MEDNARODNA PODIPLOMSKA ŠOLA JOŽEFA STEFANA
JOŽEF STEFAN INTERNATIONAL POSTGRADUATE SCHOOL

Social Network Analysis

Course: New Media and Knowledge Management

Lecturer: Prof. Dr. Nada Lavrač

Slide Design: Sergeja Sabo, David Fabjan and Miha Grčar

Slide review and correction: Jure Ferlež

Ljubljana, January 2007

Overview

1. Introduction to social network analysis
2. Domain: Co-authorship graph: ILPnet2
3. Cohesion
4. Brokerage
5. Ranking

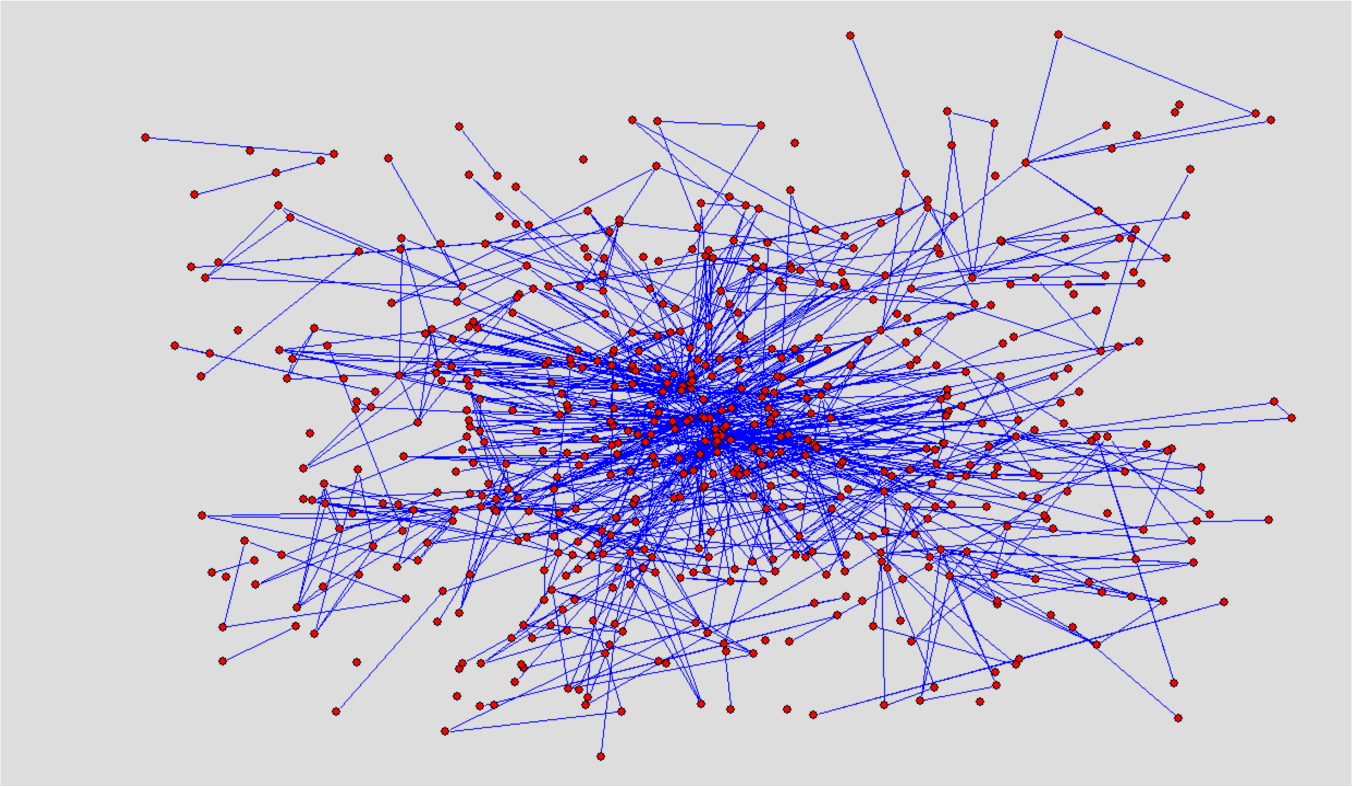
SOCIAL NETWORK ANALYSIS

- Social network analysis focuses on interpreting patterns of social ties among people, groups of people, organizations, and countries.
- A typical domain is a group of individuals and their characteristics and the structure of their ties.
- Goals**:**
  1. COHESION
     1. How many separate research groups do exits?
     2. How connected are the researchers with each other?
  2. BROKERAGE
     1. Who are the most influential authors?
     2. Who are the authors who are responsible for ideas trading across the community?
  3. RANKING
     1. Who are the most prestigeus authors in the area?
     2. Is there a census on who are the most important authors

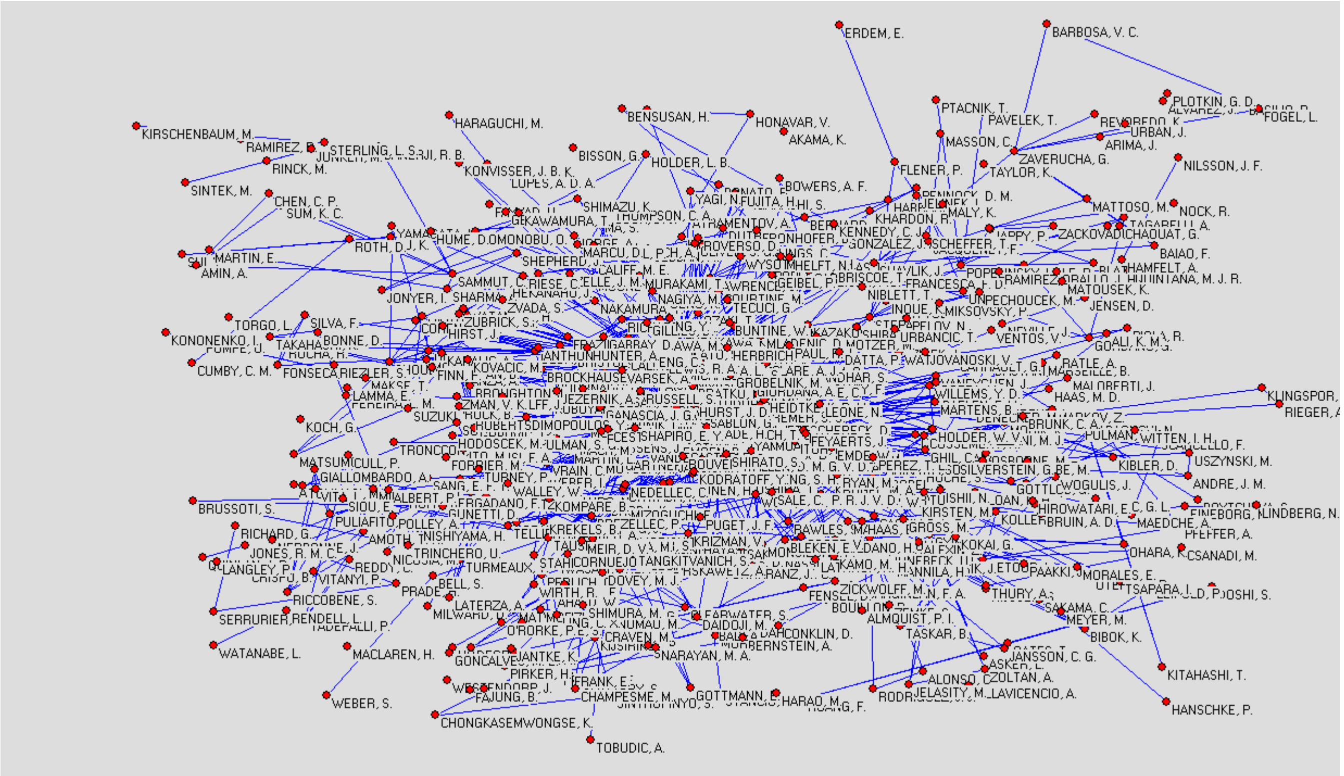
THE DOMAIN: ILPnet2

- Network of Excellence in Inductive Logic Programming (1998-2002)
- http://www.cs.bris.ac.uk/~ILPnet2/
- Basic characteristics: 589 authors, 1046 co-authorships, 1147 publications from 1970 to 2003

ILPnet2 network



ILPnet2 labeled network



COHESION

What is Cohesion?

1. Density
2. Degree
3. Components
4. Cores

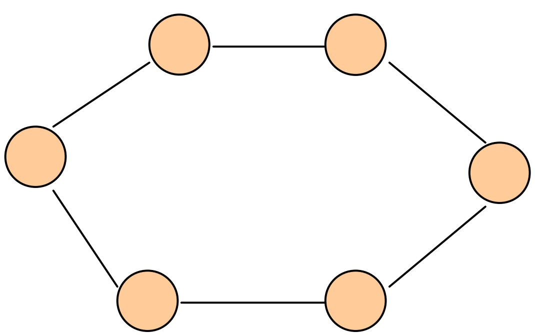
COHESION

- COHESION = an attractive “force” between individuals
- SOCIAL NETWORKS $\Rightarrow$ dense pockets of people who »stick together« = COHESIVE SUBGROUPS.
- The first concern of social network analysis $\Rightarrow$ to investigate who is related and who is not.

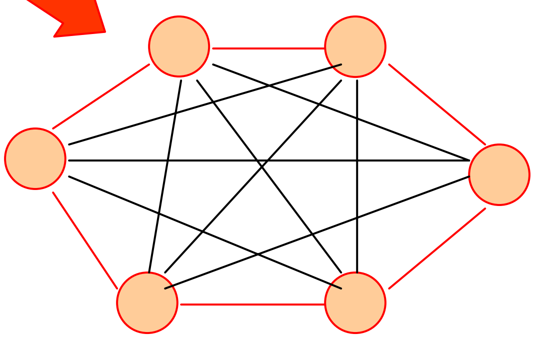
DENSITY

Density of the network = the number of lines in a simple network, expressed as a proportion of the maximum possible number of lines

- all possible lines = 15
- number of lines = 6



- **Density = 6/15 = 0.4**



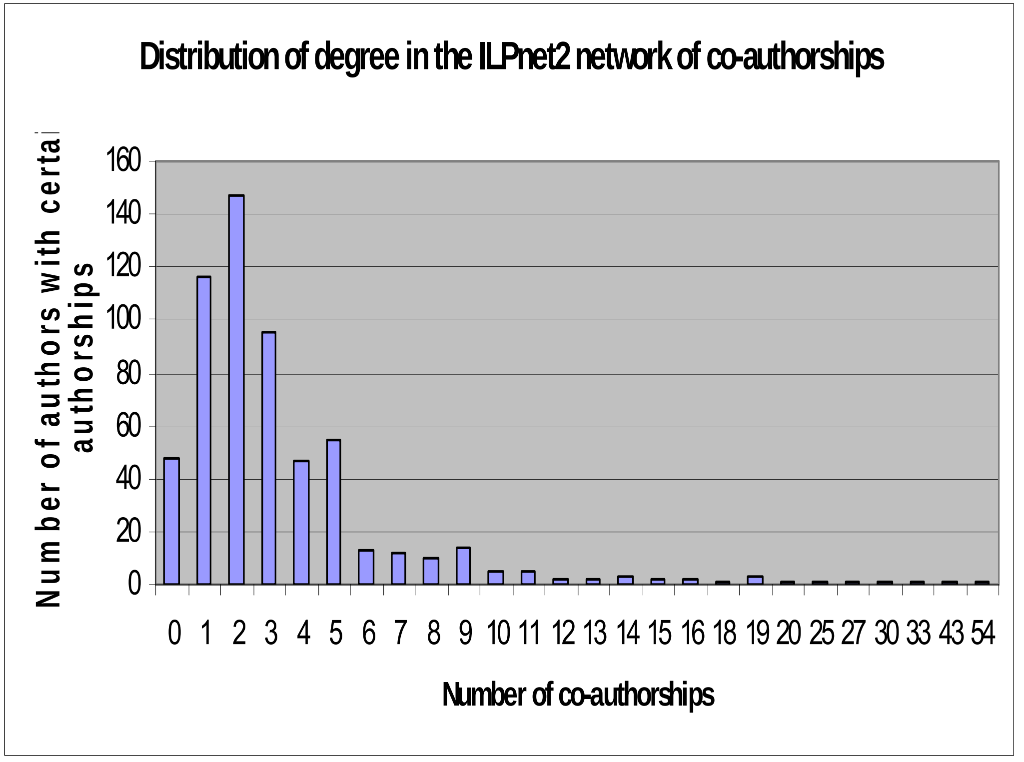
DENSITY

- inversely related to network size $\Rightarrow$ the larger the social network, the lower the density

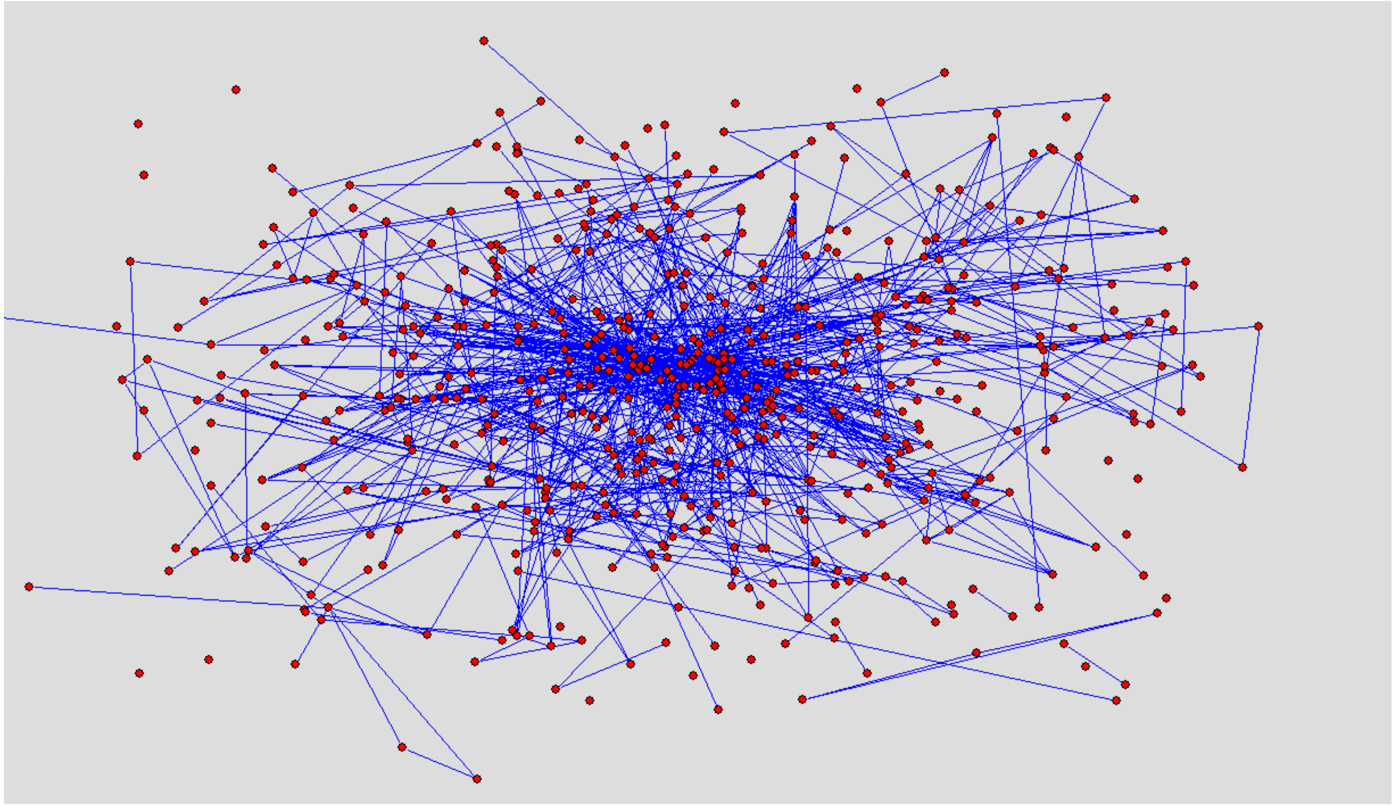
- ILPnet2 network Density = number of lines / maximum possible number of lines =
  = 1046 / 173166 = 0.0060

DEGREE

- A *degree* of a vertex = the number of lines incident with it.

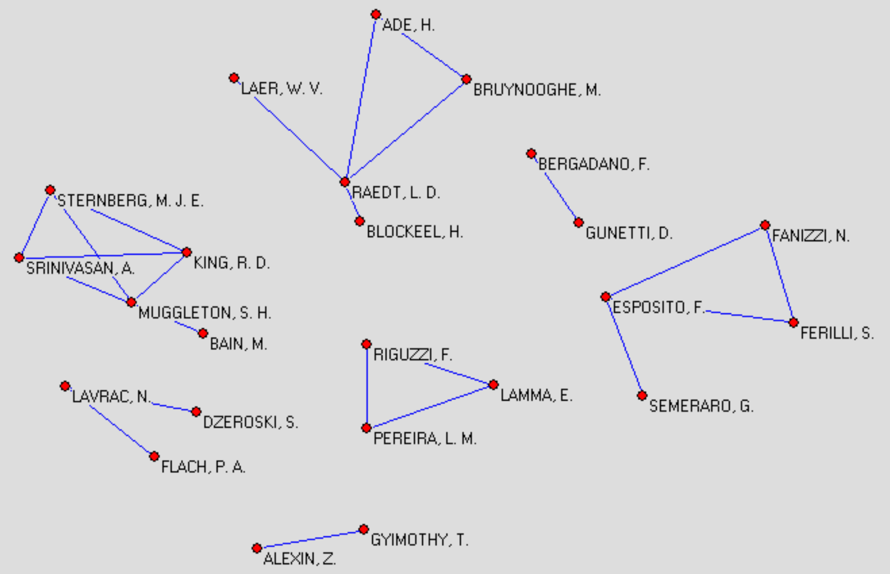


ILPnet2 network



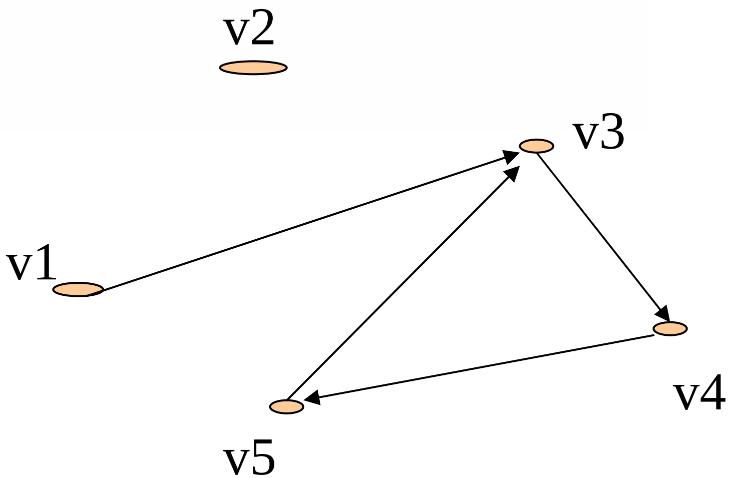
ILPnet2 network

- removed lines with value lower than 10 and reducted for vertices with degree lower than 1



WALKS AND PATHS

- semiwalk = we don`t consider the direction of the arcs (from v5→v3→v1)
- walk = we have to follow the directions of the arcs (v5→v3)
- semipath = semiwalk in which no vertex in between the first and last vertex of the semiwalk occurs more than once (v5→v3→v4→v5→v3)
- path = walk in which no vertex in between the first and last vertex of the walk occurs more than once (v5→v3)



ILPnet2 network is undirected

⇓

strongly/weakly connected network

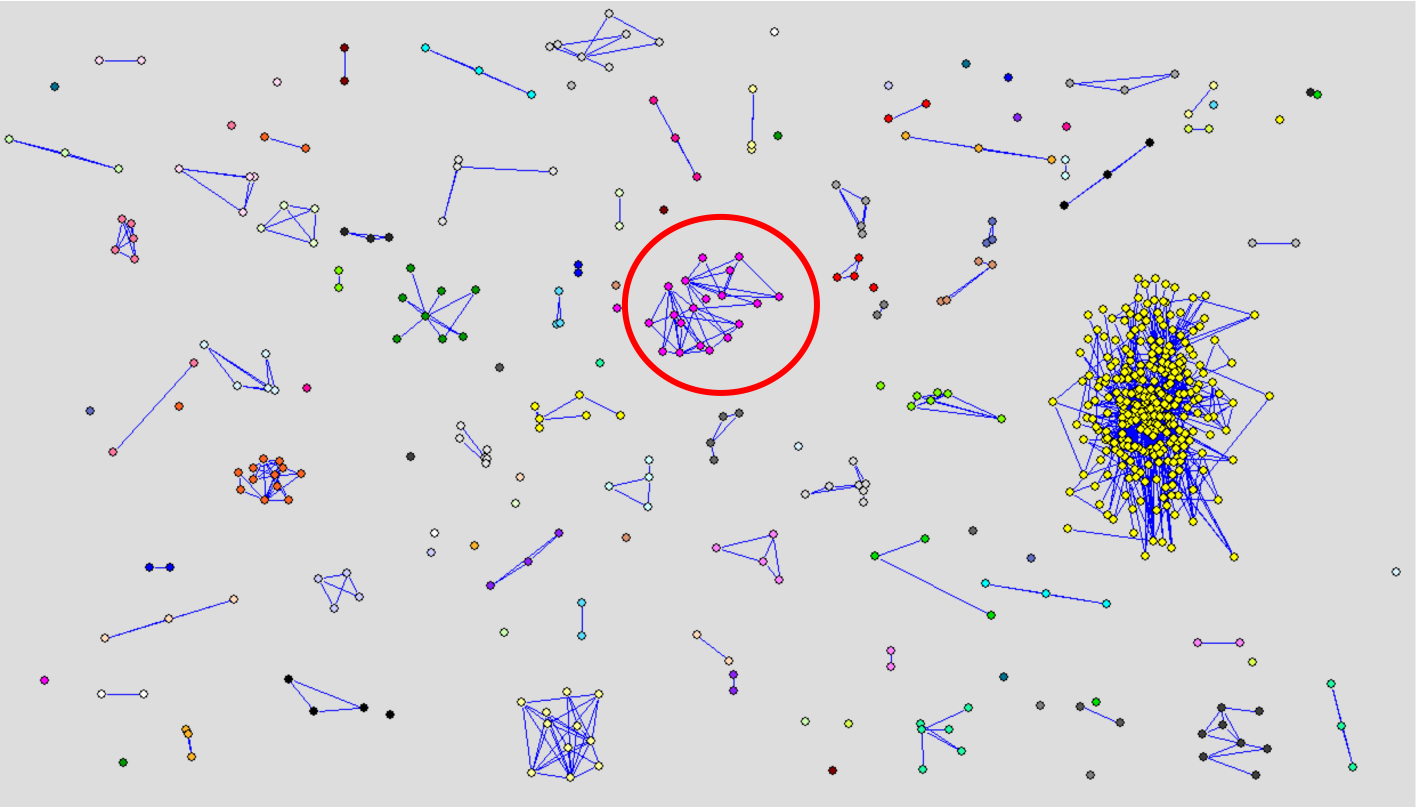
COMPONENTS

- Components identify cohesive subgroups in a straightforward manner - each vertex belongs to exactly one component.
- weakly connected networks = all vertices are connected by a semipath
- strongly connected networks = all vertices are connected by a path

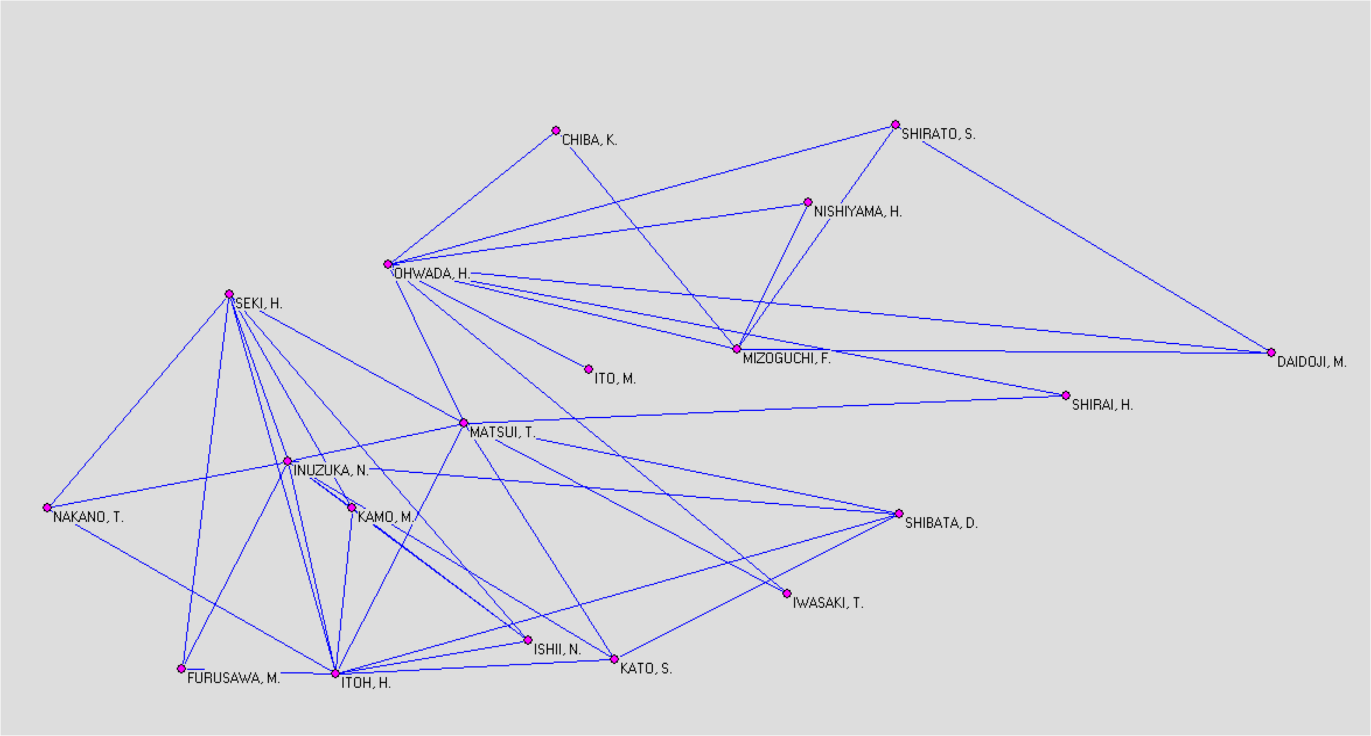
COMPONENTS

- A weak component is a maximal weakly connected subnetwork
- A strong component is a maximal strongly connected subnetwork.

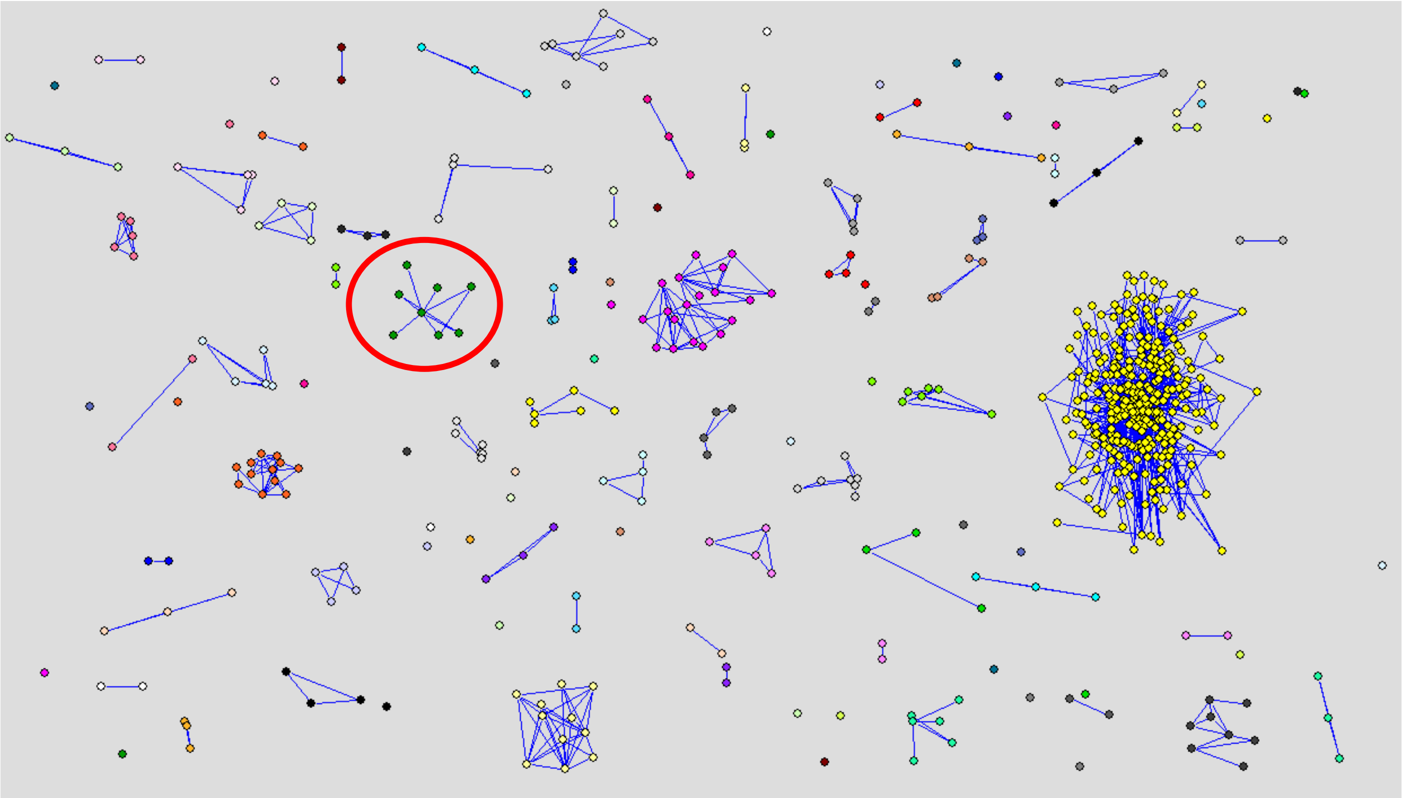
110 Components in ILPnet2 network



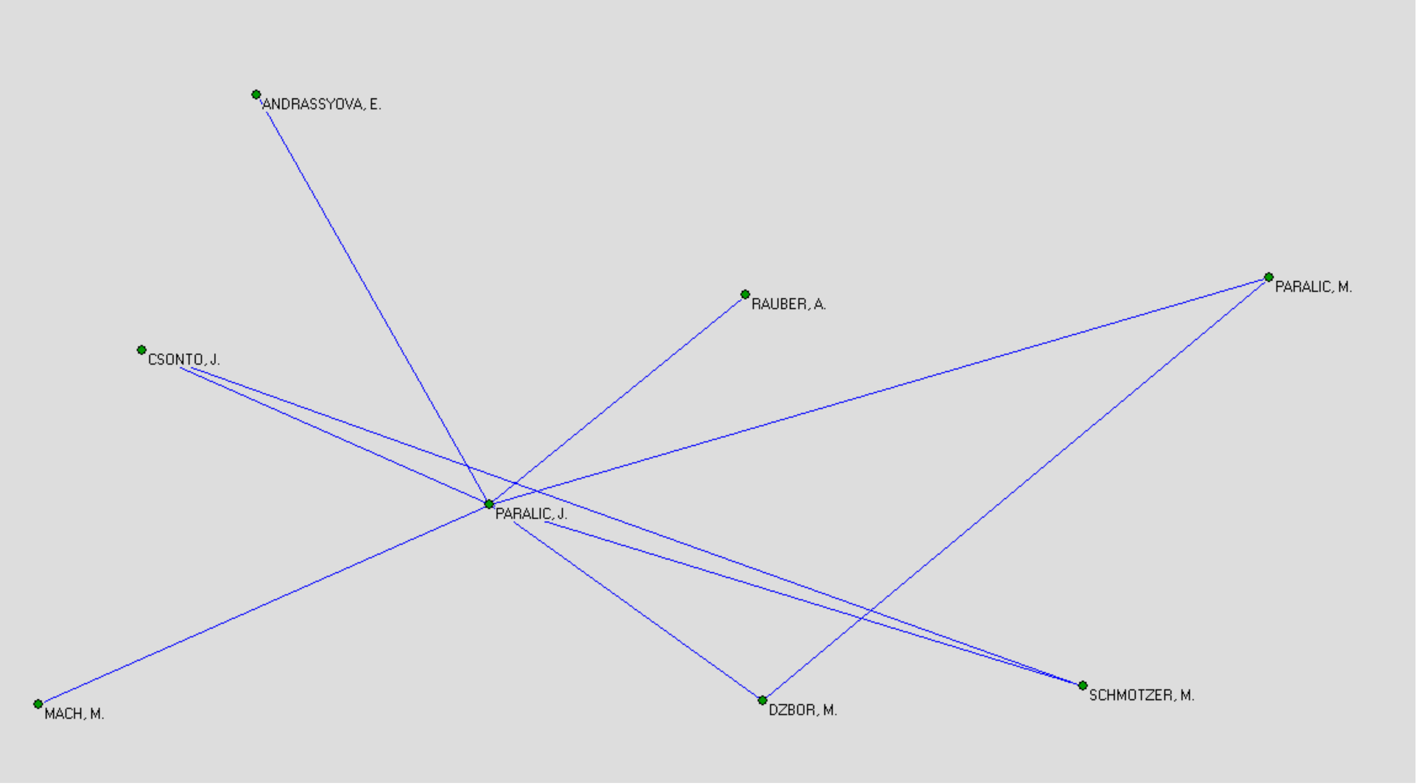
Zoomed component



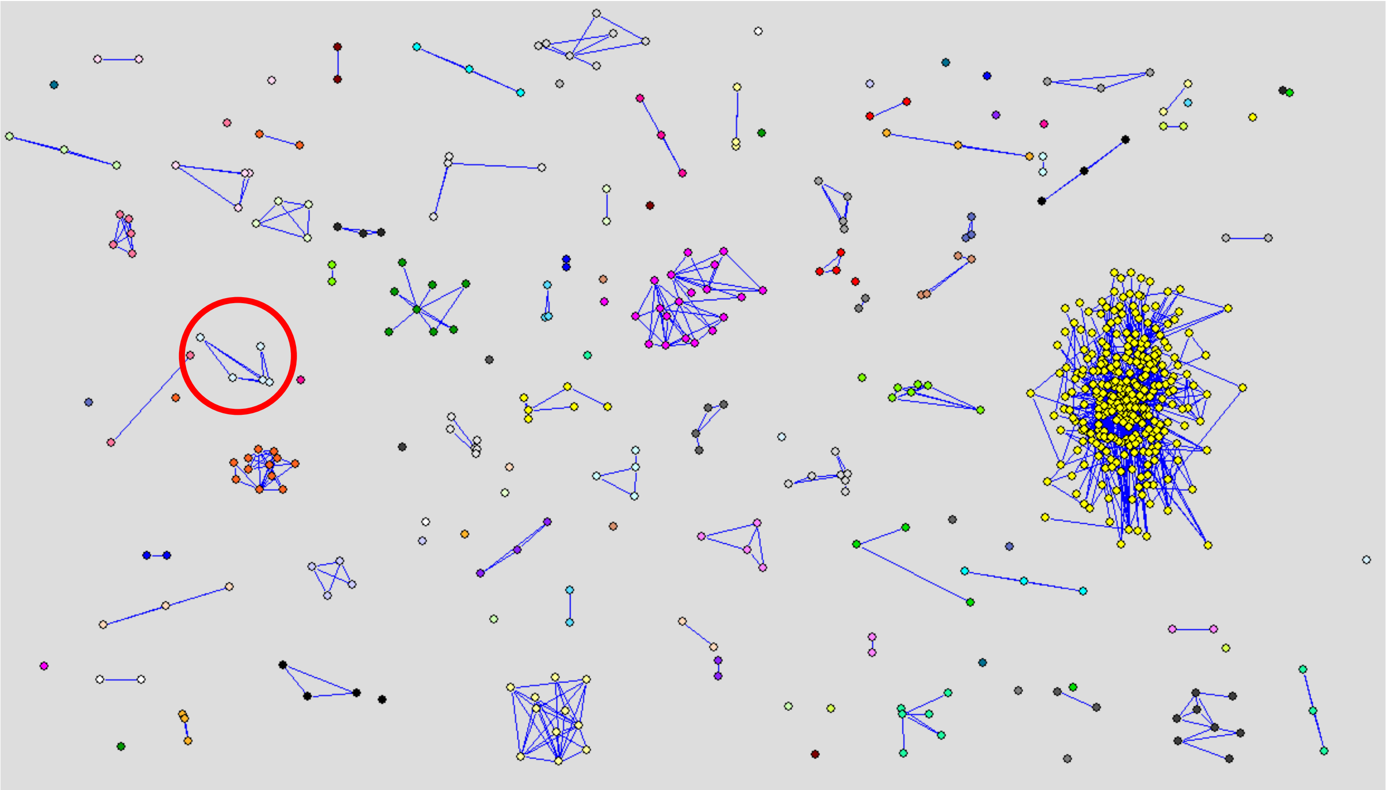
110 Components in ILPnet2 network



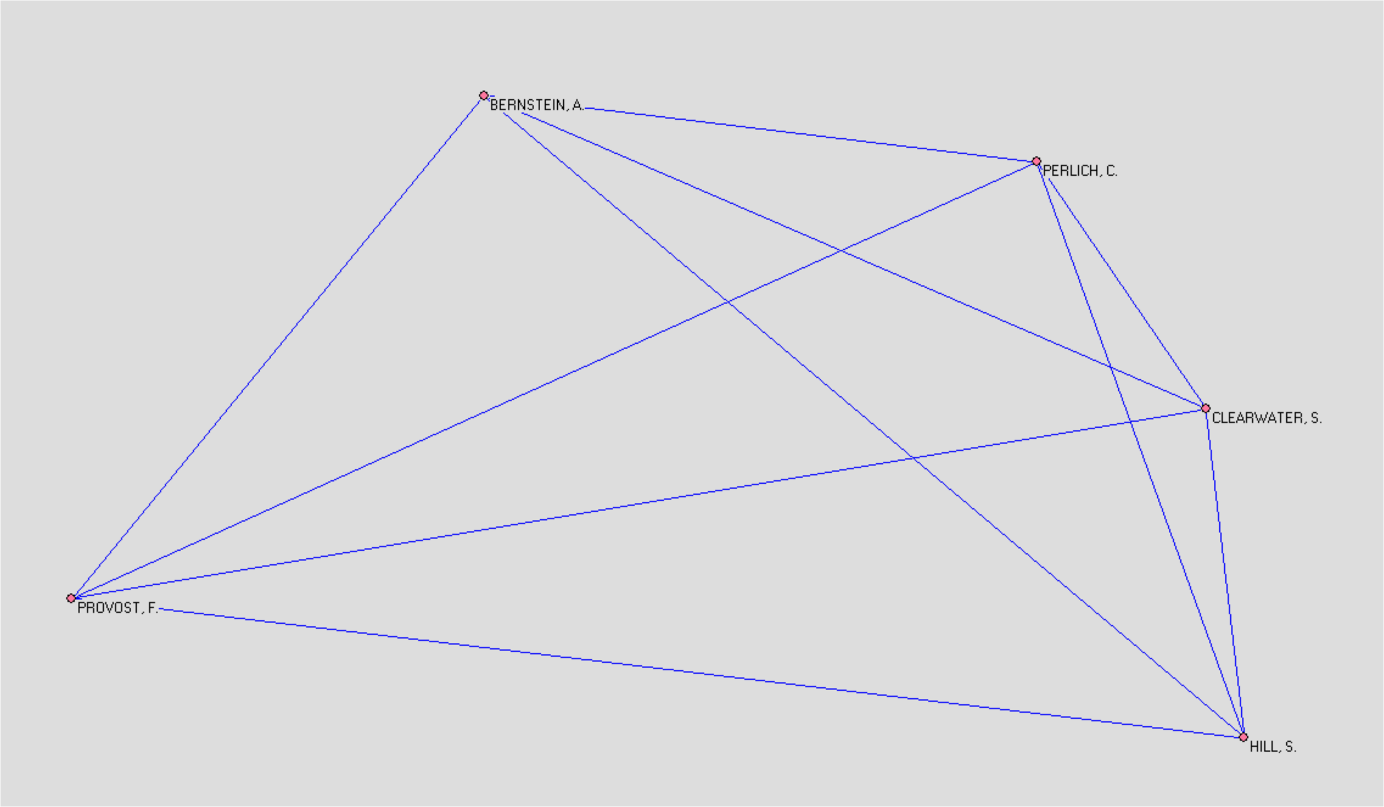
Zoomed component



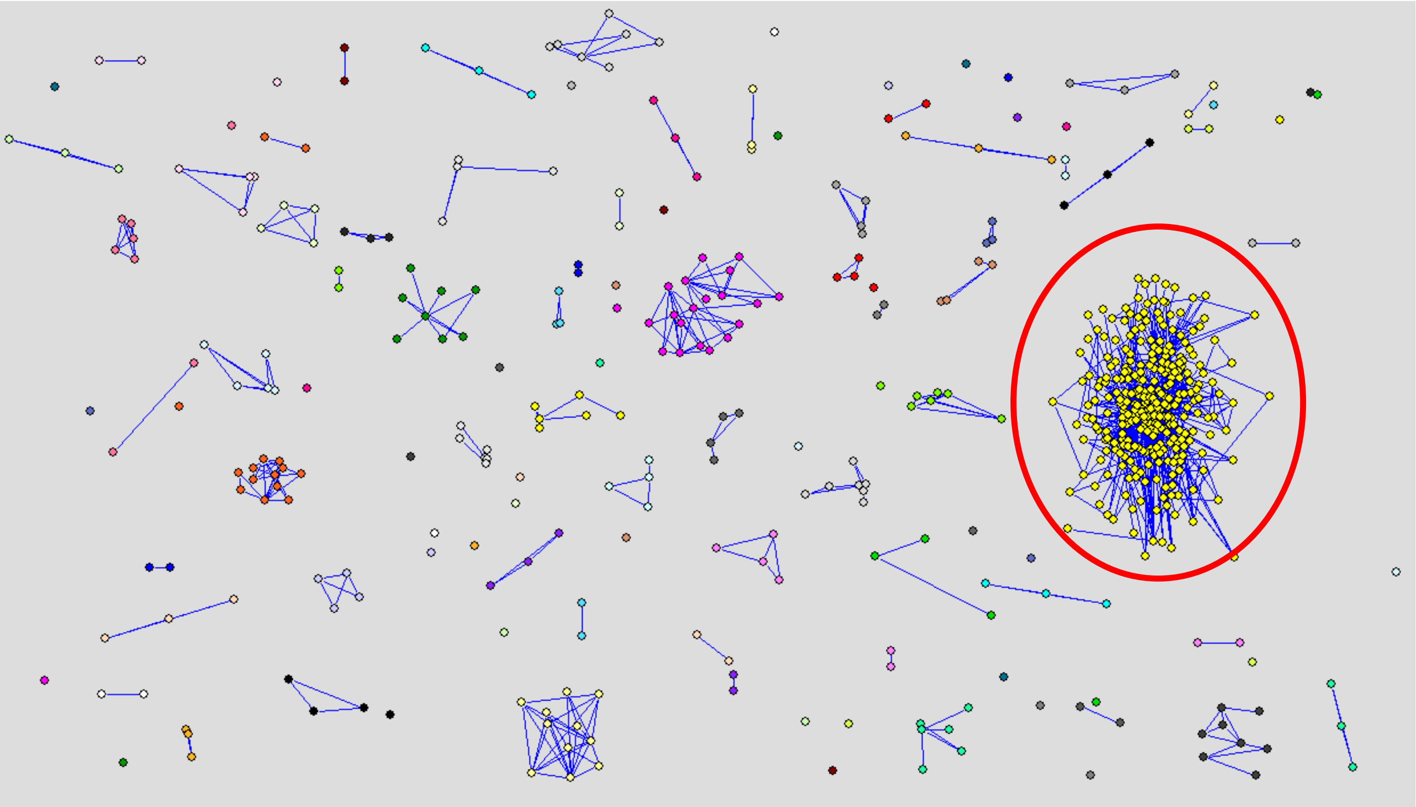
110 Components in ILPnet2 network



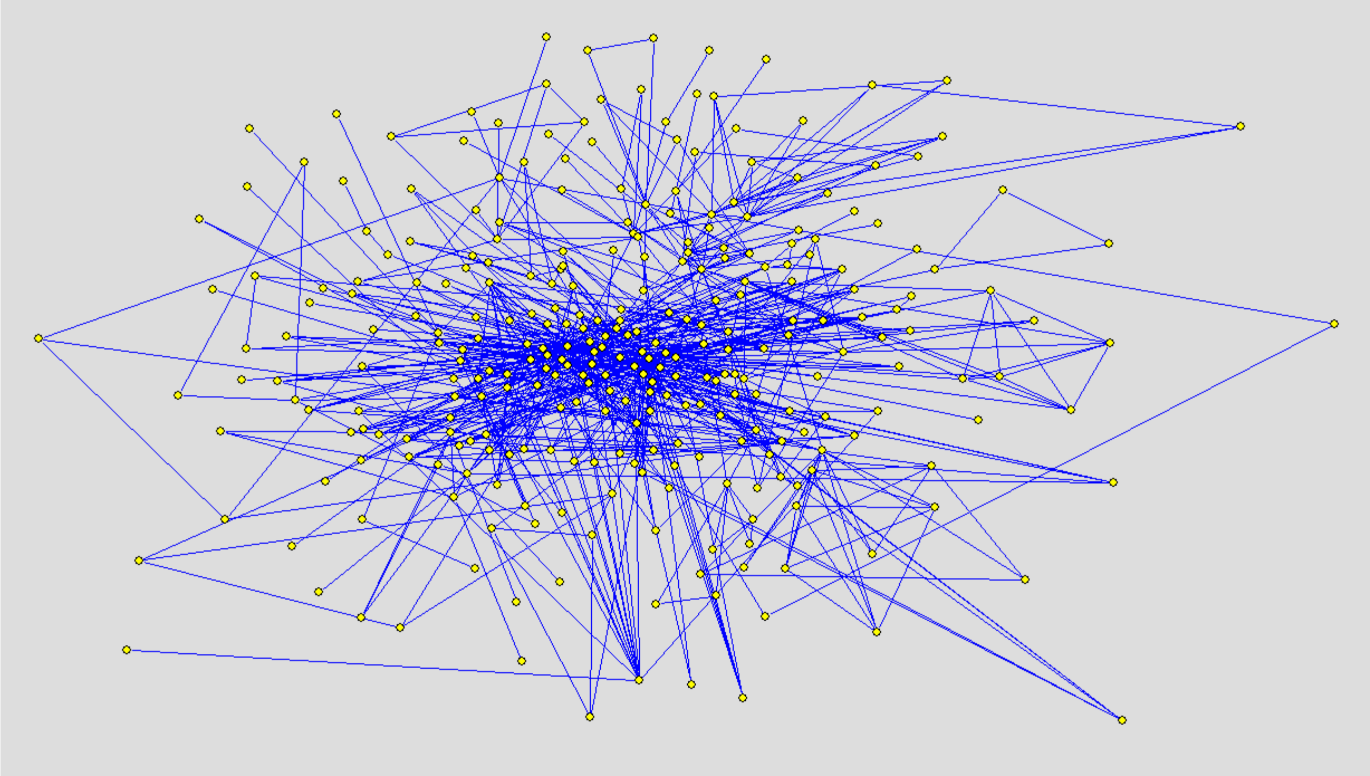
Zoomed component



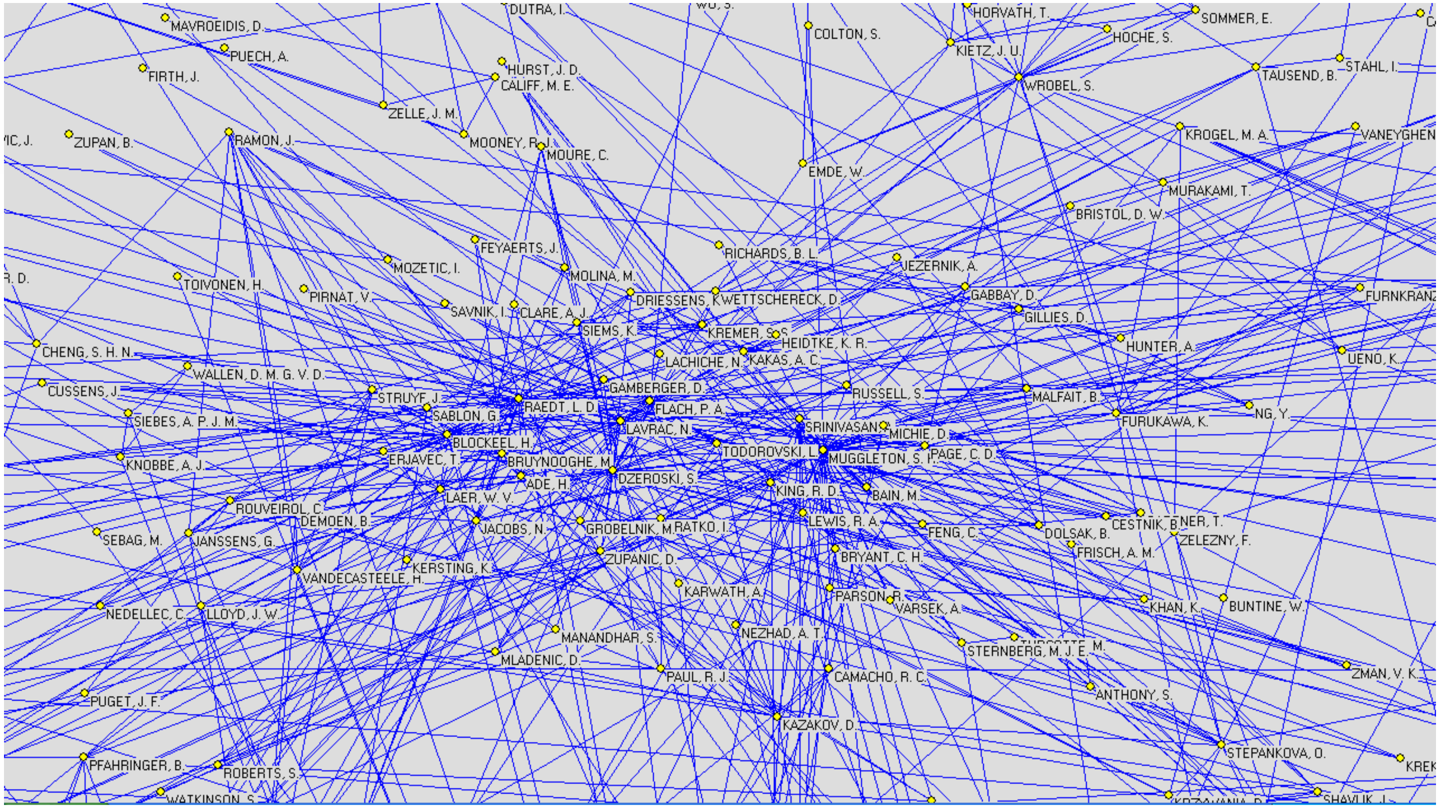
110 Components in ILPnet2 network



Zoomed component



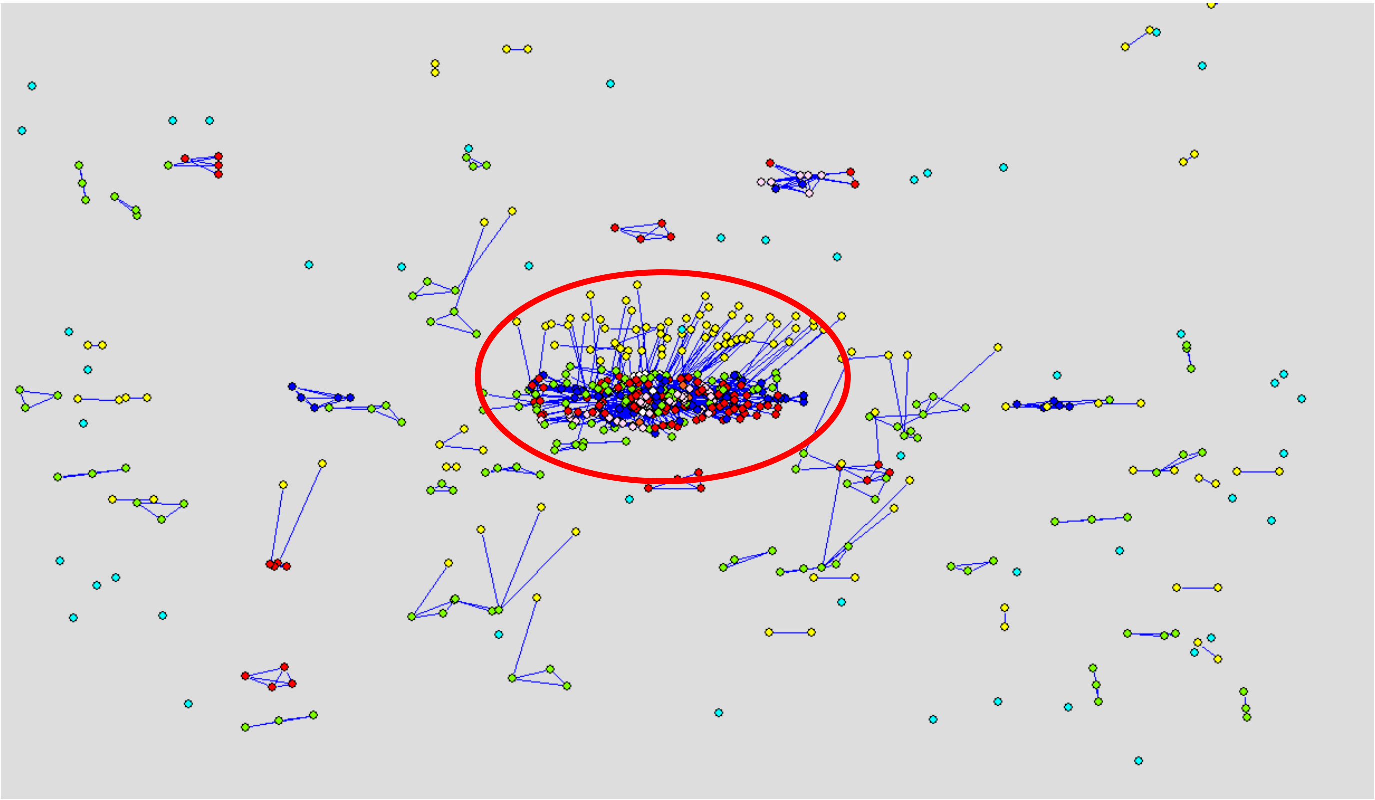
Zoomed component



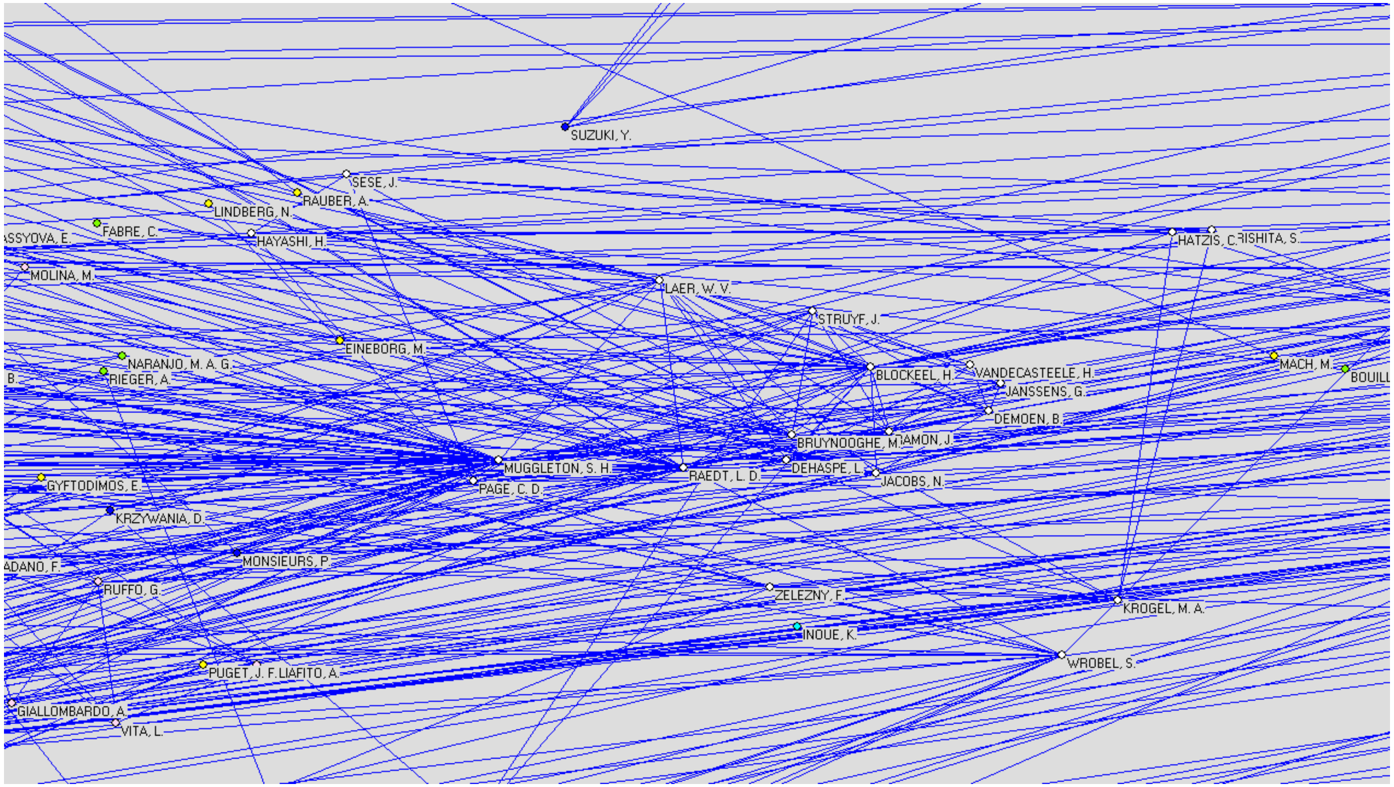
CORES

- A k-core is a maximal subnetwork in which each vertex has at least degree k within the subnetwork
- A *k*-core is not necessarily a cohesive group itself.
- Are the important people all present in the same component or are they scatered around?

ILPnet2 network with 7 cores – each color represents one core



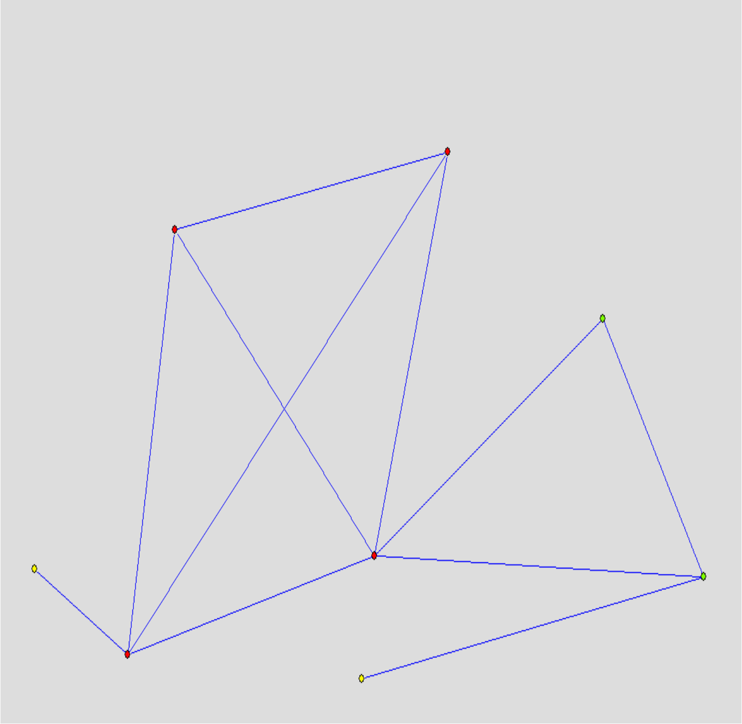
Zoomed *k*-core



CORES

Nested cores:

a vertex in a 3-core (red-colored dots) is also part of a 2-core (green-colored dot), but not all members of a 2-core belong to a 3-core



CLIQUES

- Homework:
  - What is a clique?
  - What is clique hierarchy?
  - What is the interpretation of a clique?

BROKERAGE

- **Center and Periphery:**
  - Nodes in the center have more information and are share the information easier
  - Centrality Measures
    - Degree centrality/centralization
    - Closeness centrality/centralization
    - Betweenness centrality/centralization
- **Brokers and Bridges**
  - Nodes which bridge structural holes are more important – Bridges
  - Brokers are bridges across many structural holes

Degree centrality/centralization reachability of a vertex inside a network

- The star network is the most efficient structure (given the fix number of lines)
- Network is more centralized if the vertices vary more with respect to their centrality. More variation in centrality scores of vertices yields a more centralized network.

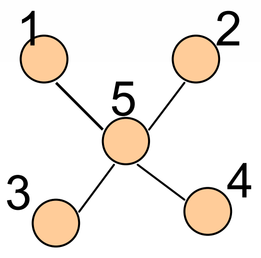
Defining degree of centralization

In social network: Who has the more sources of information at its disposal?

- The degree centrality of vertex is its degree
- Degree centralization of a network is the variation in the degrees of vertices divided by the maximum degree which is posible in the network of the same size

Degree centrality/centralization

a) star network (most efficient) degree centralizaton:



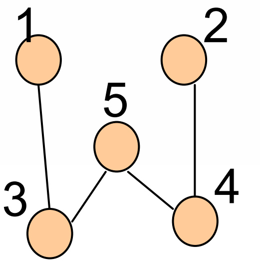
v5 degree = 4 (max degree)

v1 to v4 degree = 1 (min degree)

=> v5 contributes 1x (4-4) and v2 to v4 contributes 4x (4-1) => so **12 is the maximum degree** variations

=> 12/12 = 1 max degree centralization

b) line network:



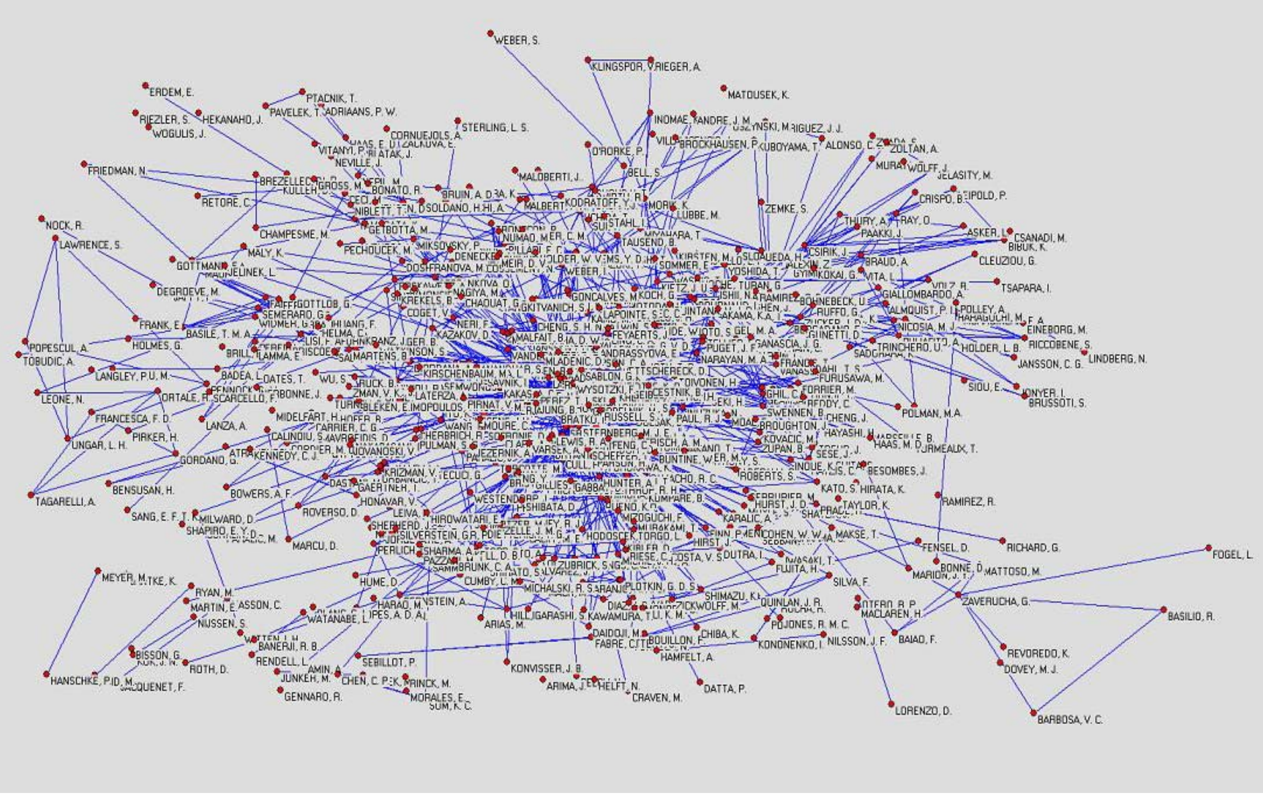
v1 and v2 degree = 1

v3, v4 and v5 degree = 2 (max degree in this network)

=> v1 and v2 contributes 2 x (2 – 1) and v3 to v5 contibutes 3 x (2 – 2)

=> 2 / 12 (**max degree in the network of the same size**) = 0,17

Degree centrality/centralization on IlpNet2 (all vertices)



Degree centrality/centralization on IlpNet2

- **Who are the most central persons in network; who has the most collaborations?**

- First we reduce number of vertices to those connected with min two neighbors
- Net > Transform > Reduction > Degree > All (min. Degree of vertices < 2)
- From 589 to 416 vertices

We remove people who wrote only one article by themselves or pairs of people that wrote one article together

Degree centrality/centralization on IlpNet2

Centralization of the network:

Net > Partitions > Degree > All

- All degree centrality of 2. All (recursive) degree reduction of N1 [2] (416)
- ---
- Working...
- --
- Network All Degree Centralization = 0.10282

Top 20 central persons in IlpNet2 (sorted using excel)

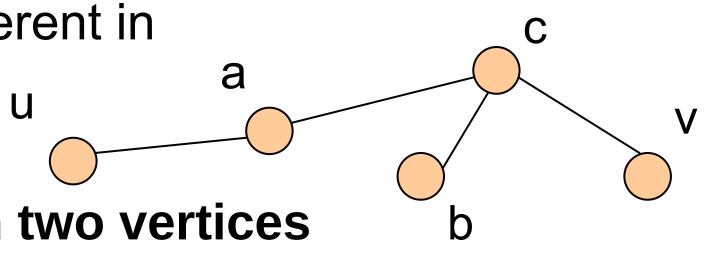
1. 0.1132530 - MUGGLETON,S.
2. 0.1036145 - DZEROSKI,S.
3. 0.0722892 - BLOCKEEL,H.
4. 0.0722892 - RAEDT,L.
5. 0.0650602 - LAVRAC,N.
6. 0.0481928 - FLACH,P.
7. 0.0457831 - LAER,W.
8. 0.0457831 - SRINIVASAN,A.
9. 0.0433735 - WROBEL,S.
10. 0.0433735 - BRUYNOOGHE,M.
11. 0.0409639 - PAGE,C.
12. 0.0385542 - KING,R.
13. 0.0385542 - JACOBS,N.
14. 0.0361446 - STEPANKOVA,O.
15. 0.0337349 - RAMON,J.
16. 0.0337349 - DEHASPE,L.
17. 0.0337349 - GYIMOTHY,T.
18. 0.0337349 - BERGADANO,F.
19. 0.0313253 - KAZAKOV,D.
20. 0.0289157 - ZUPANIC,D.

Distance – Geodesic

Two vertices (people) are connected if path exists form one to another

- In undirected network the distance is the number of lines or steps in the shortest path that connect two vertices together

- In directed network distance can be different in reverse way (one-way street example)

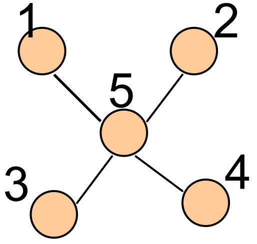


A geodesic is the shortest path between two vertices

- The distance from vertex *u* to vertex *v* is the length of the geodesic *u* to *v*.

Distance/Closeness – centrality/centralization

- The closeness centrality of a vertex is the number of all other vertices divided by the sum of all distances between the vertex and all others



v5: 4 / 4 =1

v1: 4 / 1+2+2+2 = 4 / 7

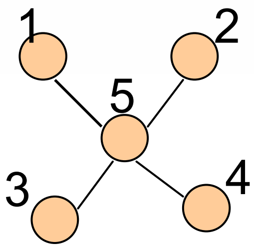
- We see that the problem arises if all vertices are not (strongly) connected!
- In social network: how easy is it for a person to notify the whole network with information.

Distance – centrality/centralization on IlpNet2

1. 0.3996198 - RAEDT,L.
2. 0.3758981 - DZEROSKI, S.
3. 0.3741894 - MUGGLETON, S.
4. 0.3415837 - LAER, W.
5. 0.3360069 - PAGE, C.
6. 0.3332861 - JACOBS,N.
7. 0.3279748 - LAVRAC, N.
8. 0.3215691 - WROBEL, S.
9. 0.3178443 - BLOCKEEL, H.
10. 0.3142049 - BRUYNOOGHE, M.
11. 0.3142049 - KAZAKOV, D.
12. 0.3106478 - DEHASPE, L.
13. 0.3071704 - FLACH, P.
14. 0.3037700 - RAMON, J.
15. 0.3015446 - CUSSENS, J.
16. 0.2993516 - BRATKO, I.
17. 0.2961211 - DRIESSENS, K.
18. 0.2919208 - WEBER, I.
19. 0.2898651 - MOURE, C.
20. 0.2898651 - MOLINA, M.

Betweenness centrality/centralization

- The betweenness centrality of a vertex is the proportion of all geodesics between pairs of other vertices that include this vertex



v5: 1

v1 to v4:0

- In social network : to what extent may a person (vertice) control the flow of informaton due to the his / her position inside the communication network?

Betweenness centrality/centralization on IlpNet2

Q: I discovered something new in the area, how likely is that the information spreading will get blocked and not the whole network will be informed

- Net > Vector > Centrality > Betweenness

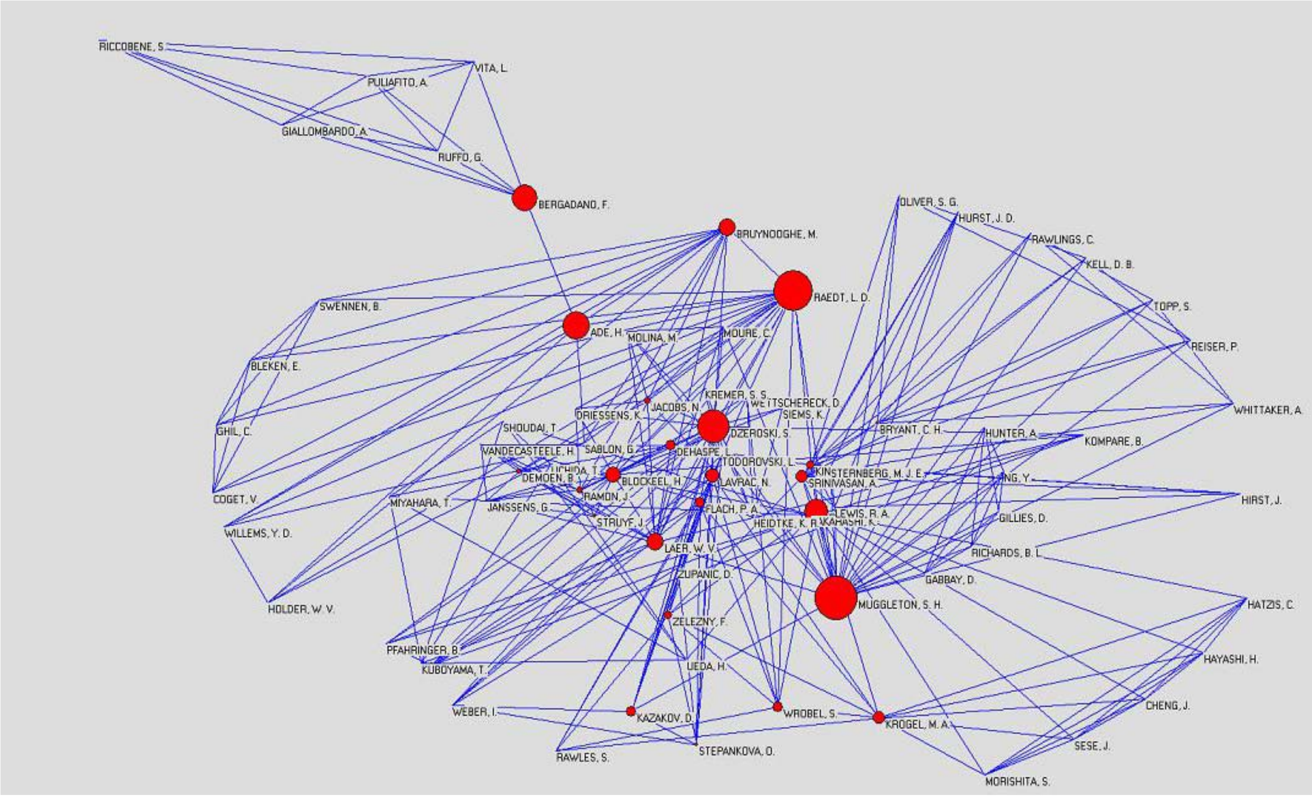
- Network Betweenness Centralization = 0.09198

Betweenness centrality/centralization on IlpNet2

- **This are top twenty persons with ability to control information spreading**

1. 0.0931813 - MUGGLETON, S.
2. 0.0742590 - RAEDT, L.
3. 0.0546139 - DZEROSKI, S.
4. 0.0459601 - WROBEL, S.
5. 0.0375969 - PAGE, C.
6. 0.0343720 - FLACH, P.
7. 0.0234424 - ADE, H.
8. 0.0218422 - BLOCKEEL, H.
9. 0.0192772 - LAVRAC, N.
10. 0.0181330 - STEPANKOVA, O.
11. 0.0178831 - ROUVEIROL, C.
12. 0.0170300 - BERGADANO, F.
13. 0.0157335 - BOSTROM, H.
14. 0.0153786 - FURUKAWA, K.
15. 0.0152876 - BAIN, M.
16. 0.0143174 - GYIMOTHY, T.
17. 0.0119656 - SHAVLIK, J.
18. 0.0110439 - CHENG, S.
19. 0.0107273 - SRINIVASAN, A.
20. 0.0106385 - LAER, W.

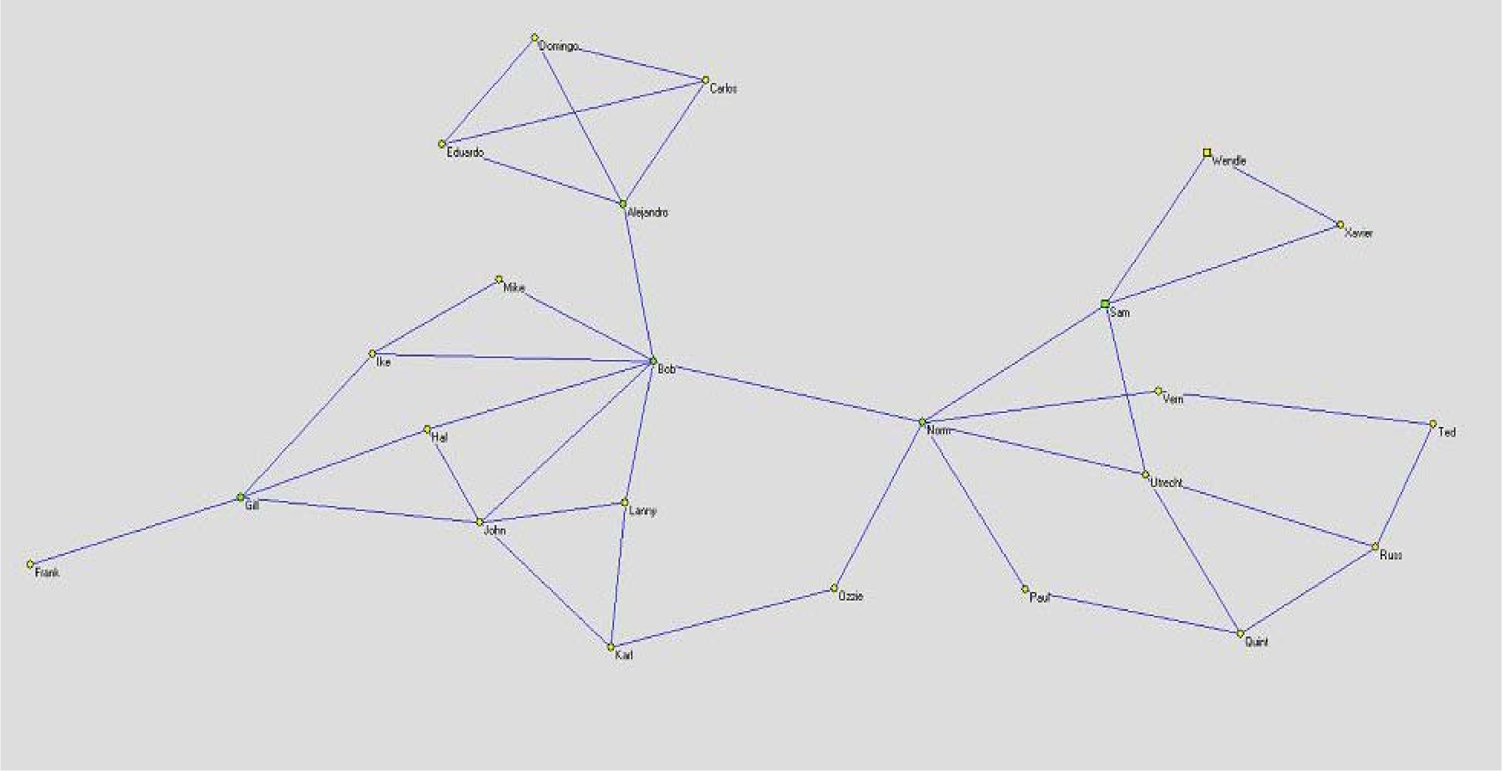
Betweenness centrality/centralization on IlpNet2 using “Pajek”,



Bridges

- The bridges and lines who bridge structural holes between other have more control and perform better
- A bridge is a line whose removal increases the number of components in the network
- Deleting a vertex from a network means that the vertex and all lines incident with this vertex are removed from the network
- A cut-vertex is a vertex whose deletion increases the number of components in the network
- A bi-component is a component of minimum size of three that does not contain a cut-vertex

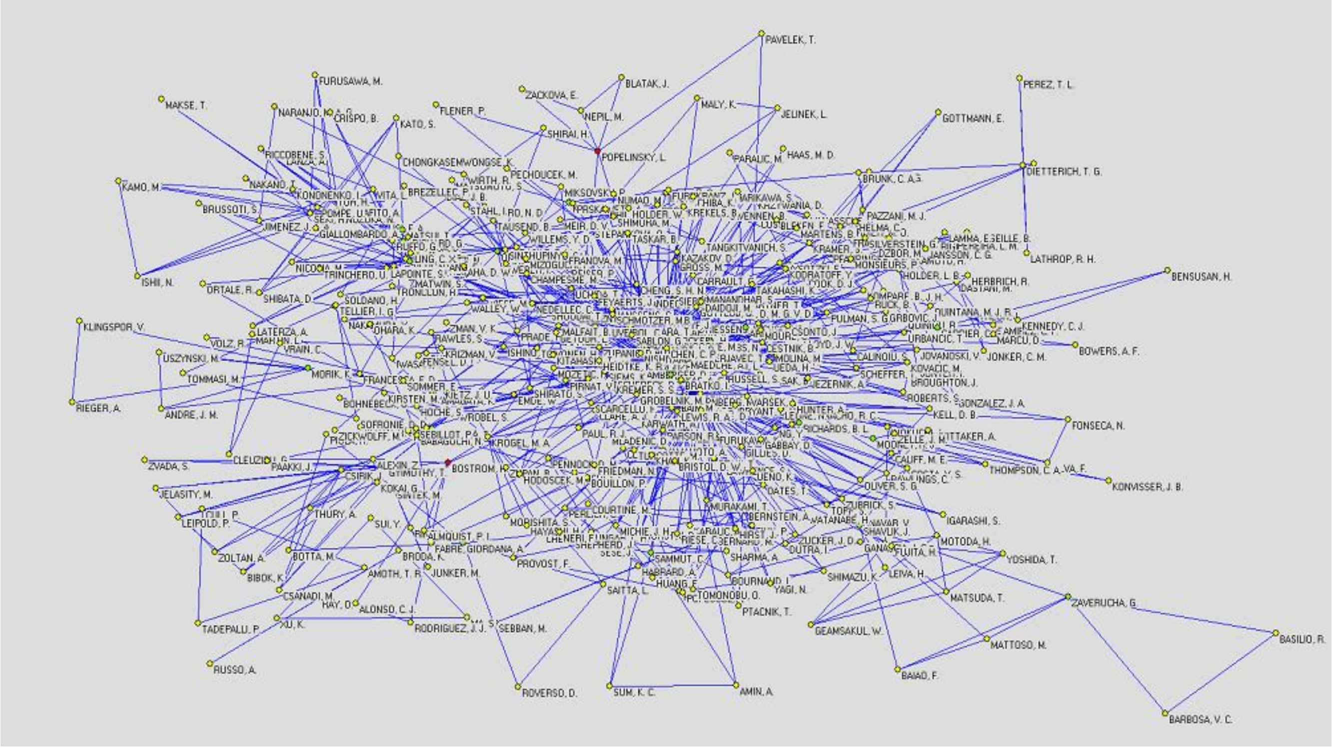
Bridges
simple example



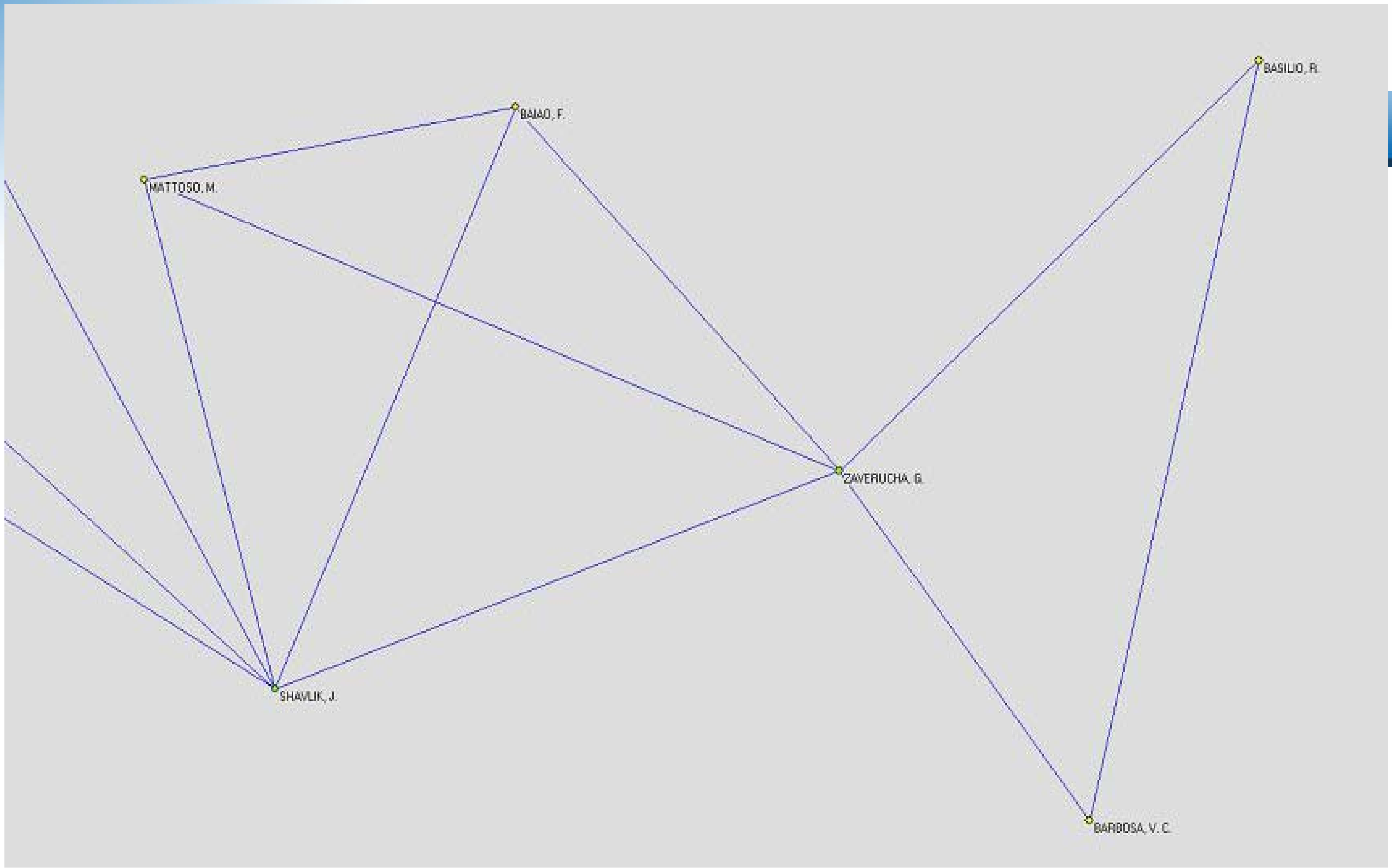
Bridges on IlpNet2

- Who are the bridges and lines in IlpNet2 who bridge structural holes
- Net > Components > Bi-Components (with a minimum size of 2 so we can look for lines that represents bridges)

Bridges
on IlpNet2



Bridges



Bridges
IlpNet2

- **Who are the bridges and lines in IlpNet2 who bridge structural holes, articles that two persons work together on?**
- Net > Components > Bi-Components (with a minimum size of 2 so we can look for lines that represents bridges)

- Root (449)
- 1 (3)
- 2 (3)
- 3 (11)
- 4 (4)
- 5 (3)
- 6 (4)
- ...
- 25(2)
- 26(2)

1. 11 - MUGGLETON, S.
2. 7 - FLACH, P.
3. 6 - BOSTROM, H.
4. 6 - CHENG, S.
5. 5 - ROUVEIROL, C.
6. 5 - PARALIC, J.
7. 4 - ZAVERUCHA, G.
8. 4 - VRAIN, C.
9. 4 - RAEDT, L.
10. 4 - PAGE, C.
11. 4 - SAMMUT, C.
12. 3 - POPELINSKY, L
13. 3 - OHWADA, H.
14. 3 - KIETZ, J.
15. 3 - SEBAG, M.
16. 3 - KRAMER, S.
17. 3 - FURUKAWA, K.
18. 3 - KAKAS, A.
19. 3 - GIORDANA, A.
20. 3 - MORIK, K.
21. 3 - PAZZANI, M.
22. 2 - RIGUZZI, F.
23. 2 - WROBEL, S.
24. 2 - HORVATH, T.
25. 2 - TURAN, G.

- Bridges are bi components of size two in an undirected network, so we can easily find them

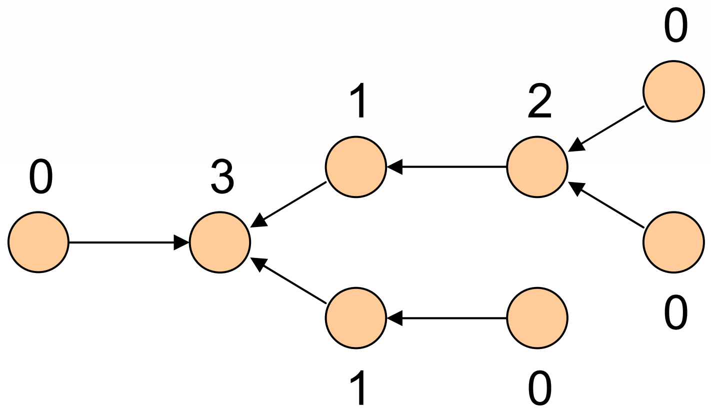
RANKING

- I. Prestige
  - Structural prestige, social prestige, correlation
  - Ways of calculating structural prestige
- II. Ranking
  - Triad census
  - Acyclic decomposition
  - Symmetric-acyclic decomposition

I. Prestige

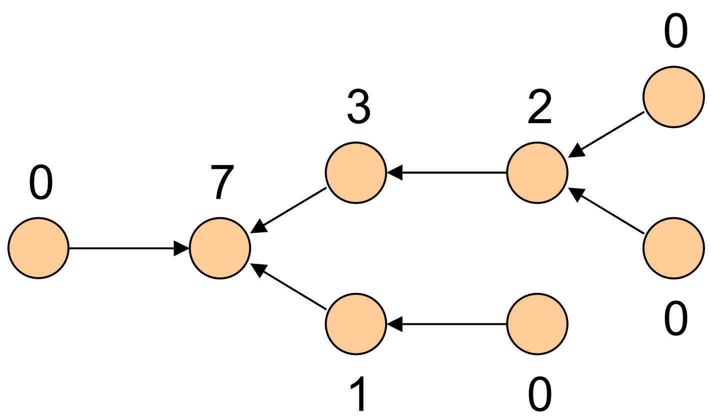
- Prestigious people
  - People who receive many positive in-links
- Structural prestige measures
  - Popularity or in-degree
  - (Restricted) input domain
  - Proximity prestige
- Structural prestige ≠ social prestige (social status)
- Correlation between structural and social prestige
  - Pearson’s correlation coefficient
  - Spearman’s rank correlation coefficient

Popularity or in-degree



Input domain

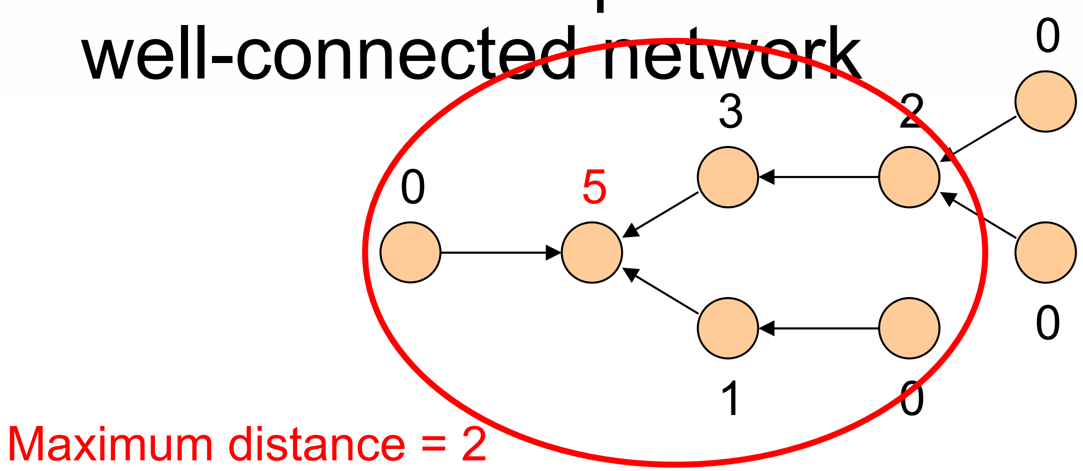
- Input domain size
  - How many nodes are path-connected to a particular node?
- Overall structure of the network is taken into account



- Problematic in a well-connected network

Restricted input domain

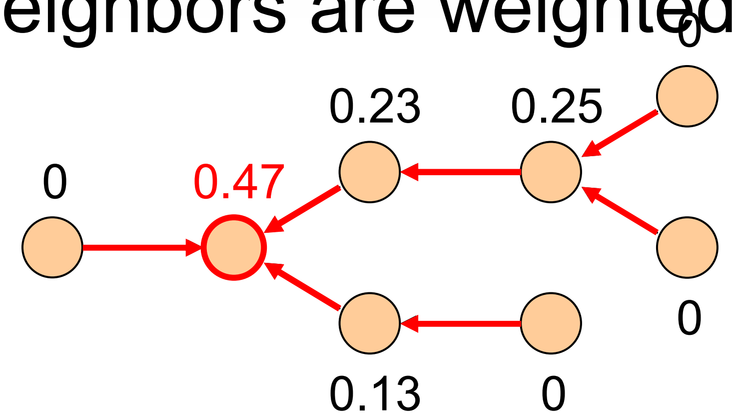
- Resolves the input-domain issue in a well-connected network



- Issue: the choice of the maximum distance is quite arbitrary

Proximity prestige

- Eliminates the maximum-distance parameter
- Closer neighbors are weighted higher



$$\text{Proximity prestige} = \frac{\text{Input domain size / Number of nodes} \quad 7/8}{\text{Average path-distance to the node} \quad (3+3+2+2+1+1+1)/7} = 0.47$$

Structural prestige
ILPnet2 dataset top 25

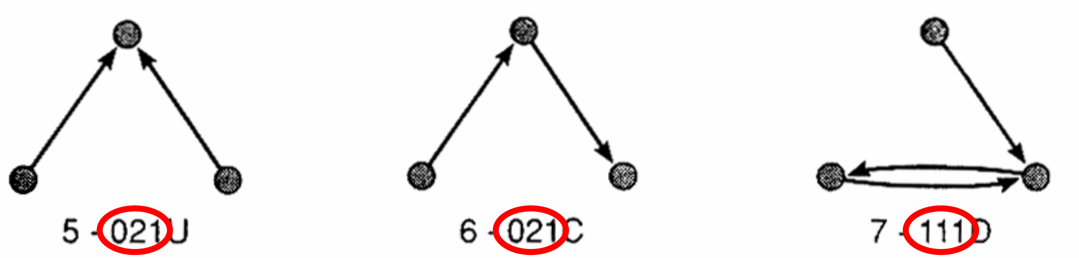
Input degree		Unrestricted input domain size		Proximity prestige	
28	MUGGLETON, S. H.	152	LAMMA, E.	0.082030307	RAEDT, L. D.
21	RAEDT, L. D.	152	RIGUZZI, F.	0.077044151	DZEROSKI, S.
20	DZEROSKI, S.	152	PEREIRA, L. M.	0.068453862	LAVRAC, N.
17	LAVRAC, N.	152	RAMON, J.	0.066777042	MUGGLETON, S. H.
17	BLOCKEEL, H.	152	FLACH, P. A.	0.064946309	ADE, H.
12	FLACH, P. A.	152	LAVRAC, N.	0.06462585	BRUYNOOGHE, M.
12	SRINIVASAN, A.	152	STRUYF, J.	0.063683172	LAER, W. V.
11	GYIMOTHY, T.	152	BLOCKEEL, H.	0.060918631	TODOROVSKI, L.
10	JACOBS, N.	152	DEHASPE, L.	0.057783113	FLACH, P. A.
10	BERGADANO, F.	152	LAER, W. V.	0.054504505	SRINIVASAN, A.
9	WROBEL, S.	152	BRUYNOOGHE, M.	0.054346497	GAMBERGER, D.
9	STEPANKOVA, O.	152	DZEROSKI, S.	0.052812523	SABLON, G.
9	ITOH, H.	152	RAEDT, L. D.	0.051974229	DEHASPE, L.
9	ADE, H.	152	GAMBERGER, D.	0.051837094	BLOCKEEL, H.
8	KING, R. D.	152	LACHICHE, N.	0.048245614	KING, R. D.
8	OHWADA, H.	152	TODOROVSKI, L.	0.048015873	STERNBERG, M. J. E.
8	BRUYNOOGHE, M.	152	KAKAS, A. C.	0.047743034	KAKAS, A. C.
8	BOSTROM, H.	152	JOVANOSKI, V.	0.047283414	LACHICHE, N.
8	KRAMER, S.	152	TURNEY, P.	0.044957113	JOVANOSKI, V.
8	FURUKAWA, K.	152	ADE, H.	0.044957113	TURNEY, P.
8	CSIRIK, J.	152	DIMOPOULOS, Y.	0.043609897	RAMON, J.
7	HORVATH, T.	152	SABLON, G.	0.043226091	STRUYF, J.
7	ESPOSITO, F.	77	KING, R. D.	0.040507749	RIGUZZI, F.
7	SHOUDAI, T.	77	MUGGLETON, S. H.	0.040341393	DIMOPOULOS, Y.
7	DEHASPE, L.	77	SRINIVASAN, A.	0.035082604	LAMMA, E.

II. Ranking

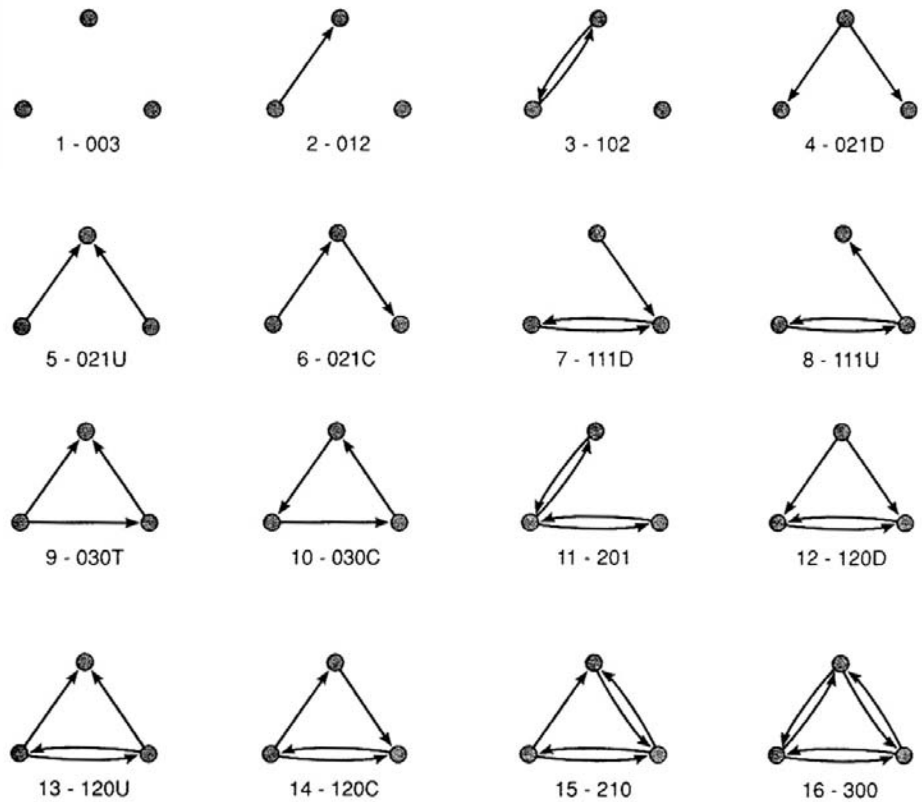
- We discuss techniques to extract discrete ranks from social relations
- Triad analysis helps us determine if our network is biased towards…
  - Unrelated clusters (cluster = clique)
  - Ranked clusters
  - Hierarchical clusters
- Recipes for determining the hierarchy
  - Acyclic decomposition
  - Symmetric-acyclic decomposition

Triad analysis

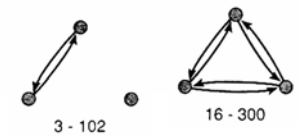
- Triads
  - Atomic network structures (local)
  - 16 different types
  - M-A-N naming convention
    - **M**utual positive
    - **A**symmetric
    - **N**ull



All 16 types of triads



Triad census

- 6 balance-theoretic models
  - Balance — Restricted to 

  - Clusterability — Unrelated clusters
  - Ranked clusters — Ranked clusters
  - Transitivity — Less and less restricted
  - Hierarchical clusters — Hierarchical clusters
  - (Theoretic model) — Unrestricted
- **Triad census**: triads found in the network, arranged by the balance-theoretic model to which they belong

Triad census
ILPnet2 dataset

Type	Number of triads (ni)	Expected (ei)	(ni-ei)/ei	Model
3 - 102	247225	1292.72	190.24	Balance
16 - 300	112	0.00	1539118270.84	Balance
1 - 003	33404551	33159112.00	0.01	Clusterability
4 - 021D	36	1292.72	-0.97	Ranked Clusters
5 - 021U	1176	1292.72	-0.09	Ranked Clusters
9 - 030T	39	9.32	3.18	Ranked Clusters
12 - 120D	91	0.02	5415.95	Ranked Clusters
13 - 120U	83	0.02	4939.74	Ranked Clusters
2 - 012	228351	717207.24	-0.68	Transitivity
14 - 120C	1	0.03	28.76	Hierarchical Clusters
15 - 210	64	0.00	528411.66	Hierarchical Clusters
6 - 021C	182	2585.44	-0.93	Forbidden
7 - 111D	719	9.32	76.14	Forbidden
8 - 111U	63	9.32	5.76	Forbidden
10 - 030C	0	3.11	-1.00	Forbidden
11 - 201	121	0.02	7201.76	Forbidden

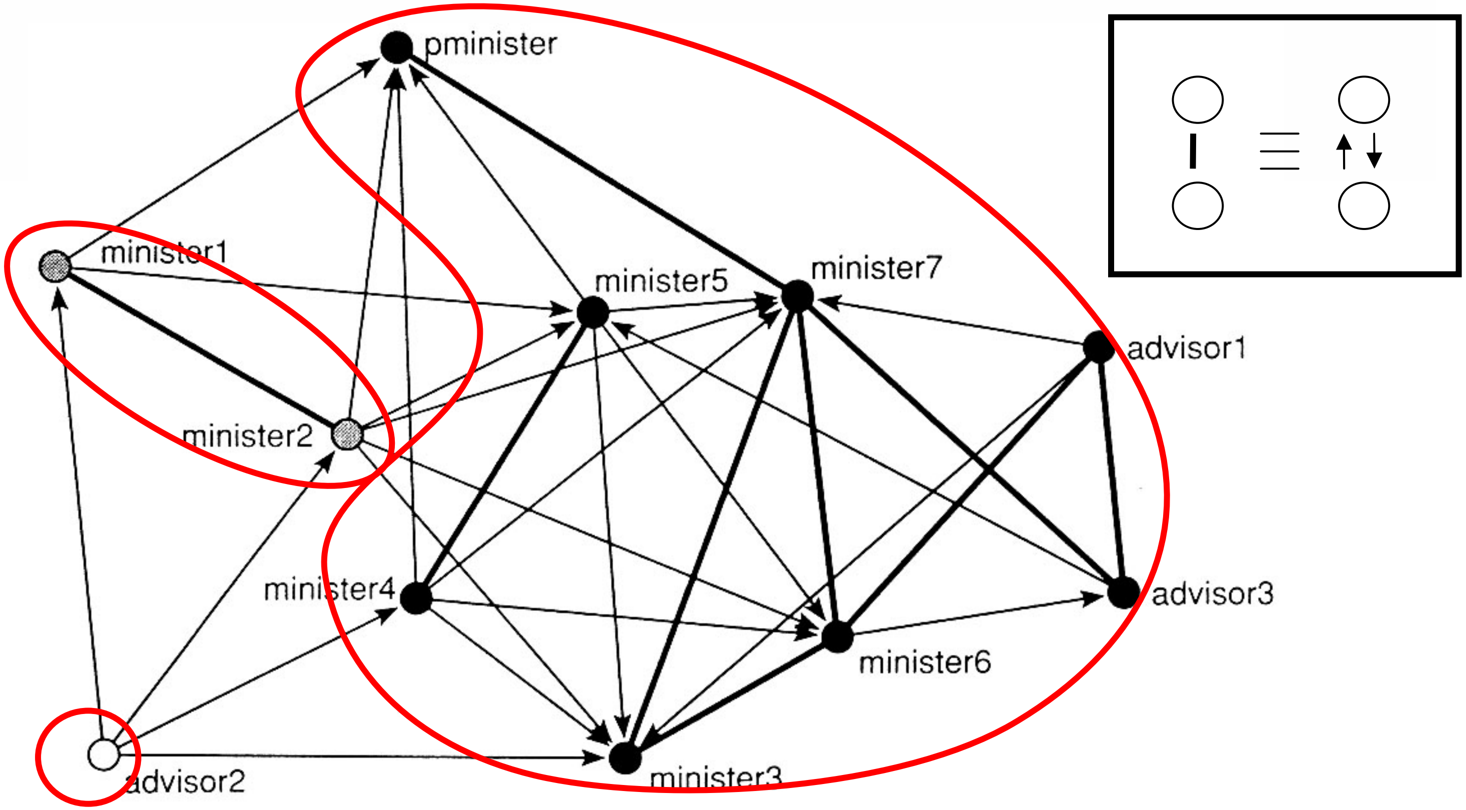
Chi-Square: 172464018511.5997***
7 cells (43.75%) have expected frequencies less than 5.
The minimum expected cell frequency is 0.00.

Acyclic decomposition

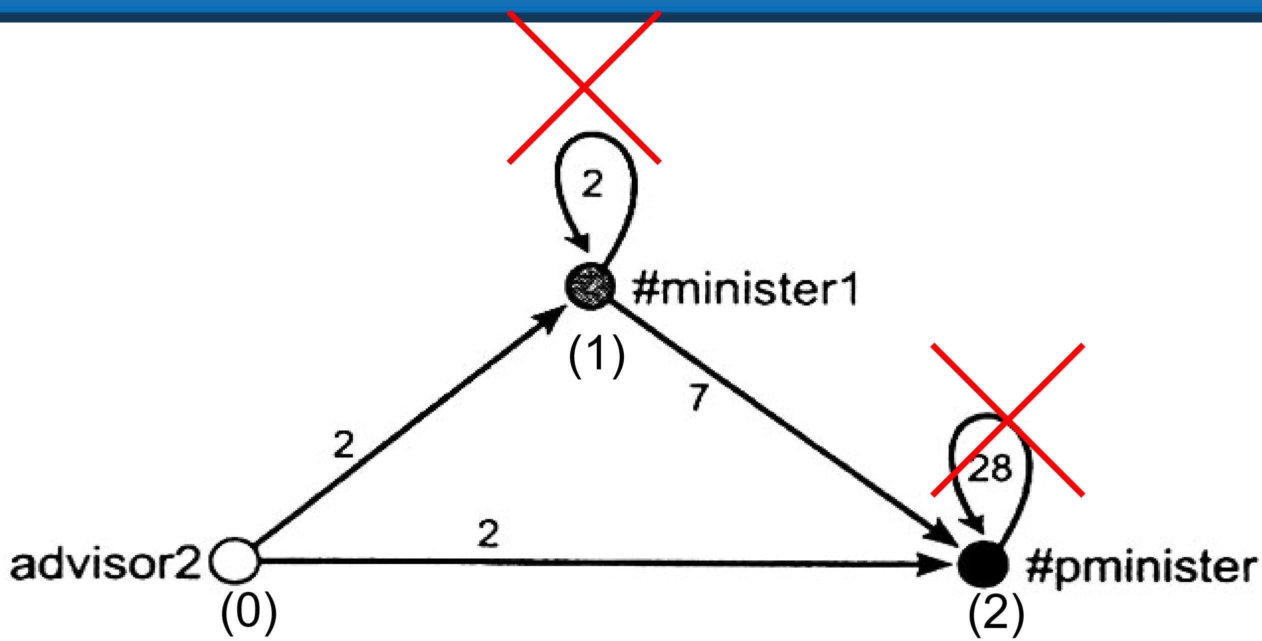
- Cyclic networks (strong components) are clusters of equals
- Acyclic networks perfectly reflect hierarchy
- Recipe
  - Partition the network into strong components (i.e. clusters of equals)
  - Create a new network in which each node represents one cluster
  - Compute the maximum depth of each node to determine the hierarchy

Acyclic decomposition

An example



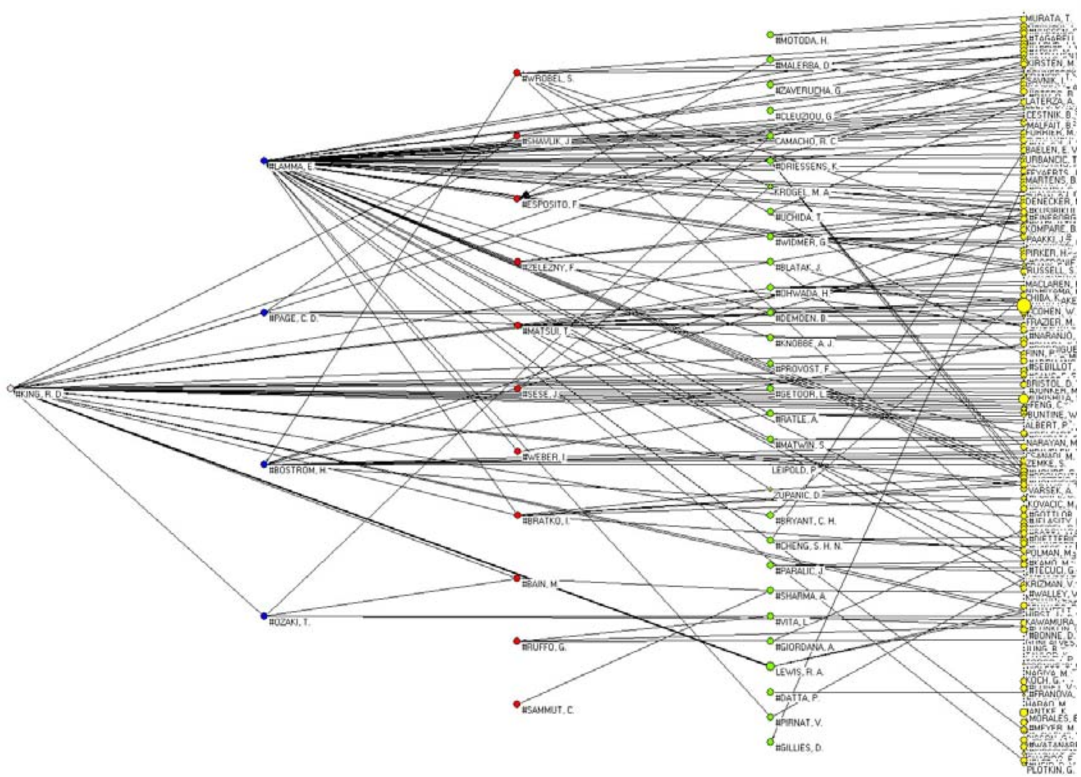
Acyclic decomposition
An example



The maximum depth of a node determines its position in the hierarchy

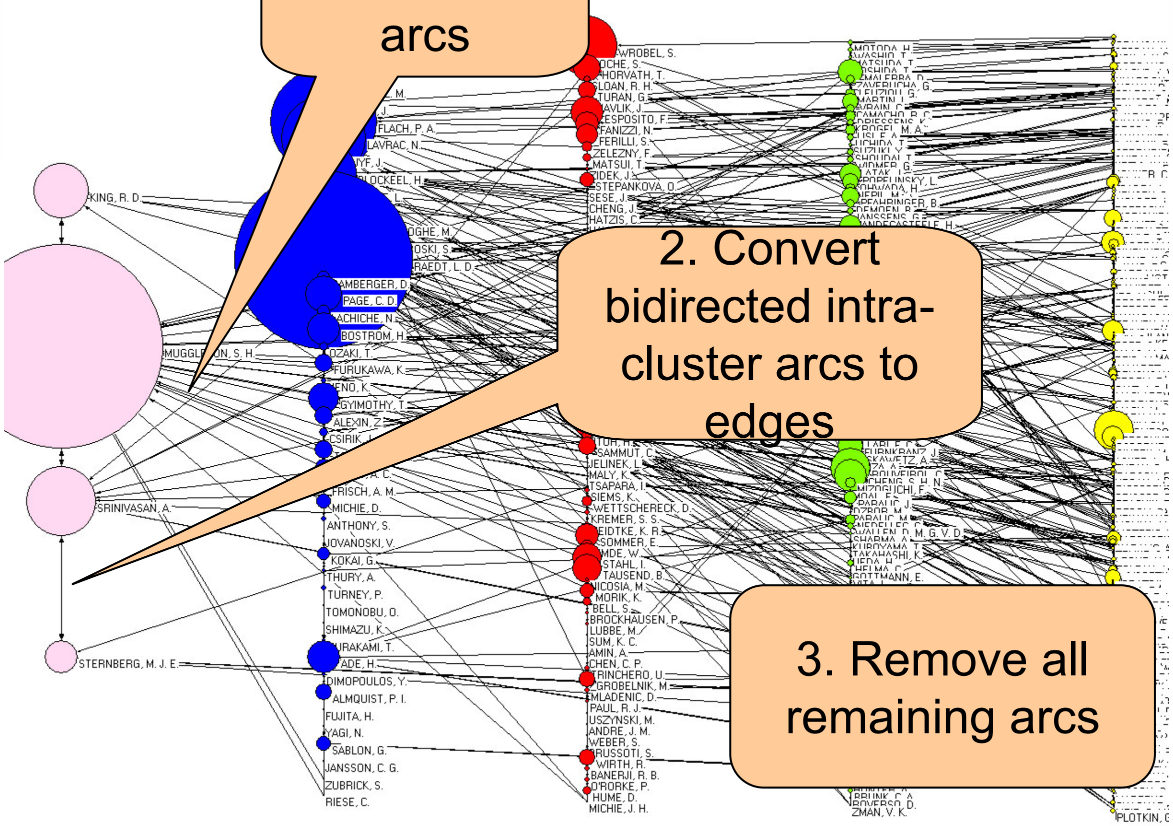
Acyclic decomposition
ILPnet2 dataset

Acyclic decomposition
ILPnet2, hierarchical view



Acyclic decomposition

ILPnet2, hierarchical view (people)

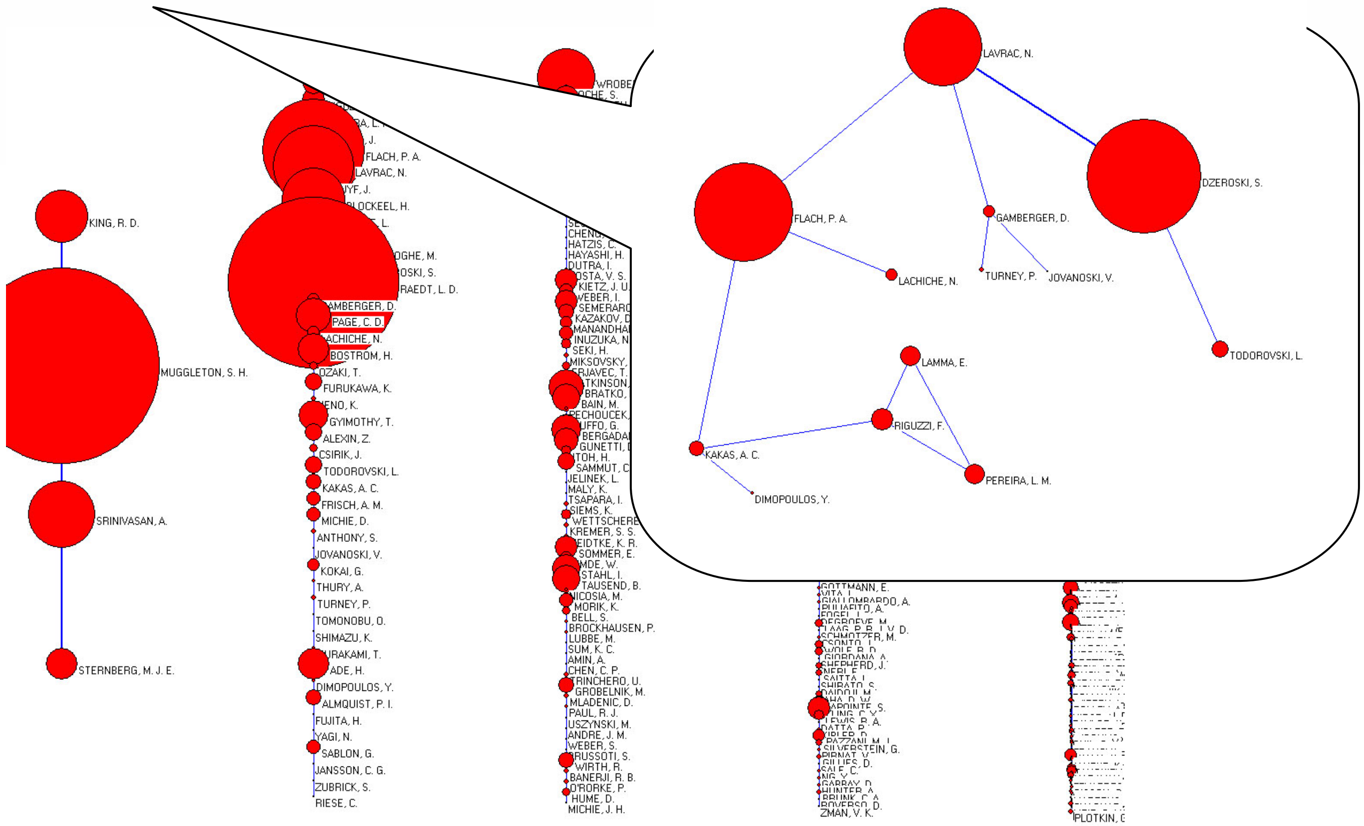


1. Remove inter-cluster arcs

2. Convert bidirected intra-cluster arcs to edges

3. Remove all remaining arcs

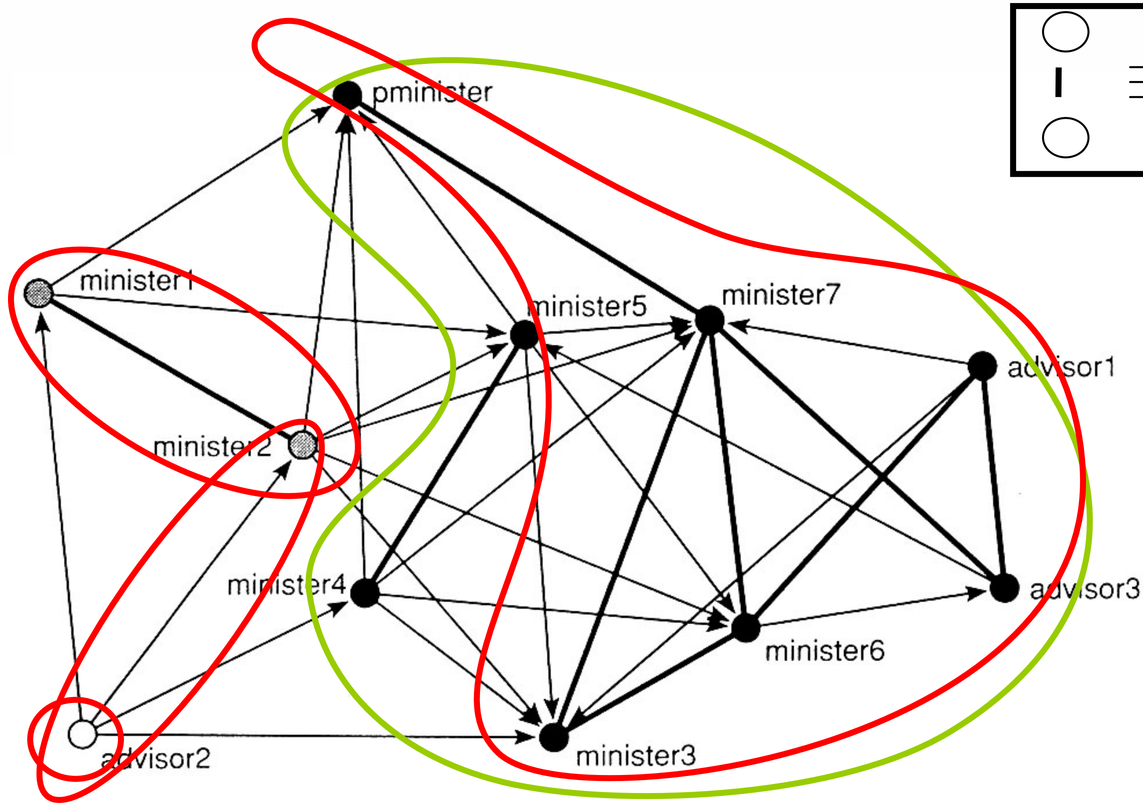
Acyclic decomposition
ILPnet2, hierarchical view (people)

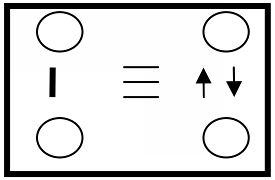


Symmetric-acyclic decomposition

- Strong components are not strict enough to detect clusters in the triad-census sense
- Symmetric-acyclic decomposition extracts clusters of vertices that are connected both ways
- After the clusters are identified, we can follow the same steps as in acyclic decomposition to determine the hierarchy

Symmetric-acyclic decomp.
An example





Reference

- Batagelj V., Mrvar A., de Nooy W. (2004): **Exploratory Network Analysis with Pajek**. Cambridge University Press
- Some figures used in the presentation are taken from this book
- Most of the presentations are taken from presentation of ILPNet2 Social Network Analysis Project of Sergeja Sabo, David Fabjan and Miha Grčar