

## Visualization of large networks and Pajek

Vladimir Batagelj

University of Ljubljana  
Slovenia

**Workshop on Complex objects visualization**

November 16-19, 2005, Koper, Slovenia

# Outline

1	Networks in social sciences	1
3	Networks	3
5	Types of networks	5
10	Large Networks	10
11	<b>Pajek</b>	11
14	Representations of properties	14
15	Analysis and Visualization	15
26	Large networks	26
31	Bipartite cores	31
35	Directed 4-rings	35
41	Dense networks	41
45	New graphical elements	45
47	Dynamic/temporal networks	47
52	Challenges	52

## Networks in social sciences

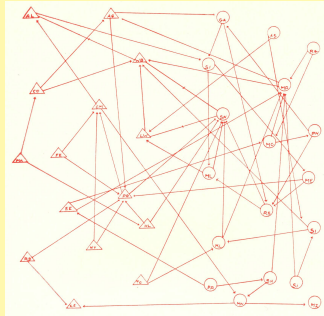


Moreno

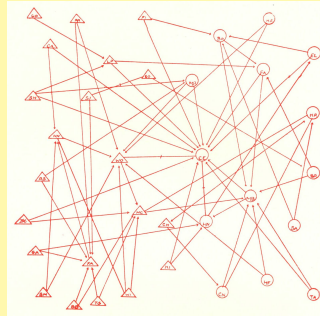
The use of networks was introduced in sociology by Moreno developing the sociometry (1934, 1953, 1960). An overview of visualization of social networks was prepared by [Lin Freeman](#) (1, 2).

# Moreno: Who shall survive?

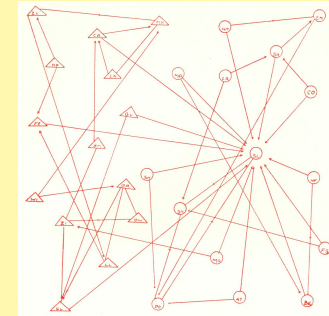
K:



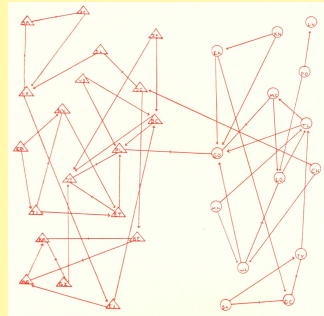
1:



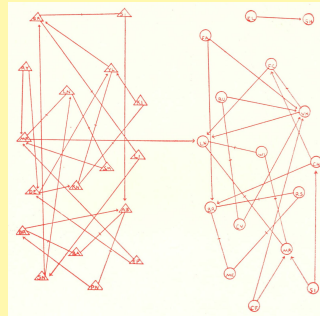
2:



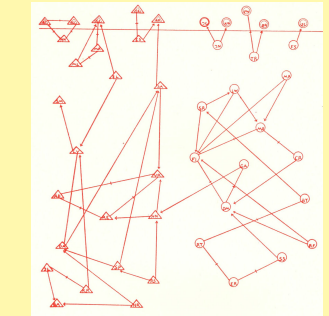
3:



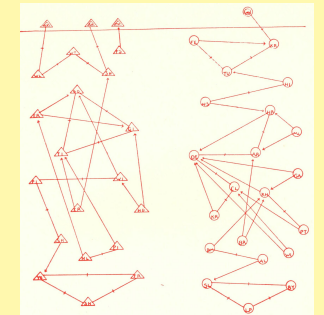
4:



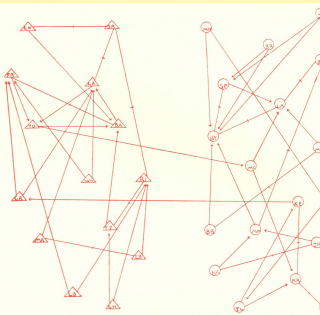
5:



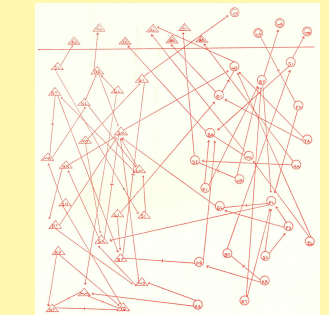
6:



7:



8:

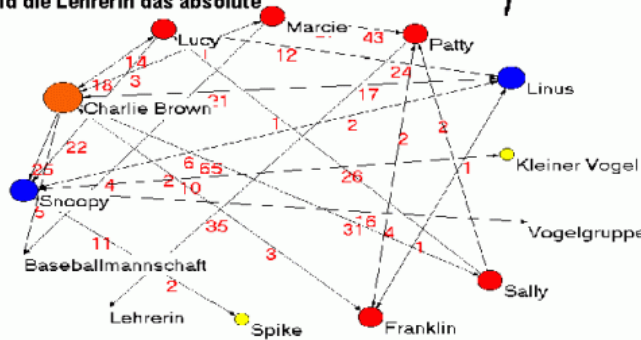




## Networks

Aber damit geben sich Soziologen noch lange nicht zufrieden. Sie wollen zum Beispiel noch wissen, wer ist hier wichtig in diesem Netzwerk?

Dabei gibt es in der Art des Wichtig-Seins Unterschiede. Beispielsweise kann man fragen, wer ist am besten eingebunden, wer hat die meisten Beziehungen? Hier wurde gemessen, wer die meisten Beziehungen zu anderen aufgenommen hat, und da ist Charlie Spitzenreiter, während die Vogelgruppe, die Baseballmannschaft und die Lehrerin das absolute Schlußlicht bilden



Alexandra Schuler/ Marion Laging-Glaser:

Analyse von Snoopy Comics

A *network* is based on two sets – set of *vertices* (nodes), that represent the selected *units*, and set of *lines* (links, ties), that represent *relations* between units. They determine a *graph*. A line can be *directed* – an *arc*, or *undirected* – an *edge*.

Additional data about vertices or lines can be known – their *properties* (attributes). For example: name/label, type, value, ...

# Network = Graph + Data

The data can be measured or computed.

## Networks / Formally

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

- a *graph*  $\mathcal{G} = (\mathcal{V}, \mathcal{L})$ , where  $\mathcal{V}$  is the set of vertices,  $\mathcal{A}$  is the set of arcs,  $\mathcal{E}$  is the set of edges, and  $\mathcal{L} = \mathcal{E} \cup \mathcal{A}$  is the set of lines.  $n = \text{card}(\mathcal{V})$ ,  $m = \text{card}(\mathcal{L})$
- $\mathcal{P}$  *vertex value functions* / properties:  $p : \mathcal{V} \rightarrow A$
- $\mathcal{W}$  *line value functions* / weights:  $w : \mathcal{L} \rightarrow B$

## Types of networks

Besides ordinary (directed, undirected, mixed) networks some extended types of networks are also used:

- *2-mode networks*, bipartite (valued) graphs – networks between two disjoint sets of vertices.
- *multi-relational networks*.
- *temporal networks*, dynamic graphs – networks changing over time.
- specialized networks: representation of genealogies as *p-graphs*; *Petri's nets*, ...

The network (input) file formats should provide means to express all these types of networks. All interesting data should be recorded (respecting privacy).

## Deep South



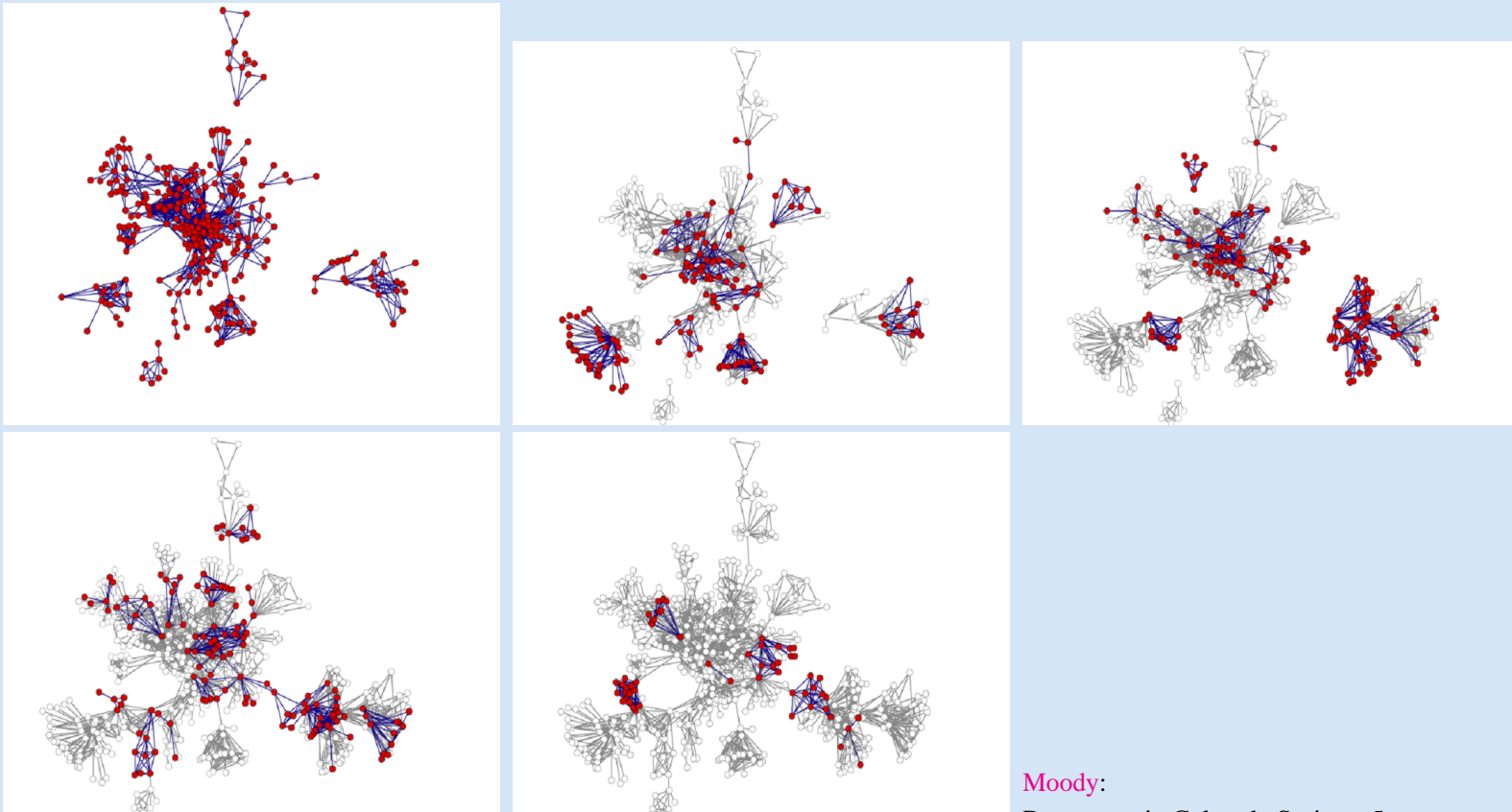
Classical example of two-mode network are Southern women (Davis 1941).

Davis.paj. Freeman's overview.

NAMES OF PARTICIPANTS OF GROUP I	CODE NUMBERS AND DATES OF SOCIAL EVENTS REPORTED IN <i>Old City Herald</i>													
	(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	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	X							
6. Miss Frances Anderson.....			X		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		X		
12. Miss Katherine Rogers.....							X	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	X	X				
17. Mrs. Olivia Carleton.....								X	X	X		X		
18. Mrs. Flora Price.....								X	X	X		X		

## Temporal networks

In a *temporal network* the presence/activity of vertex/line can change through time. Two types of descriptions of temporal networks are used – based on *presence* and based on *events*.



Moody:

Drug users in Colorado Springs, 5 years



## Multi-relational temporal network – KEDS/WEIS

```

% Recoded by WEISmonths, Sun Nov 28 21:57:00 2004
% from http://www.ku.edu/~keds/data.dir/balk.html
*vertices 325
1 "AFG" [1-*]
2 "AFR" [1-*]
3 "ALB" [1-*]
4 "ALBMED" [1-*]
5 "ALG" [1-*]
...
318 "YUGGOV" [1-*]
319 "YUGMAC" [1-*]
320 "YUGMED" [1-*]
321 "YUGMTN" [1-*]
322 "YUGSER" [1-*]
323 "ZAI" [1-*]
324 "ZAM" [1-*]
325 "ZIM" [1-*]
*arcs :0 "*** ABANDONED"
*arcs :10 "YIELD"
*arcs :11 "SURRENDER"
*arcs :12 "RETREAT"
...
*arcs :223 "MIL ENGAGEMENT"
*arcs :224 "RIOT"
*arcs :225 "ASSASSINATE TORTURE"
*arcs
224: 314 153 1 [4]           890402 YUG      KSV      224 (RIOT) RIOT-TORN
212: 314 83 1 [4]           890404 YUG      ETHALB  212 (ARREST PERSON) ALB ETHNIC JAILED IN YUG
224: 3 83 1 [4]             890407 ALB      ETHALB  224 (RIOT) RIOTS
123: 83 153 1 [4]          890408 ETHALB  KSV      123 (INVESTIGATE)  PROBING
...
42: 105 63 1 [175]         030731 GER      CYP      042 (ENDORSE)      GAVE SUPPORT
212: 295 35 1 [175]        030731 UNWCT   BOSSER  212 (ARREST PERSON) SENTENCED TO PRISON
43: 306 87 1 [175]        030731 VAT      EUR      043 (RALLY) RALLIED
13: 295 35 1 [175]        030731 UNWCT   BOSSER  013 (RETRACT)     CLEARED
121: 295 22 1 [175]       030731 UNWCT   BAL      121 (CRITICIZE)   CHARGES
122: 246 295 1 [175]     030731 SER      UNWCT   122 (DENIGRATE)   TESTIFIED
121: 35 295 1 [175]      030731 BOSSER  UNWCT   121 (CRITICIZE)   ACCUSED

```

Kansas Event Data System *KEDS*

## Size of Networks

The size of a network/graph is expressed by two numbers: number of vertices  $n = |\mathcal{V}|$  and number of lines  $m = |\mathcal{L}|$ .

In a *simple undirected* graph (no parallel edges, no loops)  $m \leq \frac{1}{2}n(n-1)$ ; and in a *simple directed* graph (no parallel arcs)  $m \leq n^2$ . The quotient  $\gamma = \frac{m}{m_{max}}$  is a *density* of graph. In large networks more intuitive density measure is the *average degree*  $\bar{d} = \frac{1}{n} \sum_{v \in V} deg(v) = \frac{2m}{n}$ .

*Small* networks (some tens vertices) – can be represented by a picture and analyzed by many algorithms (*UCINET*, *NetMiner*).

Also *middle size* networks (some hundreds vertices) can still be represented by a picture (!?), but some analytical procedures can't be used.

Till 1990 most networks were small – they were collected by researchers using surveys, observations, archival records, ... The advances in IT allowed to create networks from the data already available in the computer(s). Large networks became reality.

## Large Networks

*Large* network – several thousands or millions of vertices. Can be stored in computer's memory – otherwise *huge* network.

Usually sparse  $m \ll n^2$ ; typical:  $m = O(n)$  or  $m = O(n \log n)$ .

Examples:

network	size	$n =  V $	$m =  L $	source
ODLIS dictionary	61K	2909	18419	ODLIS online
Citations SOM	168K	4470	12731	Garfield's collection
Molecula 1ATN	74K	5020	5128	Brookhaven PDB
Comput. geometry	140K	7343	11898	BiBTeX bibliographies
English words 2-8	520K	52652	89038	Knuth's English words
Internet traceroutes	1.7M	124651	207214	Internet Mapping Project
Franklin genealogy	12M	203909	195650	RoperId.com gedcoms
World-Wide-Web	3.6M	325729	1497135	Notre Dame Networks
Actors	3.9M	392400	1342595	Notre Dame Networks
US patents	82M	3774768	16522438	Nber
SI internet	38M	5547916	62259968	Najdi Si

## Pajek

Large networks are too big to be displayed in details; special algorithms are needed for their analysis.



**Pajek** is a program, for Windows, for analysis and visualization of *large networks* having some ten or hundred of thousands of vertices.

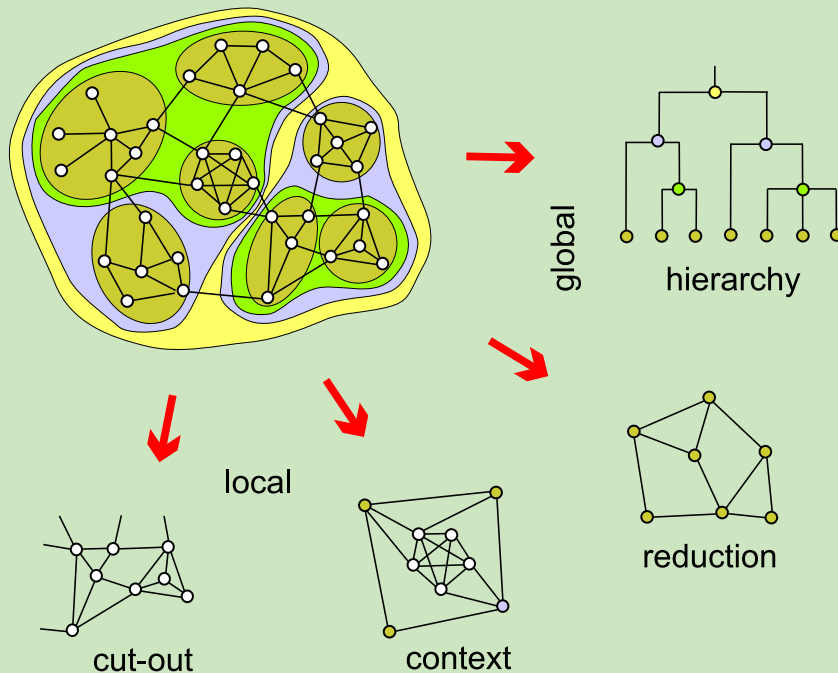
In Slovenian language *pajek* means spider.

The latest version of **Pajek** is freely available, for noncommercial use, at its home page:

<http://vlado.fmf.uni-lj.si/pub/networks/pajek/>

Book: W. de Nooy, A. Mrvar, V. Batagelj: *Exploratory Social Network Analysis with Pajek*, CUP, 2005. [Amazon](#). [ESNA page](#).

## Main design goals



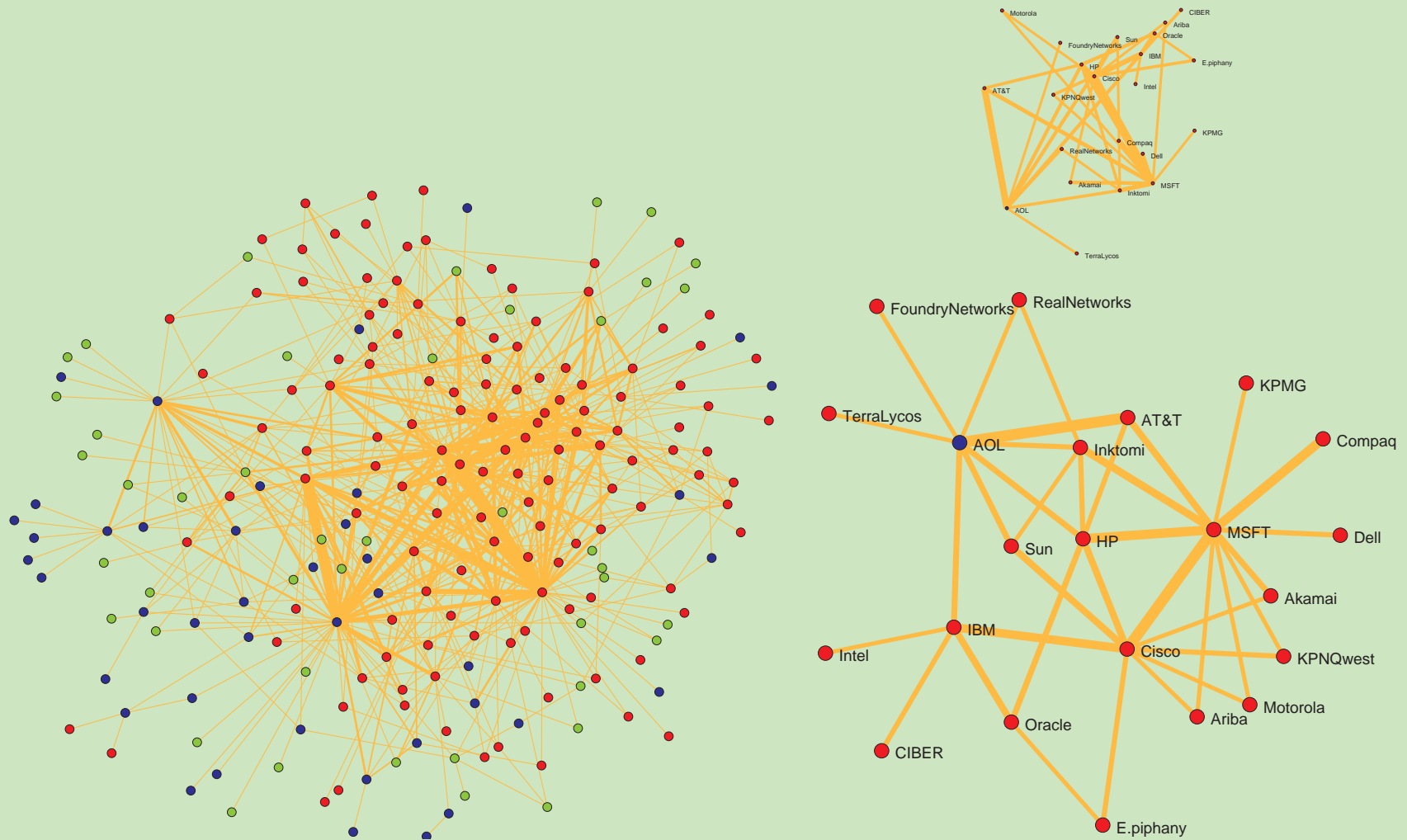
The main goals in the design of **Pajek** are:

- to support abstraction by (recursive) *decomposition* of a large network into several smaller networks that can be treated further using more sophisticated methods;
- to provide the user with some powerful *visualization* tools;
- to implement a selection of efficient *subquadratic* algorithms for analysis of large networks.

With **Pajek** we can: *find* clusters (components, neighbourhoods of ‘important’ vertices, cores, etc.) or patterns (motifs) in a network, *extract* vertices that belong to the same clusters and *show* them separately, possibly with the parts of the context (detailed local view), *shrink* vertices in clusters and show relations among clusters (global view).



# Line-cut: Krebs Internet Industries, $w_3 \geq 5$



## Representations of properties

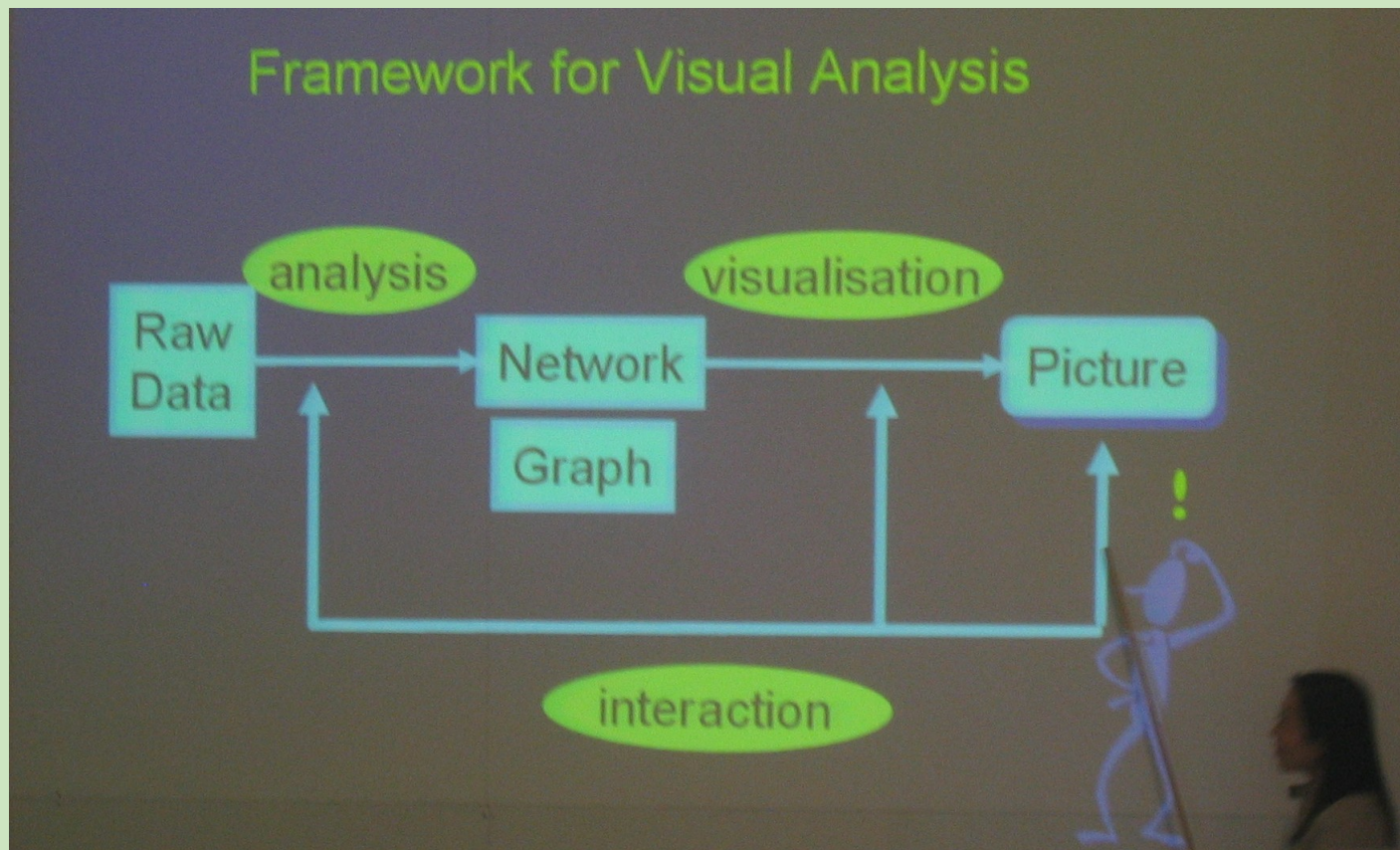
*Properties* of vertices  $\mathcal{P}$  and lines  $\mathcal{W}$  can be measured in different scales: numerical, ordinal and nominal. They can be *input* as data or *computed* from the network.

In **Pajek** numerical properties of vertices are represented by *vectors*, nominal properties by *partitions* or as *labels* of vertices. Numerical property can be displayed as *size* of vertex (figure), as its *coordinate*; and a nominal property as *color* or *shape* of the figure, or as a vertex *label*.

We can assign in **Pajek** numerical values to links. They can be displayed as *value*, *thickness* or *grey level*. Nominal values can be assigned as *label*, *color* or *line pattern* (see **Pajek manual**, section 4.3).

## Analysis and Visualization

The network visualization is an iterative process that combines analysis and creation of layouts.



## Some comments

While the technical graph drawing problems could ask for a single 'the best' picture, the social network analysis is a part of data analysis. Its goal is to get insight into the structure and characteristics of a given network, but also how it influences related social processes.

**What is the goal:** exploration of network data (*layout editor*), presentation of the obtained results (*layout viewer*), ...?

**What is a GD result:** picture or layout ?

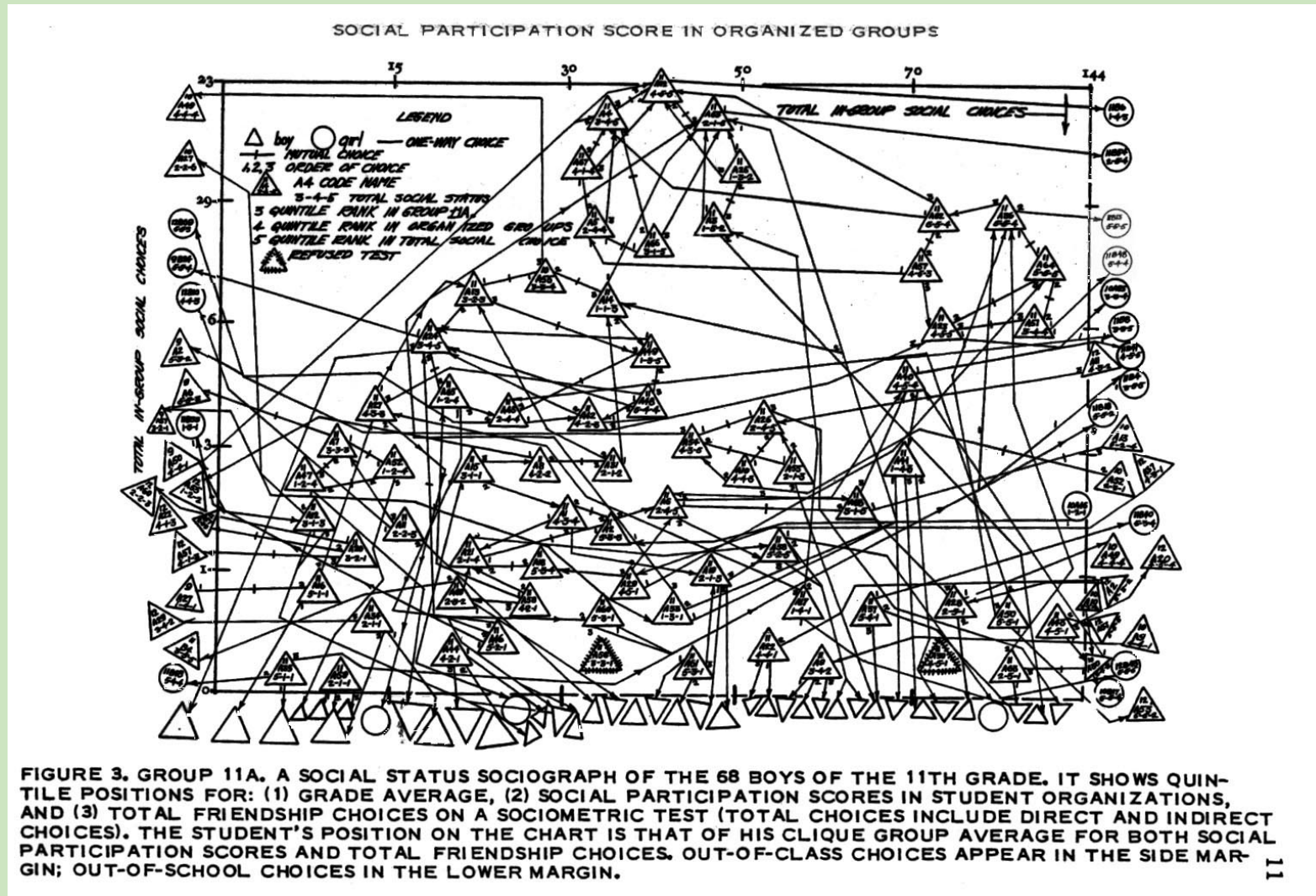
**What is the medium of the result:** static picture on 'paper', interactive layout, ...?

**What kind of user** will use the result: simple, advanced, ...?

Most methods are 'paper' oriented. In larger/denser networks there is often too much information to be presented at once. A possible answer are *interactive layouts* where the user controls what (s)he wants to see.



# Sociograph



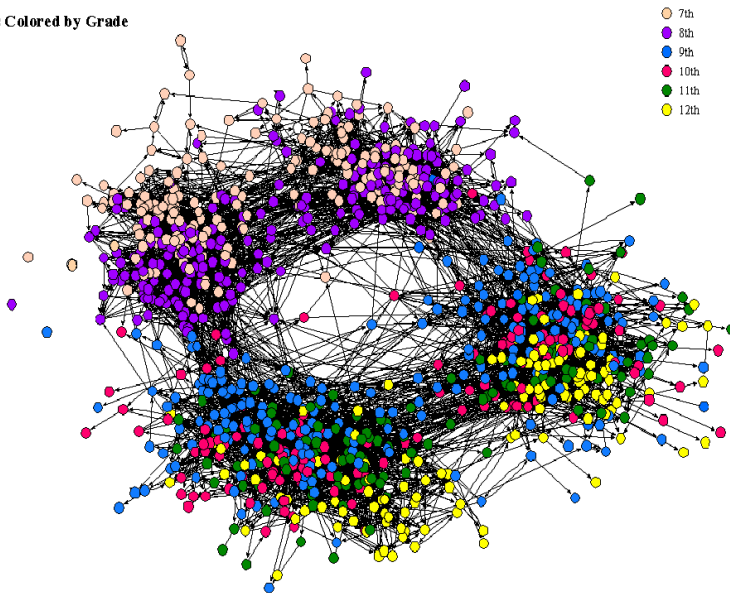
11



## James Moody: Display of properties – school

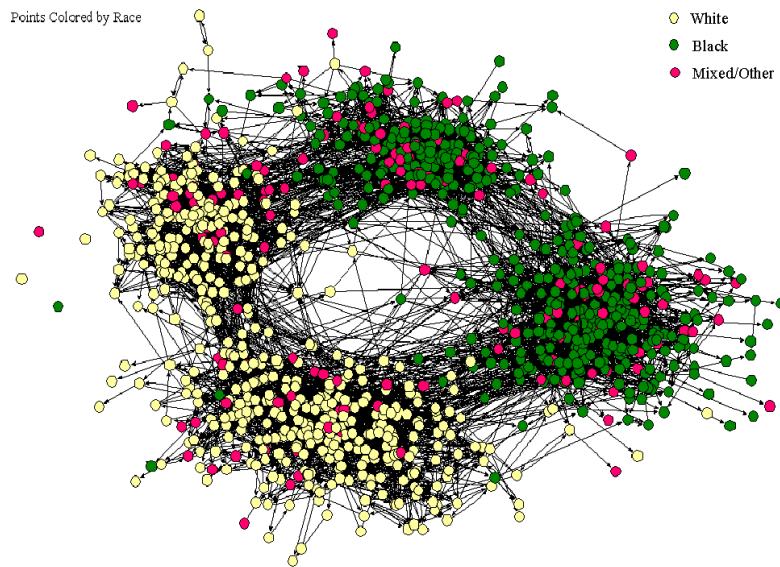
The Social Structure of “Countryside” School District

Points Colored by Grade

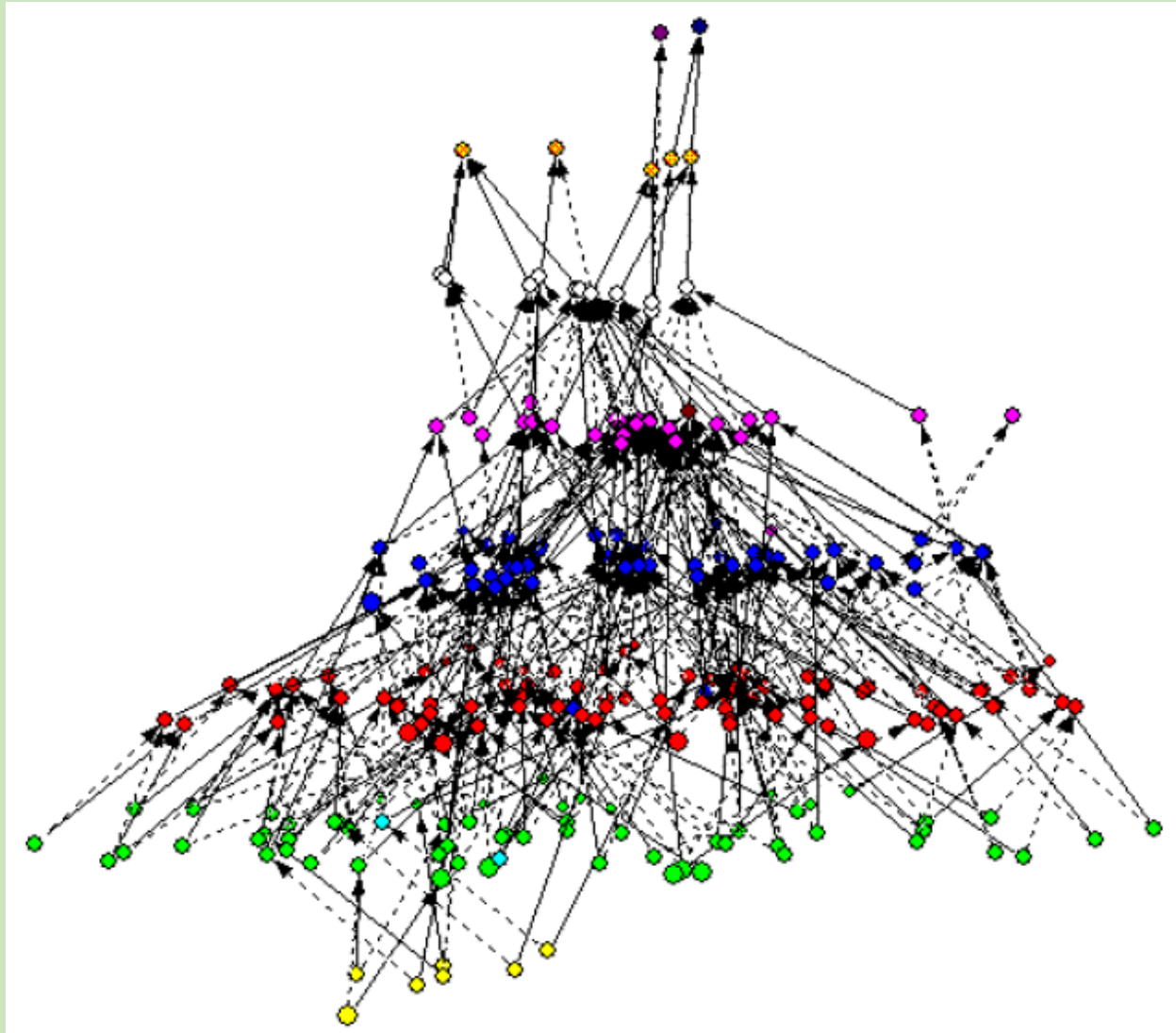


The Social Structure of “Countryside” School District

Points Colored by Race



## Douglas White: relinking marriages among the Aydinli

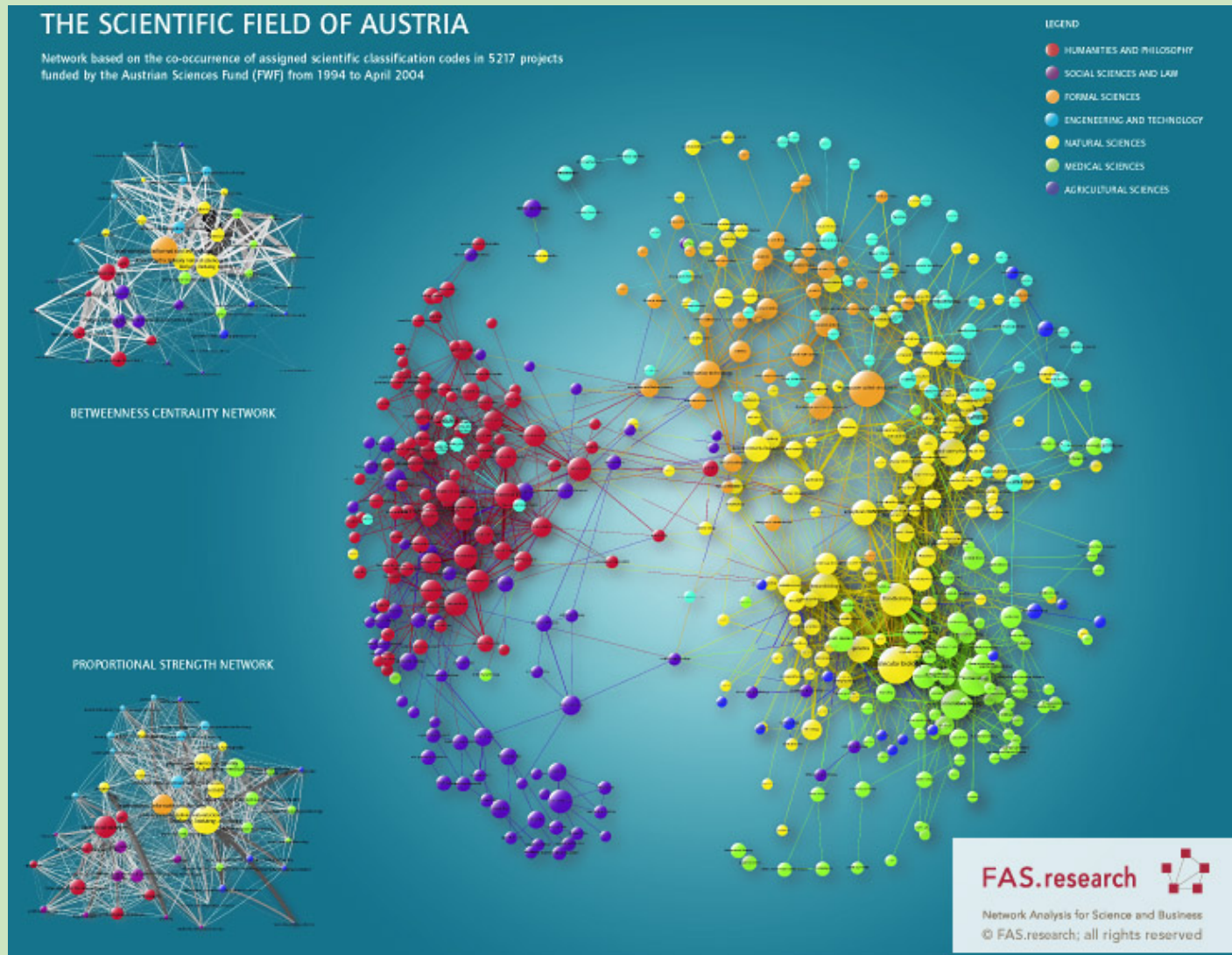


# Lothar Krempel

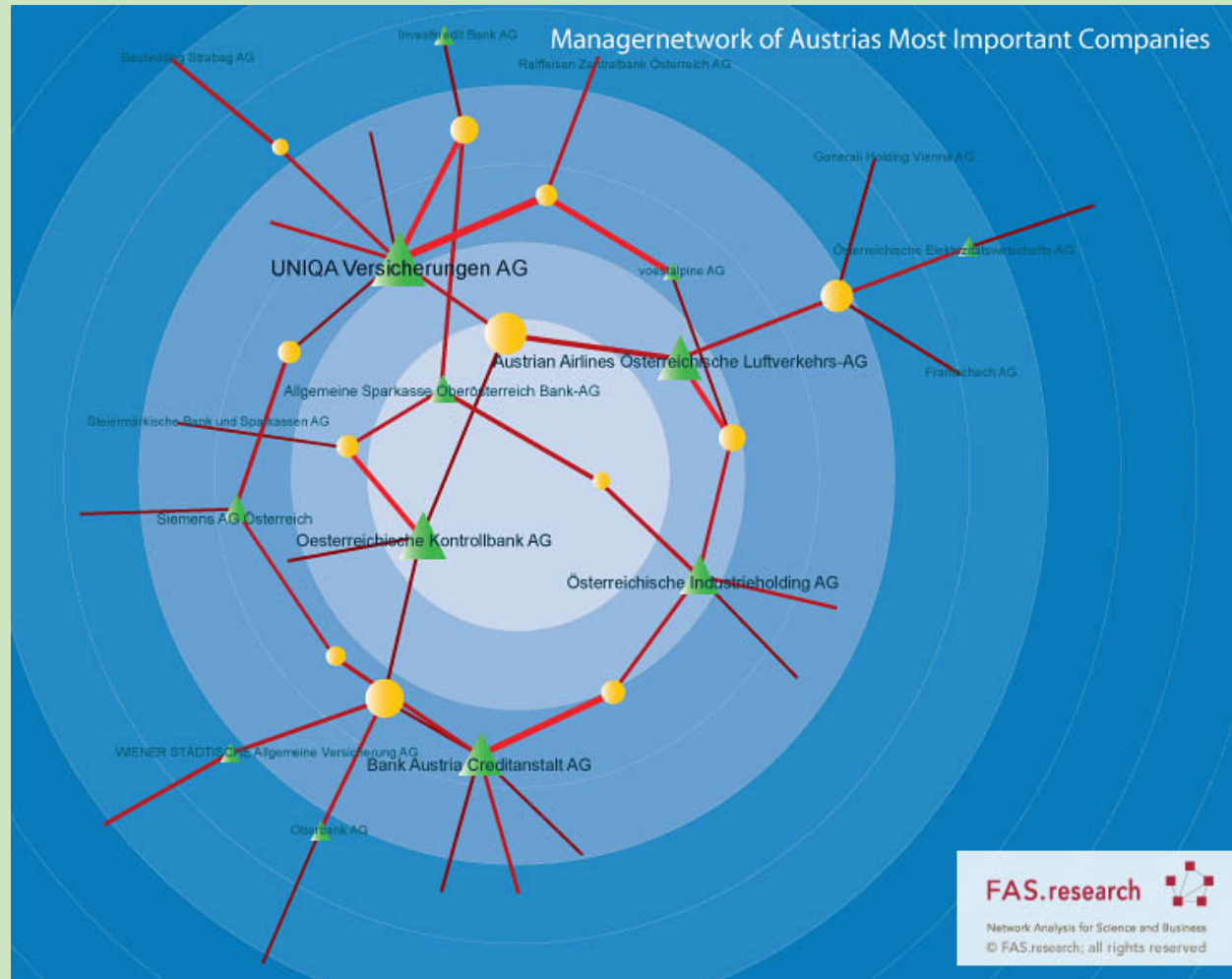




# FAS: The scientific field of Austria



## FAS: Austria's most important Companies

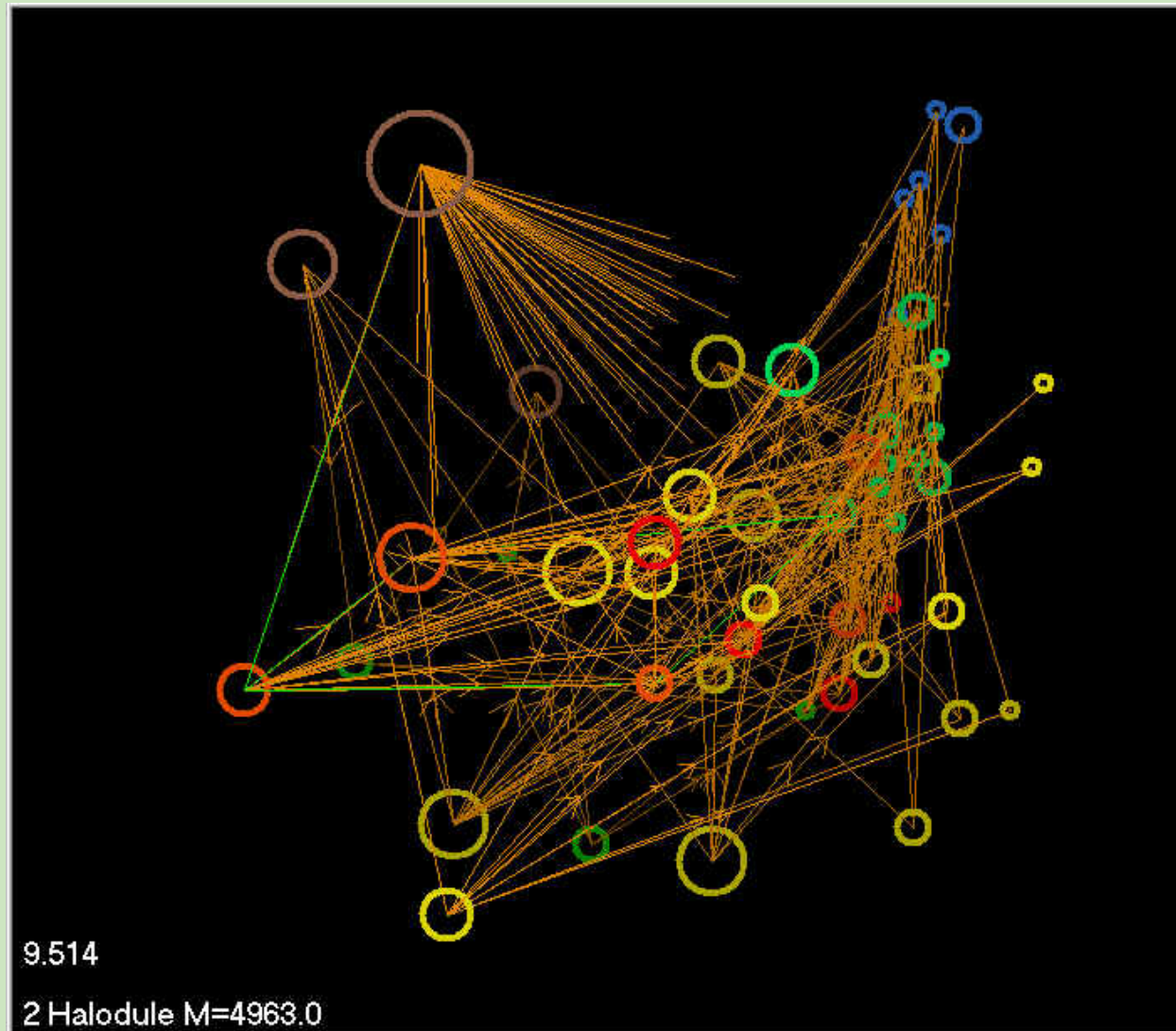








## Jeffrey Johnson: St Marks food web



## Large networks

The simplest solution would be to take very large 'paper'.

The standard solution is abstraction – clustering. We could also introduce different new graphical objects representing typical parts of network, LOD.

The other possibility is dynamic navigation through network: visualization of the neighborhood with indications of position in the global network (darker/brighter background – distribution of neighbors) complemented with a global map.



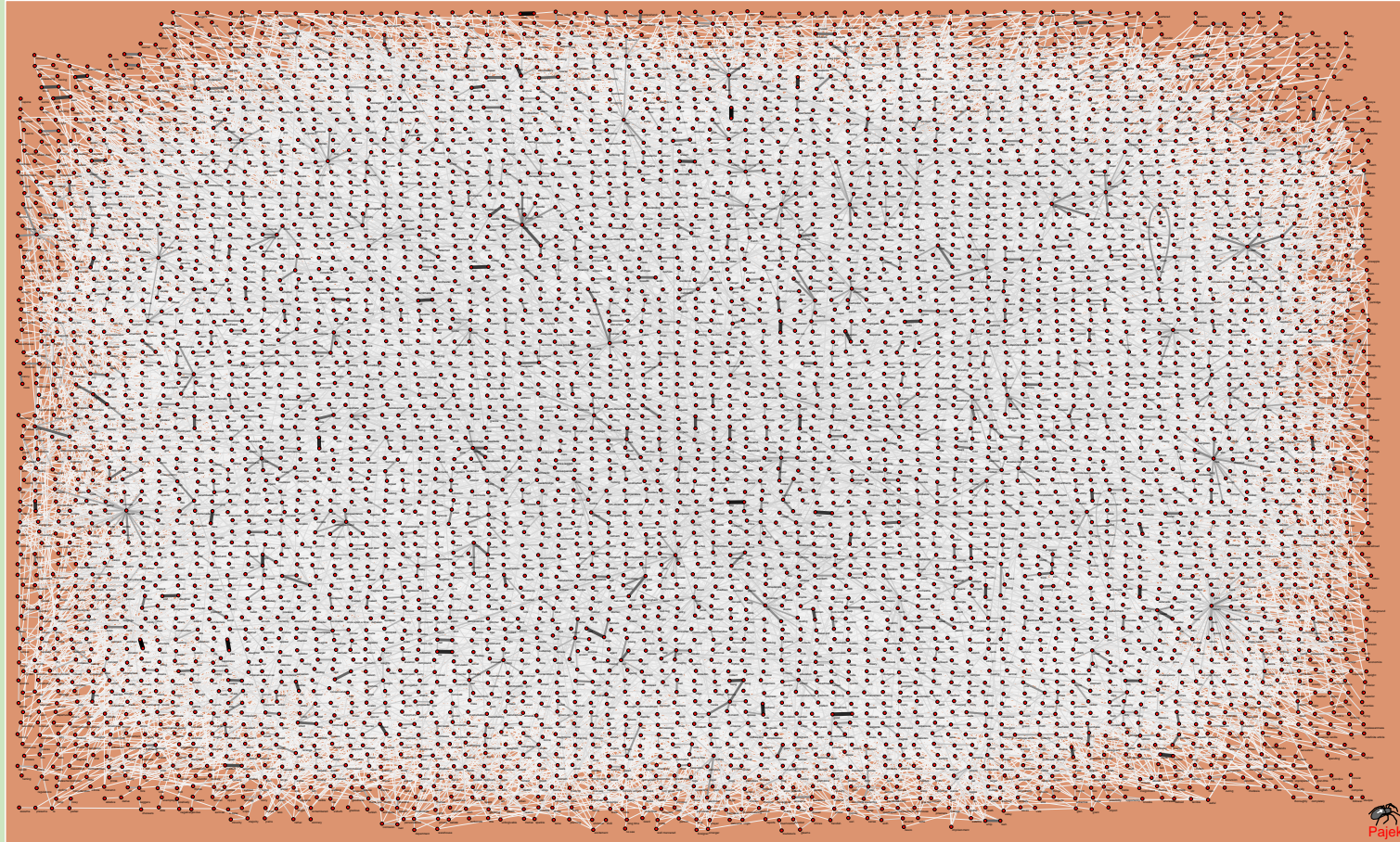
## Big picture, V. Batagelj, AE'04



subnetwork ( $n = 5952$ ,  $m = 18008$ ) of the symmetrized [Edinburgh Associative Thesaurus](#)



## Big picture





# Michael Blum



# Internet Movie Database <http://www.imdb.com/>

**IMDb**  
Earth's Biggest Movie Database™

Home | Top Movies | Photos | Independent Film | Browse | Help | Login | Register to personalize

**The Internet Movie Database**  
Visited by over **30 million** movie lovers each month!

Welcome to the Internet Movie Database, the biggest, best, most award-winning movie site on the planet. Want to make IMDb your home page? Drag [this link](#) onto your Home button.

**Honda Civic** and IMDb Want You to **"Pitch Your Picture"** Today!

**PITCH YOUR PICTURE.**

You have the idea for your movie. You even have the poster. Now, [Honda Civic](#) and IMDb want you to "Pitch Your Picture." Submit your poster for your made-up movie, along with the tagline, and you may be eligible to be [entered into](#) our "Pitch Your Picture" [competition](#) (please note [game rules and restrictions](#)). We are now accepting submissions (voting will commence on the 14th). Use only your original ideas and your original images. Do not use existing screen captures, posters, or stills from other

**Movie and TV News**  
**Wed 19 October 2005:**  
Celebrity News  

- [Kidman Photographer Wins DNA Appeal](#)
- [Sizemore Has His Probation Reinstated](#)
- [Madonna Thanks ABBA for the Music](#)

 Studio Briefing  

- ['Fog' Obscures Box Office](#)
- [Schwarzenegger Wants To Terminate Video Game Lawsuit](#)
- [Jackson Dumps 'King Kong' Music](#)

**Born Today**  
Wednesday, 19 October 2005:

**12th Annual Graph Drawing Contest, 2005.** The IMDB network is bipartite (2-mode) and has  $1324748 = 428440 + 896308$  vertices and 3792390 arcs.

## Bipartite cores

The subset of vertices  $C \subseteq V$  is a  $(p, q)$ -core in a bipartite (2-mode) network  $N = (V_1, V_2; L)$ ,  $V = V_1 \cup V_2$  iff

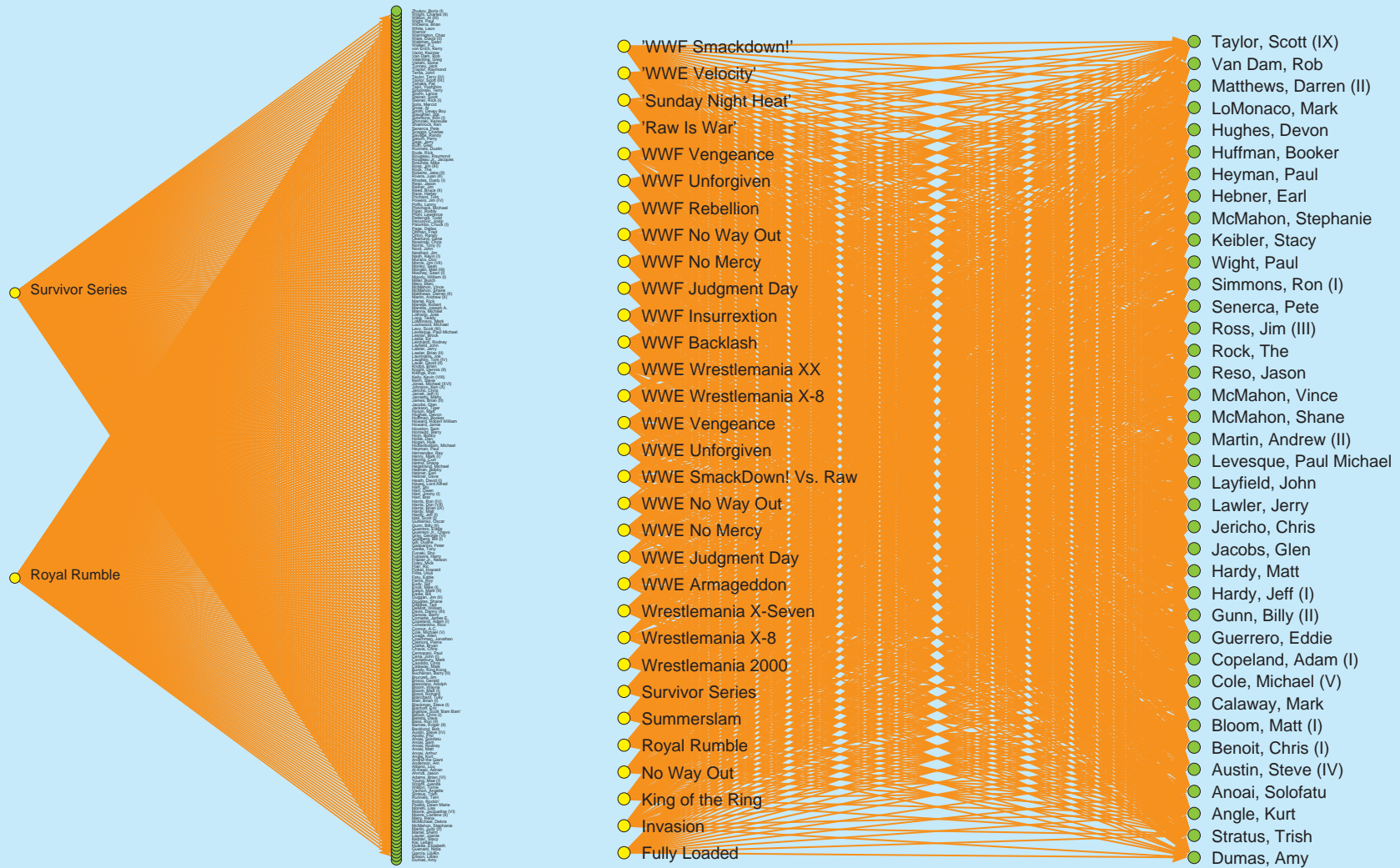
- a. in the induced subnetwork  $K = (C_1, C_2; L(C))$ ,  $C_1 = C \cap V_1$ ,  $C_2 = C \cap V_2$  it holds  $\forall v \in C_1 : \deg_K(v) \geq p$  and  $\forall v \in C_2 : \deg_K(v) \geq q$  ;
- b.  $C$  is the maximal subset of  $V$  satisfying condition a.

Properties of bipartite cores:

- $C(0, 0) = V$
- $K(p, q)$  is not always connected
- $(p_1 \leq p_2) \wedge (q_1 \leq q_2) \Rightarrow C(p_1, q_1) \subseteq C(p_2, q_2)$
- $\mathcal{C} = \{C(p, q) : p, q \in \mathbf{N}\}$ . If all nonempty elements of  $\mathcal{C}$  are different it is a lattice.
- To determine a  $(p, q)$ -core the procedure similar to the ordinary core procedure can be used.



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

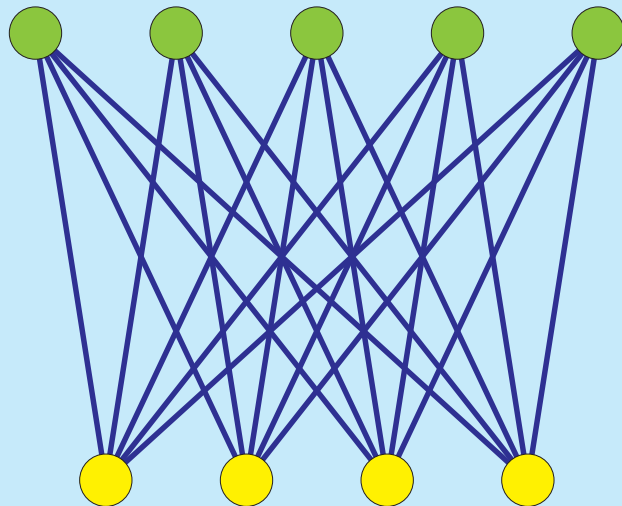






## 4-rings and analysis of 2-mode networks

In bipartite (2-mode) network there are no 3-rings. The densest substructures are complete bipartite subgraphs  $K_{p,q}$ . They contain many 4-rings.

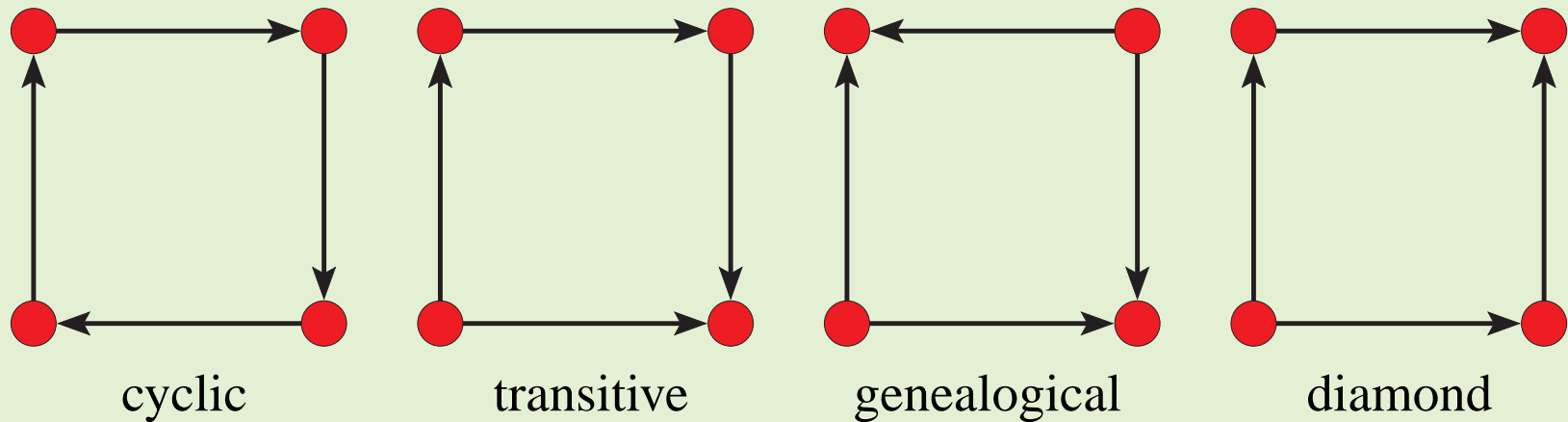


$$w_4(K_{p,q}) = (p-1)(q-1)$$

The 4-rings weights were implemented in **Pajek** only recently, in August 2005.

## Directed 4-rings

There are 4 types of directed 4-rings:



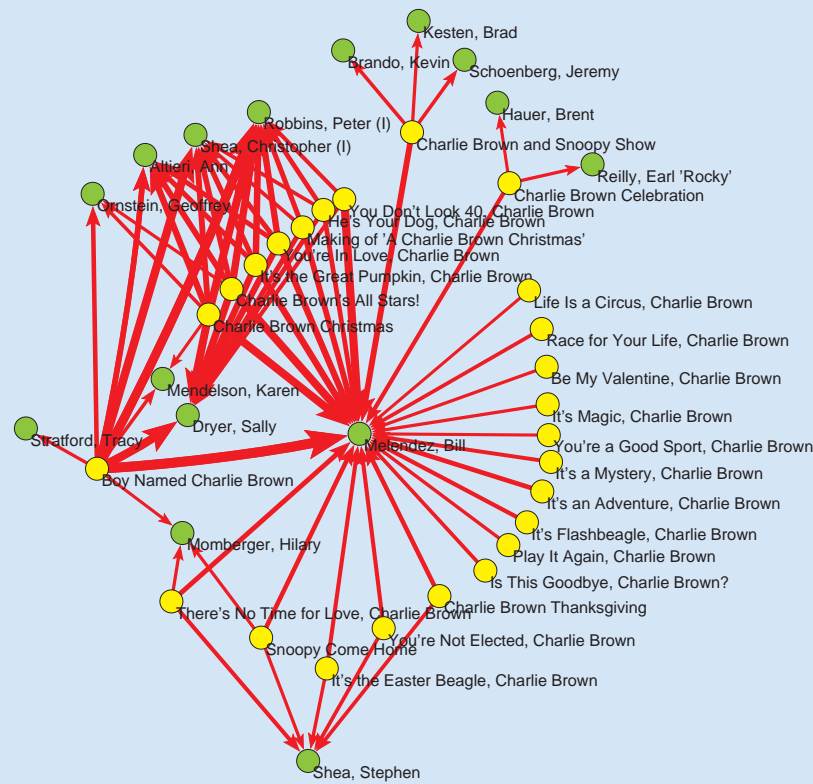
In the case of transitive rings **Pajek** provides a special weight counting on how many transitive rings the arc is a *shortcut*.

## Simple line islands in IMDB for $w_4$

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

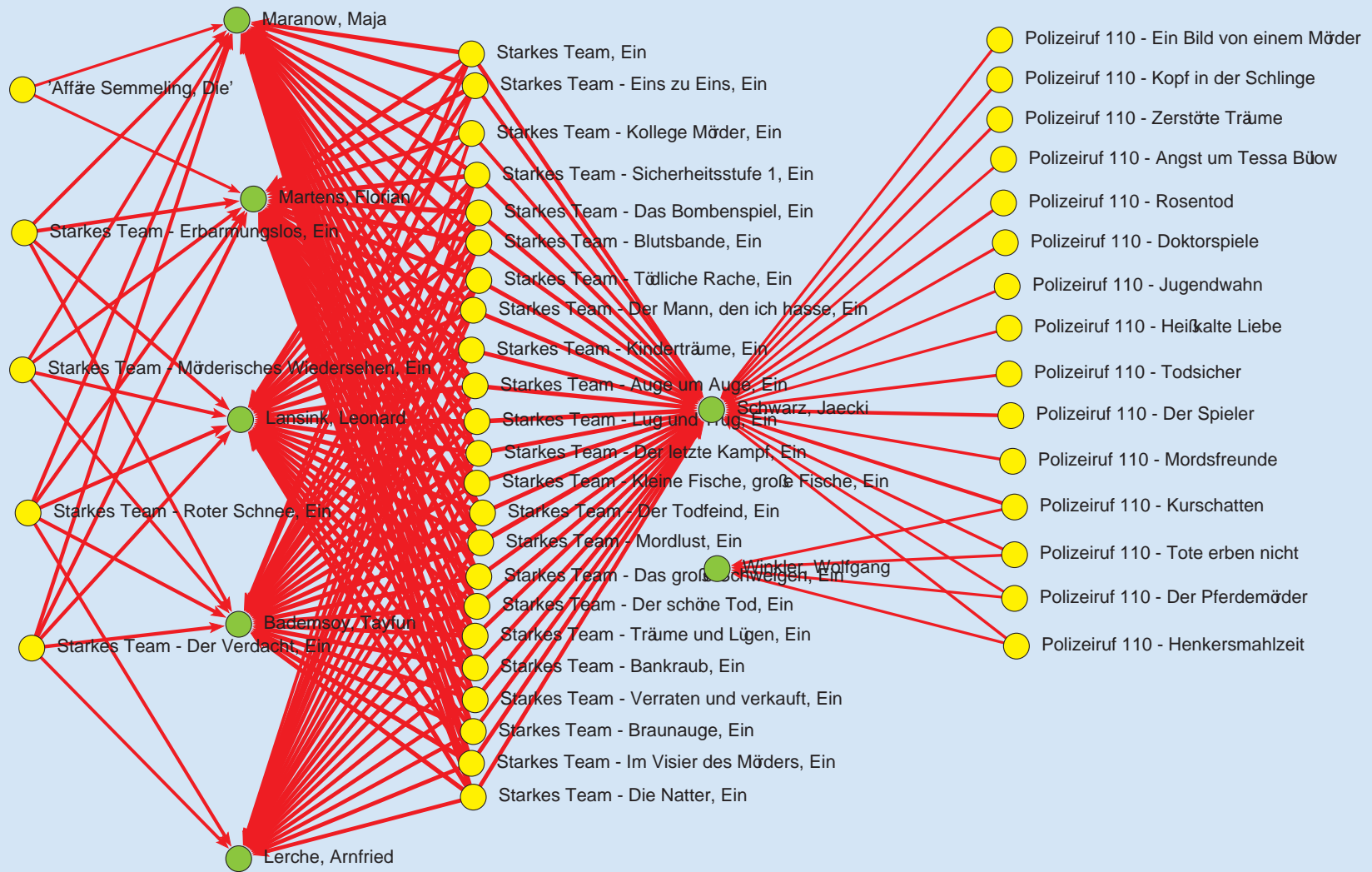
Size	Freq	Size	Freq	Size	Freq	Size	Freq
2	5512	20	19	38	4	59	2
3	1978	21	18	39	3	61	1
4	1639	22	15	40	2	64	1
5	968	23	9	42	2	67	1
6	666	24	13	43	3	70	1
7	394	25	12	45	3	73	1
8	257	26	6	46	4	76	1
9	209	27	6	47	5	82	1
10	148	28	5	48	1	86	1
11	118	29	6	49	2	106	1
12	87	30	3	50	2	122	1
13	55	31	6	51	1	135	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: Island for $w_4$ / Polizeiruf 110 and Starkes Team



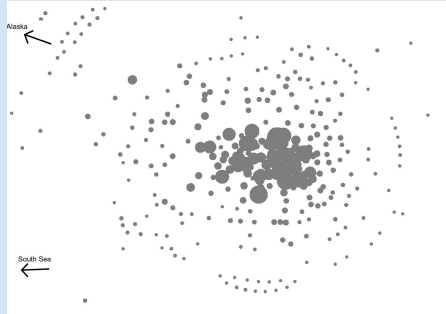
## New drawing algorithms from GD'05

Tim Dwyer, Yehuda Koren, and Kim Marriott: **Stress Majorization with Orthogonal Ordering Constraints.**

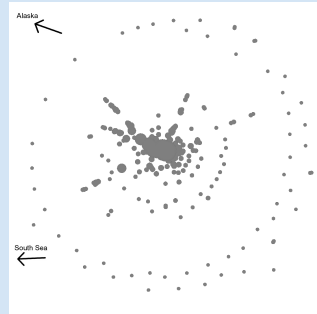
The *1138bus* graph (1138 vertices, 1458 arcs).



## New drawing algorithms from GD'05



(a) Fruchterman-Reingold model



(b) Node-repulsion LinLog model



(c) Edge-repulsion LinLog model

Andreas Noack:

Energy-Based Clustering of Graphs  
with Nonuniform Degrees.

American airports.

## Dense networks

Dense(r) networks with more than 15 vertices usually can't be clearly presented. For such networks a more appropriate display is using *matrix representation*. Properly reordering vertices can reveal *patterns* in corresponding matrix display.

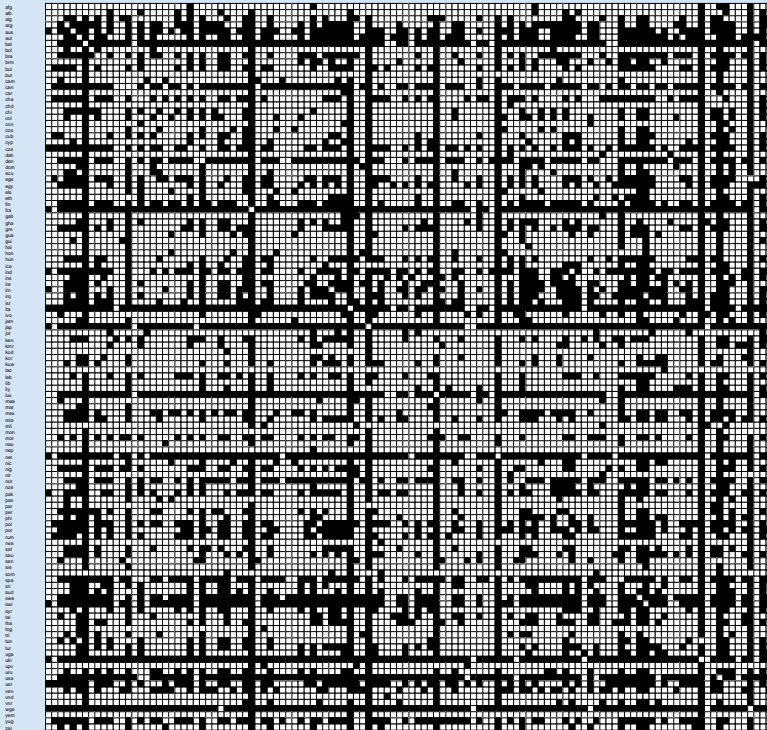
Matrix representation can be supplemented with display of properties of vertices. It is also compatible with shrinking/ expanding with respect to a given hierarchy (see papers 1, 2).

There are different approaches how to determine good orderings: weak, strong components; topological sort in acyclic networks; special heuristics (**Reverse Cuthill-McKee**); seriation and clumping (Murtagh, 1985); clustering; blockmodeling.

## Snyder & Kick's World trade network / $n = 118, m = 514$

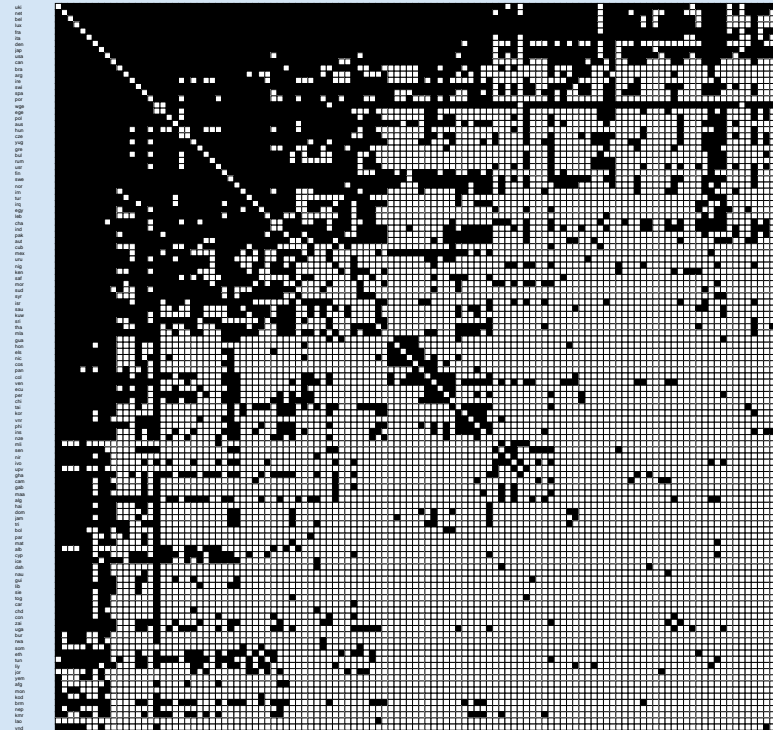
Pajek - shadow 0.00,1.00  
World trade - alphabetic order

Sep-5-1998



Pajek - shadow 0.00,1.00  
World Trade (Snyder and Kick, 1979) - cores

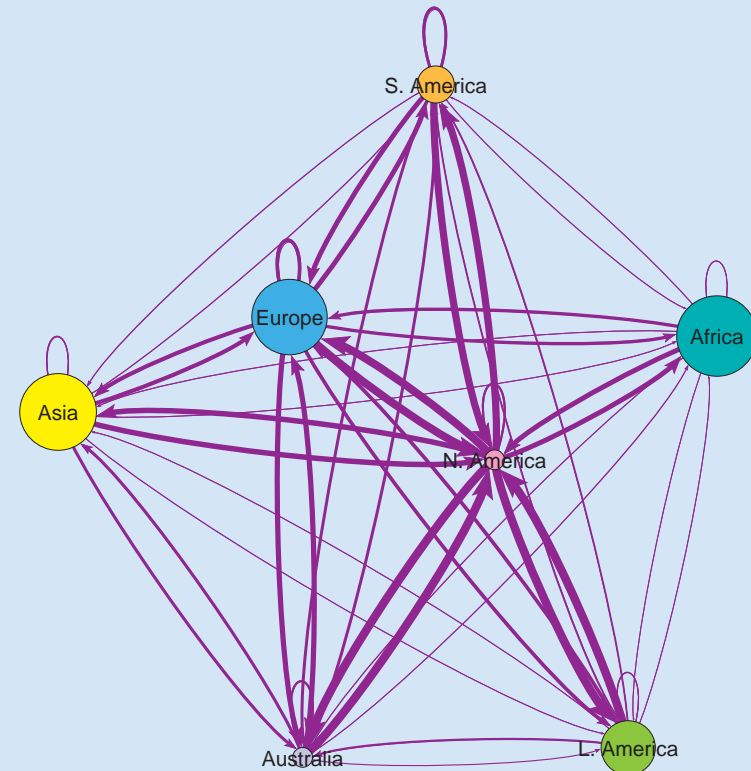
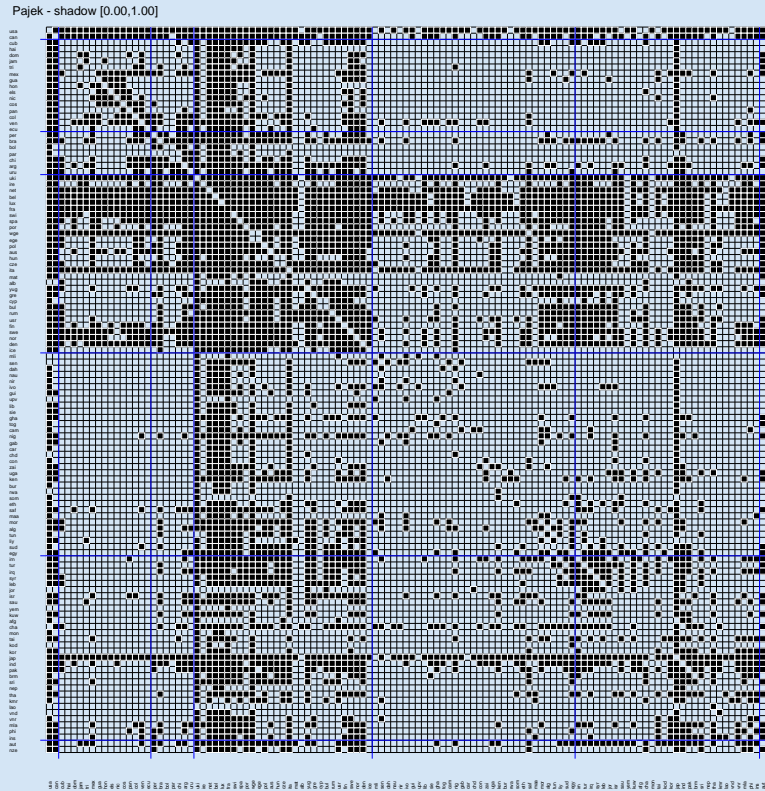
Sep-5-1998



Alphabetic order of countries (left) and rearrangement (right)



## Contracted clusters – international trade

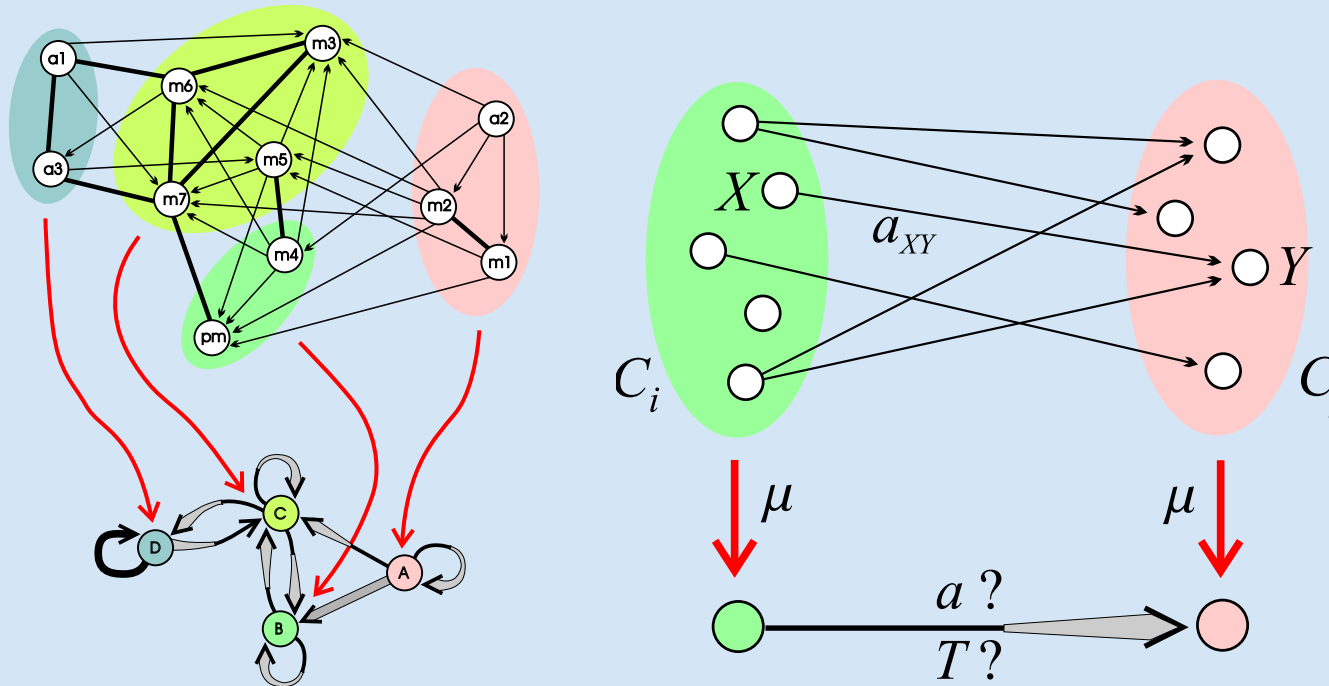


Snyder and Kick's international trade. Matrix display of dense networks.

$$w(C_i, C_j) = \frac{n(C_i, C_j)}{n(C_i) \cdot n(C_j)}$$

## Generalized Blockmodeling

A *blockmodel* consists of structures obtained by identifying all units from the same cluster of the clustering  $C$ . For an exact definition of a blockmodel we have to be precise also about which blocks produce an arc in the *reduced graph* and which do not, and of what *type*.



Book: P. Doreian, V. Batagelj, A. Ferligoj: *Generalized Blockmodeling*, CUP, 2004. Amazon.

## New graphical elements

To improve the pictures new graphical elements could be introduced:

- Almost clique or bipartite complete graph could be replaced by the circle/quadrangle with the inverse drawing of missing edges.
- 'multiedges'



## Core based layout

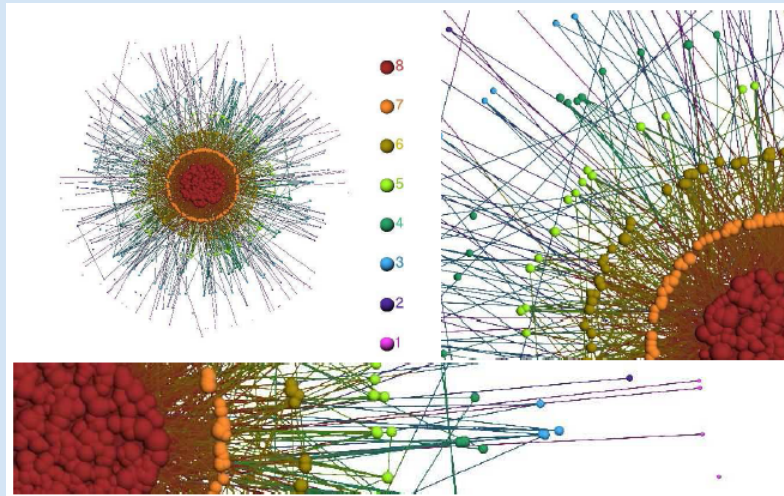


Figure 3: Graphical representation of a BA network with a random  $m \in [1 : 10]$ , and  $n=1000$

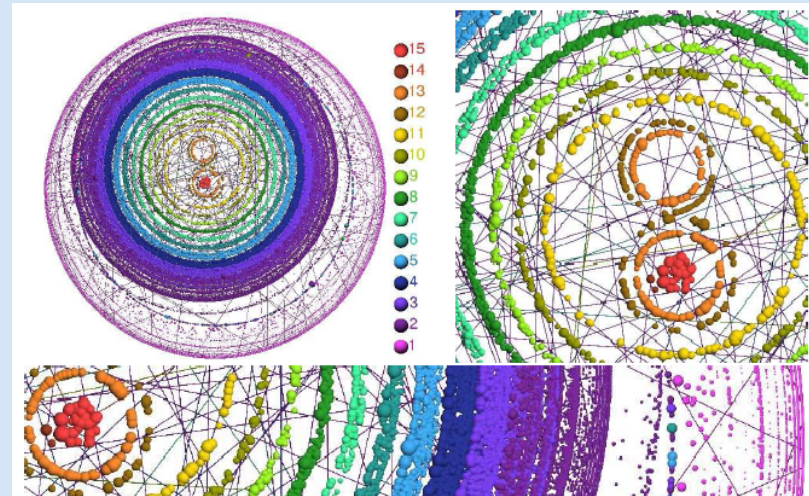


Figure 11: Graphical representation of a fraction of the .fr domain of the WWW

I. Alvarez-Hamelin, L. Dall'Asta, A. Barrat, and A. Vespignani: k-core decomposition: a tool for the visualization of large scale networks [arXiv:cs 0504107](https://arxiv.org/abs/cs/0504107)

Graphical representation should be simplified into icon omitting details.

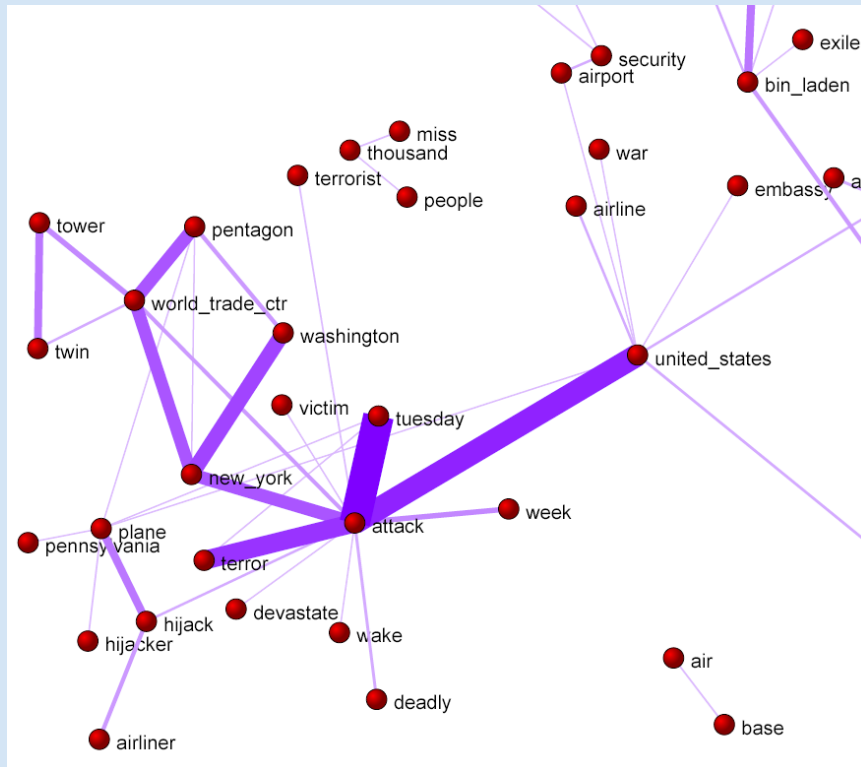


## Dynamic/temporal networks

There are different types of temporal networks:

- networks with implicit time dimension: genealogies
- (recorded) sequence of time slices: Sampson's monks
- (recorded) sequence of events changing the network: KEDS
- online networks: They rule

## Temporal networks / September 11

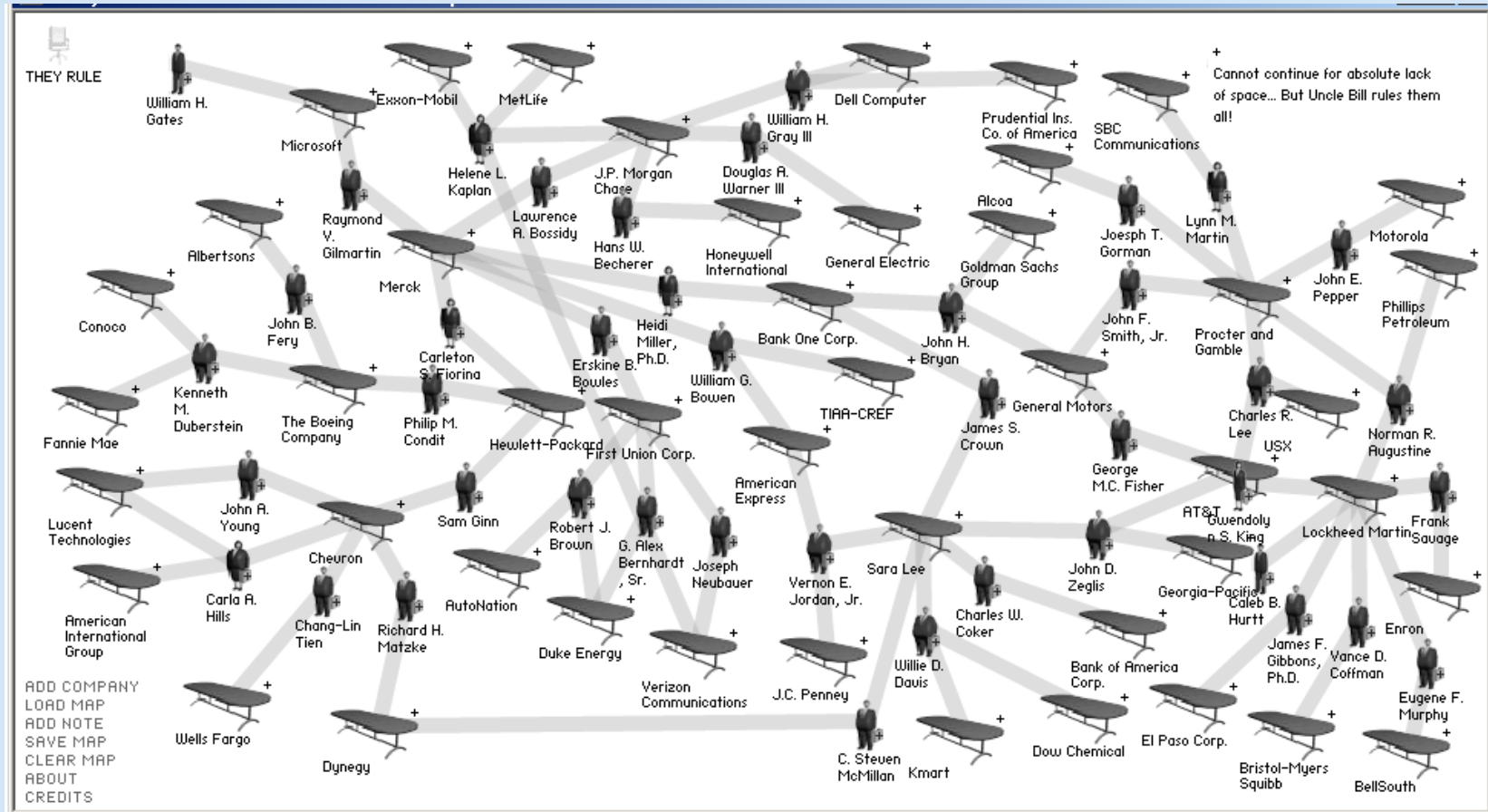


Steve Corman with collaborators from Arizona State University transformed, using his Centering Resonance Analysis (*CRA*), daily Reuters news (66 days) about September 11th into a temporal network of words coappearance. This network was a challenge network for *Viszards* 2002 session.

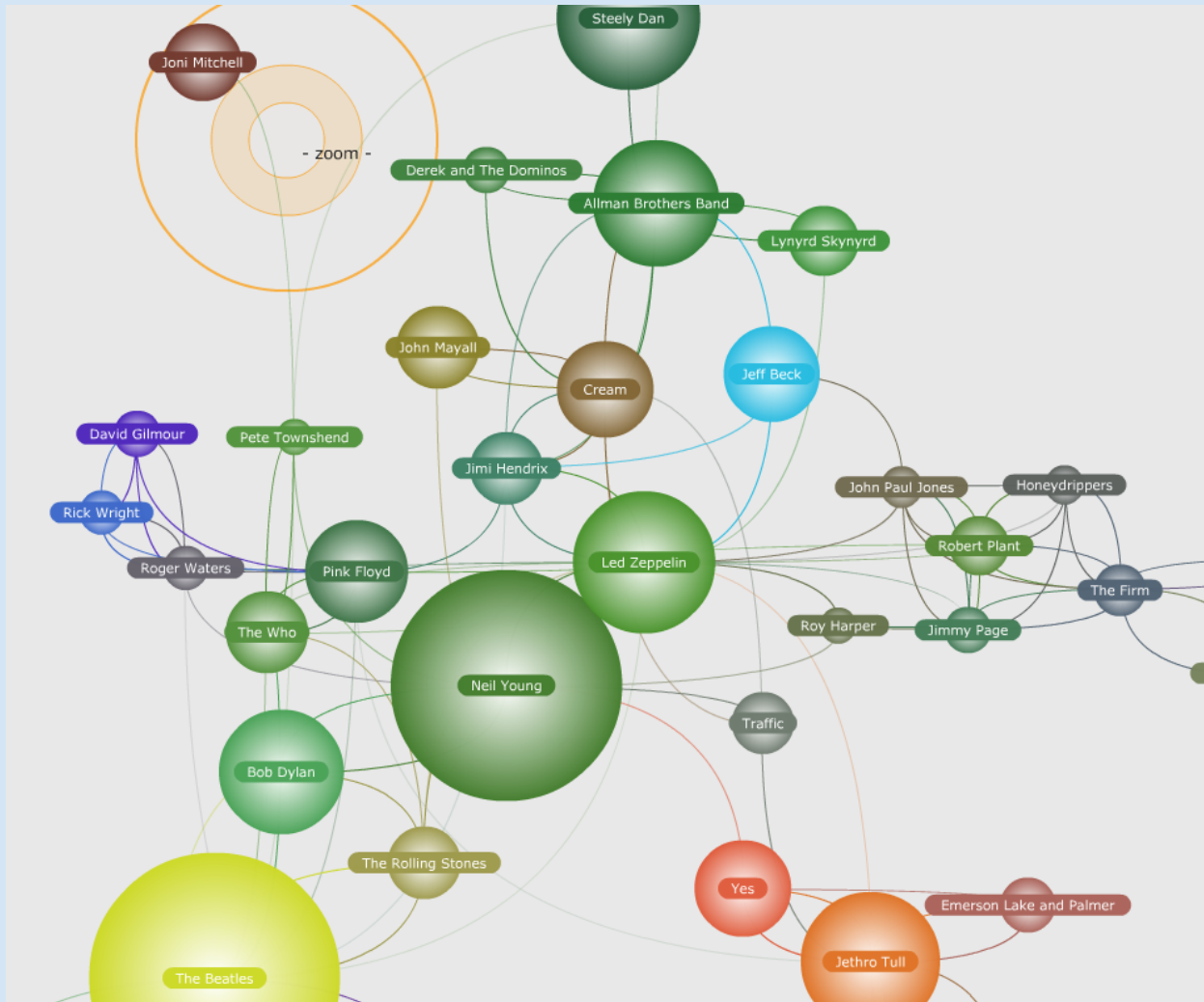
Pictures in SVG: *66 days*. ([SVG viewer](#))

Every year (from 2002) we have at the *Sunbelt conference* a *Viszards* session presenting solutions of *Viszards* group to a visualizations of selected network or type of networks.

# They Rule



# MusicPlasma



# Musicvine



## Networks from the Internet



KartOO network

*Internet Mapping Project.*

Links among WWW pages.

*KartOO, TouchGraph.*

Derived from archives of E-mail, blogs, . . . , server's logs.

*Cybergeography, CAIDA.*

## Challenges

- Visualization of properties
- Drawing/display styles
- Interactive layouts, common format(s) ?
- Matrix layout, generalized blockmodeling of large networks
- Additional graphical elements to improve layouts
- Visualization (and analysis) of multi-relational networks
- Visualization (and analysis) of dynamic/temporal networks
- Visualization of dense acyclic networks (genealogies, citation networks)





# Context

