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

Exploring Interdisciplinarity Patterns in a Generalist Journal

Juste Raimbault[1,2]

Received: date / Accepted: date

Abstract Keywords Bibliometrics · Semantic Analysis · Theoretical and Quantitative Geography

Introduction

Most of scientific disciplines seem to be in a need of more interdisciplinarity and transversal approaches, as explored in a recent special issue of Nature , for diverse reasons that may include the development of vertically integrated fields conjointly with horizontal transversal questions Bourgine et al (2009). There are naturally ongoing debates on what is exactly interdisciplinarity (many other terms such as transdisciplinarity, crossdisciplinarity also exist) and it actually depends of involved domains : recent hybrid disciplines (see e.g.)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 has no precise definition and integration is by essence horizontal is an other illustration of how transversal knowledge can be produced (leading to misunderstandings when recently introduced to non-aware physicists Dupuy and Benguigui (2015)). The question is naturally transferred into scientific communication : what are corresponding alternatives for an efficient dissemination of knowledge ? Elements of answer to such a high-level issue imply, in an evidence-based perspective, quantitative measures of interdisciplinarity.

The possible methods for quantitative insights into epistemology are numerous. 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 recently proposed in Omodei et al (2016), using bipartites

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

networks of papers and scholars, in order to produce measures of interdisciplinarity. 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 technology : 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. Shibata et al (2008) uses topological analysis of the citation network to detect emerging research fronts.

We describe here a study implementing these ideas for the particular case of a scientific journal for which bibliographical data is difficult to obtain, namely *Cybergeo*, an electronic journal in theoretical and quantitative geography. It makes a particularly interesting case study as deliberately 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.

The rest of the paper is organized as follows : we describe in section the context of the dataset, in particular the scientific purpose of the case study journal, and the data collection procedure. We then give in section 1 results on interdisciplinarity landscape obtained through network multilayer analysis of the dataset.

Database Construction

The general architecture for data collection is presented in Fig. 1.

Initial Corpus

The production database of *Cybergeo* (snapshot dump taken at date), provided by the editorial board, provides after pre-processing the initial database of articles, with basic information (title, abstract, publication year, xxx). 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.

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 not referenced in other databases[1]. We are aware of the possible biaises using this single source

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

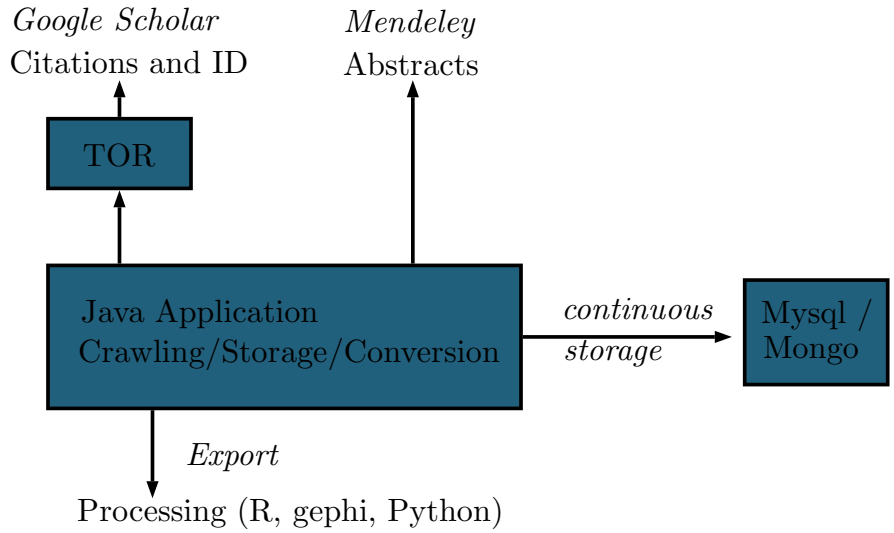

Fig. 1 Heterogeneous Bibliographical Data Collection. Architecture of the application for content (semantic data), metadata and citation data collection. The heterogeneity of tasks requires a multi-lingual approach. Source code and more precise informations on architecture are available on the `git` repository of the project at .

(see e.g. Bohannon (2014))[2], but these critics are more directed towards search results than citation counts. The automatic collection requires the use of an open source data crawling software to pipe requests, namely `TorPool` that provides a Java API allowing an easy integration into our application. Using it, a simple crawler is enough to collect 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* Cybergeo and references *citing the ones cited* by cybergeo.

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

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

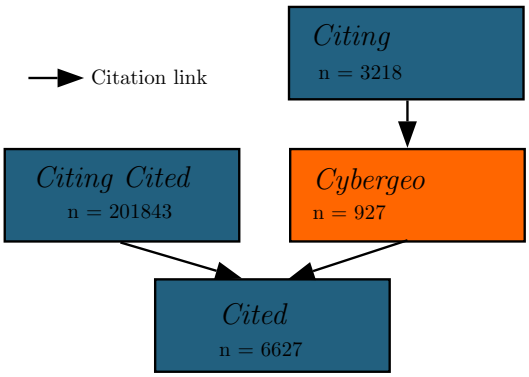

Fig. 2 Structure and content of the citation network. The original corpus of *Cybergeo* consists in 927 articles, themselves cited by a slightly larger corpus (yielding a stationary impact factor of around 3.18), cite $\simeq$ 6600 references, themselves co-cited by more than $2 \cdot 10^6$ works.

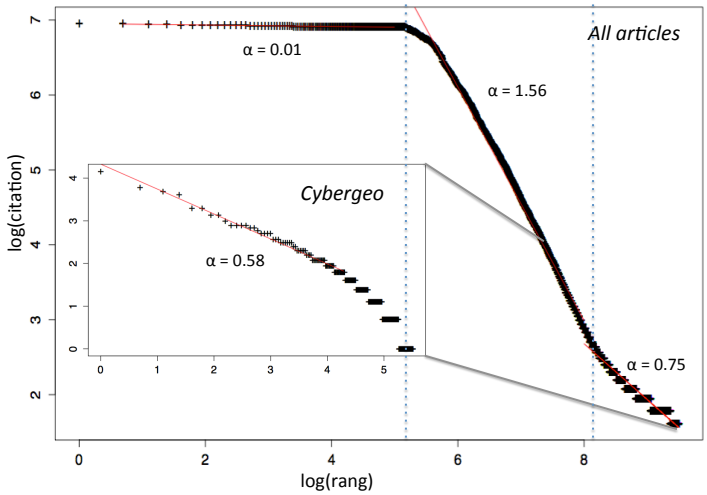

Fig. 3 Properties of the citation network. Rank-size plot of in-degrees ; three superposing successive regimes must correspond to different literature types or practices across disciplines.

1 Methods and Results

1.1 Citation Network Properties

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 $2.1 \cdot 10^6$ are retrieved with abstract text allowing semantic analysis.

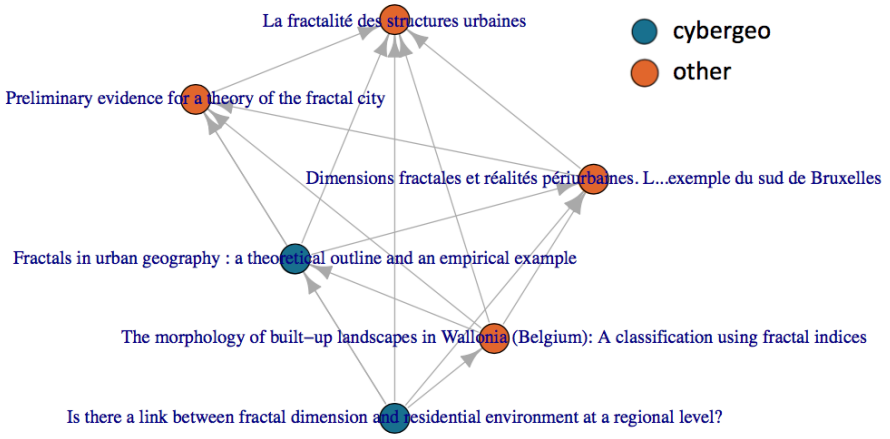

Fig. 4 Example of a maximal clique in the citation network, paper of `cybergeo` being in blue. Such topological structure reveal citation practices such as here a systematic citation of previous works in the research niche.

1.2 Semantic Communities Construction

Relevant Keywords Extraction Corpus 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 using a method adapted from **?**. We use the python library `nltk` Bird (2006) that provides state-of-the-art operations in Natural Language Processing. A particular treatment is required for language detection with *stop-words* and a specific tagger `TreeTagger` is used for other languages than english Schmid (1994). More precisely, we go 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 $1 \leq n \leq 4$) : English $\bigcap\{NN \cup VBG \cup JJ\}$; French $\bigcap\{NOM \cup ADJ\}$
4. Database insertion for instantaneous utilisation (10j $\rightarrow$ 2min)
5. Estimation of *n-grams* relevance, following co-occurrences statistical distribution

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

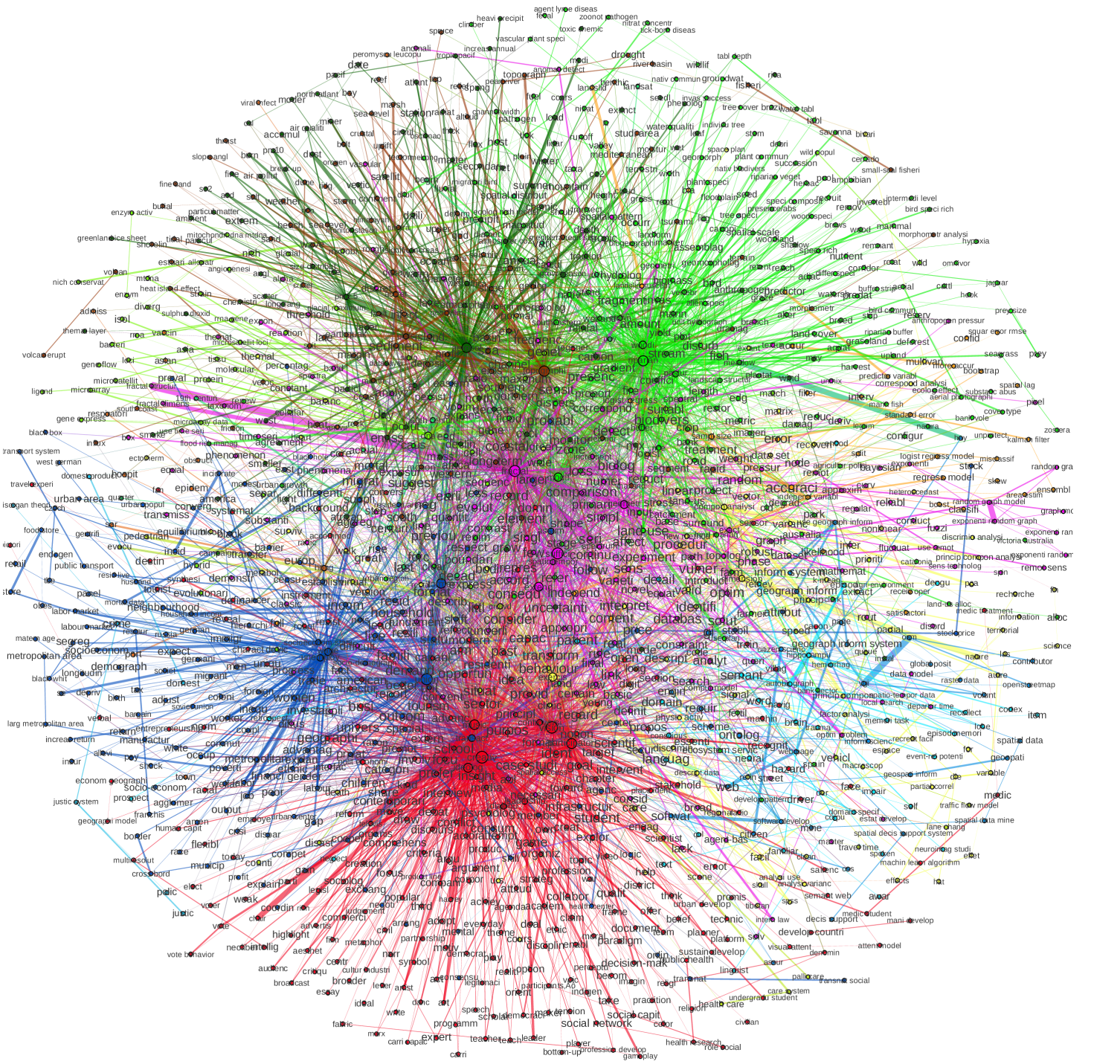

Fig. 5 Semantic network of domains linked to theoretical and quantitative geography. Network is constructed by co-occurrences of most relevant keywords. Filtering parameters are here taken according to the multi-objective optimization done in Fig. 6, i.e. ($k_{max} =$, $e_{th} =$, f_{min}, $f_{max} =$). The graph spatialization algorithm (Fruchterman-Reingold), despite its stochastic and path-dependent character, unveils information A zoomable vectorial file (`.svg`) of the network is available as Supplementary Material.

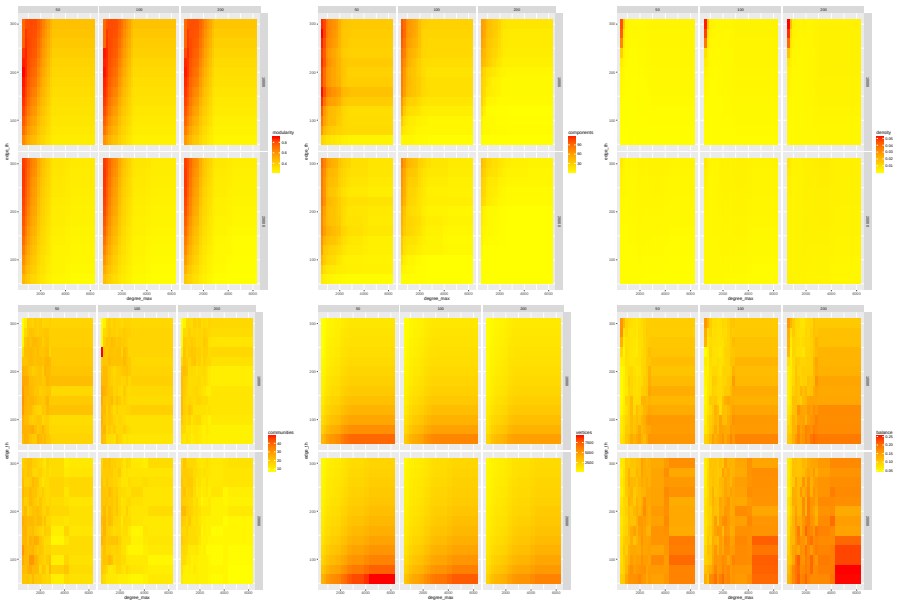

Fig. 6 Sensitivity analysis of network indicators to filtering parameters

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}. Similarly, edge with small weight must not carry significant information and are filtered according to a minimal edge weight threshold w_{min}. 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

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 x/y, i.e. z% of keywords) , communities correspond to well-defined scientific fields (and/or domains, approaches). An expert eye-ball validation provides names to these, a more complicated naming procedure would eventually be possible (as in where a chi-square test on distribution of documents in classes), but we prefer to stick here to a certain level of supervision. Table 1

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

1.3 Measures of 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.

2 Discussion

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.

3 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

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

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

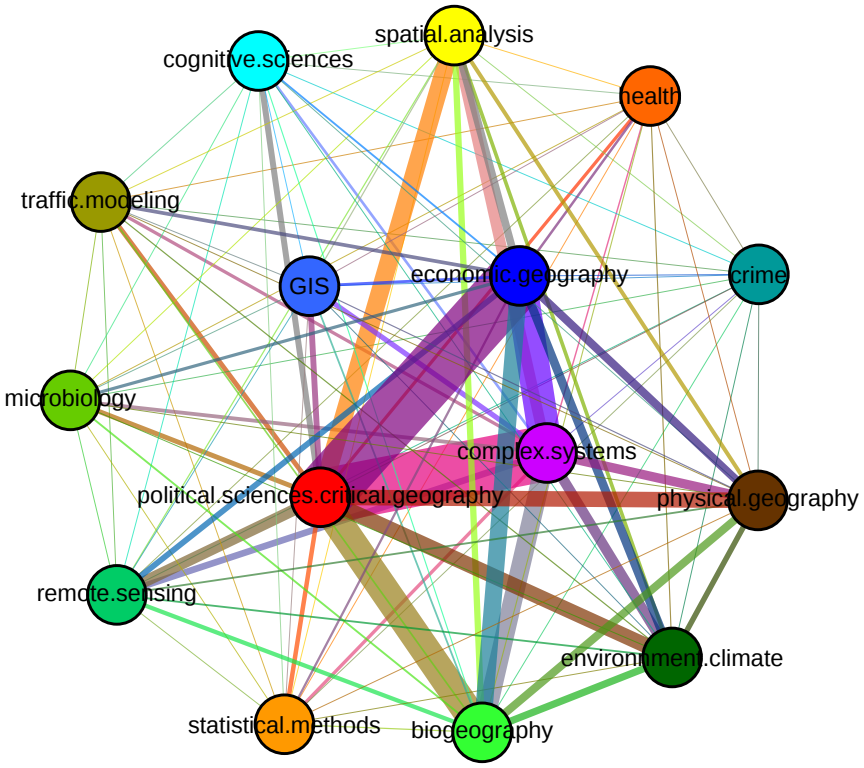

Fig. 7 Synthesis of disciplinary communities and their links.

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

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

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

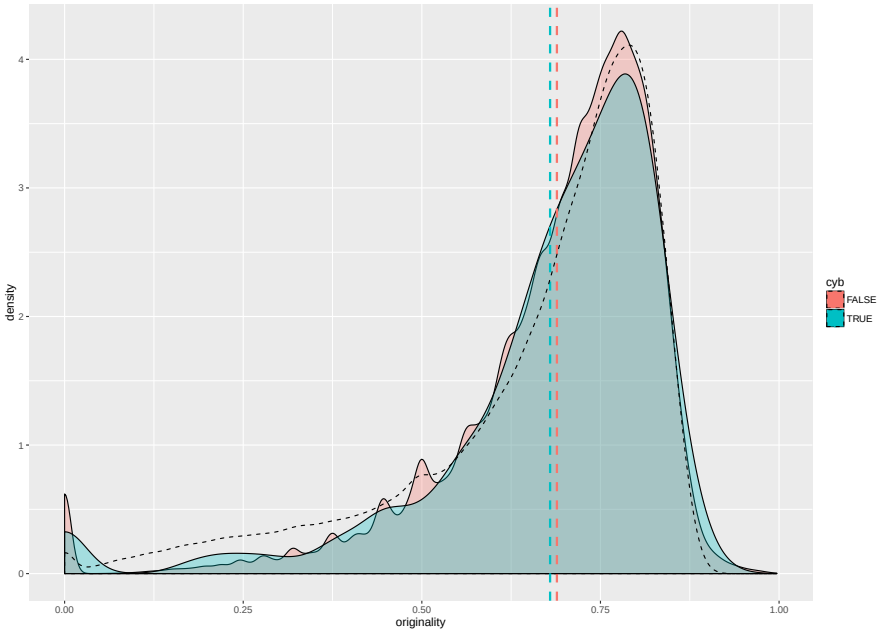

Fig. 8 Distribution of first order interdisciplinarity

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

Omodei E, De Domenico M, Arenas A (2016) Evaluating the impact of interdisciplinary research: a multilayer network approach. ArXiv e-prints 1601.06075

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

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