ADVANCES IN NETWORK CLUSTERING AND BLOCKMODELING

Patrick Doreian
University of Ljubljana, University of Pittsburgh

Vladimir Batagelj
University of Ljubljana, IMFM Ljubljana, IAM UP Koper, NRU HSE Moscow

Anuška Ferligoj
University of Ljubljana, NRU HSE Moscow

WILEY-INTERSCIENCE

A JOHN WILEY & SONS, INC., PUBLICATION

CHAPTER 2

BIBLIOMETRIC ANALYSES OF THE NETWORK CLUSTERING LITERATURE

Vladimir Batagelj[1,2], Anuška Ferligoj[3], and Patrick Doreian[3,4]

[1] IMFM Ljubljana
[2] IAM, University of Primorska, Koper
[3] FDV, University of Ljubljana
[4] University of Pittsburgh
[5] NRU HSE Moscow

2.1 Introduction

Partitioning networks is performed in many disciplines, as is evidenced by the chapters of this book. The data we consider here are from the *network clustering literature*. Our focus here is the *large* set of publications identified in the area 'graph/network clustering and blockmodeling' and included also in the Web of Science[1] (WoS) through February 2017. The two dominant approaches for clustering networks are found in the 'social' social network literature and the literature featuring physicists and other scientists examining networks. Blockmodeling is an approach that partitions the nodes of a network into *positions* (clusters of nodes) with the *blocks* being the sets of relationships within and between positions. The result is simplified image of the *whole* network. Community detection, associated with the work of physicists studying networks, aims to identify *communities* composed of nodes *having a higher probability of being connected to each other than to members of other communities*. In identifying the literature featuring the clustering of networks we ensured the inclusion of both of these approaches.

[1] The origins of, and the rationale for, collecting such data are found in the work of Garfield [15], [1]

The rest of the chapter is structured as follows: Section 2.2 outlines steps in the collection of data and cleaning them together with constructing measures and identifying specific productions. Section 2.3 presents several approaches to identifying network features including components, critical main paths, and key-route paths for analyzing citation networks. Section 2.4 examines line islands as clusters in the network clustering literature. Section 2.5 focuses on authors, productivity, collaboration, and bibliometric coupling. The chapter concludes with suggestions for future work.

2.2 Data collection and cleaning

We view scientific productions as *works* and sought the citation links connecting them. Citations from later works to prior works can be viewed as 'votes' from researchers in their scientific fields regarding the value of earlier scientific works. Given our focus on network clustering literature, we obtained data from the Web of Science (WoS) (now owned by Clarivate Analytics) by using the following terms in a general query:

```
"block model*" or "network cluster*" or "graph cluster*" or
"community detect*" or "blockmodel*" or "block-model*" or
"structural equival*" or "regular equival*"
```

We limited the search to the Web of Science Core Collection because other data bases from WoS do not permit exporting CR-fields (which contain citation information). Some works appear only in the WoS CR field as a reference and lack a description in the collected data set. We call such works *cited-only* works. Additionally, we collected, using WoS and Google, some information about cited-only nodes with large indegrees (highly cited works) to add such descriptions to the collected data set. When a description of a node was unavailable in these sources, we manually constructed a description for them.[2]

Our first WoS search was completed on May 16, 2015. It was updated on January 6, 2017 for 2014-2017. A further updating for 2015-2017 was completed on February 22, 2017. We applied the new WoS2Pajek 1.5 [3] to convert WoS data into Pajek networks[3]. Preliminary results regarding the size of the data set are shown in Table 2.1. In slightly less than two years, the number of works increased by 56%, the number of authors by 38%, the number of journals by 40%, and the number of records by 136%. Clearly, partitioning networks is a *rapidly expanding* area of research in multiple areas given the increases in the number of works, authors and journals. Of some interest is that the increase of authors was less than the number of works. The decrease of the final number of keywords is due to the replacement of keyword phrases with the constituting words.

While a citation network is simply composed of links between works treated as nodes, there is more to consider when other units are included. These include authors, journals, and keywords. As part of a more general strategy, the following two-mode networks were constructed: i) an author network, **WA** as works × authors; ii) a journal network **WJ** featuring works × journals; iii) a keyword network **WK** with works × keywords; as well as iv) a one-mode citation network **Ci** featuring only scientific productions. Additional information was obtained considering some useful partitions: i) *year* of works by publication

[2] There are two approaches to deal with the resulting data: i) manually filtering the hits and preserving only those matching the criteria or ii) using all obtained hits while considering non-topic hits as noise. Given the enormous amount of work required for the first option, we used the second one.

[3] Most of the analyses featured in the chapter were done in Pajek (see [5]) and R [26]. For a highly accessible introduction to Pajek, see [25].

Table 2.1: Sizes of networks on clustering literature

	2015/05/16	2017/01/06	2017/02/23
Number of works	75249	112114	117082
Number of authors	44787	60419	62143
Number of journals	8993	12271	12652
Number of keywords	10095	12715	10269
Number of records	2944	5472	6953

year; ii) a *DC* partition distinguishing works having a complete description (DC=1) and cited-only works (DC=0); and iii) a vector of the number of pages, *NP*. The dimensions of the studied networks (shown in the right-most column of Table 2.1): the number of works, $|W| = 117082$; the number of contributing authors, $|A| = 62143$; the number of journals where these works appear, $|J| = 12652$; and the number of keywords employed to characterize works, $|K| = 10269$. All these networks share the set of works (papers, reports, books, etc.), W.

Another problem complicating data collection is that different data sources use different conventions for their data items. The usual *ISI name* of a work (field CR), has the form:

```
LEFKOVITCH LP, 1985, THEOR APPL GENET, V70, P585
```

All its elements are upper case. AU denotes author, PY is for publication year, SO denotes journals (with an allowance for at most 20 characters), VL is for Volume, and BP denotes the beginning page. Its format is:

AU + `', '` + PY + `', '` + SO[:20] + `', V'` + VL + `', P'` + BP

In WoS, the same work can have different ISI names! To improve the precision of identification of works (entity resolution, disambiguation), the program **`WoS2Pajek`** supports also *short names* with the format:

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

For example: `LEFKOVIT_L(1985)70:585`

For last names with prefixes, e.g. `VAN`, `DE`, ... the space is deleted. Unusual names start with character `*` or `$`. A citation network, **Ci**, is based on the citing relation where $w\,\mathbf{Ci}\,z$ means work, w, cites work, z.

For correcting equivalent data items, there are two options: i) make corrections in the local copy of original data (WoS file); or ii) make the equivalence partition of nodes and shrink the set of works accordingly in all networks. We used the second option. For the works with large counts (≥ 30), we prepared lists of possible equivalent items and manually determined equivalence classes. Using a simple program in Python, we produced a Pajek partition file, `worksEQ.clu`, and shrank sets of works using Pajek. Using the partition $p = \mathit{worksEQ}$, $p : V \to C$, we used Pajek to shrink the citation network *cite* to *citeR*. As a byproduct, we obtained a partition $q : V_C \to V$, such that $q(v) = u \Rightarrow p(u) = v$. It was necessary to shrink also the partitions *year*, *DC* and the vector *NP*. This can be done in Pajek as follows. Given a general mapping $s : V \to B$, we seek a mapping $r : V_C \to B$ such that if $q(v) = u$, then $s(u) = r(v)$. Therefore, $r(v) = s(u) = s(q(v)) = q * s(v)$ or equivalently $r = q * s$.

In Pajek, given a mapping $q : V_C \to V$, the mapping r is determined for a partition s by:

```
select partition q as First partition
```

Table 2.2: Sizes of "reduced" networks

Network	Nodes	Arcs
WAr	179049 = 116906 + 62143	132776
WKr	127175 = 116906 + 10269	88965
WJr	129558 = 116906 + 12652	117044
CiR	116906	195784

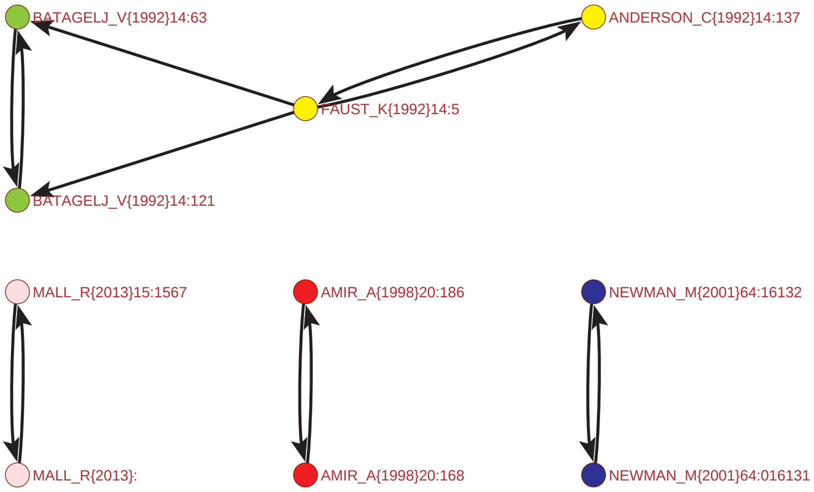

Figure 2.1: Dyadic strong components

```
select partition s as Second partition
Partitions/Functional Composition First*Second
```

or for a vector *s* by

```
select partition q as First partition
select vector s as First vector
Operations/Vector+Partition/
Functional Composition Partition*Vector
```

For the partition *worksEQ*, we computed the "reduced" networks **CiR**, **WAr**, **WKr**, **WJr** and the partitions *YearR* and *DCr* as well as the vector *NPr*. Their sizes are shown in Table 2.2. For example, the network **WAr** has 179049 nodes – 116906 works and 62143 authors.

For cited only works we have only information about their first author and no information about keywords. So, we have to limit our analysis about authors or keywords only to works with complete descriptions ($DC > 0$). The sizes of corresponding networks are shown in Table 2.3.

In principle, citation networks are *acyclic*: earlier works cannot cite later works. Yet, works appearing at the same time can cite each other. As the methods we use require a ci-

Table 2.3: Sizes of networks with complete descriptions

Network	Nodes	Arcs
WAc	19071 = 5695 + 13376	21562
WKc	15964 = 5695 + 10269	88953
WJc	7451 = 5695 + 1756	5815
CiC	5695	38400

tation network to be acyclic, such ties must be located. More generally, strong components need to be identified. There were five in the network we studied, all in the form of reciprocal dyads. These are shown in Figure 2.1. Methods for identifying strong components and ways of treating them prior to analyzing citation networks are described in [7].

2.2.1 Most cited/citing works

It is straightforward to identify the works in a citation network receiving the most citations[4]. Similarly, identifying works with the greatest outdegrees is straightforward.

Table 2.4 lists the 60 most cited works (indegree in **CiR**). Heading the list are seven works produced in the physicist approach to networks featuring community detection. The top ranked document is by Girvan and Newman[5] as a research paper, in the *Proceedings of the National Academy of Sciences* (US) in 2002. The second ranked paper written by Fortunato is a long survey paper on community detection in graphs, appearing in *Physics Reports* in 2010. In third place is a 2004 paper on community detection for very large networks by Clauset, Newman, and Moore in *Physical Review E*. The most cited paper from the social sciences, at rank 7, is by an anthropologist whose data attracted the attention of the aforementioned physicists. The next highest document from the social sciences is the Wasserman and Faust book of 1994. The other ‘social’ social network productions in this list primarily feature works devoted to blockmodeling, albeit the earlier productions in this area. This is suggestive of the domination, in recent years, by the approach adopted by physicists when studying social networks to identify communities as clusters.

In Table 2.5 the Top 10 citing works (outdegree in **CiR**) are listed. They consist of books, theses and survey papers. Only two of the items in this table come from the social sciences. The role of survey papers was studied in [7] with an emphasis on their secondary role in the *production* of scientific knowledge.

2.2.2 The boundary problem for citation networks

For any network study, the boundary of the network must be determined with great care. In some studies, the context determines the boundary in a straightforward fashion. However, for citation networks the problem is far more ambiguous in that judgments must be made.

[4] The results reported here follow in the tradition outlined in [11].

[5] We have adopted the convention of citing only methodologically relevant items for the methods we consider in this chapter. Such frequently cited paper can be identified easily in the relevant literature.

Table 2.4: The most cited works in the network clustering literature

Rank	Citations	Work	Rank	Citations	Work	Rank	Citations	Work
1	1096	GIRVAN_M(2002)99:7821	21	292	NEWMAN_M(2003)45:167	41	145	BURRIDGE_R(1967)57:341
2	969	FORTUNAT_S(2010)486:75	22	292	LANCICHI_A(2009)80:056117	42	145	LANCICHI_A(2011)6:0018961
3	712	CLAUSET_A(2004)70:066111	23	286	NEWMAN_M(2004)69:1	43	139	GREGORY_S(2010)12:103018
4	638	BLONDEL_V(2008):P10008	24	259	GUIMERA_R(2005)433:895	44	139	LESKOVEC_J(2010):
5	621	NEWMAN_M(2004)69:026113	25	251	ALBERT_R(2002)74:47	45	138	BOCCALET_S(2006)424:175
6	578	NEWMAN_M(2006)103:8577	26	244	DUCH_J(2005)72:027104	46	137	GUIMERA_R(2004)70:025101
7	553	ZACHARY_W(1977)33:452	27	236	LUSSEAU_D(2003)54:396	47	129	NEWMAN_M(2004)70:056131
8	544	PALLA_G(2005)435:814	28	216	SHI_J(2000)22:888	48	127	BRANDES_U(2008)20:172
9	489	FORTUNAT_S(2007)104:36	29	216	LORRAIN_F(1971)1:49	49	126	BREIGER_R(1975)12:328
10	416	WATTS_D(1998)393:440	30	215	REICHARD_J(2006)74:016110	50	126	NOWICKI_K(2001)96:1077
11	412	DANON_L(2005):	31	211	HOLLAND_P(1983)5:109	51	125	ROSVALL_M(2007)104:7327
12	380	NEWMAN_M(2004)38:321	32	206	WHITE_H(1976)81:730	52	124	VONLUXBU_U(2007)17:395
13	369	LANCICHI_A(2008)78:046110	33	199	AHN_Y(2010)466:761	53	122	NEWMAN_M(2001)64:026118
14	351	WASSERMA_S(1994):	34	168	KERNIGHA_B(1970)49:291	54	119	REICHARD_J(2004)93:218701
15	329	NEWMAN_M(2006)74:036104	35	163	AIROLDI_E(2008)9:1981	55	118	ARENAS_A(2008)10:053039
16	326	ROSVALL_M(2008)105:1118	36	161	NEWMAN_M(2010):	56	118	ERDOS_P(1959)6:290
17	319	RAGHAVAN_U(2007)76:036106	37	157	SCHAEFFE_S(2007)1:27	57	116	FREEMAN_L(1979)1:215
18	307	LANCICHI_A(2009)11:033015	38	155	GOOD_B(2010)81:046106	58	116	FREEMAN_L(1977)40:35
19	306	RADICCHI_F(2004)101:2658	39	150	KARRER_B(2011)83:016107	59	113	NEWMAN_M(2001)98:404
20	304	BARABASI_A(1999)286:509	40	150	LANCICHI_A(2009)80:016118	60	112	SHEN_H(2009)388:1706

Table 2.5: The most citing works of the network clustering literature

Rank	Citations	Document	Rank	Citations	Document
1	1095	PRUESSNE_G(2012):1	6	417	NEWMAN_M(2003)45:167
2	863	BOCCALET_S(2006)424:175	7	398	FORTUNAT_S(2010)486:75
3	839	FOUSS_F(2016):1	8	327	HOLME_P(2015)88:e2015-60657-4
4	476	ARABIE_P(1992)43:169	9	321	SIBLEY_C(2012)12:505
5	456	TURCOTTE_D(1999)62:1377	10	310	FRANK_K(1998)23:171

It is reasonable to exclude cited-only works with indegree 1 for this indicates minimal notice. More generally, to get rid of the influence of sporadic citations, some threshold in terms of citations received for inclusion is necessary. To examine this the following counts were established.

The network **CiR** has 116906 nodes and 195784 arcs. The counts for the lowest number of received citations are: 0 (4070); 1 (93248); 2 (10694); 3 (3352) and 4 (1610). Most nodes are cited only once (indegree=1). We 'solved' the boundary problem by including in our networks those nodes with $DCr > 0$ or indeg > 2. These criteria determined a subnetwork, denoted as **CiB**, with 13540 nodes and 82238 arcs.

With the network boundary determined, obtaining general description is straightforward prior to completing any analyses. Table 2.6 lists journals whose articles were cited the most. The left panel came from the **WJr** network while the right panel came from **WJc** (defined for *only* those documents having complete descriptions). Without surprise, the counts for the journals differ substantially as the two networks differ greatly in size. More consequentially, the orders of the journals differ. Journals from the social sciences are marked in boldface.

For the much larger network, **WJr**, the dominant journals are the *Proceedings of the National Academy of Sciences* (US) and *Nature* with over 1000 citations. Both *Lecture Notes in Computing Sciences* and *Science* contained more than 900 citations. Three Physics journals follow. The top-ranked social science journal *Social Networks* is in tenth place. The remaining journals cover many disciplines.

But for the network with only complete descriptions for the works, **WJc**, (works are the hits dealing with the research topic) there are dramatic changes. The *Proceedings of the National Academy of Sciences* drops to the ninth place and *Nature* drops to nineteenth place. Many other journals drop out of the list. In contrast, both *Physica A* and *Physics Review E* retain their high rankings. *Social Networks* moves up to fourth place. Other journals in the right panel replace those dropping out of the left panel.

These differences reinforce the importance of solving the boundary problem appropriately. While strong cases can be made for using either **WJr** or **WJc**, it is clear that setting different boundaries can lead to dramatically different outcomes. One obvious question is whether having more information about productions is worth it. In terms of *interpreting* citation patterns and, more generally, understanding science dynamics, we contend that having more information is preferred. As a general point, when results are reported, the ways in which boundaries for networks are established *must* be made clear.

Most journals demand the use of keywords which become part of the information about works. When keywords are not parts of works, they can be constructed from titles. Composite keywords were split into single words. Lemmatization was used in WoS2Pajek to deal with the 'word-equivalence problem'. Table 2.7 lists the frequency counts for keywords attached to works in the network **WKc**. Having 'network' as the most frequent keyword is trivial. The next two items, 'community' and 'detection', suggest a problem

Table 2.6: The most used journals in two works × journals networks

Rank	Frequency in **WJr**	Journal	Frequency in **WJc**	Journal
1	1058	P NATL ACAD SCI USA	223	LECT NOTES COMPUT SC
2	1014	NATURE	175	PHYS REV E
3	941	LECT NOTES COMPUT SC	151	PHYSICA A
4	908	SCIENCE	122	**SOC NETWORKS**
5	667	PHYSICA A	88	PLOS ONE
6	639	PHYS REV E	56	LECT NOTES ARTIF INT
7	616	PHYS REV LETT	56	J GEOPHYS RES-SOL EA
8	549	BIOINFORMATICS	45	P NATL ACAD SCI USA
9	548	NUCLEIC ACIDS RES	40	SCI REP-UK
10	522	**SOC NETWORKS**	39	J STAT MECH-THEORY E
11	519	J GEOPHYS RES-SOL EA	33	NEUROCOMPUTING
12	428	B SEISMOL SOC AM	30	PHYS REV LETT
13	400	TECTONOPHYSICS	28	COMM COM INF SC
14	398	GEOPHYS J INT	27	APPL MECH MATER
15	348	NEUROIMAGE	27	BMC BIOINFORMATICS
16	342	J GEOPHYS RES	27	EUR PHYS J B
17	342	J BIOL CHEM	27	GEOPHYS J INT
18	336	J MOL BIOL	25	PROCEDIA COMPUT SCI
19	330	PHYS REV B	25	BIOINFORMATICS
20	321	IEEE T PATTERN ANAL	24	INFORM SCIENCES
21	285	**AM J SOCIOL**	23	IEEE DATA MINING
22	274	PATTERN RECOGN	23	KNOWL-BASED SYST
23	272	**AM SOCIOL REV**	23	**J MATH SOCIOL**
24	260	GEOPHYS RES LETT	21	**SOC NETW ANAL MIN**
25	249	GEOLOGY	21	ADV INTELL SYST
26	239	**SCIENTOMETRICS**	20	MATH PROBL ENG
27	229	LECT NOTES ARTIF INT	20	EXPERT SYST APPL
28	224	EARTH PLANET SC LETT	19	EPL-EUROPHYS LETT
29	220	BIOCHEMISTRY-US	19	INT J MOD PHYS B
30	214	APPL ENVIRON MICROB	19	TECTONOPHYSICS
31	212	J CHEM PHYS	19	ANN STAT
32	207	J NEUROSCI	19	NATURE
33	207	J AM STAT ASSOC	18	IEEE T KNOWL DATA EN
34	205	J GEOPHYS RES-SOLID	18	PATTERN RECOGN LETT
35	201	J AM CHEM SOC	18	**AM J SOCIOL**
36	187	J PHYS A-MATH GEN	17	ADV MATER RES-SWITZ
37	185	**ADMIN SCI QUART**	17	PURE APPL GEOPHYS
38	184	CELL	16	DATA MIN KNOWL DISC
39	184	PURE APPL GEOPHYS	16	GEOPHYS RES LETT
40	181	INFORM SCIENCES	16	IEEE T PATTERN ANAL
41	171	BIOPHYS J	16	**SCIENTOMETRICS**
42	170	**PSYCHOMETRIKA**	15	INT CONF ACOUST SPEE
43	167	IEEE T KNOWL DATA EN	14	NEW J PHYS
44	165	EUR PHYS J B	14	J CLASSIF
45	159	EXPERT SYST APPL	14	IEEE T MICROW THEORY
46	159	GEOL SOC AM BULL	14	**PSYCHOMETRIKA**
47	158	EUR J OPER RES	13	SCI WORLD J
48	154	IEEE T INFORM THEORY	13	J COMPUT SCI TECH-CH
49	144	PATTERN RECOGN LETT	13	PLOS COMPUT BIOL
50	142	**J PERS SOC PSYCHOL**	13	ADV COMPLEX SYST

Table 2.7: The most used keywords

Rank	Freq.	Keyword	Rank	Freq.	Keyword	Rank	Freq.	Keyword
1	1204	network	25	291	earthquake	48	186	similarity
2	1064	community	26	281	protein	49	184	multi
3	1533	detection	27	276	stochastic	50	181	evolution
4	1499	model	28	270	overlap	51	176	mining
5	1177	graph	29	268	fault	52	166	functional
6	1135	cluster	30	265	equivalence	53	165	behavior
7	1104	algorithm	31	241	prediction	54	164	simulation
8	1082	complex	32	240	organization	55	163	state
9	1080	social	33	237	interaction	56	163	gene
10	932	structure	34	236	scale	57	160	genetic
11	900	analysis	35	229	time	58	159	centrality
12	880	base	36	227	clustering	59	157	flow
13	727	block	37	220	theory	60	156	classification
14	494	use	38	213	large	61	155	partition
15	430	datum	39	209	self	62	155	hierarchical
16	407	modularity	40	205	matrix	63	150	application
17	398	method	41	204	dynamic	64	148	slip
18	373	dynamics	42	204	identification	65	146	small
19	357	structural	43	197	modeling	66	146	design
20	317	approach	44	197	pattern	67	146	link
21	300	blockmodel	45	195	detect	68	145	web
22	294	information	46	194	local	69	144	organize
23	293	optimization	47	190	world	70	143	spectral
24	293	random						

with keywords containing two words. As a term relating to clustering, ‘cluster’ is only in the sixth place.

Many of the other frequently used terms in Table 2.7 including model, graph, and structure are generic with limited value. Other keywords – complex, social, base, use, datum, method, approach, information, fault, scale, self, local, world, gene, genetic, flow, slip, small, and organize - convey less information. Either keywords are utterly useless for understanding of scientific citation or they have to be *examined with great care in clearly defined contexts*. To this end, we identify parts of the citation network by identifying islands (see [7]) of closely related works in them. For this, keywords become very useful for discerning the major interests of the works in an island as a *focused* substantive context. The same idea is clear also when we consider bibliographic coupling.

2.3 Analyses of the citation networks

Given our focus on citation networks, we consider ways of identifying and interpreting important parts of these networks. They include components for identifying important paths through these networks based in the ideas formulated in [17], used to examine the DNA development literature in [18], applied to the network centrality literature in [19] and extended in [7].

2.3.1 Components

Our analyses of the primary ‘clustering citation network’ (**CiB**) features components, the identification of main paths through this literature, identifying islands (as clusters of related works) and bibliometric coupling. Our main use of components is for identifying networks useful for obtaining important paths and islands. The network, **CiB**, has 690 (weak) components. The largest have sizes 12702, 21, 20, 19, 17, 10, and 9. Here, we limit our analysis to the largest component, labeled **CiteMain**.

The presence of the reciprocal dyads identified in Figure 2.1 remains. To obtain an acyclic network, we applied the preprint transformation (see [7]) to **CiteMain**. The resulting network, **CiteMacy** (Cite, Main, acyclic), has 12712 nodes and 81972 arcs. The increase in the number of works is due to some of them appearing twice with one name starting with an = sign indicating the “preprint” version of a paper. We computed the SPC weights on its arcs [2]. The total flow is 1.625 10^{20}.

2.3.2 The CPM path of the main citation network

We start by identifying main paths. Figure 2.2 shows the CPM main path [7, 2] through the network clustering literature (in **CiteMacy**). At the bottom of this main path, there are seven publications, all cited by an influential paper by Cartwright and Harary appearing in 1956. They are important foundational works for social network analysis. It continues with 22 publications from the blockmodeling literature encompassing both unsigned and signed networks. This is followed by an important transition in this main path marking a transition between the social networks field and the work of natural scientists on social networks. The and (2000) publication is the last work from the area of social network analysis. It analyzed the Erdős collaboration graph. The connecting link feature this production and one by Newman in 2001. Thereafter, the rest of the main path features works from the community detection literature through 2016. We expand further on this description when discussing key-route paths.

The branching at the top of the figure reflects the end of the search period we used. The top four papers cite a work by Fortunato and Hric appearing in 2016. Were a new search used to expand this main path, undoubtedly these most recent works would be cited and the main path would continue through some of them. We note that when the network centrality literature was analyzed in [7] a similar transition between fields was identified: social networks to physics to neuroscience.

2.3.3 Key-route paths

The CPM approach yields a single main path through the literature. A more nuanced image of this feature is obtained by identifying key-routes through a network. This method, known as the Taiwan approach, was developed in [22]. The algorithm has been generalized and included in Pajek. The Pajek instruction for obtaining key-routes through 150 arcs with the largest weights is:

```
Network/Acyclic Network/Create (Sub)network/CPM/
   Global Search/Key-Route [1-150]
```

Figure 2.3 shows the results for this network clustering network. The starting and ending works are the same for both Figure 2.2 and Figure 2.3. However, between these ends,

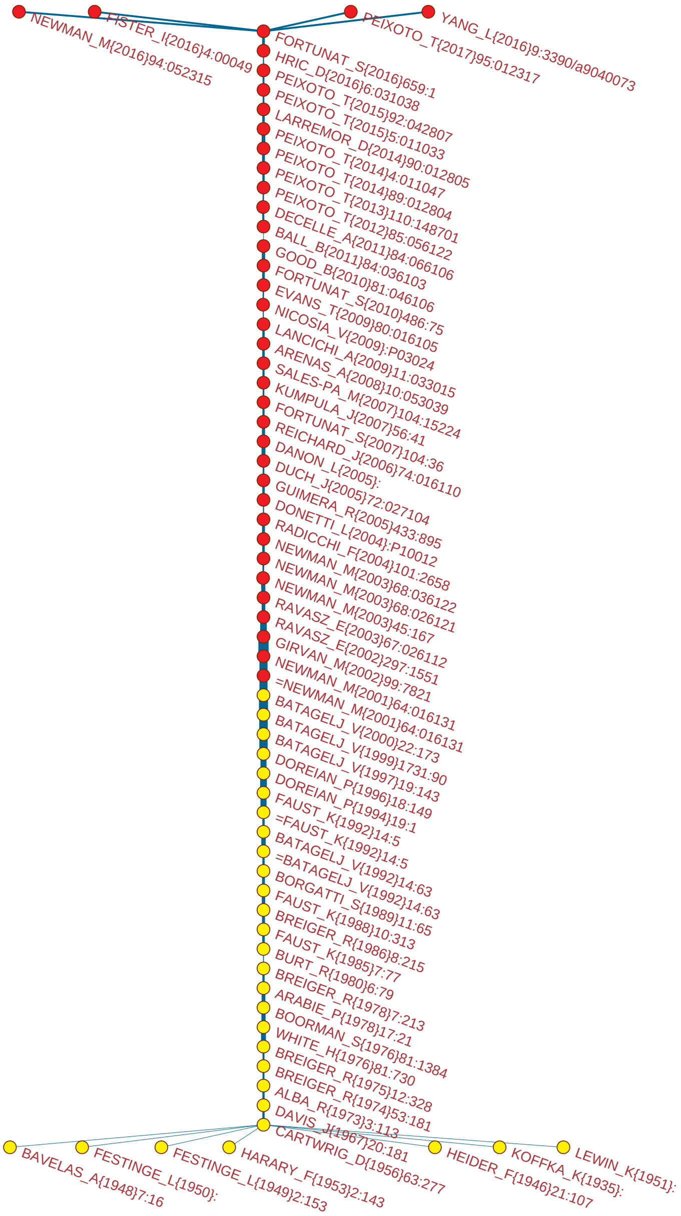

Figure 2.2: The CPM path through the network clustering literature

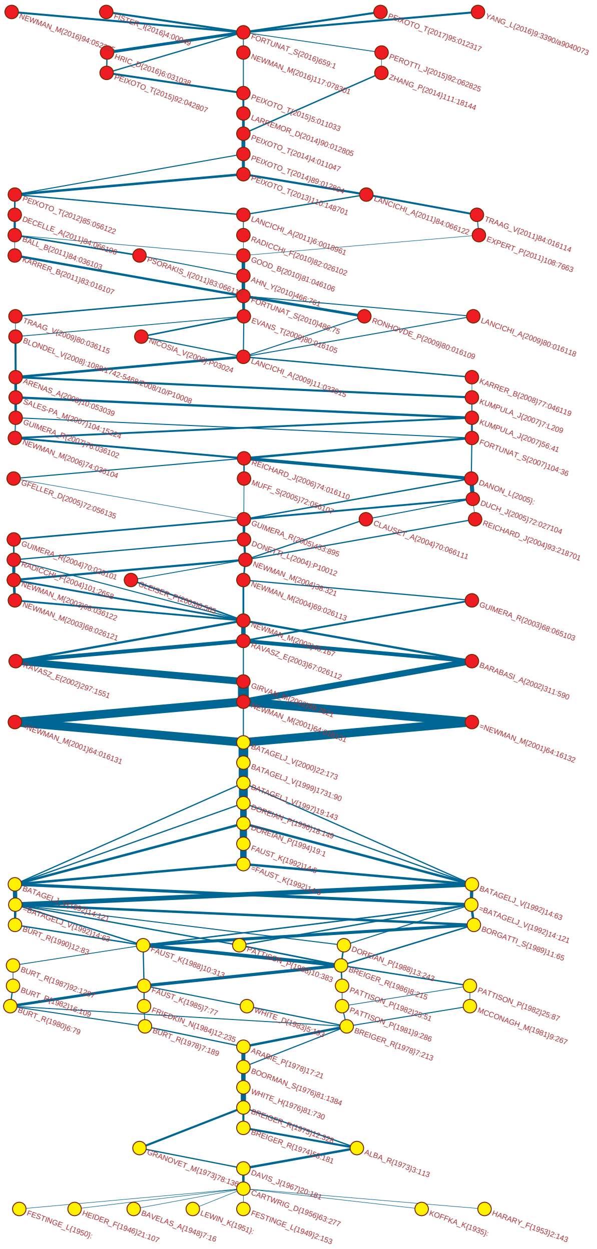

Figure 2.3: Key-route paths through the network clustering literature

additional works are included to provide a more complex view of the evolution of the clustering field(s). The basic sequence between the social network and community detection literatures remains. Indeed, the transition point between these two literatures is a cut.

We divide our expanded discussion into two temporal periods.

The period 1956–2000 Two papers by Cartwright and Harary and Davis formed the foundations for signed blockmodeling. After these two papers, we would have expected to see the foundational paper for blockmodeling of Lorrain and White (appearing in1971). But it is not on the CPM main path nor on the key-routes. We account for this below in our discussion of Tables 2.8 and 2.9. Next comes a paper of Alba discussing cliques, a conceptual dead end even though it is much studied in the social networks area. This is followed by Breiger who created the foundations for analyzing two-mode networks, a critically important development. The five papers involving Breiger, Boorman, Arabie, White, Levitt and Pattison, all important for creating the blockmodeling tradition, follow. Included is the work outlining the first algorithm, CONCOR, for blockmodeling and works with substantive interpretations of blockmodeling results involving White, Boorman, Breiger, Arabie, Levitt, and Pattison in the mid-to-late 1970s. Also appearing on the main path are papers on explanations of role structure theory in algebraic models involving Boorman and White and Breiger and Pattison. Burt proposed a rival algorithm for blockmodeling in 1976 which is not on the main path. A later paper from him, published in 1980, is on the main path.

A special issue of *Social Networks* devoted to blockmodeling appeared in 1992. Four papers from this issue are in the main path: two works by Batagelj, Doreian, Ferligoj introducing the direct approach to blockmodeling for structural and regular equivalence, and two papers by Faust and Wasserman (with one with Anderson) discussing the interpretation and evaluation of blockmodels and stochastic blockmodels. In 1994, Doreian, Batagelj and Ferligoj proposed generalized concepts of equivalence based on block types and corresponding criterion functions which provides an appropriate measure of fit of blockmodels to the empirical data.

Also on this main path is a paper by Doreian and Mrvar appearing in 1996 who used the generalized blockmodeling approach and applied it to signed networks and a paper by Batagelj (appearing in 1997) which provided a mathematical formalization of the generalized blockmodeling. The last two papers in the class of the social network contributions involve Batagelj, Mrvar and Zaveršnik who proposed several clustering procedures for large networks and applied these algorithms to the Erdős collaboration graph. As noted above, this work is the bridge to the contributions of natural scientists, mostly working on community detection problems.

The period 2001–2016 A paper by Newman, appearing in 2001 is the first production on the main path for works from the natural sciences. He presented a variety of statistical properties of scientific collaboration networks. An important contribution for the development of community detection approach is the paper of Girvan and Newman, also on the CPM main path and the key-routes through this network. Here (and in some other papers not included in the main path but are in the key-route paths and islands) they introduced the clustering coefficient. They also introduced the term community detection to avoid confusion with the clustering coefficient. We note that only recently, with the further development of stochastic blockmodels, did the social networks terminology get used again, albeit to a limited extent.

Next, two papers of Ravasz with her collaborators discuss the hierarchical organization in complex networks. Later, productions by Sales-Pardo and Arenas *et al.* also deal with this topic. Newman applied a variety of techniques and models to analyze complex networks and to examine the properties of highly clustered networks in 2003. In the same year, Newman and Park argued that social networks differ from most other types of networks. Next, four papers propose different algorithms for detecting network communities involving (Radicchi *et al.*, Donetti and Munozuch, Arenas, and Ball *et al.*). Guimera and

Amaral in *Nature* analyzed complex metabolic networks. The first paper on the main path dealing with the statistical aspects of community detection by Reichardt and Bornholdt appeared in 2006. Fortunato and Bathelemy found that modularity optimization may fail to identify smaller modules. Kumpula et al. then proposed an approach for dealing with this problem.

The following two papers involving (Lancichinetti *et al.* and Nicosia *et al.*) proposed an approach for detecting overlapping structures in complex networks. Evans and Lambiotte proposed clustering links of a network. The next paper on the main path is by Fortunato, a highly cited overview of community detection in networks. Good *et al.* studied the performance of modularity maximization. The first paper in the main path discussing stochastic blockmodels is by Decelle *et al.*. This idea was developed further by Peixoto in several papers appearing between 2012 and 2014. Larremore *et al.* studied the community structure in bipartite networks. In 2015, Peixoto used a statistical approach to large network models to discern overlapping clusters. Similarly, Hric *et al.* developed a joint generative model for data and meta-data to attempt the prediction of missing nodes. Peixoto's terminology is becoming closer to the one used in social network analysis. The last paper in the main path by Fortunato and Hric is a user guide for community detection in networks.

Tables 2.8–2.12 They provide more details regarding the authors, works, and journals for the works in the CPM path, key-routes, and islands. They are also relevant for our discussion of islands in Section 2.7. The five tables form a single extended table which are separated only for pagination reasons. In these tables the labels of the works are given in the first column, in the second column (code) it is described in which analysis the work appeared (1 – CPM path, 2 – Key-routes, and 3 – link island). The following columns give the first author of the work, the work's title and the journal in which the work was published.

The items in Table 2.8 and all but the last six items in Table 2.9 come from the social networks literature. The earliest items set the foundations of, and inspiration for, the development of social network analysis. The foundational paper for blockmodeling was published in 1971 by Lorrain and White [23]. Its absence from both the CPM main and key-routes is due to it being mathematically 'fierce' in its use of category theory. However, the 1975 Breiger *et. al.* [8], the White *et. al.* (1976) and the Boorman *et. al.* (1976) (the sixteenth through eighteenth items) provided the first algorithm for blockmodeling along with substantive interpretations of blockmodeling results. The next three papers in the table, by Burt, introduced a rival algorithm for blockmodeling, especially [9]. Other papers presented blockmodeling results, critiques of methods, and discussions of closely related topics, especially role structures.

The Heider (1946) paper (the second work in Table 2.8), along with the Harary (1953) (sixth), the Cartwright and Harary (1956) (eighth) and Davis (1967) (tenth) papers formed the foundations for the creation of signed blockmodeling by Doreian and Mrvar (1996) (the twenty third item in Table 2.9.) The basic idea is located in the structural theorems in the papers of Cartwright and Harary, and in Davis, being coupled to the direct approach to blockmodeling [13].

Examining journals as venues for works is facilitated by considering the right-hand column of Tables 2.8–2.12. Many of the journals relating to blockmodeling in Table 2.8 are from the mainstream sociological literature. They include two from *The American Journal of Sociology* and four each from *Social Forces* and *Sociological Research and Methods*. The list of journals in Table 2.9 reveals a sharp transition with *Social Networks* appearing fifteen times. It appeared just once in Table 2.8. It appears that: i) blockmodeling became

Table 2.8: List of works on CPM path (1), main paths (2) and island (3) – part 1

label	code	first author	title	journal
KOFFKA_K(1935):	12	Koffka, K	Principles of Gestalt Psychology	book
HEIDER_F(1946)21:107	12	Heider, F	Attitudes and cognitive organization	J PSYCHOL
BAVELAS_A(1948)7:16	12	Bavelas, A	A mathematical model for group structure	HUMAN ORG
FESTINGE_L(1949)2:153	12	Festinger, L	The analysis of sociograms using matrix algebra	HUMAN REL
FESTINGE_L(1950):	12	Festinger, L	Informal social communication	PSYCHO REV
LEWIN_K(1951):	12	Lewin, K	Field theory in social science	book
HARARY_F(1953)2:143	12	Harary, F	On the notion of balance of a signed graph	MICH MATH J
CARTWRIG_D(1956)63:277	123	Cartwright, D	Structural balance - a generalization of heider theory	PSYCHOL REV
HUBBELL_C(1965)28:377	3	Hubbell, CH	An input-output approach to clique identification	SOCIOMETRY
DAVIS_J(1967)20:181	123	Davis, JA	Clustering and structural balance in graphs	HUM RELAT
BOYD_J(1969)6:139	3	Boyd, JP	Algebra of group kinship	J MATH PSYCHOL
HARTIGAN_J(1972)67:123	3	Hartigan, JA	Direct clustering of a data matrix	J AM STAT ASSOC
ALBA_R(1973)3:113	123	Alba, RD	Graph-theoretic definition of a sociometric clique	J MATH SOCIOL
GRANOVET_M(1973)78:1360	23	Granovet.MS	The strength of weak ties	AM J SOCIOL
BREIGER_R(1974)53:181	123	Breiger, RL	Duality of persons and groups	SOC FORCES
BREIGER_R(1975)12:328	123	Breiger, RL	Algorithm for clustering relational data with applications to social network analysis and comparison with multidimensional-scaling	J MATH PSYCHOL
WHITE_H(1976)81:730	123	White, HC	Social-structure from multiple networks .1. Blockmodels of roles and positions	AM J SOCIOL
BOORMAN_S(1976)81:1384	123	Boorman, SA	Social-structure from multiple networks .2. Role structures	AM J SOCIOL
BURT_R(1976)55:93	3	Burt, RS	Positions in networks	SOC FORCES
BURT_R(1977)56:106	3	Burt, RS	Positions in multiple network systems .1. General conception of stratification and prestige in a system of actors cast as a social topology	SOC FORCES
BURT_R(1977)56:551	3	Burt, RS	Positions in multiple network systems .2. Stratification and prestige among elite decision-makers in community of Altneustadt	SOC FORCES
ARABIE_P(1978)17:21	123	Arabie, P	Constructing blockmodels - how and why	J MATH PSYCHOL
SAILER_L(1978)1:73	3	Sailer, LD	Structural equivalence - meaning and definition, computation and application	SOC NETWORKS
BURT_R(1978)7:189	23	Burt, RS	Cohesion versus structural equivalence as a basis for network subgroups	SOCIOL METHOD RES
BREIGER_R(1978)7:213	123	Breiger, RL	Joint role structure of 2 communities elites	SOCIOL METHOD RES
SNYDER_D(1979)84:1096	3	Snyder, D	Structural position in the world system and economic-growth, 1955-1970 - multiple-network analysis of transnational interactions	AM J SOCIOL
BREIGER_R(1979)13:21	3	Breiger, RL	Toward an operational theory of community elite structures	QUAL QUANT
BREIGER_R(1979)42:262	3	Breiger, RL	Personae and social roles - network structure of personality-types in small-groups	SOC PSYCHOL
BURT_R(1980)6:79	123	Burt, RS	Models of network structure	ANNU REV SOCIOL
MCCONAGH_M(1981)9:267	23	Mcconaghy, MJ	The common role structure - improved block-modeling methods applied to 2 communities elites	SOCIOL METHOD RES
PATTISON_P(1981)9:286	23	Pattison, PE	A reply to Mcconaghy - equating the joint reduction with block-model common role structures	SOCIOL METHOD RES
BURT_R(1982)16:109	23	Burt, RS	Testing a structural model of perception - conformity and deviance with respect to journal norms in elite sociological methodology	QUAL QUANT

Table 2.9: List of works on CPM path (1), main paths (2) and island (3) – part 2

label	code	first author	title	journal
PATTISON_P(1982)25:51	23	Pattison, PE	A factorization procedure for finite-algebras	J MATH PSYCHOL
PATTISON_P(1982)25:87	23	Pattison, PE	The analysis of semigroups of multirelational systems	J MATH PSYCHOL
MANDEL_M(1983)48:376	3	Mandel, MJ	Local roles and social networks	AM SOCIOL REV
WHITE_D(1983)5:193	23	White, DR	Graph and semigroup homomorphisms on networks of relations	SOC NETWORKS
FRIEDKIN_N(1984)12:235	23	Friedkin, NE	Structural cohesion and equivalence explanations of social homogeneity	SOCIOL METHOD RES
DOREIAN_P(1985)36:28	3	Doreian, P	Structural equivalence in a journal network	J AM SOC INFORM SCI
FAUST_K(1985)7:77	123	Faust, K	Does structure find structure - a critique of Burt use of distance as a measure of structural equivalence	SOC NETWORKS
FIENBERG_S(1985)80:51	3	Fienberg, SE	Statistical-analysis of multiple sociometric relations	J AM STAT ASSOC
BREIGER_R(1986)8:215	123	Breiger, RL	Cumulated social roles - the duality of persons and their algebras	SOC NETWORKS
BURT_R(1987)92:1287	23	Burt, RS	Social contagion and innovation - cohesion versus structural equivalence	AM J SOCIOL
FAUST_K(1988)10:313	123	Faust, K	Comparison of methods for positional analysis - structural and general equivalences	SOC NETWORKS
DOREIAN_P(1988)13:243	23	Doreian, P	Equivalence in a social network	J MATH SOCIOL
PATTISON_P(1988)10:383	23	Pattison, PE	Network models - some comments on papers in this special issue	SOC NETWORKS
WINSHIP_C(1988)10:209	3	Winship, C	Thoughts about roles and relations - an old document revisited	SOC NETWORKS
BORGATTI_S(1989)11:65	123	Borgatti, SP	The class of all regular equivalences - algebraic structure and computation	SOC NETWORKS
IACOBUCC_D(1990)55:707	3	Iacobucci, D	Social networks with 2 sets of actors	PSYCHOMETRIKA
BURT_R(1990)12:83	23	Burt, Rs	Detecting role equivalence	SOC NETWORKS
BATAGELJ_V(1992)14:63	123	Batagelj, V	Direct and indirect methods for structural equivalence	SOC NETWORKS
BATAGELJ_V(1992)14:121	23	Batagelj, V	An optimizational approach to regular equivalence	SOC NETWORKS
ANDERSON_C(1992)14:137	3	Anderson, CJ	Building stochastic blockmodels	SOC NETWORKS
FAUST_K(1992)14:5	123	Faust, K	Blockmodels - interpretation and evaluation	SOC NETWORKS
DOREIAN_P(1994)19:1	123	Doreian, P	Partitioning networks based on generalized concepts of equivalence	J MATH SOCIOL
DOREIAN_P(1996)18:149	123	Doreian, P	A partitioning approach to structural balance	SOC NETWORKS
BATAGELJ_V(1997)19:143	123	Batagelj, V	Notes on blockmodeling	SOC NETWORKS
BATAGELJ_V(1999)1731:90	123	Batagelj, V	Partitioning approach to visualization of large graphs	LECT NOTES COMPUT SC
BATAGELJ_V(2000)22:173	123	Batagelj, V	Some analyses of Erdos collaboration graph	SOC NETWORKS
NEWMAN_M(2001)64:016131	123	Newman, MEJ	Scientific collaboration networks. I. Network construction and fundamental results	PHYS REV E
NEWMAN_M(2001)64:16132	23	Newman, MEJ	Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality	PHYS REV E
GIRVAN_M(2002)99:7821	123	Girvan, M	Community structure in social and biological networks	P NATL ACAD SCI USA
RAVASZ_E(2002)297:1551	123	Ravasz, E	Hierarchical organization of modularity in metabolic networks	SCIENCE
BARABASI_A(2002)311:590	23	Barabasi, AL	Evolution of the social network of scientific collaborations	PHYSICA A
RAVASZ_E(2003)67:026112	123	Ravasz, E	Hierarchical organization in complex networks	PHYS REV E

Table 2.10: List of works on CPM path (1), main paths (2) and island (3) – part 3

label	code	first author	title	journal
NEWMAN_M(2003)45:167	123	Newman, MEJ	The structure and function of complex networks	SIAM REV
NEWMAN_M(2003)68:026121	123	Newman, MEJ	Properties of highly clustered networks	PHYS REV E
RIVES_A(2003)100:1128	3	Rives, AW	Modular organization of cellular networks	P NATL ACAD SCI USA
GUIMERA_R(2003)68:065103	23	Guimera, R	Self-similar community structure in a network of human interactions	PHYS REV E
HOLME_P(2003)19:532	3	Holme, P	Subnetwork hierarchies of biochemical pathways	BIOINFORMATICS
NEWMAN_M(2003)68:036122	123	Newman, MEJ	Why social networks are different from other types of networks	PHYS REV E
GLEISER_P(2003)6:565	23	Gleiser, PM	Community structure in jazz	ADV COMPLEX SYST
NEWMAN_M(2004)69:026113	23	Newman, MEJ	Finding and evaluating community structure in networks	PHYS REV E
NEWMAN_M(2004)38:321	23	Newman, MEJ	Detecting community structure in networks	EUR PHYS J B
REICHARD_J(2004)93:218701	23	Reichardt, J	Detecting fuzzy community structures in complex networks with a Potts model	PHYS REV LETT
ARENAS_A(2004)38:373	3	Arenas, A	Community analysis in social networks	EUR PHYS J B
CLAUSET_A(2004)70:066111	23	Clauset, A	Finding community structure in very large networks	PHYS REV E
RADICCHI_F(2004)101:2658	123	Radicchi, F	Defining and identifying communities in networks	P NATL ACAD SCI USA
DONETTI_L(2004):P10012	123	Donetti, L	Detecting network communities: a new systematic and efficient algorithm	J STAT MECH
GUIMERA_R(2004)70:025101	23	Guimera, R	Modularity from fluctuations in random graphs and complex networks	PHYS REV E
GUIMERA_R(2005)433:895	123	Guimera, R	Functional cartography of complex metabolic networks	NATURE
DUCH_J(2005)72:027104	123	Duch, J	Community detection in complex networks using extremal optimization	PHYS REV E
DANON_L(2005):	123	Danon, L	COSIN book	–
PALLA_G(2005)435:814	3	Palla, G	Uncovering the overlapping community structure of complex networks in nature and society	NATURE
MUFF_S(2005)72:056107	23	Muff, S	Local modularity measure for network clusterizations	PHYS REV E
GFELLER_D(2005)72:056135	23	Gfeller, D	Finding instabilities in the community structure of complex networks	PHYS REV E
GUIMERA_R(2005)102:7794	3	Guimera, R	The worldwide air transportation network: Anomalous centrality, community structure, and cities’ global roles	P NATL ACAD SCI USA
NEWMAN_M(2006)103:8577	3	Newman, MEJ	Modularity and community structure in networks	P NATL ACAD SCI USA
REICHARD_J(2006)74:016110	123	Reichardt, J	Statistical mechanics of community detection	PHYS REV E
BOCCALET_S(2006)424:175	3	Boccaletti, S	Complex networks: Structure and dynamics	PHYS REP
NEWMAN_M(2006)74:036104	23	Newman, MEJ	Finding community structure in networks using the eigenvectors of matrices	PHYS REV E
FORTUNAT_S(2007)104:36	123	Fortunato, S	Resolution limit in community detection	P NATL ACAD SCI USA
KUMPULA_J(2007)56:41	123	Kumpula, JM	Limited resolution in complex network community detection with Potts model approach	EUR PHYS J B
KUMPULA_J(2007)7:L209	23	Kumpula, JM	Limited resolution and multiresolution methods in complex network community detection	FLUCT NOISE LETT
GUIMERA_R(2007)3:63	3	Guimera, R	Classes of complex networks defined by role-to-role connectivity profiles	NAT PHYS
ROSVALL_M(2007)104:7327	3	Rosvall, M	An information-theoretic framework for resolving community structure in complex networks	P NATL ACAD SCI USA
GUIMERA_R(2007)76:036102	23	Guimera, R	Module identification in bipartite and directed networks	PHYS REV E

Table 2.11: List of works on CPM path (1), main paths (2) and island (3) – part 4

label	code	first author	title	journal
SALES-PA_M(2007)104:15224	123	Sales-Pardo, M	Extracting the hierarchical organization of complex systems	P NATL ACAD SCI USA
ARENAS_A(2008)10:053039	123	Arenas, A	Analysis of the structure of complex networks at different resolution levels	NEW J PHYS
CLAUSET_A(2008)453:98	3	Clauset, A	Hierarchical structure and the prediction of missing links in networks	NATURE
KUMPULA_J(2008)78:026109	3	Kumpula, JM	Sequential algorithm for fast clique percolation	PHYS REV E
KARRER_B(2008)77:046119	23	Karrer, B	Robustness of community structure in networks	PHYS REV E
BLONDEL_V(2008):P10008	23	Blondel, VD	Fast unfolding of communities in large networks	J STAT MECH-THEORY E
LEUNG_I(2009)79:066107	3	Leung, IXY	Towards real-time community detection in large networks	PHYS REV E
LANCICHI_A(2009)11:033015	123	Lancichinetti, A	Detecting the overlapping and hierarchical community structure of complex networks	NEW J PHYS
LANCICHI_A(2009)80:016118	23	Lancichinetti, A	Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities	PHYS REV E
RONHOVDE_P(2009)80:016109	23	Ronhovde, P	Multiresolution community detection for megascale networks by information-based replica correlations	PHYS REV E
GOMEZ_S(2009)80:016114	3	Gomez, S	Analysis of community structure in networks of correlated data	PHYS REV E
TRAAG_V(2009)80:036115	23	Traag, VA	Community detection in networks with positive and negative links	PHYS REV E
NICOSIA_V(2009):P03024	123	Nicosia, V	Extending the definition of modularity to directed graphs with overlapping communities	J STAT MECH
EVANS_T(2009)80:016105	123	Evans, TS	Line graphs, link partitions, and overlapping communities	PHYS REV E
LANCICHI_A(2009)80:056117	3	Lancichinetti, A	Community detection algorithms: A comparative analysis	PHYS REV E
BARBER_M(2009)80:026129	3	Barber, MJ	Detecting network communities by propagating labels under constraints	PHYS REV E
FORTUNAT_S(2010)486:75	123	Fortunato, S	Community detection in graphs	PHYS REP
GOOD_B(2010)81:046106	123	Good, BH	Performance of modularity maximization in practical contexts	PHYS REV E
LANCICHI_A(2010)81:046110	3	Lancichinetti, A	Statistical significance of communities in networks	PHYS REV E
RADICCHI_F(2010)82:026102	23	Radicchi, F	Combinatorial approach to modularity	PHYS REV E
LANCICHI_A(2010)5:0011976	3	Lancichinetti, A	Characterizing the Community Structure of Complex Networks	PLOS ONE
AHN_Y(2010)466:761	23	Ahn, YY	Link communities reveal multiscale complexity in networks	NATURE
EVANS_T(2010)77:265	3	Evans, TS	Line graphs of weighted networks for overlapping communities	EUR PHYS J B
MUCHA_P(2010)328:876	3	Mucha, PJ	Community Structure in Time-Dependent, Multiscale, and Multiplex Networks	SCIENCE
GREGORY_S(2010)12:103018	3	Gregory, S	Finding overlapping communities in networks by label propagation	NEW J PHYS
KARRER_B(2011)83:016107	23	Karrer, B	Stochastic blockmodels and community structure in networks	PHYS REV E
EXPERT_P(2011)108:7663	23	Expert, P	Uncovering space-independent communities in spatial networks	P NATL ACAD SCI USA
PSORAKIS_I(2011)83:066114	23	Psorakis, I	Overlapping community detection using Bayesian non-negative matrix factorization	PHYS REV E
TRAAG_V(2011)84:016114	23	Traag, VA	Narrow scope for resolution-limit-free community detection	PHYS REV E
DECELLE_A(2011)107:065701	3	Decelle, A	Inference and Phase Transitions in the Detection of Modules in Sparse Networks	PHYS REV LETT
BALL_B(2011)84:036103	123	Ball, B	Efficient and principled method for detecting communities in networks	PHYS REV E
DECELLE_A(2011)84:066106	123	Decelle, A	Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications	PHYS REV E

Table 2.12: List of works on CPM path (1), main paths (2) and island (3) – part 5

label	code	first author	title	journal
LANCICHI_A(2011)6:0018961	23	Lancichinetti, A	Finding Statistically Significant Communities in Networks	PLOS ONE
LANCICHI_A(2011)84:066122	23	Lancichinetti, A	Limits of modularity maximization in community detection	PHYS REV E
NADAKUDI_R(2012)108:188701	3	Nadakuditi, RR	Graph Spectra and the Detectability of Community Structure in Networks	PHYS REV LETT
PEIXOTO_T(2012)85:056122	123	Peixoto, TP	Entropy of stochastic blockmodel ensembles	PHYS REV E
PEIXOTO_T(2013)110:148701	123	Peixoto, TP	Parsimonious Module Inference in Large Networks	PHYS REV LETT
GOPALAN_P(2013)110:14534	3	Gopalan, PK	Efficient discovery of overlapping communities in massive networks	P NATL ACAD SCI USA
PEIXOTO_T(2014)89:012804	123	Peixoto, TP	Efficient Monte Carlo and greedy heuristic for the inference of stochastic block models	PHYS REV E
PEIXOTO_T(2014)4:011047	123	Peixoto, TP	Hierarchical Block Structures and High-Resolution Model Selection in Large Networks	PHYS REV X
LARREMOR_D(2014)90:012805	123	Larremore, DB	Efficiently inferring community structure in bipartite networks	PHYS REV E
ZHANG_P(2014)111:18144	23	Zhang, P	Scalable detection of statistically significant communities and hierarchies, using message passing for modularity	P NATL ACAD SCI USA
KAWAMOTO_T(2015)91:012809	3	Kawamoto, T	Estimating the resolution limit of the map equation in community detection	PHYS REV E
ZHANG_X(2015)91:032803	3	Zhang, X	Identification of core-periphery structure in networks	PHYS REV E
PEIXOTO_T(2015)5:011033	123	Peixoto, TP	Model Selection and Hypothesis Testing for Large-Scale Network Models with Overlapping Groups	PHYS REV X
JIANG_J(2015)91:062805	3	Jiang, JQ	Stochastic block model and exploratory analysis in signed networks	PHYS REV E
PEIXOTO_T(2015)92:042807	23	Peixoto, TP	Inferring the mesoscale structure of layered, edge-valued, and time-varying networks	PHYS REV E
PEROTTI_J(2015)92:062825	23	Perotti, JI	Hierarchical mutual information for the comparison of hierarchical community structures in complex networks	PHYS REV E
ZHANG_P(2016)93:012303	3	Zhang, P	Community detection in networks with unequal groups	PHYS REV E
VALLES-C_T(2016)6:011036	3	Valles-Catala, T	Multilayer Stochastic Block Models Reveal the Multilayer Structure of Complex Networks	PHYS REV X
NEWMAN_M(2016)117:078301	23	Newman, MEJ	Estimating the Number of Communities in a Network	PHYS REV LETT
HRIC_D(2016)6:031038	123	Hric, D	Network Structure, Metadata, and the Prediction of Missing Nodes and Annotations	PHYS REV X
FORTUNAT_S(2016)659:1	123	Fortunato, S	Community detection in networks: A user guide	PHYS REP
NEWMAN_M(2016)94:052315	123	Newman, MEJ	Equivalence between modularity optimization and maximum likelihood methods for community detection	PHYS REV E
FISTER_I(2016)4:00049	123	Fister, I	Toward the Discovery of Citation Cartels in Citation Networks	FRONT PHYS
YANG_L(2016)9:3390/a9040073	123	Yang, LJ	Community Structure Detection for Directed Networks through Modularity Optimisation	ALGORITHMS
PEIXOTO_T(2017)95:012317	123	Peixoto, TP	Nonparametric Bayesian inference of the microcanonical stochastic block model	PHYS REV E

more of a method for partitioning social networks with a migration to a newer journal and ii) the interest of sociologists in this research area diminished.

A similar pattern can be discerned for the subsequent community detection literature. The shift from 'social' social networks to community detection development is marked by the appearance of works produced by Mark Newman and Michelle Girvan in *Physics Review E* and *The Proceedings of the National Academy of Sciences* (the sixth, fifth and fourth items from the bottom in Table 2.9). In terms of a frequent venue, *Physics Review E* dominates with 44 works related to clustering appearing in its pages. It seems that *Physics Review E* plays the same 'venue role' as *Social Networks* regarding clustering. However, it does so to a far larger research community having many more scientists and more journals (which are also larger in size).

There is a contrast between the lists of journals in Table 2.6 and those listed in Tables 2.8–2.12. In the left panel of Table 2.6, the high ranking journals were *The Proceedings of the National Academy of Sciences*, *Nature*, *Lecture Notes in Computer Science*, *Science*, *Physica A*, *Physics Review E*, and *Physics Review Letters*. In the right panel, the top four journals are *Lecture Notes in Computer Science*, *Physics Review E*, *Physica A* and *Social Networks*. In the main, the journals heading the lists in Table 2.6 largely vanish from the list in Tables 2.8–2.12. *Lecture Notes in Computer Science* does not appear, *Physica A* appears only once, *Science* twice, *Nature* thrice and *Physics Review Letters* four times. Only *Proceedings of the National Academy of Sciences* and *Physics Review E* have works appearing with any regularity. It seems that community detection has a relatively narrow focus within the wider natural sciences literature – just as blockmodeling did in the earlier sociological literature. For researchers interested in the substantive meanings associated with partitioning with this literature, this raises interesting questions that are answered, in part, by examining islands in this literature.

2.3.4 Positioning sets of selected works in a citation network

The original main path analyses produced figures like the one shown in Figure 2.2 with a single main path. A recent extension of this approach enables a researcher to determine main paths through a selected set of nodes (works) in a citation network. This can be used to position a given set of nodes in a citation network – they can, either attach to the principal main path, or form separate streams. This is illustrated with three examples involving valued networks, signed networks, and a geophysics network.

The basic idea is to select a set of works on specific topic. For the valued networks example, we focused on works authored by Zziberna@Žiberna and by Nordlund extending blockmodeling for binary networks to valued networks. For the signed networks we selected papers by Doreian and Mrvar who have written extensively on this topic. For the geophysics network we used works selected from the network discussed in Subsection 2.4.3. The new option determines, for each work from the set of given works, the corresponding main path passing through it. There are two possible outcomes. One is that the intersection of the principal main path and the obtained main path is non-empty. If so, then the selected work is related to those in the principal main path. This allows 'branches' having ties to or from works in the principal main path. The second option is that the intersection is empty implying that the selected work is focused on different issues.

The main path of Figure 2.2 is present in both Figures 2.4 and 2.5 which differ only by considering separately valued and signed networks. We consider first works on valued networks, { ZIBERNA_A(2007)29:105, NORDLUND_C(2007)29:59, ZIBERNA_A (2008)32:57, ZIBERNA_A(2009)6:99, ZIBERNA_A(2013):, ZIBERNA_A

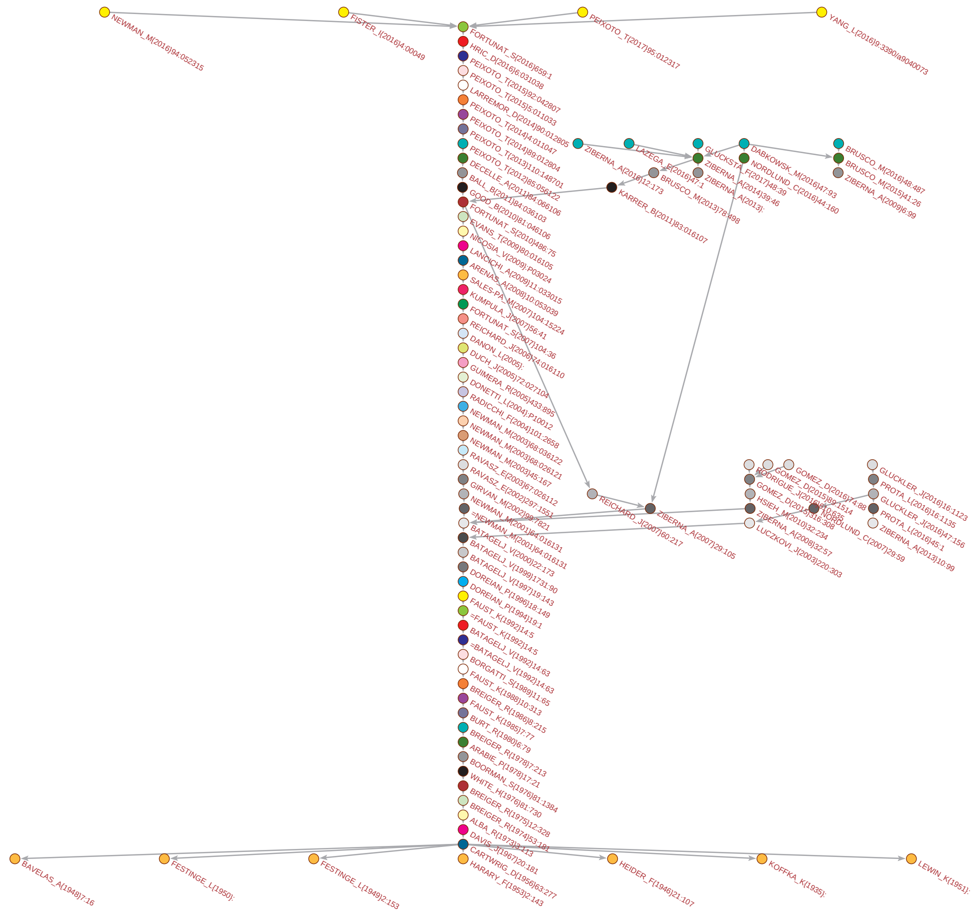

Figure 2.4: Valued networks Main Path with Branches

(2013)10:99, ZIBERNA_A(2014)39:46, ZIBERNA_A(2016)12:137, NORDLUND_C(2016)44:160 }, as shown in Figure 2.4.

All the foundational papers are cited by a paper by Cartwright and Harary published in 1956. The next 22 works are all well known in the **social** network analysis with many in the blockmodeling tradition before a branch appears. It is a pair of branches with two ties to works by Batagelj regarding valued networks. One comes from a work of his Ljubljana colleague, Zziberna@Žiberna, and the other from a work involving Luczkovich. The former has a small main path with works considering valued networks. The second branch includes productions that appeared in a special issue of the Journal of Economic Geography focused on social networks distributed in geographic space. Blockmodeling ideas were mobilized in this special issue. The second branch, actually, a double branch, involves the survey paper of Fortunato on community detection. There is a citation from that work to a work that cites one of works authored by Zziberna@Žiberna in one of the branches mentioned above. The other takes the form of a citation to the Fortunato work. The works on this branch involve works authored by social scientists regarding valued net-

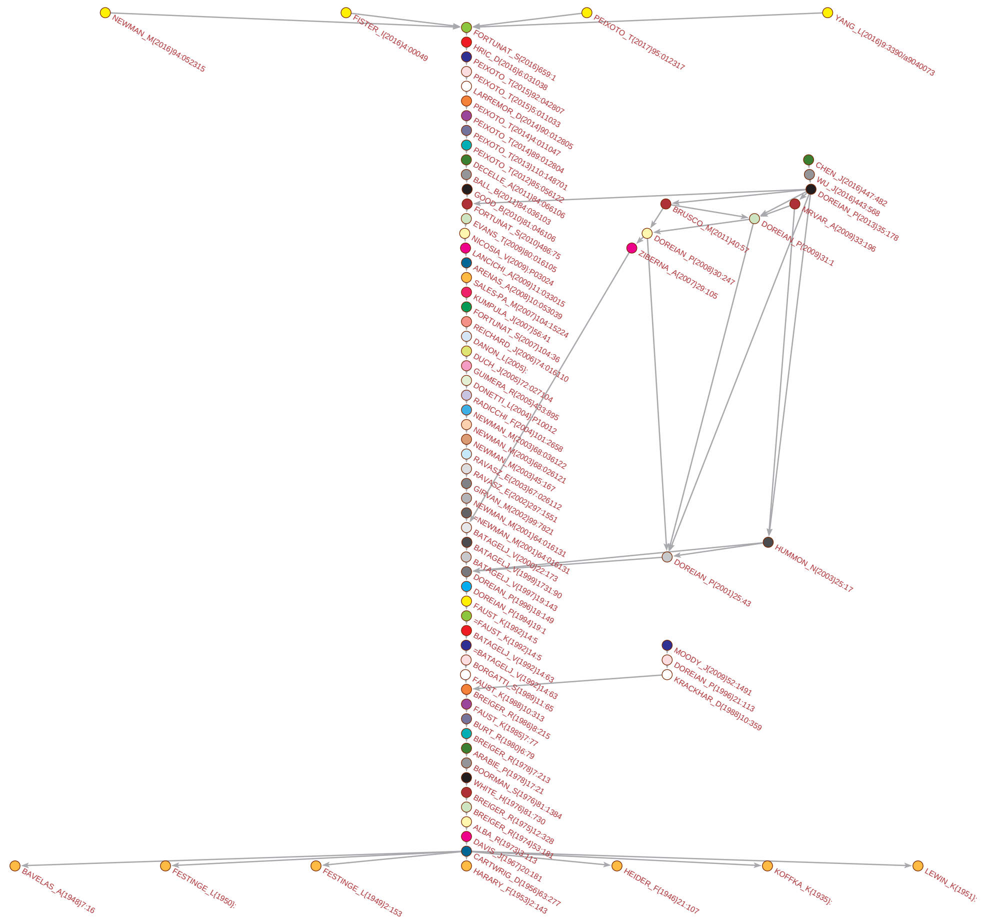

Figure 2.5: Signed network Main Path with Branches

works and algorithmic methods for partitioning networks. There is also a citation from a work involving Nordlund, a coauthor with Žiberna to another Žiberna production in the earlier branch.

We turn to consider the works on signed network, { DOREIAN_P(1996)18:149, DOREIAN_P(1996)21:113, DOREIAN_P(2001)25:43, HUMMON_N(2003)25:17, DOREIAN_P(2009)31:1, MRVAR_A(2009)33:196, BRUSCO_M(2011)40:57, DOREIAN_P(2013)35:178 }, as shown in Figure 2.5. The first branch appears with a citation to a 1986 paper authored by Breiger, firmly in the blockmodeling tradition, by Krackhardt. It appears far earlier than the first branches of Figure 2.5. This paper was cited by another production involving both Doreian and Krackhardt which in turn was cited by a paper involving Moody, a rising scholar in the social networks field. While this paper has been cited frequently, its content was not defined as being primarily within the domain defined by signed networks. This has been captured in Figure 2.5.

The next branch off the main path involves a book edited by Doreian and Stokman on the evolution of social networks. Ideas expressed there were picked up by Doreian

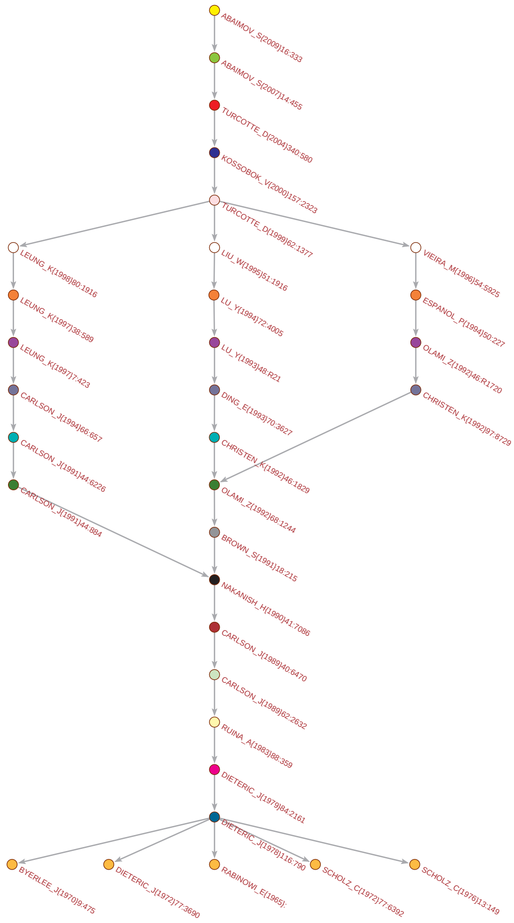

Figure 2.6: GeoPhysics Main Path with Branches

and Krackhardt in a 2001 work, which provided compelling evidence against the widely accepted idea of signed networks tending towards balance. This was reinforced by a 2003 production by Hummon and Doreian.

Another branch of the main path is due to a citation from Zziberna@Žiberna to Batagelj, colleagues at the University of Ljubljana. This is connected through citations from productions involving Doreian, Mrvar and Brusco on fitting signed blockmodels. There are ties from this group citing the 2001 and 2003 productions mentioned in the previous paragraph. There is also a tie from a work on partitioning signed 2-mode into a work involving Fortunato late in the main path. It forms the last branch of Figure 2.5.

Note that in both examples a part of the principal main path is also a part of the main path through each work from the selected set of works. These parts combine in both cases in the complete principal main path.

The foregoing two examples of ‘attaching’ branches to a main path, were triggered by considering variants of networks studied in the blockmodeling literature. Here, we examine a completely different example. In Subsection 2.4.3, we examine link islands in the network partitioning literature. We selected some of the productions in this link island, { `DIETERIC_J(1979)84:2161`, `CARLSON_J(1989)40:6470`, `CARLSON_J(1991)44:6226`, `OLAMI_Z(1992)46:R1720`, `TURCOTTE_D(1999)62:1377` }, and repeated the above analyses. The results are shown in Figure 2.6, a simpler figure than for the two previous examples.

For this example, we start our discussion at the top of the main path. A 1999 work, one involving Turcotte, a very prominent contributor to this literature, cites two works off the 'main' path. Each citation leads to a smaller and distinct main path. Both smaller main paths link back to the 'main' path, albeit at different places. The one on the right of is a 1992 production that cites a production of the same year. The one on the left sent a tie to a production published in 1990. At the bottom of Figure 2.6, there is a publication on the main path that cites earlier foundational works for the geophysical literature. This is a common pattern for identified main paths.

2.4 Link islands in the clustering network literature

A *link island* is a connected subnetwork having a higher internal cohesion than on the links to its neighbors. Identifying islands is a general and efficient approach for identifying locally 'important' subnetworks in a given network. The details for doing this were described in [7] (Chapter 2, Section 9). The method amounts to filtering networks to identify some manageable parts. In large networks, it is likely that many such islands will be identified. While islands are identified through the ties linking them, it is crucial to examine the substantive content of the islands. Just identifying topological features, while useful, is not enough. Islands are *coherent* with the coherence coming from substance and the kind of information contained in Tables 2.8-2.12.

Link islands were used extensively in [7] to examine the structure of scientific citation networks (Chapter 4, Section 7), US patent networks (Chapter 5, Section 6) and the US Supreme Court citation network (in Chapter 6, Section 2). We use the same tools here to examine the clustering network as defined above. General Pajek instructions for doing this are contained in [7].

Figure 2.7 shows the ten link islands having sizes in the range $[20, 150]$ identified in the network clustering literature. Adopting George Orwell's phrase (from *Animal Farm*) "All animals are equal, but some animals are more equal than others" we change it to "All islands are interesting, but some islands are more interesting than others". It seems that islands, labeled 10, 7, 9 and 2 have the most interest value. The other islands have much smaller maximal weights, smaller diameter and represent less important stories.

Island 10 is the largest of these islands having 150 works and a maximal weight 0.5785. It has two clear parts separated by a cut. Island 7 is next in size with 74 works having a much lower maximal weight 4.9611×10^{-18}. It also has two parts. The lower left part is centered on a single production while the upper right appears to be centered on a set of inter-linked works. Both parts are highly centralized. Island 9 has 44 works with a maximal weight 2.416×10^{-14}. Its structure suggests separate parts linked only through a cut. Island 2 has 33 nodes with a maximal weight 2.462×10^{-19}. Apart from the presence of pendants linked to the main part of the island, there are no obvious sub-parts.

The obvious question is simple: What holds these islands together in terms of substance and content? We turn to examine this next.

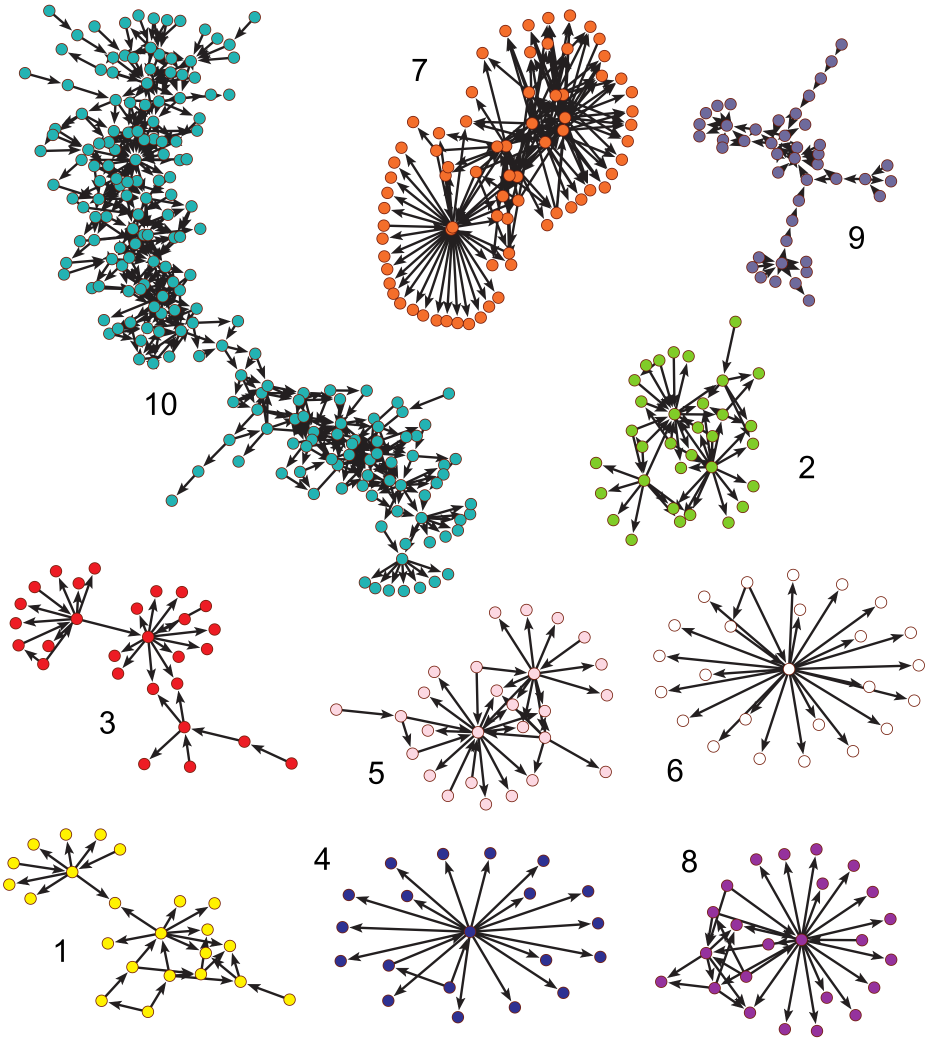

Figure 2.7: Ten identified islands in the clustering network literature

2.4.1 Island 10: Community Detection and Blockmodeling

Figure 2.8 shows Island 10 in more detail with its works identified. The upper left part is exclusively in the community detection domain while the lower right part contains works from the social network literature. The clear cut is the last node of the latter literature as identified in Figure 2.3. This link island provides a more expanded view compared to the one in Figure 2.3 which was more expanded than Figure 2.2. Given our earlier detailed consideration of the main path and key-routes, little more needs to be written at this point. The additional works in Island 10 do provide a foundation for a more detailed examination of the transition between two fields and what is featured in the two parts of the network clustering literature.

On the blockmodeling side, the first authors involved with the most works in Island 10 are Burt (9), Batagelj (7), Breiger (6), Doreian (5), Faust (4) and Pattison (4). Many are co-authored productions involving some of these researchers. For those involved in community detection analyses, the first authors involved with the most works in the island are Newman (13), Peixoto (8), Lancichinetti (7), Guimera (6), and Arenas (4). In terms of indegree, the three most cited items for the blockmodeling part of this island involve Arabie, Boorman, and Breiger. As noted above, all were involved in the foundational work for blockmodeling. Regarding community detection, the most cited researchers are Fortunato, Peixoto, and Newman. Their works appear to be either foundational or general surveys.

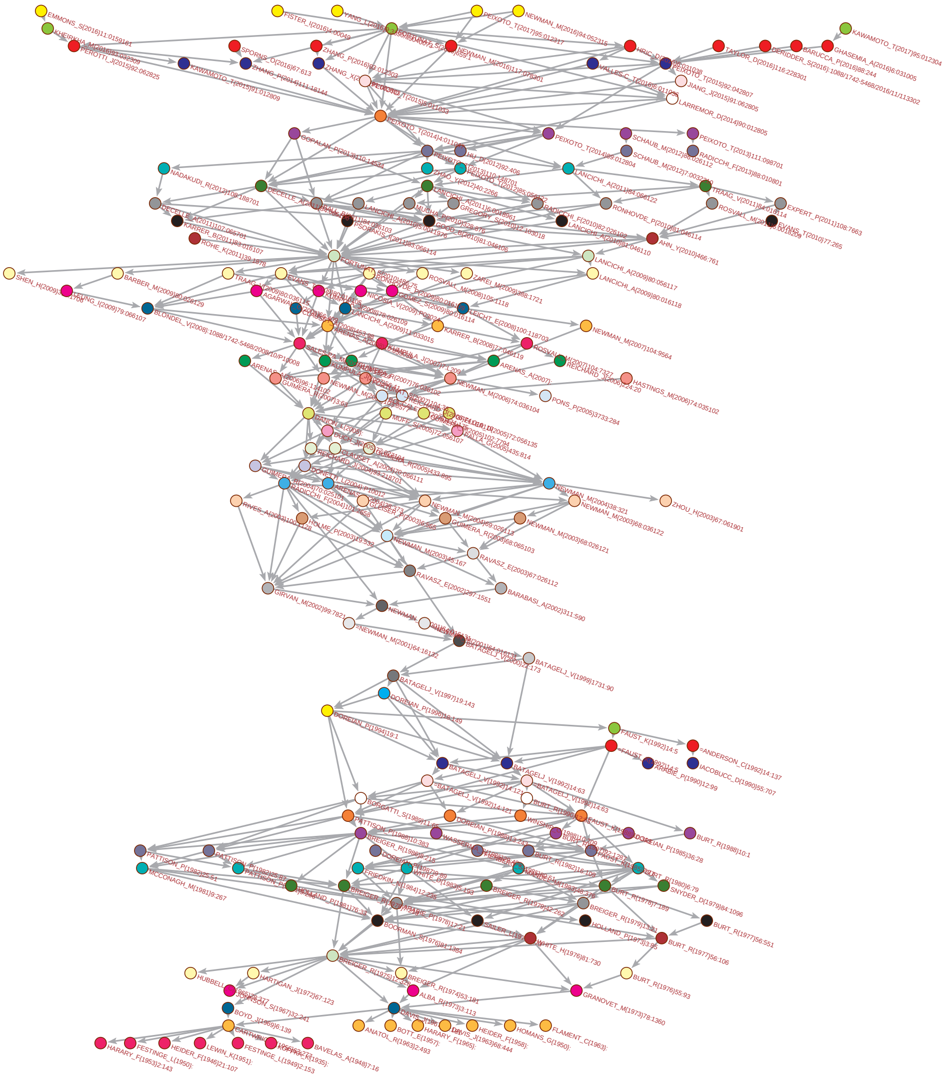

Figure 2.8: Island 10 Community Detection and Blockmodeling

2.4.2 Island 7: Engineering Geology

The publication years for works in this island span 1965 through 2017. Studying this island leads to a caution regarding the boundary problem. Its works are present in the citation network through the term ‘sliding block analysis’ used in this discipline. The earliest paper on the island appeared in *Geotechnique*. It contained a method for calculating the permanent displacements of soil slopes, embankments and dams during seismic events. The model was recognized as having great value in studying earthquakes. Two papers appeared in *Geotechnique* about a decade later proposing another method. It also is valuable, especially for analyzing earth slopes and earthen dams. Indeed, there are a variety of methods for studying seismic events related to earthquakes and landslides. The works on this island are focused on methods for measuring seismic activity and predicting their consequences.

Some works in this link island stand out in terms of the number of citations they receive and make. One paper by Jibson (of the US Geological Survey) appeared in 2007 in *Engineering Geology*. Its citations topped both the indegree and outdegree values. With colleagues, six other papers involving this scientist are in this island. Also high on the outdegree listing are papers involving Stamatopoulos. The more recent works have as the primary focus, as reflected by keywords, of predicting the dynamics of slip surfaces, saturated sands, slopes, and landslides. The methodological focus is clear also with both multi-block models and sliding-block models playing a central role.

The final paper in this island used a large database of recorded ground motions, to develop a predictive model of earth displacements. The empirical contexts for the body of work in this island are landslides and earthquakes linked through impact of seismic events. The major journals for this line of work include *Geotechnique*, *Journal of Geotechnical and Geoenvironmental Engineering*, *Bulletin of Earthquake Engineering*, *Soils and Foundations*, *Engineering Geology* and *Soil Dynamics and Earthquake Engineering*. It is clearly part of a broader field of Engineering Geology, the application of geological knowledge to improve the design of engineering projects, their construction as well as their maintenance and operation - including the impact of seismic events.

2.4.3 Island 9: Geophysics

Island 9 is shown in Figure 2.9. Its works are focused on earthquake modeling. This part of the literature is in the clustering citation network because of the term ‘spring-block model’. Again, this is another meaning of the term ‘block model’. The works in Island 9 are part of the Geophysics literature as evidenced by the journals where many of them appear. They include *Geophysical Research Letters*, *Physical Review Letters*, *Journal of Geophysics Research*, and *Physical Review A*.

One obvious question is simple to state: Why are Island 7 and Island 9 not joined? Seismic events and earthquakes are features in both of them. Surely, they must be linked? After a closer inspection of the works in these two islands, there is a very simple answer to our question. The works in Island 9 are especially focused on *temporal* and *dynamic* issues in contrast to the works in Island 7 which is entirely *static*. The difference between Island 9 and Island 7 reveals a profound similarity between two completely distinct scientific fields. It seems there is a real difference between static and dynamic approaches to studying empirical phenomena. Surprisingly, it is present in *both* the natural and social sciences. For far too long, social network analyses ignored temporal issues. One set of approaches to dealing with the evolution of social networks appeared in the edited collection [12]. Subsequent contributions appeared in special issues of *The Journal of Mathematical*

Sociology. Since then, a focus on dynamic models of social networks has become far more extensive. It remains to be seen if the static and dynamic approaches of the works studying seismic events of Islands 7 and 9 will be joined in geophysics and engineering geology.

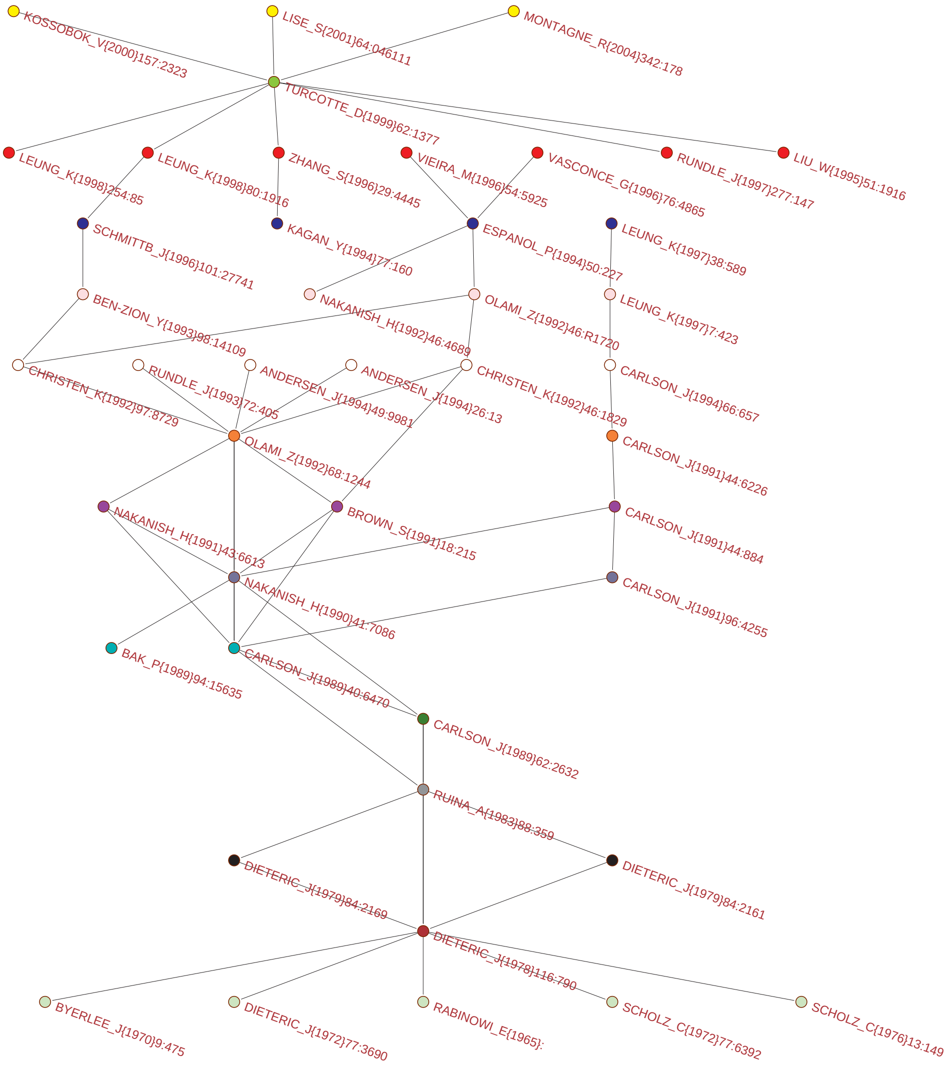

Figure 2.9: Island 9 Geophysics

2.4.4 Island 2: Electromagnetic fields and their impact on humans

The papers appearing in this island deal with numerical methods for computations relating to electromagnetic fields. Their appearance in this island is due to the term ‘block model’. In this different literature, ‘block models of people’ use a limited number of cubical cells to predict the internal electromagnetic fields and specific absorption rate distributions inside human bodies. The earliest paper on this island appeared in 1968. Other early papers involved Hagmann with his colleagues are present also. A production by Massoudi appeared in 1985 with the words ‘limitations of the cubical block model’ in the title. It has the highest indegree and the second highest outdegree. The main content concerns the meaning of a block model. One of the productions involving Hagmann (1986) has the highest outdegree. The Massoudi paper cites it, discussing the block model concept. Both papers appeared in *IEEE Transactions on Microwave Theory and Techniques*. Indeed, many of the works on this island appeared in this journal. They are strictly methodological.

A paper involving Zwamborn was published in 1991 (three papers with him as a coauthor appear on this island). It has the second highest outdegree and appeared in the *Journal of the Optical Society of America A*. It concerned the computation of electromagnetic fields inside strongly inhomogeneous objects. The most recent paper on this island, appearing in 2002, was published in *Microwave and Optical Technical Letters*. It concerned resonant frequency calculation for inhomogeneous dielectric resonators. These papers also are methodological. If human bodies are inhomogeneous objects, there is continuity of this empirical focus.

In the analyses of bibliometric networks reported in [7] (Section 4.7.3), there was an ‘optical network line island’. Many of the productions on this island involved journals published by IEEE (The Institute of Electrical and Electronics Engineers) and by the Optical Society of America. The same appears to be the case for Island 2. The analysis in [7] examined the role of the institutional dominance of large professional organizations, something that appears relevant here also, especially when their interests are coupled.

2.4.5 Limitations and extensions

Our brief examination of these four link islands implies some cautionary notes - along with suggestions for further work.

- One is that WoS is quite limited in the information it provides for individual works. Only half of the works on Island 2 had complete descriptions in WoS ($DC = 1$). This restriction is known already [7]. The problem was far less acute for the other islands. Clearly, different subject areas will have differing levels of this problem. These gaps in the information must be filled. One option is to extend the original WoS data with additional manually constructed descriptions for these works.

- The search terms used for extracting citation networks can be ambiguous. The search terms used here included block model*, blockmodel*, and block-model*. For those in the social networks sub-field, the term ‘blockmodel’ is very well known. But, for the works in Island 2, ‘block model’ means something quite different. The works in Islands 7 featured ‘sliding block analysis’ while in Island 9 the core term was ‘spring-block model’.

- Such differences in meaning for a search term can be discerned only through a careful examination of the identified literatures. Clearly, general terms have to be used to include as many potentially relevant works as possible. However, the results need to be considered carefully. We were surprised to learn of the other meanings for the term ‘block’. No doubt, researchers in geophysics and engineering geology would be surprised to find works from the social networks literature in their citation networks if searches were done using the term ‘block’. The proposed approach enabled us to identify these other meanings and consider the maximal weight of corresponding island, along with their importance.

Table 2.13: Authors involved in the largest number of works

Rank	Frequency	Author	Rank	Frequency	Author	Rank	Frequency	Author
1	66	ZHANG_X	15	35	ZHANG_Z	28	26	ZHANG_H
2	57	WANG_Y	16	35	ZHANG_J	29	26	WANG_L
3	56	LIU_J	17	34	JIAO_L	30	26	TURCOTTE_D
4	51	WANG_X	18	33	ZHANG_S	31	26	**BORGATTI_S**
5	44	LI_J	19	32	WANG_S	32	26	**EVERETT_M**
6	42	WANG_H	20	31	**BATAGELJ_V**	33	26	WANG_C
7	41	LIU_Y	21	31	CHEN_H	34	24	LI_X
8	41	LI_Y	22	29	YANG_J	35	24	LI_L
9	40	NEWMAN_M	23	28	HANCOCK_E	36	24	LIU_X
10	39	WANG_J	24	28	WANG_W	37	23	LI_S
11	39	**DOREIAN_P**	25	27	CHEN_L	38	23	ZHOU_Y
12	38	CHEN_Y	26	26	LI_H	39	23	CHEN_X
13	36	ZHANG_Y	27	26	WU_J	40	23	LEE_J
14	35	WANG_Z						

- Examining temporal shifts in the keywords used in literatures and the journals where works are published are important avenues of exploration for understanding the dynamics of scientific fields. The changes were most dramatic for the network clustering literature examined in Island 10.

- The structures of the islands shown in Figure 2.7 are quite different. An open problem is whether this has an impact on the production of knowledge and the social organization of scientific disciplines.

2.5 Authors

We consider, in more detail, the authors creating the papers in the network clustering literature. The network considered in this section is for works having complete information. We computed the networks **CiC**, **WAc**, **WKc**, and **WJc**. Their sizes are in Table 2.3.

The publication counts for authors are shown in Table 2.13 with a focus on the authors producing works in the core topic of this book. From **CiC**, it is straightforward to construct the counts of works by these authors. Authors with the largest number of papers about clustering networks are shown in Table 2.13. The large number of Chinese authors in the list may be an example of the "three Zhang, four Li" effect [28]. Lacking the resources to examine the relevant works to identify these authors, we proceed with a caution that some of the counts for these Chinese authors are not final.

The top 10 entries in Table 2.13 come from the community detection area. Only four of the (first) authors listed in Table 2.13 work in the social networks literature. Their names are bolded. As all four are involved in collaborative work, the counts by single author names is limited as a summary of individual activity. The remaining works come from researchers in other disciplines, most of whom study community detection. The same caveat regarding collaborative production holds there also. Even so, these counts of works reflect accurately the far larger number of researchers and productions from the natural sciences, consistent with our results about the main path, key-routes and Island 10.

We contend it may be more useful to examine productivity *inside* research groups and focus explicitly on collaboration. To this end, the idea of identifying cores in networks has value. A full description of k-cores and p-cores is provided in [7] (in Section 2.10.1). More importantly, for our purposes here, are P_S-cores also described in [7] (Section 4.10.1.3).

2.5.1 Productivity inside research groups

The network we use here is **Ct**, an undirected network obtained from $\mathbf{N}^T * \mathbf{N}$, where

$$\mathbf{N} = \text{diag}(\frac{1}{\max(1, \text{outdeg}(p))})\mathbf{WA}$$

with symmetrization [4].

A subset of nodes C is a P_S-core at some threshold iff for each of its nodes the sum of weights on links to other nodes from C is greater or equal to that threshold and C is the maximum such subset. Authors with the largest P_S-core values in **Ct** [6] are listed in Table 2.14. Again, bolding is used for researchers in the social network field. The number of researcher names from the social network side is now up to 14, still a minority. The values for authors equals the sum of all their fractional contributions to works with authors inside the core, a better measure than counts of publications bearing their names.

Figure 2.10 shows the links between author names with the size of vertices being proportional to their P_S-core value. For visual clarity, loops are removed. The names for researchers in the social network community are marked in boldface. The large top left P_S-core features researchers from the physical sciences with a clear central part. While Newman is connected to this P_S-core though one link, the size of his vertex is the largest. Several paths link other prominent researchers to this central part. They include one linking Peixoto to Fortunato to Lancichietti to Wang_J. There is also a path from Barabasi to Newman to Zhang_X. One surprise, at least for us, is the connection of Borgatti and Everett, having the strongest tie in Figure 2.11, to the central part of the P_S-core featuring natural scientists through their links with Boyd and his many links within this core. All three met, and worked, at the University of California at Irvine, an important center for social network analysis. This merits further attention.

Immediately to the right of this large P_S-core is a much smaller one involving Wasserman, Pattison and Breiger who worked with each other on role systems and helped create the foundations for exponential random graph modeling of networks. Below this P_S-core is one centered on Doreian who collaborated with all of the other researchers in this P_S-core. The links are strongest with Batagelj, Ferligoj and Mrvar. The next strongest tie is between him and Brusco – they worked on algorithms for blockmodeling along with Steinley. The strongest dyadic links in this core are between Batagelj and Ferligoj and between Brusco and Steinley.

We note two items: i) many of the author names in Table 2.14 involve researchers participating in collaborative work (see below for more on this); and ii) many of the names in this table have been mentioned in the above analyses, adding to the coherence of the results we report.

2.5.2 Collaboration

Collaboration is a critically important, and increasing, feature of modern scientific research. To examine this we use $\mathbf{Ct}'$, an undirected network without loops obtained from $\mathbf{N}^T * \mathbf{N}'$, where

$$\mathbf{N}' = \text{diag}(\frac{1}{\max(1, \text{outdeg}(p) - 1)})\mathbf{WA},$$

Table 2.14: Authors with the largest P_S-core values in **Ct**

Rank	P_S-core value	Author	Rank	P_S-core value	Author	Rank	P_S-core value	Author
1	21.0347	NEWMAN_M	15	6.0292	WANG_J	28	5.2589	**BRUSCO_M**
2	15.9653	**BORGATTI_S**	16	5.7500	PIZZUTI_C	29	5.2500	DIETERIC_J
3	15.9653	**EVERETT_M**	17	5.7014	STAMATOP_C	30	5.2483	LANCICHI_A
4	12.5000	**BURT_R**	18	5.6736	SUN_P	31	5.2483	FORTUNAT_S
5	12.5000	**DOREIAN_P**	19	5.6669	ZHANG_S	32	5.1111	**BOYD_J**
6	10.4722	PEIXOTO_T	20	5.6307	WANG_H	33	5.0633	WANG_X
7	10.1126	TURCOTTE_D	21	5.6307	LIU_J	34	5.0278	QIAN_X
8	8.7900	**FERLIGOJ_A**	22	5.5417	YANG_J	35	5.0208	**WASSERMA_S**
9	8.7900	**BATAGELJ_V**	23	5.5417	LESKOVEC_J	36	5.0000	OKAMOTO_H
10	6.5115	WANG_Y	24	5.5417	ZHANG_J	37	5.0000	JESSOP_A
11	6.4097	**PATTISON_P**	25	5.4432	HANCOCK_E	38	4.9881	BARABASI_A
12	6.4097	**BREIGER_R**	26	5.4432	ZHANG_Z	39	4.9775	**KRACKHAR_D**
13	6.2083	**MRVAR_A**	27	5.2589	**STEINLEY_D**	40	4.9112	ZHANG_H
14	6.0292	ZHANG_X						

through symmetrization and setting the diagonal values to 0 [10]. In the network **Ct** each work co-authored by an author contributes $\frac{1}{k^2}$ (k is the number of co-authors) to "self-collaboration" (value on the loop) of that author. The network $\mathbf{Ct}'$ describes the true collaboration with *others*.

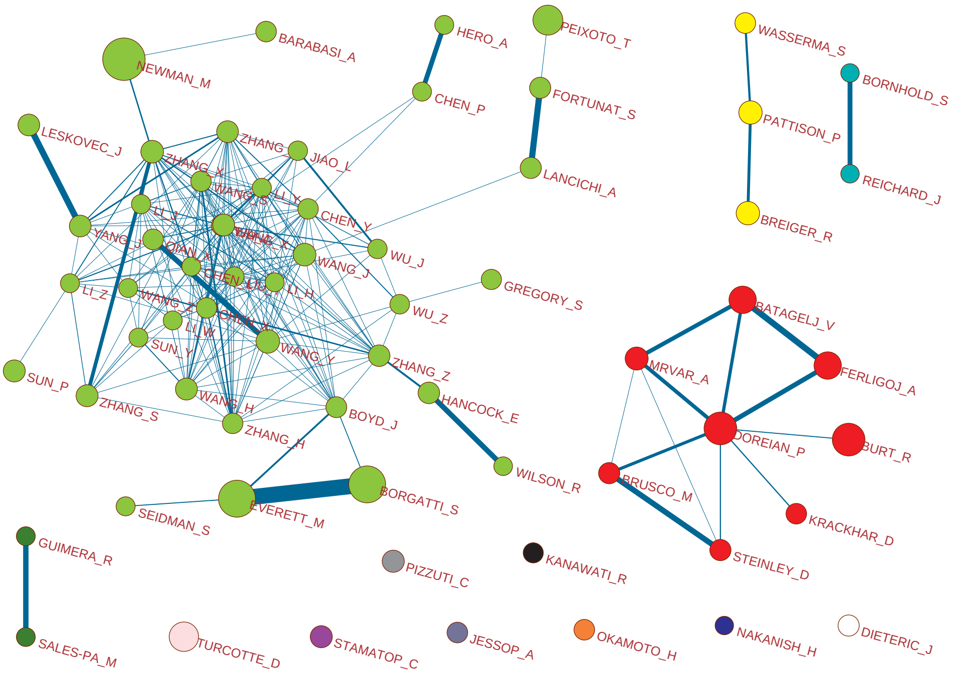

Figure 2.10: P_S-cores at level 4 in **Ct**

Table 2.15: Authors with the largest P_S-core values in $\mathbf{Ct}'$s

Rank	Value	Author	Rank	Value	Author
1	15.8333	**BORGATTI_S**	15	5.0000	AMELIO_A
1	15.8333	**EVERETT_M**	15	5.0000	BAJEC_M
2	7.6667	**FERLIGOJ_A**	15	5.0000	SUBELJ_L
2	7.6667	**BATAGELJ_V**	15	5.0000	CHEN_P
2	7.6667	**MRVAR_A**	15	5.0000	PIZZUTI_C
2	7.6667	**DOREIAN_P**	15	5.0000	REICHARD_J
7	6.4333	**STEINLEY_D**	15	5.0000	BORNHOLD_S
7	6.4333	**BRUSCO_M**	23	4.8333	SALES-PA_M
9	6.3333	YANG_J	23	4.8333	GUIMERA_R
9	6.3333	LESKOVEC_J	23	4.5833	NUSSINOV_Z
11	6.0000	LANCICHI_A	23	4.5833	RONHOVDE_P
11	6.0000	FORTUNAT_S	27	4.3333	ROSVALL_M
13	5.3333	QIAN_X	27	4.3333	BERGSTRO_C
13	5.3333	WANG_Y	27	4.3333	WILSON_R
15	5.0000	HERO_A	37	4.3333	HANCOCK_E

Authors with the largest P_S-core values in $\mathbf{Ct}'$ are listed in Table 2.15 and presented in Figure 2.11. Heading the list are Borgatti and Everett who have published together on blockmodeling for a long time. Next comes publications involving Ferligoj, Batagelj, Doreian, and Mrvar who also have worked together for an extensive period. Both Steinley and Brusco, who have collaborated with Doreian, appear next - but they also worked together on clustering problems before publishing papers with Doreian on blockmodeling. It is interesting that the leading ‘nonsocial’ authors from Table 2.14 Newman, Peixoto and Turcotte are missing in Table 2.15. The reasons are combination of publishing of single author papers and publishing with many different co-authors.

Similar analyses had been performed for social networks as a whole in [7]. The citation network studied there was far larger as a more extensive literature was studied. Many of the above names appear also in the tables and figures of [7]. Comparing the two sets of analyses makes it clear that the role of these authors in this literature largely, but not completely, involves blockmodeling.

There is always a choice regarding which links are included for further examination of the structure of any studied network. Figure 2.11 shows the network when the threshold was set at 3.5. Necessarily, the results are more fragmented with 18 smaller link islands. In the middle of Figure 2.11 is the heavy Borgatti-Everett dyad having the highest value. The top left link island also features authors working on blockmodeling, consistent with our earlier results. The remaining items belong to the community detection literature.

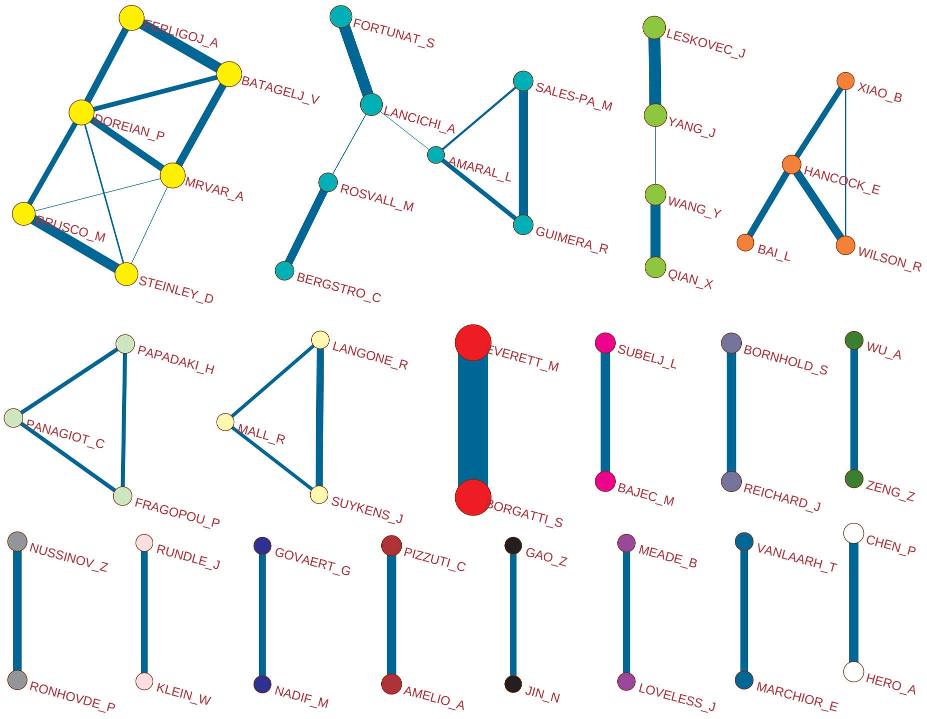

Figure 2.11: Links between authors in a P_S-core at level 3.5 in $\mathbf{Ct}'$

2.5.3 Citations among authors contributing the network partitioning literature

The network $\mathbf{Acite} = \mathbf{WAc}^T * \mathbf{CiC} * \mathbf{WAc}$ describes the citations among authors. The value of element $\mathbf{Acite}[u,v]$ is equal to the number of citations from works coauthored by u to works coauthored by v. While these numbers are inflated slightly when u and v collaborate, co-authorship is part of the citation structure. Collaboration matters greatly.

Link islands can be extracted from this network. The methods described in [7] require setting bounds for delineating islands. For this analysis they were [10,50] with 16 islands identified for this network. They have quite different structures. Each can be examined but we focus on two of them as they pertain to community detection and (non-stochastic) blockmodeling.

The community detection island shown at the bottom of Figure 2.12 is large and massively centered on Newman. By far, works involving him are cited the most. Without surprise to those in the field, a strong case can be made for him founding this research front both alone and with key collaborators. Fortunato is another highly cited author, most likely for his extensive and comprehensive summary of this research area. Note also that the terminal nodes (outdegree is 0) are the founders of complex networks approach Barabasi, Albert and Girvan.

The island contains publications about blockmodeling is smaller and is less centralized. The most central author is Doreian, but nowhere near to the extent of Newman. Moreover,

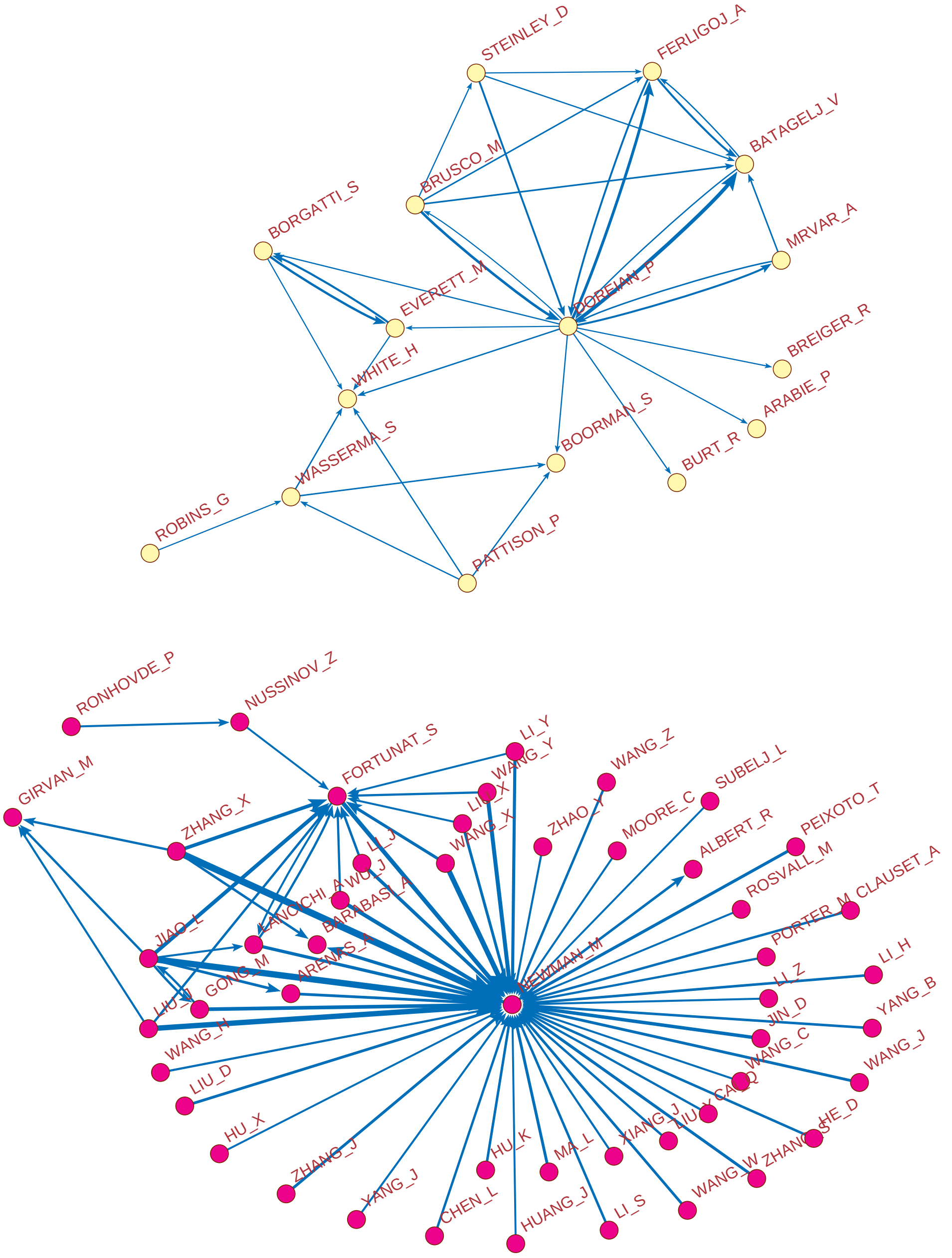

Figure 2.12: Citations among authors from two parts of the literature: Community detection and Blockmodeling

there is more nuance in the structure with citations going *from* him to authors involved in creating the foundations of blockmodeling. Also, three distinct collaborative efforts are involved. One features works featuring him with Batagelj and Ferligoj on blockmodeling. One is with Mrvar on signed networks and the third involves his work with Brusco and Steinley on algorithms for partitioning networks. Citations go also to Borgatti and Everett without any corresponding reciprocating citations. Citations from from Robins, Pattison and Wasserman, all prominent in social networks, are reminders that this island is about blockmodeling. Were the focus on probabilistic approaches to studying social networks, especially exponential random graph models, that part of the network would expand greatly with the blockmodeling part disappearing. An expanded analysis of the whole social network literature, albeit for an earlier time period is in [7] (their Figure 4.17) reinforces this point while showing links between these two areas of the literature.

No doubt, researchers more familiar with the community detection literature, could paint a more nuanced picture for their part of the network clustering literature. One feature of the islands technique is the way it determines items more closely related among themselves compared to the connections from them to elsewhere in the network. While useful, an open problem is the examination of links between such islands. When coupled to the use of keywords and placed in a temporal framework, this will facilitate an examination of the movements of ideas within and between parts of citation and collaboration networks.

We could analyze also networks $n(\mathbf{WAc})^T * \mathbf{CiC} * n(\mathbf{WAc})$ (every citation has value 1 that is distributed among authors) and $n(\mathbf{WAc})^T * n(\mathbf{CiC}) * n(\mathbf{WAc})$ (every work has value 1 that is distributed among authors).

2.5.4 Citations among journals

There is a huge literature on citation relations between journals. It origins are found in the work of Garfield starting in the 1950's. Among his many contributions were establishing the Institute for Scientific Information (ISI) and the creation of the *Science Citation Index* (SCI) making extensive use of the aggregated journal-to-journal citation data provided annually by the Journal Citation Reports (JCR). See, for example, [14]. Also created was the *Social Science Citation Index* (SSCI). Much work has followed on mapping to structure of these networks. A recent example is provided by Leydesdorff *et. al* examining structural shifts in journal-to-journal networks [21].

Our focus here has been on a sustained look at the citation network of works considering partitioning of networks. This can be extended to construct a journal-to-journal network for this literature only. Most likely, some of the works studied above will be found in the SCI. Others will be located in in the SSCI, with some overlap. Using only one of these data sources would be limited and combining them would be difficult. Our case is somewhat special because of our interest in citations in the field of network clustering and not in general citations among journals. The task is one of counting the citation links between journals featuring works in this area.

2.5.4.1 Counting To get information about citations among journals we compute the derived network

$$\mathbf{JJ} = \mathbf{WJ}^T * \mathbf{Ci} * \mathbf{WJ}$$

Its weights have the following meaning:
$jj(i,j) =$ # citations from papers published in journal i to papers published in journal j – with attention confined to the network partitioning literature.

While this network can be searched for link islands, the results are limited due the different sizes of the journals involved. To obtain more useful results we applied the fractional approach described immediately below. Note that $n(\mathbf{WJ}) = \mathbf{WJ}$.

2.5.4.2 The Fractional Approach In the fractional approach, we use the normalized citation network $n(\mathbf{Ci})$. The derived network is determined as follows:

$$\mathbf{JJf} = \mathbf{WJ}^T * n(\mathbf{Ci}) * \mathbf{WJ}$$

Its weights have the following meaning:
$jjf(i, j) =$ *fractional* contribution of citations from papers published in journal i to papers published in journal j, again with attention restricted to the network partitioning literature.

There are 12 link islands in [10, 50] range for the number of vertexes – see Figure 2.13. Examining this figure more closely, the largest link island (top left) involves the journals where work on blockmodeling and community detection appeared. This island is considered in more detail below. The subject matter of the remaining islands contains surprises. Continuing to read across the top row of this figure, the primary subjects are: dentistry; medical technologies; and surgery, reconstructive surgery, and physical therapy as follow-up treatments to surgery. Dropping to the next row, reading from the left, the subjects are: earthquakes and fluid mechanics; laser surgery in dentistry; petroleum engineering; and cardiovascular problems and treatments. Across the bottom row, the topics are: archeology and antiquity studies; marine research and ship technology; linguistics (featuring only German language journals); and soil science.

This diversity of subject matter suggests a variety of issues. One is that the network partitioning literature is spread across for more disciplines than we anticipated. Of course, this could imply that the initial search was too broad. But if multiple disciples are involved, examining to journal-to-journal structures these other disciplines has interest value. All of the islands are highly centralized having either star-like or hierarchical structures. This is suggestive of another feature of the organization of scientific production at the journal level which merits further attention.

We label the largest island as the main island. It is presented in Figure 2.14. By far, most journals on this link island are from the physics-driven approach. Indeed, as shown at the bottom left, only a small number come from the social science approach to social networks. In part, this reflects the institutional dominance of the natural sciences, especially physics. The only link from the physics literature to the social science literature is from *Physics Review E* to *Social Networks*. This is due to a link from a Newman paper in the former journal to a Batagelj paper in the latter literature, exactly the transition point between the blockmodeling literature and the community detection literature discussed in our analysis of main paths in Figure 2.2.

Figure 2.14 emphasizes its acyclic (hierarchical) structure with strong components. They are few in number. Only one is in social science part of the network (lower left of the figure). It has *Soc Networks* and *J Math Sociol* both of which featured works on blockmodeling. The largest strong component has (*Phys Rev E*, *Phys Rev A*, *Phys Rev Letters*, *Physica A*, *Eur Phys J B*, *Nature*, *Science* and *PNAS*). Note that the subgroup *Nature*, *Science*, and *PNAS* is linked back to the second strong component only with the arc between *PNAS* and *Phys Rev E*.

In the strong component with *Physics Review E*, the primary journal for works on community detection and related topics, *Physics Review Letters*, and *Physica A* formed by two reciprocated dyads involving *Physics Review E*. The fractional tie from *Physica A* to *Physics Review E* is far stronger than the reverse tie. The fractional tie from *Physics Re-*

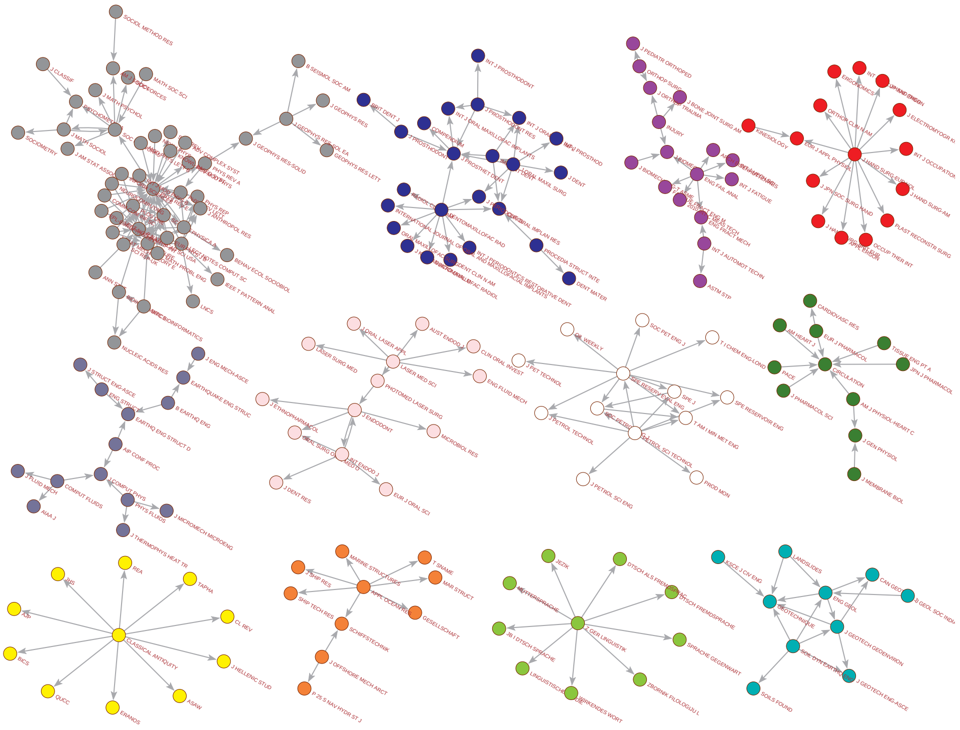

Figure 2.13: **JJf** fractional Islands

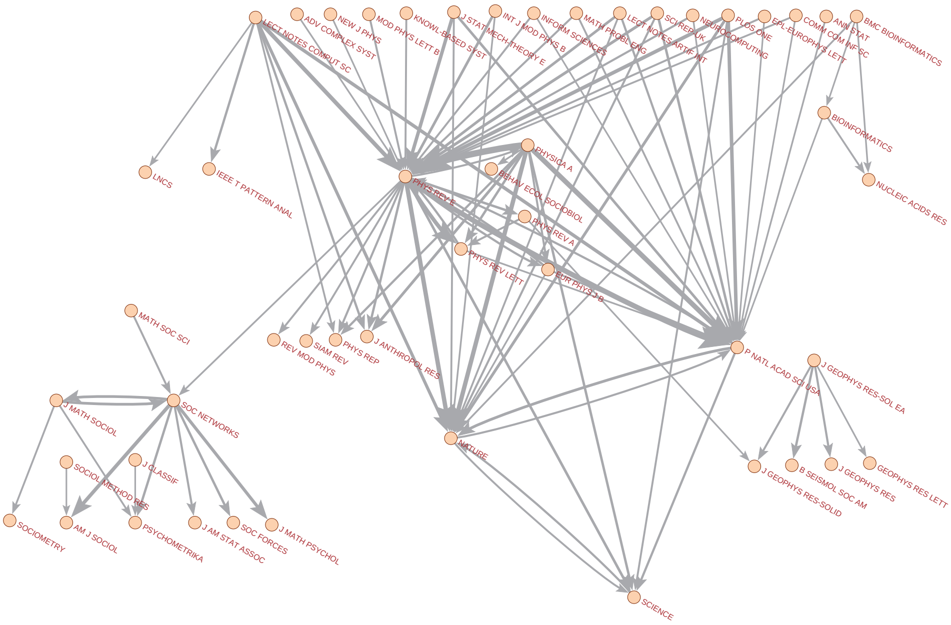

Figure 2.14: **JJf** main fractional island

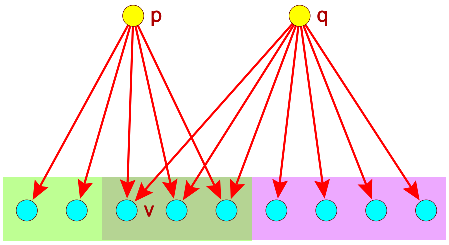

Figure 2.15: Bibliographic Coupling

view E to *Physics Review Letters* is stronger than the reverse tie, something meriting further attention. The third strong component also involves three journals with two reciprocated dyads. *Nature* has reciprocated ties with both the *Proceedings of the National Academy of Sciences* (*PNAS*) and *Science*. All three journals are highly institutionalized within the natural sciences. Both *PNAS* and *Nature* are heavily cited which reflects this institutional prominence. But these ties are not reciprocated. It seems reasonable to assume that works in the other journals sent ties to these journals as a form of validation of their ideas. *Science* is relatively peripheral in Figure 2.14. This suggests that the works involving partitioning networks are not a central part of the overall scientific literature involving the natural sciences.

2.5.5 Bibliographic Coupling

Bibliographic coupling occurs when two works each cite a third work in their bibliographies. The idea was introduced by Kessler in 1963 [20] and has been used extensively since then. See Figure 2.15 where two citing works, p and q, are shown. Work p cites five works and q cites seven works. The key idea is that there are three documents cited by both p and q. This suggests some content communality between p and q. It is thought that having more works citing pairs of prior works increases the likelihood of them sharing content. This is not unreasonable.

In **`WoS2Pajek`** the citation relation is $p\ \mathbf{Ci}\ q$ work p cites work q. Therefore the *bibliographic coupling* network **biCo** can be determined as

$$\mathbf{biCo} = \mathbf{Ci} * \mathbf{Ci}^T$$

$bico_{pq} =$ # of works cited by both works p and $q = |\mathbf{Ci}(p) \cap \mathbf{Ci}(q)|$.

Bibliographic coupling weights are symmetric: $bico_{pq} = bico_{qp}$:

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

The pairs with the largest values involve works featuring reviews (or overviews of a field) and authors citing themselves. Review papers may require closer consideration when considering bibliographic coupling as they make many citations across wide areas.

Figure 2.16 shows the bibliographic coupling of works for links above a threshold of 25. There is one large set of such coupled works in a network along with three dyads and a triple of works. They feature productions involving physicists and computer scientists.

2.5.5.1 Fractional bibliographic coupling Given the problems with works making many citations, especially with review works citing *many* works, we take a different approach. Necessarily, review papers cover a wide area (or multiple areas). That two works

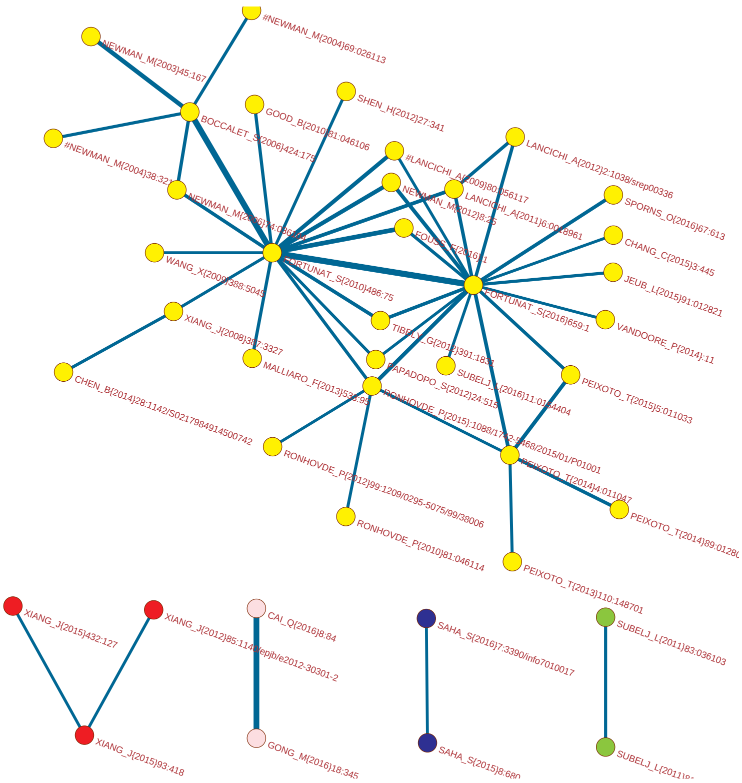

Figure 2.16: Bibliographic coupling above a threshold set at 25

are cited in a broad review paper need not imply that they have common content. Ideally, it would be useful to separate specific contributions on *research fronts* from works looking back at what was done in general. But the literature contains both types of documents. We think a different strategy is required. Neutralizing the distorting impact of review documents suggests using normalized measures designed to control for this is useful (see [16]). We first consider:

$$\mathbf{biC} = n(\mathbf{Ci}) * \mathbf{Ci}^T$$

where $n(\mathbf{Ci}) = \mathbf{D} * \mathbf{Ci}$ and $\mathbf{D} = \mathrm{diag}(\frac{1}{\max(1,\mathrm{outdeg}(p))})$. $\mathbf{D}^T = \mathbf{D}$.

$$\mathbf{biC} = (\mathbf{D} * \mathbf{Ci}) * \mathbf{Ci}^T = \mathbf{D} * \mathbf{biCo}$$

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

For $\mathbf{Ci}(p) \neq \emptyset$ and $\mathbf{Ci}(q) \neq \emptyset$ it holds (proportions)

$$\mathbf{biC}_{pq} = \frac{|\mathbf{Ci}(p) \cap \mathbf{Ci}(q)|}{|\mathbf{Ci}(p)|} \quad \text{and} \quad \mathbf{biC}_{qp} = \frac{|\mathbf{Ci}(p) \cap \mathbf{Ci}(q)|}{|\mathbf{Ci}(q)|} = \mathbf{biC}^T_{pq}$$

and $\mathbf{biC}_{pq} \in [0,1]$. $\mathbf{biC}_{pq}$ is the proportion of its references that the work p shares with the work q.

Combining $\mathbf{biC}_{pq}$ and $\mathbf{biC}_{qp}$ we can construct different normalized measures such as

$$\mathbf{biCoa}_{pq} = \frac{1}{2}(\mathbf{biC}_{pq} + \mathbf{biC}_{qp}) \quad \text{Average}$$

$$\mathbf{biCom}_{pq} = \min(\mathbf{biC}_{pq}, \mathbf{biC}_{qp}) \quad \text{Minimum}$$

Other possible measures include geometric mean, the harmonic mean and the Jaccard index. All these measures are symmetric. In the following we will use the Jaccard coefficient

$$\mathbf{biCoj}_{pq} = (\mathbf{biC}_{pq}^{-1} + \mathbf{biC}_{qp}^{-1} - 1)^{-1} = \frac{|\mathbf{Ci}(p) \cap \mathbf{Ci}(q)|}{|\mathbf{Ci}(p) \cup \mathbf{Ci}(q)|}$$

It is easy to verify that $biCoj_{pq} \in [0,1]$ and: $biCoj_{pq} = 1$ iff the works p and q are referencing the same works, $\mathbf{Ci}(p) = \mathbf{Ci}(q)$. To get a useful dissimilarity measure, use $dis = 1 - sim$ or $dis = \frac{1}{sim} - 1$ or $dis = -\log sim$. For example

$$\mathbf{biCojD}_{pq} = 1 - \mathbf{biCoj}_{pq} = \frac{|\mathbf{Ci}(p) \oplus \mathbf{Ci}(q)|}{|\mathbf{Ci}(p) \cup \mathbf{Ci}(q)|} \quad \text{Jaccard distance}$$

the proportion of the number of distinct neighbors and all neighbors of works p and q in the citation network.

2.5.5.2 Jaccard islands We computed Jaccard similarity measures for the network CiteB and determined corresponding link islands having sizes in the range [5,75]. The following table shows the distribution of sizes of 133 islands that were identified.

size	5	6	7	8	9	10	11	12	13	14	15	17	18	24	27
number	33	16	11	17	12	8	4	2	2	3	1	4	2	1	1

size	28	31	33	34	40	43	48	51	52	55	58	70	71	75
number	1	2	1	1	1	1	1	1	2	1	1	1	1	1

We examine more closely a social networks Jaccard island (shown in Figure 2.17 with 70 works), a Jaccard island featuring works of physicists (in Figure 2.18 with 58 works), and three smaller Jaccard islands having 23, 22 and 18 works (see Figure 2.20).

The social networks Jaccard island is the largest such island. It has works spread over a variety of topics linked to partitioning social networks. There are many cuts linking these areas. One the top left of Figure 2.17, the works involve stochastic blockmodeling and exponential random graph models. The work by Sailer appeared in 1974 and is a cut connecting three sub-areas including the part just described. To the right of this cut are works involving the origins of blockmodeling. Below this cut are more works on classical blockmodeling. On the lower right of Figure 2.17 are works featuring discussions of the early algorithms for blockmodeling. At the bottom of the figure are more contemporary works on blockmodeling, including generalized blockmodeling. Many of these works were featured in Section 2.3.

The Jaccard island shown in Figure 2.18 features works by physicists regarding community detection and related methods for partitioning networks. It also has many cuts. Indeed, we suggest the presence of cuts is a feature of networks formed through bibliographic coupling links. In addition, it seems that bibliographic coupling is very useful for identifying different *sub-areas* of fields and how they are connected.

It is straightforward to determine the citations received by works in these two Jaccard islands. The top numbers of received citations are shown in Table 2.18 where the relevant items from social network literature is on the left and those for the physicists are on the right.

Without surprise, most of the works appearing in both columns have appeared earlier in our narrative. There are some clear differences between the two distributions. When the

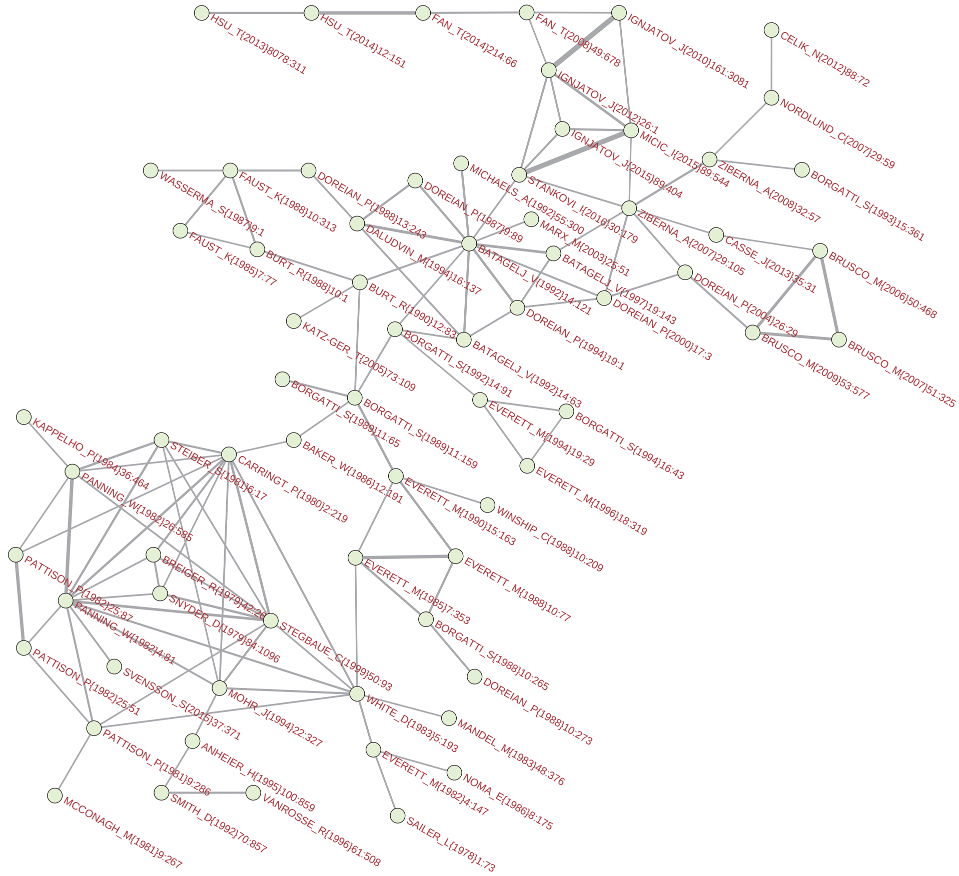

Figure 2.17: Bibliographic coupling in the social networks literature

box-plots are drawn, the distribution for the social networks literature is far more skewed, with outliers present, than for the physicist part of the literature. Also the mean and median for these limited distributions are considerably higher in the physicist literature.

When the years of the publications are examined, another clear difference emerges. The range of years for the social networks part of the literature goes from 1950 to 1992. In contrast, the physicist works have dates ranging from 2002 to 2010. (The one document on the right in Table 2.18 that appeared in 1977 was written by an anthropologist. His data were latched upon by the physicists as useful data allowing the demonstration of community detection methods.) This reflects a clear difference between these two parts of the literature on clustering networks. One was developed over a longer period of time as a 'leisurely' generation of methods, their application, and the generation of substantive results regarding the structure of *social* networks. It was merely one part of this literature that focused on many other issues regarding social networks The community detection literature exploded over a much shorter period of time with a focus on a clearly defined technical research issue.

It reflects also a difference in the social organization of science, something noted in [7]. Larger disciplines having more journals and a much longer institutionalized organization

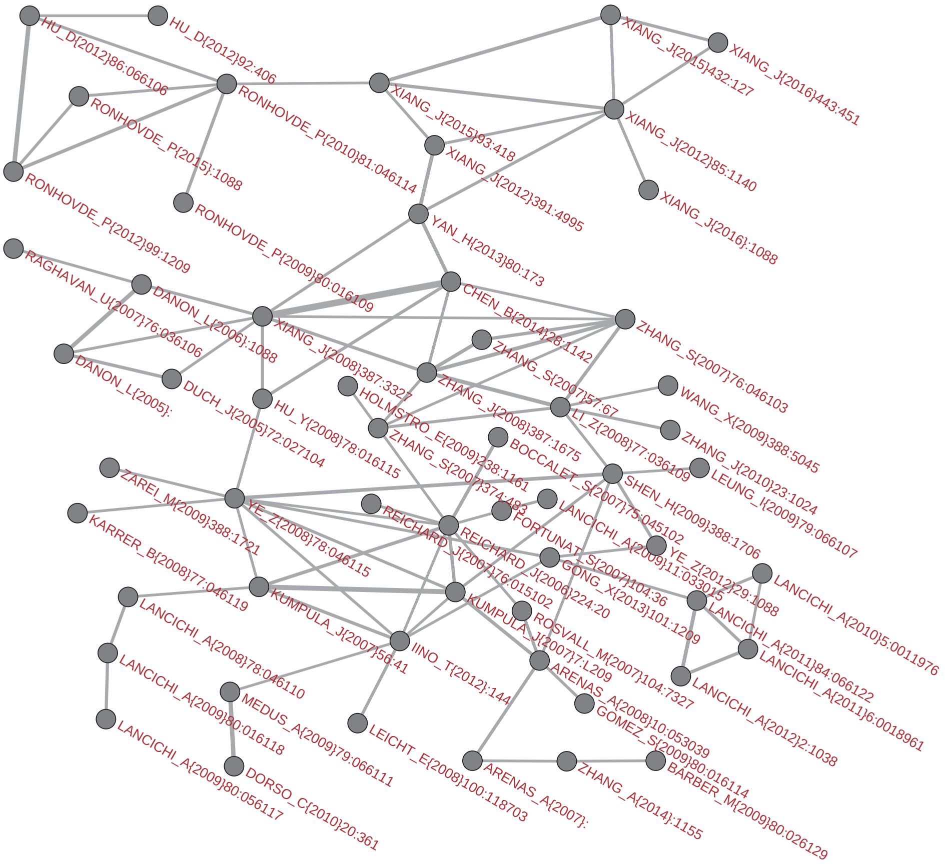

Figure 2.18: Bibliographic coupling in the physicist-driven literature

regarding professional organizations, as well as having more publication outlets, become far more visible.

We turn now to consider three smaller Jaccard islands. They are shown in Figure 2.20. The methods for determining citations are exactly the same as for the two largest islands. These three smaller islands have works focused in three domains. The first deals with a part of the physicist and mathematical literature, the second with a part of the broader clustering literature and the third with signed networks.

As indicated by the works in the left column of Table 2.17, the earliest work (by Erdős appearing in 1960 and is at rank 10 of the column) set the foundations for the development of random graph theory. Another mathematical work appeared in 1985 (rank 6 in the column). There is an early social science work at rank 15 that attracted the attention of some physicists. A social networks text appears at rank 4 with sections on random graphs. The remaining works produced by physicists building on these ideas are concentrated between 1995 and 2001.

The top ranked item in the second column of Table 2.17 appeared in *Psychometrika* in 1982. A companion paper by the same authors (Ferligoj and Batagelj) in the same journal appeared a year later. These works created the foundations for a distinctive approach to clustering relational and attribute data that was picked up by others working on general

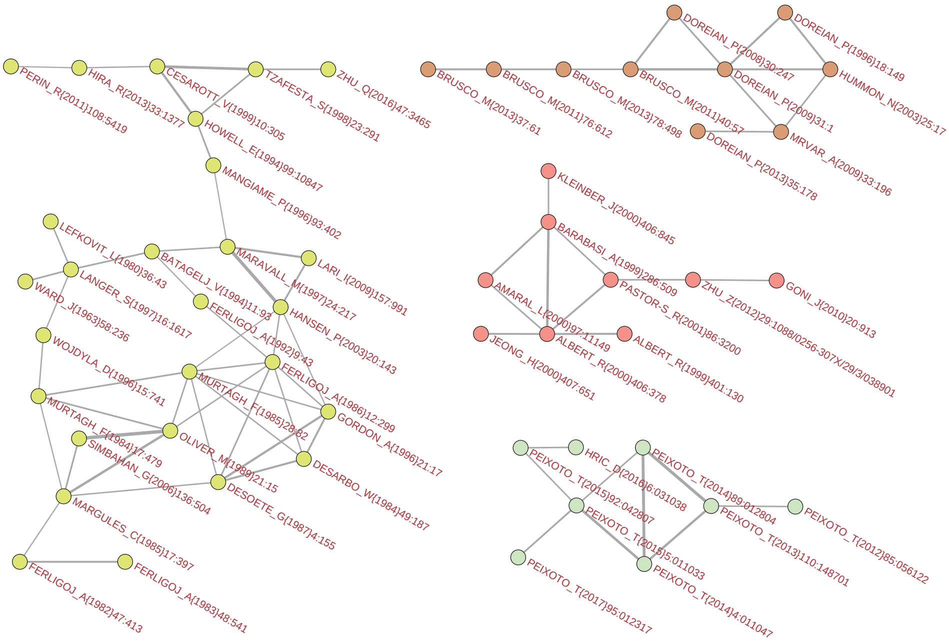

Figure 2.19: Bibliographic coupling – selected islands

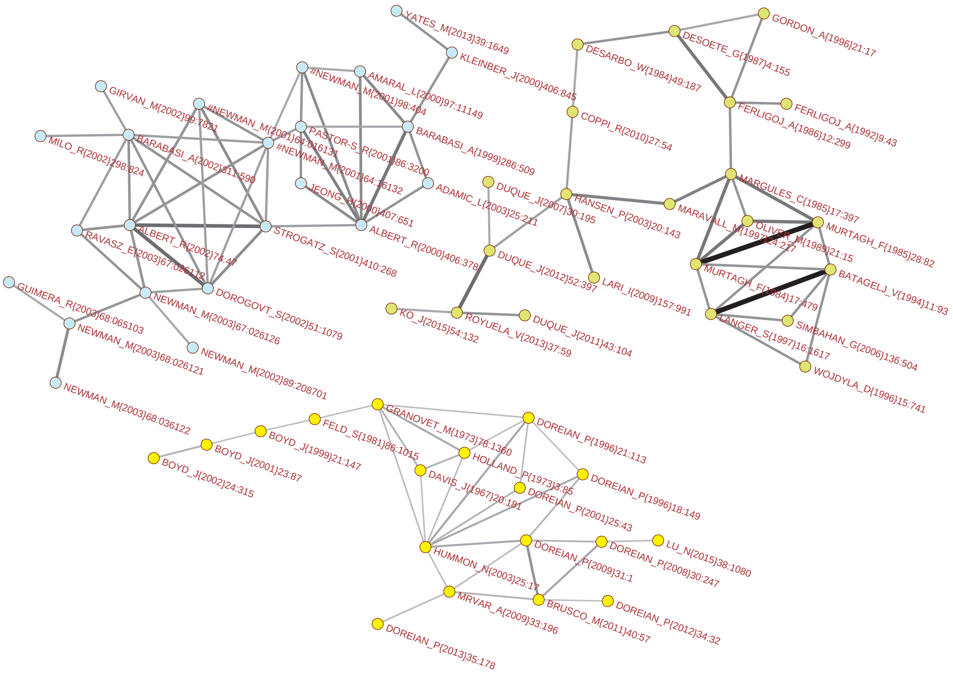

Figure 2.20: Bibliographic coupling for three smaller islands

Table 2.16: Bibliographic coupling of the most cited works from works of the two largest islands

Figure 2.17 (Social network literature)			Figure 2.18 (Physicist literature)		
Rank	Count	Work	Rank	Count	Work
1	58	LORRAIN_F(1971)1:49	1	45	GIRVAN_M(2002)99:7821
2	50	WHITE_H(1976)81:730	2	43	#NEWMAN_M(2004)69:026113
3	48	BREIGER_R(1975)12:328	3	40	CLAUSET_A(2004)70:066111
4	33	ARABIE_P(1978)17:21	4	38	DUCH_J(2005)72:027104
5	26	BOORMAN_S(1976)81:1384	5	36	GUIMERA_R(2005)433:895
6	24	SAILER_L(1978)1:73	6	35	#NEWMAN_M(2004)38:321
7	22	BURT_R(1976)55:93	7	34	RADICCHI_F(2004)101:2658
8	22	WHITE_D(1983)5:193	8	31	#DANON_L(2005):
9	15	NADEL_S(1957):	9	31	#ZACHARY_W(1977)33:452
10	14	HEIL_G(1976)21:26	10	27	FORTUNAT_S(2007)104:36
11	12	SAMPSON_S(1969):	11	25	ALBERT_R(2002)74:47
12	12	HOLLAND_P(1981)76:33	12	25	NEWMAN_M(2003)45:167
13	11	BURT_R(1983):	13	20	REICHARD_J(2006)74:016110
14	11	JOHNSON_S(1967)32:241	14	20	REICHARD_J(2004)93:218701
15	10	BURT_R(1982):	15	19	GUIMERA_R(2003)68:065103
16	10	HOMANS_G(1950):	16	19	NEWMAN_M(2006)103:8577
17	10	FAUST_K(1988)10:313	17	19	PALLA_G(2005)435:814
18	10	FREEMAN_L(1979)1:215	18	19	WU_F(2004)38:331
19	10	FIENBERG_S(1985)80:51	19	17	FLAKE_G(2002)35:66
20	9	BORGATTI_S(1989)11:65	20	17	#BLONDEL_V(2008):P10008
21	8	WHITE_H(1963):	21	17	BOCCALET_S(2006)424:175
22	8	BURT_R(1980)6:79	22	17	GLEISER_P(2003)6:565
23	8	BREIGER_R(1979)13:21	23	16	FORTUNAT_S(2010)486:75
24	8	BATAGELJ_V(1992)14:121	24	16	RAVASZ_E(2002)297:1551
25	7	MANDEL_M(1983)48:376	25	16	MEDUS_A(2005)358:593
26	7	KNOKE_D(1982):	26	16	#DONETTI_L(2004):P10012
27	7	DOREIAN_P(1988)13:243	27	15	NEWMAN_M(2006)74:036104
28	7	BREIGER_R(1978)7:213	28	13	BRANDES_U(2008)20:172
29	7	SNYDER_D(1979)84:1096	29	13	GUIMERA_R(2004)70:025101
30	7	HUBERT_L(1978)43:31	30	12	HOLME_P(2003)19:532

Table 2.17: Bibliographic Coupling in the three smaller islands

Rank	Physicist literature		Clustering literature		Signed networks	
1	23	WATTS_D(1998)393:440	21	FERLIGOJ_A(1982)47:413	13	CARTWRIG_D(1956)63:277
2	18	BARABASI_A(1999)286:509	11	LEFKOVIT_L(1980)36:43	12	HEIDER_F(1946)21:107
3	17	ALBERT_R(1999)401:130	10	PERRUCHE_C(1983)16:213	11	DAVIS_J(1967)20:181
4	15	WASSERMA_S(1994):	9	MURTAGH_F(1985)28:82	10	NEWCOMB_T(1961):
5	15	AMARAL_L(2000)97:11149	8	FERLIGOJ_A(1983)48:541	9	WHITE_H(1976)81:730
6	13	BOLLOBAS_B(1985):	6	GORDON_A(1996)21:17	8	HARARY_F(1965):
7	13	FALOUTSO_M(1999)29:251	4	DUQUE_J(2007)30:195	8	DOREIAN_P(1996)18:149
8	13	NEWMAN_M(2001)98:404	4	KIRKPATR_S(1983)220:671	7	DOREIAN_P(2005):
9	10	STROGATZ_S(2001)410:268	4	MACQUEEN_J(1967):281	7	HEIDER_F(1958):
10	10	ERDOS_P(1960)5:17	3	DESARBO_W(1984)49:187	6	BREIGER_R(1975)12:328
11	10	REDNER_S(1998)4:131	3	MARGULES_C(1985)17:397	6	HOMANS_G(1950):
12	9	JEONG_H(2000)407:651	3	HANSEN_P(2003)20:143	6	BATAGELJ_V(1998)21:47
13	9	ALBERT_R(2000)406:378	3	DUQUE_J(2011)43:104	5	BORGATTI_S(2002):
14	9	MOLLOY_M(1995)6:161	3	MARAVALL_M(1997)24:217	5	LORRAIN_F(1971)1:49
15	9	MILGRAM_S(1967)1:61	3	GAREY_M(1979):	5	WHITE_D(1983)5:193

clustering problems. The other works in this column came from researchers working on traditional clustering problems.

Most of the works appearing in the third column of Table 2.17 deal with signed networks. The top four ranked items set the foundations for a formal approach to structural balance. The conceptual approach came from Heider in 1946 is ranked second. The top rank is for an initial formal statement by Cartwright and Harary in 1956 and extended by Davis in 1967. There are some items on blockmodeling that were picked up by Doreian and Mrvar in 1996 to create an algorithm for partitioning signed networks.

Bibliographic Coupling the most frequent keywords in works of a given subnetwork For the social networks island and the physicist island identified in Figures 2.17 and 2.18, the most frequent keywords in works of these islands were extracted. They are shown in Table 2.18.

We consider first the left column featuring the social networks part of the clustering literature. The top two keywords are social and network confirming the nature of the works in this island. The next two are solidly about blockmodeling which is based on conceptions of equivalence. Additional terms include role structural, relation, sociometric, position, regular (for a specific equivalence type), direct (for one approach to blockmodeling) and block. All of these terms are recognizable as relevant terms.

The word network also heads the list of keywords for the community detection literature. It followed immediately by community. Again, the essence of the content of the island is identified. It is followed by complex, a term used far more by the physicists in the expression ‘complex networks’. The term modularity is foundational for community detection. The presence of overlap as a keyword in this island reflects another difference between the two literatures with community detection authors being far more concerned with overlapping clusters. The presence of the keywords metabolic and biological provide a hint that the physicists study a broader set of networks that those working in social networks.

There are only seven keywords common to both lists - network, analysis, structure, graph, model, algorithm and organization. Both areas are concerned with delineating structure, studying graphs, fitting models (albeit of different sorts) and mobilizing algorithms.

Co-citation is a concept with strong parallels with bibliographic coupling (see Small and Marshakova [24, 27]). The focus is on the extent to which works are co-cited by later works. The basic intuition is that the more earlier works are cited, the higher the likelihood that they have common content. The *co-citation* network **coCi** can be determined as $\mathbf{coCi} = \mathbf{Ci}^T * \mathbf{Ci}$. $coci_{pq} =$ # of works citing both works p and q. $coci_{pq} = coci_{qp}$. The same kinds of analyses can be performed for co-citation. An example of doing this is in [7] regarding the Supreme Court. However, we do not pursue this here.

2.5.6 Linking through a Jaccard network

Bibliographic coupling networks are linking works to works. Let **S** be such a network. The derived network $\mathbf{WA}^T * \mathbf{S} * \mathbf{WA}$ links authors to authors through **S**. Again, the normalization question must be addressed. Given different options, we selected the derived networks defined as:

$$\mathbf{C} = n(\mathbf{WA})^T * \mathbf{S} * n(\mathbf{WA})$$

It is easy to verify that:

- if **S** is symmetric, $\mathbf{S}^T = \mathbf{S}$, then also **C** is symmetric, $\mathbf{C}^T = \mathbf{C}$;

Table 2.18: The most frequent keywords of the two largest islands in the Jaccard bibliographic coupling network

Figure 2.17 (Social network literature)			Figure 2.18 (Physicist-driven literature)		
Rank	Count	Work	Rank	Count	Work
1	42	network	1	54	network
2	34	social	2	52	community
3	27	blockmodel	3	48	complex
4	24	equivalence	4	30	structure
5	23	analysis	5	30	modularity
6	17	structure	6	28	detection
7	17	role	7	19	algorithm
8	15	structural	8	18	graph
9	12	relation	9	17	metabolic
10	11	multiple	10	12	resolution
11	10	graph	11	12	model
12	10	datum	12	12	optimization
13	8	statistical	13	9	organization
14	7	model	14	8	detect
15	7	algorithm	15	8	cluster
16	7	sociometric	16	7	identification
17	7	position	17	6	dynamics
18	7	regular	18	6	analysis
19	6	relational	19	6	method
20	6	computation	20	5	use
21	6	two	21	5	base
22	5	organization	22	5	hierarchical
23	5	stochastic	23	4	overlap
24	5	approach	24	4	pott
25	5	direct	25	4	multi
26	4	block	26	4	maximization
27	4	similarity	27	4	world
28	4	group	28	4	information
29	4	application	29	4	biological
30	3	measure	30	4	limit

- the total of weights of **S** is redistributed in **C**:

$$\sum_{a \in L(C)} c(a) = \sum_{a \in L(S)} s(a)$$

We applied this construction to combine the Jaccard network with networks **WA**, **WJ** and **WK**. We limited our analysis to networks with complete descriptions of works (**WAc**, **WJc**, **WKc**, **CiteC**).

As an example, Figure 2.21 presents a link-cut at level 11 in the authors Jaccard coupling network $\mathbf{ACoj} = n(\mathbf{WA})^T * \mathbf{biCoj} * n(\mathbf{WA})$. There are two disjoint parts to the figure. The smaller one on the right features authors active in the blockmodeling literature. It is centered on Doreian. The larger part on the left comes from the physics driven literature and is centered on Newman. The results is very similar to the one shown in Figure 2.12. The social networks part is smaller in Figure 2.21 while the physics driven part is larger with an additional part linked through Turcotte.

In Figure 2.22 a link-cut at level 1300 in the journals Jaccard coupling network $\mathbf{JCoj} = n(\mathbf{WJ})^T * \mathbf{biCoj} * n(\mathbf{WJ})$ is presented as another example. Because the links between

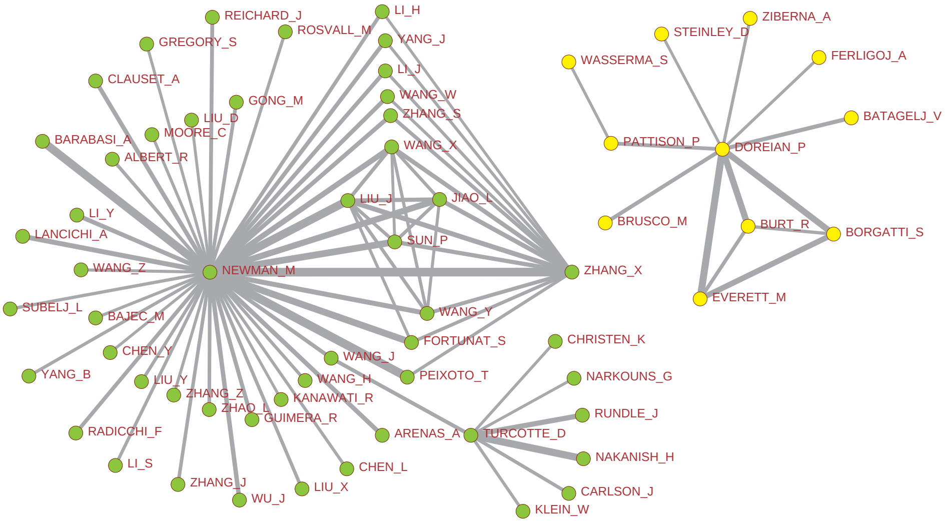

Figure 2.21: Authors Jaccard coupling – cut at level 11

journals have greater weights, a much larger link-cut is required. Overwhelmingly, the journals come from the physics driven part of the literature. Three such journals are particularly prominent: *Physica A*, *Phys Rev E* and *PLOS ONE*. The fourth prominent journal is *Lecture Notes Comput SC* from the computing science literature. Only two social science journals are present. *J Math Sociol* is linked to *Soc Networks* which is linked to only *Physica A*. Despite its name, *Soc Netw Anal Min* is focused more on data mining in large networks, reflecting more a computer science orientation.

In Figure 2.23 the main link-island in [1,30] in the keywords Jaccard coupling network $\mathbf{KCoj} = n(\mathbf{WK})^T * \mathbf{biCoj} * n(\mathbf{WK})$ is presented. Table 2.18 presented separate lists of keywords in the social network literature and the physics-driven literature. Figure 2.23 shows how *some* of these keywords are linked. That the keywords, network and community, are the most prominent is not surprising given these keywords head of the physicists-driven literature list in Table 2.18. Having ‘community’ and ‘detection’ separated is problematic and reflects the problem of two-word keywords discussed earlier. Overall, 15 of the 30 keywords from the physicist-driven literature are in Figure 2.23 while ten of the keywords from the social network literature are present (with seven common to both lists). Overall, the linkage of the keywords shown in Figure 2.23 seems more useful than the separate list in Table 2.18.

2.6 Summary and future work

We obtained citation data for the network clustering literature for a large citation network including both community detection and blockmodeling works through to February 22, 2017. The primary data source was the Web of Science. Details about recording, processing and resulting data sets were provided. In addition to having works as units, we included data on authors, journals and keywords to generate some two-mode networks fea-

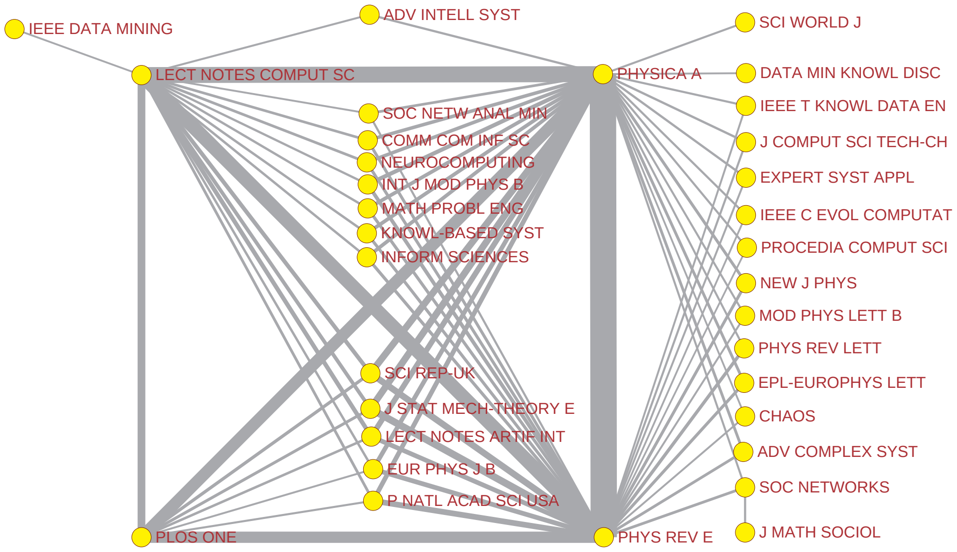

Figure 2.22: Journals Jaccard coupling – cut at level 1300

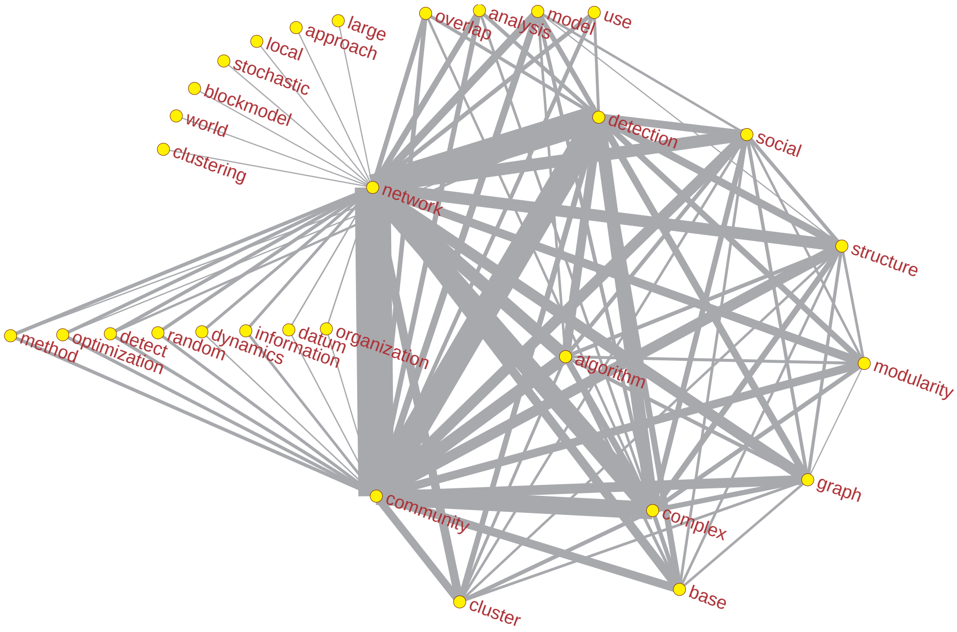

Figure 2.23: Keywords Jaccard coupling – main island

turing works × authors, works × journals, and works × keywords. The boundary problem was discussed as was a treatment ensuring the studied citation network is acyclic.

Our results included descriptions of the most cited works and the most citing works as a preliminary delineation of the content of this research area. Lists of the most prominent journals where works in the network clustering literature appeared were created. In doing so, the importance of establishing network boundaries appropriately was discussed. The nature of keywords was discussed with a proviso that many cannot be taken at face value and using them to understand science must be done with great care.

Components of the studied network were identified with attention confined to the largest one. The CPM path through this component was identified. It revealed a clear transition from the social network part of the literature to the community detection part. The key-route paths revealed the same transition but with more works and a more nuanced view of it. Link islands, as clusters, were identified. There were ten of them. Detailed discussions were provided for four including one with a clear distinction between the community detection and social networks literatures as being connected through a cut.

When attention was turned to considering authors, a listing of authors involved in the most works was provided. This was a limited result. To move beyond this, we examined productivity within research groups by using P_S-cores. A listing of authors having the largest P_S-core values was provided. To dig further into the contribution of authors, both co-authorship and collaboration were studied. This was extended to citations among the authors contributing to the network clustering literature, with close attention paid to the community detection and blockmodeling parts. Attention was paid journal-to-journal networks for only the items identified in the network clustering literature.

Bibliographic coupling was considered and extended through fractional bibliographic coupling to use a better measure of the extent to which works are coupled. A total of 15 link islands were identified in the network of bibliometrically coupled documents. Again, attention was focused on two featuring, separately, the social network and physicist-driven parts of this literature. Three more smaller link islands were examined, each with a clear sub-part of the literature. When keywords were examined in the context of link islands and bibliometric coupling, they were much more useful. Also, sub-areas were more clearly identified.

Together, these different ways of examining the network clustering literature provided a coherent and consistent understanding of its citation structure of works and the contributions of authors and journals. Future work will consider the other link islands in the citation network and those identified in the bibliometric coupling of works. Given the usefulness of bibliometric coupling, it is highly likely that the co-citation network will add additional insight into the coherence of this literature.

REFERENCES

1. ***histcomp*** – (*comp*iled *Hist*oriography program), 2010. URL `http://garfield.library.upenn.edu/histcomp/guide.html`.
2. V. Batagelj. Efficient algorithms for citation network analysis. *CoRR*, cs.DL/0309023, 2003. URL `http://arxiv.org/abs/cs.DL/0309023`.
3. V. Batagelj. *WoS2Pajek: Manual for version 1.4*. IMFM, Ljubljana, Slovenia, 2016. URL `http://vladowiki.fmf.uni-lj.si/doku.php?id=pajek:wos2pajek`.
4. V. Batagelj and M. Cerinšek. On bibliographic networks. *Scientometrics*, 96(3):845–864, 2013. doi: 10.1007/s11192-012-0940-1. URL `http://dx.doi.org/10.1007/s11192-012-0940-1`.
5. V. Batagelj and A. Mrvar. *Pajek and Pajek-XXL, Program for Analysis and Visualization of Large Networks, Reference Manual, List of commands with short explanation*, 1996-2017. version 5.01.

6. V. Batagelj and M. Zaveršnik. Fast algorithms for determining (generalized) core groups in social networks. *Adv. Data Analysis and Classification*, 5(2):129–145, 2011. doi: 10.1007/s11634-010-0079-y. URL `http://dx.doi.org/10.1007/s11634-010-0079-y`.

7. V. Batagelj, P. Doreian, A. Ferligoj, and N. Kejžar. *Understanding Large Temporal Networks and Spatial Networks: Exploration, Pattern Searching, Visualization and Network Evolution*. Wiley Series in Computational and Quantitative Social Science Series. Wiley, 2014. ISBN 9781118915370. URL `https://books.google.si/books?id=ez8xBwAAQBAJ`.

8. R. L. Breiger, S. A. Boorman, and P. Arabie. An algorithm for clustering relational data with applications to social network analysis and comparison with multidimensional scaling. *Journal of Mathematical Psychology*, 12(3):328 – 383, 1975. ISSN 0022-2496. doi: https://doi.org/10.1016/0022-2496(75)90028-0. URL `http://www.sciencedirect.com/science/article/pii/0022249675900280`.

9. R. S. Burt. Positions in networks. *Social Forces*, 55(1):93–122, 1976. doi: 10.1093/sf/55.1.93. URL `http://dx.doi.org/10.1093/sf/55.1.93`.

10. M. Cerinšek and V. Batagelj. Network analysis of zentralblatt MATH data. *Scientometrics*, 102 (1):977–1001, 2015. doi: 10.1007/s11192-014-1419-z. URL `http://dx.doi.org/10.1007/s11192-014-1419-z`.

11. S. Cole and J. R. Cole. Visibility and the structural bases of awareness of scientific research. *American Sociological Review*, 33(3):397–413, 1968. ISSN 00031224. URL `http://www.jstor.org/stable/2091914`.

12. P. Doreian and F. N. E. Stokman. *Evolution of Social Networks*. Gordon & Breach, New York, USA, 1997.

13. P. Doreian, V. Batagelj, and A. Ferligoj. *Generalized Blockmodeling*. Structural Analysis in the Social Sciences. Cambridge University Press, New York, USA, 2005. ISBN 978-0-521-84085-9.

14. E. Garfield. Citation analysis as a tool in journal evaluation. *Science*, 178:471–479, 1972.

15. E. Garfield, A. I. Pudovkin, and V. S. Istomin. Why do we need algorithmic historiography? *J. Am. Soc. Inf. Sci. Technol.*, 54(5):400–412, Mar. 2003. ISSN 1532-2882. doi: 10.1002/asi.10226. URL `http://dx.doi.org/10.1002/asi.10226`.

16. M. Gauffriau, P. O. Larsen, I. Maye, A. Roulin-Perriard, and M. von Ins. Publication, cooperation and productivity measures in scientific research. *Scientometrics*, 73(2):175–214, 2007. doi: 10.1007/s11192-007-1800-2. URL `http://dx.doi.org/10.1007/s11192-007-1800-2`.

17. N. P. Hummon and P. Doreian. Connectivity in a citation network: The development of dna theory. *Social Networks*, 13:39–63, 1989.

18. N. P. Hummon and P. Doreian. Computational methods for social network analysis. *Social Networks*, 12:273–288, 1990.

19. N. P. Hummon, P. Doreian, and L. C. Freeman. Analyzing the structure of the centrality-productivity literature. *Knowledge*, 11:460–481, 1990.

20. M. Kessler. Bibliographic coupling between scientific papers. *American Documentation*, 14(1): 10–25, 1963.

21. L. Leydesdorff, C. S. Wagner, and L. Bornmann. Betweenness and diversity in journal citation networks as measures of interdisciplinarity—a tribute to eugene garfield. *Scientometrics*, 114: 567–592, 2018.

22. J. S. Liu and L. Y. Y. Lu. An integrated approach for main path analysis: Development of the hirsch index as an example. *Journal of the American Society for Information Science and Technology*, 63(3):528–542, 2012. doi: 10.1002/asi.21692. URL `http://dx.doi.org/10.1002/asi.21692`.

23. F. Lorrain and H. C. White. Structural equivalence of individual in social networks. *Journal of Mathematical Sociology*, 1:49–80, 1971.

24. I. Marshakova. System of documentation connections based on references (sci). *Nauchno-TekhnicheskayaInformatsiya Seriya*, 2(6):3–8, 1973.

25. W. D. Nooy, A. Mrvar, and V. Batagelj. *Exploratory Social Network Analysis with Pajek*. Cambridge University Press, New York, NY, USA, 2011. ISBN 0521174805, 9780521174800.

26. R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2011. URL `http://www.R-project.org`. ISBN 3-900051-07-0.

27. H. Small. Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science*, 24(4):265–269, 1973.

28. Wikipedia. Chinese surname — Wikipedia, the free encyclopedia, 2013. URL `https://en.wikipedia.org/wiki/Chinese_surname`. [Online; accessed 22-July-2014].