Bibliométrie Indirecte par Analyse de Réseaux Complexes

J. Raimbault[1,2]

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

Séminaire *Cartha-Géo-Prisme*
Mercedi 17 février 2016

Données massives ?

Caractéristiques nécessaires possibles

- **Taille relative** : algorithmes et/ou structures de stockage non-conventionnels
- **Temps réel** : traitement d'un flux de données en temps réel
- **Hétérogénéité** : différentes sources, types, nature

Données massives et Systèmes Complexes

Système Complexe : grand nombre d'agents hétérogènes qui interagissent, emergence du comportement macro. $\rightarrow$ pas nécessairement "grand" selon l'epistemologie utilisée : cf trois corps de Poincaré.

Nouveaux moyens d'observation de systèmes complexes (sociaux entre autres) ? (cf tweetoscope de l'iscpif) Attention à ce qu'on observe !

Impact sociétaux : [Chavalarias, 2016] : Retroactions négatives et contre-productivité (ex de la dépendance au chemin du modèles de Polya)

P. Bourgine framework for Complex Adaptive Systems

Bourgine a récemment proposé un framework pour extraire des motifs internes aux Systèmes complexes adaptatifs. Par un théorème de représentation (Knight, 1975), tout processus discret stationnaire est un *Hidden Markov Model*
Etant donné la partition du système en états causaux (tels que $\mathbb{P}[future|A] = \mathbb{P}[future|B]$), un Réseau Bayesien Récurrent permet prédire l'état suivant par le présent uniquement de facon déterministe[Shalizi and Crutchfield, 2001]

$$(x_{t+1}, s_{t+1}) = F\left[(x_t, s_t)\right]$$

$\rightarrow$ *Estimation des Etats Cachés et de la fonction de récurrence par apprentissage profond fournit l'information complète sur les motifs de la dynamique*

Application en Géographie ?

- Hypothèse de stationarité obtenue par augmentation des etats (cf Bayesian Signal Processing)
- Utilisation de données hétérogènes et asynchrones pour bootstrapper des séries temporelles assez longues pour une convergence correcte de l'estimation ?

Cas d'étude

Revue scientifique *Cybergeo* : analyse bibliométrique par approches variées

→ Enjeu par rapport au positionnement de la revue : interdisciplinarité ; contre une bibliométrie quantitative pure en SHS

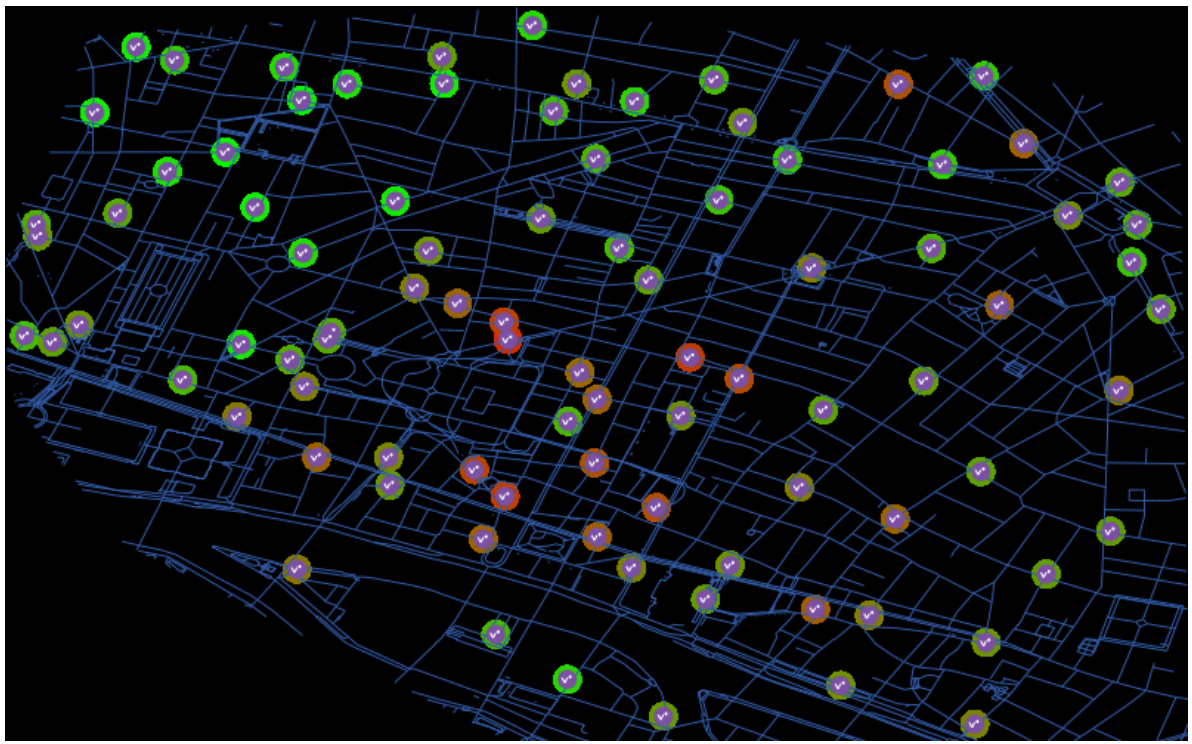
→ Elaboration d'une approche par *Hyperréseau* : croisement du réseau de citations au réseau sémantique.
Gain d'information par croisement des couches (démarche transversale, analogie avec construction scientifique : cf CS Roadmap)

→ Données difficiles d'accès : base à construire

Collecte des données

Crawling de données semi-ouvertes : exemples en géographie

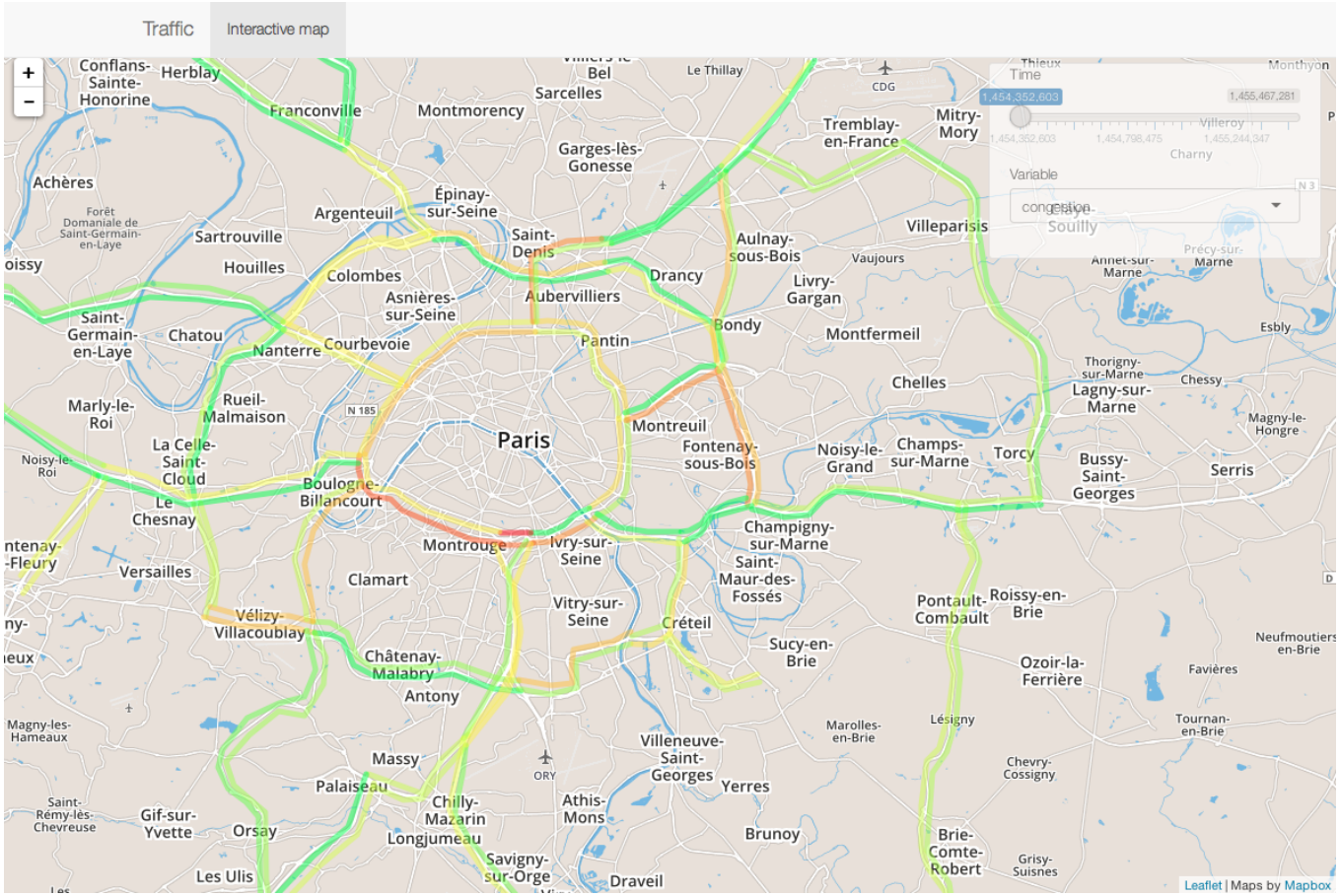
Données de mobilité : statuts des stations Vlib en temps réel (API) [Raimbault, 2015]



Collecte des données

Exemples en géographie (suite)

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



Collecte des données

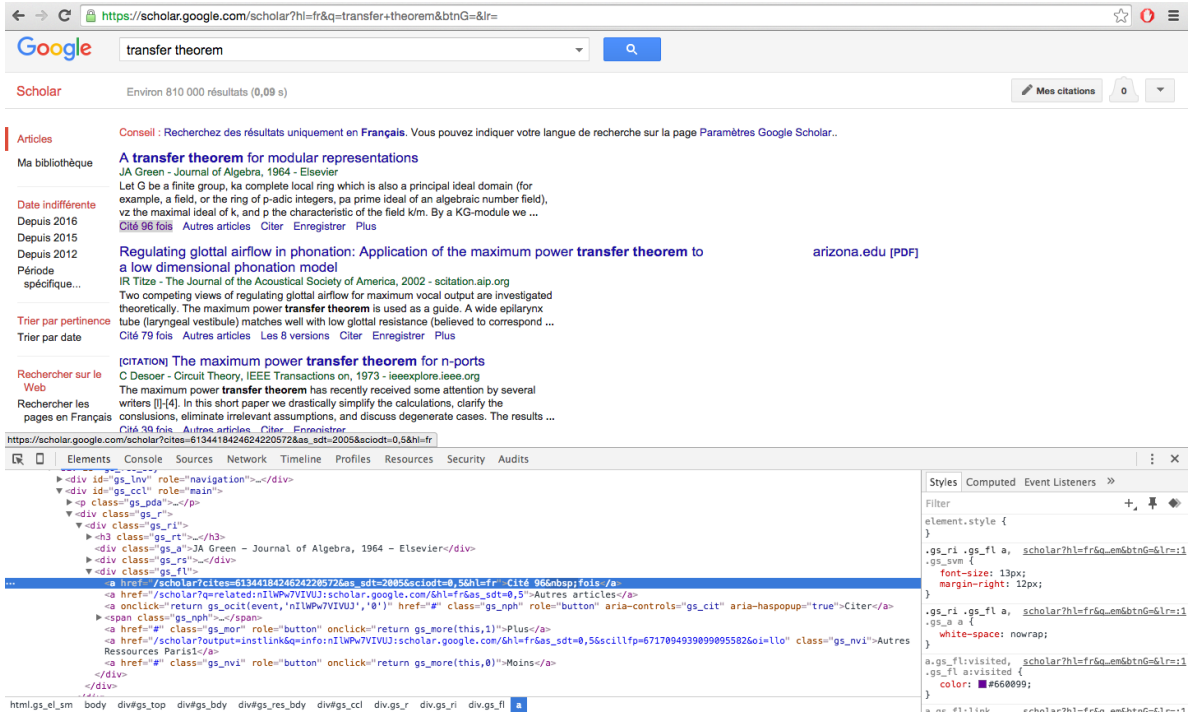
Données des articles : bases de fonctionnement (production de la revue)

- Structuration : rétro-ingénierie de la base relationelle ; architecture ; extraction.
- Nettoyage : Formatage des textes ; encodage.
- Consolidation : sources différentes sans id de lien.

Collecte des données

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]

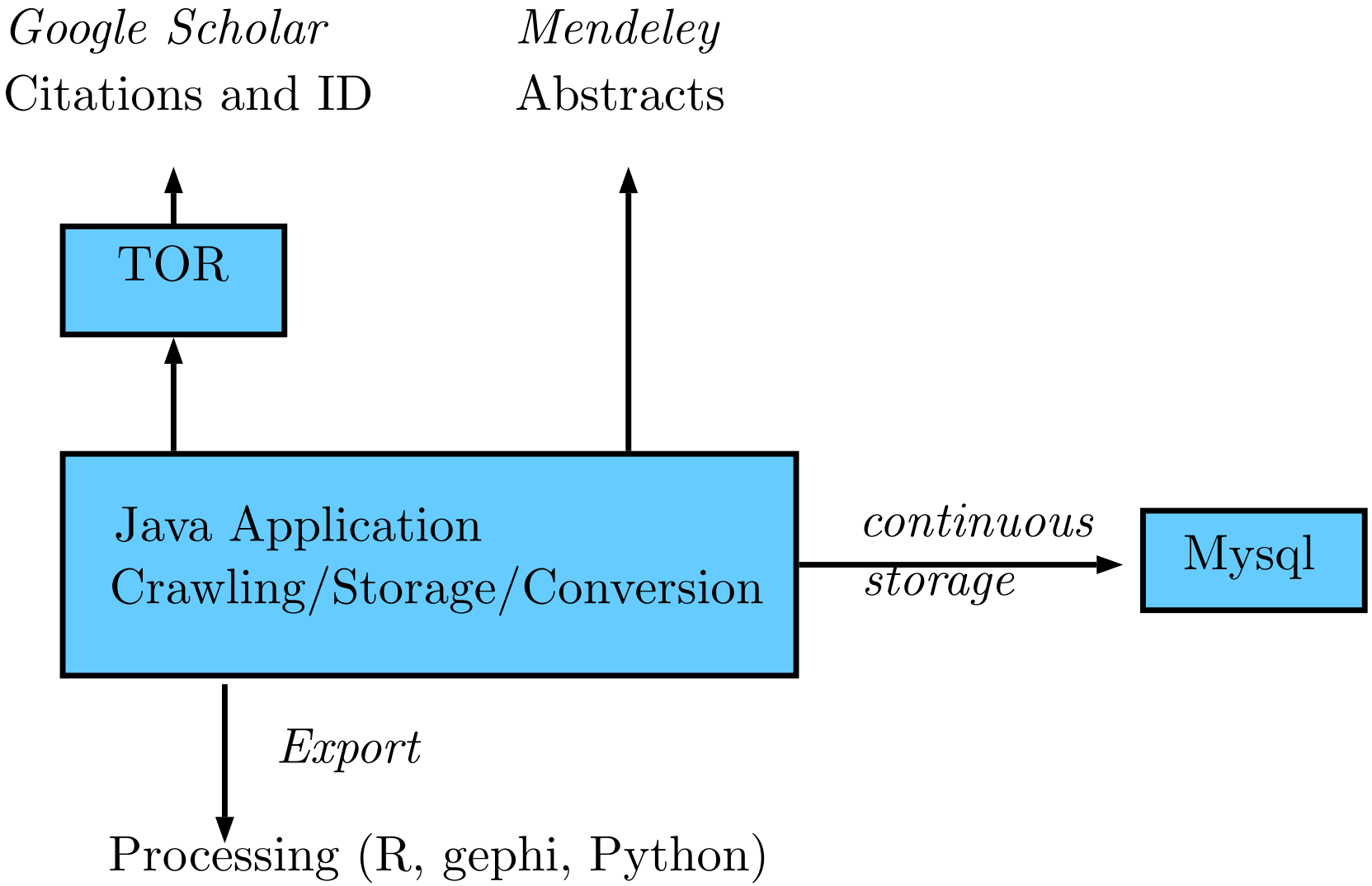


Collecte des données

Données textuelles : besoin des résumés pour l'ensemble des références liées

$\rightarrow$ utilisation de l'API Mendeley [Mendeley, 2015] (gratuite mais non ouverte).

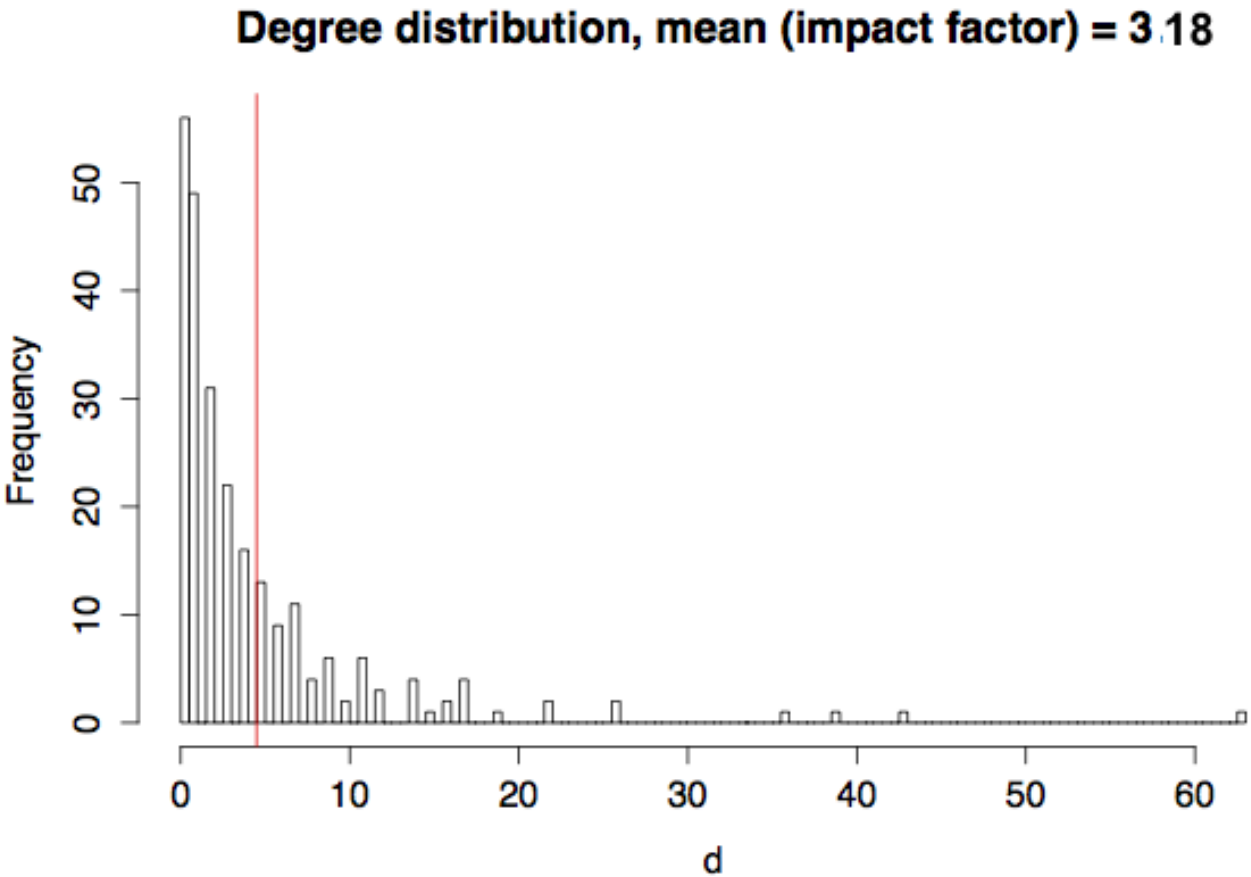
Architecture de collecte des données



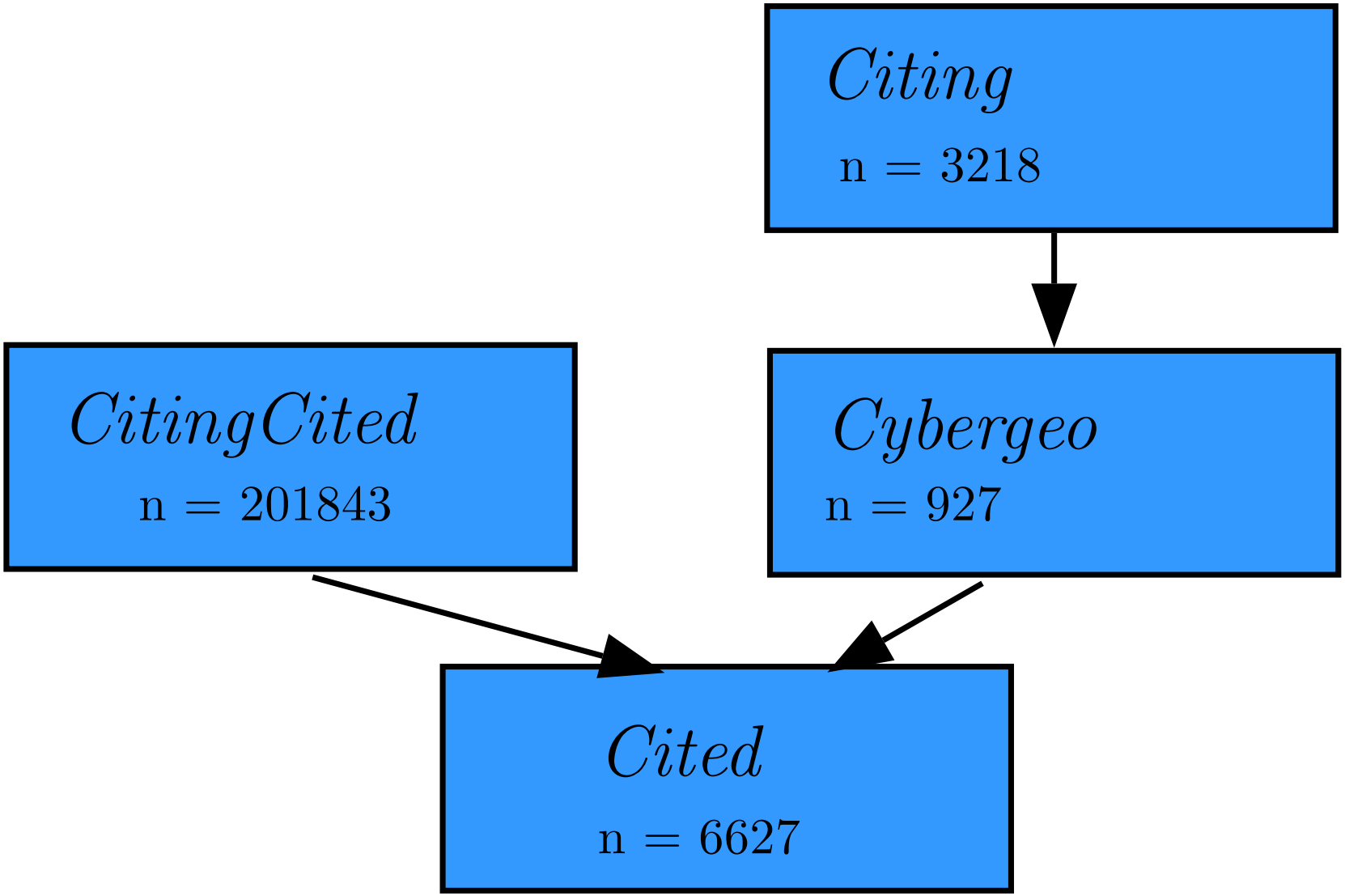
Caractéristiques du réseau

→ Après raffinement, $\simeq$ 947 références de cybergeo exploitables, sur $\simeq$ 1200 ; certaines inexistantes, d'autres mal enregistrées sur scholar

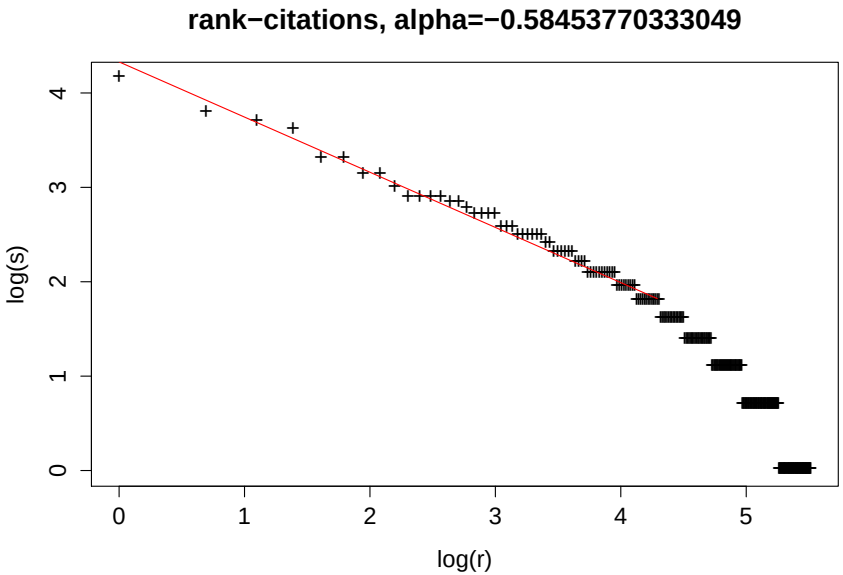
→ *418670 Noeuds et 570352 Liens ; Diamétre : 9 ; Densité : 3.25E-6 ; degré moyen : 2.724284*

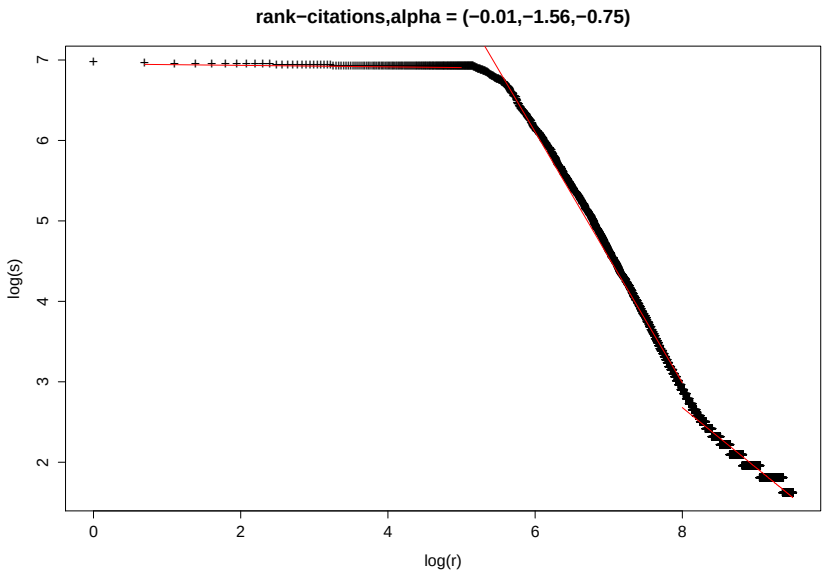


Structure du Réseau



Degrés : Loi rang-taille

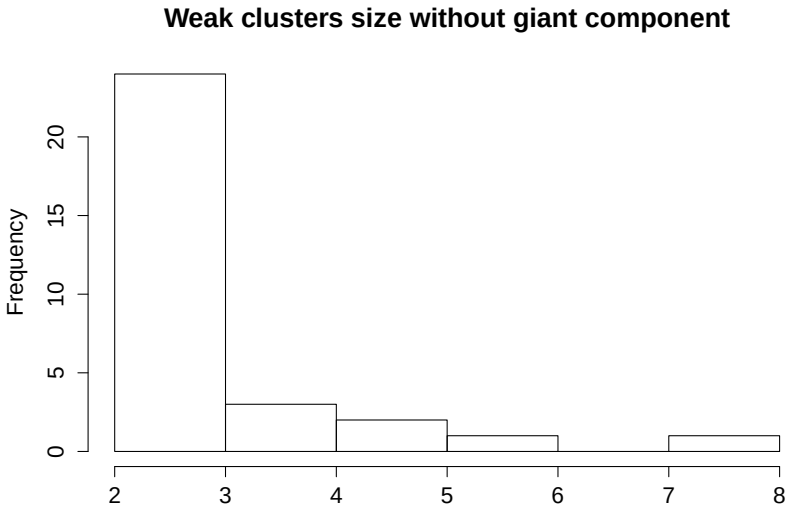


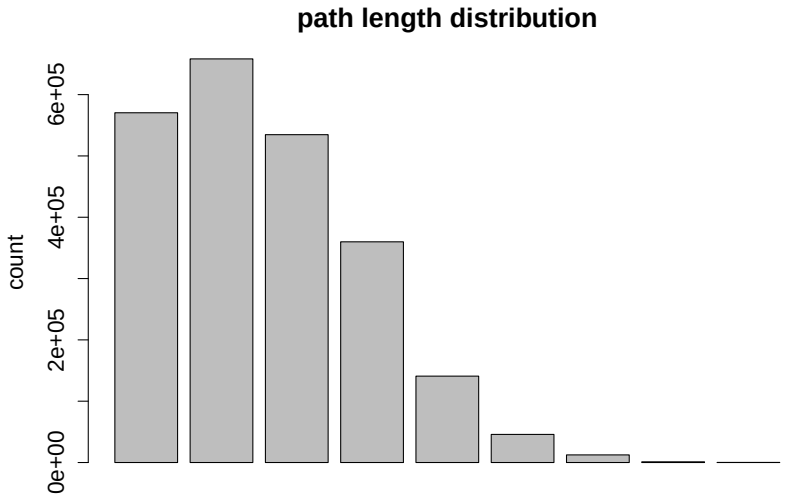


Gauche : Cybergéo ; Droite : Ensemble du réseau

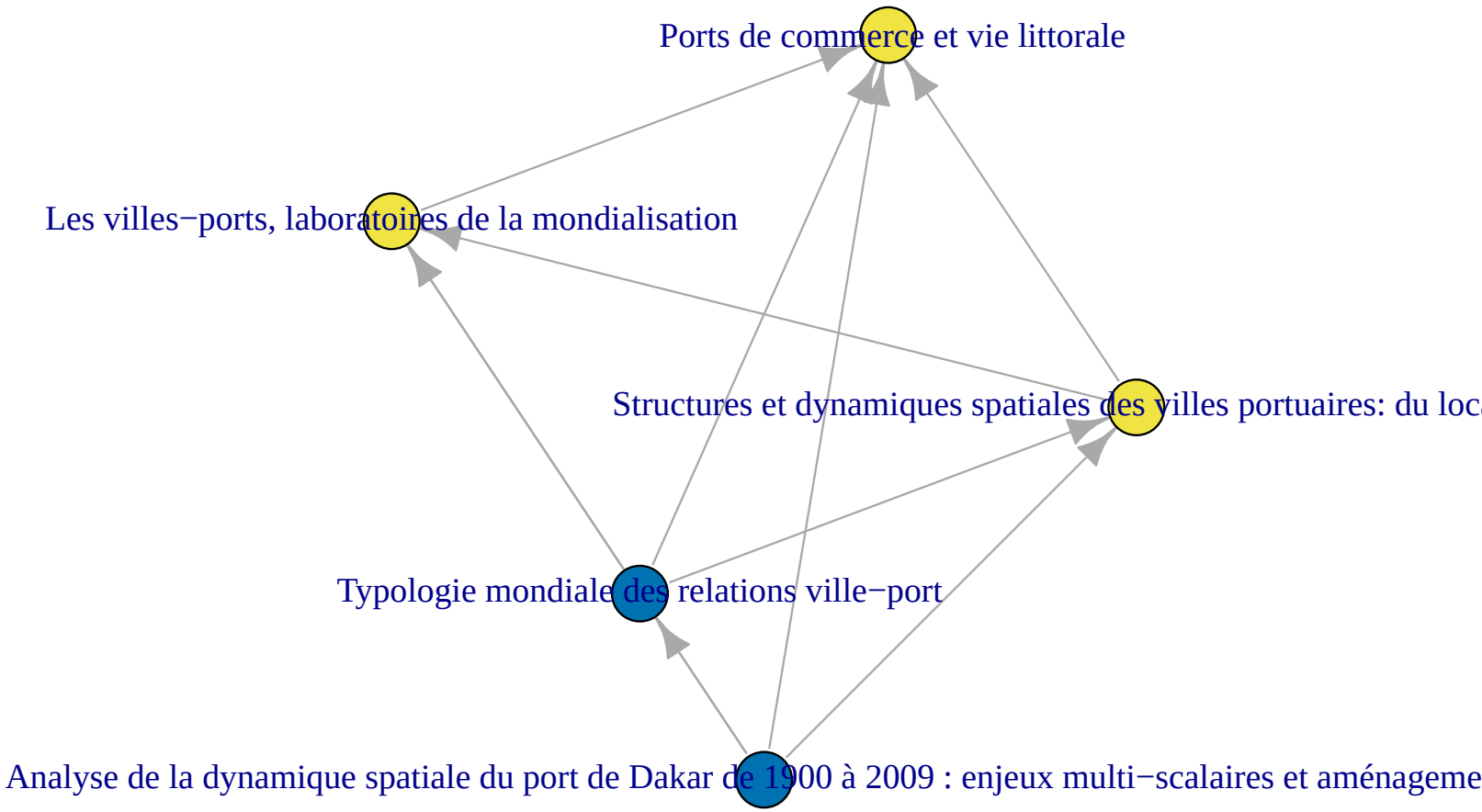
Clustering

Composante géante : plus de 99% des noeuds.

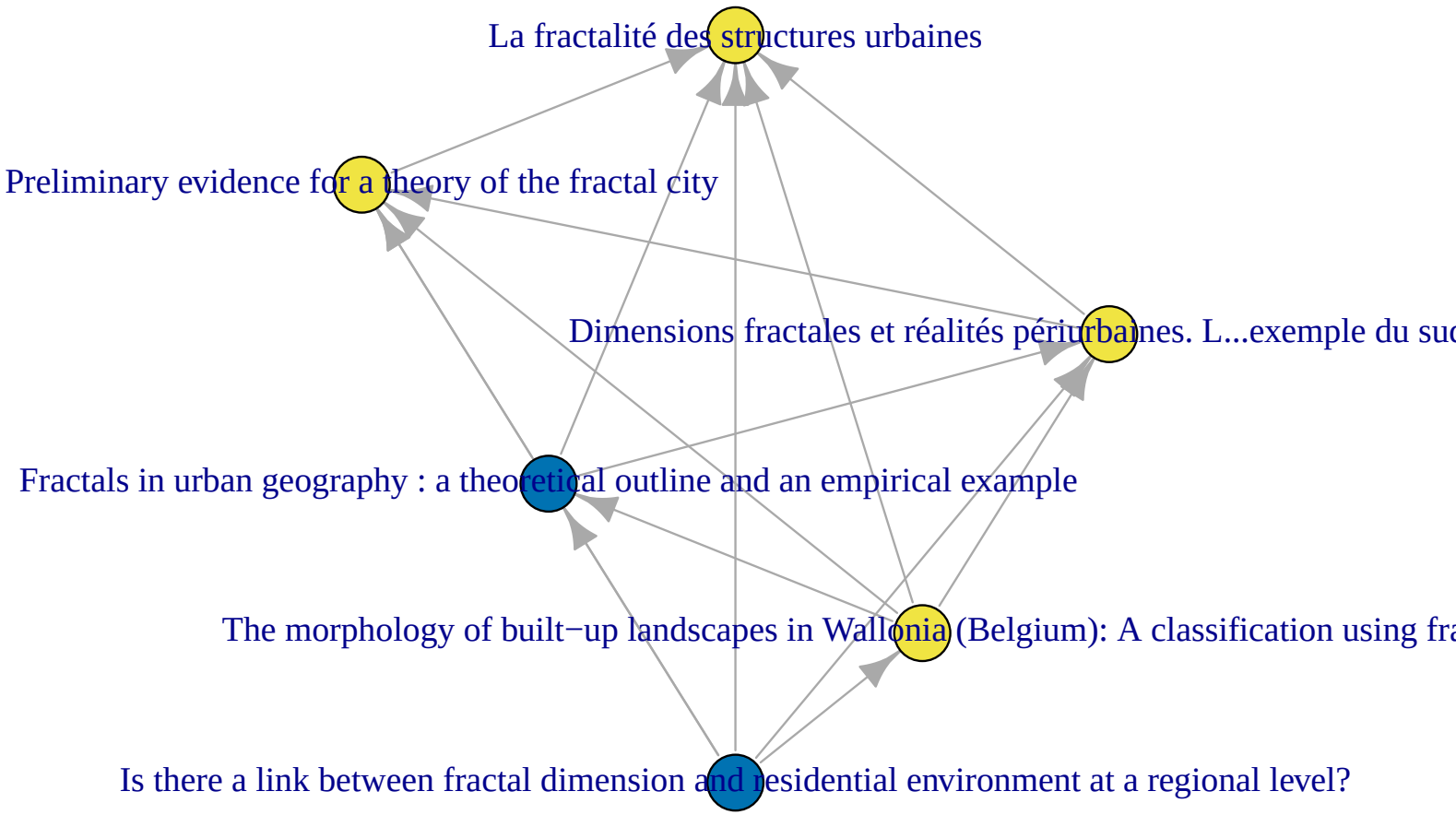




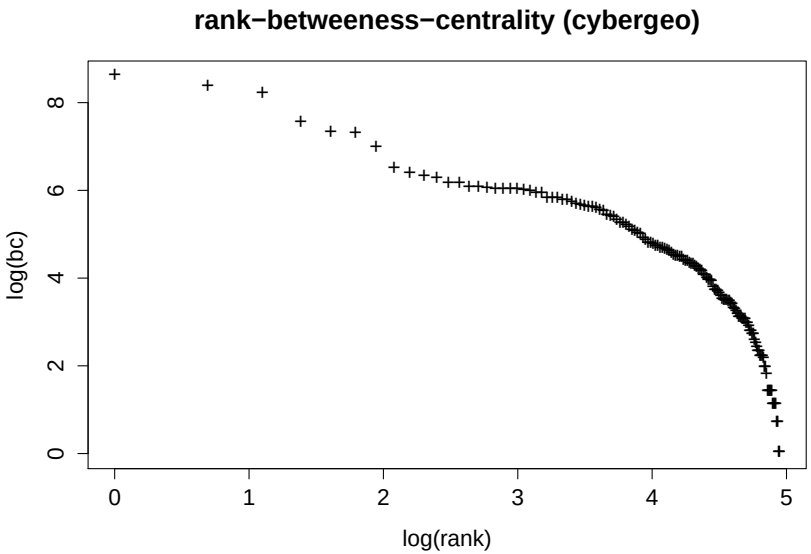
Cliques

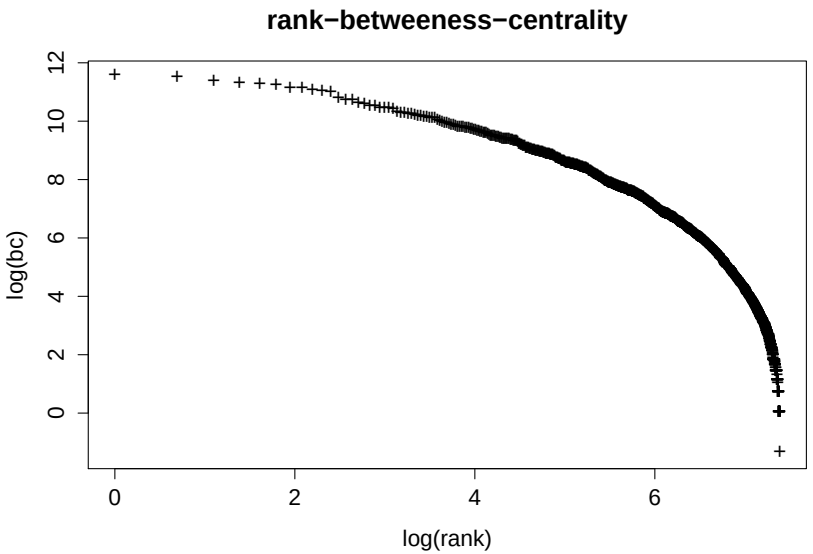

Ports de commerce et vie littorale
Les villes−ports, laboratoires de la mondialisation
Structures et dynamiques spatiales des villes portuaires: du loc
Typologie mondiale des relations ville−port
Analyse de la dynamique spatiale du port de Dakar de 1900 à 2009 : enjeux multi−scalaires et aménageme

Cliques


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

Centralité





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

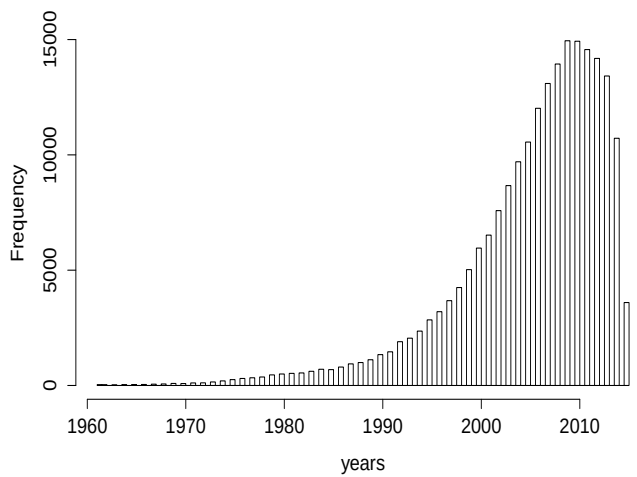
Réseau sémantique

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

Statistiques

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

Repartitions par années :



Extraction des mots-clés brute

Text-mining en python avec `nltk` *[Bird, 2006]*, méthode adaptée de [Chavalarias a

- Detection 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]
- Selection des n-grams potentiels (avec $1 \leq n \leq 4$) : anglais $\bigcap\{NN \cup VBG \cup JJ\}$; franais $\bigcap\{NOM \cup ADJ\}$
- Insertion en base pour extraction quasi-instantane plus tard (10j $\rightarrow$ 5min)

Estimation de la pertinence

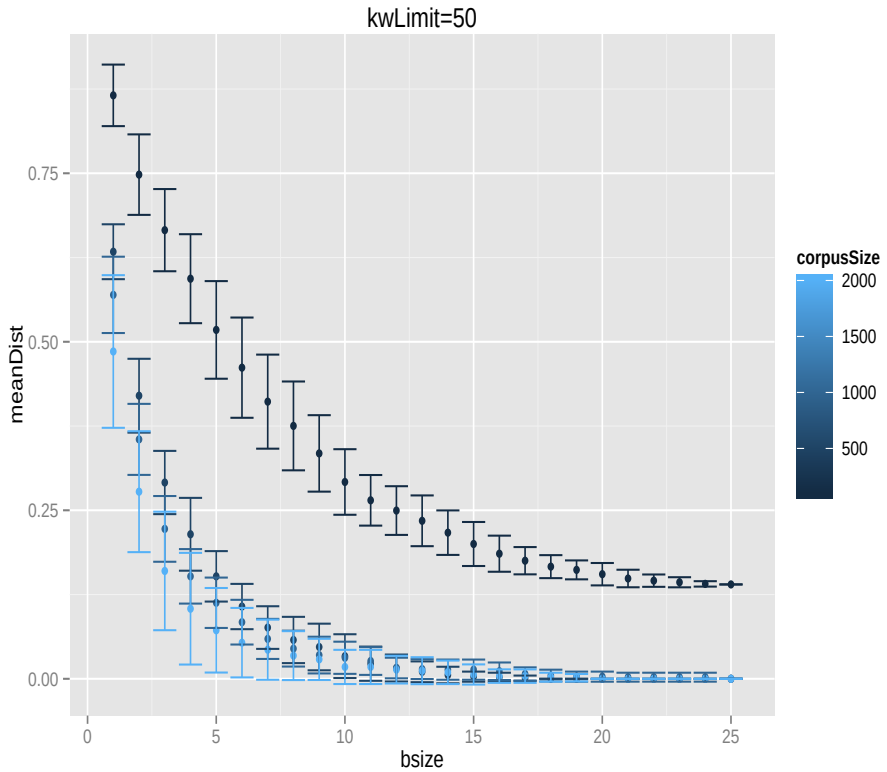
Estimation exacte de la pertinence via la repartition statistique des co-occurrences (score de χ^2) : *termhood* définie, avec M_{ij} nombre d'articles o i et j apparaissent simultanment,

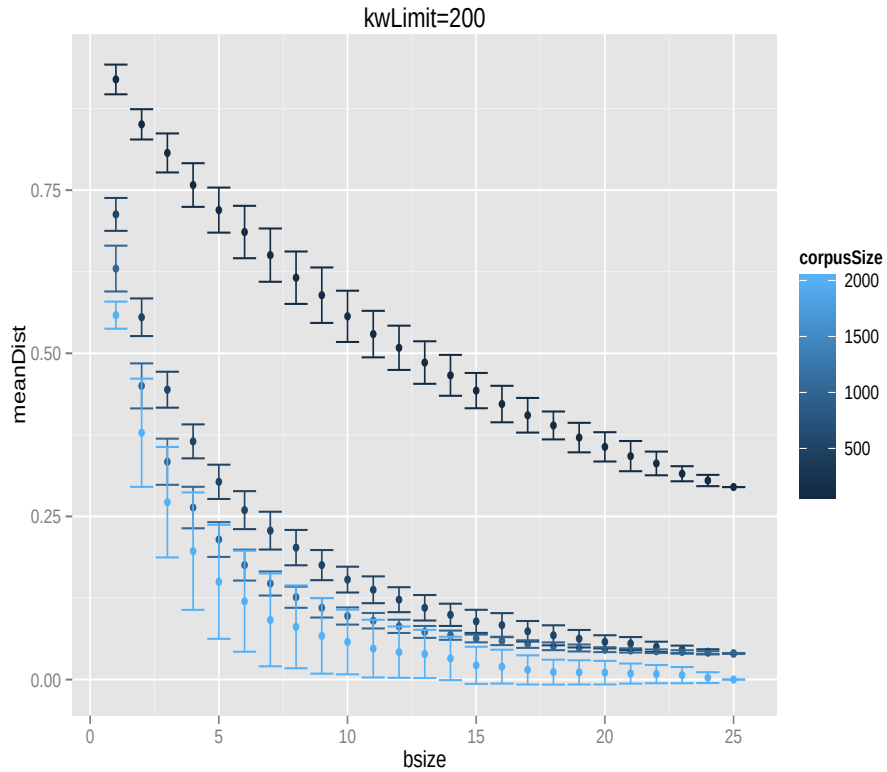
$$t_i = \sum_{j \neq i} \frac{(M_{ij} - \sum_k M_{ik} \sum_k M_{jk})^2}{\sum_k M_{ik} \sum_k M_{jk}}$$

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

Estimation de la pertinence par bootstrap

→ Estimation par *booootstrap* sur des échantillons du corpus : moyenne de la *termhood* sur B échantillons de taille C, avec nombre de mots clés K_L





Construction du réseau sémantique

Noeuds : Mots-clés avec la plus grande pertinence cumulée

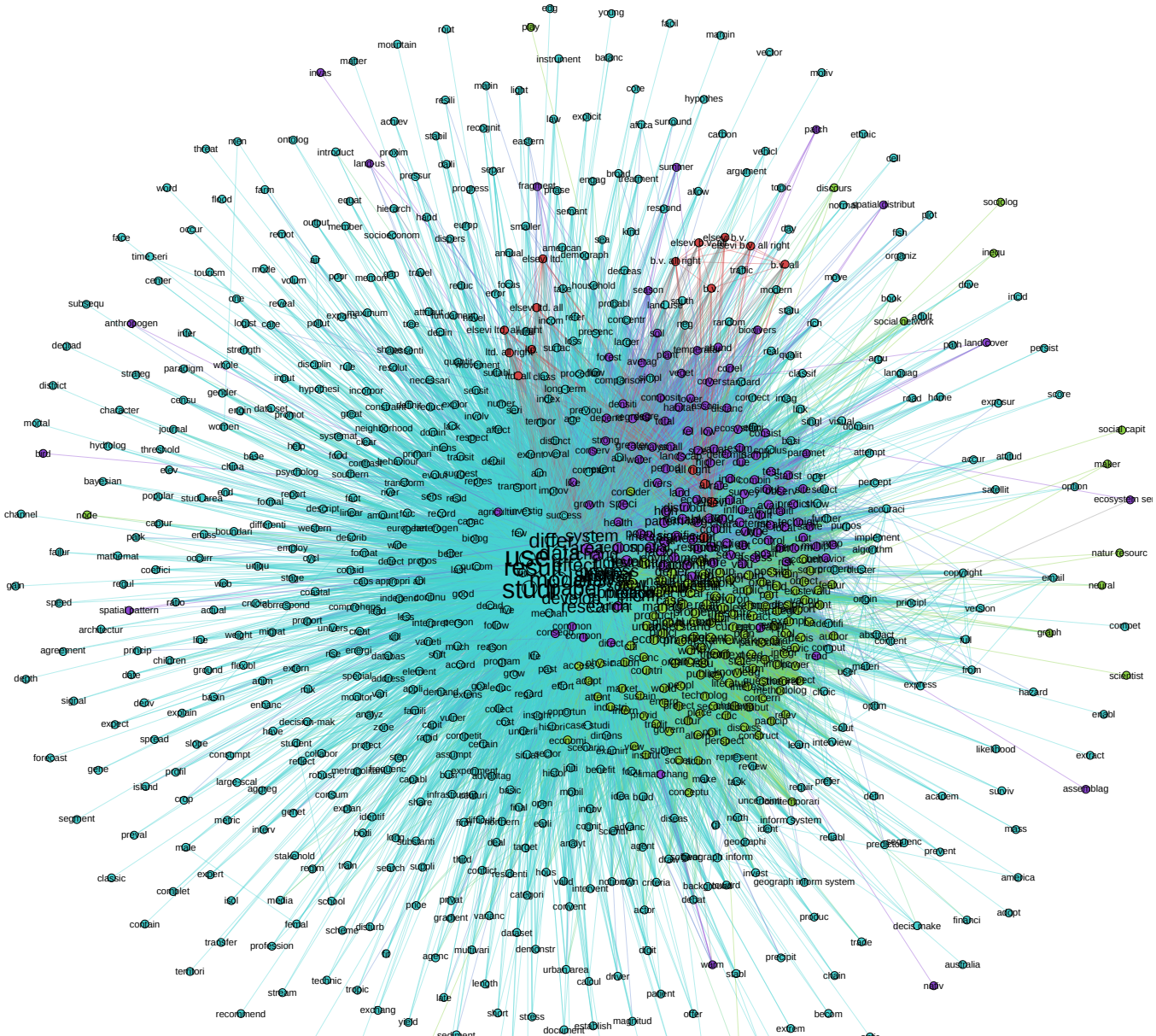
Liens : Co-occurrences pondérées

Filtrage des liens en dessous d'un seuil ; ajustement manuel de mots parasites

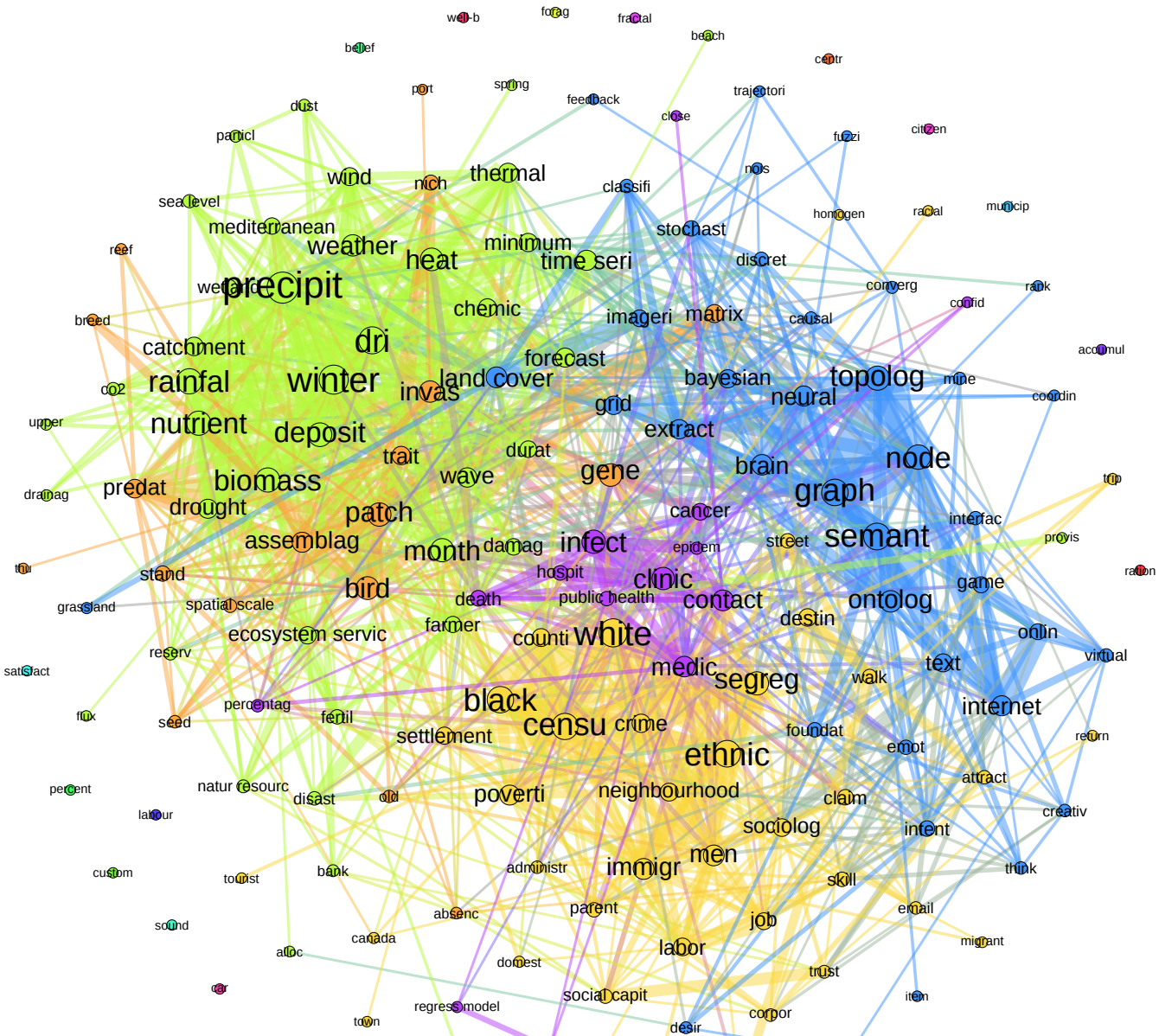
Detection de communautés par maximisation de modularité [Blondel et al., 2008] après suppression des *hubs* (`model`, `space`, `structure`, `process`)

Spatialisation de *Fruchterman-Reingold*

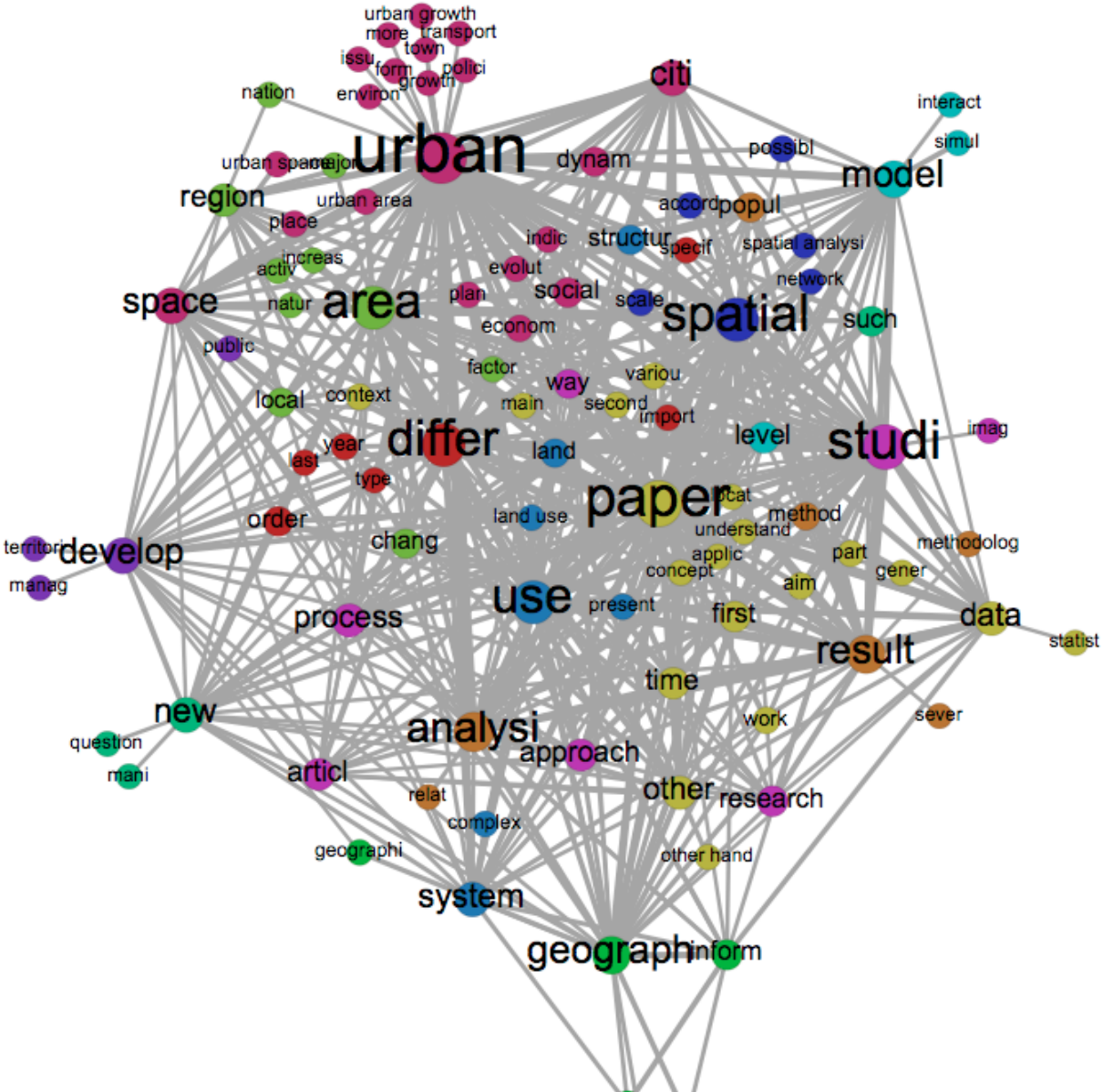
Réseau sémantique : réseau brut



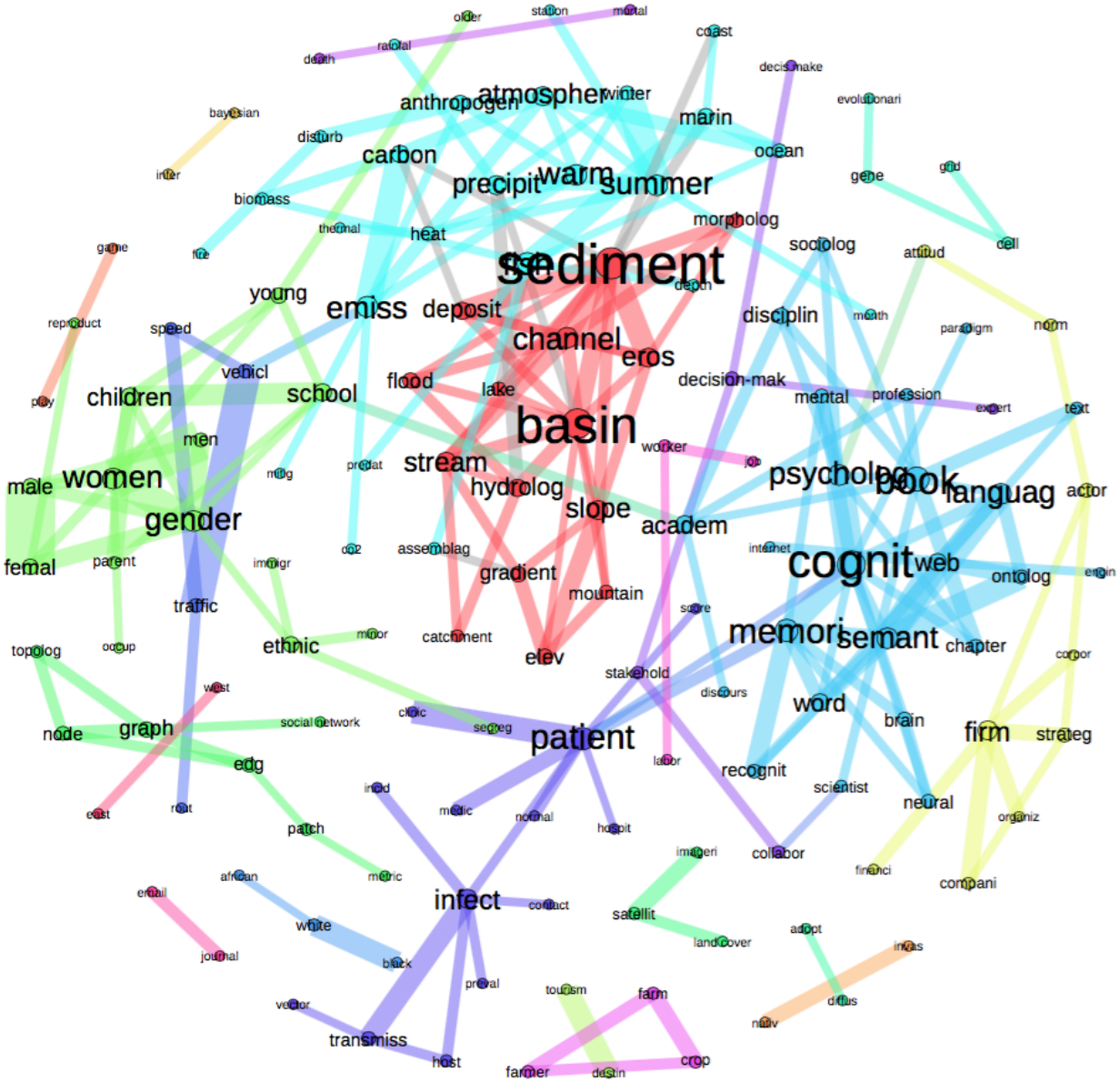
Réseau sémantique : corpus complet (avec filtrage)



Réseau sémantique : corpus cybergeo



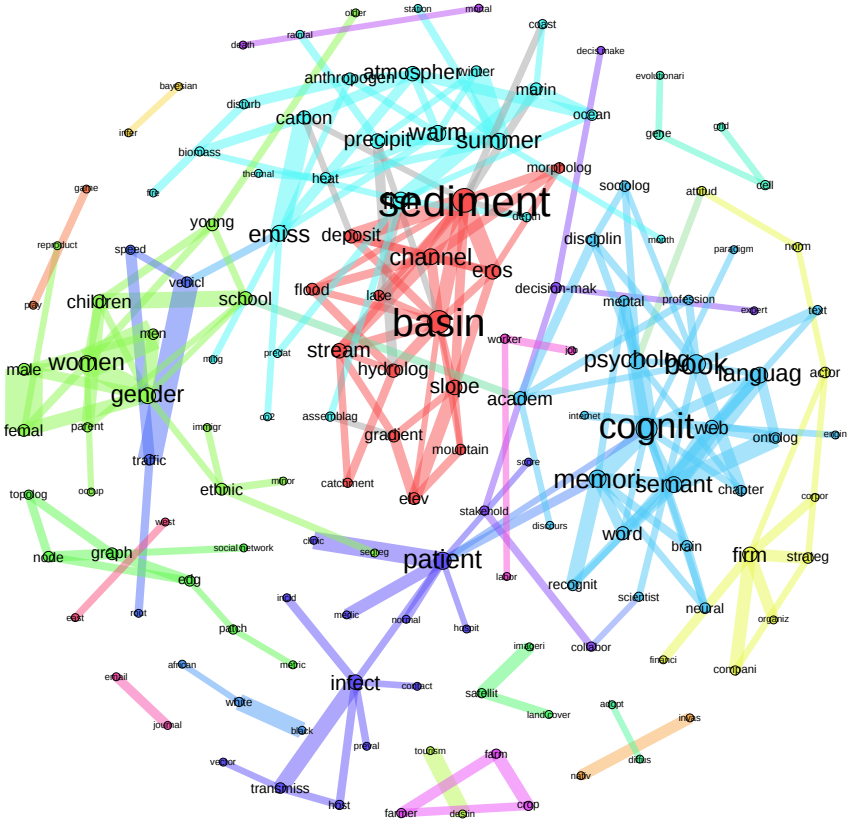
Retour au corpus complet



Influence des paramètres

Importance du règlage fin des paramètres
→ Sensibilité des modèles **et** traitements de données aux parametres. 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]

Application



Domaines extraits

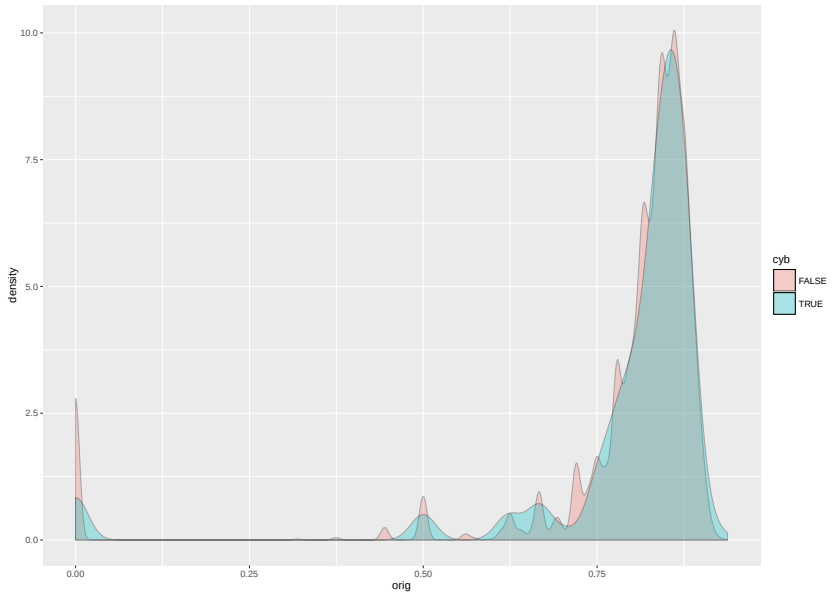
Comunautés obtenues pour $\theta_V = 1000, \theta_E = 200$ *:*

- Hydrology : `water, basin, river, capac`
- Traffic : `traffic, road, vehicl`
- Biogeography : `habitat, soil, veget, ecosystem`
- Political Science : `polit, cultur, societi, debat`
- Economy : `market, economi, privat, competit, industri`
- Transportation : `transport, travel`
- Teledetection : `cluster, imag, classif, satellit`
- Education : `educ, age, student, school`
- Health : `diseas, infect`
- GIS : `gi, geograph inform system`
- Social geography : `neighborhood, resid`

Application : degré d'interdisciplinarité

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) :

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



Mean orig : 0.79

Degré d'interdisciplinarité

Aggregation au niveau de la revue : originalité des thèmes abordés dans l'ensemble de la revue

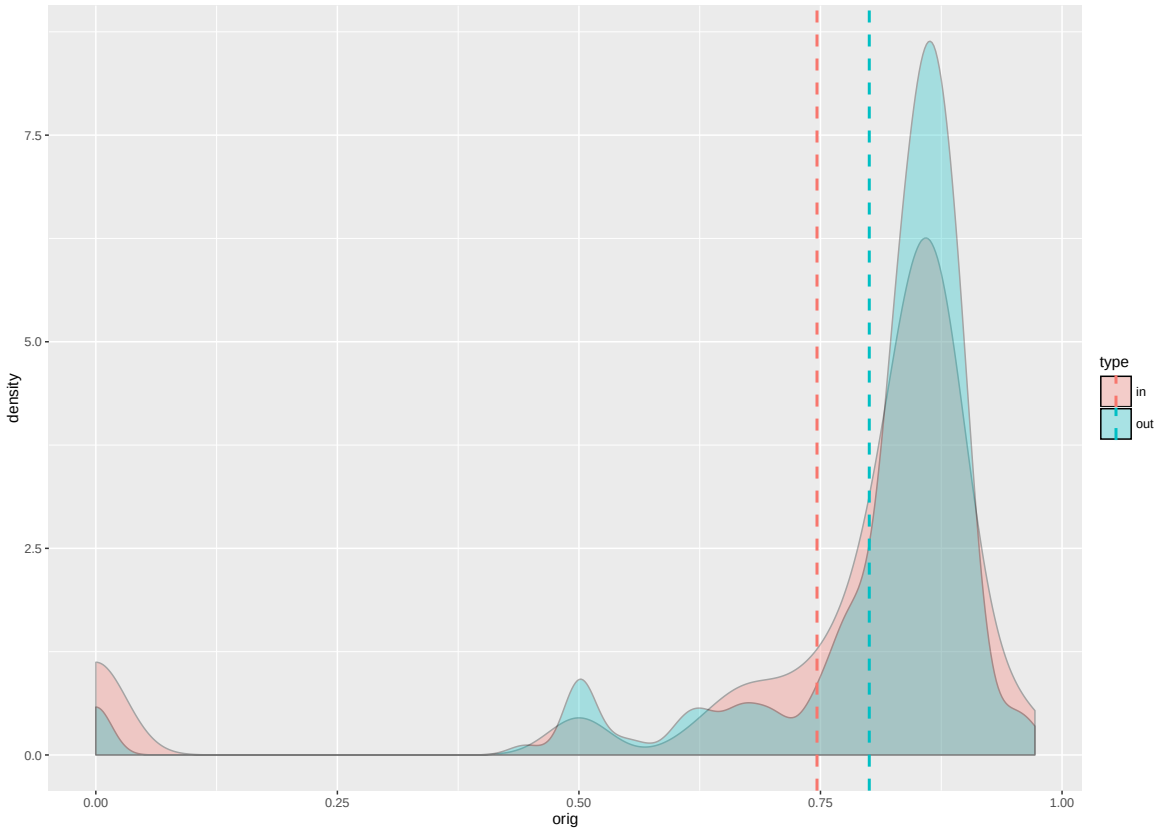
$$O = 1 - \sum_i \left[\frac{1}{K} \sum_k p_i^{(k)} \right]^2$$

$\rightarrow$ 0.890 pour cybergeo (0.8933 pour un modèle nul par tirage aléatoire)
Besoin d'autres revues pour comparaison : retour à la collecte de données

Interdisciplinarité au second ordre

Croisement des couches de l'hyperréseau.

$\rightarrow$ Originalité de l'ensemble des voisins (entrants ou sortants) dans le réseau de citation



Perspectives

Attention à la tentation des *big data* et de la simulation à outrance : garder un ancrage théorique (poser les bonnes questions) et méthodologique.

Exemple : travail en cours sur interactions Ville/Transports : données "simples" et classiques (densité population et OSM), statiques, fournissent de l'information sur processus dynamiques sous-jacents. Nécessité du cadre théorique (théorie évolutive des villes) et du travail méthodologique pour relier statique-dynamique.

Conclusion (à retenir)

- "Big Data" plus que relatif
- Des données partout, à vous de les collecter et créer des bases **ouvertes** (pas de science sans ouverture : multimodeling et open science)
- Géographie à la pointe de par les connaissances déjà présentes : à vous de jouer !

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. (2016).
The unlikely encounter between von foerster and snowden: When second-order cybernetics sheds light on societal impacts of big data.

Big Data & Society, 3(1):2053951715621086.

References II


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


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.

References III


Raimbault, J. (2015).
User-based solutions for increasing level of service in bike-sharing transportation systems.
In Complex Systems Design & Management, pages 31–44. Springer.


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.


Shalizi, C. R. and Crutchfield, J. P. (2001).
Computational mechanics: Pattern and prediction, structure and simplicity.
Journal of statistical physics, 104(3-4):817–879.