The structure of the citation network for the literature on clustering networks

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

Anuška Ferligoj
University of Ljubljana and NRU HSE, Moscow

Patrick Doreian
University of Pittsburgh and University of Ljubljana

XXXVIII Sunbelt 2018
Utrecht, The Netherlands, June 26 – July 1, 2018

Outline

Goal

The goal is to study the network clustering literature including both blockmodeling and community detection works included in Web of Science till Feberuary 2017. The questions to be answered are:

- Which publications are the most cited?
- Which are the main authors and journals publishing papers on network clustering?
- Which are the main keywords in papers on network clustering?
- Which are the most influential publications in the field of network clustering?

For answering these questions several social network analysis approaches are applied on several two-mode networks and on large citation network obtained from WoS. The most useful ones are the 'main path' analyses and the 'islands' procedure.

Data

From the Web of Science (WoS), using the queries:

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

we downloaded in February 2017 the corresponding data set. We manually improved it.

The publications that appear in descriptions are of two types:

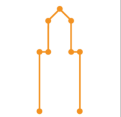
- publications with a WoS description (hits);
- cited-only publications (listed in CR fields of descriptions, but not contained in hits).

WoS networks

Using the program WoS2Pajek we transformed it into a collection of networks: the citation network, the authorship network, the journalship network, and the keywordship network.

```
number of publications =  117 082
number of authors      =   62 143
number of journals     =   12 652
number of keywords     =   36 279
number of records      =   10 269
```

Since the same work can have different names we identified all such works with the citation input degree larger than 30 and shrank the set of works (116 906). We removed multiple links and loops from networks. This networks are labeled by r.



The boundary problem for citation network

Most works are cited only once (indegree=1). We 'solved' the boundary problem by including in our networks those works with full description (hits, labeled by names ending by c) or with indegree > 2 (these are labeled by b). These criteria determined a subnetwork, denoted as **CiB**, with 13540 nodes and 82238 arcs.

Most cited publications (hits)

R	Cit	Work	R	Cit	Work	R	Cit	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)6:00189
3	712	CLAUSET_A(2004)70:066111	23	286	NEWMAN_M(2004)69:1	43	139	GREGORY_S(2010)12:1030
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)424:17
6	578	NEWMAN_M(2006)103:8577	26	244	DUCH_J(2005)72:027104	46	137	GUIMERA_R(2004)70:0251
7	553	**ZACHARY**_W(1977)33:452	27	236	LUSSEAU_D(2003)54:396	47	129	NEWMAN_M(2004)70:056
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)104:732
12	380	NEWMAN_M(2004)38:321	32	206	**WHITE_H**(1976)81:730	52	124	VONLUXBU_U(2007)17:39
13	369	LANCICHI_A(2008)78:046110	33	199	AHN_Y(2010)466:761	53	122	NEWMAN_M(2001)64:026
14	351	**WASSERMA**_S(1994):	34	168	KERNIGHA_B(1970)49:291	54	119	REICHARD_J(2004)93:218
15	329	NEWMAN_M(2006)74:036104	35	163	AIROLDI_E(2008)9:1981	55	118	ARENAS_A(2008)10:05303
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

The most used journals in two works × journals networks

Rank	Freq. (**WJr**)	Journal	Freq. (**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

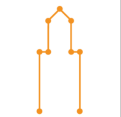
The most used keywords

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						

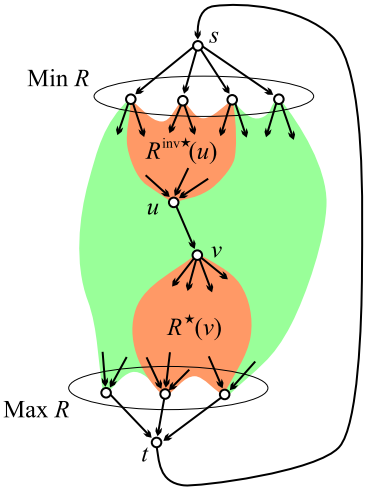
Citation networks

Our analyses of the primary 'clustering citation network' (**CiB**) features components, the identification of main paths through this literature and identifying islands (as clusters of related works). 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 remains. To obtain an acyclic network, we applied the preprint transformation to **CiteMain**. The resulting network, **CiteMacy** (Cite, Main, acyclic), has 12712 nodes and 81972 arcs. We computed the SPC weights on its arcs.



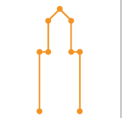
The CPM path of the main citation network (**CiteMacy**)



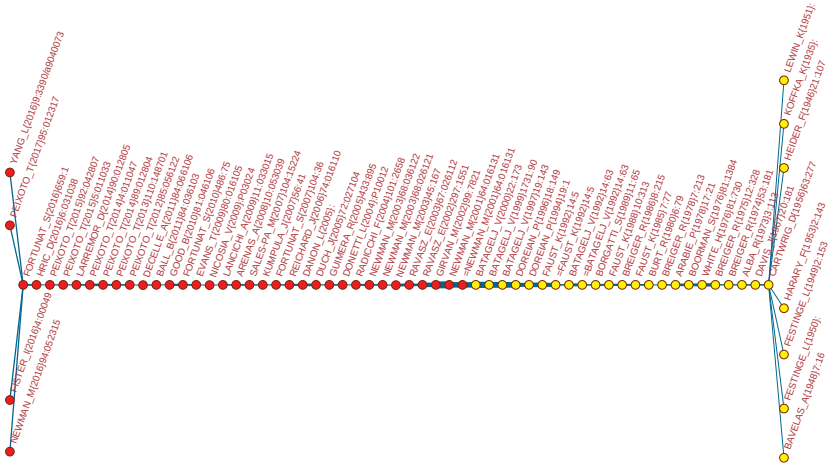
The *search path count* (SPC) method is based on counters $n(u, v)$ that count the number of different paths from s to t through the arc (u, v) (Batagelj et al. 2014).

The *Main path* starts in a link with the largest SPC weight and expands in both directions following the adjacent link with the largest SPC weight.

The *CPM path* is determined using the Critical Path Method from Operations Research (the sum of SPC weights in a path is maximal).



The CPM path through the network clustering literature



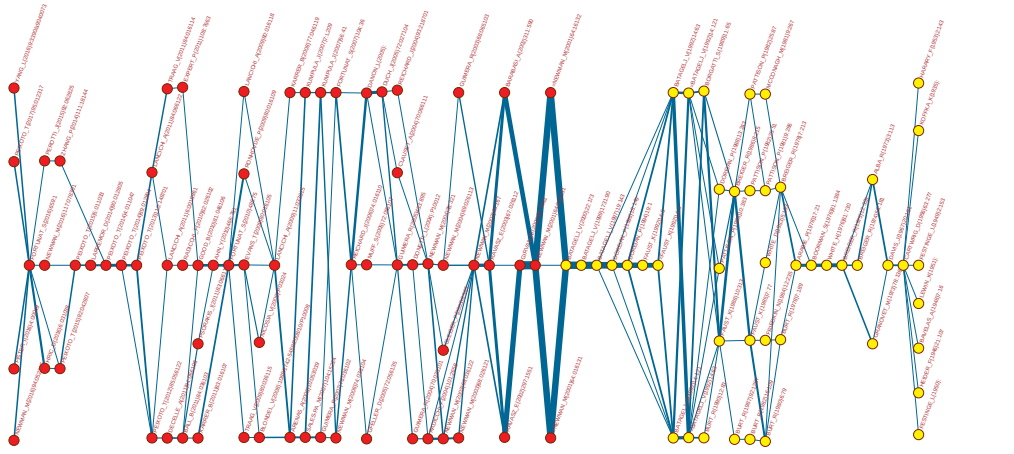
List of works on CPM path, 1

first author	title	journal
Cartwright, D (1956)	Structural balance - a generalization of heider theory	PSYCHOL REV
Davis, JA (1967)	Clustering and structural balance in graphs	HUM RELAT
Alba, RD (1973)	Graph-theoretic definition of a sociometric clique	J MATH SOCIOL
Breiger, RL (1974)	Duality of persons and groups	SOC FORCES
Breiger, RL (1975)	Algorithm for clustering relational data with applications to social network analysis and comparison with multidimensional-scaling	J MATH PSYCHOL
White, HC (1976)	Social-structure from multiple networks .1. Blockmodels of roles and positions	AM J SOCIOL
Boorman, SA (1976)	Social-structure from multiple networks .2. Role structures	AM J SOCIOL
Arabie, P (1978)	Constructing blockmodels - how and why	J MATH PSYCHOL
Breiger, RL (1978)	Joint role structure of 2 communities elites	SOCIOL METHOD RES
Burt, RS (1980)	Models of network structure	ANNU REV SOCIOL
Faust, K (1985)	Does structure find structure - a critique of Burt use of distance as a measure of structural equivalence	SOC NETWORKS
Breiger, RL (1986)	Cumulated social roles - the duality of persons and their algebras	SOC NETWORKS
Faust, K (1988)	Comparison of methods for positional analysis - structural and general equivalences	SOC NETWORKS
Borgatti, SP (1989)	The class of all regular equivalences - algebraic structure and computation	SOC NETWORKS
Batagelj, V (1992)	Direct and indirect methods for structural equivalence	SOC NETWORKS
Faust, K (1992)	Blockmodels - interpretation and evaluation	SOC NETWORKS
Doreian, P (1994)	Partitioning networks based on generalized concepts of equivalence	J MATH SOCIOL
Doreian, P (1996)	A partitioning approach to structural balance	SOC NETWORKS
Batagelj, V (1997)	Notes on blockmodeling	SOC NETWORKS
Batagelj, V (1999)	Partitioning approach to visualization of large graphs	LECT NOTES COMPUT S
Batagelj, V (2000)	Some analyses of Erdos collaboration graph	SOC NETWORKS
Newman, MEJ (2001)	Scientific collaboration networks. I. Network construction and fundamental results	PHYS REV E
Girvan, M (2002)	Community structure in social and biological networks	P NATL ACAD SCI USA
Ravasz, E (2002)	Hierarchical organization of modularity in metabolic networks	SCIENCE
Ravasz, E (2003)	Hierarchical organization in complex networks	PHYS REV E
Newman, MEJ (2003)	The structure and function of complex networks	SIAM REV

List of works on CPM path, 2

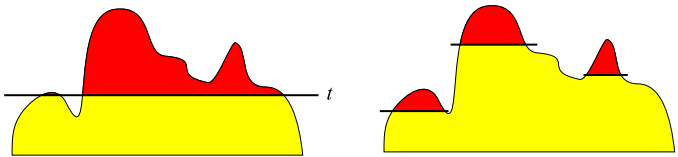
first author	title	journal
Newman, MEJ (2003)	Properties of highly clustered networks	PHYS REV E
Newman, MEJ (2003)	Why social networks are different from other types of networks	PHYS REV E
Radicchi, F (2004)	Defining and identifying communities in networks	P NATL ACAD SCI USA
Donetti, L (2004)	Detecting network communities: a new systematic and efficient algorithm	J STAT MECH
Guimera, R (2005)	Functional cartography of complex metabolic networks	NATURE
Duch, J (2005)	Community detection in complex networks using extremal optimization	PHYS REV E
Danon, L (2005)	COSIN book	–
Reichardt, J (2006)	Statistical mechanics of community detection	PHYS REV E
Fortunato, S (2007)	Resolution limit in community detection	P NATL ACAD SCI USA
Kumpula, JM (2007)	Limited resolution in complex network community detection with Potts model approach	EUR PHYS J B
Sales-Pardo, M (2007)	Extracting the hierarchical organization of complex systems	P NATL ACAD SCI USA
Arenas, A (2008)	Analysis of the structure of complex networks at different resolution levels	NEW J PHYS
Lancichinetti, A (2009)	Detecting the overlapping and hierarchical community structure of complex networks	NEW J PHYS
Nicosia, V (2009)	Extending the definition of modularity to directed graphs with overlapping communities	J STAT MECH
Evans, TS (2009)	Line graphs, link partitions, and overlapping communities	PHYS REV E
Fortunato, S (2010)	Community detection in graphs	PHYS REP
Good, BH (2010)	Performance of modularity maximization in practical contexts	PHYS REV E
Ball, B (2011)	Efficient and principled method for detecting communities in networks	PHYS REV E
Decelle, A (2011)	Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications	PHYS REV E
Peixoto, TP (2012)	Entropy of stochastic blockmodel ensembles	PHYS REV E
Peixoto, TP (2013)	Parsimonious Module Inference in Large Networks	PHYS REV LETT
Peixoto, TP (2014)	Efficient Monte Carlo and greedy heuristic for the inference of stochastic block models	PHYS REV E
Peixoto, TP (2014)	Hierarchical Block Structures and High-Resolution Model Selection in Large Networks	PHYS REV X
Larremore, DB (2014)	Efficiently inferring community structure in bipartite networks	PHYS REV E
Peixoto, TP (2015)	Model Selection and Hypothesis Testing for Large-Scale Network Models with Overlapping Groups	PHYS REV X
Peixoto, TP (2015)	Inferring the mesoscale structure of layered, edge-valued, and time-varying networks	PHYS REV E
Hric, D (2016)	Network Structure, Metadata, and the Prediction of Missing Nodes and Annotations	PHYS REV X
Fortunato, S (2016)	Community detection in networks: A user guide	PHYS REP

Key-rout paths (150 largest weights)



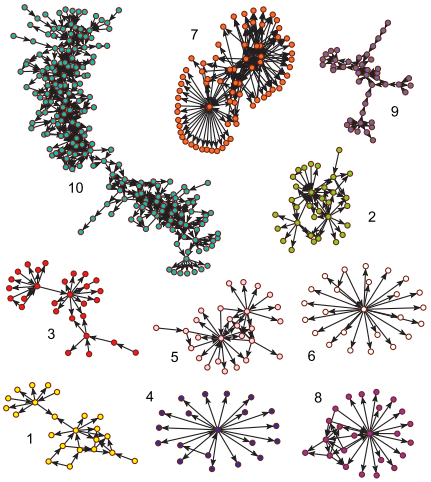
Islands

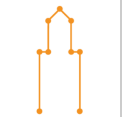
If we represent a given or computed property of nodes / lines as a height of nodes / lines and we immerse the network into a water up to selected property level we get *cuts*. Varying the level we get different *islands* (connected subnetworks).



Batagelj et al. (2003) developed very efficient algorithms to determine the islands hierarchy and to list all the islands of selected sizes.

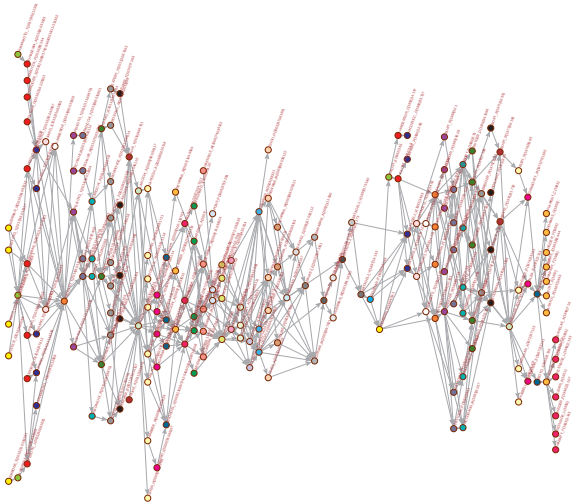
Ten identified SPC line islands [20, 150] in the clustering network literature





SPC – Line island10 on blockmodeling and community detection

$w_{max} = 0.5785$

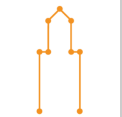


This island is very similar to the key-routh paths for 150 largest weights and includes the main path.

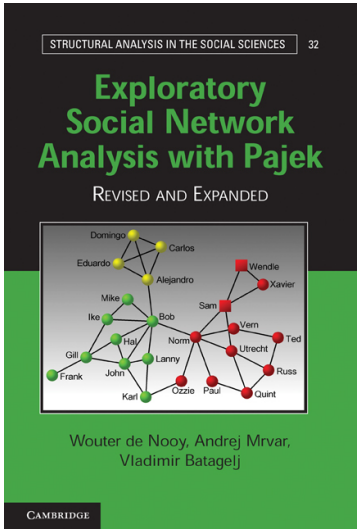
All other islands have maximal SPC weight 10^{-14} or less. Between them there are islands on: engineering geology, earthquake modeling, electromagnetic fields and their impact on humans.

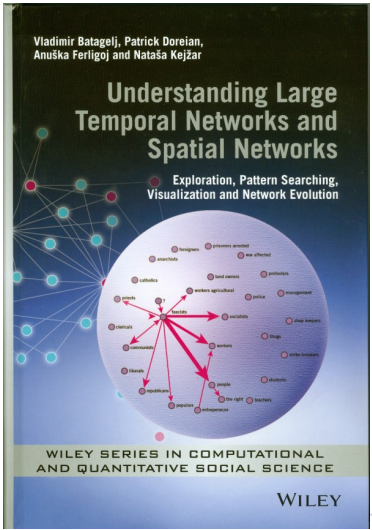
Limitations and extensions

- WoS is quite limited in the information it provides for individual works. High proportion of the works does not have complete descriptions in WoS. 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. For those in the social networks field, the term 'blockmodel' is very well known. But, for the works 'block model' means something quite different. E.g., 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'.
- Examining temporal shifts in the keywords and the journals where works are published are important avenues of exploration for understanding the dynamics of scientific fields.
- The structures of the islands are quite different. An open problem is whether this has an impact on the production of knowledge and the social organization of scientific disciplines.



Support


STRUCTURAL ANALYSIS IN THE SOCIAL SCIENCES
32
Exploratory Social Network Analysis with Pajek
REVISED AND EXPANDED
Domingo
Carlos
Eduardo
Alejandro
Wendle
Xavier
Mike
Sam
Ike
Bob
Vern
Hal
Norm
Ted
Utrecht
Gil
Lanny
Frank
John
Russ
Karl
Ozzie
Paul
Quint
Wouter de Nooy, Andrej Mrvar, Vladimir Batagelj
CAMBRIDGE


Vladimir Batagelj, Patrick Doreian, Anuška Ferligoj and Nataša Kejžar
Understanding Large Temporal Networks and Spatial Networks
Exploration, Pattern Searching, Visualization and Network Evolution
WILEY SERIES IN COMPUTATIONAL AND QUANTITATIVE SOCIAL SCIENCE
WILEY

An extended version of this presentation will be published as Chapter 2 in the book:

Doreian, P., Batagelj, V. and Ferligoj, A. (Eds):
Advances in Network Clustering and Blockmodeling. Wiley, 2018