Fractional approach to bibliographic coupling and co-citation

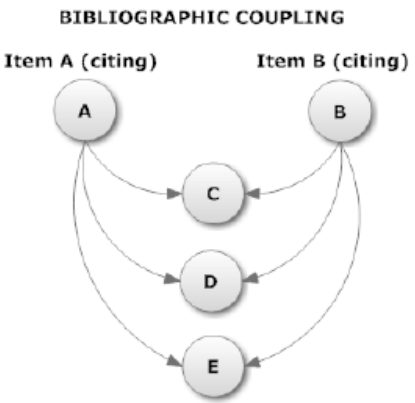
Vladimir Batagelj

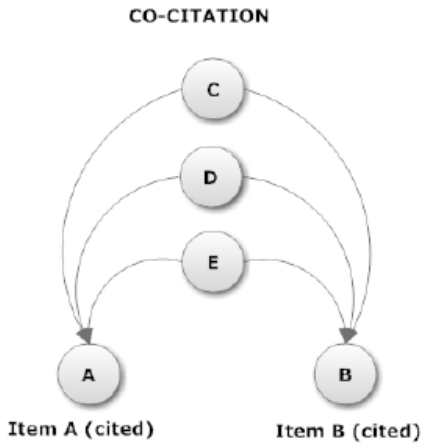
IMFM Ljubljana and IAM UP Koper

1275. Sredin seminar

Ljubljana, 19. April 2017

Outline

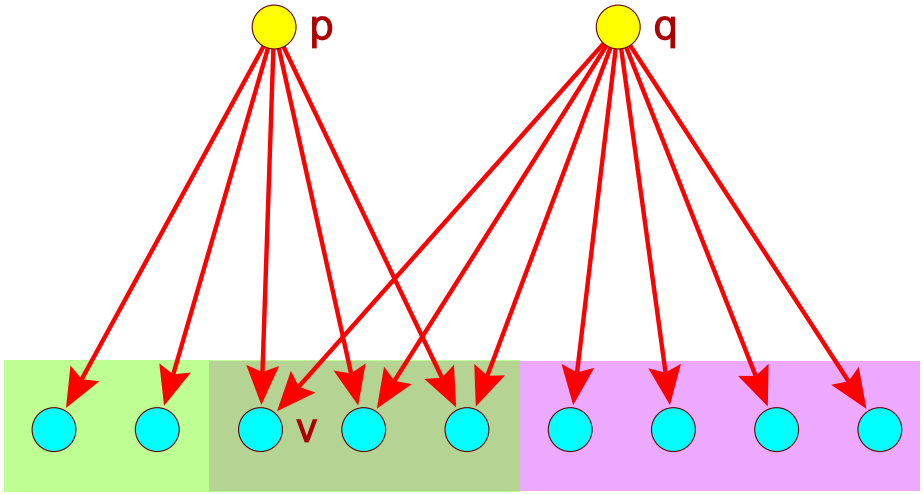




Vladimir Batagelj: vladimir.batagelj@fmf.uni-lj.si

Version (April 23, 2017, 04 : 27): sreda1275.pdf

Bibliographic Coupling



In `WoS2Pajek` the citation relation means p **Ci** $q \equiv$ work p cites work q.

Therefore the *bibliographic coupling* (Kessler, 1963) 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: $bico_{pq} = bico_{qp}$:

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

Bibliographic Coupling

macro biCo

```
select / read the citation network
Network/Create new network/Transform/Transpose 1-mode
select transposed as Second network
select citation network as First
Networks/Multiply networks
Network/Create new network/Transform/Remove/Loops
Network/Create new network/Transform/Arcs->Edges/Bidirect only/Min
```

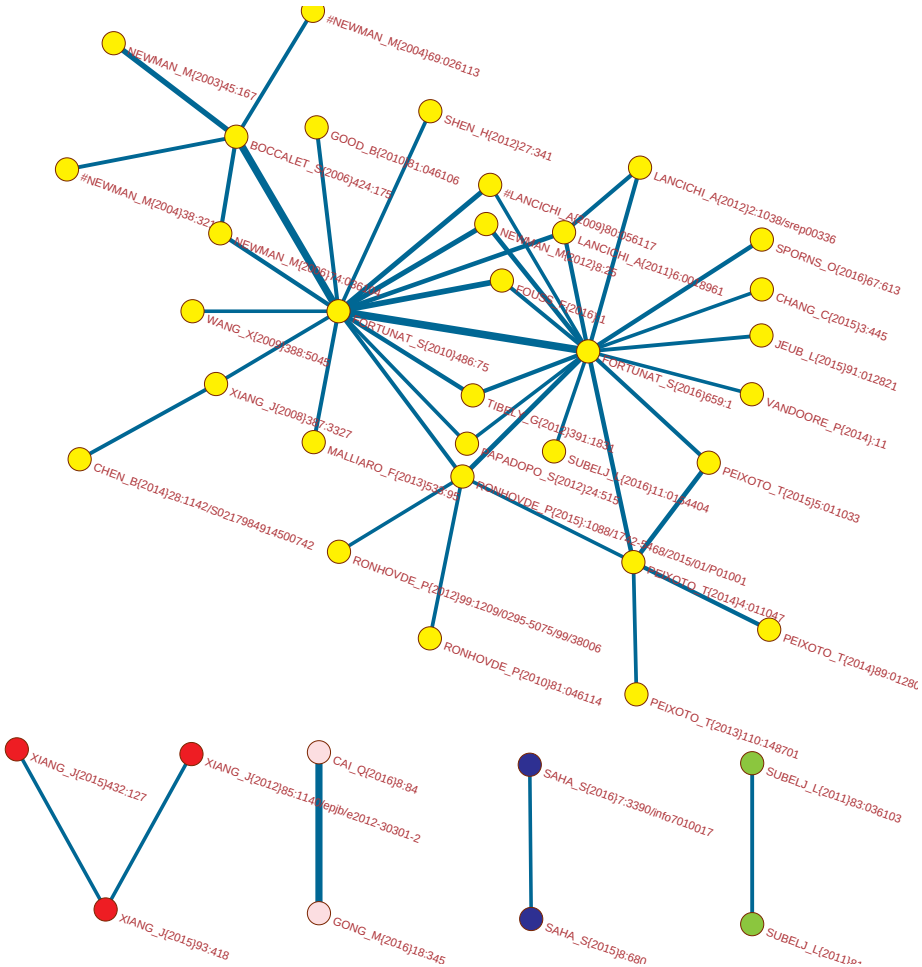
Pairs with the largest value

- overview works
- same author works

w(FORTUNAT_S2010486:75, FORTUNAT_S2016659:1) = 53
w(FORTUNAT_S2010486:75, BOCCALET_S2006424:175) = 51
w(CAI_Q20168:84, GONG_M201618:345) = 50
w(FORTUNAT_S2010486:75, FOUSS_F2016:1) = 40
w(BOCCALET_S2006424:175, NEWMAN_M200345:167) = 38

Bibliographic Coupling

cut at level 25



Bibliographic Coupling

fractional approach

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 $\mathbf{biC}_{pq} \in [0, 1]$.

Bibliographic Coupling

fractional approach

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.

Bibliographic Coupling

fractional approach

$\mathbf{biC}_{pq}$ is the proportion of its references the work p shares with the work q.

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

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

Question: compute bibliographic coupling on original Cite or boundary CiteB ? Essentially we can use the reduced network with original outdegrees.

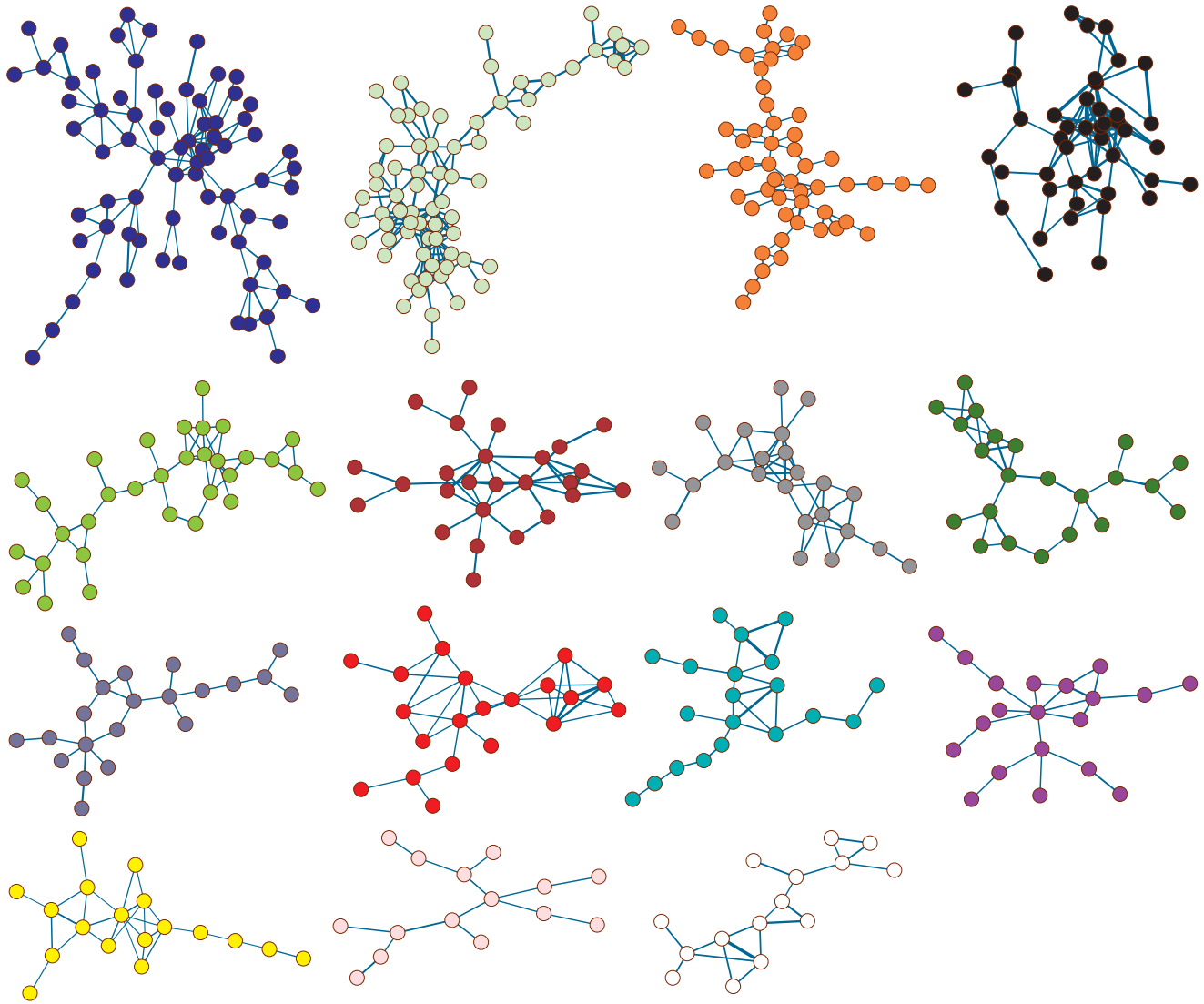
Bibliographic Coupling

macro biCon

```
select citation network Cite
Network/Create Vector/Centrality/Degree/Output = V1
Vector/Create Constant Vector [n,1] = V2
select V1 as Second vector
Vectors/Max(First,Second)
Vector/Transform/Invert
Network/Create new network/Transform/Transpose 1-mode = CiteT
select network Cite as First
select network CiteT as Second
Networks/Multiply networks = biCo
Operations/Network+Vector/Vector#Network/Output
Network/Create new network/Transform/Remove/Loops = biC
Network/Create new network/Transform/Line values/Power [-1]
Network/Create new network/Transform/Arcs->Edges/Bidirected only/Sum
Network/Create new network/Transform/Line values/Add constant [-1]
Network/Create new network/Transform/Line values/Power [-1] = Jaccard
Network/Create new network/Transform/Line values/Multiply by [-1]
Network/Create new network/Transform/Line values/Add constant [1] = Hamm
```

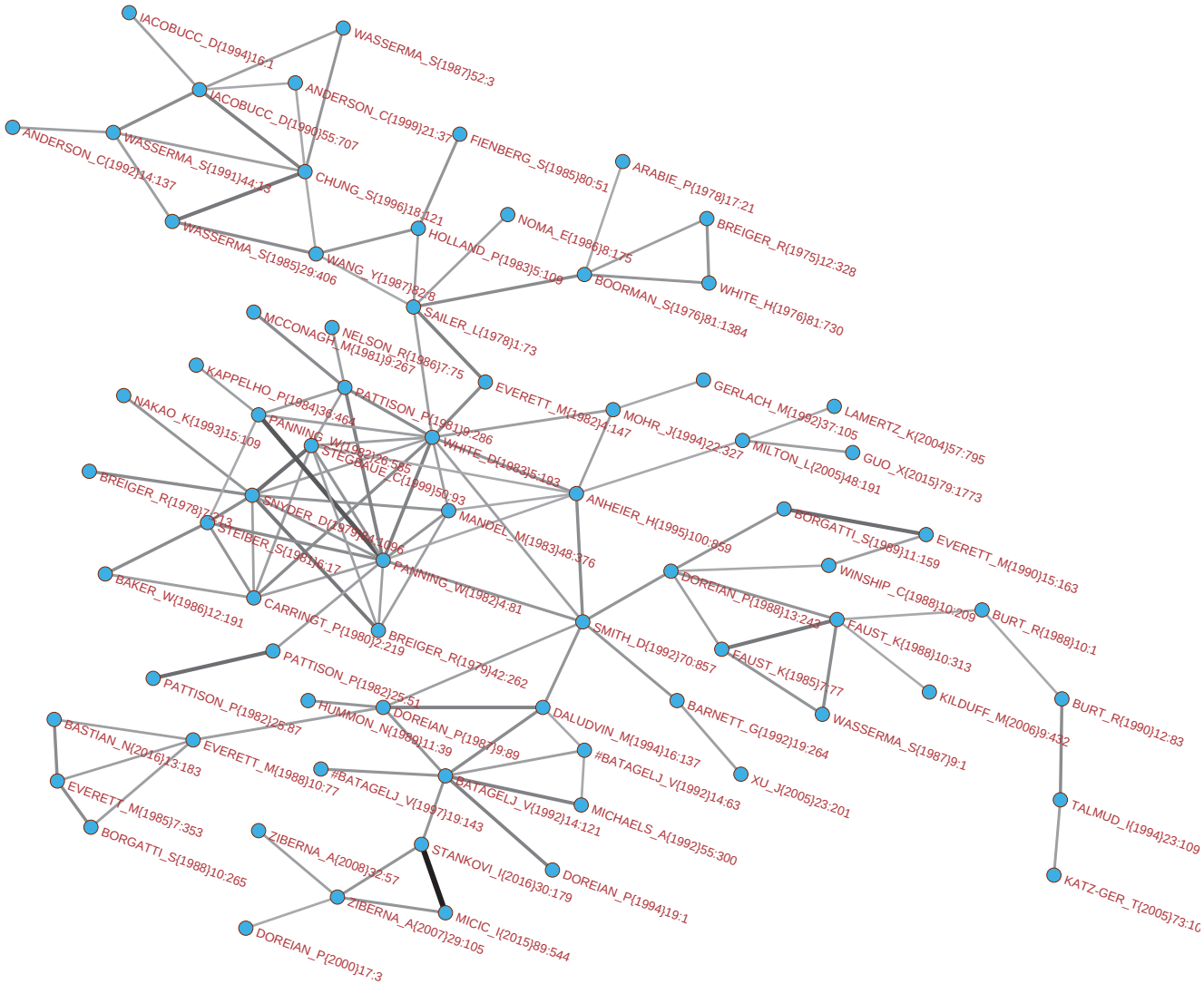
Bibliographic Coupling

Jaccard islands [15, 75]



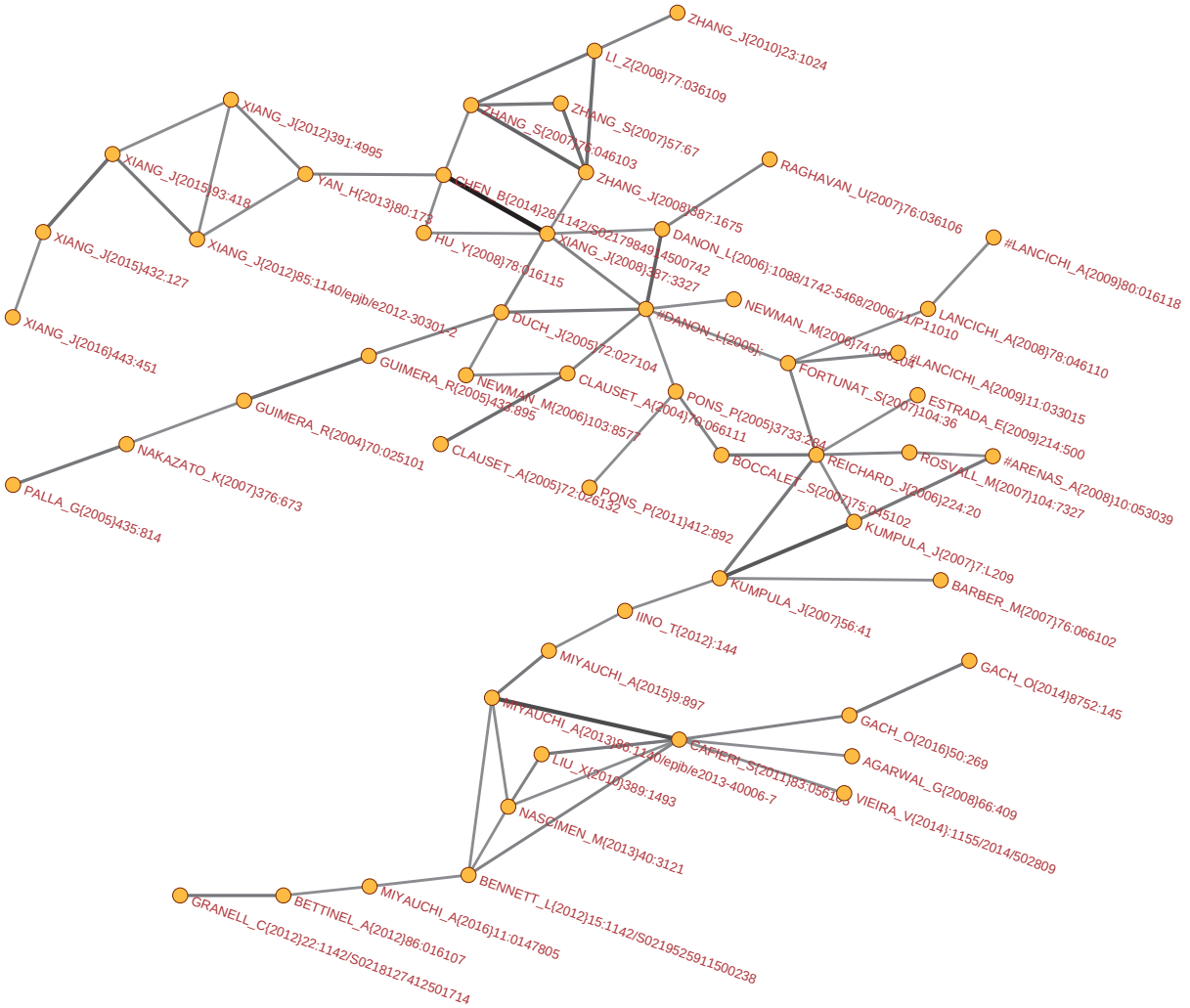
Bibliographic Coupling

Jaccard island 4 (74)



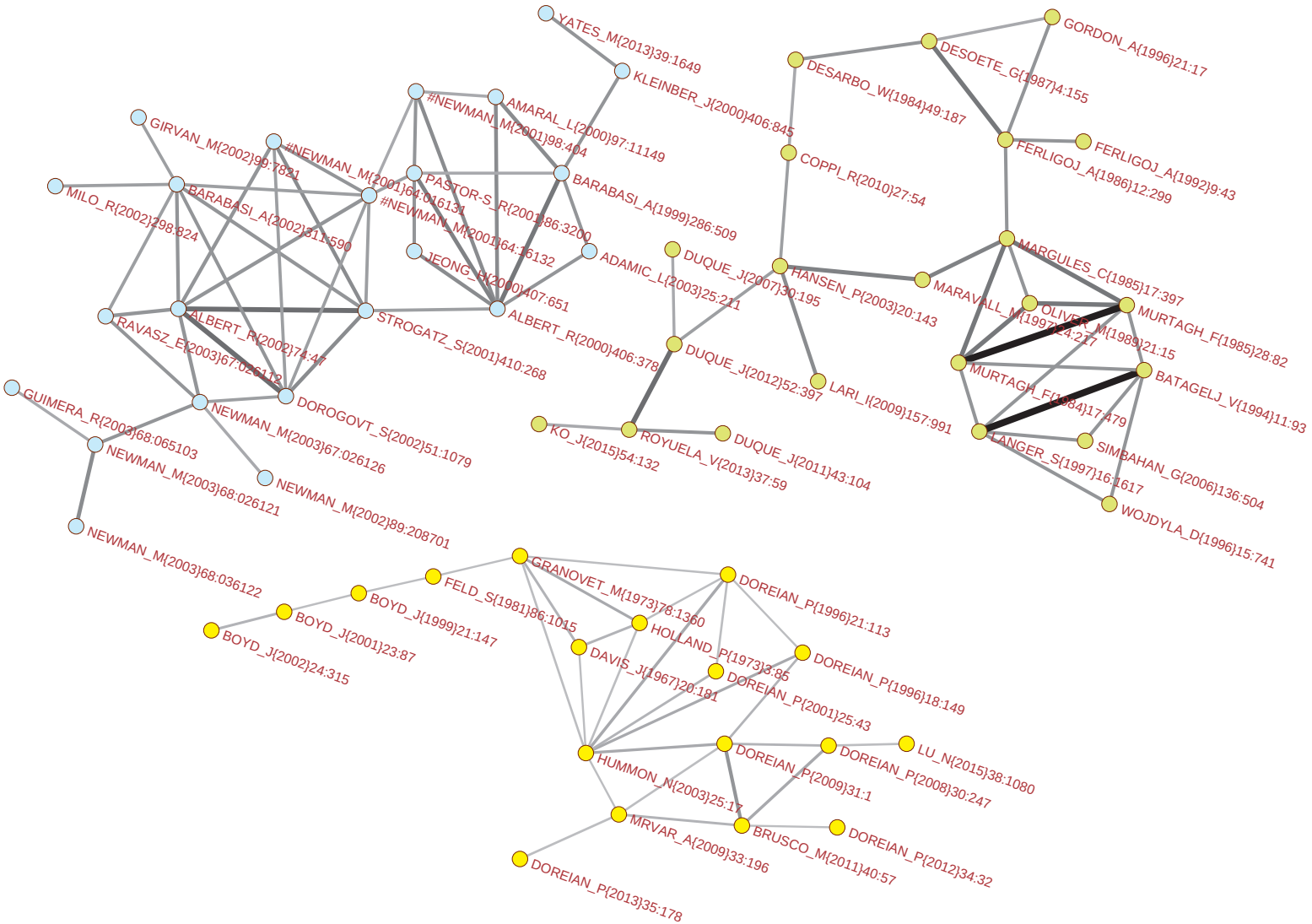
Bibliographic Coupling

Jaccard island 7 (54)



Bibliographic Coupling

Jaccard islands 12 (23), 11 (22), 1 (18)



Bibliographic Coupling

the most cited works from works of a given subnetwork

For titles of works from an island see the CSV files obtained in R using the function description

```
setwd("C:/Users/batagelj/work/Python/WoS/BM/results/jaccard")
source("C:\\Users\\batagelj\\work\\Python\\WoS\\peere1\\description.R")
T <- read.csv('../../titles.csv',sep=";",colClasses="character")
T$code <- 1
dim(T)
d <- description("Jisland4.net","Jisland4.csv",T)
head(d)
d <- description("Jisland7.net","Jisland7.csv",T)
d <- description("Jisland12.net","Jisland12.csv",T)
```

```
select Island network as First
select citation network Cite as Second
Networks/Match vertex labels
select partition Positions of Second network in First
Partition/Binarize Partition [1-*]
Partition/Copy to Vector
select transposed network Cite
Operations/Network+Vector/Network*Vector [1]
info Vector [+30]
```

Bibliographic Coupling

the most cited works from works of a given island

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

Bibliographic Coupling

the most cited works from works of a given island

	Jisland 12		Jisland 11		Jisland 1
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

Bibliographic Coupling

the most frequent keywords in works of a given subnetwork


```
select Island network as First
select  network WK as Second
Networks/Match vertex labels
select partition Positions of Second network in First
Partition/Binarize Partition [1-*]
Partition/Copy to Vector
select WK
Network/Two-mode network/Partition into 2 Modes
Operations/Vector+Partition/Extract Subvector [1]
Network/Two-mode network/Transpose 2-mode
Operations/Network+Vector/Network*Vector [1] = V1
Vector/Constant [n1,0] = V2
select V1 as First
select V2 as Second
Vectors/Fuse vectors
info Vector [+50]
```

The same approach can be applied to WA network.

Bibliographic Coupling

the most frequent keywords in works of a given island

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

Bibliographic Coupling

clustering

We restrict our analysis to the largest connected component HammondMain of Hammond network.

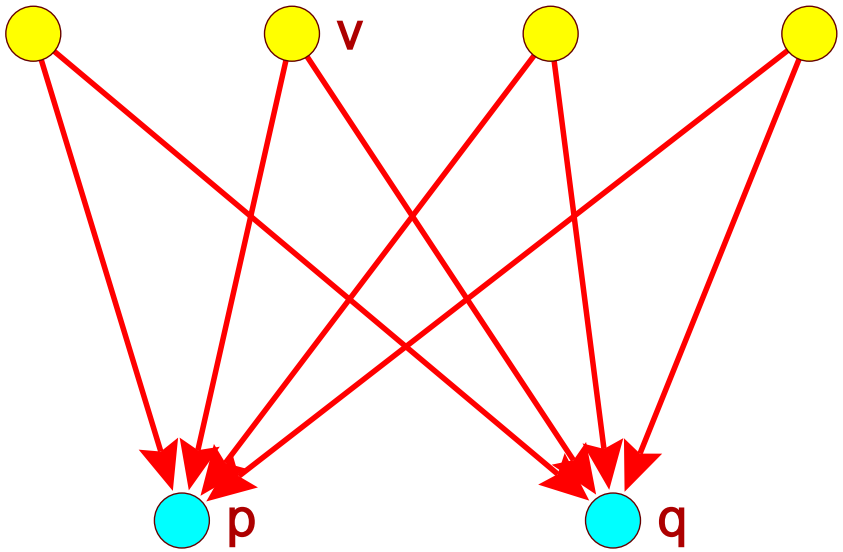
```
Network/Create Hierarchy/Clustering with Relational Constraints/Run [Max Tolerant]
save partition Clustering with relational constraint (tree) [Max/Tolerant] to MaxTol.clu
save vector Clustering with relational constraint (heights) [Max/Tolerant] to heightMaxTol.vec
save vector Clustering with relational constraint (size) [Max/Tolerant] to sizeMaxTol.vec
select HammondMain
Network/Create new network/Transform/Remove/All edges
save network to HammondNam.net
```

We continue in R.

```
> setwd("C:\\Users\\batagelj\\work\\Python\\WoS\\BM\\results\\cluster")
> source("C:\\Users\\batagelj\\work\\Python\\WoS\\BM\\results\\cluster\\Pajek2R.R")
> source("C:\\Users\\batagelj\\work\\Python\\WoS\\BM\\results\\cluster\\varCutTree.R")
> RC <- Pajek2R("MaxTol.clu")
> n <- RC$n; nm <- n-1; np <- n+1
> rCount <- varCutree(RC,rep(1,n),5,400)
> t <- table(rCount$part)
> out <- file("CMaxTot1.clu","w")
> cat(paste("*vertices ",n),rCount$part,sep="\n",file=out); close(out)
> t
  1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20
375  85  35  28  65 354  94  48  55 386  23 228  34   6  80 103  16  11   8  27
 21  22  23  24  25  26  27  28  29  30  31  32  33  34
 82 335  17 265   6  37 204  28 234  68  80 159  23 209
```

!!! make clusters from JaccardMain !!!

Co-Citation



The *co-citation* (Small, Marshakova, 1973) network **coCi** can be determined as

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

.

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

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

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

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

Others

The weight $w(a, p)$ in the *author citation* network

$$\mathbf{ACi} = \mathbf{AW} * \mathbf{Ci}$$

counts the number of times author a cited work p.
The *author co-citation* network can be obtained as

$$\mathbf{ACo} = b(\mathbf{ACi}) * t(b(\mathbf{ACi}))$$

Authors using keywords $\mathbf{AK} = \mathbf{AW} * \mathbf{WK}$.