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

Quantifying Interdisciplinarity of a Generalist Journal

Juste Raimbault

Received: date / Accepted: date

Abstract Insert your abstract here. Include keywords, PACS and mathematical subject classification numbers as needed.

Keywords Bibliometrics · Semantic Analysis

1 Introduction

Existing works in quantitative epistemology using various types of networks have shown interesting potentialities. For the citation network, 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 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, that is `cybergeo`, an electronic journal in theoretical and quantitative geography, that is concerned with open science issues such as peer-review ethics transparency Wicherts (2016). Our approach combine semantic communities analysis (as done in Palchykov et al (2016) for papers in physics but with keyword

J. Raimbault
UMR CNRS 8504 Géographie-cités
E-mail: juste.raimbault@polytechnique.edu

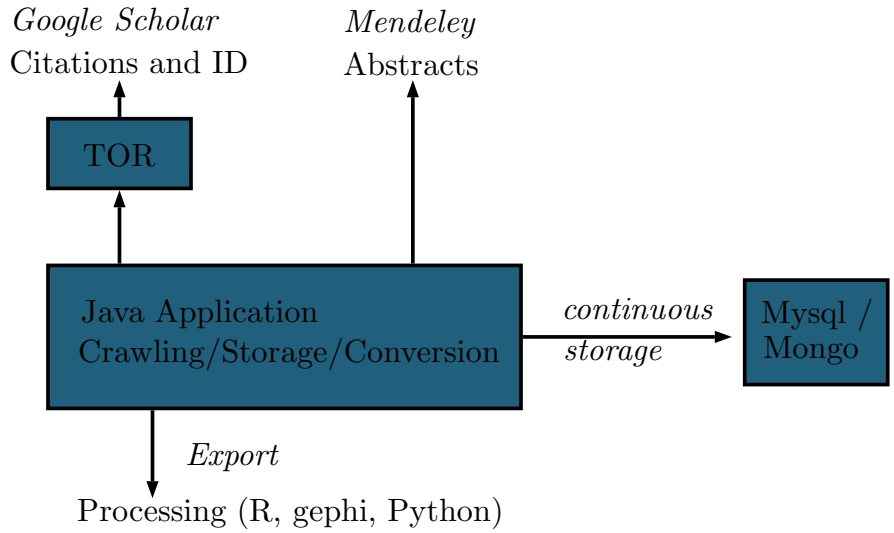

Fig. 1 Heterogeneous Bibliographical Data Collection. Architecture of the application for content (semantic data), metadata and citation data collection.

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 2 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 3 results on interdisciplinarity landscape obtained through network multilayer analysis of the dataset.

2 Data Collection

2.1 Implementation

The general architecture for data collection is presented in Fig. **??**. 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. We are aware of the possible biaises using this single source Bohannon (2014)[1], but these critics are more directed towards search results than citation counts.

Text processing is done the same way as in previous section, expect that a particular treatment is done to language detection using *stop-words* and a specific tagger `TreeTagger` is used for other languages than english Schmid (1994).

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

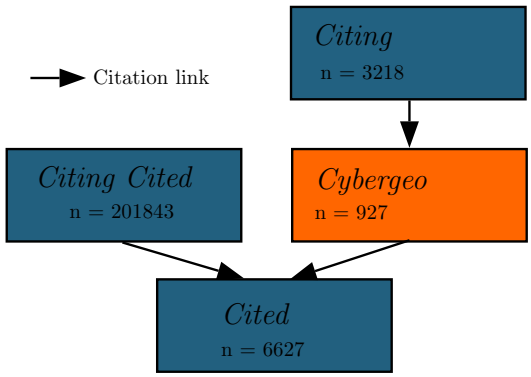

Fig. 2 Structure of the citation network.

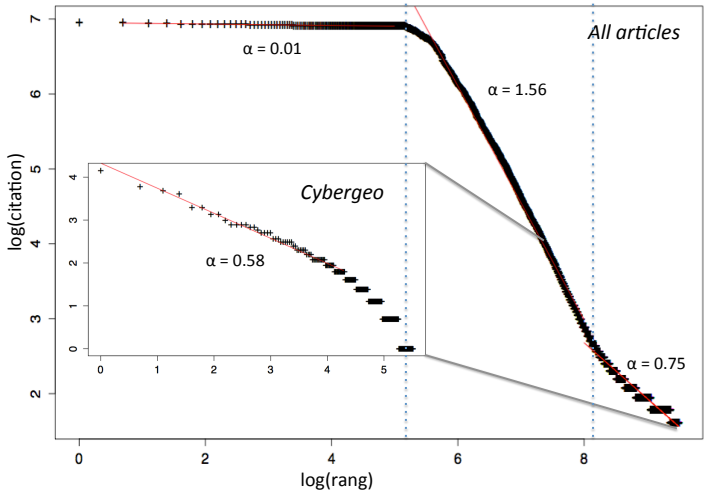

Fig. 3 Properties of the citation network. Rank-size plot of in-degrees ; three superposing successive regimes must correspond to different literature types or practices across disciplines.

3 Results

3.1 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.

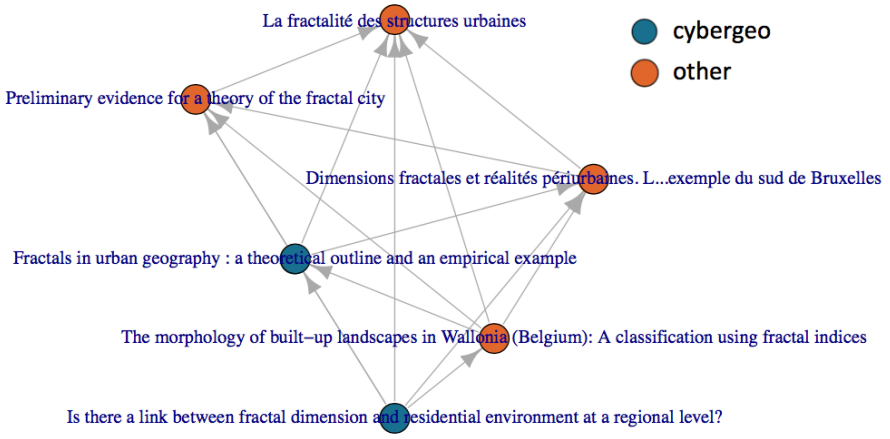

Fig. 4 Example of a maximal clique in the citation network, paper of `cybergeo` being in blue.

3.2 Semantic Communities Construction

Relevant Keywords Extraction *Text-mining in python with* ***`nltk`*** *Bird (2006)*, method adpated from **?**

- Language detection using *stop-words*
- 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)
- Selection of potential *n-grams* (with $1 \leq n \leq 4$) : English $\bigcap\{NN \cup VBG \cup JJ\}$; French $\bigcap\{NOM \cup ADJ\}$
- Database insertion for instantaneous utilisation (10j $\rightarrow$ 2min)
- Estimation of *n-grams* relevance, following co-occurrences statistical distribution

Semantic Network

Communities Construction We retrieve by community detection in the semantic network typical geographical disciplines, such as :

- 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`

3.3 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. 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.

4 Discussion

5 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

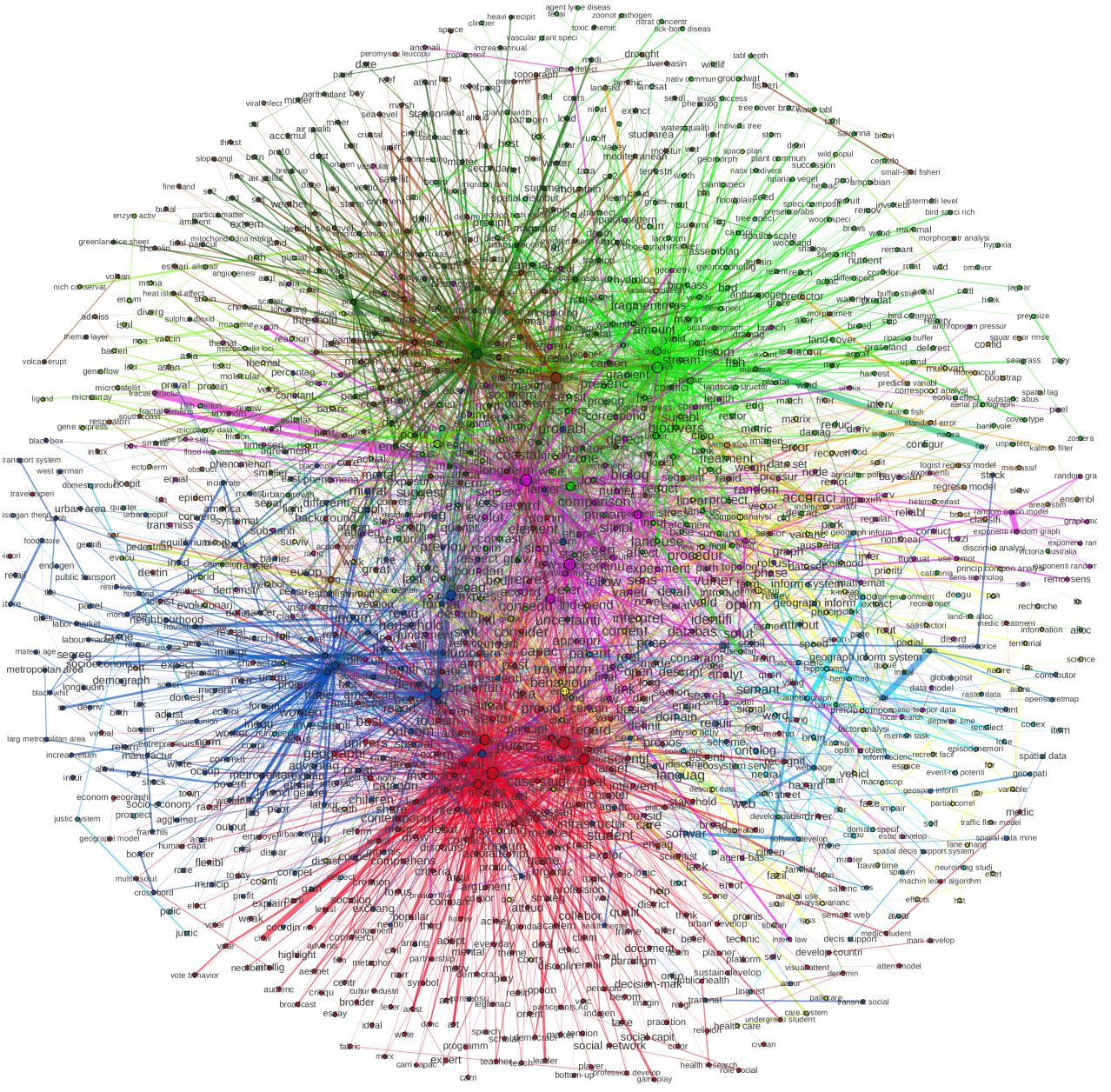

Fig. 5 Semantic network of concepts in quantitative geography. Corpus consists of around $2 \cdot 10^5$ abstracts of publications at a topological distance shorter than 2 from the journal `cybergeo` in the citation network. Relevance of keywords were estimated with a bootstrap method, and semantic network is constructed by co-occurrences of keywords (cut at larger degrees, 10% here to delete hubs such as `model` or `space` and efficiently reveal communities).

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

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

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

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