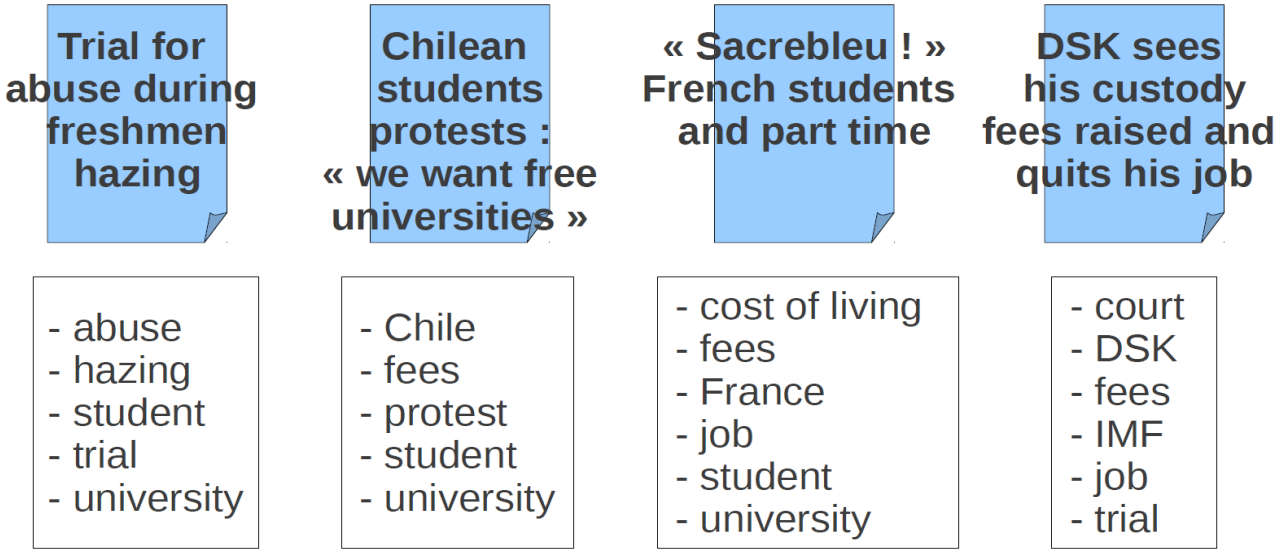
Sunbelt, May, 28th 2013, Hamburg

Document Corpus Analysis based on Term Entanglement

Benjamin Renoust*,**
Guy Melançon*
Marie-Luce Viaud**
*LaBRI, Université de Bordeaux I, France
**Institut National de l'Audiovisuel, France

Document Visualization and Analysis

Indexed news Corpus from the National Institute for Audiovisual*



Quality of the index (terms given from a thesaurus)
→ *topics / concepts / named entities / controlled vocabulary*

**www.ina.fr*

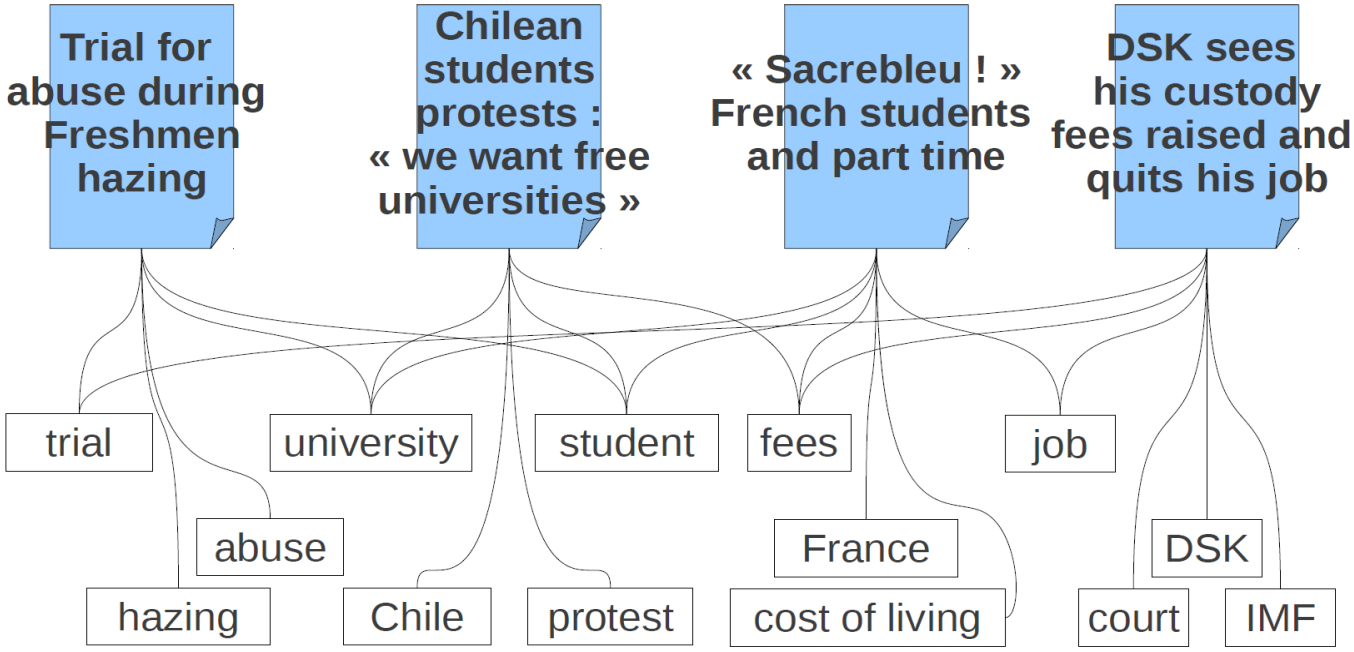
Corpus Analysis

- Query databases
- Group documents according to content into semantic units (*terms, topics, concepts, …*)
- Visually organize documents

- Provide feedback on group cohesion
  - → Grouping suffers from linguistic ambiguities (homonymy, polysemy...)
  - → Geometric distances do not nessarily reflect semantic proximities

Document Visualization

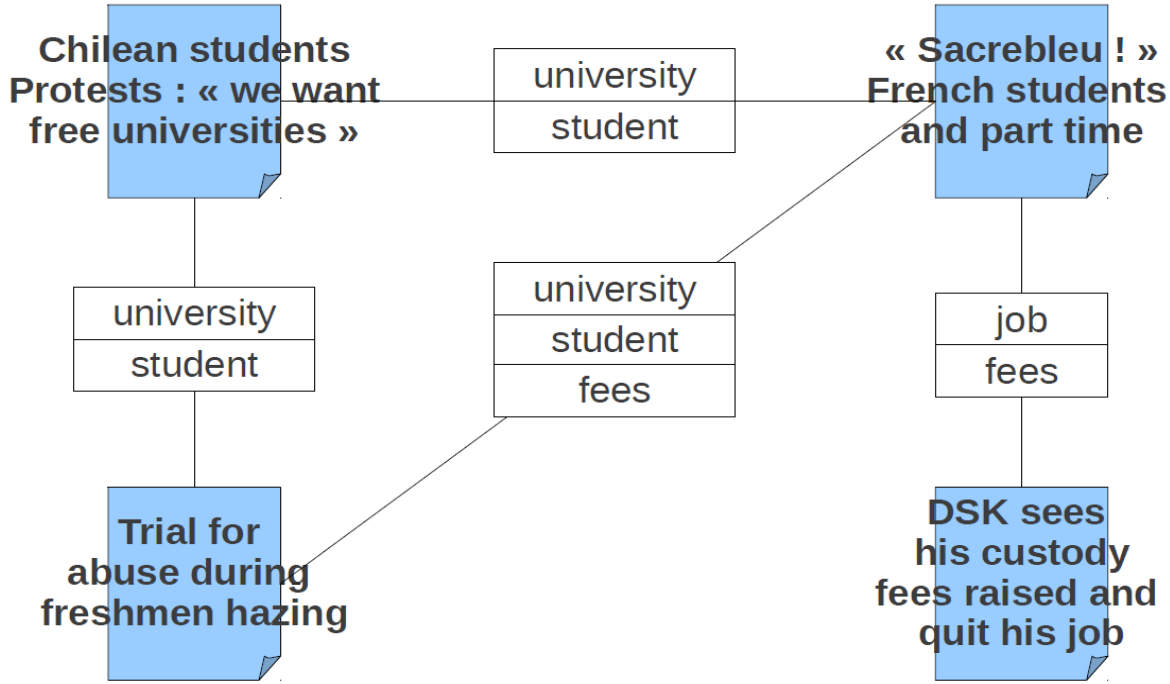
Model data by graphs, bipartite graphs (2-mode)



Pajek, Bataglej et al. 2001, NetLens, Kang et al. 2006, 3D-SE, Usui et al. 2007, Table-based Viz, Schulz et al. 2008 ...

Multiple Content Relations

Project bipartite graph onto a *Document-Document* graph
→ Terms are turned into edge types



(Document-Document projection, limited to multiple content relations, no loop)

Working with the Graphs

Most approaches focus on :

- Turn index into weights
- Work around the *Document-Document* (or *Term-Term*) *Graph*, and exploit its topology
- Improve on clustering methods
- Topic detection and tracking, topic ranking

Allan 2002, Leskovec 2009, Eler, Paulovich, et al. (2009), Wei et al. (2010)…

Working with Groups

Cohesion of a group

→ When all documents group around common concepts (*terms*)

→ When involved concepts uniformely overlap with one another

Measure

→ How much concepts overlap between documents

→ How much/well concepts interact through documents

Ambiguity in a Group

Ambiguity in Relation Content

→ Borrow idea from Burt & Schott
→ Originally aimed at clarifying *ambiguity* in (social) relation content

Relations : *friendship, advice, intimacy, …*

(“*ad hoc* distinctions between relations' content increase likelihood of equivocal research conclusions”)

Burt, Schott (1985), Relation Content in Multiple Networks

Semantic entanglement

With documents, *ambiguity* is turned into *semantic entanglement*

→ Measure how much a concept embraces the scope of a document group

→ Look at how it relates with concepts having a similar behavior

Measuring entanglement : Co-occurrences matrix

Edges carry terms $s, t, \ldots \in T$

N total number of edges (in projected graph)

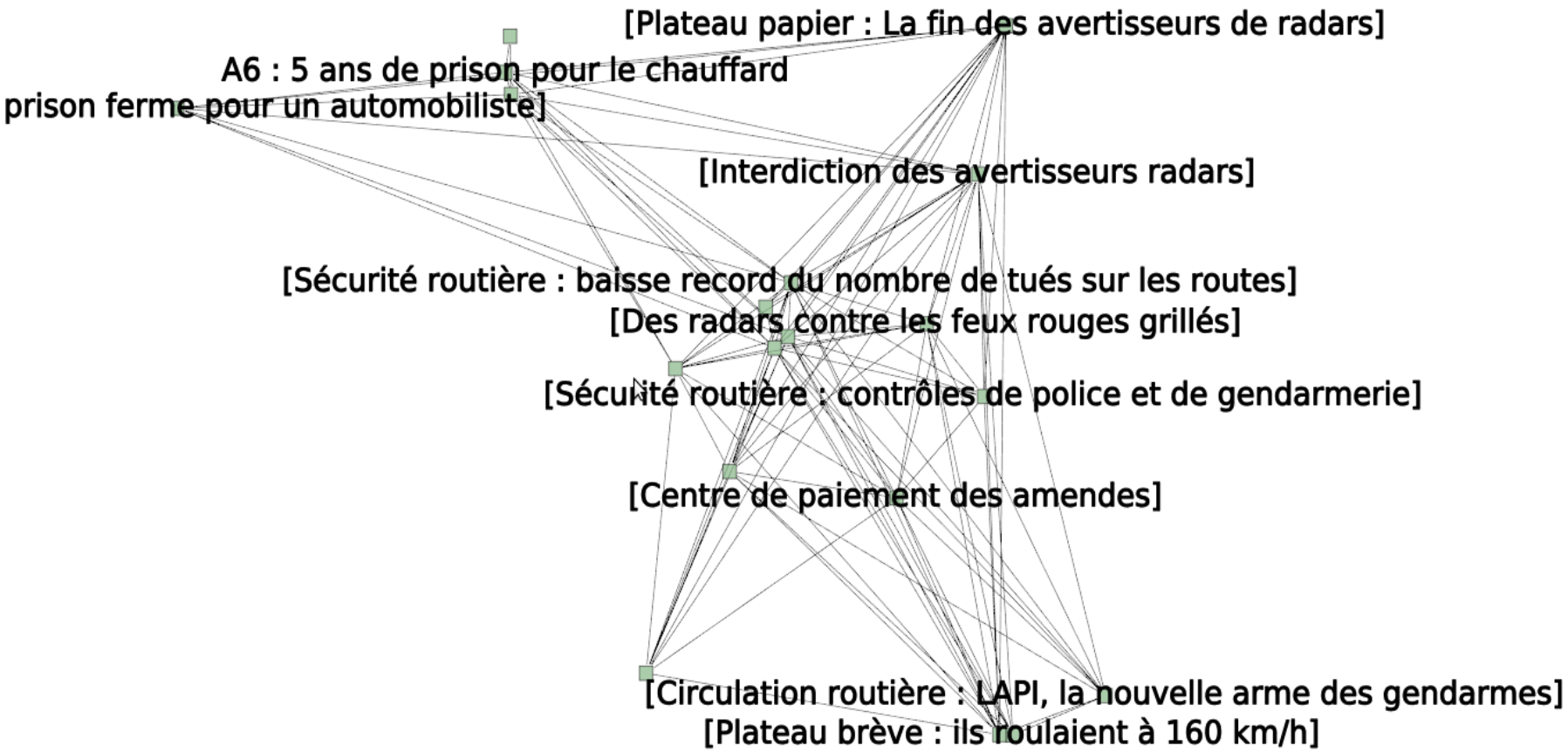
n_s number of edges of term s

$n_{s,t}$ number of edges of term s and t

$$\begin{bmatrix} n_s & n_{s,t} & \cdots \\ n_{t,s} & n_t & \\ \vdots & & \ddots \end{bmatrix}_{T \times T}$$

Measuring entanglement : Co-occurrences matrix

(Document-Document graph)

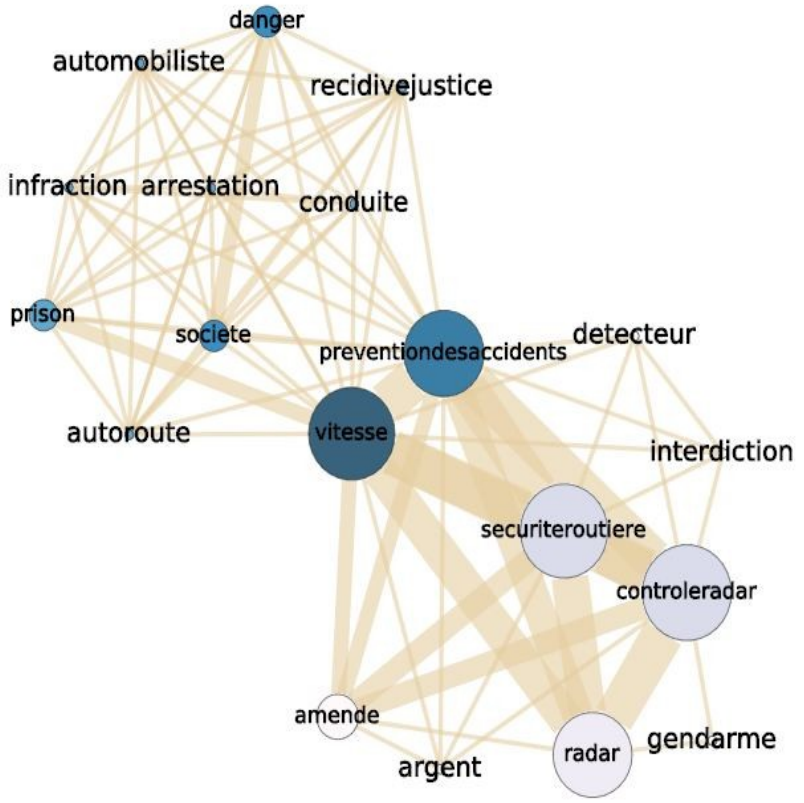


19 documents / 105 links

Measuring entanglement : Co-occurrences matrix

Resulting matrix and term-interaction network
Node and edge size correspond to (co)occurrence

1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0
1	3	1	1	3	1	1	1	1	1	1	0	0	0	0	0	0	0	0
1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0
1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0
1	3	1	1	3	1	1	1	1	1	1	0	0	0	0	0	0	0	0
1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0
1	1	1	1	1	1	3	1	1	1	3	0	0	0	0	0	0	0	0
1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0
1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0
1	1	1	1	1	1	1	1	1	48	41	15	35	35	1	1	0	3	1
1	1	1	1	1	1	3	1	1	41	67	21	42	46	1	1	0	3	1
0	0	0	0	0	0	0	0	0	15	21	45	45	28	0	0	1	1	0
0	0	0	0	0	0	0	0	0	35	42	45	78	61	1	1	1	6	1
0	0	0	0	0	0	0	0	0	35	46	28	61	71	1	1	0	6	1
0	0	0	0	0	0	0	0	0	1	1	0	1	1	1	1	0	0	0
0	0	0	0	0	0	0	0	0	1	1	0	1	1	1	1	0	0	0
0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	1	0	0
0	0	0	0	0	0	0	0	0	3	3	1	6	6	0	0	0	6	1
0	0	0	0	0	0	0	0	0	1	1	0	1	1	0	0	0	1	1



Measuring entanglement : Interaction matrix

Edges carry terms $s, t, \ldots \in T$

c_s fraction of edges of term s, $c_s = n_s / N$

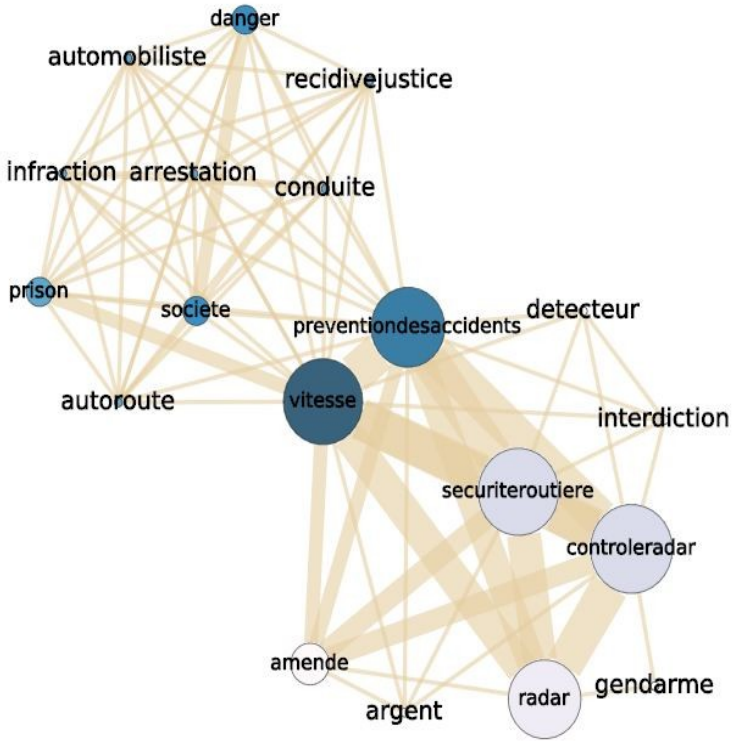
$c_{s,t}$ fraction of edges of term t given they are of type s (cond. freq.), $c_{s,t} = n_{s,t} / n_s$

$$\begin{bmatrix} c_s & c_{s,t} & \cdots \\ c_{t,s} & c_t & \\ \vdots & & \ddots \end{bmatrix}_{T \times T}$$

Note: non symmetric matrix

Measuring entanglement : Interaction matrix

0,01	1	1	0	0	0	1	1	0,33	0,01	0,02	0,33	0,33	0	0	0	0	0	1
1	0,01	1	0	0	0	1	1	0,33	0,01	0,02	0,33	0,33	0	0	0	0	0	1
1	1	0,01	0	0	0	1	1	0,33	0,01	0,02	0,33	0,33	0	0	0	0	0	1
1	1	1	0,01	0	0	1	1	1	0,01	0,02	0,33	0,33	0	0	0	0	0	1
0	0	0	0	0,01	0	0	0	0	0	0	0	0	0	0,01	0,02	0	0	0
0	0	0	0	0	0,01	0	0	0	0,01	0	0	1	0	1	0	0,02	0	0
1	1	1	0	0	0	0,01	1	0,33	0,01	0,02	0,33	0,33	0	0	0	0	0	1
1	1	1	0	0	0	1	0,01	0,33	0,01	0,02	0,33	0,33	0	0	0	0	0	1
1	1	1	0	0	0	1	1	0,03	0,01	0,02	0,33	1	0	0	0	0	0	1
1	1	1	1	0	1	1	1	0,33	0,64	0,85	1	0,33	0,65	0,54	0,47	1	0,5	1
1	1	1	0	0	0	1	1	0,33	0,61	0,46	0,33	0,33	0,49	0,45	0,33	1	0,5	0
1	1	1	0	0	0	1	1	0,33	0,04	0,02	0,03	0	0	0	0	0	0	1
0	0	0	0	0	1	0	0	0	0,04	1	0	0,03	0	0,08	0,02	0,06	0	0,08
0	0	0	0	0	0	0	0	0	0,69	0,73	0	0	0,01	0,78	0,62	1	1	0
0	0	0	0	1	0,74	0	0	0	0,63	1	0	1	0	0,01	1	0,73	1	0,86
0	0	0	0	1	0	0	0	0	0,31	0,31	0	0	0,39	0,58	0,43	0	0,17	0
0	0	0	1	0	1	0	0	0	0,01	0,01	0	0	0,68	0,01	0	0,02	0,17	1
0	0	0	0	0	0	0	0	0	0,01	0,02	0	0	0,01	0,01	0	0	0,06	0
1	1	1	1	0	1	1	1	0,33	0,01	0	0,33	0,33	0,01	0	0	0,02	0	0,01



Measuring entanglement : entanglement index

With documents, ambiguity is turned into *semantic entanglement*

- compute an entanglement index for each term (to what extent does a term interwins with all others?)
- let u denote the *highest entanglement index* among all terms
- denote by g_s the *fraction* computing *entanglement index* for terms s
- entanglement index for s is equal to $g_s\, u$

Measuring entanglement : entanglement index

- “term entanglement is re-inforced through interactions with other tangled terms”

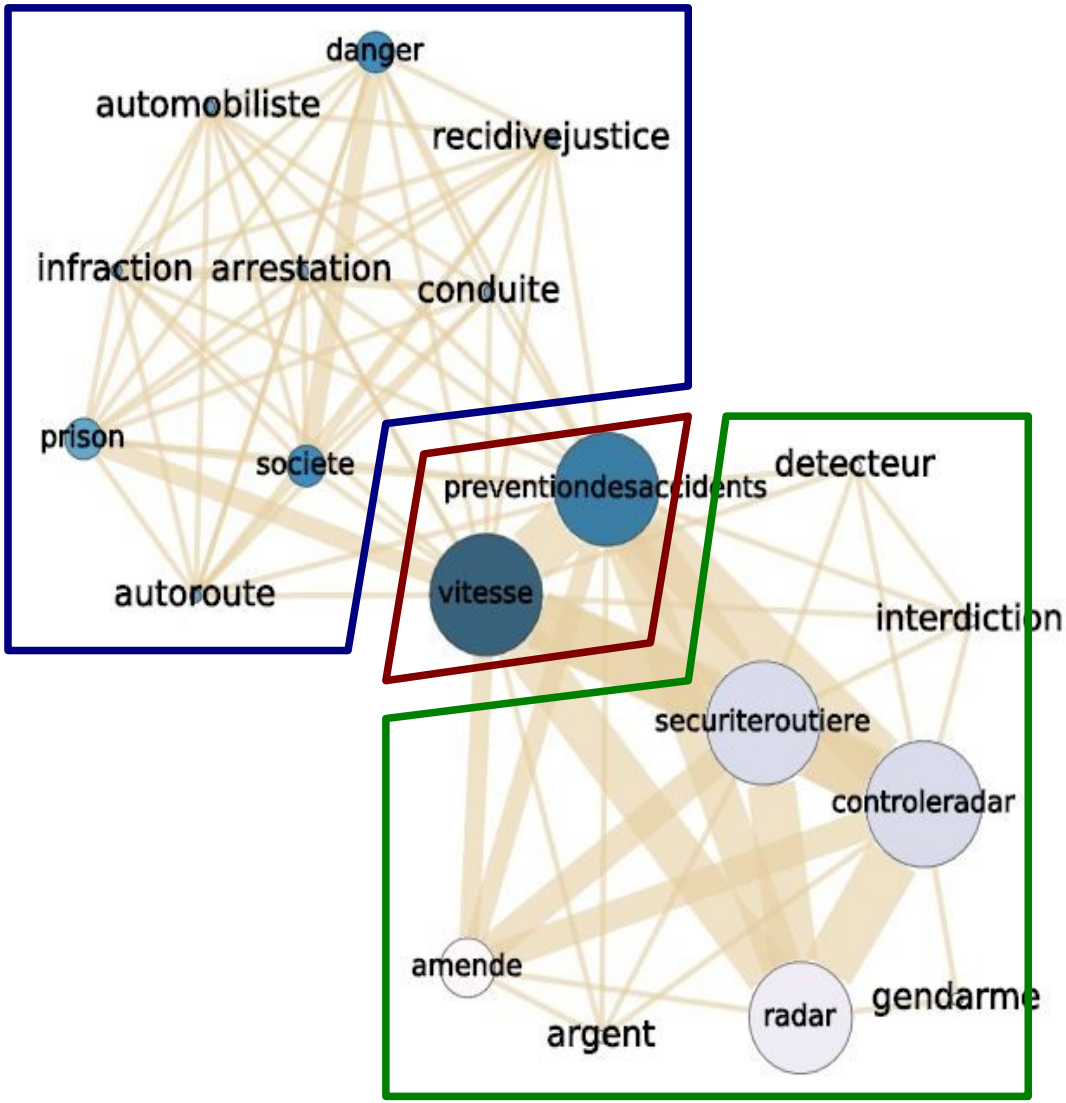
$$g_s u = \sum_t c_{t,s} g_t$$

- g is thus an eigenvector of the interaction matrix

$$gu = C'g \qquad C = \begin{bmatrix} c_s & c_{s,t} & \cdots \\ c_{t,s} & c_t & \\ \vdots & & \ddots \end{bmatrix}$$

Measuring entanglement : entanglement index

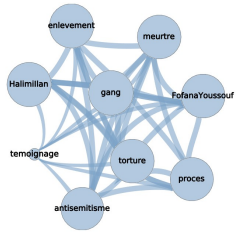
vitesse	0.441454
preventiondesaccidents	0.379679
danger	0.302022
societe	0.302022
prison	0.273354
automobiliste	0.24568
recidivejustice	0.24568
infraction	0.24568
arrestation	0.24568
autoroute	0.24568
conduite	0.24568
controleradar	0.137461
securiteroutiere	0.134463
radar	0.0671915
amende	0.0113085
interdiction	0.00345846
detecteur	0.00345846
argent	0.00320664
gendarme	0.00052092



So *what* is a cohesive group?

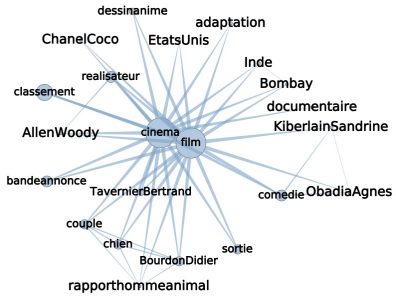
Different entanglement *profiles*

All terms mix equally



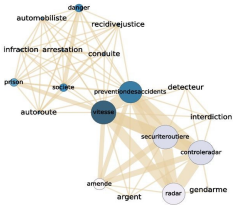
- A clique with the same number of documents edges on which terms co-occur

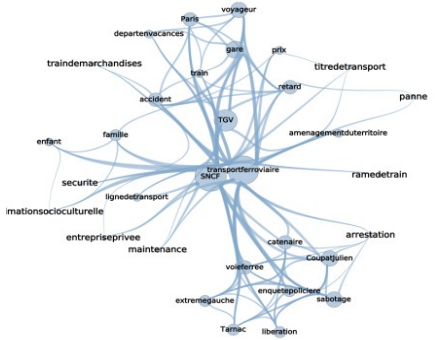
Some terms dominate



- But do co-occur on most edges
- Second order terms occur here and there (but do contribute to the overall cohesion)

In between





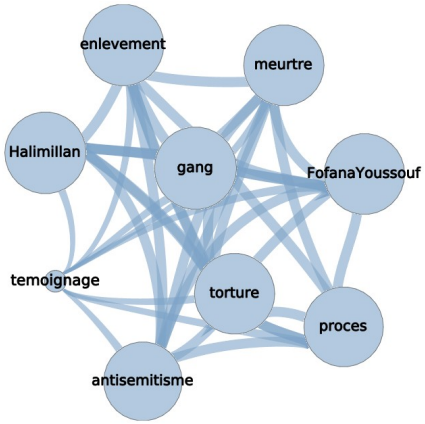
Measuring entanglement : entanglement homogeneity

All terms co-occur equally on the documents (and have the same entanglement index)

The vector g is close to the diagonal

The cosine value $\langle g,1 \rangle$ is high

→ defines the *entanglement homogeneity*



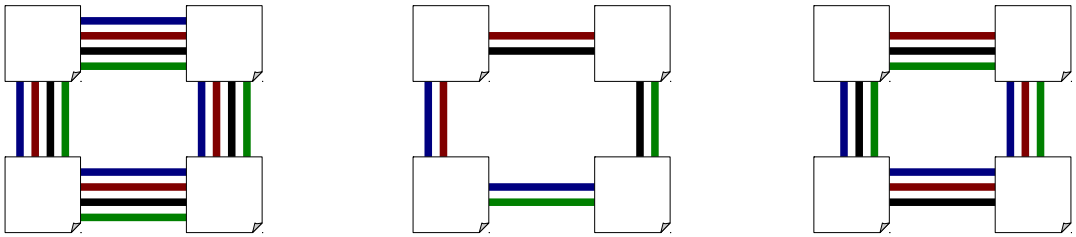
(we could just as well consider the Shannon entropy, for instance)

Measuring entanglement : entanglement homogeneity

The cosine value (*homogeneity*) $\langle g,1 \rangle \approx 1$

You need to take the number of terms into account (dimension of the entanglement space)

Many identical configurations :



Measuring entanglement : entanglement intensity

All terms co-occur on the maximum number of documents (all entries of the interaction matrix equal 1, the interaction graph is a clique)

non-negative matrices theory

g is the Perron vector of matrix C

we know $\min_s \sum_t c_{s,t} \leq \lambda \leq \max_s \sum_t c_{s,t}$

so $\lambda \leq |T|$

then $\lambda / |T| \in [0,1]$

defines a *group entanglement intensity*

How do you measure entanglement in a group?

When λ is maximum :

we then have:

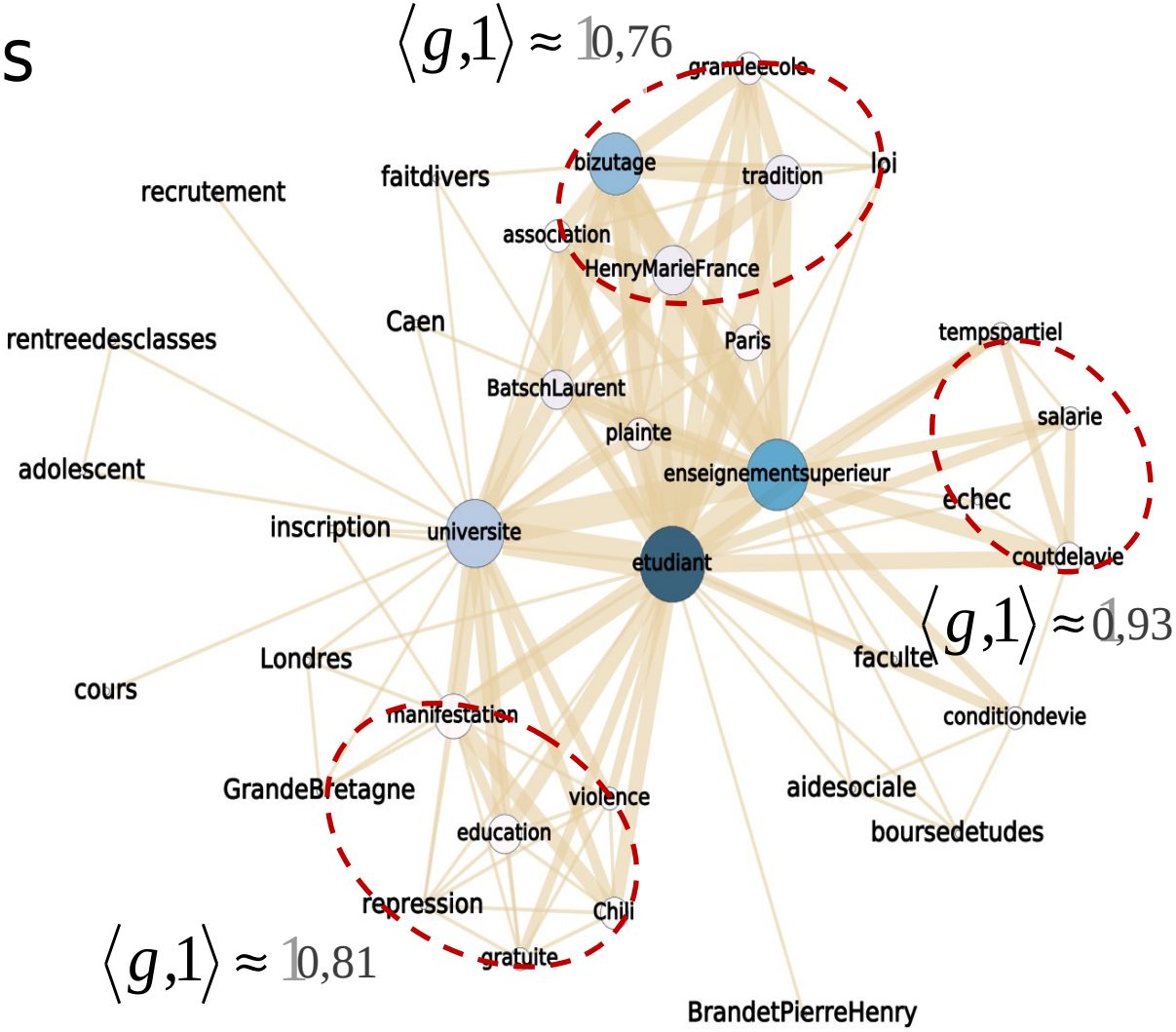
$$C = \begin{bmatrix} 1 & 1 & \cdots \\ 1 & 1 & \\ \vdots & & \ddots \end{bmatrix} \qquad g = \frac{1}{\sqrt{|T|}}(1,\ldots,1)$$

$$\lambda = |T| \qquad \langle g, 1 \rangle = 1$$

Maximum intensity and homogeneity

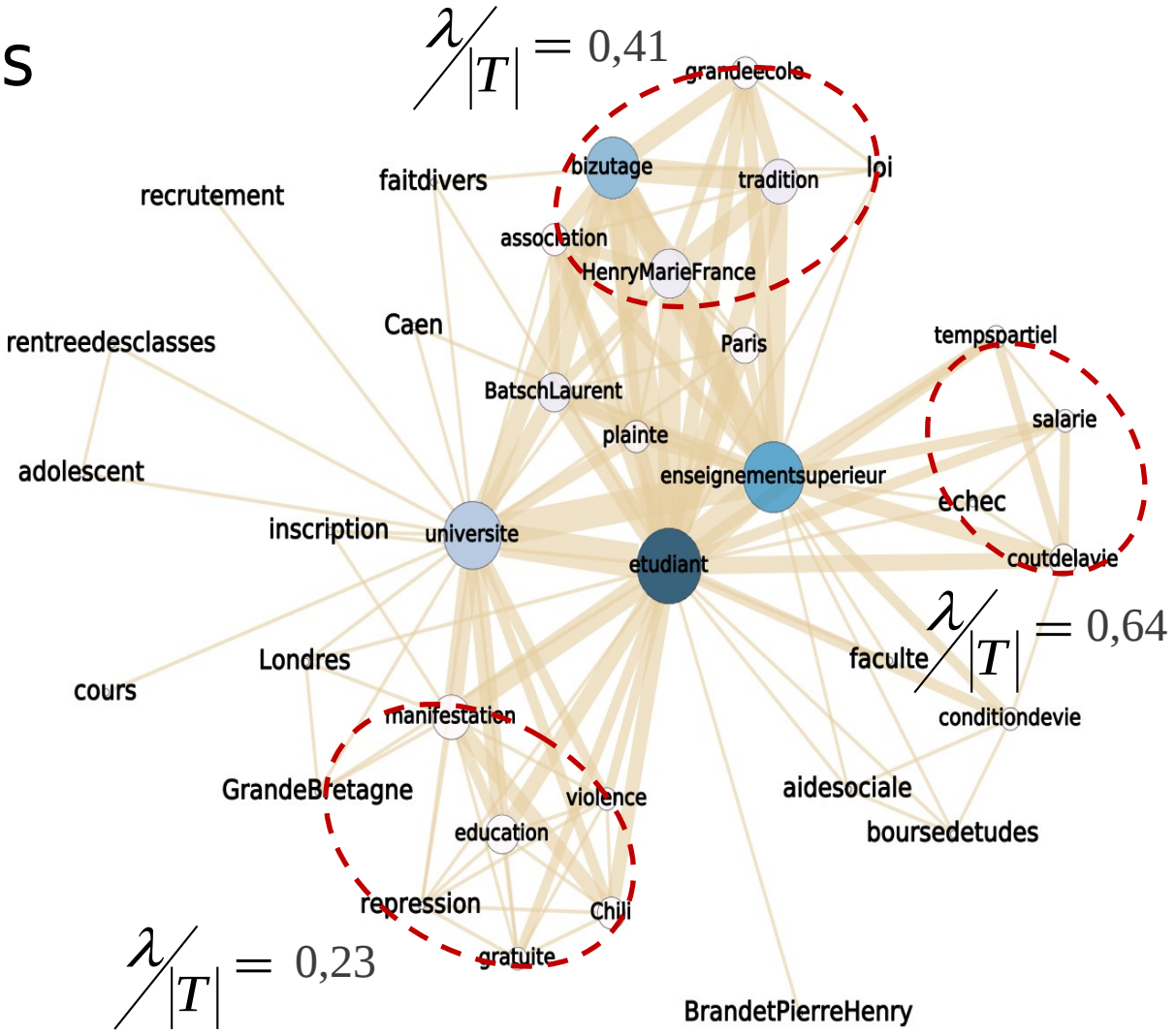
Tangled subregions

Local ‘zones’ means less global entanglement homogeneity but larger local homogeneity



Tangled subregions

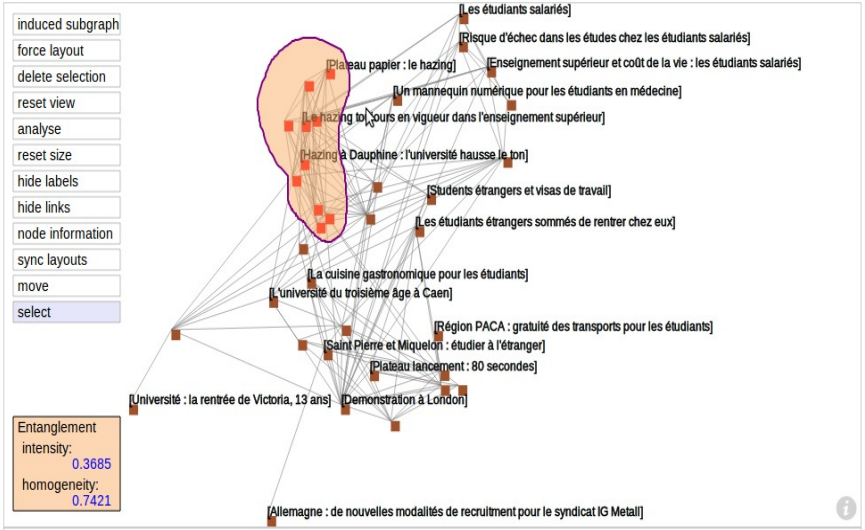
Local ‘zones’ means less global entanglement intensity but larger local intensity

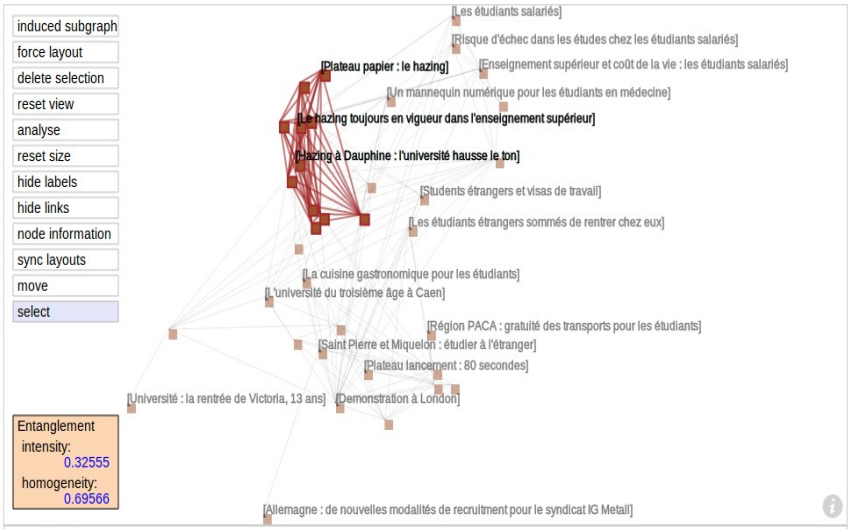


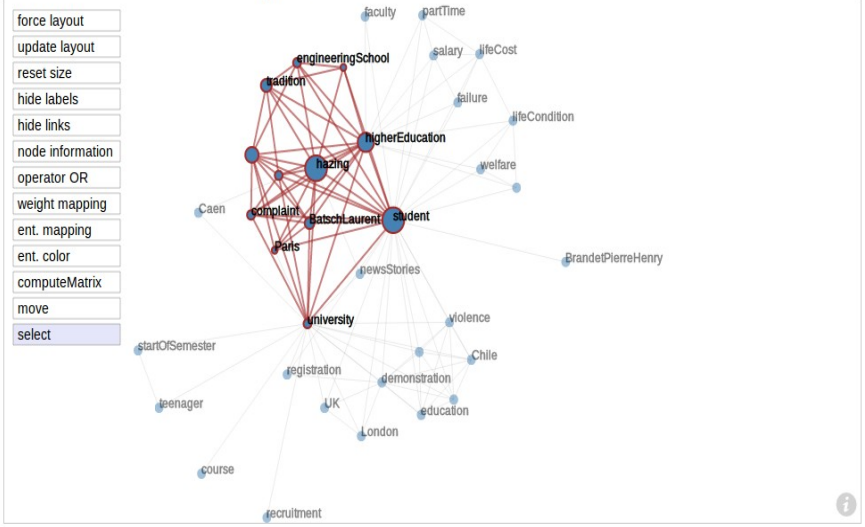
$\lambda/|T| = 0{,}09$

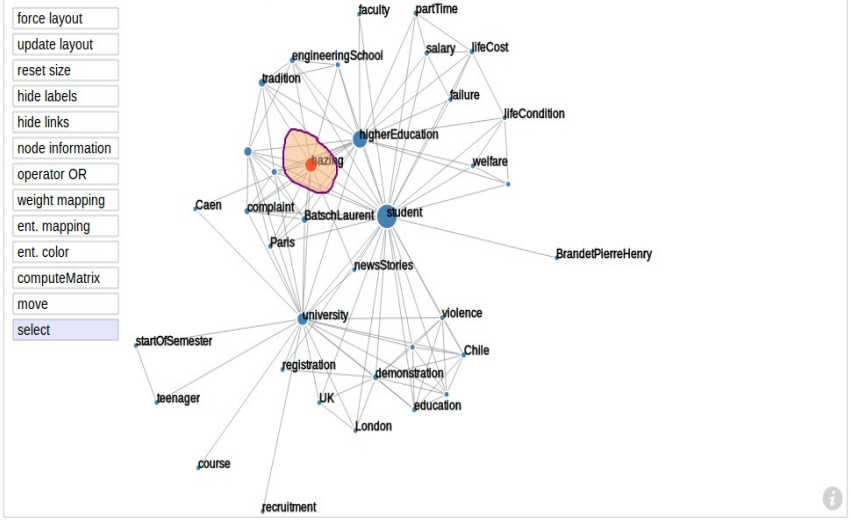
$\lambda/|T| = 0{,}23$

Entanglement embedded in interaction



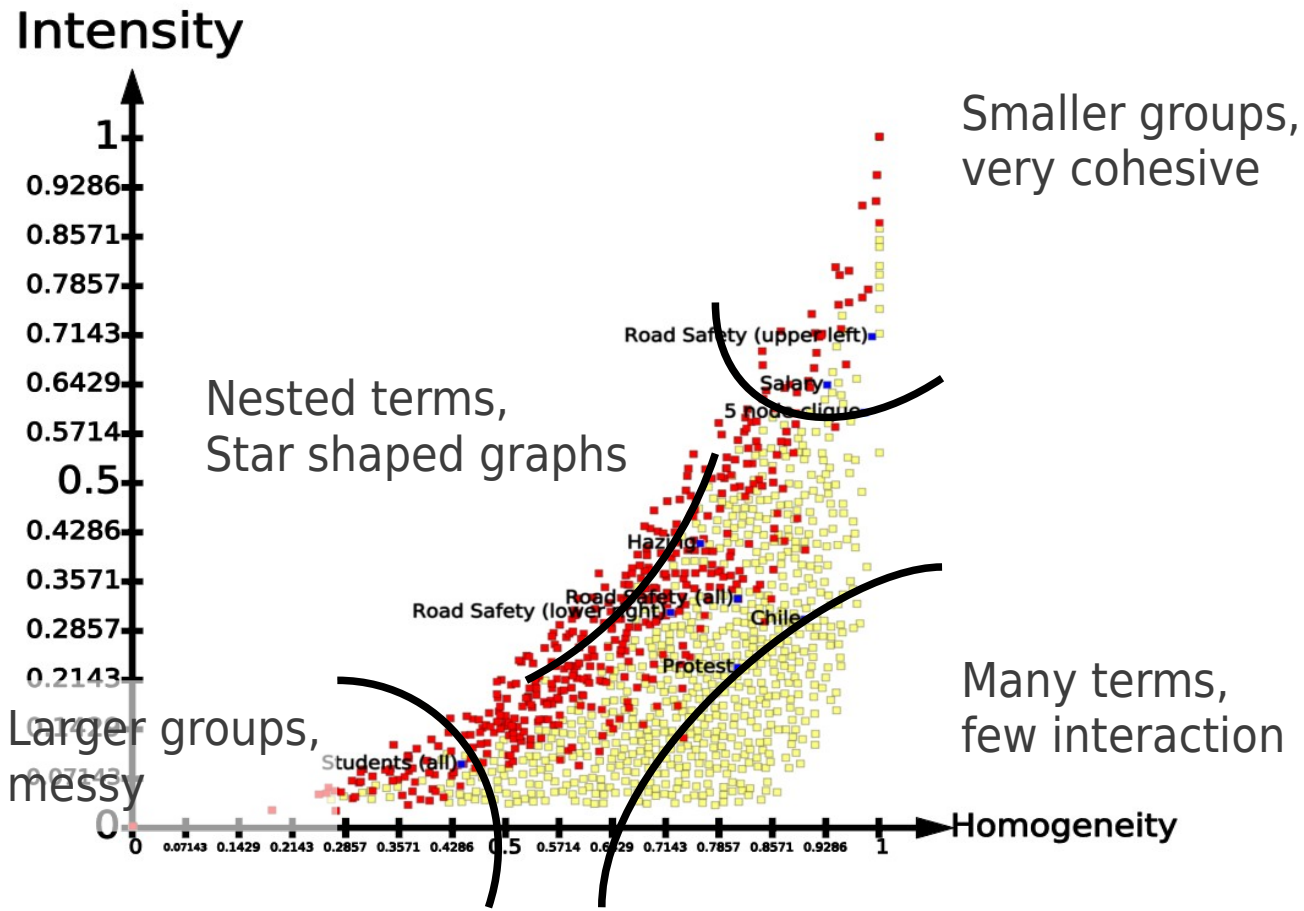






Entanglement profiles

4 entanglement profiles tendancies depending on entanglement intensity and homogeneity



Conclusions & Future (on-going) work

Hierarchize terms

- Find terms bringing in maximum entanglement
- Find terms fragilizing overall entanglement

Weightenned model
Generalization to multivariate graphs

Embed entanglement calculus into different visualization context

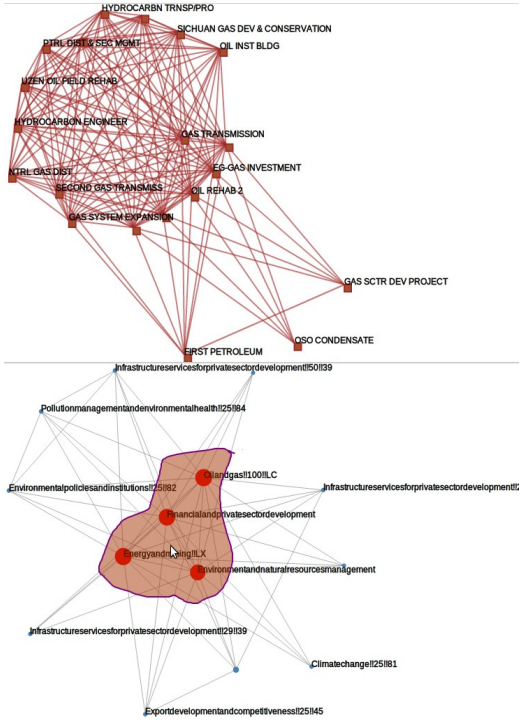
Use entanglement index in order to improve document layout and/or term interaction network layout

Transfer methodology to other areas

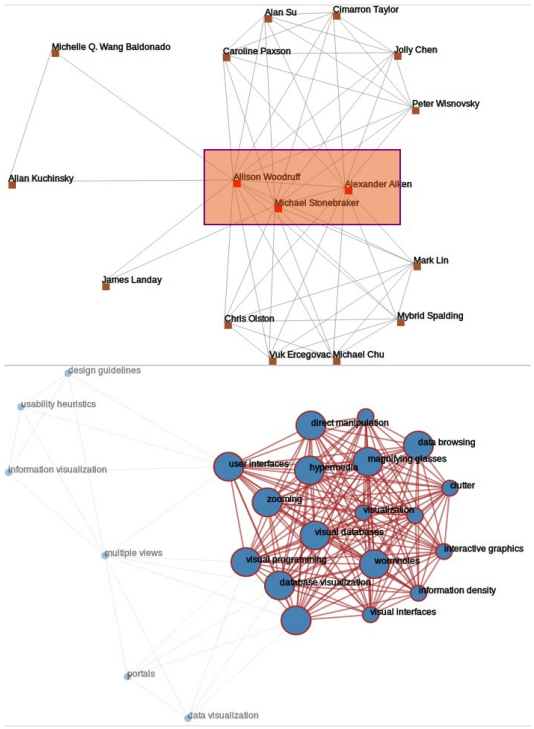
Conclusions & Future (on-going) work

Other applications :

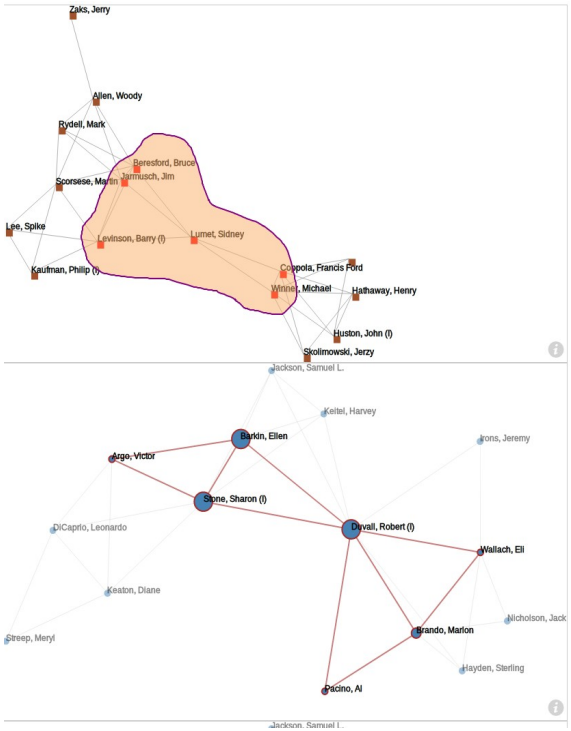
WorldBank projects



InfoVis2004 contest



IMDB



Questions

Benjamin Renoust
Guy Melançon
Marie-Luce Viaud

Université de Bordeaux, France
Institut National de l'Audiovisuel, France

Benjamin.Renoust@labri.fr
Guy.Melancon@labri.fr
mlviaud@ina.fr