Clustering of attribute and/or relational data

Anuška Ferligoj

Faculty of Social Sciences, University of Ljubljana, Slovenia

NRU HSE Moscow, Russia

17th Applied Statistics 2021, 20 – 22 September, 2021

My collaborators on this topic

Vladimir Batagelj

Patrick Doreian

Luka Kronegger

Marjan Cugmas

Outline

Introduction

- A large class of *clustering problems* can be formulated as an *optimizational problem* in which the best clustering is searched among all *feasible clusterings* according to a selected *criterion function*.
- This approach can be applied to a variety of interesting clustering problems. It is possible to adapt it to a concrete clustering problem by an appropriate specification of the criterion function and by the definition of the set of feasible clusterings.
- Both, the *blockmodeling problem* (clustering of the relational data) and *clustering with relational constraint problem* (clustering of the attribute and relational data) can be successfully treated by this approach.

Cluster Analysis

Grouping units into clusters so that those within a cluster are as similar to each other as possible according to the selected attributes (variables), while units in different clusters as dissimilar as possible, is a very old problem.

Although the clustering problem is intuitively simple and understandable, providing solution(s) remains an exciting activity.

Organization of the field

The field of cluster analysis

- has its society, the *International Federation of Classification Societies* IFCS), formed in 1985 from several national classification societies;
- organizes every second year its conference;
- publishes two journals: the *Journal of Classification* (from 1984) and the journal *Advances in Data Analysis and Classification* (from 2007).

Clustering problem

Cluster analysis (known also as classification and taxonomy) deals mainly with the following general problem: given a set of *units*, $\mathcal{U}$, determine subsets, called *clusters*, C, which are homogeneous and/or well separated according to the measured variables. The set of clusters forms a *clustering*.

This problem can be formulated as an *optimization problem*:

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

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

where $\mathbf{C}$ *is a clustering of a given set of units,* $\mathcal{U}$*,* Φ *is the set of all feasible clusterings and* $P : \Phi \to R$ *is a criterion function.*

Clustering

There are several types of clusterings, e.g., partition, hierarchy, pyramid, fuzzy clustering, clustering with overlaping clusters. The most frequently used clusterings are partitions and hierarchies.

A clustering $\mathbf{C} = \{C_1, C_2, \dots C_k\}$ is a *partition* of the set of units $\mathcal{U}$ if

$$\bigcup_i C_i = \mathcal{U}$$

$$i \neq j \Rightarrow C_i \cap C_j = \emptyset$$

A clustering $\mathbf{H} = \{C_1, C_2, \dots C_k\}$ is a *hierarchy* if for each pair of clusters C_i and C_j from $\mathbf{H}$ it holds

$$C_i \cap C_j \in \{C_i, C_j, \emptyset\}$$

Clustering criterion function

Clustering criterion functions can be constructed *indirectly*, e.g., as a function of a suitable (dis)similarity measure between pairs of units (e.g., euclidean distance) or *directly*.

For partitions into k clusters, the *Ward criterion function*

$$P(\mathbf{C}) = \sum_{C \in \mathbf{C}} \sum_{x \in C} d(x, t_C)$$

is usually used, where d is the squared euclidean distance and t_C is the center of the cluster C. It is defined as

$$t_C = (\overline{u}_{1C}, \overline{u}_{2C}, \ldots, \overline{u}_{mC})$$

where $\overline{u}_{iC}$ is the average of the (interval or ratio) variable U_i, $i = 1, \ldots m$, for the units from the cluster C.

Solving the clustering problem

- As the nonempty set of feasible clusterings is finite a solution of the clustering problem always exists.
- Since the set of feasible clusterings is usually very large it is not easy to find an optimal solution.
- In general, most of the clustering problems are *NP-hard*.
- For this reason, different efficient *heuristic* algorithms are used. Among these, the *agglomerative* (hierarchical) and the *relocation* approach are most often used.

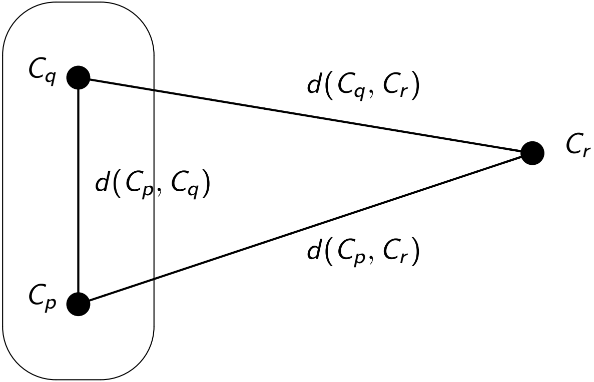
Agglomerative approach

The agglomerative clustering approach usually assumes that all relevant information on the relationships between n units from the set $\mathcal{U}$ is summarized by a symmetric pairwise *dissimilarity matrix* $D = [d_{ij}]$.

Each unit is a cluster;
repeat while there exist at least two clusters:
    determine the nearest pair of clusters C_p and C_q;
    fuse the clusters C_p and C_q to form a new cluster C_r;
    replace C_p and C_q by the cluster C_r;
    determine the dissimilarities between C_r and other clusters.

The result is a hierarchy that is usually presented by a clustering tree – a *dendrogram*.

Dissimilarities between three clusters



Some agglomerative clustering methods

- The **Minimum method**, or single linkage, (Florek et al., 1951; Sneath, 1957):

$$d(C_p \cup C_q, C_r) = \min(d(C_p, C_r), d(C_q, C_r))$$

- The **Maximum method**, or complete linkage, (McQuitty, 1960):

$$d(C_p \cup C_q, C_r) = \max(d(C_p, C_r), d(C_q, C_r))$$

- The **Ward method** (Ward, 1963):

$$d(C_p \cup C_q, C_r) = \frac{(n_p + n_q)n_r}{(n_p + n_q + n_r)} d^2(t_{pq}, t_r)$$

where t_{pq} denotes the centroid of the fused cluster $C_p \cup C_q$ and t_r the center of the cluster C_r. n_i denotes the number of units in the cluster C_i.

Relocation approach (local optimization)

This approach assumes that the user can specify the number of clusters in the partition.

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;
- *interchanging* units x_u and x_v from different clusters C_p and C_q.

Benefits from the optimizational approach

The *optimizational approch* to clustering problem offers two possibilities to adapt to a concrete clustering problem: the definition of the *criterion function* P and the specification of the *set of feasible clusterings* Φ.

- Classical clustering deals with *clustering of attribute data*.
- *Blockmodeling* is based on a clustering according to the *relational data only* and the solution can be obtained by an appropriatelly defined *criterion function*.
- *Clustering with relational constraint* is searching for a clustering according to the *attribute and the relational data*. The relational constraint is considered by an appropriately defined *set of feasible clusterings*.

Clustering of relational data - Blockmodeling

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 rests on the assumption that the role structure of the positions of individuals exists. The key tasks here are (Faust and Wasserman, 1992):

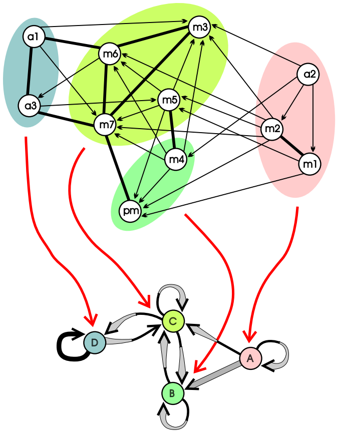
- identifying *social positions* (clusters) as collections of units who are similar in their relationships to the others, and
- modeling *social roles* as system of relationships among positions (clusters).

Blockmodeling deals with these two aspects.

Blockmodeling

The goal of *blockmodeling* is to reduce a larger network to a smaller comprehensible structure.

Blockmodeling, as an empirical procedure, is based on the idea that units in a network can be grouped 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 $\mathcal{N} = (\mathcal{U}, R)$, $R \subseteq \mathcal{U} \times \mathcal{U}$, *clusters* 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 *partition* $\mathbf{C} = \{C_1, C_2, \ldots, C_k\}$ of the set $\mathcal{U}$.

Each partition determines an equivalence relation, denoted by $\sim$ (and vice versa).

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

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

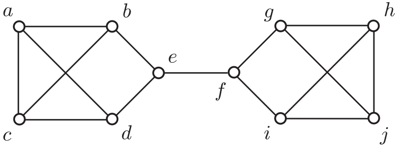
Each such block consists of units belonging to clusters C_i and C_j and all arcs leading from cluster C_i to cluster C_j.

If $i = j$, a block $R(C_i, C_i)$ is called a *diagonal* block.

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.
There are many other definitions of equivalences.

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 blocks.

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

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

Regular equivalence

Integral to all attempts to generalize structural equivalence is the idea that units are equivalent if they link in equivalent ways to other units that are also equivalent.

White and Reitz (1983): The equivalence relation $\approx$ on $\mathcal{U}$ is a *regular equivalence* on the network $\mathcal{N} = (\mathcal{U}, R)$ if and only if for all $\mathrm{x}, \mathrm{y}, \mathrm{z} \in \mathcal{U}$, $\mathrm{x} \approx \mathrm{y}$ implies both

R1. $\mathrm{x} R \mathrm{z} \Rightarrow \exists \mathrm{w} \in \mathcal{U} : (\mathrm{y} R \mathrm{w} \wedge \mathrm{w} \approx \mathrm{z})$

R2. $\mathrm{z} R \mathrm{x} \Rightarrow \exists \mathrm{w} \in \mathcal{U} : (\mathrm{w} R \mathrm{y} \wedge \mathrm{w} \approx \mathrm{z})$

. . . Regular equivalence

The blocks for regular equivalence are null or 1-covered (Batagelj, Doreian, Ferligoj, 1992):

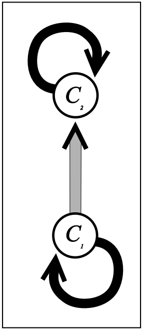
$$\begin{matrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{matrix} \qquad \begin{matrix} 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \end{matrix}$$

1-covered blocks are usually called regular blocks.

Generalized blockmodeling

$$\begin{array}{cccc|cccc}
1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 \\
1 & 1 & 1 & 1 & 0 & 1 & 0 & 1 \\
1 & 1 & 1 & 1 & 0 & 0 & 1 & 0 \\
1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\
\hline
0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\
0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 \\
0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 \\
0 & 0 & 0 & 0 & 1 & 1 & 1 & 0
\end{array}$$

	C_1	C_2
C_1	complete	regular
C_2	null	complete



Generalized equivalence / Block types

$$\begin{array}{c|ccccc} & & & Y & & \\ \hline X & 1 & 1 & 1 & 1 & 1 \\ & 1 & 1 & 1 & 1 & 1 \\ & 1 & 1 & 1 & 1 & 1 \\ & 1 & 1 & 1 & 1 & 1 \end{array}$$

complete

$$\begin{array}{c|ccccc} & & & Y & & \\ \hline X & 0 & 1 & 0 & 0 & 0 \\ & 1 & 1 & 1 & 1 & 1 \\ & 0 & 0 & 0 & 0 & 0 \\ & 0 & 0 & 0 & 1 & 0 \end{array}$$

row-dominant

$$\begin{array}{c|ccccc} & & & Y & & \\ \hline X & 0 & 0 & 1 & 0 & 0 \\ & 0 & 0 & 1 & 1 & 0 \\ & 1 & 1 & 1 & 0 & 0 \\ & 0 & 0 & 1 & 0 & 1 \end{array}$$

col-dominant

$$\begin{array}{c|ccccc} & & & Y & & \\ \hline X & 0 & 1 & 0 & 0 & 0 \\ & 1 & 0 & 1 & 1 & 0 \\ & 0 & 0 & 1 & 0 & 1 \\ & 1 & 1 & 0 & 0 & 0 \end{array}$$

regular

$$\begin{array}{c|ccccc} & & & Y & & \\ \hline X & 0 & 1 & 0 & 0 & 0 \\ & 0 & 1 & 1 & 0 & 0 \\ & 1 & 0 & 1 & 0 & 0 \\ & 0 & 1 & 0 & 0 & 1 \end{array}$$

row-regular

$$\begin{array}{c|ccccc} & & & Y & & \\ \hline X & 0 & 1 & 0 & 1 & 0 \\ & 1 & 0 & 1 & 0 & 0 \\ & 1 & 1 & 0 & 1 & 1 \\ & 0 & 0 & 0 & 0 & 0 \end{array}$$

col-regular

$$\begin{array}{c|ccccc} & & & Y & & \\ \hline X & 0 & 0 & 0 & 0 & 0 \\ & 0 & 0 & 0 & 0 & 0 \\ & 0 & 0 & 0 & 0 & 0 \\ & 0 & 0 & 0 & 0 & 0 \end{array}$$

null

$$\begin{array}{c|ccccc} & & & Y & & \\ \hline X & 0 & 0 & 0 & 1 & 0 \\ & 0 & 0 & 1 & 0 & 0 \\ & 1 & 0 & 0 & 0 & 0 \\ & 0 & 0 & 0 & 1 & 0 \end{array}$$

row-functional

$$\begin{array}{c|cccc} & & Y & & \\ \hline X & 1 & 0 & 0 & 0 \\ & 0 & 1 & 0 & 0 \\ & 0 & 0 & 1 & 0 \\ & 0 & 0 & 0 & 0 \\ & 0 & 0 & 0 & 1 \end{array}$$

col-functional

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*:

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

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

Here an appropriate criterion function has to be defined.

Criterion function

Criterion functions can be constructed

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

Direct approach

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

In the direct approach the 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

First the set of the ideal blocks $\mathcal{B}$ according to the considered equivalence has to be defined.

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

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

where the term $d(R(C_u, C_v), B(C_u, C_v))$ measures the difference (the number of errors or inconsistencies) between the block $R(C_u, C_v)$ and the ideal block $B(C_u, C_v)$.

d is constructed on the basis of characterizations of type of blocks.

Example for structural equivalence

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 errors for each block

	A	B
A	0	1
B	1	2

The value of the criterion function is the sum of all errors (inconsistencies), $P = 4$.

Relocation approach

For solving the blockmodeling problem we use the relocation algorithm:

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;
- *interchanging* units x_u and x_v from different clusters C_p and C_q.

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 case given a network, set of types of ideal blocks, and a blockmodel, a clustering can be determined which minimizes the criterion function.

Types of blockmodels

The basic types of blockmodels are:

*	*	*
*	0	0
*	0	0

core - periphery

*	0	0
*	*	0
?	*	*

hierarchy

*	0	0
0	*	0
0	0	*

cohesive blockmodel

Advances in blockmodeling

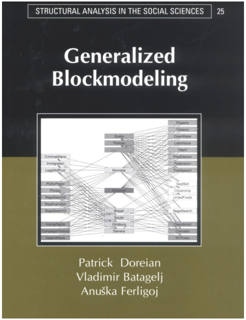
In our recent book

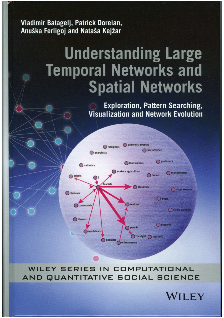
> *Doreian P., Batagelj V., Ferligoj A. (2020) (Eds.): Advances in Network Clustering and Blockmodeling, Wiley*

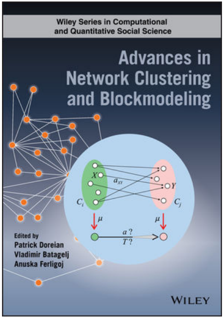
some new developments in blockmodeling are presented:

- blockmodeling of valued networks (Nordlund, Žiberna)
- blockmodeling of signed networks (Traag, Doreian, Mrvar)
- blockmodeling of multimode networks (Everett, Borgatti)
- blockmodeling of linked networks (Žiberna)
- stochastic blockmodeling (Peixoto)
- missing data and blockmodeling (Žnidaršič, Ferligoj, Doreian)

Books on blockmodeling and social network analysis (2005, 2014, 2020)







Example – Clustering of co-authorship networks

Kronegger, Ferligoj, Doreian (2011) studied the dynamics of co-authorship netwoks of Slovenian researchers.
Based on **four scientific disciplines in four five years long periods** (1986 - 2005);

- they concluded that the most typical form of co-authorship structure consist of the **multi core – semi-periphery – periphery** structure,
- they **studied the stability of cores** of obtained blockmodels of co-authorship networks visually.

Goal of the study

To analyze the stability of research teams for each Slovenian scientific discipline in two time periods (Cugmas, Ferligoj, Kronegger, 2020).

- To apply the pre-specified multi-core–semi-pheriphery–periphery blockmodeling **to most of the Slovenian scientific disciplines** for two time periods.
- The Modified Adjusted Rand Indices (MARI) to **measure core stability in two time periods** were defined and apllied it to the analyzed disciplines.

Data

The data set used here was obtained from two connected sources in Slovenia:

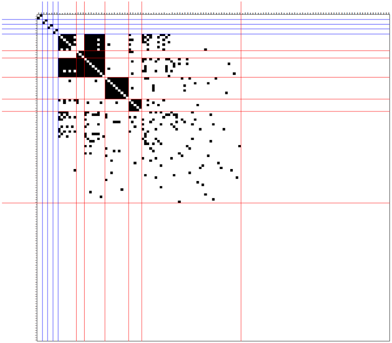
- the Current Research Information System (SICRIS) which includes the information on all active researchers registered at the Slovenian Research Agency and
- the Co-operative On-Line Bibliographic System & Services (COBISS) which contains a database of all publications available in Slovenian libraries.

. . . Data

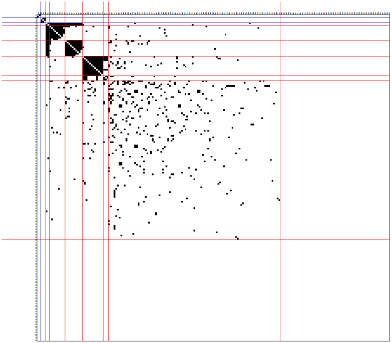
- Two researchers are linked, if they have published at least one scientific bibliographic unit together in a time period (co-authorship network).
- Scientific bibliographic output is defined by Slovenian Research Agency's (SRA).
- Two time periods: 1991 - 2000 and 2001 - 2010.
- The analysis included 43 out of 72 scientific disciplines.

An example of a blockmodel – Sociology

1991 - 2000

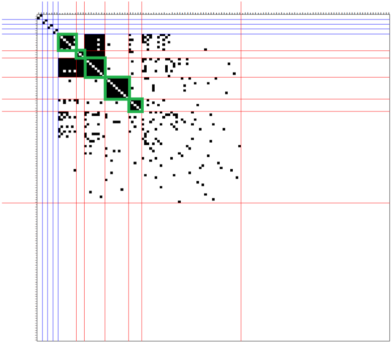


2001 - 2010

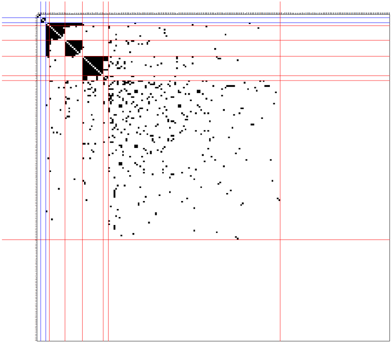


An example of a blockmodel – Cores

1991 - 2000

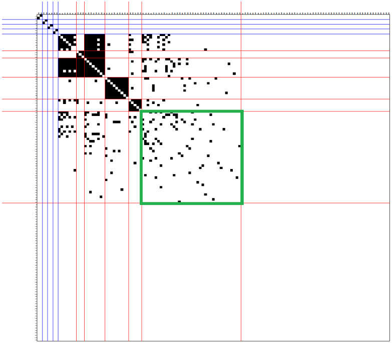


2001 - 2010

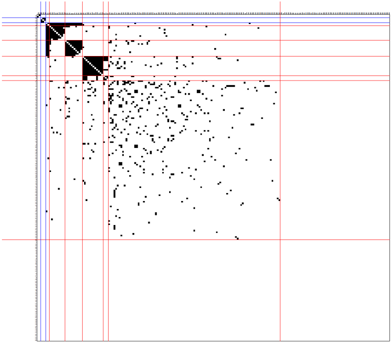


An example of a blockmodel – Semi-periphery

1991 - 2000

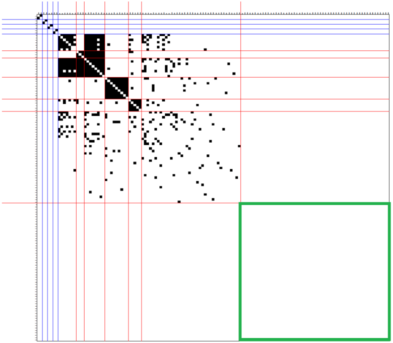


2001 - 2010

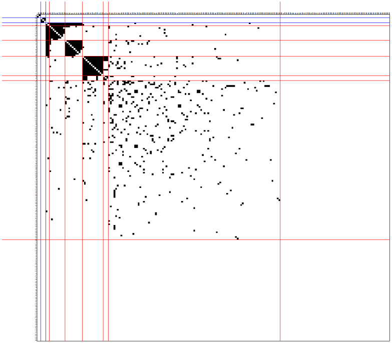


An example of a blockmodel – Periphery

1991 - 2000



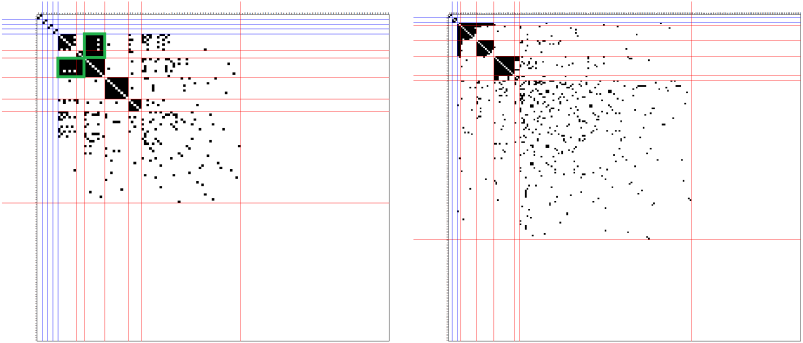
2001 - 2010



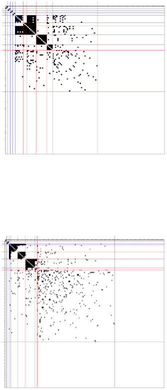
An example of a blockmodel – Bridging core

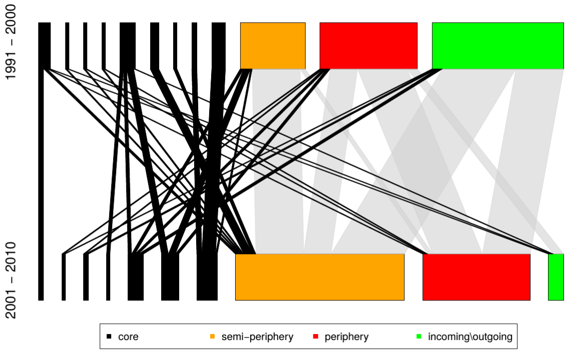
1991 - 2000

2001 - 2010

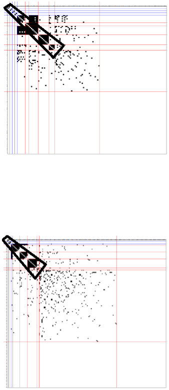


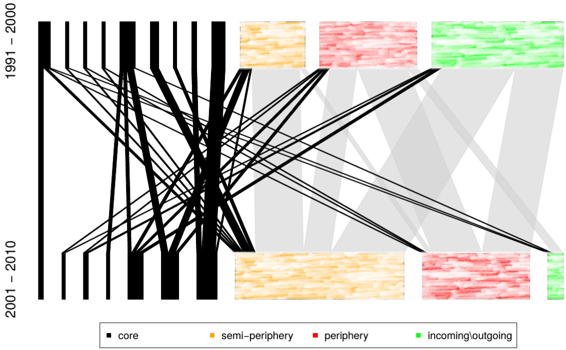
Visualization of blockmodels in two time periods



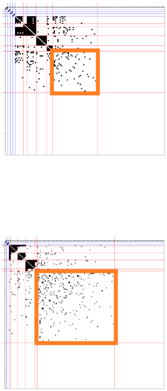


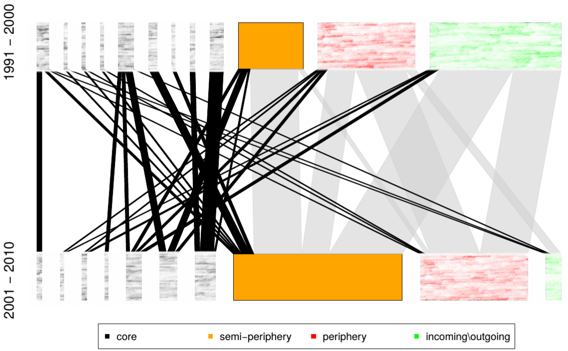
Visualization of blockmodels in two time periods



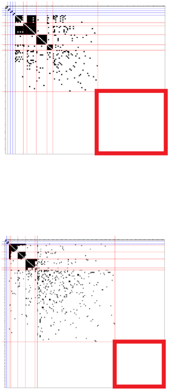


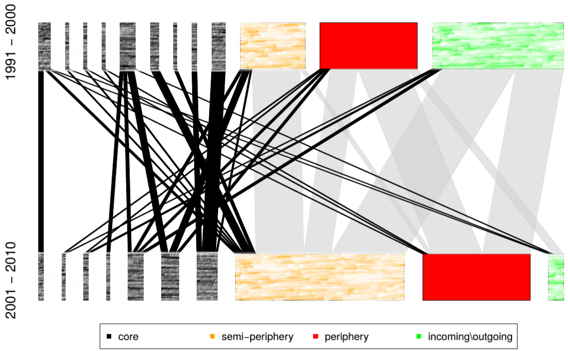
Visualization of blockmodels in two time periods



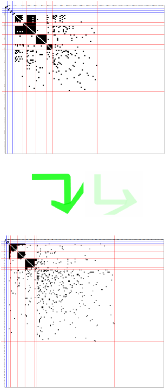


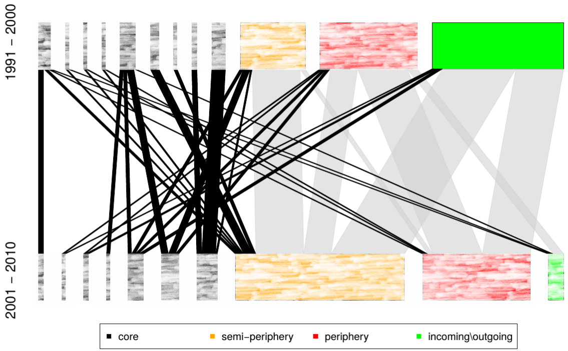
Visualization of blockmodels in two time periods



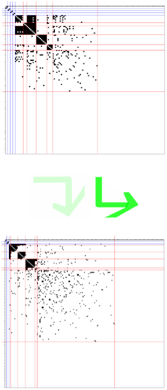


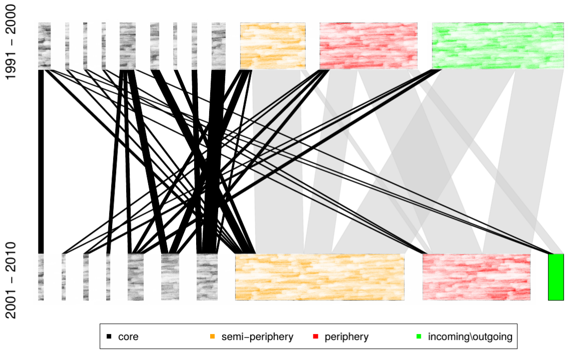
Visualization of blockmodels in two time periods





Visualization of blockmodels in two time periods



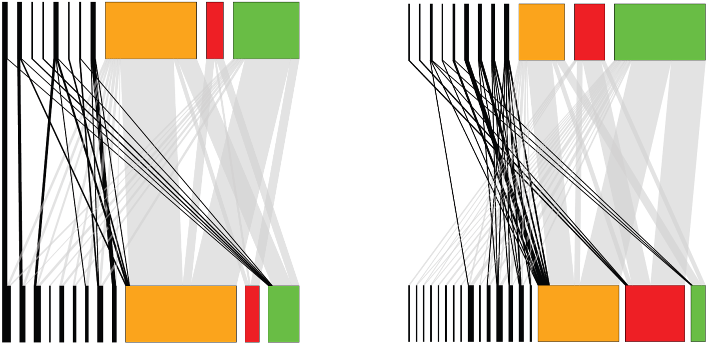


Transitions of researchers between two time periods for each scientific disciplines

The discipline with the most stable cores and with the least stable cores

Plant production
(MARI 1 = 0.49, N=312)

Biochem. and mol. biology
(MARI 1 = 0.01, N=301)



Clustering of attribute and relational data

Suppose that the units $\mathcal{U}$ are described by *attribute data* (variables) and related by a binary relation R that determines the *relational data*.

We want to cluster similar units according to the selected variables (treated by the criterion function), but also considering the relation R. The relation imposes *constraints* on the set of feasible clusterings $\mathbf{\Phi}$, usually in the following form:

$$\mathbf{\Phi} = \{\mathbf{C}\colon \text{ each cluster } C \in \mathbf{C} \text{ induces a subnetwork in the network } (\mathcal{U}, R) \text{ of the required type of connectedness}\}$$

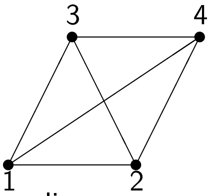
. . . Clustering with relational constraint

We can define different types of sets of feasible clusterings for the same relation R. Some examples of *types of relational constraint* Φ^i are

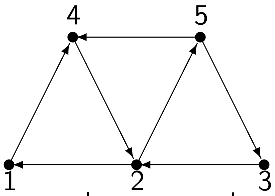
type of clusterings	type of connectedness
Φ^1	weakly connected units
Φ^2	weakly connected units that contain at most one center
Φ^3	strongly connected units
Φ^4	clique
Φ^5	the existence of a trail containing all the units of the cluster

Trail – all arcs are distinct.

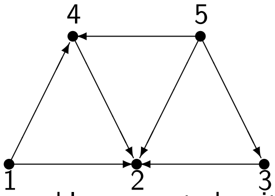
Some networks of different types



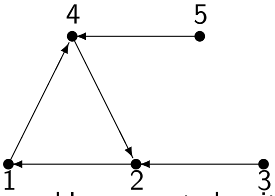
a clique



strongly connected units



weakly connected units



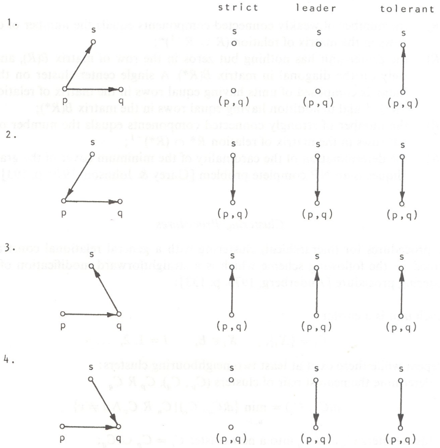
weakly connected units with a center $\{1, 2, 4\}$

Agglomerative approach for relational constraint

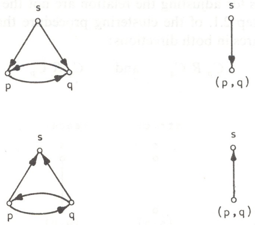
We can use both agglomerative and relocation approaches for solving some types of clustering problems with relational constraint (Ferligoj, Batagelj 1983). Here we present the agglomerative algorithm.
The algorithm begins with two data matrices: the dissimilarity matrix calculated from the attribute data and the relational matrix.

1. each unit is a cluster
2. **repeat while** there exist at least two related clusters:
2.1. determine the nearest pair of clusters (C_p, C_q), $C_p R C_q$
2.2. fuse clusters C_p and C_q into a new cluster $C_s = C_p \cup C_q$
2.3. replace the clusters C_p and C_q by the cluster C_s
2.4. determine the dissimilarities between the new cluster C_s and the other clusters
2.5. determine the relationships between the new cluster C_s and the other clusters according to the clustering type.

Adjusting relation after fusion



Φ^1 – tolerant
Φ^2 – leader
Φ^4 – two-way
Φ^5 – strict



Fast algorithm

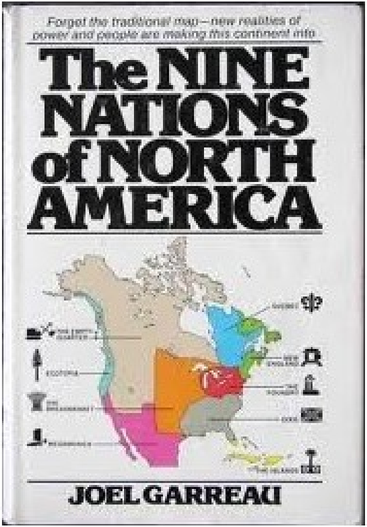
In the original approach, a complete dissimilarity matrix is needed. To obtain fast algorithms we proposed to *consider only the dissimilarities between linked units* (Batagelj, Doreian, Ferligoj, 2014, Chapter 9).

By this fast algorithm the following agglomerative methods can be used:

- minimum
- maximum
- average

This algorithm is available in the program `Pajek`.

Example: The nine nations of North America

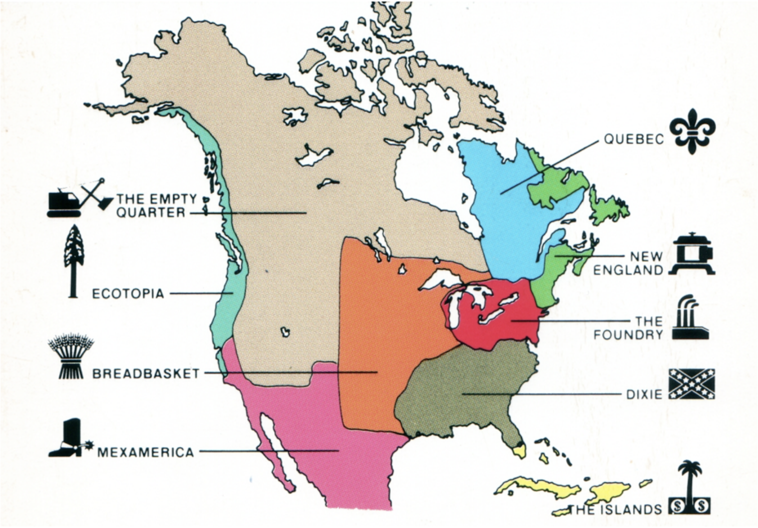


Joel Garreau suggested (1981) that North America can be divided into nine regions or "nations":

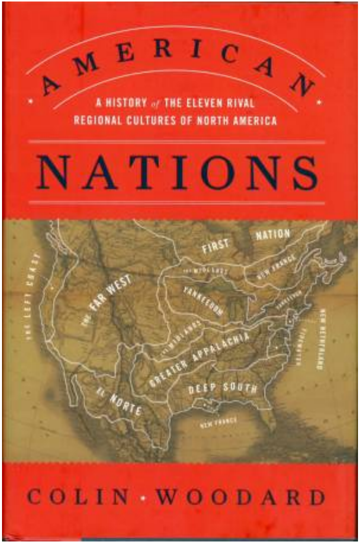
- New England
- The Foundry
- Dixie
- The Islands
- MexAmerica
- Ecotopia
- The Empty Quarter
- The Breadbasket
- Quebec

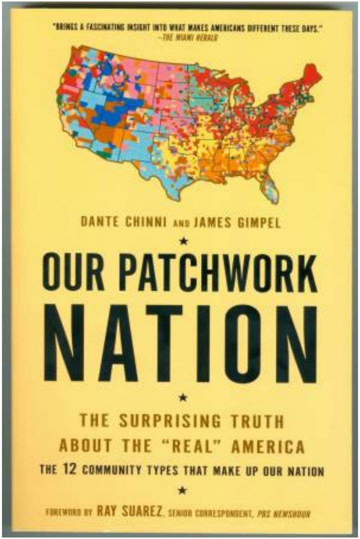
that represent the true nature of North American society.

The nine Nations of North America



Some other books on North America

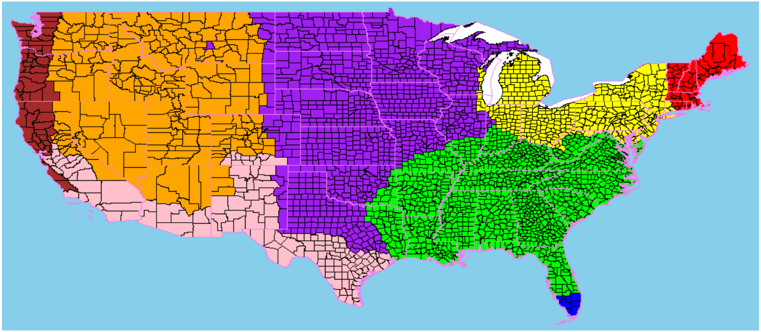




Research problem

Can we empirically reproduce the map that was proposed by Garreau using methods with constrained clustering?

Regionalization problem



Group given territorial units (counties) into regions such that units inside the region will be similar according to selected *attributes* (variables) and form *contiguous* parts of the territory (relational constraint).

Here, the relation is symmetric and we search for clusters with (strongly) connected units (counties) according to the contiguity relation.
In the obtained dendrogram at each cut of the dendrogram the connected units (regions) are obtained.

Data on US counties

Contiguity relation between 3111 mainland US counties was produced by Luc Anselin http://sal.uiuc.edu/weights/index.html
and some additional data were obtained from
ftp://spo.nos.noaa.gov/datasets/CADS/Data/references/reference_cnty.zip.
Because it is not included in other data sets South Boston [2916] was removed. We analyze 3110 counties.

Maps (shape files) for US states and counties were obtained from
http://gadm.org/data/shp/USA_adm.zip.
For visualization we used R library `maptools`.

The statistical data about US counties are available at US Census Bureau http://www.census.gov/support/USACdataDownloads.html.
It was a lot of work to make all three data sets compatible.
From 6849 variables we selected 1125 variables (+ unit name and Id) and stored them on the single CSV file.

Variables

We selected variables that are congruent with the monograph *Nine Nations of North America* and are available for the year 2000 (with some exeptions). Some variables were excluded because of muticolinearity. At the end we selected 42 variables for the following topics:

- demography
- age
- education
- poverty
- race
- income
- labor force
- employment
- housing
- crime
- land, water
- political

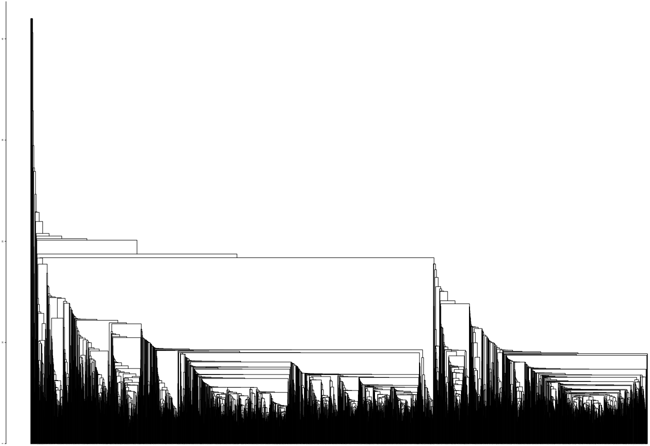
Variables

```
 [1] "median age"                            [2] "civilian labor force unemplyment rate"
 [3] "% bachelor's degree or hi. 25 +"       [4] "% change housing units 90-00"
 [5] "median household income"               [6] "% people in poverty"
 [7] "per capita personal income"            [8] "% change population 90-00"
 [9] "population per square mile"           [10] "% female population"
[11] "% Black or African American"          [12] "% Am. Indian or Alaska Native"
[13] "% Asian"                              [14] "% Hispanic or Latino"
[15] "birth rate"                           [16] "death rate"
[17] "infant mortality"                     [18] "water use per capita"
[19] "% pop.under18"                        [20] "% pop. 85+"
[21] "% land.farms"                         [22] "% emply.ind.CONSTRUCTION"
[23] "% emply.ind.MANUFACTORING"            [24] "% emply.ind.TRANSPORT.WAREHOUSING"
[25] "% emply.ind.FINANC.INSUR"             [26] "% emply.ind.PROFscientTECH"
[27] "% emply.ind.EDUC.HEALTH"              [28] "% 25overLESS9thGRADE"
[29] "% employ.FARMING"                     [30] "% employ.GOV.stateLoc"
[31] "% OWNERoccupiedHousingUnits"          [32] "% occupiedHousingUnitsLackingPlumb"
[33] "% RURALpopul"                         [34] "% CHANGE urban 90to00"
[35] "CHANGEperCapitaIncome89to99"          [36] "GroundWaterUsePerCapita"
[37] "% NET.DOMESTIC.MIGRATIONS"            [38] "% NativePopulationBornInStateOfRes"
[39] "R.LABOR.FORCEmaleFemale"              [40] "% VOTING.DEMOCRATESoverREPUBLICANS"
[41] "% PUBLIC.SCHOOL.ENROLNEMT"            [42] "% CHANGE poverty 95to00"
```

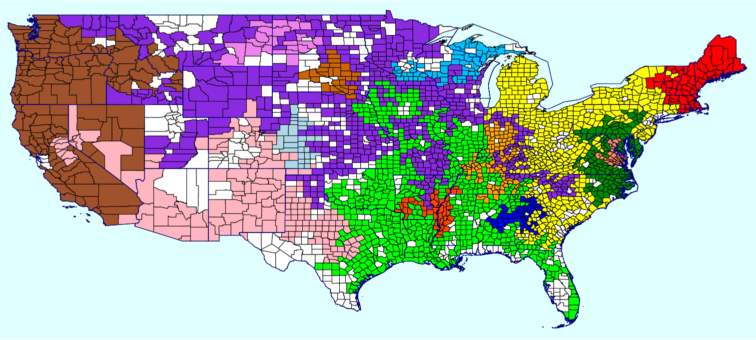
Empirically obtained regions

- Regionalization was obtained according to the selected 42 variables and the contiguity relation.
- The variables were standardized and the Euclidean distances between linked counties were computed.
- The `tolerant` strategy (each cluster induces a connected subnetwork) with the `Maximum` agglomerative method applied.

The obtained dendrogram

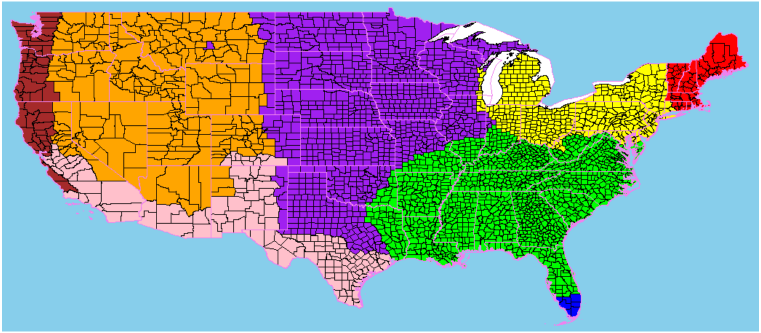


US counties / Maximum



We obtain many clusters. On the picture, we preserved only the largest 15 clusters (colored) with at least 20 counties and all others are considered as unclassified counties (white).

US counties / Nine nations



Some comments

- Counties that contain large cities and university cities are quite different from their neighborhoods (the white cluster). We can assign these counties by a post-processing to the corresponding neighborhoods.
- Some variables that are mentioned by Garreau are not available.
- Some differences between the obtained regions and the regions from Garreau's book can be the result of changes in 20 years.

Conclusions

- Each clustering problem has to be properly defined.
- Specific clustering problems can be treated by an appropriately defined set of feasible clusterings and the criterion function or dissimilarity measure.
- Here, two types of clustering problems were discussed: clustering of relational data – blockmodeling and the combination of attribute and relational data. Several other constraints can be treated, e.g., constraining variable(s) (Ferligoj 1986), and combined.
- Some presented algorithms are adapted for large datasets (clustering with relational constraint) some are not yet (generalized blockmodeling).

References I

- Batagelj V., Ferligoj A., Doreian P. (1992): *Direct and indirect methods for structural equivalence*. *Social Networks*, **14**, 63–90.
- Batagelj V., Doreian P., Ferligoj A. (1992): *An optimizational approach to regular equivalence*. *Social Networks*, **14**, 121–135.
- Batagelj V., Doreian P., Ferligoj A. and Kejzar N. (2014): Understanding large temporal networks and spatial networks: Exploration, pattern searching, visualization and network evolution, Chichester: Wiley.
- Batagelj V. Mrvar A.(1996-): *Pajek– program for analysis and visualization of large network*, home page, data sets.
- Cugmas M., Ferligoj A., Kronegger L. (2020): Scientific co-authorship networks. In Doreian P., Batagelj V., Ferligoj A. (Eds.), Advances in Network Clustering and Blockmodeling, Chichester: Wiley 363–387.

References II

- Doreian P., Batagelj V., Ferligoj A. (2005): Generalized Blockmodeling, Cambridgge: Cambridge University Press.
- Doreian P., Batagelj V., Ferligoj A. (2020): Advances in Network Clustering and Blockmodeling, Chichester: Wiley.
- Ferligoj A., Batagelj V. (1982): *Clustering with relational constraint*. *Psychometrika*, **47**(4), 413–426.
- Ferligoj A., Batagelj V. (1983): *Some types of clustering with relational constraints*. *Psychometrika*, **48**(4), 541-552.
- Ferligoj A. (1986): *Clustering with a constraining variable*. *The Journal of mathematical sociology*, **12**(3), 299-313.
- Garreau J. (1981): The Nine Nations of North America. Houghton Mifflin.
- Kronegger L., Ferligoj A., Doreian P. (2011): *On the dynamics of national scientific systems*. *Quality & Quantity*, **45**(5), 989–1015.