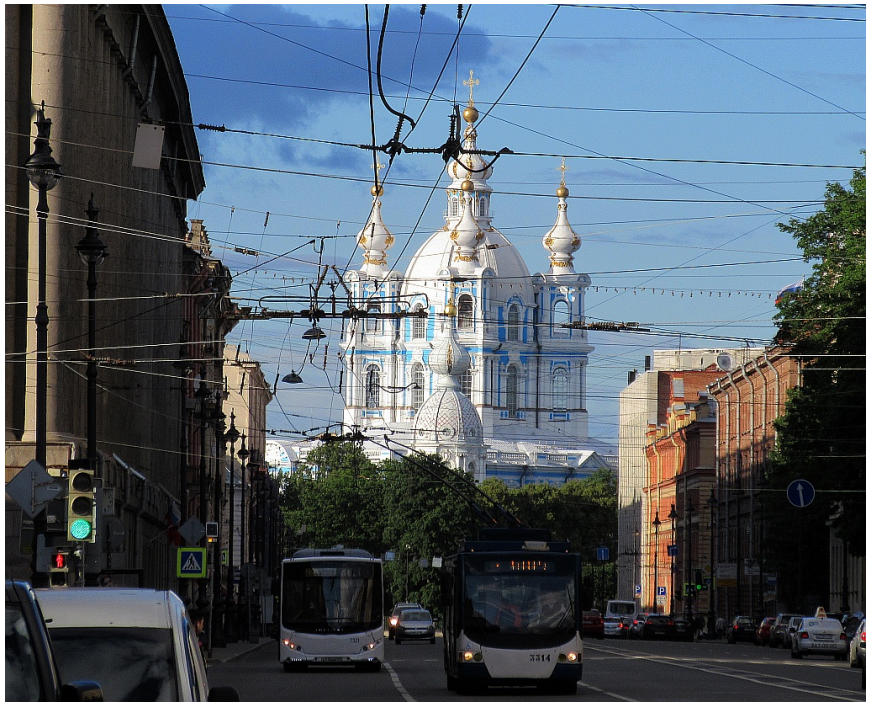
Fractional bibliographic coupling and co-citation

Vladimir Batagelj
IMFM Ljubljana and IAM UP Koper

NetGloW 2018
St Petersburg, July 4-6, 2018

Outline

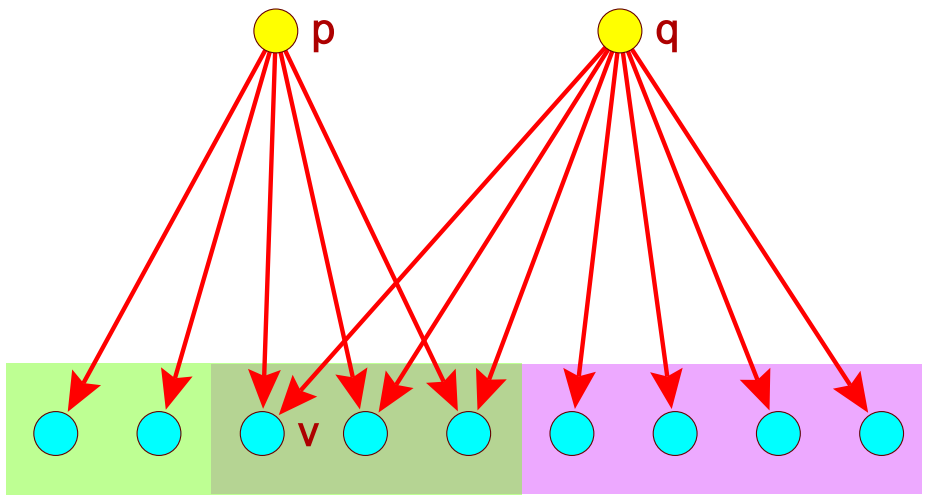


Vladimir Batagelj: vladimir.batagelj@fmf.uni-lj.si

Current version of slides (July 4, 2018 at 15 : 36): slides PDF
https://github.com/bavla/biblio/blob/master/doc/WS/fractional.pdf

Bibliographic Coupling

Bibliographic coupling occurs when two works each cite a third work in their bibliographies. The idea was introduced by Kessler (1963) and has been used extensively since then. See figure 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 for the three works cited by both p and q. Having more works citing pairs of prior works increases the likelihood of them sharing content.



Bibliographic Coupling

In `WoS2Pajek` the citation relation means

p **Ci** $q \equiv$ work p cites work r.

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: $bico_{pq} = bico_{qp}$:

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

Example: clustering networks

We obtained bibliographic data from the Web of Science (WoS) 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*"
```

Using WoS2Pajek we created the correspponding collection of networks – 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. Number of works with complete description (hits) is 5695.

Example

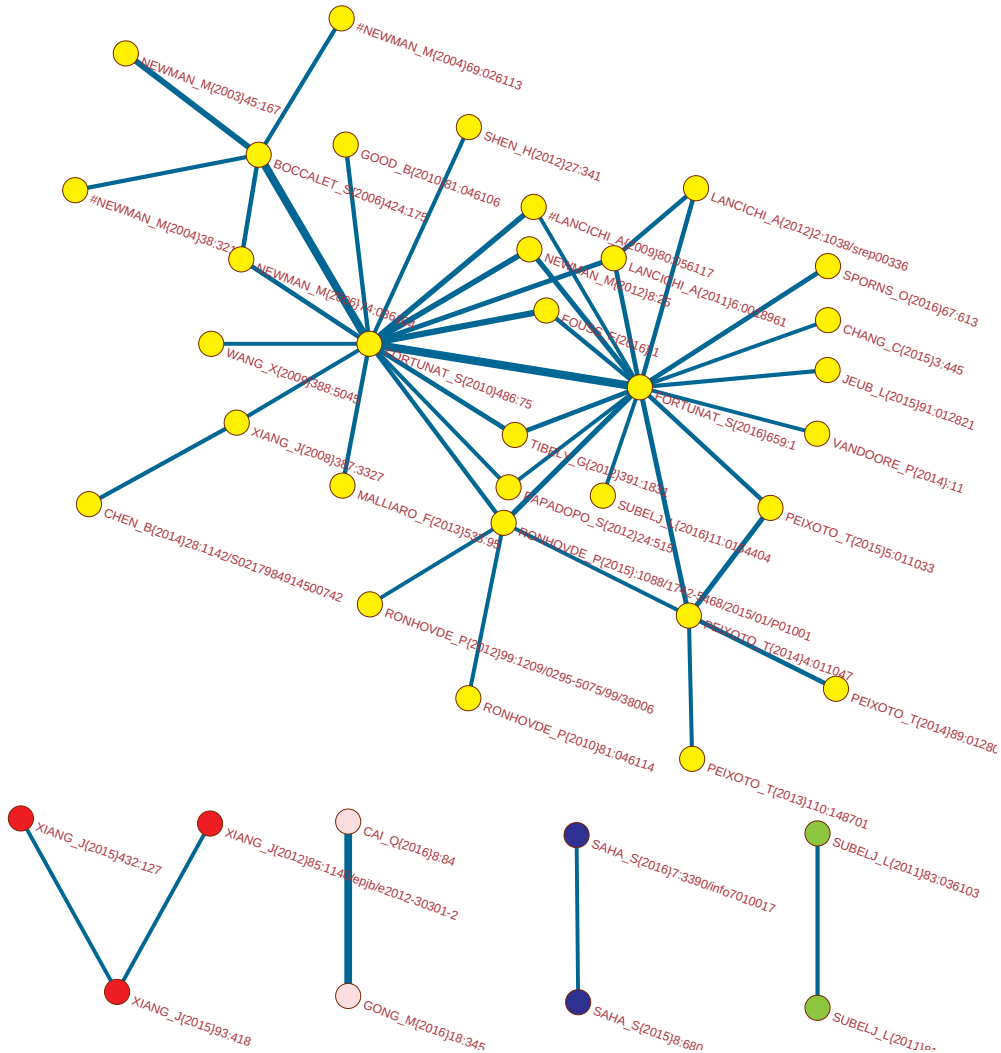
Pairs with the largest value

- overview works
- same author works

w(FORTUNAT_S(2010)486:75, FORTUNAT_S(2016)659:1) = 53
w(FORTUNAT_S(2010)486:75, BOCCALET_S(2006)424:175) = 51
w(CAI_Q(2016)8:84, GONG_M(2016)18:345) = 50
w(FORTUNAT_S(2010)486:75, FOUSS_F(2016):1) = 40
w(BOCCALET_S(2006)424:175, NEWMAN_M(2003)45:167) = 38

Bibliographic Coupling

cut at level 25



Fractional bibliographic coupling

Again we have problems with works with many citations, especially with review papers. To neutralize their impact we can introduce normalized measures. Let's first look at

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

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)

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

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

Fractional bibliographic coupling

Using **biC** we can construct different normalized measures such as

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

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

or, may be more interesting

$$\mathbf{biCog}_{pq} = \sqrt{\mathbf{biC}_{pq} \cdot \mathbf{biC}_{qp}} = \frac{|\mathbf{Ci}(p) \cap \mathbf{Ci}(q)|}{\sqrt{|\mathbf{Ci}(p)| \cdot |\mathbf{Ci}(q)|}} \quad \begin{matrix}\text{Geometric mean} \\ \text{Salton cosinus}\end{matrix}$$

$$\mathbf{biCoh}_{pq} = 2 \cdot (\mathbf{biC}_{pq}^{-1} + \mathbf{biC}_{qp}^{-1})^{-1} = \frac{2|\mathbf{Ci}(p) \cap \mathbf{Ci}(q)|}{|\mathbf{Ci}(p)| + |\mathbf{Ci}(q)|} \quad \text{Harmonic mean}$$

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

All these measures are symmetric.

Fractional bibliographic coupling

It is easy to verify that $biCoX_{pq} \in [0,1]$ and: $biCoX_{pq} = 1$ iff the works p and q are referencing the same works, $\mathbf{Ci}(p) = \mathbf{Ci}(q)$.

From $H \leq G \leq A$ and $J = \frac{H}{2-H}$, $2 - H \geq 1$ we get

$$\mathbf{biCom}_{pq} \leq \mathbf{biCoj}_{pq} \leq \mathbf{biCoh}_{pq} \leq \mathbf{biCog}_{pq} \leq \mathbf{biCoa}_{pq} \leq \mathbf{biCoM}_{pq}$$

The equalities hold iff $\mathbf{Ci}(p) = \mathbf{Ci}(q)$.
To get a dissimilarity use $dis = 1 - sim$ or $dis = \frac{1}{sim} - 1$ or $dis = -\log sim$. For example

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

Jaccard islands

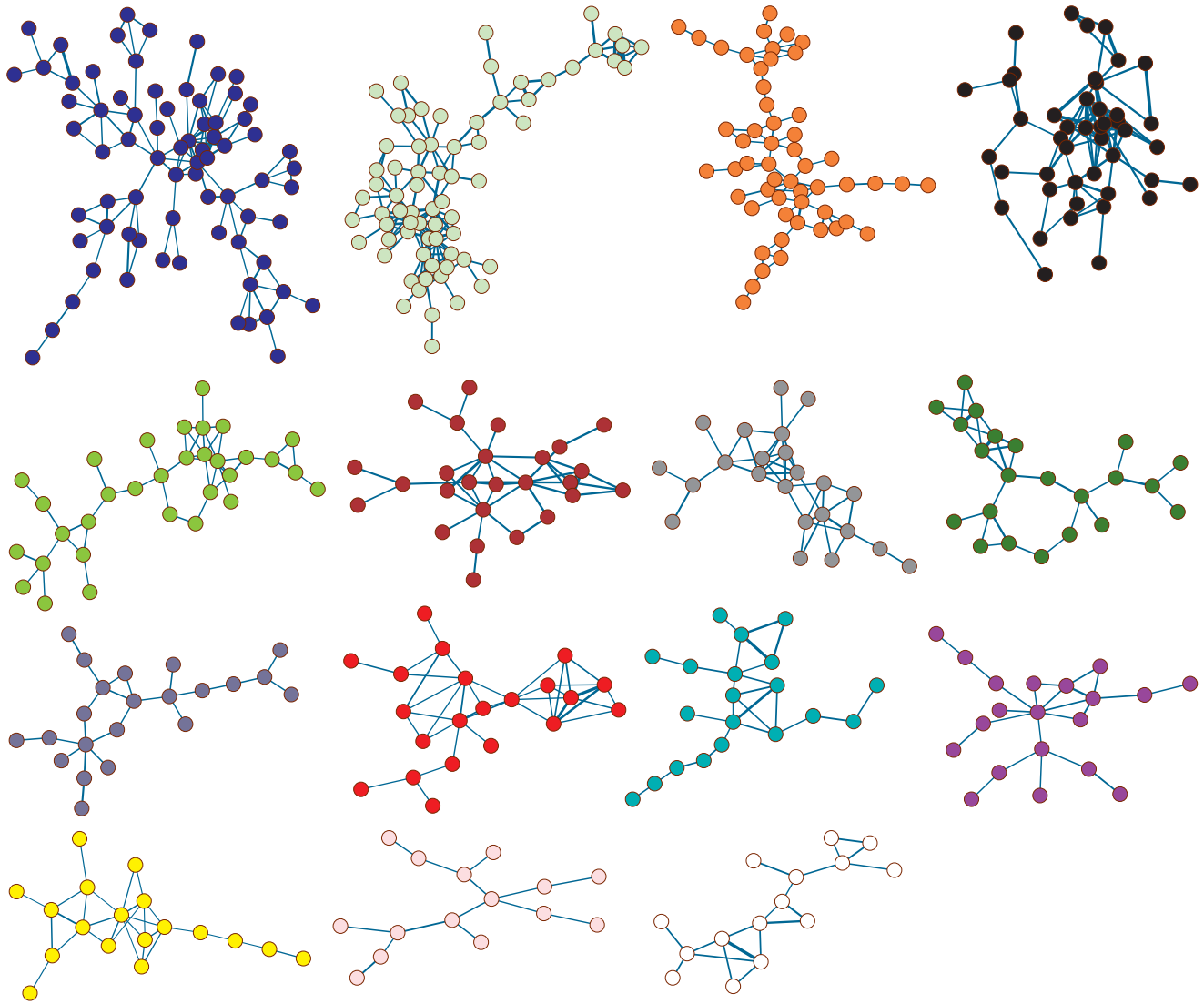
We computed Jaccard similarity measures for the network CiteB and determined corresponding link islands having sizes in the range [5,75]. The following table shows the distribution of sizes of 133 islands that were identified.

size	5	6	7	8	9	10	11	12	13	14	15	17	18	24	27
num	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
num	1	2	1	1	1	1	1	1	2	1	1	1	1	1

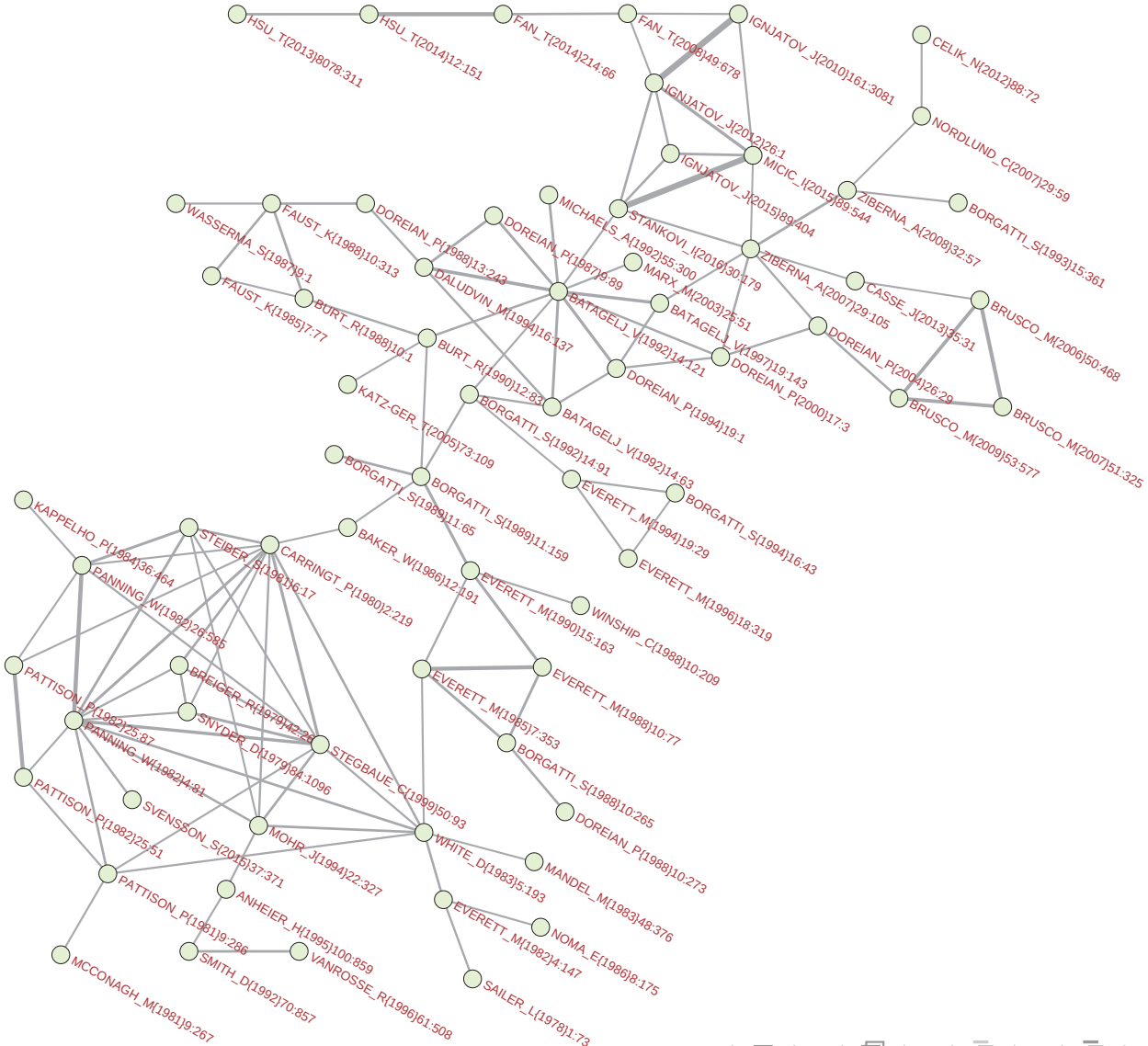
Fractional bibliographic coupling

some Jaccard islands



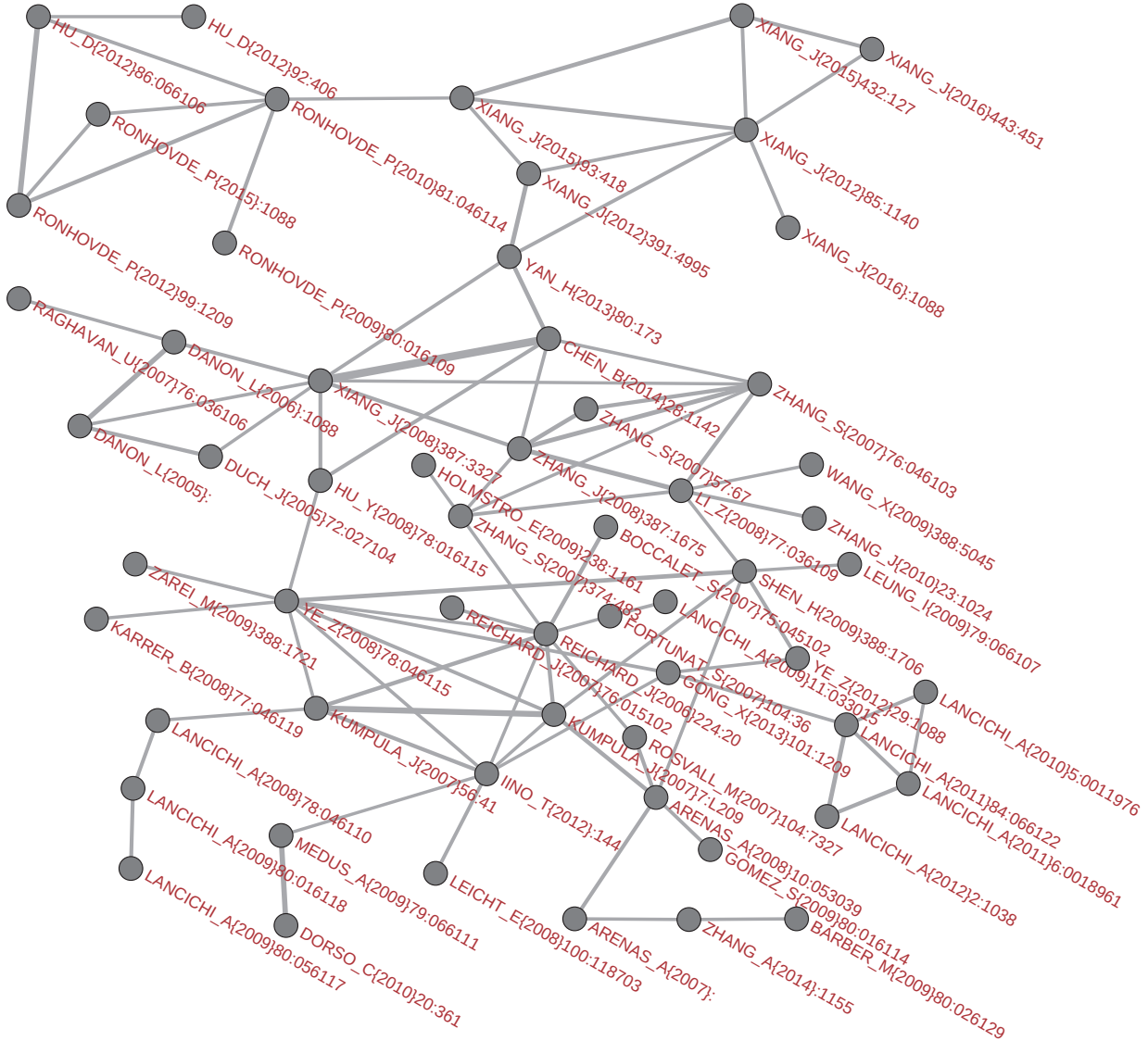
Fractional bibliographic coupling

in the social networks literature



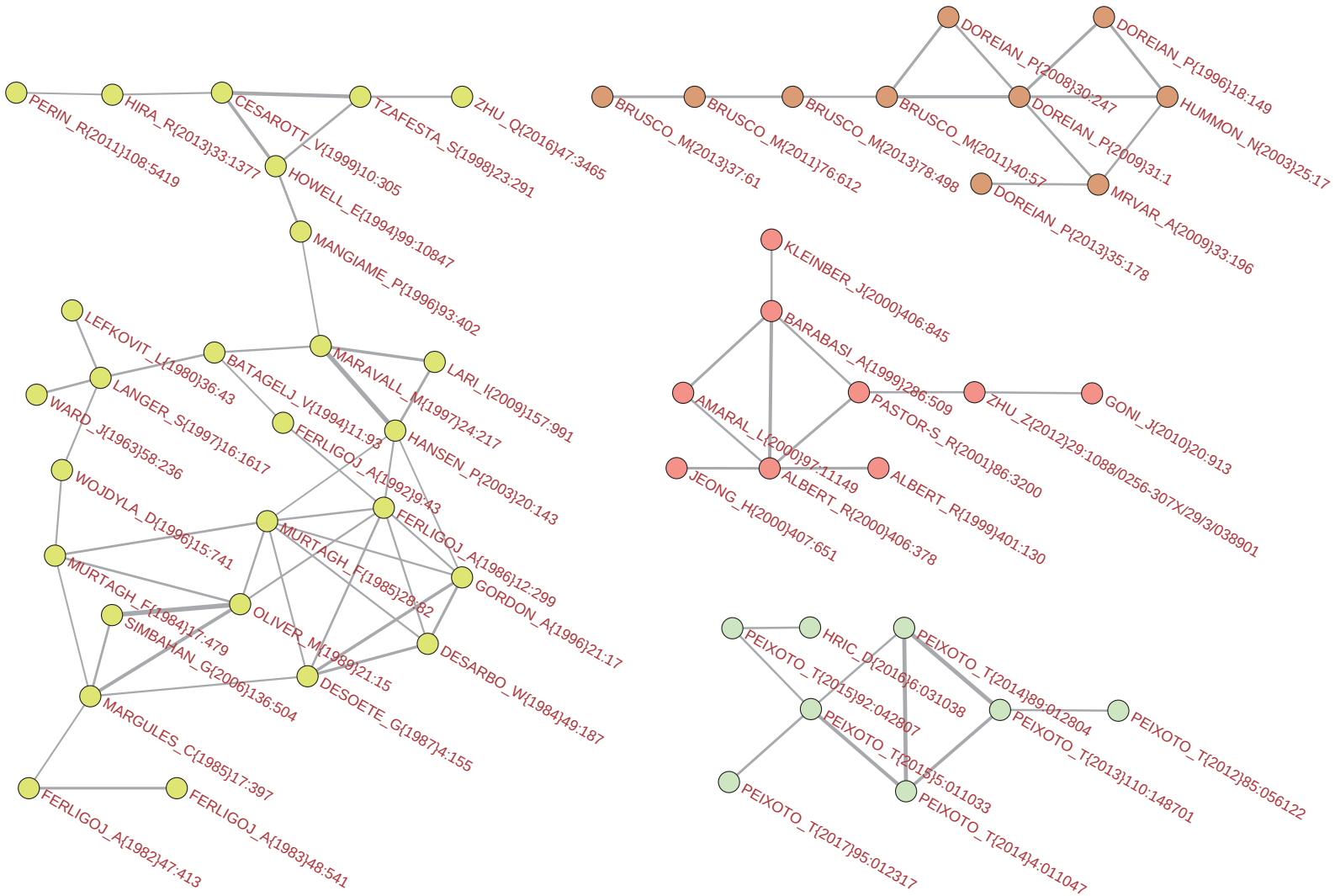
Fractional bibliographic coupling

in the physicist-driven literature



Fractional bibliographic coupling

selected islands



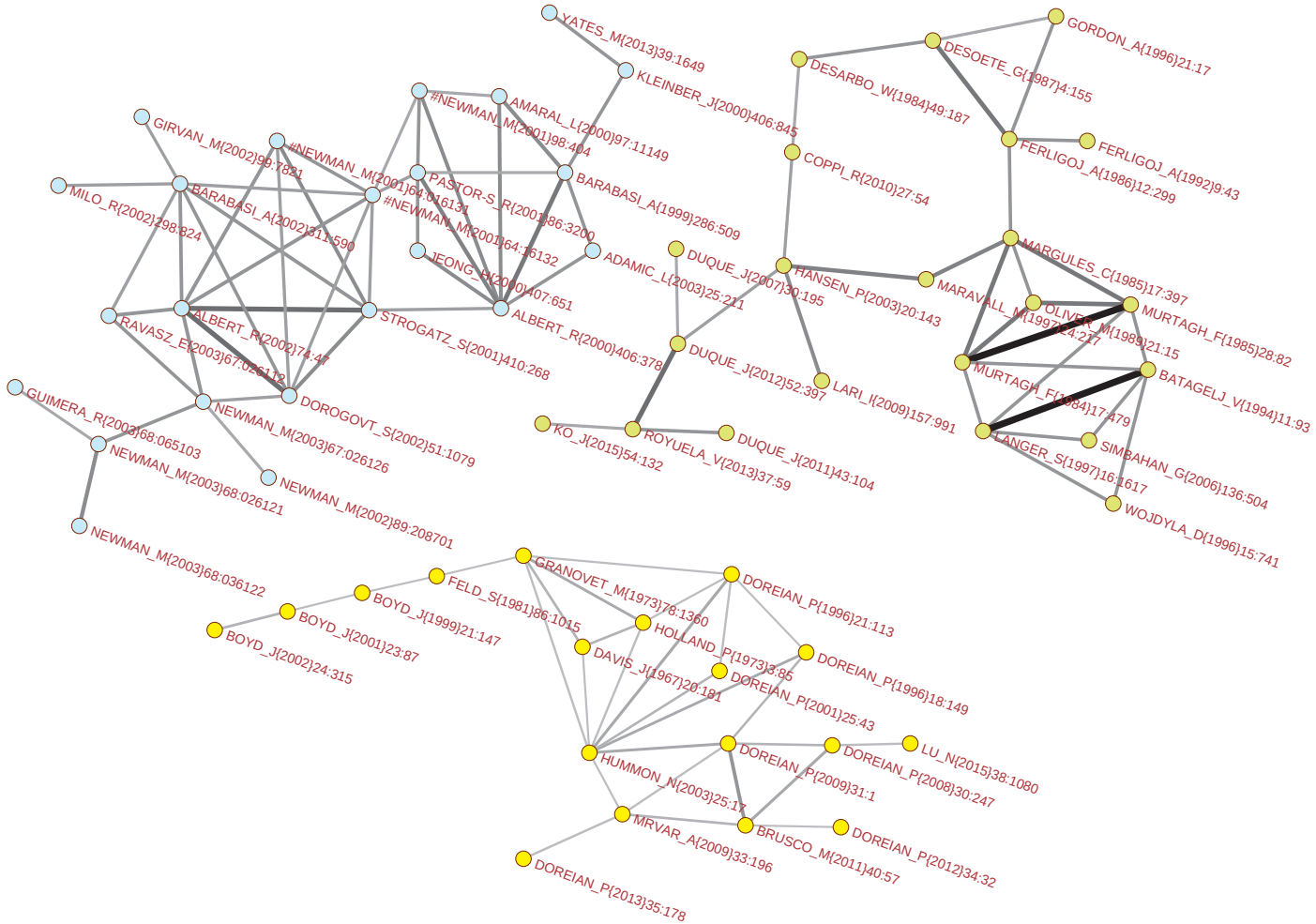
Fractional bibliographic coupling

the most cited works from works of the two largest islands

Social network literature			(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

Fractional bibliographic coupling

for three smaller islands



Fractional bibliographic coupling

the most cited works from works from three smaller islands

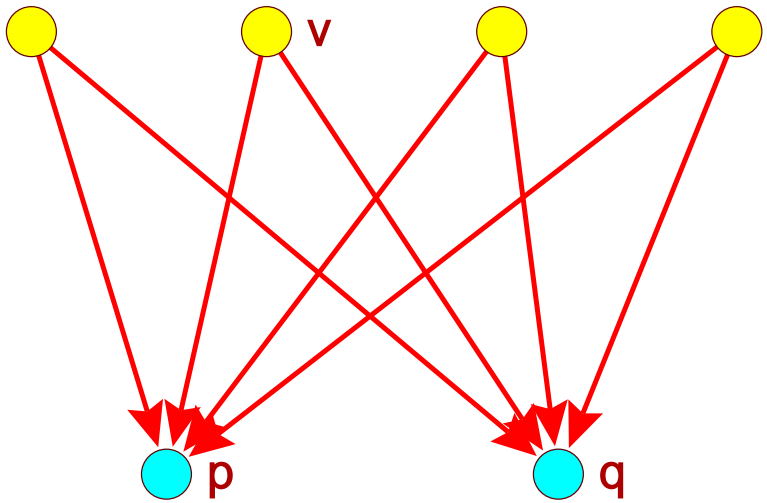
n		Physicist literature		Clustering literature		Signed networ
1	23	WATTS_D(1998)393:440	21	FERLIGOJ_A(1982)47:413	13	CARTWRIG_D(1
2	18	BARABASI_A(1999)286:509	11	LEFKOVIT_L(1980)36:43	12	HEIDER_F(1946
3	17	ALBERT_R(1999)401:130	10	PERRUCHE_C(1983)16:213	11	DAVIS_J(1967)2(
4	15	WASSERMA_S(1994):	9	MURTAGH_F(1985)28:82	10	NEWCOMB_T(1
5	15	AMARAL_L(2000)97:11149	8	FERLIGOJ_A(1983)48:541	9	WHITE_H(1976)
6	13	BOLLOBAS_B(1985):	6	GORDON_A(1996)21:17	8	HARARY_F(196
7	13	FALOUTSO_M(1999)29:251	4	DUQUE_J(2007)30:195	8	DOREIAN_P(199
8	13	NEWMAN_M(2001)98:404	4	KIRKPATR_S(1983)220:671	7	DOREIAN_P(200
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(197
11	10	REDNER_S(1998)4:131	3	MARGULES_C(1985)17:397	6	HOMANS_G(195
12	9	JEONG_H(2000)407:651	3	HANSEN_P(2003)20:143	6	BATAGELJ_V(19
13	9	ALBERT_R(2000)406:378	3	DUQUE_J(2011)43:104	5	BORGATTI_S(20
14	9	MOLLOY_M(1995)6:161	3	MARAVALL_M(1997)24:217	5	LORRAIN_F(197
15	9	MILGRAM_S(1967)1:61	3	GAREY_M(1979):	5	WHITE_D(1983)

Fractional bibliographic coupling

the most frequent keywords in works of a given subnetworks

Social network literature			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

Co-Citation



Co-citation is a concept with strong parallels with bibliographic coupling (Small and Marshakova 1973). 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

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

.

Co-Citation

$coci_{pq} = \#$ of works citing both works p and q.

$coci_{pq} = coci_{qp}$.

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