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

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

Corresponding Author: Jian Qin, Ph.D., School of Information Studies, Syracuse University, Syracuse, NY 13244, United States

Co-Authors: Jeff Hemsley, Ph.D., School of Information Studies, Syracuse University, Syracuse, NY 13244, United States; Sarah Bratt, Ph.D. student, School of Information Studies, Syracuse University, Syracuse, NY 13244, United States

Corresponding Author Contact Information: jqin@syr.edu; (315)443-5642

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 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, 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 from 1990-2013. The data contain details about sequences, submission date, submitter(s), and associated publications and authors. Based on the collaboration capacity framework, 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 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. Another example are the data repositories at the National Center for Biotechnology Information (NCBI), which 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). After more than three decades' investment by the federal government into building data repositories and related services, the fast growth of CI-enabled data repositories and services has played a significant role in the paradigmatic 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 has been subsequently articulated by Szalay & Blakeley (2009). Science today, small- or large-scale, is increasingly carried out through distributed global collaborations enabled by these cyberinfrastructures.

Molecular sequencing is an important research field at the most fundamental level of biological research. Technological advances in molecular sequencing have resulted in innovations, increased capabilities in sequencing while decreased costs, and data and computational intensive biological science (Heather & Chain, 2016). GenBank, as the sequence data repository, records not only genetic sequences and related publications, but also the history of collaboration and research productivity in the molecular biology research community. The metadata in GenBank includes information about the DNA sequences—the authors and their affiliations, locus and source of sequence, taxon lineage, publications associated with the sequence, and submission date. The GenBank metadata not only offers a great case to study team science in the context of a large research community that is (relatively) disciplinarily homogeneous while geographically and sectorially diverse, but can also can provide useful leads for in-depth, qualitative inquiries for further study of this large community.

Many factors have contributed to the growth in GenBank data, but this paper focuses on team science, and specifically on the collaboration as a factor in productivity and knowledge diffusion. We argue that the ability a scientist to collaborate with others in conducting research, i.e., collaboration capacity, is an indicator of her or his *scientific and technical human capital (S&THC)*, a concept defined by Bozeman et al. (2001). As an operationalization of the S&THC theory, we use collaboration capacity as a framework of metrics to measure the CI-enabled collaboration networks as represented in the GenBank data repository, and the impact of data-driven collaboration networks on scientific capacity and knowledge diffusion. In the rest of this paper, we will first elaborate on the collaboration capacity framework, and then describe the methods and findings, which followed by a discussion of the findings and implications for team science research.

2 A Framework 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 can be sliced by author's disciplinary fields to measure the interdisciplinarity of collaboration and by geographical locations to measure inter-institutional and international collaborations (Qin et al, 1997; Porter & Rafols, 2009). Co-author data also have been used to examine the effect of team assembly mechanisms on network structure and team performance (Guimerà et al., 2005), quantitative modeling of the structure of collaboration

networks (Newman, 2001; Newman, 2003), and the evolution of collaboration networks (Barabási et al., 2002).

The fact that these studies used primarily coauthorship data in research papers poses at least two limitations for studying team science. Firstly, research papers are the end products of a research lifecycle and the coauthors of a paper only presents the fact that they have collaborated during the research. The posterior nature of publication authorships cannot offer insights into the collaboration picture during the research lifecycle. Secondly, in many science fields, especially molecular sequencing, collaboration behavior does not take place only in publications. Our study of GenBank collaboration has found that authors in associated publications for a sequence dataset do not necessarily appear in the data submitter list, or vice versa. As biological sciences are becoming increasingly data-driven, excluding from the analysis the missing authors who contributed as data researchers would create an incomplete picture of science collaboration in today's data-intensive research landscape. Finally, collaboration capacity is a broader concept than publication coauthorship; it encompasses all of the collaboration activities that occurred over the whole research lifecycle, from data production to paper publication and patent generation. The ability to garner collaborators early on in the data production stage may be a factor in a team's ability to maintain a high level of productivity and deliver impactful research.

We think of collaboration in research as the "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).

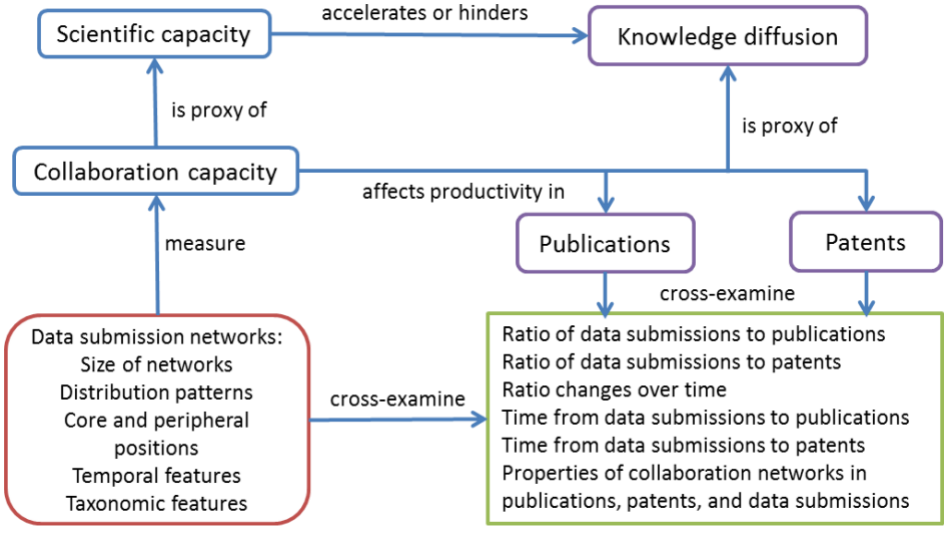

Figure 1. An illustration of the concepts and relationships in the theoretic framework for the impact assessment of collaboration capacity

Within the context of S&T Human Capital, we use the term *Collaboration Capacity* as a framework to operationalize the measurement of S&T Human Capital. Figure 1 illustrates the framework of

collaboration capacity and its relationships with specific metrics and with the larger context of science research enterprise. We define collaboration capacity as the ability of an individual researcher or a team of researchers to collaborate throughout the data production and publication lifecycle and sustain a network of collaborators over time. As biological sciences become data-intensive, data creation is being recognized with increasing credit weight in the academic research evaluation system. Growing the level of collaboration capacity is not only a necessity in data-intensive science but now also has incentives because of the increasing significance of data creation. Given this, collaboration capacity can be considered as a proxy for scientific capacity, which is considered as the extent and quality of postgraduate research education, output in basic science, and the caliber of research universities (Thomson, 2012). Analogously, we use publications and patents as proxies for knowledge diffusion. In the GenBank repository case, the size, centrality, degree distribution, and clustering coefficient of data submission networks and the ratio and overlap ratio between data production teams and publication teams can provide empirical evidence that simple publication-based coauthorship data alone is unable to offer.

This framework suggests that scientific capacity levels can accelerate or hinder the extent of knowledge diffusion and this assumption can be measured by using collaboration capacity as the proxy of scientific capacity. The metrics that we derived from GenBank metadata include those for data submission networks and cross-examined through publication and patent data. They are only sample measures that may be used measure collaboration capacity and its impact and by no means complete.

3 Methods

The framework of collaboration capacity uses network science 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 the ability and opportunity to 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 not only offers new empirical sources, but also can 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), was not needed for the purposes of this study. As such, we dropped the sequence data.

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). We dropped the data in 2013 because they were only up to June, not the whole calendar year.

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 their data are, 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).

Based on the collaboration capacity framework, we generated datasets for the following measures:

- Size of collaboration networks for data submission: number of authors in data submissions and number of data submissions. Edge lists were created based on coauthorship in data submissions.
- Extent of knowledge diffusion: number of patent applications by year and sector, ratio of data submissions to publications at individual author level, and ratio of data submissions to patents at individual author level. The data is limited to U.S. patents that are associated with GenBank data only.
- Rate of knowledge diffusion: time from data submission to patent application

As of the time at this writing, we have completed preliminary analysis. Although we developed a number of hypotheses to be tested with the datasets generated, we were unable to complete the hypothesis testing for this paper. Hypotheses and the testing of them will be reported in future publications.

4 Findings

The analysis results are organized based on two themes: the connectedness of collaboration networks and the ratio of data submissions to publications. Both themes are measured with metrics in the framework with an emphasis on changes over time, aiming to lay the empirical ground for testing the first hypothesis.

4.1 Connectedness of collaboration networks

Between 1994 and 2012[1], the number of data submissions maintained a steady increase faster than that of publications (Figure 2). 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 (the network of publication and submission networks together with all isolated nodes and non-connected clusters removed) increased from 48.1% of all scientists in 1994 to 80.8% in 2012 (Figure 3). That is, the largest group of connected nodes in the network grew by more than 30%. 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 (Figure 4), 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. This network process

1 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.

has been called network densification (Leskovec, Kleinberg, & Faloutsos, 2007).

The mean degree for the publication network 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 (Figure 5). However, the clustering coefficient, which measures how clustered the nodes are in the network, had a sharp drop after 2007. This implies that actors tended to collaborate with more people over time. 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 teams or 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, or more loosely connected networks formed and connected to the hubs through a few edges, rather than historically highly connected and clustered around hubs (Figure 3).

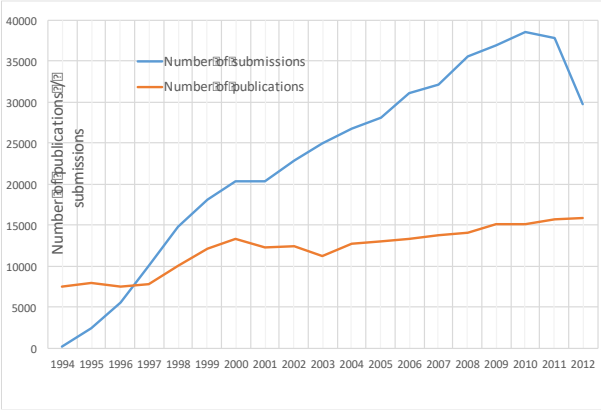

Figure 2. Frequency distribution of data submissions and publications in GenBank 1994-2012

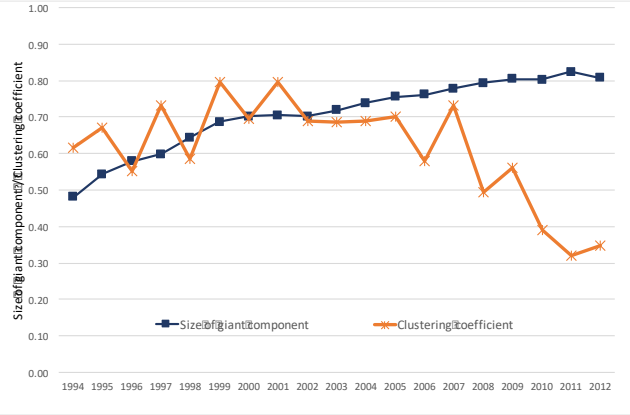

Figure 3. The size of giant component and clustering coefficient in GenBank 1994-2012

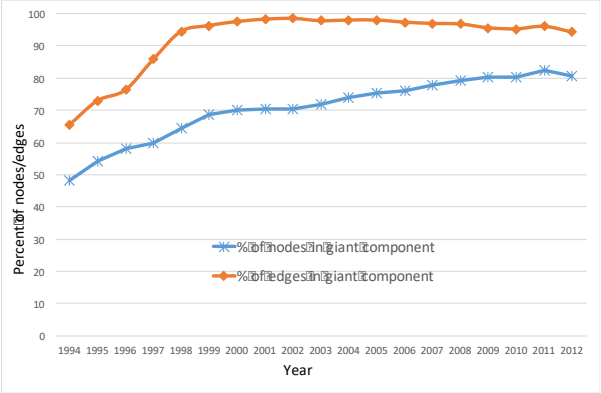

Figure 4. Percentage distribution of nodes / edges in giant component: 1994-2012

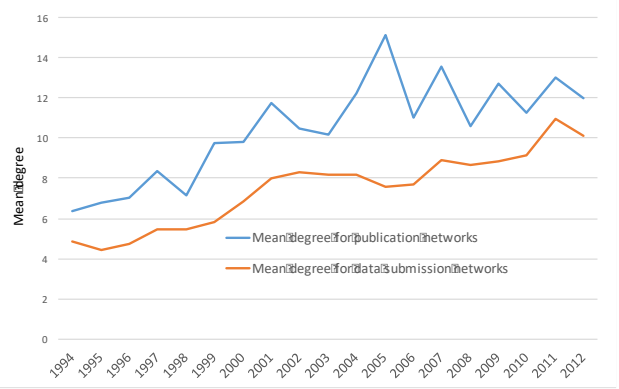

Figure 5. Mean degree distribution by year: 1994-2012

The L-shaped degree distribution for both publication and submission networks (Figure 7) confirms this pattern, which can be interpreted as a few nodes being highly connected to many nodes, while nodes tended to be connected to a relatively small number of nodes. Those nodes in the long tail were much less

connected with others. This property illustrates a preferential attachment process (Barabási & Albert, 1999), or the "Matthew effect", in the GenBank collaboration networks.

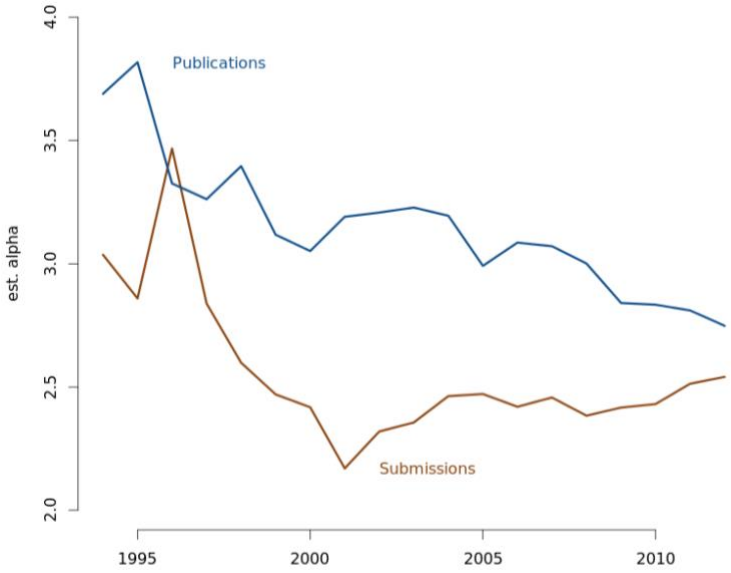

Figure 6. Change of Alpha value in power law distribution over time

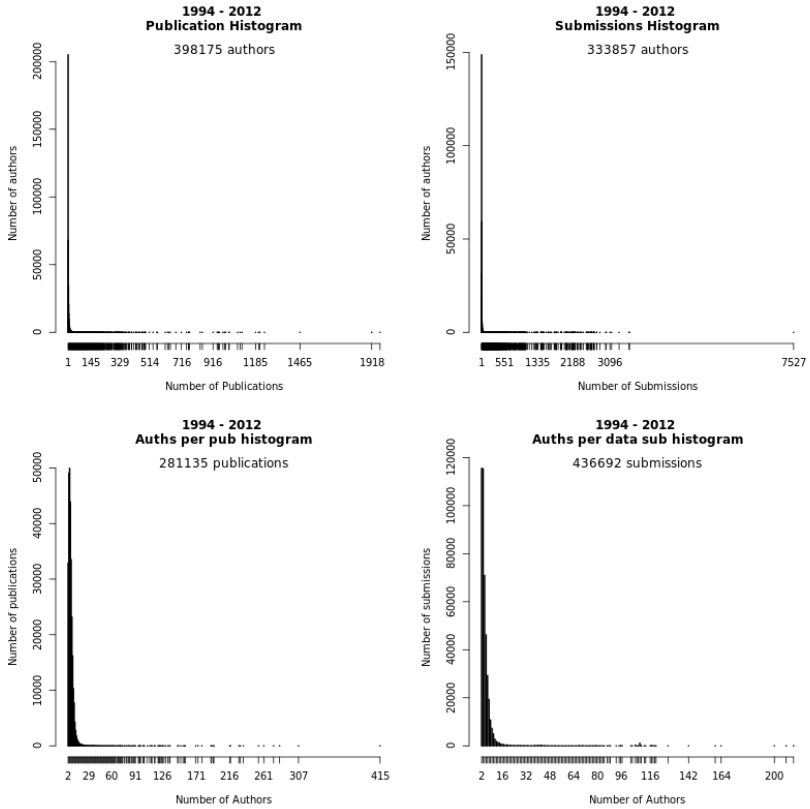

Figure 7. Histograms of authors vs. publications and submissions for all years

4.2 Ratio of data submissions to publications

Sequencing technologies have gone through three generations since Sanger developed the chain-termination method in 1975 (Sanger & Coulson, 1975) and "led the way in revolutionizing the field of genomics and beyond, motivating an astonishing number of scientific advances" (Schadt et al., 2010, p. R228). The most recent statistics released by NCBI shows that, although there has been a steady increase in both the numbers of bases and sequences since GenBank's inception in 1982, the numbers accelerated exponentially after 1995 (NCBI-b, 2017). The lowered cost and increased efficiency in sequencing allows "the reading of DNA hundreds of basepairs in length, massively parallelized to produce gigabases of data in one run. Researchers moved from the lab to the computer, from pouring gels to running code" (Heather & Chain, 2016, p. 6). This shift suggests that data production in genomics research has taken an increasingly significant role, which makes it compelling to examine the ratio of data submissions to publications as well as its impact.

As mentioned earlier, the size of giant component in both data submission and publication networks steadily increased over the 19 years covered by this dataset. This increase trend, however, had evolved with different patterns over time. In early years (up to 1998), there were more authors in publications than in data submission. Whether this phenomenon was due to the amount of data production or the traditional view that data production was not considered as privileged as publications will be the topic of another study. This trend, however, took a turn to a different direction. From 1999, the distribution shape of data production vs. publication began tilting toward data submission. Not only were more authors involved in data submission, but also more authors became more productive in data submission. Before 1998, a majority of authors concentrated in the range of 20 publications and 50 data submissions. After 1999, while the publication range remained the same for most authors, the data submission range has climbed to 100. It is noticeable that since 2008, a sizable number of authors maintained a high productivity in both publications and data submission, approximately in the range between 50~100 publications and 100~300 data submissions.

Another observation we had was that a small number of authors had extremely large numbers of publications but lower number of data submissions. Similarly, a small group of authors who contributed large numbers of data submissions had only very few publications. While it is obvious that not all authors in data submission networks were in the publication networks, or vice versa, the average ratio of submissions to publications showed an upward trend, with a sharp increase from less than one (1) to around 4.01 by 2005 (Figure 9). This was perhaps a turning point for microbiology to become a data-intensive science.

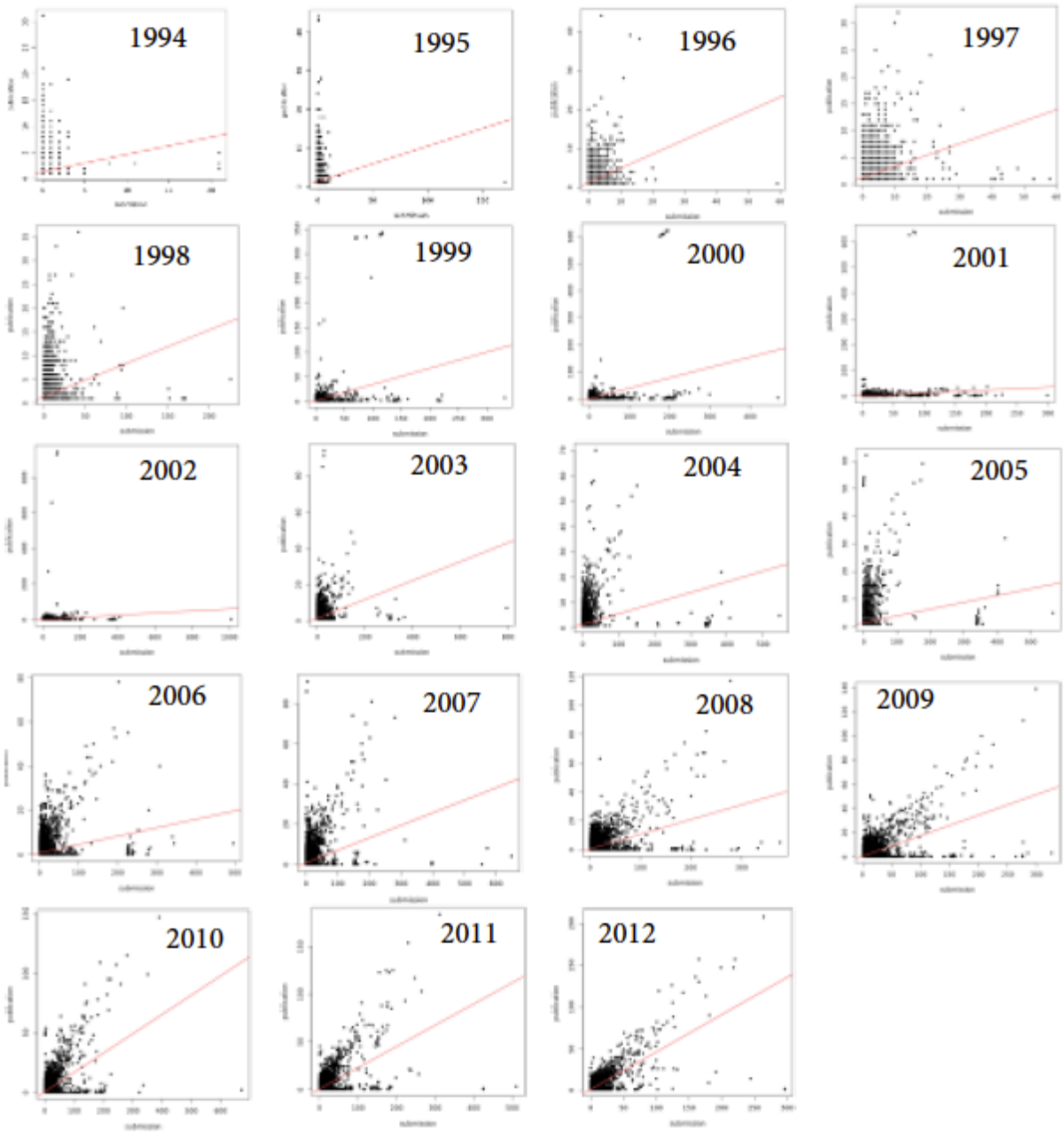

Figure 8. Distribution of number of authors with trend line for 1994 to 2012. The x axis represents the number of authors who submitted sequence data and y axis represents those who had published a paper associated with the submissions.

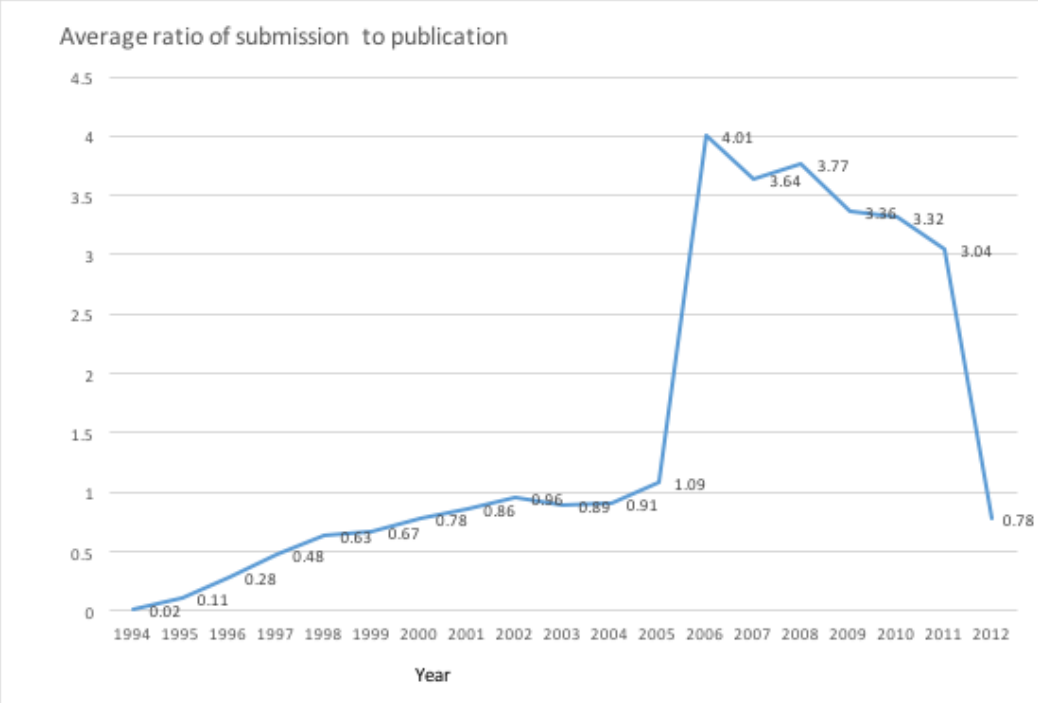

Figure 9. Average ratio of data submissions to publications in GenBank: 1994-2012

5 Discussion and Future Study

In this paper we proposed a framework of collaboration capacity and demonstrated how the metrics under this framework may be used to assess the changes of collaboration networks in the data production stage – the pre-publication phase of the research life cycle – and the impact of such changes on research productivity and knowledge diffusion. The framework operationalizes the S&T human capital theory and makes it possible to tie the collaboration network analysis with theory so that the interpretation of collaboration networks analysis can be built on a solid, meaningful theoretical grounding. This is the major contribution of this paper. The collaboration capacity framework is still a work in progress and needs to explore further. The findings from our GenBank metadata mining project so far provide a number of directions for further exploration.

The first direction lies in deeper mining of the collaboration networks in this very large research 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 new approach to examining and evaluating 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 GenBank metadata proves to be a rich data source for studying team science. Although our analysis shows some interesting distribution patterns in the ratio of data submission vs. publication, more questions are raised from the findings: What are the common team formation by sector, institution, and research field? What are the team sizes that sustained a high productivity? How did the team sizes and relationships correlate with productivity and knowledge diffusion? What other measures can be applied to study collaboration capacity? 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 address the questions raised above, we need to collect events in economic, technological, policy, and other domains that correspond to the GenBank data submission history to tell the whole story of the rise and change of data submission and publication networks in this community, which will add new knowledge to the repertoire of team science.

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 perhaps be seen as evidence for the point in time when microbiology transformed 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: The authors thank the support of NSF SciSIP grant #1561348 and Chensen Wang and Suchitra Deekshitula for their technical assistance.

References

Barabási, A.L., Jeong, H., Néda, Z., Ravasz, E., Schubert, A., & Vicsek, T. (2002). Evolution of the social network of scientific collaboration. *Physica A*, 311: 590-614.

Barabási, A.L. & Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286(5439): 509-512. DOI: 10.1126/science.286.5439.509

Bozeman, B., Dietz, J., & Gaughan, M. (2001). Scientific and technical human capital: an alternative model for research evaluation. *International Journal of Technology Management*, 22: 636–655.

Bozeman, B. & Boardman, C. (2014). *Research collaboration and team science: A state-of-the-art review and agenda*. Heidelberg: Springer.

Costa, M., Qin, J., & Bratt, S. (2016). Emergence of collaboration networks around large-scale data repositories: A study of the genomics community using GenBank. *Scientometrics*, 108(1): 21-40.

Gray, J., Liu, D.T., Nieto-Santisteban, M.A., Szalay, A.S., Heber, G., & DeWitt, D. (2005). 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.

Gray, J. (2007). 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.

Guimerà, R., Uzzi, B., Spiro, J., & Nunes Amaral, L.A. (2005). Team assembly mechanisms determine collaboration network structure and team performance. *Science*, 308(5722): 697-702.

Hara, N., Solomon, P., Seung-Lye, K., Sonnenwald, D. H. (2003). 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.

Heather, J. M. & Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. *Genomics*, 107(1): 1-8. http://dx.doi.org/10.1016/j.ygeno.2015.11.003

KNB. (2017). The Knowledge Network for Biocomplexity, https://knb.ecoinformatics.org/.

Leskovec, J., Kleinberg, J., & Faloutsos, C. (2007). Graph evolution: Densification and shrinking diameters. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1): 2. Doi:10.1145/1217299.1217301

NCBI-a. (2017). GenBank overview, http://www.ncbi.nlm.nih.gov/genbank/.

NCBI-b. (2017). Growth of GenBank and WGS, http://www.ncbi.nlm.nih.gov/genbank/statistics.

NLM. (2015). 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.

Newman, M.E.J. (2003). The structure and function of complex networks. *SIAM Review*, 45: 167-256.

Newman, M.E.J. 2001). The structure of scientific collaboration networks. *Proceedings of National Academy of Science*, 98(2): 404-409.

Porter, A.L. & Rafols, I. (2009). Is science becoming more interdisciplinary? Measuring and mapping six research fields over time. *Scientometrics*, 81(3): 719-745.

Qin, J., Lancaster, F.W., & Allen, B. (1997). Levels and types of collaboration in interdisciplinary research. *Journal of the American Society for Information Science*, 48(10): 893-916.

Sanger, F. & Coulson, A.R. (1975) A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of Molecular Biology*, 94: 441–448.

Schadt, E. E., Turner, S., & Kasarskis, A. (2010). A window into third-generation sequencing. Human Molecular Genetics, 19(Review Issue 2): R227–R240. doi:10.1093/hmg/ddq416

Szalay, A.S. & Blakeley, J. A. (2009). Grey's laws: Database-centric computing in science. In: T. Hey & S. Tansley (eds.) *The Fourth Paradigm: Data-Intensive Scientific Discovery*, pp. 5-11. Redmond, WA: Microsoft Research.

Thomson, R. (2012). National scientific capacity and R&D offshoring. *Research Policy*, 42: 517-528. http://dx.doi.org/10.1016/j.respol.2012.07.003

Tukey, J. W. (1977). Exploratory data analysis. Reading, Mass.: Pearson.

Wasserman, S., & Faust, K. (1994). *Social Network Analysis: Methods and Applications*. Cambridge, UK: Cambridge University Press.