Photo: Stefan Ernst, ***Gartenkreuzspinne* / *Araneus diadematus***

Some new procedures in `Pajek`

Vladimir Batagelj

joint work with Andrej Mrvar

University of Ljubljana

Slovenia

Dagstuhl Seminar 05361

September 5 - 9, 2005

Outline

Standard matrix multiplication

$C_{N \times M} := A_{N \times K} * B_{K \times M}$

$$c_{i,j} = \sum_{k=1}^{K} a_{i,k} \cdot b_{k,j}$$

for i:=1 **to** N **do**
  for j:=1 **to** M **do begin**
    $s := 0$;
    for k:=1 **to** K **do** $s := s + a_{i,k} * b_{k,j}$;
    $c_{i,j} := s$;
  end;

Complexity $O(N \cdot K \cdot M)$.

Fast sparse matrix multiplication

for k:=1 **to** K **do**
  for $i \in N_A^-(k)$ **do**
    for $j \in N_B^+(k)$ **do**
      if $\exists c_{i,j}$ **then** $c_{i,j} := c_{i,j} + a_{i,k} * b_{k,j}$
      else new $c_{i,j} := a_{i,k} * b_{k,j}$

$N_A^-(k)$: input neighbors of vertex k in network A

$N_B^+(k)$: output neighbors of vertex k in network B

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

Complexity of fast sparse matrix multiplication

A and B matrices of directed networks $\mathbf{N}_A = (\mathcal{N}, \mathcal{K}, \mathcal{A}_A)$ and $\mathbf{N}_B = (\mathcal{K}, \mathcal{M}, \mathcal{A}_B)$.

Assume that the body of the loops can be computed in the constant time c. Then the complexity of product is

$$C = \sum_{k \in \mathcal{K}} \sum_{i \in N_A^-(k)} \sum_{j \in N_B^+(k)} c = c \cdot \sum_{k \in \mathcal{K}} \mathrm{indeg}_A(k) \cdot \mathrm{outdeg}_B(k)$$

Let $\Delta_{in}^A = \max_{k \in \mathcal{K}} \mathrm{indeg}_A(k)$ and $\Delta_{out}^B = \max_{k \in \mathcal{K}} \mathrm{outdeg}_B(k)$ and consider the well known equality

$$\sum_{k \in \mathcal{K}} \mathrm{indeg}_A(k) = \sum_{i \in \mathcal{N}} \mathrm{outdeg}_A(k) = |\mathcal{A}_A|$$

We get $C \leq c \cdot \min(|\mathcal{A}_A| \cdot \Delta_{out}^B, |\mathcal{A}_B| \cdot \Delta_{in}^A)$.

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

More detailed complexity analysis

Let $d_{min}(k) = \min(\mathrm{indeg}_A(k), \mathrm{outdeg}_B(k))$, $\Delta_{min} = \max_{k\in\mathcal{K}} d_{min}(k)$,
$d_{max}(k) = \max(\mathrm{indeg}_A(k), \mathrm{outdeg}_B(k))$,

$\mathcal{K}_L = \{k \in \mathcal{K} : d_{max}(k) \in O(n)\}$ and $\mathcal{K}_S = \mathcal{K} \setminus \mathcal{K}_L$.

Then $C = c \cdot \sum_{k\in\mathcal{K}} \mathrm{indeg}_A(k) \cdot \mathrm{outdeg}_B(k) =$

$$= c \cdot (\sum_{k\in\mathcal{K}_L} \mathrm{indeg}_A(k) \cdot \mathrm{outdeg}_B(k) + \sum_{k\in\mathcal{K}_S} \mathrm{indeg}_A(k) \cdot \mathrm{outdeg}_B(k))$$

$$\leq c \cdot (n \cdot \sum_{k\in\mathcal{K}_L} d_{min}(k) + \Delta_{min} \min(\sum_{k\in\mathcal{K}_S} \mathrm{indeg}_A(k), \sum_{k\in\mathcal{K}_S} \mathrm{outdeg}_B(k)))$$

$$\leq c \cdot \Delta_{min} \cdot (n \cdot |\mathcal{K}_L| + \min(|\mathcal{A}_A|, |\mathcal{A}_B|))$$

If for the sparse networks $\mathbf{N}_A$ and $\mathbf{N}_B$ the quantities Δ_{min} and $|\mathcal{K}_L|$ are small then also the resulting product network $\mathbf{N}_C$ is sparse.

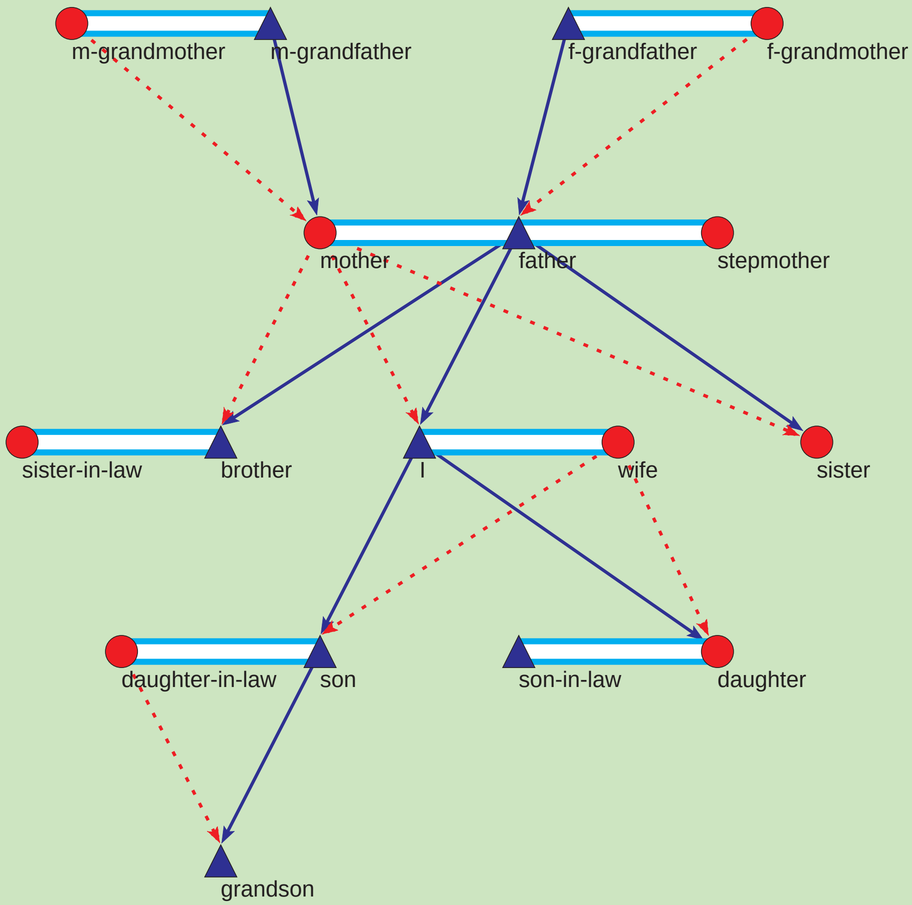
Example: Kinship relations

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

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

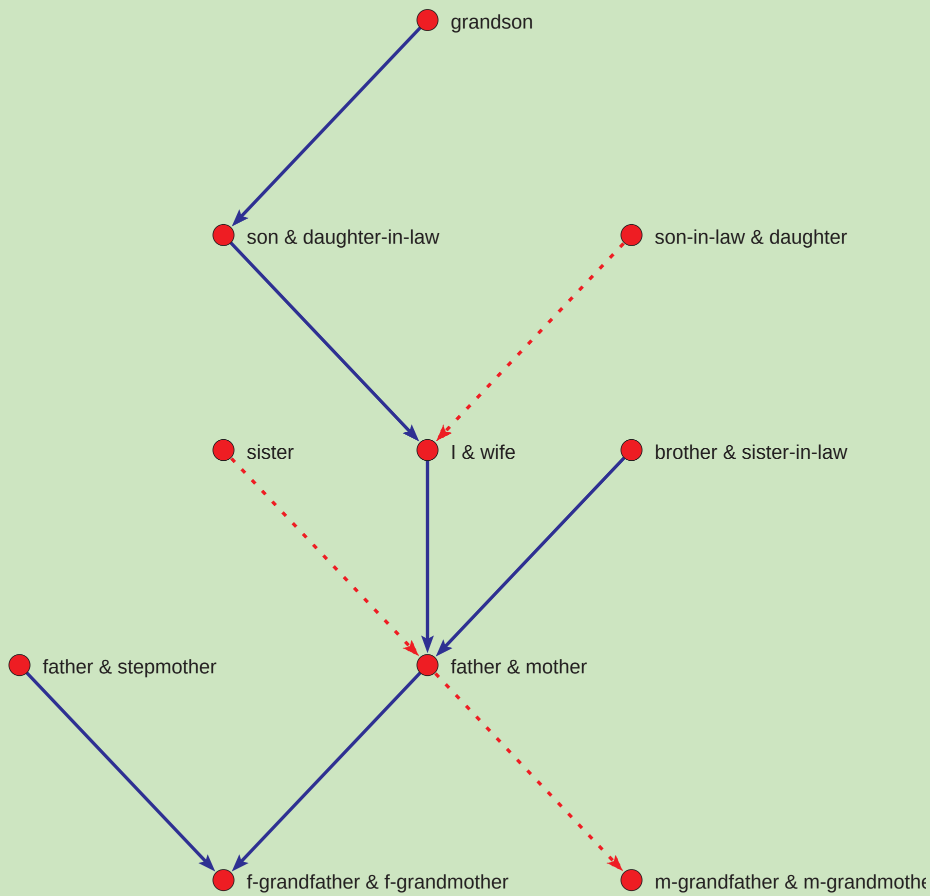
The genealogies are usually described in GEDCOM format. Examples family, Bouchards.

Ore-graph



In Ore-graph every person is represented by a vertex, marriages, relation _ ***is a spouse of*** _, are represented with edges and relations _ ***is a mother of*** _ and _ ***is a father of*** _ as arcs pointing from parents to their children.

p-graph



In p-graph vertices represent individuals or couples. In the case that a person is not married yet (s)he is represented by a vertex, otherwise person is represented with the partner in a common vertex. There are only arcs in p-graphs – they point from children to their parents, representing the relations ***FiC*** _ ***is a daughter of*** _ and ***MiC*** _ ***is a son of*** _; where ***FiC*** $\equiv$ **f**emale **i**n the **c**ouple; and ***MiC*** $\equiv$ **m**ale **i**n the **c**ouple.

Calculating kinship relations

Pajek generates three relations when reading genealogy as Ore graph:

F: _ ***is a father of*** _

M: _ ***is a mother of*** _

E: _ ***is a spouse of*** _

Additionally we must generate two binary diagonal matrices, to distinguish between male and female:

L: _ ***is a male*** _ / 1-male, 0-female

J: _ ***is a female*** _ / 1-female, 0-male

Derived kinship relations

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

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

and using them other relations can be determined

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

Relative sizes of kinship relations in genealogies

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

Product of temporal networks

The notion of product can be extended to temporal networks along the following lines (to be elaborated):

The expression $c_{i,j} := c_{i,j} + a_{i,k} * b_{k,j}$ should be replaced by ($T_{i,k}$ denotes the time span of $a_{i,k}$ ...) :

$S := T_{i,k} \cap T_{k,j}$
if $S \neq \emptyset$ **then begin**
  if $T_{i,j} \setminus S \neq \emptyset$ **then** $c_{i,j}[T_{i,j} \setminus S] := c_{i,j}[T_{i,j}]$;
  if $S \setminus T_{i,j} \neq \emptyset$ **then** $c_{i,j}[S \setminus T_{i,j}] := a_{i,k} * b_{k,j}$;
  $c_{i,j}[S] := c_{i,j} + a_{i,k} * b_{k,j}$
end;

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
- $(p_1 \leq p_2) \wedge (q_1 \leq q_2) \Rightarrow C(p_1,q_1) \subseteq C(p_2,q_2)$
- $\mathcal{C} = \{C(p,q) : p, q \in \mathbb{N}\}$. If all nonempty elements of $\mathcal{C}$ are different it is a lattice.

Algorithm for bipartite 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 vertices of degree less than p,
    and from the second set all vertices of degree less than q
until no vertex was deleted

It can be implemented to run in $O(m)$ time.

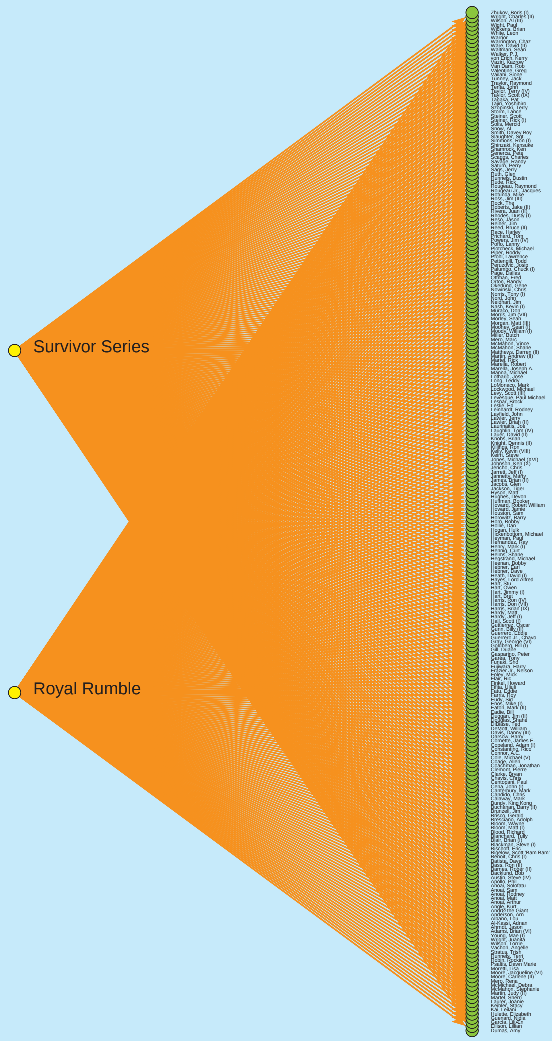
Interesting (p, q)-cores? Table of cores' characteristics $n_1 = |C_1(p, q)|$, $n_2 = |C_2(p, q)|$ and k – number of components in $K(p, q)$:

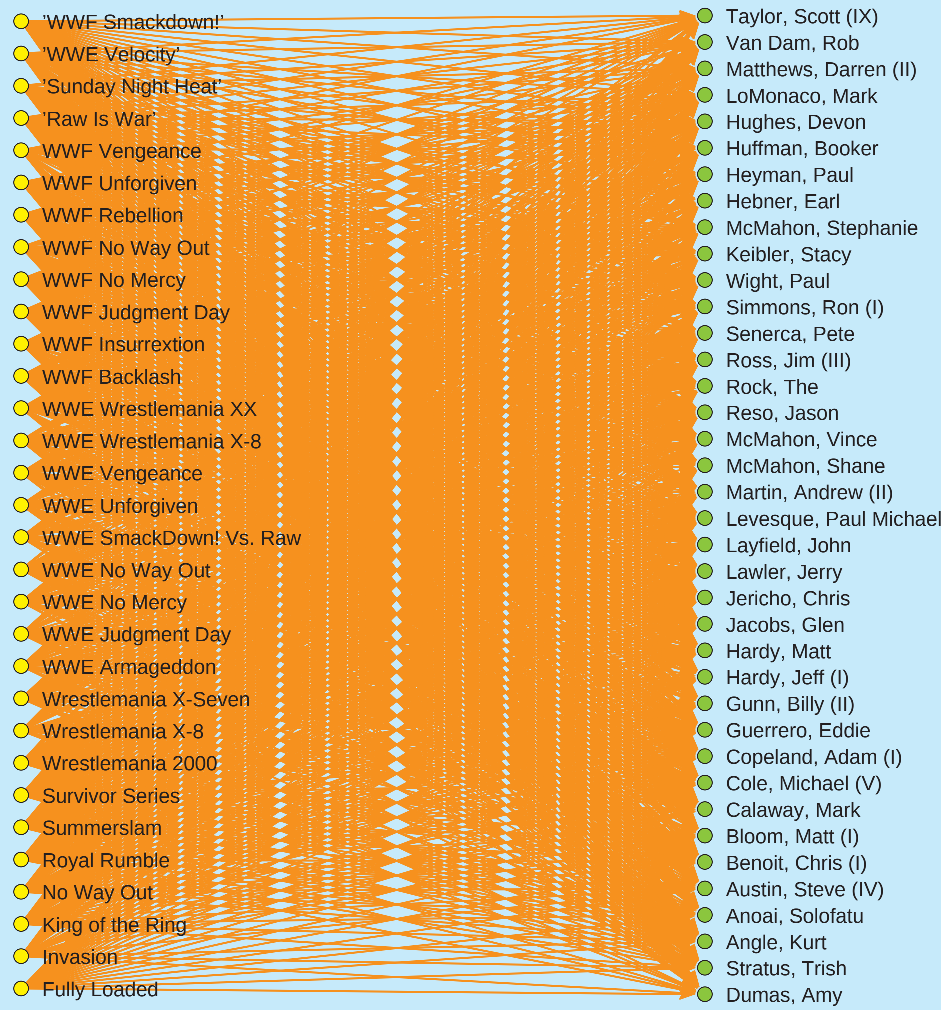
- $n_1 + n_2 \leq$ selected threshold
- big jumps from $C(p - 1, q)$ and $C(p, q - 1)$ to $C(p, q)$.

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

1	1590:	1590	1	|	22	24:	1854	1153	|	43	14:	29	83
2	516:	788	3	|	23	23:	47	56	|	44	14:	29	83
3	212:	1705	18	|	24	23:	34	39	|	45	13:	30	95
4	151:	4330	154	|	25	22:	42	53	|	46	13:	29	94
5	131:	4282	209	|	26	22:	31	38	|	47	12:	29	101
6	115:	3635	223	|	27	22:	31	38	|	48	12:	28	100
7	101:	3224	244	|	28	20:	36	53	|	49	12:	26	95
8	88:	2860	263	|	29	20:	35	52	|	50	11:	27	111
9	77:	3467	393	|	30	19:	35	59	|	51	11:	26	110
10	69:	3150	428	|	31	19:	35	59	|	52	11:	16	79
11	63:	2442	382	|	32	19:	34	57	|	53	10:	35	162
12	56:	2479	454	|	33	18:	34	62	|	54	10:	35	162
13	50:	3330	716	|	34	18:	34	62	|	55	10:	34	162
14	46:	2460	596	|	35	18:	33	61	|	56	10:	34	162
15	42:	2663	739	|	36	17:	33	65	|	57	9:	35	187
16	39:	2173	678	|	37	16:	33	75	|	58	9:	33	180
17	35:	2791	995	|	38	16:	30	73	|	59	9:	33	180
18	32:	2684	1080	|	39	16:	29	70	|	60	9:	32	178
19	30:	2395	1063	|	40	15:	29	77	|	61	9:	31	177
20	28:	2216	1087	|	41	15:	28	76	|	62	9:	31	177
21	26:	1988	1087	|	42	15:	28	76	|	63	8:	31	202

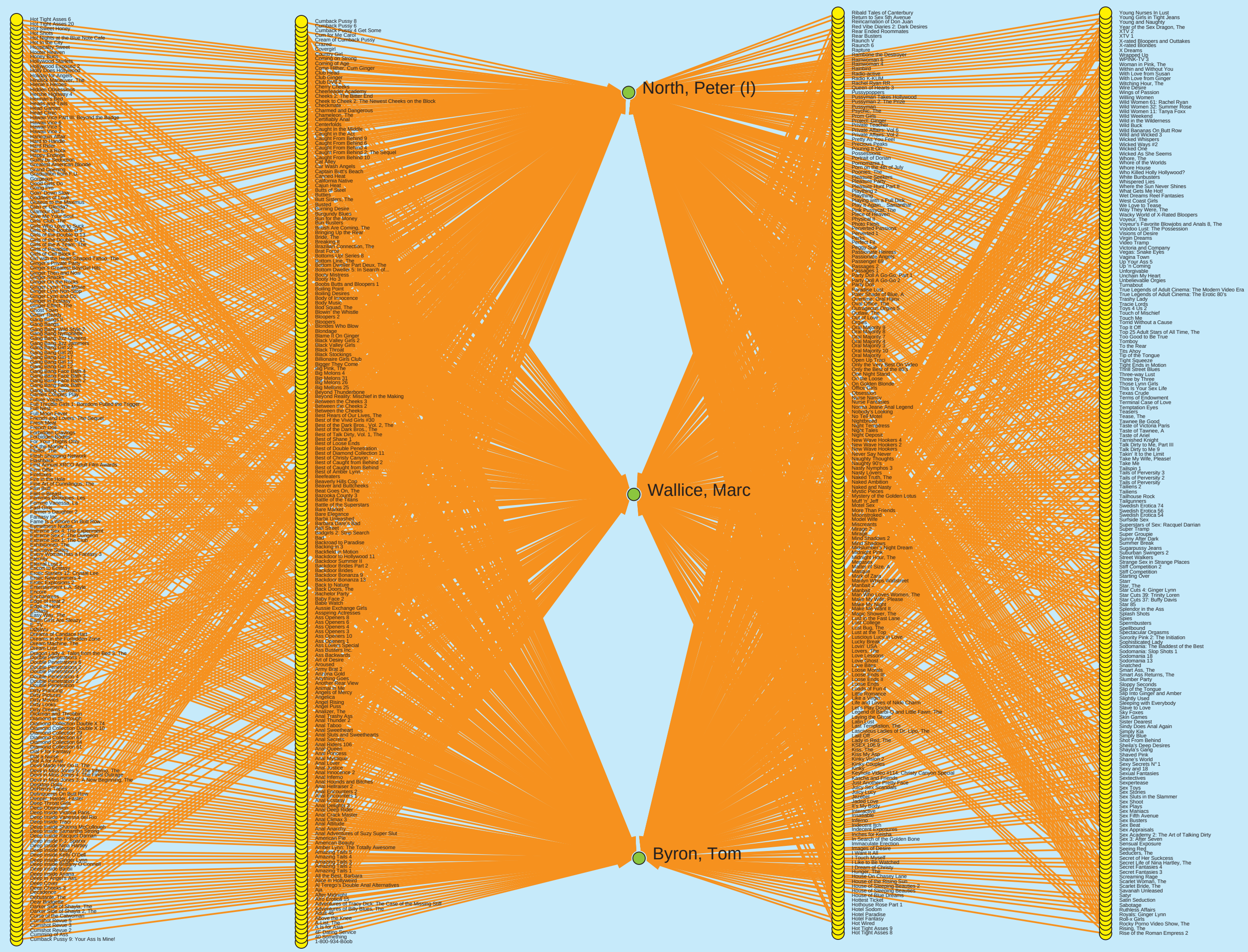
(247,2)-core and (27,22)-core





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

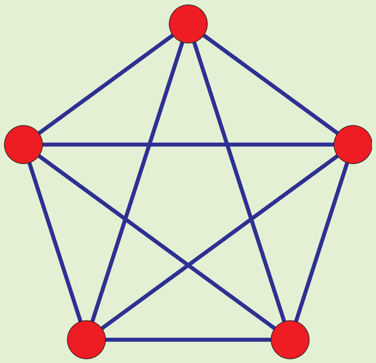
(2,516)-Hard core



k-rings

A *k-ring* is a simple closed chain of length k. Using k-rings we can define a weight of edges as

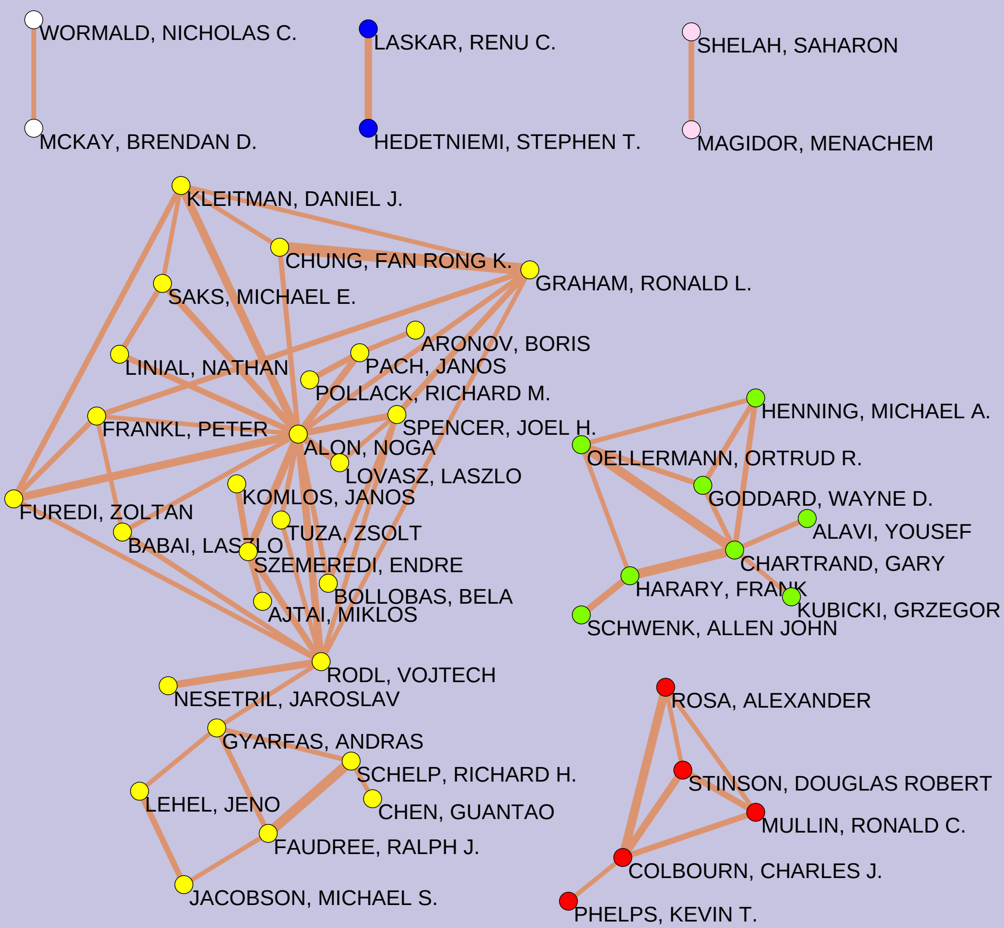
$w_k(e) = \#$ of different k-rings containing the edge $e \in E$



Since for a complete graph K_r, $r \geq k \geq 3$ we have $w_k(K_r) = (r-2)!/(r-k)!$ the edges belonging to cliques have large weights. Therefore these weights can be used to identify the dense parts of a network.

For example: all r-cliques of a network belong to $r-2$-edge cut for the weight w_3.

Edge-cut at level 16 of triangular network of Erdős collaboration graph



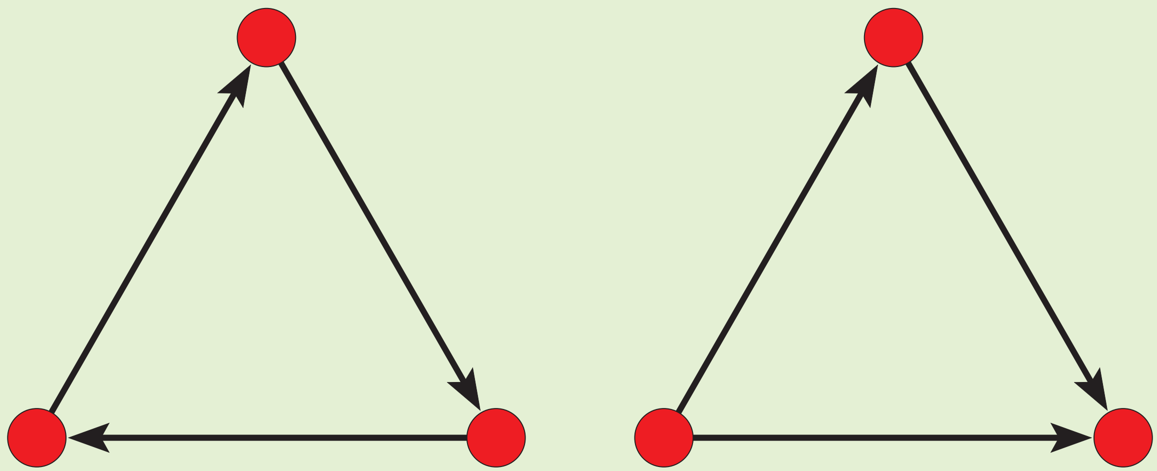
without Erdős,

$n = 6926$,

$m = 11343$

Directed 3-rings

In directed networks there are two types of 3-rings:

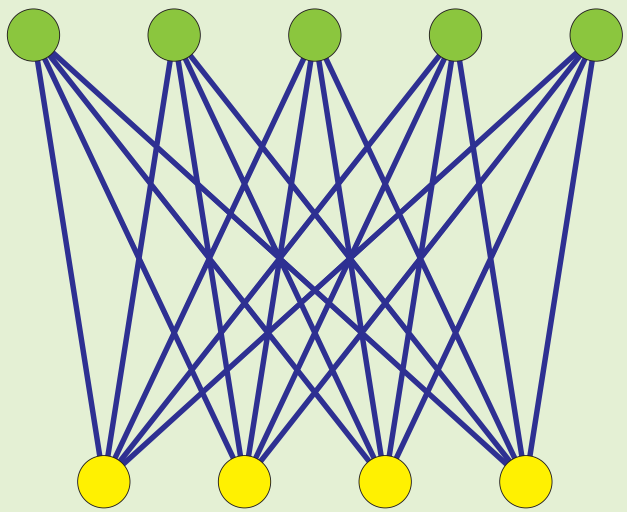


cyclic transitive

The 3-rings weights were implemented in **Pajek** in May 2002.

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.

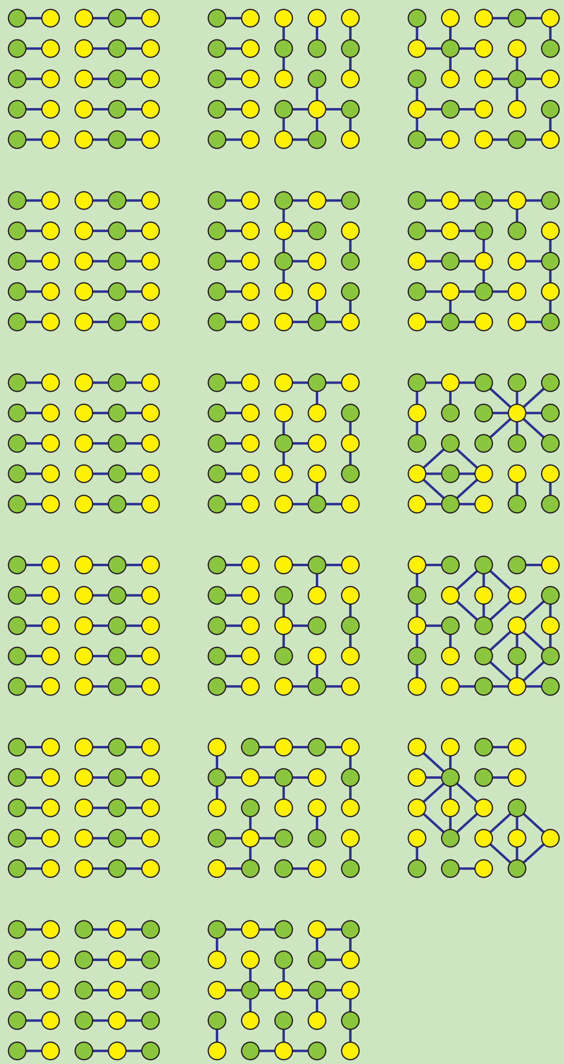


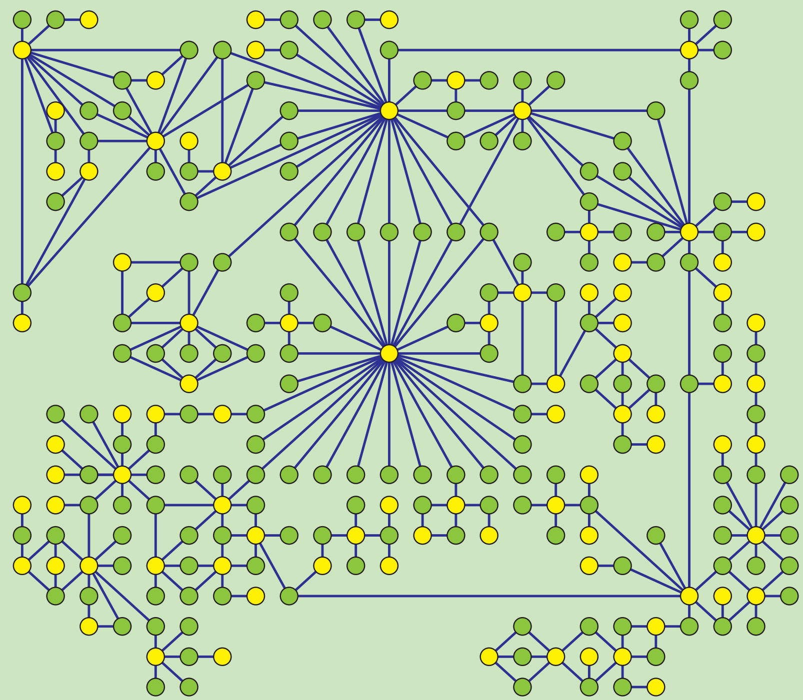
$$w_4(K_{p,q}) = (p-1)(q-1)$$

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

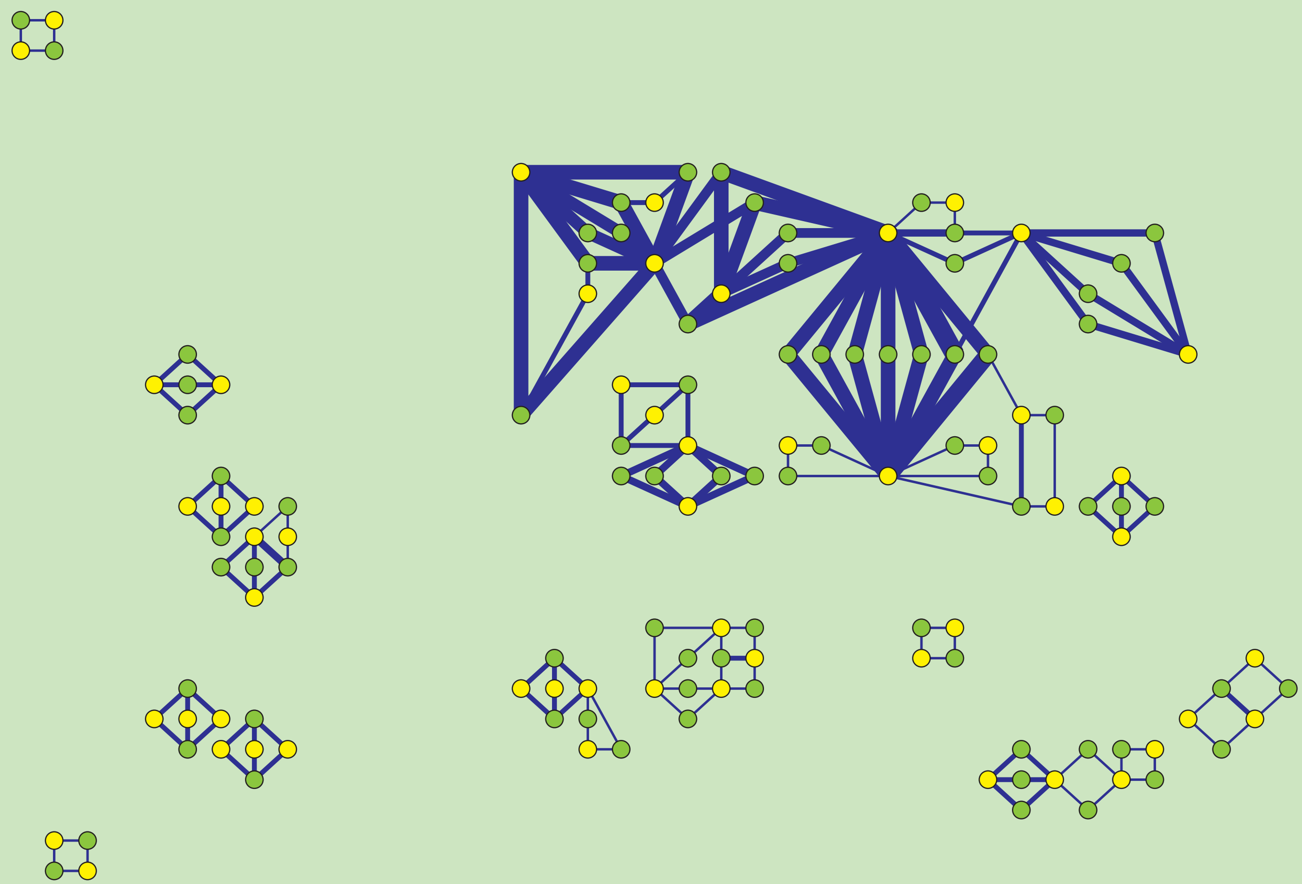
Example: Bibliography from W. Imrich, S. Klavžar: ***Product graphs: structure and recognition***, JohnWiley & Sons, New York, USA, 2000. (PDF), (net – 2-mode 674×314 network).

Example: 4-rings in a 2-mode network

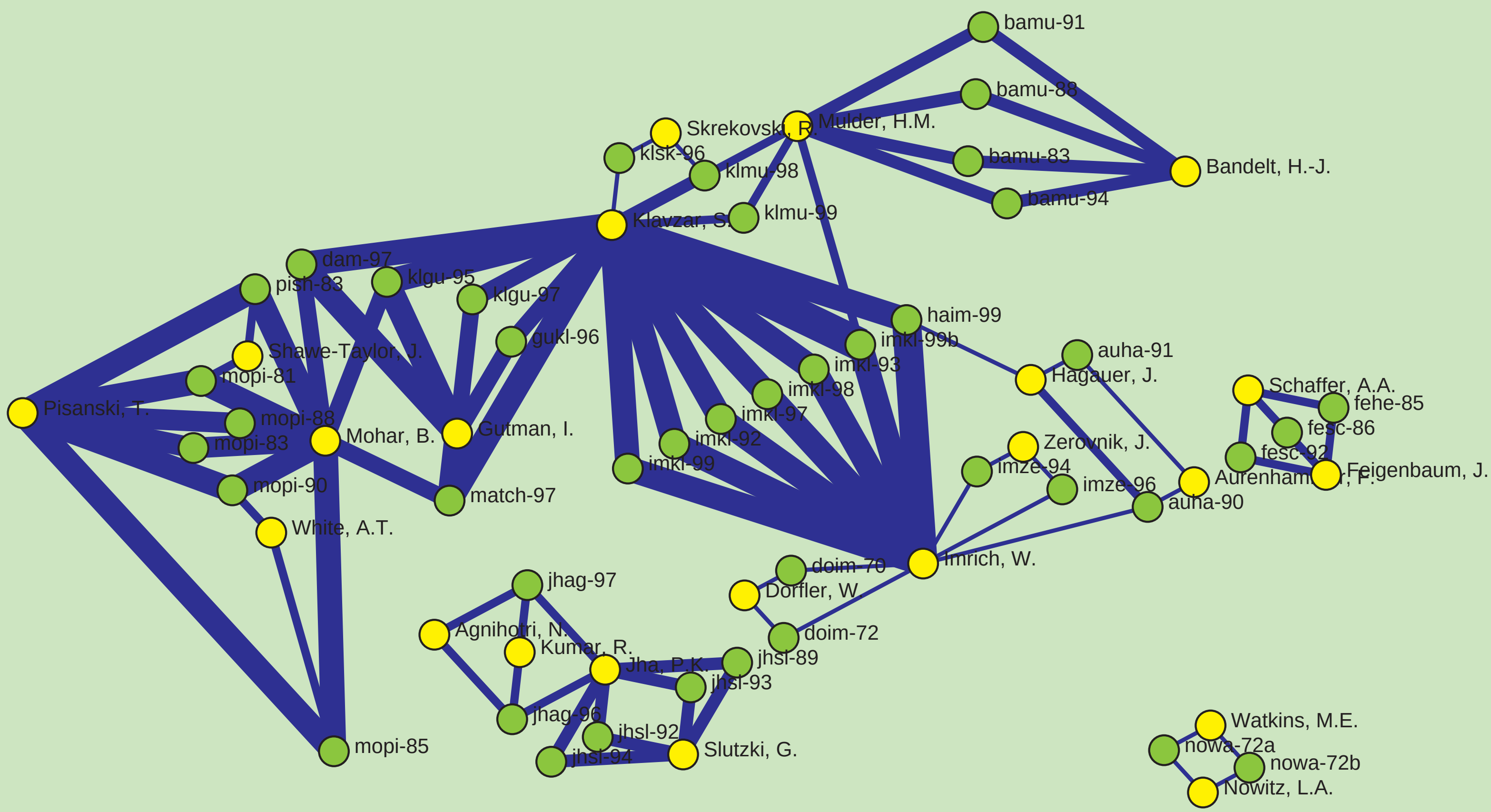




Example: 1-edge cut for w_4

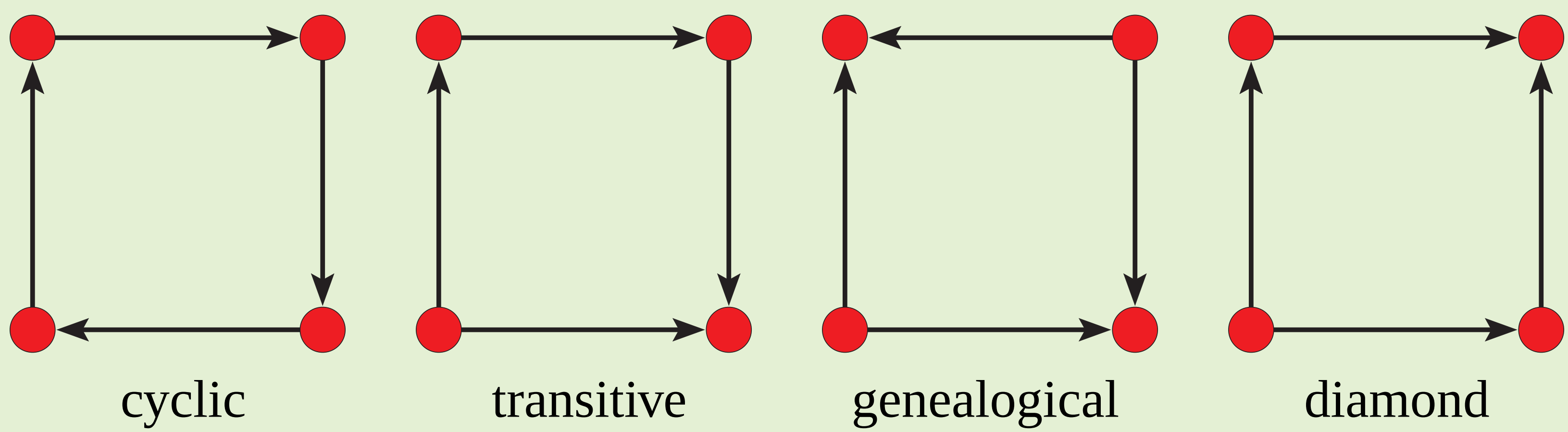


Example: labeled main part of 1-edge cut for w_4



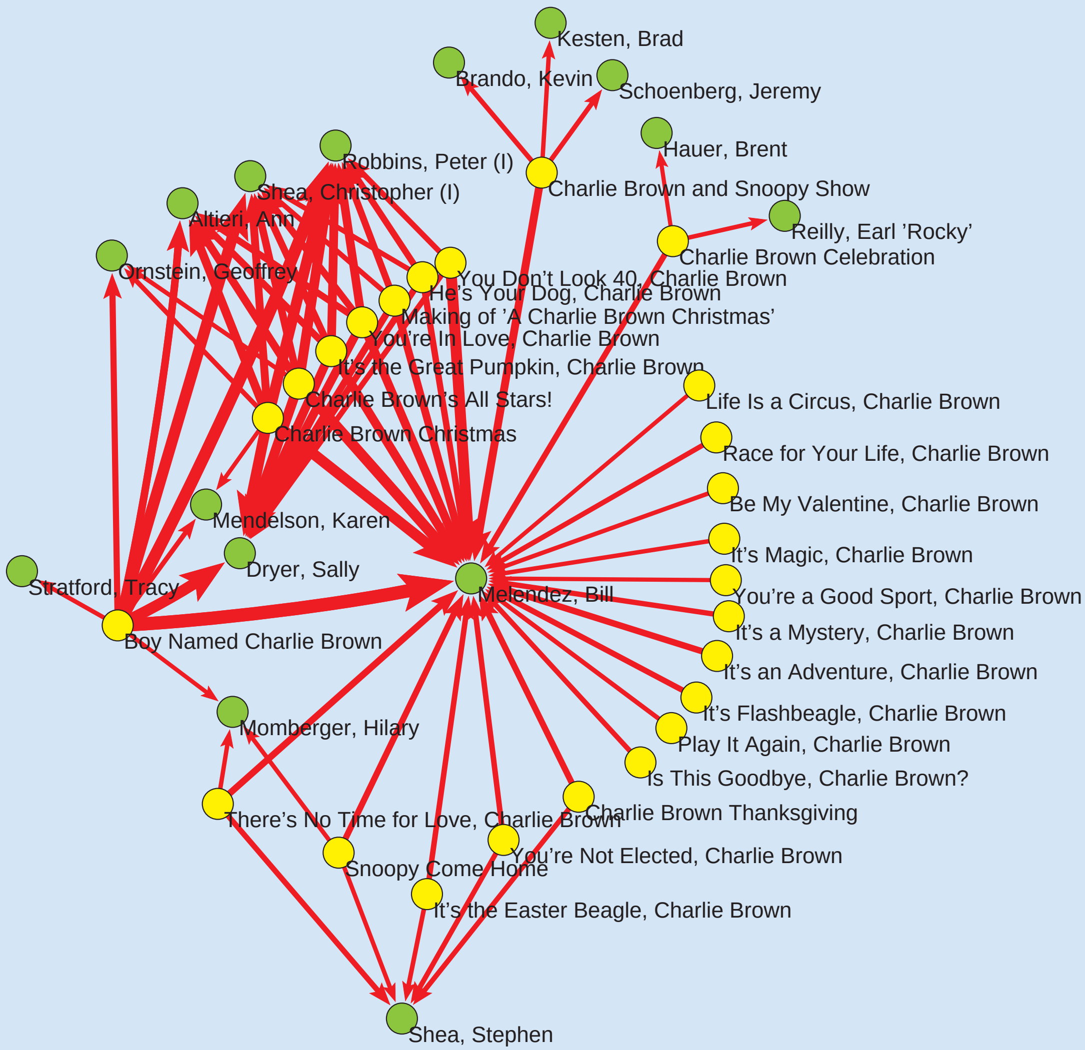
Directed 4-rings

There are 4 types of directed 4-rings:



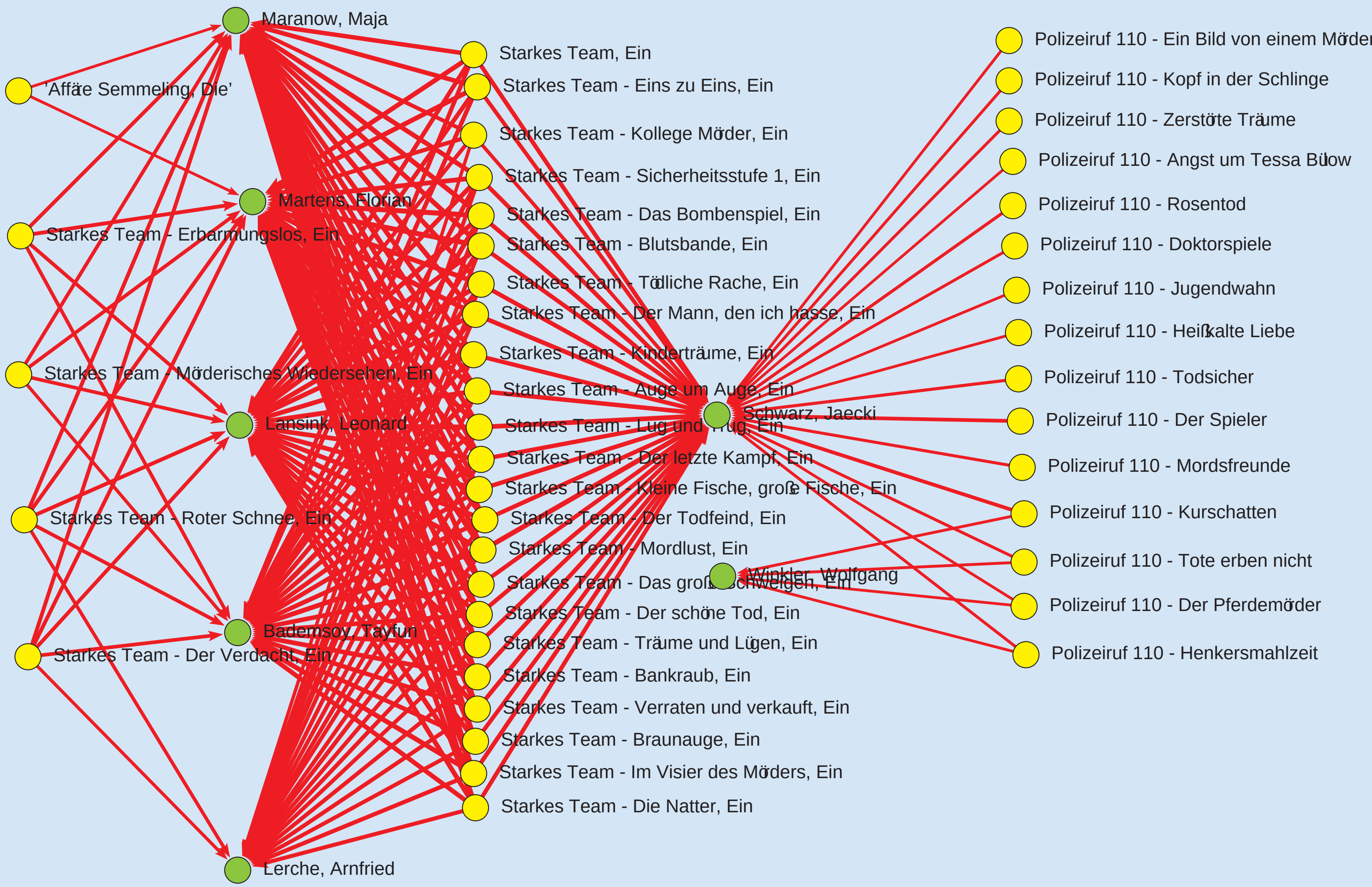
In the case of transitive rings **Pajek** provides a special weight counting on how many transitive rings the arc is a *shortcut*.

Example: Island for w_4 / Charlie Brown



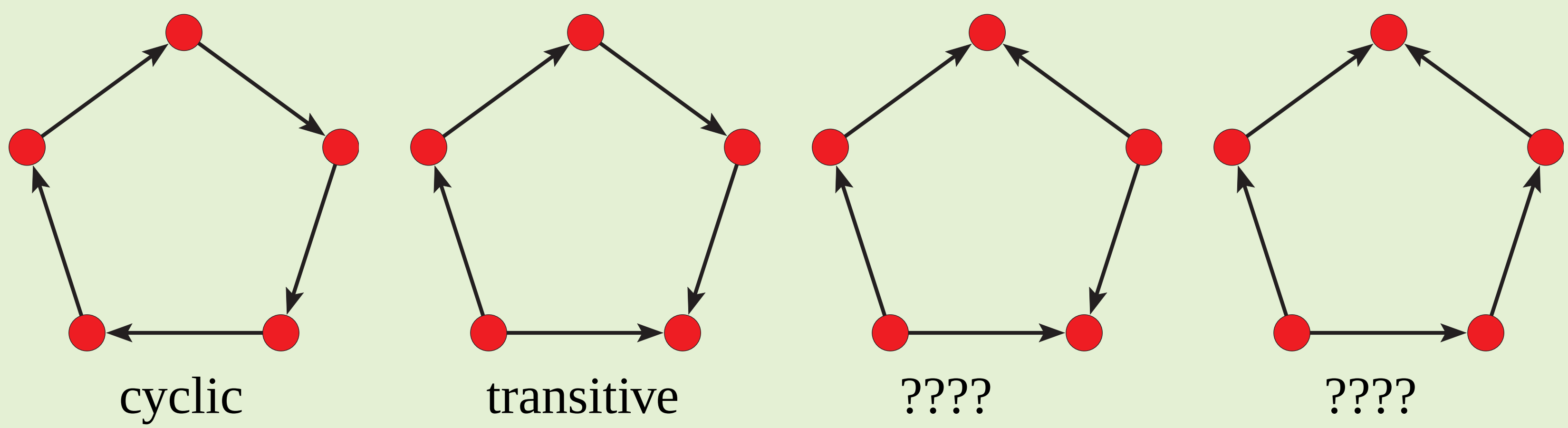
Charlie Brown

Example: Island for w_4 / Polizeiruf 110 and Starkes Team



Short cycle connectivity

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



These notions can be generalized to short cycle connectivity (see paper).

References

1. Batagelj, V. and Mrvar, A.(1996-): ***Pajek**– program for analysis and visualization of large network*, home page, data sets.

2. Batagelj, V. and Zaveršnik, M.(2002): *Generalized Cores*, arxiv cs.DS/0202039

3. Batagelj, V. and Zaveršnik, M.(2003): *Short cycle connectivity*. arxiv cs.DS/0308011

4. Dremelj, P., Mrvar, A., and Batagelj, V. (2002): *Analiza rodoslova dubrovačkog vlasteoskog kruga pomoču programa **Pajek***. *Anali Zavoda povij. znan. Hrvat. akad. znan. umjet. Dubr.* **40**, 105-126.

5. Fischer, M.D.: *Representing Anthropological Knowledge: Calculating Kinship*. http://www.era.anthropology.ac.uk/Era_Resources/Era/Kinship/prologTerm2.html

6. Mrvar, A. and Batagelj, V. (2004): *Relinking Marriages in Genealogies*. *Metodološki zvezki - Advances in Methodology and Statistics*, **1**, Ljubljana: FDV, 407-418.

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

8. White, D.R., Batagelj, V., and Mrvar, A. (1999): *Analyzing Large Kinship and Marriage Networks with Pgraph and Pajek*. *Social Science Computer Review – SSCORE*, **17**, 245-274.

9. Zaveršnik, M. and Batagelj, V. (2004): *Islands*. Slides from *Sunbelt XXIV, Portorož, Slovenia, 12.-16. May 2004*, PDF