Microsoft Academic Graph

Viszards session

Vladimir Batagelj

IMFM Ljubljana and IAM UP Koper

XXXVI Sunbelt 2016

Newport Beach, California; April 5–10, 2016

Outline

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

Current version of slides (April 11, 2016, 16 : 06):
http://vlado.fmf.uni-lj.si/pub/slides/vbMAG16.pdf

Microsoft Academic Graph

The Microsoft Academic Graph (MAG) is a heterogeneous graph containing scientific publication records, citation relationships between those publications, as well as authors, institutions, journals and conference "venues" and fields of study. The first version was published on June 5, 2015; the last updated version is from February 5, 2016.
Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, and Kuansan Wang, An Overview of Microsoft Academic Service (MAS) and Applications, WWW – World Wide Web Consortium (W3C), 18 May 2015.

MAG – entities and sizes

Entity name	Entity Count
Papers	> 83 million
Authors	> 20 million
Institutions	$> 770,000$
Journals	$> 22,000$
Conference series	> 900
Conference instances	$> 26,000$
Fields of study	$> 50,000$

The ZIP containing all data files has size 28.2 GB.
Searching, machine learning, recomendation tasks.

MAG – data files structure

```
Affiliations
1  Affiliation ID
2  Affiliation name

Authors
1  Author ID
2  Author name

FieldsOfStudy
1  Field of study ID
2  Field of study name

FieldOfStudyHierarchy
1  Child field of study ID
2  Child field of study level
3  Parent field of study ID
4  Parent field of study level
5  Confidence
```

```
ConferenceSeries
1  Conference series ID
2  Short name (abbreviation)
3  Full name

ConferenceInstances
1  Conference series ID
2  Conference instance ID
3  Short name (abbreviation)
4  Full name
5  Location
6  Official conference URL
7  Conference start date
8  Conference end date
9  Conference abstract registration date
10 Conference submission deadline date
11 Conference notification due date
12 Conference final version due date
```

MAG – data files structure

```
Papers
1  Paper ID
2  Original paper title
3  Normalized paper title
4  Paper publish year
5  Paper publish date
6  Paper Document Object Identifier
   (DOI)
7  Original venue name
8  Normalized venue name
9  Journal ID mapped to venue name
10 Conference series ID
   mapped to venue name
11 Paper rank

PaperKeywords
1  Paper ID
2  Keyword name
3  Field of study ID mapped to keyword
```

```
PaperAuthorAffiliations
1  Paper ID
2  Author ID
3  Affiliation ID
4  Original affiliation name
5  Normalized affiliation name
6  Author sequence number

PaperReferences
1  Paper ID
2  Paper reference ID

PaperUrls
1  Paper ID
2  URL

Journals
1  Journal ID
2  Journal name
```

MAG into a collection of networks

MAG is similar to data from bibliographic data bases (Web of Science, Scopus, DBLP, ZB Math, etc.).
In our paper On bibliographic networks we proposed to transform such data into a collection of one-mode and two-mode networks – in the case of MAG into:

Cite, **WA**, **WK**, **WV**, **AC**,
where: **W** – works (papers, books, etc.), **A** – authors, **K** – keywords, **V** – venues (conferences, journals, publishers), **C** - companies or institutions, **F** - field.

and some properties of nodes:

year – publication year of a work.

An important fact about these networks is that many pairs share a common set – using the network multiplication we can get *derived* networks.

Problems

- the networks obtained from the complete MAG are very large and require substantial time for construction and analysis. We decided:
  - to limit in the first phase the analysis to some smaller subset of data on which the analyses can be performed fast.
  - to explore the data an see what are the problems
  - to identify problems and develop solutions.
- transforming and cleaning the data
  - identifying problems
  - missing "standard" bibliographic data such as Volume and First page.

We selected as the subset the data related to SNA. Extraction was done by Juergen Pfeffer.

MAG/SNA – sizes

W – works (papers, books, etc.)	634552
A – authors	1048433
K – keywords	24535
V – venues (conferences, journals, publishers)	
C – companies or institutions	
F – field	

Cleaning

Authors.txt

```
7CF2C714    joe hennessey
7CF2C724    p w brown
7CF2C758    r kumar
7CF2C83F    j decastro
7CF2C87B    a g bonchosmolovskii
7CF2C8A2    n romano
7CF2C90F    r kh makhmudov
7CF2C920    正義 上原
7CF2C980    h m ruijter
7CF2CA3D    eszter k vladar
7CF2CAE3    isis brook
7CF2CB1B    jeremy cardin
7CF2CB96    jose angel calderon
7CF2CBF3    blachowski stefan
7CF2CC11    mitsuru shinoda
7CF2CC32    真生 飯山
7CF2CC58    federico herrera
7CF2CCA7    jones paul r
7CF2CD18    博之 伊藤
7CF2CD1C    л в карпенко
7CF2CD29    長谷川
7CF2CD8B    cristieli sergio de menezes oliveira
7CF2CDAD    炳滿 田
7CF2CE5B    adir j perez
7CF2CEE6    doris r schwartz
7CF2D0AE    irene bartsch
7CF2D0B4    chong ren
7CF2D0D8    نیا صفاری مجید
7CF2D1A9    reno camilleri
7CF2D1BD    kresz bierut
7CF2D1C8    石原
7CF2D202    carrol a alvarez
```

The networks are too large to do individual cleaning in general. We can identify some problems that can be corrected using (short) programs. For example, the same author appears several times in the list of authors – the *identity problem*.

We produced a partition that puts all authors with the same name into the same class. The application of it to shrink the set of authors can be risky – in MathSciNet there exist 697 chinese mathematicians with the name

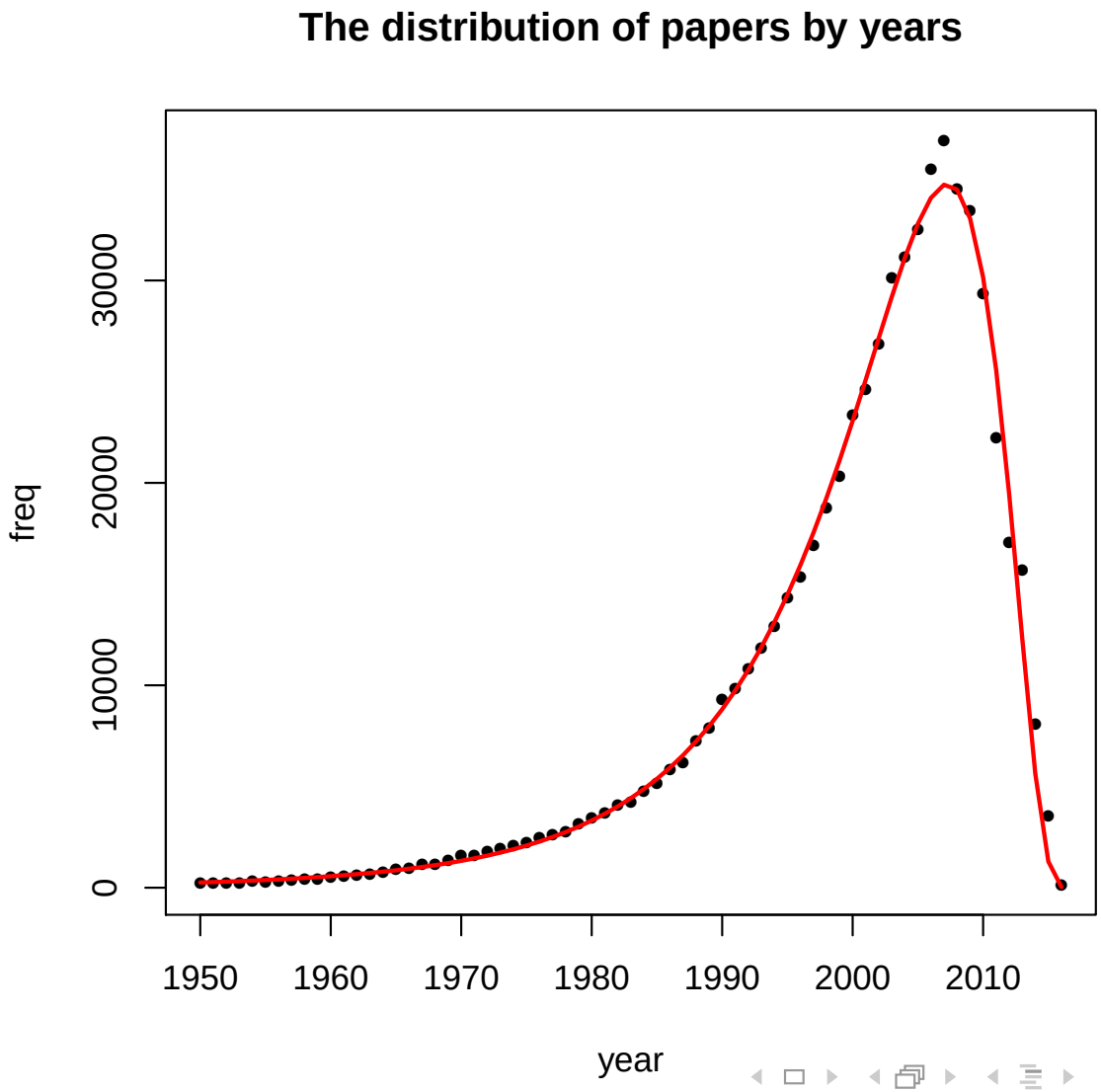
Wang, Li.

MAG – entities and sizes

Another such partition is the partition DOI the puts into the same class all works with the same DOI. In this case it is reasonable to assume that they identify the same work.

In general we treat the remaining inconsistencies in data as a noise. If they show up also in results we correct the data in an appropriate way and repeat the analysis.

MAG/SNA – The distribution of papers by years



MAG/SNA – The distribution of papers by years


```
> setwd("c:/users/Batagelj/work/Python/MAG")
> years <- read.table(file="Year.clu",header=FALSE,skip=2)$V1
> t <- table(years)
> min(years)
[1] 1803
> max(years)
[1] 2016
> year <- as.integer(names(t))
> freq <- as.vector(t[1950<=year & year<=2016])
> y <- 1950:2016
> model <- nls(freq~c*dlnorm(2017-y,a,b),start=list(c=500000,a=2.5,b=0.7
> model
Nonlinear regression model
  model: freq ~ c * dlnorm(2017 - y, a, b)
   data: parent.frame()
        c         a         b
6.317e+05 2.655e+00 6.164e-01
 residual sum-of-squares: 51166952

Number of iterations to convergence: 6
Achieved convergence tolerance: 9.371e-06
> plot(y,freq,pch=16,cex=0.75,main="The distribution of papers by years"
+ xlab="year",ylab="freq")
> lines(y,predict(model,list(x=2017-y)),col='red',lw=2)
```

WK – keywords with the largest indegree

rank	freq	Id
1	24104	Social network
2	10349	Network analysis
3	9726	internet
4	8974	genetics
5	8921	bioinformatics
6	8919	computer model
7	8203	Flow network
8	8094	developing countries
9	8066	computer network
10	7688	mathematical model
11	7359	Network model
12	7240	neural network model
13	7043	algorithms
14	6741	human factors
15	6257	indexing terms
16	6232	biomedical research
17	6140	occupational safety
18	6036	signal transduction
19	5939	injury prevention
20	5937	suicide prevention
21	5736	research methodology
22	5310	biological sciences
23	5303	higher education
24	5138	medicine
25	5128	data mining

WK – outdegree distribution

d	f	d	f	d	f	d	f
0	185261	25	144	50	93	75	4
1	82195	26	109	51	94	76	5
2	69677	27	106	52	62	77	2
3	61829	28	72	53	44	78	2
4	54083	29	51	54	39	79	3
5	43478	30	47	55	24	80	1
6	34880	31	45	56	19	82	2
7	27853	32	49	57	15	83	1
8	22266	33	36	58	16	85	1
9	17855	34	27	59	14	86	1
10	14140	35	51	60	8	88	1
11	10905	36	41	61	16	92	1
12	8480	37	31	62	11	93	1
13	6465	38	37	63	14	100	1
14	4975	39	31	64	14	102	1
15	3397	40	44	65	11	106	1
16	2325	41	112	66	6	110	1
17	1739	42	258	67	6	112	1
18	1104	43	377	68	6		
19	789	44	337	69	5		
20	510	45	339	70	4		
21	419	46	304	71	3		
22	268	47	232	72	3		
23	233	48	187	73	2		
24	171	49	162	74	2		

Derived network **AK**

$$\mathbf{AK} = \mathbf{WA}^T * \mathbf{WK}$$

ak_{ak} = number of works authored by the author a taged by the keyword k

In the following picture we present the link-cut in **AK** at level 40 – we preserve only links with value at least 40.

Other possibilities: collaboration network

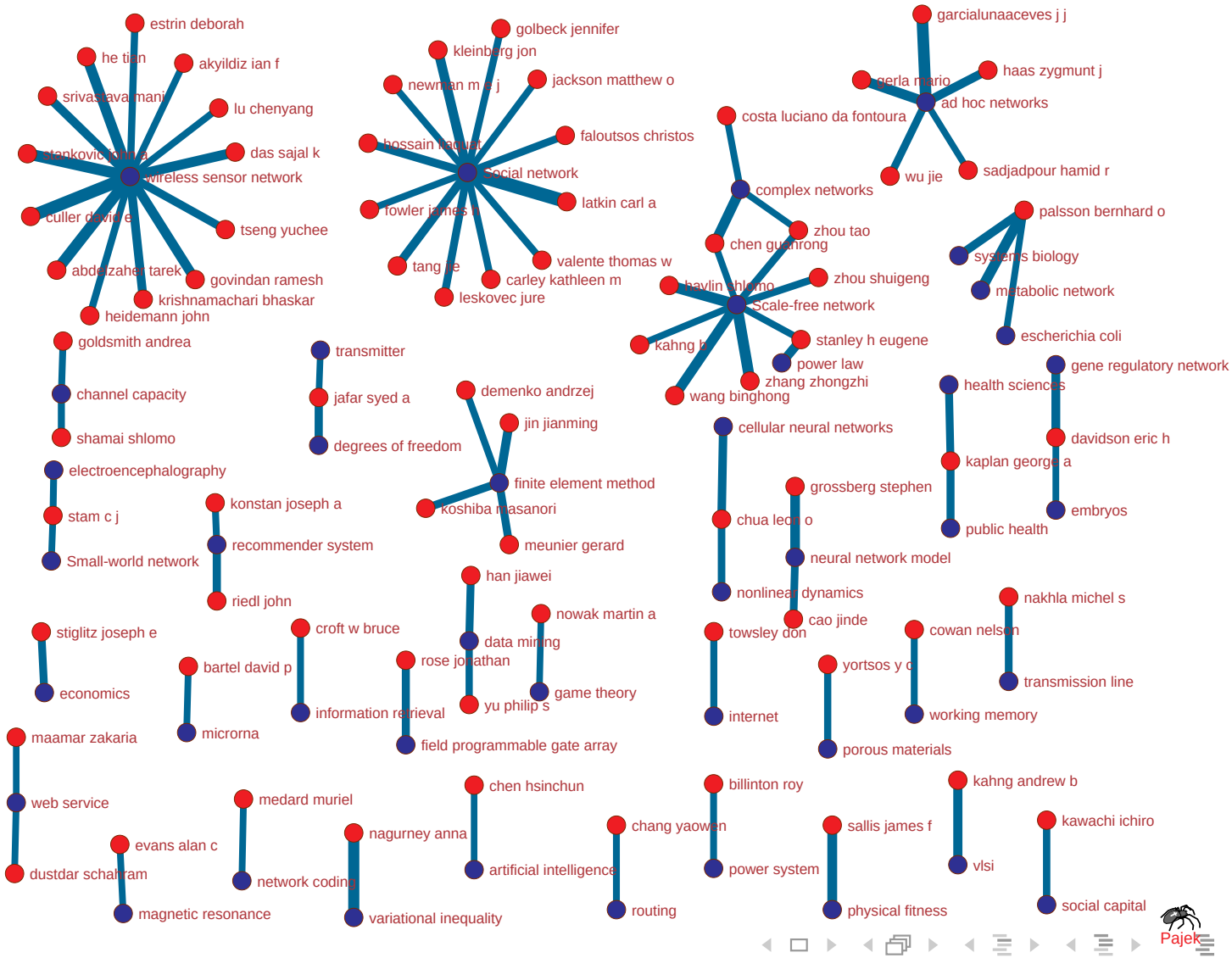
$$\mathbf{AA}_W = \mathbf{WA}^T * \mathbf{WA}$$

co-taging network

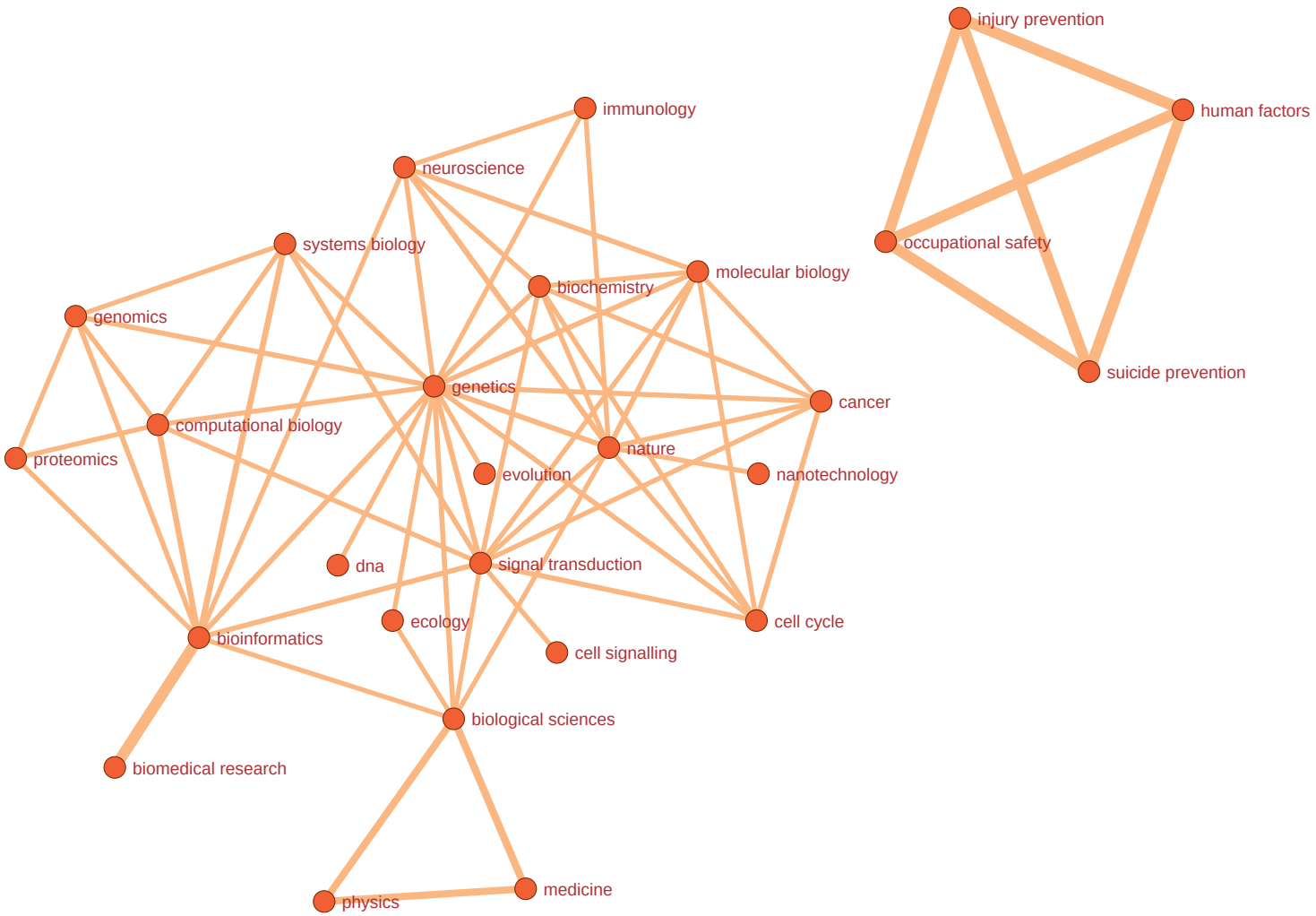
$$\mathbf{KK}_W = \mathbf{WK}^T * \mathbf{WK}$$

Problem with nodes of large degree – contributing large complete subgraphs (overrepresented). The solution is to use the fractional approach.

AK link cut at level 40



KK link cut at level 2500



AA link cut at level 40

Cite citation network

By its nature: citing work is usually citing an older work, a citation network is usually almost acyclic. In acyclic networks we can compute the importance of arcs using Hummon-Doreian's SPC weights.

A first analysis of **Cite**/SNA network revealed some quite large strong components – there are some inconsistent arcs. In general, it is very hard to detect them. But in MAG we have a publication year for each work. This allows us to split the set of arcs to the set of *inconsistent* arcs (year(citing work) $<$ year(cited work)) and *consistent* arcs (year(citing work) $\geq$ year(cited work)). The set of consistent arcs still contains some very small strong components that we remove using the preprint transformation. In this subnetwork we compute the SPC weights and analyze it.

Cite nodes with largest indegree – the most cited

rank	ideg	Id	title
1	2733	7DE3F24E	1998:Collective dynamics of 'small-world' network
2	1778	7DFD00FF	1998:Collective dynamics of |[lsquo]|small-world|
3	1591	5F4231F7	1999:Emergence of scaling in random networks
4	1522	7AE62067	1994:Social network analysis : methods and applic
5	1518	801120F4	2006:The Structure and Function of Complex Networ
6	1111	7A9A7CE3	1978:Centrality in social networks conceptual cla
7	938	7EA36534	2001:Statistical mechanics of complex networks
8	917	7C4E1302	1988:Social capital in the creation of human capi
9	807	7EFAA2E1	2003:Birds of a Feather: Homophily in Social Netw
10	601	5F4C44DB	1985:Network Externalities, Competition, and Comp
11	562	5DCAEA41	1967:The small world problem
12	555	7AE8C51A	1992:Structural Holes: The Social Structure of Co
13	554	7CE3A440	2003:Finding and evaluating community structure i
14	548	758182E5	2002:Community Structure in Social and Biological
15	479	78201C0E	1991:Social network analysis : a handbook
16	475	7FD85A5E	2000:The large-scale organization of metabolic ne
17	474	08F73288	2002:Ucinet for Windows: Software for Social Netw
18	440	805DB3F6	2002:Network Motifs: Simple Building Blocks of Co
19	412	797C66A2	2001:Epidemic spreading in scale-free networks
20	409	074A990C	1999:Diameter of the World Wide Web
21	408	7672CE5D	1983:THE STRENGTH OF WEAK TIES: A NETWORK THEORY
22	405	7F014945	2001:Lethality and centrality in protein networks
23	399	7AE4E1EC	2003:Maximizing the spread of influence through a
24	392	0E9A2F6A	1993:Social Network Analysis
25	381	7B21241E	2000:Error and attack tolerance of complex networ

Cite CPM main path

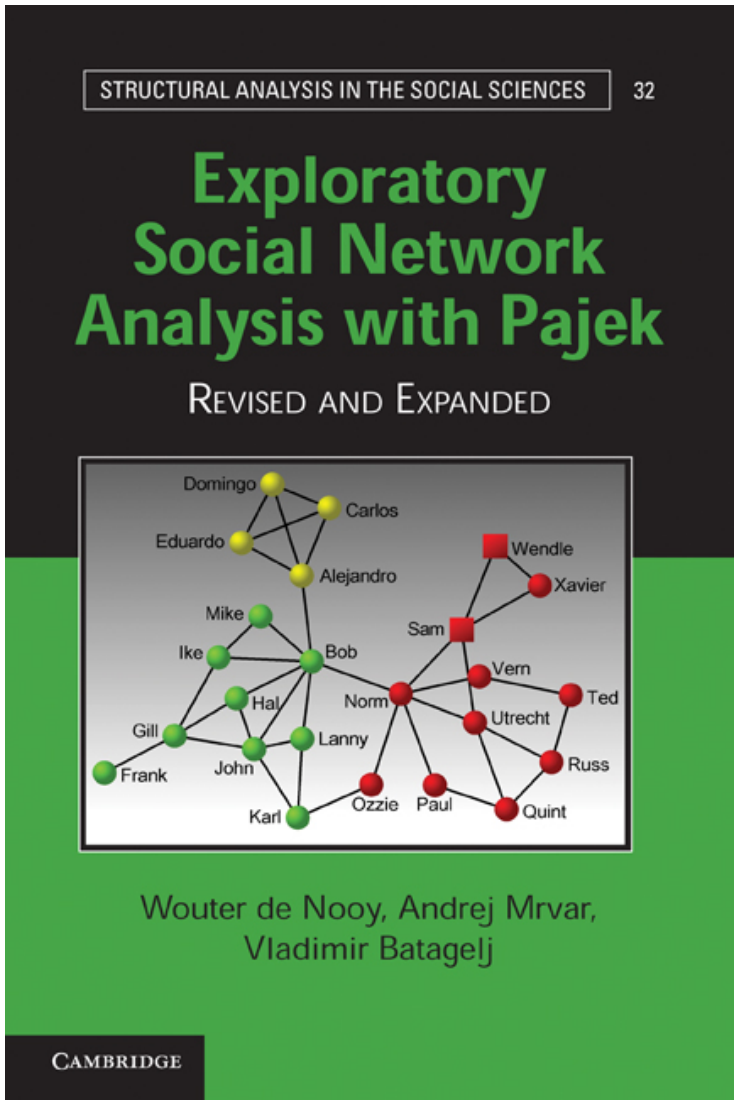
Works on the CPM main path

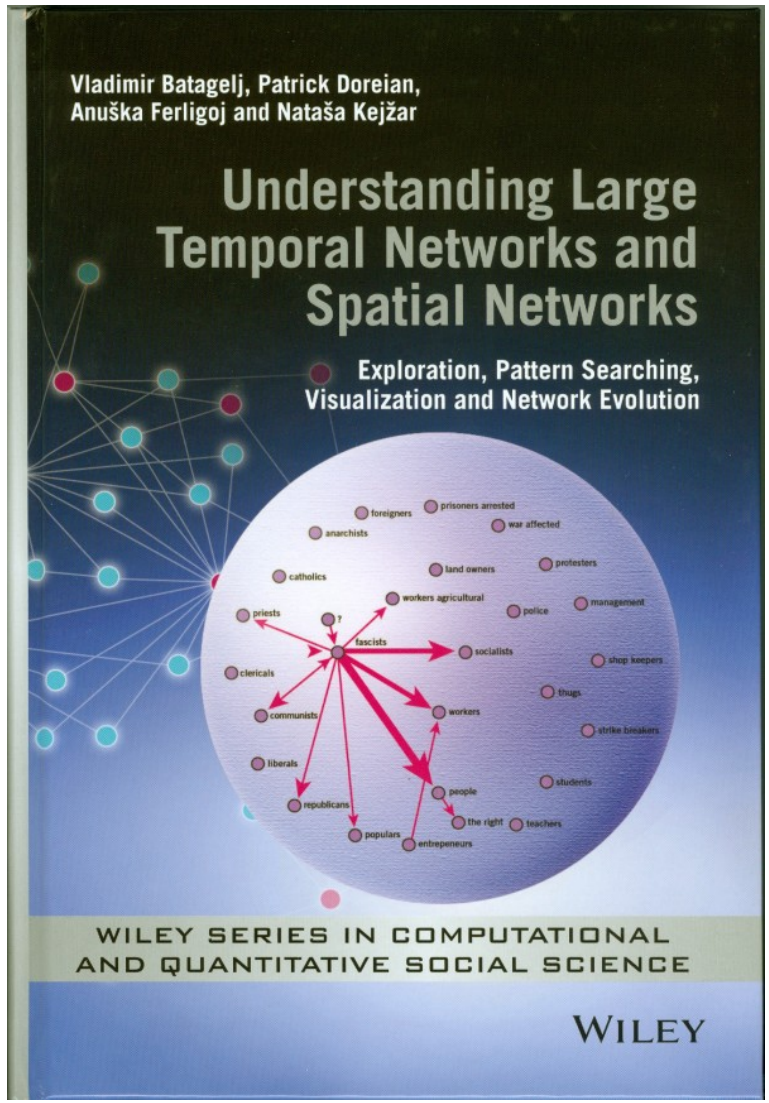
```
7180C1D6  2016:Influence maximization in social networks under an independent cascade-based mode
796367FF  2015:A fast algorithm for finding most influential people based on the linear threshol
7BD90FAA  2014:Conformity-aware influence maximization in online social networks
7A0295DE  2013:Confluence: conformity influence in large social networks
074F8859  2013:Mining structural hole spanners through information diffusion in social networks
76E3785A  2013:Learning to predict reciprocity and triadic closure in social networks
7892819F  2012:Inferring social ties across heterogenous networks
807589F1  2011:Who will follow you back?: reciprocal relationship prediction
7D3DB51F  2010:What is Twitter, a social network or a news media?
7E35209C  2009:Characterizing user behavior in online social networks
80574CC0  2009:On the evolution of user interaction in Facebook
7A09829C  2009:User interactions in social networks and their implications
7EA5C7A7  2008:Comparison of online social relations in volume vs interaction: a case study of c
7DFD6839  2008:Planetary-Scale Views on an Instant-Messaging Network
7FA740C8  2008:Yes, there is a correlation: - from social networks to personal behavior on the w
7CEFD341  2007:Model-based clustering for social networks
7F4E4D82  2007:Recent developments in exponential random graph ( p *) models for social networks
80C31505  2007:An introduction to exponential random graph ( p * ) models for social networks
7F5B174D  2006:NEW SPECIFICATIONS FOR EXPONENTIAL RANDOM GRAPH MODELS
801120F4  2006:The Structure and Function of Complex Networks
7B58E93A  2001:The risk environment for HIV transmission: results from the Atlanta and Flagstaff
78866E79  2000:The Atlanta Urban Networks Study: a blueprint for endemic transmission
78687B67  1998:Social network dynamics and HIV transmission
7C05C659  1995:Choosing a centrality measure: Epidemiologic correlates in the Colorado Springs s
79B75E43  1994:Social networks and infectious disease: the Colorado Springs Study
7AD762F3  1985:Social networks and the spread of infectious diseases: The AIDS example
75D61FE9  1980:Social networks: A promising direction for research on the relationship of the so
7D317928  1978:Social Networks and Schizophrenia*
```

Conclusions

- add the networks **WV**, **WF** and **AC** and analyze them
- Fractional analysis of **KK** and **AA**
- Find a good (content based) identifier for works and analyze **Cite** using main multi-paths and islands
- repeat the analyses on MAG

Support





References I


Vladimir Batagelj: WoS2Pajek


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.


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.


Wikipedia: Peer review