SCIENCE OF TEAM SCIENCE (SCITS) 2018 CONFERENCE
Galveston, Texas: Moody Gardens Hotel and Conference Center
May 21-24, 2018
Hosted by the University of Texas Medical Branch Institute for Translational Sciences

Collaboration Capacity: Measuring the Impact of Cyberinfrastructure-Enabled Collaboration Networks

Jian Qin[1], Jeff Hemsley, Sarah Bratt

School of Information Studies, Syracuse University, Syracuse, NY 13244

Keywords: Collaboration capacity; Collaboration networks; Big metadata analytics; GenBank data repository; Research impact assessment.

Abstract: This paper reports a study of the incremental impact of evolving cyberinfrastructure (CI)-enabled collaboration networks on scientific capacity and knowledge diffusion. While ample research shows how collaboration contributes to greater productivity, higher-quality scientific outputs, and increased probability of breakthroughs, it is unclear how the early stages of collaboration on data creation supports knowledge generation and diffusion. Further, it is not known whether the ability to garner larger inputs[2] increases collaboration capacity and subsequently accelerates the rate of knowledge diffusion.

Given that the collaboration capacity of a science team is largely dependent upon the Scientific and Technical (S&T) Human Capital[3], the greater a researcher's S&T human capital, the greater the opportunity to collaborate and access resources. We use "*Collaboration Capacity*" to refer to this measure of S&T human capital.

In this study, we collected metadata for molecular sequences in GenBank[4] [5] from 1990-2013. The data contain details about sequences, submission date, submitter(s), and associated publications and authors. Based on the collaboration capacity framework (Figure 1), we focused on the relationship between collaboration network size and research productivity and the role of CI-enabled data repositories in accelerating collaboration capacity. Our preliminary results show that the size of CI-enabled collaboration networks at data creation stage was positively related to research productivity as measured by sequence data production, and the extent and rate of knowledge diffusion, represented by patent applications. Shrinking time gaps between data submissions and patent applications support the hypothesis that CI-enabled data repositories are an accelerating factor in incremental collaboration capacity.

[1] Corresponding Author Contact Information: jqin@syr.edu; (315)443-5642
[2] Stephan, P. (2012). How Economics Shapes Science. Cambridge, MA: Harvard University Press.
[3] Bozeman, B., Dietz, J., & Gaughan, M.: Scientific and technical human capital: an alternative model for research evaluation. International Journal of Technology Management, 22: 636–655 (2001).
[4] NCBI-a. GenBank overview, http://www.ncbi.nlm.nih.gov/genbank/.
[5] NCBI-b. Growth of GenBank and WGS, http://www.ncbi.nlm.nih.gov/genbank/statistics.

1 Introduction

Cyberinfrastructure (CI)-enabled data repositories are the systems that store and manage scientific data and provide data submission and discovery services for long-term curation, sharing, and reuse of scientific data. The Knowledge Network for Biocomplexity (KNB, 2017) and GenBank (NCBI-a, 2017) are examples of such data repositories. The fast growth of CI-enabled data repositories and services in the last four decades has played s crucial role in the paradigm shift in science from empiricism, theory, and simulation to data (a.k.a. the fourth paradigm), as Jim Gray envisioned (Gray et al., 2005; Gray, 2007) and subsequently articulated by Szalay & Blakeley (2009). Science today, small- or large-scale, is increasingly carried out through distributed global collaborations enabled by these cyberinfrastructures. The CI-enabled data repositories administered at the National Center for BiologiInformation (NCBI) store "massive amounts of genetic sequence data generated from evolving high-throughput sequencing technologies" and serve "more than 30 terabytes of biomedical data to more than 3.3 million users every day" (NLM, 2015).

While published papers in science keep growing in quantity over time, the amounts of data increase at an even faster pace. One of the NCBI data repositories, GenBank, has had a steady increase in the number of submissions of genetic sequences, doubling roughly every 18 months since 1982 (NCBI-b, 2017). The rapid increase in science data is supported by CI-enabled tools and services – the large number of tools for using the vast biomedical data available on NCBI's website underlines the importance of CI-enabled tools and services in data-driven science. What is unclear in this grand picture of data-driven science is how this changing climate of science research has affected the productivity and rate of knowledge diffusion.

Traditional measures for research productivity and knowledge diffusion typically include the number of publications, citations, and patents. Data collected or generated from observations, lab experiments, and simulations have rarely been included in research impact assessment, much less for playing a part in science of science and policy research. As data-intensive science increasingly becomes the norm in the digital era, it becomes necessary to reexamine the impact assessment metrics and introduce scientific data production into the impact assessment equation.

This paper reports preliminary findings from an NSF funded project that investigates how CI-enabled collaboration networks evolved, as represented in a data repository, and how the data-driven collaboration networks made an incremental impact on scientific capacity and knowledge diffusion. The preliminary findings focus on providing a general picture of collaboration networks in both publications and data submissions and highlighting patterns and/or characteristics of the CI-enabled collaboration networks for potential future exploration.

2 Theory of Collaboration Capacity

Research on collaboration looks at individual scientists and their interaction with one another at various levels (individual, institutional, national, international, community, cross-community, cross-discipline). This body of work also looks at the impact of those interactions on research productivity and science policy. In an effort to understand the nature and properties of scientific collaboration networks, both quantitative and qualitative approaches have been applied. In the quantitative stream of research on scientific networks, the most frequently used measure is co-authorship, which measures the interdisciplinarity of collaboration (Qin et al, 1997; Porter & Rafols, 2009). Others have examined team assembly mechanisms' effect on network structure and team performance (Guimerà et al., 2005), quantitatively modeled the structure of collaboration networks (Newman, 2001; Newman, 2003), and the evolution of collaboration networks (Barabási et al., 2002).

Collaboration in research is considered "social processes in which researchers pool their experience, knowledge, and social skills with the objective of producing new knowledge, including knowledge embedded in technology" (Bozeman & Boardman, 2014, p. 2). The occurrence, scale, and success or failure of collaboration may be affected by many factors, including compatibility of work style, work connections, incentives, and social-technical infrastructures (Hara et al., 2003). The ability of researchers to engage in different types of collaboration, whether it is within or outside of one's workplace or discipline, is determined not only by the abovementioned factors, but also by the *Scientific and Technical (S&T) Human Capital*, a concept defined as the sum of scientific, technical and social knowledge, skills and resources embodied in a particular individual (Bozeman et al., 2001). As collaboration is mainly about S&T human capital, we assume that the greater the S&T human capital a researcher has, the more opportunity and resources he or she can garner to collaborate with other researchers. For this notion, we use the term "*Collaboration Capacity*" as a measure for one's S&T human capital. In other words, a researcher with rich S&T human capital will have higher collaboration capacity than those having sparse S&T human capital.

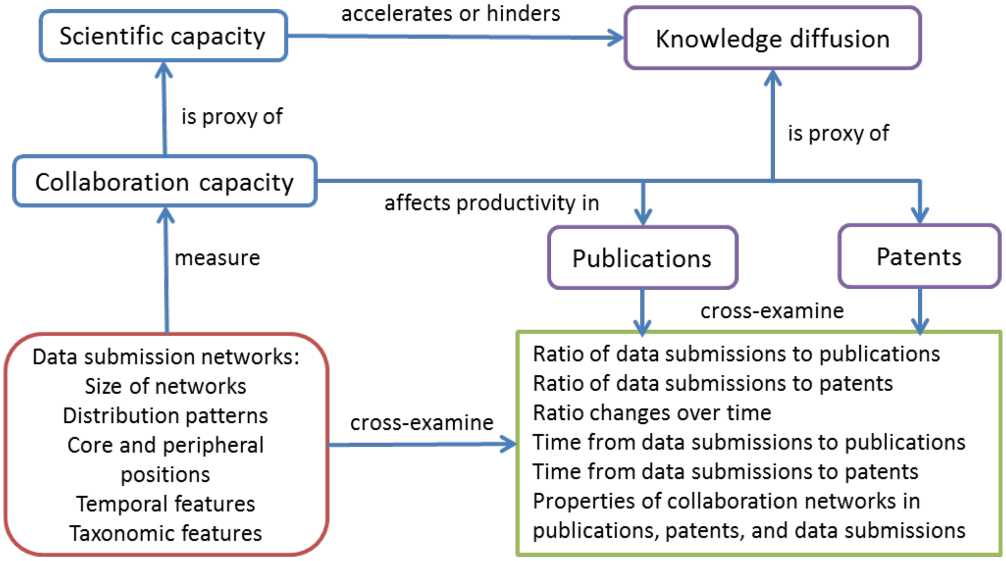
Scientific capacity
accelerates or hinders
Knowledge diffusion
is proxy of
is proxy of
Collaboration capacity
affects productivity in
Publications
Patents
measure
cross-examine
Data submission networks:
Size of networks
Distribution patterns
Core and peripheral positions
Temporal features
Taxonomic features
cross-examine
Ratio of data submissions to publications
Ratio of data submissions to patents
Ratio changes over time
Time from data submissions to publications
Time from data submissions to patents
Properties of collaboration networks in publications, patents, and data submissions

Figure 1. An illustration of the concepts and relationships in the theory framework for the impact assessment of collaboration capacity

3 Methods

One way to operationalize collaboration capacity is to use collaboration network measures to uncover the sizes, patterns, and status for individuals, groups, institutions, or communities in collaboration networks. As any research activity cannot take place without funding, knowledge, ability, facilities, and materials, the ability to acquire large amounts of these needed inputs as well as work in teams becomes a signature characteristic of CI-enabled research. In data-driven science, productivity and innovative output are critically dependent upon collaborative work prior to the final stage of a research lifecycle: publication of a paper or a patent application. Capturing data on collaboration networks prior to paper publications and patent applications will not only offer new empirical sources but also generate new insights into how collaboration at the data creation stage affects research productivity and innovative discoveries.

GenBank is an international data repository for DNA/RNA sequence datasets. Each annotation record in this repository consists of a metadata section and a sequence data section. The metadata section includes information on the sequence data submission as well as publication(s) associated with the DNA sequence data submission. The annotation records are submitted by researchers from around the world and in a semi-structured format. We downloaded the GenBank Release 191.0 on 8/16/2013 from the FTP server, which covers data from its beginning in 1982 to June 2013. We only needed the metadata for our research questions; the sequence data, which comprises the bulk of the file size (over 90% of the file in many cases) and not needed for the purposes of this study, were therefore dropped.

From our pretest of data collection strategies, we adopted a workflow that was computationally efficient for collecting the metadata needed for this study. The workflow includes the following steps: download one compressed sequence file from the FTP server → decompress the file → extract the metadata section from each record in the file → save the metadata records to a buffer space → delete the downloaded file → parse the metadata into database→ repeat the workflow for next compressed file on the FTP server. A computer program was created to automatically complete these steps in a batch process. We set up a data server with the necessary software and storage space for

the GenBank metadata extractions as per Costa et al. (2014 & 2016).
Our data analysis includes exploratory data analysis (EDA) (Tukey, 1977) and social network analysis (Wasserman & Faust, 1994). EDA uses descriptive statistics, tables, aggregation and data visualization techniques to make sense of the data and can easily be scaled to very large datasets. The objective is to explore the data looking for patterns, structures, problems and both the expected and un-expected. Correct use of EDA requires that practitioners have a willingness to work with and explore the data in different ways. EDA can provide researchers with both a broad and deep understanding of what is their data, the kinds of questions that the data can answer, and the quality of those answers. Social network analysis is a collection of methods that allows researchers to study the patterns in social networks. In our case, scientists are the nodes in the network and coauthoring a publication or making a data submission together, are the links between them. We use centrality measurements (calculated attributes of the nodes based on the linking structure of the network and their place in it) and centralization measures (network wide measures at a given point in time).

4 Findings

4.1 Connectedness of collaboration networks

Between 1994 and 2012[6], the number of data submissions maintained a steadily faster increase than that of publications (Fig. 1). Collaboration networks in GenBank show a number of characteristics. First, there was a consistent increase in collaborative and connected networks over the years. The measures for network connectedness show a clear trend of increase: the size of the giant component (publication and submission networks together) increased from 48.1% of all scientists in 1994 to 80.8% in 2012 (Fig. 2). The increase in the percentage of edges (connection between nodes) in the giant component grew at a rate consistently larger than that of the nodes (Fig. 3), which indicates that those who collaborated with others tended to gain more connections over time than those who worked alone or in small, isolated groups. The mean degree for publication networks increased from 6.345 in 1994 to 11.98 in 2012 with some fluctuations while the mean degree for submission networks grew from 4.844 in 1994 to 10.12 in 2012 (Fig. 4). However, the clustering coefficient, which measures how clustered nodes are in the network, had a sharp drop after 2007. While this phenomenon is worth further exploring, we speculate that two possible factors may have contributed to the drop: one is the availability and lowered cost of advanced sequencing technology made it possible for individuals or smaller networks to pursue diverse studies, and the other is the ending of large scale sequencing projects such as Human Genome Project. The result could have been a flattening of the collaboration networks where a larger number of smaller networks formed and connected to the hubs through a few edges, rather than historically highly connected and clustered around hubs (Fig. 2).

4.2 Preferential attachment

Based on the scale-free networks theory, networks that have power law degree distributions bear two characteristics: the number of nodes in the network grows over time and the nodes having high degree are more likely to receive new links (Barabási et al., 2002). We computed the alpha value, which indicates the steepness of the power law, in the power law degree distribution for each year. The first three years had an alpha value greater than 3, while for the rest of the years, alpha fell under 3, suggesting that more medium sized hubs accumulated in the network in later years (fig. 5).

[6] The data prior 1994 were extremely sparse for both DNA sequence submissions and publications associated with the sequence. We aggregated all data from 1982-1994 in the 1994 group. Because data in 2013 were only up to June, all records in 2013 were also dropped.

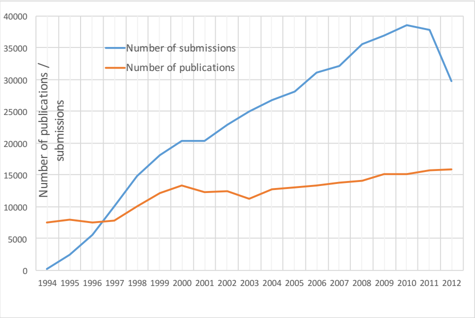

Fig. 1. Frequency distribution of data submissions and publications in GenBank 1994-2012

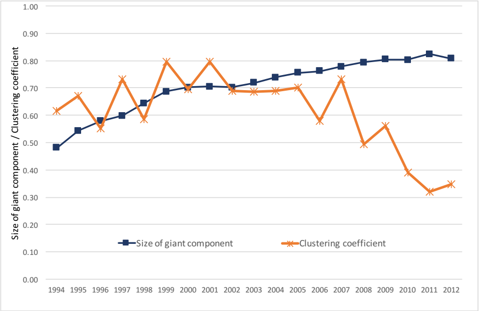

Fig. 2. The size of gian component and clustering coeffificent in GenBank 1994-2012

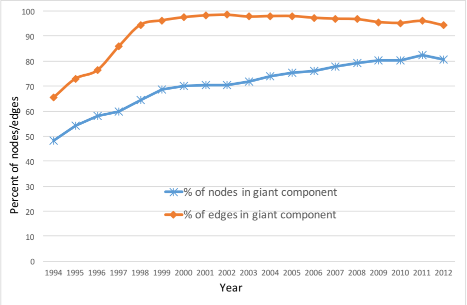

Fig. 3. Percentage distribution of nodes / edges in giant component: 1994-2012

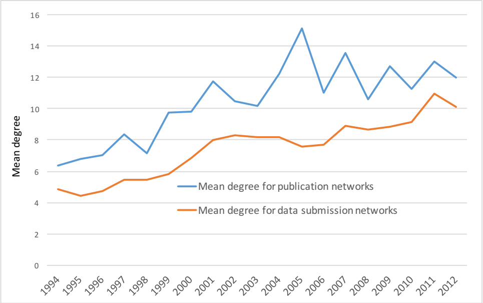

Fig. 4. Mean degree distribution by year: 1994-2012

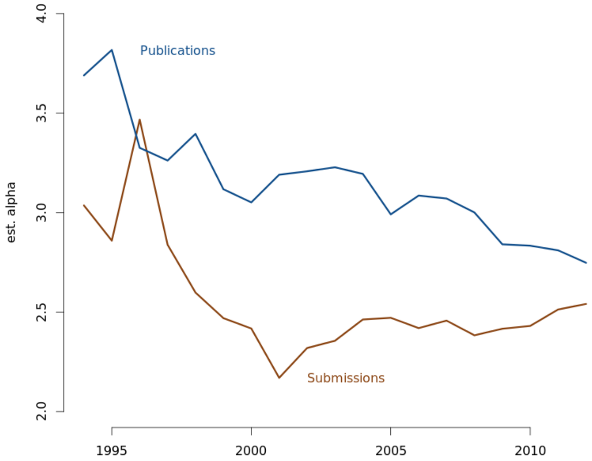

Fig. 5. Change of Alpha value in power law distribution over time

The L-shaped degree distribution for both publication and submission networks (Fig. 6) confirms this pattern, which can be interpreted as that a large number of nodes were highly connected to a small number of nodes while those nodes in the long tail were much less connected with others. This property illustrates a preferential attachment process, or the "Matthew effect", in the GenBank collaboration networks.

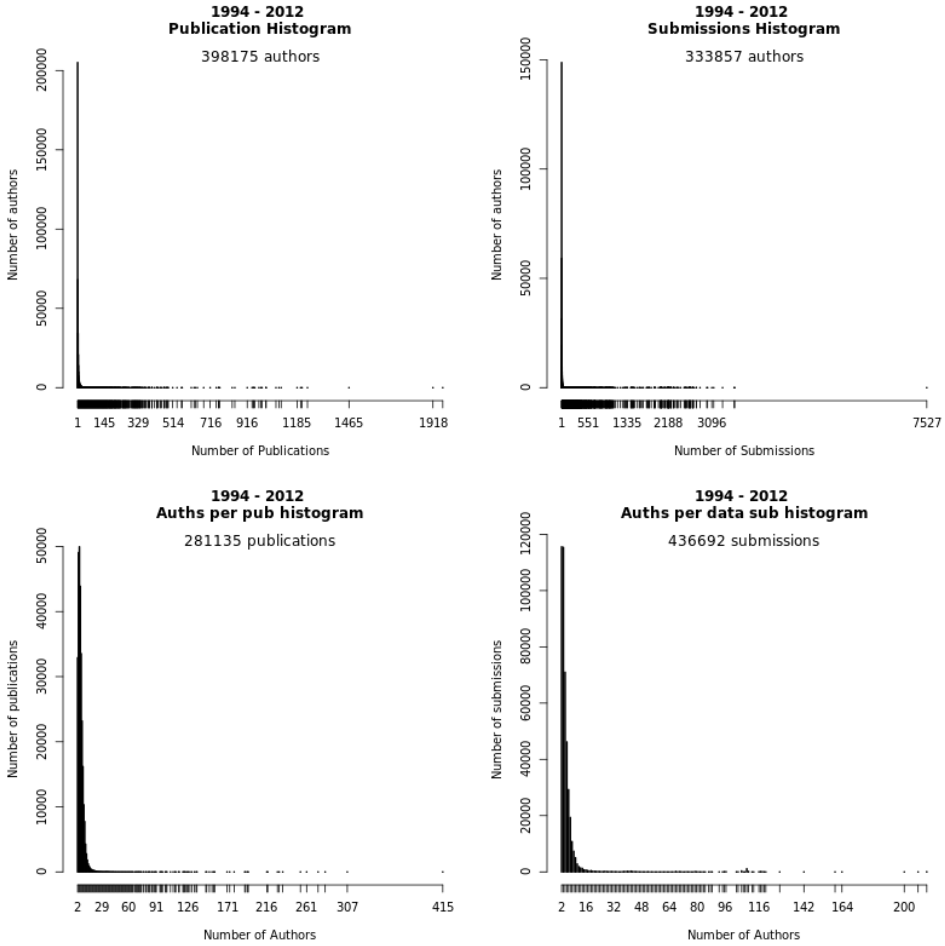

Fig. 6. Histograms of authors vs. publications and submissions for all years

4.3 Ratio of data submissions to publications

In cyberinfrastructure-enabled data-intensive science, the size of data submission networks (which consist of authors and datasets submitted among other measures), distribution patterns of such networks, core and peripheral positions of nodes, as well as temporal and taxonomic features can be used to measure the collaboration capacity. In this paper, we focus on the ratio of data submissions to publications.

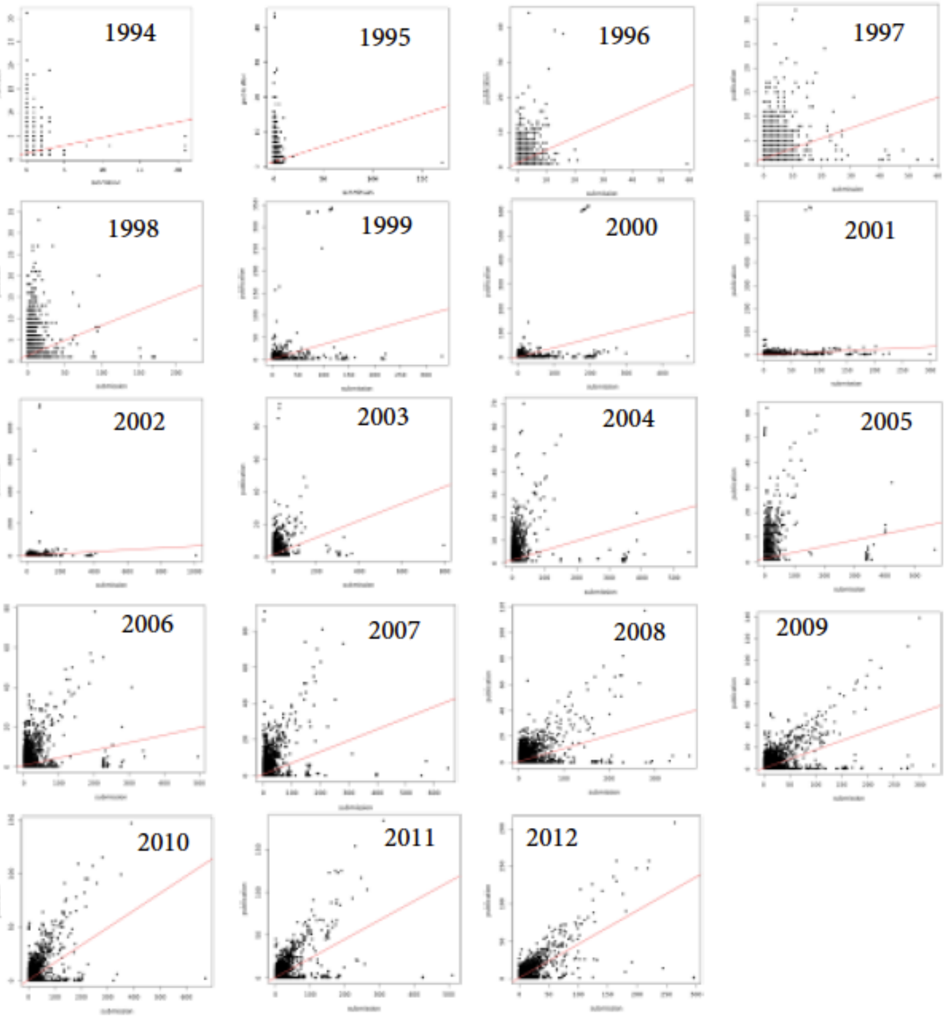

Fig. 7. Distribution of number of authors with trend line for 1994 to 2012. The x axis represents the number of authors who submitted sequence data and y axis represents those who had published a paper associated with the submissions.

As mentio**ned** earlier, the size of giant component in both data submission and publication networks steadily increased over the 19 years covered by this dataset. The year to year distributions in Fig. 7 illustrates that a majority of authors concentrated in the range of approximately 20 publications and 100 submissions after 2000. It is noticeable that authors with extremely large numbers of publications had lower number of data submissions. Similarly, authors who contributed large numbers of data submissions were not among those with high number of publications. This pattern started to change around 2008 whereas highly productive authors appeared to scatter diagonally in the submission-publication space.

While it is obvious that not all authors in data submission networks were also in the publication networks, or vice versa, the average ratio of submission to publication showed an upward trend, with a sharp increase from less than one (1) to 4.01 around 2005 (Fig. 8), which was perhaps a turning point for microbiology to become a data-intensive science.

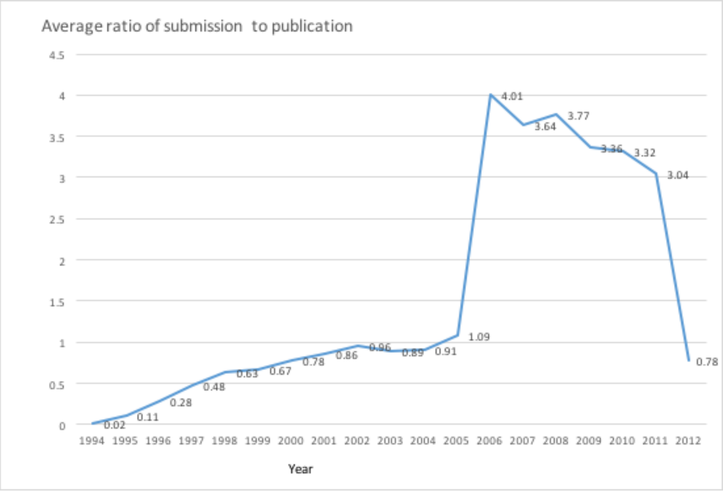

Fig. 8. Average ratio of data submission to publication in GenBank: 1994-2012

5 Discussion and Future Study

Collaboration capacity as a measure for assessing the impact of collaboration networks on research productivity and knowledge diffusion operationalizes the S&T human capital theory and makes it possible tie the collaboration network analysis with theory so that the interpretation of collaboration networks analysis can be built on a solid, meaningful theory ground. This is the major contribution of this paper. Even though several important concepts may have not been given full explanation due to the page limit of this paper, the findings provide a number of directions for further exploration.

The first direction lies in deeper mining of the collaboration networks in this very large community. In CI-enabled data-intensive science, the size of data submission networks (which consist of authors and datasets submitted among other measures), distribution patterns of such networks, core and peripheral positions of nodes, as well as temporal and taxonomic features can be used to measure the collaboration capacity as well as the impact of collaboration capacity on productivity and scientific and innovative capacity. We envision that collaboration capacity as a theory-backed measure will provide a useful way to examine and evaluate CI-enabled collaboration networks and their impact on scientific and innovative capacity at different levels.

Another direction is to uncover more specific features and patterns of data submission networks and the relationships between data submission and publication networks. The data show how a disciplinary field such as genetics or microbiology evolved from a publication-centric to data-intensive paradigm. Our ongoing analysis found that, while “super-hubs” emerged and remained consistent throughout the whole period of time, the data submission networks had been increasingly branching out. This trend provides evidence for explaining the decrease in clustering coefficient. To gain more insights into the quantitative phenomenon, we need to collect events in economic, technological, policy, and other domains that correspond to the GenBank to tell the whole story of the rise and change of data submission and publication networks in this community.

While not all authors in data submission networks were in the publication networks, the average ratio of submission to publication appeared to be on an upward trend. The ratio of data submissions to publications can be perhaps seen as evidence for when microbiology turned into a data-intensive science. The changes in the ratio of data submission to publication raise further questions for future research: to what extent data submission networks accelerated and/or facilitated the creation of new knowledge as represented by publications and patents? More broadly, how have data-intensive biology impacted the emergence and evolution of new research areas such as precision medicine? The ratio of submission to publication will be a metric worth further analysis and development for assessing the impact of cyberinfrastructure-enabled data-intensive science.

Acknowledgement: This project is supported by [Funding information omitted for blind review].

References

1. Barabási, A.L., Jeong, H., Néda, Z., Ravasz, E., Schubert, A., & Vicsek, T.: Evolution of the social network of scientific collaboration. Physica A, 311, 590-614 (2002).
2. Bozeman, B., Dietz, J., & Gaughan, M.: Scientific and technical human capital: an alternative model for research evaluation. International Journal of Technology Management, 22: 636–655 (2001).
3. Bozeman, B. & Boardman, C.: Research collaboration and team science: A state-of-the-art review and agenda. Springer, Heidelberg (2014).
4. Costa, M., Qin, J., & Bratt, S.: Emergence of collaboration networks around large-scale data repositories: A study of the genomics community using GenBank. Scientometrics, 108(1), 21-40 (2016).
5. Gray, J., Liu, D.T., Nieto-Santisteban, M.A., Szalay, A.S., Heber, G., & DeWitt, D.: Scientific data management in the coming decade. Microsoft Research Technical Report. MSR-TR-2005-10, https://www.microsoft.com/en-us/research/wp-content/uploads/2005/01/tr-2005-10.pdf, last accessed 2017/09/17.
6. Gray, J.: Jim Gray on eScience: A transformed scientific method. In: T. Hey, S. Tansley, & K. Tolle (eds.), The Fourth Paradigm: Data-Intensive Scientific Discovery, pp. xvii-xxxi. Redmond, WA: Microsoft (2007).
7. Guimerà, R., Uzzi, B., Spiro, J., & Nunes Amaral, L.A.: Team assembly mechanisms determine collaboration network structure and team performance. Science, 308(5722), 697-702 (2005).
8. Hara, N., Solomon, P., Seung-Lye, K., Sonnenwald, D. H.: An emerging view of scientific collaboration: Scientists' perspectives on collaboration and factors that impact collaboration. Journal of the American Society for Information Science and Technology 54(10), 952-965 (2003).
9. KNB. The Knowledge Network for Biocomplexity, https://knb.ecoinformatics.org/, last accessed 2017/09/17.
10. NCBI-a. GenBank overview, http://www.ncbi.nlm.nih.gov/genbank/, last accessed 2017/09/17.
11. NCBI-b. Growth of GenBank and WGS, http://www.ncbi.nlm.nih.gov/genbank/statistics, last accessed 2017/09/17.
12. NLM. Congressional Justification FY2015: Department of Health and Human Services, National Institute of Health, National Library of Medicine, http://www.nlm.nih.gov/about/2015CJ.html, last accessed 2017/09/17.
13. Newman, M.E.J.: The structure and function of complex networks. SIAM Review, 45, 167-256 (2003).
14. Newman, M.E.J.: The structure of scientific collaboration networks. Proceedings of National Academy of Science, 98(2), 404-409 (2001).
15. Porter, A.L. & Rafols, I.: Is science becoming more interdisciplinary? Measuring and mapping six research fields over time. Scientometrics, 81(3), 719-745 (2009).
16. Qin, J., Lancaster, F.W., & Allen, B.: Levels and types of collaboration in interdisciplinary research. Journal of the American Society for Information Science, 48(10), 893-916 (1997).
17. Szalay, A.S. & Blakeley, J. A.: Grey's laws: Database-centric computing in science. In: T. Hey & S. Tansley (eds.) The Fourth Paradigm: Data-Intensive Scientific Discovery, pp. 5-11. Microsoft Research, Redmond, WA (2009).
18. Tukey, J. W.: Exploratory data analysis. Pearson, Reading, Mass. (1977).
19. Wasserman, S., & Faust, K.: Social network analysis: Methods and applications. Cambridge Univ Pr., Cambridge, UK (1994).