

Bibliométrie Indirecte par Analyse de Réseaux Complexes

J. Raimbault^{1,2}

¹Géographie-cités (UMR 8504 CNRS)

²LVMT (UMR-T 9403 IFSTTAR)

Cybergeo : 20 ans déjà !
Jeudi 26 mai 2016



Contexte

"You are what you cite" : **Quels sont les environnements disciplinaires voisins de Cybergeo ? Sont-ils différents des contenus des articles (PO) et des contenus déclarés (HC) ?**

- Enjeu par rapport à la ligne éditoriale de la revue : interdisciplinarité et ouverture
- Contre une bibliométrie quantitative pure en SHS : approche semi-qualitative nécessaire



1996-2016 : 20 ans de cybergeo

1996-2016 : 20 years of cybergeo

Objectif

Question de recherche : *Dans quelle mesure le croisement des données de citation et des données sémantiques peut-il permettre d'analyser le contexte disciplinaire de la revue ?*

- Elaboration d'une approche par *Hyperréseau* : croisement du réseau de citations au réseau sémantique. Gain d'information par croisement des couches (démarche transversale)
- Données difficiles d'accès : base à construire



Collecte des données

Données des articles : bases de fonctionnement (production de la revue)
→ Structuration et Consolidation

Données de citation : revue non référencée par base “classiques” (de plus non libres !)
→ *crawling* de google scholar par utilisation de l'option “*cité par*” [Noruzi, 2005]

Données textuelles : besoin des résumés pour l'ensemble des références liées
→ utilisation de l'API Mendeley [Mendeley, 2015] (gratuite mais non ouverte).

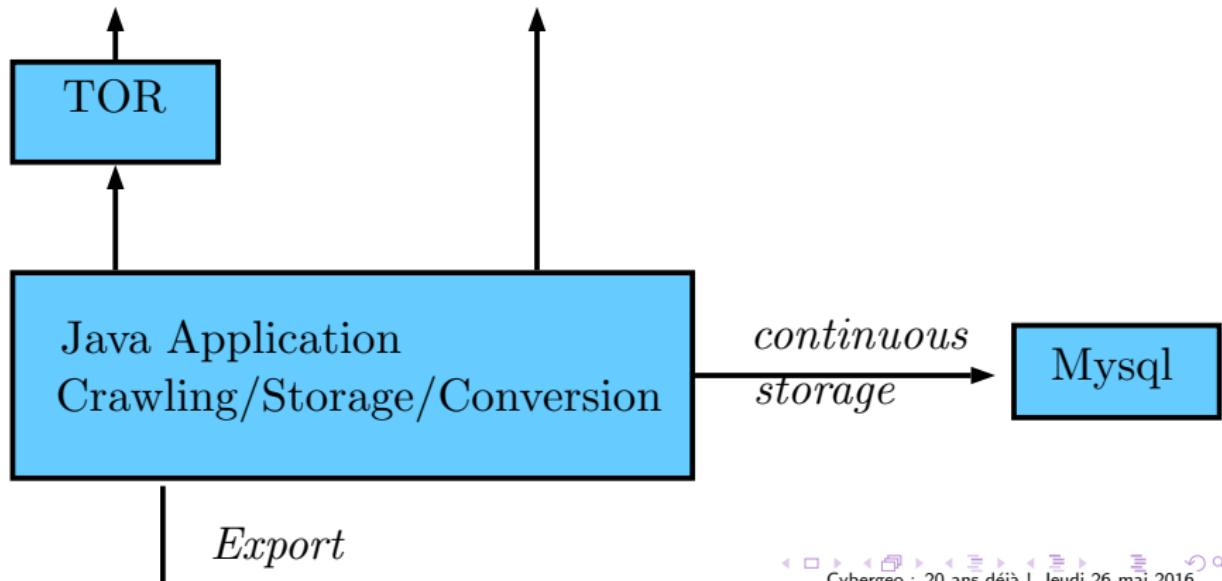
Architecture de collecte des données

Google Scholar

Citations and ID

Mendeley

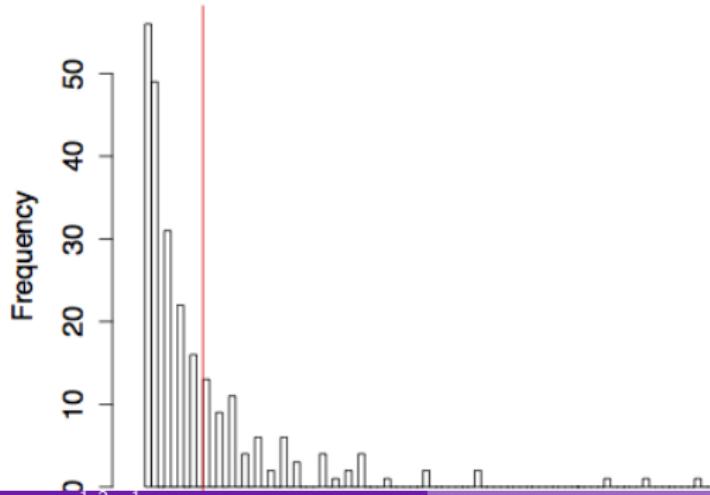
Abstracts



Caractéristiques du réseau

- Après raffinement, $\simeq 947$ références de cybergeo exploitables, sur $\simeq 1200$
- 418670 Noeuds et 570352 Liens ; Diamètre : 9 ; Densité : $3.25E-6$; degré moyen : 2.724284

Degree distribution, mean (impact factor) = 3.18

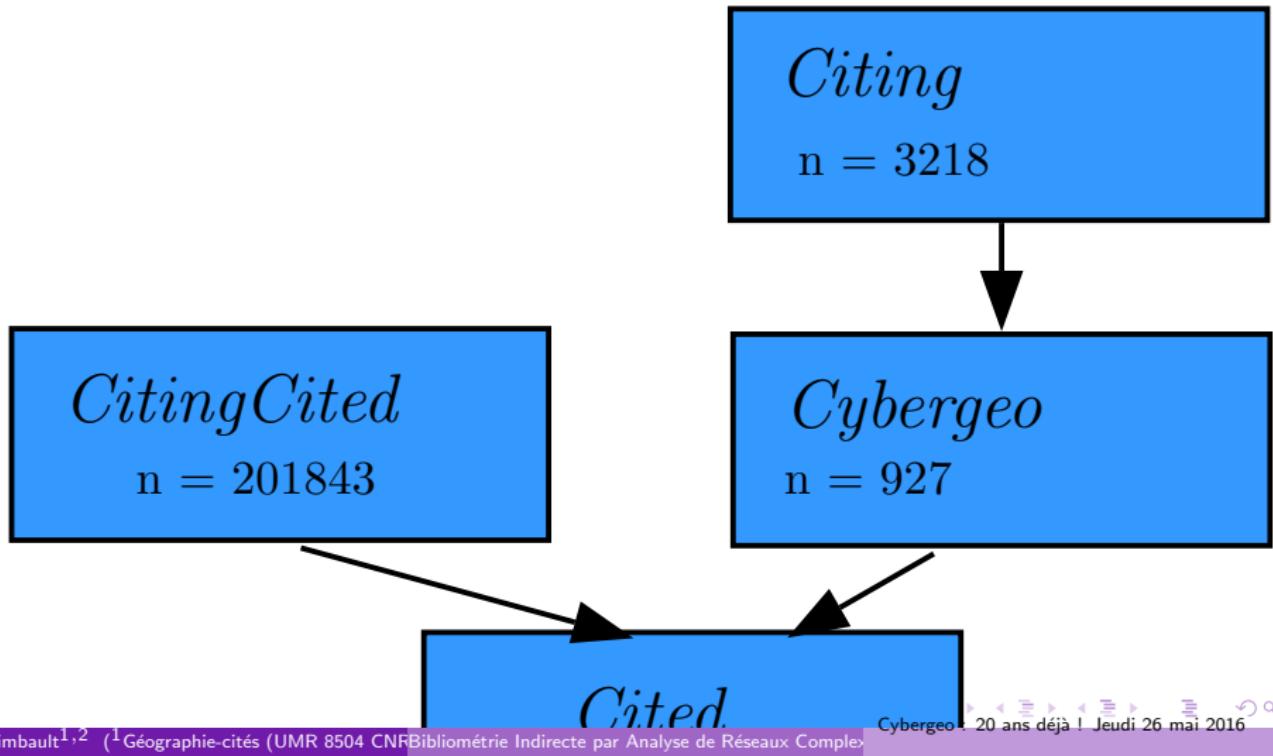




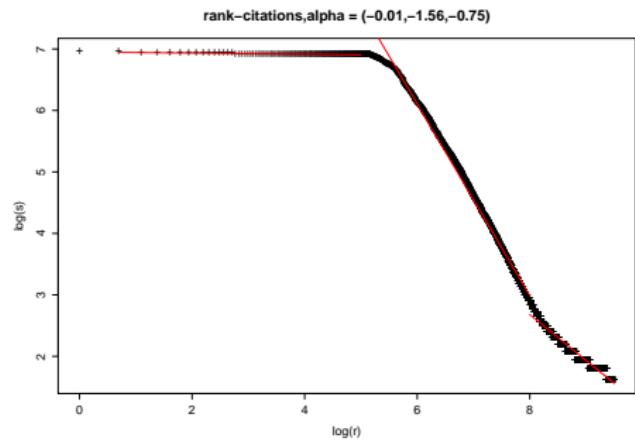
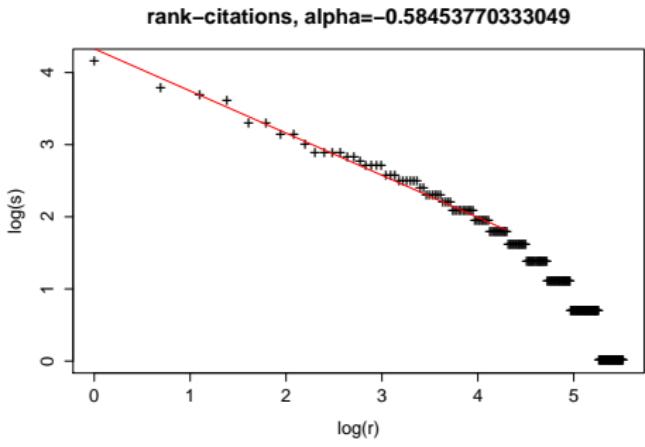
1996-2016 : 20 ans de cybergeo

1996-2016 : 20 years of cybergeo

Structure du Réseau

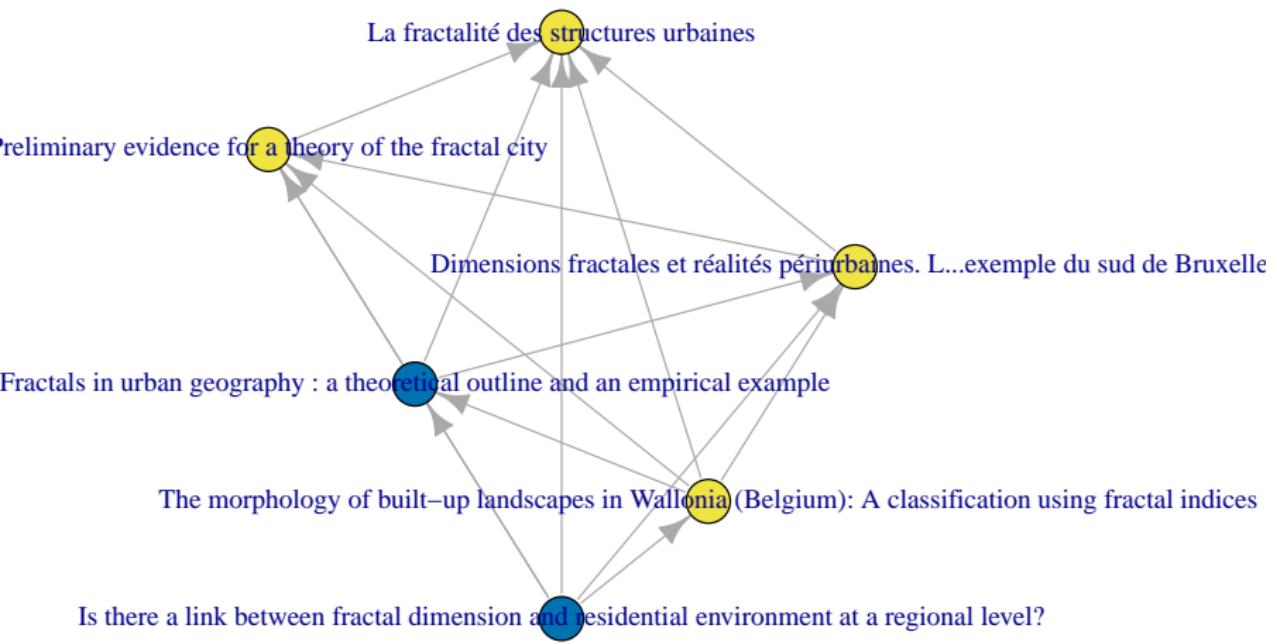


Degrés : Loi rang-taille



Gauche : Cybergéo ; Droite : Ensemble du réseau

Cliques



Is there a link between fractal dimension and residential environment at a regional level?

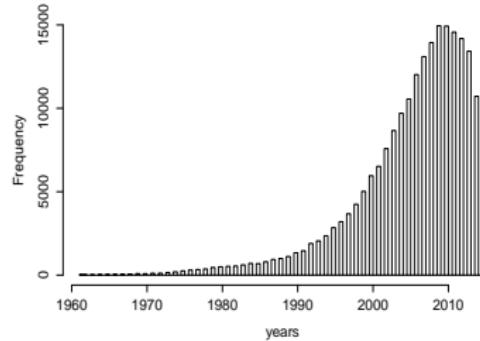
Réseau sémantique

Réseau sémantique. Collection des résumés/années/auteurs/mots-clés pour les 400000 références via l'API Mendeley → ~ 215000 références avec données complètes.

Statistiques

Langues : anglais 206607, français 4109, espagnol 2029, allemand 892, portugais 891, néerlandais 124, autres 182

Repartitions par années :





Extraction des mots-clés

Text-mining en python avec nltk [Bird, 2006], méthode adaptée de [Chavalarias and Cointet, 2013]

- Détection de la langue par *stop-words* (mots vides de sens)
- Parsing et tokenizing (isolation des mots) /pos-tagging (fonction des mots) /stemming (extraction de la racine) effectués différemment selon la langue :
 - ▶ Anglais : pos-tagger intégré à `nltk`, combiné à un PorterStemmer
 - ▶ Français ou autre : utilisation de TreeTagger [Schmid, 1994]
- Sélection des *n-grams* potentiels (avec $1 \leq n \leq 4$) : anglais
 $\cap\{NN \cup VBG \cup JJ\}$; français $\cap\{NOM \cup ADJ\}$
- Insertion en base pour extraction quasi-instantanée plus tard (10j → 5min)
- Estimation de la pertinence des *n-grams* selon répartition statistique des co-occurrences



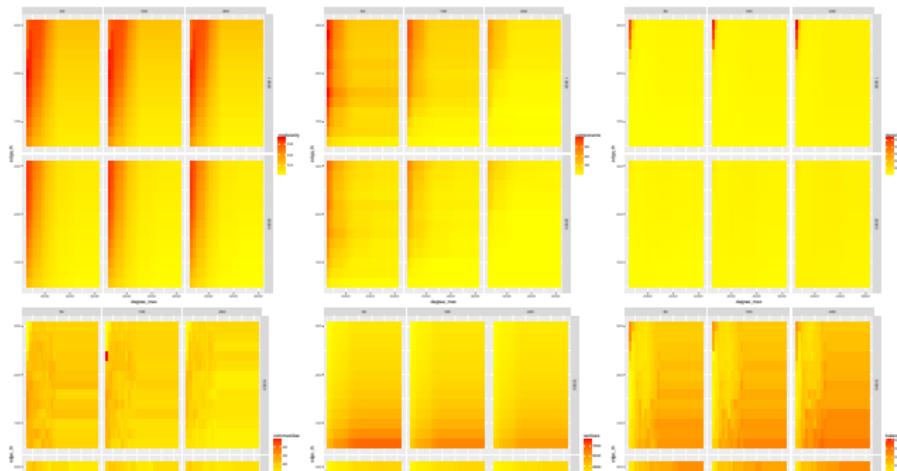
Construction du réseau sémantique

- **Noeuds** : Mots-clés avec la plus grande pertinence cumulée
- **Liens** : Co-occurrences pondérées
- Filtrage des liens en dessous d'un seuil ; ajustement manuel de mots parasites
- Détection de communautés par maximisation de modularité
[Blondel et al., 2008] après suppression des *hubs* (ex. model, space, structure, process)

Influence des paramètres

Importance du réglage fin des paramètres

- 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”
- Sensibilité aux conditions initiales : *Space matters* [Cottineau et al., 2015]





Domaines extraits

Communautés obtenues pour $\theta_V = 1200, \theta_E = 50$:

- 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



1996-2016 : 20 ans de cybergeo

1996-2016 : 20 years of cybergeo

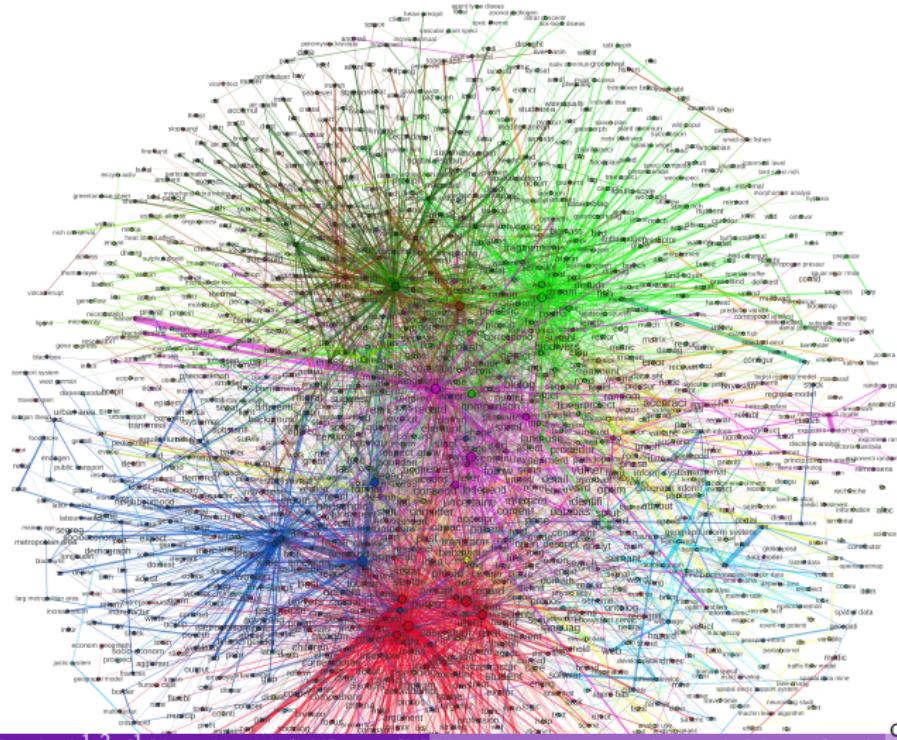
Domaines extraits (suite)

- 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

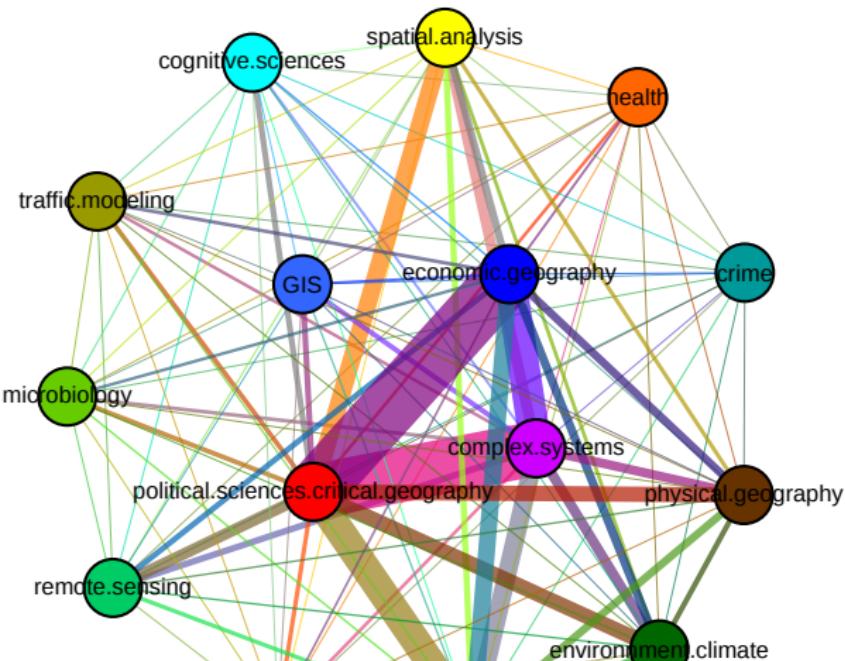


1996-2016 : 20 ans de cybergeo

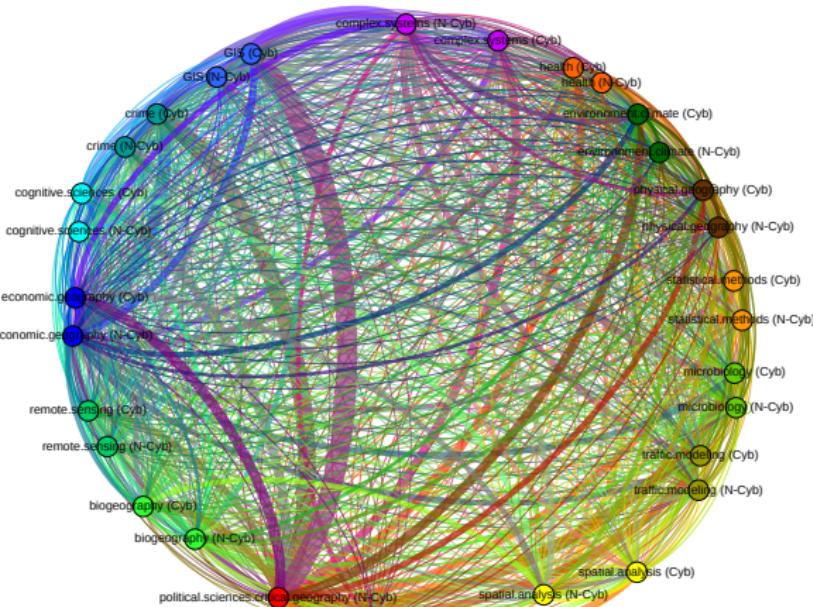
Réseau



Interdisciplinarité



Interdisciplinarité de citation

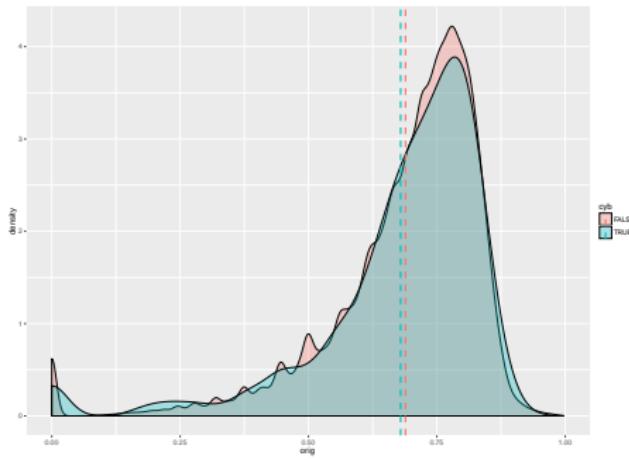


Degré d'interdisciplinarité par articles

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) :

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



Mean orig : 0.70

Conclusion

- Un environnement disciplinaire très varié et une interdisciplinarité affirmée
- 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 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)

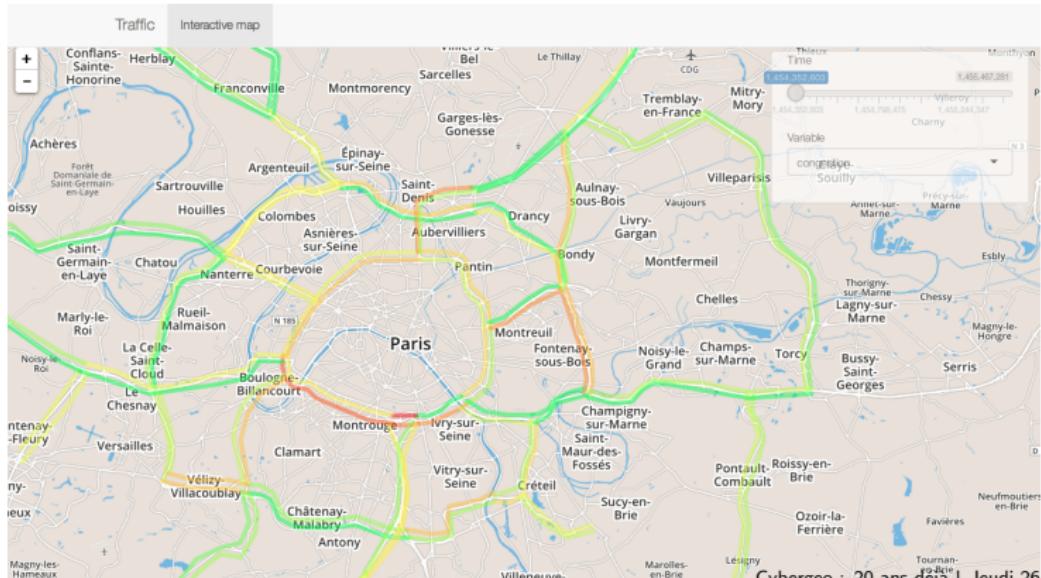
[?]



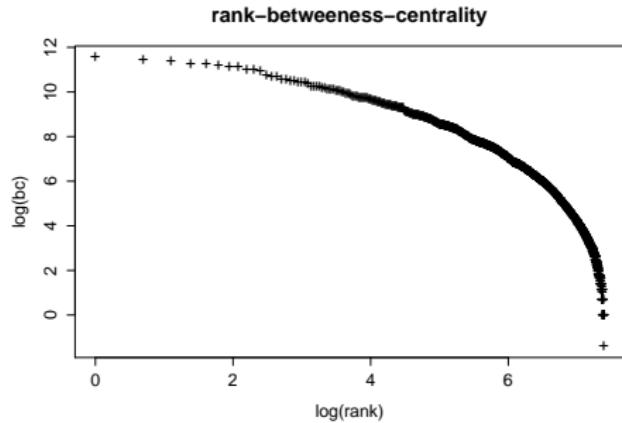
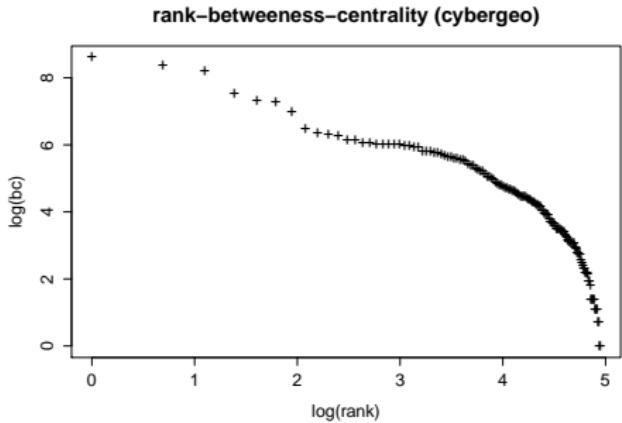
Collecte des données

Exemples en géographie (suite)

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



Centralité (citation)

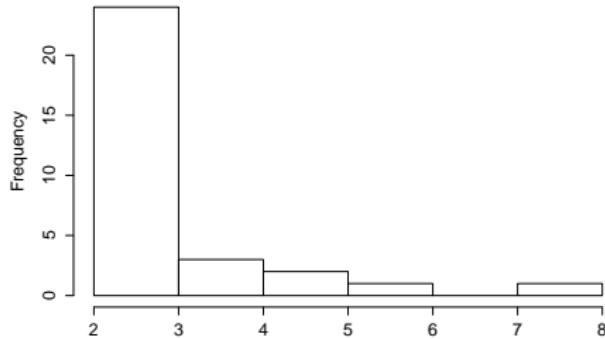


Centralités faibles (*rq* : impossibilité des clusters forts pour des citations car causalité temporelle). Gauche : Cybergeo ; Droite : Ensemble du réseau

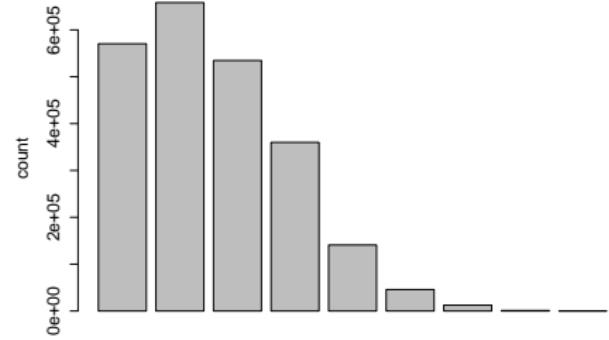
Clustering (citation)

Composante géante : plus de 99% des noeuds.

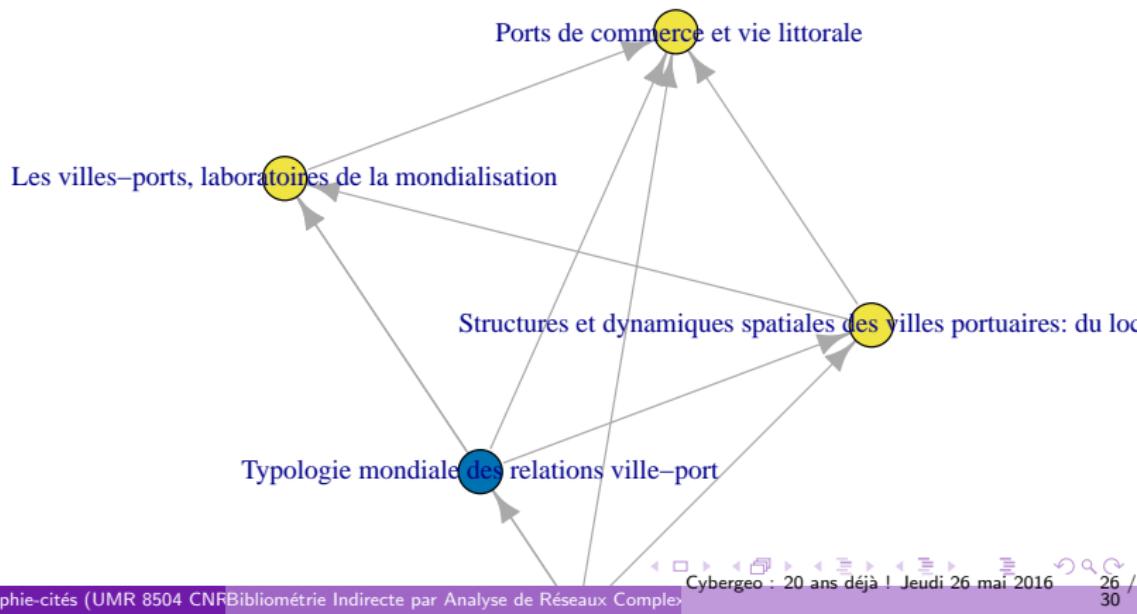
Weak clusters size without giant component



path length distribution



Cliques(citation)





Estimation de la pertinence

Estimation exacte de la pertinence via la répartition statistique des co-occurrences (score de χ^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 I

-  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).
Phylogenetic 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.



1996-2016 : 20 ans de cybergeo

1996-2016 : 20 years of cybergeo

References III



- 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.