Report from 2015 Brainhack Montreal

Automatic Extraction of Academic Collaborations in Neuroimaging

Project URL: http://github.com/sderygithub/clubs-of-science

Sebastien Dery

1 Introduction

Our ability to quantitatively study large-scale social and behavioral phenomena such as peer influence and confirmation bias within scientific circles rest on quality and relevant data [1]. Yet the compilation of specific coauthorship databases are often restricted to certain well-defined fields of study or publication resources, limiting the extent and depth by which investigations can be performed. Related work ([2], [3]) have focused on aggregating manually curated data from the community. Ultimately, we aim to understand how the social construct and its underlying dynamics influence the trajectories of scientific endeavors [4]. This work is motivated by an interest in observing social patterns, monitoring their evolution, and possibly understanding the emergence and spreading of ideas and their biases in the neuroimaging community; central themes to deciphering facts from opinions. However, before being able to fully investigate and address these fundamental and inherently complex questions, we need to address the extraction and validation of data. The goal of this project was to leverage publicly available information on Google Scholar (GS) to automatically extract coauthorship networks.

2 Approach

The tool can be accessed through a public website at (http://cos.dery.xyz). The site is constructed using a set of openly accessible libraries allowing the display of coauthorship networks as interactive graphs [5]. Visitors can peruse a set of pre-computed networks extracted using custom Python scripts designed to crawl GS based on a set of predefined constraints (e.g. search topic, publication journal). The proposed interface offers seamless manipulation to keep interaction straightforward and easy to use. The simplicity of the design aims to reach a maximum number of users, assuming a minimal level of technical knowledge.

GraphConstruction: Scholarly citations are commonly found in standardized format, suggesting the structure can be reliably used within an automatic procedure. Moreover, while the result of typical search engines are not structured towards data mining (i.e. mixture of natural language embedded in semi-structured tags and page links), particular combinations of HTML tags and CSS identifiers can be leveraged to extract specific information. This simple scheme allows the reconstruction of large-scale networks of collaborations. Interestingly, Google Scholar also hosts individual pages for authors' rich with pre-computed metrics of scientific productivity and impact (e.g. cumulative number of citations, h-index, i10-index). This data can be further exploited to structure and highlight part of the network.

CommunityDetection: Scientific communities were detected using a greedy agglomerative modularity optimization process [6].

Validation: To assess the recovered network's reliability we performed a spot check on its content. First we examined the accuracy of 100 randomly selected researchers from the network and sought after their departmental affiliation and publication journals to confirm their belonging to the broad field of neuroimaging. The dependence on profiles availability injects a strong negative bias. To better appreciate the crawling ability to construct network we further compare with the number of members having a Google Scholar page in the form of a corrected accuracy.

Correspondence: sebastien.dery@mail.mcgill.ca
Montreal Neurological Institute, McGill University, Montreal, 3801 University Street, H3A 2B4, QC, Canada
Full list of author information is available at the end of the article

Table 1 Completeness study: accuracy between the faculty roster of five major neuroimaging institutes and the neuroimaging network.

Institute	Total Count	Recovered	On Google Scholar	Accuracy	Corrected Accuracy
McConnell Brain Imaging Center (Montreal Neurological Institute)	12	7	9	58.33%	77.77%
Martinos Center for Biomedical Imaging (Harvard University)	39	12	22	30.76%	54.54%
Cognitive-Neuroimaging Unit (INSERM-CEA, France)	15	7	8	46.66%	87.50%
Wellcome Trust Center for Neuroimaging (University College London)	16	10	11	62.50%	90.90%
FMRIB (Oxford University)	17	8	11	47.05%	72.72%
Totals	99	44	61	49.06%	76.69%

3 Results

96 researchers were confirmed to have direct institutional affiliation to neuroscience, psychology, or biomedical engineering departments. The remaining 4 randomly selected researchers were found to work in the fields of human genome sequencing, image analysis, nano particles, and pharmacology. Note that these individuals were located on the outskirts of the main graph. To further assess completeness of the network, we compared results with faculty rosters of 5 major neuroimaging institutes (Table 1).

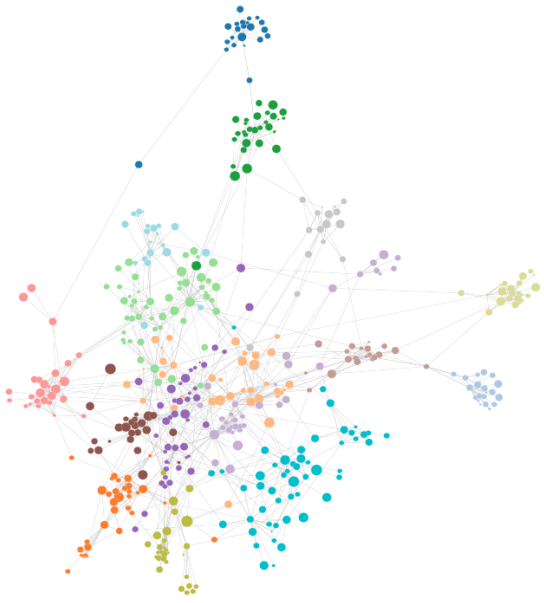

Figure 1 Coauthorship network for the field of neuroimaging. Each disk represent a single researcher with its radius encoding log10(Nc), where Nc is the number of citations. Edges stand for a binary relation of coauthorship between two researchers.

4 Conclusions

Accuracy results suggest a sufficient number of individuals are registered through GS to make it a useful platform of discovery. Meticulous inspection of the grouping suggest that communities typically embed either a geographical or a topical component, that is to say, certain communities are seemingly brought together by either proximity or similarity of interest. With the increasing complexity of science, finding accurate and relevant information on specific topics is a challenging task. We feel that a better appreciation of the wealth and variety of opinions within scientific communities may help enforcing the notion that grand claims require grand evidence.

Availability of Supporting Data

More information about this project can be found at: http://github.com/sderygithub/clubs-of-science. Further data and files supporting this project are hosted in the *GigaScience* repository REFXXX.

Competing interests

None

Author's contributions

SD wrote the software, performed tests, and wrote the report.

Acknowledgements

The authors would like to thank the organizers and attendees of Brainhack Montreal.

References

1. Freeman, L.C.: The Development of Social Network Analysis: A Study in the Sociology of Science. Empirical Press, ??? (2004)
2. Rybacki, H., Spies, J.R., Carp, J.M.: OSF SciNet. https://osf.io/4ujpn/ Accessed 2016-03-24
3. David, S.V., Hayden, B.Y.: Neurotree: A collaborative, graphical database of the academic genealogy of neuroscience. PLoS ONE **7**(10), 1–12 (2012). doi:10.1371/journal.pone.0046608
4. Sarigöl, E., Pfitzner, R., Scholtes, I., Garas, A., Schweitzer, F.: Predicting scientific success based on coauthorship networks. EPJ Data Science **3**(1), 9 (2014). doi:10.1140/epjds/s13688-014-0009-x
5. Holten, D., van Wijk, J.J.: Force-directed edge bundling for graph visualization. In: Proceedings of the 11th Eurographics / IEEE - VGTC Conference on Visualization. EuroVis'09, pp. 983–998. The Eurographs Association & John Wiley & Sons, Ltd., Chichester, UK (2009). doi:10.1111/j.1467-8659.2009.01450.x. http://dx.doi.org/10.1111/j.1467-8659.2009.01450.x
6. Blondel, V.D., Guillaume, J.-l., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks (2008)