Hierarchical thematic model visualising algorithm

Arsenty Kuzmin, Alexander Aduenko and Vadim Strijov

Moscow Institute of Physics and Technology
Department of Control and Applied Mathematics

EURO 2013

EURO 2013 Thematic visualisation

We must offer a Decision Support System for thematic clustering.

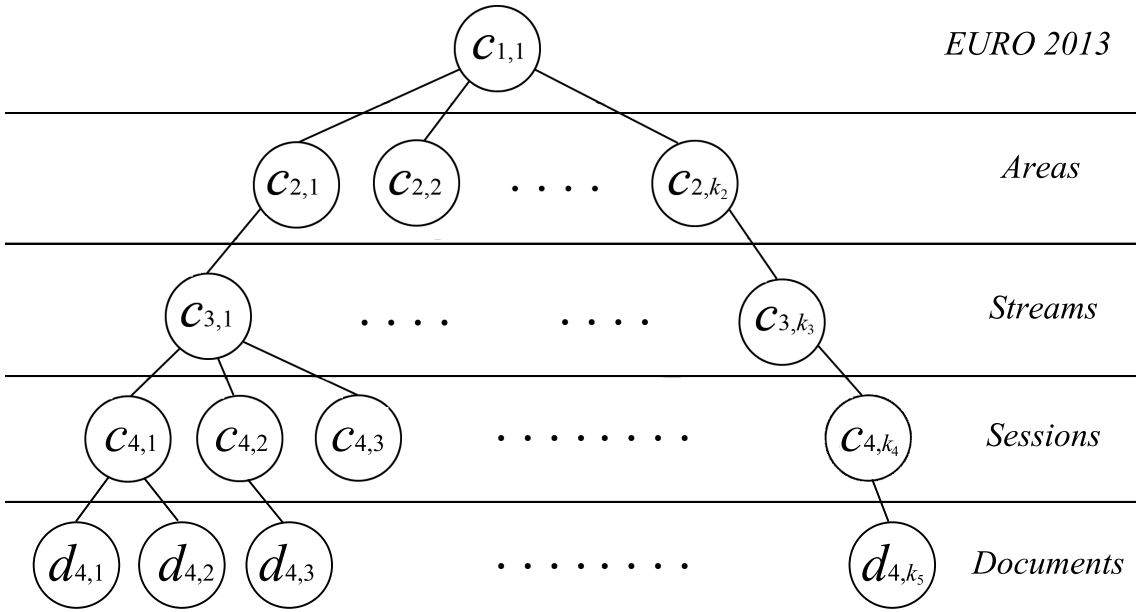
The goals:

- to construct a thematic model of the conference,
- to reveal the inconsistencies between the constructed and the expert model,
- to visualise the expert model and the revealed inconsistencies.

Call for algorithmic model:

- To be similar with the expert model,
- To rank the inconsistencieses,
- To have a plain representation of the conference structure.

EURO 2013 Conference structure



Create the expert thematic model

1. A group of experts is responsiable for each Area,
2. Participants send documents to the collection,
3. The experts distribute the documents over the Streams,
4. The documents are organised into Sessions.

Challenges

Causes of the problems

1. Great number of the experts (more than 200),
2. expert classification could be controversial,
3. there is no base thematic model.

The problems

1. To verify thematic consistency,
2. to detect inconsistencies in the hierarchical model,
3. to detect the unclaimed Streams and Sessions,
4. to assess quality of the expert hierarchical model.

Matrix "document/terms"

Let the theme of the document be determined by its terms.

$W = \{w_1, \ldots, w_n\}$ is the dictionary of the conference.

Let the document be the bag of words.

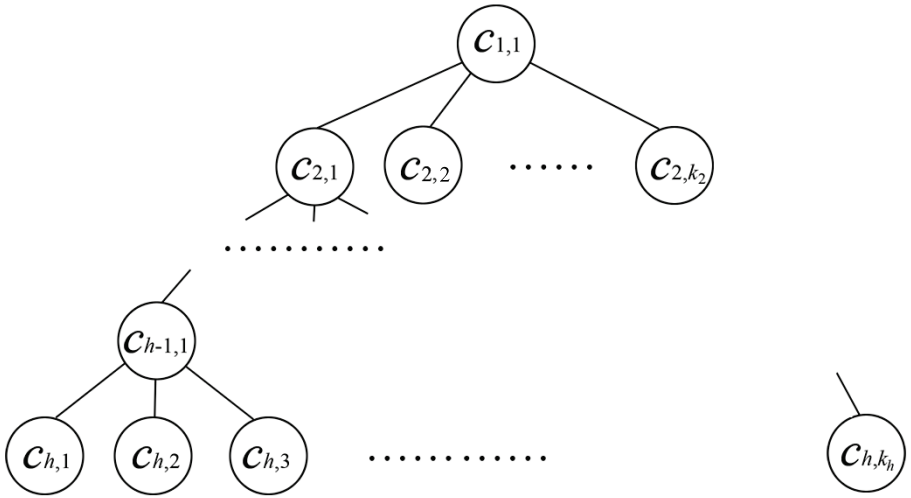
The document d of the collection D is an unordered set of words of the dictionary W, $d = \{w_j\}, j \in \{1, \ldots, n\}$.

$$\mathbf{x}_s \mapsto \frac{\mathbf{x}_s}{\sqrt{\mathbf{x}_s^{\mathsf{T}} \mathbf{x}_s}}, \qquad \mathbf{X} = \begin{pmatrix} x_{1,1} & \cdots & x_{1,n} \\ \cdots & \cdots & \cdots \\ x_{|D|,1} & \cdots & x_{|D|,n} \end{pmatrix}.$$

Hierarchical representation of the thematic model

Each leaf (h, i) of the tree corresponds to the document d_i.

Each node (ℓ, i), $\ell \neq h$ corresponds to the cluster $c_{\ell,i}$, which consists of corresponding documents.



Here ℓ is a conference level, $h = 5$ is the number levels and i is the index of a node given level.

Similarity function

Define the similarity function $s(\cdot, \cdot)$ between documents $\mathbf{x}_i$ and $\mathbf{x}_j$ as:

$$s(\mathbf{x}_i, \mathbf{x}_j) = \frac{\mathbf{x}_i^{\mathsf{T}} \mathbf{x}_j}{\|\mathbf{x}_i\|_2 \|\mathbf{x}_j\|_2} = \mathbf{x}_i^{\mathsf{T}} \mathbf{x}_j.$$

Define the similarity function $S(\cdot, \cdot)$ between clusters $c_{\ell,i}$ and $c_{\ell,j}$ as the mean $s(\mathbf{x}, \mathbf{y})$ between their documents $\mathbf{x} \in c_{\ell,i}, \mathbf{y} \in c_{\ell,j}$

$$S(c_{\ell,i}, c_{\ell,j}) = \frac{1}{|A|} \sum_{(\mathbf{x}, \mathbf{y}) \in A} s(\mathbf{x}, \mathbf{y}),$$

where A is the set of all document pairs from clusters $c_{\ell,i}$ and $c_{\ell,j}$, $\mathbf{x} \in c_{\ell,i}$, $\mathbf{y} \in c_{\ell,j}$, $\mathbf{x} \neq \mathbf{y}$.

The clustering quality function

Suppose F_0 is a mean intra-cluster similarity: $F_0 = \frac{1}{k_\ell} \sum_{i=1}^{k_\ell} S(c_{\ell,i},\ c_{\ell,i})$,

and F_1 is a mean inter-cluster similarity: $F_1 = \frac{2}{k_\ell(k_\ell - 1)} \sum_{i<j} S(c_{\ell,i},\ c_{\ell,j})$

Clustering quality criterion

$F = \frac{F_1}{F_0} \to \min.$

The expert hierarchical model is the origin for the algorithmic thematic model.

Distance functions

Euclidean distance:

$$\rho(\mathbf{x},\mathbf{y}) = \sqrt{\sum_{k=1}^{n}(x_k - y_k)^2}$$

Hellinger distance:

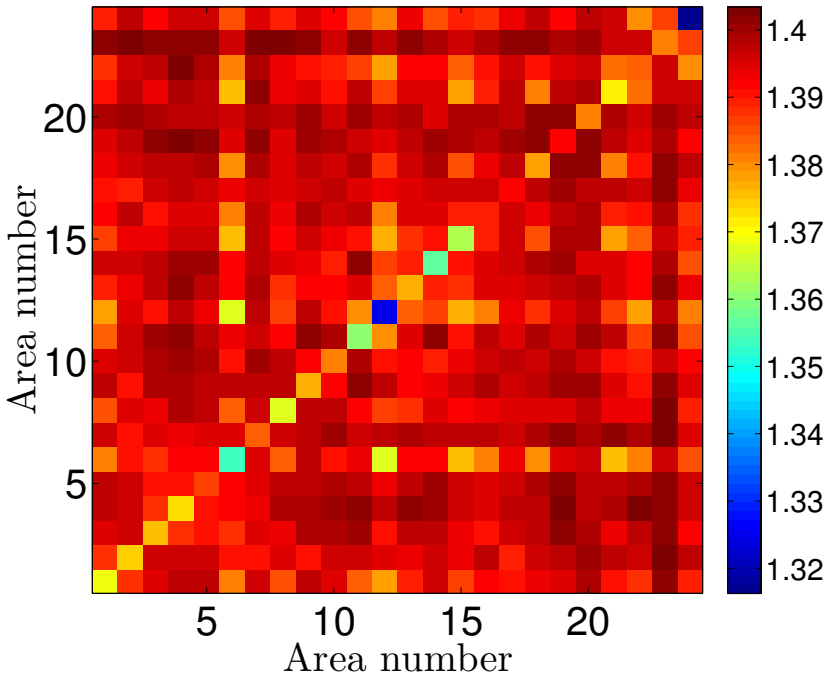
$$H(\mathbf{x},\mathbf{y}) = \frac{1}{\sqrt{2}}||\sqrt{\mathbf{x}} - \sqrt{\mathbf{y}}||_2$$

Jenson-Shannon distance:

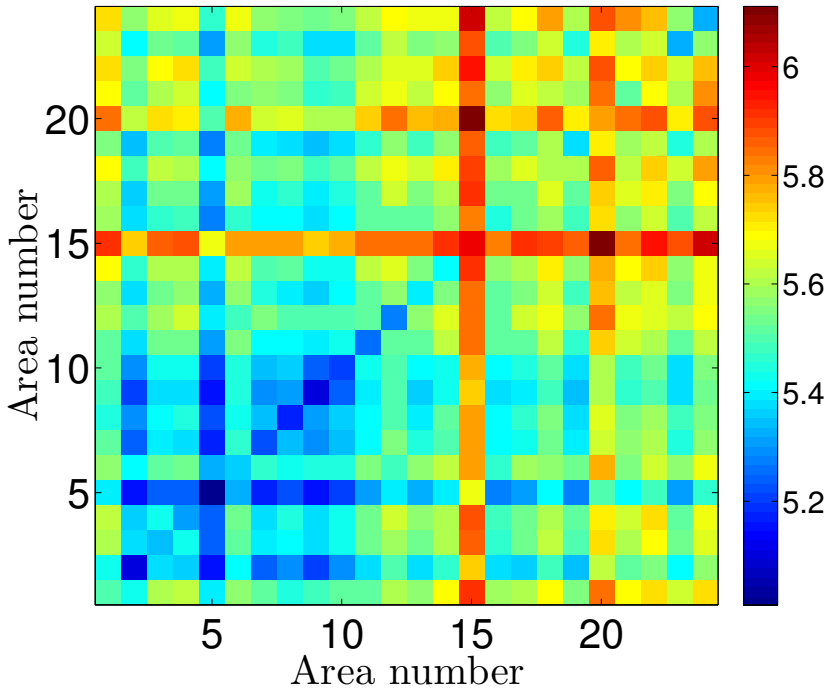
$$JSD(\mathbf{x}||\mathbf{y}) = \frac{1}{2}D(\mathbf{x}||M) + \frac{1}{2}D(\mathbf{y}||M),\ M = \frac{1}{2}(\mathbf{x}+\mathbf{y}),\ \text{where}$$

$$D(\mathbf{x}||\mathbf{y}) = \sum_i \ln\left(\frac{x_i}{y_i}\right)x_i \quad \text{is a Kullback–Leibler distance}$$

Distance and similarity functions comparison

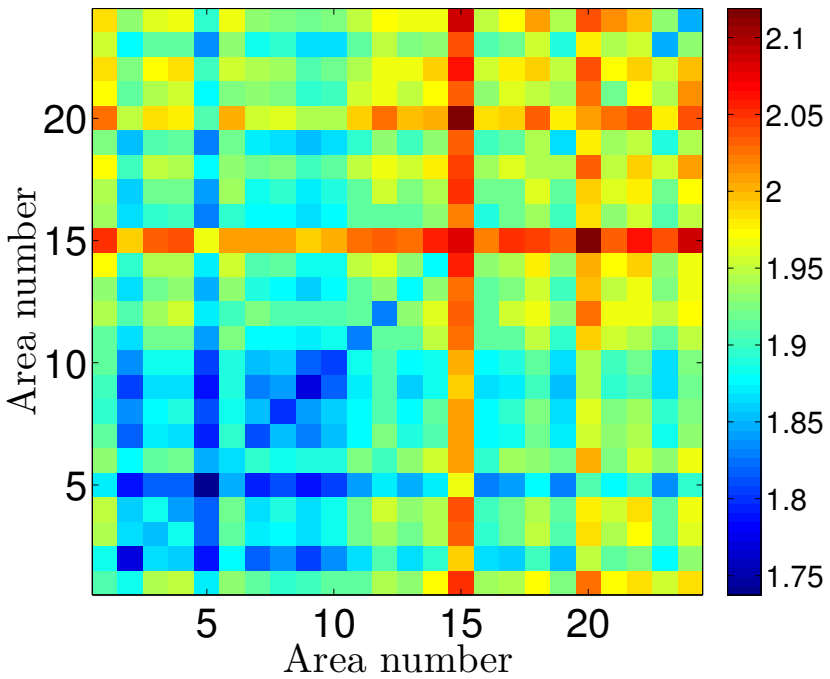


Euclidean distance

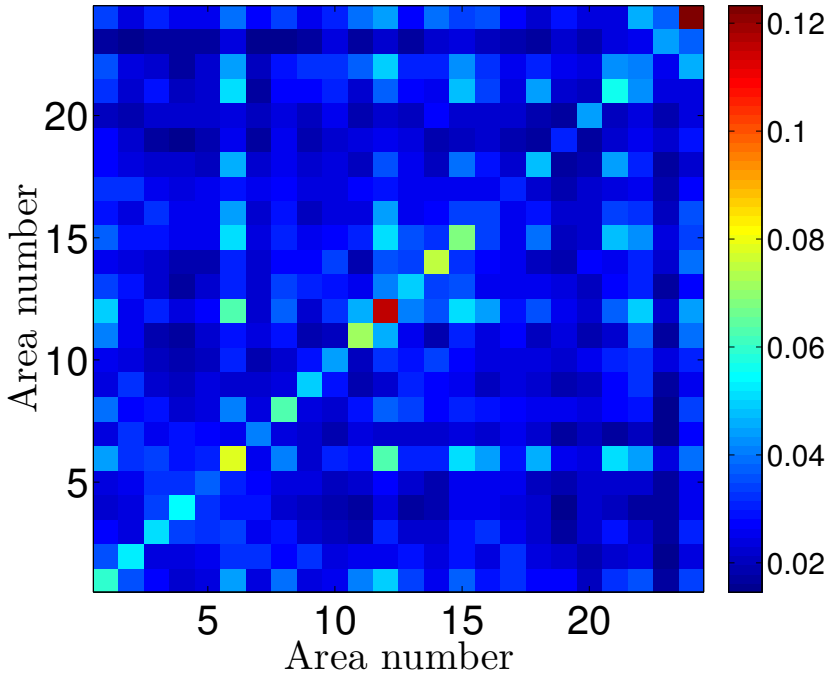


Hellinger distance

Distance and similarity functions comparison



Jenson-Shannon distance



Proposed similarity function

The quality function of the hierarchical model

The linear combination of intra- and inter-cluster similarities:

$$Q(\bar{\mathbf{x}}_1, \ldots, \bar{\mathbf{x}}_k) =$$

$$= \sum_{\ell=2}^{h-1} \left[\frac{1-\alpha}{k_\ell} \sum_{i=1}^{k_\ell} |c_{\ell,i}| S(c_{\ell,i}, c_{\ell,i}) - \frac{2\alpha}{k_\ell(k_\ell-1)} \sum_{i<j} S(c_{\ell,i}, c_{\ell,j}) \right] \to \max$$

$\alpha \in [0,1]$ is the weights coefficient; it determines the clustering priority.
k_ℓ is the quantity of clusters on level ℓ.

Creating an algorithmic model, similar to the expert model

Penalty matrix.

From \ To	$(+,+)$	$(+,-)$	$(-,-)$
$(+,+)$	δ_{11}	δ_{12}	δ_{13}
$(+,-)$	δ_{21}	δ_{22}	δ_{23}
$(-,-)$	δ_{31}	δ_{32}	δ_{33}

Move a document from its expert cluster to an algorithmic cluster if

$$Q_2 - Q_1 \geq \delta$$

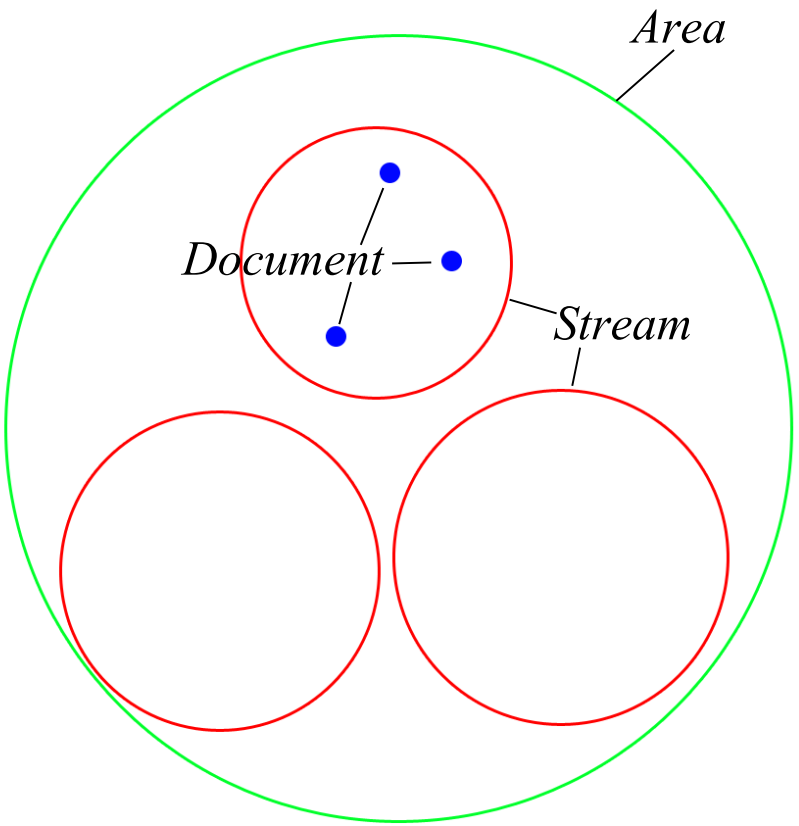
the algorithmic model quality drastically increases.

Nested visualisation

Visualisation requirements

1. to keep hierarchical structure of the expert model
2. to preserve the relative distances

- $\boldsymbol{\mu}(c_{\ell,i})$ is the coordinates of the cluster $c_{\ell,i}$ center
- $\rho(\cdot,\ \cdot)$ is the distance between documents in $\mathbb{R}^{|W|}$
- $\rho_2(\cdot,\ \cdot)$ is the distance between projections of the documents



Nested model creation

Let the cluster $c_{\ell,i}$ with the plain radius R is already placed on the plain; $C_1, \ldots, C_q$ are the clusters of the level $\ell+1$, which are in $c_{\ell,i}$, $\boldsymbol{\mu}(C_1), \ldots, \boldsymbol{\mu}(C_q)$ are their centers, $r_1, \ldots, r_q$ are their radiuses.

1. Make a Sammon projection of the centers $\boldsymbol{\mu}(C_1), \ldots, \boldsymbol{\mu}(C_q)$.
2. Find the plain radiuses $\hat{r}_1, \ldots, \hat{r}_q$ of the clusters $C_1, \ldots, C_q$ as:
$$\hat{r}_j = \min_{i \neq j} \frac{r_j}{r_j + r_i} \rho_2\big(\boldsymbol{\mu}(C_i), \boldsymbol{\mu}(C_j)\big).$$
3. Find $\hat{\rho} = \max\limits_{j \in \{1, \ldots, q\}} \rho_2\big(\boldsymbol{\mu}(C_j), \boldsymbol{\mu}_{\ell,i}\big) + \hat{r}_j$is the distance to the border of the projection, taking into consideration sizes of clusters.
4. Translate the homothety of the ratio $\dfrac{R}{\hat{\rho}}$ and center $\boldsymbol{\mu}(c_{\ell,i})$

Documents collection

The purpose of the experiment

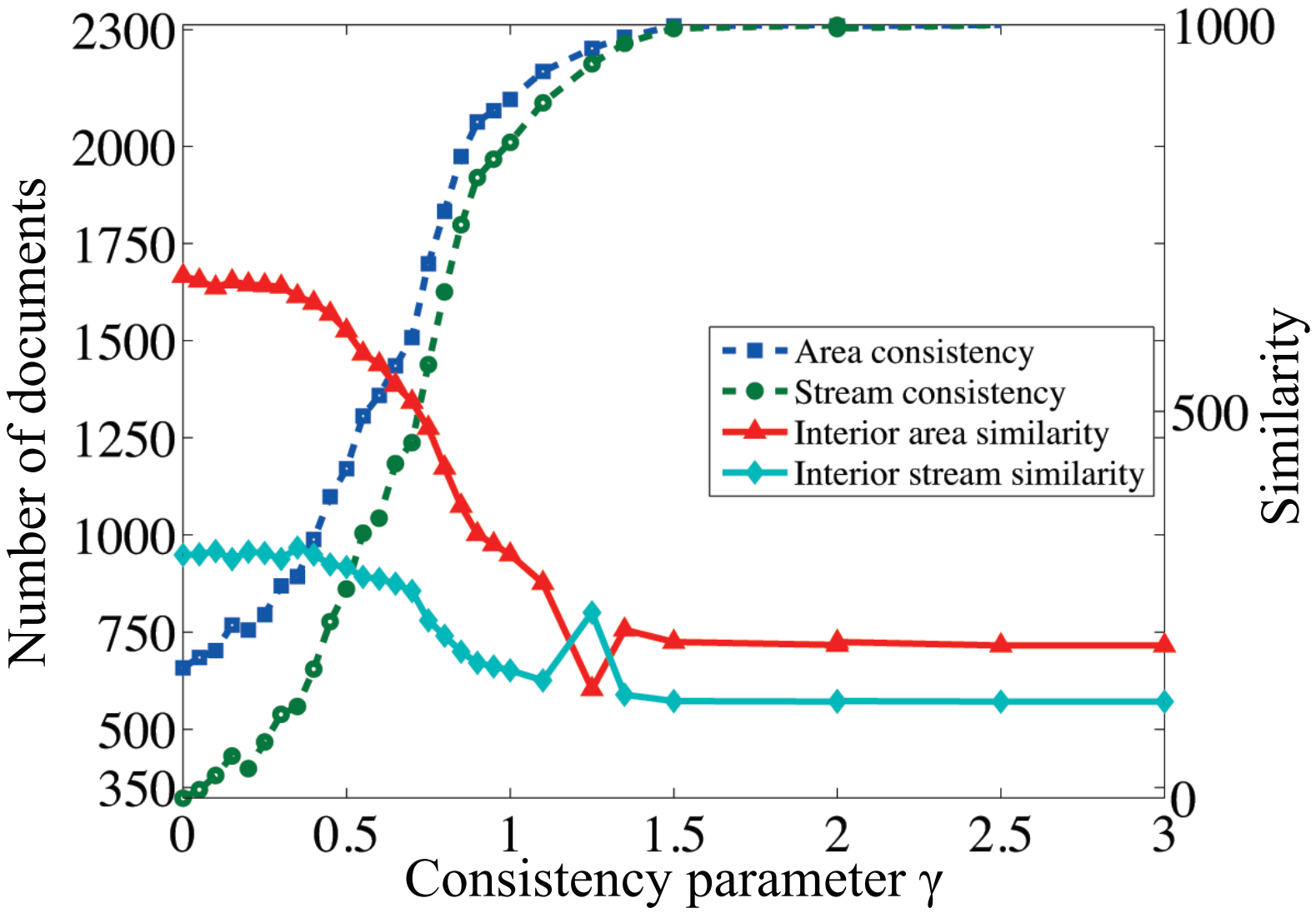
Visualisation of hierarchical thematic model of EURO 2013.

- Documents quantity: $|D| = 2313$, Areas: 24, Streams: 137.
- Size of the dictionary: $|W| = 1261$.
- Penalties are determined by the parameter of nonconformance with the expert model: $\gamma \geq 0$, $\mathbf{F} = \gamma\tilde{\mathbf{F}}$.

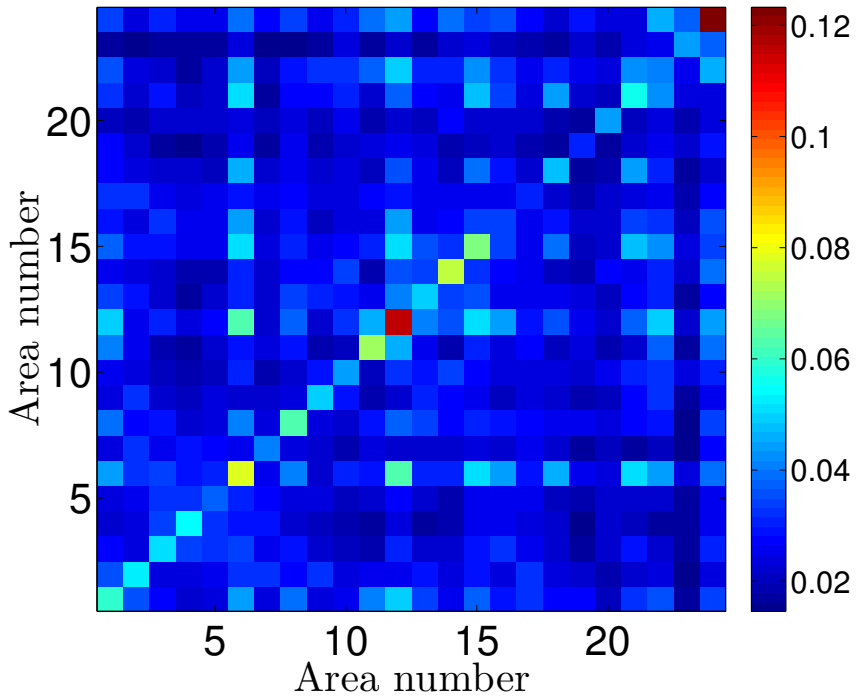
Penalties matrix $\tilde{\mathbf{F}}$.

From \ To	$(+,+)$	$(+,-)$	$(-,-)$
$(+,+)$	0	0.002	0.005
$(+,-)$	-0.001	0	0.003
$(-,-)$	-0.003	-0.002	0

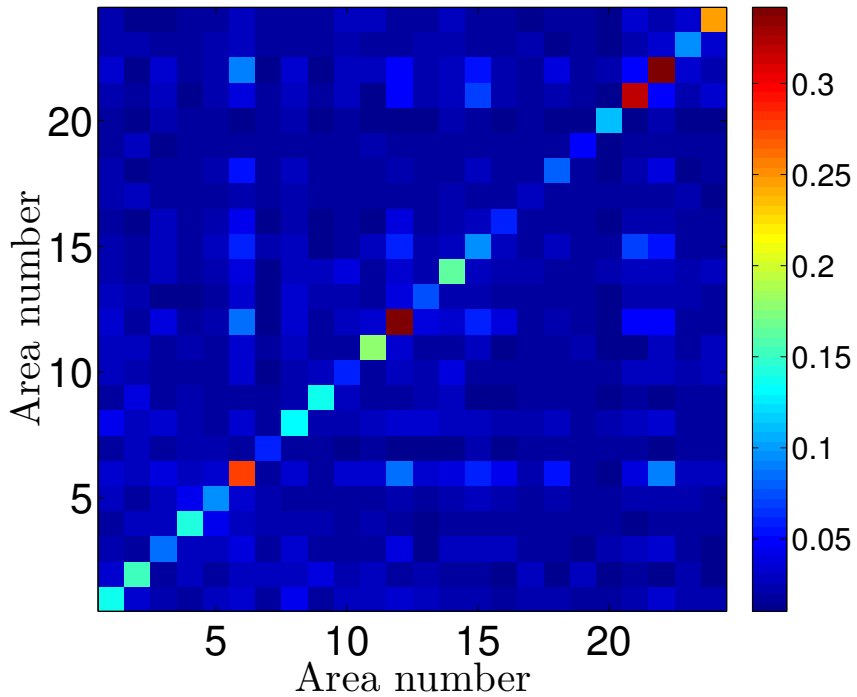
Clustering results with various penalties γ



Area similarity

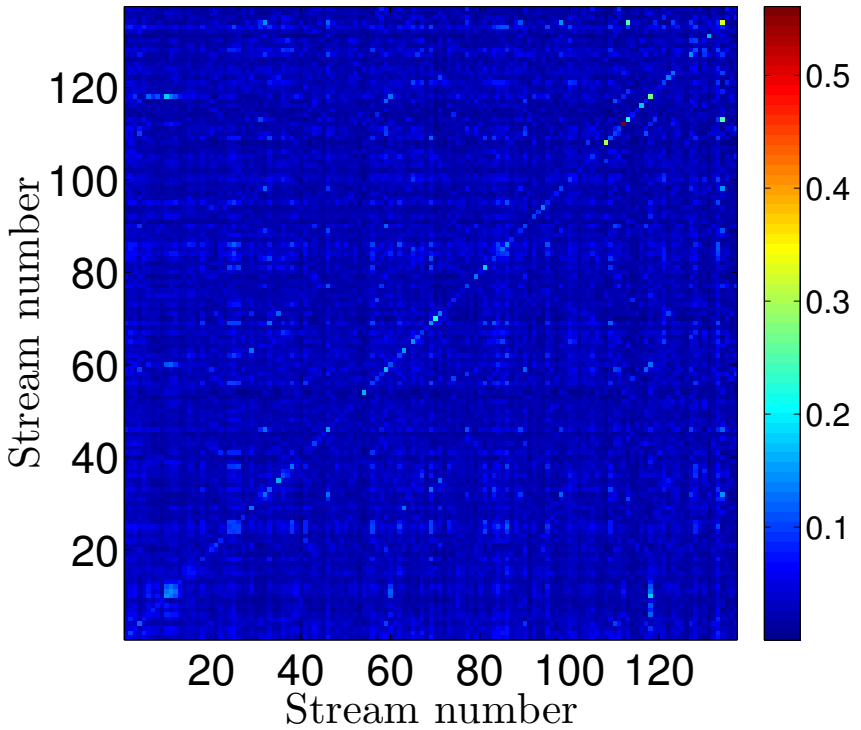


Expert clustering

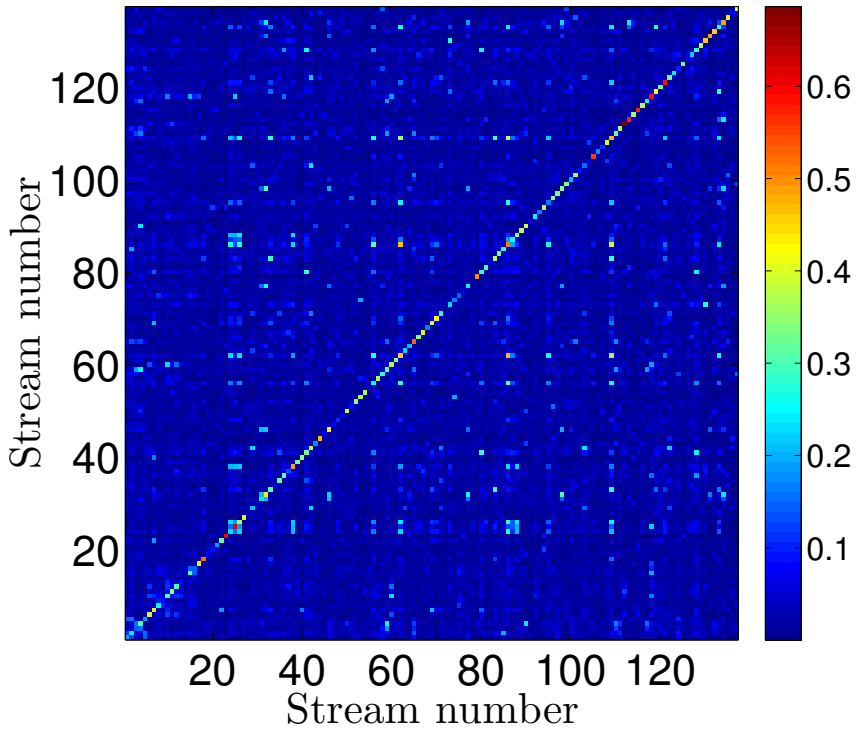


Algorithmic clustering

Streams similarity

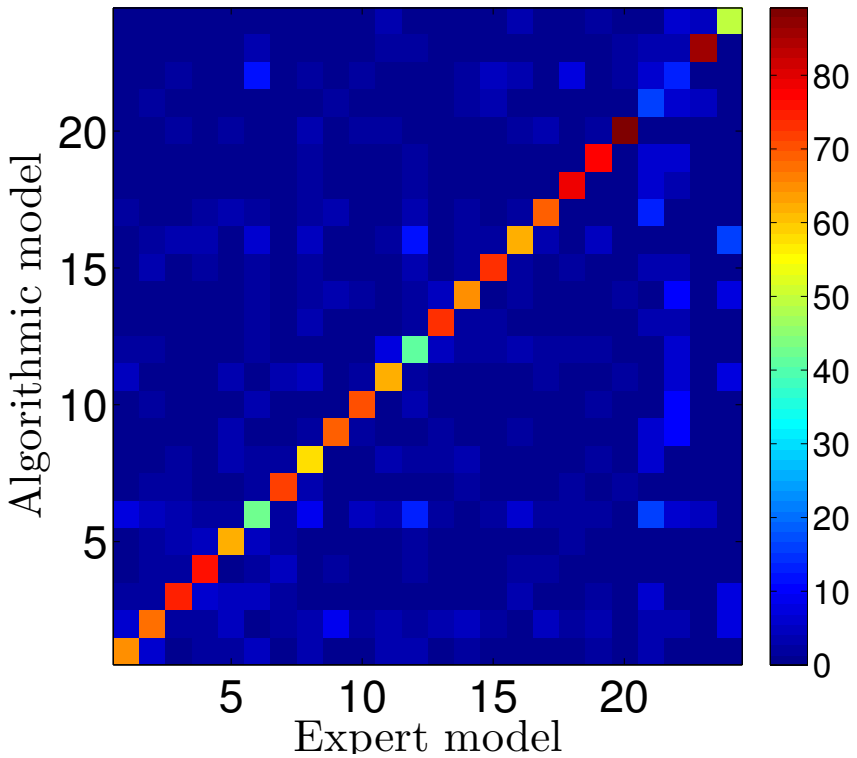


Expert clustering

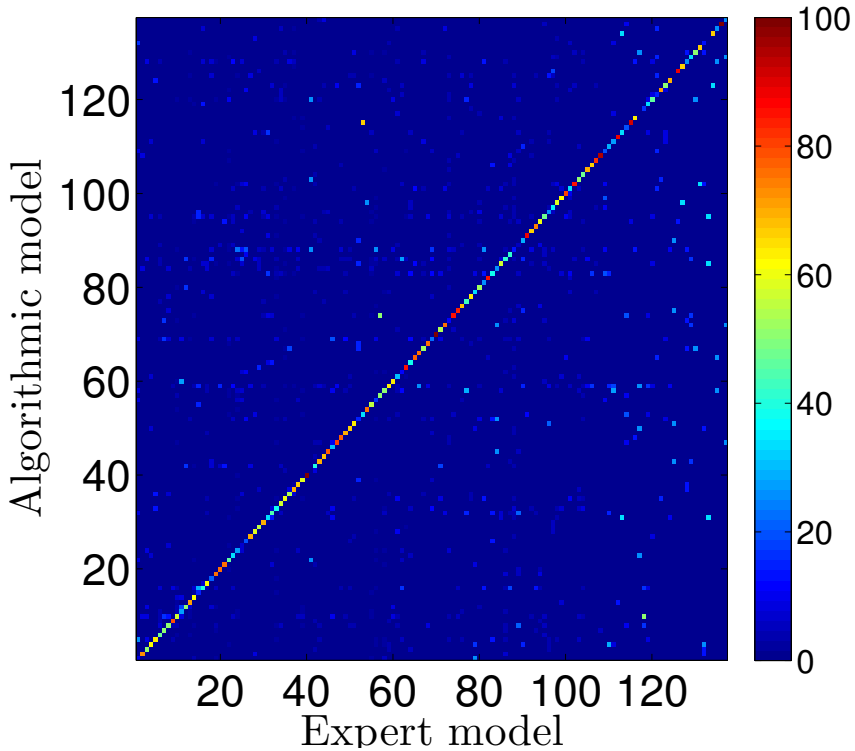


Algorithmic clustering

Inconsistencies between models, middling penalties



Area level



Stream level

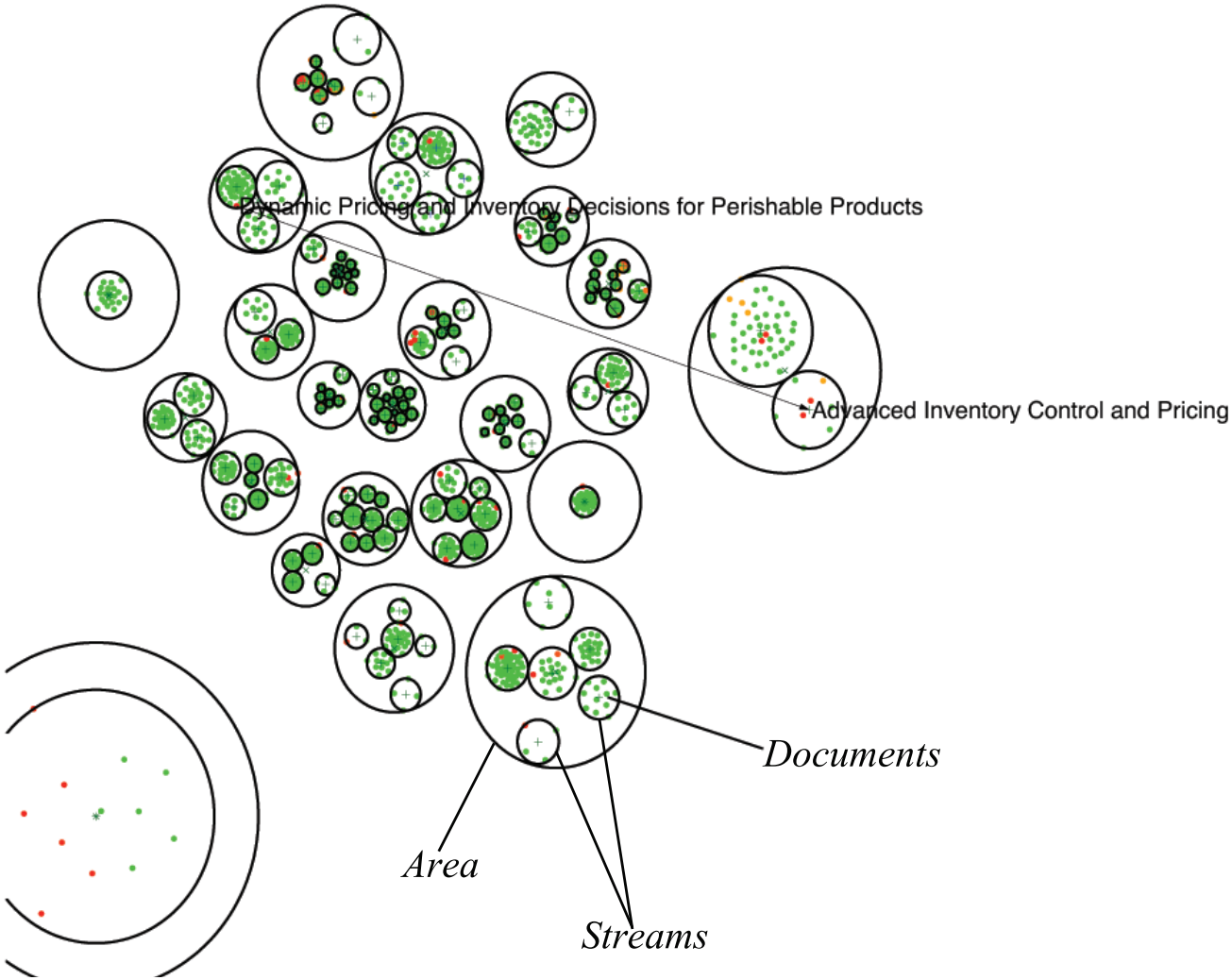
Degree of the document inconsistency

For each document the degree of inconsistency between the expert and the algorithmic model is determined by

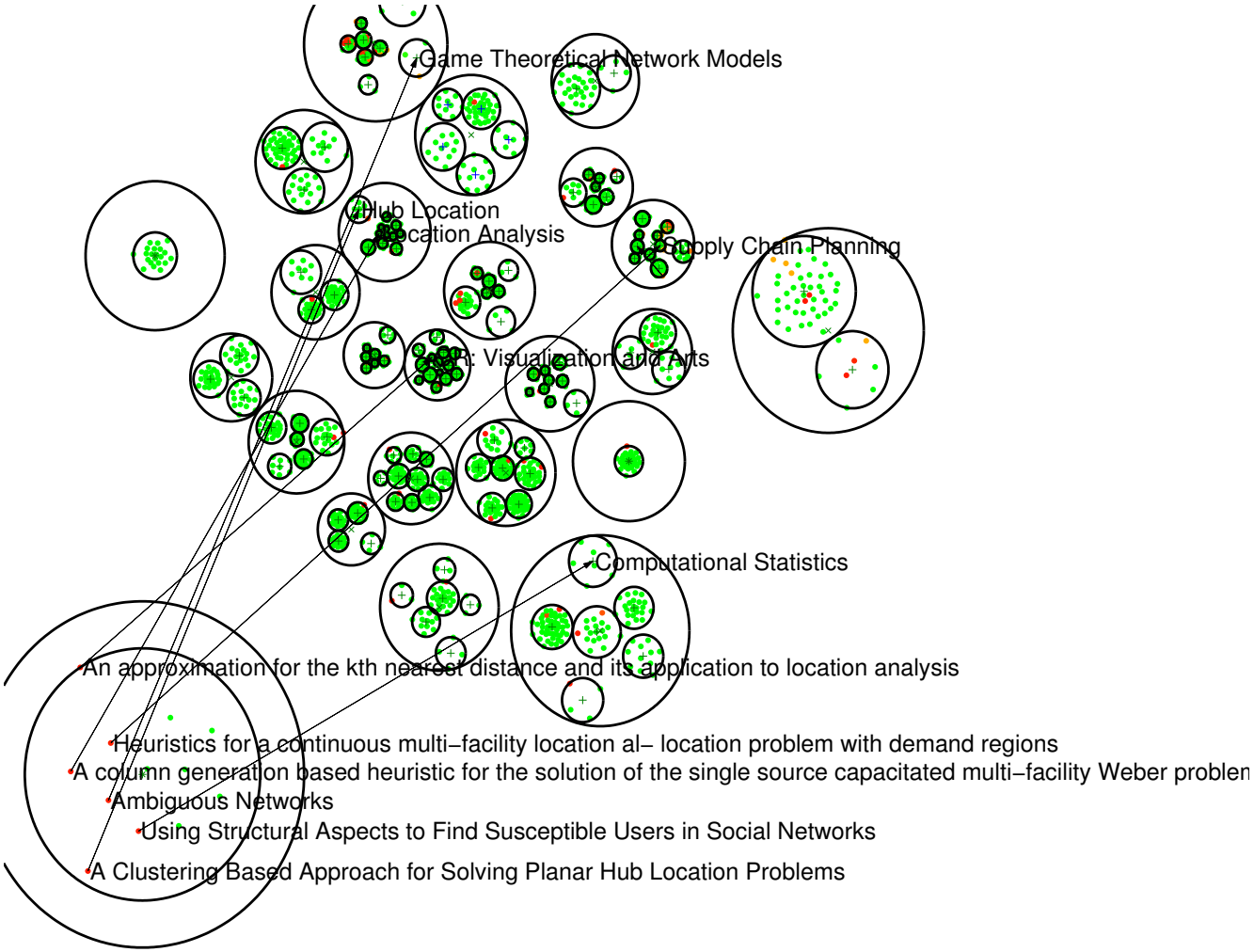
1. the level number, where the models differ,
2. the distance between the expert and the algorithmic cluster on this level.

Map the range of inconsistency to the color scale:

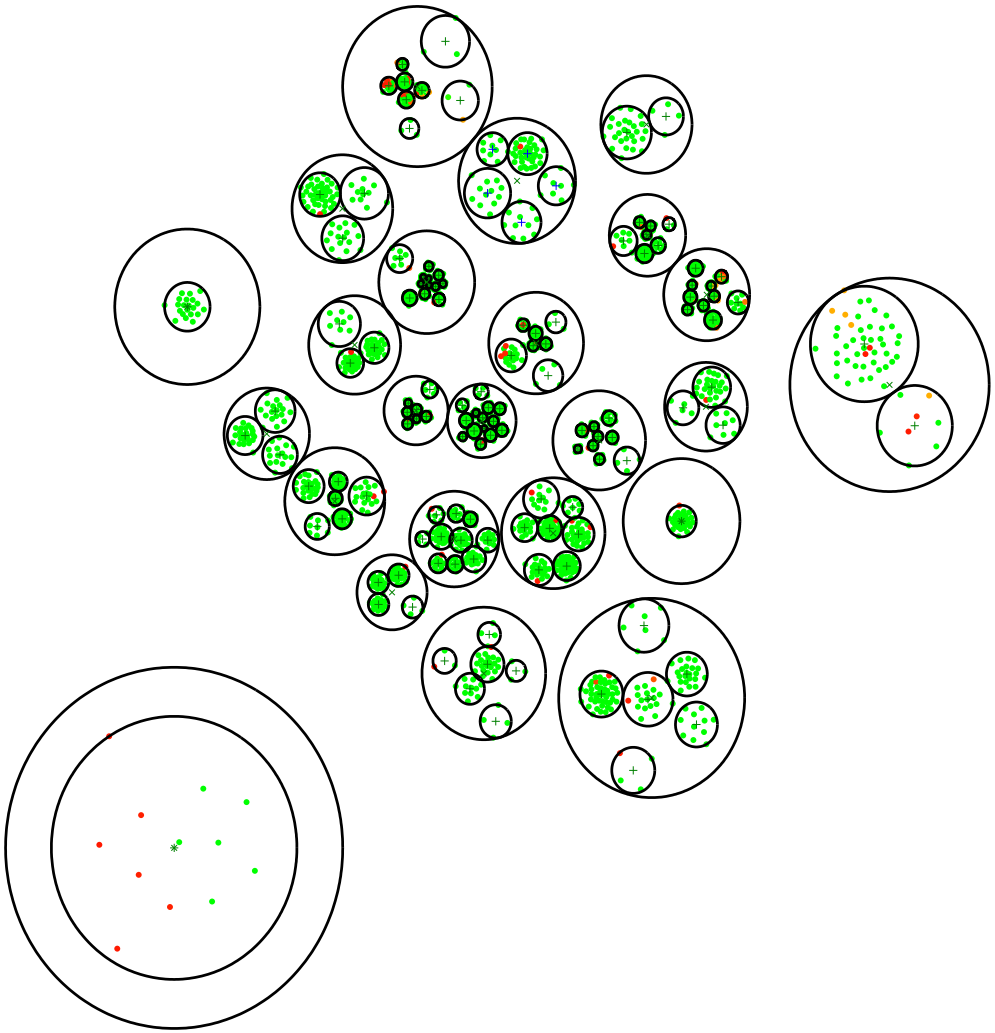
Visualisation of inconsistencieses, $\gamma = 1.25$



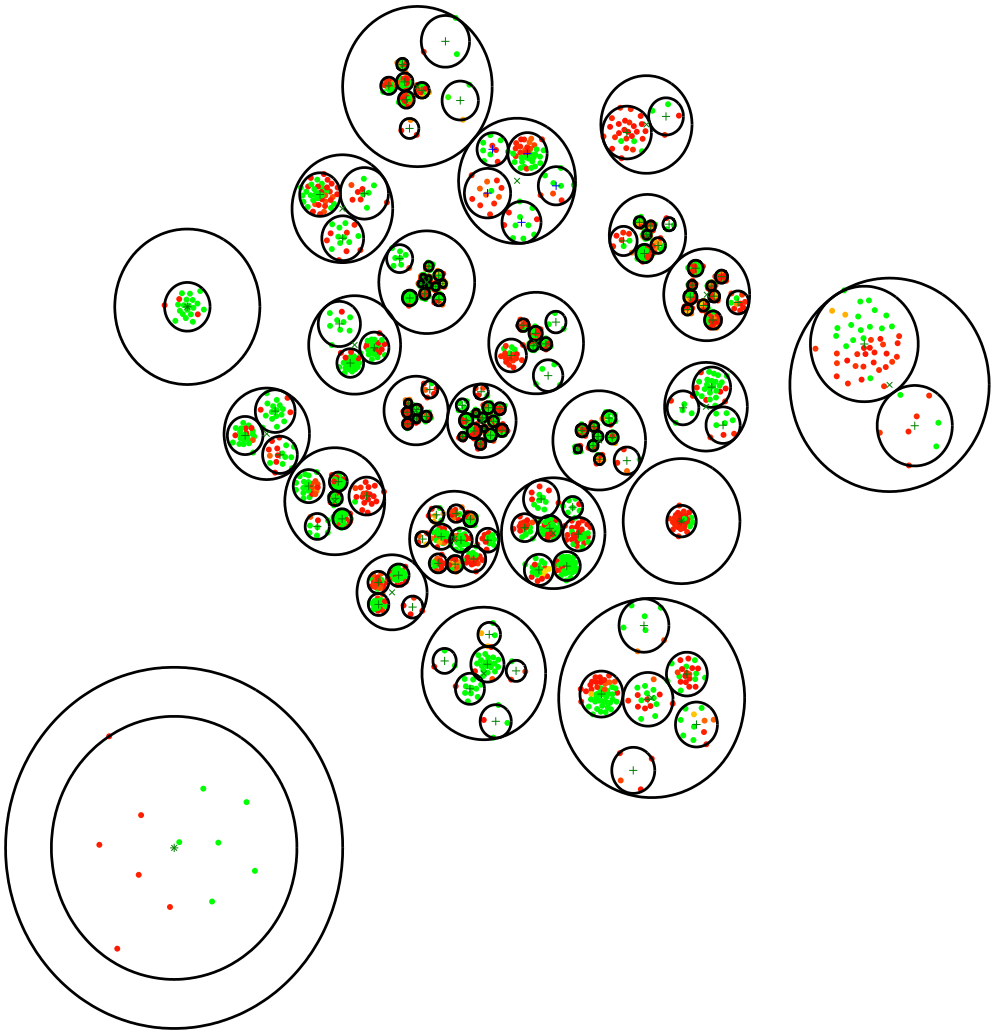
Visualisation of inconsistencieses, $\gamma = 1.25$



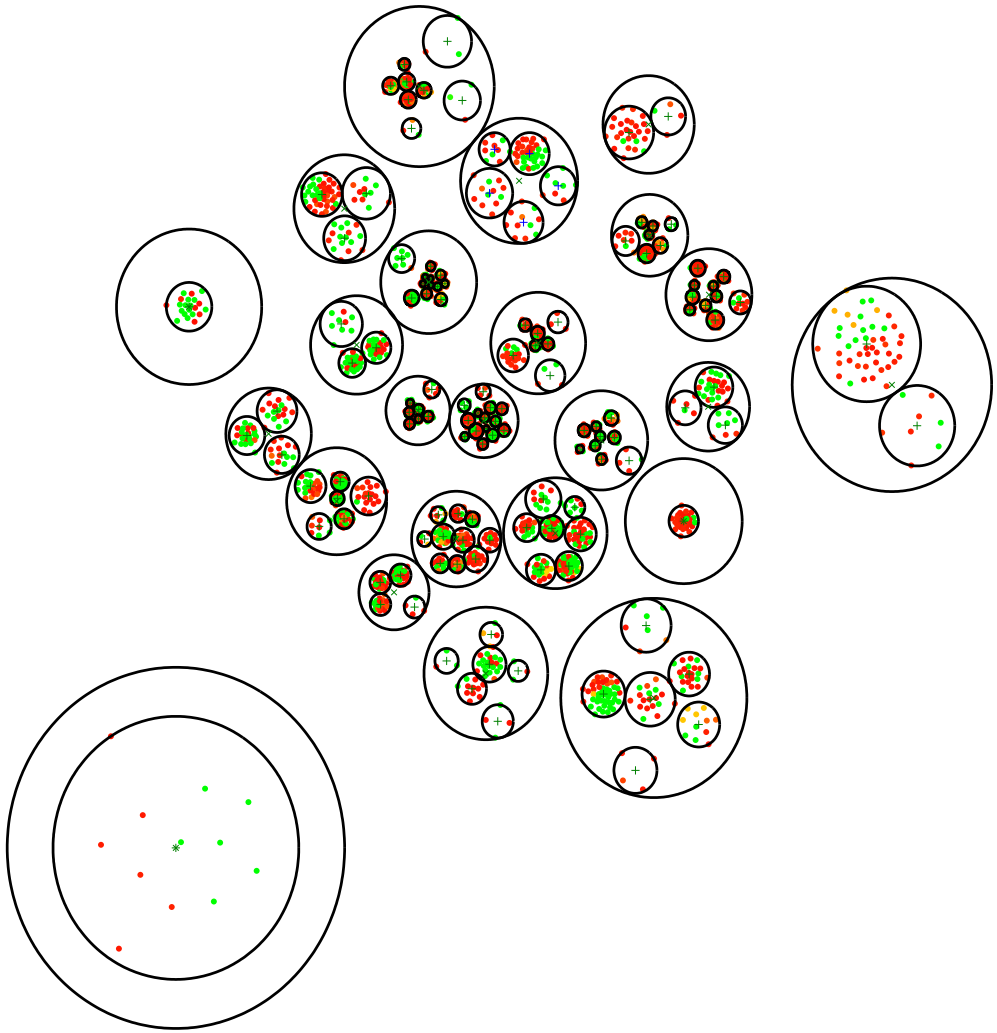
Visualisation of inconsistencieses, $\gamma = 1.25$



Visualisation of inconsistencieses, $\gamma = 0.7$



Visualisation of inconsistencieses, $\gamma = 0.5$



Conclusion

- The similarity function between clusters is proposed.
- A method for thematic hierarchical model creation is proposed.
- The way of visualising a huge thematic model on the plain is proposed.
- The way of visualising inconsistencieses between the expert and the algorithmic model is proposed.