8th International Conference of the ERCIM WG on
Computational and Methodological Statistics (CMStatistics 2015)
12-14 December 2015, Senate House, University of London, UK
ERCIM WG
CMStatistics

Exploratory data analysis and contiguity relations: An outlook

Giuseppe Giordano[1], Gilbert Saporta[2], Maria Prosperina Vitale[1]

[1] **Dept. of Economics and Statistics, University of Salerno, Italy**
{ggiordan; mvitale}@unisa.it
[2] **Conservatoire National des Arts et Métiers, Paris, France**
gilbert.saporta@cnam.fr

London – December 13, 2015

Outline

- Setting: Multidimensional Data Analysis (MDA) and Social Network Analysis (SNA)
- Theoretical frameworks: Notion of contiguity, Homophily principle, Social Influence, ...
- Aim 1: to present a framework for the treatment of SNA data structures with explorative techniques of MDA
- Methods: Smooth factorial analysis-SFA; Factorial Analysis of Local Differences-FALD (PCA, MCA) Benali, H., Escofier, B. (1990
- Aim 2: To define ad hoc relational data structures highlighting the effect of external information on networks
- Methods: Factorial Contiguity Maps and Auxiliary information in SNA

  Giordano G., Vitale M.P. (2007) (2011)
- Illustrative example on Scientific Collaboration

General aims

- use of Contiguity Analysis to synthesize and visualize the patterns of social relationships in a metric space
- explore the effect of external information on relational data looking for groups of structurally equivalent actors obtained through clustering methods
- illustrative example in the framework of scientific collaboration gives a major insight into the proposed strategy of Multidimensional Data Analysis in the framework of Social Network Analysis

Background and Aim

Background

- **SNA** focuses on **ties** among **interacting units** (Dyad, Triad, Subgroups) to describe the pattern of the social relationships in a network

- the techniques of **MDA** consider statistical observations (at individual level) to obtain syntheses of variables and units

Aim

to present a framework for the analysis of relational data and attribute variables through the **explorative techniques of Multidimensional Data Analysis**

Exploratory data analysis (EDA) is a detective work, finding and revealing the clues, i.e. uncovering structures in the data. EDA uses numerical as well as visual and graphical techniques to accomplish its aims. (Tukey, 1977)

Data Structures

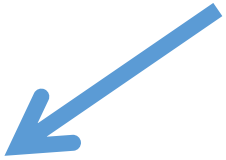
Relational data (pairwise links joining two units)

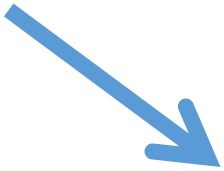
=> SNA

Attribute data (qualitative or quantitative variables)

=> MDA

Two Perspectives **(to put together the 2 data structures)**





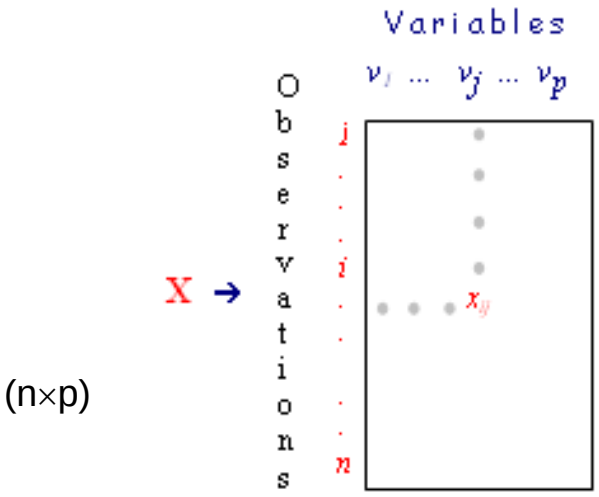
Multidimensional Data Analysis - MDA

Social Network Analysis - SNA

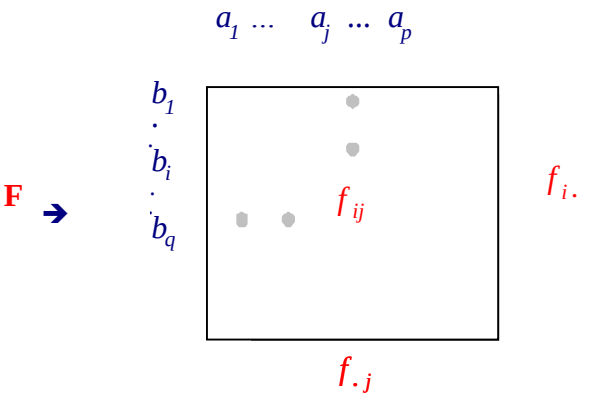
MDA and SNA: Data structures

Attribute Data Matrix

MDA



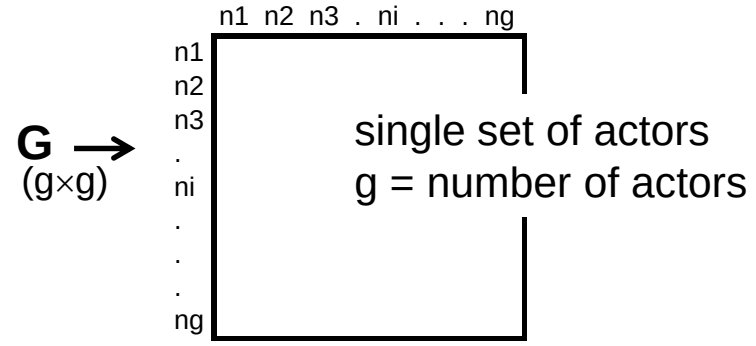
Contingency Table



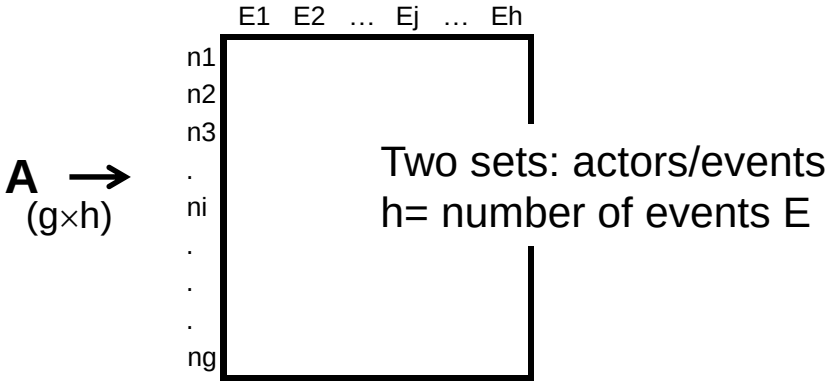
Relational Data Matrix

SNA

Adjacency matrix: 1-mode network



Affiliation matrix: 2-mode network



MDA in SNA

such as…

Usually different techniques of MDA have been used to visualise and explore the relationships in the net structure

- ***Multidimensional Scaling*** : representation of similarity or dissimilarity measures among the actors onto a factorial map (Freeman, 2005)
- ***Canonical correlation***: analysis of associations among actor characteristics, (i.e. network composition) and the pattern of social relationships (i.e. network structure) (Wasserman and Faust, 1989)
- ***Correspondence Analysis and Multiple Factor Analysis***: analysis of 2-mode networks (Roberts, 2000; de Nooy, 2003; Faust, 2005; D'Esposito *et al*., 2014; Ragozini et al., 2015)
- ***Clustering techniques for network data***

(Batagelj, Ferligoj, 1982, 2000)

Aim 1

Contiguity analysis

How to take into account relational data in MDA?

Answer

Contiguity analysis is a generalization of linear discriminant analysis in which the partition of elements is replaced by a more general **graph structure** defined a priori of the set of the observations (disjoint cliques, chain structures, undirected graph).

(Lebart 1969, 2006; Lebart *et al.,* 2000)

→ **G** (n,n) **contiguity matrix** holding the n vertices in a contiguity graph (symmetric binary matrix); $g_{ii'} = 1$ if i is a neighbour of i' and $g_{ii'} = 0$ if not

Contiguity analysis in MDA: SFA and FALD

Contiguity analysis in MDA
- Smooth factorial analysis – SFA
- Factorial Analysis of Local Differences - FALD

(Benali *et al*. 1990) (Benali, Escofier, 1990)

Smooth factorial analysis - SFA
Analysis of the general pattern in the data by removing local variations
(replacing each point with the centre of gravity of its neighbors)

Factorial Analysis of Local Differences - FALD
Analysis of the local variations
(replacing each point with the differences from the barycentre of its neighbors)

SFA and FALD: matrices definition

Qualitative Variables:
***Multiple Correspondence Analysis* - MCA**

Quantitative Variables:
***Principal Component Analysis* - PCA**

$Q_{(n, k)}$ = full disjunctive coding matrix (0/1)

$X_{(n,p)}$ attribute matrix information

of p characteristics on n statistical units (vertices)

$G_{(n,n)}$ contiguity matrix (network structure)

$N_{(n,n)}$ diagonal matrix [**N**= diag (**G'G**)] holding the degree of each vertices

Analysis of relational data and auxiliary information

Given the triplet **Q**, **G**, **N** for a MCA multiply the **Q** matrix by $\mathbf{N^{-1}G}$

SFA $\quad \mathbf{N^{-1}GQ}$

FALD $\quad \mathbf{Q - N^{-1}GQ + \frac{q_{i.}q'_{.j}}{q_{..}}}$

Analysis of relational data and auxiliary information

Given the triplet **X**, **G**, **N** for a PCA multiply the **X** matrix by $\mathbf{N^{-1}G}$

SFA $\mathbf{N^{-1}GX}$

FALD $\mathbf{X - N^{-1}GX}$

SFA and FALD in SNA

Entries in adjacency matrix can be seen as a particular case of contiguity relation among statistical units defined in **G**. It produces a fuzzy partitioning of the units.

Decomposition of the total variance/inertia into two components:

- **local variance between the adjacent units**
  to discover patterns in the data

  ↳ **SFA:** variability explained by the presence of a contiguity structure

- **residual variance**
  analysis of cohesive sub-group variations

  ↳ **FALD:** actors with a prominent role in contiguous groups

Illustrative Example

Real data set: Scientific collaboration among scientists

… the process generating ties (e.g. co-authorship) in a collaboration network is somewhat affected by attribute data on authors (*academic position, research specialty, geographical proximity* …) or on publications (*type, scientific relevance*, …).

Expected results:

- Are authors with different academic positions in their institution (phd. student, assistant professor, full professor) more likely to collaborate in writing a publication than authors sharing the same position?

- Are authors who work in the same research specialty (e.g. Social statistics) more likely to collaborate in writing a publication than authors from different specialties?

*Co-authorship network in a scientific community**

Co-authorship patterns in Statistics, focusing on academic statisticians in Italy (792 grouped in 5 subfields, at March 2010):

- target population involved in a discipline which is not yet fully explored in terms of its scientific collaboration (i.e., co-authorship) behaviour

- no unique bibliographic archive for collecting their publications

- interest to trace co-authorship relations in distinct data sources

	Years	# of publications	Author coverage rate
WoS	1989–2010	2289	60.7%
CIS	1975–2010	3459	73.4%
PRIN projects	*2000–2008*	5054	70.2%

* De Stefano D., Fuccella V., Vitale M.P., Zaccarin S. (2013). The use of different data sources in the analysis of co-authorship networks and scientific performance, *Social Networks*, 35, pp. 370-381

Data source: Italian academic statisticians whit publications in WoS archive

Relational Data

A = affiliation matrix 481 x 2289
481 scientists, 2289 publications and the cells are = 1 if authors (in rows) authored a paper (in columns)

Data Definition (categorical variables)

For ***authors***

X = matrix 481 x 3
3 columns expanded in dummy coding of the 3 + 5 + 4 categories:
Academic position: *Assistant, Associate, full professor*
Scientific sub-sector (italian classification): *S01-Statistics; S02-Statistics for experimental and technological research; S03- Economic statistic; S04-Demography; S05-Social statistics*
Geo-localization of university*: North-west; North-East; Centre; South Italy*

For ***publications***:

Z = matrix 2 x 2289
2 rows expanded in 3 + 3 dummy coding
Number of authors per publication <4 Auth; 4-10 Auth; >10 Auth;
Year of publication (1989-96; 1997-03; 2004-10)

DATA PRE-TREATMENT

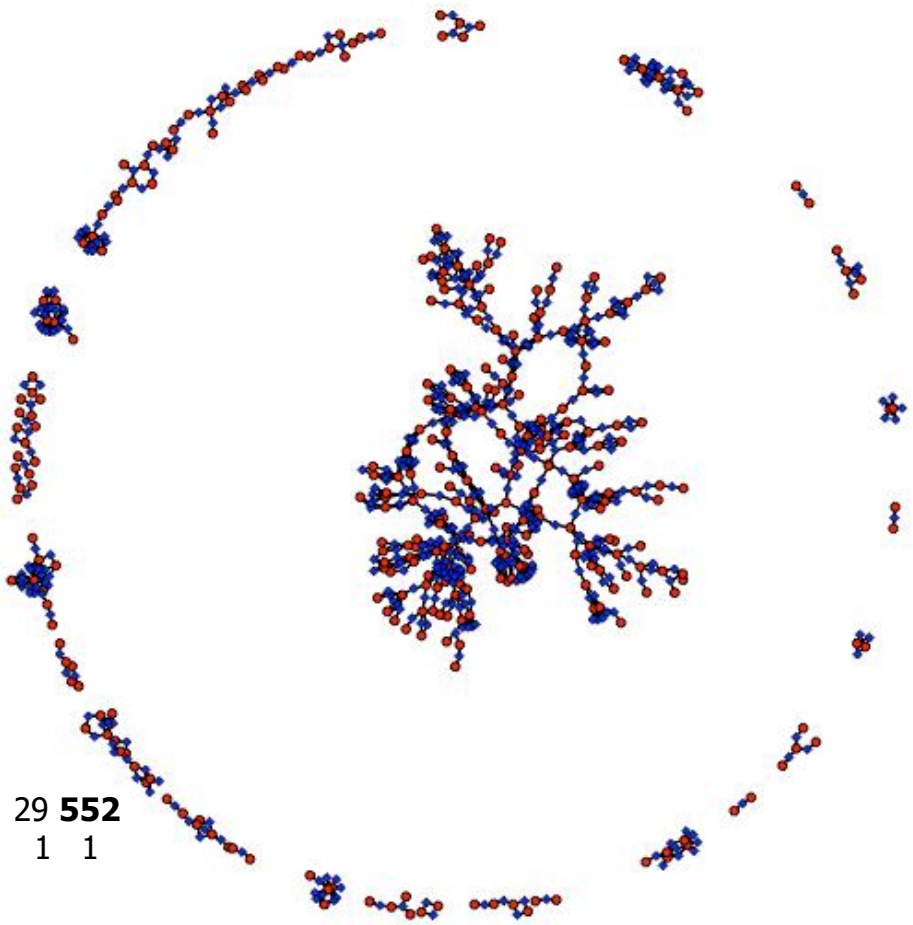
Since we are interested in co-authorship among italian scholars we cut out external link reducing the initial dataset to an affiliation matrix of

333 Authors and 526 Papers

Bipartite network

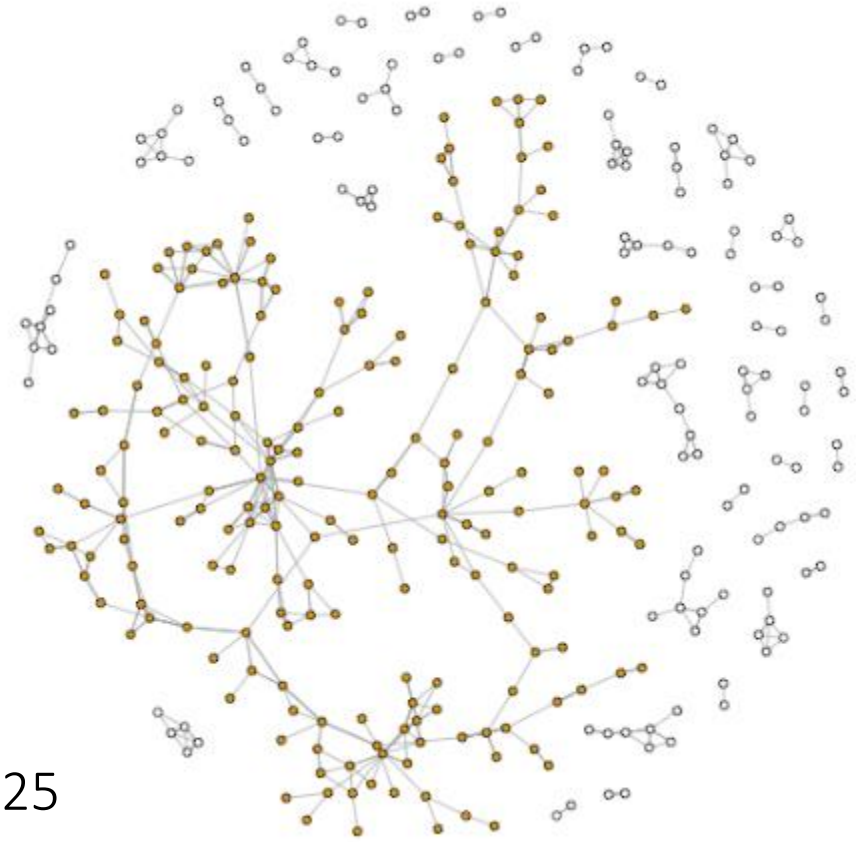
Papers

Authors



Component Size	3	4	5	6	7	8	11	13	14	16	19	23	25	29	**552**
# of components	14	4	4	2	3	4	1	1	2	1	1	1	1	1	1

Authors x Authors projection (size: 333)

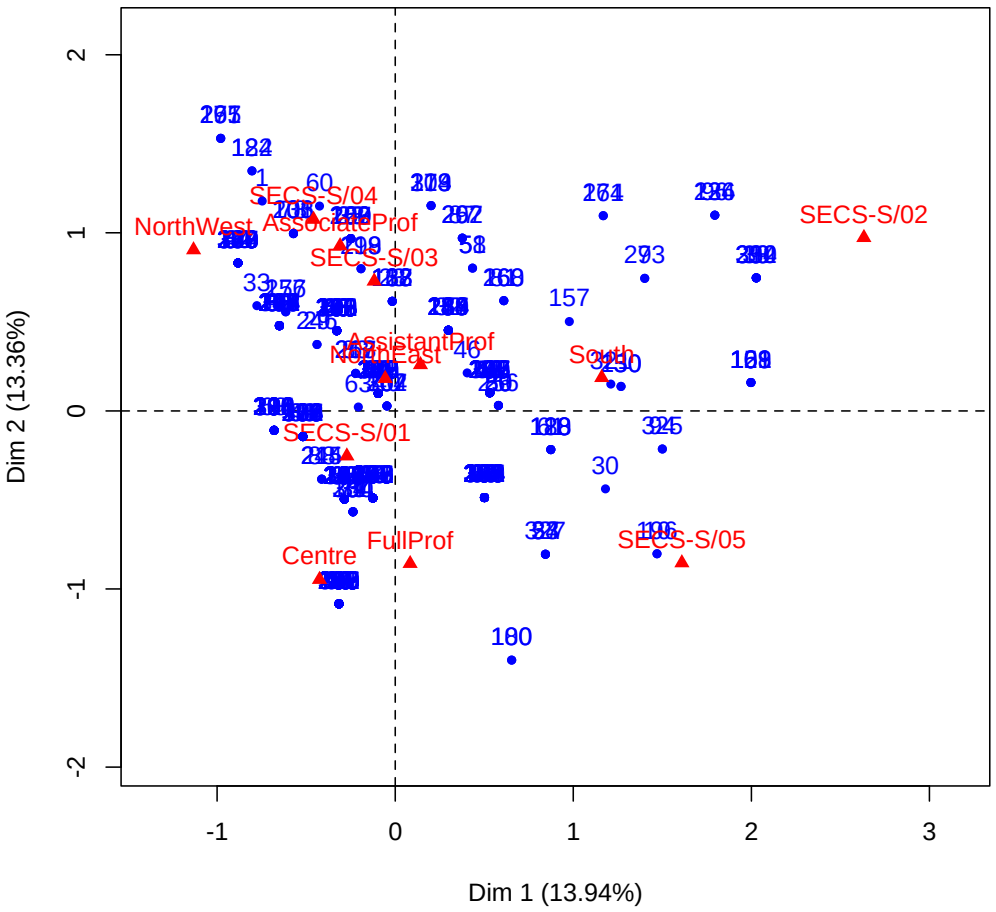


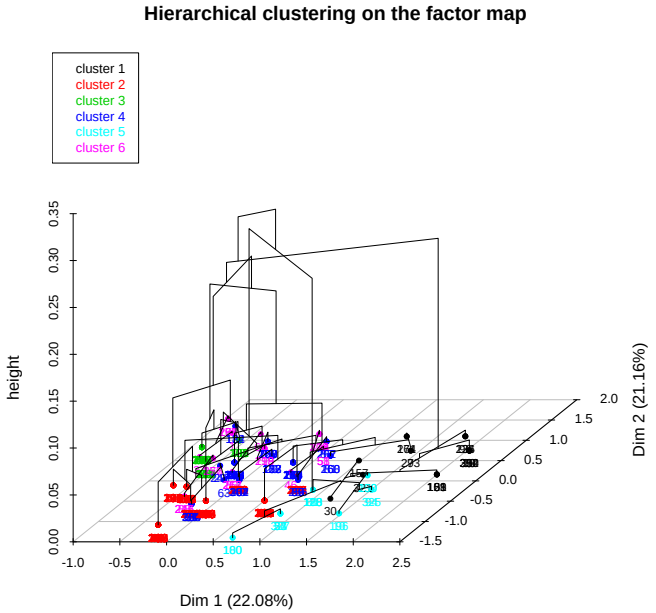
Graph density = 0.013

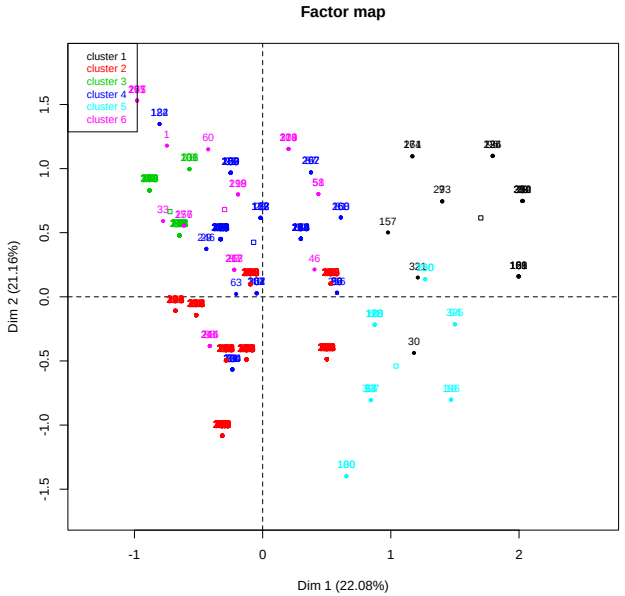
Largest Component density = 0.025

Component Size	2	3	4	5	6	7	8	197
# of components	20	5	5	5	1	2	2	1

Analysis of Attribute Data in X(MCA):

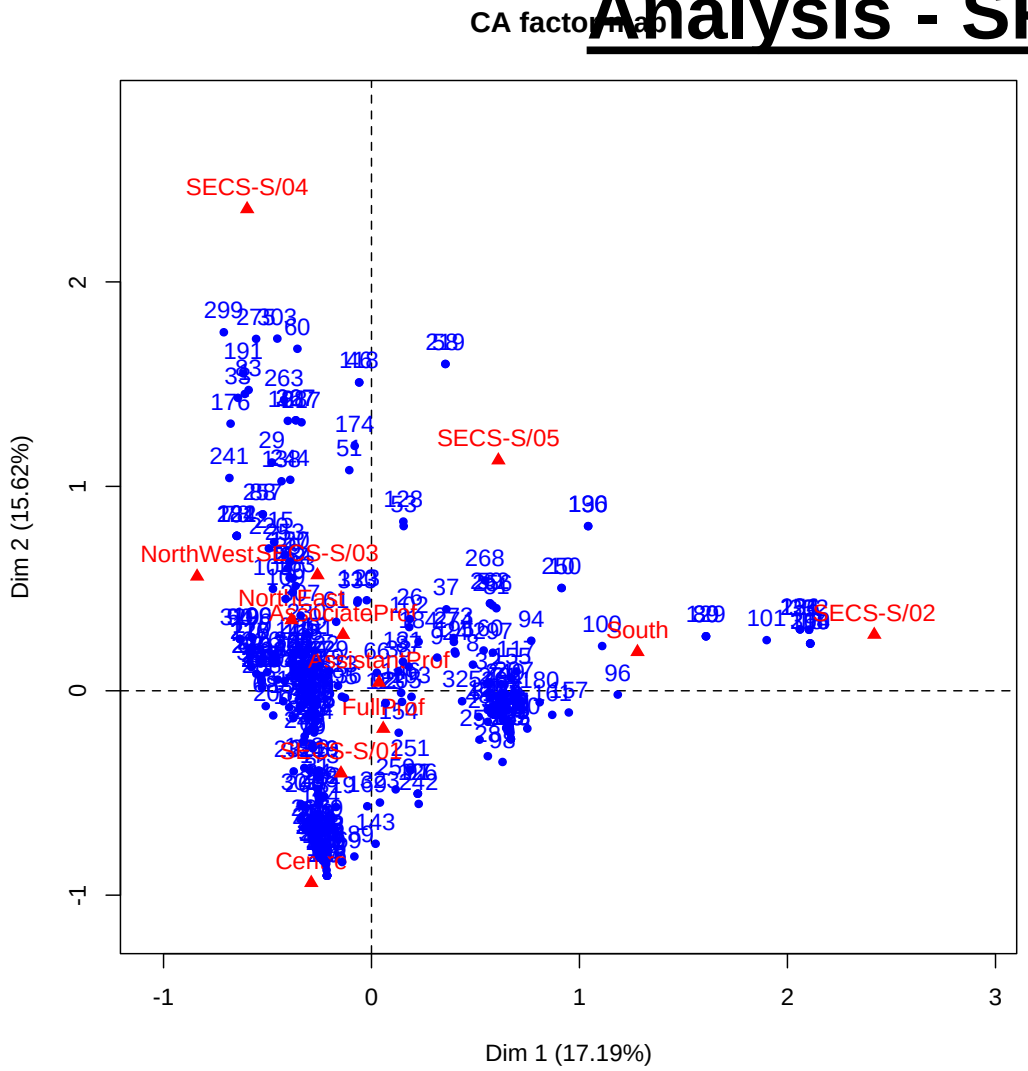






Aim: to explore the association structure between authors characteristics (**attribute variables on actors**)

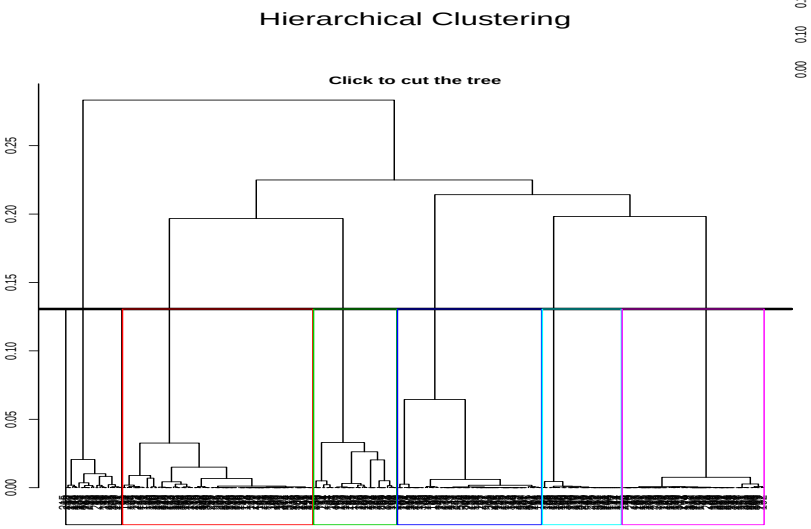
Looking at association structure on attribute variables: Smoothed Multiple Correspondence Analysis - SFA

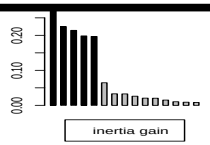


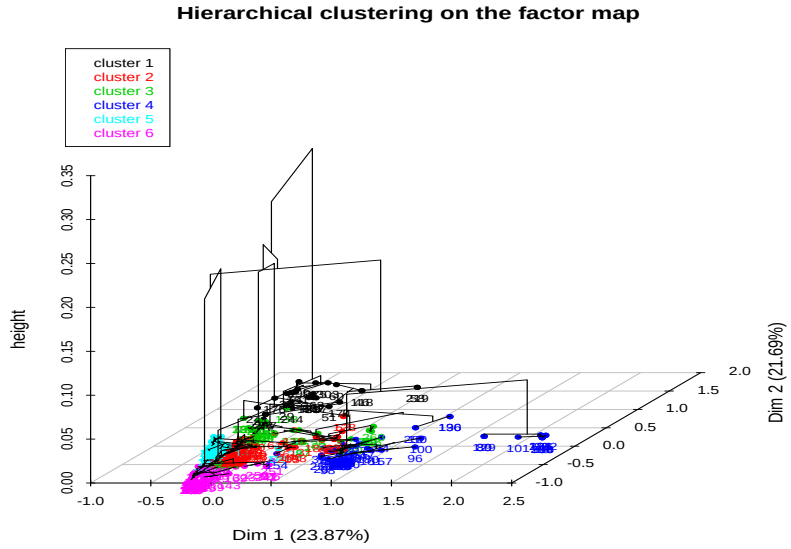
Analysis of $\mathbf{N^{-1}GQ}$

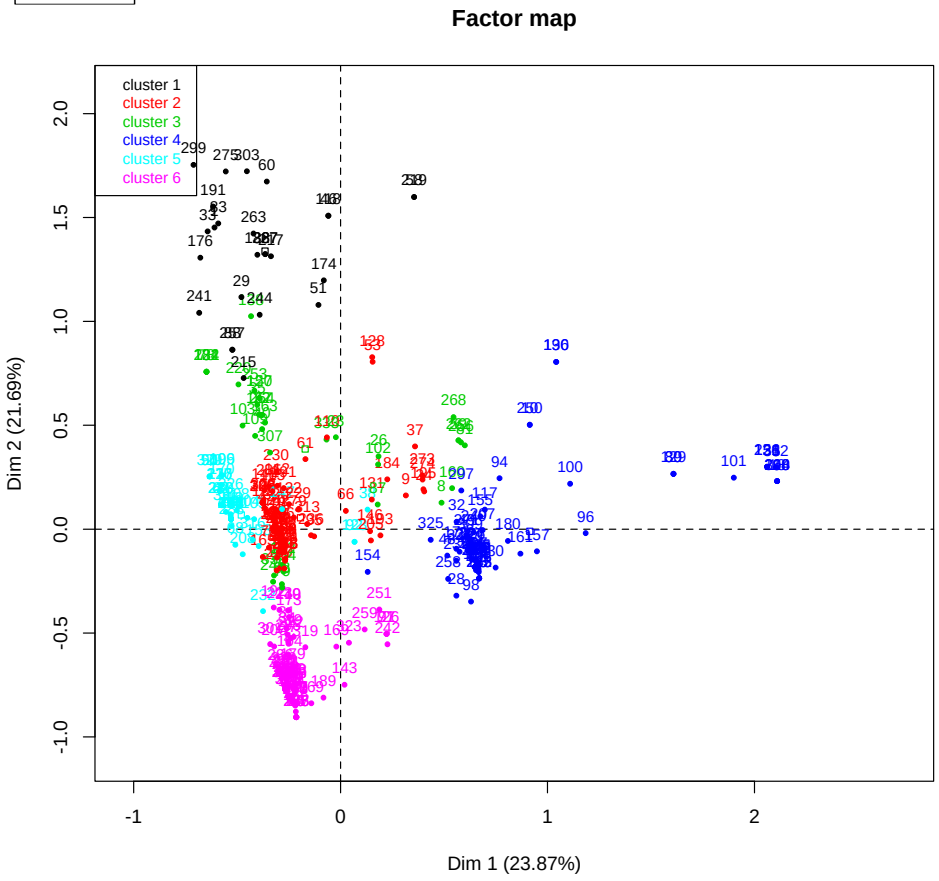
The factorial map highlights the characteristics of contiguous scholars. For instance the nodes are the barycentre of their neighbours and lye close to characteristics which are peculiar of their ego-centered network (Sector, Role,..)

Cluster Analysis on the Results of the S-MCA

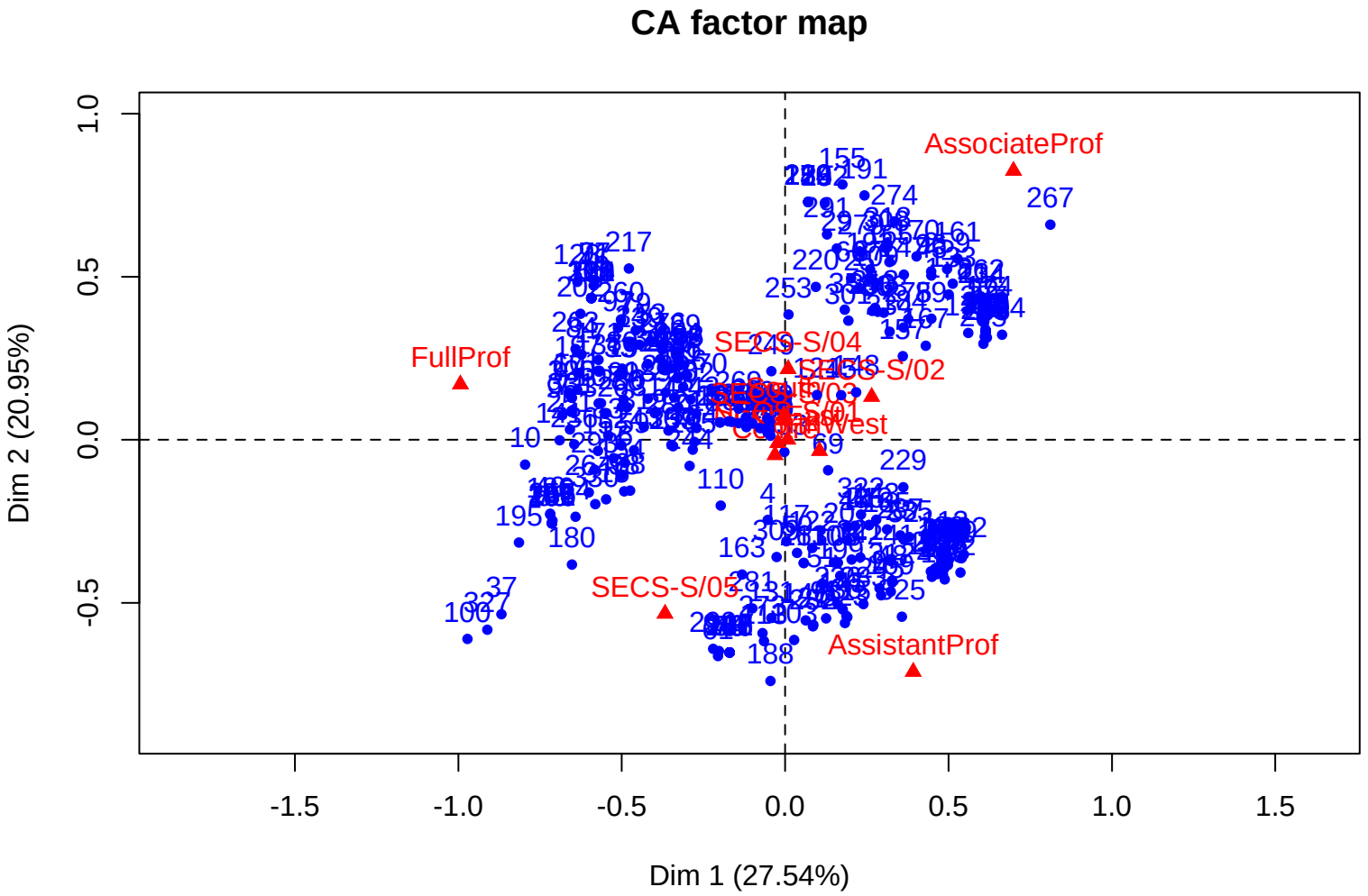








FALD: Correspondence Analysis of $\mathbf{Q} - \mathbf{M}^{-1}\mathbf{G}\mathbf{Q} + \frac{\mathbf{q}_{i.}\mathbf{q}'_{.j}}{\mathbf{q}_{..}}$



Residual information after leaving out the contiguity relationships

Attribute far from the origin explain the residual variance: They **are not important in explaining a network effect**

CONCLUDING REMARKS – AIM 1

SFA and FALD can be used to explore the relationships between the network structure and attribute variables…

The factorial maps show the author position as a function of the attributes in Q and the contiguity structure…

How to combine Network Data and external information in SNA?

Aim 2

AIM 2 -Notation and Definition

A $_{(nxp)}$ affiliation matrix (binary); *n* actors; *p* events

a_{ij} = 1 -< *i*-th actor is present at the *j*-th event

(i=1, …, n; j=1, …, p)

X $_{(nxm)}$ *n actors; r* nominal variables expanded into *m* dummy variables

x_{ik} = 1 -< *i*-th actor belongs to the *k*-th category

(i=1, …, n; k=1, …, m)

Z$_{(qxp)}$ *p* events; *s* nominal variables expanded into *q* dummy variables

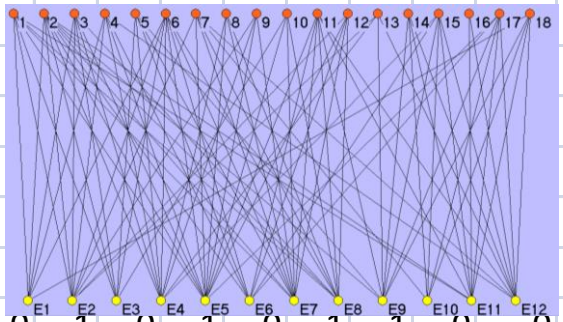
z_{hj} = 1 -< *j*-th event belong to the *h*-th category

(h=1, …, q; j=1, …,p)

Data structure

Z	E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	E11	E12
Z11	1	1	1	1	0	0	0	0	0	0	0	0
Z12	0	0	0	0	1	1	1	1	0	0	0	0
Z13	0	0	0	0	0	0	0	0	1	1	1	1

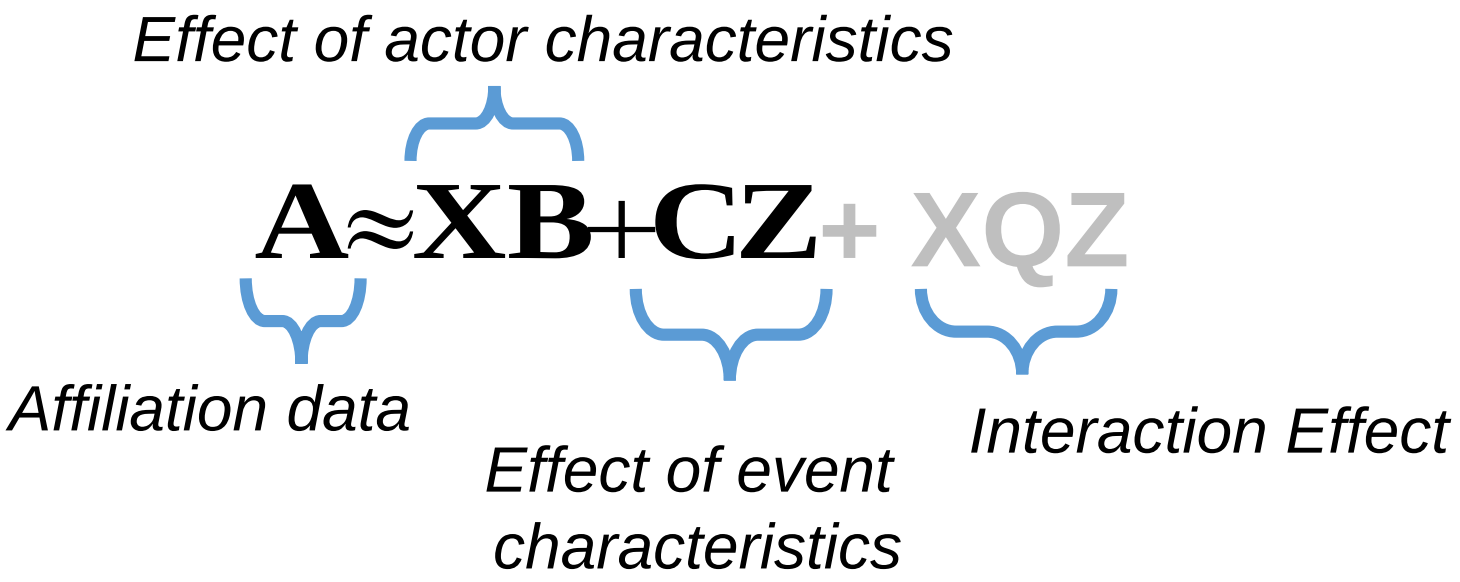
A	E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	E11	E12
1	1	1	1	0	1	0	1	1	0	0	1	0
2	1	1	1	1	1	1	1	1	1	0	1	0
3	1	1	1	1	1	0	0	1	0	0	0	0
4	1										0	1
5	0										0	0
6	1										0	0
7	0										0	1
8	0										0	0
9	0										0	0
10	0	0	1	0	1	0	1	1	0	0	0	0
11	0	0	0	0	1	1	1	1	1	0	1	1
12	0	1	0	0	1	1	1	1	0	0	0	0
13	0	1	0	1	0	0	0	0	1	0	0	1
14	0	0	0	0	0	0	0	0	1	1	1	1
15	0	0	0	1	0	1	0	0	1	1	1	1
16	0	0	0	0	0	0	0	0	0	0	1	1
17	0	0	0	0	1	1	0	0	1	1	1	1
18	1	0	0	0	0	0	0	0	1	1	1	1



X	X11	X12	X13	X21	X22
1	1	0	0	1	0
2	1	0	0	1	0
3	1	0	0	1	0
4	1	0	0	1	0
5	1	0	0	1	0
6	1	0	0	1	0
7	0	1	0	1	0
8	0	1	0	1	0
9	0	1	0	1	0
10	0	1	0	0	1
11	0	1	0	0	1
12	0	1	0	0	1
13	0	0	1	0	1
14	0	0	1	0	1
15	0	0	1	0	1
16	0	0	1	0	1
17	0	0	1	0	1
18	0	0	1	0	1

… The Underlying Decomposition Model

(Takane, Shibayama, 1991)



Effect of actor characteristics on relational data

$$\mathbf{A} \approx \mathbf{XB}$$

$$J(\mathbf{B}) = \underset{\mathbf{B}}{argmin} \left\| \mathbf{A} - \mathbf{XB} \right\|^2$$

$$\hat{\mathbf{B}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{A}$$

particular case of orthogonal design X :

$$\hat{\mathbf{B}} = [diag(\mathbf{X}'\mathbf{X})]^{-1}\mathbf{X}'\mathbf{A}$$

statistical interpretation as conditional frequency

$b_{k,j} \in [0,1]$

=> Affiliation effect of the *k*-th category to the *j*-th event.

Effect of events characteristics in determining the presence of actors at events

$$\mathbf{A} \approx \mathbf{CZ}$$

$$\mathbf{J}(C) = \underset{\mathbf{C}}{argmin} \left\| \mathbf{A} - \mathbf{CZ} \right\|^2$$

we retain the particular solution

$$\hat{\mathbf{C}} = [diag(\mathbf{Z}'\mathbf{Z})]^{-1}\mathbf{Z}'\mathbf{A}$$

C_{hi} affiliation effect due to the presence of actors at event given the *h*-th event's category

Use of the coefficients B and C in SNA

The **B** e **C** coefficients have been used to approximate the original **affiliation matrix A** and to derive the **adjacency matrices G_x** and **G_z** (*actors* x *actors*)

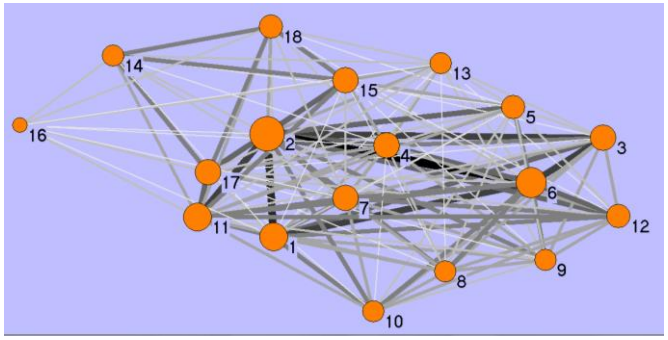
$$\hat{\mathbf{B}} \Rightarrow \hat{\mathbf{A}}_x \Rightarrow \mathbf{G}_x$$

$$\hat{\mathbf{C}} \Rightarrow \hat{\mathbf{A}}_z \Rightarrow \mathbf{G}_z$$

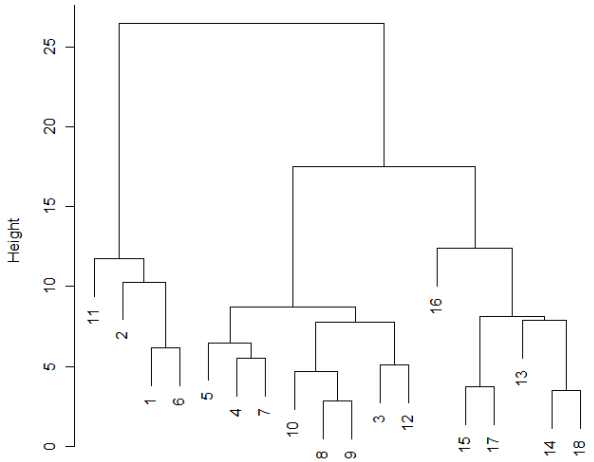
G_x and **G_z** are then analyzed by SNA methods specially to look for peculiar patterns of ties and homogeneous groups of actors induced by the effect of the external information.

From A to G: Clustering of G

$$\mathbf{G} = \mathbf{A} \cdot \mathbf{A}'$$

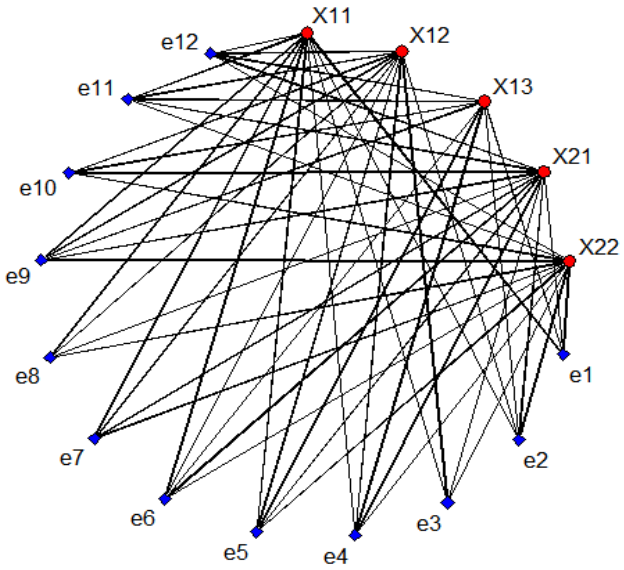


Hierarchical clustering of network positions



Analysis of X and A

B (m x p)

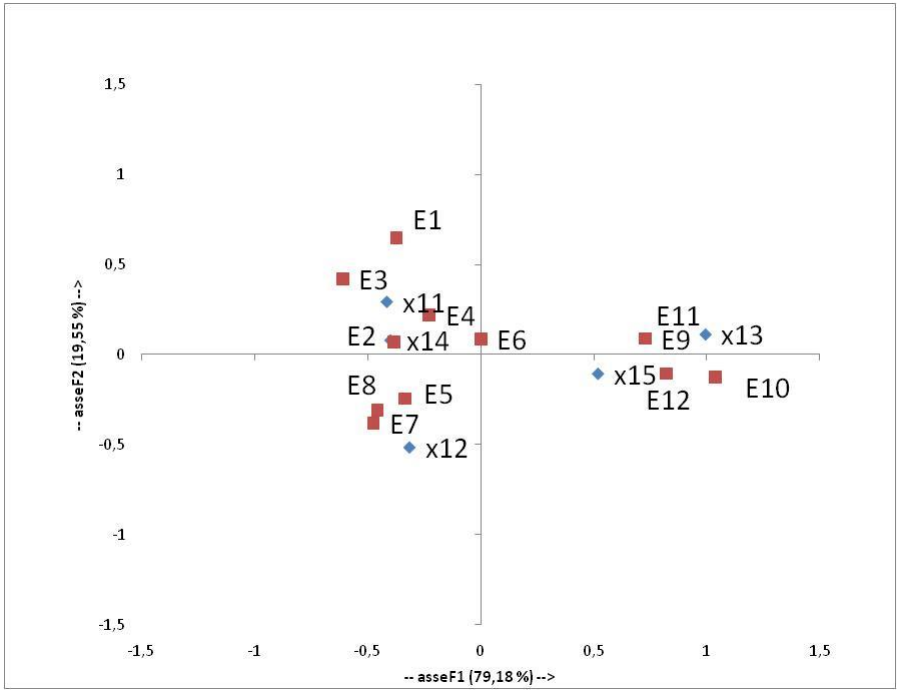


maximum value of the columns marginal of **B** = *r* x *m*

(all *n* actors are present at the *j*-th event)

maximum value of rows marginal of **B** = *p*
(a category fully characterizes all *p* events)

Proximity categories and actors



$$\hat{\mathbf{A}}_{\mathbf{X}} = \mathbf{X}\hat{\mathbf{B}}$$

column maximum value = 1
row marginal maximum theoretical value = r x p

Analysis of X and A

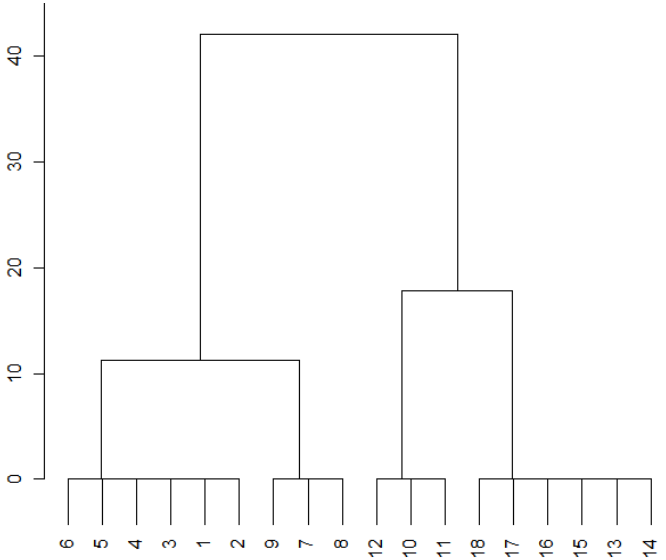
$$\mathbf{G_X} = \hat{\mathbf{A}}_\mathbf{X} \cdot \hat{\mathbf{A}}'_\mathbf{X}$$

whose generic element weights the joint presence at events of actors having similar characteristics. Maximum theoretical value in G_X

$$\sum_{j=1}^{n} r^2 = p \cdot r^2$$

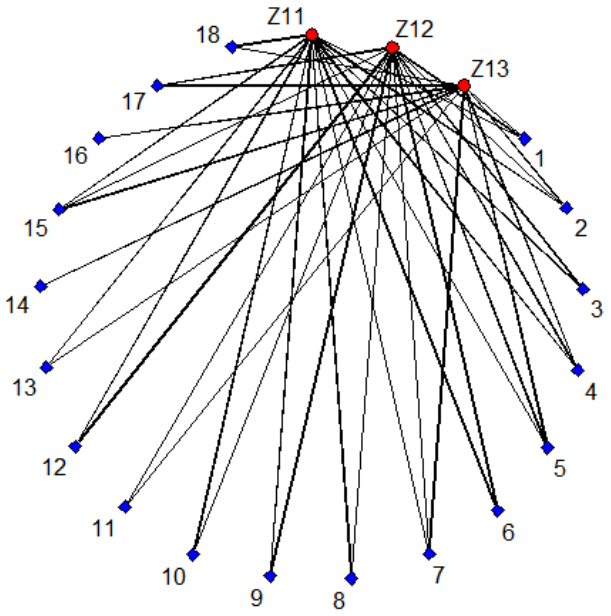
If a *homophily effect* is present, then G_X shows a weighting system coherent with the capability of all actors characteristics to jointly explain the participation in events.

- Clustering / Blockmodeling

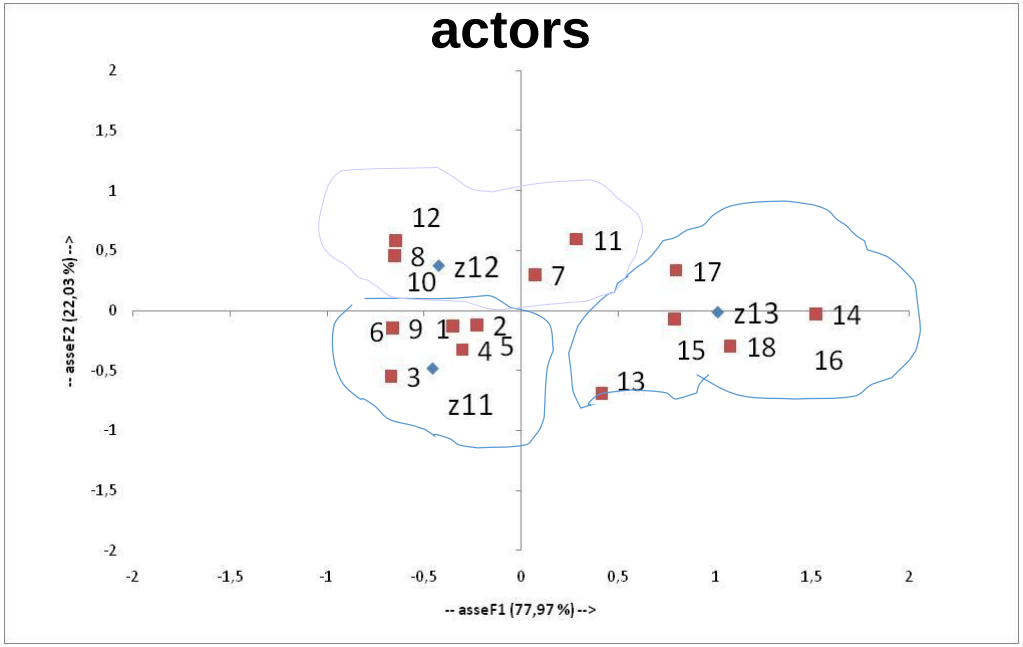


Analysis of Z and A

C (n x q)



Proximity between Z categories and actors



Results …

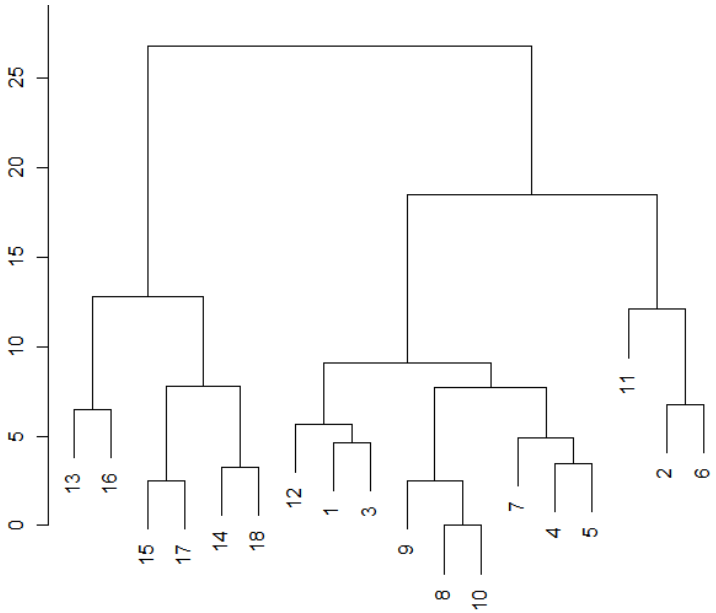
$$\hat{\mathbf{A}}_Z = \hat{\mathbf{C}}\mathbf{Z}$$

Analysis of Z and A

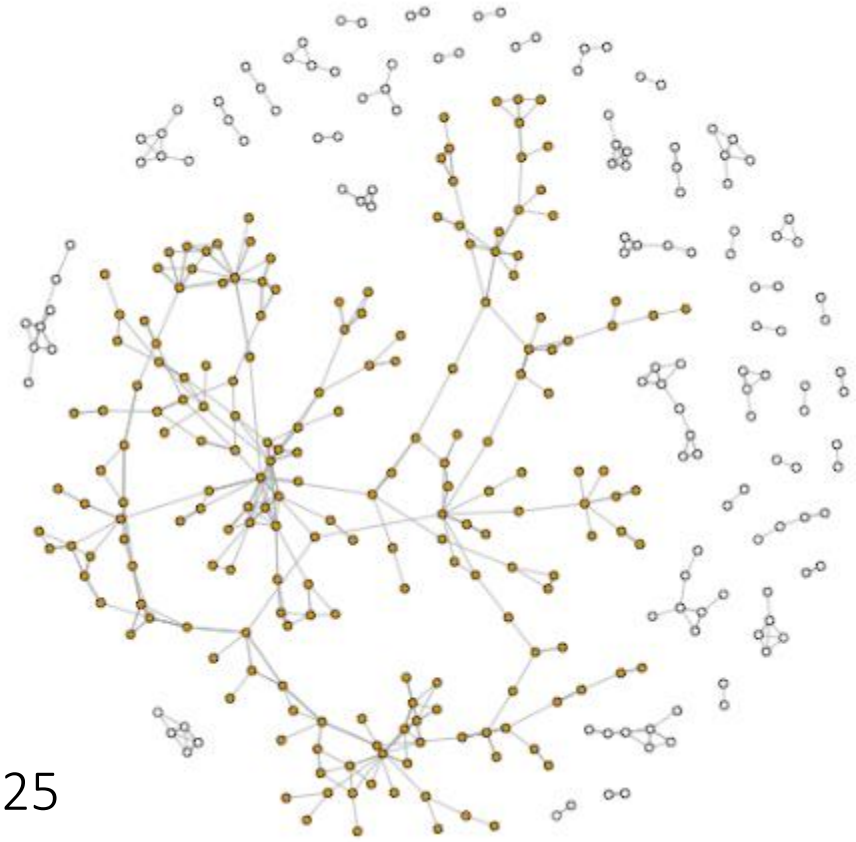
To derive from $\hat{\mathbf{A}}_Z$ the weighted adjacency $\mathbf{G_Z}$: $\mathbf{G}_Z = \hat{\mathbf{A}}_Z \cdot \hat{\mathbf{A}}'_Z$

where the generic element of $\mathbf{G_Z}$ is the weight of the tie between a pair of actors related to their presence at the same categories of events

Clustering (or Blockmodeling)



Authors x Authors projection (size: 333)



Graph density = 0.013

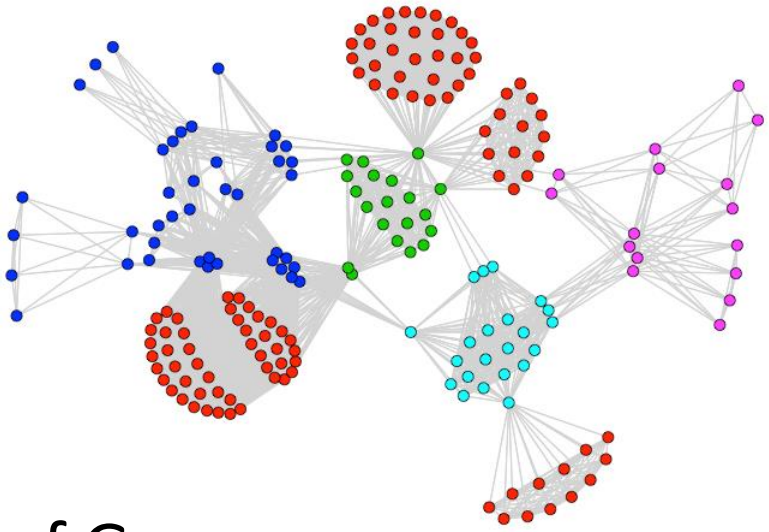
Largest Component density = 0.025

Component Size	2	3	4	5	6	7	8	197
# of components	20	5	5	5	1	2	2	1

Authors Attributes in X and Coefficients B

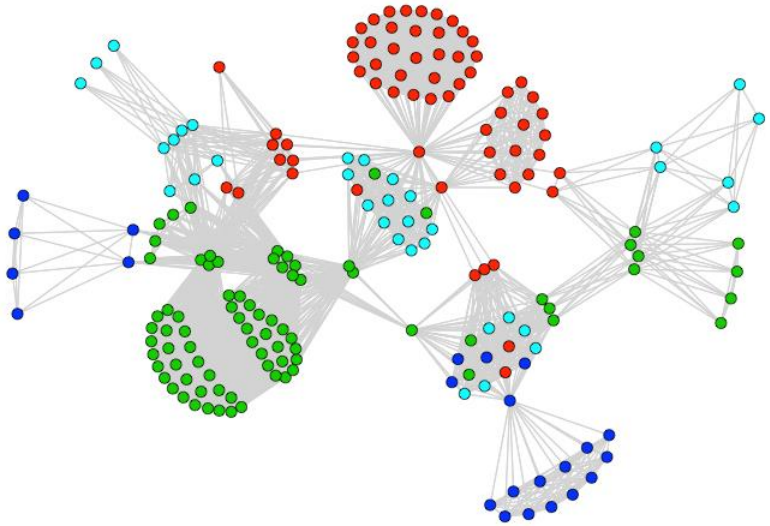
$$\mathbf{G}_{\mathbf{X}} = \hat{\mathbf{A}}_{\mathbf{X}} \cdot \hat{\mathbf{A}}'_{\mathbf{X}}$$

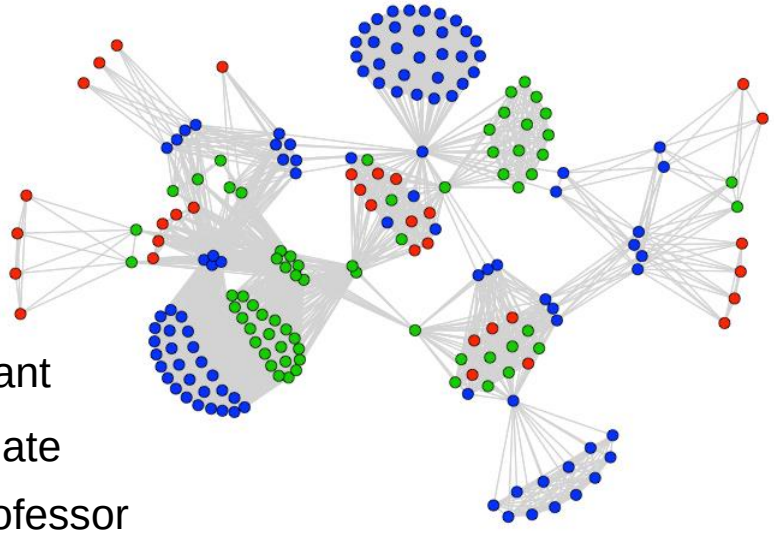
- Secs-S01
- Secs-S02
- Secs-S03
- Secs-S04
- Secs-S05



- North-west
- North-east
- Centre
- South

Analysis of G_x

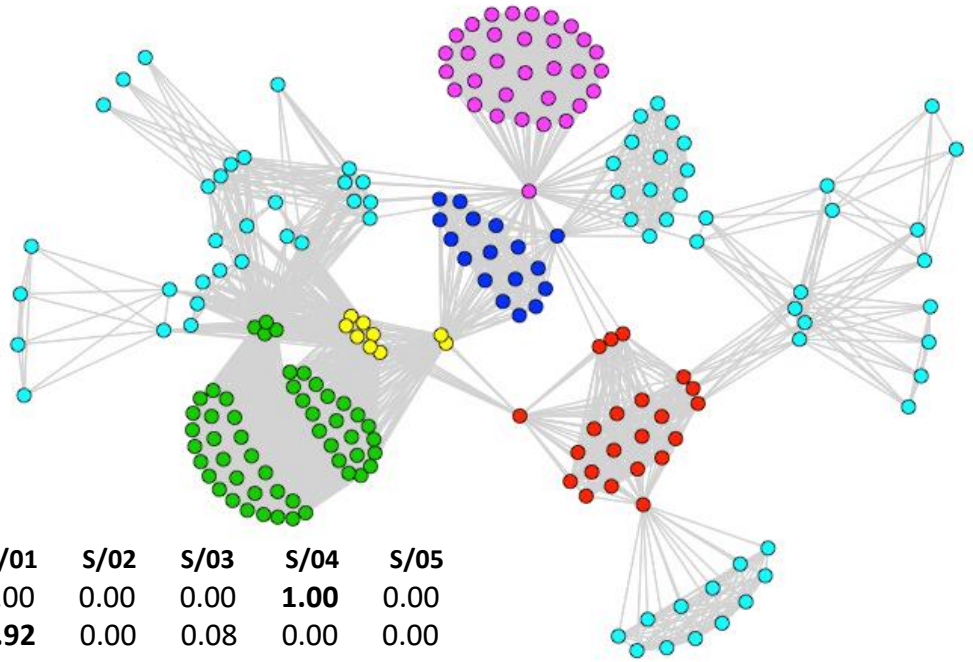




- Assistant
- Associate
- Full professor

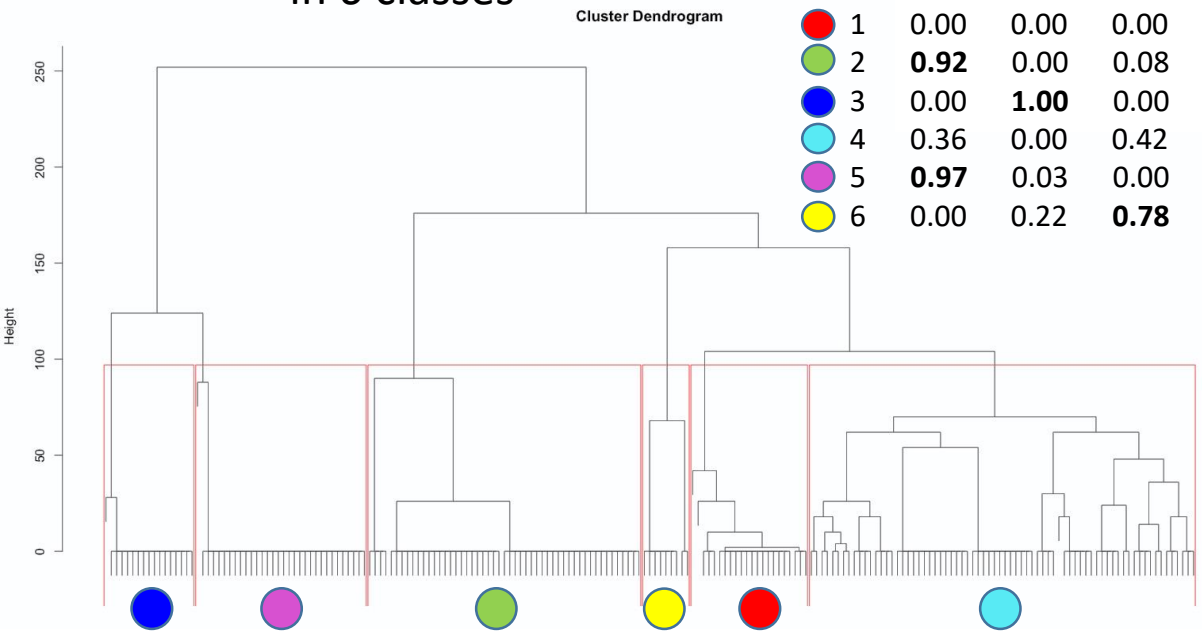
Structural Equivalence - Cluster Analysis

K	Centre	NorthEast	NorthWest	South
1	0.23	0.27	0.23	0.27
2	0.00	**1.00**	0.00	0.00
3	0.12	0.12	0.00	**0.76**
4	0.35	0.18	0.25	0.22
5	**1.00**	0.00	0.00	0.00
6	0.00	**1.00**	0.00	0.00



Analysis of G_x

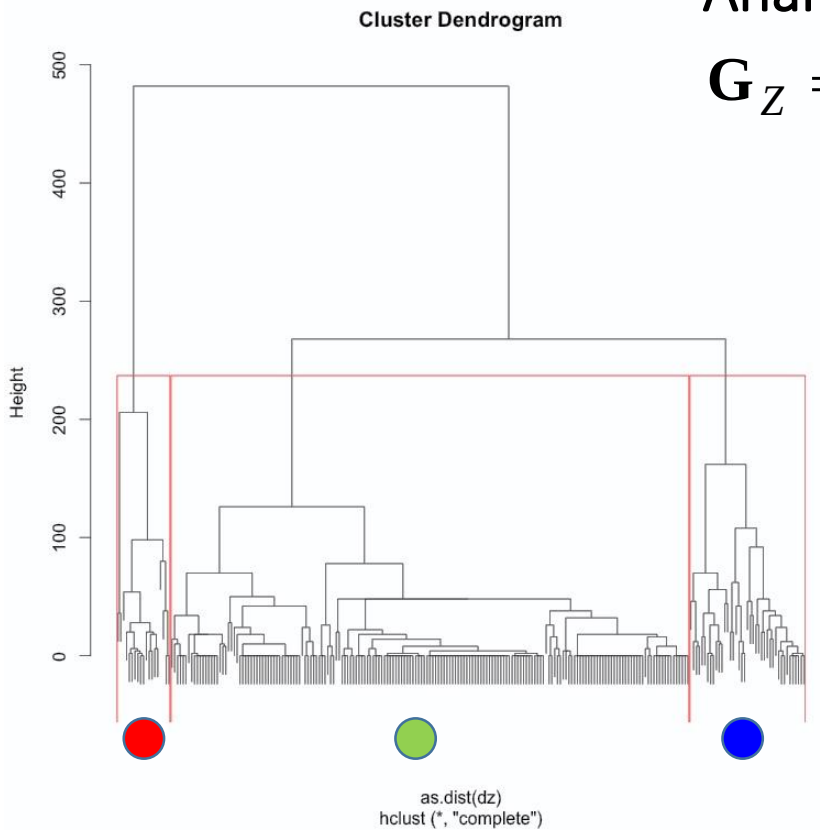
Importance of Coefficients B in 6 classes



K	S/01	S/02	S/03	S/04	S/05
1	0.00	0.00	0.00	**1.00**	0.00
2	**0.92**	0.00	0.08	0.00	0.00
3	0.00	**1.00**	0.00	0.00	0.00
4	0.36	0.00	0.42	0.00	0.22
5	**0.97**	0.03	0.00	0.00	0.00
6	0.00	0.22	**0.78**	0.00	0.00

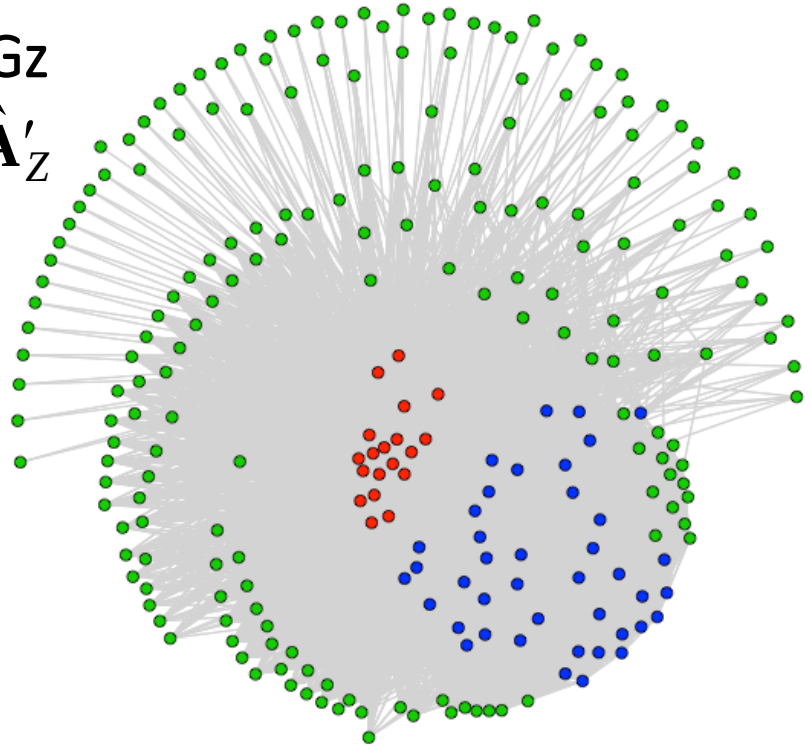
k	AssistantProf	AssociateProf	FullProf
1	0.23	0.41	0.36
2	0.00	**0.41**	**0.59**
3	**0.53**	0.24	0.24
4	0.26	0.32	0.42
5	0.00	0.00	**1.00**
6	0.00	**1.00**	0.00

Papers Attributes in Z and coefficients C



Analysis of Gz

$$\mathbf{G}_Z = \hat{\mathbf{A}}_Z \cdot \hat{\mathbf{A}}'_Z$$



Weighted importance
of coefficients **C** for the 3 classes

k	1989-96	1997-03	2004-10	Auth <4	Auth 4-10	Auth >10
1	**0.24**	0.09	0.08	0.07	**0.27**	0.12
2	0.05	0.03	0.03	0.03	0.06	0.11
3	0.08	0.07	0.05	0.05	0.06	**0.26**

Some concluding remarks

Relational and attribute data

- to derive ad hoc relational data structures (affiliation and adjacency matrices)

- to enhance the interpretation of traditional network analysis from a different point of view:

i) the complementary use of valued graphs defined according to observed auxiliary information;

ii) the possibility to introduce explicative measures joining external information and relational data

iii) the interpretation of the results as complex data where groups of actors are defined and interpreted as "second order" individuals

Some references

Aluja, Banet T., Lebart, L. (1984). Local and Partial Principal Component Analysis and Correspondence Analysis. In: Havranek, T., Sidak, Z., Novak, M. (Eds.), *COMPSTAT Proceedings*, Phisyca-Verlag, Vienna. pp. 113-118.

Benali, H., Escofier, B. (1990). Analyse factorielle lissèe et analyse factorielle des diffèrences locales. *Revue de Statistique Appliquèe.* **38**, pp. 55-76.

De Stefano D., Fuccella V., Vitale M.P., Zaccarin S. (2013). The use of different data sources in the analysis of co-authorship networks and scientific performance. *Social Networks.* **35**, pp. 370-381.

D'Esposito, M. R., De Stefano, D., Ragozini, G. (2014). On the use of Multiple Correspondence Analysis to visually explore affiliation networks. *Social Networks*. **38**. pp. 28–40.

Ferligoj, A., Kronegger, L. (2009). Clustering of Attribute and/or Relational Data. *Metodoloski Zvezki.* **6**, 135-153.

Giordano G., Vitale M.P. (2007). Factorial Contiguity Maps to Explore Relational Data Patterns. *Statistica applicata*. **19**, pp. 297-306.

Giordano G., Vitale M.P. (2011). On the use of auxiliary information in Social Network Analysis. *Advances in Data Analysis and Classification* (ADAC). **5**, pp. 95-112.

Lazarsfeld, P., Merton, R. K. (1954). Friendship as a Social Process: A Substantive and Methodological Analysis. In*: Freedom and Control in Modern Society*, Berger, M., Abel, T., H. Page, C. (eds.). Van Nostrand, New York, 18-66.

Maddala, G.S. (1991). A Perspective on the Use of Limited-Dependent and Qualitative Variables Models in Accounting Research. *The Accounting Review .* **66**, pp. 788-807.

McPherson, M., Smith-Lovin, L., Cook., J. (2001). Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology*. **27**, pp. 415-44.

Opsahl, T., Panzarasa, P. (2009). Clustering in weighted networks. *Social Networks.* **31**, pp. 155-163

Ragozini, G., De Stefano, D., D'Esposito, M. R., (2015). Multiple factor analysis for time-varying two-mode networks. *Network Science.* **3**, pp. 18-36.

Robins, G.L., Pattison, P., Kalish, Y., Lusher,D. (2007). An introduction to exponential random graph (p*) models for social networks. *Social Networks.* **29**, pp. 173-191.

Takane Y. Shibayama T. (1991). Principal component analysis with external information on both subjects and variables, Psychometrika. **56**, issue 1, pp. 97-120.