Scientific Collaboration in Slovenian Scientific Communities: The stability of co-authorship structures

Anuška Ferligoj
Marjan Cugmas, Luka Kronegger

Faculty of Social Sciences, University of Ljubljana, Slovenia
NRU Higher School of Economics, Moscow, Russia

International Summer School on Social Network Analysis, Moscow July, 2019

Overview of studies on scientific collaboration in Slovenia

In the last ten years we have studied the scientific collaboration (SC) using

- bibliometric analysis (co-authorship structures - Nuša, key factors driving scientific collaboration - Luka)
- survey analysis,
- qualitative approach

of co-authorship networks. We used longitudinal data on the Slovenian science system in order to explore and explain their dynamics in the scientific fields and the scientific disciplines.

Bibliometric analysis – co-authorship structures

Using **blockmodeling**, we will show in this presentation how co-authorship structures change in time in all Slovenian scientific disciplines.

The aim of this presentation

- To characterize the blockmodel structures of the Slovenian scientific co-authorship networks.
- To study the stability of the blockmodels of co-authorship networks in time.
- To reveal the differences in the global network structrues between different scientific disciplines (natural and technical sciences vs. social sciences and humanitites) and time periods (1990–2000 vs. 2001–2010).

Bockmodeling

One of the major goals of social network analysis is to discern fundamental structures of networks in ways that allow us to get insight into the structure of a network and to facilitate our understanding of network phenomena.

Positional analysis of social network data rests on the assumption that the role structure of the positions of individuals in the groups exist. The key tasks here are (Faust and Wasserman, 1992):

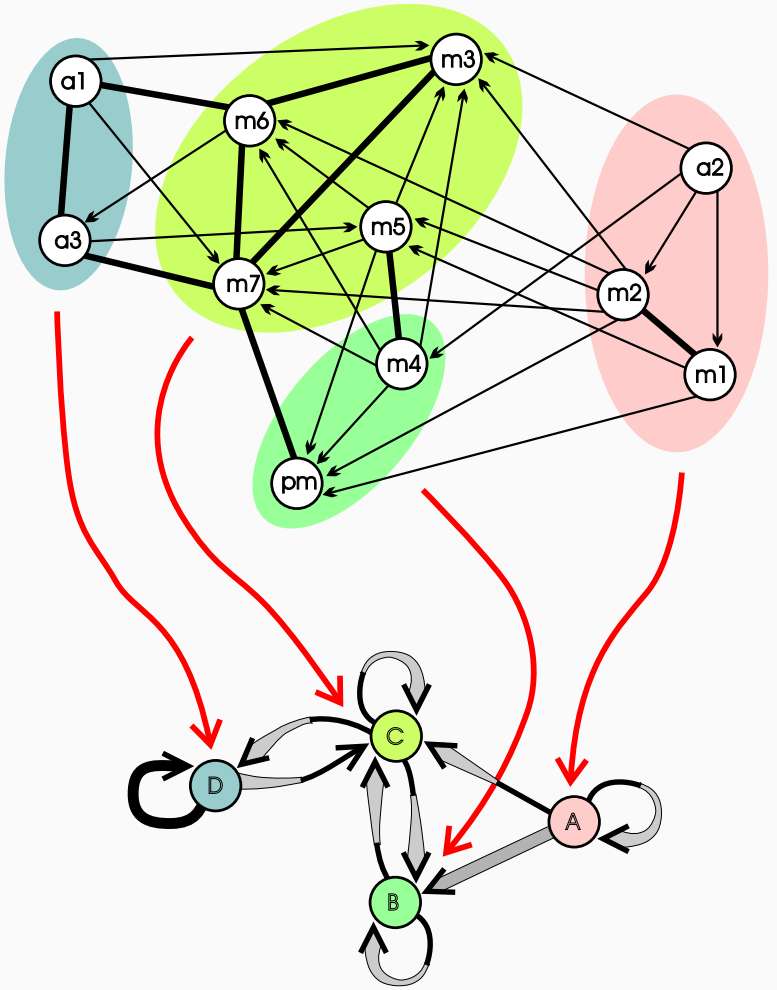
- identifying *social positions* as collections (groups, clusters) of units (actors) who are similar in their relationships to the others, and
- modeling *social roles* as system of relationships among positions.

Blockmodeling is dealing with these two aspects.

The goal of bockmodeling

The goal of *blockmodeling* is to reduce a large, potentially incoherent network to a smaller comprehensible structure that can be interpreted more readily.

Blockmodeling, as an empirical procedure, is based on the idea that units in a network can be grouped and form the positions according to the extent to which they are equivalent, according to some *meaningful* definition of equivalence.



Cluster, Clustering, Blocks

One of the main procedural goals of blockmodeling is to identify, in a given network $\mathbf{N} = (\mathcal{U}, R)$, $R \subseteq \mathcal{U} \times \mathcal{U}$, *clusters* (classes) of units that share structural characteristics defined in terms of R. The units within a cluster have the same or similar connection patterns to other units. They form a *clustering* $\mathbf{C} = \{C_1, C_2, \ldots, C_k\}$ which can be a *partition* of the set $\mathcal{U}$. Each partition determines an equivalence relation (and vice versa). Let us denote by $\sim$ the relation determined by partition $\mathbf{C}$.

A clustering $\mathbf{C}$ partitions also the relation R into *blocks*

$$R(C_i, C_j) = R \cap C_i \times C_j$$

Each such block consists of units belonging to clusters C_i and C_j and all arcs leading from cluster C_i to cluster C_j. If $i = j$, a block $R(C_i, C_i)$ is called a *diagonal* block.

An example: The Everett Network

	a	b	c	d	e	f	g	h	i	j
a	0	1	1	1	0	0	0	0	0	0
b	1	0	1	0	1	0	0	0	0	0
c	1	1	0	1	0	0	0	0	0	0
d	1	0	1	0	1	0	0	0	0	0
e	0	1	0	1	0	1	0	0	0	0
f	0	0	0	0	1	0	1	0	1	0
g	0	0	0	0	0	1	0	1	0	1
h	0	0	0	0	0	0	1	0	1	1
i	0	0	0	0	0	1	0	1	0	1
j	0	0	0	0	0	0	1	1	1	0

	a	c	h	j	b	d	g	i	e	f
a	0	1	0	0	1	1	0	0	0	0
c	1	0	0	0	1	1	0	0	0	0
h	0	0	0	1	0	0	1	1	0	0
j	0	0	1	0	0	0	1	1	0	0
b	1	1	0	0	0	0	0	0	1	0
d	1	1	0	0	0	0	0	0	1	0
g	0	0	1	1	0	0	0	0	0	1
i	0	0	1	1	0	0	0	0	0	1
e	0	0	0	0	1	1	0	0	0	1
f	0	0	0	0	0	0	1	1	1	0

	A	B	C
A	1	1	0
B	1	0	1
C	0	1	1

Equivalences

Regardless of the definition of equivalence used, there are two basic approaches to the equivalence of units in a given network (Faust, 1988):

- the equivalent units have the same connection pattern to the *same* neighbors;
- the equivalent units have the same or similar connection pattern to (possibly) *different* neighbors.

The first type of equivalence is formalized by the notion of structural equivalence and the second by the notion of regular equivalence with the latter a generalization of the former.

Structural Equivalence

Units are equivalent if they are connected to the rest of the network in *identical* ways (Lorrain and White, 1971). Such units are said to be *structurally equivalent*.

In other words, x and y are structurally equivalent iff:

s1. $\mathrm{x}R\mathrm{y} \Leftrightarrow \mathrm{y}R\mathrm{x}$

s2. $\mathrm{x}R\mathrm{x} \Leftrightarrow \mathrm{y}R\mathrm{y}$

s3. $\forall \mathrm{z} \in \mathcal{U} \setminus \{\mathrm{x}, \mathrm{y}\} : (\mathrm{x}R\mathrm{z} \Leftrightarrow \mathrm{y}R\mathrm{z})$

s4. $\forall \mathrm{z} \in \mathcal{U} \setminus \{\mathrm{x}, \mathrm{y}\} : (\mathrm{z}R\mathrm{x} \Leftrightarrow \mathrm{z}R\mathrm{y})$

. . . Structural Equivalence

The blocks for structural equivalence are null or complete with variations on diagonal in diagonal blocks.

$$\begin{matrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{matrix} \qquad \begin{matrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{matrix}$$

$$\begin{matrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{matrix} \qquad \begin{matrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{matrix}$$

Establishing Blockmodels

The problem of establishing a partition of units in a network in terms of a selected type of equivalence is a special case of *clustering problem* that can be formulated as an optimization problem (Φ, P) as follows:

Determine the clustering $\mathbf{C}^\star \in \Phi$ *for which*

$$P(\mathbf{C}^\star) = \min_{\mathbf{C} \in \Phi} P(\mathbf{C})$$

where Φ is the set of *feasible clusterings* and P is a *criterion function*.

Criterion function

Criterion functions can be constructed

- *indirectly* as a function of a *compatible* (dis)similarity measure between pairs of units, or
- *directly* as a function measuring the *fit* of a clustering to an ideal one with perfect relations within each cluster and between clusters according to the considered types of connections (equivalence).

Indirect Approach

RELATION

R

DESCRIPTIONS OF UNITS

Q

original relation

path matrix

triads

DISSIMILARITY MATRIX

D

STANDARD CLUSTERING ALGORITHMS

hierarchical algorithms,

relocation algorithm, k-means algorithm, etc.

Dissimilarities

The dissimilarity measure d is *compatible* with a considered equivalence $\sim$ if for each pair of units holds

$$x_i \sim x_j \Leftrightarrow d(x_i, x_j) = 0$$

Not all dissimilarity measures typically used are compatible with structural equivalence. For example, the *corrected Euclidean-like dissimilarity*

$$d(x_i, x_j) = \sqrt{(r_{ii} - r_{jj})^2 + (r_{ij} - r_{ji})^2 + \sum_{\substack{s=1 \\ s \neq i,j}}^{n} ((r_{is} - r_{js})^2 + (r_{si} - r_{sj})^2)}$$

is compatible with structural equivalence.

The indirect clustering approach does not seem suitable for establishing clusterings in terms of regular equivalence since there is no evident way how to construct a compatible (dis)similarity measure.

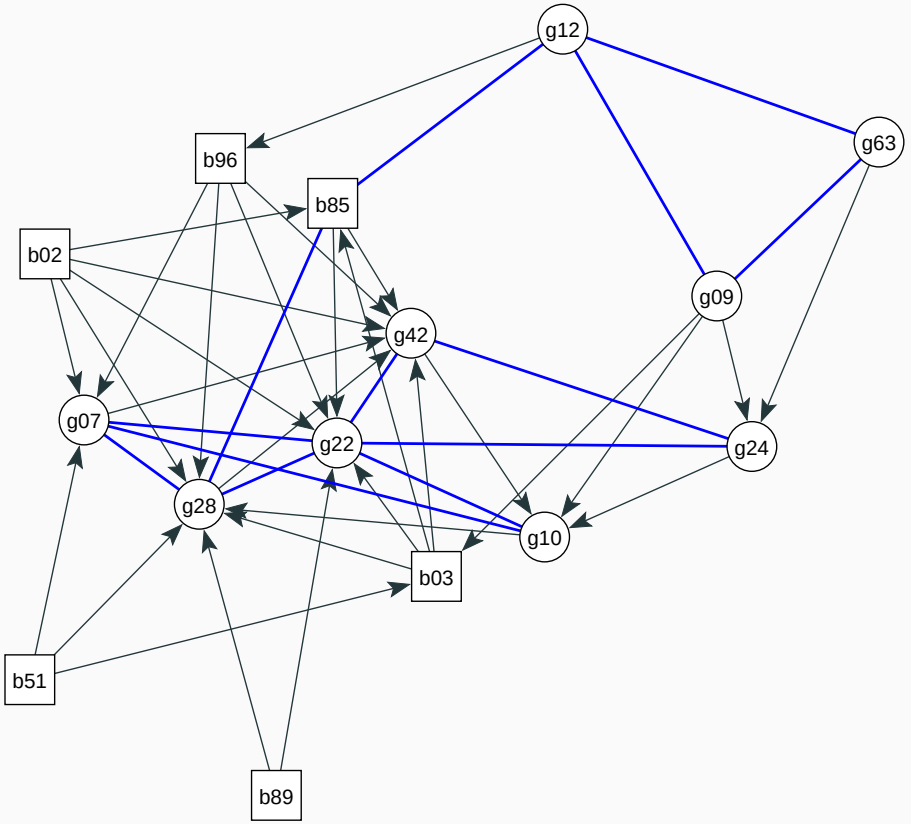
An example: Support network among informatics students

The analyzed network consists of social support exchange relation among fifteen students of the Social Science Informatics fourth year class (2002/2003) at the Faculty of Social Sciences, University of Ljubljana. Interviews were conducted in October 2002.

Support relation among students was identified by the following question:

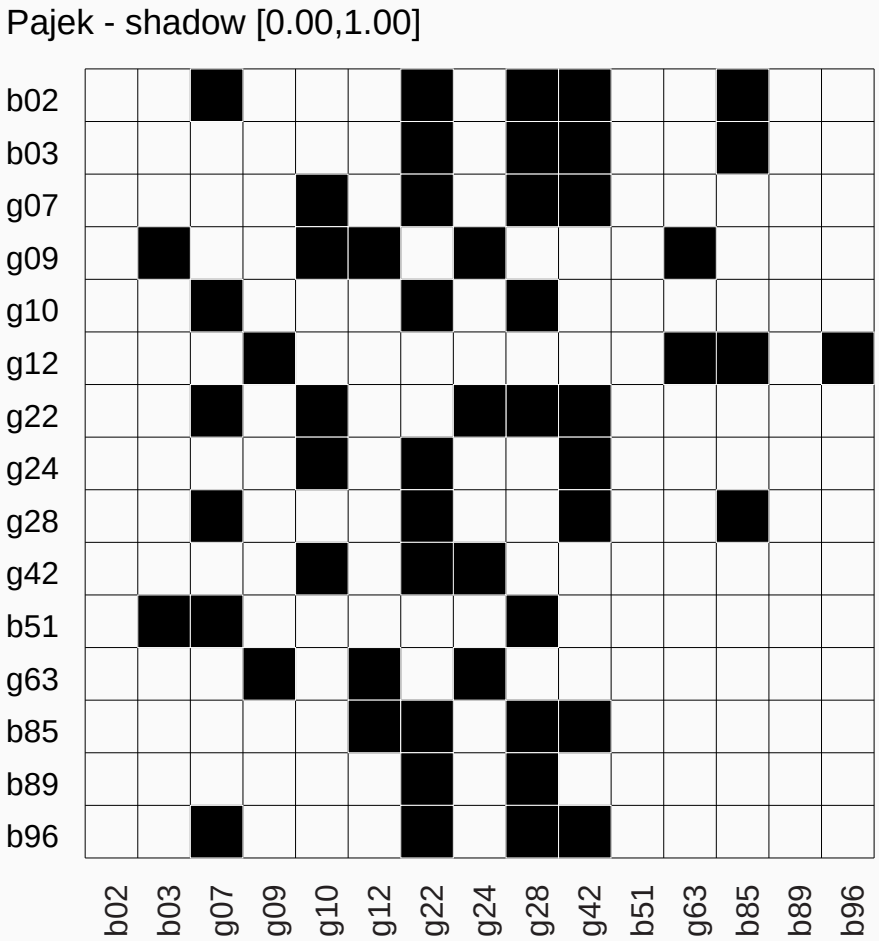
> *Introduction: You have done several exams since you are in the second class now. Students usually borrow studying material from their colleagues.*
> *Enumerate (list) the names of your colleagues that you have most often borrowed studying material from. (The number of listed persons is not limited.)*

Class network - graph

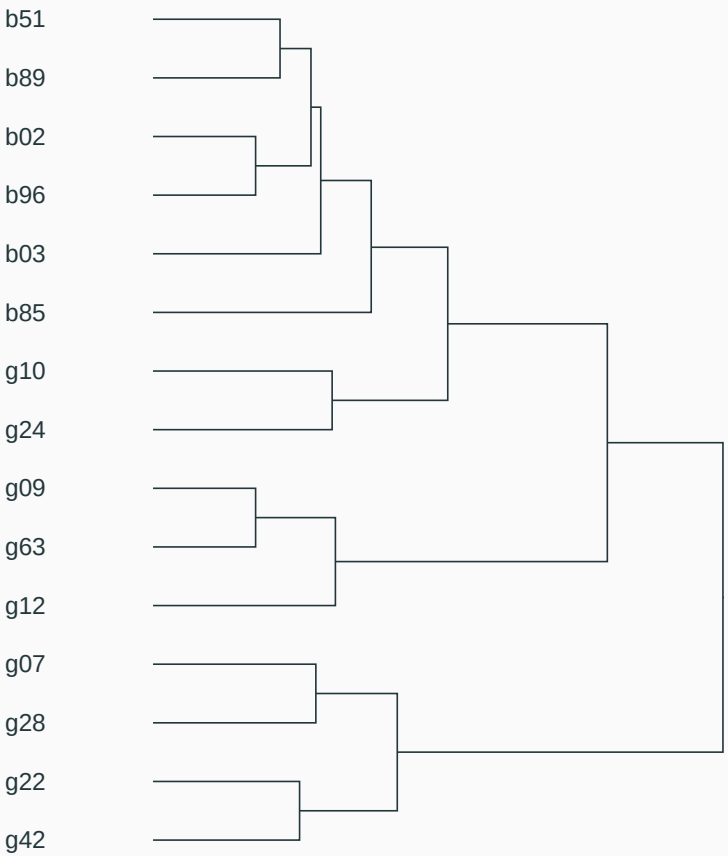


Vertices represent students in the class:
circles – girls, squares – boys.
Reciprocated arcs are represented by edges.

Class network – matrix



Indirect Approach



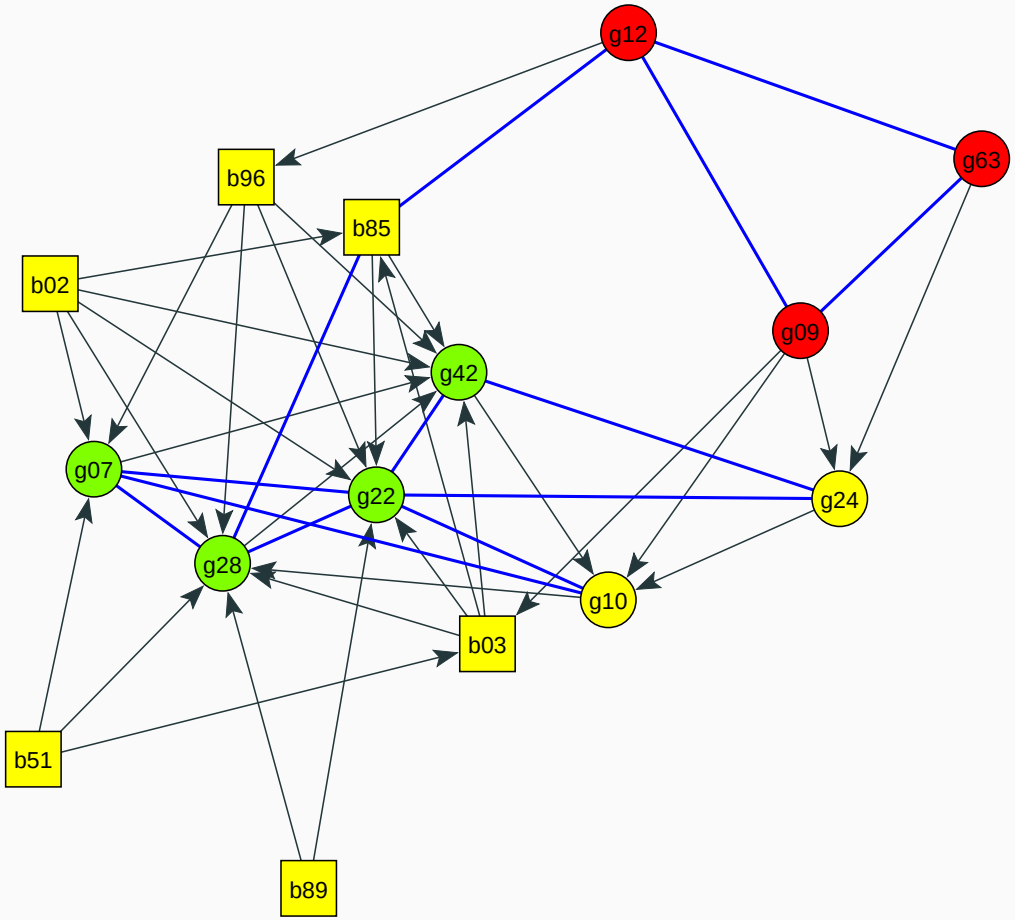
Using *Corrected Euclidean-like dissimilarity* and *Ward clustering method* we obtain the following dendrogram.

From it we can determine the number of clusters: 'Natural' clusterings correspond to clear 'jumps' in the dendrogram.

If we select 3 clusters we get the partition $\mathbf{C}$.

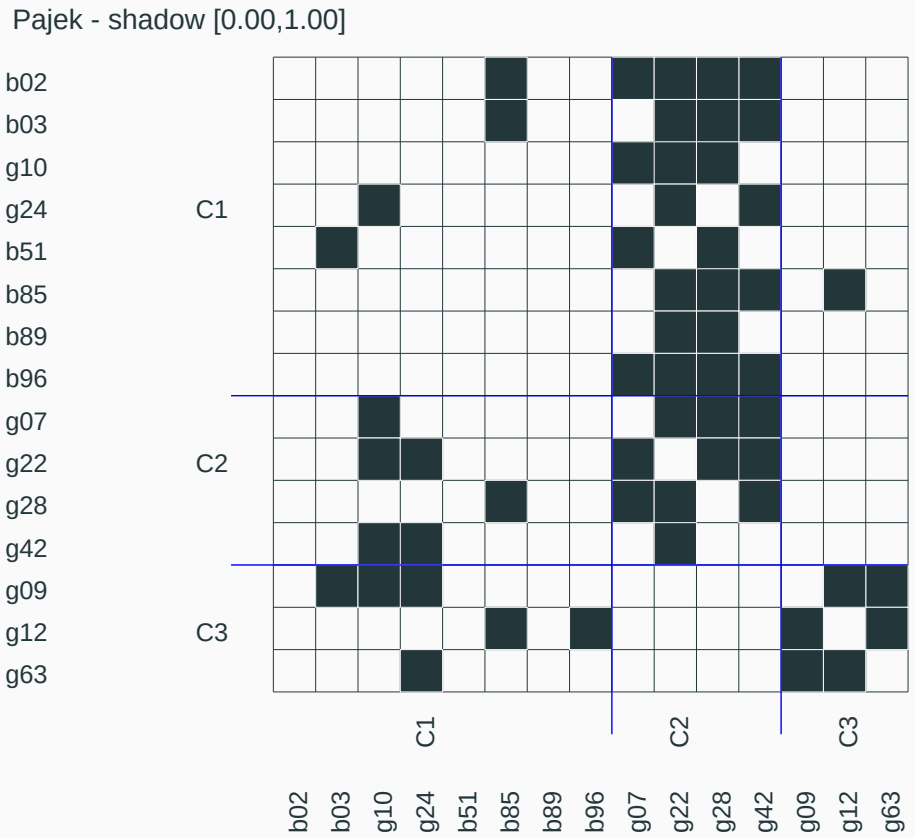
$$\begin{aligned}\mathbf{C} \quad = \quad & \{\{b51, b89, b02, b96, b03, b85, g10, g24\}, \\ & \{g09, g63, g12\}, \{g07, g28, g22, g42\}\}\end{aligned}$$

Partition into three clusters (Indirect approach)



On the picture, vertices in the same cluster are of the same color.

Matrix



The partition can be used also to reorder rows and columns of the matrix representing the network. Clusters are divided using blue vertical and horizontal lines.

Direct Approach

The second possibility for solving the blockmodeling problem is to construct an appropriate criterion function directly and then use a local optimization algorithm to obtain a 'good' clustering solution.

Criterion function $P(\mathbf{C})$ has to be *sensitive* to considered equivalence:

$$P(\mathbf{C}) = 0 \Leftrightarrow \mathbf{C} \text{ defines considered equivalence.}$$

Criterion Function

One of the possible ways of constructing a criterion function that directly reflects the considered equivalence is to measure the fit of a clustering to an ideal one with perfect relations within each cluster and between clusters according to the considered equivalence.

Given a clustering $\mathbf{C} = \{C_1, C_2, \ldots, C_k\}$, let $\mathcal{B}(C_u, C_v)$ denote the set of all ideal blocks corresponding to block $R(C_u, C_v)$. Then the global error of clustering $\mathbf{C}$ can be expressed as

$$P(\mathbf{C}) = \sum_{C_u, C_v \in \mathbf{C}} \min_{B \in \mathcal{B}(C_u, C_v)} d(R(C_u, C_v), B)$$

where the term $d(R(C_u, C_v), B)$ measures the difference (error) between the block $R(C_u, C_v)$ and the ideal block B. d is constructed on the basis of characterizations of types of blocks. The function d has to be compatible with the selected type of equivalence.

Example

Empirical blocks

	a	b	c	d	e	f	g
a	0	1	1	0	1	0	0
b	1	0	1	0	0	0	0
c	1	1	0	0	0	0	0
d	1	1	1	0	0	0	0
e	1	1	1	0	0	0	0
f	1	1	1	0	1	0	1
g	0	1	1	0	0	0	0

Ideal blocks

	a	b	c	d	e	f	g
a	0	1	1	0	0	0	0
b	1	0	1	0	0	0	0
c	1	1	0	0	0	0	0
d	1	1	1	0	0	0	0
e	1	1	1	0	0	0	0
f	1	1	1	0	0	0	0
g	1	1	1	0	0	0	0

. Number of inconsistencies for each block

	A	B
A	0	1
B	1	2

$P = 4$.

Local Optimization

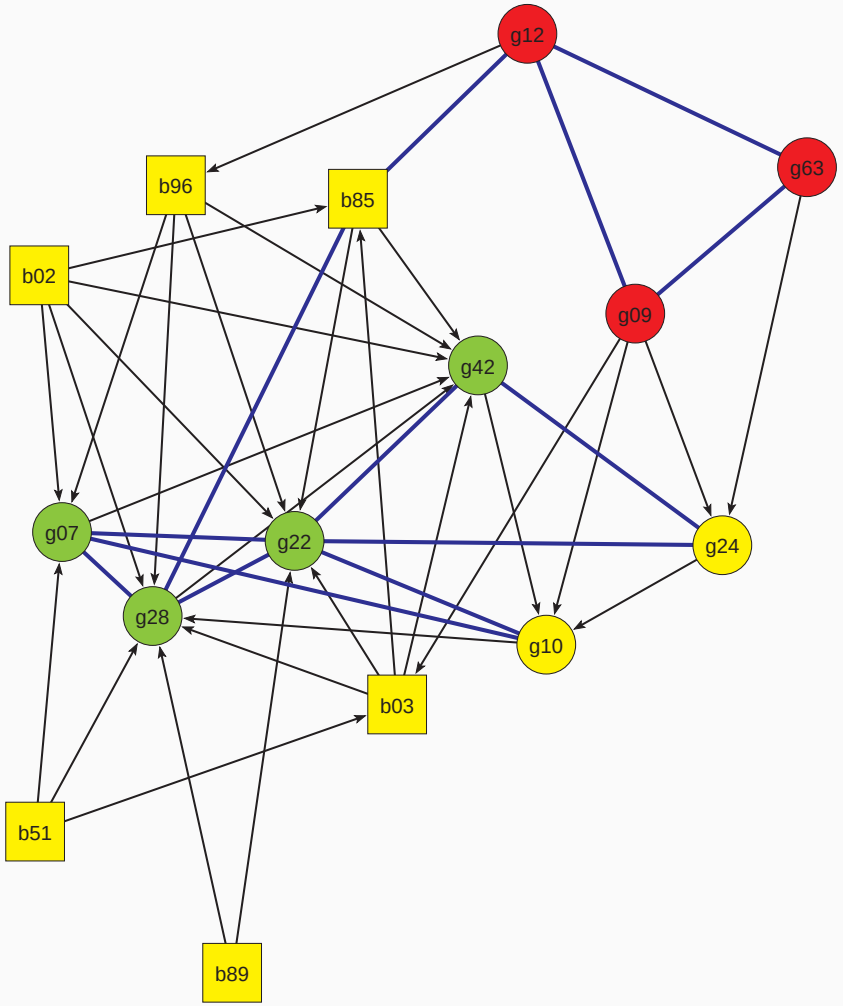
For solving the blockmodeling problem the *relocation algorithm* can be used:

Determine the initial clustering $\mathcal{C}$;
repeat:
    if in the neighborhood of the current clustering $\mathcal{C}$
    there exists a clustering $\mathcal{C}'$ such that $P(\mathcal{C}') < P(\mathcal{C})$
    then move to clustering $\mathcal{C}'$.

The neighborhood in this local optimization procedure is determined by the following two transformations:

- *moving* a unit x_k from cluster C_p to cluster C_q (*transition*);
- *interchanging* units x_u and x_v from different clusters C_p and C_q (*transposition*).

Partition into three clusters: Direct solution



This is the same partition and has the number of inconsistencies.

Pre-specified Blockmodeling

In the previous slides the *inductive* approaches for establishing blockmodels for a set of social relations defined over a set of units were discussed. Some form of equivalence is specified and clusterings are sought that are consistent with a specified equivalence.

Another view of blockmodeling is *deductive* in the sense of starting with a blockmodel that is specified in terms of substance prior to an analysis.

In this pre-specified case given

- a network,
- an equivalence (e.g. structural equivalence or a set of types of ideal blocks), and
- a blockmodel (reduced model)

a clustering can be determined which minimizes the criterion function.

Types of Pre-specified Blockmodels

The basic types of pre-specified blockmodels are:

*	*	*
*	0	0
*	0	0

core - periphery

*	0	0
*	*	0
?	*	*

hierarchy

*	0	0
0	*	0
0	0	*

cohesive

Global network structure of co-authorship networks

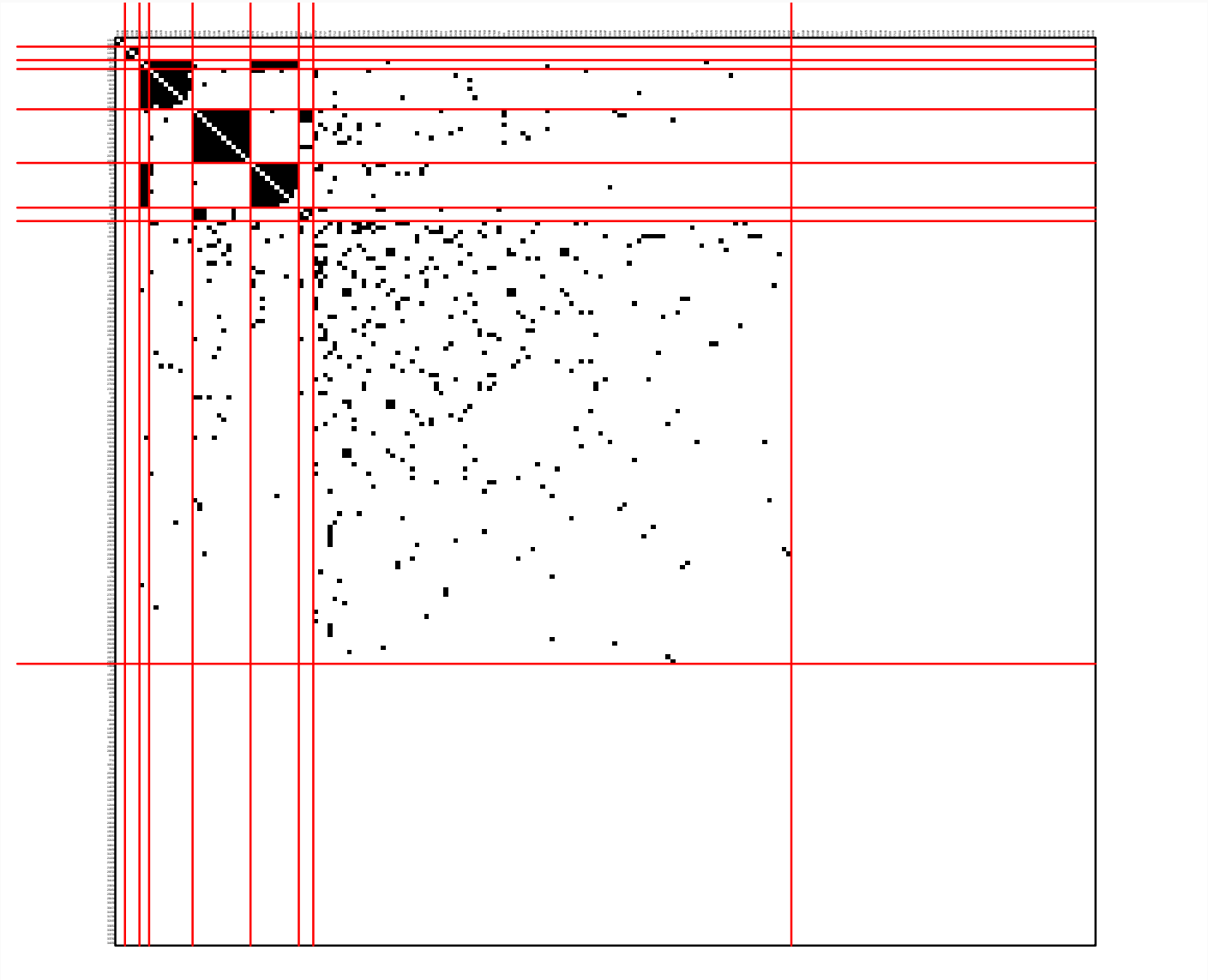
The global network structure of co-authorship networks can be obtained by blockmodeling.

The following structure is assumed in the case co-authorship networks (Ferligoj and Kronegger 2009; Kronegger et al. 2011):

- **multi-cores** consist of the groups of researchers from a scientific discipline who published scientific bibliographic units together;
- **semi-periphery** consists of researchers who co-author in a less systematic way;
- **periphery** consists of authors who published just as a single author or with authors outside the boundary of the defined disciplinary network.

The assumed global network structure appears in most of scientific disciplines.

Example



The aim

- To characterize the blockmodel structures of the Slovenian scientific co-authorship networks in most of the scientific disciplines.
- To study the stability of the blockmodels of co-authorship networks in time.
- To reveal the differences in the global network structrues between different scientific disciplines (natural and technical sciences vs. social sciences and humanitites) and time periods (1990–2000 vs. 2001–2010).

Data

Current Research Information System (SICRIS) includes information on all researchers registered with the Slovenian Research Agency (SRA) and co-operative On-Line Bibliographic System & Services (COBISS) which is an officially maintained database of all publications available in Slovenian libraries.

From this system, we collected complete scientific bibliographies of all Slovenian researchers.

A tie was defined if two researchers appeared together as authors in at least one scientific bibliographic unit (binary networks).

Scientific bibliographic unit

Scientific bibliographic unit is defined by SRA classification scheme.

Type of scientific bibliographic unit	1991—2000 (independence)	2001—2010 (joining the EU)
Original sci. article	26531	47905
Review article	4895	5738
Short sci. article	969	2530
Published sci. conference contribution (inv. lecture)	3427	5279
Published sci. conference contribution	28670	41138
Independent sci. component part in monograph	6417	14759
Sci. monograph	1725	2912
Sci. or documentary films, sound or video recording	44	133
Complete sci. database or corpus	73	182
Patent	381	710
Total	73132	121286

Scientific fields and disciplines

ID	Scientific field	No. of disciplines
1	Natural sciences and mathematics	9
2	Engineering sciences and technologies	19
3	Medical sciences	9
4	Biotechnical sciences	6
5	Social sciences	11
6	Humanities	12
7	Interdisciplinary studies	2

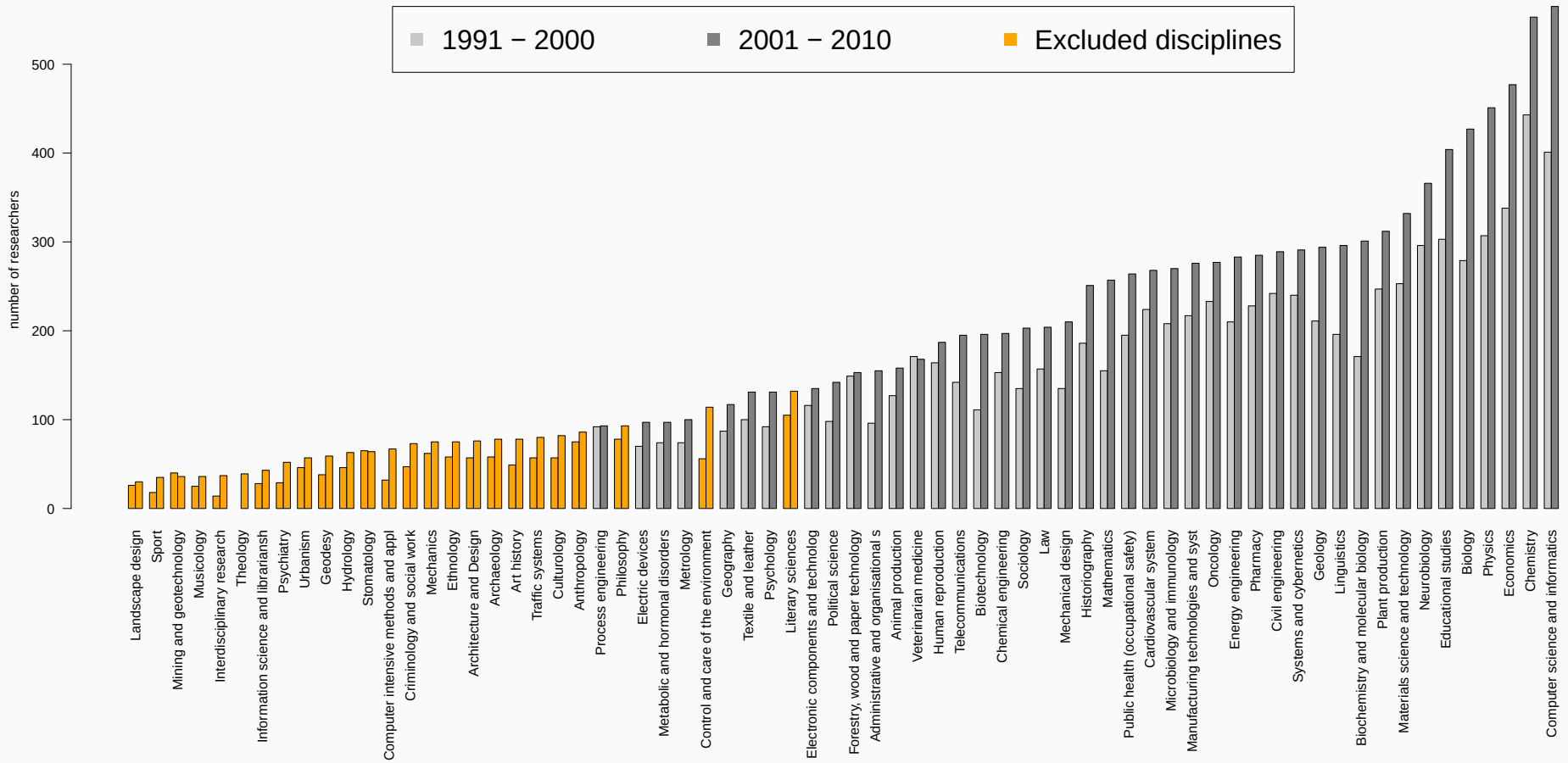
Time period

- **1991–2000**: the independence of Slovenia, meaning that the country had started adopting and implementing its own science policies.
- **2001–2010**: joining the European Union and adopting European Union standards. By the end of this period, Slovenia had already partly integrated its national science system into the European one.

The level of analysis

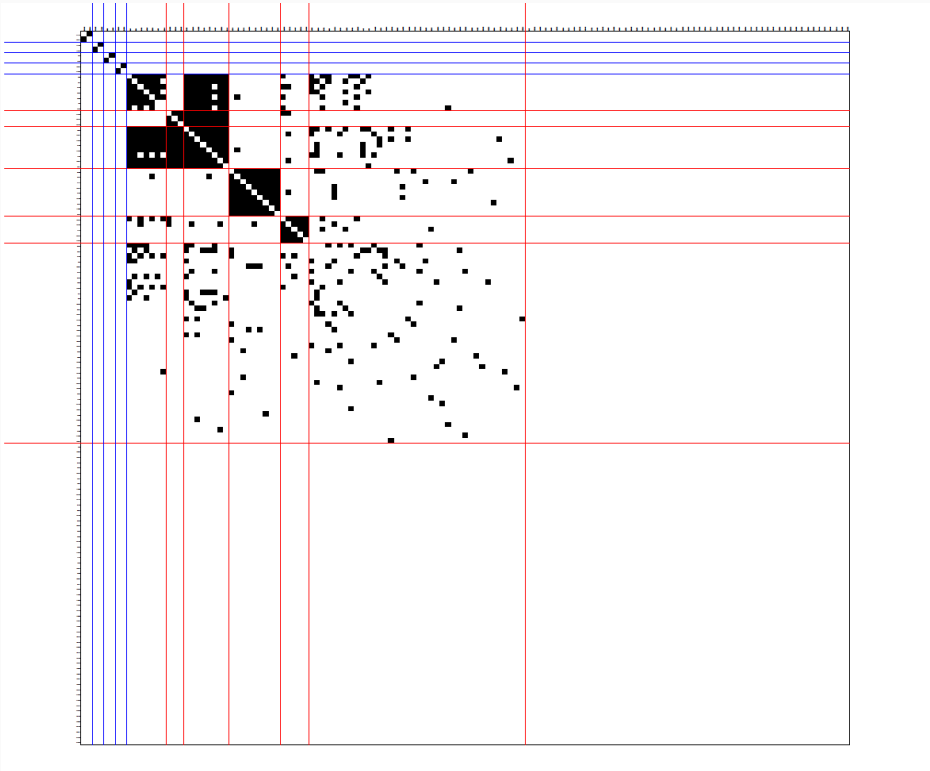
The analysis is done on the level of scientific disciplines which are defined by ARRS. The analysis includes 43 out of 72 scientific disciplines.

The number of researchers (with at least one bibliographic unit) by scientific disciplines

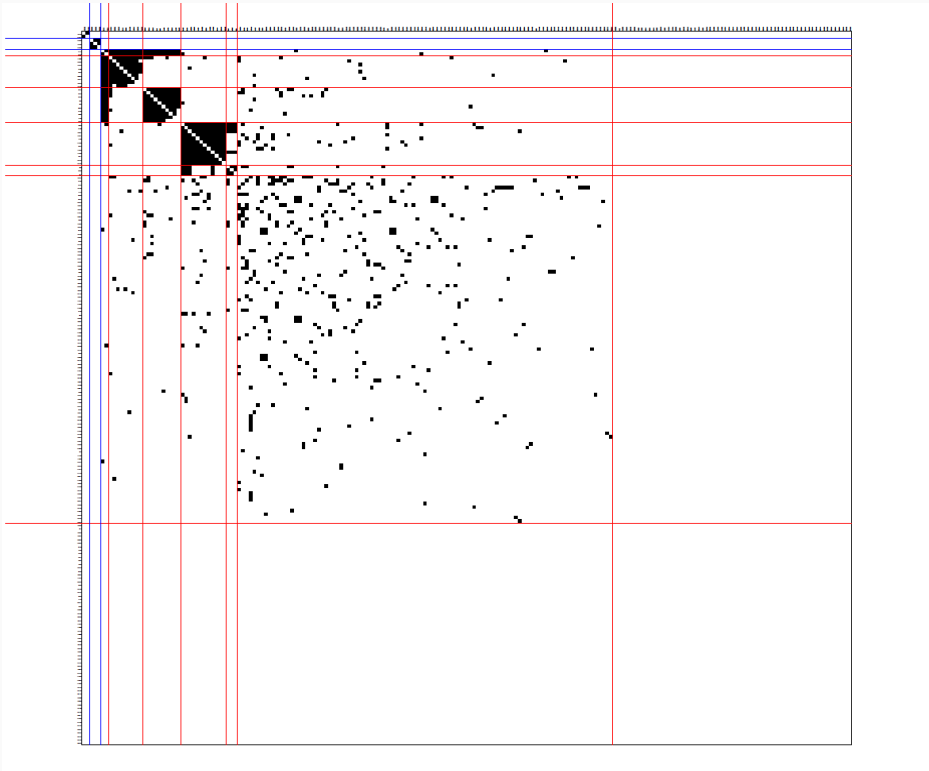


An example of a blockmodel – Sociology

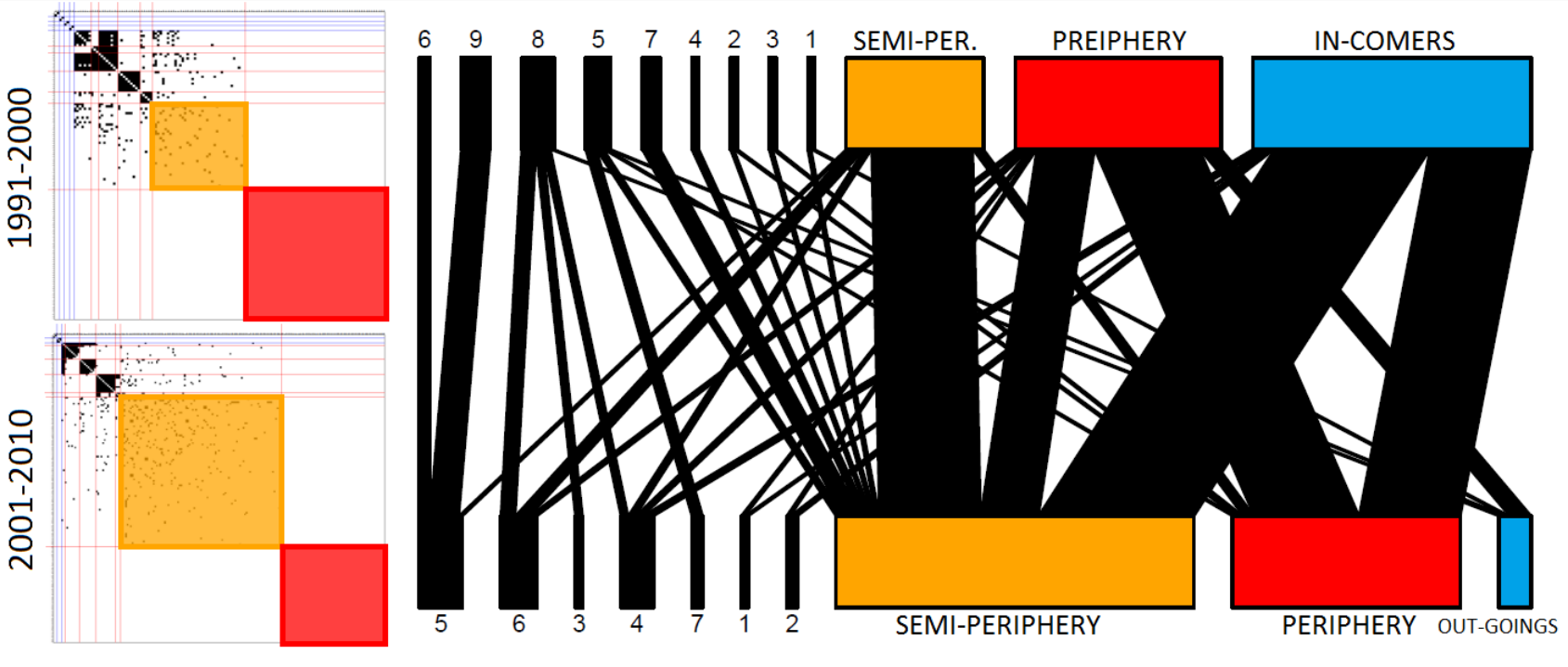
1991 - 2000



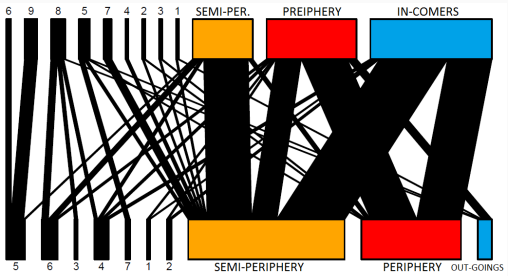
2001 - 2010

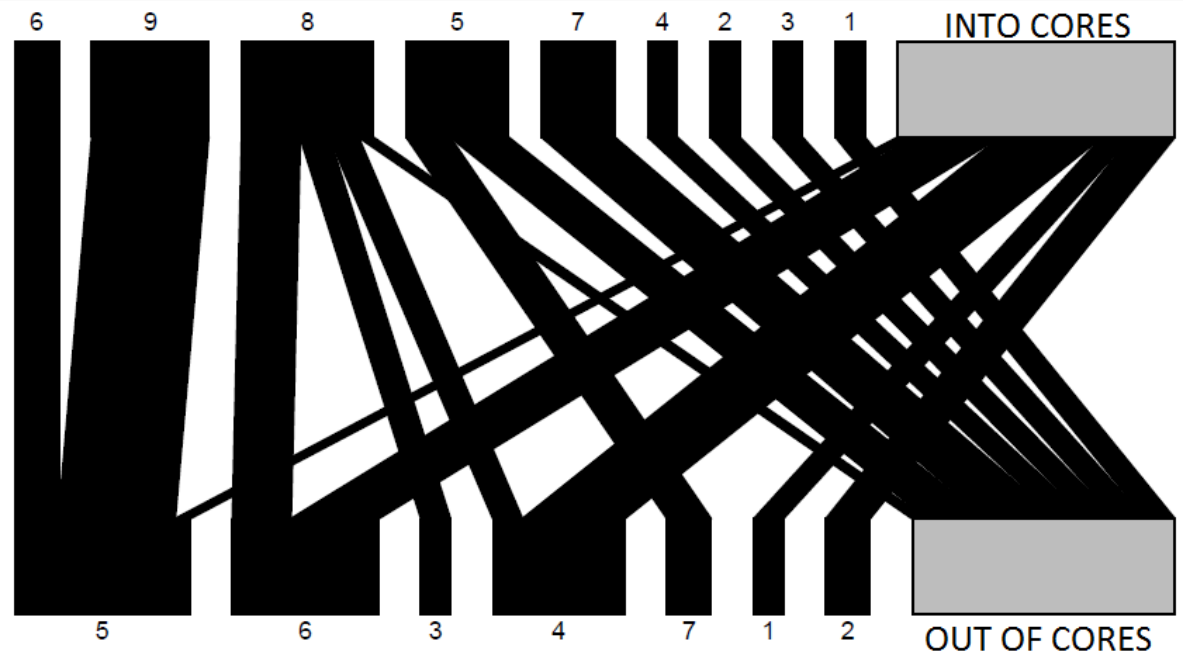


The visualisation of two blockmodels – Sociology



The visualisation of cores – Sociology





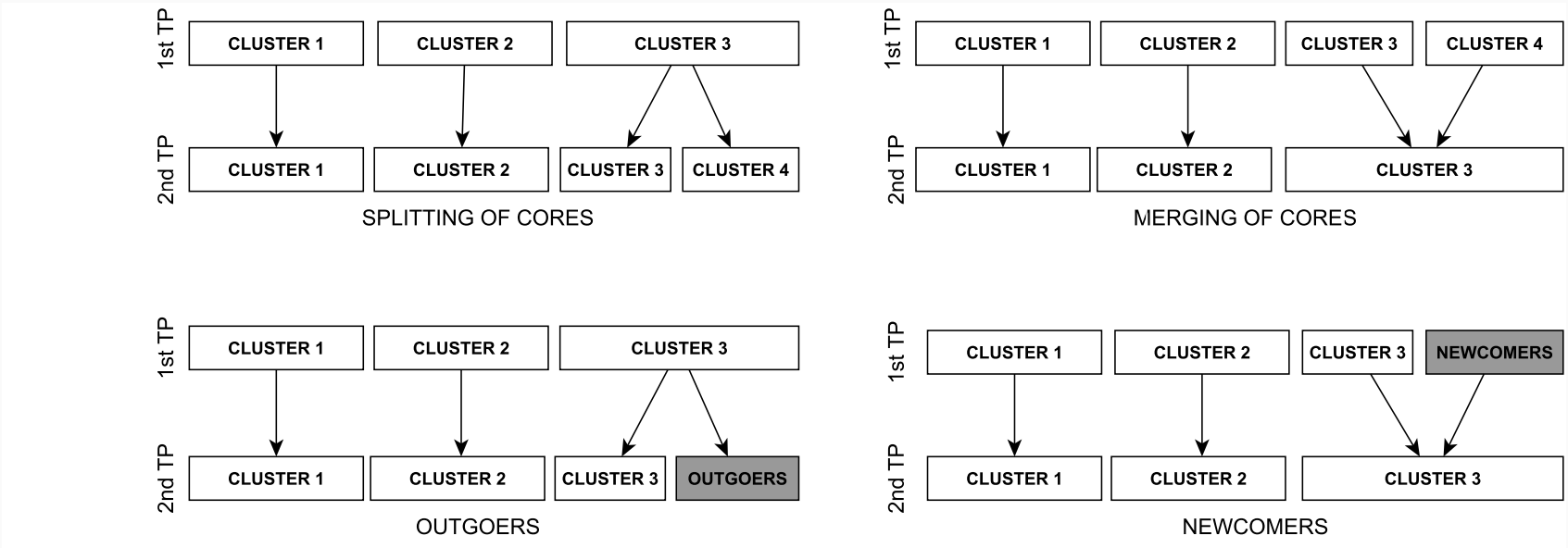
Comparing partitions

By blockmodeling the partition of units and links between clusters are obtained. The goal is to study the stability of obtained partitions in two time periods.

- There is a need to compare two partitions obtained on two different sets of units with a non-empty intcersection.
- Several modified indices are presented. They allow to compare two partitions, obtained on the different sets of units. The units that are not in the intersection of two sets of units are considered as a factor influencing the value of the measure. Furthermore, the splitting and merging of clusters have a different effect on the value of the measure.
- For all presented indices, the correction for chance is also considered. The expected value in the case of two random and independent partitions is 0.

Comparing partitions

- A partition assigns a unit into a certain cluster.
- When the units are observed in, e.g., two time points:
  - clusters can merge and/or split
  - new units can join (newcomers) while some old units can leave (outgoers)



Standard indices to compare partitions

Pair types

The following table can be calculated:

V / U	Pairs in same cluster	Pairs in different cluster
Pairs in same cluster	a	b
Pairs in different cluster	c	d

Rand Index (RI)

The proportion of all possible pairs of units that are in the same or in different clusters in both partitions U and V in comparison to all possible pairs.

$$RI = \frac{a+d}{a+b+c+d} \in [0,1]$$

Standard indices to compare partitions

Wallace Index 1 (W1)

The proportion of all possible pairs that are classified into the same cluster in both partitions, compared to all possible pairs that are classified into the same cluster in partiiton U.

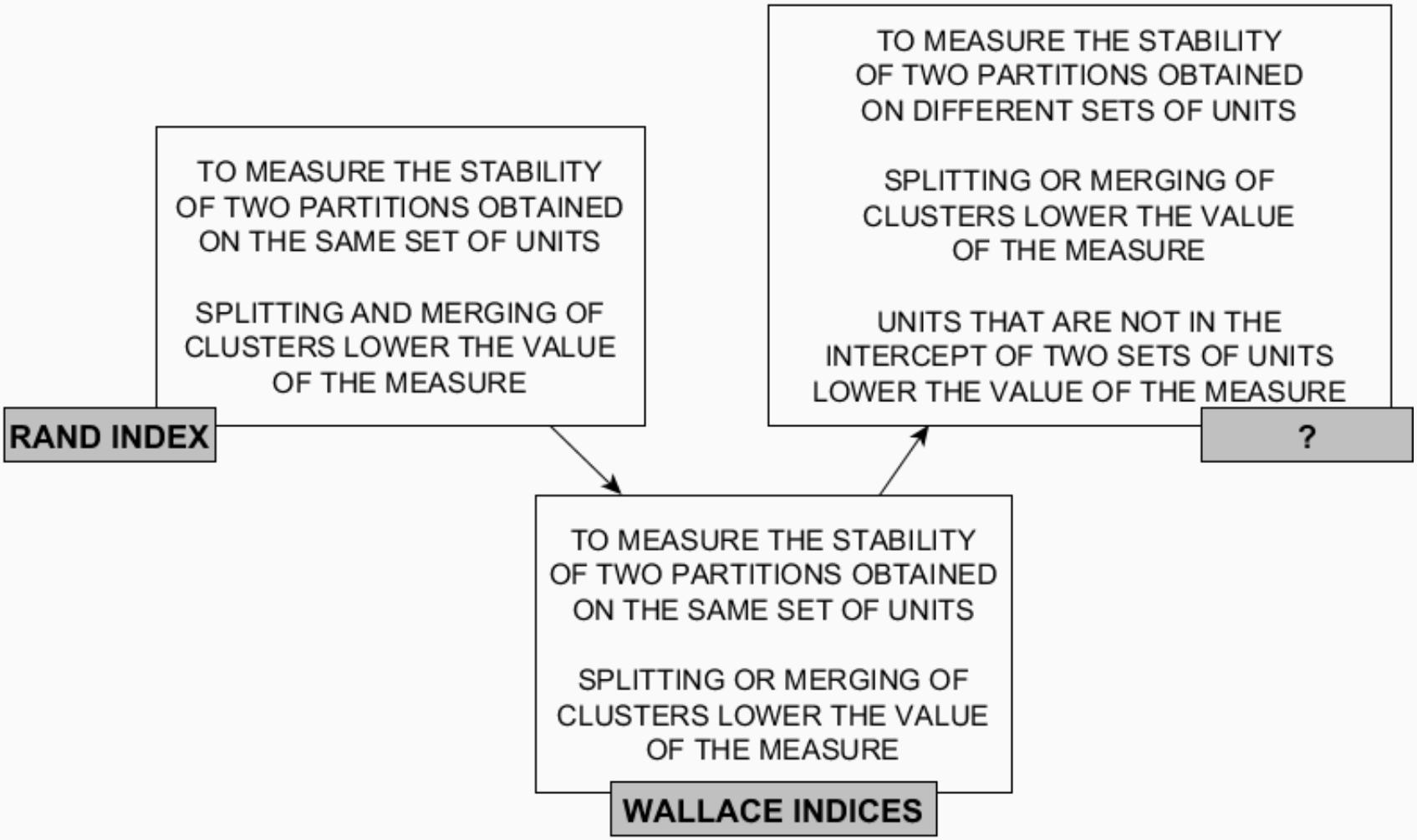
$$W1 = \frac{a}{a+b} \in [0,1]$$

Wallace Index 2 (W2)

The proportion of all possible pairs that are classified into the same cluster in both partitions, compared to all possible pairs that are classified into the same cluster in partiiton V.

$$W2 = \frac{a}{a+c} \in [0,1]$$

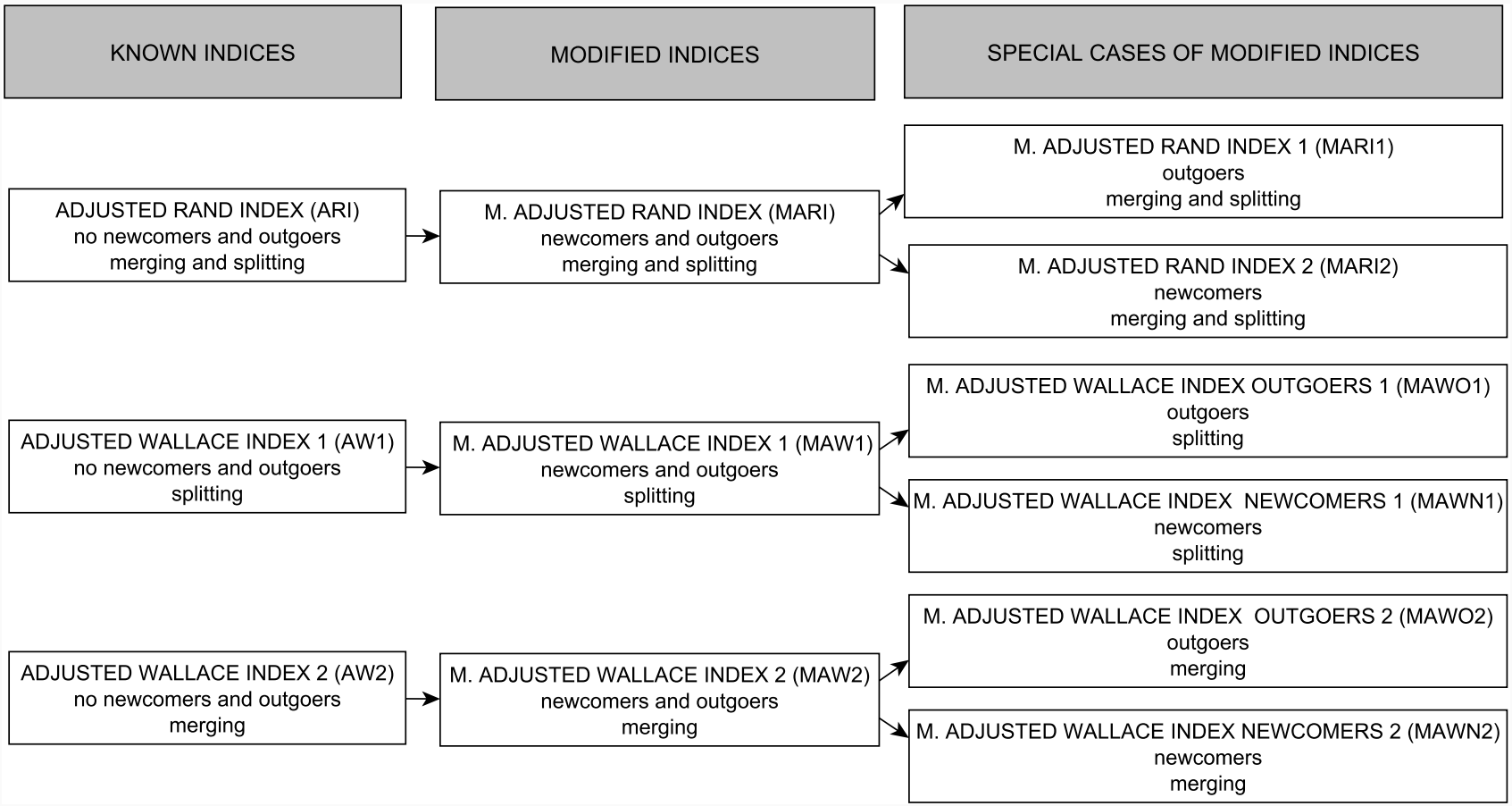
Indices to compare partitions



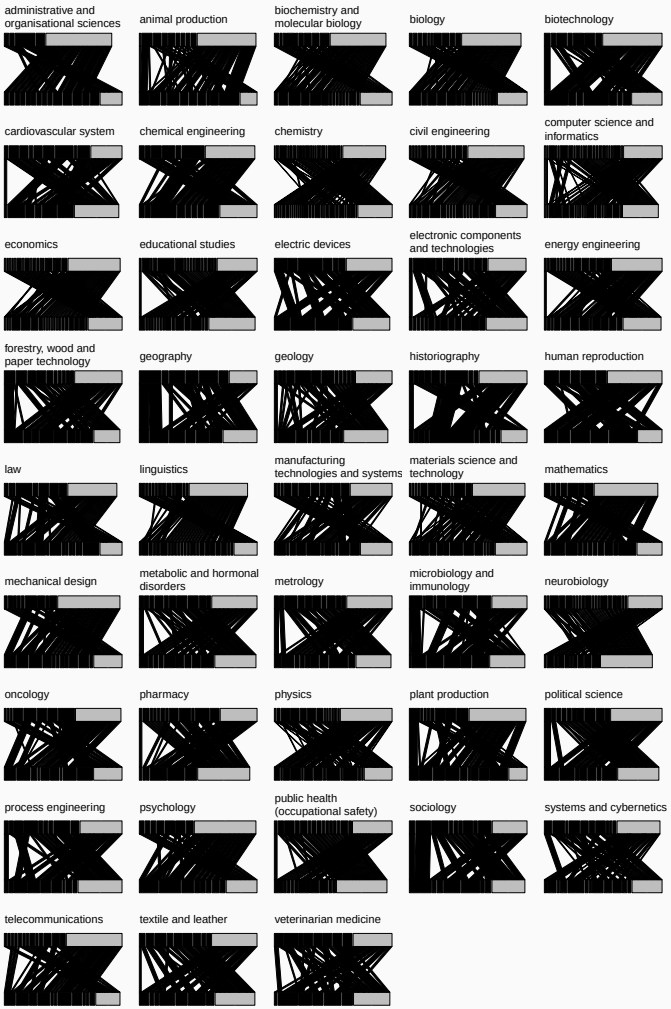
Indices to compare partitions from different sets of units

Several modified indices were proposed by Cugmas and Ferligoj (2018)

- partitions are obtained on different sets of the units (intersection of two sets is not empty)
- merging and splitting of clusters (can) has a different effect to the value of the indices

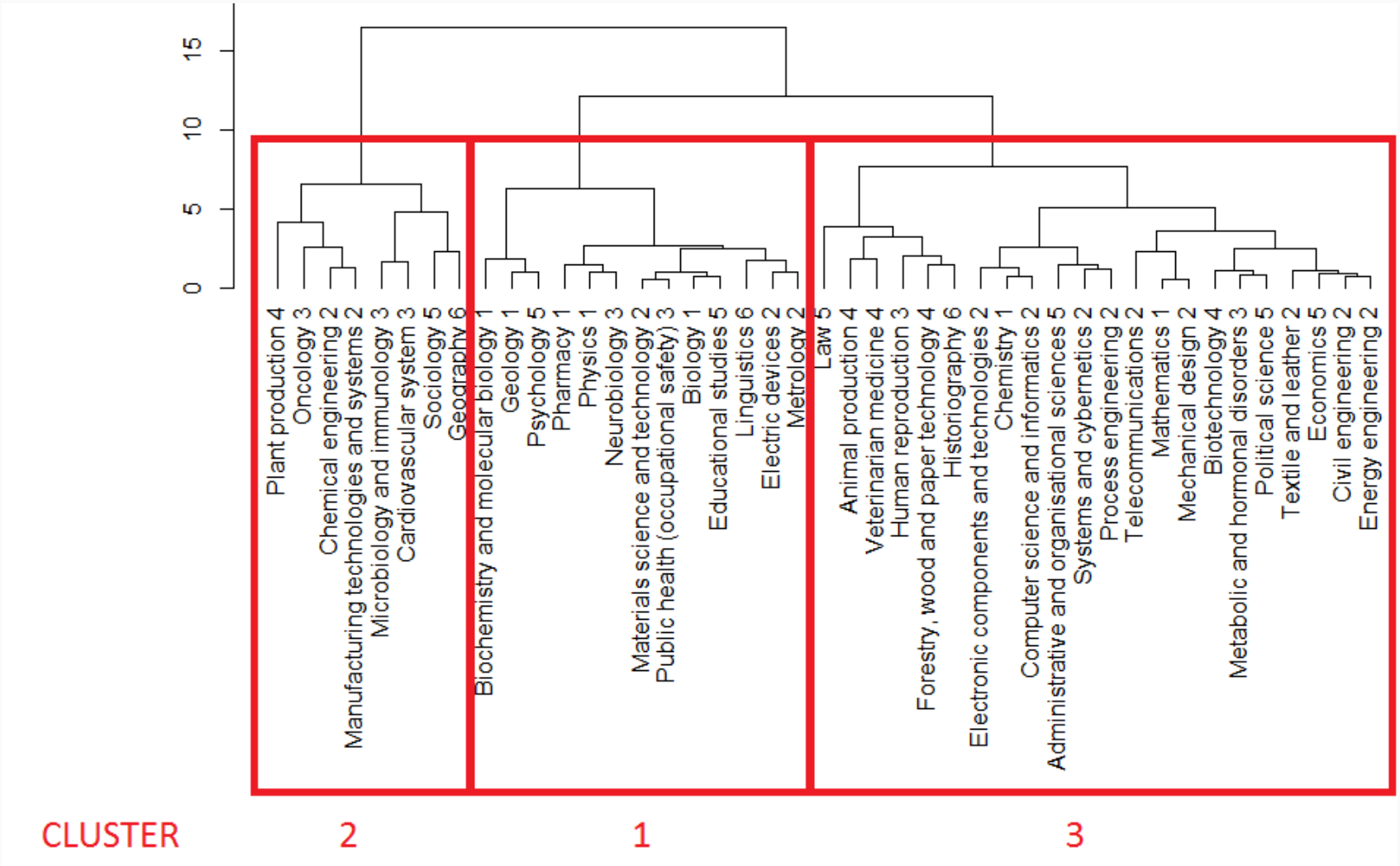


The stability of partitions for scientific disciplines



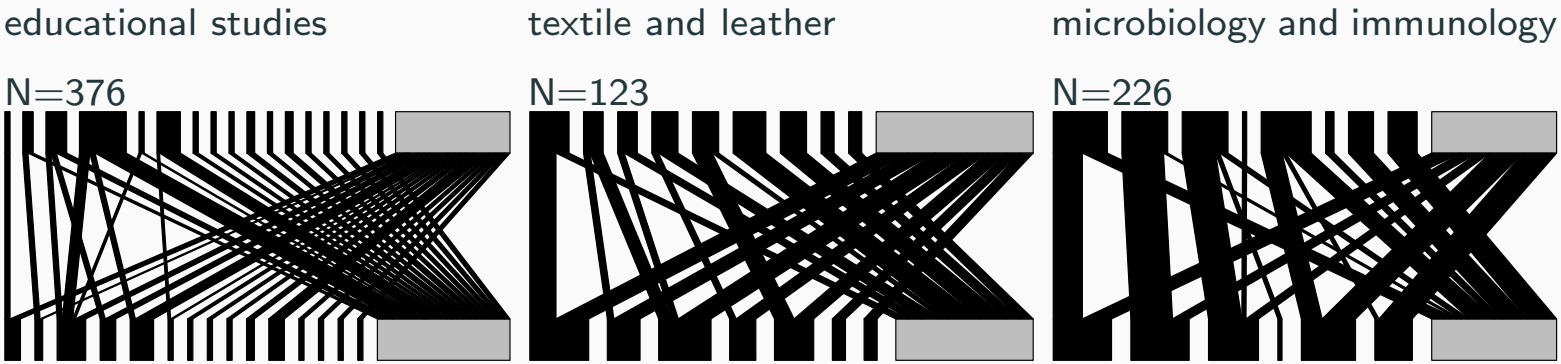
	ARI	AW'	AW''	MARI1	MAWIS1	MAWIM1	MARI2	MAWIS2	MAWIM2
Mathema	1.00	1.00	1.00	0.05	0.27	0.08	0.01	0.01	0.09
Civil e	0.86	0.76	1.00	0.01	0.14	0.02	0.01	0.01	0.06
Chemica	1.00	1.00	1.00	0.08	0.33	0.14	0.09	0.08	0.30
Energy	0.81	0.75	0.88	0.01	0.15	0.02	0.05	0.01	0.09
Materia	0.33	0.26	0.45	0.01	0.06	0.02	0.02	0.01	0.04
Systems	0.70	0.57	0.92	0.04	0.19	0.09	0.12	0.04	0.15
Compute	0.63	0.57	0.71	0.01	0.19	0.03	0.04	0.02	0.14
Telecom	0.69	0.89	0.56	0.04	0.35	0.08	0.06	0.02	0.09
Electro	0.62	0.47	0.91	0.01	0.13	0.03	0.11	0.03	0.19
Manufac	0.95	0.90	1.00	0.06	0.43	0.12	0.11	0.07	0.25
Physics	0.40	0.67	0.28	0.03	0.15	0.04	0.03	0.02	0.08
Mechani	1.00	1.00	1.00	0.04	0.23	0.06	0.04	0.01	0.09
Electri	0.50	0.46	0.56	0.03	0.11	0.06	0.04	0.03	0.07
Process	0.84	0.72	1.00	0.02	0.14	0.04	0.1	0.04	0.17
Textile	0.80	0.80	0.80	0.03	0.18	0.04	0.00	0.03	0.10
Metrolo	0.42	0.44	0.40	0.01	0.16	0.04	0.00	0.02	0.09
Biology	0.38	0.48	0.31	0.01	0.09	0.02	-0.02	0.00	0.04
Microbi	0.88	0.86	0.91	0.09	0.32	0.16	0.25	0.16	0.26
Neurobi	0.43	0.68	0.32	0.00	0.08	0.01	0.03	0.01	0.09
Oncolog	0.89	0.85	0.93	0.11	0.42	0.23	0.12	0.11	0.35
Human r	0.40	0.93	0.26	0.08	0.28	0.14	0.12	0.06	0.14
Cardiov	1.00	1.00	1.00	0.06	0.30	0.10	0.32	0.18	0.30
Metabol	0.73	1.00	0.57	0.02	0.07	0.01	-0.02	0.01	0.06
Public	0.37	0.32	0.42	0.00	0.03	0.00	0.00	0.00	0.02
Chemist	0.60	0.46	0.89	0.01	0.17	0.02	0.04	0.01	0.17
Forestr	0.64	0.69	0.60	0.06	0.34	0.14	0.10	0.05	0.15
Animal	0.49	0.51	0.47	0.06	0.19	0.13	0.06	0.01	0.06
Plant p	0.90	0.84	0.97	0.15	0.45	0.34	0.11	0.05	0.19
Veterin	0.52	0.68	0.43	0.04	0.15	0.05	0.13	0.05	0.09
Biotech	0.73	1.00	0.57	0.04	0.14	0.05	-0.01	0.01	0.04
Educati	0.32	0.34	0.31	0.00	0.04	0.01	-0.02	0.01	0.07
Economi	0.71	0.64	0.80	0.01	0.14	0.01	0.01	0.01	0.07
Sociolo	0.52	0.55	0.50	0.06	0.36	0.16	0.25	0.14	0.23
Biochem	-0.16	-0.11	-0.27	-0.02	0.00	0.00	-0.03	0.00	0.00
Adminis	0.80	0.67	1.00	0.06	0.11	0.09	0.05	0.01	0.19
Law 5	0.58	0.80	0.45	0.09	0.29	0.17	-0.14	0.06	0.22
Politic	0.86	1.00	0.75	0.01	0.13	0.02	-0.03	0.01	0.05
Psychol	0.04	0.07	0.03	0.00	0.02	0.00	-0.09	0.00	0.01
Geology	0.09	0.11	0.07	0.00	0.02	0.00	0.00	0.01	0.03
Histori	0.57	0.68	0.49	0.08	0.39	0.20	0.05	0.07	0.09
Linguis	0.43	0.29	0.82	0.00	0.16	0.03	0.03	0.00	0.08
Geograp	0.36	0.29	0.49	0.03	0.15	0.12	0.22	0.12	0.21
Pharmac	0.50	0.69	0.39	0.01	0.03	0.01	0.09	0.01	0.03

Cluster analysis dendrogram
(Euclidean distances, Ward clustering method)



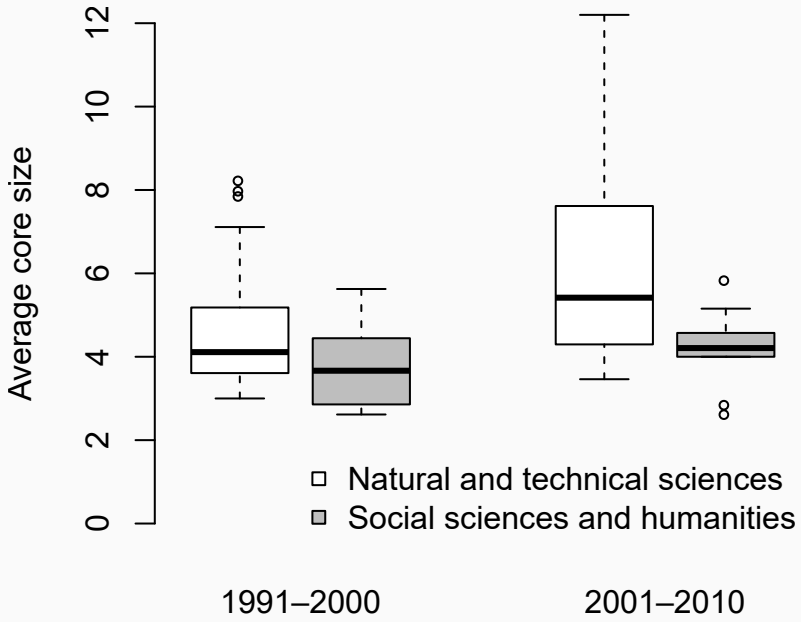
Representatives of three obtained clusters

A scientific discipline from each cluster (the closest one to the centroid) is chosen to represent the cluster.



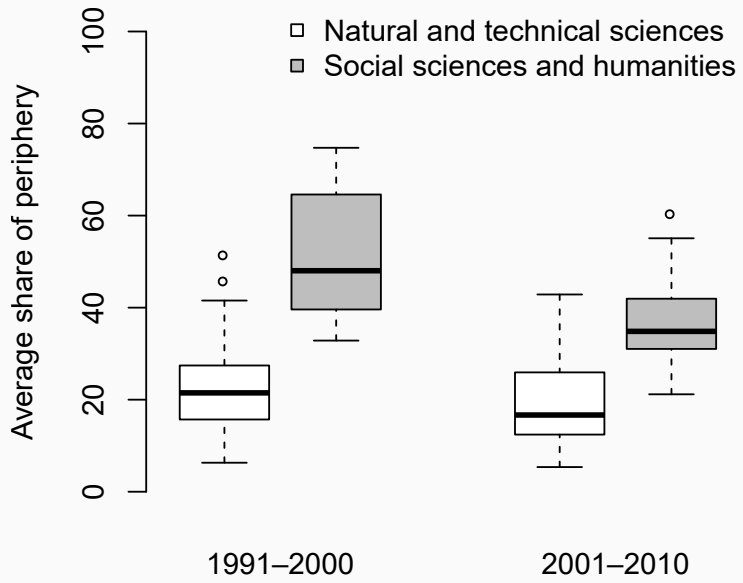
- **unstable (educational studies)**: many into-cores and out-of-cores researchers, most pairs of researchers classified in the same core at the first time point are not classified in the same core in the second period
- **average (textile and leather)**: the share of out-of-cores and into-cores researchers is lower, some relatively large cores remain relatively stable in the second period
- **stable (microbiology and immunology)**: most of cores remain stable in the second time period

The size of cores



- The average core size is higher in natural and technical sciences.
- The average core size is increasing in the natural and technical sciences.

The size of peripheries



- The average size of the periphery is higher in social sciences and humanities.
- The average size of the periphery is decreasing in social sciences and humanities.

Conclusions on the stability of the co-authorship networks

- Scientific disciplines can be classified into three clusters regading different operationalizations of the stability of cores.
- The most stable cluster of scientific disciplines is characterized by the lower mean percentage of into-cores and out-of-cores researchers and higher mean core size.

An extended version of this presentation will be published as Chapter 13 in the book:

Doreian, P., Batagelj, V. and Ferligoj, A. (Eds): **Advances in Network Clustering and Blockmodeling**. Wiley, 2019.