Photo: V. Batagelj

Hierarchical Clustering in Large Networks

Vladimir Batagelj
Anuška Ferligoj
Andrej Mrvar
University of Ljubljana

Detection, evolution and visualization of communities in complex networks

Louvain-la-Neuve, Belgium, March 13-14, 2008.

version: March 13, 2008 / 01 : 49

Outline

Based on *Hierarchical clustering with relational constraints of large data sets* presented at **6th Slovenian International Conference on Graph Theory**, Bled, Slovenia, 24 – 30 June 2007.

Regionalization problem

Departements and regions of France

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

In Ferligoj and Batagelj (1982 and 1983) we generalized this problem to *clustering with relational constraints problem* and proposed some algorithms to solve it.

Clustering with relational constraint

Suppose that the units are described by attribute data $a: \mathcal{U} \rightarrow [\mathcal{U}]$ and related by a binary *relation* $R \subseteq \mathcal{U} \times \mathcal{U}$ that determine the *relational data* $(\mathcal{U}, R, a)$.

We want to cluster the units according to the similarity of their descriptions, but also considering the relation R – it imposes *constraints* on the set of feasible clusterings, usually in the following form:

$$\begin{aligned}\mathbf{\Phi}(R) = \; & \{\mathbf{C} \in P(\mathcal{U}) : \text{each cluster } C \in \mathbf{C} \text{ induces a subgraph} \\ & (C, R \cap C \times C) \text{ in the graph } (\mathcal{U}, R) \text{ of the required type} \\ & \text{of connectedness}\}\end{aligned}$$

. . . Clustering with relational constraints

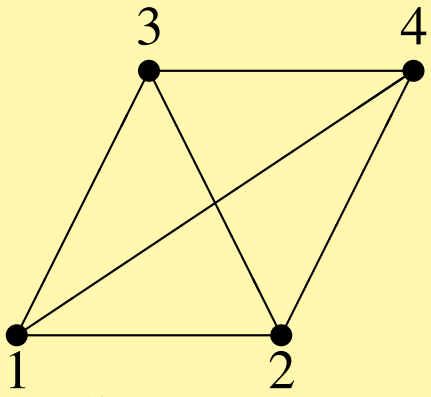
We can define different types of sets of feasible clusterings for the same relation R. Some examples of *types of relational constraint* $\Phi^i(R)$ are

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

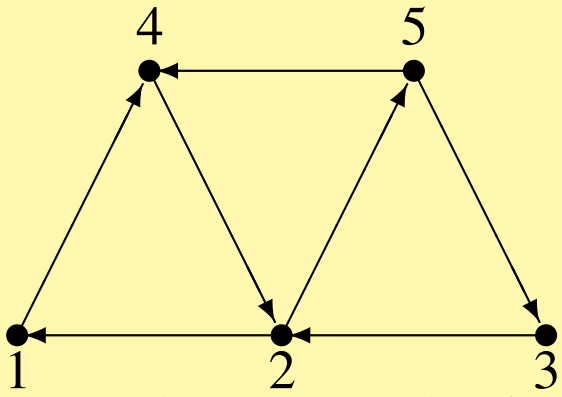
Trail – all arcs are distinct.

A set of units $L \subseteq C$ is a *center* of cluster C in the clustering of type $\Phi^2(R)$ iff the subgraph induced by L is strongly connected and $R(L) \cap (C \setminus L) = \emptyset$.

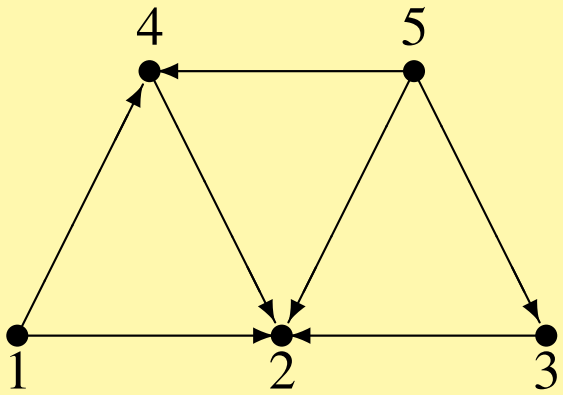
Some graphs of different types



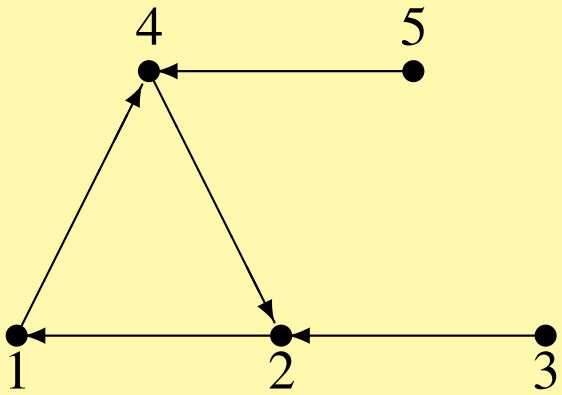
a clique



strongly connected units



weakly connected units



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

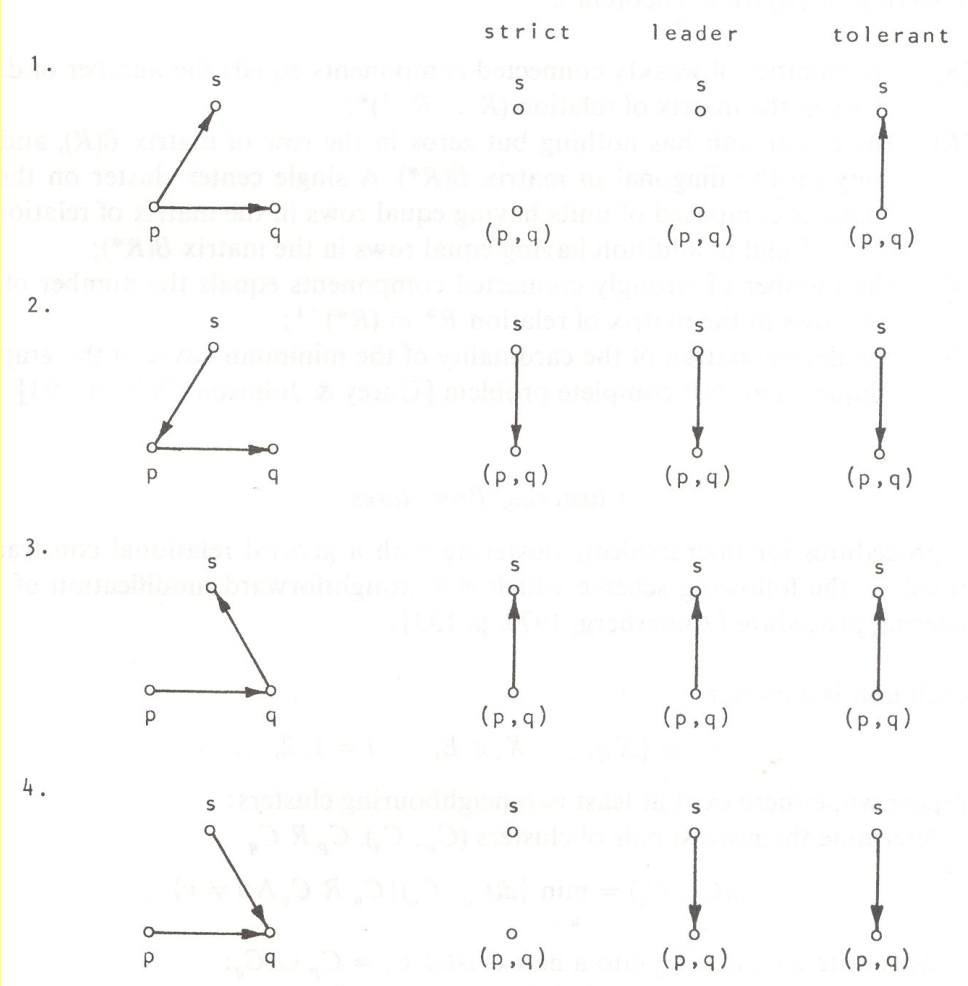
Agglomerative method for relational constraints

We can use both hierarchical and local optimization methods for solving some types of problems with relational constraint (Ferligoj, Batagelj 1983).

1. $k := n$; $\mathbf{C}(k) := \{\{\mathrm{X}\} : \mathrm{X} \in \mathcal{U}\}$;
2. **while** $\exists C_i, C_j \in \mathbf{C}(k)$: $(i \neq j \wedge \psi(C_i, C_j))$ **repeat**
   - 2.1. $(C_p, C_q) := \mathrm{argmin}\{D(C_i, C_j) : i \neq j \wedge \psi(C_i, C_j)\}$;
   - 2.2. $C := C_p \cup C_q$; $k := k - 1$;
   - 2.3. $\mathbf{C}(k) := \mathbf{C}(k+1) \setminus \{C_p, C_q\} \cup \{C\}$;
   - 2.4. determine $D(C, C_s)$ for all $C_s \in \mathbf{C}(k)$
   - 2.5. adjust the relation R as required by the clustering type
3. $m := k$

The *fusibility condition* $\psi(C_i, C_j)$ is equivalent to $C_i R C_j$ for tolerant, leader and strict method; and to $C_i R C_j \wedge C_j R C_i$ for two-way method.

Adjusting relation after joining

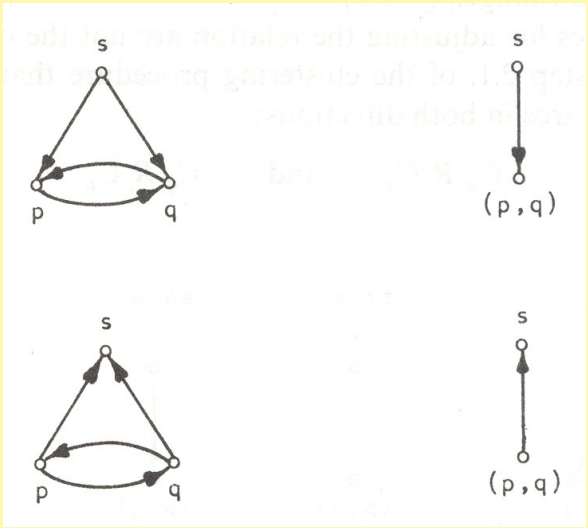


Φ^1 – tolerant

Φ^2 – leader

Φ^4 – two-way

Φ^5 – strict



Dissimilarities between clusters

In the original approach a complete dissimilarity matrix is needed. To obtain fast algorithms we propose to *consider only the dissimilarities between linked units*.

Let $(\mathcal{U}, R)$, $R \subseteq \mathcal{U} \times \mathcal{U}$ be a graph and $\emptyset \subset S, T \subset \mathcal{U}$ and $S \cap T = \emptyset$.

We call a *block* of relation R for S and T its part $R(S,T) = R \cap S \times T$.

The *symmetric closure* of relation R we denote with $\hat{R} = R \cup R^{-1}$. It holds: $\hat{R}(S,T) = \hat{R}(T,S)$.

For all dissimilarities between clusters $D(S,T)$ we set:

$$D(\{s\},\{t\}) = \begin{cases} d(s,t) & s\hat{R}t \\ \infty & \text{otherwise} \end{cases}$$

where d is a selected dissimilarity between units.

Minimum

$$D_{\min}(S,T) = \min_{(s,t)\in\hat{R}(S,T)} d(s,t)$$

$$D_{\min}(S, T_1 \cup T_2) = \min(D_{\min}(S,T_1), D_{\min}(S,T_2))$$

Maximum

$$D_{\max}(S,T) = \max_{(s,t)\in\hat{R}(S,T)} d(s,t)$$

$$D_{\max}(S, T_1 \cup T_2) = \max(D_{\max}(S,T_1), D_{\max}(S,T_2))$$

Average

$w : V \rightarrow \mathbb{R}$ – is a weight on units; for example $w(v) = 1$, for all $v \in \mathcal{U}$.

$$D_{\mathrm{a}}(S,T) = \frac{1}{w(\hat{R}(S,T))} \sum_{(s,t)\in\hat{R}(S,T)} d(s,t)$$

$$w(\hat{R}(S,T_1 \cup T_2)) = w(\hat{R}(S,T_1)) + w(\hat{R}(S,T_2))$$

$$D_{\mathrm{a}}(S,T_1 \cup T_2) = \frac{w(\hat{R}(S,T_1))}{w(\hat{R}(S,T_1 \cup T_2))} D_{\mathrm{a}}(S,T_1) + \frac{w(\hat{R}(S,T_2))}{w(\hat{R}(S,T_1 \cup T_2))} D_{\mathrm{a}}(S,T_2)$$

Reducibility

The dissimilarity D has the *reducibility* property (Bruynooghe, 1977) iff

$$D(C_p, C_q) \leq \min(D(C_p, C_s), D(C_q, C_s)) \Rightarrow$$

$$\min(D(C_p, C_s), D(C_q, C_s)) \leq D(C_p \cup C_q, C_s)$$

or equivalently

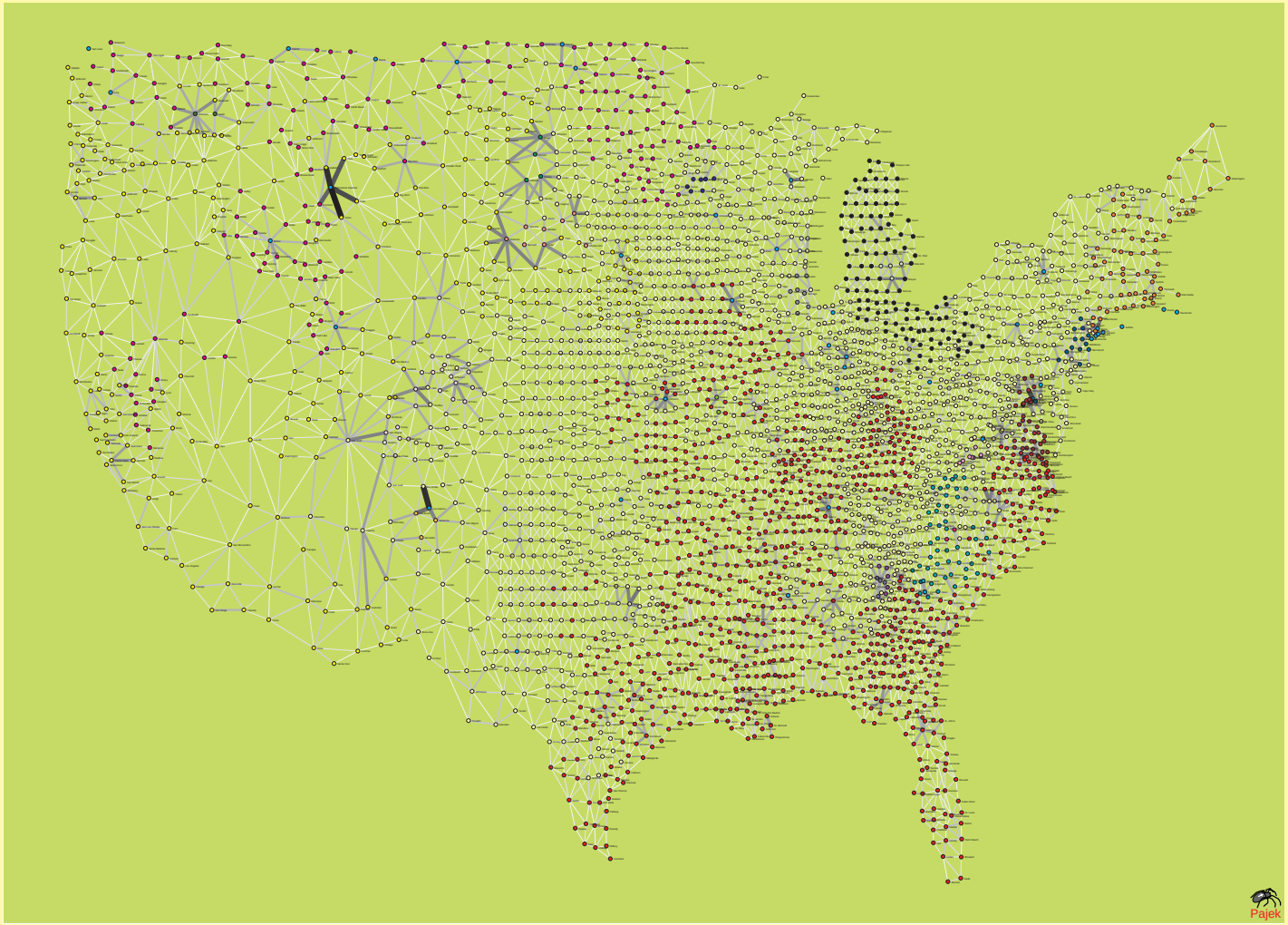
$$D(C_p, C_q) \leq t,\ D(C_p, C_s) \geq t,\ D(C_q, C_s) \geq t \Rightarrow D(C_p \cup C_q, C_s) \geq t$$

Theorem 1 *If a dissimilarity D has the reducibility property then h_D is a level function.*

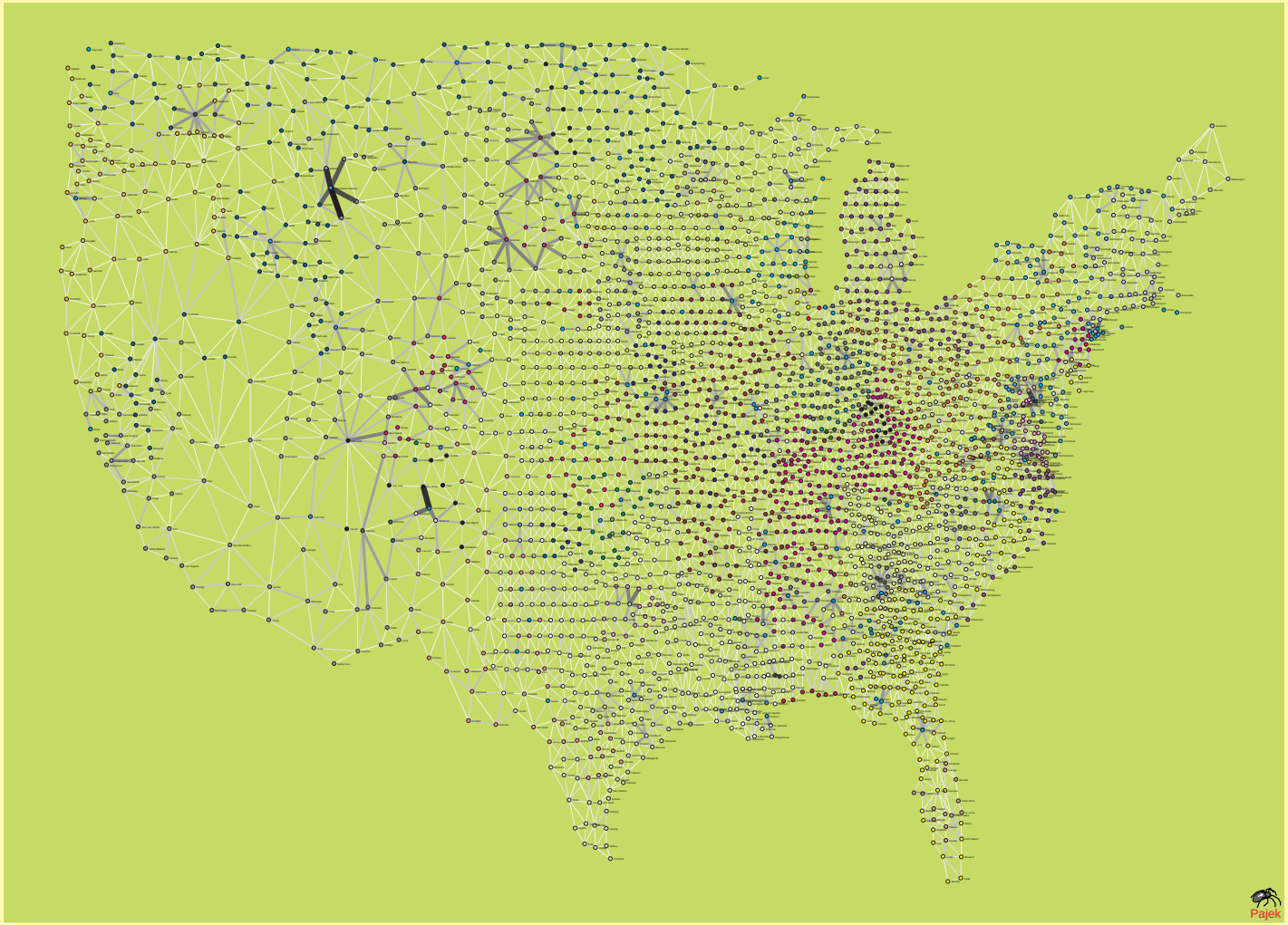
All three disimilarities have the reducibility property. In this case also the *nearest neighbors network* for a given network is preserved after joining the nearest clusters. This allows us to develop a very fast agglomerative hierarchical clustering procedure. It is available in program **Pajek** .

Example: US counties / maximum, $t = 1400$

US Census 2000: V1 – Area, V2 – Population, V47 – Percent of White, V125 – Educational attainment 1990, V126 – Household income; standardized



Example: US counties / maximum, $t = 200$



Hierarchical clustering in two-mode networks

At the Bertinoro workshop on graph drawing (9-14. March 2008) Katharina Zweig (Lehmann) asked for a method for clustering bipartite graphs. It turned out that our method can be easily adapted to solve also this problem.

Let $((U,V),L,d)$, $d : L \rightarrow \mathbb{R}_0^+$ be a weighted two-mode (bipartite) network. d is a dissimilarity measure between linked units (vertices). As an example of such dissimilarity we can take $d(u,v) = d(v,u) = w^* - w_4(u,v)$ where $w_4(u,v)$ is the number of different 4-cycles containing the line (u,v) and $w^* = \max_{(u,v)\in L} w_4(u,v)$.

We call a *two-mode partition* a set

$$\mathbf{C} = \{(C_1,D_1),(C_2,D_2),\ldots,(C_k,D_k)\}$$

such that the sets $\{C_i\} \setminus \emptyset$ and $\{D_i\} \setminus \emptyset$ are partitions of sets U and V; and $(\emptyset,\emptyset) \notin \mathbf{C}$.

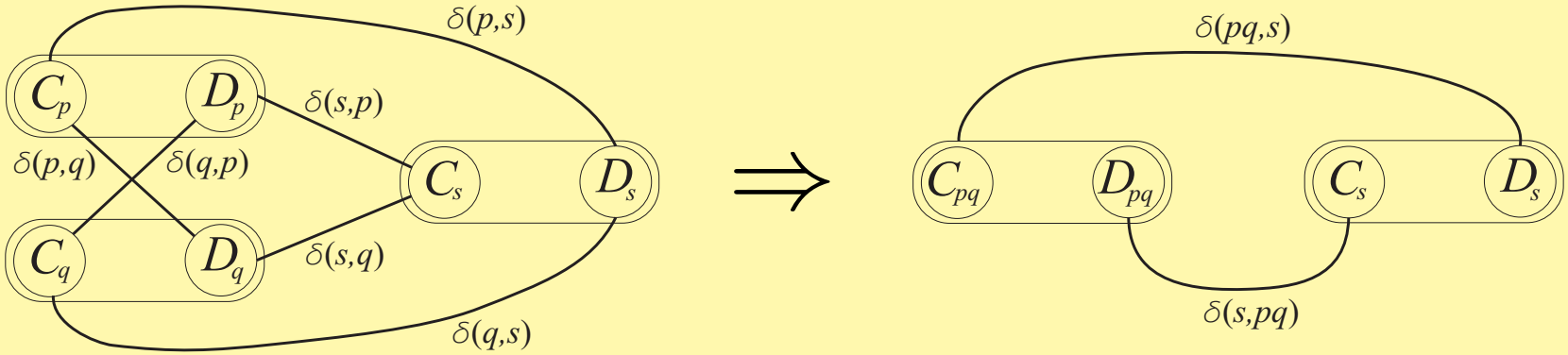
. . . Hierarchical clustering in two-mode networks

We start the hierarchical clustering with the singeltons partition

$$\mathbf{C}_N = U^{(1)} \times \{\emptyset\} \cup \{\emptyset\} \times V^{(1)} = \{(\{u_1\}, \emptyset), (\{u_2\}, \emptyset), \ldots, (\emptyset, \{v_n\})\}$$

where $U^{(1)} = \{\{u\} : u \in U\}$ and $N = |U \cup V|$.

The fusibility condition is now $\psi((C_i, D_i), (C_j, D_j)) = C_i R D_j \vee C_j R D_i$
where $C_i R D_j = \exists u \in C_i \exists v \in D_j : (u, v) \in L$.



The relation R update rules are

$$C_{pq} R C_s = C_p R C_s \vee C_q R C_s \quad \text{and} \quad C_s R C_{pq} = C_s R C_p \vee C_s R C_q$$

. . . Hierarchical clustering in two-mode networks

For a pair of two-mode clusters (C_p, D_p) and (C_q, D_q), such that C_pRD_q, let $\delta(p, q)$ denote the difference of cluster p from cluster q. δ is determined as follows

$$\delta(\{u\}, \{v\}) = d(u, v)$$

and the update formulae for δ after merging are: if $C_pRD_q \wedge C_qRD_p$ then

$$\delta((pq), s) = \oplus(\delta(p, s), \delta(q, s)) \quad \delta(s, (pq)) = \oplus(\delta(s, p), \delta(s, q))$$

where $\oplus \in \{\min, \max, \text{ave}\}$; otherwise only values of existing links are considered.

Using δ the clustering dissimilarity $D(p, q)$ is defined as

$$D(p, q) = \oplus(\delta(p, q), \delta(q, p))$$

Again the link needs not to exist in both 'directions'.

Conditions for hierarchical methods

The set of feasible clusterings $\mathbf{\Phi}$ determines the *feasibility predicate* $\Phi(\mathbf{C}) \equiv \mathbf{C} \in \mathbf{\Phi}$ defined on $\mathcal{P}(\mathcal{P}(\mathcal{U}) \setminus \{\emptyset\})$; and conversely $\mathbf{\Phi} \equiv \{\mathbf{C} \in \mathcal{P}(\mathcal{P}(\mathcal{U}) \setminus \{\emptyset\}) : \Phi(\mathbf{C})\}$.

In the set $\mathbf{\Phi}$ the relation of *clustering inclusion* $\sqsubseteq$ can be introduced by

$$\mathbf{C}_1 \sqsubseteq \mathbf{C}_2 \equiv \forall C_1 \in \mathbf{C}_1, C_2 \in \mathbf{C}_2 : C_1 \cap C_2 \in \{\emptyset, C_1\}$$

we say also that the clustering $\mathbf{C}_1$ is a *refinement* of the clustering $\mathbf{C}_2$.

It is well known that $(\Pi(\mathcal{U}), \sqsubseteq)$ is a partially ordered set (even more, semimodular lattice). Because any subset of partially ordered set is also partially ordered, we have: Let $\mathbf{\Phi} \subseteq \Pi(\mathcal{U})$ then $(\mathbf{\Phi}, \sqsubseteq)$ is a partially ordered set.

The clustering inclusion determines two related relations (on $\mathbf{\Phi}$):

$\mathbf{C}_1 \sqsubset \mathbf{C}_2 \equiv \mathbf{C}_1 \sqsubseteq \mathbf{C}_2 \wedge \mathbf{C}_1 \neq \mathbf{C}_2$ – strict inclusion, and

$\mathbf{C}_1 \boxdot \mathbf{C}_2 \equiv \mathbf{C}_1 \sqsubset \mathbf{C}_2 \wedge \neg\exists \mathbf{C} \in \mathbf{\Phi} : (\mathbf{C}_1 \sqsubset \mathbf{C} \wedge \mathbf{C} \sqsubset \mathbf{C}_2)$ – predecessor.

Conditions on the structure of the set of feasible clusterings

We shall assume that the set of feasible clusterings $\mathbf{\Phi} \subseteq \Pi(\mathcal{U})$ satisfies the following conditions:

F1. $\mathbf{O} \equiv \{\{\mathrm{X}\} : \mathrm{X} \in \mathcal{U}\} \in \mathbf{\Phi}$

F2. The feasibility predicate Φ is *local* – it has the form $\Phi(\mathbf{C}) = \bigwedge_{C \in \mathbf{C}} \varphi(C)$ where $\varphi(C)$ is a predicate defined on $\mathcal{P}(\mathcal{U}) \setminus \{\emptyset\}$ (clusters).

The intuitive meaning of $\varphi(C)$ is: $\varphi(C) \equiv$ the cluster C is 'good'. Therefore the locality condition can be read: a 'good' clustering $\mathbf{C} \in \mathbf{\Phi}$ consists of 'good' clusters.

F3. The predicate Φ has the property of *binary heredity* with respect to the *fusibility* predicate $\psi(C_1, C_2)$, i.e.,

$$C_1 \cap C_2 = \emptyset \wedge \varphi(C_1) \wedge \varphi(C_2) \wedge \psi(C_1, C_2) \Rightarrow \varphi(C_1 \cup C_2)$$

This condition means: in a 'good' clustering, a fusion of two 'fusible' clusters produces a 'good' clustering.

. . . conditions

F4. The predicate ψ is *compatible* with clustering inclusion $\sqsubseteq$, i.e.,

$$\forall \mathbf{C}_1, \mathbf{C}_2 \in \mathbf{\Phi} : (\mathbf{C}_1 \sqsubset \mathbf{C}_2 \wedge \mathbf{C}_1 \setminus \mathbf{C}_2 = \{C_1, C_2\} \Rightarrow \psi(C_1, C_2) \vee \psi(C_2, C_1))$$

F5. The *interpolation* property holds in $\mathbf{\Phi}$, i.e., $\forall \mathbf{C}_1, \mathbf{C}_2 \in \mathbf{\Phi}$:

$$(\mathbf{C}_1 \sqsubset \mathbf{C}_2 \wedge \text{card}(()\mathbf{C}_1) > \text{card}(()\mathbf{C}_2) + 1 \Rightarrow \exists \mathbf{C} \in \mathbf{\Phi} : (\mathbf{C}_1 \sqsubset \mathbf{C} \wedge \mathbf{C} \sqsubset \mathbf{C}_2))$$

These conditions provide a framework in which the hierarchical methods can be applied also for constrained clustering problems $\mathbf{\Phi}_k(\mathcal{U}) \subset \Pi_k(\mathcal{U})$.

In the ordinary problem both predicates $\varphi(C)$ and $\psi(C_p, C_q)$ are always true – all conditions F1-F5 are satisfied.

References

1. Batagelj, V. and Mrvar, A.(1996-): ***Pajek***– *program for analysis and visualization of large network*, home page, data sets.
2. Batagelj V., Ferligoj A. (2000): Clustering relational data. Data Analysis (Eds.: W. Gaul, O. Opitz, M. Schader), Springer, Berlin, 3–15.
3. Bruynooghe, M. (1977), Méthodes nouvelles en classification automatique des données taxinomiques nombreuses. Statistique et Analyse des Données, **3**, 24–42.
4. Ferligoj A., Batagelj V. (1982): *Clustering with relational constraint*. *Psychometrika*, **47**(4), 413–426.
5. Ferligoj A., Batagelj V. (1983): *Some types of clustering with relational constraints*. *Psychometrika*, **48**(4), 541–552.
6. Murtagh, F. (1985), Multidimensional Clustering Algorithms, *Compstat lectures*, **4**, Vienna: Physica-Verlag.
7. US census 2000 / counties. http://fisher.lib.virginia.edu/collections/stats/ccdb/county2000.html