Analysis of bibliographic networks

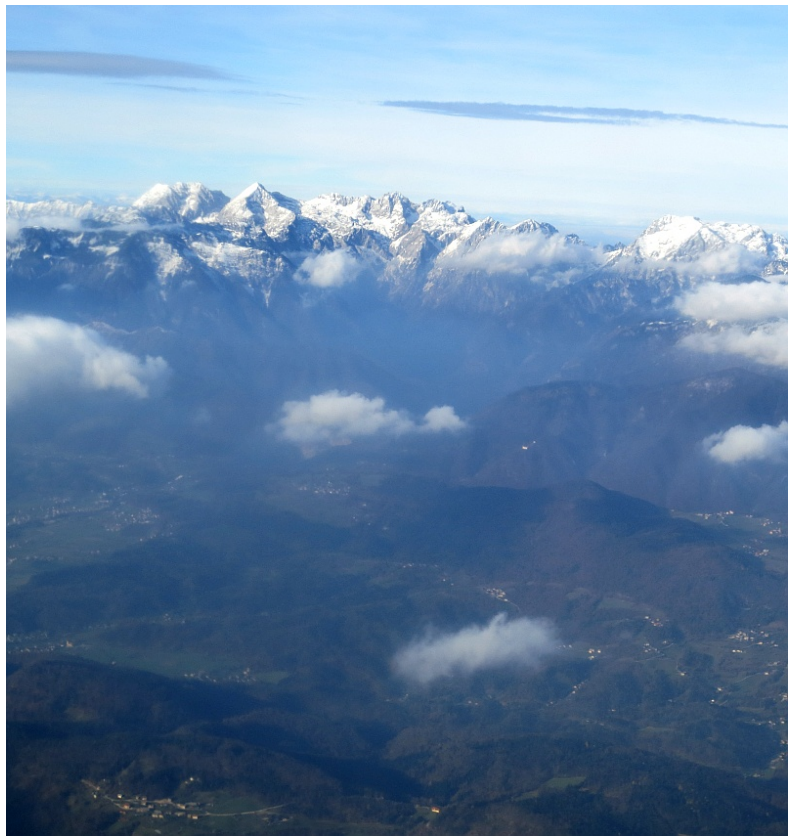
Vladimir Batagelj

IMFM Ljubljana and IAM UP Koper

Guest lecture, PhD course on networks

Uppsala University, November 18, 2016

Outline



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

Current version of slides (November 20, 2016 at 01 : 06): slides PDF
http://vlado.fmf.uni-lj.si/pub/slides/bibnet16.pdf

Networks

A *network* is based on two sets – a set of *nodes* (vertices), that represent the selected *units*, and a set of *links* (lines), that represent *ties* between units. They determine a *graph*. A link can be *directed* – an *arc*, or *undirected* – an *edge*.
Additional data about nodes or links may be known – their *properties* (attributes). For example: name/label, type, age, value, ...

Network = Graph + Data

A *network* $\mathcal{N} = (\mathcal{V}, \mathcal{L}, \mathcal{P}, \mathcal{W})$ consists of:

- a *graph* $\mathcal{G} = (\mathcal{V}, \mathcal{L})$, where $\mathcal{V}$ is the set of nodes, $\mathcal{A}$ is the set of arcs, $\mathcal{E}$ is the set of edges, and $\mathcal{L} = \mathcal{E} \cup \mathcal{A}$ is the set of links.
  $n = |\mathcal{V}|$, $m = |\mathcal{L}|$
- $\mathcal{P}$ *vertex value functions* / properties: $p : \mathcal{V} \to A$
- $\mathcal{W}$ *link value functions* / weights: $w : \mathcal{L} \to B$

Types of networks

In a *two-mode* network $\mathcal{N} = ((\mathcal{V}_1, \mathcal{V}_2), \mathcal{L}, \mathcal{P}, \mathcal{W})$ its set of nodes is split to two subsets. Each link has its end-nodes in both sets.

In a *multi-relational* network $\mathcal{N} = (\mathcal{V}, (\mathcal{L}_i, i \in I), \mathcal{P}, \mathcal{W})$ the set of its links is partitioned into several mutally disjoint subsets – relations. (Subject Verb Object).

In a *temporal* network $\mathcal{N} = (\mathcal{V}, \mathcal{L}, \mathcal{T}, \mathcal{P}, \mathcal{W})$ the time $\mathcal{T}$ is added. To each node and to each link its *activity* set is assigned. Also properties and weights can change through time – temporal quantities.

A *collection* of networks consists of some networks with common subsets of nodes.

Types of networks can be combined – for example: a temporal two-mode multi-relational network.

Description of networks

How to describe a network $\mathcal{N}$? In principle the answer is simple – we list its components $\mathcal{V}$, $\mathcal{L}$, $\mathcal{P}$, and $\mathcal{W}$.

The simplest way is to describe a network $\mathcal{N}$ by providing $(\mathcal{V}, \mathcal{P})$ and $(\mathcal{L}, \mathcal{W})$ in a form of two tables.

As an example, let us describe a part of network determined by the following works:
Generalized blockmodeling, Clustering with relational constraint, Partitioning signed social networks, The Strength of Weak Ties

There are nodes of different types (modes): persons, papers, books, series, journals, publishers; and different relations among them: author_of, editor_of, contained_in, cites, published_by.

Both tables are often maintained in Excel. They can be exported as text in CSV (Comma Separated Values) format.

bibNodes.csv

```
name;mode;country;sex;year;vol;num;fPage;lPage;x;y
"Batagelj, Vladimir";person;SI;m;;;;;;809.1;653.7
"Doreian, Patrick";person;US;m;;;;;;358.5;679.1
"Ferligoj, Anuška";person;SI;f;;;;;;619.5;680.7
"Granovetter, Mark";person;US;m;;;;;;145.6;660.5
"Moustaki, Irini";person;UK;f;;;;;;783.0;228.0
"Mrvar, Andrej";person;SI;m;;;;;;478.0;630.1
"Clustering with relational constraint";paper;;;1982;47;;413;426;684.1;3
"The Strength of Weak Ties";paper;;;1973;78;6;1360;1380;111.3;329.4
"Partitioning signed social networks";paper;;;2009;31;1;1;11;408.0;337.8
"Generalized Blockmodeling";book;;;2005;24;;1;385;533.0;445.9
"Psychometrika";journal;;;;;;;;741.8;086.1
"Social Networks";journal;;;;;;;;321.4;236.5
"The American Journal of Sociology";journal;;;;;;;;111.3;168.9
"Structural Analysis in the Social Sciences";series;;;;;;;;310.4;082.8
"Cambridge University Press";publisher;UK;;;;;;;534.3;238.2
"Springer";publisher;US;;;;;;;884.6;174.0
```

bibNodes.csv

In large networks, to avoid the empty cells, we split a network to some subnetworks – a collection.

bibLinks.csv

```
from;relation;to
"Batagelj, Vladimir";authorOf;"Generalized Blockmodeling"
"Doreian, Patrick";authorOf;"Generalized Blockmodeling"
"Ferligoj, Anuška";authorOf;"Generalized Blockmodeling"
"Batagelj, Vladimir";authorOf;"Clustering with relational constraint"
"Ferligoj, Anuška";authorOf;"Clustering with relational constraint"
"Granovetter, Mark";authorOf;"The Strength of Weak Ties"
"Granovetter, Mark";editorOf;"Structural Analysis in the Social Sciences
"Doreian, Patrick";authorOf;"Partitioning signed social networks"
"Mrvar, Andrej";authorOf;"Partitioning signed social networks"
"Moustaki, Irini";editorOf;"Psychometrika"
"Doreian, Patrick";editorOf;"Social Networks"
"Generalized Blockmodeling";containedIn;"Structural Analysis in the Soci
"Clustering with relational constraint";containedIn;"Psychometrika"
"The Strength of Weak Ties";containedIn;"The American Journal of Sociolo
"Partitioning signed social networks";containedIn;"Social Networks"
"Partitioning signed social networks";cites;"Generalized Blockmodeling"
"Generalized Blockmodeling";cites;"Clustering with relational constraint
"Structural Analysis in the Social Sciences";publishedBy;"Cambridge Univ
"Psychometrika";publishedBy;"Springer"
```

bibLinks.csv

Large networks

The *size* of a network is usually measured by the number of nodes n and the number of links m. In a simple network (no parallel links) it holds $m \leq n^2$.

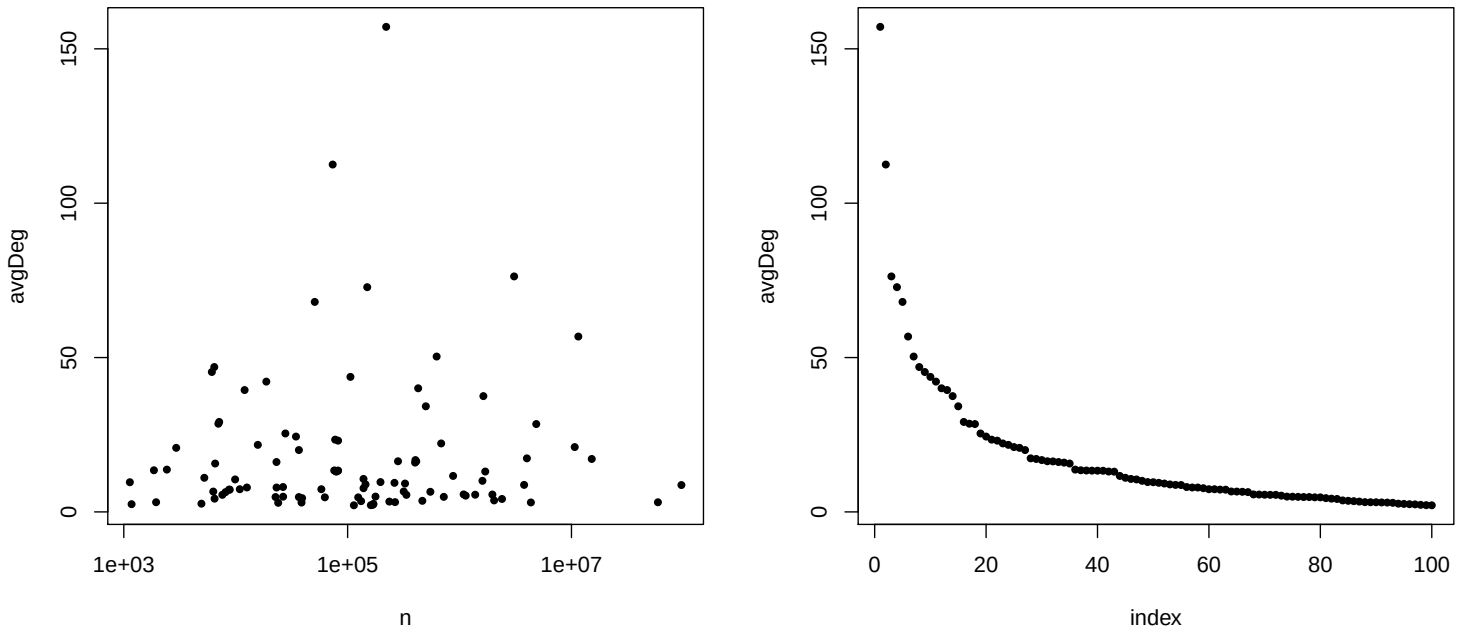
Large networks are networks with at least some thousands of nodes that can be stored entirely in the computer's memory. Huge networks.

Large networks are usually *sparse*, $m \leq k \cdot n$, where $k << n$ (see Dunbar's number). This is a crucial property that often allows us to develop efficient (subquadratic) algorithms for analysis of large networks.

To support the analysis of large networks we (Andrej Mrvar and me) started in 1996 to develop a program `Pajek`.

Dunbar's number

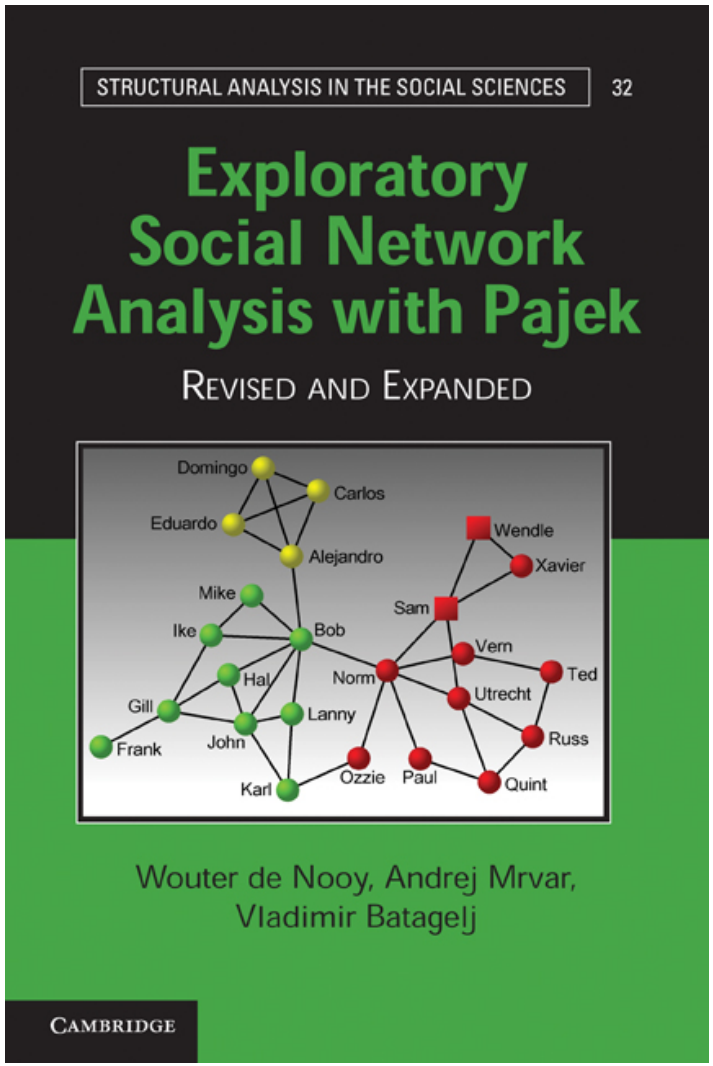
Average degrees of the SNAP and Konect networks



Average degree $\bar{d} = \frac{1}{n}\sum_{v \in V} \deg(v) = \frac{2m}{n}$. Most real-life large networks are *sparse* – the number of nodes and links are of the same order. This property is also known as a Dunbar's number.

The basic idea is that if each vertex has to spend for each link certain amount of "energy" to maintain the links to selected other vertices then, since it has a limited "energy" at its disposal, the number of links should be limited. In human networks the Dunbar's number is between 100 and 150.

ESNA Pajek



An introduction to social network analysis with Pajek is available in the book ESNA (de Nooy, Mrvar, Batagelj 2005). Second extended edition in September 2011.

ESNA in Japanese was published by Tokyo Denki University Press in 2010; and in Chinese by Beijing World Publishing in November 2012.

Pajek – program for analysis and visualization of large networks is freely available, for noncommercial use, at its web site.

http://pajek.imfm.si/
Pajek 2.* → Pajek 3.*

Factorization and description of large networks

To save space and improve the computing efficiency we often replace values of categorical variables with integers. In R this encoding is called a *factorization*.

We enumerate all possible values of a given categorical variable (coding table) and afterwards replace each its value by the corresponding index in the coding table.

This approach is used in most programs dealing with large networks. Unfortunately the coding table is often a kind of meta-data.

CSV2Pajek.R

```
# transforming CSV file to Pajek files
# by Vladimir Batagelj, June 2016
setwd("C:/Users/batagelj/work/Python/graph/SVG/EUSN")
colC <- c(rep("character",4),rep("integer",7)); nas <- c("","NA","NaN")
nodes <- read.csv2("bibNodes.csv",encoding='UTF-8',colClasses=colC,na.strings=nas)
n <- nrow(nodes); M <- factor(nodes$mode); S <- factor(nodes$sex)
mod <- levels(M); sx <- levels(S); S <- as.numeric(S); S[is.na(S)] <- 0
links <- read.csv2("bibLinks.csv",encoding='UTF-8',colClasses="character")
F <- factor(links$from,levels=nodes$name,ordered=TRUE)
T <- factor(links$to,levels=nodes$name,ordered=TRUE)
R <- factor(links$relation); rel <- levels(R)
net <- file("bib.net","w"); cat('*vertices ',n,'\n',file=net)
clu <- file("bibMode.clu","w"); sex <- file("bibSex.clu","w")
cat('%',file=clu); cat('%',file=sex)
for(i in 1:length(mod)) cat(' ',i,mod[i],file=clu)
cat('\n*vertices ',n,'\n',file=clu)
for(i in 1:length(sx)) cat(' ',i,sx[i],file=sex)
cat('\n*vertices ',n,'\n',file=sex)
for(v in 1:n) {
  cat(v,' "',nodes$name[v],'"\n',sep='',file=net);
  cat(M[v],'\n',file=clu); cat(S[v],'\n',file=sex)
}
for(r in 1:length(rel)) cat('*arcs :',r,' "',rel[r],'"\n',sep='',file=net)
cat('*arcs\n',file=net)
for(a in 1:nrow(links))
  cat(R[a],': ',F[a],' ',T[a],' 1 1 "',rel[R[a]],'"\n',sep='',file=net)
close(net); close(clu); close(sex)
```

CSV2Pajek.R

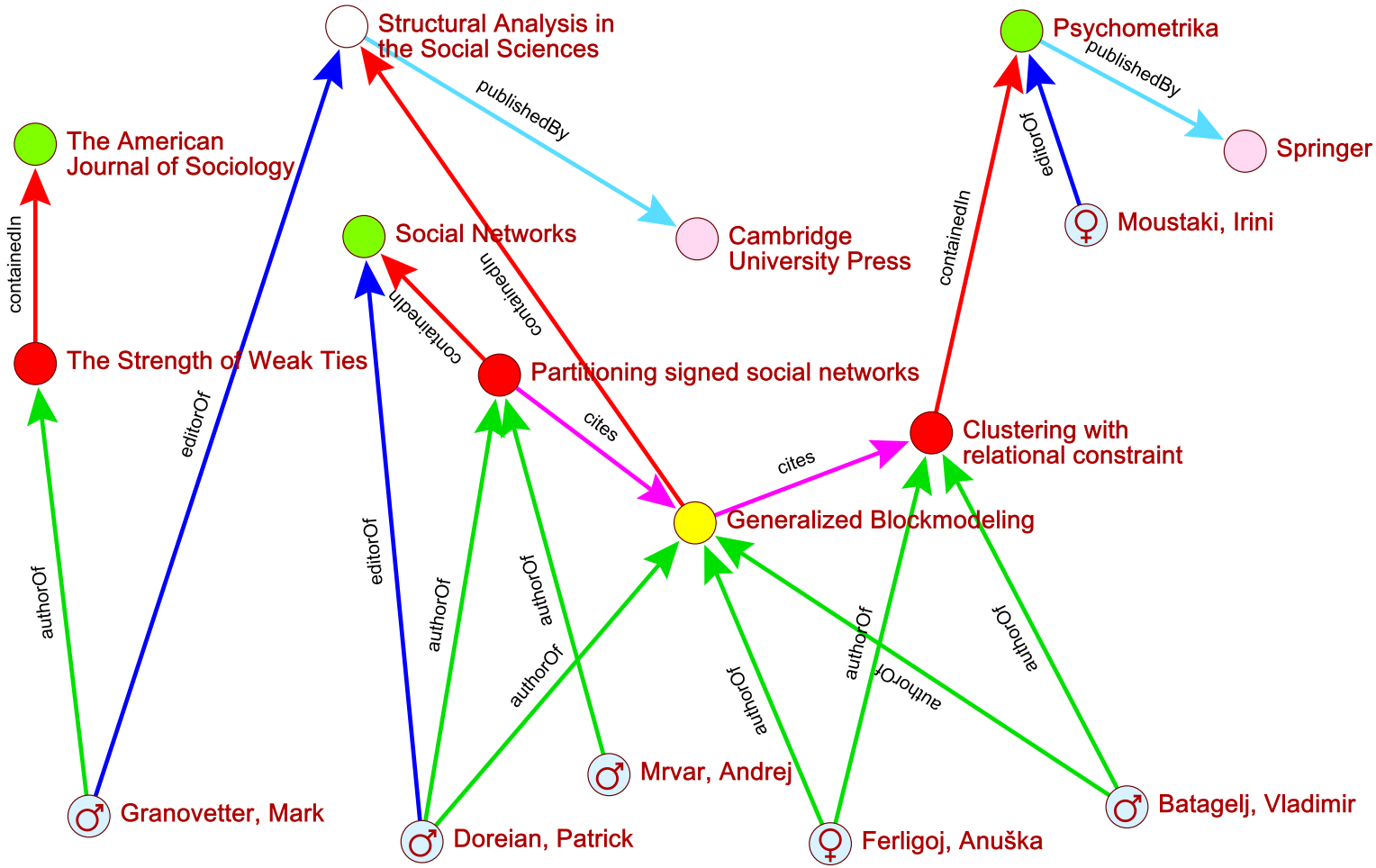
bib.net

```
*vertices  16
1 "Batagelj, Vladimir"
2 "Doreian, Patrick"
3 "Ferligoj, Anuška"
4 "Granovetter, Mark"
5 "Moustaki, Irini"
6 "Mrvar, Andrej"
7 "Clustering with relational constraint"
8 "The Strength of Weak Ties"
9 "Partitioning signed social networks"
10 "Generalized Blockmodeling"
11 "Psychometrika"
12 "Social Networks"
13 "The American Journal of Sociology"
14 "Structural Analysis in the Social Sciences"
15 "Cambridge University Press"
16 "Springer"
*arcs :1 "authorOf"
*arcs :2 "cites"
*arcs :3 "containedIn"
*arcs :4 "editorOf"
*arcs :5 "publishedBy"
```

```
*arcs
1: 1 10 1 1 "authorOf"
1: 2 10 1 1 "authorOf"
1: 3 10 1 1 "authorOf"
1: 1 7 1 1 "authorOf"
1: 3 7 1 1 "authorOf"
1: 4 8 1 1 "authorOf"
4: 4 14 1 1 "editorOf"
1: 2 9 1 1 "authorOf"
1: 6 9 1 1 "authorOf"
4: 5 11 1 1 "editorOf"
4: 2 12 1 1 "editorOf"
3: 10 14 1 1 "containedIn"
3: 7 11 1 1 "containedIn"
3: 8 13 1 1 "containedIn"
3: 9 12 1 1 "containedIn"
2: 9 10 1 1 "cites"
2: 10 7 1 1 "cites"
5: 14 15 1 1 "publishedBy"
5: 11 16 1 1 "publishedBy"
```

bib.net, bibMode.clu, bibSex.clu; bib.paj, bib.ini.

Bibliographic network – picture / Pajek



Networks from bibliographic data

From special bibliographies (BibTEX) and bibliographic services (Web of Science, Scopus, SICRIS, CiteSeer, Zentralblatt MATH, Google Scholar, DBLP Bibliography, US patent office, IMDb, and others) we can derive some two-mode networks on selected topics:

works $\times$ authors (**WA**),
works $\times$ keywords (**WK**);

and from some data also the network

works $\times$ classification (**WC**)

and the one-mode citation network

works $\times$ works (**Ci**);

where works include papers, reports, books, patents etc.

Besides this we get also at least the partition of works by the journal or publisher, the partition of works by the publication year, and the vector of number of pages.

Records from BiBTEX

```
@Article{int:Mizuno1,
  author =        "S. Mizuno",
  title =         "An ${O(n^{3}L)}$ algorithm using a sequence for
                  linear complementarity problems",
  journal =       "Journal of the Operations Research Society of Japan",
  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 scale
                  optimisation",
  booktitle =     "Logistics, Where Ends Have to Meet~: Proceedings of
                  the Shell Conference on Logistics in Apeldoorn, The
                  Netherlands, November 1988",
  editor =        "{C. F. H. van} Rijn",
  year =          "1989",
  pages =         "112--119",
  publisher =     "Pergamon Press",
  address =       "Oxford, United Kingdom",
}
```

Bib2Pajek.py

Records from DBLP

```
<article mdate="2004-01-15" key="journals/arscom/BeinekeGL97">
<author>Lowell W. Beineke</author>
<author>Wayne Goddard</author>
<author>Marc J. Lipman</author>
<title>Graphs with Maximum Edge-Integrity.</title>
<year>1997</year>
<volume>46</volume>
<journal>Ars Comb.</journal>
<url>db/journals/arscom/arscom46.html#BeinekeGL97</url>
</article>

<inproceedings mdate="2004-12-09" key="conf/sigcse/BermanD96">
<author>A. Michael Berman</author>
<author>Robert C. Duvall</author>
<title>Thinking about binary trees in an object-oriented world.</title>
<pages>185-189</pages>
<year>1996</year>
<crossref>conf/sigcse/1996</crossref>
<booktitle>SIGCSE</booktitle>
<ee>http://doi.acm.org/10.1145/236536</ee>
<url>db/conf/sigcse/sigcse1996.html#BermanD96</url>
</inproceedings>
```

DBLP2Pajek.py

Records from Zentralblatt

```
an  00549739
ai  gross.mark-d
is  ISSN 0025-5874; ISSN 1432-1823
au  Gross, Mark
py  1993
cc  *14M15 14J15
ti  Surfaces of bidegree $(3,n)$ in $\text{Gr}(1,\bbfP\sp 3)$.
ut  congruence; family of lines
so  Math. Z. 212, No.1, 73-106 (1993).
an  01488230
ai  tiras.yuecel; harmanci.abdullah; -
is  ISSN 0092-7872; ISSN 1532-4125
au  T{\i}ra\D{s}, Y\"ucel; Harmanc{\i}, Abdullah; Smith, P.F.
py  2000
cc  *13A15 13C05
ti  Some remarks on dense submodules of multiplication modules.
ut  multiplication module; dense submodule
so  Commun. Algebra 28, No.5, 2291-2296 (2000).
se  00000057  Communications in Algebra  Commun. Algebra 0092-7872; 1532
```

ZBml.py

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 ONE, 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

Problems in producing networks

Most of the source bibliographic data are semi-structured – they are available in the form of records from some data base. Selected fields in the record represent different units: names of people, names of journals, keywords, IDs of works, countries, institutions . . . Unfortunately the names of these units are usually not stored in a standardized way.

Synonymy: Unit names meaning the same. Make a partition. Identify (shrink) equivalent units.

Homonymy: Same unit names having different meanings. Correct the data in your copy of the data base.

When the unit names are extracted from the text the so called *stopwords* are omitted. The equivalence is automatically determined using stemming or lemmatization.
Names: many ways to write the name. Some data bases are trying to standardize the names (DBLP, ZB, ResearcherId). Chinese, 100 names.
Keywords: provided in data or extracted from the text (title, abstract). Key phrases.
There can be also errors (typos) in the data base – correct them in your copy of the data base data.

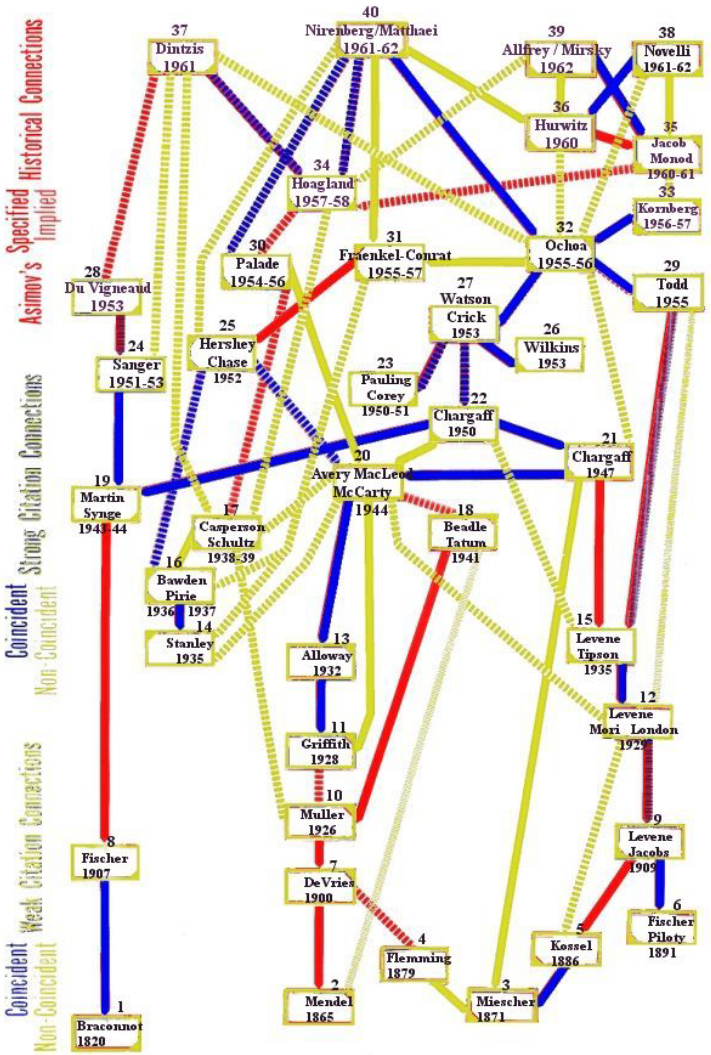
Analyses

The saved records from a data base can still contain some inconsistencies. Some of them are detected as results of the analyses. The simplest way to deal with them is to correct them in the saved data base file and rerun the creation of `Pajek`'s files and analyses.

To improve the quality of the data some tools for detecting (possible) inconsistencies could be developed.

Check (in `Pajek`) the obtained networks for multiple lines and remove them, if they exist. Remove also the loops from the citation network.

Citation networks



The citation network analysis started in 1964 with the paper of Garfield et al. In 1989 Hummon and Doreian proposed three indices – weights of arcs that provide us with automatic way to identify the (most) important part of the citation network. We developed an algorithm to efficiently compute SPC weights, provided the citation network is acyclic.

. . . Citation networks

In a given set of works/nodes W (articles, books, works, etc.) we introduce a *citing relation*/set of arcs $\mathbf{Ci} \subseteq W \times W$

$$u\,\mathbf{Ci}\,v \equiv u \text{ cites } v$$

which determines a *citation network* $\mathcal{N} = (W, \mathbf{Ci})$.

A citing relation is usually *irreflexive* (no loops) and (almost) *acyclic*. We shall assume that it has these two properties. Since in real-life citation networks the strong components are small (usually 2 or 3 nodes) we can transform such network into an acyclic network by shrinking strong components and deleting loops.

Preparing the citation network

Using on a file *PR*`cite.net` the commands

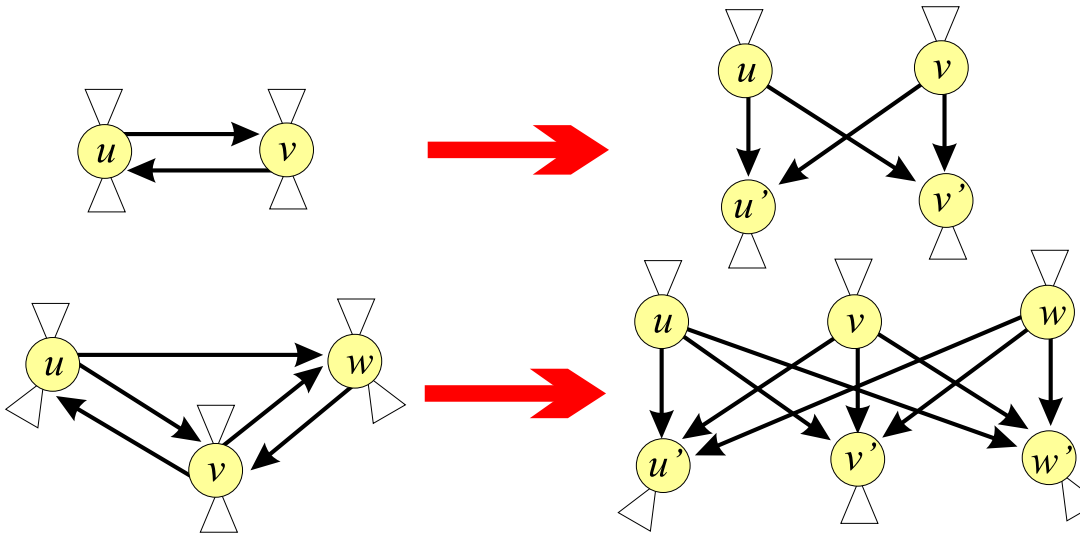
```
Info/Network/General
Net/Transform/Remove/Loops
Net/Transform/Remove lines/Single line
```

we get the information about the number of loops and multiple links. Remove loops, and replace multiple links with single links. The obtained network we save (`Options - Save coordinates [OFF]`) to file *PR*`citeR.net`. For further analysis the citation network has to be acyclic – has no nontrivial strong component. To identify nontrivial strong components and extract them use the commands:

```
Net/Components/Strong [2]
Operations/Extract from Network/Partition [1-*]
Operations/Transform/Remove Lines/Between Clusters
```

Save the obtained network to a file *PR*`strong.net`.

. . . Preparing the citation network



To transform the network *PR*citeR.net into an acyclic network using the preprint transformation use the Pajek's command

`Network/Acyclic Network/Transform/Preprint Transformation`

Network boundary problem

Networks obtained from a WoS file using the program `WoS2Pajek` are in the 'raw' form. We still have to resolve in some way the *network boundary problem*. The first option is to limit the network to the works with complete descriptions – records from the WoS file, $DC = 1$. Since for cited-only works only the first author (no keywords, ...) is known this option is used for most analyses.

We can get a richer network if we decide to include also some cited-only works that are cited often – at least k times; we delete nodes for which it holds $(0 < \text{indeg}(v) < k) \wedge (\text{outdeg}(v) = 0)$.

```
Net/Partition/Degree/Input
Partition/Binarize  [1-(k-1)]
Net/Partition/Degree/Output
Partition/Binarize  [0]
[select partition 1]
[select partition 2]
Partitions/Min(V1,V2)
Operations/Extract from Network/Partition [0]
```

For some (most frequent) of these additional works we can augment the WoS file with their descriptions (without CR data).

PEERE network

To the *Web of Science* (WoS) we put the query `"peer review*"`. In May and June 2015 we got (from Web of Science Core Collection) 17053 hits, and additional 2867 hits for the query `refereeing`.
In March 2016 we updated the data by adding hits for the years 2015 and 2016 and manually prepared short descriptions for the most cited works (fields: AU, PU, TI, PY, PG, KW; but without CR data).
The analysis in 2015 revealed many papers without WoS descriptions having large indegrees in the citation network. We manually searched in WoS for each of them (with indegree larger or equal to 20) and, if found, we added them into the data set. Important earlier papers often did not use the now established terminology and were therefore overlooked by our queries.
The final run of the program `WoS2Pajek` produced networks with sets of the following sizes: works $|W| = 721547$, authors $|A| = 295849$, journals $|J| = 39988$, and keywords $|K| = 36279$. In both phases 22981 records were collected. There were 887 duplicates (considered only once).
We removed multiple links and loops (resulting from homonyms) from the networks. The cleaned citation network **CiteAll** has $n = 721547$ nodes and $m = 869821$ arcs.

PEERE – most cited works

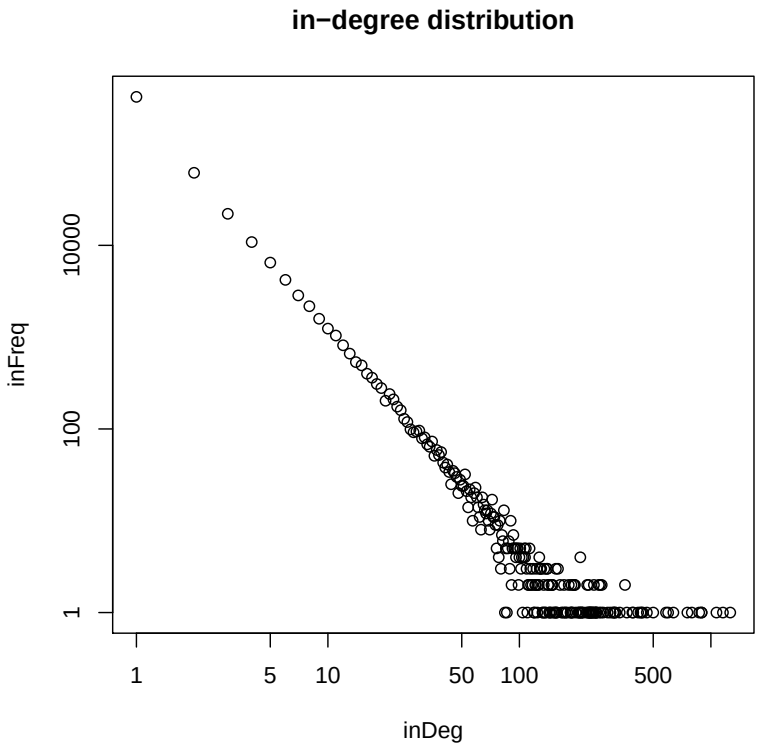
n	freq	first author	title
1	173	Cohen, J	Statistical Power Analysis for the Behavioral Sciences. Routledge,1988
2	164	Peters, DP	Peer-review practices of psychological journals - the fate of . . . Behav Brain S
3	151	Egger, M	Bias in meta-analysis detected by a simple, graphical test. Brit Med J, 1997
4	150	Stroup, DF	Meta-analysis of observational studies in epidemiology - A proposal for repor
5	135	Dersimonian, R	Metaanalysis in clinical-trials. Control Clin Trials, 1986
6	130	Zuckerma.H	Patterns of evaluation in science - institutionalisation, structure and function
7	130	Higgins, JPT	Cochrane Handbook for Systematic Reviews of Interventions. Cochrane, 201
8	126	Moher, D	Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The P
9	125	Higgins, JPT	Measuring inconsistency in meta-analyses. Brit Med J, 2003
10	121	Cicchetti, DV	The reliability of peer-review for manuscript and grant submissions - . . . Beha
11	119	Hirsch, JE	An index to quantify an individual's scientific research output. P Natl Acad S
12	114	Mahoney, M	Publication prejudices: An experimental study of confirmatory bias . . . Cogni
13	114	van Rooyen, S	Effect of open peer review on quality of reviews and on reviewers' recommen
14	114	Easterbrook, PJ	Publication bias in clinical research. Lancet, 1991
15	110	Landis, JR	Measurement Of Observer Agreement For Categorical Data. Biometrics, 197
16	109	Godlee, F	Effect on the quality of peer review of blinding reviewers and asking them to
17	108	Horrobin, DF	The philosophical basis of peer-review and the suppression of innovation. JA
18	107	Moher, D	Preferred Reporting Items for Systematic Reviews and Meta-Analyses: PRISI
19	107	Jadad, AR	Assessing the quality of reports of randomized clinical trials: Is blinding nece
20	105	Mcnutt, RA	The effects of blinding on the quality of peer-review - a randomized trial. JA
21	104	Cole, S	Chance and consensus in peer-review. Science, 1981
22	103	Moher, D	Improving the quality of reports of meta-analyses of randomised controlled tr
23	98	Justice, AC	Does masking author identity improve peer review quality? - A randomized c
24	97	Lock, S	A Difficult Balance: Editorial Peer Review in Medicine. Nuffield Trust, 1985
25	95	van Rooyen, S	Effect of blinding and unmasking on the quality of peer review - A randomize
26	92	Black, N	What makes a good reviewer and a good review for a general medical journa
27	91	Scherer, RW	Full publication of results initially presented in abstracts - a metaanalysis. JA
28	90	Higgins, JPT	Quantifying heterogeneity in a meta-analysis. Stat Med, 2002
29	90	Smith, R	Peer review: a flawed process at the heart of science and journals. J Roy Soc
30	87	Goodman, SN	Manuscript quality before and after peer-review and editing at annals of inte

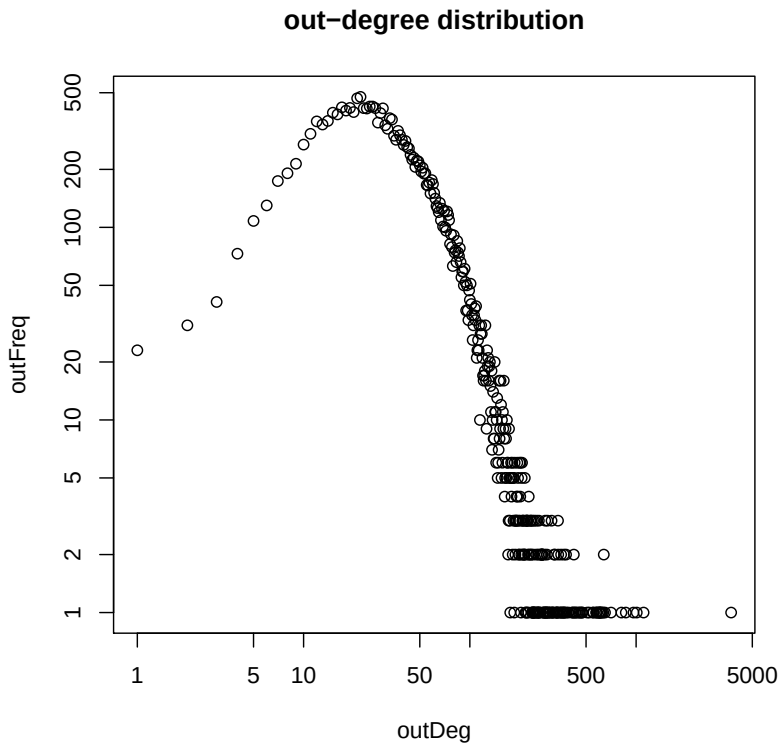
Distributions

We read in Pajek the citation network cite.net for 'centrality' literature. First we remove loops and multiple lines. Then we determine the indegrees and outdegrees. We dispose the normalized degree vectors, transform both partitions into vectors and call R from Pajek submitting all vectors.

```
##########################################
 R called from Pajek
 The following vectors read:
 v3   :  From partition 1 (548600)
 v4   :  From partition 2 (548600)
------------------------------------------------------------
> inTab <- table(v3)
> indeg <- as.integer(names(inTab))
> inDeg <- indeg[indeg>0]
> inFreq <- as.vector(inTab[indeg>0])
> plot(inDeg,inFreq,log='xy',main="in-degree distribution")
> ouTab <- table(v4)
> outdeg <- as.integer(names(ouTab))
> outDeg <- outdeg[outdeg>0]
> outFreq <- as.vector(ouTab[outdeg>0])
> plot(outDeg,outFreq,log='xy',main="out-degree distribution")
```

Distributions





The in-degree distribution is "scale-free"-like. The parameters can be determined using the package of Clauset, Shalizi and Newman. See also Stumpf, et al.: Critical Truths About Power Laws.

Distributions

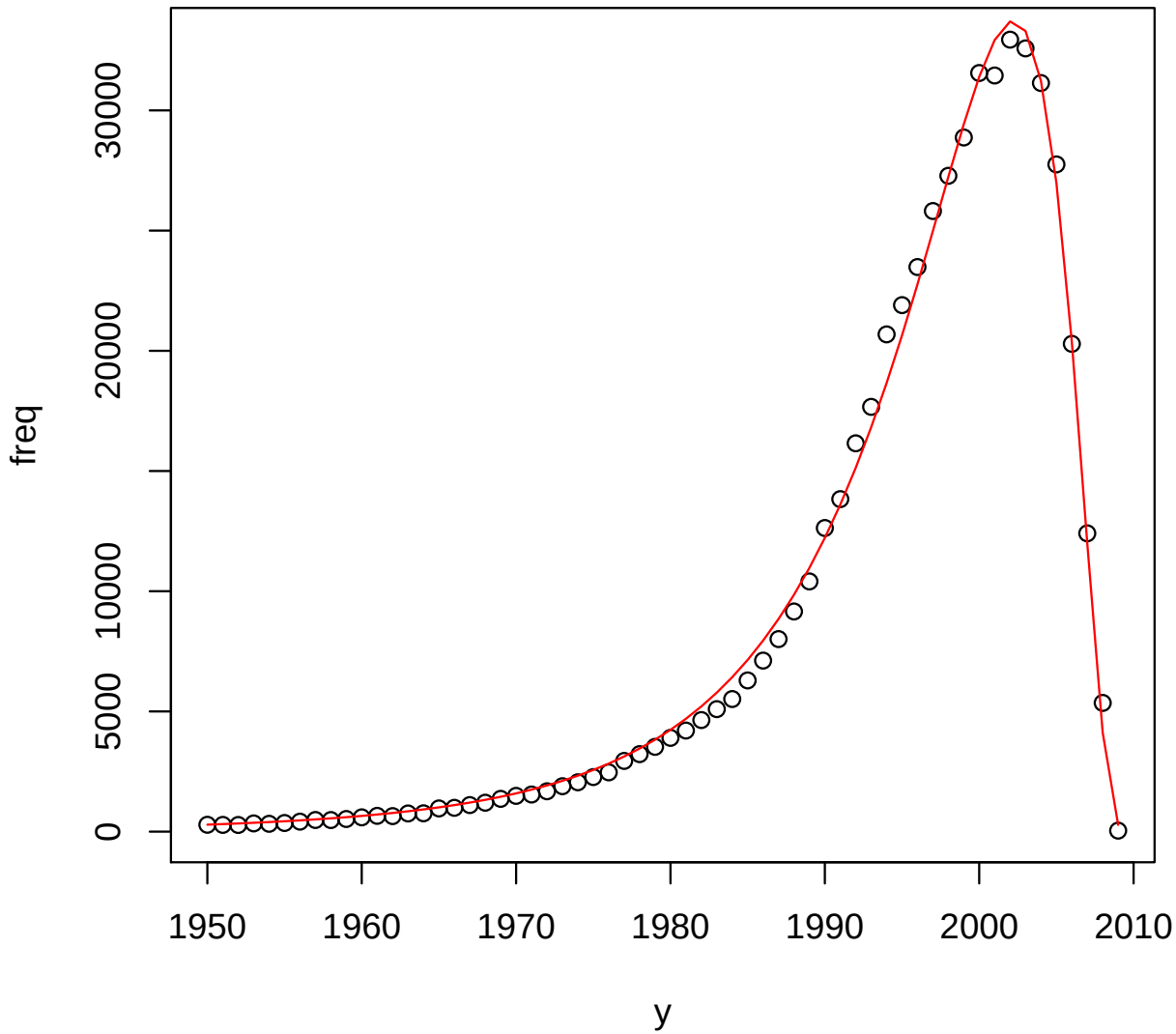
From the file `Year.clu` we can get the distribution of *citations by years*. For the centrality network we get:

```
> setwd("C:/Users/Batagelj/work/Python/WoS/Central")
> years <- read.table(file="Year.clu",header=FALSE,skip=2)$V1
> t <- table(years)
> year <- as.integer(names(t))
> freq <- as.vector(t[1950<=year & year<=2009])
> y <- 1950:2009
> plot(y,freq)
> model <- nls(freq~c*dlnorm(2010-y,a,b),start=list(c=350000,a=2,b=0.7))
> model
Nonlinear regression model
  model:  freq ~ c * dlnorm(2010 - y, a, b)
   data:  parent.frame()
        c         a         b
5.427e+05 2.491e+00 6.624e-01
 residual sum-of-squares: 20474181

Number of iterations to convergence: 7
Achieved convergence tolerance: 3.978e-06
> lines(y,predict(model,list(x=2010-y)),col='red')
```

It can be well approximated by the *lognormal distribution*, but also by the *generalized reciprocal power exponential curve* $c * (x + d)^{\frac{a}{b+x}}$.

Distributions



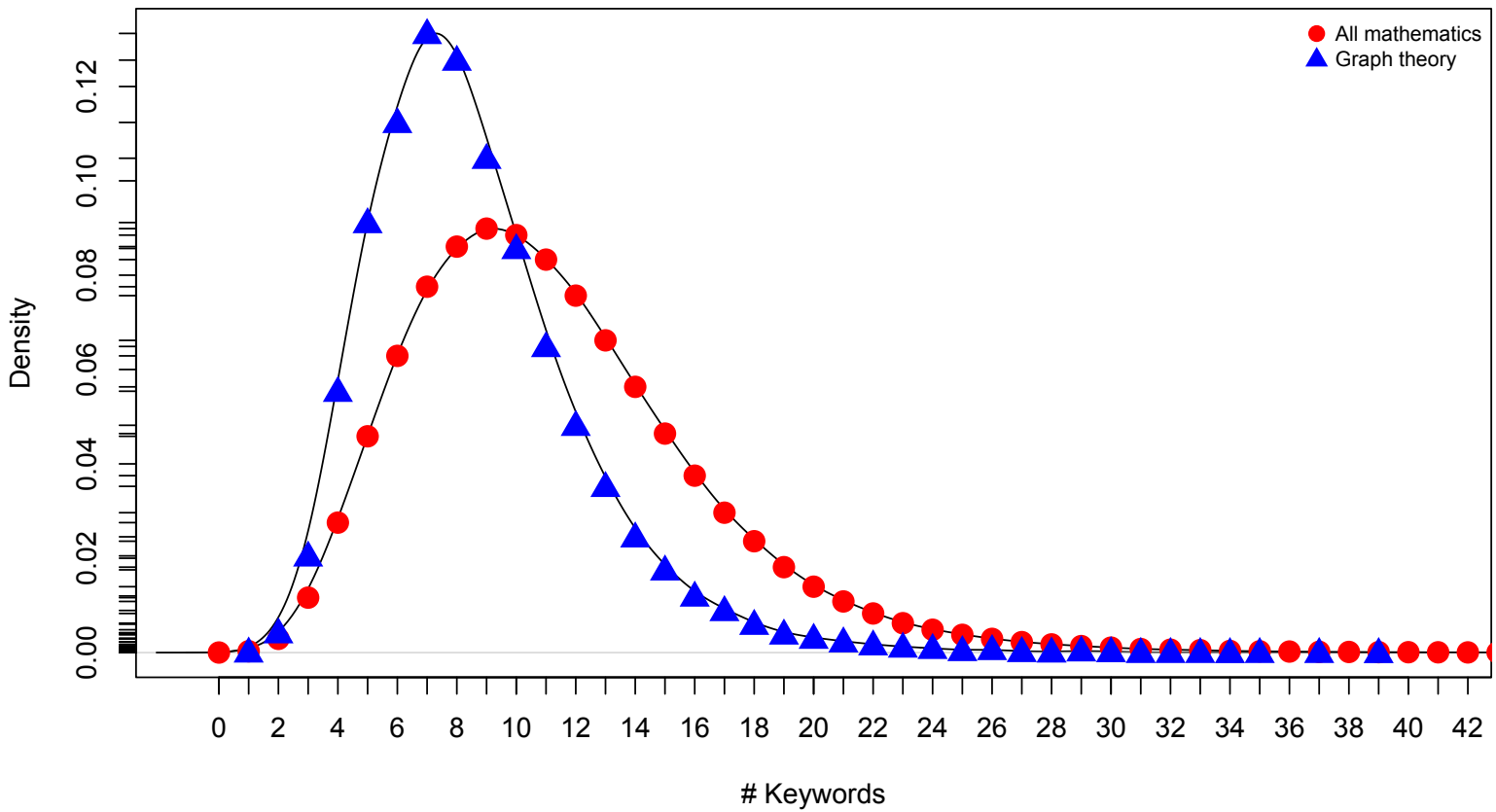
Zentralblatt MATH networks

Data from the ZB data base for years 1990–2010.

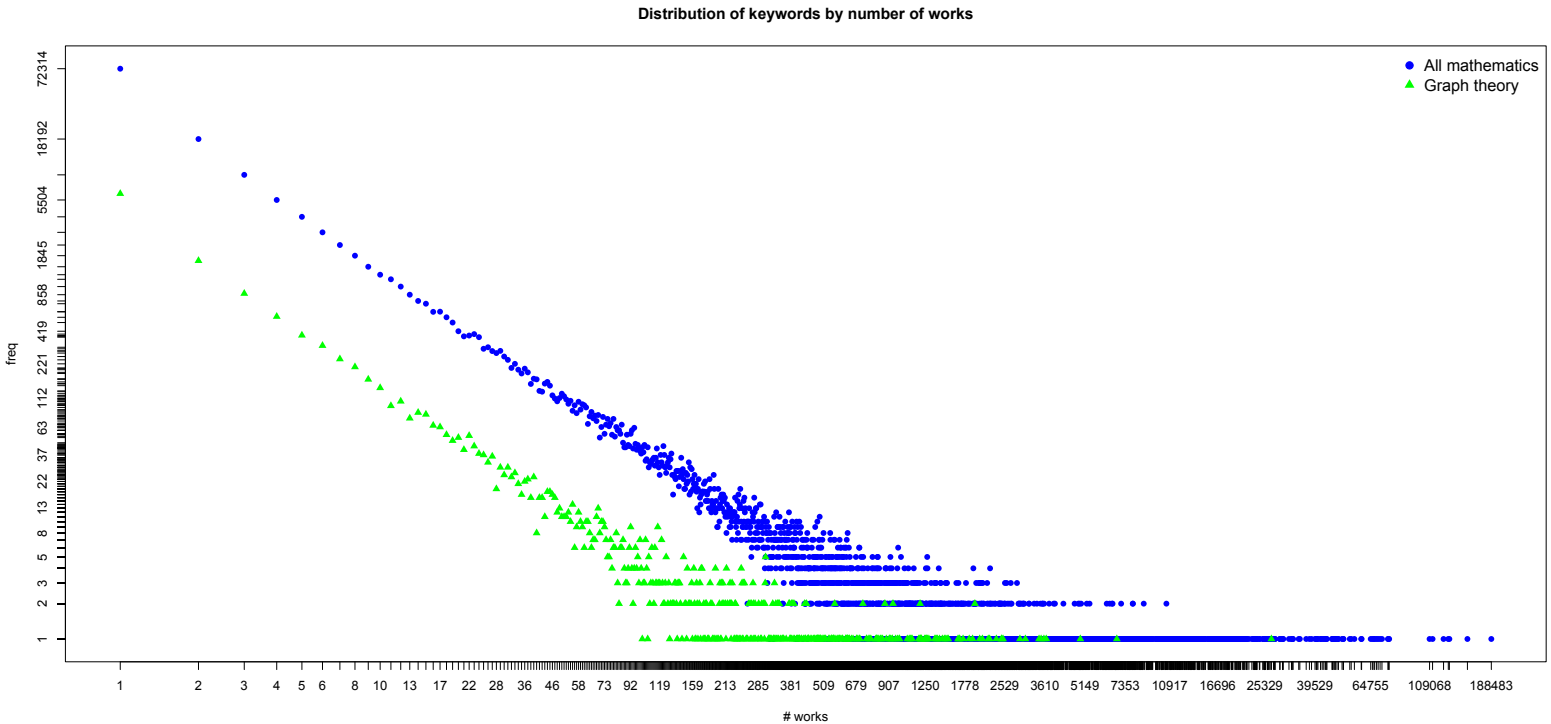
Network	**WA**	**WJ**	**WK**	**WM**
Size of the first set	1339201	1339201	1339201	1339201
Size of the second set	557104	3158	143513	12390
Number of arcs	2550437	1331036	15062377	3370820

See the paper.

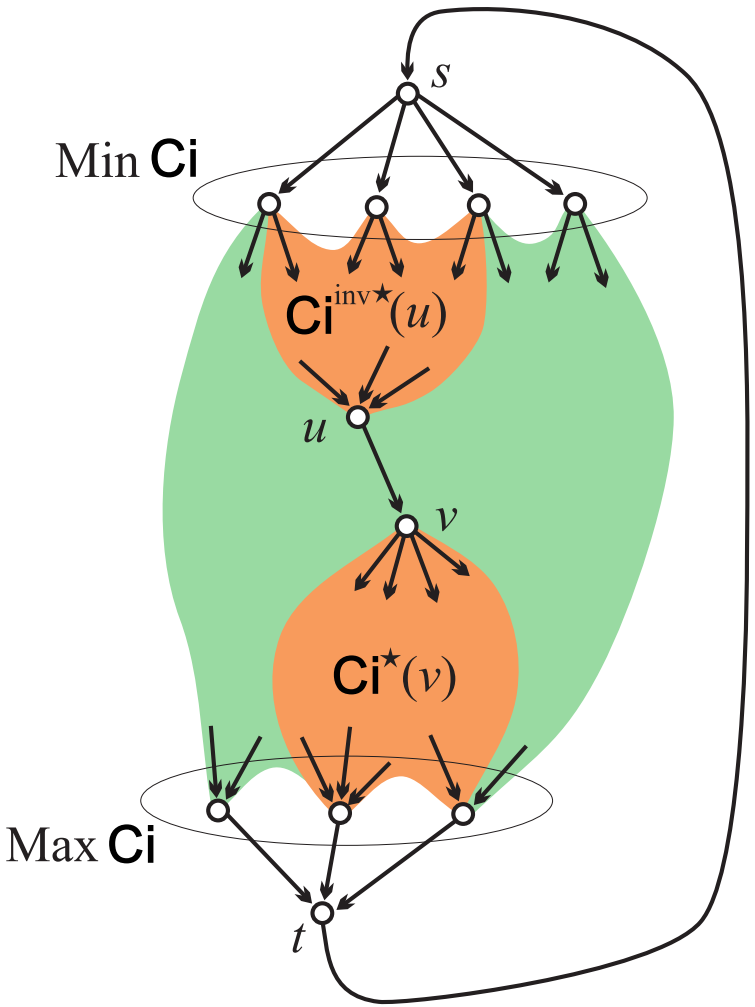
Distributions – number of keywords



Distributions – keywords by the number of works using a keyword in their description



Standardized citation network



We assume that the citation relation **Ci** is acyclic. It is useful to transform a citation network to its *standardized* form by adding a common *source* node $s \notin W$ and a common *sink* node $t \notin W$. The source s is linked by an arc to all minimal elements of **Ci**; and all maximal elements of **Ci** are linked to the sink t. We add also the 'feedback' arc (t, s).

Search path count method

Hummon and Doreian

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). To compute $n(u, v)$ we introduce two auxiliary quantities: $n^-(v)$ counts the number of different paths from s to v, and $n^+(v)$ counts the number of different paths from v to t. Then

$$n(u, v) = n^-(u) \cdot n^+(v)$$

There exists a very efficient algorithm to compute counters $n^-(v)$ and $n^+(v)$.

The quantities used to compute the arc weights n can be used also to define the corresponding *node weight* t_c

$$t_c(u) \quad = \quad n^-(u) \cdot n^+(u)$$

It is counting the number of paths through the node u.

Properties of SPC weights

The values of counters $n(u,v)$ form a flow in the citation network – the *Kirchoff's vertex law* holds: For every node u in a standardized citation network *incoming flow* = *outgoing flow*:

$$\sum_{v:v\,\mathbf{Ci}\,u} n(v,u) = \sum_{v:u\,\mathbf{Ci}\,v} n(u,v) = n^-(u)\cdot n^+(u) = t_c(u)$$

The weight $n(t,s)$ equals to the total flow through network and provides a natural normalization of weights

$$w(u,v) = \frac{n(u,v)}{n(t,s)} \quad \Rightarrow \quad 0 \le w(u,v) \le 1$$

and if C is a minimal arc-cut-set $\sum_{(u,v)\in C} w(u,v) = 1$.
In large networks the values of weights can grow very large. This should be considered in the implementation of the algorithms.

Cuts

The standard approach to find interesting groups inside a network was based on properties/weights – they can be *measured* or *computed* from network structure (for example Kleinberg's hubs and authorities).

The *node-cut* of a network $\mathcal{N} = (\mathcal{V}, \mathcal{L}, p)$, $p : \mathcal{V} \rightarrow \mathbb{R}$, at selected level t is a subnetwork $\mathcal{N}(t) = (\mathcal{V}', \mathcal{L}(\mathcal{V}'), p)$, determined by the set

$$\mathcal{V}' = \{v \in \mathcal{V} : p(v) \geq t\}$$

and $\mathcal{L}(\mathcal{V}')$ is the set of links from $\mathcal{L}$ that have both endnodes in $\mathcal{V}'$.

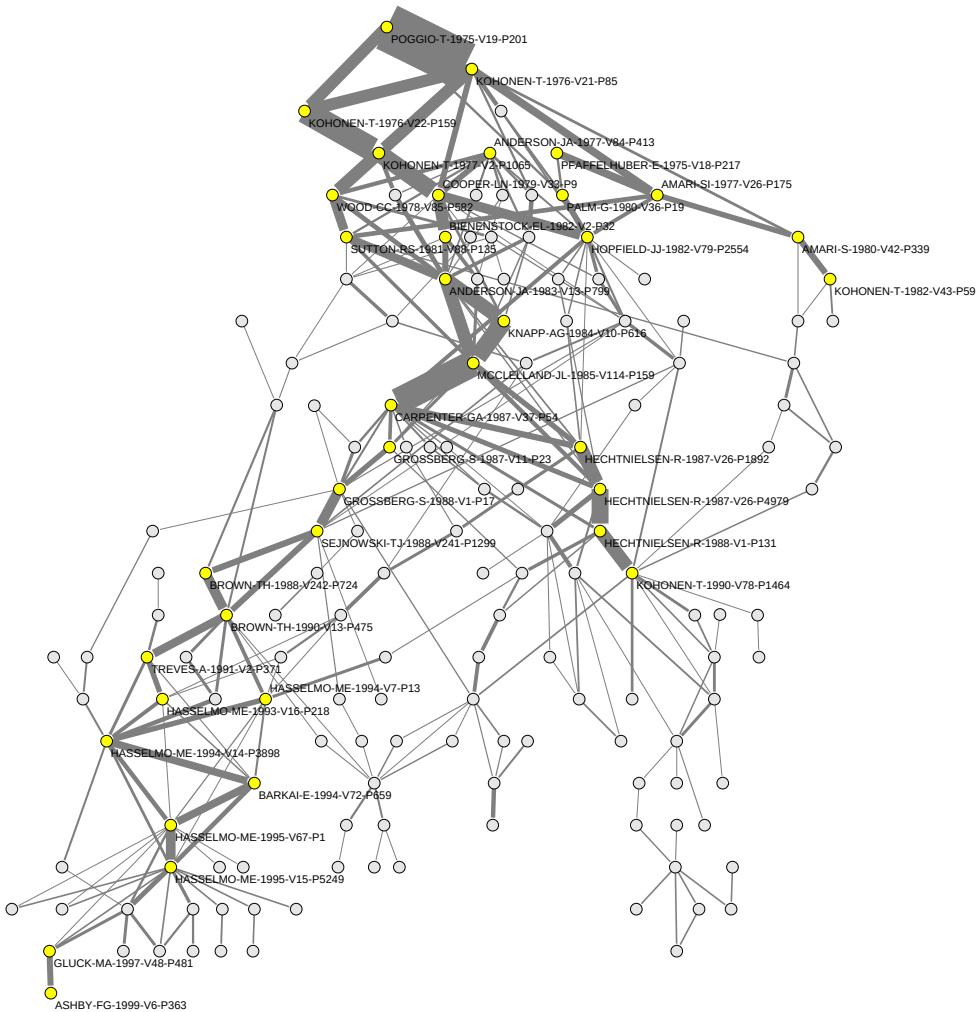
The *link-cut* of a network $\mathcal{N} = (\mathcal{V}, \mathcal{L}, w)$, $w : \mathcal{L} \rightarrow \mathbb{R}$, at selected level t is a subnetwork $\mathcal{N}(t) = (\mathcal{V}(\mathcal{L}'), \mathcal{L}', w)$, determined by the set

$$\mathcal{L}' = \{e \in \mathcal{L} : w(e) \geq t\}$$

and $\mathcal{V}(\mathcal{L}')$ is the set of all endnodes of the links from $\mathcal{L}'$.

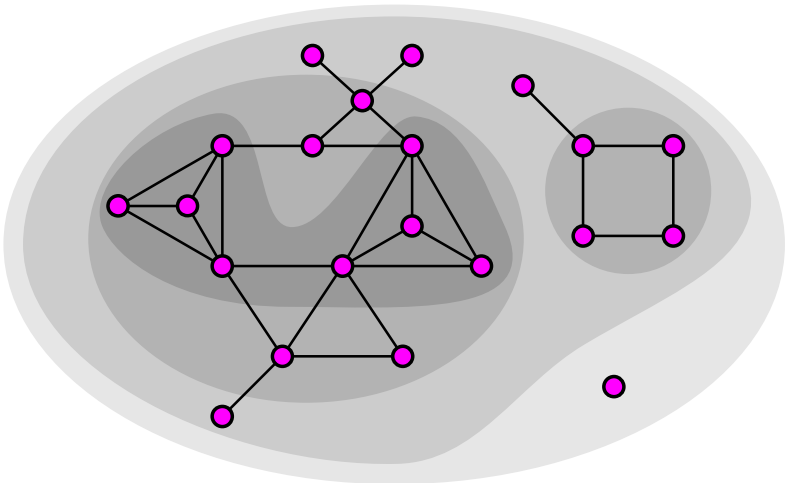
```
File/Network/read  eatRS.net
Info/Network/Line values        ... >= 70
Net/Transform/Remove/Lines with Value/lower than   70
Net/Partitions/Degree/All
Operations/Extract from Network/Partition   1-*
Net/Components/Weak
Draw/Draw-Partition
```

Citation weights



Main subnetwork (arc cut at level 0.007) of the SOM (selforganizing maps) citation network (4470 nodes, 12731 arcs). See paper.

Cores and generalized cores



The notion of core was introduced by Seidman in 1983. Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a graph. A subgraph $\mathcal{H} = (C, \mathcal{E}|C)$ induced by the set C is a *k-core* or a *core of order* k iff $\forall v \in C$: $\deg_{\mathcal{H}}(v) \geq k$, and $\mathcal{H}$ is a maximal subgraph with this property. The core of maximum order is also called the *main* core.

The *core number* of a node v is the highest order of a core that contains this node. The degree $\deg(v)$ can be: in-degree, out-degree, in-degree + out-degree, etc., determining different types of cores.

Properties of cores

From the figure, representing 0, 1, 2 and 3 core, we can see the following properties of cores:

- The cores are nested: $i < j \implies \mathcal{H}_j \subseteq \mathcal{H}_i$
- Cores are not necessarily connected subgraphs.

An efficient algorithm for determining the cores hierarchy is based on the following property:

> *If from a given graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ we recursively delete all nodes, and edges incident with them, of degree less than k, the remaining graph is the k-core.*

. . . Properties of cores

The cores, because they can be determined very efficiently, are one among few concepts that provide us with meaningful decompositions of large networks. We expect that different approaches to the analysis of large networks can be built on this basis. For example: we get the following bound on the chromatic number of a given graph $\mathcal{G}$

$$\chi(\mathcal{G}) \leq 1 + \mathrm{core}(\mathcal{G})$$

Cores can also be used to localize the search for interesting subnetworks in large networks since: if it exists, a k-component is contained in a k-core; and a k-clique is contained in a k-core.
For details see the paper.

Generalized cores

The notion of core can be generalized to networks. Let $\mathcal{N} = (\mathcal{V}, \mathcal{E}, w)$ be a network, where $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is a graph and $w : \mathcal{E} \rightarrow \mathbb{R}$ is a function assigning values to edges. A *node property function* on **N**, or a *p-function* for short, is a function $p(v, U)$, $v \in \mathcal{V}$, $U \subseteq \mathcal{V}$ with real values. Let $N_U(v) = N(v) \cap U$. Besides degrees and (corrected) clustering coefficient, here are some examples of p-functions:

$$
\begin{aligned}
p_S(v, U) &= \sum_{u \in N_U(v)} w(v, u), \text{ where } w : \mathcal{E} \rightarrow \mathbb{R}_0^+ \\
p_M(v, U) &= \max_{u \in N_U(v)} w(v, u), \text{ where } w : \mathcal{E} \rightarrow \mathbb{R} \\
p_k(v, U) &= \text{number of cycles of length } k \text{ through the node } v \text{ in } (U,
\end{aligned}
$$

The subgraph $\mathcal{H} = (C, \mathcal{E}|C)$ induced by the set $C \subseteq \mathcal{V}$ is a *p-core at level* $t \in \mathbb{R}$ iff $\forall v \in C : t \leq p(v, C)$ and C is a maximal such set.

Additional p-functions

relative density

$$p_\gamma(v,\mathcal{C}) = \frac{\deg(v,\mathcal{C})}{\max_{u\in N(v)} \deg(u)}, \text{ if } \deg(v) > 0;\ 0, \text{ otherwise}$$

diversity

$$p_\delta(v,\mathcal{C}) = \max_{u\in N^+(v,\mathcal{C})} \deg(u) - \min_{u\in N^+(v,\mathcal{C})} \deg(u)$$

average weight

$$p_a(v,\mathcal{C}) = \frac{1}{|N(v,\mathcal{C})|} \sum_{u\in N(v,\mathcal{C})} w(v,u), \text{ if } N(v,\mathcal{C}) \neq \emptyset;\ 0, \text{ otherwise}$$

Generalized cores algorithm

The function p is *monotone* iff it has the property

$$C_1 \subset C_2 \Rightarrow \forall v \in \mathcal{V} : (p(v, C_1) \leq p(v, C_2))$$

The degrees and the functions p_S, p_M and p_k are monotone. For a monotone function the p-core at level t can be determined, as in the ordinary case, by successively deleting nodes with value of p lower than t; and the cores on different levels are nested

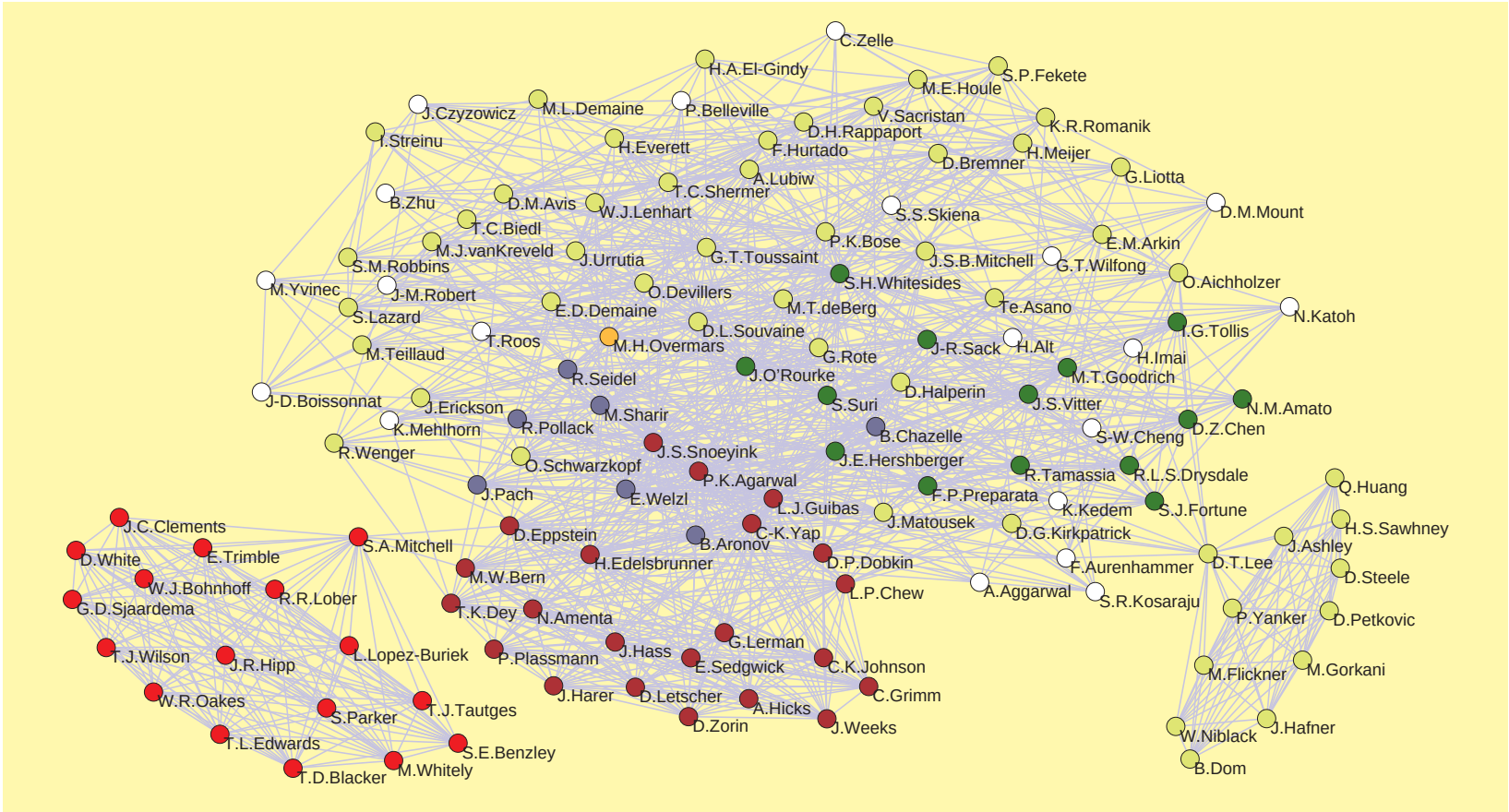
$$t_1 < t_2 \Rightarrow \mathcal{H}_{t_2} \subseteq \mathcal{H}_{t_1}$$

The p-function is *local* iff $p(v, U) = p(v, N_U(v))$.
The degrees, p_S and p_M are local; but p_k is **not** local for $k \geq 4$. For a local p-function an $O(m \max(\Delta, \log n))$ algorithm for determining the p-core levels exists, assuming that $p(v, N_C(v))$ can be computed in $O(\deg_C(v))$.
For details see the paper.

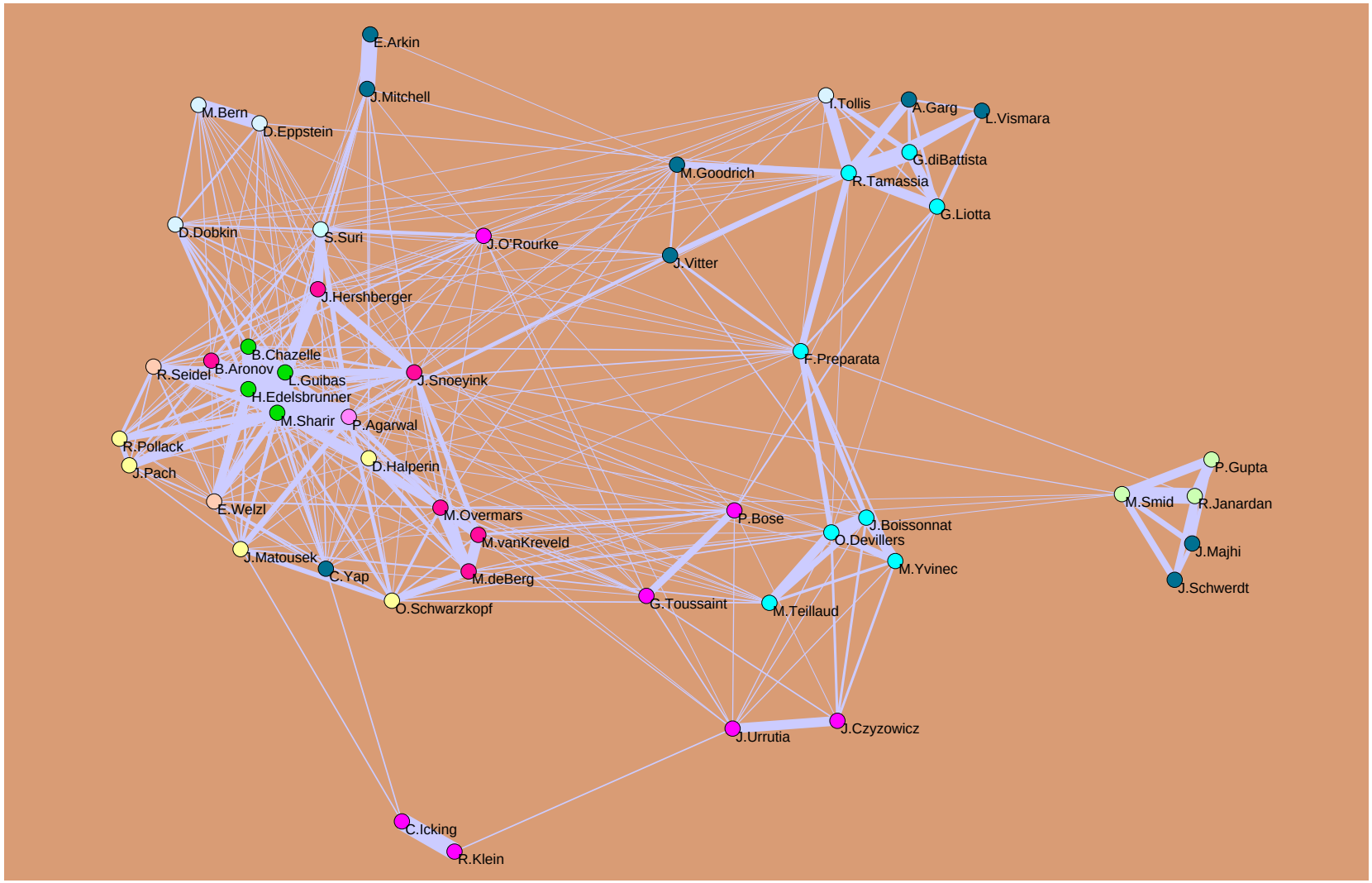
Cores and generalized cores / Pajek commands

```
File/Network/Read  [Geom.net]
Net/Partitions/Core/All
Info/Partition
Operations/Extract from Network/Partition [13-*]
Draw/Draw-Partition
Layout/Energy/Kamada-Kawai
Options/Values of lines/Similarities
Layout/Energy/Kamada-Kawai
Operations/Extract from Network/Partition [21]
Draw
Layout/Energy/Kamada-Kawai
Options/Values of lines/Forget
Layout/Energy/Kamada-Kawai
[select Geom.net]
Net/Vector/PCore/Sum/All
Info/Vector
Vector/Make Partition/by Intervals/Selected Thresholds [45]
Info/Partition
Operations/Extract from Network/Partition [2]
Draw
Options/Values of lines/Similarities
Layout/Energy/Fruchterman-Reingold
```

Cores of orders 10–21 in Computational Geometry

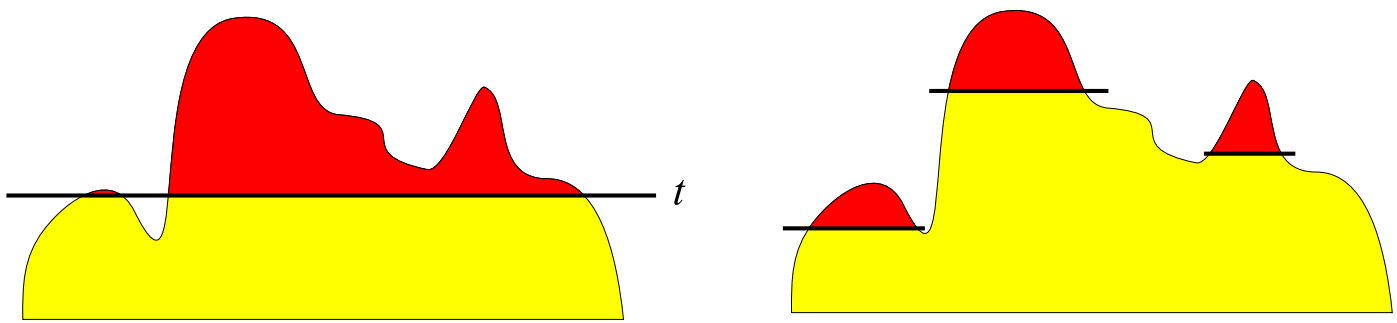


p_S-core at level 46 in Computational Geometry network



Islands

If we represent a given or computed value of nodes / links as a height of nodes / links and we immerse the network into a water up to selected level we get *islands*. Varying the level we get different islands.



We developed very efficient algorithms to determine the islands hierarchy and to list all the islands of selected sizes.
See details.

. . . Islands

Islands are very general and efficient approach to determine the 'important' subnetworks in a given network.
We have to express the goals of our analysis with a related property of the nodes or weight of the links. Using this property we determine the islands of an appropriate size (in the interval k to K).

In large networks we can get many islands which we have to inspect individually and interpret their content.
An important property of the islands is that they identify locally important subnetworks on different levels. Therefore they detect also emerging groups.

The set of nodes $\mathcal{C} \subseteq \mathcal{V}$ is a *local node peak*, if it is a regular node island and all of its nodes have the same value. Node island with a single local node peak is called a *simple node island*. In similar way we define simple link island.

. . . Islands

A set of nodes $C \subseteq \mathcal{V}$ is a *regular node island* in a network $\mathcal{N} = (\mathcal{V}, \mathcal{L}, p)$, $p : \mathcal{V} \to \mathbb{R}$ iff it induces a connected subgraph and the nodes from the island are 'higher' than the neighboring nodes

$$\max_{u \in N(C)} p(u) < \min_{v \in C} p(v)$$

A set of nodes $C \subseteq \mathcal{V}$ is a *regular link island* in a network $\mathcal{N} = (\mathcal{V}, \mathcal{L}, w)$, $w : \mathcal{L} \to \mathbb{R}$ iff it induces a connected subgraph and the links inside the island are 'stronger related' among them than with the neighboring nodes – in $\mathcal{N}$ there exists a spanning tree $\mathcal{T}$ over C such that

$$\max_{(u,v) \in \mathcal{L}, u \notin C, v \in C} w(u, v) < \min_{(u,v) \in \mathcal{T}} w(u, v)$$

`Network/Create Partition/Islands/Line Weights`
`Operations/Network+Vector/Islands/Vertex Property`

Some properties of islands

- The sets of nodes of connected components of node/link-cut at selected level t are regular node/link islands.
- The set $\mathcal{H}_p(\mathcal{N})$ of all regular node islands of a network $\mathcal{N}$ is a complete hierarchy:
  - two islands are disjoint or one of them is a subset of the other
  - each node belongs to at least one island
- The set $\mathcal{H}_w(\mathcal{N})$ of all nondegenerated regular link islands of a network $\mathcal{N}$ is a hierarchy (not necessarily complete):
  - two islands are disjoint or one of them is a subset of the other
- Node/link islands are invariant for the strictly increasing transformations of the property p / weight w.
- Two linked nodes cannot/may belong to two disjoint regular node/link islands.

A *simple island* is an island with only one peak.

US patents

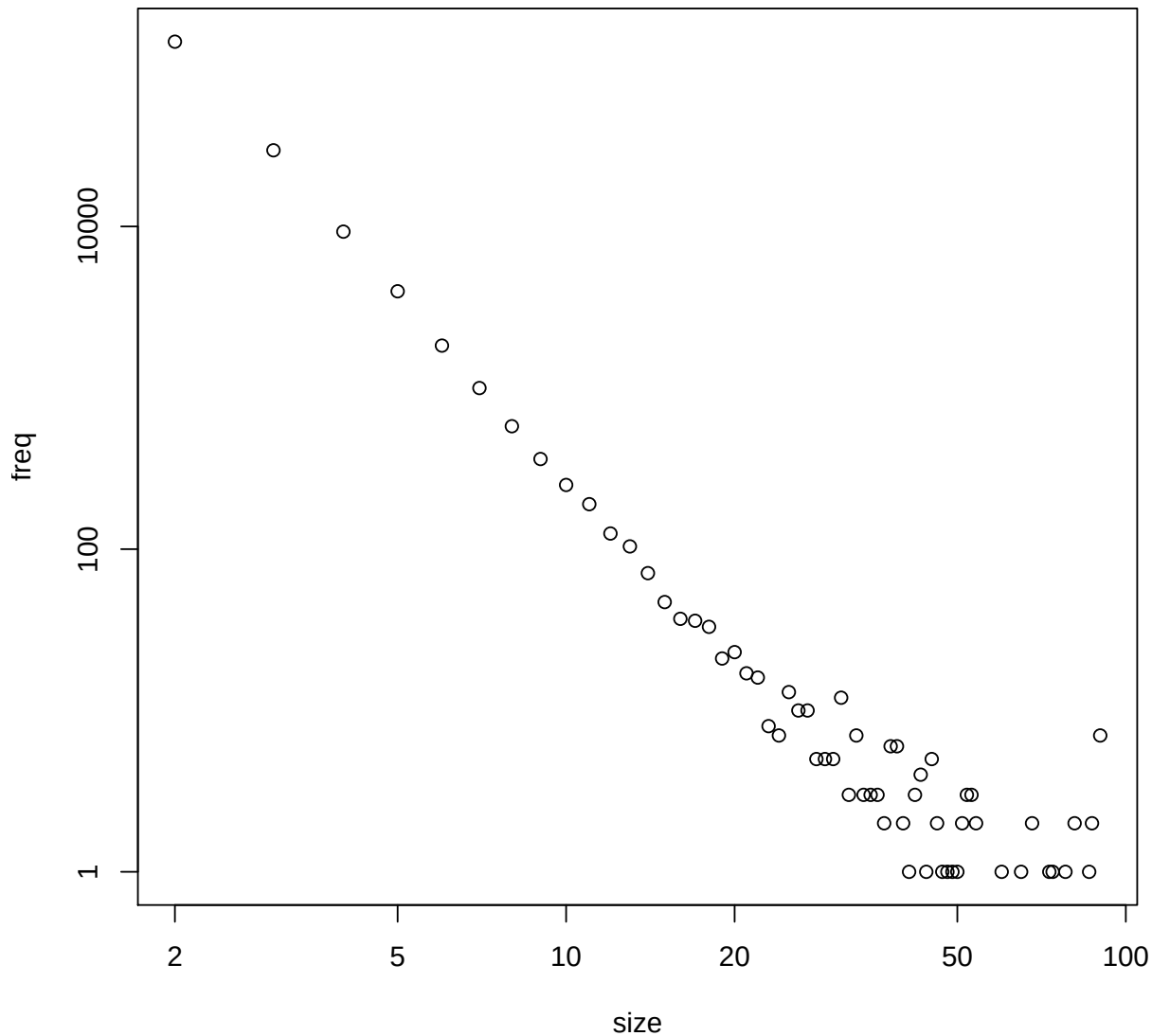
US patents network (Nber, US Patents) has 3774768 nodes and16522438 arcs (1 loop). Without the loop it is acyclic. The weight of an arc is the proportion of paths through the arc from some initial node to some terminal node. We determined al (2,90)-link islands. The corresponding subnetwork has 470137 nodes, 307472 arcs and different k: C_2 =187610, C_5 =8859, C_{30} =101, C_{50} =30, ... islands. Rolex

```
 [1]   0 139793  29670 9288 3966 1827 997 578 362 250
[11] 190    125    104   71   47   37  36  33  21  23
[21]  17     16      8    7   13   10  10   5   5   5
[31]  12      3      7    3    3    3   2   6   6   2
[41]   1      3      4    1    5    2   1   1   1   1
[51]   2      3      3    2    0    0   0   0   0   1
[61]   0      0      0    0    1    0   0   2   0   0
[71]   0      0      1    1    0    0   0   1   0   0
[81]   2      0      0    0    0    1   2   0   0   7
```

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.

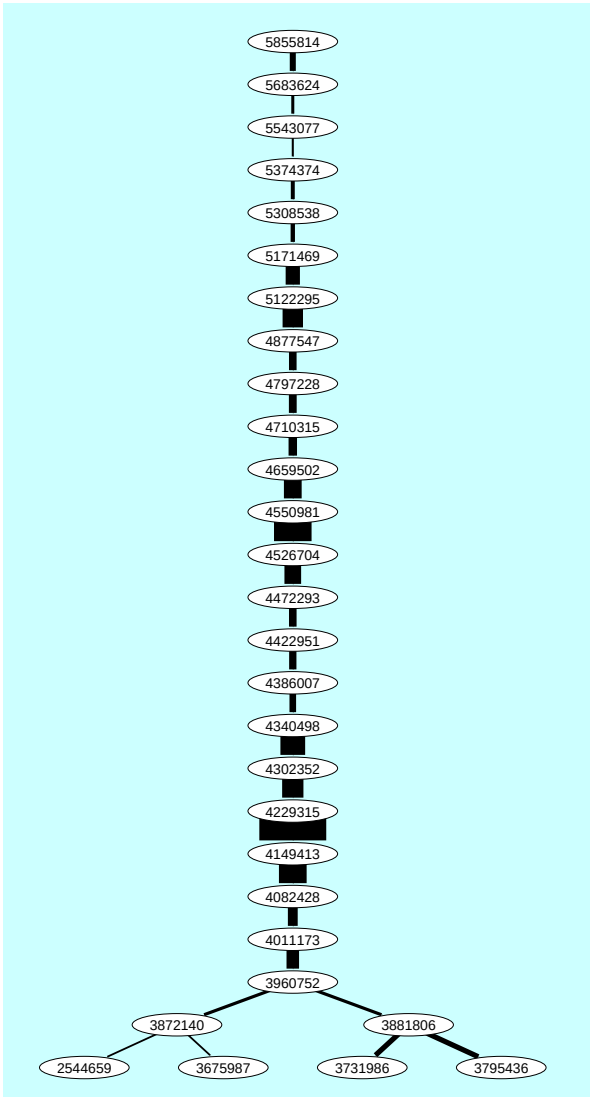
The *CPM path* is determined using the Critical Path Method from Operations Research (the sum of SPC weights along a path is maximal).

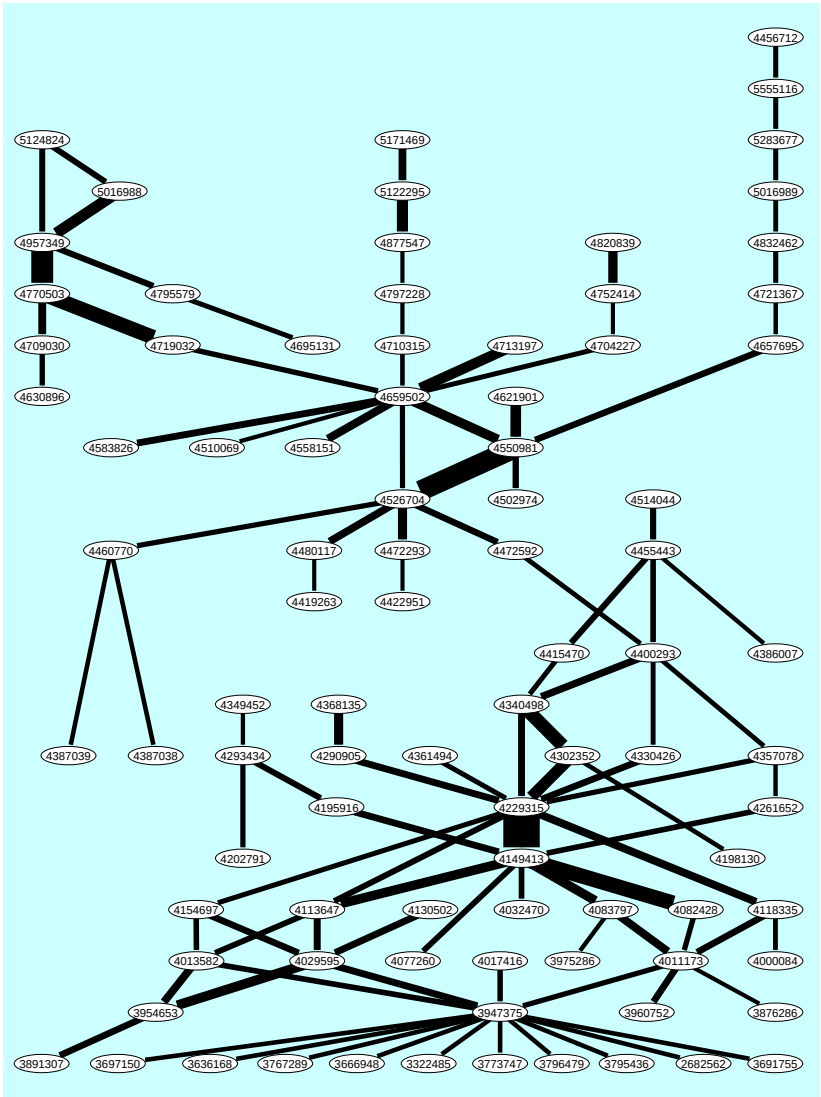
Distribution of island size



Main path and main island in US Patents

Nber, US Patents; $n = 3774768$, $m = 16522438$





Main island – Liquid crystal display

Table 1: Patents on the liquid-crystal display

patent	date	author(s) and title
2544659	Mar 13, 1951	Dreyer. Dichroic light-polarizing sheet and the like and the formation and use thereof
2682562	Jun 29, 1954	Wender, et al. Reduction of aromatic carbinols
3322485	May 30, 1967	Williams. Electro-optical elements utilazing an organic nematic compound
3636168	Jan 18, 1972	Josephson. Preparation of polynuclear aromatic compounds
3666948	May 30, 1972	Mechlowitz, et al. Liquid crystal termal imaging system having an undisturbed image on a disturbed background
3675987	Jul 11, 1972	Rafuse. Liquid crystal compositions and devices
3691755	Sep 19, 1972	Girard. Clock with digital display
3697150	Oct 10, 1972	Wysochi. Electro-optic systems in which an electrophoretic-like or dipolar material is dispersed throughout a liquid crystal to reduce the turn-off time
3731986	May 8, 1973	Fergason. Display devices utilizing liquid crystal light modulation
3767289	Oct 23, 1973	Aviram, et al. Class of stable trans-stilbene compounds, some displaying nematic mesophases at or near room temperature and others in a range up to 100°C
3773747	Nov 20, 1973	Steinstrasser. Substituted azoxy benzene compounds
3795436	Mar 5, 1974	Boller, et al. Nematogenic material which exhibit the Kerr effect at isotropic temperatures
3796479	Mar 12, 1974	Helfrich, et al. Electro-optical light-modulation cell utilizing a nematogenic material which exhibits the Kerr effect at isotropic temperatures
3872140	Mar 18, 1975	Klanderman, et al. Liquid crystalline compositions and method
3876286	Apr 8, 1975	Deutscher, et al. Use of nematic liquid crystalline substances
3881806	May 6, 1975	Suzuki. Electro-optical display device
3891307	Jun 24, 1975	Tsukamoto, et al. Phase control of the voltages applied to opposite electrodes for a cholesteric to nematic phase transition display
3947375	Mar 30, 1976	Gray, et al. Liquid crystal materials and devices
3954653	May 4, 1976	Yamazaki. Liquid crystal composition having high dielectric anisotropy and display device incorporating same
3960752	Jun 1, 1976	Klanderman, et al. Liquid crystal compositions
3975286	Aug 17, 1976	Oh. Low voltage actuated field effect liquid crystals compositions and method of synthesis
4000084	Dec 28, 1976	Hsieh, et al. Liquid crystal mixtures for electro-optical display devices
4011173	Mar 8, 1977	Steinstrasser. Modified nematic mixtures with positive dielectric anisotropy
4013582	Mar 22, 1977	Gavrilovic. Liquid crystal compounds and electro-optic devices incorporating them
4017416	Apr 12, 1977	Inukai, et al. P-cyanophenyl 4-alkyl-4'-biphenylcarboxylate, method for preparing same and liquid crystal compositions using same
4029595	Jun 14, 1977	Ross, et al. Novel liquid crystal compounds and electro-optic devices incorporating them
4032470	Jun 28, 1977	Bloom, et al. Electro-optic device
4077260	Mar 7, 1978	Gray, et al. Optically active cyano-biphenyl compounds and liquid crystal materials containing them
4082428	Apr 4, 1978	Hsu. Liquid crystal composition and method

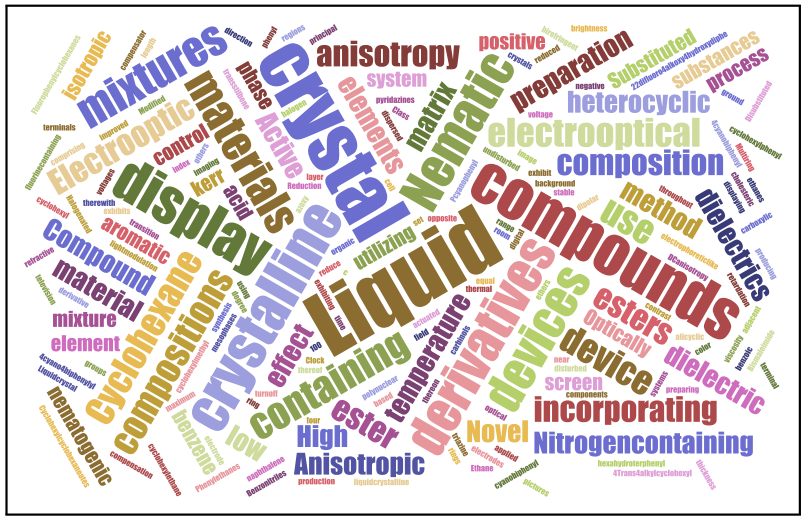
Table 2: Patents on the liquid-crystal display

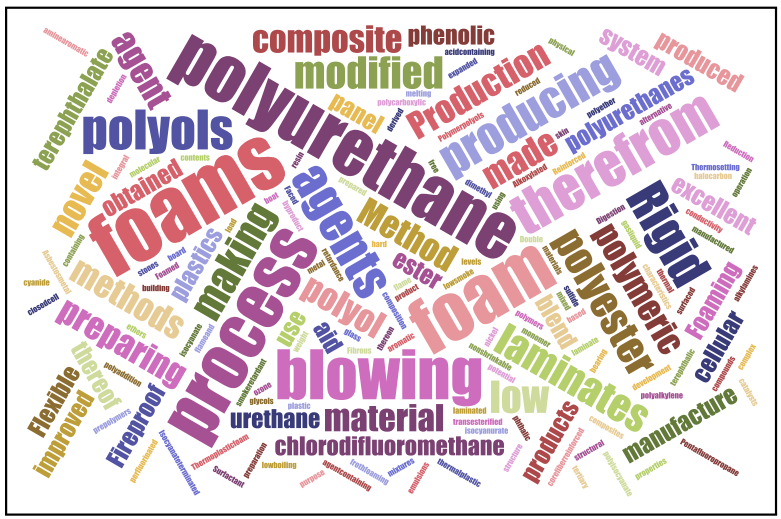
patent	date	author(s) and title
4083797	Apr 11, 1978	Oh. Nematic liquid crystal compositions
4113647	Sep 12, 1978	Coates, et al. Liquid crystalline materials
4118335	Oct 3, 1978	Krause, et al. Liquid crystalline materials of reduced viscosity
4130502	Dec 19, 1978	Eidenschink, et al. Liquid crystalline cyclohexane derivatives
4149413	Apr 17, 1979	Gray, et al. Optically active liquid crystal mixtures and liquid crystal devices containing them
4154697	May 15, 1979	Eidenschink, et al. Liquid crystalline hexahydroterphenyl derivatives
4195916	Apr 1, 1980	Coates, et al. Liquid crystal compounds
4198130	Apr 15, 1980	Boller, et al. Liquid crystal mixtures
4202791	May 13, 1980	Sato, et al. Nematic liquid crystalline materials
4229315	Oct 21, 1980	Krause, et al. Liquid crystalline cyclohexane derivatives
4261652	Apr 14, 1981	Gray, et al. Liquid crystal compounds and materials and devices containing them
4290905	Sep 22, 1981	Kanbe. Ester compound
4293434	Oct 6, 1981	Deutscher, et al. Liquid crystal compounds
4302352	Nov 24, 1981	Eidenschink, et al. Fluorophenylcyclohexanes, the preparation thereof and their use as components of liquid crystal dielectrics
4330426	May 18, 1982	Eidenschink, et al. Cyclohexylbiphenyls, their preparation and use in dielectrics and electrooptical display elements
4340498	Jul 20, 1982	Sugimori. Halogenated ester derivatives
4349452	Sep 14, 1982	Osman, et al. Cyclohexylcyclohexanoates
4357078	Nov 2, 1982	Carr, et al. Liquid crystal compounds containing an alicyclic ring and exhibiting a low dielectric anisotropy and liquid crystal materials and devices incorporating such compounds
4361494	Nov 30, 1982	Osman, et al. Anisotropic cyclohexyl cyclohexylmethyl ethers
4368135	Jan 11, 1983	Osman. Anisotropic compounds with negative or positive DC-anisotropy and low optical anisotropy
4386007	May 31, 1983	Krause, et al. Liquid crystalline naphthalene derivatives
4387038	Jun 7, 1983	Fukui, et al. 4-(Trans-4'-alkylcyclohexyl) benzoic acid 4'''-cyano-4''-biphenylyl esters
4387039	Jun 7, 1983	Sugimori, et al. Trans-4-(trans-4'-alkylcyclohexyl)-cyclohexane carboxylic acid 4'''-cyanobiphenyl ester
4400293	Aug 23, 1983	Romer, et al. Liquid crystalline cyclohexylphenyl derivatives
4415470	Nov 15, 1983	Eidenschink, et al. Liquid crystalline fluorine-containing cyclohexylbiphenyls and dielectrics and electro-optical display elements based thereon
4419263	Dec 6, 1983	Praefcke, et al. Liquid crystalline cyclohexylcarbonitrile derivatives
4422951	Dec 27, 1983	Sugimori, et al. Liquid crystal benzene derivatives
4455443	Jun 19, 1984	Takatsu, et al. Nematic halogen Compound
4456712	Jun 26, 1984	Christie, et al. Bismaleimide triazine composition
4460770	Jul 17, 1984	Petrzilka, et al. Liquid crystal mixture
4472293	Sep 18, 1984	Sugimori, et al. High temperature liquid crystal substances of four rings and liquid crystal compositions containing the same
4472592	Sep 18, 1984	Takatsu, et al. Nematic liquid crystalline compounds
4480117	Oct 30, 1984	Takatsu, et al. Nematic liquid crystalline compounds
4502974	Mar 5, 1985	Sugimori, et al. High temperature liquid-crystalline ester compounds
4510069	Apr 9, 1985	Eidenschink, et al. Cyclohexane derivatives

Table 3: Patents on the liquid-crystal display

patent	date	author(s) and title
4514044	Apr 30, 1985	Gunjima, et al. 1-(Trans-4-alkylcyclohexyl)-2-(trans-4'-(p-substituted phenyl) cyclohexyl)ethane and liquid crystal mixture
4526704	Jul 2, 1985	Petrzilka, et al. Multiring liquid crystal esters
4550981	Nov 5, 1985	Petrzilka, et al. Liquid crystalline esters and mixtures
4558151	Dec 10, 1985	Takatsu, et al. Nematic liquid crystalline compounds
4583826	Apr 22, 1986	Petrzilka, et al. Phenylethanes
4621901	Nov 11, 1986	Petrzilka, et al. Novel liquid crystal mixtures
4630896	Dec 23, 1986	Petrzilka, et al. Benzonitriles
4657695	Apr 14, 1987	Saito, et al. Substituted pyridazines
4659502	Apr 21, 1987	Fearon, et al. Ethane derivatives
4695131	Sep 22, 1987	Balkwill, et al. Disubstituted ethanes and their use in liquid crystal materials and devices
4704227	Nov 3, 1987	Krause, et al. Liquid crystal compounds
4709030	Nov 24, 1987	Petrzilka, et al. Novel liquid crystal mixtures
4710315	Dec 1, 1987	Schad, et al. Anisotropic compounds and liquid crystal mixtures therewith
4713197	Dec 15, 1987	Eidenschink, et al. Nitrogen-containing heterocyclic compounds
4719032	Jan 12, 1988	Wachtler, et al. Cyclohexane derivatives
4721367	Jan 26, 1988	Yoshinaga, et al. Liquid crystal device
4752414	Jun 21, 1988	Eidenschink, et al. Nitrogen-containing heterocyclic compounds
4770503	Sep 13, 1988	Buchecker, et al. Liquid crystalline compounds
4795579	Jan 3, 1989	Vauchier, et al. 2,2'-difluoro-4-alkoxy-4'-hydroxydiphenyls and their derivatives, their production process and their use in liquid crystal display devices
4797228	Jan 10, 1989	Goto, et al. Cyclohexane derivative and liquid crystal composition containing same
4820839	Apr 11, 1989	Krause, et al. Nitrogen-containing heterocyclic esters
4832462	May 23, 1989	Clark, et al. Liquid crystal devices
4877547	Oct 31, 1989	Weber, et al. Liquid crystal display element
4957349	Sep 18, 1990	Clerc, et al. Active matrix screen for the color display of television pictures, control system and process for producing said screen
5016988	May 21, 1991	Iimura. Liquid crystal display device with a birefringent compensator
5016989	May 21, 1991	Okada. Liquid crystal element with improved contrast and brightness
5122295	Jun 16, 1992	Weber, et al. Matrix liquid crystal display
5124824	Jun 23, 1992	Kozaki, et al. Liquid crystal display device comprising a retardation compensation layer having a maximum principal refractive index in the thickness direction
5171469	Dec 15, 1992	Hittich, et al. Liquid-crystal matrix display
5283677	Feb 1, 1994	Sagawa, et al. Liquid crystal display with ground regions between terminal groups
5308538	May 3, 1994	Weber, et al. Supertwist liquid-crystal display
5374374	Dec 20, 1994	Weber, et al. Supertwist liquid-crystal display
5543077	Aug 6, 1996	Rieger, et al. Nematic liquid-crystal composition
5555116	Sep 10, 1996	Ishikawa, et al. Liquid crystal display having adjacent electrode terminals set equal in length
5683624	Nov 4, 1997	Sekiguchi, et al. Liquid crystal composition
5855814	Jan 5, 1999	Matsui, et al. Liquid crystal compositions and liquid crystal display elements

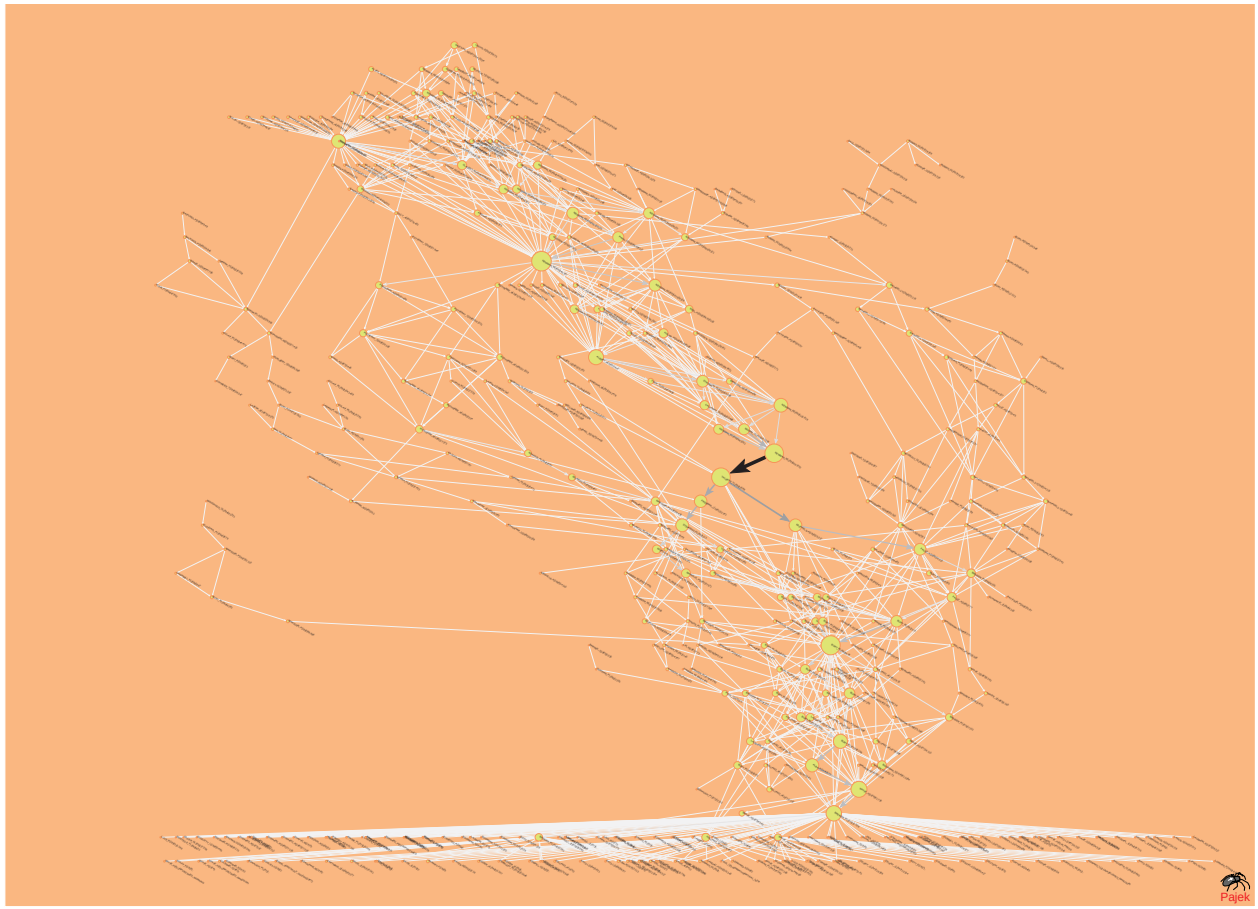
Word clouds for LCD island and foam island



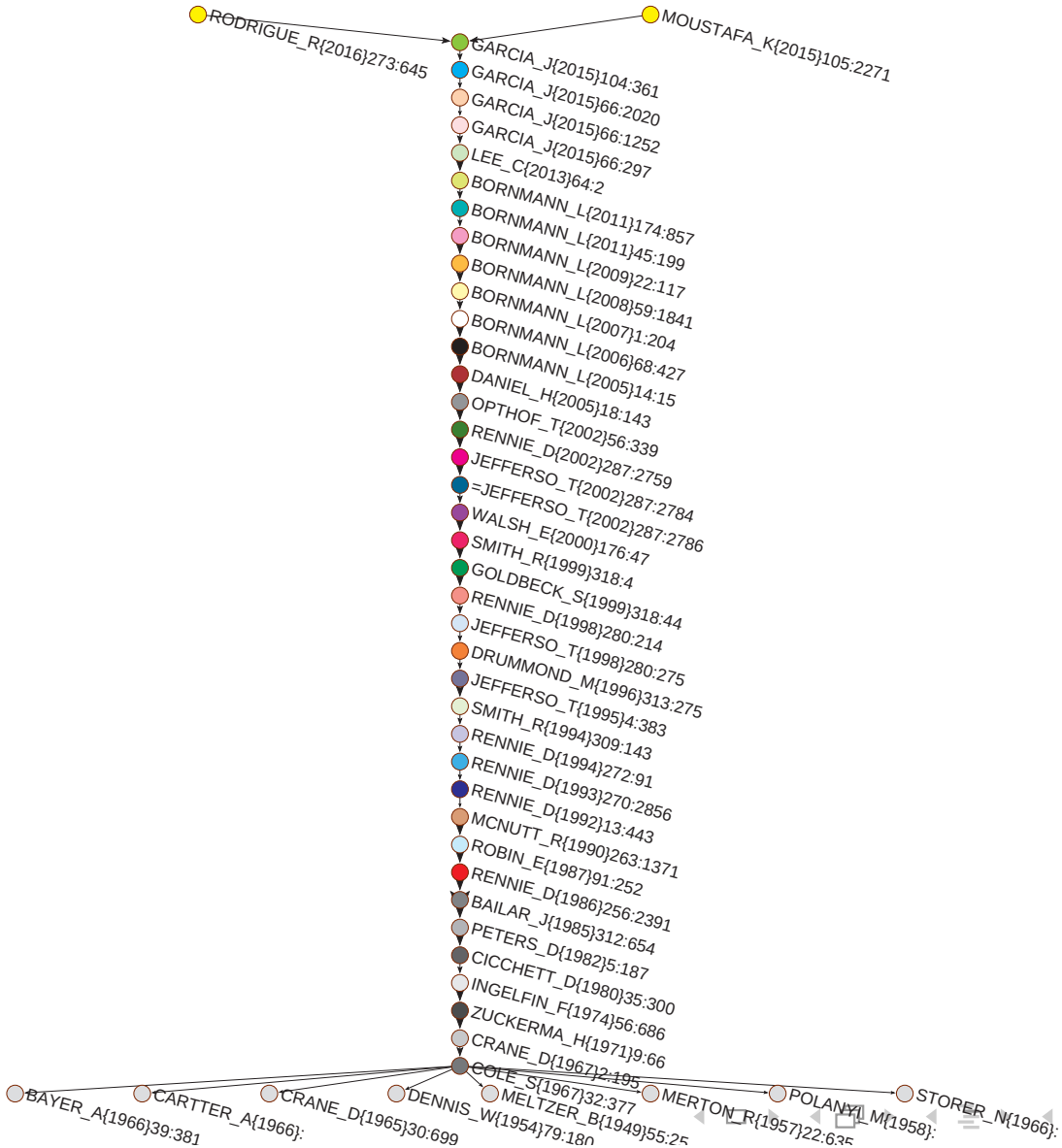


Main SPC island in SN5

Network SN5 (2008): for "social network*" + most frequent references + around 100 social networkers; $|W| = 193376$, $|C| = 7950$, $|A| = 75930$, $|J| = 14651$, $|K| = 29267$. Citation networks are acyclic. Acyclic networks can be displayed by levels – run macro Layers1.mcr from the map Geneo.



PEERE – Main path



List of publications on main path

year	first author	title	journal
1949	Meltzer BN	The productivity of social scientists	AmJSociol
1954	Dennis W	Bibliographies of eminent scientists	ScientificM
1957	Merton RK	Priorities in scientific discovery - a chapter in the sociology of sci...	AmSociolRev
1958	Polanyi M	Personal Knowledge: Towards a Post-Critical Philosophy	UPChicago
1965	Crane D	Scientists at major and minor universities	AmSociolRev
1966	Bayer AE	Some correlates of citation measure of productivity in science	SociolEduc
1966	Storer NW	The Social System of Science	HRW
1966	Cartter A	An Assessment of Quality in Graduate Education	ACE
1967	Crane D	Gatekeepers of science - some factors affecting selection...	AmSociol
1967	Cole S	Scientific output and recognition - study in operation of reward...	AmSociolRev
1971	Zuckerman H	Patterns of evaluation in science - institutionalisation, struct...	Minerva
1974	Ingelfinger FJ	Peer review in biomedical publication	AmJMed
1980	Cicchetti DV	Reliability of reviews for the american-psychologist	AmPsychol
1982	Peters DP	Peer-review practices of psychological journals - the fate...	BehavBrainS
1985	Bailar JC	Journal peer-review - the need for a research agenda	NewEnglJMe
1986	Rennie D	Guarding the guardians - a conference on editorial peer-review	Jama
1987	Robin ED	Peer-review in medical journals	Chest
1990	Mcnutt RA	The effects of blinding on the quality of peer-review	Jama
1992	Rennie D	Suspended judgment - editorial peer-review - let us put it on trial	ControlClinT
1993	Rennie D	More peering into editorial peer-review	Jama
1994	Rennie D	The 2nd international-congress on peer-review in biomedical...	Jama
1994	Smith R	Promoting research into peer-review	BritMedJ
1995	Jefferson T	Are guidelines for peer-reviewing economic evaluations necessary	HealthEcon
1996	Drummond M	Guidelines for authors and peer reviewers of economic submis...	BritMedJ
1998	Jefferson T	Evaluating the BMJ guidelines for economic submissions	Jama
1998	Rennie D	Peer review in Prague	Jama

List of works on main path cont.

year	first author	title	journal
1999	Smith R	Opening up BMJ peer review - A beginning that should lead to...	BritMedJ
1999	Goldbeck-W. S	Evidence on peer review - scientific quality control or smokescreen?	BritMedJ
2000	Walsh E	Open peer review: a randomised controlled trial	BritJPsych
2002	Jefferson T	Effects of editorial peer review - A systematic review	Jama
2002	Rennie D	Fourth International Congress on Peer Review in Biomedical Pub...	Jama
2002	Opthof T	The significance of the peer review process against . . . bias	Cardiovasc
2002	Jefferson T	Measuring the quality of editorial peer review	Jama
2005	Bornmann L	Committee peer review at an international research foundation	ResEvaluat
2005	Daniel HD	Publications as a measure of scientific advancement and of...	LearnPubl
2006	Bornmann L	Selecting scientific excellence through committee peer review	Scientomet
2007	Bornmann L	Convergent validation of peer review decisions using the h index	JInformetr
2008	Bornmann L	Selecting manuscripts for a high-impact journal through peer review	JAmSocInf
2009	Bornmann L	The luck of the referee draw: the effect of exchanging reviews	LearnPubl
2011	Bornmann L	Scientific Peer Review	AnnuRevIn
2011	Bornmann L	A multilevel modelling approach to investigating the predictive...	JRStatSoc
2013	Lee CJ	Bias in peer review	JAmSocInf
2015	Garcia JA	The Principal-Agent Problem in Peer Review	JAssocInfS
2015	Garcia JA	Adverse selection of reviewers	JAssocInfS
2015	Garcia JA	Bias and effort in peer review	JAssocInfS
2015	Garcia JA	The author-editor game	Scientomet
2015	Moustafa K	Don't infer anything from unavailable data	Scientomet
2016	Rodriguez-S. R	Evolutionary games between authors and their editors	ApplMathC

The main path publications

Phases

- before 1982: social science journals;
- from 1982 to 2002: biomedical journals;
- after 2002: specialized journals on science studies.

The main path publications

till 1982

Journals: social science journals (sociological, psychological, educational,...) and three books.

The most **influential authors:** Meltzer (1949), Dennis (1954), Merton (1957), Polany (1958), Crane (1965, 1967), Bayer and Folger (1966), Storer (1966), Cartter (1966), Cole and Cole (1967), Zuckerman and Merton (1971), Ingelfinger (1974), Cicchetti (1980) and Peters and Ceci (1982).

Topics: scientific productivity, bibliographies, knowledge, citation measures as measures of scientific accomplishement, scientific output and recognition, evaluation in science, referee system, journal evaluation, peer-evaluation system, review process, peer review practices.

The main path publications

from 1983 to 2002

Journals: biomedical journals, mainly JAMA. From 1986 the International Congress on Peer Review and Biomedical Publication is organized every four years.

The most **influential authors:** Rennie (1986, 1992, 1993, 1994, 1998, 2002), Smith (1994, 1999), Jefferson (1995, 1998, 2002), and their collaborators.

Topics: the effects of blinding on review quality, research into peer review, guidelines for peer reviewing, monitoring the peer review performance, open peer review, bias in peer review system, measuring the quality of editorial peer review. Development of meta-analysis and systematic reviews studies of peer-review.

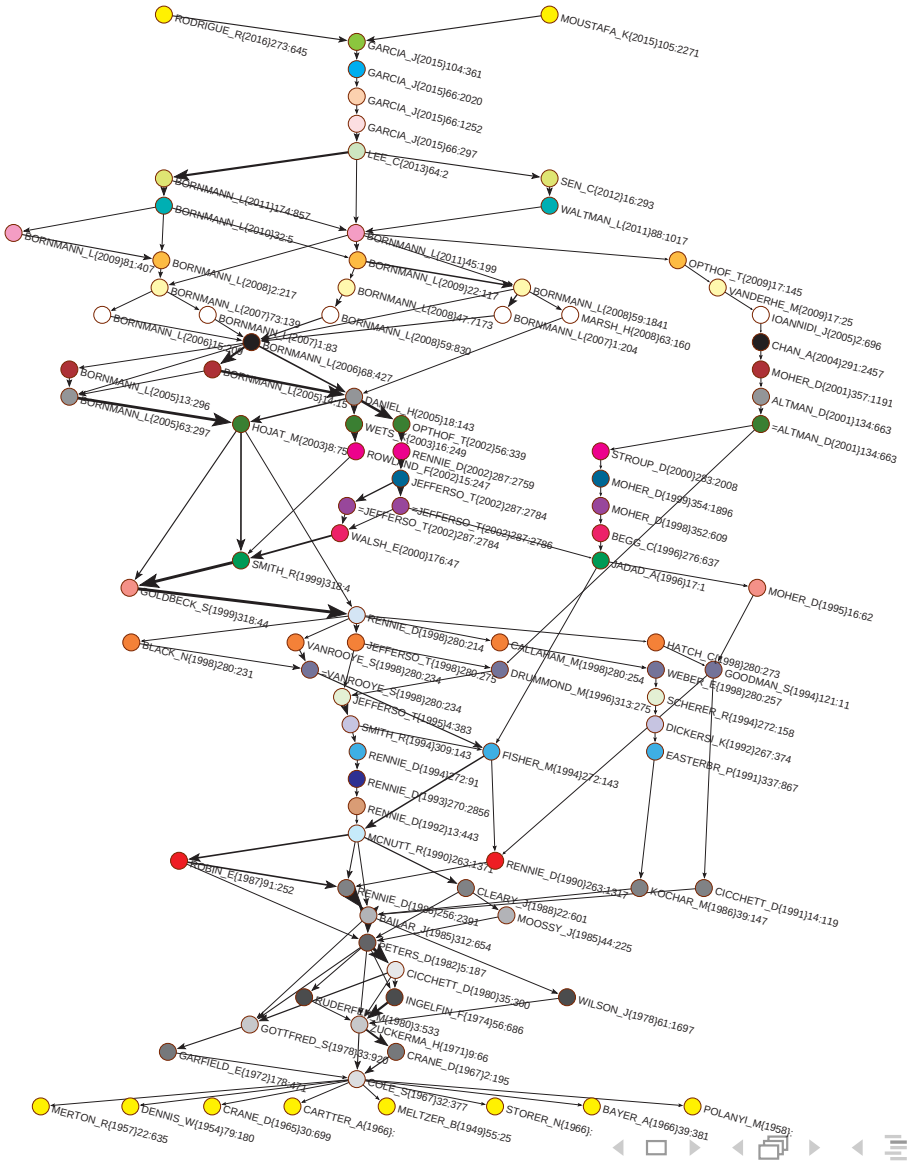
The main path publications
from 2003 on

Journals: specialized journals on science studies: Scientometrics, Research Evaluation, Journal of Informetrics, JASIST.

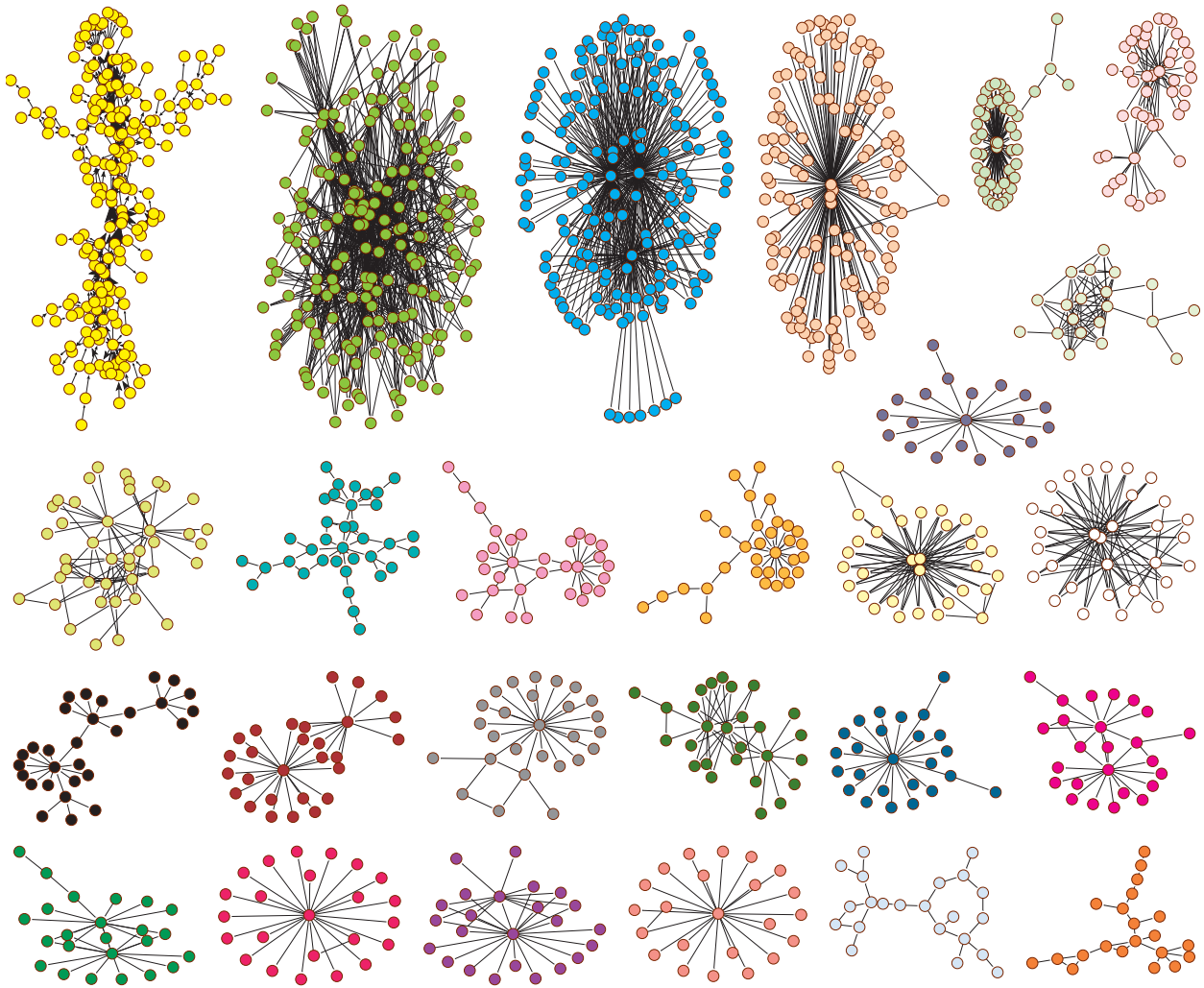
The most **influential authors:** Bornmann and Daniel (2005, 2006, 2007, 2008, 2009, 2011) and Garcia, Rodriguez-Sanchez and Fdez-Valdivia (4 papers in 2015, 2016). Others are Lee et al. (2013) and Moustafa (2015).

Topics: Bornmann and Daniel studied the validity of commitee peer review process for awarding long-term fellowship to post-graduate researchers, the use of h-index and pre-screening of applications at Boehringer Ingelheim Fonds. They also analysed citations of accepted and rejected papers at a prime chemistry journal, the effect of exchanging reviews, the peer review process in this journal, the validity of its editorial decisions. The other papers studied bias in peer review, selection of reviewers, and the author-editor game.

PEERE – Main paths for 100 largest weights

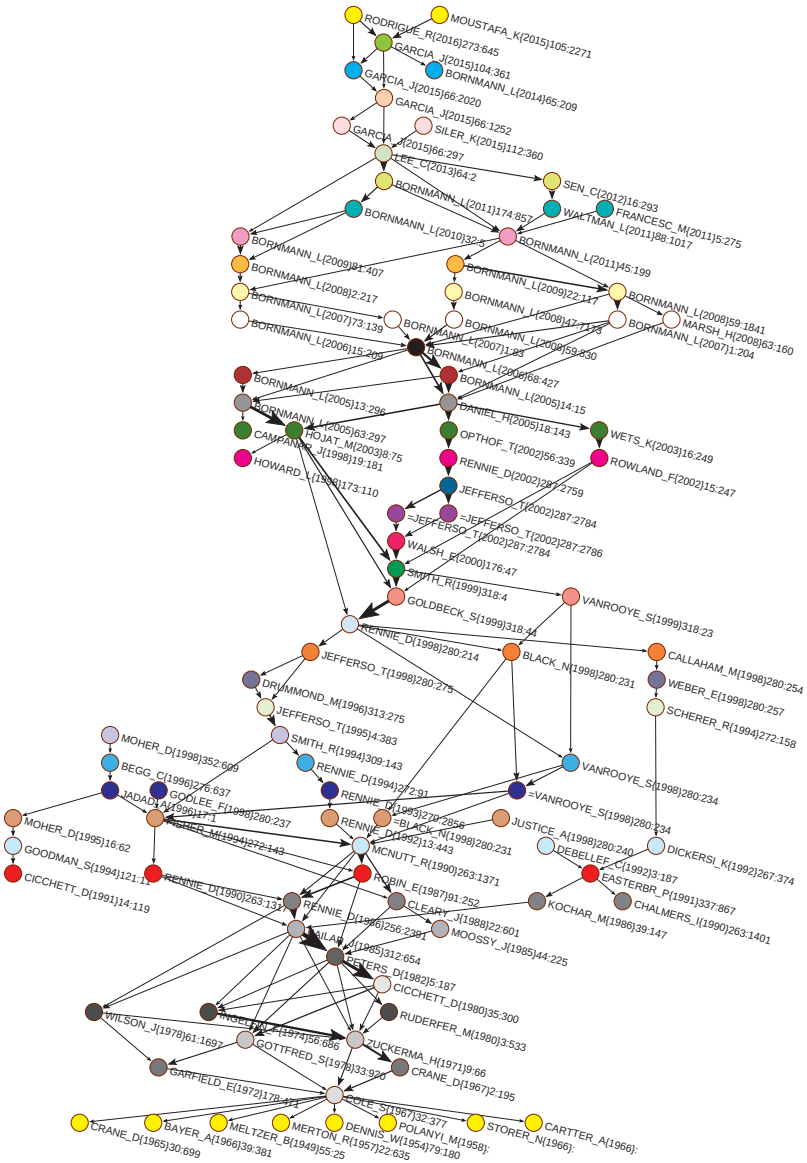


PEERE – SPC link islands [20 200]



PEERE – SPC – Link island1

$w_{max} = 0.297$



This island is very similar to the main paths for 100 largest weights and includes main path.

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 links from $\mathcal{L}$ have one end-node in $\mathcal{U}$ and the other in $\mathcal{V}$. Often also a *weight* $w : \mathcal{L} \rightarrow \mathbb{R} \in \mathcal{W}$ 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	X	X	X	X	X	X	.	X	X	.	.	.	.	.
2. Miss Laura Mandeville	X	X	X	.	X	X	X	X	.	.	.	.	.	.
3. Miss Theresa Anderson	.	X	X	X	X	X	X	X	X	.	.	.	.	.
4. Miss Brenda Rogers	X	.	X	X	X	X	X	X	.	.	.	.	.	.
5. Miss Charlotte McDowd	.	.	X	X	X	.	X	.	.	.	.	.	.	.
6. Miss Frances Anderson	.	.	X	.	X	X	.	X	.	.	.	.	.	.
7. Miss Eleanor Nye	.	.	.	.	X	X	X	X	.	.	.	.	.	.
8. Miss Pearl Oglethorpe	.	.	.	.	.	X	.	X	X	.	.	.	.	.
9. Miss Ruth DeSand	.	.	.	.	X	.	X	X	X	.	.	.	.	.
10. Miss Verne Sanderson	.	.	.	.	.	.	X	X	X	.	.	X	.	.
11. Miss Myra Liddell	.	.	.	.	.	.	.	X	X	X	.	X	.	.
12. Miss Katherine Rogers	.	.	.	.	.	.	.	X	X	X	.	X	X	X
13. Mrs. Sylvia Avondale	.	.	.	.	.	.	X	X	X	X	.	X	X	X
14. Mrs. Nora Fayette	.	.	.	.	.	X	X	.	X	X	X	X	X	X
15. Mrs. Helen Lloyd	.	.	.	.	.	.	X	X	.	X	X	X	.	.
16. Mrs. Dorothy Murchison	.	.	.	.	.	.	.	X	X	.	.	.	.	.
17. Mrs. Olivia Carleton	.	.	.	.	.	.	.	.	X	.	X	.	.	.
18. Mrs. Flora Price	.	.	.	.	.	.	.	.	X	.	X	.	.	.

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 $\mathbf{x} = \mathbf{A}^{\mathsf{T}}\mathbf{y}$.

`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 $\mathcal{K}(p, q)$:

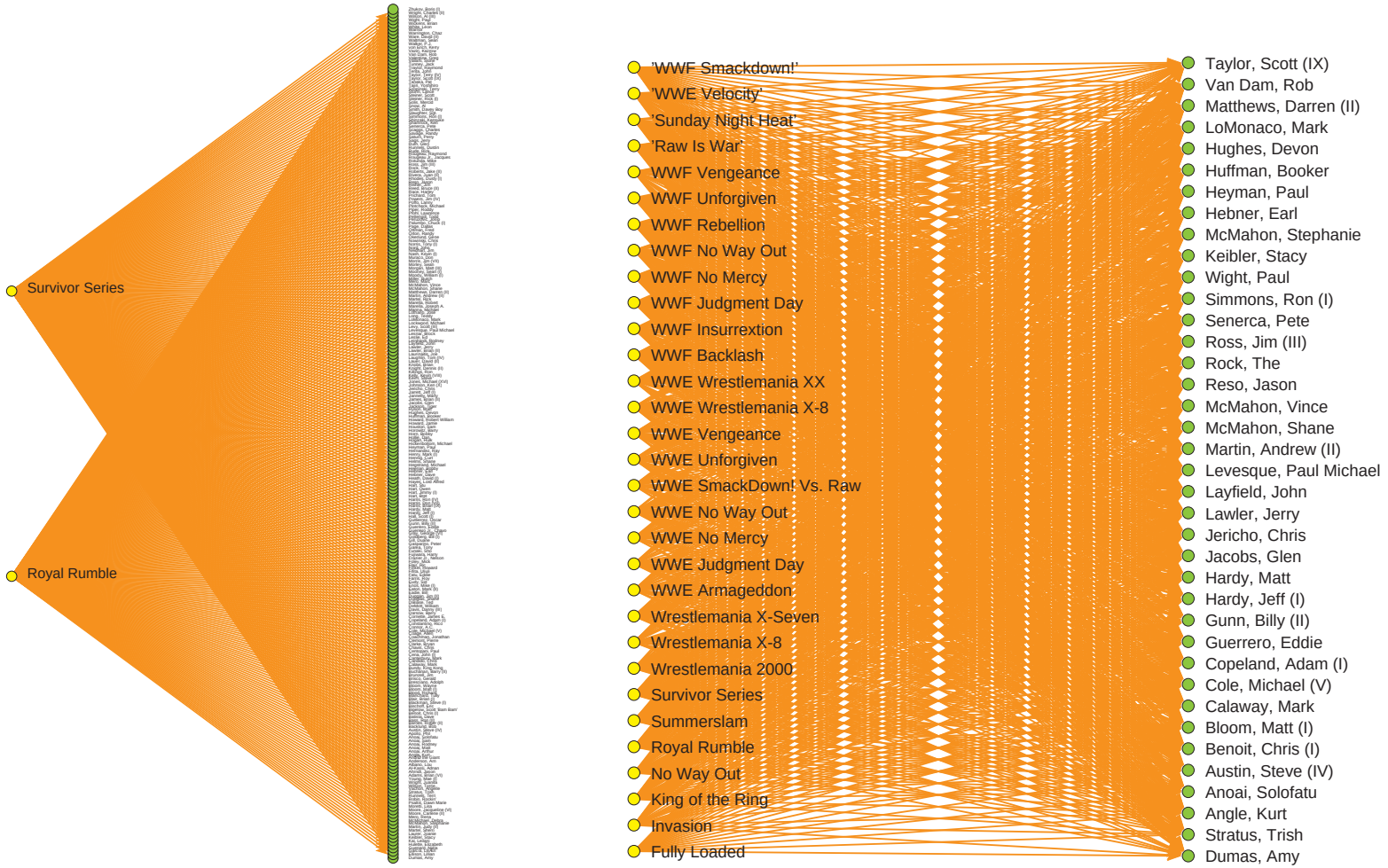
- $n_1 + n_2 \leq$ selected threshold
- 'border line' in the (p, q)-table.

Table $(p, q : n_1, n_2)$ for Internet Movie Database

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

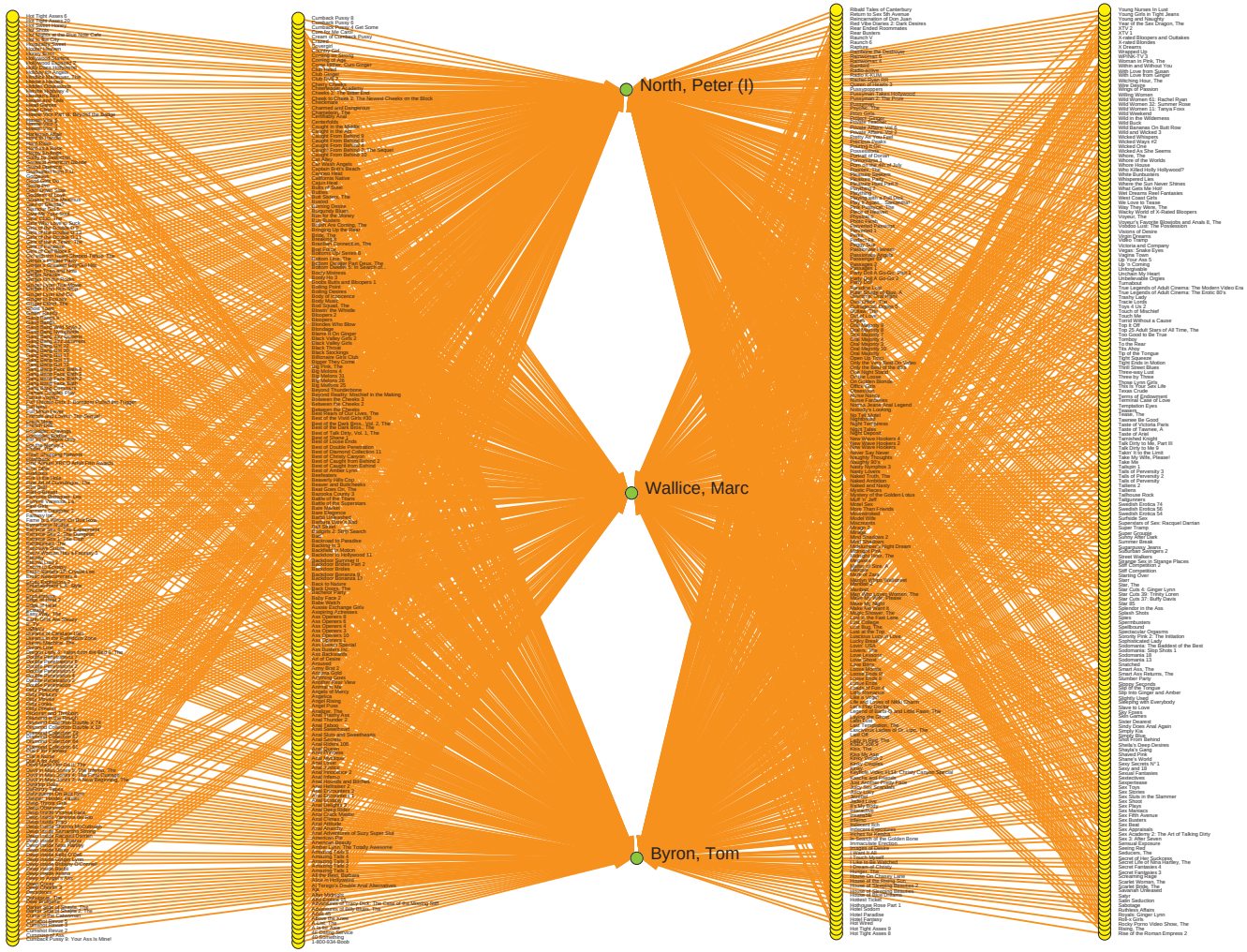
1	1590:	1590	1	16	39:	2173	678	44	14:	29	83
2	516:	788	3	17	35:	2791	995	46	13:	29	94
3	212:	1705	18	18	32:	2684	1080	49	12:	26	95
4	151:	4330	154	19	30:	2395	1063	52	11:	16	79
5	131:	4282	209	20	28:	2216	1087	56	10:	34	162
6	115:	3635	223	21	26:	1988	1087	62	9:	31	177
7	101:	3224	244	22	24:	1854	1153	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



IMDb: $n_1 = 428440$, $n_2 = 896308$, $m = 3792390$.

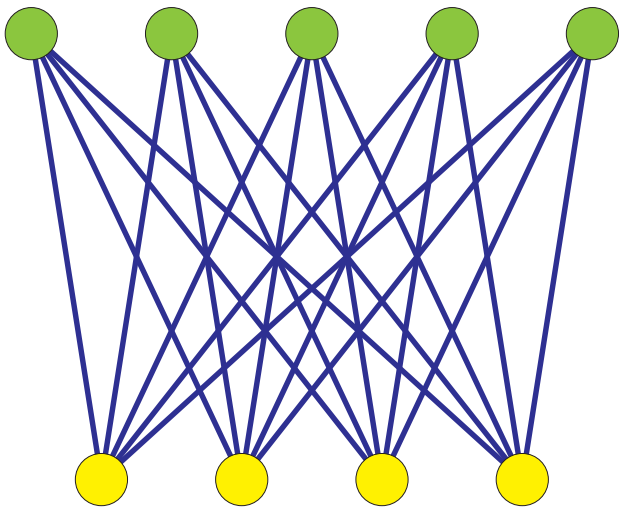
(2,516)-Hard core



4-rings and analysis of two-mode networks

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 $e \in \mathcal{E}$

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

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

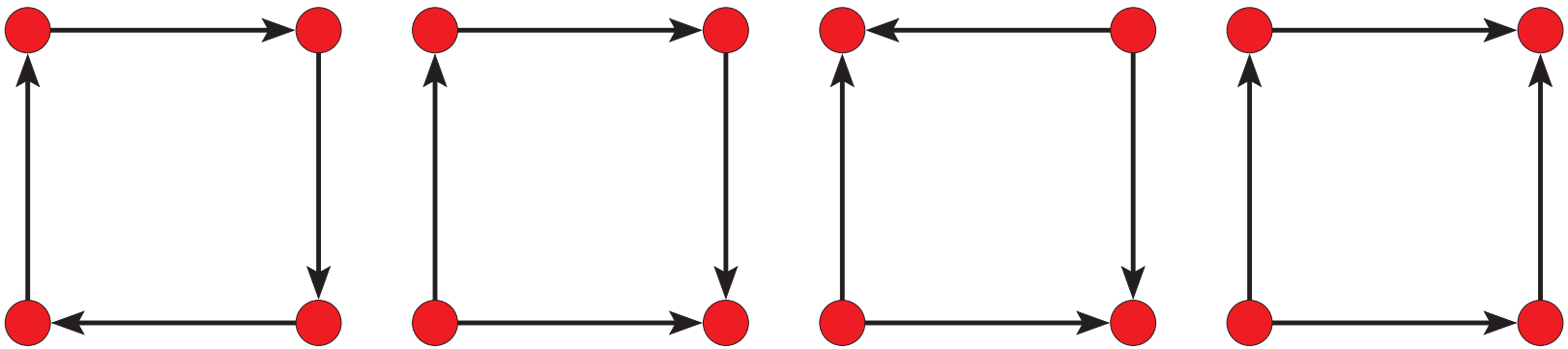
4-rings in $K_{p,q}$; and each of its edges e has weight

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

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

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*.

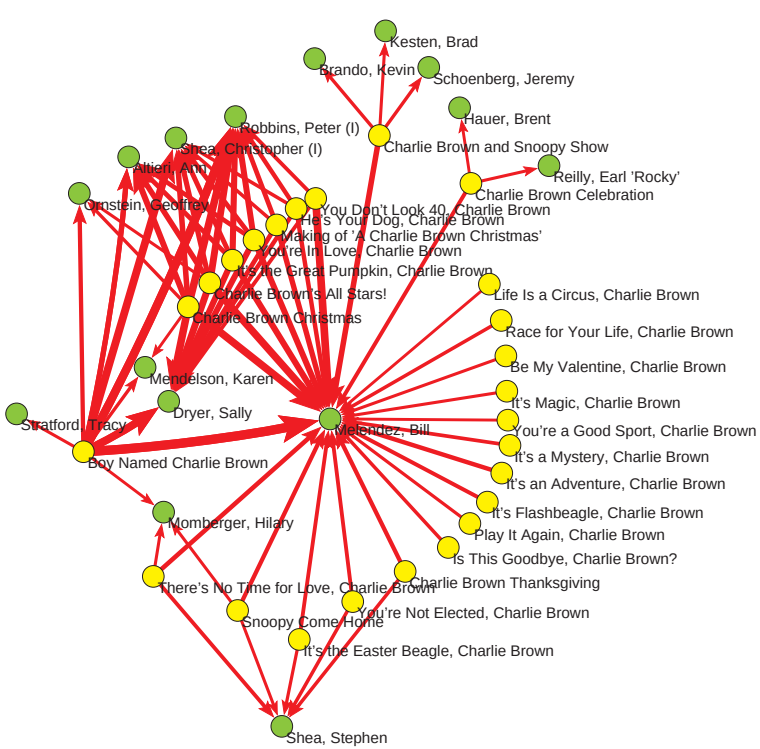
`Network/Create New Network/with Ring Counts/4-Rings/Directed`

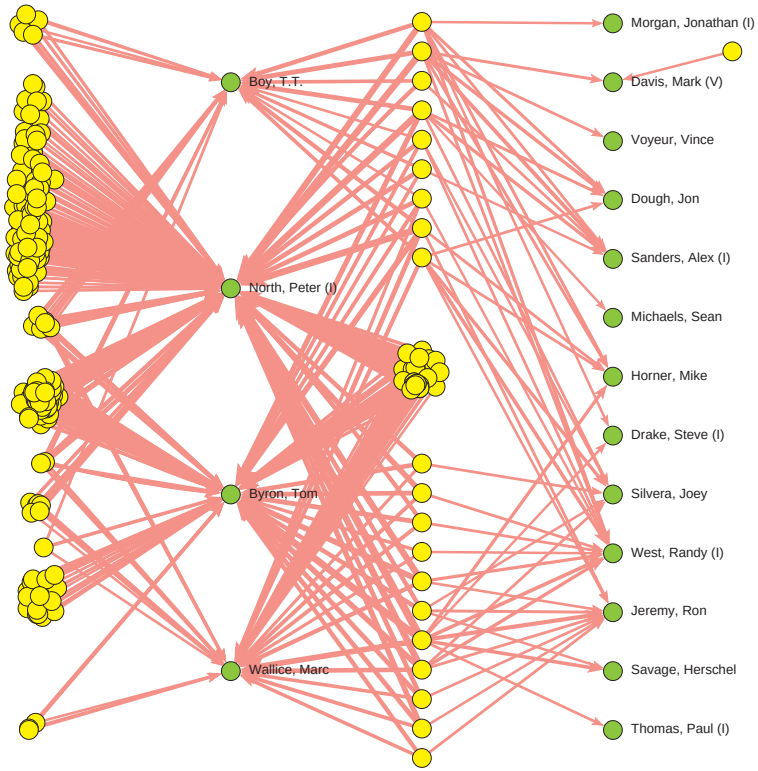
Simple link islands in IMDb for w_4

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

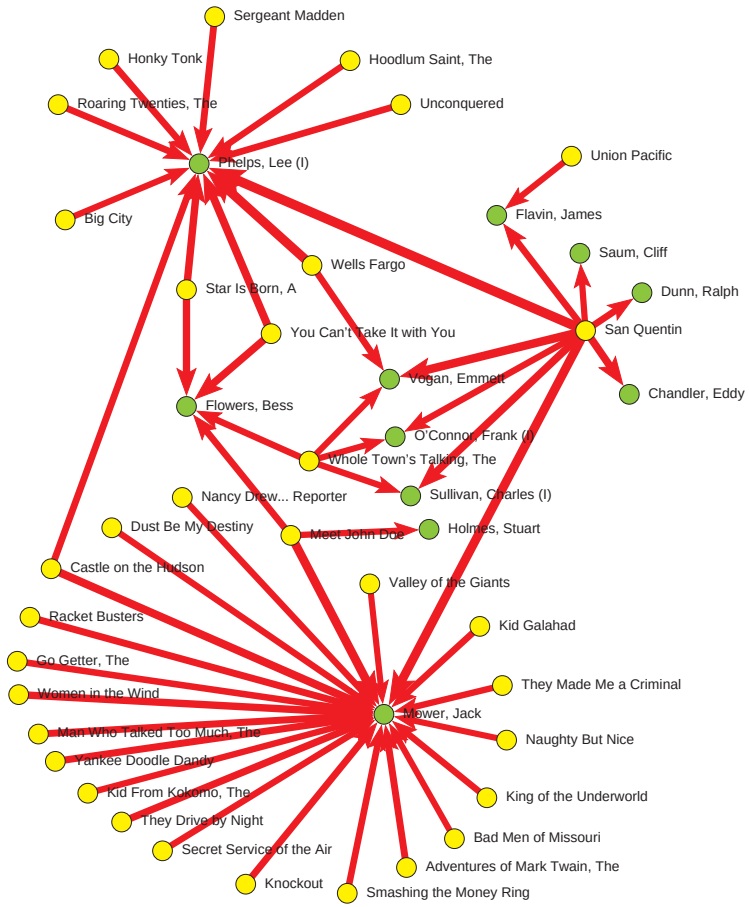
Size	Freq	Size	Freq	Size	Freq	Size	Freq
2	5512	20	19	38	4	59	2
3	1978	21	18	39	3	61	1
4	1639	22	15	40	2	64	1
5	968	23	9	42	2	67	1
6	666	24	13	43	3	70	1
7	394	25	12	45	3	73	1
8	257	26	6	46	4	76	1
9	209	27	6	47	5	82	1
10	148	28	5	48	1	86	1
11	118	29	6	49	2	106	1
12	87	30	3	50	2	122	1
13	55	31	6	51	1	135	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

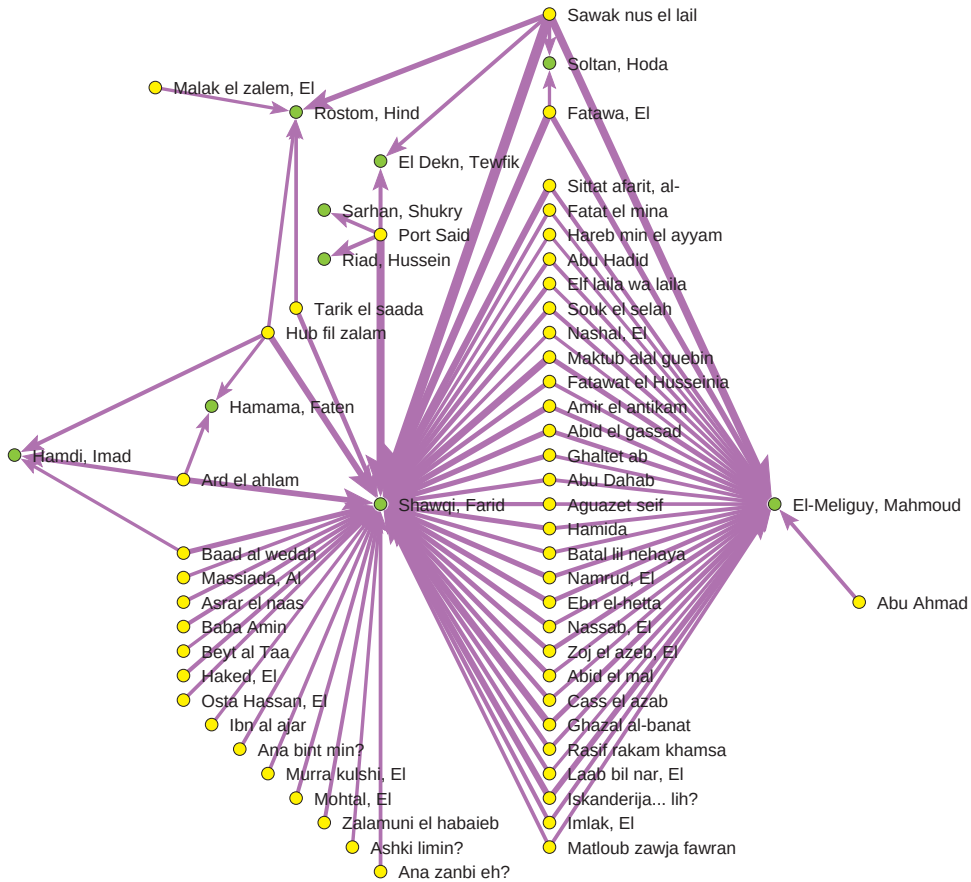
Example: Islands for w_4 Charlie Brown and Adult



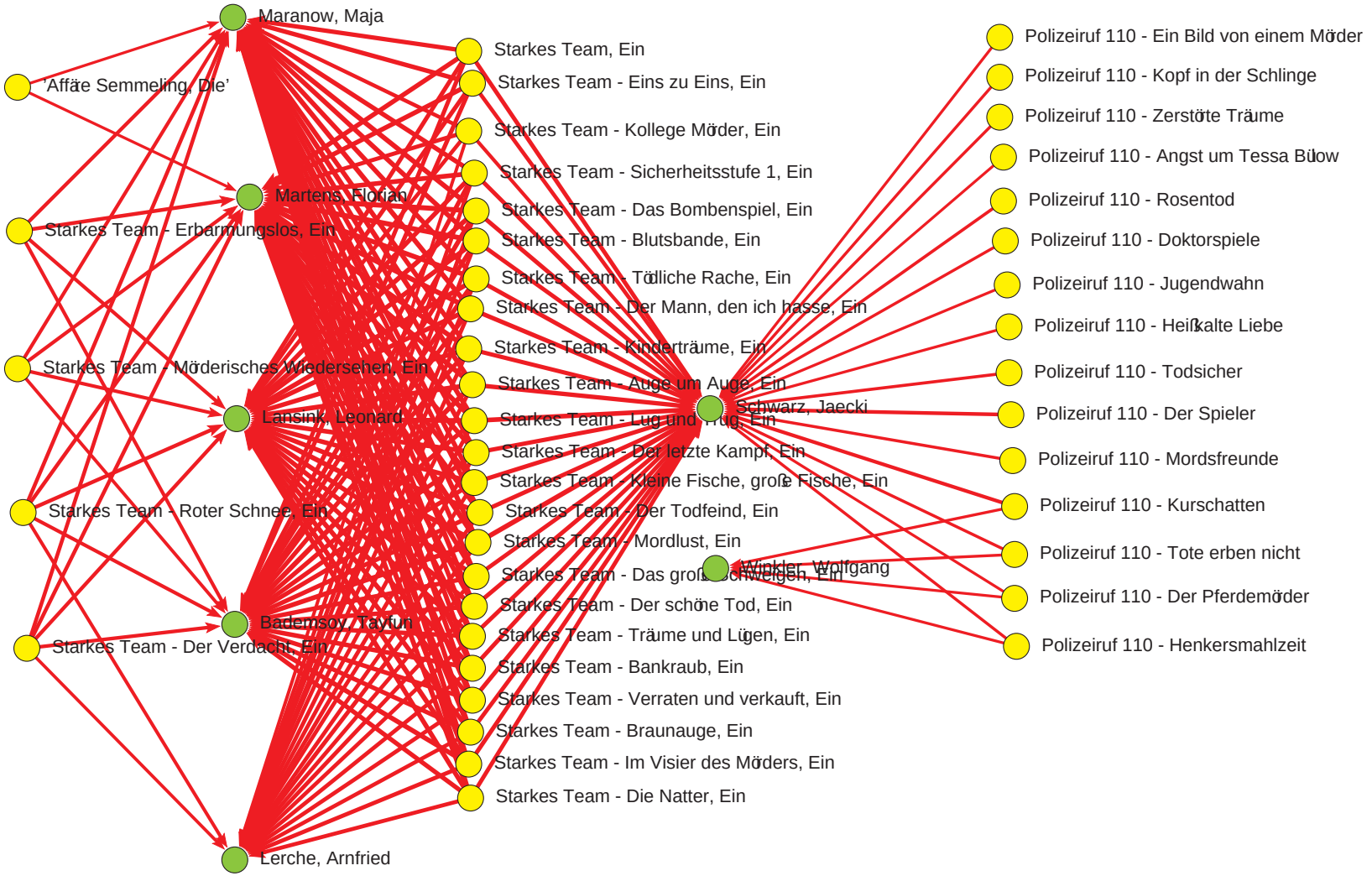


Example: Islands for w_4 Mark Twain and Abid





Example: Island for w_4 Polizeiruf 110 and Starkes Team



Multiplication of networks

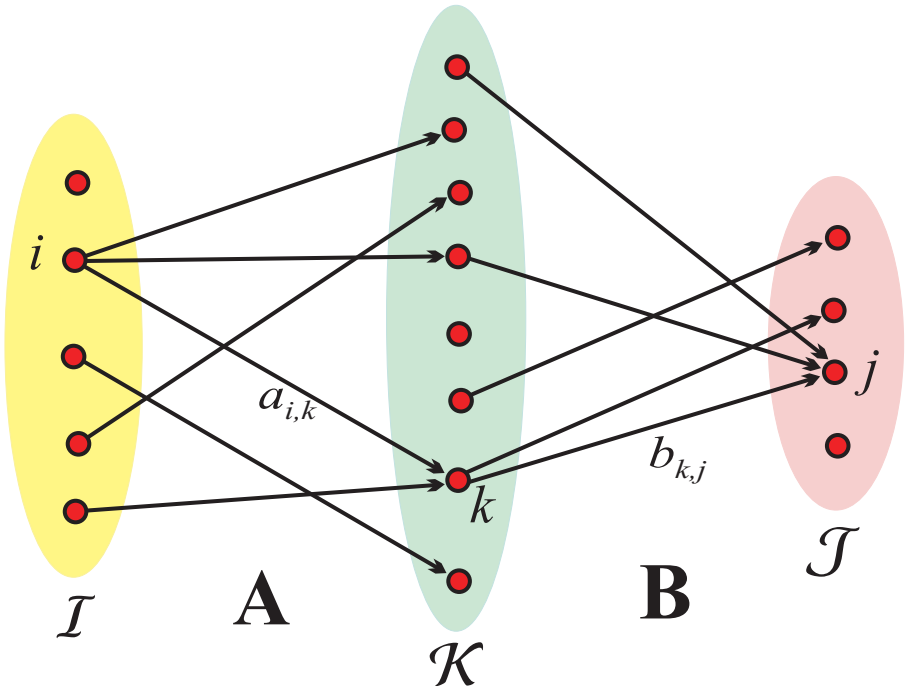
To a simple (no parallel arcs) two-mode *network* $\mathcal{N} = (\mathcal{I}, \mathcal{J}, \mathcal{A}, w)$; where $\mathcal{I}$ and $\mathcal{J}$ are sets of *nodes*, $\mathcal{A}$ is a set of *arcs* linking $\mathcal{I}$ and $\mathcal{J}$, and $w : \mathcal{A} \to \mathbb{R}$ (or some other semiring) 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 $\mathcal{K}$.

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

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.

Multiplication of networks – details

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'.

For more details see the paper.

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

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`

Authorship networks

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.

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

Let **N** be its normalized version

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

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

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 node 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

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

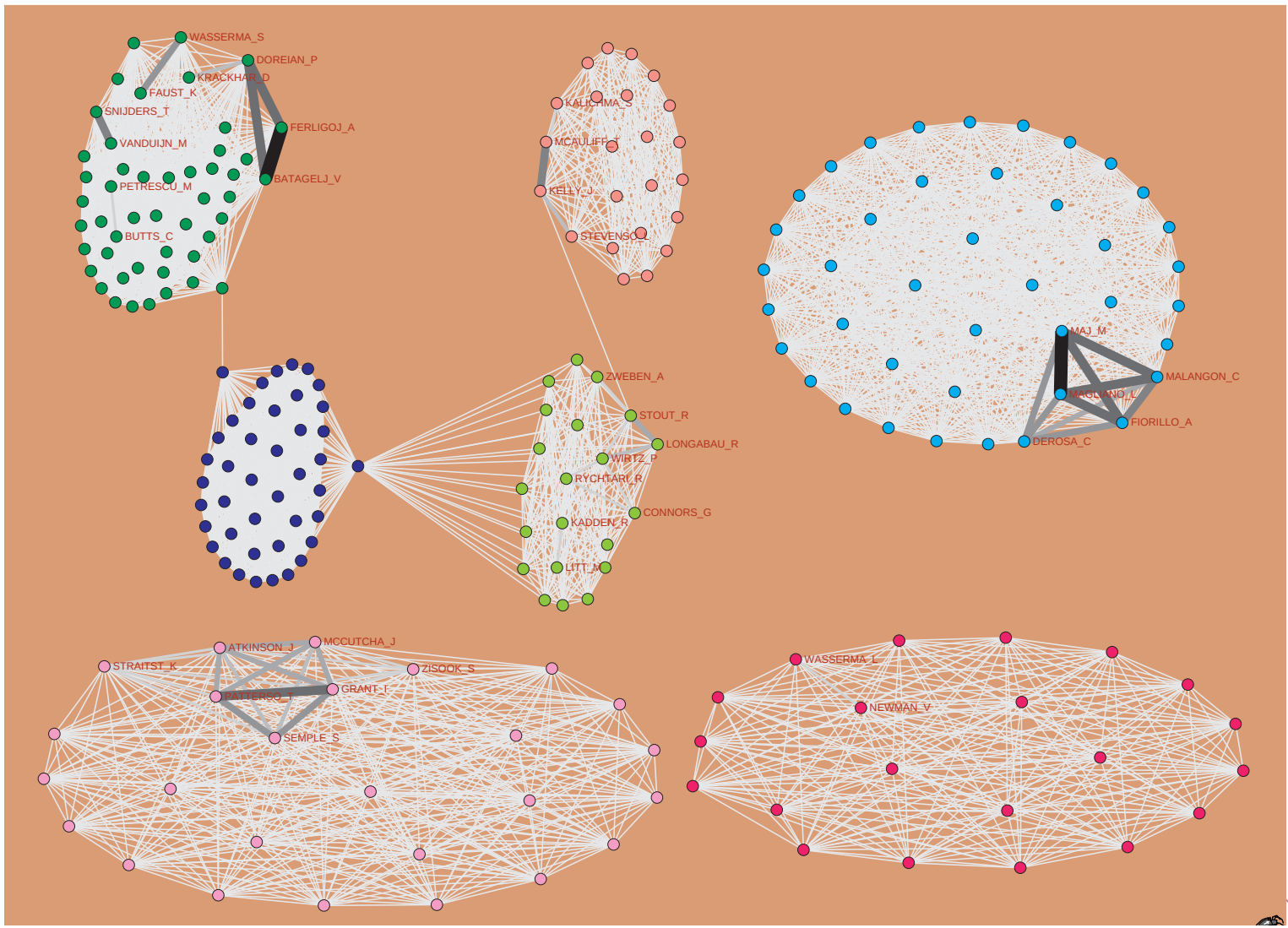
$$co_{ij} = \sum_{p \in W} wa_{pi} wa_{pj} = \sum_{p \in N^-(i) \cap N^-(j)} 1$$

co_{ij} = the number of works that authors i and j wrote together
co_{ii} = the total number of works that author i wrote
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

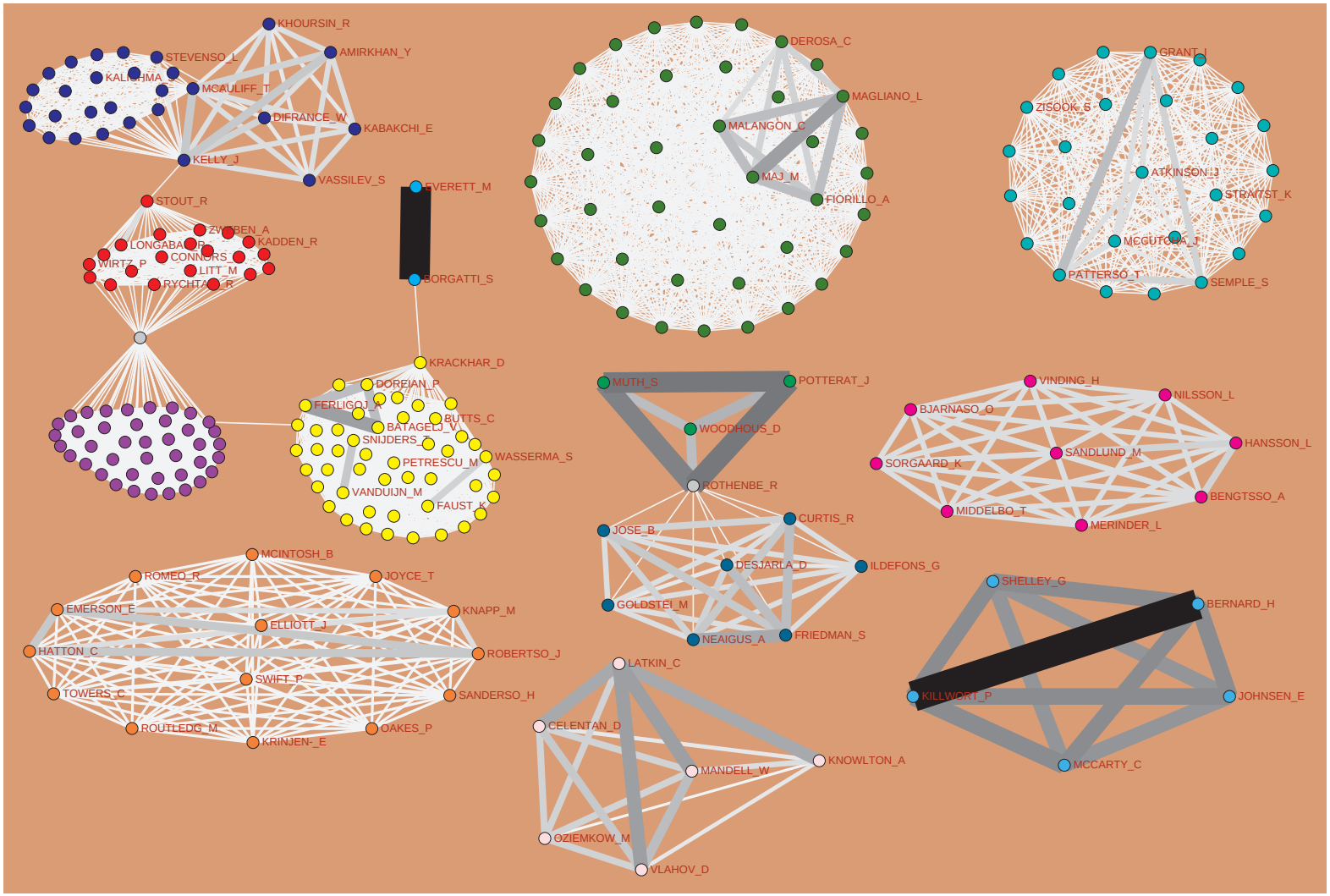
Problem: The **Co** network is composed of complete graphs on the set of work's authors. Works with many authors produce large complete subgraphs and are over-represented, thus bluring the collaboration structure.

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**(SN5)



Second co-authorship network

Cn = **AW** $*$ **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 $\sum_{j \in A} \sum_{j \in A} wa_{pi}\, n_{pj} = \text{outdeg}_{WA}(p)$ and $\sum_{j \in A} cn_{ij} = \text{indeg}_{WA}(i)$

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

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

Collaborativness: $K_i = 1 - S_i$

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

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 $ct_{ij} = ct_{ji}$ and

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

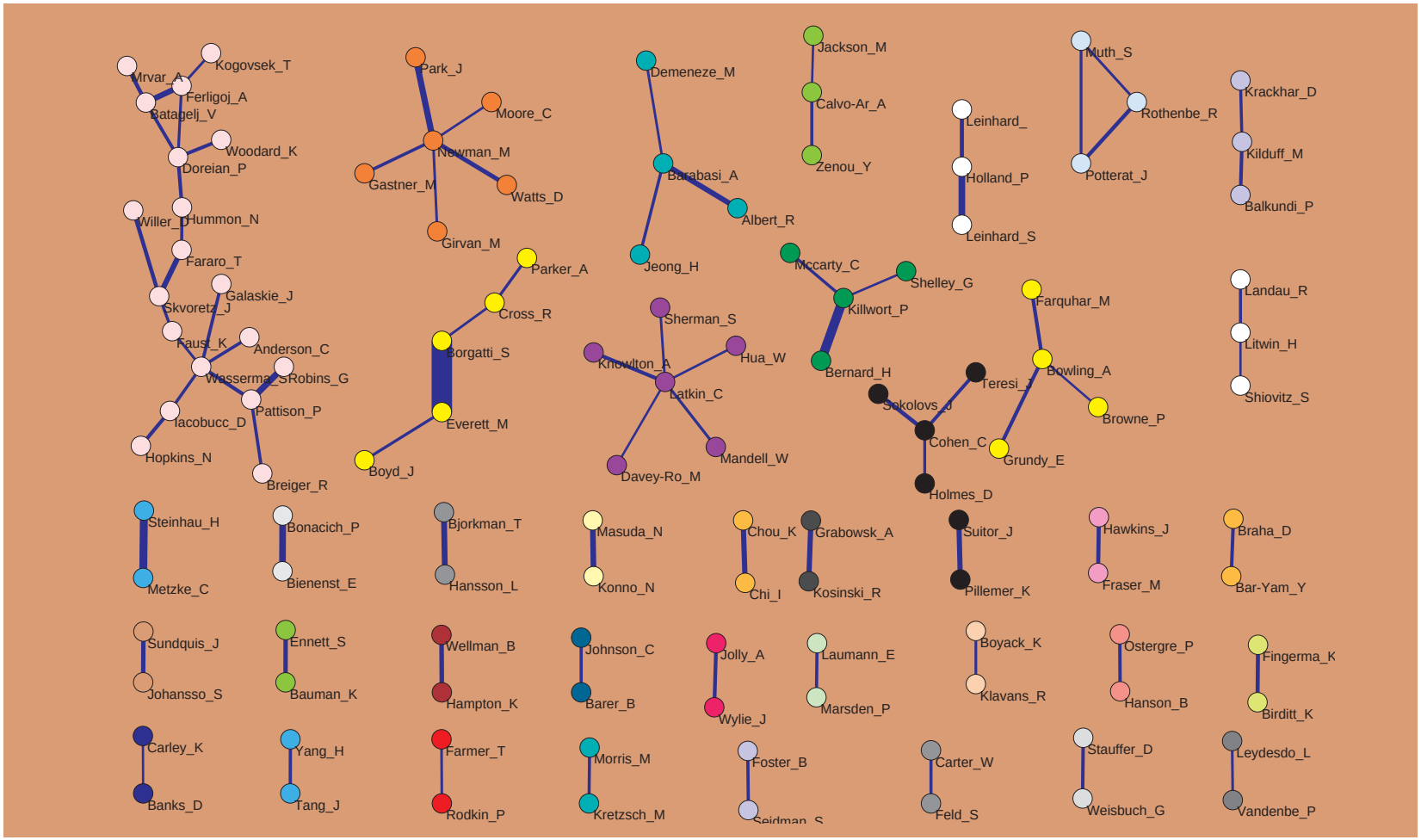
The total contribution of a complete subgraph corresponding to the authors of a work p is 1.

$\sum_{j \in A} ct_{ij} = \sum_{p \in W} n_{pi}$ = the total contribution of author i to works from W.

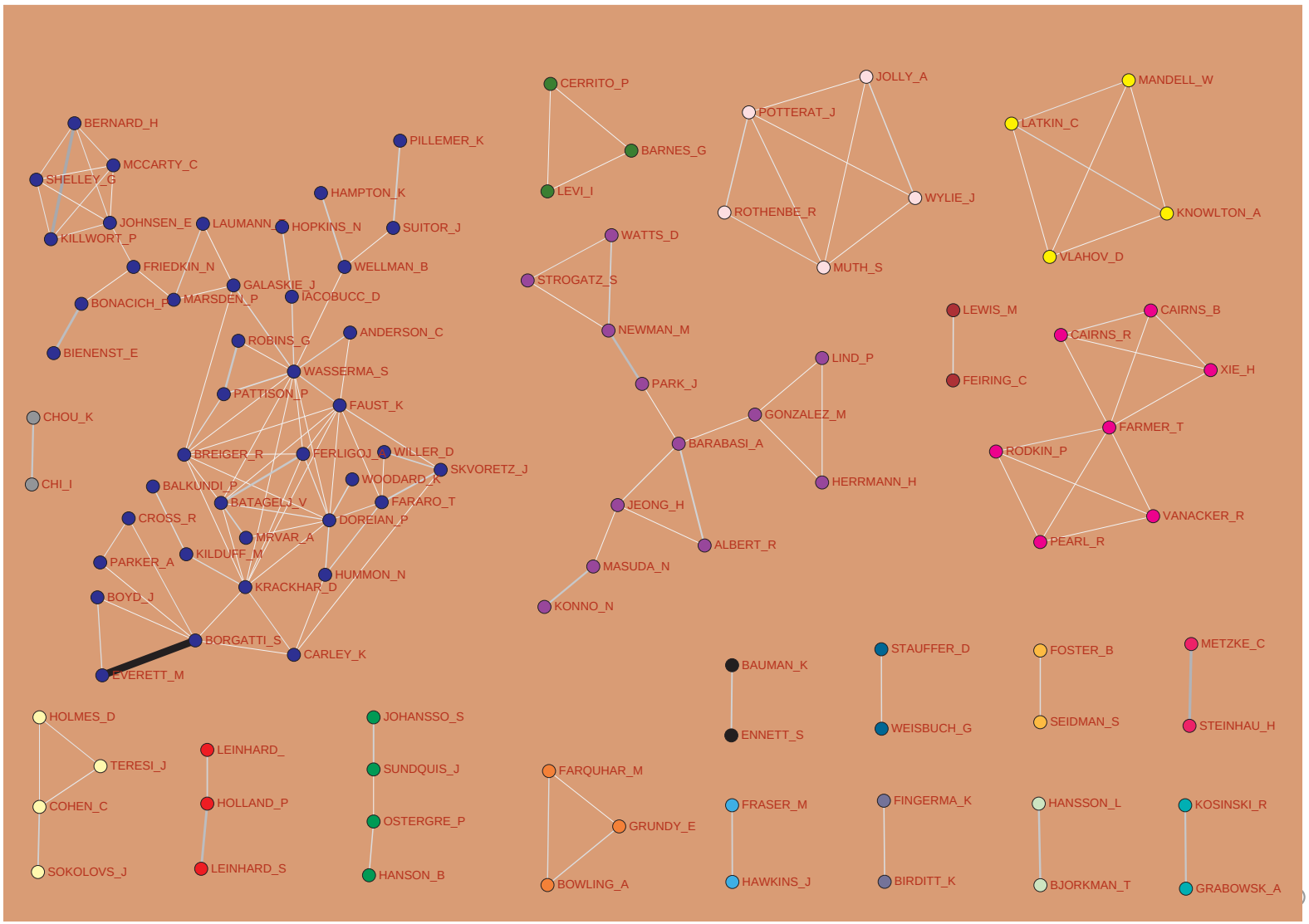
$$\sum_{i \in A} \sum_{j \in A} ct_{ij} = |W|$$

Components in **Ct**(SN5) cut at level 0.5

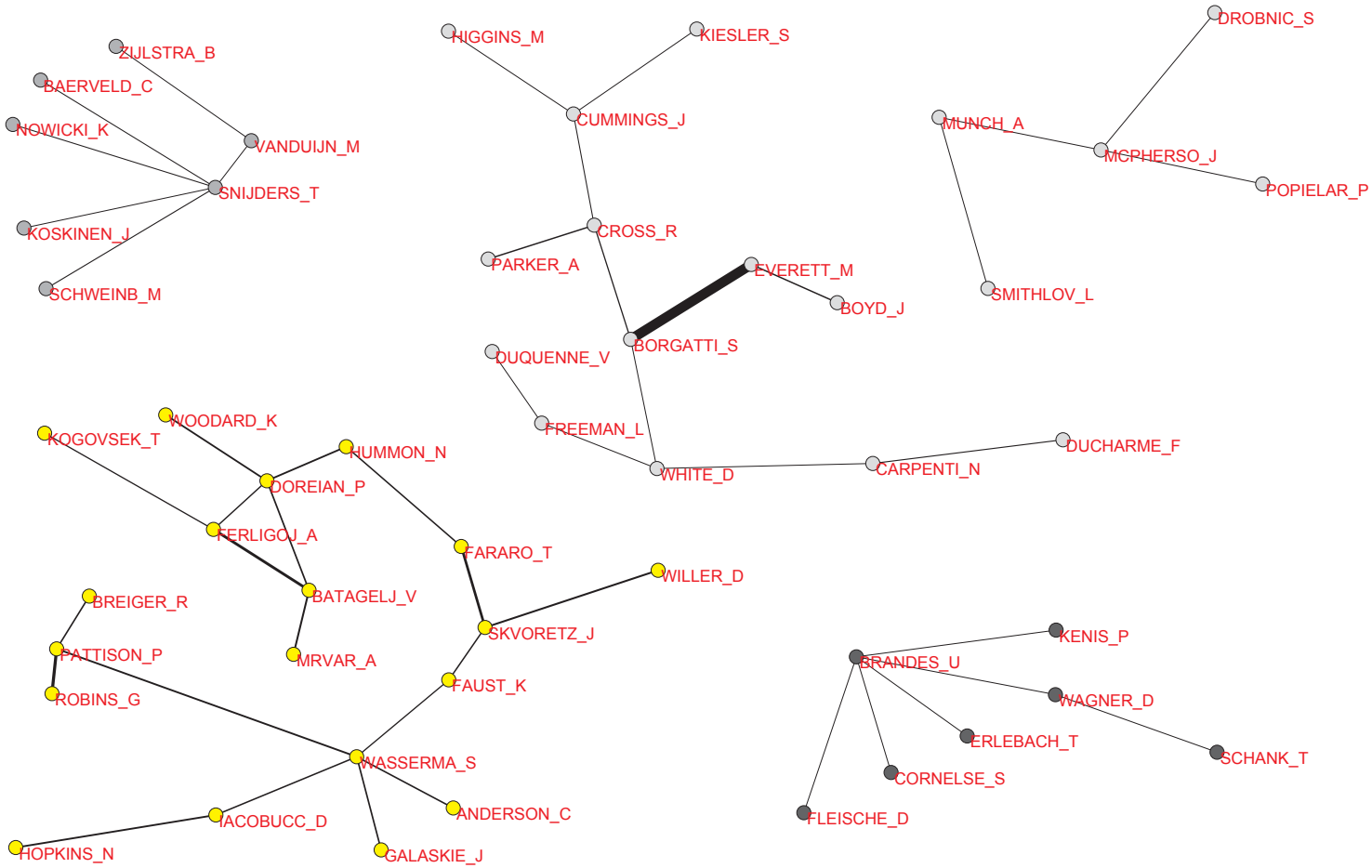
Network SN5 (2008): for "social network*" + most frequent references + around 100 social networkers;
$|W| = 193376$, $|C| = 7950$, $|A| = 75930$, $|J| = 14651$, $|K| = 29267$



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 $n'_{pi} = wa_{pi} / \max(1, \text{outdeg}_{WA}(p) - 1)$

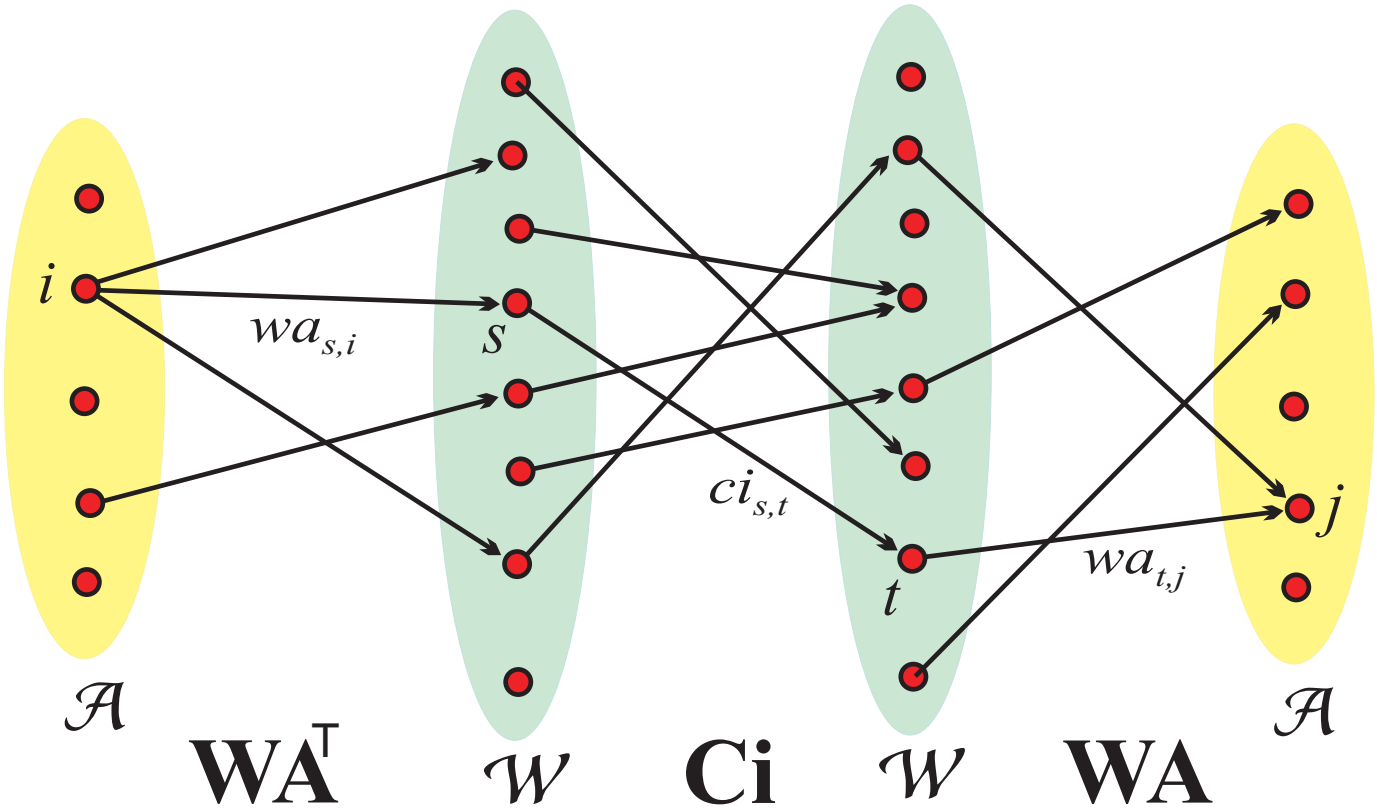
ct'_{ij} = the total contribution of 'strict collaboration' of authors i and j to works.

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 $i \neq j$)

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

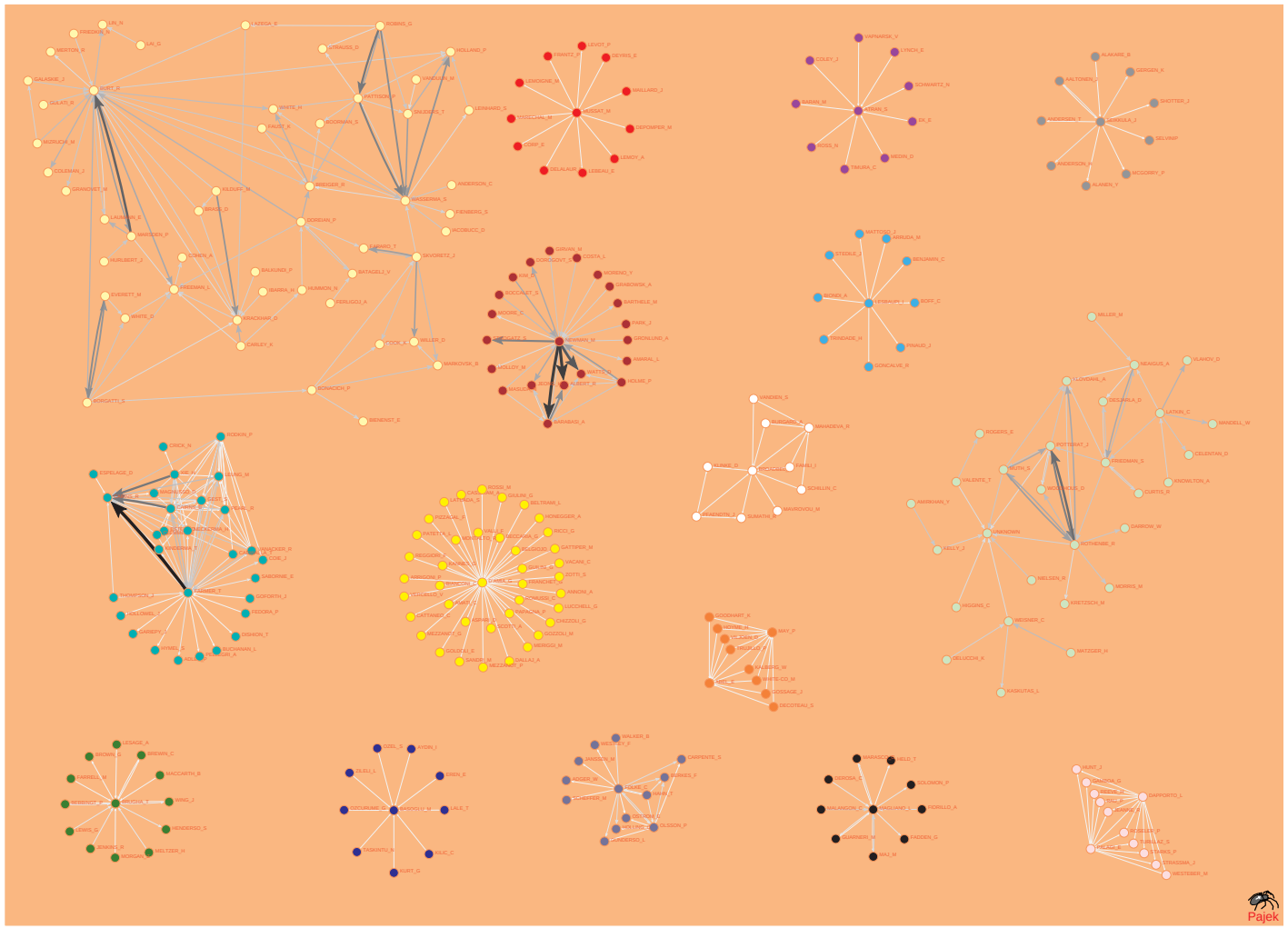
Authors' citations network



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 $p\mathbf{Ci}q \equiv$ work p cites work q. Therefore the *bibliographic coupling* network **biCo** can be determined as

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

$bico_{pq} = \#$ of works cited by both works p and q. $bico_{pq} = bico_{qp}$. Again we have problems with works with many citations, especially with review papers. To neutralize their impact we can introduce a normalized measure such as

$$\mathbf{biCon} = \frac{1}{2}(n(\mathbf{Ci}) * \mathbf{Ci}^T + \mathbf{Ci} * n(\mathbf{Ci})^T)$$

It is easy to verify that $bicon_{pq} \in [0, 1]$ and $bicon_{pq} = bicon_{qp}$ (symmetry). It also holds: $bicon_{pq} = 1$ iff the works p and q are referencing the same works.

Co-Citation and others

The *co-citation* network **coCi** can be determined as

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

$coci_{pq} = \#$ of works citing both works p and q.
$coci_{pq} = coci_{qp}$.
The 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 $\mathbf{AK} = \mathbf{AW} * \mathbf{WK}$.

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 $A_{\text{⌘}}(\mathcal{T})$ denote the set of all temporal quantities over $A_{\text{⌘}}$ 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)]
```

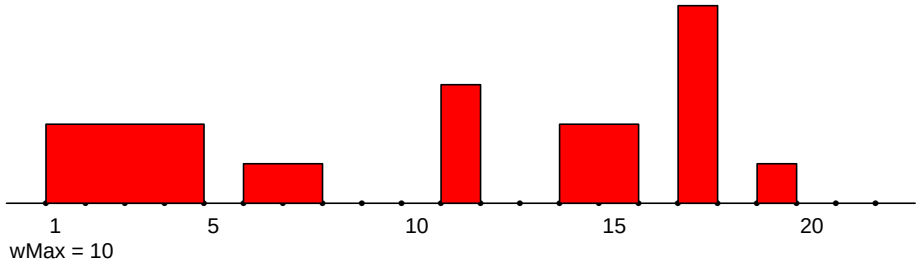
The following are the sum $s = a + b$ and the product $p = a \cdot b$ of temporal quantities a and b over combinatorial semiring.

```
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)]
```

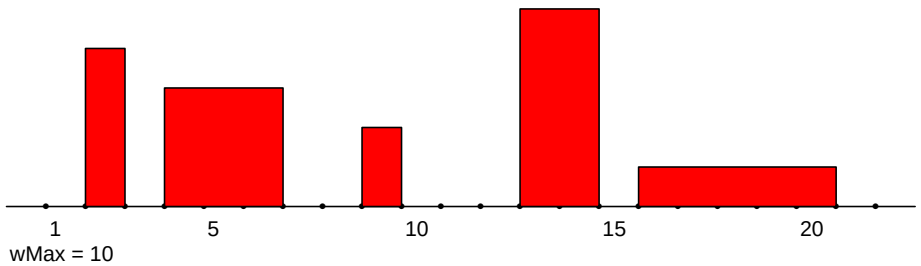
They are visually displayed at the bottom half of figures on the following slides.

Addition of temporal quantities.

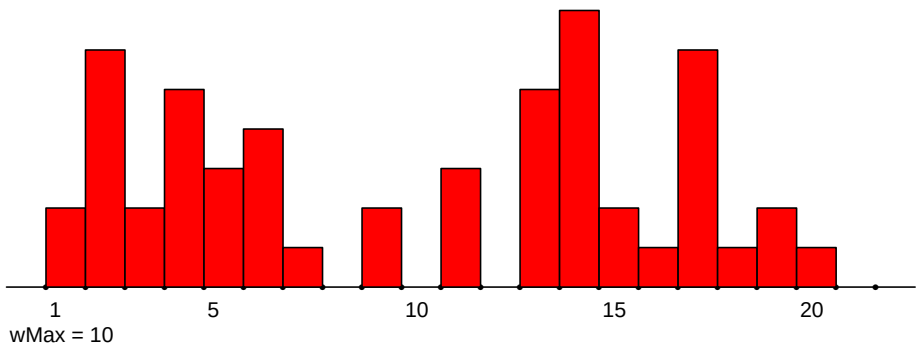
a :



b :

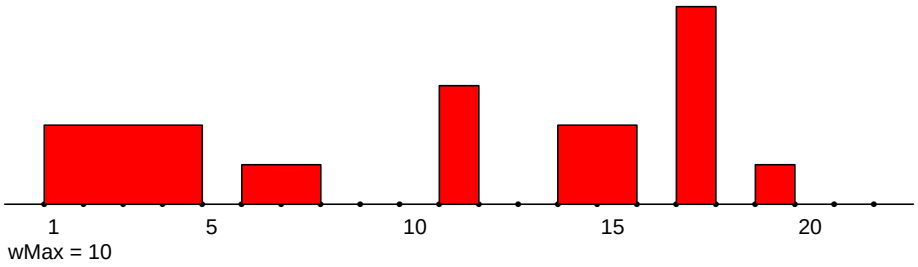


$a + b$:

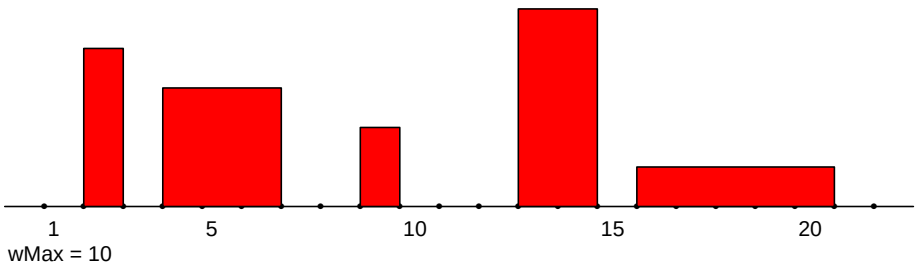


Multiplication of temporal quantities.

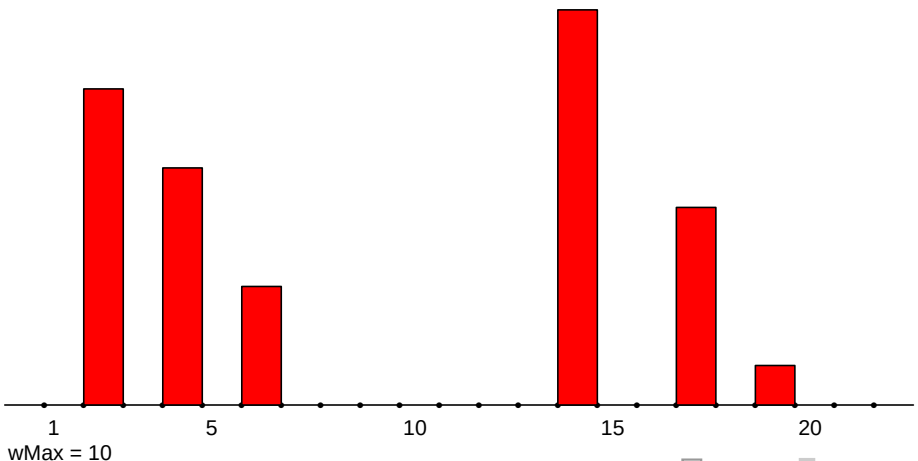
a :



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. $\mathcal{T} = [first, last] \subset \mathbb{N}$. 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

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

Multiplication of temporal affiliation networks

Instantaneous

Instantaneous $\mathbf{A}$ on $P \times A$ and $\mathbf{B}$ 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 $\mathbf{A}$ and $\mathbf{B}$ 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 $\mathbf{A}$ on $P \times A$ and $\mathbf{B}$ 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 $\mathbf{A}$ and $\mathbf{B}$ 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.

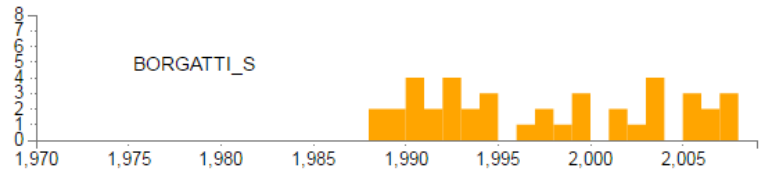
multiply.py

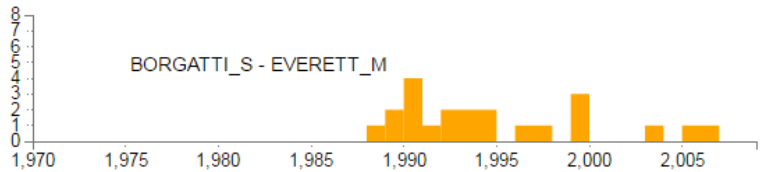
```
gdir = 'c:/users/batagelj/work/python/graph/graph'
wdir = 'c:/users/batagelj/work/python/graph/JSON/SN5'
cdir = 'c:/users/batagelj/work/python/graph/chart'
import sys, os, datetime, json
sys.path = [gdir]+sys.path; os.chdir(wdir)
import TQ
from GraphNew import Graph
# file = 'C:/Users/batagelj/work/Python/graph/JSON/WAtest.json'
file = 'C:/Users/batagelj/work/Python/graph/JSON/SN5/WAcInst.json'
# file = 'C:/Users/batagelj/work/Python/graph/JSON/SN5/WAcCum.json'
# file = 'C:/Users/batagelj/work/Python/graph/JSON/Gisela/papInst.json'
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']
```

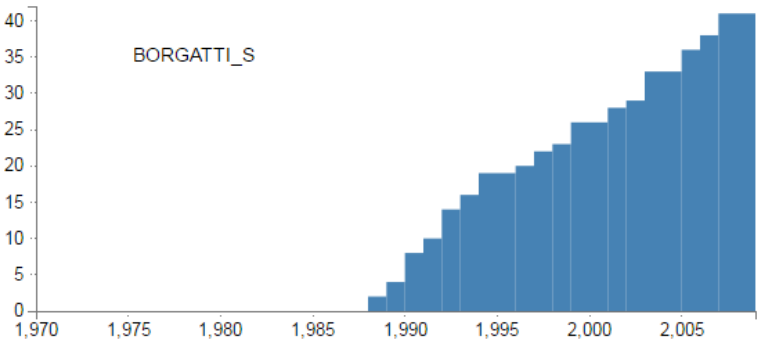
Temporal co-authorship network for SN5

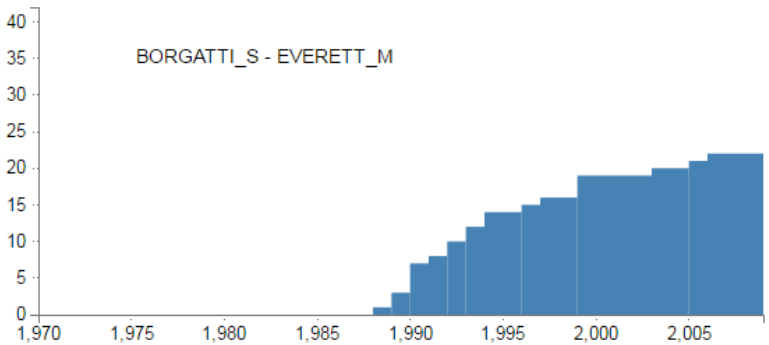
```
======= RESTART: C:\Users\batagelj\work\Python\graph\graph\multiply.py =======
started:  Sun Nov 20 00:26:51 2016
loaded:  Sun Nov 20 00:26:51 2016
time used:  0:00:00.425024
computed:  Sun Nov 20 00:26:52 2016
time used:  0:00:01.165066
>>> BB = CC._links[(ia['BORGATTI_S'],ia['BORGATTI_S'])][4]['tq']
>>> 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')
```

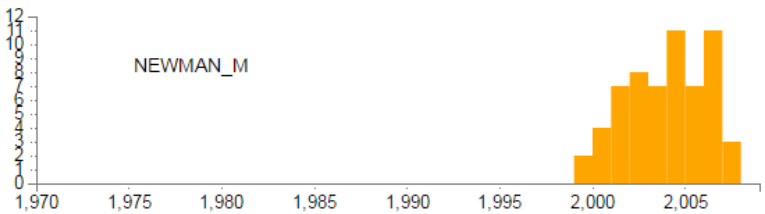
Visualization



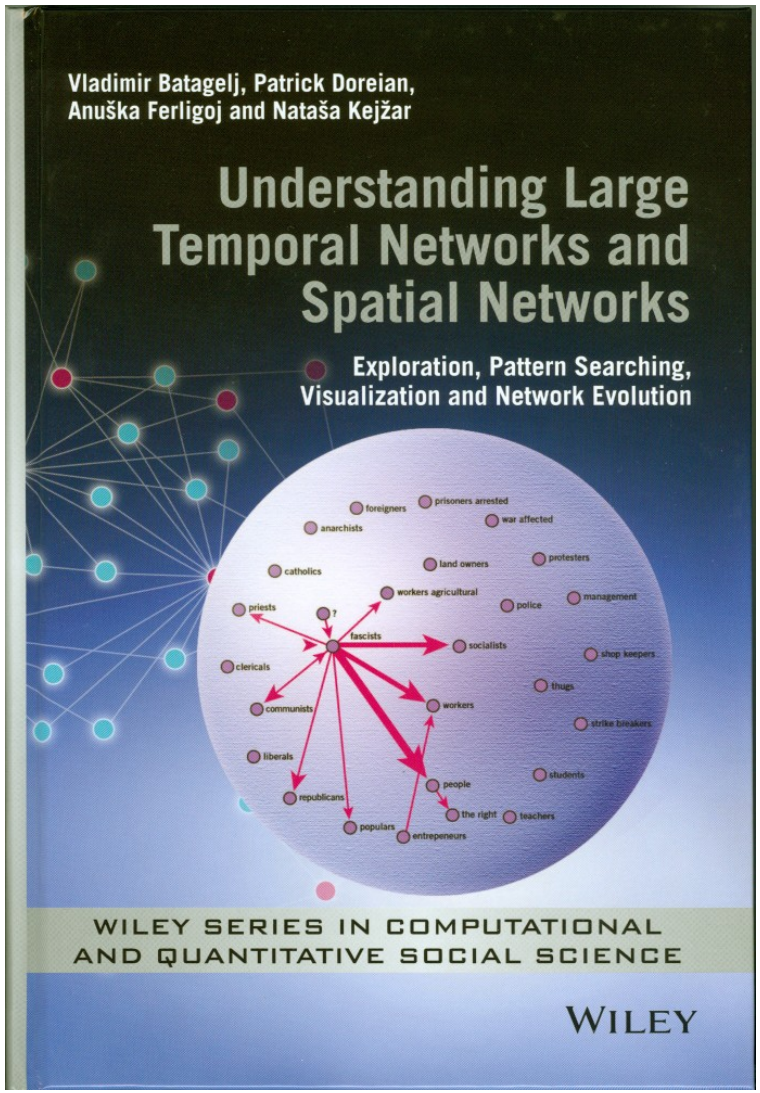








Understanding large networks



This lecture 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.

References I

Ahmed, A., Batagelj, V., Fu, X., Hong, S.-H., Merrick, D., Mrvar, A.: Visualisation and analysis of the Internet movie database. Asia-Pacific Symposium on Visualisation 2007 (IEEE Cat. No. 07EX1615), 2007, p 17-24.

Barabasi, A.L., Jeong, H., Neda, Z., Ravasz, E., Schubert, A., Vicsek, T.: Evolution of the social network of scientic collaborations. Physica **311** (2002) 590–614

Batagelj, V.: Wos2pajek – networks from web of science (2007). http://vladowiki.fmf.uni-lj.si/doku.php?id=pajek:wos2pajek

Batagelj, V, Cerinšek, M: On bibliographic networks. Scientometrics 96 (2013) 3, 845-864.

Batagelj, V., Praprotnik, S.: An algebraic approach to temporal network analysis based on temporal quantities. Social Network Analysis and Mining, 6(2016)1, 1-22.

Batagelj, V., Zaveršnik, M.: Fast algorithms for determining (generalized) core groups in social networks. Advances in Data Analysis and Classification, 2011. Volume 5, Number 2, 129-145.

References II


Cerinšek, M., Batagelj, V.: Network analysis of Zentralblatt MATH data. Scientometrics, 102(2015)1, 977-1001.


Cerinšek, M., Batagelj, V.: Generalized two-mode cores. Social Networks 42 (2015), 80–87.


De Nooy, W., Mrvar, A., Batagelj, V.: Exploratory Social Network Analysis with Pajek; Revised and Expanded Second Edition. Structural Analysis in the Social Sciences, Cambridge University Press, September 2011.


Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.: From data mining to knowledge discovery in databases. American Association for Artificial Intelligence Magazine (1996), 37–54


J. G. Fletcher, "A more general algorithm for computing closed semiring costs between vertices of a directed graph," *CACM* (1980), pp. 350-351.


Kejžar, N., Korenjak Černe, S., Batagelj, V.: Network Analysis of Works on Clustering and Classification from Web of Science. Classification as a Tool for Research. Hermann Locarek-Junge, Claus Weihs eds. Proceedings of IFCS 2009. Studies in Classification, Data Analysis, and Knowledge Organization, 525-536, Springer, Berlin, 2010.

References III


Newman, M.E.: The structure of scientific collaboration communities. Proceedings of the National Academy of Science (PNAS) **98** (2001) 404–409


Newman, M.E.J., Girvan, M.: Finding and evaluating community structure in networks. Physical Review E **69** (2004)


Palla, G., Derényi, I., Farkas, I., Vicsek, T.: Uncovering the overlapping community structure of complex networks in nature and society. Nature **435** (2005) 814


Palla, G., Barabasi, A.L., Vicsek, T.: Quantifying social group evolution. Nature **446** (2007) 664-667


Perianes-Rodriguez, A., Waltman, L., Van Eck, N.J. (2016). Constructing bibliometric networks: A comparison between full and fractional counting. Journal of Informetrics, 10(4), 1178-1195.


Wasserman, S., Faust, K.: Social network analysis: methods and applications. Cambridge Univ. Press, Cambridge, 1997.


Zaveršnik, M., Batagelj, V.: Islands. In: XXIV International Sunbelt Social Network Conference, Portorož, Slovenia (2004)

References IV

Pajek's wiki. http://pajek.imfm.si

Vladimir Batagelj, Andrej Mrvar: Pajek manual.

Wouter De Nooy, Andrej Mrvar, Vladimir Batagelj: Exploratory Social Network Analysis with Pajek; Revised and Expanded Second Edition. Structural Analysis in the Social Sciences, Cambridge University Press, September 2011.

Names of works in WoS2Pajek

The usual *ISI name* of a work (field CR)

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

has the following structure

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

All its elements are in upper case.

In WoS the same work can have different ISI names. To improve the precission the program WoS2Pajek supports also *short names* (similar to the names used in HISTCITE output). They have the format:

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

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

From the last names with prefixes `VAN`, `DE`, ... the space is deleted.

Unusual names start with character * or $.

. . . Names of works

In the CR field other forms of ISI names and several errors and inconsistencies can be found:

```
NEWMAN MEJ, 2004, PHYS REV E 2, V69, ARTN 066133
PALLA G, 2005, NATURE, V435, P814, DOI 10.1038/nature03607
PAPIN JA, 2004, TRENDS BIOCHEM SCI, V29, P641, DOI
10.1016/j.tibs.2004.10.001
DOLCINI MM, 2005, J ADOLESCENT HEALTH, V36, UNSP 267.E6-15
EVANS JD, 2001, GENOME BIOL, V2, UNSP RESEARCH0001
NEWMAN MEJ, 2001, IN PRESS COMPLEX NETUNSP 215239
GRANOVET.MS, 1973, AM J SOCIOL, V78, P1360
GRANOVETTER M, 1983, SOCIOLOGICAL THEORY, V1, P203
BORGATTI SP, 2002, UGINET WINDOWS SOFTW
BORGATTI S, 1999, UCINET V USERS GUIDE
CANTANZARO M, 2005, PHYS REV E, V71, UNSP 027103
CANTAZARO M, 2005, PHYS REV E, V71, UNSP 056104
CATANZARO M, 2005, PHYS REV E 2, V71, ARTN 056104
BRICKER  PD, 1968, OCT M PSYCH SOC ST L  :  BRICKER
```

We decided to treat in short names the ARTN and UNSP values as BP values. We also remove the DOI parts. There are also irregular names in AU field:

```
AU BENSON, , C
   KULHAVY, , W
AU SCHONEMA.PH
```

The user can correct the typing errors and nonuniformities on the WoS file.

Program WoS2Pajek

For converting WoS file into networks in Pajek's format a program WoS2Pajek was developed (in Python). It produces the following files:

- citation network: works × works, **Ci**;
- authorship (two-mode) network: works × authors, **WA**; for works without complete description only the first author is known;
- keywords (two-mode) network: works × keywords, **WK**; only for works with complete description;
- journals (two-mode) network: works × journals, **WJ**; field J9;
- partition of works by the publication year, **year**;
- partition of works, **DC** – complete description (1) / ISI name only (0);
- vector number of pages, **np**; PG or EP − BP +1.

Program WoS2Pajek

The keywords are obtained from the fields TI (title), ID, DE and AB (abstract). From the text the stopwords are removed and a list of words is produced. The words are lemmatized using MontyLingua package.
In future versions aditional networks can be derived: works $\times$ discipline, works $\times$ countries, ...
In version 0.7 a GUI support (based on Tkinter) for specifying the program parameters was implemented.

Similar package of programs was produced also by Loet Leydesdorff.

Program WoS2Pajek

The current version of WoS2Pajek requires 7 parameters to be given by the user:

- MontyLingua directory: path to the directory in which the MontyLingua package is installed (put it also in the PATH env-variable);
- project directory: where the output files are saved;
- WoS file;
- maxnum – estimate of the number of all vertices (number of records + number of cited Works) – 30∗ number of records;
- step – prints info about each k*step record as a trace; step = 0 – no trace.
- use ISI name / short name;
- make a clean WoS file without duplicates;
- boolean list [DE, ID, TI, AB] specifying which fields are sources of keywords.

Collecting the data from Web of Science

When preparing for exploration of selected topic we first search the WoS for records matching selected keys. WoS allows storing the hits on a file, but only 500 at once.

For the list of hits we can in the **Citation report**, using the option **View citing articles**, get also the list of all works citing the works from the first list. We also save these records and join all the files into a single file to which we apply `WoS2Pajek`.

Using `Pajek` we identify the important only cited works – out-degre $= 0$ and large in-degree. We search for them in WoS and try to add their descriptions.

In this way we produce the final source data set for our analysis.

Types on DC file

When we combine partial files with saved records from WoS into a single file required by the program `WoS2Pajek` we can include into this file some additional lines:

Comments have the form

```
** comment
```

Besides this we can specify diffent types of input records using the lines of the form

```
*T n
```

where n is a type number (1, 2, ...). Since the same record can appear in different parts of the file its class is determined as the set of all corresponding types transformed in integer. For example: $\{3, 1\} \rightarrow 2^2 + 2^0 = 5$.