Scientometrics manuscript No.
(will be inserted by the editor)

Exploration of an Interdisciplinary Scientific Landscape

Juste Raimbault[1,2]

Received: date / Accepted: date

Abstract Patterns of interdisciplinarity in science can be quantified through diverse complementary dimensions. This paper studies as a case study the scientific environment of a generalist journal in Geography, *Cybergeo*, in order to introduce a novel methodology combining citation network analysis and semantic analysis. We collect a large corpus of 200,000 articles with their abstracts and the corresponding citation network. Relevant keywords are extracted for each article through text-mining, allowing to construct a semantic classification. We show the complementarity of the citation and of the semantic classifications and their associated interdisciplinarity measures. The tools we develop accordingly are open and reusable for similar large scale studies of scientific environments.

Keywords Citation Network · Semantic Analysis · Interdisciplinarity · Geography

Introduction

The development of interdisciplinary approaches is increasingly necessary for most of disciplines, both for further knowledge discovery but also societal impact of discoveries, as it was recently by the special issue of Nature (Nature, 2015). Banos (2013) suggests they must occur within a subtle spiral between and inside disciplines. An other way to understand this phenomenon is through the emergence of vertically integrated fields conjointly with horizontal questions as detailed in the Complex Systems roadmap (Bourgine et al

J. Raimbault
[1] UMR CNRS 8504 Géographie-cités
[2] UMR-T IFSTTAR 9403 LVMT
Tel.: +33140464000
E-mail: juste.raimbault@polytechnique.edu

(2009)). There are naturally ongoing debates on what is exactly interdisciplinarity (many other terms such as trans-disciplinarity, cross-disciplinarity also exist) and it actually depends of involved domains : recent hybrid disciplines (see e.g. the ones underlined by Bais (2010) such as astro-biology) are a good illustration of the case where entanglement is strong and new discoveries are vertically deep, whereas more loose fields such as "urbanism", which have no precise definition and where integration is by essence horizontal, are an other illustration of how transversal knowledge can be produced. Interaction between disciplines are not always smooth, as shows the misunderstandings when urban issues were recently introduced to physicists as Dupuy and Benguigui (2015) recalls.

These concerns are part of an understanding of processes of knowledge production, i.e. the *Knowledge of the knowledge* as Morin (1986) puts it, in which evidence-based perspectives, involving quantitative approaches, play an important role. These paradigms can be understood as a *quantitative epistemology.* Quantitative measures of interdisciplinarity would therefore be part of a multidimensional approach of the study of science that is in a way "beyond bibliometrics" (Cronin and Sugimoto, 2014). The focus of this paper is positioned within this stream of research. We first review existing approaches to the measure of interdisciplinarity.

The possible methods for quantitative insights into epistemology are numerous. A good illustration of the variety of approaches is given by network analysis Using citation network features, a good predicting power for citation patterns is for example obtained by Newman (2013). Co-authorship networks can also be used for predictive models (Sarigöl et al, 2014). A multilayer network approach was proposed in Omodei et al (2017), using bipartites networks of papers and scholars, in order to produce measures of interdisciplinarity using generalized centrality measures. Disciplines can be stratified into layers to reveal communities between them and therein collaboration patterns (Battiston et al, 2015). Keyword networks are used in other fields such as economics of innovation: for example, Choi and Hwang (2014) proposes a method to identify technological opportunities by detecting important keywords from the point of view of topological measures. In a similar manner, Shibata et al (2008) uses topological analysis of the citation network to detect emerging research fronts.

We develop in this paper a case study coupling citation network exploration and analysis with text-mining, aiming at mapping the scientific landscape in the neighborhood of a particular journal. The choice of the journal yield several challenges and issues that make it particularly relevant for our study. It is an electronic journal in theoretical and quantitative geography, named *Cybergeo*[1]. First of all, the discipline of Geography is very broad and by essence interdisciplinary: the spectrum ranges from Human and Critical geography to physical geography and geomorphology. Secondly, bibliographical data is difficult to obtain, raising the concern of how the perception of a scientific landscape may be shaped by actors of the dissemination and thus far from

[1] http://cybergeo.revues.org/

objective, making technical solutions as the ones consequently developed here crucial tools for an open and neutral science. Finally it makes a particularly interesting case study as the editorial policy is generalist and concerned with open science issues such as peer-review ethics transparency Wicherts (2016), data and model practices, etc.

Our approach combine semantic communities analysis (as done in Palchykov et al (2016) for papers in physics but with keyword extraction ; Gurciullo et al (2015) analyses semantic networks of political debates) with citation network to extract e.g. interdisciplinarity measures. Our contribution differs from the previous works quantifying interdisciplinarity as it does not assume predefined domains nor classification of the considered papers, but reconstructs from the bottom-up the fields with the endogenous semantic information. Nichols (2014) already introduced a close approach, using Latent Dirichlet Allocation topic modeling to characterize interdisciplinarity of awards in particular sciences. Larivière and Gingras (2014) quantifies interdisciplinarity over a long time range by looking at the field of references of publications.

The rest of the paper is organized as follows : we describe in section the nature of the dataset used and the data collection procedure. We then give in section results on interdisciplinarity landscape obtained through network multilayer analysis of the dataset, which are finally discussed in section .

Database Construction

Our approach imposes some requirements on the dataset used, namely: (i) cover a certain neighborhood of the studied journal in the citation network in order to have a consistent view on the scientific landscape; (ii) have at least a textual description for each node. For these to be met, we need to gather and compile data from heterogeneous sources, using therefore a specific application, which general architecture is given in Fig. 1. Source code is available on the `git` repository of the project at .

For the sake of simplicity, we will denote by *reference* any standard scientific production that can be cited by another (journal paper, book, book chapter, conference paper, communication, etc.) and contains basic records (title, abstract, authors, publication year). We will work in the following on networks of references. Note that one significant contribution of this paper is the construction of such an hybrid dataset from heterogeneous sources, and the development of associated tools that can be reused and further developed for similar purposes.

Initial Corpus

The production database of *Cybergeo* (snapshot taken in February 2016), provided by the editorial board, provides after pre-processing the initial database of articles, with basic information (title, abstract, publication year, authors).

The processed version used is available together with the full database constructed, as a `mysql` dump, at . This base provide also bibliographical records of articles that give all references cited by the initial base (*forward citations* for the initial corpus).

Citation Data

Citation data is collected from `Google Scholar`, that is the only source for incoming citations Noruzi (2005) in our case as the journal is poorly referenced in other databases[2]. We are aware of the possible biaises using this single source (see e.g. Bohannon (2014))[3], but these critics are more directed towards search results or possible manipulations than citation counts. The automatic collection requires the use of an open source data crawling software to pipe requests, namely `TorPool` (Raimbault, 2016) that provides a Java API allowing an easy integration into our application of data collection. A crawler can therethrough retrieve html pages and get backward citation data, i.e. all citing articles for a given initial article. We retrieve that way two sub-corpuses: references citing papers in *Cybergeo* and references *citing the ones cited* by *Cybergeo*. At this stage, the full corpus contains around $4 \cdot 10^5$ references.

Text Data

A textual description for all references is necessary for a complete semantic analysis. We use for this an other source of data, that is the online catalog of *Mendeley* reference manager software Mendeley (2015). It provides a free API allowing to get various records under a structured format. Although not complete, the catalog provides a reasonable coverage (over 55%), yielding a final corpus with full abstracts of size $2.1 \cdot 10^5$. The structure and descriptive statistics of the corresponding citation network is recalled in Fig. 2.

Methods and Results

Citation Network Properties

Properties As detailed above, we are able by the reconstruction of the citation network at depth ±1 from the original 1000 references of the journal to retrieve around $45 \cdot 10^6$ references, on which $2.1 \cdot 10^5$ are retrieved with abstract text allowing semantic analysis. A first glance on citation network properties provides useful insights. Mean in-degree (that can be interpreted as a stationary integrated impact factor) on references where it is defined has

[2] or was just added as in the case of *Web of Science*, indexing *Cybergeo* since May 2016 only

[3] or `http:iscpif.frblog201602the-strange-arithmetic-of-google-scholars`

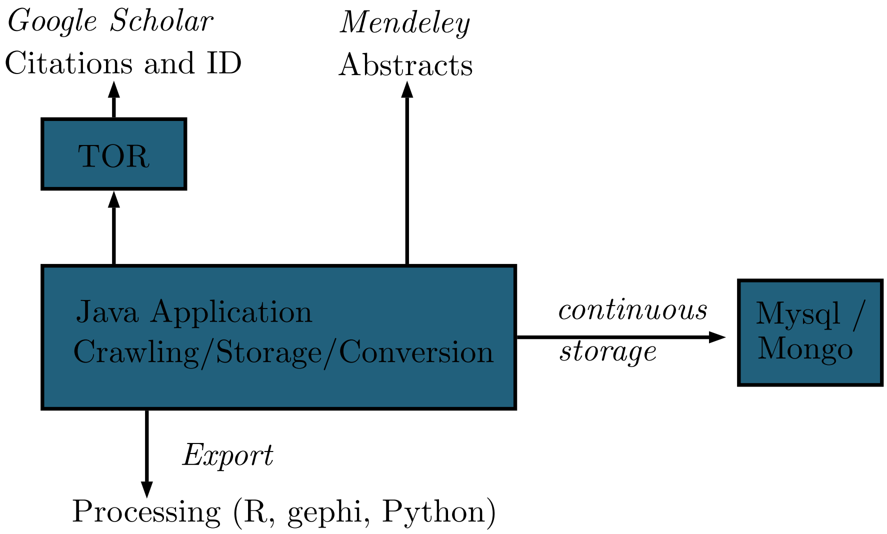

Fig. 1 Heterogeneous Bibliographical Data Collection. Architecture of the application for content (semantic data), metadata and citation data collection. The heterogeneity of tasks requires a multi-lingual approach.

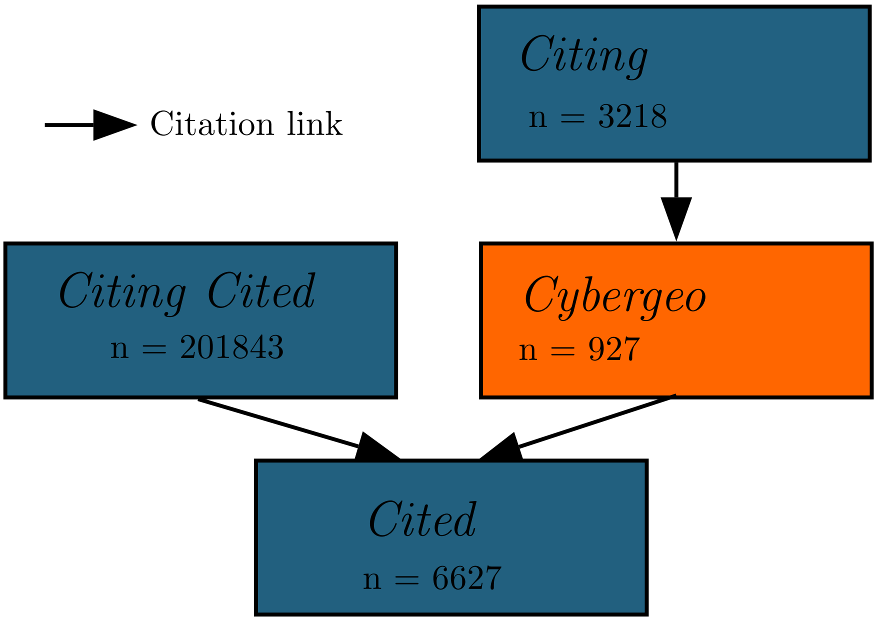

Fig. 2 Structure and content of the citation network. The original corpus of *Cybergeo* is composed by 927 articles, themselves cited by a slightly larger corpus (yielding a stationary impact factor of around 3.18), cite $\simeq$ 6600 references, themselves co-cited by more than $2 \cdot 10^5$ works.

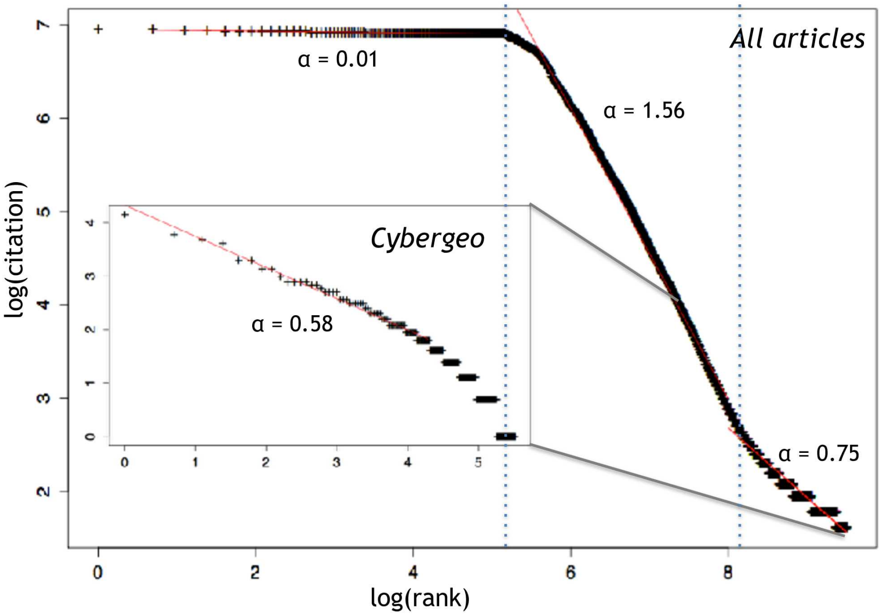

Fig. 3 Rank-size plot of in-degrees in the citation network ; three superposing successive regimes must correspond to different literature types or practices across disciplines.

a value of $\bar{d} =$, whereas for articles in *Cybergeo* we have $\bar{d} = 3.18$. This difference suggests a variety for status of references, which is confirmed by the hierarchical organisation showed in Fig. 3 with the three superposed regimes. Other topological properties of the citation network reveal typical patterns: for example, the existence of high-order cliques implies certain citation practices which compatibility with the cumulative nature of knowledge may be questionable Pumain (2005).

Citation communities The citation network is a first opportunity to construct endogenous disciplines, by extracting citation communities. More precisely, this step aims at finding patterns

Semantic Communities Construction

Relevant Keywords Extraction We recall that our corpus with available text consists of around $2 \cdot 10^5$ abstracts of publications at a topological distance shorter than 2 from the journal *Cybergeo* in the citation network.

Text processing is done with the method used by Bergeaud et al (2017). We use the python library `nltk` Bird (2006) that provides state-of-the-art

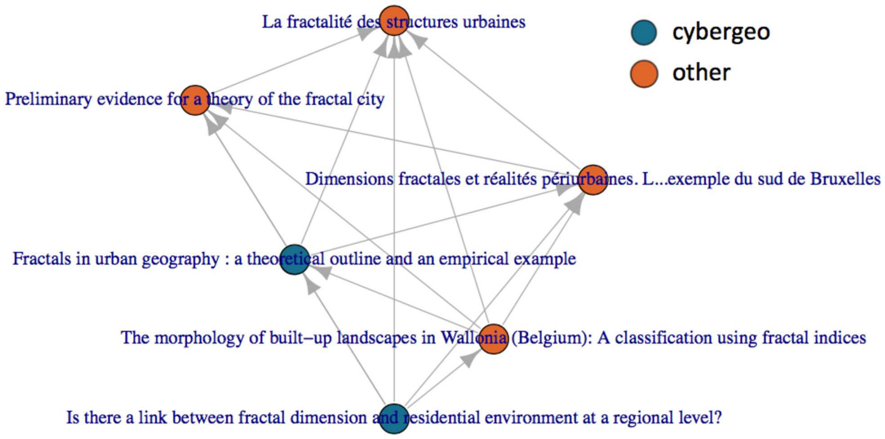
La fractalité des structures urbaines
Preliminary evidence for a theory of the fractal city
Dimensions fractales et réalités périurbaines. L...exemple du sud de Bruxelles
Fractals in urban geography : a theoretical outline and an empirical example
The morphology of built–up landscapes in Wallonia (Belgium):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

Fig. 4 Example of a maximal clique in the citation network, paper of `cybergeo` being in blue. Such topological structure reveal citation practices such as here a systematic citation of previous works in the research niche.

operations in Natural Language Processing. We add a particular treatment for language detection with *stop-words* Baldwin and Lui (2010).

We then use a specific tagger, `TreeTagger` Schmid (1994), for languages other than english. To summarize, the relevant keywords extraction workflow goes through the following steps :

1. Language detection using *stop-words*
2. 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)
3. Selection of potential *n-grams* (with $1 \leq n \leq 4$) following the given patterns: for English $\bigcap\{NN \cup VBG \cup JJ\}$, and for French $\bigcap\{NOM \cup ADJ\}$. Other languages are a negligible proportion of the corpus and are discarded.
4. Estimation of *n-grams* relevance, following co-occurrences statistical distribution

Semantic Network We keep a fixed number K_W of keywords, based on their relevance score.

most relevant keywords yield the co-occurrence matrix that can be directly interpreted as a weighted adjacency matrix.

Sensitivity Analysis The topology of raw networks does not allow the extraction of clear communities, in particular because of the presence of hubs that correspond to frequent terms common to many fields (e.g. `model`, `space`). We assume these highest degree terms do not carry specific information on particular classes and can be thus filtered given a maximal degree threshold k_{max}.

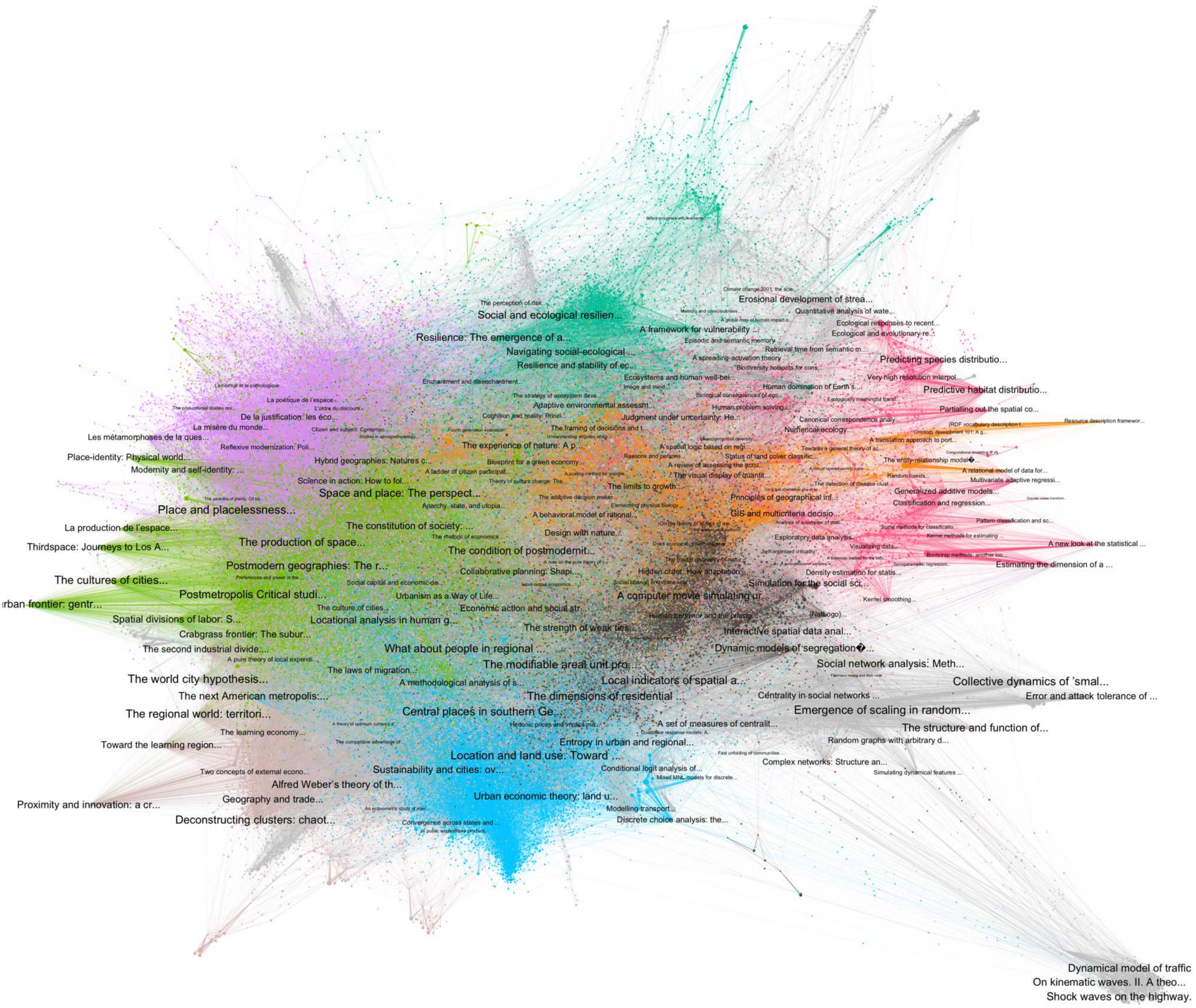
Climate change 2001: the scie...
Erosional development of strea...
The perception of risk...
Social and ecological resilien...
Memory and consciousness...
Quantitative analysis of wate...
A global map of human impact o...
Ecological responses to recent...
A framework for vulnerability ...
Ecological and evolutionary re...
Resilience: The emergence of a...
Episodic and semantic memory ...
Retrieval time from semantic m...
Navigating social-ecological ...
A spreading-activation theory ...
Predicting species distributio...
Resilience and stability of ec...
Biodiversity hotspots for cons...
Ecosystems and human well-bei...
Very high resolution interpol...
Le normal et le pathologique...
Enchantment and disenchantment...
Image and mind...
Human domination of Earth's ...
Predictive habitat distributio...
The strategy of ecosystem deve...
Biological consequences of eco...
La poétique de l'espace...
Adaptive environmental assessm...
Ecologically meaningful transf...
The post-colonial studies rea...
L'ordre du discours...
Human problem solving...
Partialling out the spatial co...
De la justification: les éco...
Cognition and reality: Princi...
Judgment under uncertainty: He...
Canonical correspondence analy...
Resource description framewor...
La misère du monde...
Citizen and subject: Contempo...
Fourth generation evaluation
The framing of decisions and t...
Numerical ecology...
(RDF vocabulary description l
Studies in ethnomethodology...
Understanding attitudes and p...
Ontology development 101: A g...
Les métamorphoses de la ques...
Measurement of diversity...
A translation approach to port...
Reflexive modernization: Poli...
The experience of nature: A p...
A spatial logic based on regi...
Towards a general theory of ac...
Place-identity: Physical world...
Hybrid geographies: Natures c...
Blueprint for a green economy...
Reasons and persons...
Status of land cover classific...
Computational modeling of...
Modernity and self-identity: ...
A ladder of citizen participat...
A scaling method for priorite...
A review of assessing the accu...
A robust layered control syste...
The entity-relationship model�...
Theory of culture change: The...
The visual display of quantit...
Random forests...
A relational model of data for...
Science in action: How to fol...
The limits to growth...
The detection of disease clust...
Multivariate adaptive regressi...
Space and place: The perspect...
The adaptive decision maker...
Principles of geographical inf...
Generalized additive models...
The paradox of plenty: Oil bo...
Anarchy, state, and utopia...
Elements of physical biology...
Classification and regression...
Place and placelessness...
A behavioral model of rational...
GIS and multicriteria decisio...
Pattern classification and sc...
The constitution of society: ...
Analysis of a complex of stati...
Some methods for classificatio...
La production de l'espace...
The rhetoric of economics...
Design with nature...
Exploratory data analysis...
Kernel methods for estimating ...
Thirdspace: Journeys to Los A...
The production of space...
The condition of postmodernit...
Does economic growth improve
Visualizing data...
A new look at the statistical ...
Self-organized criticality...
A Bayesian method for the indu...
Bootstrap methods: another loo...
Postmodern geographies: The r...
A note on the pure theory of c...
The fractal geometry of natur...
Semiparametric regression...
Estimating the dimension of a ...
Collaborative planning: Shapi...
Hidden order: How adaptation...
Density estimation for statis...
The cultures of cities...
Preferences and power in the...
Social capital and economic de...
Input-output economics...
Social change and crime rate t...
Simulation for the social sci...
Postmetropolis Critical studi...
Urbanism as a Way of Life...
A computer movie simulating ur...
Kernel smoothing...
rban frontier: gentr...
The culture of cities...
Economic action and social str...
Human behavior and the princip...
{NetLogo}...
Spatial divisions of labor: S...
Locational analysis in human g...
The strength of weak ties...
Crabgrass frontier: The subur...
Interactive spatial data anal...
The second industrial divide:...
What about people in regional ...
Dynamic models of segregation�...
A pure theory of local expendi...
The modifiable areal unit pro...
Social network analysis: Meth...
The laws of migration...
The world city hypothesis...
A methodological analysis of s...
Local indicators of spatial a...
Fibonacci heaps and their uses
Collective dynamics of 'smal...
The next American metropolis:...
The dimensions of residential ...
Centrality in social networks ...
Error and attack tolerance of ...
The regional world: territori...
Central places in southern Ge...
Emergence of scaling in random...
A theory of optimum currency a...
Hedonic prices and implicit ma...
A set of measures of centralit...
The learning economy...
Qualitative response models: A...
The structure and function of...
Toward the learning region...
The competitive advantage of...
Entropy in urban and regional...
Random graphs with arbitrary d...
Location and land use: Toward ...
Fast unfolding of communities ...
Complex networks: Structure an...
Two concepts of external econo...
Sustainability and cities: ov...
Conditional logit analysis of...
Simulating dynamical features ...
Alfred Weber's theory of th...
Mixed MNL models for discrete...
Geography and trade...
Urban economic theory: land u...
Proximity and innovation: a cr...
An econometric study of inter...
Modelling transport...
Deconstructing clusters: chaot...
Convergence across states and ...
Is public expenditure producti...
Discrete choice analysis: the...
Dynamical model of traffic
On kinematic waves. II. A theo...
Shock waves on the highway.

Fig. 5 Citation Network. We show only the "core" of the citation network, composed by references with a degree larger than one ($|V| = 107164$ and $|E| = 309778$). The community detection algorithm provides 29 communities with a modularity of 0.69. Nodes and edges color gives the main community (for example ecology in magenta, GIS in orange, Socio-ecology in turquoise, Social geography in green, Spatial analysis in blue). Node labels give shortened titles of most cited papers, size is scaled according to their in-degree.

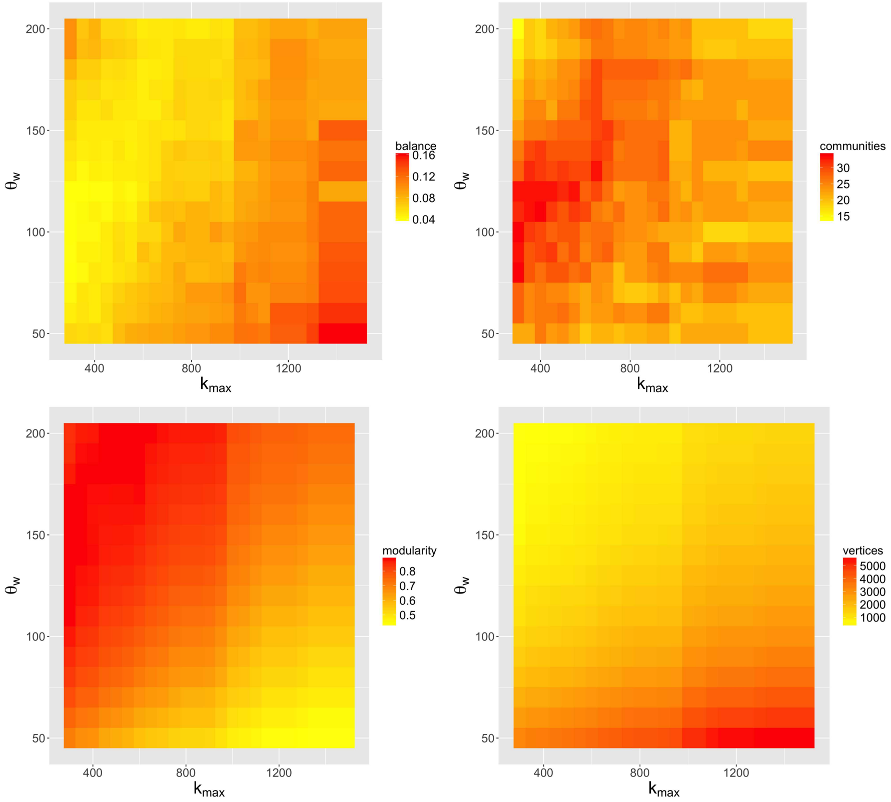

Fig. 6 Sensitivity analysis of network indicators to filtering parameters. We show here 4 indicators (balance between community sizes, modularity of the decomposition, number of communities, number of vertices), as a function of parameters k_{max} and θ_w.

Similarly, edges with low weight (i.e. rare co-occurrences) are considered as noise and are filtered accordingly to a minimal edge weight threshold θ_w. Keywords are preliminary filtered by a document frequency window $[f_{min}, f_{max}]$ which is slightly different from network filtering and complementary.

A sensitivity analysis of resulting network topology to these parameters is presented in Fig. 6. We choose parameter values that maximize modularity under the constraint of a community number and size distribution of same magnitude as technological classes. This multi-objective optimization does not have a unique solution as objectives are somehow contradictory, and a compromise point must be chosen. We take

Semantic Communities We then retrieve communities in the semantic network (using standard Louvain algorithm, with the optimized filtering parameters). At the exception of a small proportion apparently resulting from noise (representing less than 10 keywords, i.e. 0.33% of keywords), communities correspond to well-defined scientific fields (and/or domains, approaches). An expert vali-

Table 1 Disciplines/domains/fields reconstructed from community detection in the semantic network

Name	Size	Keywords
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`
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`

dation allow us to give names to these, in order to stick here to a certain level of supervision. Table 1 summarizes the communities, giving their names, sizes, and corresponding most relevant keywords.

Semantic composition of citation communities

Measuring interdisciplinarity

Distribution of keywords within reconstructed disciplines provides an article-level interdisciplinarity, and we can construct various measures at the journal level. Combination of citation and semantic layers in the hyper-network provide second order interdisciplinarity measures.

More precisely, a reference can be viewed as a probability vector on semantic classes

Given this setting, we simply measure interdisciplinarity using Herfindhal concentration index Porter and Rafols (2009)

Discussion

Comparison of journals The construction of null models for comparison and the collection of currently missing data (journals for other papers) are currently ongoing so these results are not presented here.

Performance of the semantic classification A further validation of the relevance of using complementary information contained in the semantic classification could be done by the analysis of modularities within the citation

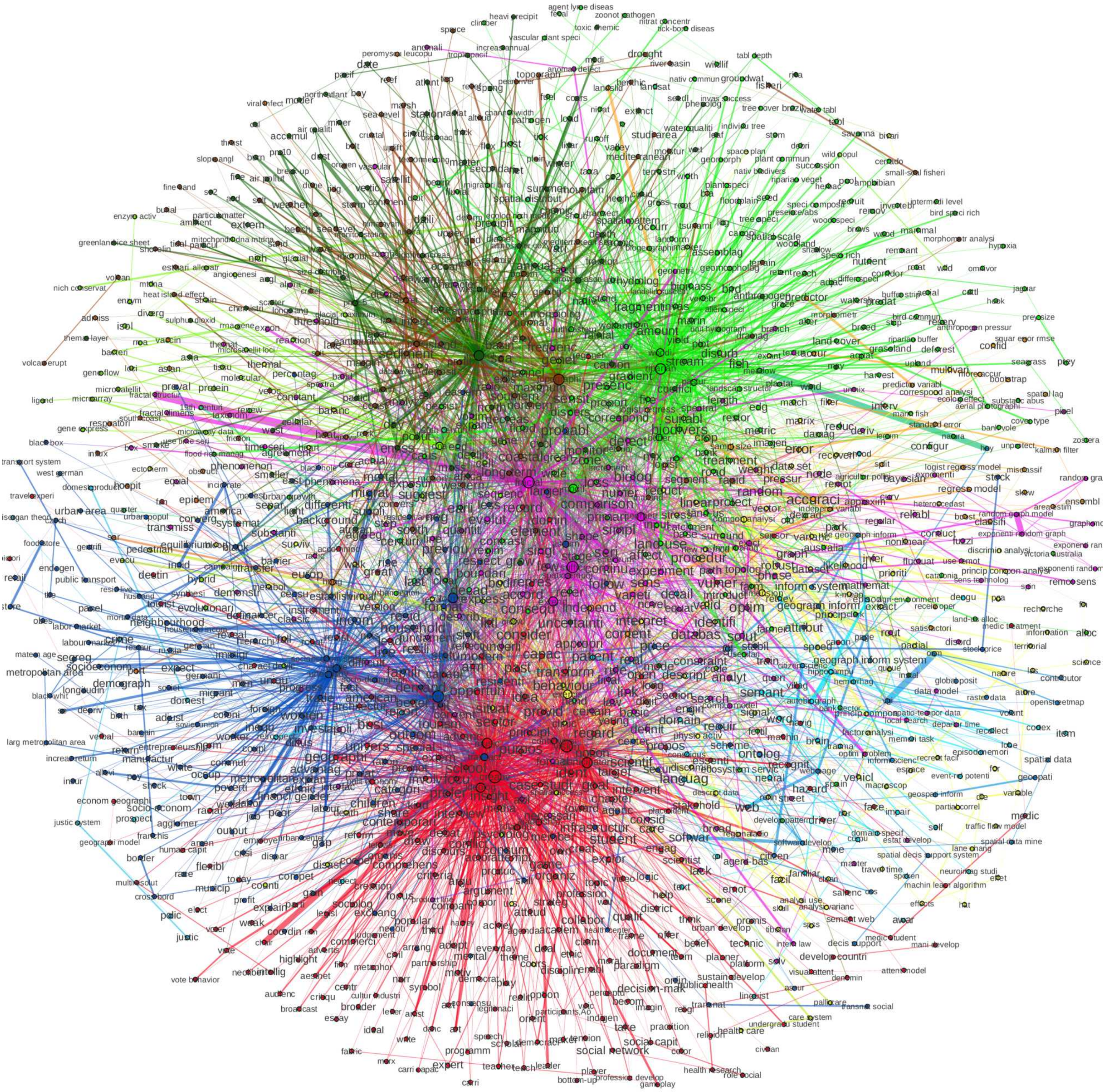

Fig. 7 Semantic network of domains linked to theoretical and quantitative geography. Network is constructed by co-occurrences of most relevant keywords. Filtering parameters are here taken according to the multi-objective optimization done in Fig. 6, i.e. ($k_{max} =, e_{th} = , f_{min}, f_{max} =$). The graph spatialization algorithm (Fruchterman-Reingold), despite its stochastic and path-dependent character, unveils information on the relative positioning of communities.

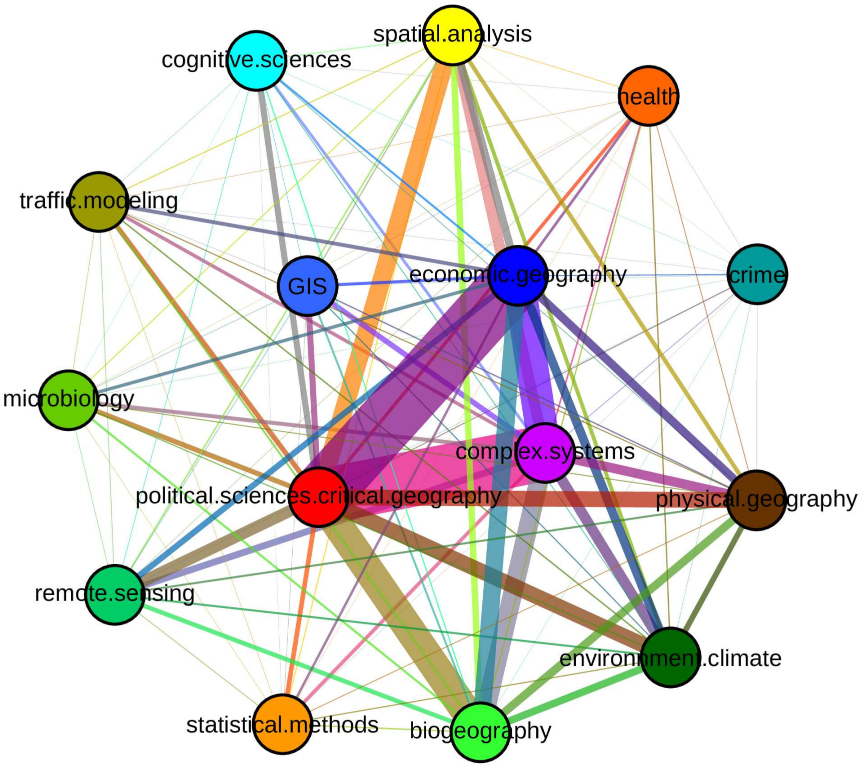

Fig. 8 Synthesis of disciplinary communities and their links.

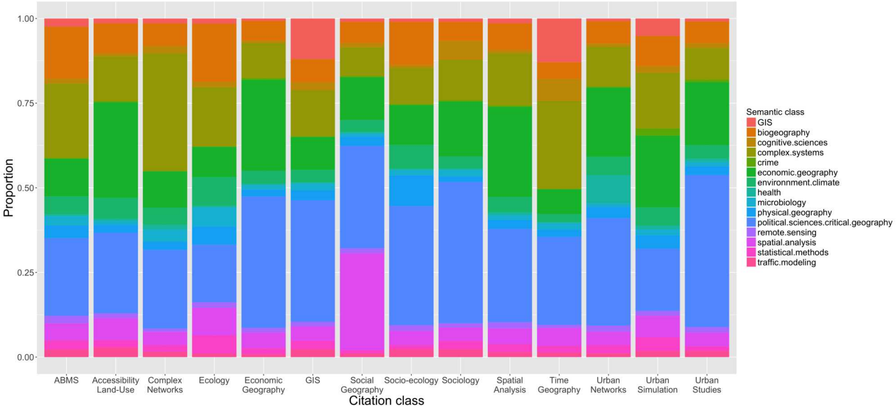

Fig. 9 Synthesis of disciplinary communities and their links.

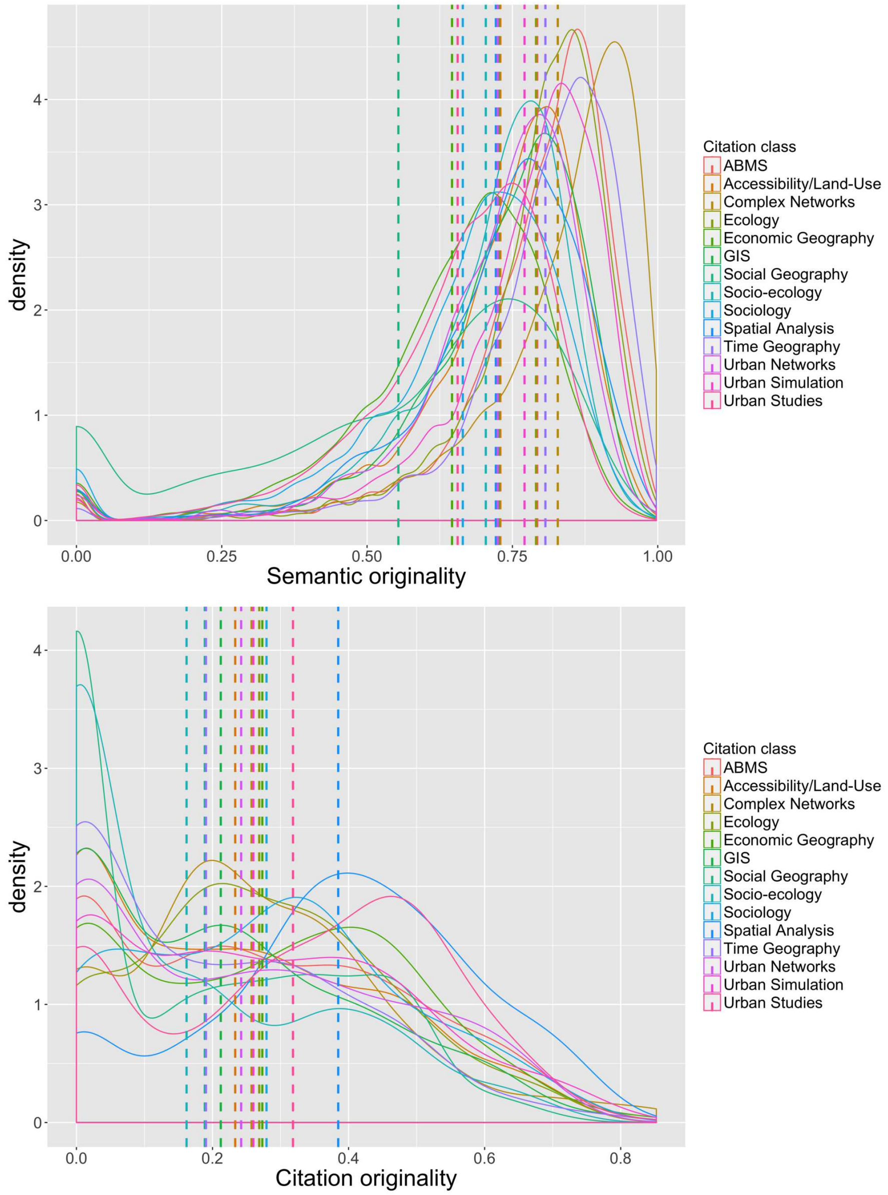

Fig. 10 Distribution of originalities, by citation class.

network, as done in Bergeaud et al (2017). This would however require a baseline classification to compare with, which is not available in the type of data we use. Open repository such as arXiv or Repec provide API to access metadata including abstracts, and could be starting points for such targeted case studies.

Further Developments

Towards an Empowerment of Authors: Open-source Tools for Future Communication Practices

Conclusion

Acknowledgements The author would like to thank the editorial board of Cybergeo, and more particularly Denise Pumain, for having offered the opportunity to work on that subject and provided the production database of the journal.

References

Bais S (2010) In Praise of Science: Curiosity, Understanding, and Progress. MIT Press

Baldwin T, Lui M (2010) Language identification: The long and the short of the matter. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, pp 229–237

Banos A (2013) Pour des pratiques de modélisation et de simulation libérées en géographies et shs. HDR Université Paris 1

Battiston F, Iacovacci J, Nicosia V, Bianconi G, Latora V (2015) Emergence of multiplex communities in collaboration networks. ArXiv e-prints 1506.01280

Bergeaud A, Potiron Y, Raimbault J (2017) Classifying patents based on their semantic content. PloS one 12(4):e0176,310

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

Bohannon J (2014) Scientific publishing. google scholar wins raves–but can it be trusted? Science (New York, NY) 343(6166):14

Bourgine P, Chavalarias D, al (2009) French Roadmap for complex Systems 2008-2009. ArXiv e-prints 0907.2221

Choi J, Hwang YS (2014) Patent keyword network analysis for improving technology development efficiency. Technological Forecasting and Social Change 83:170–182

Cronin B, Sugimoto CR (2014) Beyond bibliometrics: Harnessing multidimensional indicators of scholarly impact. MIT Press

Dupuy G, Benguigui LG (2015) Sciences urbaines: interdisciplinarités passive, naïve, transitive, offensive. Métropoles (16)

Gurciullo S, Smallegan M, Pereda M, Battiston F, Patania A, Poledna S, Hedblom D, Tolga Oztan B, Herzog A, John P, Mikhaylov S (2015) Complex Politics: A Quantitative Semantic and Topological Analysis of UK House of Commons Debates. ArXiv e-prints 1510.03797

Larivière V, Gingras Y (2014) 10 measuring interdisciplinarity. Beyond bibliometrics: Harnessing multidimensional indicators of scholarly impact p 187

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

Morin E (1986) La méthode 3. la connaissance de la connaissance. Essais, Seuil

Nature (2015) Interdisciplinarity, nature special issue. Nature 525(7569):289–418

Newman MEJ (2013) Prediction of highly cited papers. ArXiv e-prints 1310.8220

Nichols LG (2014) A topic model approach to measuring interdisciplinarity at the national science foundation. Scientometrics 100(3):741–754

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

Omodei E, De Domenico M, Arenas A (2017) Evaluating the impact of interdisciplinary research: A multilayer network approach. Network Science 5(2):235–246

Palchykov V, Gemmetto V, Boyarsky A, Garlaschelli D (2016) Ground truth? Concept-based communities versus the external classification of physics manuscripts. ArXiv e-prints 1602.08451

Porter A, Rafols I (2009) Is science becoming more interdisciplinary? measuring and mapping six research fields over time. Scientometrics 81(3):719–745

Pumain D (2005) Cumulativité des connaissances. Revue européenne des sciences sociales European Journal of Social Sciences (XLIII-131):5–12

Raimbault J (2016) Torpool v1.0, doi : 10.5281/zenodo.53739

Sarigöl E, Pfitzner R, Scholtes I, Garas A, Schweitzer F (2014) Predicting Scientific Success Based on Coauthorship Networks. ArXiv e-prints 1402.7268

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

Shibata N, Kajikawa Y, Takeda Y, Matsushima K (2008) Detecting emerging research fronts based on topological measures in citation networks of scientific publications. Technovation 28(11):758–775

Wicherts JM (2016) Peer review quality and transparency of the peer-review process in open access and subscription journals. PLoS ONE 11(1):e0147,913, DOI 10.1371/journal.pone.0147913