

# Indirect Bibliometrics by Complex Network Analysis

J. Raimbault<sup>1,2</sup>

<sup>1</sup>Géographie-cités (UMR 8504 CNRS)

<sup>2</sup>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 a purely quantitative bibliometric 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

→ 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)

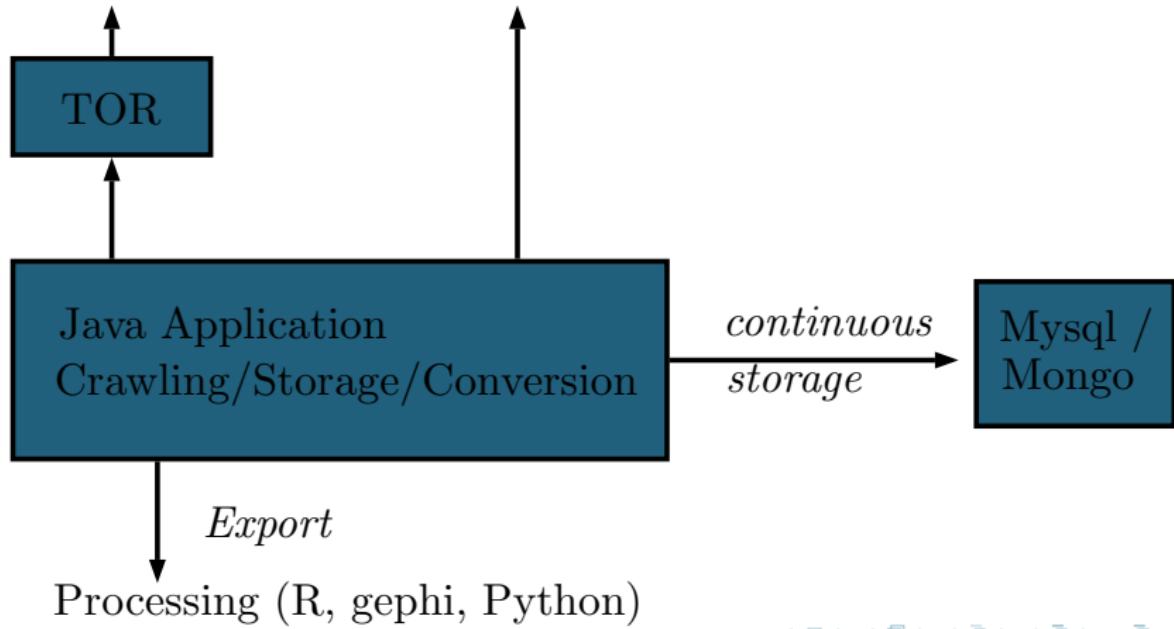
## Data Collection Architecture

*Google Scholar*

Citations and ID

*Mendeley*

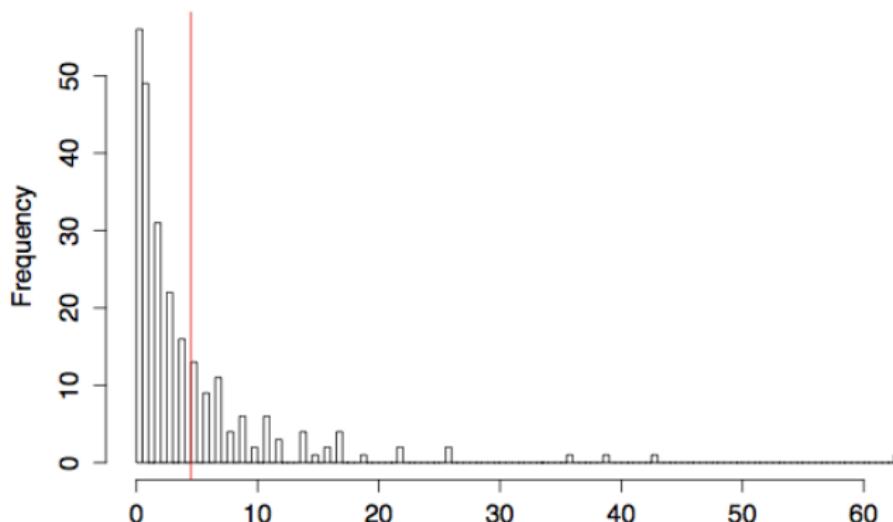
Abstracts



## Network Properties

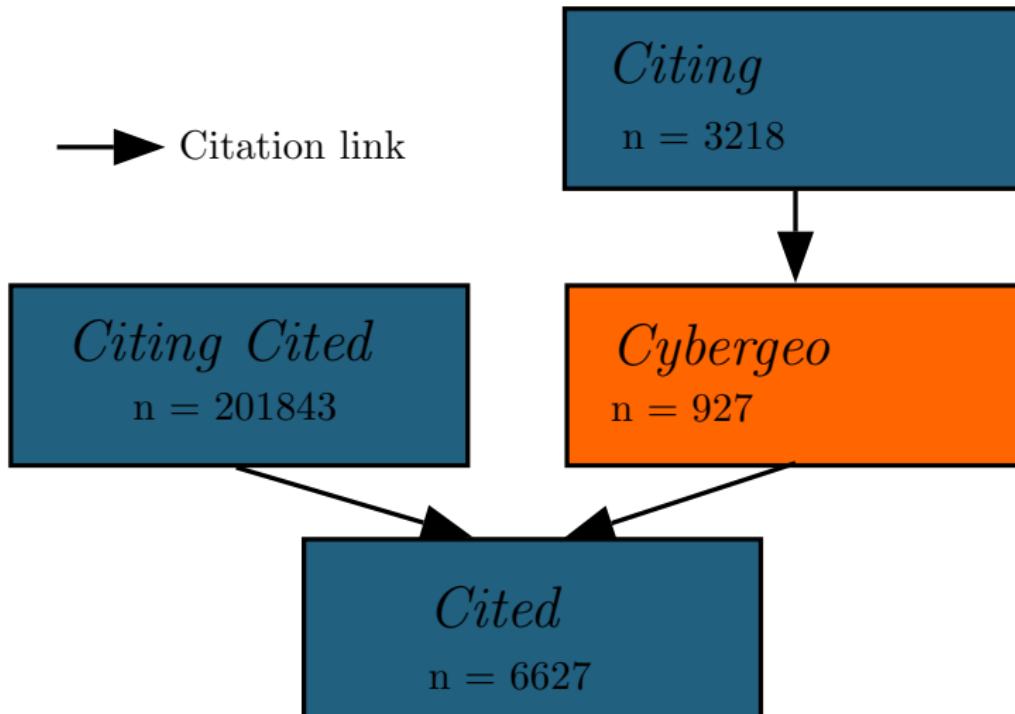
- $\simeq 947$  cybergeo articles can be studied, among  $\simeq 1200$
- $418670$  Nodes et  $570352$  Links ; Diameter : 9 ; Density :  $3.25E-6$  ; average degree : 2.724284

**Degree distribution, mean (impact factor) = 3.18**

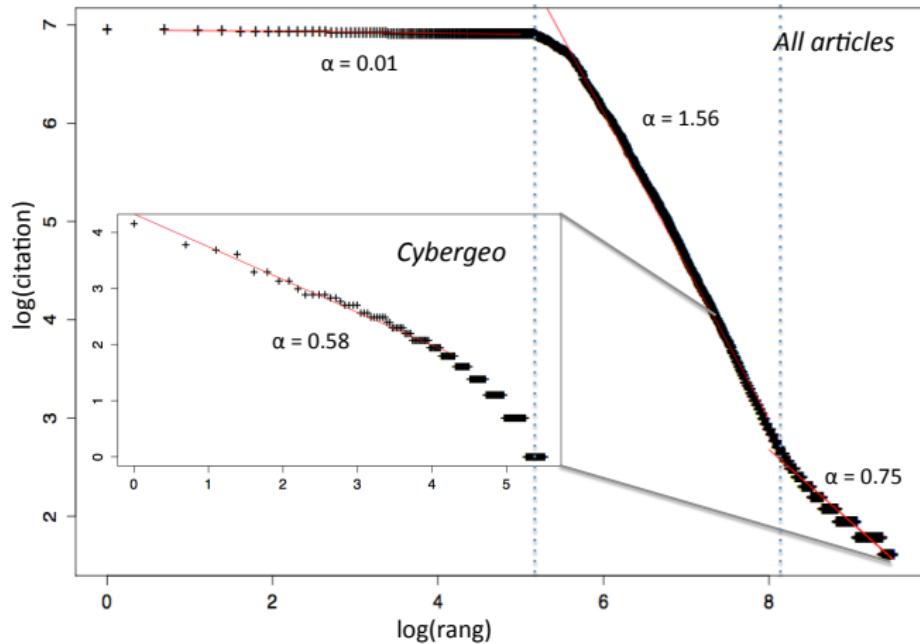


## Citation Network Structure

→ Citation link

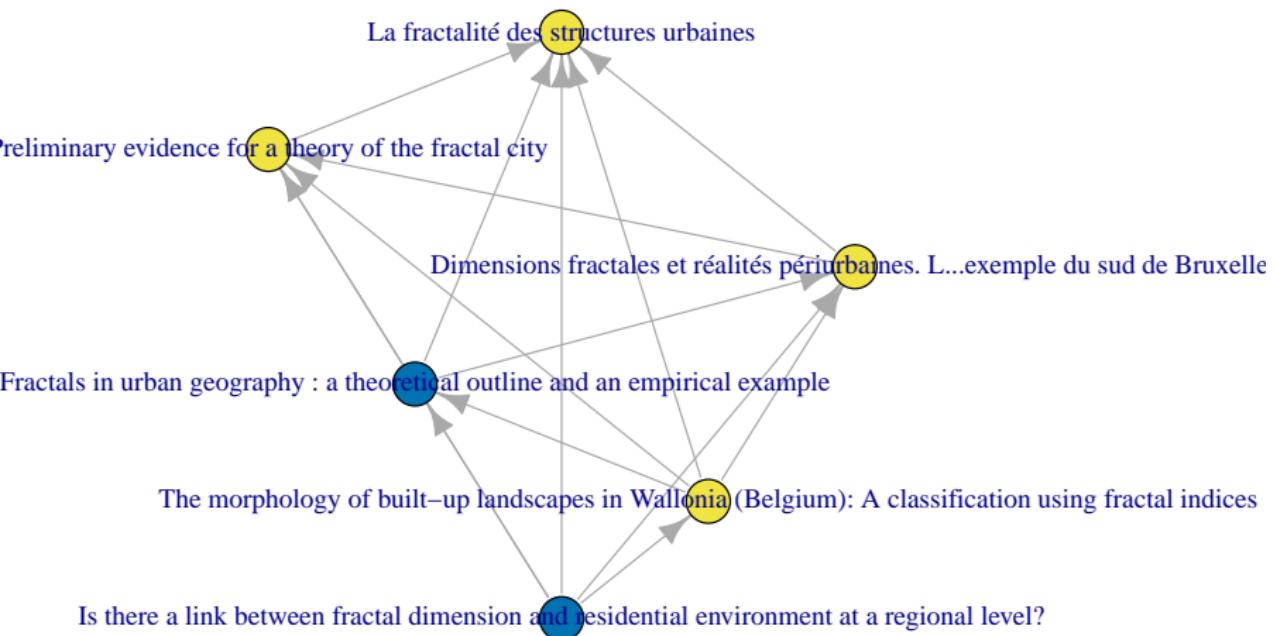


## Hierarchy in citations



*Superposition of  
different hierarchical regimes*

# Cliques



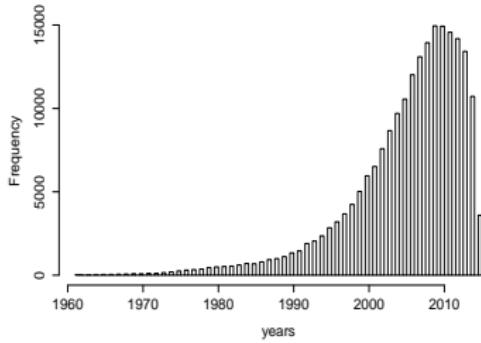
## Semantic Network

**Semantic Data :** Collection of abstract/date/authors/keywords for the 400000 references via Mendeley API → ~ 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 adapted from [Chavalarias and Céleste, 2006]

- 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 → 2min)
- Estimation of *n-grams* relevance, following co-occurrences statistical distribution

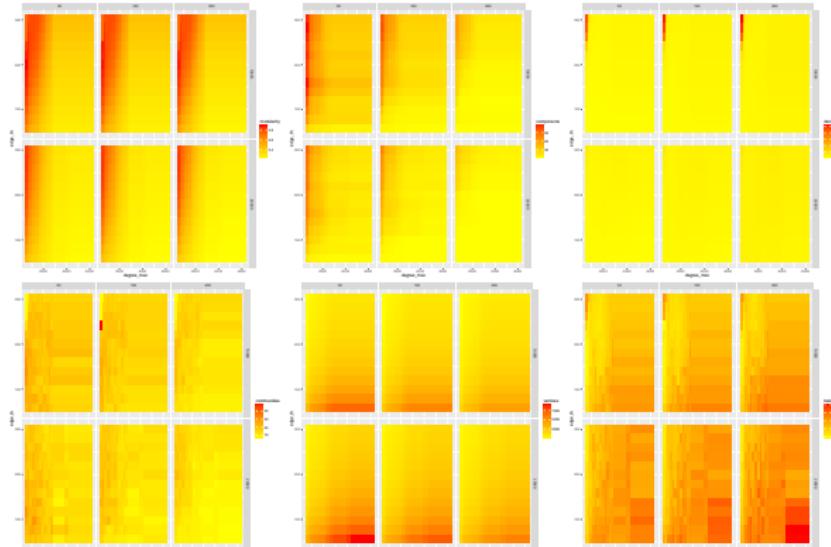
## Construction of Semantic Network

- **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 d'un réglage fin :*

- Sensibilité des modèles **et** traitements de données aux paramètres. Exploration systématique via OpenMole par exemple.
- Importance du jugement d'expert : pas de dichotomie “quanti-quali”



## Obtained disciplines

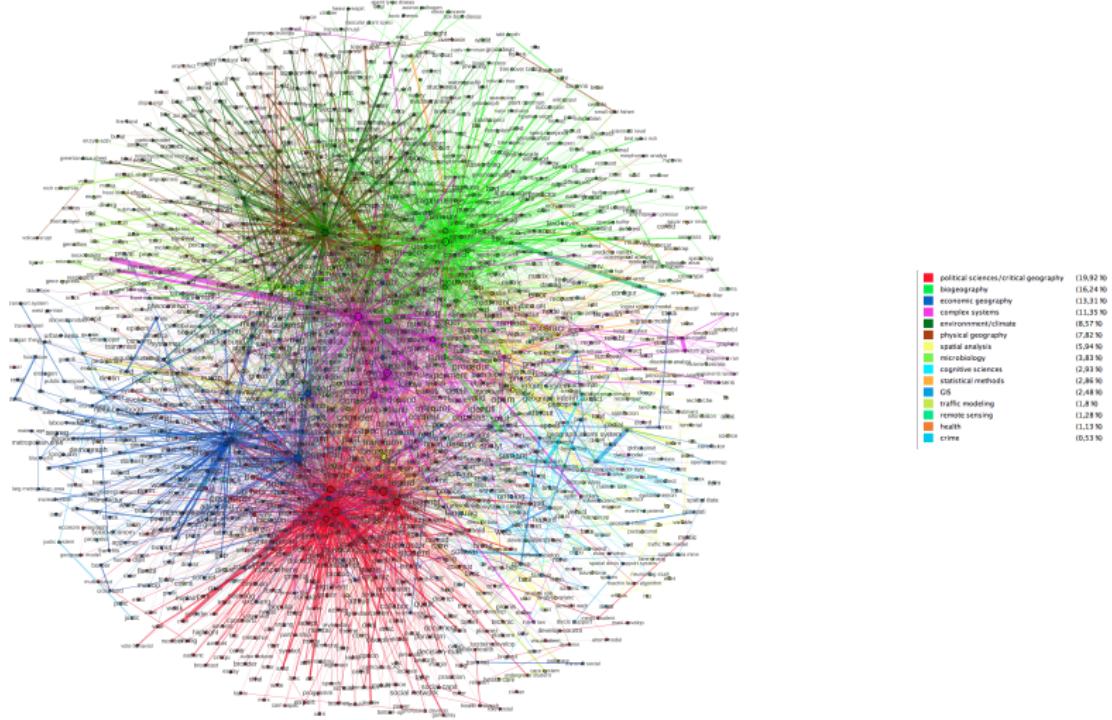
*Communities obtained with  $\theta_V = 1200, \theta_E = 50$*

- 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
- Environment/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

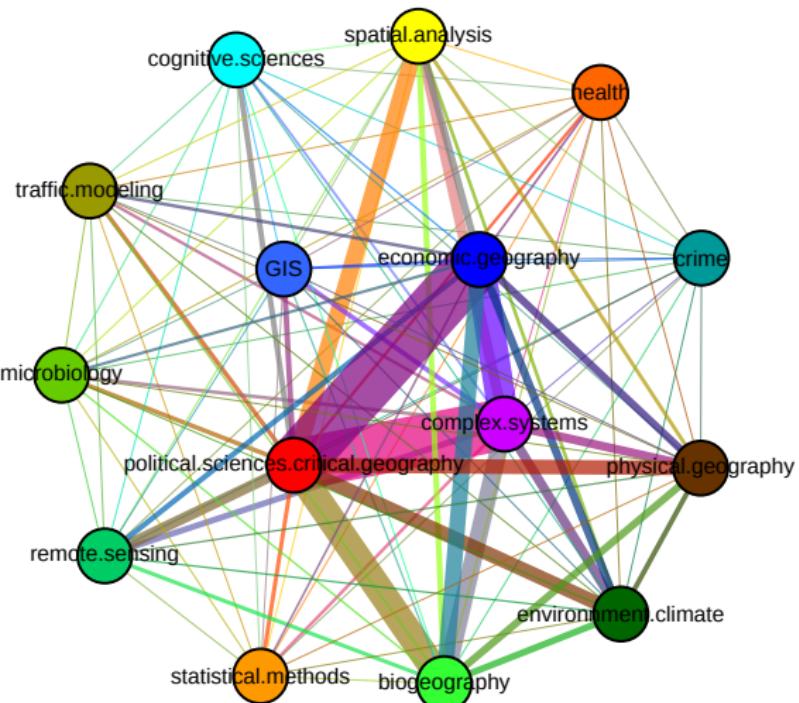
## Obtained disciplines (continued)

- 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

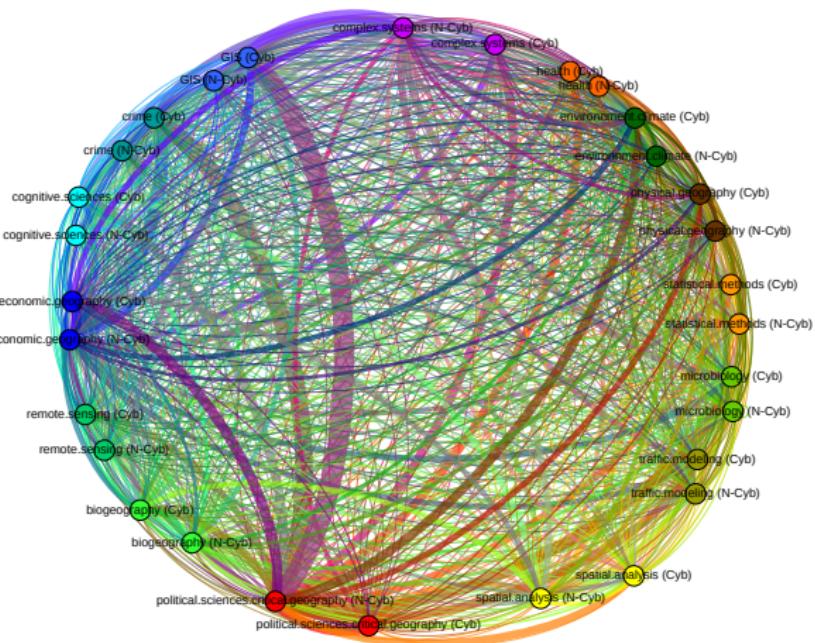
# Network



## Interdisciplinarity



... 20 ans de cybergeo

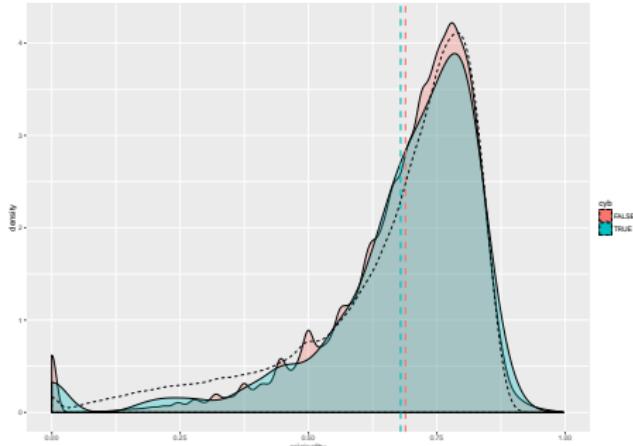


## Degré d'interdisciplinarité par articles

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) :

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





1996-2016 : 20 ans de cybergeo

1996-2016 : 20 years of cybergeo

## Citation interdisciplinarity

\relax .png

.png

\relax .png

.png

\relax .png

[width=]figures/citin;ntardisc

\relax .png

.png

## Conclusion

→

- Approche à croiser avec autres types de classifications (thématique (POC), mots-clés (HC), géographique (CC) pour en apprendre plus sur la revue et la pratique de la géographie qui lui est associée
- Méthode générique pouvant s'appliquer à tout réseau dont les noeuds ont une description textuelle



1996-2016 : 20 ans de cybergeo

1996-2016 : 20 years of cybergeo

## Reserve Slides

*Reserve Slides*

## Collecte des données

*Crawling de données semi-ouvertes : exemples en géographie*

Données de mobilité : statuts des stations Vlib en temps réel (API)

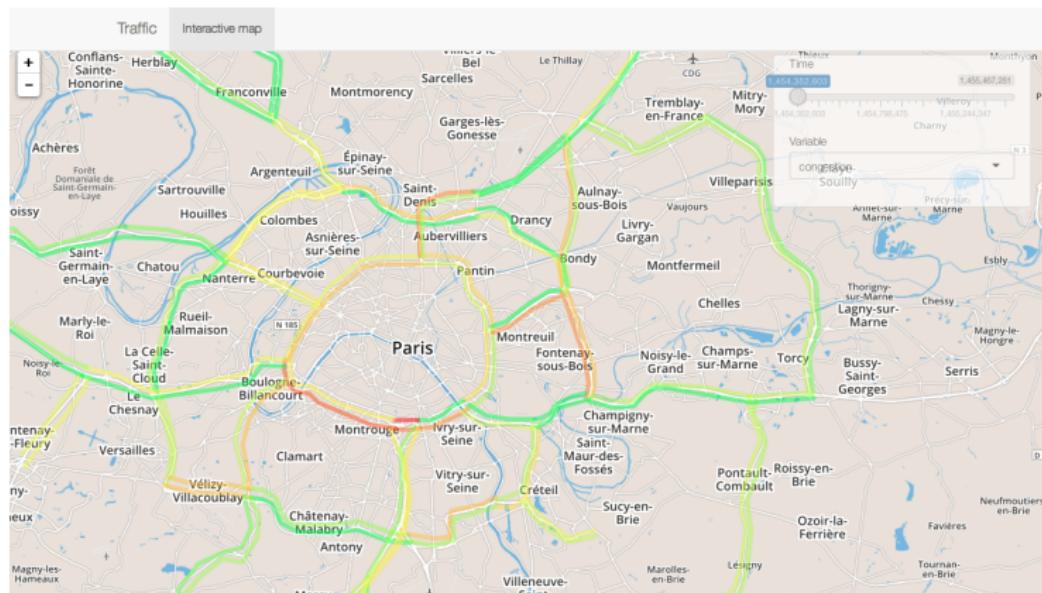
[?]



# Collecte des données

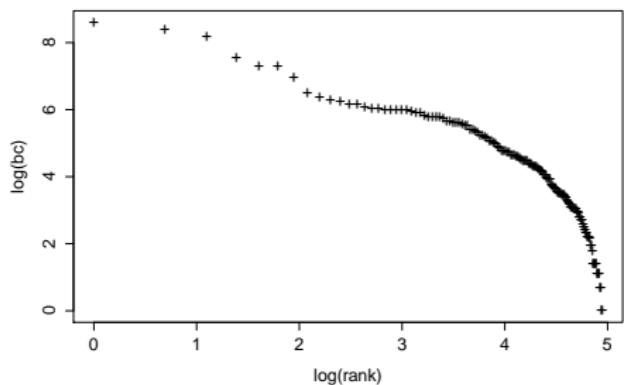
## Exemples en géographie (suite)

Traffic routier : collecte de sytadin (pas d'API : scrapping nécessaire)

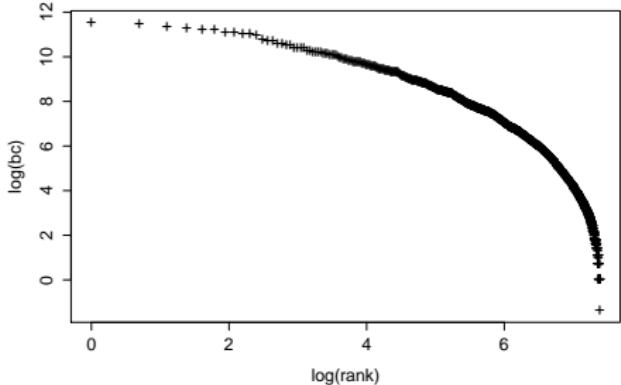


## Centralité (citation)

rank-betweenness-centrality (cybergeo)



rank-betweenness-centrality

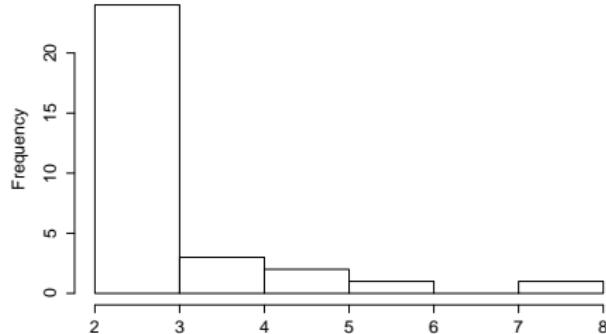


Centralités faibles (*rq* : impossibilité des clusters forts pour des citations car causalité temporelle). Gauche : Cybergéo ; Droite : Ensemble du réseau

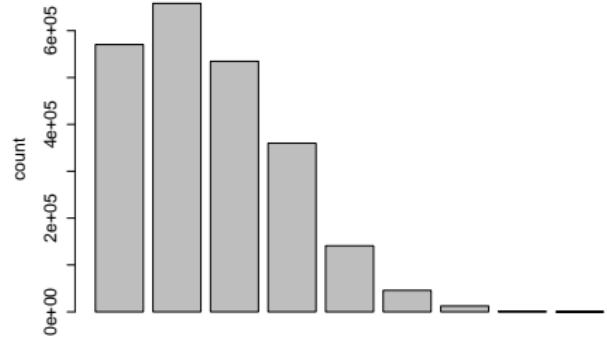
## Clustering (citation)

Composante géante : plus de 99% des noeuds.

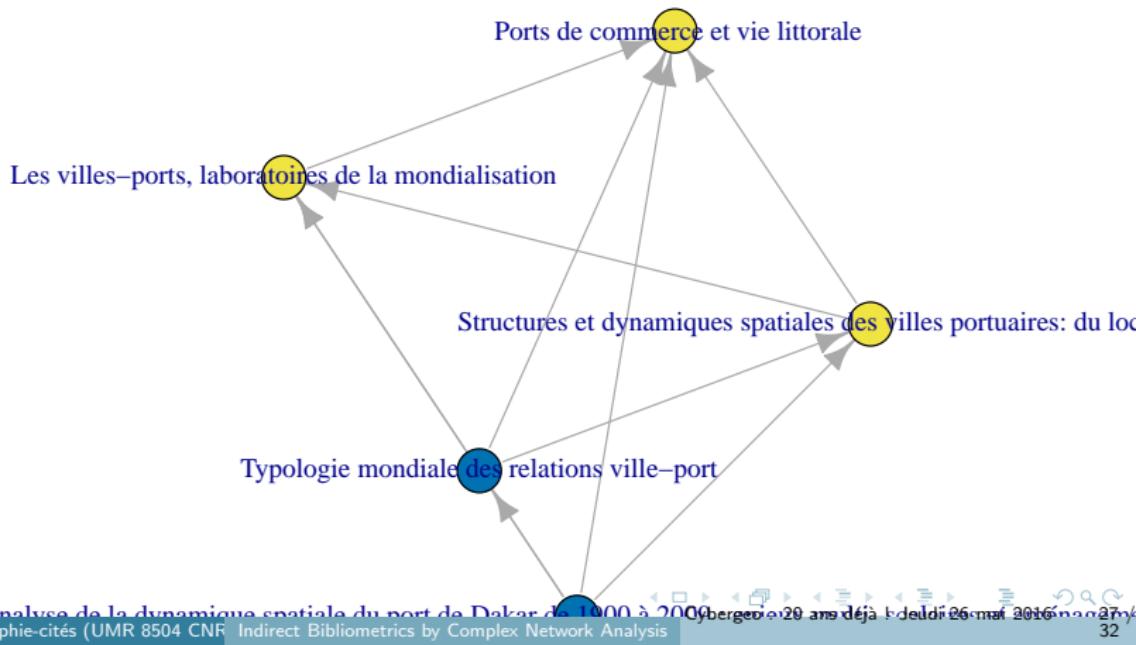
Weak clusters size without giant component



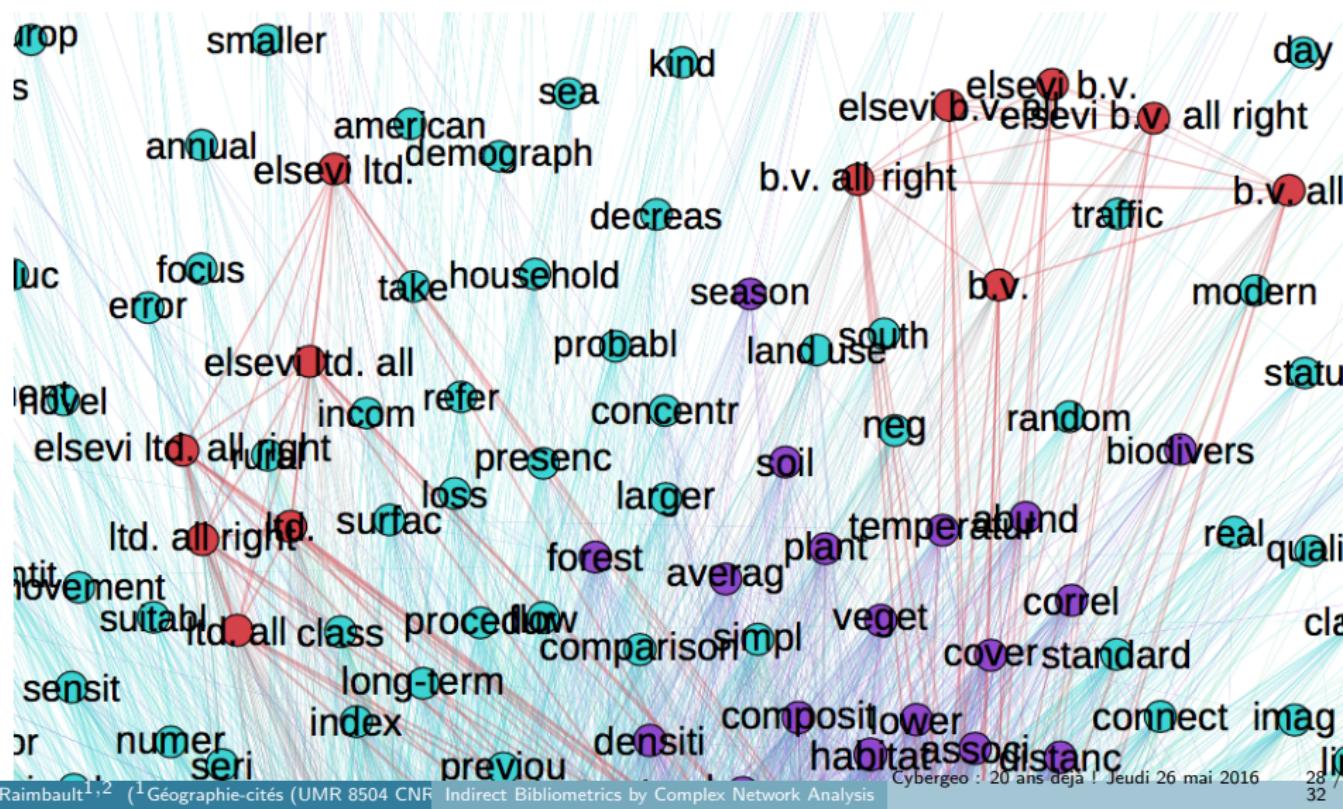
path length distribution



## Cliques(citation)



## Network Perturbation



## Estimation de la pertinence

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

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

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 |

-  Bird, S. (2006).  
NLTK: 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.



1996-2016 : 20 ans de cybergeo

1996-2016 : 20 years of cybergeo

## References III