Cite as: Costa, M.R., Qin, J., & Wang, J. (2014). Research networks in data repositories. In: JCDL '14 14th ACM/IEEE-CS Joint Conference on Digital Libraries, London, England, September 22 - 26, 2014.

Research Networks in Data Repositories

Mark R. Costa
School of Information Studies
Syracuse University
mrcosta@syr.edu

Jian Qin
School of Information Studies
Syracuse University
jqin@syr.edu

Jun Wang
School of Information Studies
Syracuse University
jwang72@syr.edu

ABSTRACT

This paper reports our ongoing work investigating the structural features of scientific collaboration based on metadata collected from a scientific data repository (SDR). The background literature is reviewed in supporting our claim that metadata collected from SDRs offer a complimentary data source to traditional publication metadata collected from digital libraries. Methodological considerations are discussed in association with using metadata from SDRs, including author name disambiguation and data parsing. Initial findings show that the network has some unique macro-level structural features while also in agreement with existing networks theories. Challenges due to inconsistent metadata quality control procedures are also discussed in an attempt to reinforce claims that metadata should be designed to support both domain specific retrieval and evaluation and assessment needs.

Categories and Subject Descriptors

H.3.7 [**Digital Libraries**]: *User Issues, Collections.*

General Terms

Measurement, Design, Theory.

Keywords

Scientific data repositories, scientific collaboration, scientific collaboration networks, complex network analysis, metadata

1. INTRODUCTION

The emergence of scientific data repositories (SDRs) as tools for supporting the sharing, archiving, and processing of scientific data for research fields is thought to be changing the scale and structure of scientific collaboration [1]. However, little empirical work has been done on the structure of networks emerging around SDRs. In this paper, we discuss: some of the differences between traditional digital libraries designed for publications and SDRs, how SDRs may serve as a complementary source of data for studying scientific collaboration, and methodological challenges of working with metadata from SDRs. Finally, we report on our preliminary findings studying one network emerging around a specific SDR.

1.1 Scientific Data Repositories

Over the past two decades, there has been a shift to cyberinfrastructures (CI) enabled science. The National Science Foundation instituted an Advanced Cyberinfrastructure Division to provide guidance and develop a framework for maximizing the use of CI to promote scientific inquiry. Data sets are considered key components of 21st Century science and engineering [2]; as such, the generation, analysis, storage, sharing and reuse [3], [4] of data are critical policy priorities.

There are two approaches to storing, searching, and accessing data for reuse and integration. On one hand, scientific data repositories support the storage, dissemination, analysis and linking of data sets, and are being employed in a variety of fields from ecology to physics. Examples include the World Wide Molecular Matrix, ALADDIN for atomic and nuclear processes data, Brain Biodiversity Bank, and GenBank for genetic data. On the other hand, there are attempts to develop systems which support the search and use of smaller data sets produced by scientists for smaller, somewhat independent projects [5].

In this research we explore one example of an SDR – GenBank, the international nucleotide sequence databank. GenBank is run by the National Center for Biotechnical Information, and has been in operation since 1984. GenBank is now part of the international nucleotide sequencing consortium in which each member exchanging information daily.

GenBank houses genome sequencing data for over 130 billion base pairs covering 250 000 different species [6]. GenBank has doubled in size roughly every eighteen months. In addition to storing genetic information, GenBank provides a number of tools for discovery and analysis.

The repository has been well adopted by its respective community, which now requires all scientists to submit genetic data to the repository prior to publication. In addition to storing genetic information for scientific publications, GenBank serves as a repository for patented genetic information.

As stated earlier, SDRs are one aspect of cyberinfrastructure enabled (CI) science, and they are a key component of the “Fourth paradigm of science” proposed by Jim Gray [7]. The fourth paradigm refers to a data intensive approach to science, requiring the integration of skill sets from many domains [8]. As a result, it is argued that this data intensive approach is affecting the structure and scale of scientific collaboration [1]. Before going into detail about the structure of collaboration, we provide a brief justification for continued research on the phenomenon of scientific collaboration.

1.2 Scientific Collaboration

Scientific collaboration is a critical component of science along many dimensions; for example, it is a mechanism for socializing [9] and training [10] new researchers, sharing resources or expertise [11], and to lesser extent, a form of diplomacy [12]. Scientific collaboration is thought to spur economic development [13], particularly when integrated with government and industry partners [14]. Given the importance of collaboration to the scientific endeavor, Bozeman and colleagues [15] argued that each scientist should be looked at as having scientific and technical human capital. That is to say, a scientist’s value is based on both their domain knowledge and professional social network.

Consequently, there is considerable interest at the policy level for understanding and fostering scientific collaboration. The Science

978-1-4799-5569-5/14/$31.00 ©2014 IEEE

of Science Policy Program considers the study of the structure and evolution of networks, along with the development of subsequent methodology, tools, and techniques, to be key components of the program [16].

1.3 The structure of scientific collaboration

Although scientific collaboration has been studied for over five decades, the use of network analysis frameworks to study the phenomenon has only recently begun to be popular. Advances in graph theory and computational power have enabled their widespread application, which has resulted in rapid developments in network analysis frameworks.

We will not provide a review of the literature on scientific collaboration networks here, as our purpose is instead to look at this problem more from the intersection between methodological considerations and digital repositories. There are several good reasons for using network analysis framework for studying scientific collaboration. The first is that the reconstruction of collaboration networks helps us explore the theories behind the social construction of knowledge. The second reason is that reconstructions of networks can help us find subject matter experts [17]. Network analysis also helps us understand, visualize, and quantify different forms of status or roles within communities [18]–[21].

In the next section we will cover the roles digital libraries have played in the study of scientific collaboration networks and some of the challenges associated with using digital libraries as a source of data for the study of scientific networks. We will then discuss how scientific data repositories may be mined for collaboration data, the benefits of using SDRs as a source of data, and the challenges associated with their use.

2. METHODOLOGY

Although collaboration can be broadly construed, it is often operationalized as co-authorship, where any two scientists are considered to have collaborated if they wrote a paper together. There are known limitations of this approach to operationalization, both as it relates to under-representing and over-representing collaboration.

In terms of under-representing collaboration, using co-authorship as a measure of collaboration misses informal feedback which may be critical for the success of a project [22], and masks the shifting nature of contributions. To some extent, the former concern is well documented and addressed by the growth in use of acknowledgements to give credit to those who have made contributions that do not quite warrant co-authorship [23].

With respect to co-authorship over-representing collaboration, the primary concern is the tendency for researchers to give honorary co-authorship credit to other scientists for social reasons [24].

Although these concerns remain valid, co-authorship is still considered a useful measure for collaboration [25]. Given the fact that co-authorship is the most frequently used measure of collaboration, metadata from scientific articles is the primary source of data for collaboration studies. In turn, digital libraries are the primary source of metadata. It is important to note that some researchers use a combination of qualitative or ethnographic studies to augment the data collected from digital libraries [26].

Although co-authorship is still useful for measuring collaboration, it only represents a summary of contributions. That is, it obscures who contributed what to the creation of the product. For example, who is responsible for generating data for analysis? Traditionally this has been the work of the subject matter expert. However, more recently there has been an increasing role of specialists in the creation of data [27], [28]. How can we tell who contributed to the generation of the data, as most data products are mid-stream creations in the scientific process? One approach is to develop a contributory taxonomy which would allow scientists to match roles to authors when submitting research [29]. We can also look at knowledge products other than final publications (e.g., data sets), because in many cases scientists still lay intellectual claim to those products. We demonstrate here that this holds true when they submit data products to SDRs.

Therefore, we argue that SDRs are complimentary sources of data to explore scientific collaboration. If it is true, then metadata extracted from SDRs can be used as a source of data for validating and testing models and theories developed on article-based co-authorships. The remainder of this section will cover some of the challenges we have encountered using metadata from an SDR to study scientific collaboration.

Our source of data is GenBank (see above). In addition to making genetic data searchable via a web interface, NCBI hosts an ftp site that has all the data and metadata in semi-structured text files. In August 2013 we downloaded the entire contents of GenBank, parsed out all metadata from the text files, dropped the genetic sequence data, and then parsed the metadata into a MySQL database.

The next phase involved data cleanup and resolution. As with most sources of data, GenBank’s metadata quality is mediocre, with some inconsistencies in the formatting of organizational and consortium affiliations, and date entries. The costs associated with developing tools to parse the data is non-trivial, adds considerable costs to each assessment exercise, and speaks to the need for developing automated tools that assist in metadata quality improvement.

Additionally, named entity resolution continues to be a problem [30]. Almost all studies using co-authorship have this problem. An increasing number of studies use disambiguation algorithms to address this issue and ours is no exception. We too are employing disambiguation algorithms, and are able to leverage a subset of the data collected as ground-truth data because it is attached to well-curated metadata from PubMed.

One major challenge we face is understanding what constitutes a contribution. When working with publication information, researchers assume that each publication required a roughly equal amount of effort to produce, and therefore each publication is given equal weight. Based on that assumption, researchers can opt to give full-credit or fractional-credit [31], [32] for each publication scientists are listed as an co-author for. However, there is no easy way to apply either full or fractional counting methods to data contributions.

Data sets are uploaded as submissions, which include some set of annotations with distinct *gi* numbers. The annotations cover a subset of the total number of base pairs for the organism. Most submissions are associated with one or more *references*. References include publications, direction submissions, and unpublished references. The ratio of references to annotations for non-patented sequences is roughly 1:112.

However, things change when we include patents, with the ratio of annotations to patents being 1:1. This means that scientists will aggregate and report on many sub-sections of sequenced DNA in one publication while patenting each sub-section. This makes it difficult to understand or measure productivity. One possible solution is to use the number of base pairs per reference or patent as a guide, but this may differ from species to species making it difficult to apply across the board.

One additional consideration is how to select appropriate social foci for extracting sub-sections of the population for more in-depth analysis. Our intuition is that the species serves as the primary focal point for community formation, but we have not been able to empirically test that hypothesis yet.

In the next section we report on our macro-level analysis of the database. This includes measuring standard network analysis concepts such as the clustering coefficient, degree distribution,

3. PRELIMINARY RESULTS

After extracting the metadata from GenBank records, we have identified almost 176 million annotations to the database, covering some 1.35 million references and 25 million patents. As of the time of submission of this report, we have identified 545,000 individual authors after our first pass at name disambiguation. This does not include scientists associated with patents, as that data has not been fully parsed yet. There are 931 consortiums listed in the metadata, with an unresolved number of institutions (Table 1).

Table 1. Macro-level statistics

No. of scientists	545,534
No. of annotations	175,889,683
No. of references	1,351,049
Mean authors per reference	5.18
Mean references per author	12.5
No. of components	2,699
Size of giant component	98.3%
Clustering coefficient	.172

Like many previous studies, we see a power law distribution of connections and productivity. On the average, a scientist produced 12.5 submissions ($\mu = 12.5$, $\sigma = 194.8$) and a maximum value of 39,747. While the very large deviations may be the results of various reasons, it seems to also reflect the fact that some lab managers affix their name to every submission, which makes them appear to be exceptionally productive [33].

As for the number of scientists per reference, we have a μ=5.18, σ= 11.9 and a maximum of 415. Although very few references have that many scientists, we can see why researchers are concerned about the tendency for co-authorship to over-represent collaboration.

The size of a giant component is another finding worth highlighting. The giant component refers to the percentage of scientists in the community who are all directly or indirectly connected to one another. In some cases, there are multiple components that are not connected in any way, essentially indicating that the populations are isolated from one another. One of the initial concerns of using metadata as a source of data for studying collaboration was that it might not contain enough information to reconstruct a network. Yet our analysis indicates an exceptionally high level of connectivity, with it being much higher than results from other data sources, which often range from 57-88%, and one or two exceeding 90%.

The clustering coefficient, or the tendency of scientists to collaborate with their collaborators' collaborators, indicates that the community is only moderately well-connected, at least as far as social networks go. Coupled with the observed degree distribution we can say that this is a relatively hierarchical system with a few scientists controlling much of interactions. Finally, based on the average path length, the small-world theory that the average path length increases in proportion to the log of the number of scientists in the network holds true for this network as well [34].

4. DISCUSSION

In this report we discussed the increasing emphasis on developing and employing quantitative measures to track and assess scientific collaboration. More specifically, advancing theories, techniques and methodologies associated with studying scientific collaboration networks is a critical component of the Science of Science Policy Program [16].

In the past, digital libraries, including proprietary databases and open pre-print archives have served as the traditional source of data for studies on scientific collaboration. There are known limitations of using this data for studying collaboration. We have argued that scientific data repositories can serve as a complimentary source of data, both for verifying and advancing scientific theory.

We covered a number of methodological challenges associated with using SDRs as a source of data for collaboration studies. We would like to highlight the fact that metadata quality dramatically impacts the costs of using article archives and research data repositories as sources of data for these studies. As a result we suggest designing metadata schemes to support both domain specific information retrieval and evaluation and assessment needs.

Our initial analysis indicates that the metadata obtained from GenBank reasonably represent a community, as evidenced by the size of the giant component, clustering coefficient and average path length. Furthermore, the exhibited degree distribution conforms to our expectations, as it is similar to many other findings.

Looking to the future, SDRs will continue to serve as an important infrastructural component for the scientific process. Scientists will continue to share and, perhaps to a greater degree reuse, data. We also see data from SDRs being integrated with other sources into a more complete story of collaboration, supporting the inquiry of researchers and policy makers who are interested in learning who contributes what to the production of knowledge [29].

5. ACKNOWLEDGMENTS

This research is sponsored by the NSF's Science of Science Policy Program, grant number 1262535.

6. REFERENCES

[1] A. Szalay and J. Blakeley, “Gray's laws: Database-centric computing in science,” in *The fourth paradigm: Data-intensive scientific discovery*, 1.1 ed., T. Hey, S. Tansley, and K. Tolle, Eds. United States: Microsoft Corporation, 2009, pp. 5–12.

[2] Advanced Cyberinfrastructure Division, “Cyberinfrastructure framework for 21st Century science and engineering: A vision and strategy for data in science, engineering, and education,” National Science Foundation, Apr. 2012.

[3] I. M. Faniel and A. Zimmerman, “Beyond the data deluge: A research agenda for large-scale data sharing and reuse,” *Int. J. Digit. Curation*, vol. 6, no. 1, pp. 58–69, Aug. 2011.

[4] I. M. Faniel and T. E. Jacobsen, “Reusing scientific data: How earthquake engineering researchers assess the reusability of colleagues’ data,” *Comput. Support. Coop. Work CSCW*, vol. 19, no. 3–4, pp. 355–375, Aug. 2010.

[5] A. Rajasekar, H.-C. Kum, M. Crosas, J. Crabtree, S. Sankaran, H. Lander, T. Carsey, G. King, and J. Zhan, “The DataBridge,” *SCIENCE*, vol. 2, no. 1, pp. pp. 1–14, Sep. 2013.

[6] D. A. Benson, M. Cavanaugh, K. Clark, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and E. W. Sayers, “GenBank.,” *Nucleic Acids Res.*, vol. 41, no. Database issue, pp. D36–42, Jan. 2013.

[7] T. Hey, S. Tansley, and K. Tolle, Eds., “Jim Gray on eScience: A transformed scientific method,” in *The fourth paradigm: Data-intensive scientific discovery*, 1.1 ed., United States: Microsoft Corporation, 2009, pp. xvii–xxxi.

[8] J. R. Hunt, D. D. Baldocchi, and C. Van Ingen, “Redefining Ecological Science using data,” in *The fourth paradigm: Data-intensive scientific discovery*, 1.1 ed., T. Hey, S. Tansley, and K. Tolle, Eds. United States: Microsoft Corporation, 2009, pp. 21–26.

[9] D. Beaver and R. Rosen, “Studies in scientific collaboration: Part II. Scientific co-authorship, research productivity and visibility in the French scientific elite, 1799-1830,” *Scientometrics*, vol. 1, no. 2, pp. 133–149, Jan. 1979.

[10] B. Bozeman and E. Corley, “Scientists’ collaboration strategies: implications for scientific and technical human capital,” *Res. Policy*, vol. 33, no. 4, pp. 599–616, 2004.

[11] G. Melin, “Pragmatism and self-organization,” *Res. Policy*, vol. 29, no. 1, pp. 31–40, 2000.

[12] S. Choi, “Core-periphery, new clusters, or rising stars?: international scientific collaboration among ‘advanced’ countries in the era of globalization,” *Scientometrics*, vol. 90, no. 1, pp. 25–41, Sep. 2011.

[13] A. Abbasi, L. Hossain, S. Uddin, and K. J. R. Rasmussen, “Evolutionary dynamics of scientific collaboration networks: multi-levels and cross-time analysis,” *Scientometrics*, vol. 89, no. 2, pp. 687–710, Aug. 2011.

[14] L. Leydesdorff and H. Etzkowitz, “Emergence of a triple helix of university-industry-government relations,” *Sci. Public Policy*, vol. 23, no. 5, pp. 279–286, 1996.

[15] B. Bozeman, J. S. Dietz, and M. Gaughan, “Scientific and technical human capital: an alternative model for research evaluation,” *Int. J. Technol. Manag.*, vol. 22, no. 7–8, pp. 716–740, 2001.

[16] B. Valdez and J. Lane, “The science of science policy: A federal roadmap: A report to the Subcommittee on Social, Behavioral and Economic Sciences, Committee on Science, Office of Science and Technology Policy,” Office of Science and Technology Policy, Washington, D.C., DOE/SC-106, Nov. 2008.

[17] C.-Y. Lin, N. Cao, S. X. Liu, S. Papadimitriou, J. Sun, and X. Yan, “SmallBlue: Social Network Analysis for Expertise Search and Collective Intelligence,” in *IEEE 25th International Conference on Data Engineering, 2009. ICDE ’09*, 2009, pp. 1483–1486.

[18] H.-W. Chang and M.-H. Huang, “Prominent institutions in international collaboration network in astronomy and astrophysics,” *Scientometrics*, vol. 97, no. 2, pp. 443–460, Mar. 2013.

[19] D. H. Lee, I. W. Seo, H. C. Choe, and H. D. Kim, “Collaboration network patterns and research performance: the case of Korean public research institutions,” *Scientometrics*, vol. 91, no. 3, pp. 925–942, Jan. 2012.

[20] M. E. J. Newman, “Who is the best connected scientist? A study of scientific coauthorship networks,” *Complex Netw.*, pp. 337–370, Jan. 2004.

[21] X. Wang, S. Xu, D. Liu, and Y. Liang, “The role of Chinese–American scientists in China–US scientific collaboration: a study in nanotechnology,” *Scientometrics*, vol. 91, no. 3, pp. 737–749, Mar. 2012.

[22] K. Subramanyam, “Bibliometric studies of research collaboration : A review,” *J. Inf. Sci.*, vol. 6, no. 1, pp. 33–38, 1983.

[23] B. Cronin, D. Shaw, and K. La Barre, “A cast of thousands: Coauthorship and subauthorship collaboration in the 20th century as manifested in the scholarly journal literature of psychology and philosophy,” *J. Am. Soc. Inf. Sci. Technol.*, vol. 54, no. 9, pp. 855–871, 2003.

[24] J. S. Katz and B. R. Martin, “What is research collaboration?,” *Res. Policy*, vol. 26, no. 1, pp. 1–18, Mar. 1997.

[25] W. Glänzel and A. Schubert, “Analysing scientific networks through co-authorship,” in *Handbook of Quantitative Science and Technology Research: The Use of Publication and Patent Statistics in Studies of S & T Systems*, Dordrecht: Kluwer, 2005, pp. 257–276.

[26] T. Velden, A. Haque, and C. Lagoze, “A new approach to analyzing patterns of collaboration in co-authorship networks - mesoscopic analysis and interpretation,” *Scientometrics*, vol. 85, no. 1, pp. 1–37, 2010.

[27] A. Swan and S. Brown, “The skills, role and career structure of data scientists and curators: An assessment of current practice and future needs,” Sep-2008. [Online]. Available: http://eprints.soton.ac.uk/266675/. [Accessed: 29-Jan-2014].

[28] G. Pryor and M. Donnelly, “Skilling Up to Do Data: Whose Role, Whose Responsibility, Whose Career?,” *Int. J. Digit. Curation*, vol. 4, no. 2, pp. 158–170, Oct. 2009.

[29] L. Allen, A. Brand, J. Scott, M. Altman, and M. Hlava, “Credit where credit is due,” *Nature*, vol. 508, pp. 312–313, 2014.

[30] C. A. D. Angelo, C. Giuffrida, and G. Abramo, “A Heuristic Approach to Author Name Disambiguation in Bibliometrics Databases for Large-Scale Research Assessments,” *J. Am. Soc. Inf. Sci. Technol.*, vol. 62, no. 2, pp. 257–269, 2011.

[31] D. J. de Solla Price and D. Beaver, “Collaboration in an invisible college,” *Am. Psychol.*, vol. 21, no. 11, pp. 1011–1018, 1966.

[32] X. Liu, J. Bollen, M. L. Nelson, and H. Van de Sompel, “Co-authorship networks in the digital library research community,” *Inf. Process. Manag.*, vol. 41, no. 6, pp. 1462–1480, 2005.

[33] M. E. J. Newman, “The structure of scientific collaboration networks,” *Proc. Natl. Acad. Sci.*, vol. 98, no. 2, pp. 404–409, Jan. 2001.

[34] D. J. Watts and S. H. Strogatz, “Collective dynamics of ‘small-world’ networks,” *Nature*, vol. 393, no. 6684, pp. 440–442, Jun. 1998.