# Indirect Bibliometrics by Complex Network Analysis

J. Raimbault[1,2]

[1]Géographie-cités (UMR 8504 CNRS)
[2]LVMT (UMR-T 9403 IFSTTAR)

Cybergeo : 20 ans déjà !
Jeudi 26 mai 2016

# Context

*"You are what you cite"* : **Which disciplines populate the scientific neighborhood of cybergeo ? Are they different from the ones obtained through article content (POC) and declared contents (HC) analysis ?**

→ Important for editorial policy : interdisciplinarity and Open Science

→ Semi-qualitative approach, against purely quantitative bibliometrics harmful to humanities

# Objective

**Research question :** *How does the combination of a citation network approach with a semantic analysis unveil disciplinary context of the journal ?*

→ Hypernetwork methodology : superposition of a citation network with a semantic network, in the spirit of a transversal approach

→ Data difficult to access : database to construct

# Data collection

**Cybergeo data** : journal production base

→ Structuration and Consolidation

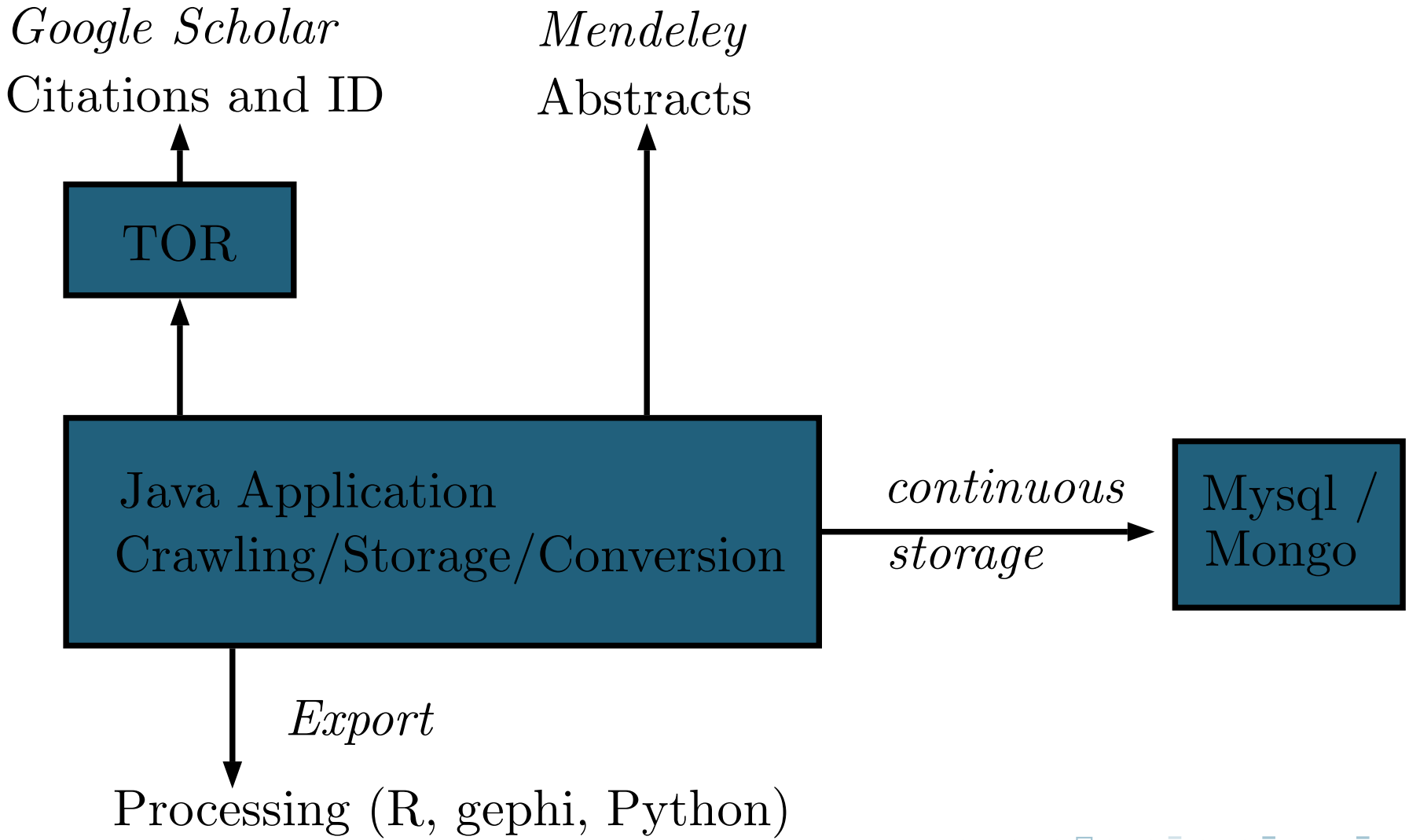
**Citation data** : cybergeo not indexed by "classical" bases (such as Web of Science©, which are furthermore not open)

→ `google scholar` crawling by using "*cited by*" option [Noruzi, 2005]

**Text data** : need abstracts for all linked articles

→ use of Mendeley API [Mendeley, 2015] (free but not open)

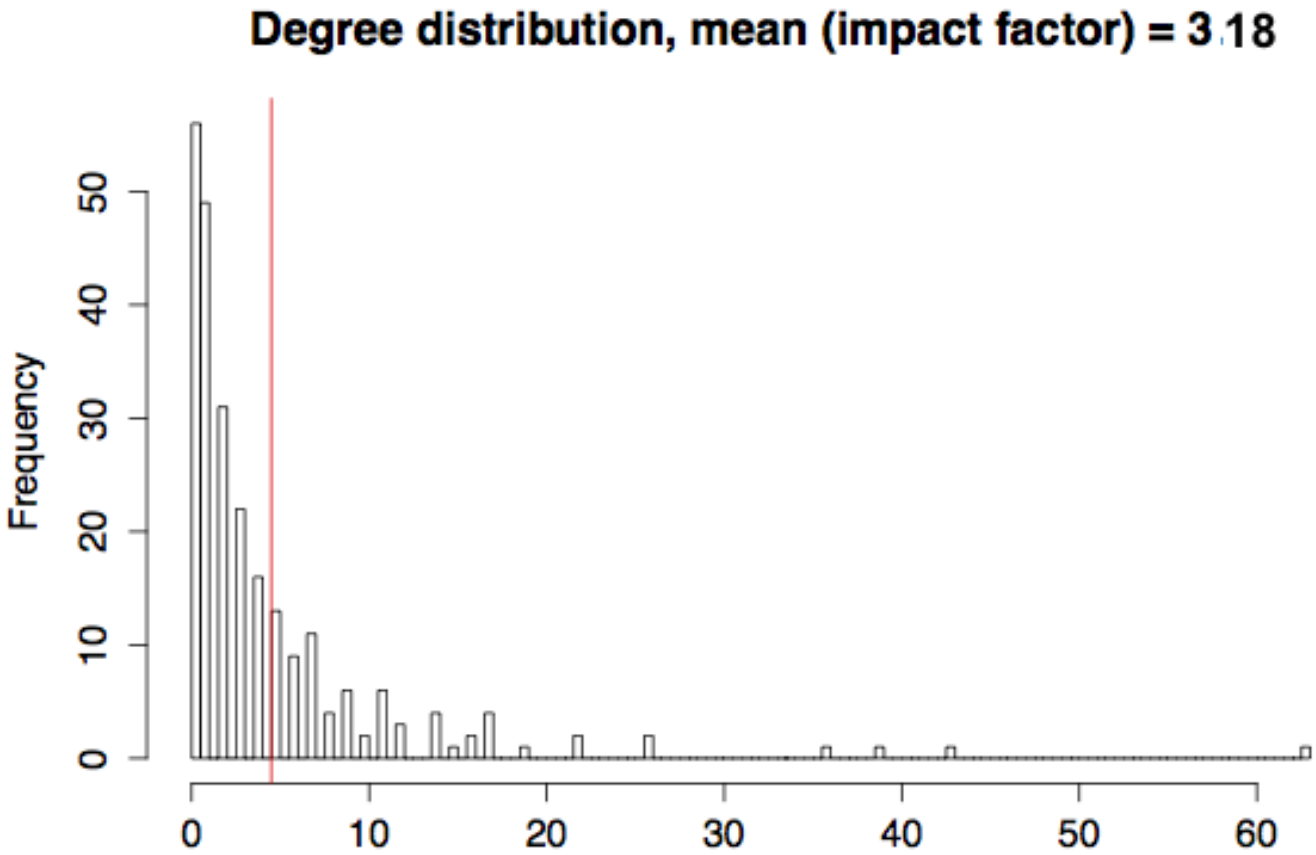
# Data Collection Architecture



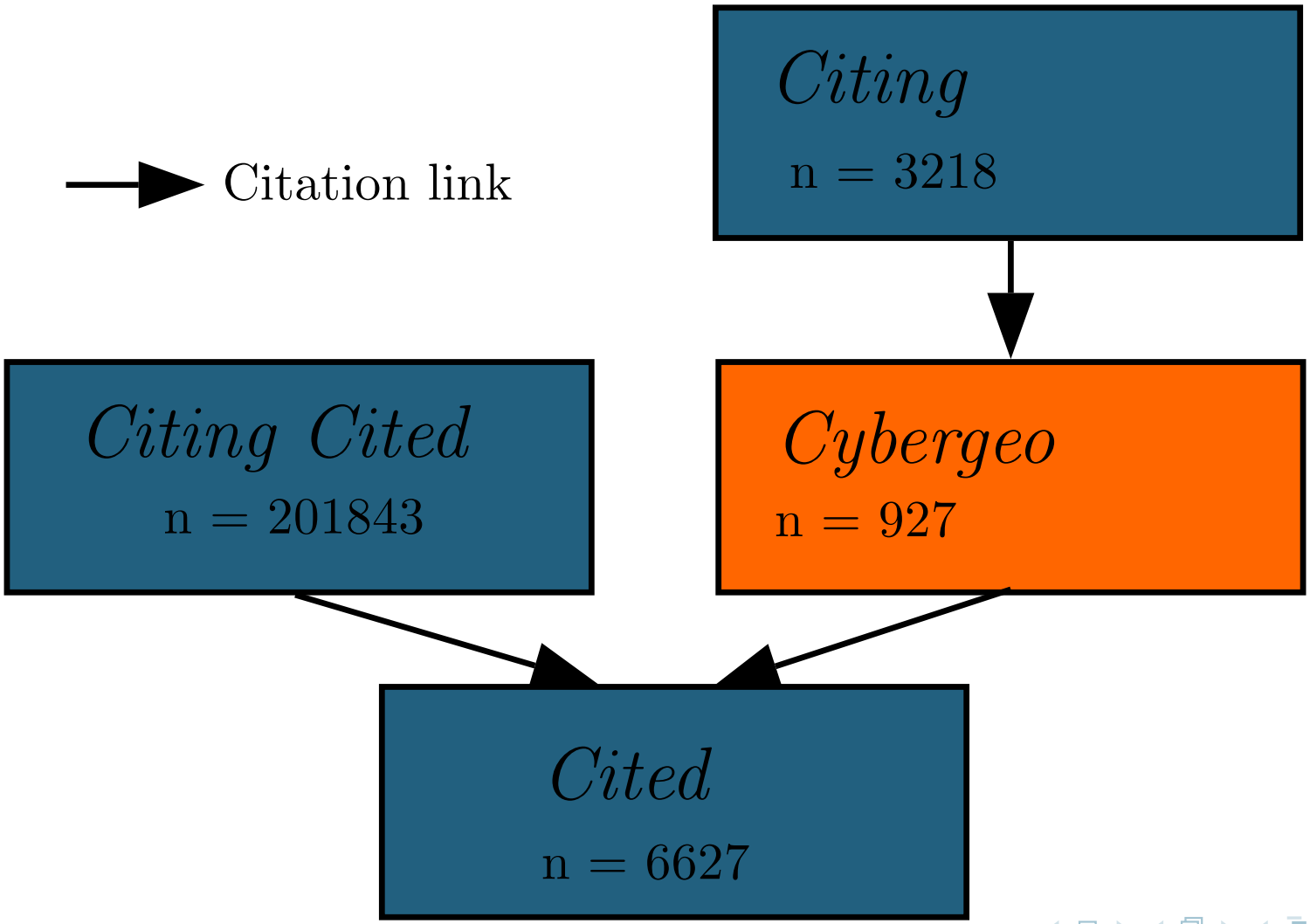
# Network Properties

→ ≃ 947 cybergeo articles can be studied, among ≃ 1200

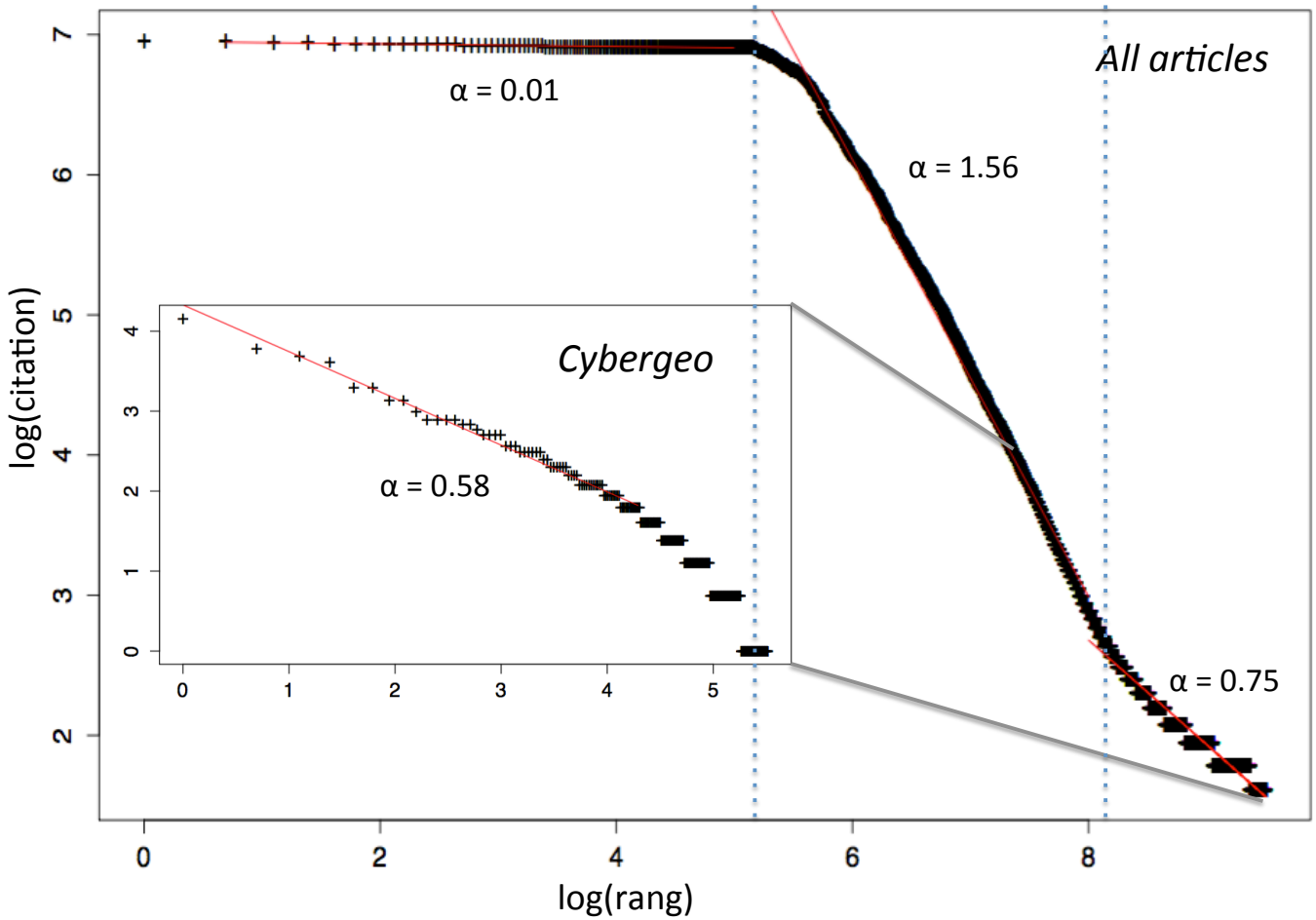
→ *418670 Nodes et 570352 Links ; Diameter : 9 ; Density : 3.25E-6 ; average degree : 2.724284*



# Citation Network Structure

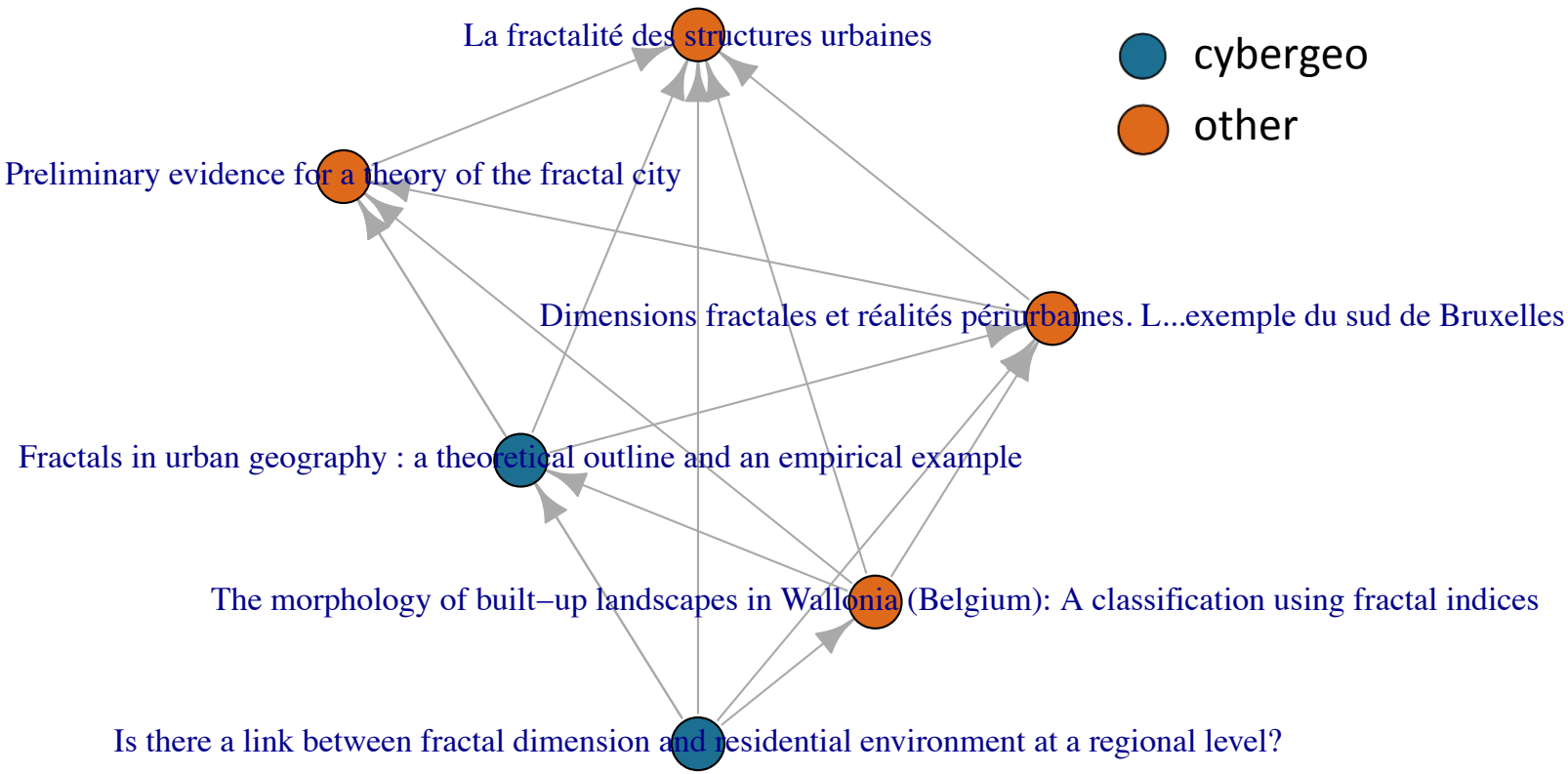


# Hierarchy in citations




*Superposition of different hierarchical regimes*

# Cliques


La fractalité des structures urbaines
Preliminary evidence for a theory of the fractal city
Dimensions fractales et réalités périurbaines. L...exemple du sud de Bruxelles
Fractals in urban geography : a theoretical outline and an empirical example
The morphology of built–up landscapes in Wallonia (Belgium): A classification using fractal indices
Is there a link between fractal dimension and residential environment at a regional level?
cybergeo
other

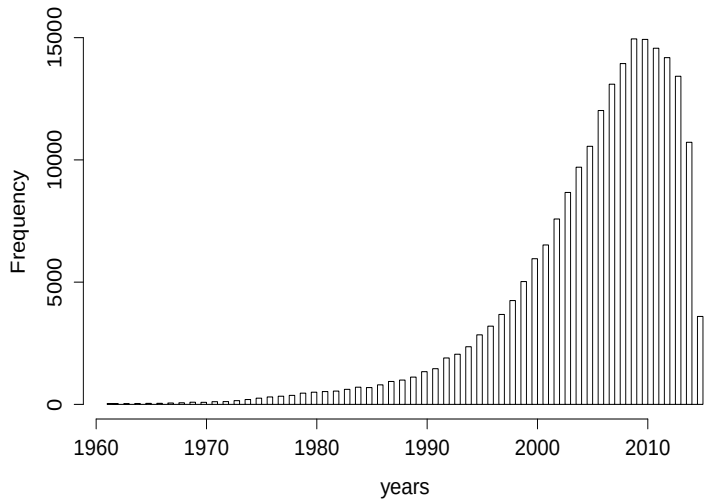
# Semantic Network

**Semantic Data :** Collection of abstract/date/authors/keywords for the 400000 references via Mendeley API $\rightarrow \sim$ 215000 references with full data.

**Summary Statistics**

*Language :* English 206607, French 4109, Spanish 2029, German 892, Portuguese 891, Dutch 124, others 182

*Yearly count*



# Keywords Extraction

*Text-mining in python with* `nltk` *[Bird, 2006]*, method adpated from [Chavalarias and

- Language detection using *stop-words*
- Parsing and tokenizing / pos-tagging (word functions) / stemming done differently depending on language :
  - English : `nltk` built-in pos-tagger, combined to a PorterStemmer
  - French or other : use of `TreeTagger` [Schmid, 1994]
- Selection of potential *n-grams* (with $1 \leq n \leq 4$) : English $\bigcap\{NN \cup VBG \cup JJ\}$ ; French $\bigcap\{NOM \cup ADJ\}$
- Database insertion for instantaneous utilisation (10j $\rightarrow$ 2min)
- Estimation of *n-grams* relevance, following co-occurrences statistical distribution

# Construction of Semantic Network

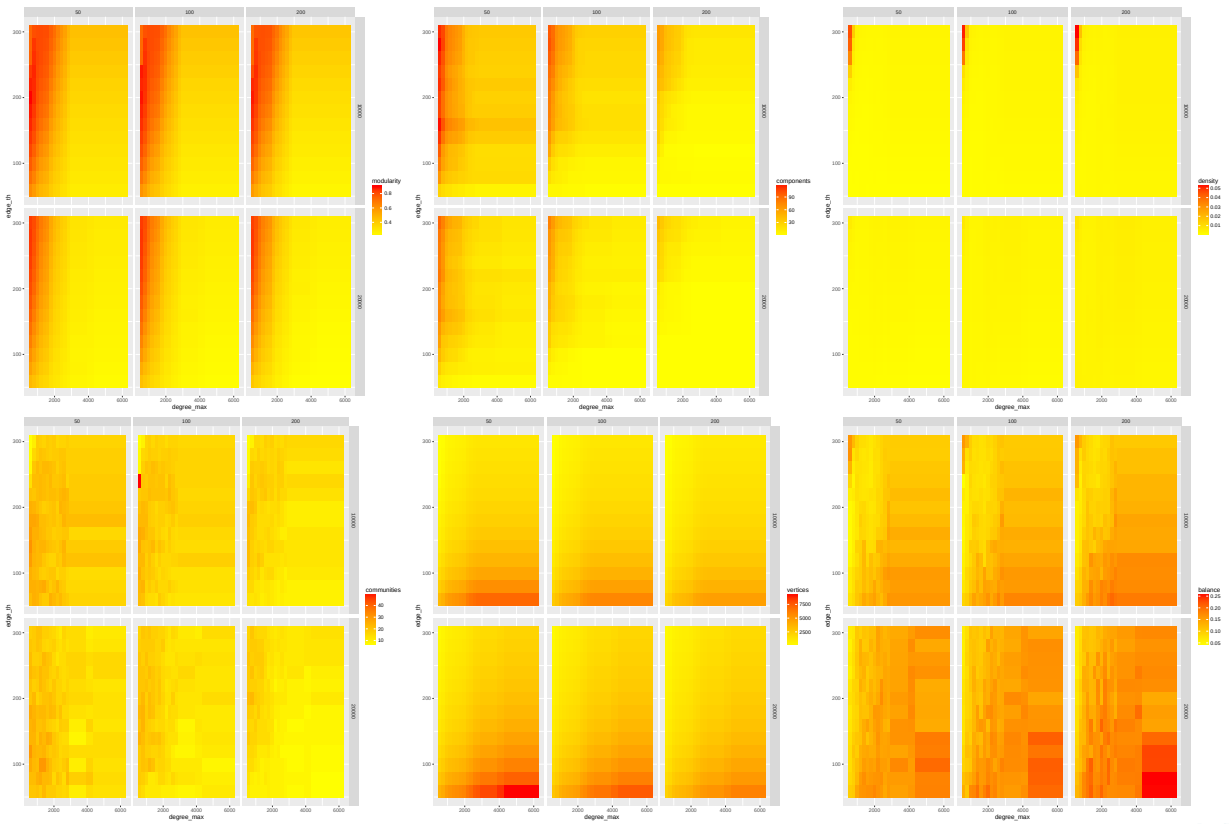
- **Nodes :** Keywords with largest relevance
- **Links :** Weighted co-occurrences
- Manual suppression of parasite words (e.g. : copyright statements !)
- Low weight link filtering
- Suppression of *hubs* (ex. `model`, `space`, `structure`, `process`) that suppress community structure
- Community detection by greedy modularity maximization (Louvain method [Blondel et al., 2008])

# Parameters influence

*Importance of* **fine tuning** *:*

→ Sensitivity of models **and** data analysis to parameters. Systematic exploration mandatory, via OpenMole for example.

→ Place of expert decision-making : no qualitative-quantitative dichotomy




*Multi-criteria optimization (modularity, size, balance) on network construction parameters*

# Obtained disciplines

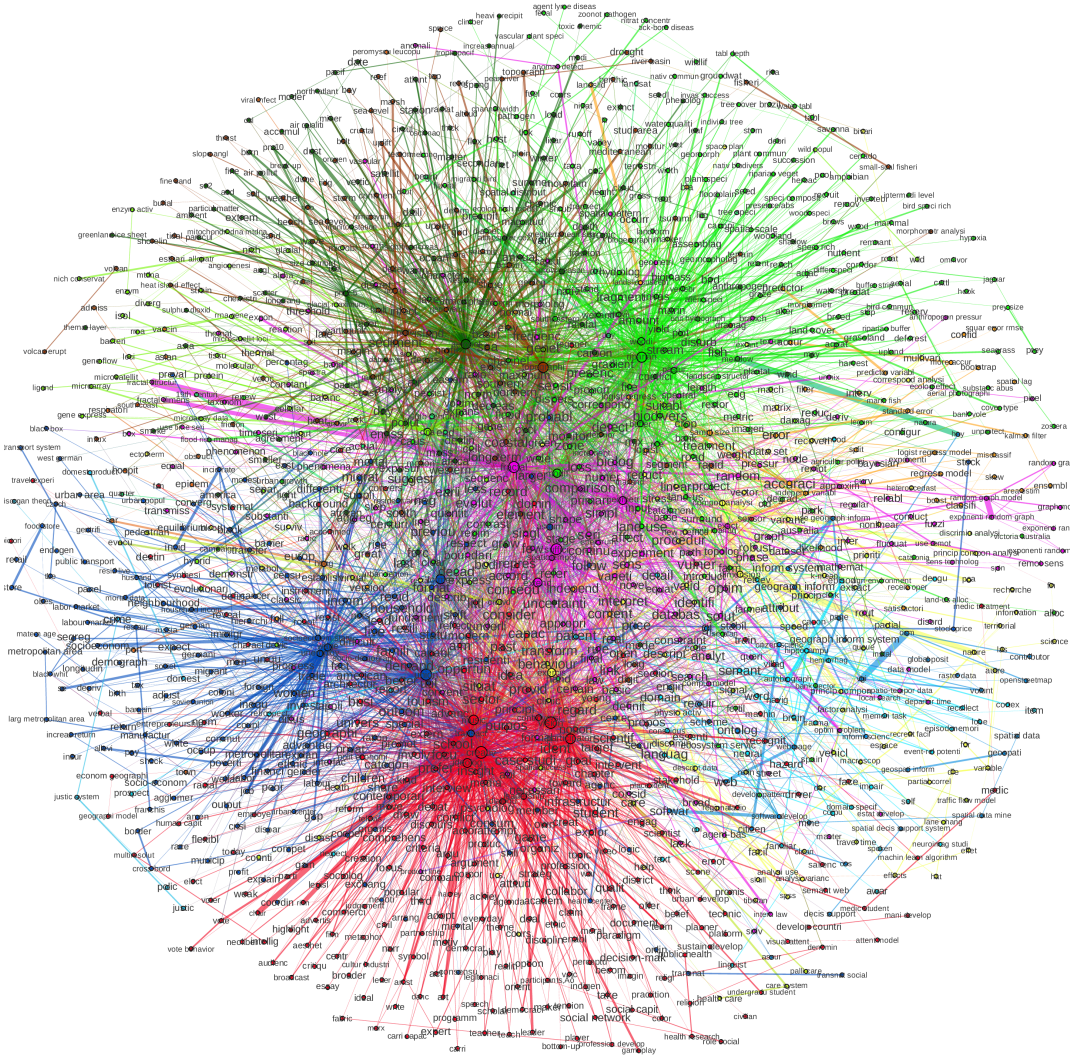
*Communities obtained with* $\theta_V = 1200, \theta_E = 50$

- Political sciences/critical geography (535) : `decision-mak, polit ideolog, democraci, stakehold, neoliber`
- Biogeography (394) : `plant densiti, wood, wetland, riparian veget`
- Economic geography (343) : `popul growth, transact cost, socio-econom, household incom`
- Environnment/climate (309) : `ice sheet, stratospher, air pollut, climat model`
- Complex systems (283) : `scale-fre, multifract, agent-bas model, self-organ`
- Physical geography (203) : `sedimentari, digit elev model, geolog, river delta`
- Spatial analysis (175) : `spatial analysi, princip compon analysi, heteroscedast, factor analysi`

# Obtained disciplines (continued)

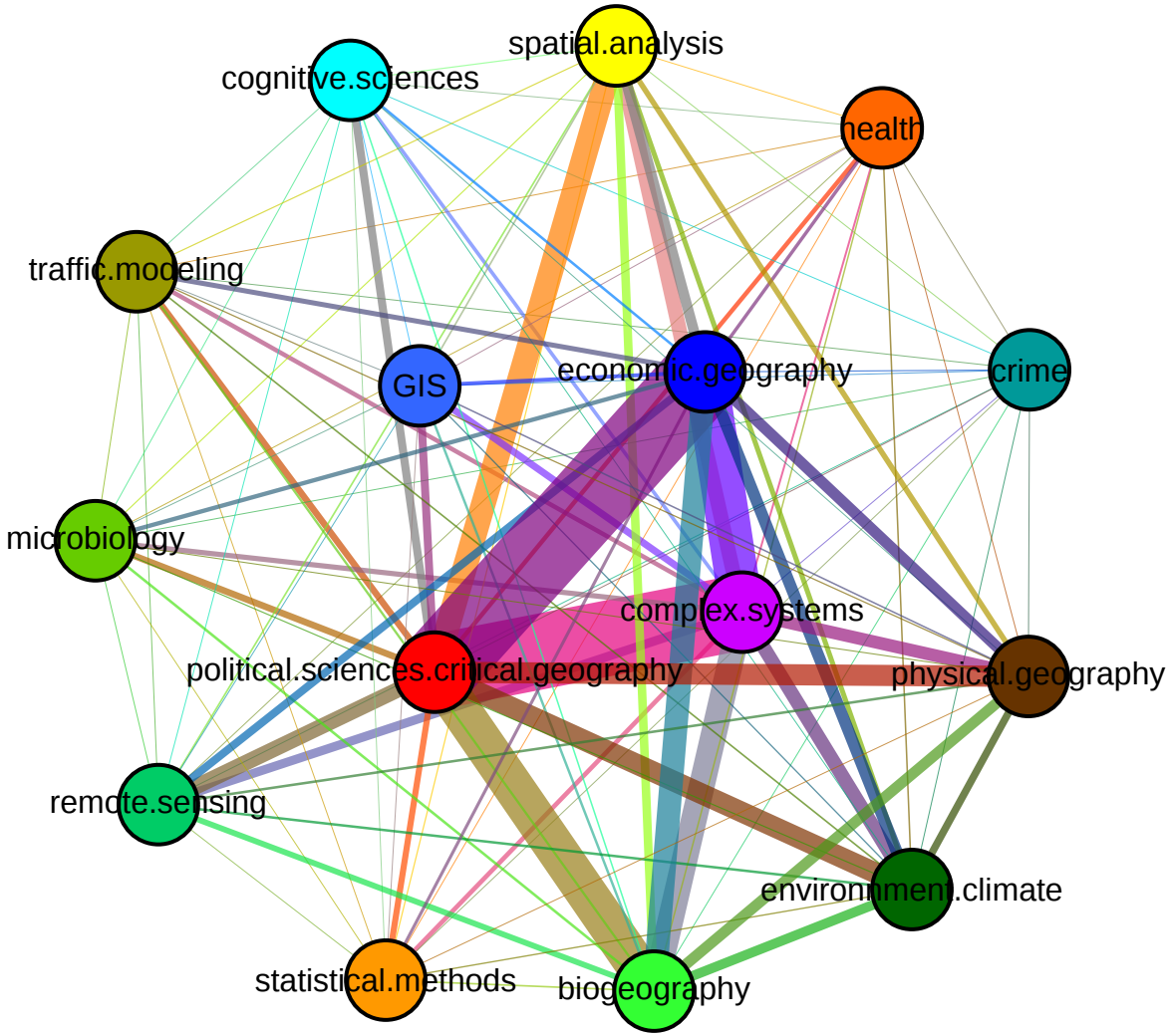
- Microbiology (118) : `chromosom, phylogenet, borrelia`
- Statistical methods (88) : `logist regress, classifi, kalman filter, sampl size`
- Cognitive sciences (81) : `semant memori, retrospect, neuroimag`
- GIS (75) : `geograph inform scienc, softwar design, volunt geograph inform, spatial decis support`
- Traffic modeling (63) : `simul model, lane chang, traffic flow, crowd behavior`
- Health (52) : `epidem, vaccin strategi, acut respiratori syndrom, hospit`
- Remote sensing (48) : `land-cov, landsat imag, lulc`
- Crime (17) : `crimin justic system, social disorgan, crime`

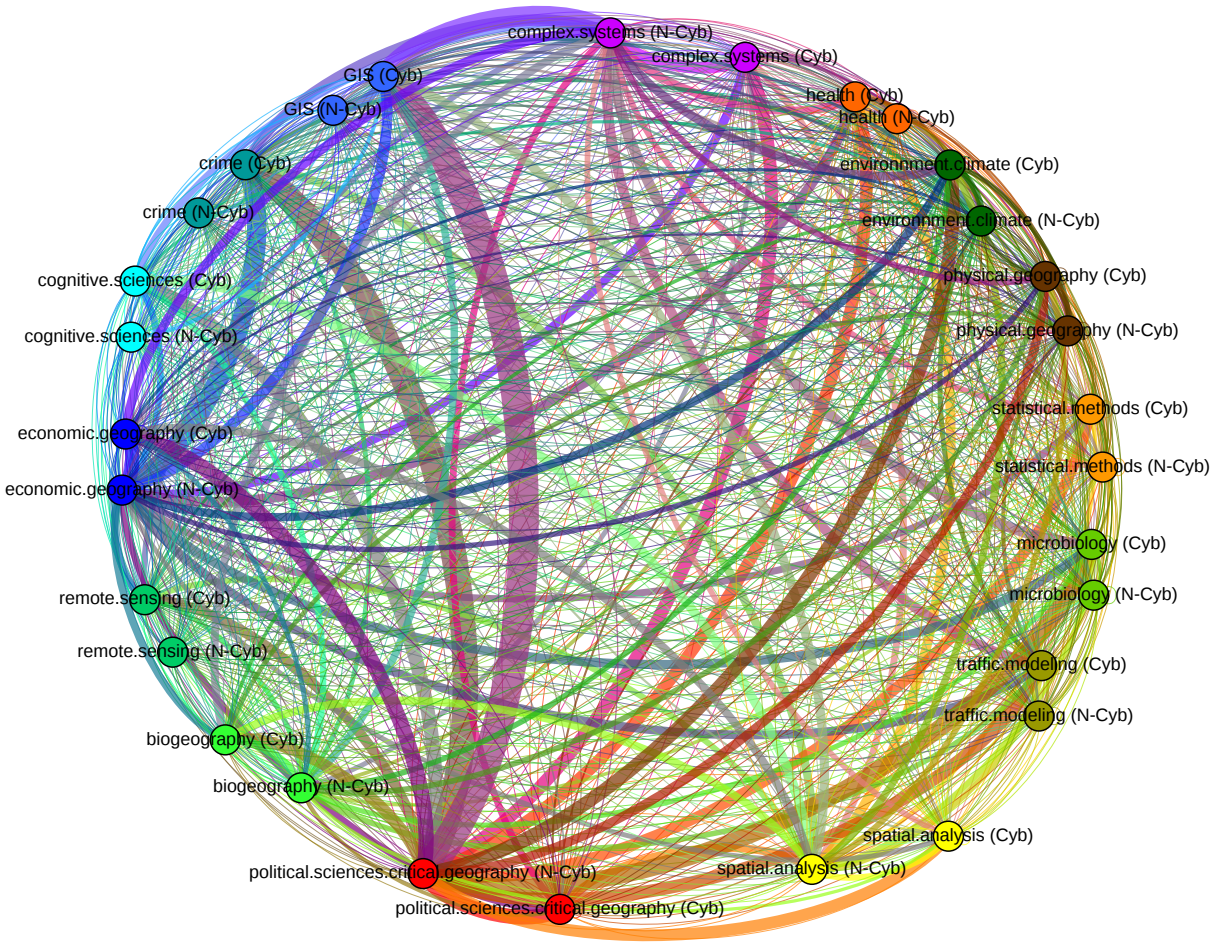
# Network



| | | |
|---|---|---|
| | political sciences/critical geography | (19,92 %) |
| | biogeography | (16,24 %) |
| | economic geography | (13,31 %) |
| | complex systems | (11,35 %) |
| | environnment/climate | (8,57 %) |
| | physical geography | (7,82 %) |
| | spatial analysis | (5,94 %) |
| | microbiology | (3,83 %) |
| | cognitive sciences | (2,93 %) |
| | statistical methods | (2,86 %) |
| | GIS | (2,48 %) |
| | traffic modeling | (1,8 %) |
| | remote sensing | (1,28 %) |
| | health | (1,13 %) |
| | crime | (0,53 %) |

# Interdisicplinarity

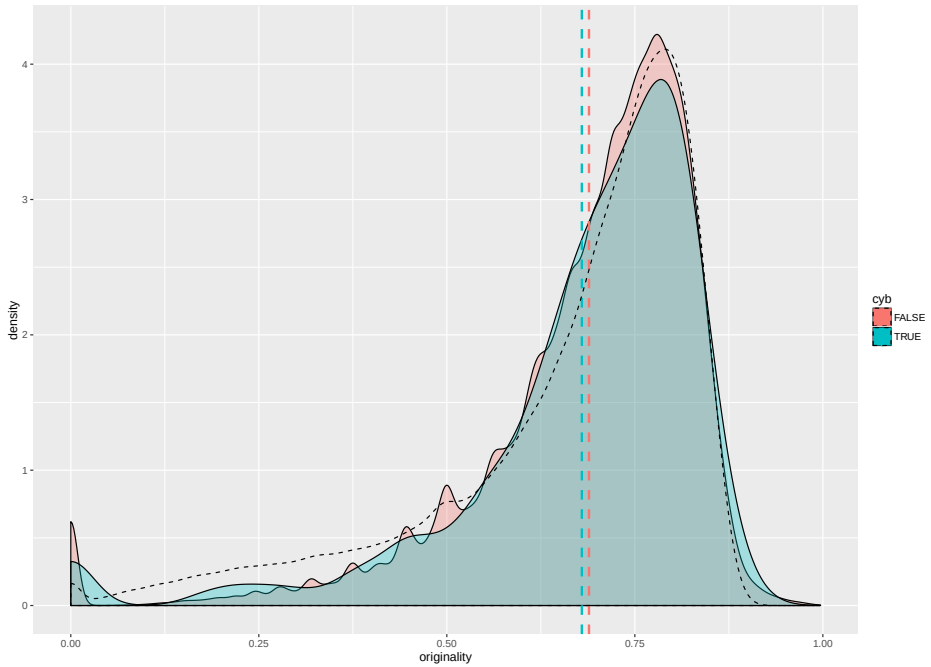



complex.systems (N-Cyb)
complex.systems (Cyb)
GIS (Cyb)
GIS (N-Cyb)
health (Cyb)
health (N-Cyb)
crime (Cyb)
environment.climate (Cyb)
crime (N-Cyb)
environment.climate (N-Cyb)
cognitive.sciences (Cyb)
physical.geography (Cyb)
cognitive.sciences (N-Cyb)
physical.geography (N-Cyb)
statistical.methods (Cyb)
economic.geography (Cyb)
statistical.methods (N-Cyb)
economic.geography (N-Cyb)
microbiology (Cyb)
microbiology (N-Cyb)
remote.sensing (Cyb)
remote.sensing (N-Cyb)
traffic.modeling (Cyb)
traffic.modeling (N-Cyb)
biogeography (Cyb)
biogeography (N-Cyb)
spatial.analysis (Cyb)
spatial.analysis (N-Cyb)
political.sciences.critical.geography (N-Cyb)
political.sciences.critical.geography (Cyb)

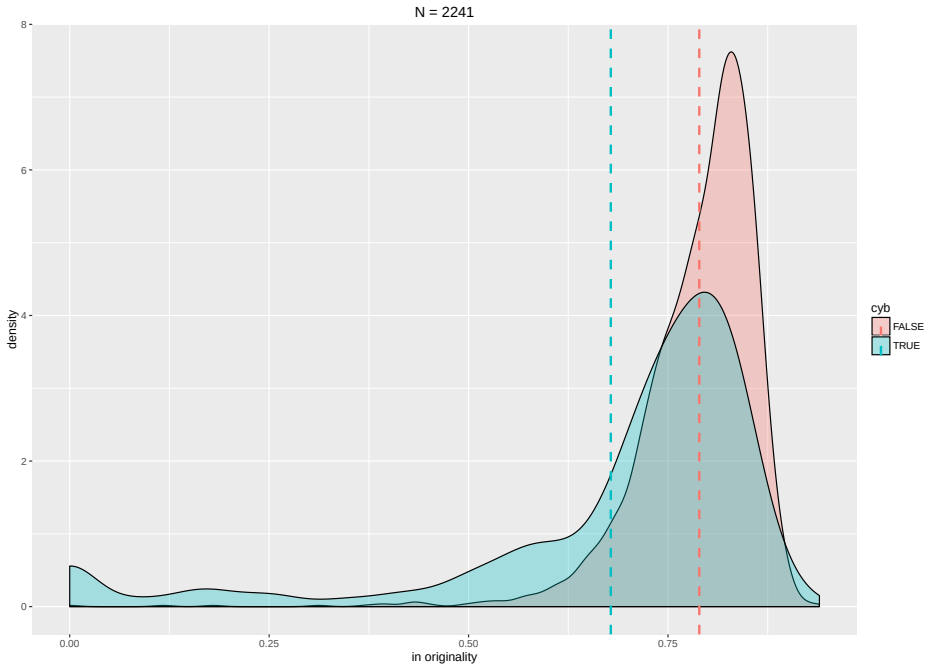
# Degré d'interdisciplinarité par articles

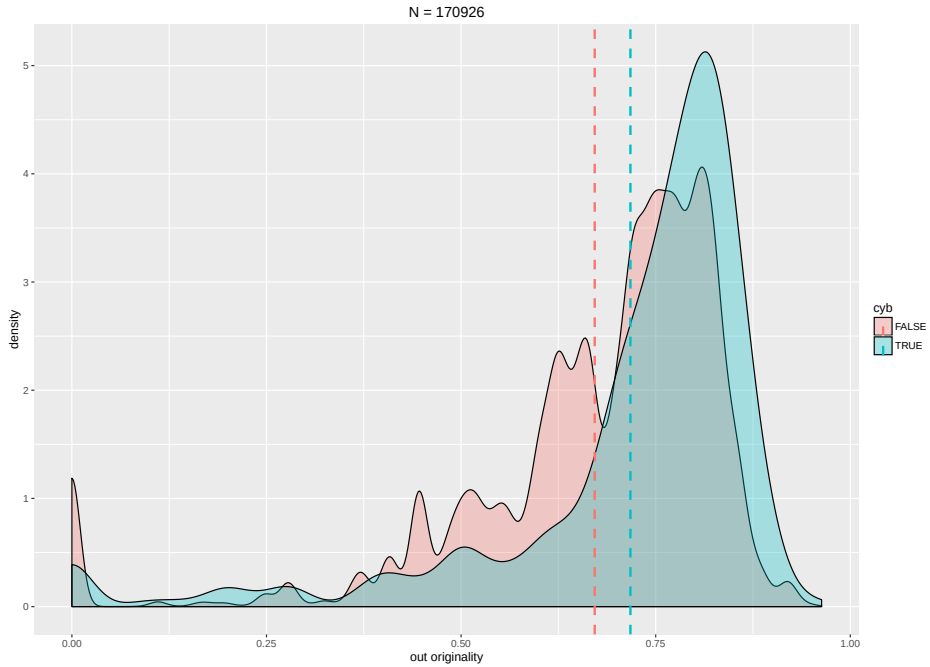
Un article peut être associé aux communautés sémantiques par ses mots clés : probas $p_i$ pour chaque communauté.
Mesure d'interdisciplinarité (pour un article, au premier ordre) :

$$o = 1 - \sum p_i^2$$



# Citation interdisciplinarity



# Conclusion

→ A very rich scientific environment and a certain interdisciplinarity

→ Approach to be combined with other classifications (thematic (POC), keywords (HC), geographical (CC)) to unveil patterns in geographical practices around the journal

→ Generic method that can be applied to any network whose nodes have a textual description

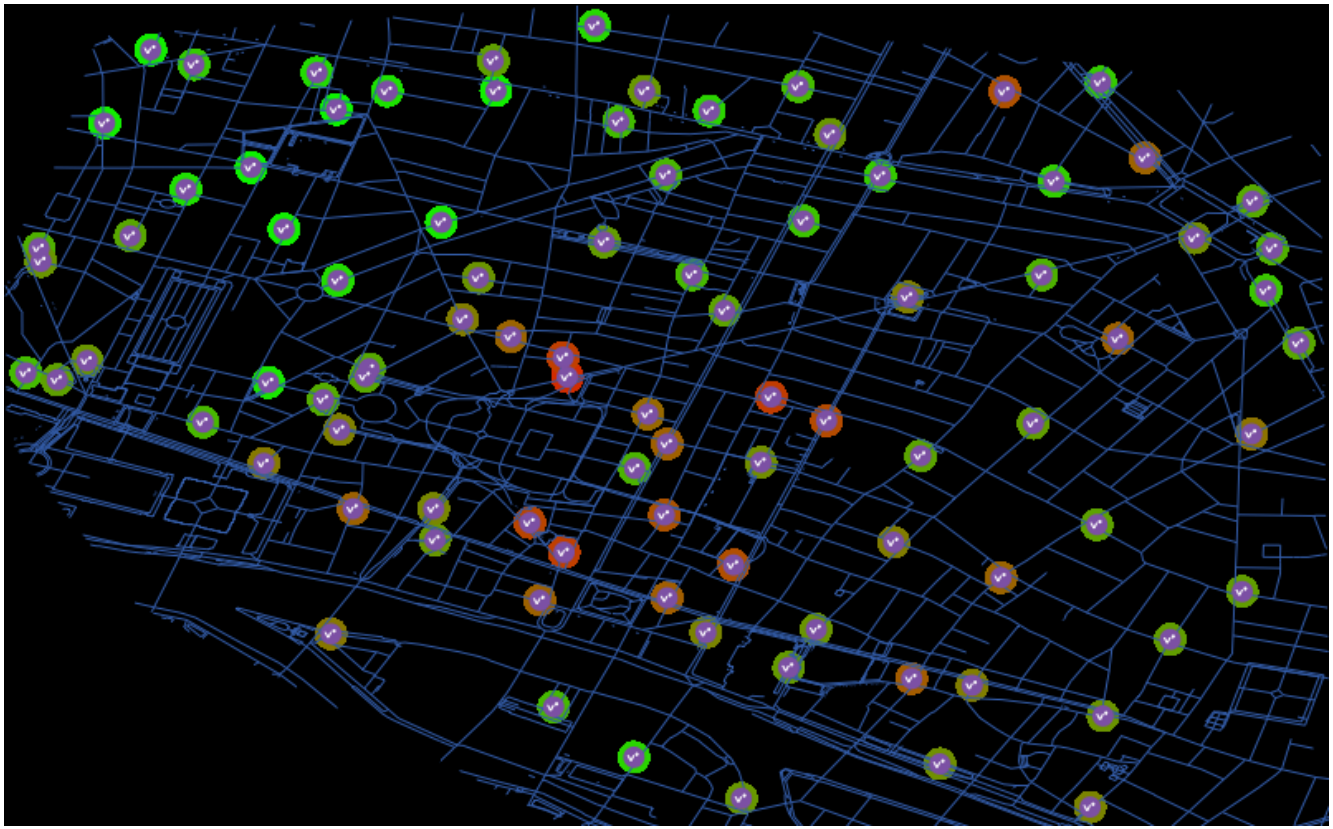
# Reserve Slides

*Reserve Slides*

# Data Collection

*Crawling of semi-open data : examples in geography*

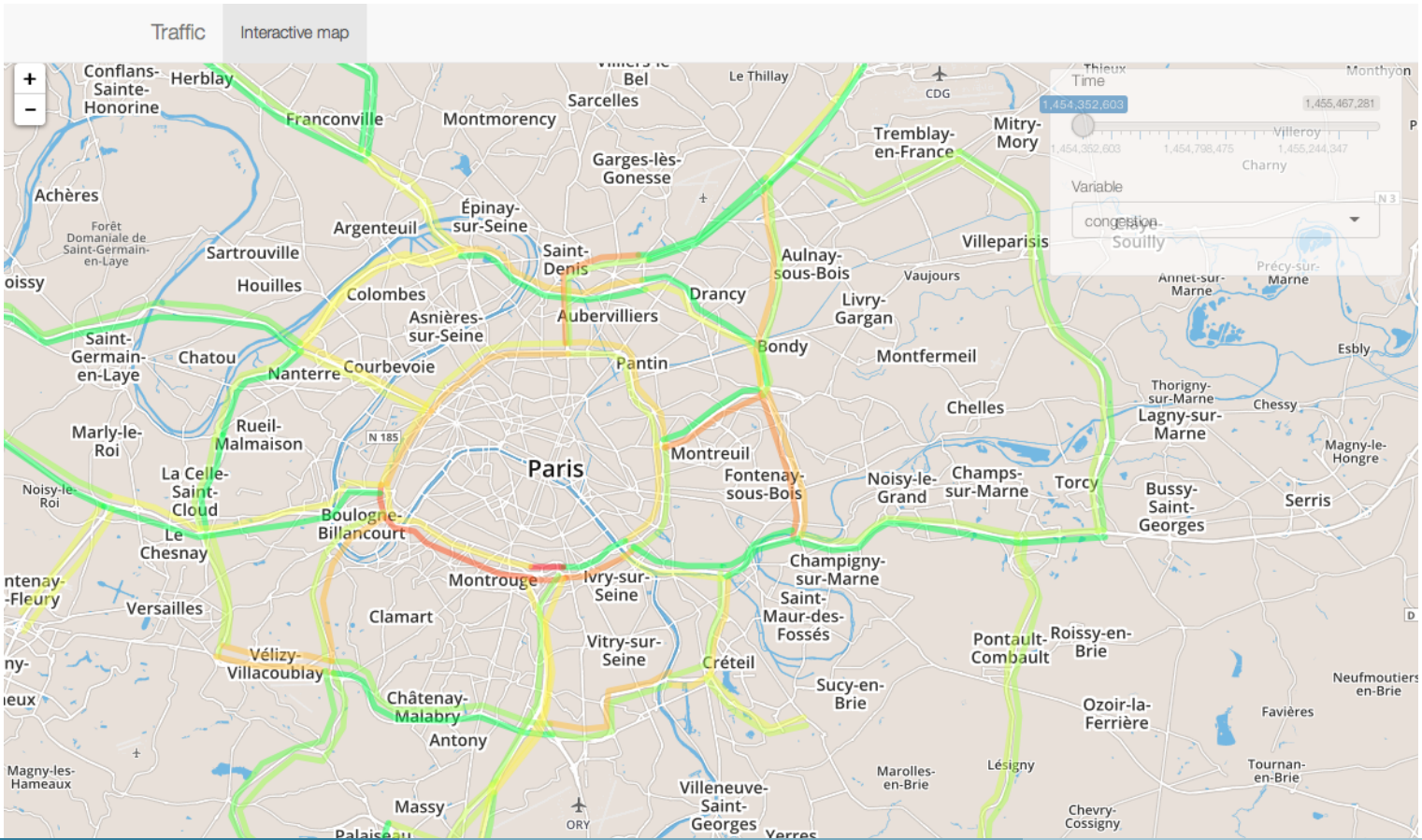
Mobility data : bike-sharing docking stations status (API)
[?]



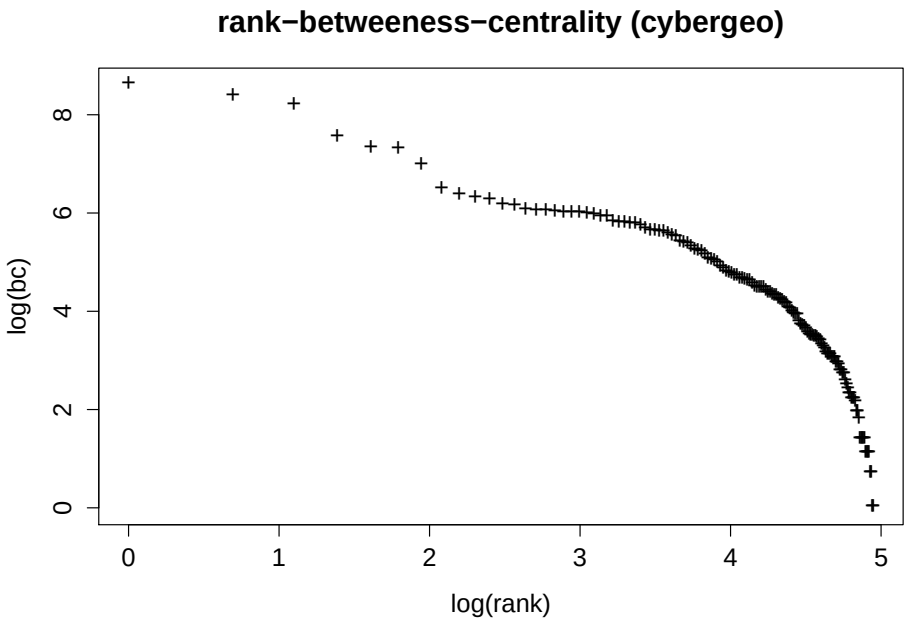
# Data Collection

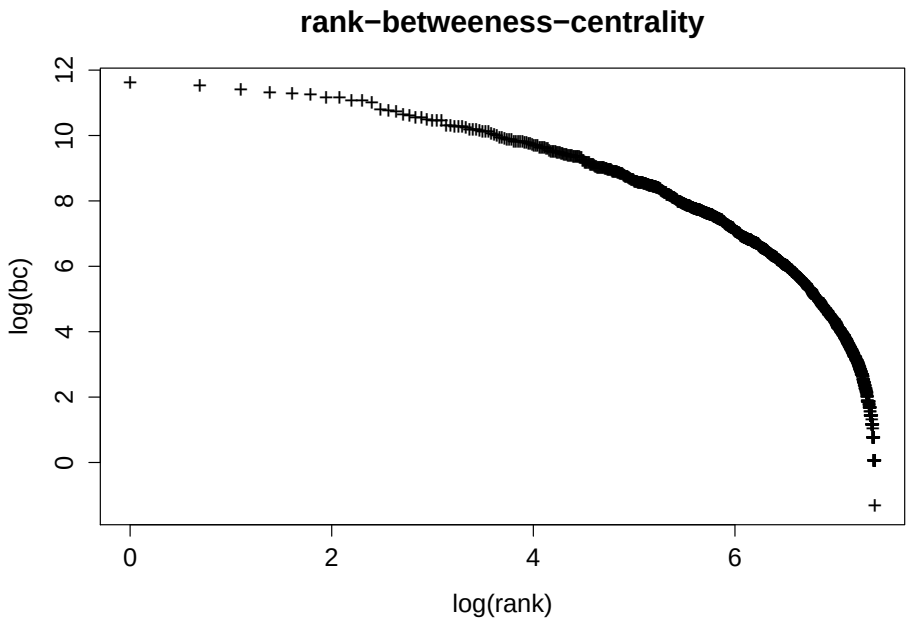
*Exemples in geography (continued)*

Road traffic : collect of *sytadin* data (no API : *scrapping* is necessary)



# Centrality (citation)

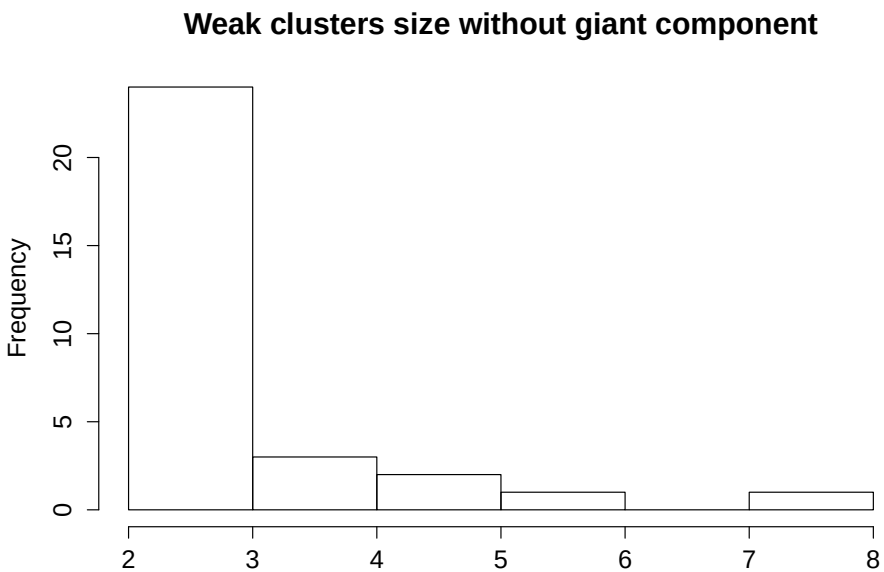


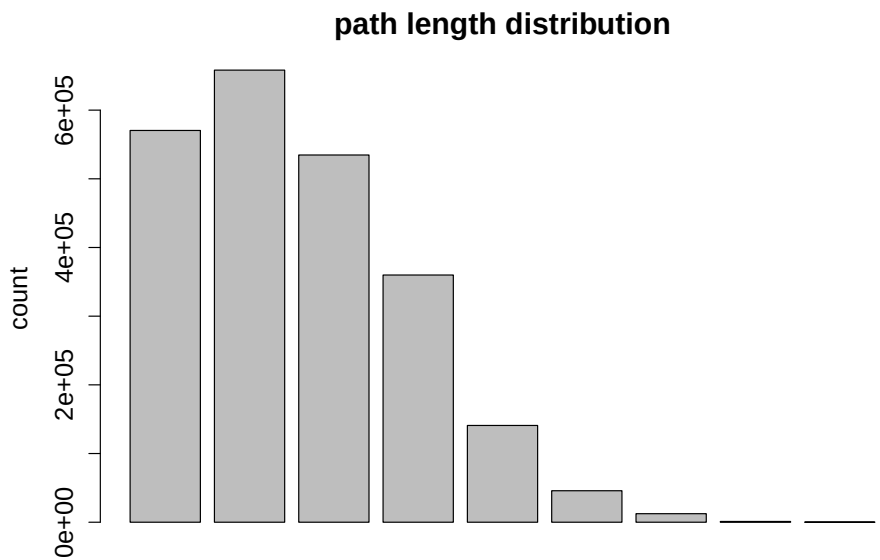




*Weak centralities (rq : impossibility of having strong clusters because of temporal causality). Left : Cybergeo ; Right : Whole Network*

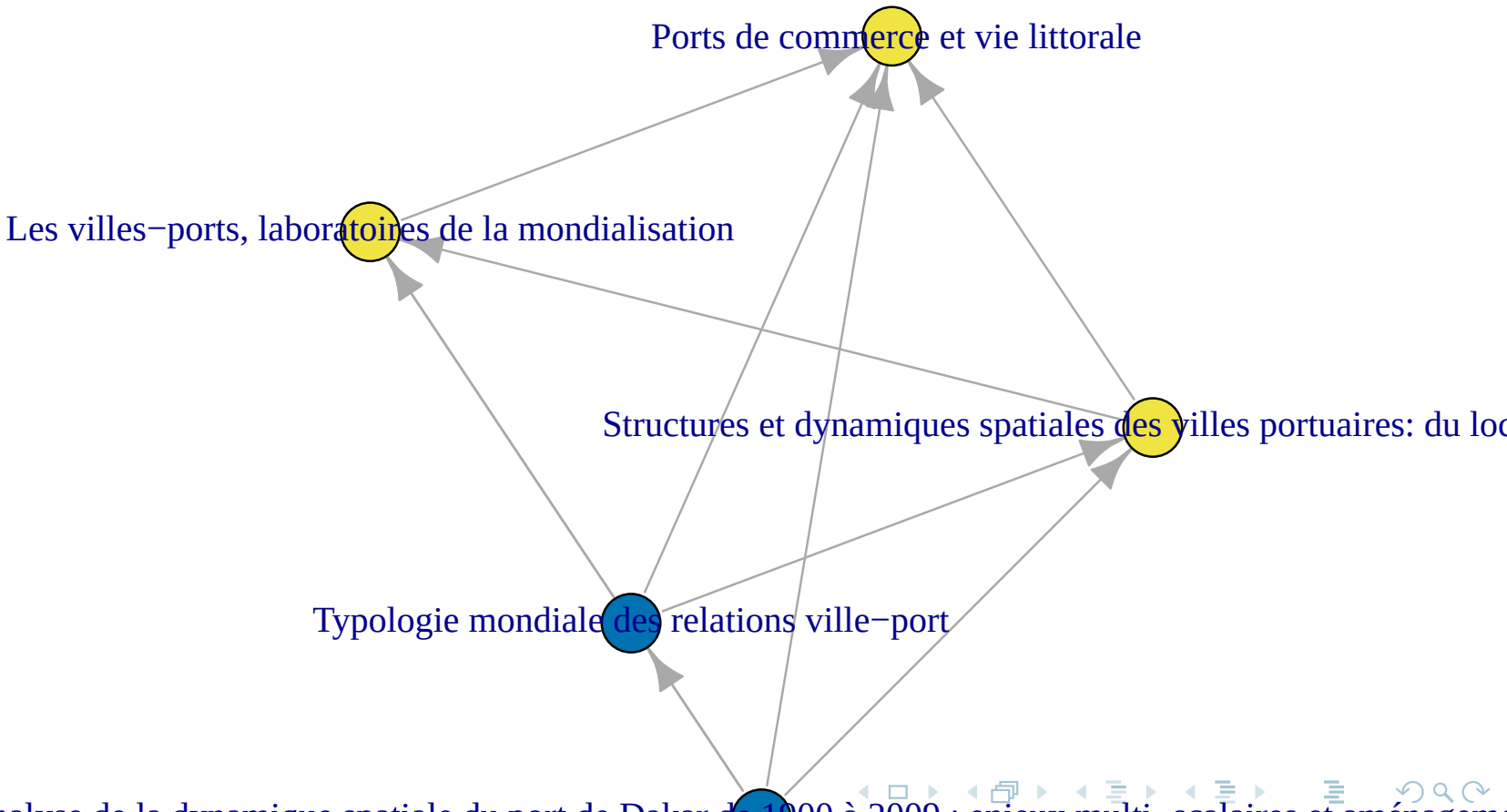
# Clustering (citation)

Giant component : more than 99% of nodes.

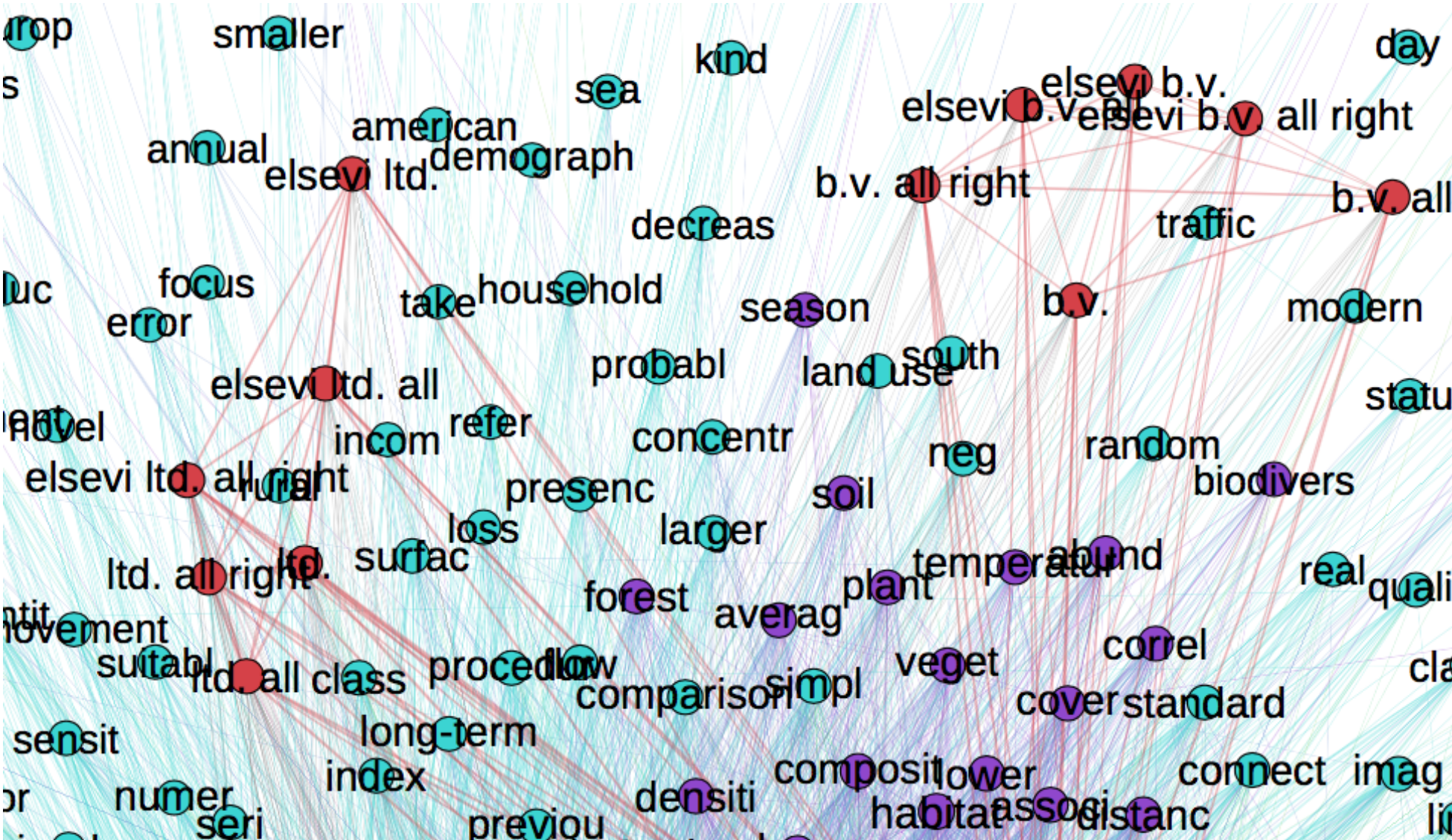


# Cliques(citation)



# Network Perturbation



# Relevance estimation

Estimation exacte de la pertinence via la repartition statistique des co-occurrences (score de $\chi^2$) : *termhood* définie, avec $M_{ij}$ nombre d'articles o $i$ et $j$ apparaissent simultanment,

$$t_i = \sum_{j \neq i} \frac{(M_{ij} - \sum_k M_{ik} \sum_k M_{jk})^2}{\sum_k M_{ik} \sum_k M_{jk}}$$

en $\Theta(\sum_i N_i^2)$ ($N_i$ taille des résumés) : difficile sur un corpus où $\sum_i N_i^2 \simeq N < N_i >^2 \simeq 8 \cdot 10^7$

# References I


Bird, S. (2006).
NItk: the natural language toolkit.
In Proceedings of the COLING/ACL on Interactive presentation sessions, pages 69–72. Association for Computational Linguistics.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008).
Fast unfolding of communities in large networks.
Journal of statistical mechanics: theory and experiment, 2008(10):P10008.

Chavalarias, D. and Cointet, J.-P. (2013).
Phylomemetic patterns in science evolution—the rise and fall of scientific fields.
Plos One, 8(2):e54847.

# References II


Cottineau, C., Le Néchet, F., Le Texier, M., and Reuillon, R. (2015).
Revisiting some geography classics with spatial simulation.
In Plurimondi. An International Forum for Research and Debate on Human Settlements, volume 7.

Mendeley (2015).
Mendeley reference manager.
http://www.mendeley.com/.

Noruzi, A. (2005).
Google scholar: The new generation of citation indexes.
Libri, 55(4):170–180.

Schmid, H. (1994).
Probabilistic part-of-speech tagging using decision trees.
In Proceedings of the international conference on new methods in language processing, volume 12, pages 44–49. Citeseer.

# References III