Introduction to Network Analysis using **Pajek**

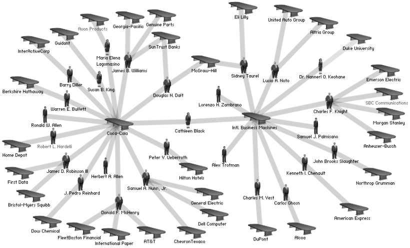
7. Two-mode networks and multiplication

Vladimir Batagelj

IMFM Ljubljana and IAM UP Koper

PhD and MS program in Statistics
University of Ljubljana, 2022

Outline



Josh On: They rule 2004

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

Current version of slides (March 23, 2022 at 00 : 33): slides PDF

Two-mode networks

In a *two-mode* network $\mathcal{N} = (\mathcal{U}, \mathcal{V}, \mathcal{L}, \mathcal{P}, \mathcal{W})$ the set of nodes consists of two disjoint sets of nodes $\mathcal{U}$ and $\mathcal{V}$, and all the lines from $\mathcal{L}$ have one end-node in $\mathcal{U}$ and the other in $\mathcal{V}$. Often also a *weight* $w : \mathcal{L} \to \mathbb{R} \in \mathcal{W}$ is given; if not, we assume $w(u, v) = 1$ for all $(u, v) \in \mathcal{L}$.
A two-mode network can also be described by a rectangular matrix $\mathbf{A} = [a_{uv}]_{\mathcal{U}\times\mathcal{V}}$.

$$a_{uv} = \begin{cases} w_{uv} & (u, v) \in \mathcal{L} \\ 0 & \text{otherwise} \end{cases}$$

Examples: (persons, societies, years of membership), (buyers/consumers, goods, quantity), (parlamentarians, problems, positive vote), (persons, journals, reading), (papers, keywords, is described by), etc.

Deep South

Classical example of two-mode network are the Southern women (Davis 1941).
Davis.paj. Freeman's overview.

NAMES OF PARTICIPANTS OF GROUP I	CODE NUMBERS AND DATES OF SOCIAL EVENTS REPORTED IN *Old City Herald*													
	(1) 6/27	(2) 3/2	(3) 4/12	(4) 9/26	(5) 2/25	(6) 5/19	(7) 3/15	(8) 9/16	(9) 4/8	(10) 6/10	(11) 2/23	(12) 4/7	(13) 11/21	(14) 8/3
1. Mrs. Evelyn Jefferson	×	×	×	×	×	×		×	×					
2. Miss Laura Mandeville	×	×	×		×	×	×	×						
3. Miss Theresa Anderson		×	×	×	×	×	×	×	×					
4. Miss Brenda Rogers	×		×	×	×	×	×	×						
5. Miss Charlotte McDowd			×	×	×		×							
6. Miss Frances Anderson			×		×	×		×						
7. Miss Eleanor Nye					×	×	×	×						
8. Miss Pearl Oglethorpe						×		×	×					
9. Miss Ruth DeSand					×		×	×	×					
10. Miss Verne Sanderson							×	×	×			×		
11. Miss Myra Liddell								×	×	×		×		
12. Miss Katherine Rogers								×	×	×		×	×	×
13. Mrs. Sylvia Avondale							×	×	×	×		×	×	×
14. Mrs. Nora Fayette						×	×		×	×	×	×	×	×
15. Mrs. Helen Lloyd							×	×		×	×	×		
16. Mrs. Dorothy Murchison								×	×					
17. Mrs. Olivia Carleton									×		×			
18. Mrs. Flora Price									×		×			

Approaches to two-mode network analysis

The usual approach to analyze a two-mode network is to transform it to a one-mode network and use standard methods on it.

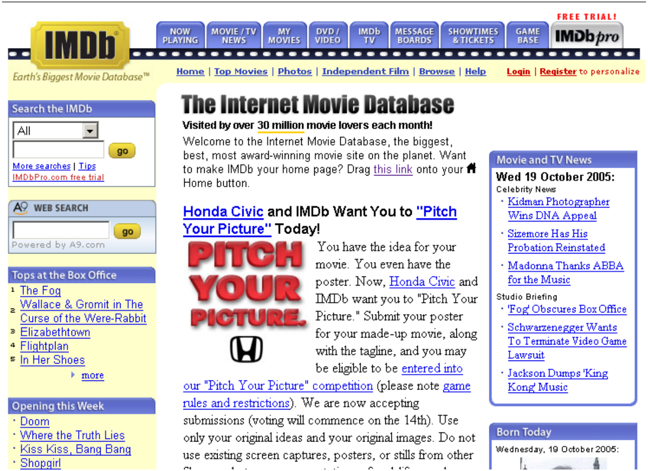
For direct analysis of two-mode networks we can use the *eigen-vector approach* – a two-mode variant of Kleinberg's hubs and authorities. The weight vector $(\mathbf{x}, \mathbf{y})$ on $\mathcal{U} \cup \mathcal{V}$ is determined by relations $\mathbf{y} = \mathbf{A}\mathbf{x}$ and $\mathbf{x} = \mathbf{A}^T\mathbf{y}$.

```
Network/2-Mode Network/Important Vertices
```

There are also special methods for *clustering* and *blockmodeling* in two-mode networks.

In this lecture we will present two additional direct methods: *two-mode cores* and *4-rings*.

Internet Movie Database http://www.imdb.com/



12th Annual Graph Drawing Contest, 2005. The IMDB network is two-mode and has $1324748 = 428440 + 896308$ nodes and 3792390 arcs.

Two-mode cores

The subset of nodes $C \subseteq \mathcal{V}$ is a *(p, q)-core* in a two-mode network $\mathcal{N} = (\mathcal{V}_1, \mathcal{V}_2; \mathcal{L})$, $\mathcal{V} = \mathcal{V}_1 \cup \mathcal{V}_2$ iff

a. in the induced subnetwork $\mathcal{K} = (C_1, C_2; \mathcal{L}(C))$, $C_1 = C \cap \mathcal{V}_1$, $C_2 = C \cap \mathcal{V}_2$ it holds $\forall v \in C_1 : \deg_{\mathcal{K}}(v) \geq p$ and $\forall v \in C_2 : \deg_{\mathcal{K}}(v) \geq q$;

b. C is the maximal subset of $\mathcal{V}$ satisfying condition **a.**

Properties of two-mode cores:

- $C(0, 0) = \mathcal{V}$
- $\mathcal{K}(p, q)$ is not always connected
- $(p_1 \leq p_2) \wedge (q_1 \leq q_2) \Rightarrow C(p_1, q_1) \subseteq C(p_2, q_2)$
- $\mathcal{C} = \{C(p, q) : p, q \in \mathbb{N}\}$. If all nonempty elements of $\mathcal{C}$ are different it is a lattice.

Algorithm for two-mode cores

To determine a (p, q)-core the procedure similar to the ordinary core procedure can be used:

repeat
remove from the first set all nodes of degree less than p,
and from the second set all nodes of degree less than q
until no node was deleted

It can be implemented to run in $O(m)$ time.
Interesting (p, q)-cores? Table of cores' characteristics $n_1 = |C_1(p, q)|$, $n_2 = |C_2(p, q)|$ and k – number of components in $\mathcal{K}(p, q)$:

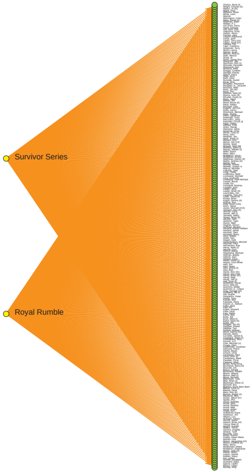
- $n_1 + n_2 \leq$ selected threshold
- 'border line' in the (p, q)-table.

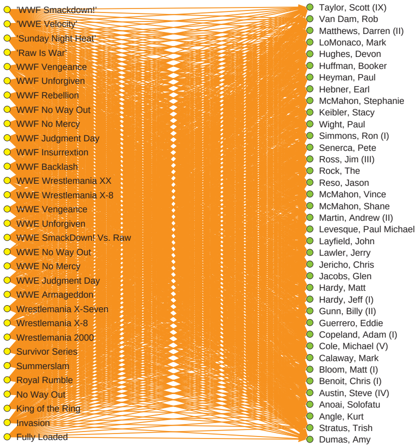
Table $(p, q : n_1, n_2)$

for Internet Movie Database

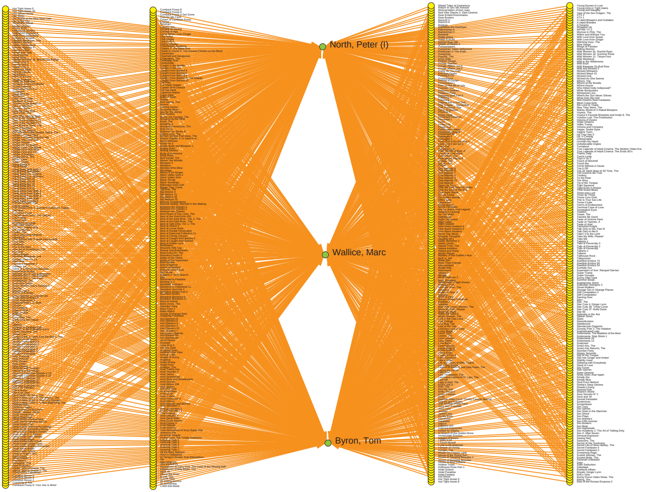
```
Network/2-Mode Network/Core/2-Mode Border

 1 1590: 1590    1 | 16 39: 2173  678 |   44 14: 29   83
 2  516:  788    3 | 17 35: 2791  995 |   46 13: 29   94
 3  212: 1705   18 | 18 32: 2684 1080 |   49 12: 26   95
 4  151: 4330  154 | 19 30: 2395 1063 |   52 11: 16   79
 5  131: 4282  209 | 20 28: 2216 1087 |   56 10: 34  162
 6  115: 3635  223 | 21 26: 1988 1087 |   62  9: 31  177
 7  101: 3224  244 | 22 24: 1854 1153 |   66  8: 29  198
 8   88: 2860  263 | 24 23:   34   39 |   72  7: 22  203
 9   77: 3467  393 | 27 22:   31   38 |   96  6:  7  114
10   69: 3150  428 | 29 20:   35   52 |  119  5:  6  137
11   63: 2442  382 | 32 19:   34   57 |  141  4:  8  258
12   56: 2479  454 | 35 18:   33   61 |  186  3:  3  186
13   50: 3330  716 | 36 17:   33   65 |  247  2:  2  247
14   46: 2460  596 | 39 16:   29   70 | 1334  1:  1 1334
15   42: 2663  739 | 42 15:   28   76 |
```

(247,2)-core and (27,22)-core





(2,516)-Hard core

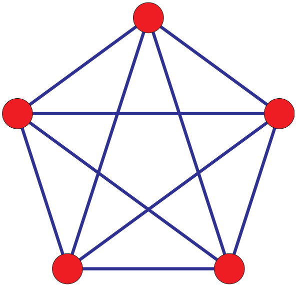


IMDB cores / Pajek commands

```
Options/Read-Write/Read-Save vertices labels [Off]
Read/Network [IMDB.net]  1:40
Info/Memory
Network/2-Mode Network/Core/2-Mode Review
Network/2-Mode Network/Core/2-Mode [27 22]
Info/Partition
Operations/Network+Partition/Extract Subnetwork [Yes 1]
Network/2-Mode Network/Partition into 2 Modes
Network/Create New Network/Transform/Add/Vertex Labels/
  from File(s) [IMDB.nam]
Draw/Network+First Partition
Layers/in y direction
Options/Transform/Rotate 2D [90]
```

k-rings

A *k-ring* is a simple closed chain of length k. Using k-rings we can define a weight of edges as
$w_k(e) = \#$ of different k-rings containing the edge $e \in \mathcal{E}$



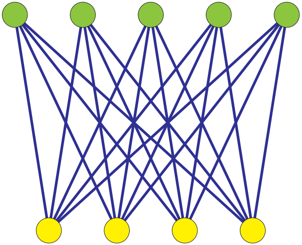
Complete graph K_5

Since for each eadge e of a complete graph K_r, $r \geq k \geq 3$ we have $w_k(e) = (r-2)!/(r-k)!$ the edges belonging to cliques have large weights. Therefore these weights can be used to identify the dense parts of a network.
The k-rings can be efficiently determined only for small values of k – 3, 4, 5.

On the k-rings we can also base the notion of short cycle connectivity which provides us with another decomposition of networks. paper

4-rings and analysis of two-mode networks

In two-mode network there are no 3-rings. The densest substructures are complete bipartite subgraphs $K_{p,q}$. They contain many 4-rings.



There are

$$\binom{p}{2}\binom{q}{2} = \frac{1}{4}p(p-1)q(q-1)$$

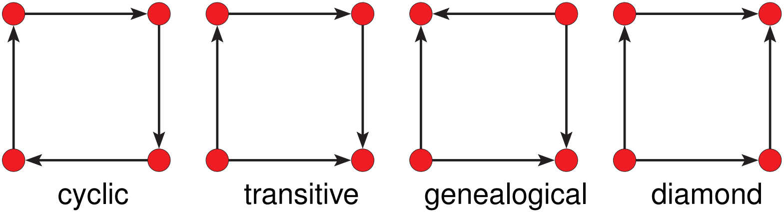
4-rings in $K_{p,q}$; and each of its edges e has weight

$$w_4(e) = (p-1)(q-1)$$

`Network/Create New Network/with Ring Counts.../4-Rings/Undirecte`

Directed 4-rings

There are 4 types of directed 4-rings:



In the case of transitive rings **Pajek** provides a special weight counting on how many transitive rings the arc is a *shortcut*.

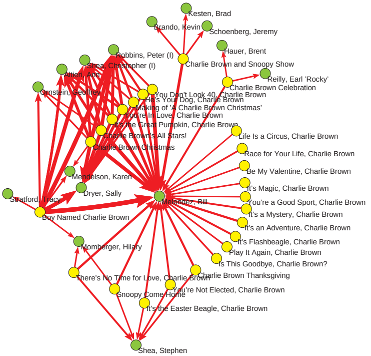
`Network/Create New Network/with Ring Counts/4-Rings/Directed`

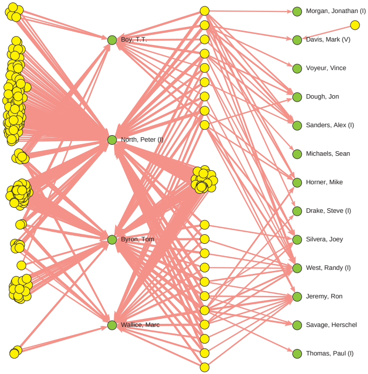
Simple line islands in IMDB for w_4

We obtained 12465 simple line islands on 56086 nodes. Here is their size distribution.

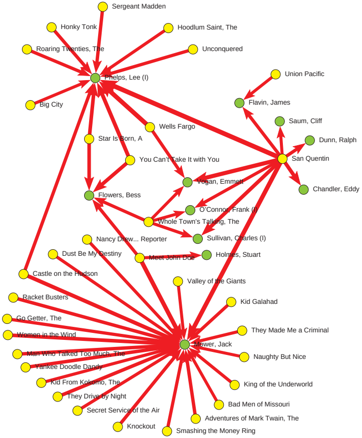
Size	Freq	Size	Freq	Size	Freq	Size	Freq
2	5512	20	19	38	4	59	2
3	1978	21	18	39	3	61	1
4	1639	22	15	40	2	64	1
5	968	23	9	42	2	67	1
6	666	24	13	43	3	70	1
7	394	25	12	45	3	73	1
8	257	26	6	46	4	76	1
9	209	27	6	47	5	82	1
10	148	28	5	48	1	86	1
11	118	29	6	49	2	106	1
12	87	30	3	50	2	122	1
13	55	31	6	51	1	135	1
14	62	32	5	52	2	144	1
15	46	33	3	53	1	163	1
16	39	34	1	54	2	269	1
17	27	35	5	55	1	301	1
18	28	36	4	57	1	332	2
19	29	37	7	58	1	673	1

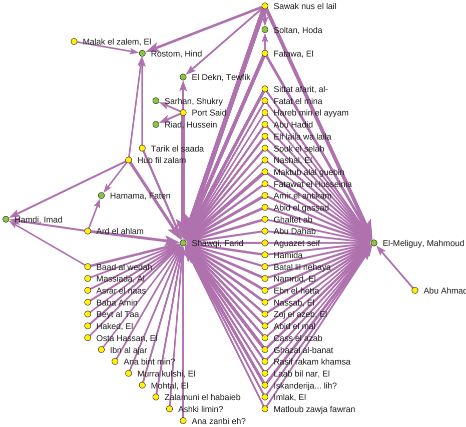
Example: Islands for w_4 Charlie Brown and Adult



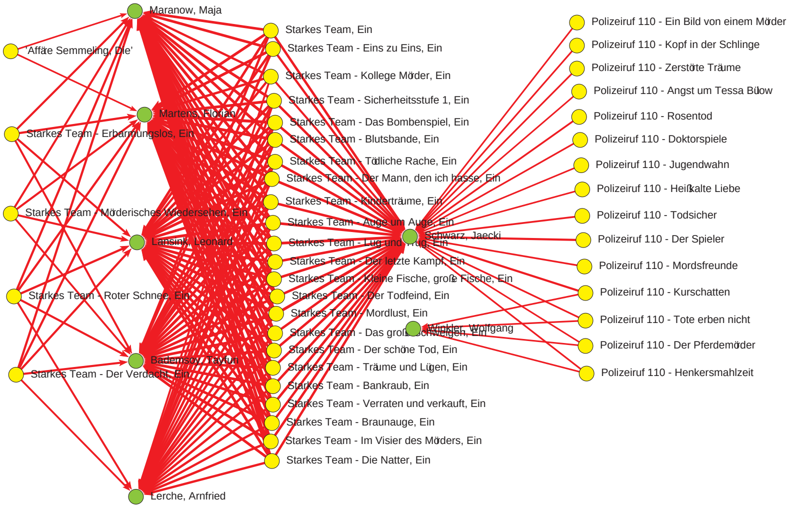


Example: Islands for w_4 Mark Twain and Abid



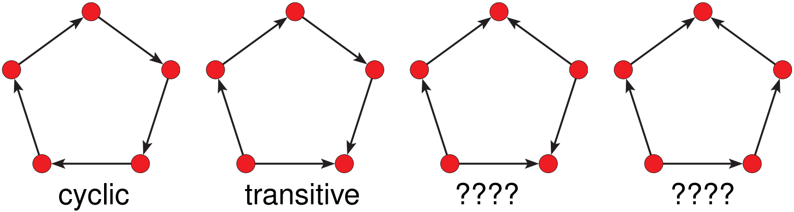


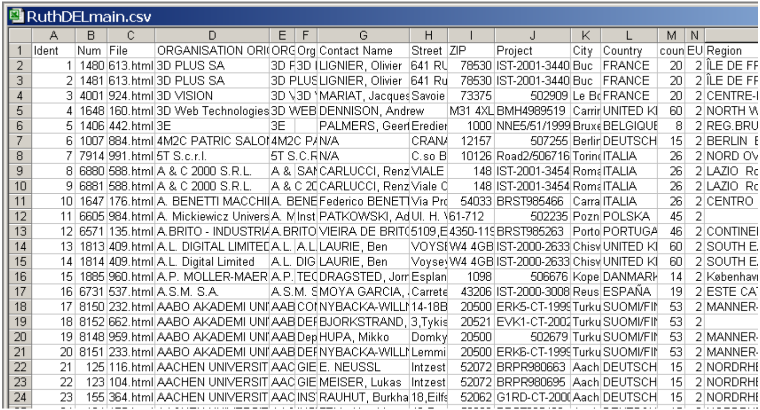
Example: Island for w_4 Polizeiruf 110 and Starkes Team



5-rings

In the future we intend to implement in **Pajek** also weights w_5. Again there are only 4 types of directed 5-rings.





RuthDELmain.csv

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	
1	Ident	Num	File	ORGANISATION ORI(	ORG	Org	Contact Name	Street	ZIP	Project	City	Country	coun	EU	Region
2	1	1480	613.html	3D PLUS SA	3D F	3D I	LIGNIER, Olivier	641 RU	78530	IST-2001-3440	Buc	FRANCE	20	2	ÎLE DE FF
3	2	1481	613.html	3D PLUS SA	3D PLUS		LIGNIER, Olivier	641 Ru	78530	IST-2001-3440	Buc	FRANCE	20	2	ÎLE DE FF
4	3	4001	924.html	3D VISION	3D V	3D '	MARIAT, Jacques	Savoie	73375	502909	Le Bc	FRANCE	20	2	CENTRE-I
5	4	1648	160.html	3D Web Technologies	3D WEB		DENNISON, Andrew		M31 4XL	BMH4989519	Carrir	UNITED KI	60	2	NORTH W
6	5	1406	442.html	3E	3E		PALMERS, Geert	Eredier	1000	NNE5/51/1999	Bruxe	BELGIQUE	8	2	REG.BRU
7	6	1007	884.html	4M2C PATRIC SALOI	4M2C PA		N/A	CRANA	12157	507255	Berlir	DEUTSCH	15	2	BERLIN E
8	7	7914	991.html	5T S.c.r.l.	5T S.C.R		N/A	C.so B	10126	Road2/506716	Torinc	ITALIA	26	2	NORD OV
9	8	6880	588.html	A & C 2000 S.R.L.	A &	SAN	CARLUCCI, Renz	VIALE	148	IST-2001-3454	Roma	ITALIA	26	2	LAZIO Rc
10	9	6881	588.html	A & C 2000 S.R.L.	A & C 2(		CARLUCCI, Renz	Viale C	148	IST-2001-3454	Roma	ITALIA	26	2	LAZIO Rc
11	10	1647	176.html	A. BENETTI MACCHII	A. BENE		Federico BENETI	Via Prc	54033	BRST985466	Carra	ITALIA	26	2	CENTRO
12	11	6605	984.html	A. Mickiewicz Univers	A. M	Inst	PATKOWSKI, Ad	Ul. H. \	61-712	502235	Pozn	POLSKA	45	2	
13	12	6571	135.html	A.BRITO - INDUSTRIA	A.BRITO		VIEIRA DE BRIT(	5109,E	4350-119	BRST985263	Porto	PORTUGA	46	2	CONTINEI
14	13	1813	409.html	A.L. DIGITAL LIMITED	A.L.	A.L.	LAURIE, Ben	VOYSE	W4 4GB	IST-2000-2633	Chisv	UNITED KI	60	2	SOUTH E.
15	14	1814	409.html	A.L. Digital Limited	A.L. DIG		LAURIE, Ben	Voysey	W4 4GB	IST-2000-2633	Chisv	UNITED KI	60	2	SOUTH E.
16	15	1885	960.html	A.P. MOLLER-MAER	A.P.	TEC	DRAGSTED, Jorr	Esplan	1098	506676	Kope	DANMARK	14	2	Københavi
17	16	6731	537.html	A.S.M. S.A.	A.S.M. S		MOYA GARCIA, .	Carrete	43206	IST-2000-3008	Reus	ESPAÑA	19	2	ESTE CAT
18	17	8150	232.html	AABO AKADEMI UNI	AAB	COI	NYBACKA-WILLI	14-18B	20500	ERK5-CT-1999	Turku	SUOMI/FII	53	2	MANNER-
19	18	8152	662.html	AABO AKADEMI UNI	AAB	DEF	BJORKSTRAND,	3,Tykis	20521	EVK1-CT-2002	Turku	SUOMI/FII	53	2	
20	19	8148	959.html	AABO AKADEMI UNI	AAB	Dep	HUPA, Mikko	Domky	20500	502679	Turku	SUOMI/FII	53	2	MANNER-
21	20	8151	233.html	AABO AKADEMI UNI	AAB	DEF	NYBACKA-WILLI	Lemmi	20500	ERK6-CT-1999	Turku	SUOMI/FII	53	2	MANNER-
22	21	125	116.html	AACHEN UNIVERSIT	AAC	GIE	E. NEUSSL	Intzest	52072	BRPR980663	Aach	DEUTSCH	15	2	NORDRHE
23	22	123	104.html	AACHEN UNIVERSIT	AAC	GIE	MEISER, Lukas	Intzest	52072	BRPR980695	Aach	DEUTSCH	15	2	NORDRHE
24	23	155	364.html	AACHEN UNIVERSIT	AAC	INS	RAUHUT, Burkha	18,Eilfs	52062	G1RD-CT-2000	Aach	DEUTSCH	15	2	NORDRHE

A *data table* $\mathcal{T}$ is a set of *records* $\mathcal{T} = \{T_k : k \in \mathcal{K}\}$, where $\mathcal{K}$ is the set of *keys*. A record has the form $T_k = (k, q_1(k), q_2(k), \ldots, q_r(k))$ where $q_i(k)$ is the value of the *property* (attribute) $\mathbf{q}_i$ for the key k.

. . . Two mode networks from data tables

Suppose that the property **q** has the range $2^{\mathcal{Q}}$. For example:

Authors(SNA) = { S. Wasserman, K. Faust },
PubYear(SNA) = { 1994 },
Keywords(SNA) = { network, centrality, matrix, . . . }, . . .

If $\mathcal{Q}$ is finite (it can always be transformed into such set by partitioning the set $\mathcal{Q}$ and recoding the values) we can assign to the property **q** a two-mode network $\mathcal{K} \times \mathbf{q} = (\mathcal{K}, \mathcal{Q}, \mathcal{E}, w)$ where $(k, v) \in \mathcal{E}$ iff $v \in q(k)$, and $w(k, v) = 1$.

	. . .	Batagelj	Faust	de Nooy	Kejžar	Korenjak	Mrvar	Wasserman	Zaveršnik	. . .
. . .										
GenCores		1							1	
Islands		1							1	
ESNA2		1		1			1			
IFCS09		1			1	1				
SNA			1					1		
. . .										

Single-valued properties can be represented by a partition.
We can always transform the partition into corresponding network.

Record from Web of Science

```
PT J
AU Dipple, H
   Evans, B
TI The Leicestershire Huntington's disease support group: a social network
   analysis
SO HEALTH & SOCIAL CARE IN THE COMMUNITY
LA English
DT Article
C1 Rehabil Serv, Troon Way Business Ctr, Leicester LE4 9HA, Leics, England.
RP Dipple, H, Rehabil Serv, Troon Way Business Ctr, Sandringham
   Suite,Humberstone Lane, Leicester LE4 9HA, Leics, England.
CR BORGATTI SP, 1992, UCINET 4 VERSION 1 0
   FOLSTEIN S, 1989, HUNTINGTONS DIS DISO
   SCOTT J, 1991, SOCIAL NETWORK ANAL
NR 3
TC 3
PU BLACKWELL SCIENCE LTD
PI OXFORD
PA P O BOX 88, OSNEY MEAD, OXFORD OX2 0NE, OXON, ENGLAND
SN 0966-0410
J9 HEALTH SOC CARE COMMUNITY
JI Health Soc. Care Community
PD JUL
PY 1998
VL 6
IS 4
BP 286
EP 289
PG 4
SC Public, Environmental & Occupational Health; Social Work
GA 105UP
UT ISI:000075092200008
ER
```

WoS2Pajek

Records from BiBTEX

```
@Article{int:Mizuno1,
  author =         "S. Mizuno",
  title =          "An ${O(n^{3}L)}$ algorithm using a sequence fo
                   linear complementarity problems",
  journal =        "Journal of the Operations Research Society of
  volume =         "33",
  year =           "1990",
  pages =          "66--75",
}

@InCollection{int:Vorst1,
  author =         "{J. G. G. van de} Vorst",
  title =          "An attempt to use parallel computing in large
                   optimisation",
  booktitle =      "Logistics, Where Ends Have to Meet~: Proceedin
                   the Shell Conference on Logistics in Apeldoorn,
                   Netherlands, November 1988",
  editor =         "{C. F. H. van} Rijn",
  year =           "1989",
  pages =          "112--119",
  publisher =      "Pergamon Press",
  address =        "Oxford, United Kingdom",
}
```

Bib2Pajek.py

Two mode networks from data tables

For data from the Web of Science (Knowledge) we can obtain the corresponding networks using the program **`WoS2Pajek`**:

- citation network **Ci**: works $\times$ works;
- authorship network **WA**: works $\times$ authors, for works without complete description only the first author is known;
- keywords network **WK**: works $\times$ keywords, only for works with complete description;
- journals network **WJ**: works $\times$ journals;
- partition of works by the publication year;
- partition of works – complete description (1) / ISI name only (0);

Similar programs exist also for other bibliographic sources/formats: Scopus, BibTEX, Zentralblatt Math, Google Scholar, DBLP, IMDB, etc.

Linked / multi-modal networks

Linked or multi-modal networks are collections of networks over at least two sets of nodes (modes) and consist of some one-mode networks and some two-mode networks linking different modes. For example: modes are Persons and Organizations. Two one-mode networks describe collaboration among Persons and among Organizations. The linking two-mode network describes membership of Persons to different Organizations.
An important approach in analysis of linked networks is the use of derived networks obtained by network multiplication.

- Krackhardt, D., Carley, K.M. 1998. A PCANS Model of Structure in Organization. In Proceedings of the 1998 International Symposium on Command and Control Research and Technology Evidence Based Research: 113-119, Vienna, VA. MetaMatrix, paper
- Kathleen M. Carley (2003). Dynamic Network Analysis. in the Summary of the NRC workshop on Social Network Modeling and Analysis, Ron Breiger and Kathleen M. Carley (Eds.), National Research Council. preprint

MetaMatrix

Carley and Diesner

Meta-Matrix Entities	Agent	Knowledge	Resources	Tasks/ Event	Organizations	Location
Agent	Social network	Knowledge network	Capabilities network	Assignment network	Membership network	Agent location network
Knowledge		Information network	Training network	Knowledge requirement network	Organizational knowledge network	Knowledge location network
Resources			Resource network	Resource requirement Network	Organizational Capability network	Resource location network
Tasks/ Events				Precedence network	Organizational assignment network	Task/Event location network
Organizations					Inter-organizational network	Organizational location network
Location						Proximity network

Multiplication of networks

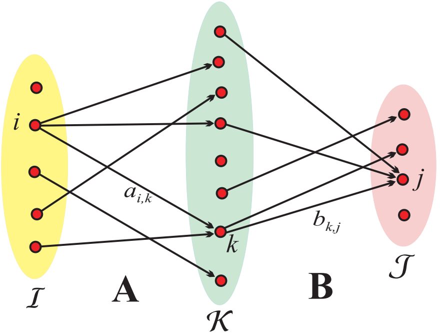
To a simple (no parallel arcs) two-mode *network* $\mathcal{N} = (\mathcal{I}, \mathcal{J}, \mathcal{A}, w)$; where $\mathcal{I}$ and $\mathcal{J}$ are sets of *nodes*, $\mathcal{A}$ is a set of *arcs* linking $\mathcal{I}$ and $\mathcal{J}$, and $w : \mathcal{A} \to \mathbb{R}$ (or some other semiring) is a *weight*; we can assign a *network matrix* $\mathbf{W} = [w_{i,j}]$ with elements: $w_{i,j} = w(i,j)$ for $(i,j) \in \mathcal{A}$ and $w_{i,j} = 0$ otherwise.

Given a pair of compatible networks $\mathcal{N}_A = (\mathcal{I}, \mathcal{K}, \mathcal{A}_A, w_A)$ and $\mathcal{N}_B = (\mathcal{K}, \mathcal{J}, \mathcal{A}_B, w_B)$ with corresponding matrices $\mathbf{A}_{\mathcal{I}\times\mathcal{K}}$ and $\mathbf{B}_{\mathcal{K}\times\mathcal{J}}$ we call a *product of networks* $\mathcal{N}_A$ and $\mathcal{N}_B$ a network $\mathcal{N}_C = (\mathcal{I}, \mathcal{J}, \mathcal{A}_C, w_C)$, where $\mathcal{A}_C = \{(i,j) : i \in \mathcal{I}, j \in \mathcal{J}, c_{i,j} \neq 0\}$ and $w_C(i,j) = c_{i,j}$ for $(i,j) \in \mathcal{A}_C$. The product matrix $\mathbf{C} = [c_{i,j}]_{\mathcal{I}\times\mathcal{J}} = \mathbf{A} * \mathbf{B}$ is defined in the standard way

$$c_{i,j} = \sum_{k \in \mathcal{K}} a_{i,k} \cdot b_{k,j}$$

In the case when $\mathcal{I} = \mathcal{K} = \mathcal{J}$ we are dealing with ordinary one-mode networks (with square matrices).

Multiplication of networks



$$c_{i,j} = \sum_{k \in N_A(i) \cap N_B^-(j)} a_{i,k} \cdot b_{k,j}$$

If all weights in networks $\mathcal{N}_A$ and $\mathcal{N}_B$ are equal to 1 the value of $c_{i,j}$ counts the number of ways we can go from $i \in \mathcal{I}$ to $j \in \mathcal{J}$ passing through $\mathcal{K}$, $c_{i,j} = |N_A(i) \cap N_B^-(j)|$.

Multiplication of networks

The standard matrix multiplication has the complexity $O(|\mathcal{I}| \cdot |\mathcal{K}| \cdot |\mathcal{J}|)$ – it is too slow to be used for large networks. For sparse large networks we can multiply much faster considering only nonzero elements.

> **for** k **in** $\mathcal{K}$ **do**
>   **for** (i, j) **in** $N_A^-(k) \times N_B(k)$ **do**
>   **if** $\exists c_{i,j}$ **then** $c_{i,j} := c_{i,j} + a_{i,k} \cdot b_{k,j}$
>   **else** new $c_{i,j} := a_{i,k} \cdot b_{k,j}$

`Networks/Multiply Networks`

In general the multiplication of large sparse networks is a 'dangerous' operation since the result can 'explode' – it is not sparse.

Multiplication of networks

From the network multiplication algorithm we see that each intermediate node $k \in \mathcal{K}$ adds to a product network a complete two-mode subgraph $K_{N_A^-(k),N_B(k)}$ (or, in the case $\mathcal{I} = \mathcal{J}$, a complete subgraph $K_{N(k)}$). If both degrees $\deg_A(k) = |N_A^-(k)|$ and $\deg_B(k) = |N_B(k)|$ are large then already the computation of this complete subgraph has a quadratic (time and space) complexity – the result 'explodes'.

If at least one of the sparse networks $\mathcal{N}_A$ and $\mathcal{N}_B$ has small maximal degree on $\mathcal{K}$ then also the resulting product network $\mathcal{N}_C$ is sparse.

If for the sparse networks $\mathcal{N}_A$ and $\mathcal{N}_B$ there are in $\mathcal{K}$ only few nodes with large degree and no one among them with large degree in both networks then also the resulting product network $\mathcal{N}_C$ is sparse.

Kinship relations

Anthropologists typically use a basic vocabulary of kin types to represent genealogical relationships. One common version of the vocabulary for basic relationships:

```
Kin Type     English Type
 P           Parent
   F         Father
   M         Mother
 C           Child
   D         Daughter
   S         Son
 G           Sibling
   Z         Sister
   B         Brother
 E           Spouse
   H         Husband
   W         Wife
```

The genealogies are usually described in GEDCOM format.

Examples family, Bouchards. Paper

Calculating kinship relations

Pajek generates three relations when reading genealogy as Ore graph:
F: _ ***is a father of*** _
M: _ ***is a mother of*** _
E: _ ***is a spouse of*** _
Additionally we must generate two binary diagonal matrices, to distinguish between male and female:
L: _ ***is a male*** _ / 1-male, 0-female
J: _ ***is a female*** _ / 1-female, 0-male

$$\mathbf{F} \cap \mathbf{M} = \emptyset, \qquad \mathbf{L} \cup \mathbf{J} \subseteq \mathbf{I}, \qquad \mathbf{L} \cap \mathbf{J} = \emptyset$$

Derived kinship relations

Other basic relations can be obtained using macros based on identities:

_ ***is a parent of*** _	P	$=$	$F \cup M$
_ ***is a child of*** _	C	$=$	P^T
_ ***is a son of*** _	S	$=$	$L * C$
_ ***is a daughter of*** _	D	$=$	$J * C$
_ ***is a husband of*** _	H	$=$	$L * E$
_ ***is a wife of*** _	W	$=$	$J * E$
_ ***is a sibling of*** _	G	$=$	$((F^T * F) \cap (M^T * M)) \setminus I$
_ ***is a brother of*** _	B	$=$	$L * G$
_ ***is a sister of*** _	Z	$=$	$J * G$
_ ***is an uncle of*** _	U	$=$	$B * P$
_ ***is an aunt of*** _	A	$=$	$Z * P$
_ ***is a semi-sibling of*** _	G_e	$=$	$(P^T * P) \setminus I$

and using them other relations can be determined

_ ***is a grand mother of*** _	M_2	$=$	$M * P$
_ ***is a niece of*** _	Ni	$=$	$D * G$

Relative sizes of kinship relations in genealogies

Kin Type	Turks	Ragusa	Loka	Silba	Royal
P-Parent	1.000	1.000	1.000	1.000	1.000
F-Father	0.514	0.532	0.504	0.519	0.540
M-Mother	0.486	0.468	0.496	0.481	0.460
C-Child	1.000	1.000	1.000	1.000	1.000
D-Daughter	0.431	0.384	0.480	0.469	0.427
S-Son	0.569	0.616	0.520	0.531	0.573
G-Sibling	1.250	0.943	1.019	0.811	0.767
Z-Sister	1.135	0.746	0.983	0.760	0.707
B-Brother	1.366	1.140	1.055	0.861	0.828
E-Spouse	0.205	0.215	0.208	0.230	0.306
H-Husband	0.205	0.215	0.208	0.230	0.306
W-Wife	0.205	0.215	0.208	0.230	0.306
U-Uncle	1.920	1.789	1.200	1.181	0.927
A-Aunt	1.750	1.143	1.190	1.097	0.798
Ge-Semi-sibling	1.473	1.155	1.128	0.932	0.905
n	1269	5999	47956	6427	3010
mE = Spouse	407	2002	14154	2217	1138
mA = Parent	1987	9315	68052	9627	3724

Two-mode network analysis by conversion to one-mode network

Often we transform a two-mode network $\mathcal{N} = (\mathcal{U}, \mathcal{V}, \mathcal{E}, w)$ into an ordinary (one-mode) network $\mathcal{N}_1 = (\mathcal{U}, \mathcal{E}_1, w_1)$ or/and $\mathcal{N}_2 = (\mathcal{V}, \mathcal{E}_2, w_2)$, where $\mathcal{E}_1$ and w_1 are determined by the matrix $\mathbf{W}^{(1)} = \mathbf{W}\mathbf{W}^T$, $w^{(1)}_{uv} = \sum_{z\in\mathcal{V}} w_{uz} \cdot w^T_{zv}$. Evidently $w^{(1)}_{uv} = w^{(1)}_{vu}$. There is an edge $(u : v) \in \mathcal{E}_1$ in $\mathcal{N}_1$ iff $N(u) \cap N(v) \neq \emptyset$. Its weight is $w_1(u, v) = w^{(1)}_{uv}$.

The network $\mathcal{N}_2$ is determined in a similar way by the matrix $\mathbf{W}^{(2)} = \mathbf{W}^T\mathbf{W}$.
The networks $\mathcal{N}_1$ and $\mathcal{N}_2$ are analyzed using standard methods.

`Network/2-Mode Network/2-Mode to 1-Mode/Rows`

Normalizations

The *normalization* approach was developed for quick inspection of (1-mode) networks obtained from two-mode networks – a kind of network based data-mining.

In networks obtained from large two-mode networks there are often huge differences in weights. Therefore it is not possible to compare the vertices according to the raw data. First we have to normalize the network to make the weights comparable.

There exist several ways how to do this. Some of them are presented in the following table. They can be used also on other networks.

In the case of networks without loops we define the diagonal weights for undirected networks as the sum of out-diagonal elements in the row (or column) $w_{vv} = \sum_u w_{vu}$ and for directed networks as some mean value of the row and column sum, for example $w_{vv} = \frac{1}{2}(\sum_u w_{vu} + \sum_u w_{uv})$. Usually we assume that the network does not contain any isolated node.

. . . Normalizations

$$\mathrm{Geo}_{uv} = \frac{w_{uv}}{\sqrt{w_{uu} w_{vv}}} \qquad \mathrm{GeoDeg}_{uv} = \frac{w_{uv}}{\sqrt{\deg_u \deg_v}}$$

$$\mathrm{Input}_{uv} = \frac{w_{uv}}{w_{vv}} \qquad \mathrm{Output}_{uv} = \frac{w_{uv}}{w_{uu}}$$

$$\mathrm{Min}_{uv} = \frac{w_{uv}}{\min(w_{uu}, w_{vv})} \qquad \mathrm{Max}_{uv} = \frac{w_{uv}}{\max(w_{uu}, w_{vv})}$$

$$\mathrm{MinDir}_{uv} = \begin{cases} \frac{w_{uv}}{w_{uu}} & w_{uu} \leq w_{vv} \\ 0 & otherwise \end{cases} \qquad \mathrm{MaxDir}_{uv} = \begin{cases} \frac{w_{uv}}{w_{vv}} & w_{uu} \leq w_{vv} \\ 0 & otherwise \end{cases}$$

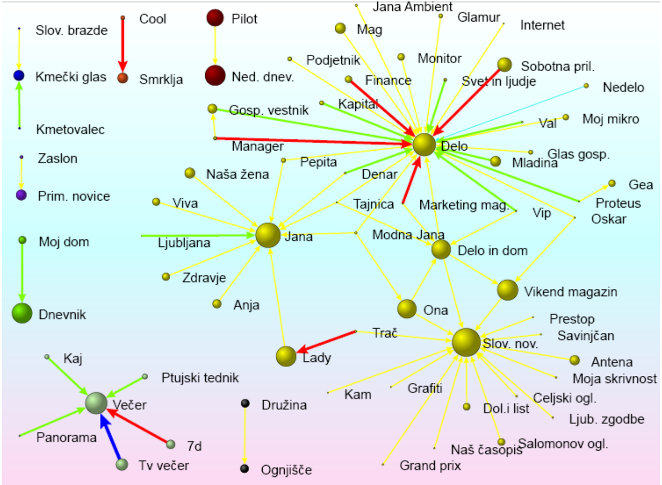
After a selected normalization the important parts of network are obtained by link-cuts or islands approaches.

```
Network/2-Mode Network/2-Mode to 1-Mode/Normalize 1-Mode/
```

Reuters Terror News: GeoDeg, MaxDir, MinDir.

Slovenian journals and magazins.

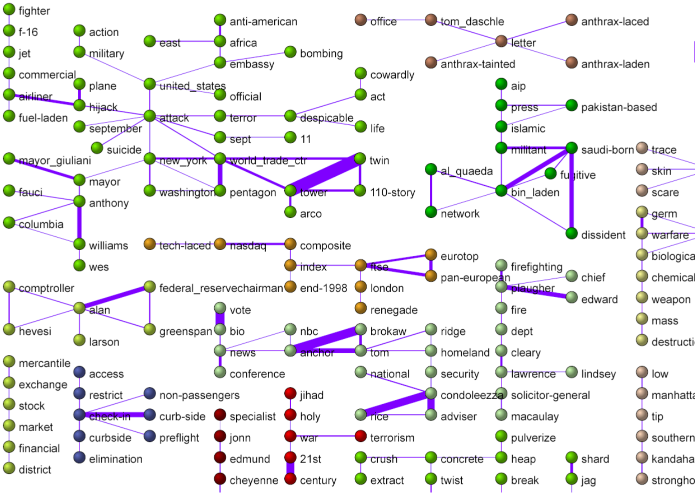
MinDir of Slovenian journals 2000



Over 100000 people were asked in the years 1999 and 2000 about the journals they read.
They mentioned 124 different journals. (source Cati)

GeoDeg normalization

of Reuters terror news network



Authorship network

Let **WA** be the works $\times$ authors two mode authorship network; $wa_{pi} \in \{0, 1\}$ is describing the authorship of author i of work p.

$$\forall p \in W : \sum_{i \in A} wa_{pi} = \mathrm{outdeg}_{WA}(p) = \text{ \# authors of work } p$$

Let **N** be its normalized version

$$\forall p \in W : \sum_{i \in A} n_{pi} \in \{0, 1\}$$

obtained from **WA** by $n_{pi} = wa_{pi} / \max(1, \mathrm{outdeg}_{WA}(p))$, or by some other rule determining the author's contribution.

Some transformations of networks

Binarization $b(\mathcal{N})$ is a network obtained from the $\mathcal{N}$ in which all weights are set to 1.

Transposition $\mathcal{N}^T$ or $t(\mathcal{N})$ is a network obtained from $\mathcal{N}$ in which to all arcs their direction is reversed. $\mathbf{AW} = \mathbf{WA}^T$, $\mathbf{KW} = \mathbf{WK}^T$, ...

(Out) normalization $n(\mathcal{N})$ is a network obtained from $\mathcal{N}$ in which the weight of each arc a is divided by the sum of weights of all arcs having the same initial vertex as the arc a. For binary networks

$$n(\mathbf{A}) = \mathrm{diag}\left(\frac{1}{\max(1, \mathrm{outdeg}_{WA}(i))}\right)_{i \in \mathcal{I}} * \mathbf{A}$$

$\mathbf{N} = n(\mathbf{WA})$, $\mathbf{WA} = b(\mathbf{N})$

First co-authorship network

$\mathbf{Co} = \mathbf{AW} * \mathbf{WA}$

$$co_{ij} = \sum_{p \in W} wa_{pi} wa_{pj} = \sum_{p \in N^-(i) \cap N^-(j)} 1$$

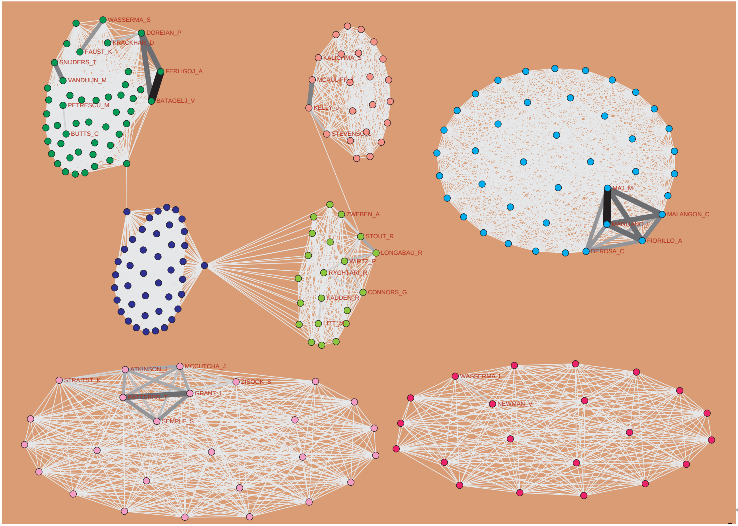
co_{ij} = the number of works that authors i and j wrote together
It holds: $co_{ij} = co_{ji}$.

Using the weights co_{ij} we can determine the Salton's cosine similarity or Ochiai coefficient between authors i and j as

$$cos(i,j) = \frac{co_{ij}}{\sqrt{co_{ii} co_{jj}}}, \qquad \text{for } co_{ij} > 0$$

Cores of orders 20–47 in **Co**(SN5)

Network SN5 (2008): for `"social network*"` + most frequent references + around 100 social networkers; $|W| = 193376$, $|C| = 7950$, $|A| = 75930$, $|J| = 14651$, $|K| = 29267$



Papers by number of authors

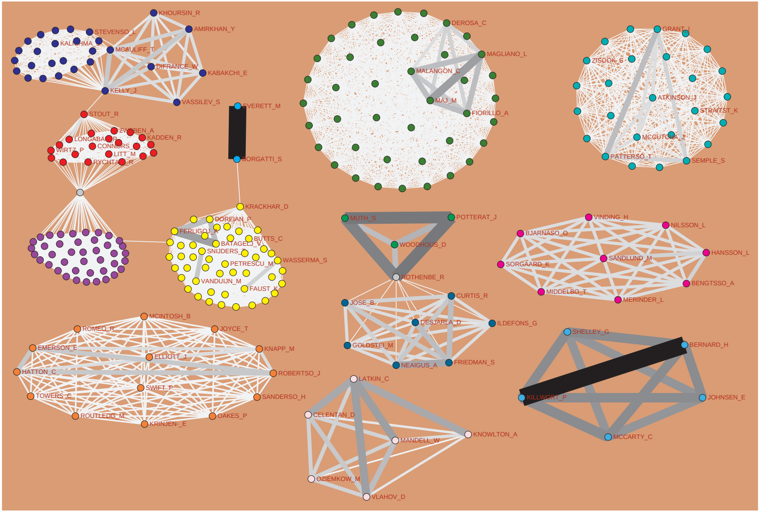
Problem: The **Co** network is composed of complete graphs on the set of work's authors. Works with many authors produce large complete subgraphs and are over-represented, thus bluring the collaboration structure.

outdeg	frequency	outdeg	frequency	paper
1	2637	12	8	
2	2143	13	4	
3	1333	14	3	
4	713	15	2	
5	396	21	1	Pierce et al. (2007)
6	206	22	1	Allen et al. (1998)
7	114	23	1	Kelly et al. (1997)
8	65	26	1	Semple et al. (1993)
9	43	41	1	Magliano et al. (2006)
10	24	42	1	Doll et al. (1992)
11	10	48	1	Snijders et al. (2007)

Snijders et al. (2007)

Snijders et al.(2007): Snijders, T.A.B., Robinson, T., Atkinson, A.C., Riani, M., Gormley, I.C., Murphy, T.B., Sweeting, T., Leslie, D.S., Longford, N.T., Kent, J.T., Lawrance, T., Airoldi, E.M., Besag, J., Blei, D., Fienberg, S.E., Breiger, R., Butts, C.T., Doreian, P., Batagelj, V., Ferligoj, A., Draper, D., van Duijn, M.A.J., Faust, K., Petrescu-Prahova, M., Forster, J.J., Gelman, A., Goodreau, S. M., Greenwood, P.E., Gruenberg, K., Francis, B., Hennig, C., Hoff, P.D., Hunter, D.R., Husmeier, D., Glasbey, C., Krackhardt, D., Kuha, J., Skrondal, A., Lawson, A., Liao, T. F., Mendes, B., Reinert, G., Richardson, S., Lewin, A., Titterington, D.M., Wasserman, S., Werhli, A.V. and Ghazal, P.. *Discussion on the paper by Handcock, Raftery and Tantrum*. Journal of the Royal Statistical Society: Series A - Statistics in Society, 170 (2007), pp. 322-354.

p_S-core at level 20 of **Co**$(\mathtt{SN5})$



Second co-authorship network

$\mathbf{Cn} = \mathbf{AW} * \mathbf{N}$

$$cn_{ij} = \sum_{p \in W} wa_{pi} n_{pj} = \sum_{p \in N^-(i) \cap N^-(j)} n_{pj}$$

cn_{ij} = contribution of author j to works, that (s)he wrote together with the author i.

It holds $\sum_{j \in A} \sum_{j \in A} wa_{pi} n_{pj} = \mathrm{outdeg}_{WA}(p)$ and $\sum_{j \in A} cn_{ij} = \mathrm{indeg}_{WA}(i)$

$cn_{ii} = \sum_{p \in N(i)} n_{pi}$ is the contribution of author i to his/her works.

Self-sufficiency: $S_i = \dfrac{cn_{ii}}{\mathrm{outdeg}_{WA}(i)}$

Collaborativness: $K_i = 1 - S_i$

$$\sum_{i \in A} \sum_{j \in A} cn_{ij} = \sum_{i \in A} \mathrm{indeg}_{WA}(i) = m_{WA}$$

To compute the table we prepared a macro in **Pajek**.

The "best" authors in Social Networks

i	author	cn_{ii}	total	K_i	i	author	cn_{ii}	total	K_i
1	Burt,R	43.83	53	0.173	26	Latkin,C	10.14	37	0.726
2	Newman,M	36.77	60	0.387	27	Morris,M	9.98	20	0.501
3	Doreian,P	34.44	47	0.267	28	Rothenberg,R	9.82	28	0.649
4	Bonacich,P	30.17	41	0.264	29	Kadushin,C	9.75	11	0.114
5	Marsden,P	29.42	37	0.205	30	Faust,K	9.72	18	0.460
6	Wellman,B	26.87	41	0.345	31	Batagelj,V	9.69	20	0.516
7	Leydesdorf,L	24.37	35	0.304	32	Mizruchi,M	9.67	15	0.356
8	White,H	23.50	33	0.288	33	[Anon]	9.00	9	0.000
9	Friedkin,N	20.00	23	0.130	34	Johnson,J	8.89	21	0.577
10	Borgatti,S	19.20	41	0.532	35	Fararo,T	8.83	16	0.448
11	Everett,M	16.92	31	0.454	36	Lazega,E	8.50	12	0.292
12	Litwin,H	16.00	21	0.238	37	Knoke,D	8.33	11	0.242
13	Freeman,L	15.53	20	0.223	38	Ferligoj,A	8.19	19	0.569
14	Barabasi,A	14.99	35	0.572	39	Brewer,D	8.03	11	0.270
15	Snijders,T	14.99	30	0.500	40	Klovdahl,A	7.96	17	0.532
16	Valente,T	14.80	34	0.565	41	Hammer,M	7.92	10	0.208
17	Breiger,R	14.44	20	0.278	42	White,D	7.83	15	0.478
18	Skvoretz,J	14.43	27	0.466	43	Holme,P	7.42	14	0.470
19	Krackhardt,D	13.65	25	0.454	44	Boyd,J	7.37	13	0.433
20	Carley,K	12.93	28	0.538	45	Kilduff,M	7.25	16	0.547
21	Pattison,P	12.10	27	0.552	46	Small,H	7.00	7	0.000
22	Wasserman,S	11.72	26	0.549	47	Iacobucci,D	7.00	12	0.417
23	Berkman,L	11.21	30	0.626	48	Pappi,F	6.83	10	0.317
24	Moody,J	10.83	15	0.278	49	Chen,C	6.78	12	0.435
25	Scott,J	10.47	15	0.302	50	Seidman,S	6.75	9	0.250

Third co-authorship network

$\mathbf{Ct} = \mathbf{N}^T * \mathbf{N}$
ct_{ij} = the total contribution of collaboration of authors i and j to works.
It holds $\quad ct_{ij} = ct_{ji} \quad$ and

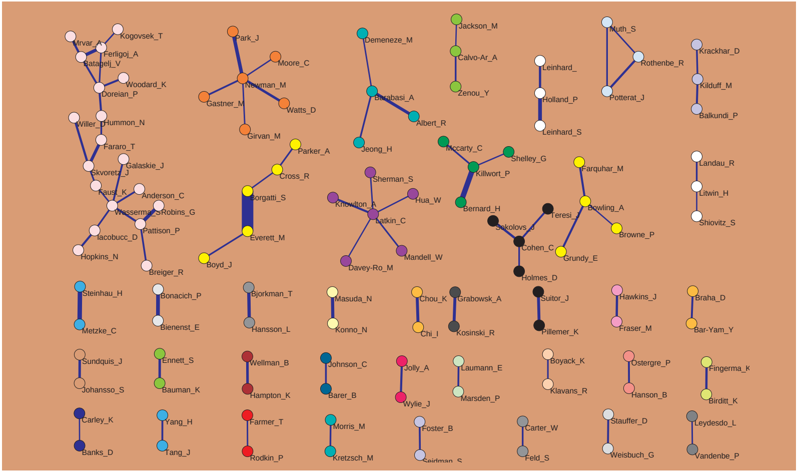
$$\sum_{i \in A} \sum_{j \in A} n_{pi} n_{pj} = 1$$

The total contribution of a complete subgraph corresponding to the authors of a work p is 1.
$\sum_{j \in A} ct_{ij} = \sum_{p \in W} n_{pi}$ = the total contribution of author i to works from W.

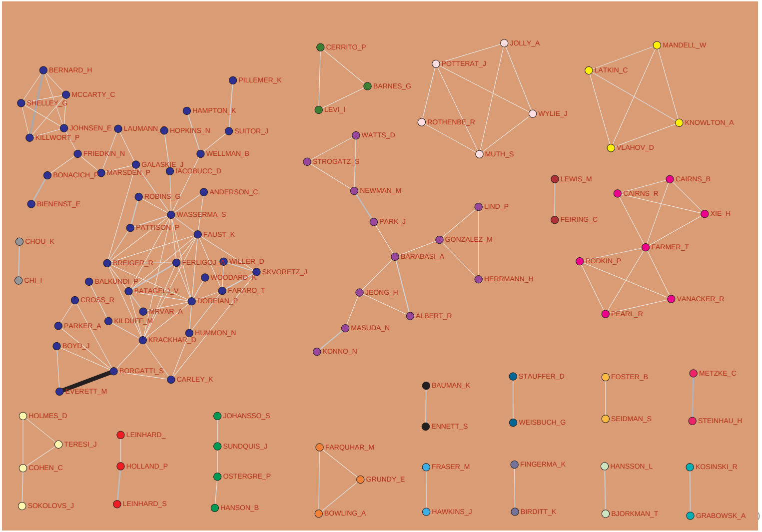
$$\sum_{i \in A} \sum_{j \in A} ct_{ij} = |W|$$

Components in **Ct**($\mathtt{SN5}$) cut at level 0.5

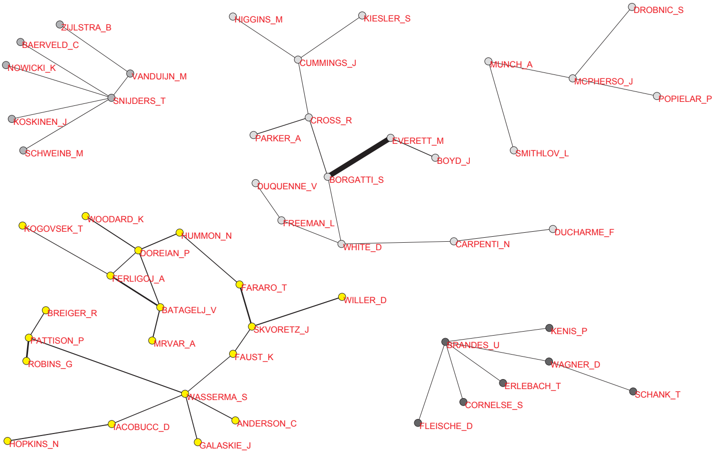
Network SN5 (2008): for "`social network*`" + most frequent references + around 100 social networkers; $|W| = 193376$, $|C| = 7950$, $|A| = 75930$, $|J| = 14651$, $|K| = 29267$



p_S-core at level 0.75 in **Ct**(SN5)



Some link islands [5,20] in **Ct**(SN5)



Fourth co-authorship network

$\mathbf{Ct}' = \mathbf{N}^T * \mathbf{N}'$, where $n'_{pi} = wa_{pi} / \max(1, \mathrm{outdeg}_{WA}(p) - 1)$

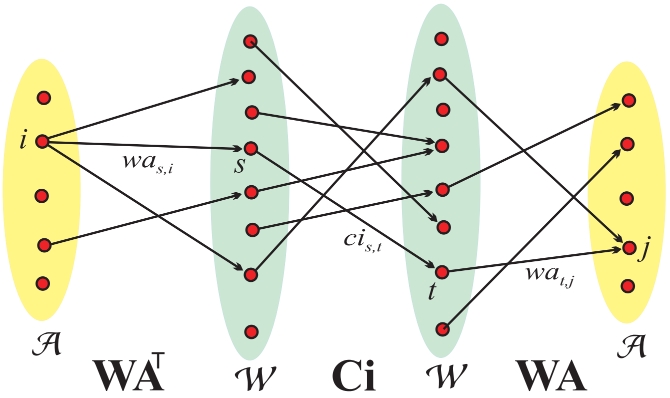
ct'_{ij} = the total contribution of 'strict collaboration' of authors i and j to works.

In **Pajek** we can use macros to save sequences of commands to produce different co-authorship networks.

The final result is returned as an undirected simple network with weights (for $i \neq j$)

$$ct'_{ij} = \sum_p \frac{2 \cdot wa_{pi} \cdot wa_{pj}}{\max(1, \mathrm{outdeg}_{WA}(p)) \cdot \max(1, \mathrm{outdeg}_{WA}(p) - 1)}$$

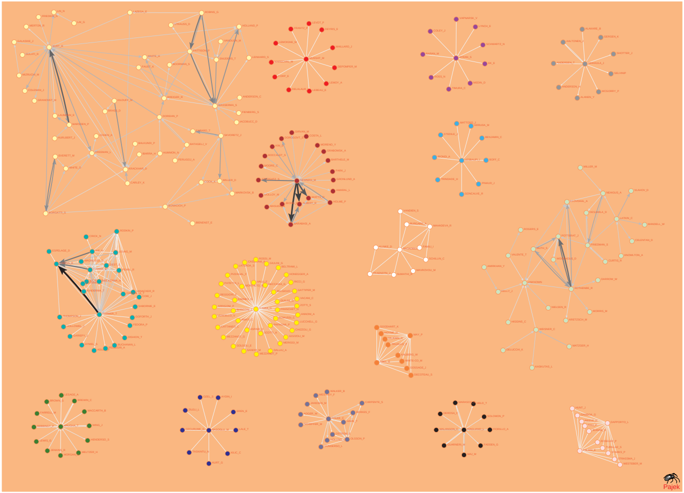
Authors' citations network



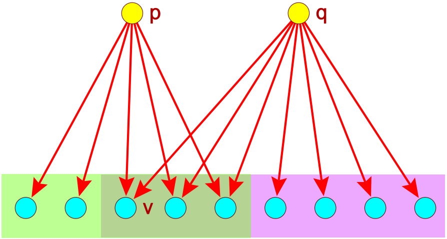
Ca $=$ **AW** $*$ **Ci** $*$ **WA** is a network of citations between authors. The weight $w(i, j)$ counts the number of times a work authored by i is citing a work authored by j.

Islands in SN5 authors citation network

Network SN5 (2008): for `"social network*"` + most frequent references + around 100 social networkers;
$|W| = 193376$, $|C| = 7950$, $|A| = 75930$, $|J| = 14651$, $|K| = 29267$



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

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)

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

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

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

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

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

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

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

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] = Distance
```

Bibliographic Coupling interpretation

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\\descript
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]
```

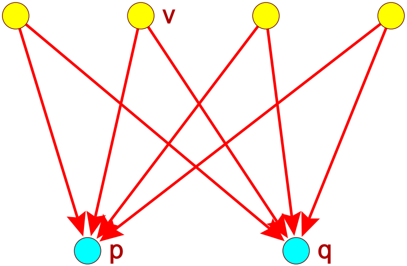
Bibliographic Coupling

the most 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.

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

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

$$\begin{aligned} n(\mathbf{Ci})^T * \mathbf{Ci} &= (\mathbf{D} * \mathbf{Ci})^T * \mathbf{Ci} = \mathbf{Ci}^T * (\mathbf{D} * \mathbf{Ci}) \\ &= \mathbf{Ci}^T * n(\mathbf{Ci}) = (n(\mathbf{Ci})^T * \mathbf{Ci})^T \end{aligned}$$

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

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 $\mathbf{AK} = \mathbf{AW} * \mathbf{WK}$.

EU projects on simulation

For the meeting ***The Age of Simulation*** at Ars Electronica in Linz, January 2006 a dataset of EU projects on simulation was collected by FAS research, Vienna and stored in the form of Excel table (`SimPro.csv`).
The rows are the projects participants (idents) and colomns correspond to different their properties. Three two-mode networks were produced from this table using Jürgen Pfeffer's **`Text2Pajek`** program:

- `project.net` – $\mathbf{P} = [\text{idents} \times \text{projects}]$
- `country.net` – $\mathbf{C} = [\text{idents} \times \text{countries}]$
- `institution.net` – $\mathbf{U} = [\text{idents} \times \text{institutions}]$

$|\text{idents}| = 8869$, $|\text{projects}| = 933$, $|\text{institutions}| = 3438$, $|\text{countries}| = 60$.

EU projects – derived networks

Since all three networks have the common set (idents) we can derive from them using *network multiplication*

```
Nets/Multiply First*Second
```

several interesting networks:

- `ProjInst.net` – $\mathbf{W}$ = [projects × institutions] = $\mathbf{P}^T * \mathbf{U}$
- `Countries.net` – $\mathbf{S}$ = [countries × countries] = $\mathbf{C}^T * \mathbf{C}$
- `Institutions.net` – $\mathbf{Q}$ = [institutions × institutions] = $\mathbf{W}^T * \mathbf{W}$
- ...

```
Network/2-Mode Network/2-Mode to 1-Mode/Rows
Network/2-Mode Network/2-Mode to 1-Mode/Columns
```

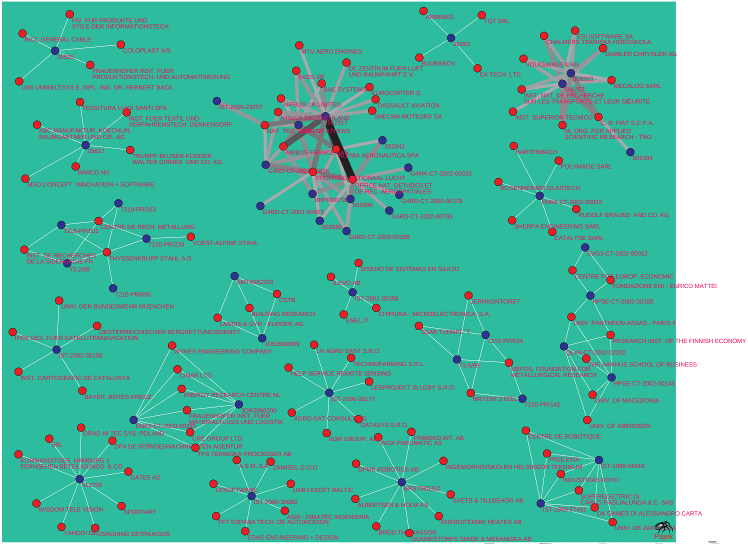
Analysis of `ProjInst.net`

For identifying important parts of `ProjInst.net` we first computed the 4-rings weights and in the obtained network we determined the line islands

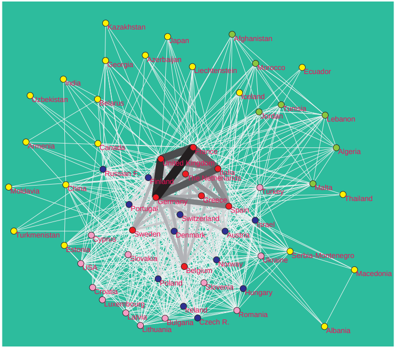
```
Network/Create New Network/With Ring Counts .../4-Rings/Undirect
Network/Create Partition/Islands/Line Weights[Simple] [2,200]
```

We obtain 101 islands. We extracted 18 islands of the size at least 5. There are two most important islands: aviation companies and car companies.
In labels we used the option `\n`.

Analysis of `ProjInst.net`



Analysis of `Countries.net`



To obtain picture in which the stronger links cover weaker links we have to sort them

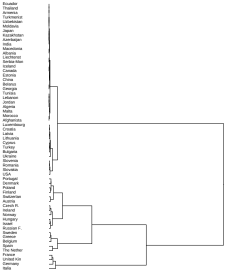
`Network/Create New Network/ Transform/Sort lines/ Line values/Ascending`

For dense (sub)networks we get better visualization by using matrix display. In this case we also recoded values (2,10,50).

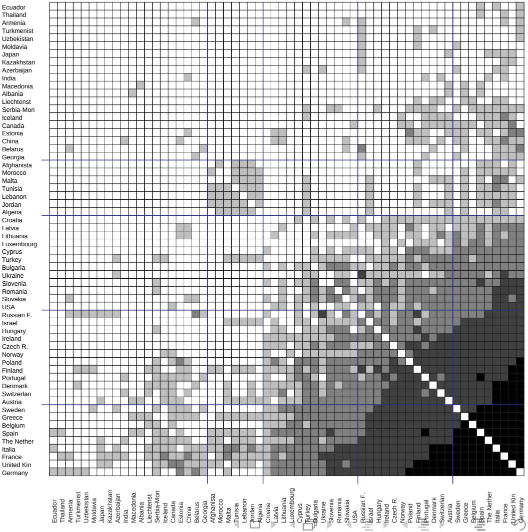
To determine clusters we used Ward's clustering procedure with dissimilarity measure d_5 (corrected Euclidean distance).
The permutation determined by hierarchy can often be improved by changing the positions of clusters. We get a typical center-periphery structure.

Analysis of `Countries.net`

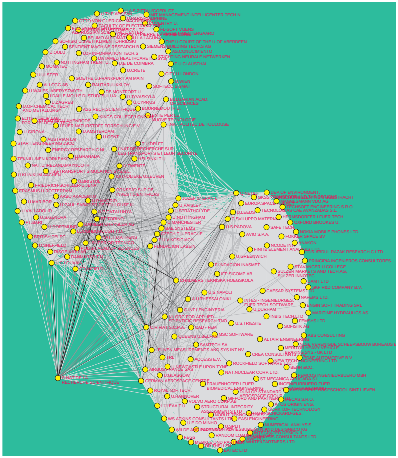
Pajek - Ward [0.00,4785.14]



Pajek - shadow [0.00,4.00]

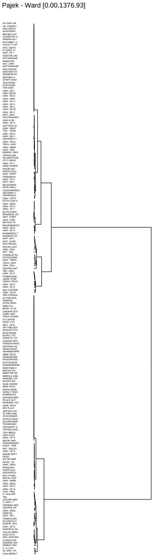


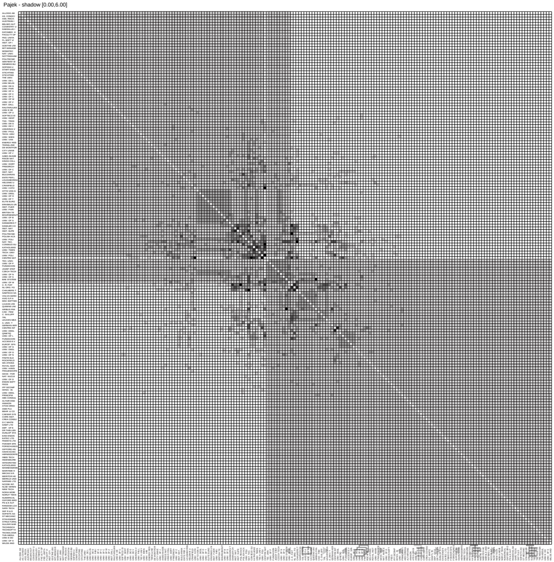
Analysis of `Institutions.net`



To identify the most important institutions we first computed p_S-cores vector and use it to determine the corresponding node islands. We got essentially one large island. Again the corresponding subnetwork is very dense. We prepared also a matrix display.

Analysis of `Institutions.net`





Temporal network and Levels of analysis

We can also transform the citation network (and other WoS networks) into temporal network using the partition of works by publication year.

Using the time slices also the temporal sequences of corresponding derived networks can be obtained.
Note that most of the obtained derived networks are one-mode networks. To analyze them standard SNA methods can be used.
In the analysis of the obtained networks the comparability of units could/should be considered.

We are developing a special approach to temporal networks based on temporal quantities. paper

`Pajek` allows analyses on different levels specified by a partition of the corresponding set of units and obtained using the *shrinking* of classes.
For example: partition of authors by institutions, or partition of institutions by countries, partitions of authors by discipline/ field/ subfield, etc.
Using the *extraction* of selected classes we can reduce the network to the area of our interest.

Temporal quantities

We introduce a notion of a *temporal quantity*

$$a(t) = \begin{cases} a'(t) & t \in T_a \\ \text{⌘} & t \in \mathcal{T} \setminus T_a \end{cases}$$

where T_a is the *activity time set* of a and $a'(t)$ is the value of a in an instant $t \in T_a$, and ⌘ denotes the value *undefined*.

We assume that the values of temporal quantities belong to a set A which is a *semiring* $(A, +, \cdot, 0, 1)$ for binary operations $+ : A \times A \to A$ and $\cdot : A \times A \to A$.
Let $A_{⌘}(\mathcal{T})$ denote the set of all temporal quantities over $A_{⌘}$ in time $\mathcal{T}$. To extend the operations to networks and their matrices we first define the *sum* (parallel links) $a + b$ as

$$(a+b)(t) = a(t) + b(t) \quad \text{and} \quad T_{a+b} = T_a \cup T_b.$$

The *product* (sequential links) $a \cdot b$ is defined as

$$(a \cdot b)(t) = a(t) \cdot b(t) \quad \text{and} \quad T_{a \cdot b} = T_a \cap T_b.$$

Sum and product of temporal quantities


```
a = [(1, 5, 2), (6, 8, 1), (11, 12, 3), (14, 16, 2),
     (17, 18, 5), (19, 20, 1)]
b = [(2, 3, 4), (4, 7, 3), (9, 10, 2), (13, 15, 5),
     (16, 21, 1)]
```

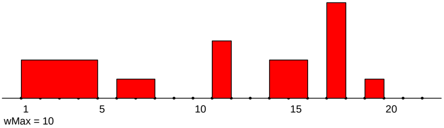
The following are the sum $s = a + b$ and the product $p = a \cdot b$ of temporal quantities a and b over combinatorial semiring.

```
s = [(1, 2, 2), (2, 3, 6), (3, 4, 2), (4, 5, 5), (5, 6, 3),
     (6, 7, 4), (7, 8, 1), (9, 10, 2), (11, 12, 3),
     (13, 14, 5), (14, 15, 7), (15, 16, 2), (16, 17, 1),
     (17, 18, 6), (18, 19, 1), (19, 20, 2), (20, 21, 1)]
p = [(2, 3, 8), (4, 5, 6), (6, 7, 3), (14, 15, 10),
     (17, 18, 5), (19, 20, 1)]
```

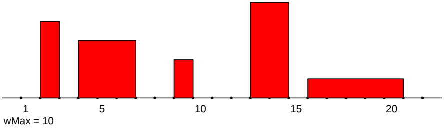
They are visually displayed at the bottom half of figures on the following slides.

Addition of temporal quantities.

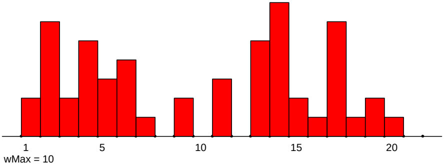
a :



b :

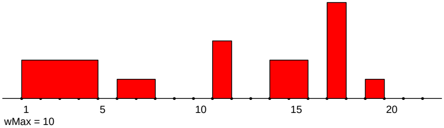


$a + b$:

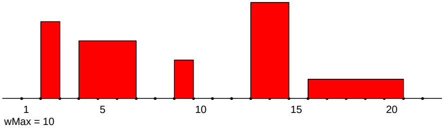


Multiplication of temporal quantities.

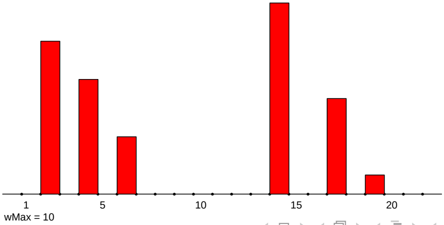
a :



b :



$a \cdot b$:



Temporal affiliation networks

Let the binary *affiliation* matrix $\mathbf{A} = [a_{ep}]$ describe a two-mode network on the set of events E and the set of of participants P:

$$a_{ep} = \begin{cases} 1 & p \text{ participated in the event } e \\ 0 & \text{otherwise} \end{cases}$$

The function $d : E \to \mathcal{T}$ assigns to each event e the date $d(e)$ when it happened. $\mathcal{T} = [\mathit{first}, \mathit{last}] \subset \mathbb{N}$. Using these data we can construct two temporal affiliation matrices:

- **instantaneous Ai** $= [ai_{ep}]$, where

$$ai_{ep} = \begin{cases} [(d(e), d(e)+1, 1)] & a_{ep} = 1 \\ [\,] & \text{otherwise} \end{cases}$$

- **cumulative Ac** $= [ac_{ep}]$, where

$$ac_{ep} = \begin{cases} [(d(e), \mathit{last}+1, 1)] & a_{ep} = 1 \\ [\,] & \text{otherwise} \end{cases}$$

Multiplication of temporal affiliation networks
Instantaneous

Instantaneous **A** on $P \times A$ and **B** on $P \times B$. $\mathbf{C} = \mathbf{A}^T.\mathbf{B}$ on $A \times B$.

$$c_{ij}(t) = \sum_{p \in P} a_{pi}(t)^T \cdot b_{pj}(t)$$

$a_{pi} = [(d_{pi}, d_{pi} + 1, v_{pi})]$ and $b_{pj} = [(d_{pj}, d_{pj} + 1, v_{pj})]$
for $t = d$ we get

$$c_{ij} = [(d, d+1, \sum_{p \in P: d_{pi} = d_{pj} = d} v_{pi}.v_{pj})]_{d \in \mathcal{T}}$$

for $v_{pi} = v_{pj} = 1$ we finally get

$$v_{ij}(d) = |\{p \in P : d_{pi} = d_{pj} = d\}|$$

For binary temporal two-mode networks **A** and **B** the value $v_{ij}(d)$ of the product $\mathbf{A}^T.\mathbf{B}$ is equal to the number of different members of P with which both i and j have contact in the instant d.

Multiplication of temporal affiliation networks

Cumulative

Cumulative **A** on $P \times A$ and **B** on $P \times B$. $\mathbf{C} = \mathbf{A}^T.\mathbf{B}$ on $A \times B$.

$$c_{ij}(t) = \sum_{p \in P} a_{pi}(t)^T \cdot b_{pj}(t)$$

$a_{pi} = [(d_{pi}, last + 1, v_{pi})]$ and $b_{pj} = [(d_{pj}, last + 1, v_{pj})]$
for $t = d$ we get

$$c_{ij} = [(d, d+1, \sum_{p \in P : (d_{pi} \leq d) \wedge (d_{pj} \leq d)} v_{pi}.v_{pj})]_{d \in \mathcal{T}}$$

for $v_{pi} = v_{pj} = 1$ we finally get

$$v_{ij}(d) = |\{p \in P : (d_{pi} \leq d) \wedge (d_{pj} \leq d)\}|$$

For binary temporal two-mode networks **A** and **B** the value $v_{ij}(d)$ of the product $\mathbf{A}^T.\mathbf{B}$ is equal to the number of different members of P with which both i and j have contact in all instants up to including the instant d.

Temporal co-authorship networks

Using the multiplication of temporal matrices over the combinatorial semiring we get the corresponding instantaneous and cumulative co-occurrence matrices

$$\mathbf{Ci} = \mathbf{Ai}^T \cdot \mathbf{Ai} \qquad \text{and} \qquad \mathbf{Cc} = \mathbf{Ac}^T \cdot \mathbf{Ac}$$

A typical example of such a matrix is the papers authorship matrix **WA** where E is the set of papers W, P is the set of authors A and d is the publication year.

The triple (s, f, v) in a temporal quantity ci_{pq} tells that in the time interval $[s, f)$ there were v events in which both p and q took part.

The triple (s, f, v) in a temporal quantity cc_{pq} tells that in the time interval $[s, f)$ there were in total v accumulated events in which both p and q took part.

The diagonal matrix entries ci_{pp} and cc_{pp} contain the temporal quantities counting the number of events in the time intervals in which the participant p took part.

Temporal co-authorship network for SN5

BibTime

SN5 (2008)

	W	A	K	J
raw	193376	75930	29267	14651
DC=1	7950	12458		

In **Pajek** we extract a subnetwork **WAc** and a corresponding partition **SN5yearC**. Using a program `twoMode2netJSON` we transform them into temporal network in the netJSON format.

Bibliographic networks are usually sparse. The network **WAcInst** has 19488 arcs. The co-authorship network **CoInst** $=$ **WAcInst**$^T *$ **WAcInst** has 64980 edges; the corresponding matrix in the package **TQ** has $12458^2 = 155201764$ entries. Using a package **Graph** the co-authorship network is computed in a second and half – a big speed-up.

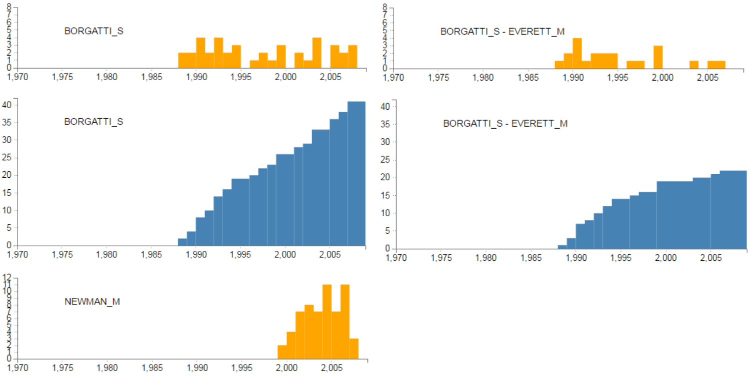
multiply.py

```
gdir = 'c:/users/batagelj/work/python/graph/graph'
wdir = 'c:/users/batagelj/work/python/graph/JSON/SN5'
cdir = 'c:/users/batagelj/work/python/graph/chart'
import sys, os, datetime, json
sys.path = [gdir]+sys.path; os.chdir(wdir)
import TQ
from GraphNew import Graph
# file = 'C:/Users/batagelj/work/Python/graph/JSON/WAtest.json'
file = 'C:/Users/batagelj/work/Python/graph/JSON/SN5/WAcInst.jsc
# file = 'C:/Users/batagelj/work/Python/graph/JSON/SN5/WAcCum.js
# file = 'C:/Users/batagelj/work/Python/graph/JSON/Gisela/papIns
t1 = datetime.datetime.now()
print("started: ",t1.ctime(),"\n")
G = Graph.loadNetJSON(file)
t2 = datetime.datetime.now()
print("\nloaded: ",t2.ctime(),"\ntime used: ", t2-t1)
# T = G.transpose()
# Co = Graph.TQmultiply(T,G,True)
# CR = G.TQtwo2oneRows()
CC = G.TQtwo2oneCols()
t3 = datetime.datetime.now()
print("\ncomputed: ",t3.ctime(),"\ntime used: ", t3-t2)
ia = { v[3]['lab']: k for k,v in CC._nodes.items() }
# CC._links[(ia['BORGATTI_S'],ia['EVERETT_M'])][4]['tq']
# CC._links[(ia['IDI/B'],ia['HCL/B'])][4]['tq']
```

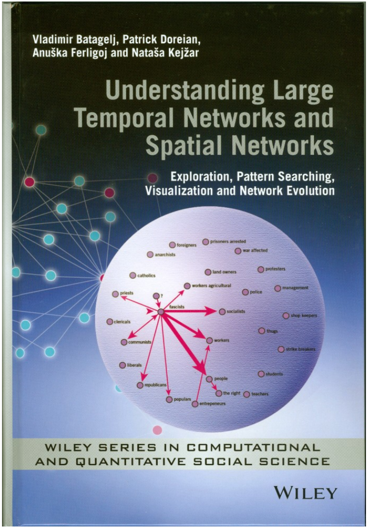
Temporal co-authorship network for SN5

```
======= RESTART: C:\Users\batagelj\work\Python\graph\graph\multiply.py =======
started:  Sun Nov 20 00:26:51 2016
loaded:  Sun Nov 20 00:26:51 2016
time used:  0:00:00.425024
computed:  Sun Nov 20 00:26:52 2016
time used:  0:00:01.165066
>>> BB = CC._links[(ia['BORGATTI_S'],ia['BORGATTI_S'])][4]['tq']
>>> BE = CC._links[(ia['BORGATTI_S'],ia['EVERETT_M'])][4]['tq']
>>> BB
[(1988, 1990, 2), (1990, 1991, 4), (1991, 1992, 2), (1992, 1993, 4),
 (1993, 1994, 2), (1994, 1995, 3), (1996, 1997, 1), (1997, 1998, 2),
 (1998, 1999, 1), (1999, 2000, 3), (2001, 2002, 2), (2002, 2003, 1),
 (2003, 2004, 4), (2005, 2006, 3), (2006, 2007, 2), (2007, 2008, 3)]
>>> BE
[(1988, 1989, 1), (1989, 1990, 2), (1990, 1991, 4), (1991, 1992, 1),
 (1992, 1995, 2), (1996, 1998, 1), (1999, 2000, 3), (2003, 2004, 1),
 (2005, 2007, 1)]
>>> TQmax = 8; Tmin = 1970; Tmax = 2009; w = 600; h = 120
>>> tit = 'BORGATTI_S'
>>> Graph.TQshow(BB,cdir,TQmax,Tmin,Tmax,w,h,tit,fill='orange')
>>> tit = 'BORGATTI_S - EVERETT_M'
>>> Graph.TQshow(BE,cdir,TQmax,Tmin,Tmax,w,h,tit,fill='orange')
>>> NN = CC._links[(ia['NEWMAN_M'],ia['NEWMAN_M'])][4]['tq']
>>> NN
[(1999, 2000, 2), (2000, 2001, 4), (2001, 2002, 7), (2002, 2003, 8),
 (2003, 2004, 7), (2004, 2005, 11), (2005, 2006, 7), (2006, 2007, 11),
 (2007, 2008, 3)]
>>> tit = 'NEWMAN_M'; TQmax = 12; h = 150
>>> Graph.TQshow(NN,cdir,TQmax,Tmin,Tmax,w,h,tit,fill='orange')
```

Visualization



Understanding large networks



This course is closely related to chapters 2 and 3 in the book:

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, October 2014.