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 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 (2009)). There are naturally multiple views on what is exactly interdisciplinarity (many

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

other terms such as trans-disciplinarity, cross-disciplinarity also exist) and it actually depends on 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.

Definitions of interdisciplinarity itself and indicators to measure it have already been tackled by a large body of literature. Huutoniemi et al (2010) recall the difference between *multidisciplinary* (an aggregate of works from different disciplines) and *interdisciplinary* (implying a certain level of integration) approaches. They construct a qualitative framework to classify types of interdisciplinarity, and for example distinguish empirical, theoretical and methodological interdisciplinarities. The multidimensionnal aspect of interdisciplinarity is confirmed even within a specific field such as literature (Austin et al, 1996). A first way to quantify interdisciplinarity of a set of publications is to look at the proportion of disciplines outside a main discipline in which they are published, as Rinia et al (2002) do for the evaluation of projects in physics, complementary with judgement of experts. Porter et al (2007) designate this measure as *specialization*, and compares it with a measure of *integration*, given by the spread of citations done by a paper within the different Subject Categories (classification of the Web of Knowledge), which is also called the *Rao-Stirling* index. Larivière and Gingras (2010) uses it on a Web of Science corpus to show the existence of an optimal intermediate level of interdisciplinarity for the citation impact within a five year window. A similar work is done in (Larivière and Gingras, 2014), focusing on the evolution of measures on a long time range. The influence of missing data on this index is studied by Moreno et al (2016), providing an extended framework taking into account uncertainty. The use of networks has also been proposed : Porter and Rafols (2009) combine the integration index with a mapping technique which consists in visualisation of synthetic networks constructed by co-citations between disciplines. Leydesdorff (2007) shows that the betweenness centrality is a relevant indicator of interdisciplinarity, when considering appropriate citation neighborhood.

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. We choose to study an electronic journal in Geography, named *Cybergeo*[1], that publishes articles within all subfields of Geography and is in that way multidisciplinary. The choice is initially due to data availability, but ensures several constraints making it highly relevant to the context given above. First of all, the "discipline" of Geography is very broad and by essence interdisciplinary Bracken (2016) : the spectrum ranges from Human and Critical geography to physical geography and geomorphology, and interactions between these subfields are numerous. 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 objective, and making technical solutions as the ones we will consequently develop 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), open data and model practices, as recalled by Pumain (2015), and this work contributes to these by fostering the opening of reflexivity.

Our approach combine semantic communities analysis with citation network to extract features such as 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. Palchykov et al (2016) takes a similar approach for papers in physics based on concept extraction from full texts, and show that the endogenous classes differ from the top-down subjects classification. Semantic networks are otherwise well studied in social sciences, such as for

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

example Gurciullo et al (2015) that analyze semantic networks of political debates.

Our contribution is original and significant on at least two aspects :

1. we combine endogenous classifications in a network multilayer fashion, using semantic information ;
2. a large dataset is constructed from scratch to study a journal not referenced in main databases, tackling both data retrieval and large scale data processing issues.

The rest of the paper is organized as follows : we describe in the next section the dataset used and the data collection procedure. We then study properties of the citation network and describe the procedure to construct the semantic classification through text-mining. We finally study complementary measures of interdisciplinarity obtained with the different classifications.

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. We use therefore an application specifically designed, which general architecture is given in Fig. 1. Source code of the application and all scripts used in this paper are available on the open `git` repository of the project[2]. Raw and processed data are also openly available on Dataverse[3]. We recall that an important 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 the address given above. This base provide also bibliographical records of articles that give all references cited by the initial base (*forward citations* for the initial corpus).

[2] at `https://github.com/JusteRaimbault/HyperNetwork`
[3] at `http://dx.doi.org/10.7910/DVN/VU2XKT`

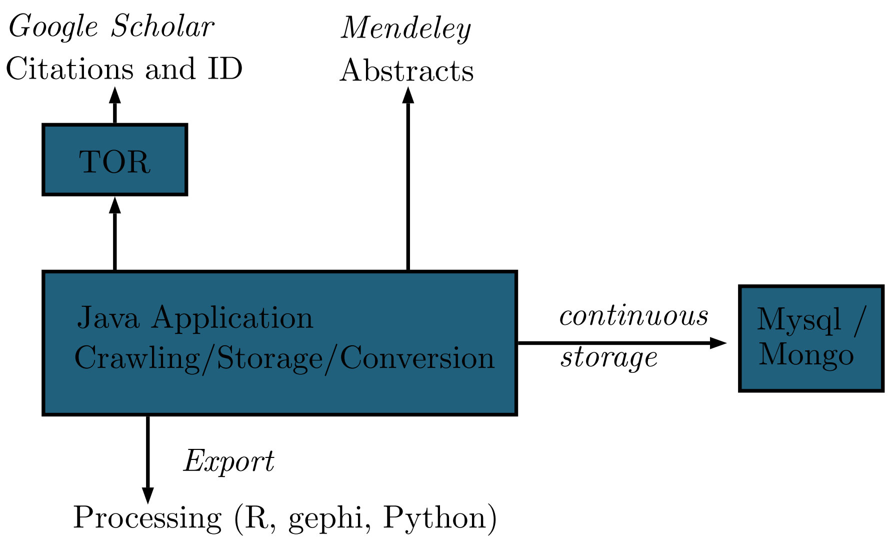

Fig. 1 Heterogeneous Bibliographical Data Collection and processing. Architecture of the application for content (semantic data), metadata and citation data collection. The heterogeneity of tasks requires the use of multiple languages : data collection and management is done in Java, and data stored in databases (Mysql and MongoDB) ; data processing is done in python for Natural Language Processing and in R for statistical and network analyses; graph visualizations are done with Gephi software.

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[4]. We are aware of the possible biaises using this single source (see e.g. Bohannon (2014))[5], but these critics are more directed towards search results or possible targeted manipulations than the global structure of the citation network. The automatic collection requires the use of a 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.

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

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

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

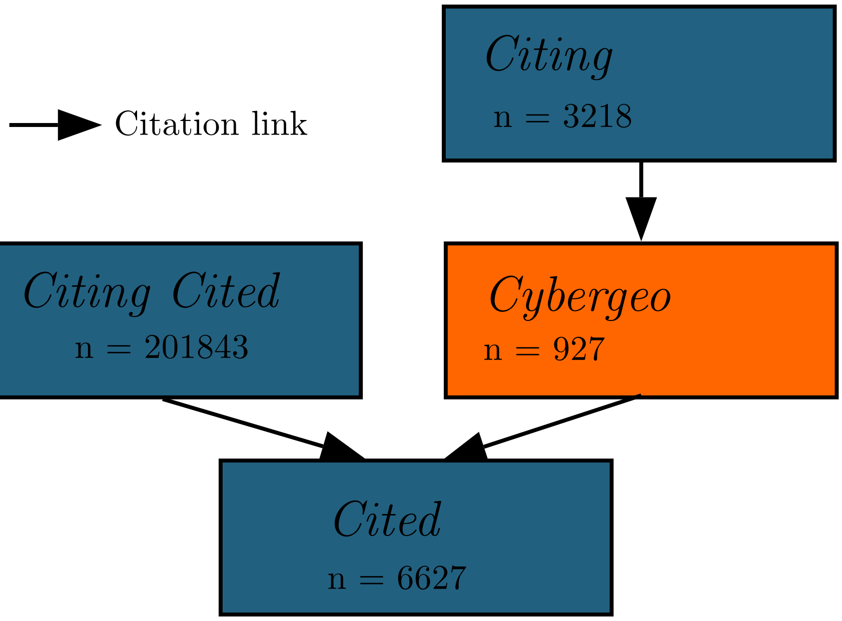

Fig. 2 Structure and content of the citation network. The original corpus of *Cybergeo* is composed by 927 articles, themselves cited by a slightly larger corpus (yielding a stationary impact factor of around 3.18), cite $\simeq 6600$ references, themselves co-cited by more than $2 \cdot 10^5$ works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

(title, abstract, authors, publication year). We work in the following on networks of references, linked by citations.

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 in our case, around 55% of the full citation network. This yields 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.

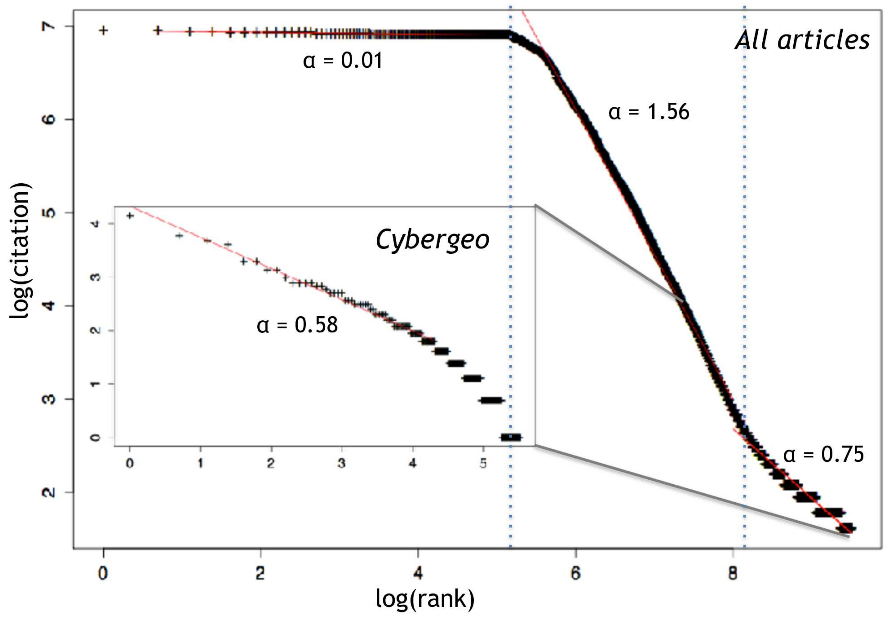

Fig. 3 Rank-size plot of citations received. The plot unveils three superposed citations regimes, corresponding to power laws with different levels of hierarchy. The references in *Cybergeo* (inset plot) are themselves in the tail and less hierarchical.

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 927 references of the journal to retrieve around $45 \cdot 10^6$ references, on which $2.1 \cdot 10^5$ have an 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 for which it can be defined has a value of $\bar{d} = 121.6$, whereas for articles in *Cybergeo* we have $\bar{d} = 3.18$. This difference suggests a variety for status of references, from old classical works (the most cited has 1051 incoming citations) to recent less influential works.

This diversity is confirmed by the hierarchical organisation examined in Fig. 3 that unveils three superposed regimes. More precisely, we look at the rank-size plot, given by the logarithm of the number of citations received as a function of the rank of the paper. We find, as expected (Redner, 1998), localized power-law behaviors. A first set of around 150 references shows a

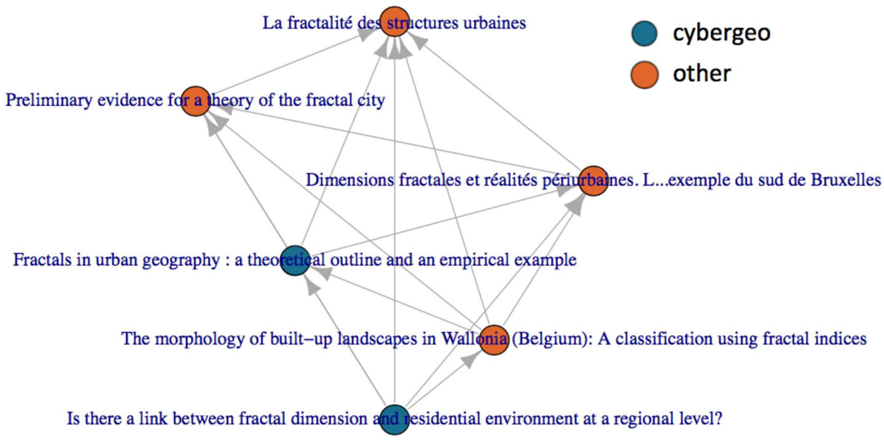

Fig. 4 Example of a maximal clique in the citation network, paper of `cybergeo` being in blue. Such topological structure reveal citation practices such as here a systematic citation of previous works in the research niche.

very low hierarchy (rank-size exponent $\alpha = 0.01$) and corresponds to classical references in different disciplines. A second regime ($\alpha = 1.56$) is much more hierarchized, followed by a last regime less hierarchical ($\alpha = 0.75$) containing more recent papers (average publication year mid-2005, against mid-1998 for the second and 1983 for the first).

Other topological properties reveal typical patterns of citation practices: for example, the existence of high-order cliques (complete sub-networks) implies citation practices which compatibility with the cumulative nature of knowledge may be questionable Pumain (2005), since these need always to source back the production of knowledge in the most recent works. An exemple of such a clique in shown in Fig. 4.

Citation communities

The citation network is a first opportunity to construct endogenous disciplines, by extracting citation communities. More precisely, this step aims at finding recurrent patterns in citations that would define a field by its citation practices. In order to be consistent with the particular data structure we have (missing incoming citations for sub-corpuses at maximal depth), we filter the network by removing all nodes with degree smaller than one. This ensures that kept nodes are either at least cited by an other node (and thus there are no missing edges for these nodes) or cite at least two other nodes, what can make “bridges” between sub-communities. The resulting network has a size of $|V| = 107164$ nodes and $|E| = 309778$ edges. It is visualized in Fig. 5.

We use a standard modularity optimization algorithm to identify communities (Blondel et al, 2008) in this citation network. It provides 29 communities with a modularity of 0.71. In comparison, a bootstrap of 100 randomisations

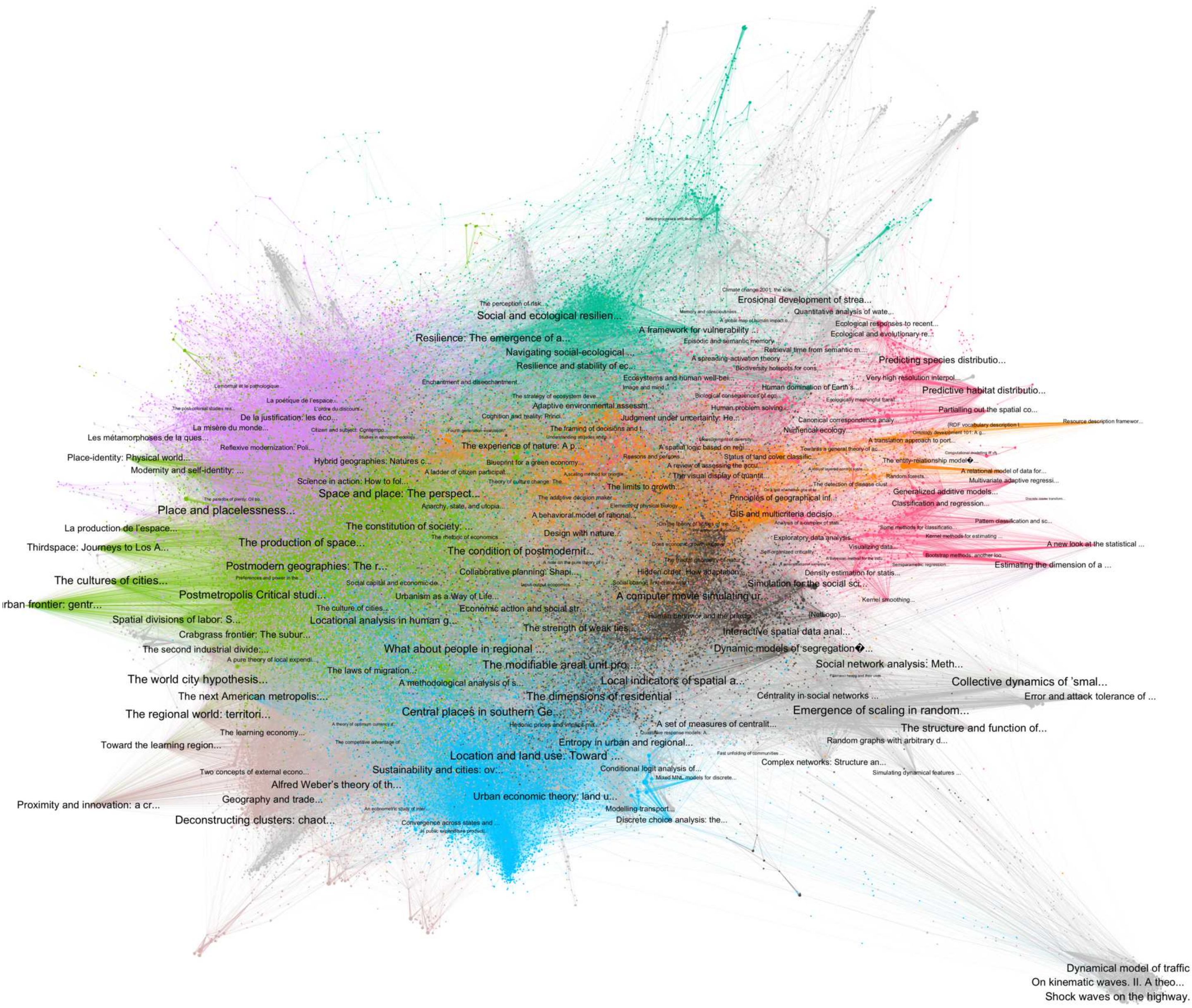

Fig. 5 Citation Network. We show only the "core" of the citation network, composed by references with a degree larger than one ($|V| = 107164$ and $|E| = 309778$). The community detection algorithm provides 29 communities with a modularity of 0.71. Nodes and edges color gives the main community (for 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. The graph is spatialized using a Force-Atlas algorithm.

of links in the network gives an average modularity of $-1.0 \cdot 10^{-4} \pm 4.4 \cdot 10^{-4}$ which means that communities are highly significant.

We name the communities by inspection of the titles of most cited references in each. The 14 communities that have a size larger than 2.5% of the network are : Complex Networks, Ecology, Social Geography, Sociology, GIS, Spatial Analysis, Agent-based Modeling and Simulation (ABMS), Socio-ecology, Urban Networks, Urban Simulation, Urban Studies, Economic Geography, Accessibility/Land-use, Time Geography. These categories do not directly correspond to well-defined disciplines, as some correspond more to methods (ABMS), objects of study (Urban Studies), or paradigms (Complex Networks). Some are "specializations" of others : most papers in Urban Studies can also be classified as Critical and Social geography. This way, we construct endogenous disciplines that correspond to *scientific practices* (what is cited) more than their representation (the "official" disciplines). The relative positioning of communities in Fig. 5, obtained with a Force-Atlas algorithm, tells a lot about their respective relations : for example, social geography makes a bridge between Urban Studies and Economic Geography, whereas the connection between Socio-ecology and Urban simulations is done by GIS (what can be expected as geomatics is an interdisciplinary field). GIS also separates and connects two subfield of Ecology, on one side more thematic studies on ecological habitats, and on the other sides statistical methods. These relations already inform qualitatively patterns of interdisciplinarity, in the sense of integration measures. We will also in the following use these communities to situate the semantic classification.

Semantic Communities Construction

We now turn to the methodological details for the construction of the semantic classification. This step adapts the methodology described by Bergeaud et al (2017), who construct a semantic classification on patent data.

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. The first important step is to extract relevant keywords from abstracts. Text processing is done with the python library `nltk` (Bird, 2006). We add a particular treatment to the method of Bergeaud et al (2017), as our corpus is multilingual: language detection is done with the technique of *stop-words* (Baldwin and Lui, 2010). We also use a specific tagger (the function allowing the attribution of grammatical function to words), `TreeTagger` (Schmid, 1994), for languages other than English.

To summarize, the keyword extraction workflow goes through the following steps :

1. Language detection is done using *stop-words*

2. Pos-tagging (detection of word functions) and stemming (extraction of the *stem*) are 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* (keywords of length n with $1 \leq n \leq 4$) following the given grammatical rules: 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 the relevance *n-grams*, by attributing a score following the deviation of the statistical distribution of co-occurrences to a random distribution.

Semantic Network

We keep at this stage a fixed number K_W of *n-grams*, based on their relevance score, that will be designated as the relevant keywords. We find that for large values of K_W, results are not sensitive to the total number of keywords, and take a reasonably large value for computational performance, $K_W = 50,000$. We construct the co-occurrence matrix of the relevant keywords. This co-occurrence matrix provides the semantic network as its adjacency matrix : nodes are keywords, and they are linked according to their co-occurrences.

Sensitivity Analysis We observe the same phenomenon than in Bergeaud et al (2017), that is the existence of nodes with large degree and not specific to a particular field : for example `model` and `space` are used in most of subfields of Geography. We also adapt the original filtering procedure, as we do not have here an exogenous information to calibrate parameters. We assume the highest degree terms do not carry specific information on particular classes and can be thus filtered given a maximal degree threshold k_{max}. We keep the second filter on a minimal edge weight threshold θ_w. We add the supplementary constraint that keywords are also filtered on a document frequency window $[f_{min}, f_{max}]$ (number of references in which they appear), what is slightly different from network filtering.

A sensitivity analysis of resulting network topology to these four parameters is presented in Fig. 6. Given a filtered network, we detect communities using modularity optimization as before for the citation network. Various properties of the network can be optimized, and we look in particular at its size (number of keywords after filtering), the optimal modularity, the number of communities, and the balance between their sizes (defined as a concentration index $\sum_k s_k^2/(\sum_k s_k)^2$). This multi-objective optimization problem does not have a unique solution as objectives are contradictory in a complex way, and a compromise point must be chosen. We take a compromise point between modularity and network size, with a high balance and a reasonable number of communities, given by $k_{max} = 1200, \theta_w = 100, f_{min} = 50, f_{max} = 10000$.

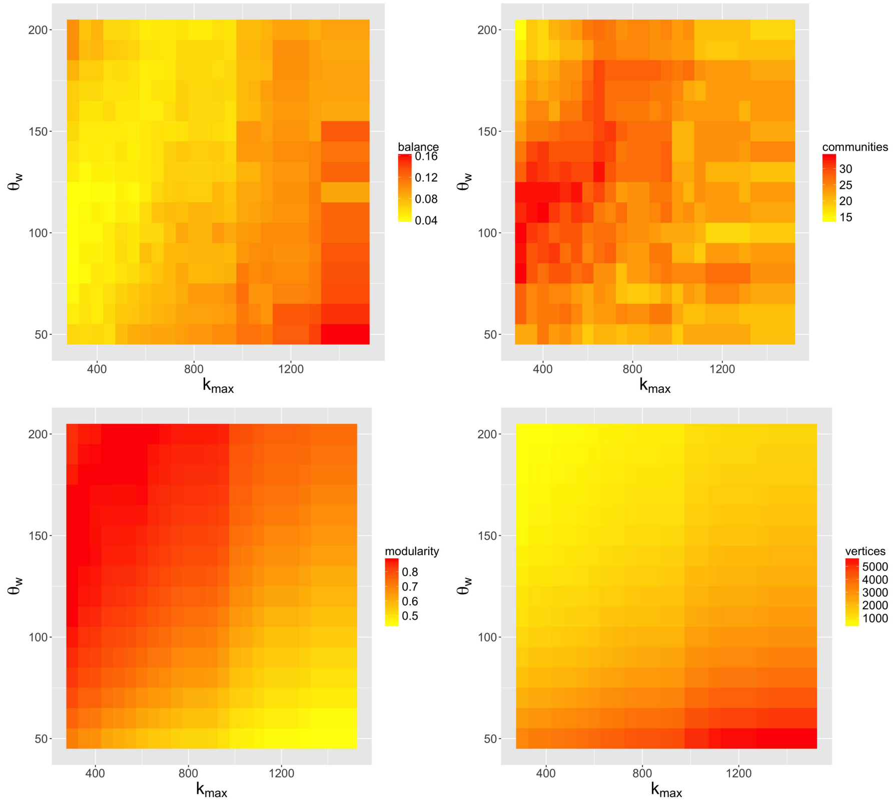

Fig. 6 Sensitivity analysis of network indicators to filtering parameters. We show here 4 indicators (balance between community sizes, modularity of the decomposition, number of communities, number of vertices), as a function of parameters k_{max} and θ_w, at fixed $f_{min} = 50, f_{max} = 10000$. Close values for these two last parameters (in a reasonable range) give similar behavior.

Semantic Communities

We obtain therein communities in the semantic network 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, domains, or approaches. Naming is also done by inspection of the most relevant keywords in each community, in order to stick here to a certain level of supervision.

Table 1 summarizes the communities, giving their names, sizes, and corresponding keywords. The most important community is related to issues in political science and critical geography, what could have been expected as several previously obtained citations communities (Social geography, Urban studies) deal with these issues. We then obtain a large cluster of terms related to biogeography, that must correspond to publications in Ecology and

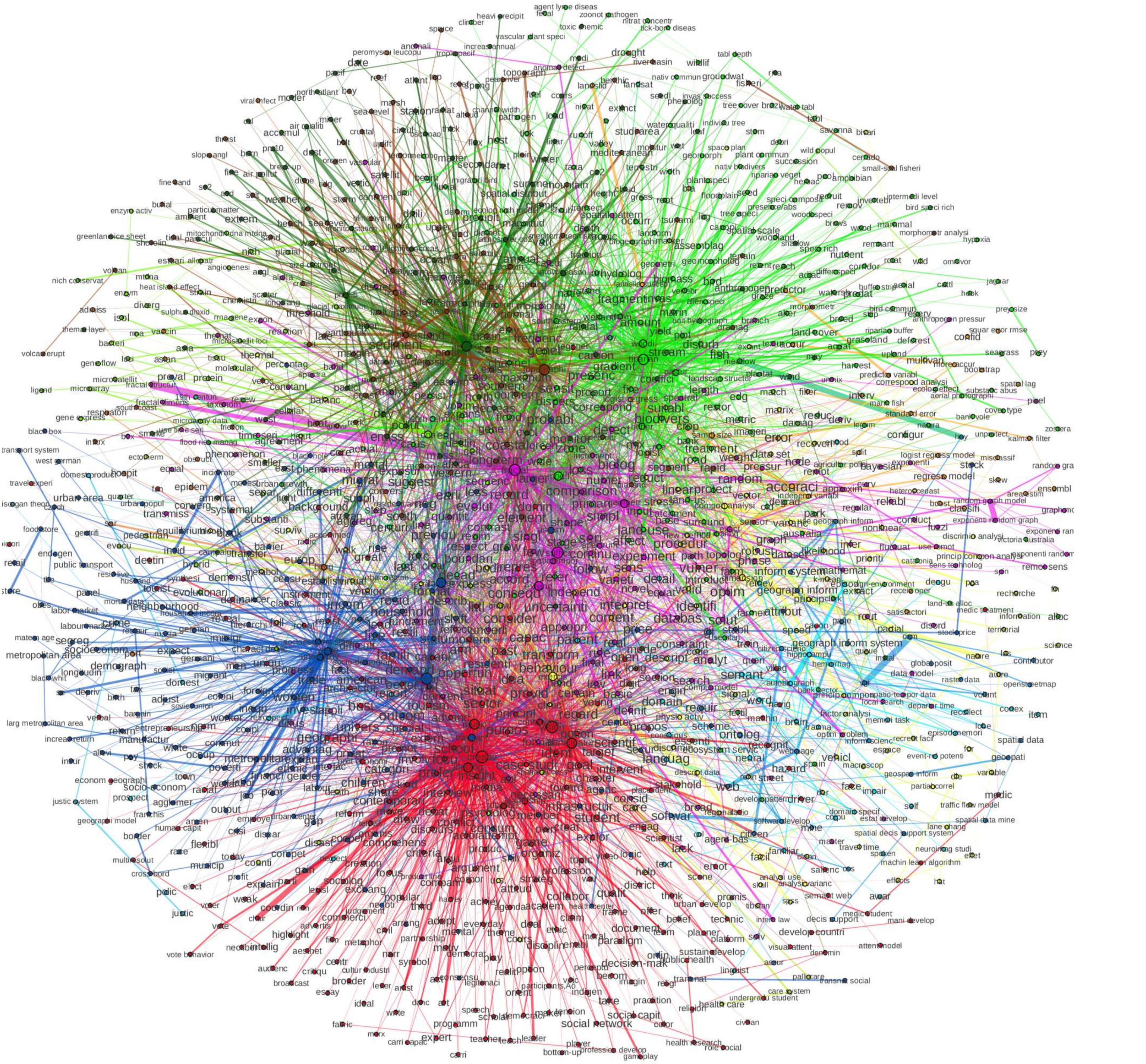

Fig. 7 Visualization of the semantic network. Network is constructed by co-occurrences of most relevant keywords. Filtering parameters are here taken according to the multi-objective optimization done in Fig. 6, i.e. ($k_{max} = 1200, \theta_w = 100, f_{min} = 50, f_{max} = 10000$). The graph spatialization algorithm (Fruchterman-Reingold), despite its stochastic and path-dependent character, unveils information on the relative positioning of communities.

Table 1 Semantic communities 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`

Socio-ecology identified before, together with a community in Environment and Climate.

In a way similar to the citation communities, but more pronounced here, we obtain endogenous “disciplines” that can correspond to real disciplines, to methodologies, to object of studies. This classification thus also unveil *effective scientific practices*, here in terms of semantic content. A class here related to complex systems can be associated to a paradigm and various approaches that were separated in the citation communities : agent-based models and complex networks for example. On the contrary, some studies that were gathered in a large domain before can be precisely differentiated in the semantic network, such as microbiology and health here that are used by studies related to socio-ecology or ecology in the citation network. Some very specific domains appear here as they have very few connections in their actual semantic content.

We show in Fig. 7 a visualisation of the semantic network, in which the positioning of communities, induced by a Fruchterman-Reingold algorithm (that we use here to have a more precise layout in the relative positioning compared to Force Atlas (Jacomy et al, 2014))

In terms of overlaps between communities, in the sense of co-occurrences of corresponding keywords within texts of references, we show a synthesis of links between semantic communities in Fig. 9.

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

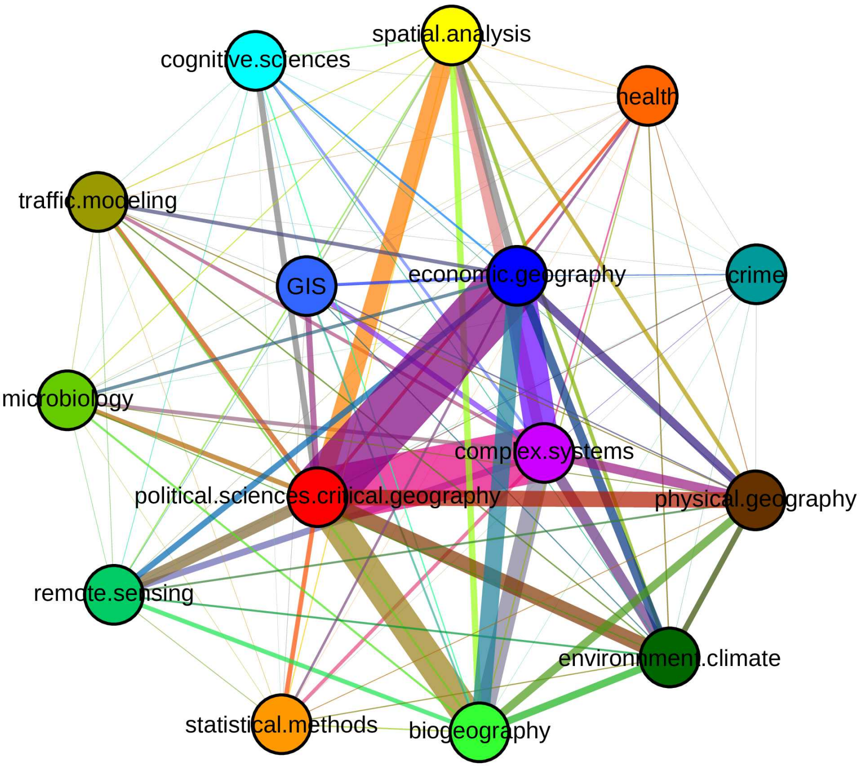

Fig. 8 Synthesis of semantic communities and their links. Weights of links are computed as probabilities of co-occurrences of corresponding keywords within references.

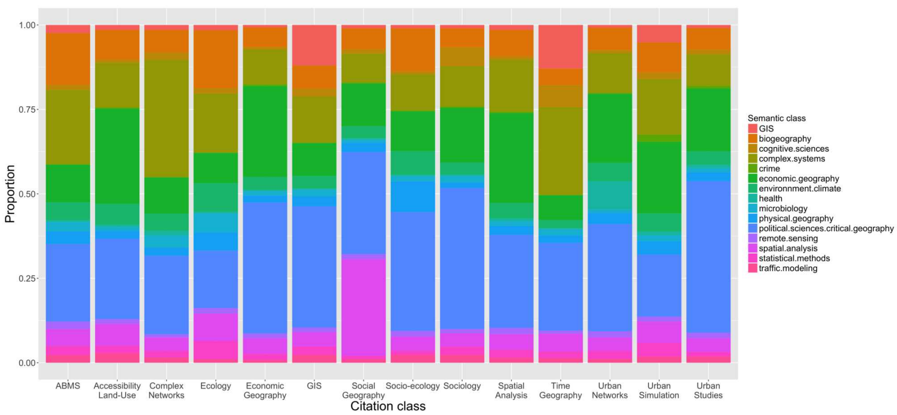

Fig. 9 Composition of citation communities in terms of semantic content.

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 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

quantification of coevolution of knowledge domains.

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

Austin TR, Rauch A, Blau H, Yudice G, van Den Berg S, Robinson LS, Henkel J, Murray T, Schoenfield M, Traub V, et al (1996) Defining interdisciplinarity. Publications of the Modern Language Association of America pp 271–282

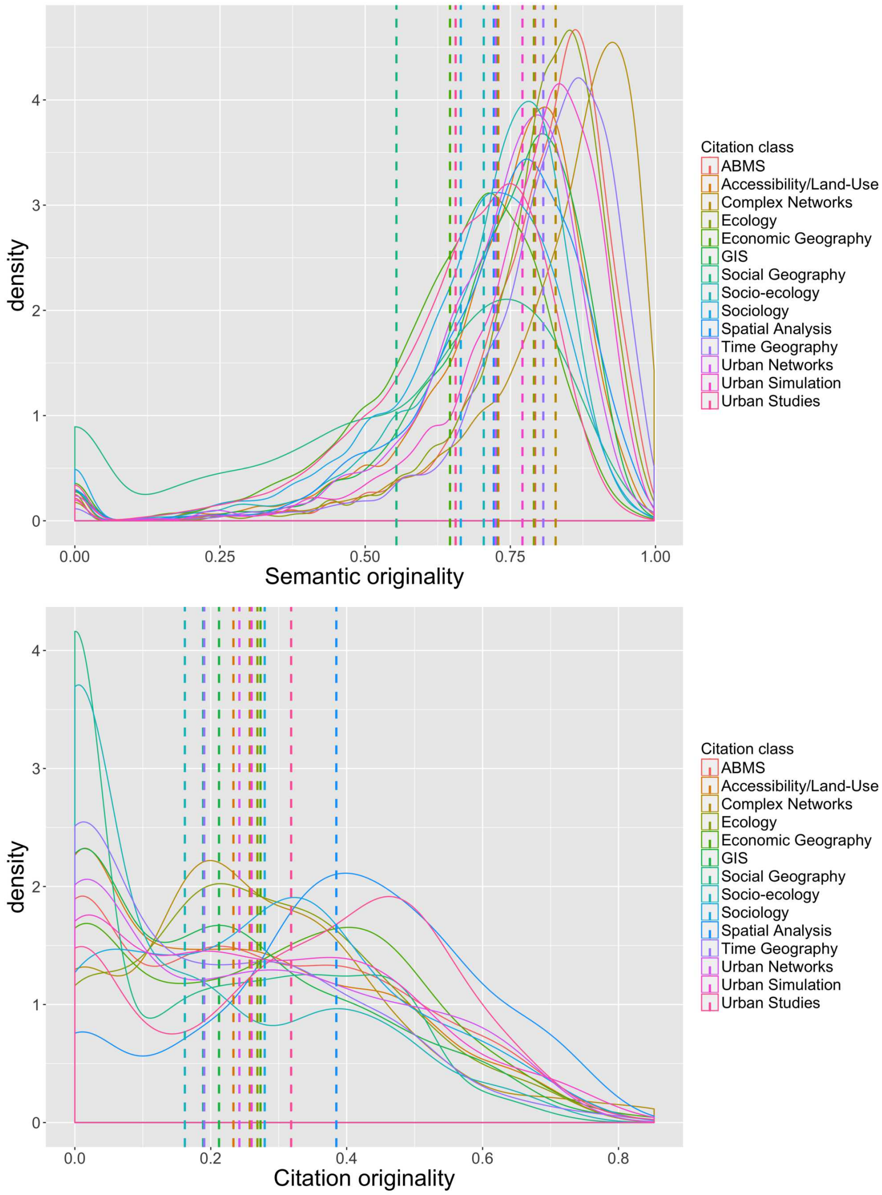

Fig. 10 Distribution of originalities, by citation class.

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

Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E (2008) Fast unfolding of communities in large networks. Journal of statistical mechanics: theory and experiment 2008(10):P10,008

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

Bracken LJ (2016) Interdisciplinarity and Geography. Wiley Online Library

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

Huutoniemi K, Klein JT, Bruun H, Hukkinen J (2010) Analyzing interdisciplinarity: Typology and indicators. Research Policy 39(1):79–88

Jacomy M, Venturini T, Heymann S, Bastian M (2014) Forceatlas2, a continuous graph layout algorithm for handy network visualization designed for the gephi software. PloS one 9(6):e98,679

Larivière V, Gingras Y (2010) On the relationship between interdisciplinarity and scientific impact. Journal of the Association for Information Science and Technology 61(1):126–131

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

Leydesdorff L (2007) Betweenness centrality as an indicator of the interdisciplinarity of scientific journals. Journal of the Association for Information Science and Technology 58(9):1303–1319

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

Moreno MdCC, Auzinger T, Werthner H (2016) On the uncertainty of interdisciplinarity measurements due to incomplete bibliographic data. Scientometrics 107(1):213–232

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. EPJ Data Science 5(1):28
Porter A, Rafols I (2009) Is science becoming more interdisciplinary? measuring and mapping six research fields over time. Scientometrics 81(3):719–745
Porter AL, Cohen AS, Roessner JD, Perreault M (2007) Measuring researcher interdisciplinarity. Scientometrics 72(1):117–147
Pumain D (2005) Cumulativité des connaissances. Revue européenne des sciences sociales European Journal of Social Sciences (XLIII-131):5–12
Pumain D (2015) Adapting the model of scientific publishing. Cybergeo: European Journal of Geography
Raimbault J (2016) Torpool v1.0, doi : 10.5281/zenodo.53739
Redner S (1998) How popular is your paper? an empirical study of the citation distribution. The European Physical Journal B-Condensed Matter and Complex Systems 4(2):131–134
Rinia E, van Leeuwen T, van Raan A (2002) Impact measures of interdisciplinary research in physics. Scientometrics 53(2):241–248
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