Centrality Mining

PFIA – DECADE Workshop
28 June, 2022

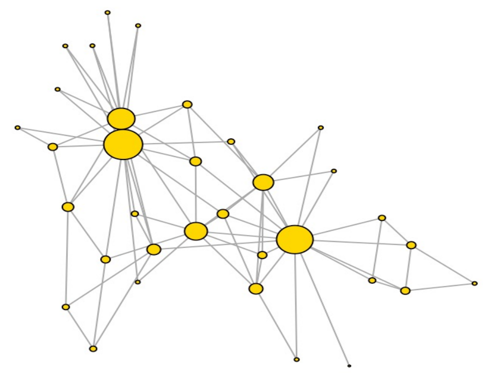
Rushed Kanawati

kanawati@sorbonne-paris-nord.fr

https://www.kanawati.fr

Université Sorbonne Paris Nord

Complex networks



Graphs modelling **direct**/**indirect** interactions among actors.

Direct interactions:
- Friendship
- Proximity
- Message exchange
- …

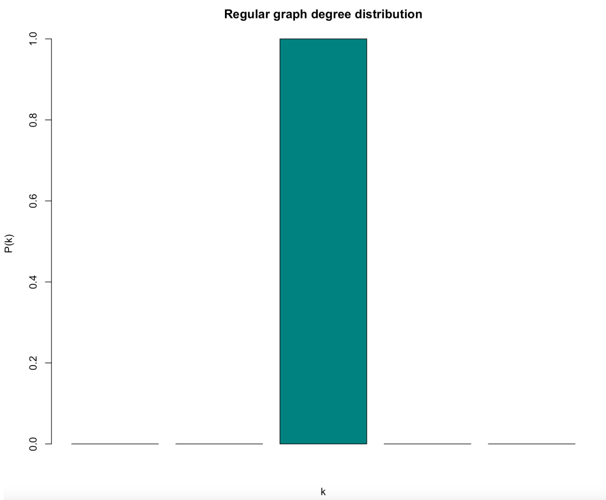
Indirect interactions:
- *Affiliation share*
- Preference share
- **Similarity**
- …

Basic topological features

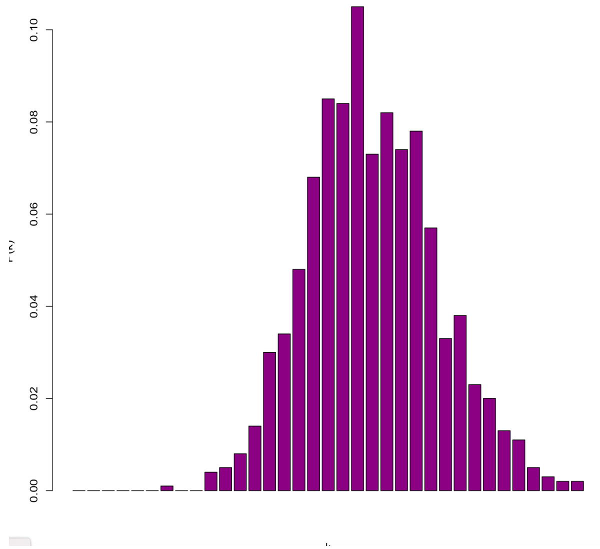
Degree distribution

$$P(k) = \frac{|\{v_i \in V(G) : d_{v_i} = k|}{n_G}$$

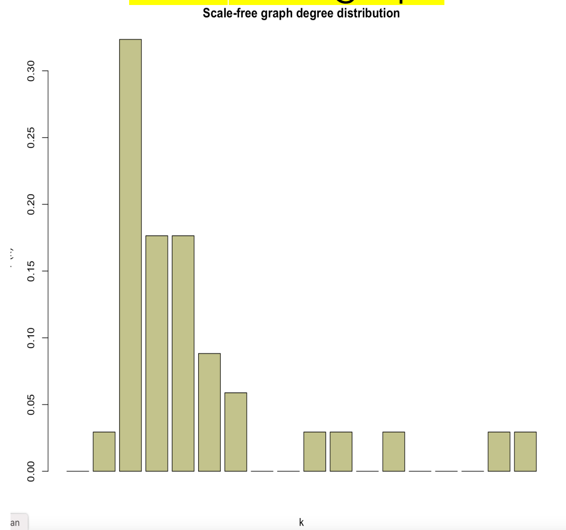
Regular graph



Random graph



Scale-Free graph



Clustering coefficient

Probability of having a link between two nodes that share a common neighbour

What is the probability that two friends of a given person are friends themselves ?

Global version

$$CC(G) = \frac{3 \times \# \Delta}{\# \wedge}$$

Local version

$$CC(G, v) = \frac{\# links\ between\ neighbours\ of\ v}{\# potential\ links\ between\ neighbors\ of\ v}$$

Social networks

Advogato

ρ_G

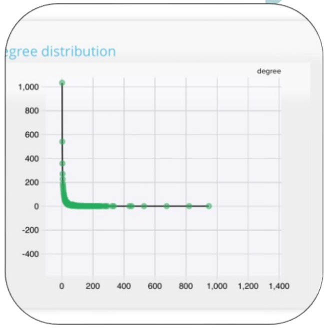
$2\ 10^{-3}$

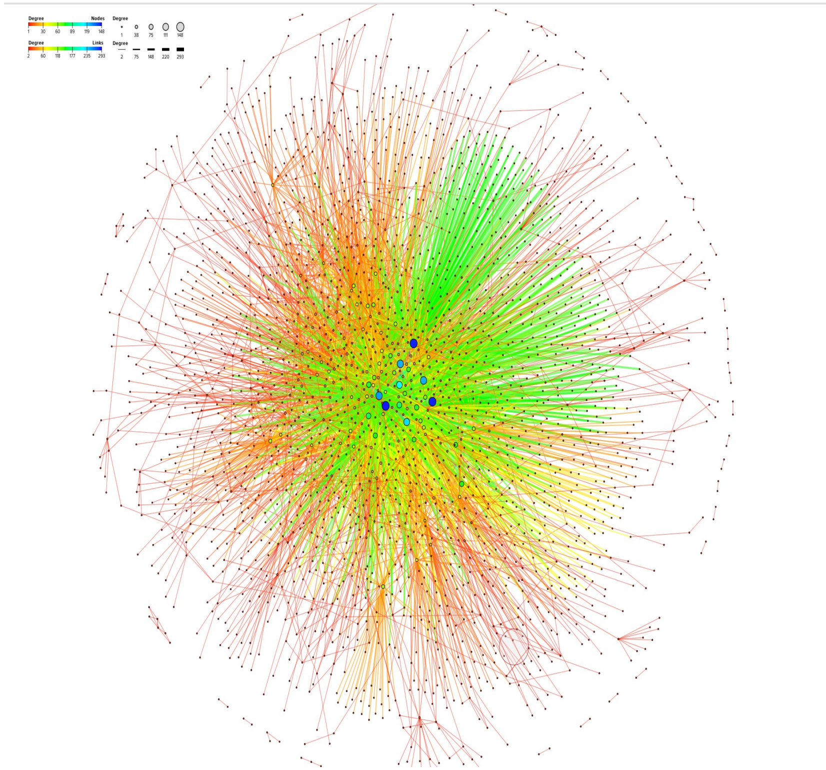
$Diam(G)$

11

CC

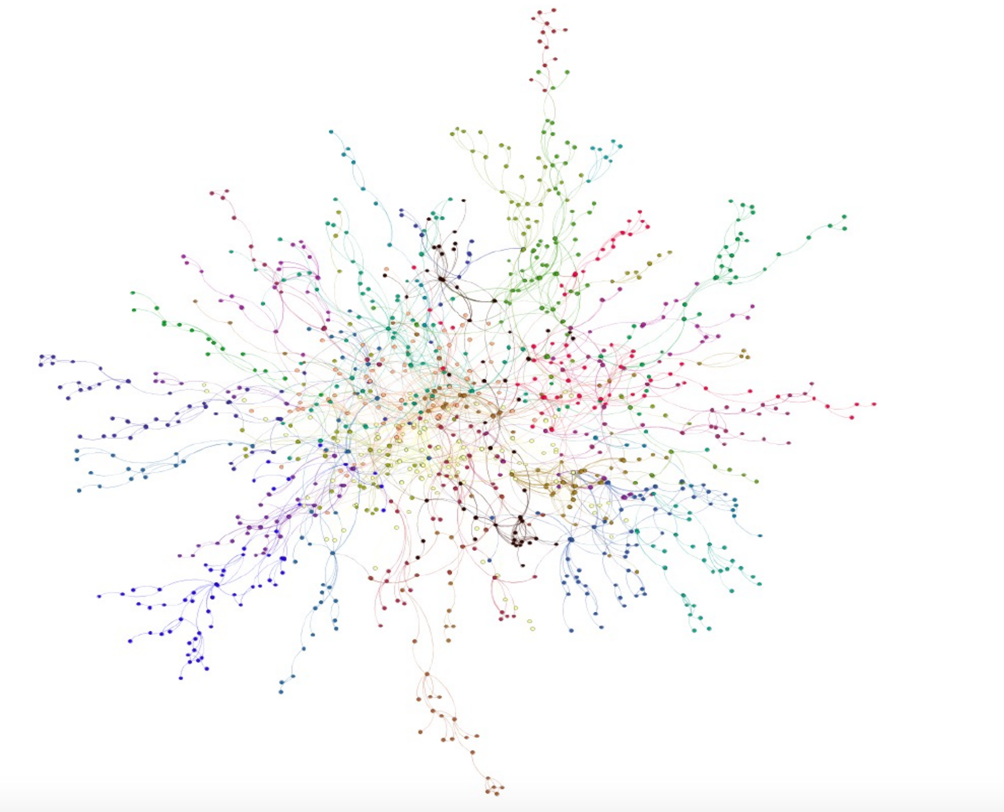
$9\ 10^{-2}$





Collaboration networks

indirect interaction



DBLP co-authorship network (1980-1984)

ρ_G

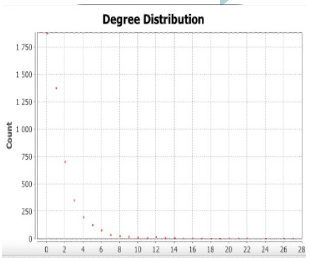
10^{-4}

$Diam(G)$

24

CC

0.67



Computer connection network

$\rho(G)$

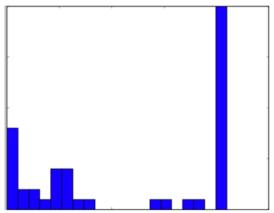
0.06

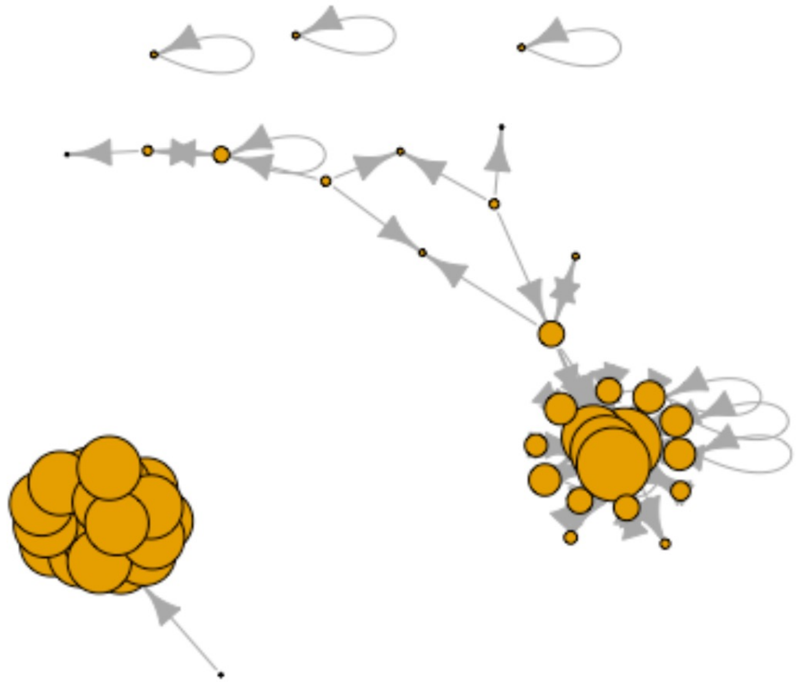
$Diam(G)$

5

CC

0.71





UNSW – connection network dataset

Probabilistic Safety Assessment inferred network

$\rho(G)$

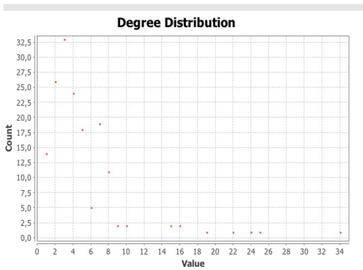
0.009

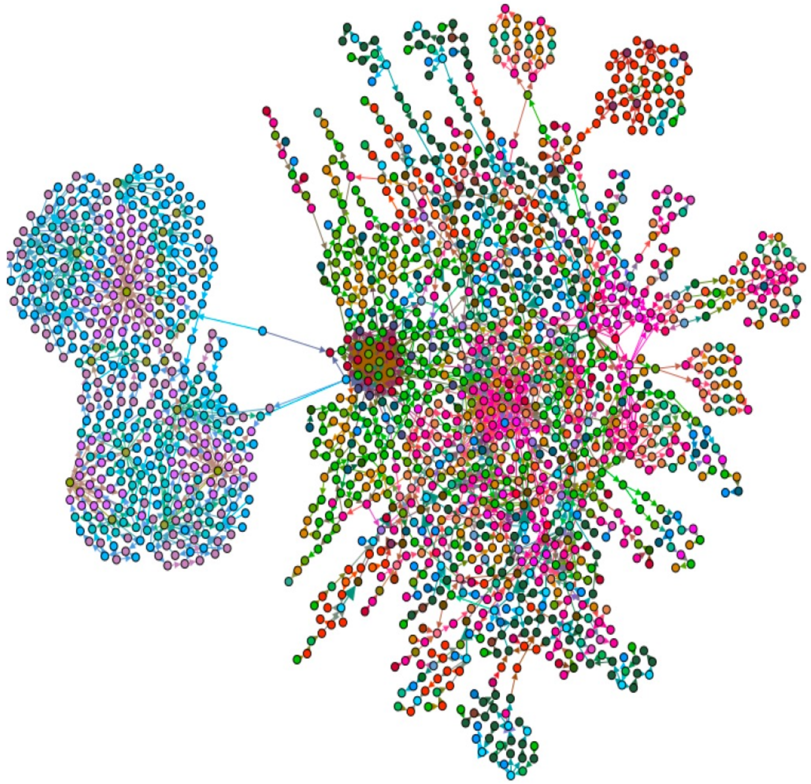
$Diam(G)$

33

CC

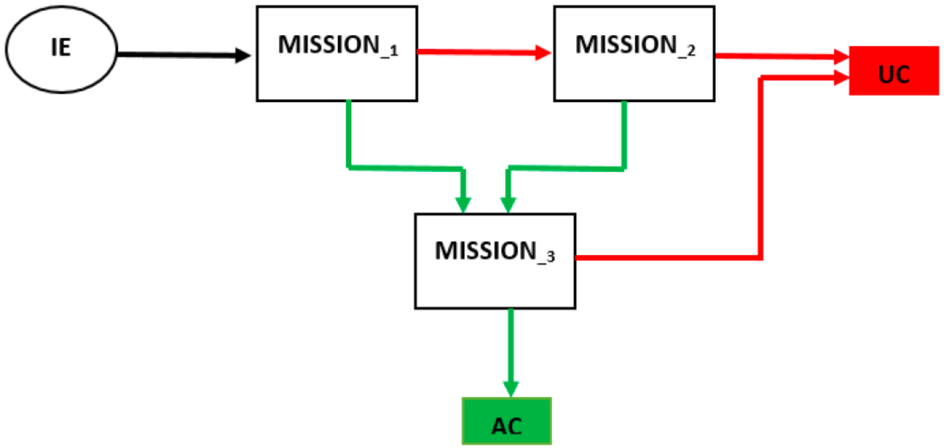
0.97





Uncontrolled level drop in EPR Nuclear Plant [RIFI, 2019]

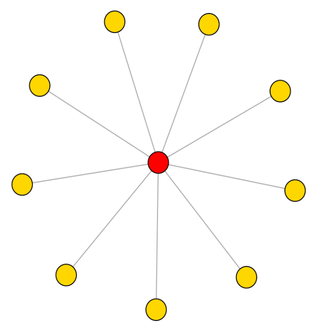
Probabilistic Safety Assessment inferred network : construction



IE : Initiating event
UC : Undesiered consequence
AC : Aceptable consequence

The network of each **Functional Requirement Diagram** is expanded by modelling missions as networks connecting involved components (Pumps, valves, etc.) with different type of links : (fluid, electrical, signal)

Centrality ?

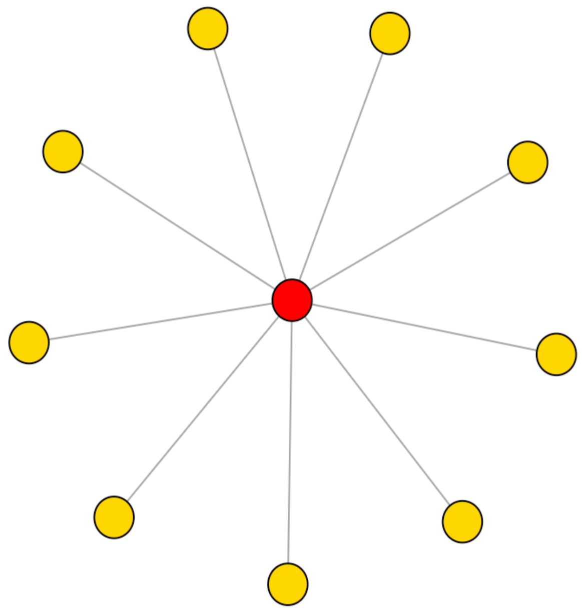


Centrality: A measure of the relative **importance** of a node (or an edge) in a (complex) network.

Influential nodes **Vulnerability nodes** **Control nodes** **…**

Intuitive example

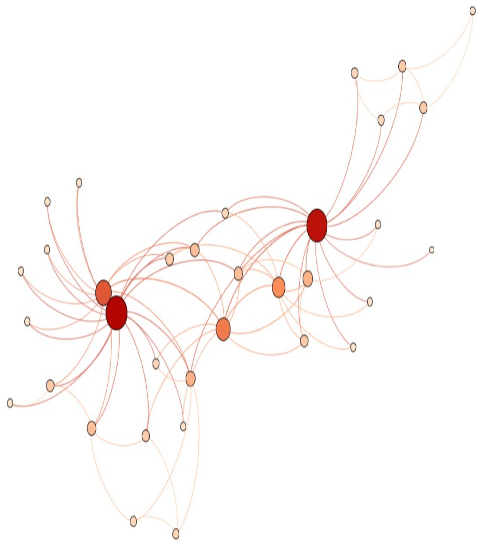
Why is the central node in a star is the most important node ?



- It has the largest degree
- it has the smallest average distance to other nodes
- It is at the intersection of all shortest paths in the network
- It is the node that maximizes the dominant eigenvector of A_G
- ....

Centrality types

- Degree-based
  - In degree, out degree, Leverage, H-Index, coreness
- Distance based
  - Closeness, Katz, Subgraph,
- Path based
  - Betweenness, Communicability, Information
- Spectral measures
  - Hub, Authority, PageRank, Eigenvector,..



Centrality mining ?

Leveraging centrality exploration for gaining new insights .

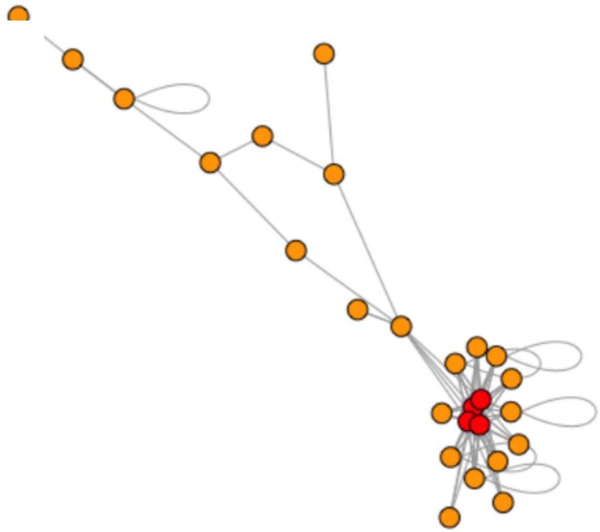
Case studies :

Node classification

Complex networks similarity computation

Node classification #1
Attacker classification

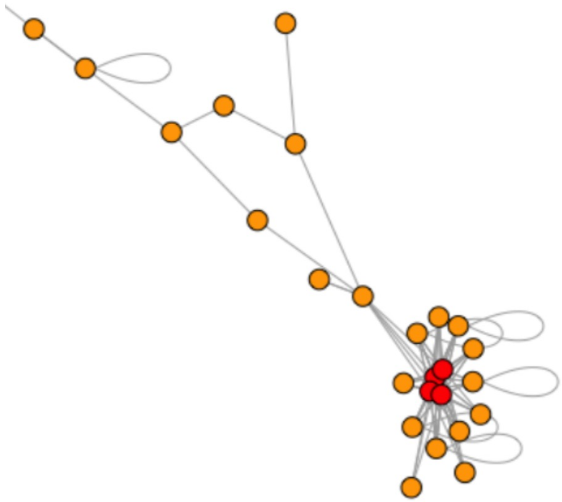
Centrality Name	Characteristic of a central node	Formula
Out degree $c_i^{D^{out}}$	Pointing out to many other nodes	$c_i^{D^{out}} = \sum_{j=1}^{n} A_{ji}$
In Degree $c_i^{D^{in}}$	Pointed to by many other nodes	$c_i^{D^{in}} = \sum_{j=1}^{n} A_{ij}$
Closeness c_i^C	Low average shortest path to other nodes in the network	$c_i^C = \frac{n}{\sum_j sp_{ij}}$
Betweenness c_i^B	Lies on many shortest paths in the network	$c_i^B = \sum_{i \neq j, i \neq k, j \neq k} \frac{\sigma_{ij}(k)}{\sigma_{ij}}$
Eigen $c_i E$	Connected to many other high degree nodes	$c_i^E = \frac{1}{\lambda_1} \sum_j A_{ji} v_j$
Subgraph c_i^S	Involved in many closed short-rang walks	$c_i^S = [e^A]_{ii}$
PageRank c_i^{PR}	Nodes popularity according to random walkers	$c_i^{PR} = \alpha \sum_j A_{ji} \frac{v_j}{c_j^{Dout}}$
Coreness c_i^{COR}	The highest k for which the node belong to non-empty $k-core$[1]	-



Biggest connected component UNSW-15 dataset

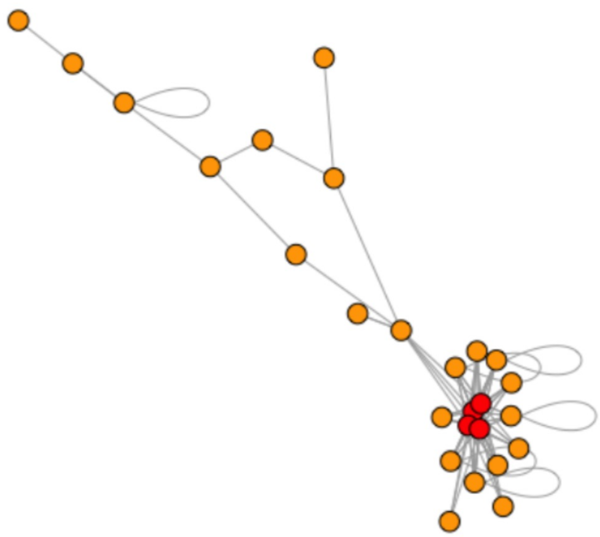
Node classification #1
Attacker classification

Vertex	degree_out	degree_in	closeness	betweenness	authority	eigen	subgraph	coreness	pageRank	Label
12	2.00000	3.00000	0.00166	2.00000	0.00123	0.00018	1.02872	2.00000	0.04374	0.00000
18	12.00000	11.00000	0.00360	42.76149	0.62485	1.00000	52.94414	10.00000	0.09709	1.00000
19	5.00000	5.00000	0.00350	1.77289	1.00000	0.65008	22.90553	10.00000	0.04246	0.00000
20	11.00000	9.00000	0.00357	28.25000	0.56006	0.90595	52.84913	10.00000	0.07212	1.00000
21	2.00000	4.00000	0.00345	0.66897	0.92872	0.45480	8.19897	6.00000	0.03524	0.00000
22	12.00000	10.00000	0.00360	36.22701	0.59872	0.96477	52.84913	10.00000	0.08211	1.00000
28	5.00000	5.00000	0.00350	1.77289	1.00000	0.65008	22.94256	10.00000	0.04246	0.00000
30	5.00000	5.00000	0.00350	1.77289	1.00000	0.65008	22.92412	10.00000	0.04246	0.00000
31	3.00000	4.00000	0.00347	1.14789	0.92872	0.52954	16.56859	7.00000	0.03524	0.00000
32	4.00000	4.00000	0.00350	1.77289	0.92872	0.59972	22.93106	8.00000	0.03524	0.00000
33	4.00000	4.00000	0.00350	1.77289	0.92872	0.59972	23.07744	8.00000	0.03524	0.00000
34	12.00000	11.00000	0.00360	42.76149	0.62485	1.00000	52.94414	10.00000	0.09709	1.00000
35	5.00000	5.00000	0.00350	1.77289	1.00000	0.65008	22.92052	10.00000	0.04246	0.00000
36	3.00000	0.00000	0.00189	0.00000	0.00000	0.00205	7.42773	2.00000	0.01010	0.00000
37	0.00000	2.00000	0.00154	0.00000	0.05376	0.02426	0.00000	2.00000	0.01643	0.00000
38	6.00000	2.00000	0.00472	35.00000	0.00003	0.31110	25.16968	4.00000	0.02449	0.00000
39	4.00000	4.00000	0.00350	1.77289	0.92872	0.59972	22.93735	8.00000	0.03524	0.00000
40	5.00000	5.00000	0.00350	1.77289	1.00000	0.65008	22.93735	10.00000	0.04246	0.00000
42	1.00000	1.00000	0.00437	0.00000	0.05259	0.04820	1.68118	2.00000	0.01357	0.00000
43	0.00000	4.00000	0.00154	0.00000	0.92872	0.29986	0.00000	4.00000	0.03524	0.00000
44	0.00000	3.00000	0.00154	0.00000	0.70764	0.22968	0.00000	3.00000	0.02967	0.00000
45	0.00000	2.00000	0.00154	0.00000	0.00120	0.00205	0.00000	2.00000	0.01582	0.00000
47	3.00000	0.00000	0.00645	0.00000	0.00000	0.02441	7.08960	2.00000	0.01010	0.00000
48	0.00000	1.00000	0.00154	0.00000	0.00003	0.00189	0.00000	1.00000	0.01296	0.00000
49	2.00000	1.00000	0.00167	2.00000	0.00003	0.00003	1.02872	2.00000	0.02869	0.00000
50	0.00000	1.00000	0.00154	0.00000	0.00003	0.00000	0.00000	1.00000	0.02229	0.00000



Node classification #1
Attacker classification

Centrality	Prediction precision
Out degree	100%
In degree	100%
Closeness	25%
Betweenness	75%
Eigen	100%
Subgraph	100 %
PageRank	100%
Coreness	44%



Node classification : case study #2

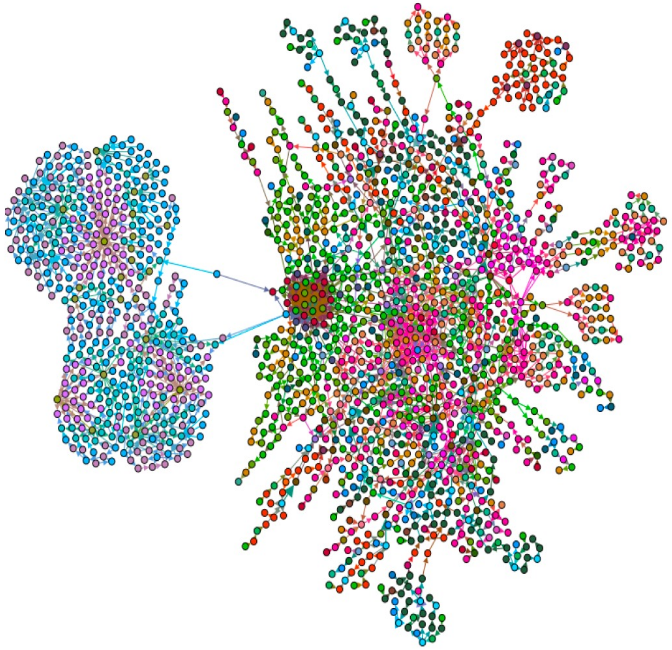
Risk Increase Factor (RIF) prediction

RIF(x_i) = $\frac{Risk(x_i=1)}{Risk(x_i=0)}$

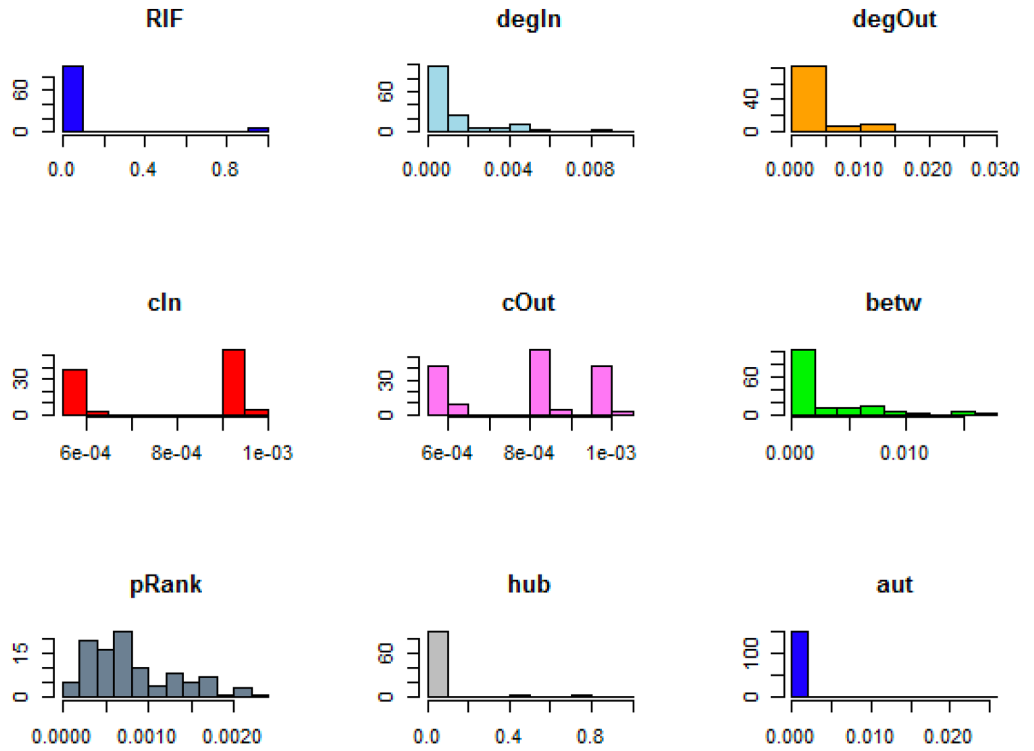
RIF computation is computationally hard

Can we predict RIF class (High/Low) from the network ?

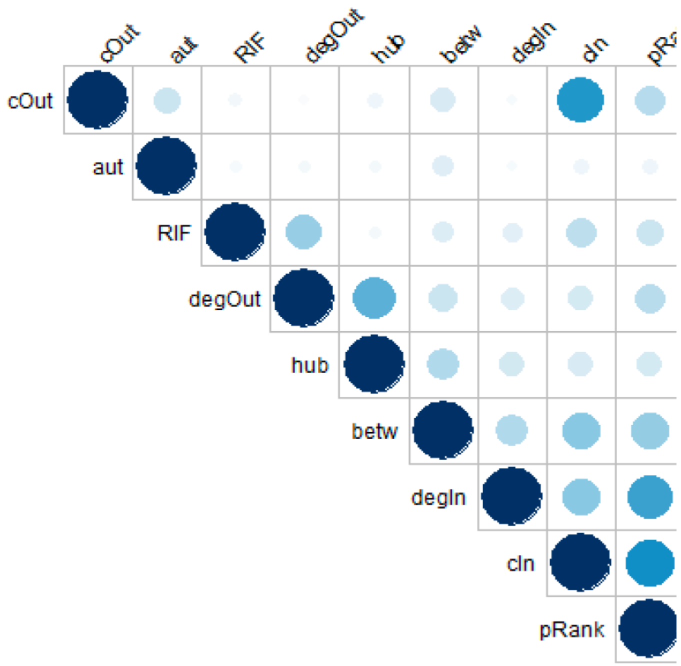
Only 5% of nodes have High RIF value



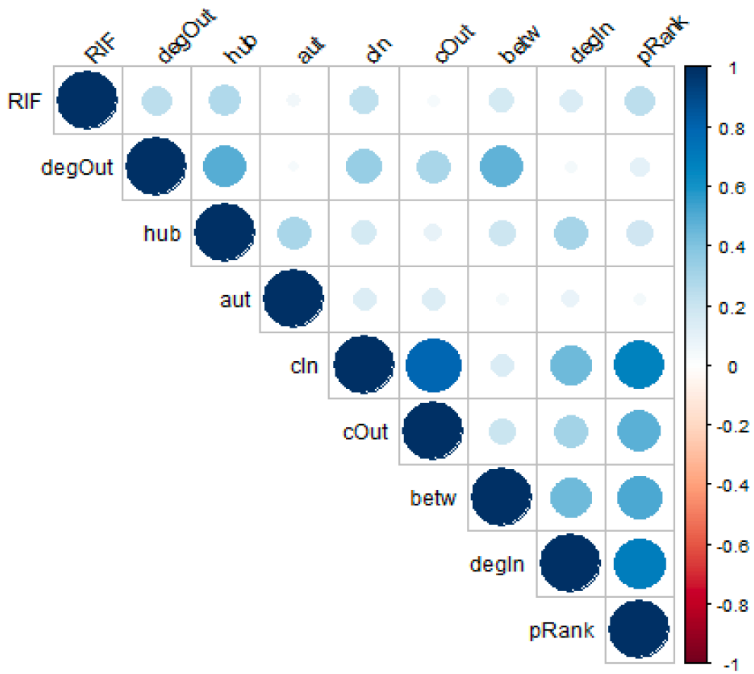
RIF Prediction



RIF Prediction



Pearson correlogram _train sample



Spearman correlogram _train sample

RIF Prediction : supervised classification

Decision Tree

sample	specificity	sensitivity	precision	F-meas	AUC
train	0.979	0.400	0.500	0.444	0.690
test	0.981	0.333	0.500	0.400	0.657

Random Forest

sample	specificity	sensitivity	precision	F-meas	AUC
train	1	0.600	1	0.750	0.800
test	1	0.333	1	0.500	0.667

Gradient Boosted Machine

sample	specificity	sensitivity	precision	F-meas	AUC
train	1	0.600	1	0.750	0.800
test	1	0.667	1	0.800	0.833

Node classification : case study #3

Local modularity selection for ego-centred community identification

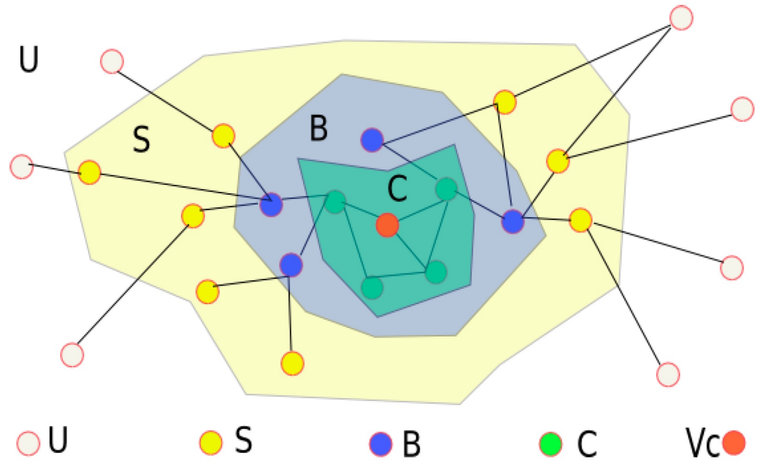
1 $C \leftarrow \{\phi\}, B \leftarrow \{n_0\}\ S \leftarrow \Gamma(n_0)$

2 $Q \leftarrow 0$ /* a community **quality function** */

3 While Q can be enhanced Do

- 1 $n \leftarrow argmax_{n \in S} Q$
- 2 $S \leftarrow S - \{n\}$
- 3 $D \leftarrow D + \{n\}$
- 4 update B, S, C

4 Return D



Local modularity functions

Local modularity R [Cla05]

$$R = \frac{B_{in}}{B_{in}+B_{out}}$$

Local modularity M [LWP08]

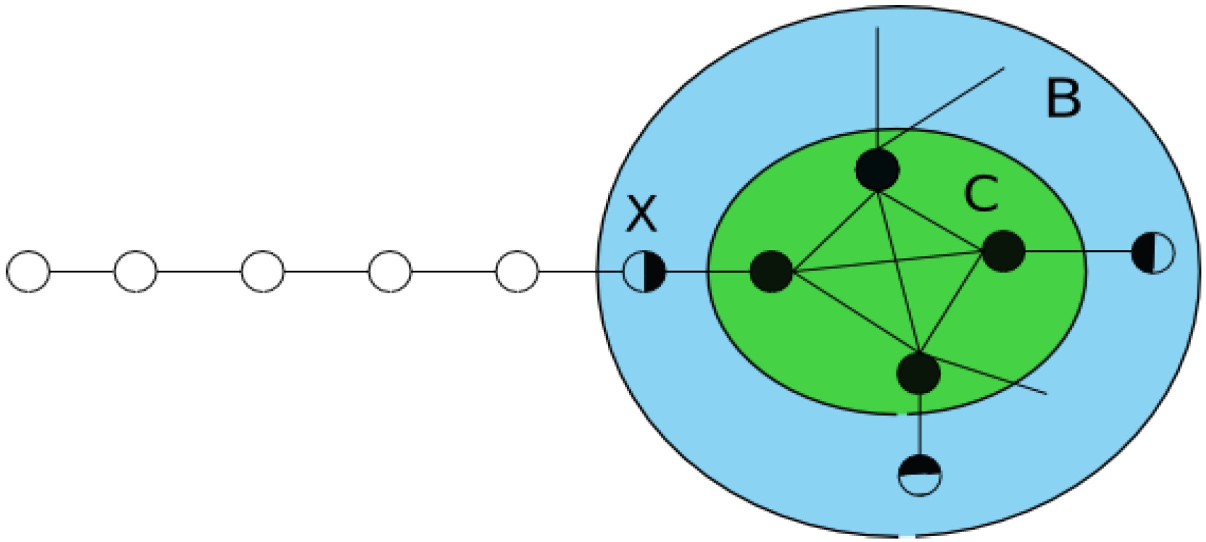
$$M = \frac{D_{in}}{D_{out}}$$

Local modularity L [CZG09]

$$L = \frac{L_{in}}{L_{ex}} \text{ where : } L_{in} = \frac{\sum_{i \in D} ||\Gamma(i) \cap D||}{||D||}, L_{ex} = \frac{\sum_{i \in B} ||\Gamma(i) \cap S||}{||B||}$$

And many many others . . . [YL12]

Local modularity limitation

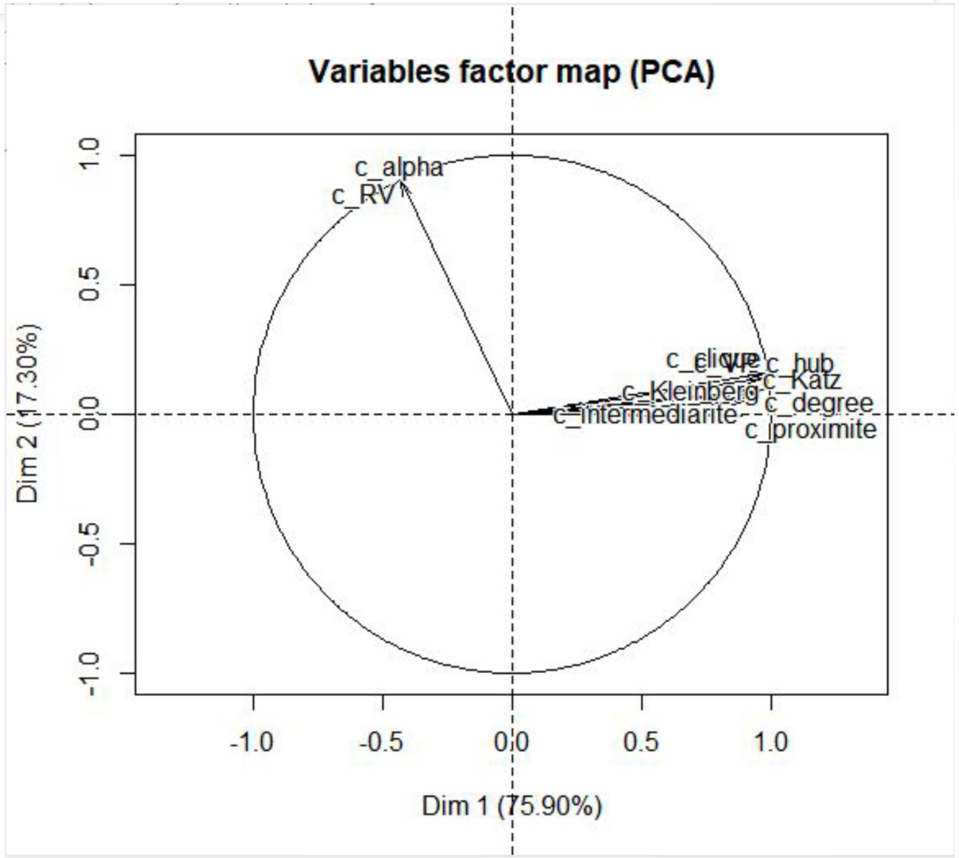


Solutions :

Ensemble Clustering
Ensemble Ranking
…
Local modularity selection

Local modularity selection

- Multi-label supervised classification problem
- Node features : centrality measures
- Applying PCA for feature selection
- Applying different classification algorithms

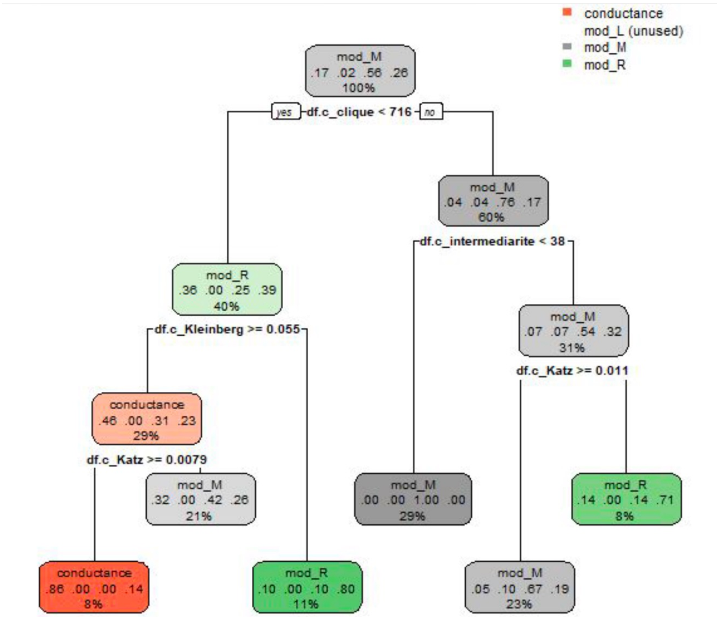


Local modularity selection : experiments

Experiments on benchmark networks with community Ground-truth information

Zachary, Football, PolBooks, Dolphine, etc.

Random forest : precision 83.33%

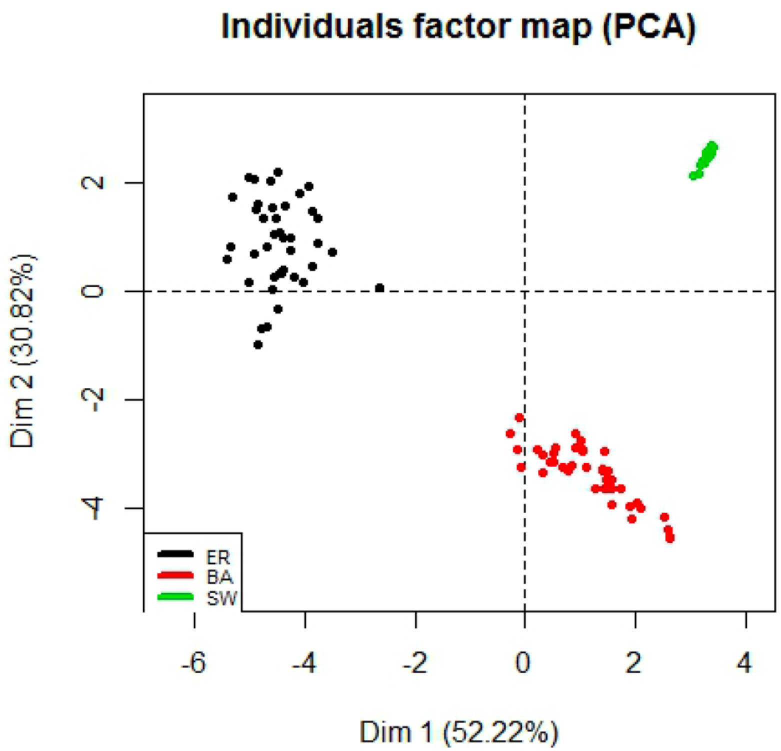


Complex networks distance function

- Goal : providing a complex network distance function
- A simple Graph embedding approach
- Let G =<V_G,E_G> be a network
- $C_i(V_G)$: $Ranked\ vector\ of\ G\ vertices\ in\ function\ of\ centrality\ i$
- $V_{top} = \ \cup\, Top_{\alpha}\ (\ C_i(V_G))$
- K_{cor}(G) = <cor(C_1(V), C_2(V)),…, cor(C_n(V), C_{n-1}(V))>
- Dist(G_i , G_j) = d(K_{cor}(G_i) , K_{cor}(G_j))

Experiment #1

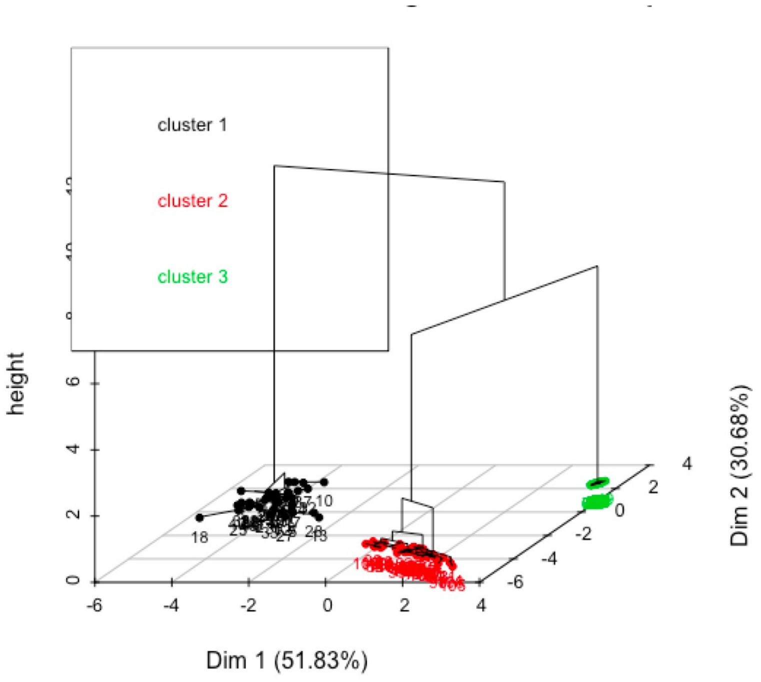
- Generating 120 networks : 40 Erdös-Renyi (0.05), 40 Watts, 40 (0.05) Scale-Free (0.1)
- Number of nodes : 1000



Experiment #1

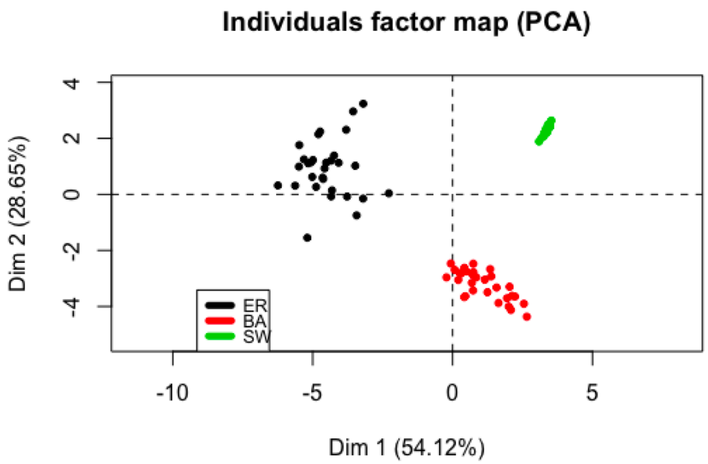
- Generating 120 networks : 40 Erdös-Renyi (0.05), 40 Watts, 40 (0.05) Scale-Free (0.1)
- Number of nodes : 1000

clusters.acp-cah	1	2	3
ER	40	0	0
BA	0	39	0
SW	0	0	40



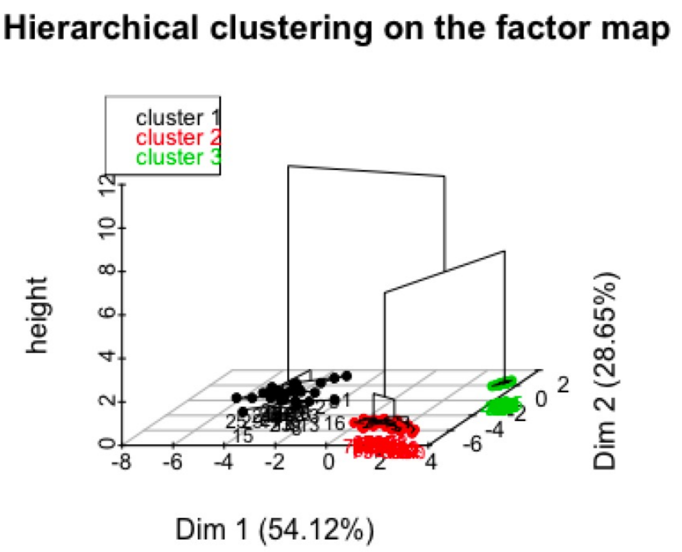
Experiment #2

- Generating 120 networks : 40 Erdös-Renyi (0.05), 40 Watts, 40 (0.05) Scale-Free (0.1)
- Number of nodes : 1000, 2000, 4000



Experiment #2

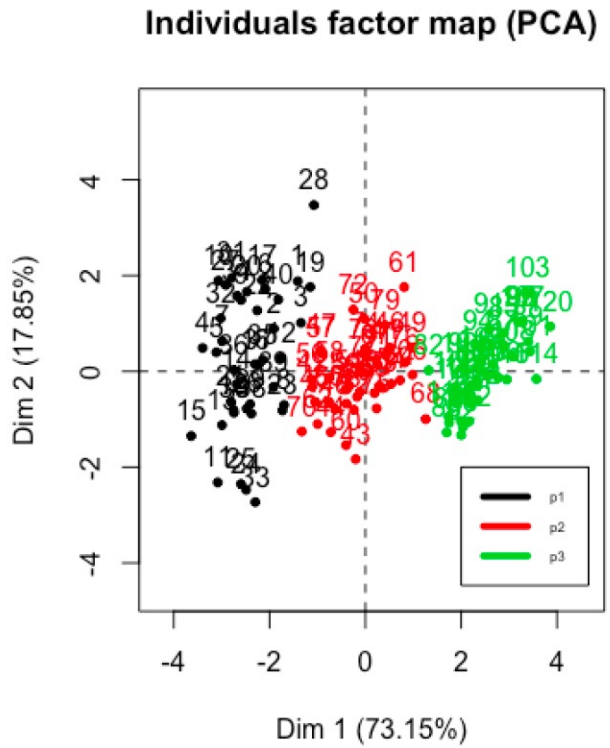
- Generating 120 networks : 40 Erdös-Renyi (0.05), 40 Watts, 40 (0.05) Scale-Free (0.1)
- Number of nodes : 1000, 2000, 4000



clusters.cah	1	2	3
BA	0	0	30
SW	0	30	0
ER	30	0	0

Experiment #3

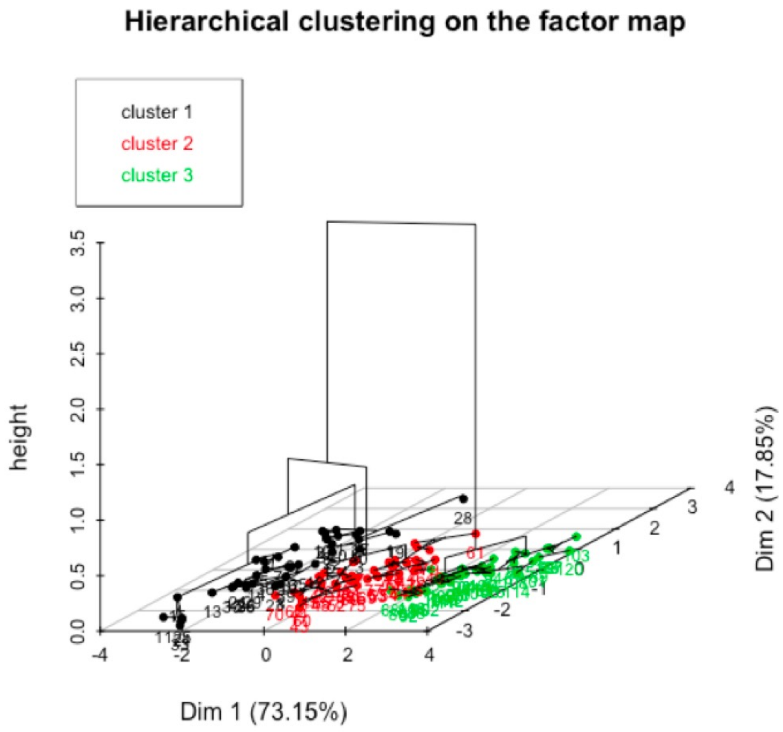
- Generating 120 Watts networks : 40 perturbations of 3 seed networks :
- P : [0,0075, 0,0125, 0,0275]



Experiment #3

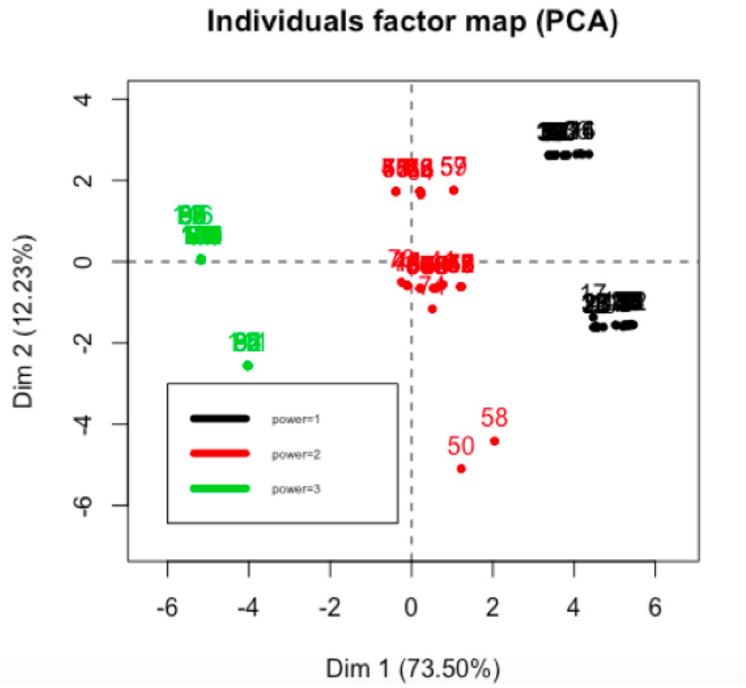
- Generating 120 Watts networks : 40 perturbations of 3 seed networks :
- P : [0,0075, 0,0125, 0,0275]

clusters.acp-cah	1	2	3
Cl1	40	0	0
Cl2	0	39	1
Cl3	0	0	40



Experiment #4

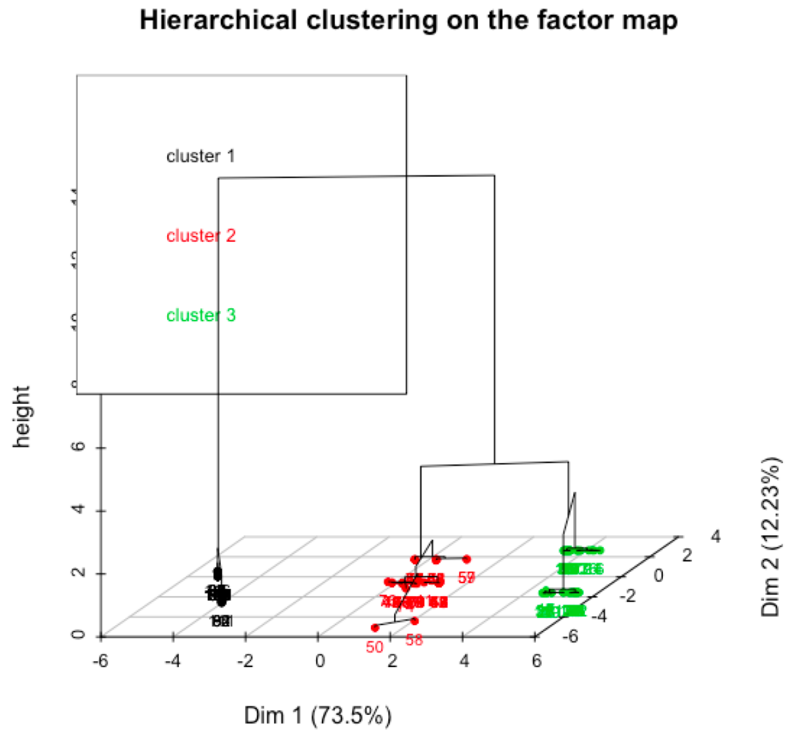
- Generating 120 PA networks : 40 perturbations of 3 seed networks :
- Power : 1,2,3



Experiment #4

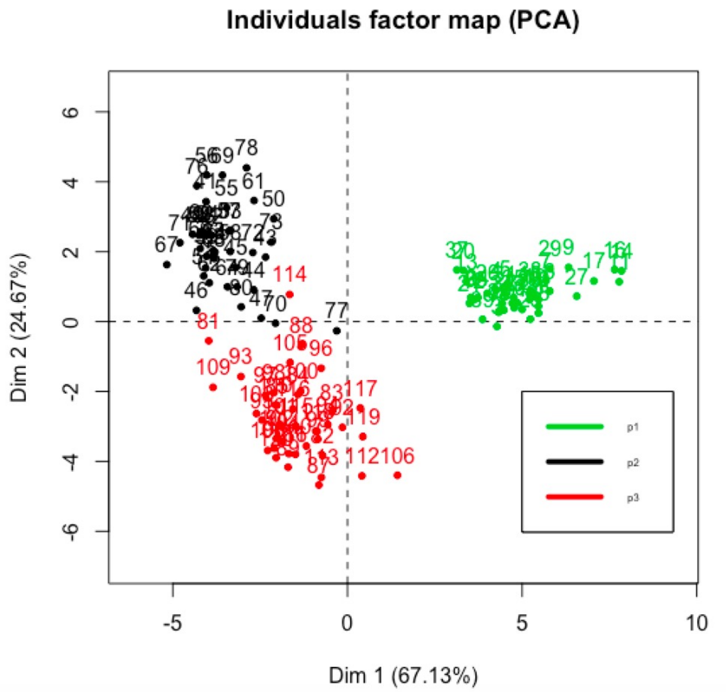
- Generating 120 PA networks : 40 perturbations of 3 seed networks :
- Power : 1,2,3

clusters.acp-cah	1	2	3
Cl1	0	0	40
Cl2	0	39	0
Cl3	40	0	0



Experiment #5

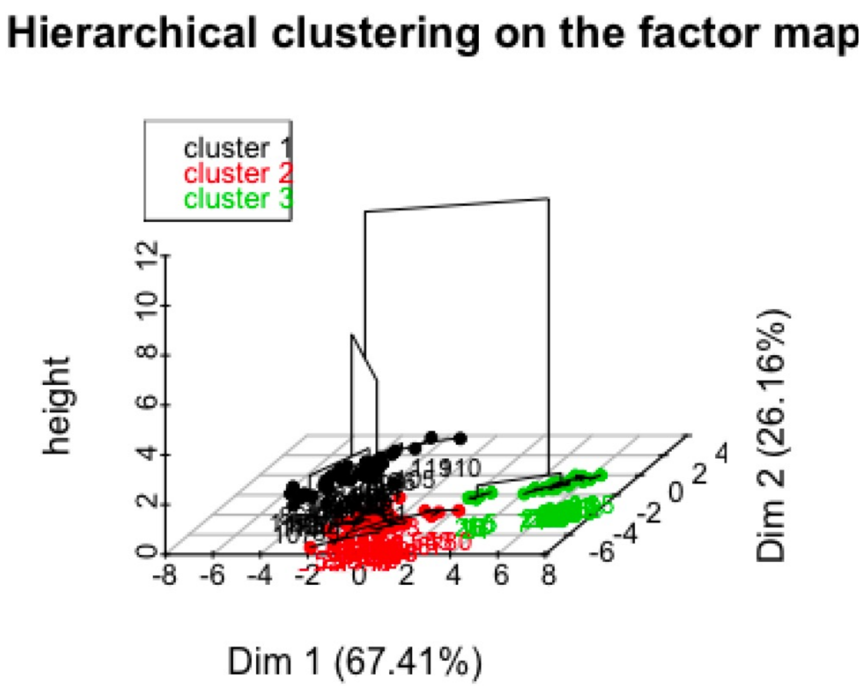
- Generating 120 ER networks : 40 perturbations of 3 seed networks :
- P : 0,01 0,03 0,05



Experiment #5

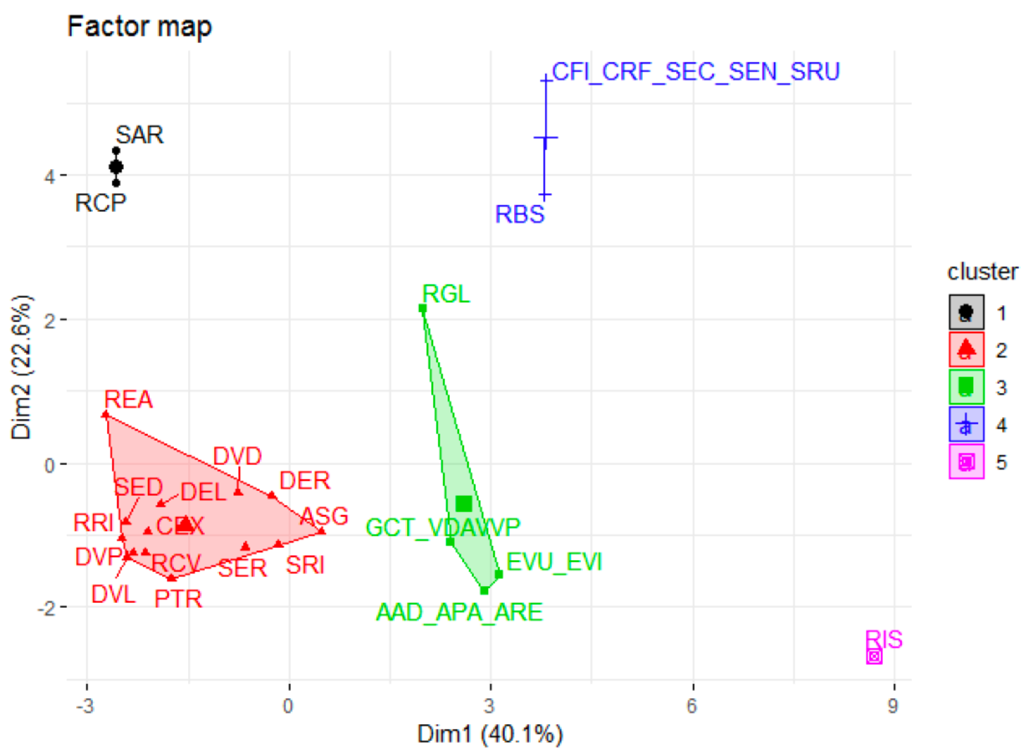
- Generating 120 ER networks : 40 perturbations of 3 seed networks :
- P : 0,01 0,03 0,05

clusters.acp-cah	1	2	3
Cl1	40	0	0
Cl2	0	39	1
Cl3	0	0	40



Experiment #6

Application on all EPR nuclear plant PSA networks



Conclusions

- Centrality mining can enhance different basic complex network analysis tasks : Node classification, community detection
- A new simple complex network distance function
  - Change and anomaly detection
  - Network influence estimation
- Centrality induced rank computation is crucial
  - Estimation function ?
- Effects of selecting the Top α-ranked nodes ?
- And a lot of applications … (Ex. Network analysis for cyber security !!)