAU_B(1999)55:1, RAMBAUT_A(1997)13:235, #MIYAMOTO_M(1995)44:64, HUELSENB_J(1995)44:17, BULL_J(1993)42:384, DEQUEIRO_A(1993)42:368, DOYLE_J(1992)17:144, PAGE_R(1990)6:119, PAGE_R(1989)5:167, PAGE_R(1988)37:254, PENNY_D(1986)3:403, PENNY_D(1985)34:75, DAY_W(1983)66:97, DAY_W(1983)103:429, ROHLF_F(1982)59:131, ROHLF_F(1981)30:459, #MICKEVIC_M(1981)30:351, SOKAL_R(1981)30:309, SCHUH_R(1980)29:1, FARRIS_J(1979)28:483, MICKEVIC_M(1978)27:143, MICKEVIC_M(1976)25:260, FARRIS_J(1972)106:645, (FARRIS_J(1970)19:172, FARRIS_J(1970)19:83, KLUGE_A(1969)18:1, ESTABROO_G(1968)21:421, THROCKMO_L(1968)17:355, FARRIS_J(1967)16:44, HENNIG_W(1966):, CAMIN_J(1965)19:311, WILSON_E(1965)14:214, SOKAL_R(1963):).

Main topics on the CPM path are *phylogenetic analysis, evolutionary trees and genome trees* and most of the works on this path are published in Systematic Biology.

Although most works are related with biology, the only common works on both paths are BULL_J(1993)42:384 and DEQUEIRO_A(1993)42:368. All other works are different. Both paths can be found also in the main island in Figure 4.

Line islands in citation network

To detect connected subnetworks (clusters) with stronger internal cohesion relatively to its neighbors we used line islands. A line island of size $[k, K]$ is a weakly connected subnetwork of the selected size in the interval $[k, K]$

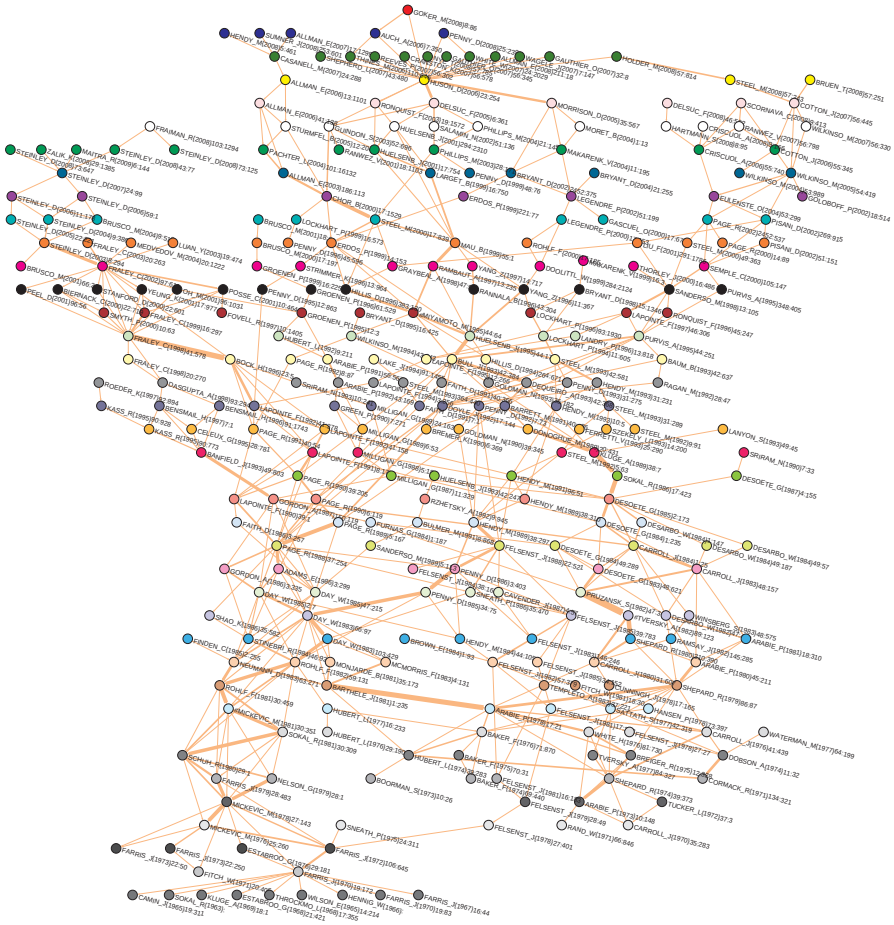

Fig. 4. Main line island of size [5, 300]. Colors represent the same topological level of vertices in the acyclic network.

where arcs linking vertices from the island to their neighbors outside island have weights lower than are values of arcs of a spanning tree inside the island [20]. Figure 4 represents the largest island of size [5, 300]. All other islands are much smaller.

In this island two main branches can be noticed. The first branch contains the CPM path. It starts with nine works from sixties (Sokal, 1963; Farris, 1967, 1970, etc.) at the bottom left. Then it follows works about taxonomic studies by Mickevitch and Rohlf, and about consensus index methods by Day, followed by Penny and Page. From there on, this branch is further divided into two parts: one goes in the middle and further on coincides with the CPM path. The other part (left of it) can be also seen as two branches: one on

the far left includes works from Page, Gordon and LaPointe on hierarchical-classification and its applications in the biosciences, continuing with Bock on probabilistic models in cluster analysis, Fraley and Raftery with review of general methodology for model-based clustering, and Steinley on cluster validation. The other branch include works from Milligan with methodology review, Arabie and Brusco on unidimensional and multidimensional scaling, and others, that are related also with works on the second main branch.

The second main branch on the right side of the figure is formed along the main path. It starts with works of Carroll and Arabie. At the bottom, both branches are connected with the strong arc from BARTHELE_J(1981)1:235 to ARABIE_P(1978)17:21 . Further, the second main branch continues with works of Shepard, Carroll, DeSarbo, De Soete et al. on multidimensional scaling, additive clustering, tree representation, and meets the first main branch with the article *Partitioning and combining data in phylogenetic analysis* by Bull (BULL_J(1993)42:384) , published in Systematic Biology, and with article *For Consensus (sometimes)* by Dequeiroz (DEQUEIRO_A(1993)42:368) . After them both branches separate again. The second main branch continues with works of Purvis and Sanderson on phylogenetic supertrees, then splits into two parts: one consisting of the works of Legendre on reticulate evolution, and the other that follows the main path with works from Pisani, Wilkinson and others on combining phylogenetic trees.

3.3 Citations between authors

By multiplying $\mathbf{Ca} = \mathbf{WA}^T * \mathbf{Ci} * \mathbf{WA}$, the authors citation network can be obtained. In this authors $\times$ authors network the arc weight corresponds to the number of citations that the first author makes to the second.

Line islands $[10, 400]$ – authors citations

There are 47 simple (one peak) line islands in authors citations network. The largest of them have sizes: 11 – Bezdek, Hathaway, et al.; 10 – Felsenstein, Penny, Hendy, Steel, et al.; 10 – Priebe, Wierman, et al.; 9 – Sokal, Gower, Legendre, et al.; 6 – Maharaj, et al.; 5 – Rohlf, et al. The strongest arcs are in the islands: Brusco $\rightarrow$ Hubert $\leftarrow$ Arabie; DeSarbo $\rightarrow$ Caroll $\leftarrow$ De Soete; and Steinley $\leftarrow$ Milligan. Increasing the upper bound K of island size, the islands with the strongest links join into a single island and the other islands are joining this island. This indicates that there is essentially a single main topic in the network.

Figure 5 presents the largest island where most of the well known names from the IFCS community can be found.

The main groups (clusters) that can be visually identified in the main island can be found also in a part of dendrogram, see Figure 6, corresponding to hierarchical clustering of vertices of the network using Ward's method on corrected Euclidean distance.

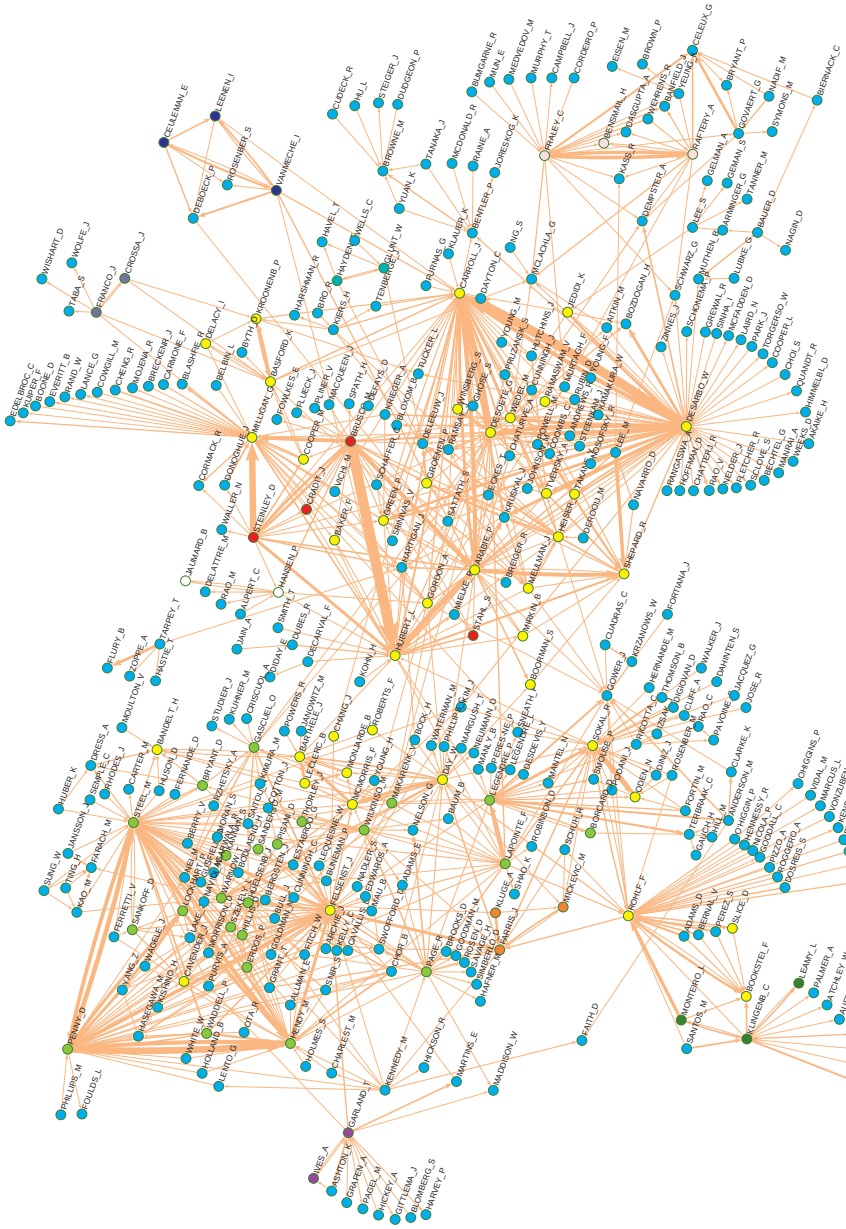

Fig. 5. Largest line island of authors citations.

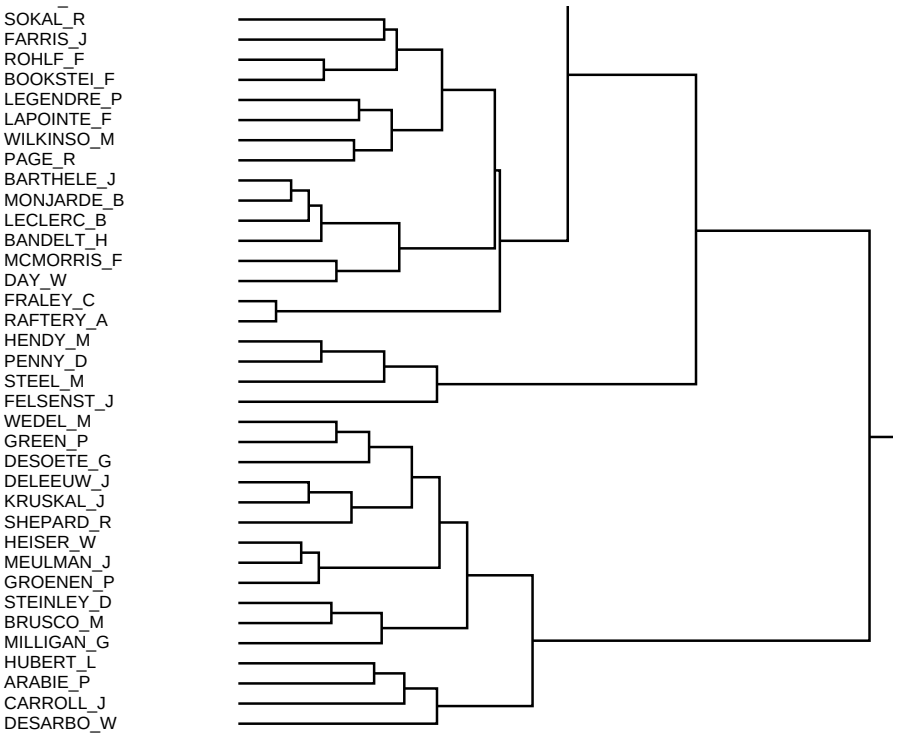

Fig. 6. Part of dendrogram.

To find out what is the topic of selected group of authors we considered the 2-mode network $\mathbf{AK} = \mathbf{WA}^T * \mathbf{WK}$. Its arc weight counts how many times the author A used the keyword K. From it we extract the subnetwork group × keywords and analyze it using methods presented in [1].

4 Conclusion

In the paper we presented the network analysis approach to analysis of bibliographic data. Program WoS2Pajek transforms the original data from Web of Science to a 1-mode citation network and three 2-mode networks (works × authors, works × journals, works × keywords) that can be analyzed separately or in combination with the citation network as derived networks. Using program Pajek we can identify important subnetworks in them and analyze their characteristics.

Because of limited space available for this paper some pictures are rather small and can be read in details only with a magnifying glass. The original color pictures in pdf format can be seen on the web page

`http://pajek.imfm.si/doku.php?id=examples`

References

1. A. Ahmed, V. Batagelj, X. Fu, S.H. Hong, D. Merrick, and A. Mrvar. Visualisation and analysis of the Internet movie database. Asia-Pacic Symposium on Visualisation 2007 (IEEE Cat. No. 07EX1615), pages 17–24, 2007.
2. A.L. Barabási, H. Jeong, Z. Néda, E. Ravasi, A. Schubert, and T. Vicsek. Evolution of the social network of scientic collaborations. *Physica A*, 331: 590–614, 2002.
3. V. Batagelj. Efficient Algorithms for Citation Network Analysis, 2003. `http://arxiv.org/abs/cs/0309023`
4. V. Batagelj. WoS2Pajek, 2008. `http://pajek.imfm.si/doku.php?id=wos2pajek`
5. V. Batagelj, and A. Mrvar. Some Analyses of Erdős Collaboration Graph. *Social Networks*, **22**: 173–186, 2000.
6. V. Batagelj, and A. Mrvar. Pajek – Program for large network analysis, 2008. `http://pajek.imfm.si`
7. V. Batagelj, and M. Zaveršnik. Generalized Cores, 2002. `http://arxiv.org/abs/cs.DS/0202039`
8. K. Börner, C. Chen, K. Boyack. Visualizing Knowledge Domains. In B. Cronin, *Annual Review of Information Science & Technology*, Volume 37, pages 179–255, Medford, NJ: Information Today, Inc./American Society for Information Science and Technology, 2003.
9. K. Börner. Atlas of Science: Guiding the Navigation and Management of Scholarly Knowledge. ESRI Press, 2009.
10. K. Boyack, R. Klavans, K. Börner. Visualizing Knowledge Domains. Mapping the backbone of science. *Scientometrics*, 64 (3), pages 351–?374, 2005.
11. S.N. Dorogovtsev, and J.F.F. Mendes. Evolution of networks. *Advanced Physics*, 51: 566–584, 2002.
12. E. Garfield. HistCite: Bibliometric Analysis and Visualization Software. `http://www.histcite.com`
13. N.P. Hummon, and P. Doreian. Connectivity in a Cittion Network: The Development of DNA Theory. *Social Networks*, 11: 39–63, 1989.
14. L. Leydesdorff, T. Schank, A. Scharnhorst, W. De Nooy. Animating the development of Social Networks over time using a dynamic extension of multidimensional scaling. *El Profesional de Información*, 17(6), 2008.
15. D. Lucio-Arias, L. Leydesdorff. Main-path analysis and path-dependent transitions in HistCite-based historiograms. *Journal of the American Society for Information Science and Technology*, 59 (12), pages 1948–1962, 2008.
16. M.E.J. Newman. The structure of scientific collaboration networks. *Proc. Natl. Acad. Sci. USA* 98: 404–409, 2001.
17. J. On, and A. Balkin. They Rule, 2004. `http://www.theyrule.net/html/about.php`
18. S.B. Seidman. Network structure and minimum degree. *Social Networks*, 5, pages 269–287, 1983.
19. Web of Science `http://isiknowledge.com/`
20. M. Zaveršnik, and V. Batagelj. Islands. Slides from Sunbelt XXIV, Portorož, Slovenia, 12–16 May, 2004.