Validating Science's Power Players: Scientometric Mixed Methods for Data Verification in Identifying Influential Scientists in a Genetics Collaboration Community

Sarah Elaine Bratt[1], Jian Qin[1], Jeffrey Joe Hemsley[1], Mark Raymond Costa[1], Jun Wang[1]
[1] Syracuse University, School of Information Studies

Abstract

The emergence of large international scientific data repositories has allowed cyber-enabled science a reflective look at itself through the lens of big scientometric analytics. Yet the drawbacks of using digital trace data from large social networks are well-documented. This poster reports the iterative process of interpreting complex network analysis (CNA) metrics of an international protein sequence data repository's metadata (*GenBank*) to identify influential "power players" in the genomics community. We describe preliminary work in developing approaches for operationalizing the mapping of CNA measures to establish a gold standard for node influence, arguing for the necessity of a confirmatory feedback loop in informing data interpretation and science funding- and submission- policy. Concrete examples and network visualizations are presented to illustrate the challenges in identifying influential scientists using GenBank metadata.

Keywords: cyber-infrastructure-enabled science; Scientometric data analytics; science of science policy; trace data validity
doi: 10.9776/16541

Acknowledgements: This research is sponsored by the NSF's Science of Science Policy Program.
Contact: Sarah Bratt {sebratt@syr.edu}

1 Introduction

There are many challenges in validating the trace data used in scientometric analytics. Known drawbacks of complex network analysis (CNA) have been discussed by (Howison, 2011), while methodological and technical challenges of database design and ensuring data quality were described by (Qin, Costa, & Wang, 2015). They underscore the importance of mixed methods, especially given the emergence of new forms of intellectual property units, knowledge products yet to be rigorously defined. Previous work on biodiversity data repositories developed data usage index (DUI) to incentivize data publishing amongst scientists (Ingwersen & Chavan, 2011). However, the operationalization collaboration in an international data repository is fundamentally distinct from measuring data usage and, to the best of our knowledge, has not been conducted.

This poster presents approaches for and examples of validating CNA data in scientific collaboration networks, arguing for the necessity of a partnership between CNA and periodic confirmatory analyses in the verification of network hubs to ground the construct of influence. In an effort to inform the science of science policy, we present preliminary results of an analysis of a unique data source: the National Center for Biotechnology Information (NCBI) data repository *GenBank* (http://www.ncbi.nlm.nih.gov/genbank), one of the largest protein sequence data repositories in the world.

Informing science policy and deriving recommendations from CNA data requires an iterative process of confirming that scientometric findings correspond with, and are reflective of, the reality of scientists' status and career development. We ask: Is the determination by centrality measures of *influential* scientists accurate? Are influential nodes, as defined by traditional complex network metrics, actually hubs in *the real world*? We use traditional scientometric and scale-free network metrics, identify items for verification such as Nobel Prize winners and disease outbreaks such as West Nile Virus, and confirm the results of analytics with web research.

2 Background of GenBank and CNA

Data Source: GenBank

Genetics underwent a revolution with the advent of international scientific data repositories, enabling scientists to connect across borders and with new technologies that enable development of sequencing techniques and species-based discoveries. Big data repositories like *Genbank* enable rapid and interdisciplinary collaboration, but also present methodological challenges. In the authors' project thus far, collection of metadata (Qin et al., 2015) and complex network analytics have been conducted to describe the collaboration dynamics and network characteristics.

Methods

Scientometrics is a domain that investigates science, innovation, and technology from a quantitative perspective. Researchers in this field have examined productivity, scholarly communication, impact, and the history of these phenomena in the sciences. Scientific practice and science policy are influenced by findings this methodological approach ("The Science of Science Policy," n.d.).

Complex Network Science Scale-free network analytics have a well-established history with a variety of metrics to summarize and describe the dynamics and structural properties of a network. In the metadata of a scientific community genetics repository, we understand *hubs* as influential scientists who have many links, and with the potential to influence the genetics community because of their highly-connected structural position. Centrality measures such as degree and eigenvector allow us to calculate and operationalize the influential scientists and powerful individuals (e.g., in terms of funding and knowledge contributions) over time, while community detection and plotting the community visually with multiple iterations of layout algorithms allows us to triangulate findings, conferring greater validity.

3 Verifying CNA methods: Hubs

Trace data has a track record of falling prey to validity worries (Howison, 2011). Drawbacks of complex social network analysis include, but are not limited to, confirming interpretation of findings (e.g., understanding sequencing submission processes and community detection vagaries) and the retrospective and temporal nature of the data. For example, the selection of a community detection layout algorithm requires additional online research to establish a ground truth of which nodes belong to community clusters. Semi-supervised and/or unsupervised learning computation (e.g., k-means clustering, Naive Bayes) classifies nodes in terms of links and CNA measures--classification decisions do not necessarily correspond to a scientist's species domain, or to her informal and formal co-authorship network clique(s). Clearly, establishing a "gold standard," i.e. true positives and negatives of influential authors and scientists requires "methods triangulation" (Howison, 2011).

Further, we assume that a Nobel Prize laureate in Chemistry will appear in the dataset. However, this is not the case, as found in a comparison of Nobel Prize winners over time in GenBank. Further, renowned scientists with high citation counts and those involved in large-scale sequencing projects such as the Human Genome Project may not be the primary submitter of sequence data. Future work is

required to correct for the idiosyncrasies associated with particular types of submissions to GenBank and their representation as metadata.

3.1 Influential Scientists

Scientists who are influential and prolific members of the genetics community are considered hubs; they constitute central parts of the GenBank collaboration network topography and often have greater access to funding resources, are readily positioned to disseminate information, drive development of new pharmaceutical discoveries, and have high social capital.

The network's macroscopic identification of bridges, hubs, and other influential nodes, and rise to their current network position in the network, are based chiefly on network metrics, used in our confirmatory workflow:

a) Centrality: Degree, Between-ness, Eigenvector
b) Map Equation: calculated with an information theoretic approach, using flow volume of a random walker to determine rank and scientists' network position) (Rosvall, Axelsson, & Bergstrom, 2009)
c) Network visualization: layout algorithms in R's iGraph package to identify patterns and central points of high connectivity

However, when these algorithms disagree about the hub status of a node, how does one establish a ground truth, or the "gold standard"? Operationalizing what *counts* as an influential person or hub is non-trivial. Methodological validity requires a convergence around all methods: quantitative, algorithmic, and qualitative alike.

3.2 GenBank Example

Specifically, we conduct research to collect supportive or disconfirming evidence of key indicators of hub-status, such as: successful funding, Nobel prize winning, current job title, number of publications, directorship of a genetics center, and Google Scholar and/or Pub Med profile.

Comparing degree centrality and eigenvector ranking of GenBank's influential authors: submission and publication network

eigen_degree_centrality_authorid	rank	infomap_author ID
1075129	1	1059043
1037638	2	968945
991145	3	830777
625261	4	1075129
1059043	5	815517

Table 1 *Identifying Influential Author Ranking Comparison: CNA methods and Map Equation (Infomap)*
We identified the top 25 influential authors (hubs) using eigenvector and degree centrality and the community detection algorithm InfoMap to identify influential "power player" nodes based on information theory, random walker-determined metric of flow volume of network links (Rosvall et al., 2009). Blue and yellow highlights identify authors who co-appear in both algorithms' ranking of the top 5 influential authors, suggesting levels of convergence between the two measures.

With the identification of these two well-connected and prolific nodes, we now use a visualization layout to inspect and compare the position in the network of the authors, both ranked highly by Infomap and iGraph's eigenvector and degree centrality measures.

Author ID	Reference year	Degree count	Between-ness centrality	Eigenvector centrality
1059043	2009	252	9714072	0.037123
1075129	2009	179	1492366	0.838831

Table 2 *Standard CNA metrics for 2 influential authors*
Authors from 25 influential authors subset are featured with traditional network measurements of degree, between-ness, and eigenvector centrality, assuming preferential attachment in the scientific collaboration community (i.e., the "rich get richer"). The R iGraph package was used for analysis.

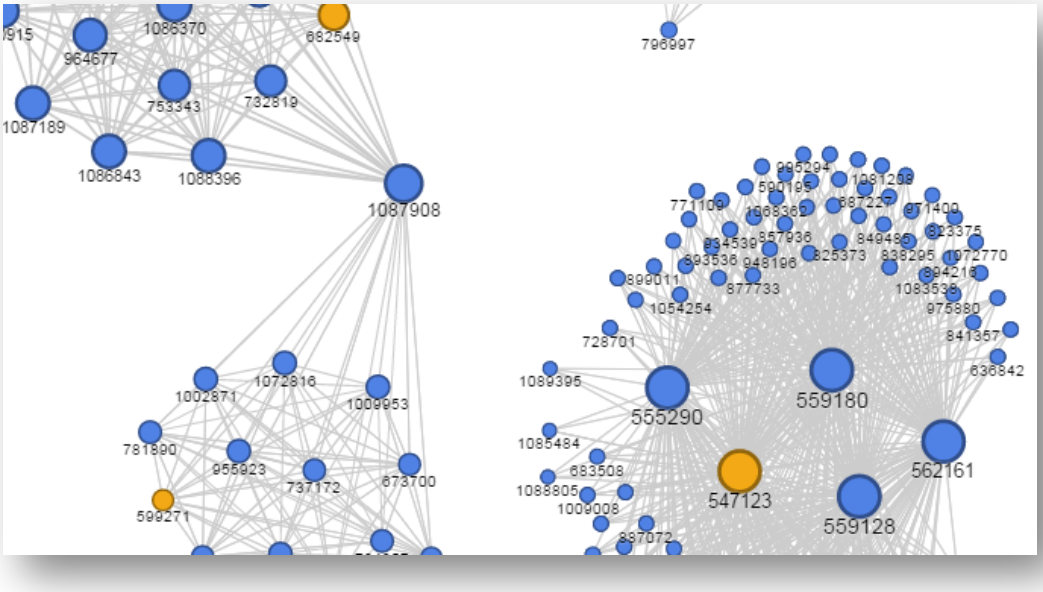

Figure 1 *Genbank submission and publication co-authorship network: Two influential nodes and their neighborhoods (2009)* Another approach to identifying influential nodes is network visualization. We compare multiple runs of the layout algorithm (we use iGraph's reingold.tilford layout) and Google Fusion table's chart visualization layout (above), (the latter weighted by frequency of reference submitted to GenBank). Note the structures of bridging communities and position of influential nodes with collaborators (colored gold and blue, respectively).

An additional challenge is author name ambiguity. Correctly identifying our two exemplar hub-scientists as the prolific hubs we have identified in web research is also a threat to efficient validation. Our two exemplar hub-scientists have common names. For example, after the name disambiguation performed by our team on the metadata, these hubs were identified as "J Wang" and "Y Xue," not infrequent names in the genetics publication arena.

4 Conclusion

Preliminary verification of samples of metadata is critical. Before developing more sophisticated algorithmic methods in CNA (e.g., node role analysis) to address the impracticality of manually verifying hubs in a dataset that contains more than 130,000,000 submission records, taxonomic terms, and a vast array of temporal data covering nearly 20 years of input, it is crucial to validate initial findings (Benson et al., 2013).

Granted, these confirmatory analyses are specific to our data set (GenBank), the community (genetics), and ultimate policy goals (the science of science policy). Yet deriving a model for a sample-workflow and establishing mixed methods approaches for identifying influential scientists in trace sets, especially to address the unique validity challenges posed by this type of data, is valuable and generalizable to analogous communities.

5 References

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2013). GenBank. *Nucleic Acids Research*, *41*(Database issue), D36–42. http://doi.org/10.1093/nar/gks1195

Howison, J. (2011). Validity issues in the use of social network analysis with digital trace data. *Journal of the Association for Information Systems*, *12*(12).

Ingwersen, P., & Chavan, V. (2011). Indicators for the Data Usage Index (DUI): an incentive for publishing primary biodiversity data through global information infrastructure. *BMC Bioinformatics*, *12*(Suppl 15), S3. http://doi.org/10.1186/1471-2105-12-S15-S3

Qin, J., Costa, M., & Wang, J. (2015). Methodological and Technical Challenges in Big Scientometric Data Analytics. *iConference 2015 Proceedings*. Retrieved from https://www.ideals.illinois.edu/handle/2142/73756

Rosvall, M., Axelsson, D., & Bergstrom, C. T. (2009). The map equation. *The European Physical Journal Special Topics*, *178*(1), 13–23. http://doi.org/10.1140/epjst/e2010-01179-1

The Science of Science Policy: A Handbook | Edited by Kaye Husbands Fealing, Julia I. Lane, John H. Marburger III, and Stephanie S. Shipp. (n.d.). Retrieved October 5, 2015, from http://www.sup.org/books/title/?id=18746