Collaboration Networks and Career Trajectories: What Do Metadata from Data Repositories Tell Us?

Hemsley, Jeff Syracuse University, USA | jjhemsle@syr.edu
Qin, Jian Syracuse University, USA | jqin@syr.edu
Bratt, Sarah Syracuse University, USA | sebratt@syr.edu
Smith, Alexander Syracuse University, USA | aosmith@syr.edu

ABSTRACT
Science is increasingly carried out through scientific collaborations, allowing researchers pool their experience, knowledge, and skills. In this work we identify factors related to a scientist's collaboration capacity, their ability accumulate new collaborations over their career. To do this offer a new collaboration capacity framework and begin the work of validating it empirically by testing a number of hypotheses. We use data from GenBank, a cyberinfrastructure (CI)-enabled data repository that stores and manages scientific data. The data allow us to construct longitudinal networks, thereby giving us yearly scientific collaboration maps. We find that a scientist's network position at an early stage is related to their capacity to build new collaborations and that researchers who manage an upward trend in productivity tend to have higher collaboration capacity. Our work makes a contribution to science of science studies by offering a collaboration capacity framework and providing partial empirical support for it.

KEYWORDS
Metadata analytics, Collaboration networks, Collaboration capacity, Research performance assessment, Assessment metrics

INTRODUCTION
Much of science today, small- or large-scale, is increasingly carried out through scientific collaborations. In this work we think of collaboration as the "social processes in which researchers pool their experience, knowledge, and social skills with the objective of producing new knowledge" (Bozeman & Boardman, 2014, p. 2). Researchers have noted that successful *team-science*, particularly in the biomedical field, is supported by three things: scientific policy, cyberinfrastructures, and Scientific and Technical (S&T) Human Capital (Qin et al., 2018). While cyberinfrastructures are the computer systems that support the long-term storage, curation, discovery, sharing, and reuse of scientific data (Costa et al., 2016), S&T human capital focuses on the career trajectories of scientists and their ongoing ability to enhance their capabilities and make contributions (Bozeman et al., 2001). Collaboration with other scientists is one way scientists gain S&T human capital (Bozeman & Corley, 2004). Thus, collaborations beget S&T human capital which supports team science, a form of collaboration.

In this work we seek to identify factors related to the capacity of scientists to accumulate new collaborations with other scientists in the biomedical field over the span of a career. That is, we look at scientist's *collaboration capacity* (CC). Our assumption is that a high (low) collaboration capacity, whether it rises (falls) over time or remains consistently, reflects a researcher's career trajectory and can be measured by how many new collaborators a researcher accumulates in their collaboration networks. To do this work, we collected metadata for molecular sequences from GenBank, a cyberinfrastructure (CI)-enabled data repository that stores and manages scientific data for later discovery and reuse (NCBI, 2019). Our data spans from 1990 to 2018. The metadata allows us to construct longitudinal networks of data submission co-authors and co-author networks of the related publications. The data submission represents a stage in a research lifecycle earlier than publication, and often includes PhD students and post-docs working in a lab setting. Examining the metadata about sequence data submissions and subsequent publications provides may give us insight into how collaboration networks evolve.

Our work makes a contribution to science of science studies by providing partial support for the collaboration capacity framework initially proposed by Qin, Hemsley and Bratt (2021). We do this by showing that a scientist's position in both the co-author data submission and the co-author publication networks, as assessed by various centrality measures, are related to scientists' collaboration capacity over their career. We also find that scientists who are consistent high performers, as measured by yearly publication count, or those who start out as low performers but become high performers, tend to have higher collaboration capacity, while low performers tend to have lower collaboration capacity. Finally, our work makes a contribution by offering a new way to measure collaboration capacity over a long timeframe. The implication of this study lies in that CI, science policy, and S&T human capital as the enablers of CC play a significant role in the increment (or decrement) of a scientist's collaboration capacity and the study of the rises and falls in scientists' collaboration capacity can provide evidence for evaluating the effectiveness of CI, science policy, and S&T human capital.

85th Annual Meeting of the Association for Information Science & Technology | Oct. 29 – Nov. 1, 2022 | Pittsburgh, PA.

This paper will first review collaboration research to provide a theoretical background for the collaboration capacity framework, on which the rationale and mechanisms of collaboration capacity framework will be elaborated and discussed. The metrics identified from the collaboration capacity framework will be tested with the GenBank metadata we collected. The metrics used for this test can give us a clearer understanding about how the enablers of collaboration played a role in how productive scientists are.

THE THEORETICAL BACKGROUND AND PRIOR RESEARCH

Collaboration and scientific capacity

Collaboration networks are a factor in *scientific capacity*, which is the aggregation of the knowledge, skills, abilities, and technical facilities of individual scientists and their networks of collaborative relationships (Bozeman et al., 2001; Dietz & Bozeman, 2005). We refer to scientific capacity as *Scientific and Technical (S&T) Human Capital* when the focus is on the career trajectories of scientists, their ability to enhance their skills and make contributions in a sustained way. The idea is based on the assumption that scientists' technical and human capacity are honed through conducting research, and that the modern system of science requires collaboration among groups of people to conduct research (Hara et al., 2003). S&T human capital is enhanced by government investment, such as funding for research. It is constantly changing and generates scientific capacity and capabilities at the individual scientist, project, discipline field, or network levels. The factors affecting the generation of scientific capacity vary at each of these levels (Bozeman et al., 2001). At the project level, factors that contribute to the growth or decline of the generation of scientific capacity include project members, new levels and types of physical and economic resources. At the individual level, scientific capacity is generated through new skills and ties to new collaborators.

Research on scientific collaboration networks extends work on S&T human capital (including social capital, the sum of actual or virtual resources with a network of institutionalized relationships). Past research has investigated why scientists collaborate (Melin, 2000), and their strategies for selecting collaboration partners, as well as how they initiate and foster collaborative opportunities (Bozeman & Corley, 2004). The traces of these collaborations are historical indicators of who has access to whom with respect to opportunities to collaborate, and can be a measure of social capital (Costa, 2014). In the aggregate, these relationships give some insight into the capacity of the community to bring scientists together to solve problems. Naturally, the structures and dynamics of these networks are intertwined with the use of S&T human capital. Growth (or decline) in the generation of scientific capacity critically relies on the use of S&T human capital. Analyzing the structures, dynamics and evolutionary history of collaboration networks allow us to identify the drivers and effective use (or otherwise) of S&T human capital. It also helps us gain a new understanding of the factors that contributed to the use of S&T human capital, and further, the generation of scientific capacity.

Collaboration and research impact

The impact of research is determined by three factors: the disciplinary extent to which research outputs have been diffused, the rate of adoption of research outputs, and the societal benefits as results of diffusion and adoption (Qin, 2010). These three separate, yet related components, form a research impact cycle in which research outputs are impactful through knowledge diffusion (Figure 1). According to Qin (2010), the units and implications of this statement can be formally expressed as:

$$I = E \times A \times B \qquad \text{[Equation 1]}$$

where the overall impact (I) is defined as the product of the extent (E) of knowledge diffusion, the rate of adoption (A) as represented by the proportion of intellectual property that has been licensed or patented among all produced, and the benefits (B) to society in both quantitative and qualitative terms (Qin, 2010).

Applying Equation 1 to assess the impact of GenBank collaboration networks, where we have both sequence data submissions and publication networks, we could operationalize the extent (E) as the ratio of data submissions to publications. In Qin et al's (2021) framework, authors supported by a large data submission network would have a higher degree of collaboration capacity. A higher degree of collaboration capacity can lead to higher productivity and better science. Adoption (A) refers to the transfer of knowledge into new applications and/or products, for example, the proportion of new drugs born out of genetic sequence patents and number of genetic therapeutic methods resulted from patents. Although Adoption (A) is beyond the scope of this paper, measures for this type of impact have both theoretical and metric values. The focus in this paper is the collaboration capacity and their correlation with productivity and knowledge diffusion.

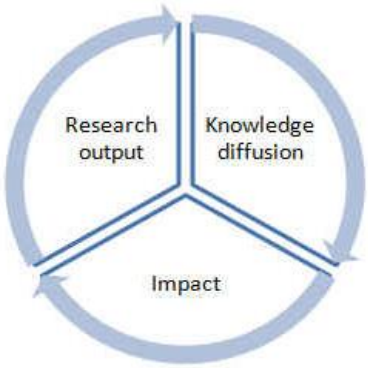

Figure 1. The research impact cycle. From Qin 2010

The impact assessment of collaboration capacity requires resolving some fundamental issues in data analysis. The first one is the role of collaboration capacity at the data production stage and how they are related to knowledge diffusion. Collaboration networks are dynamic and usually characterized by the behavior of four basic categories of network nodes: transients, continuants, newcomers, and terminators (Braun et al., 2001; De Solla Price & Gursey, 1976). Transience and its counter-concept, continuance, refer to the temporal stability of actors in a network. When we look at the publication history of a scientist in a network, a continuant is someone who has contributed to the network in the past, contributed to the network in any given year, and continues to contribute in future years. Transients are those who only contribute for a very short time (usually once). Two additional categories—newcomers and terminators—were also identified. Taken together, these concepts help describe recruitment, retention, and attrition of scientists with respect to an area. While collaboration networks are found to lead to greater knowledge flows (Singh, 2005), it is unclear what role the collaboration capacity at the data production stage played and to what extent it contributed to the rate and scope of knowledge diffusion.

The large number of longitudinal studies of collaboration networks provides insights into the evolving of interpersonal relationships in communities and sub-communities, and how this evolution is related to collaboration capacity and knowledge diffusion. However, few past longitudinal studies of collaboration networks have studied research questions from the collaboration capacity and knowledge diffusion perspectives. Discussed among the publications covering more than a decade of collaboration networks are topics such as the decay of collaboration networks and the structure of collaboration networks, as represented through all historical links between collaborators, regardless of whether or not those relationships are maintained, or if the participating parties are still active in the network (Nahapiet & Ghoshal, 1998; Velden et al., 2010). This often results in misleading representations of the current structure of the network, which not only can cause validity issues (Howison et al., 2011), but also may mislead program managers and policy makers as to the current state of the connectivity of, and interactions in, the network.

It makes sense that the evolution of collaboration networks from data production to knowledge diffusion is affected by the use of S&T human capital at the individual, project, discipline, and network levels. We speculate that the use of S&T human capital (which implies the ability to accumulate new collaborators), effectively or otherwise, is correlated with research productivity, funding, and the role and position of nodes in networks, which we define as *collaboration capacity* and has further impact on the diffusion of knowledge. Patents have been used as an indicator to study innovation and technology advances (Nagaoka et al., 2010), as well as the impact of academic-industry partnerships on productivity, and career paths. Linking genetic data submissions with publications in the context of collaboration capacity is a new approach to discovering whether there is any correlation between collaboration capacity and productivity as well as knowledge diffusion in the CI-enabled research environment.

A COLLABORATION CAPACITY FRAMEWORK

Research on collaboration investigates individual scientists and their interactions with peers at various levels (individual, team, institutional, national, international, community, cross-community, cross-discipline). This body of work also examines the impact of these interactions on research performance and science policy. The nature and properties of scientific collaboration have been studied from both quantitative and qualitative approaches. The quantitative stream of research commonly uses co-authorship and citation data based on publications to look for collaboration network properties and patterns, e.g., scale-free network (Barabási & Albert, 1999) that was developed based on the preferential attachment model by Price (1976), team assembly mechanisms' effect on network structure and team performance (Guimera et al., 2005), the structure of collaboration networks (Newman, 2001a, 2001b, 2003), and the evolution of collaboration networks (Barabâsi et al., 2002). However, these type of studies have

typically used authorship data from publications, which are limited in addressing questions about how collaboration dynamics changed prior to publication, what drove the changes, and what impact those changes may have generated.

The qualitative research on scientific collaboration may compensate for the limitations of quantitative study by using a different lens to examine collaboration. Social scientists consider collaboration as "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 can 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 S&T Human Capital, a concept mentioned above and defined as the sum of scientific, technical and social knowledge, skills and resources embodied in a particular individual (Bozeman et al., 2001). Cultural aspects have recently been added to an updated version of this model (Corley et al., 2019). While this concept provides a theory for explaining the social and cultural aspects of scientific collaboration, specific metrics are not yet operationalized to empirically test and validate it.

The limitations of publication author data and the need for operationalization of the S&T human capital theory motivated the conceptualization of a collaboration capacity framework. We define "*Collaboration Capacity*" as the ability of an individual, a group, or institution to assemble and effectively use S&T human capital in collaborative research. We assume that the greater the S&T human capital a researcher can accumulate or assemble, the more opportunity and resources he or she can accumulate, which may lead to more opportunities to collaborate with other researchers and the more likely the S&T human capital is used more effectively. This means that collaboration capacity measures not only how much S&T human capital one may accrue but more importantly, how effectively he or she can utilize the S&T human capital that may lead to an increase or decrease in productivity, innovations, and new discoveries. To accomplish this kind of research, the data sources used in such a study must include traces from pre-publication process, such as data submissions. It could also include post-publication data, such as might be represented by patents, to allow for a more comprehensive investigation into collaboration dynamics prior and post paper publication.

Collaboration has been found to be positively related to productivity (Lee & Bozeman, 2005; Qin, 1995) and to produce better science as reflected in increased number of citations over solo authored papers (Wuchty et al., 2007). As Stephan (2012) points out, collaboration combines inputs, such as effort, knowledge, equipment, materials, and space, to produce research, though different fields use the inputs in different proportions. While there is ample evidence that collaboration produces better science and increases the possibility of having breakthroughs, it is unclear how collaboration in the early stage of data creation supports knowledge creation and diffusion. It is also unclear whether the ability to accumulate larger inputs (as in Stephan's definition (2012)) increases collaboration capacity, which in turn accelerates the rate of knowledge diffusion.

Figure 2 illustrates a collaboration capacity framework (Qin et al., 2021) in which S&T human capital, cyberinfrastructure, and science policy are considered as enablers for collaboration capacity. Cyberinfrastructure includes data and publication repositories, software tools, and discovery services supporting data-intensive research. Science policy ensures resource allocation and dissemination of research outputs. While collaboration capacity is impacted by three enablers, the evaluation of it, or to be more precise, the evaluation of its *impact*, cannot be done by one single measure, but rather, requires a set of metrics that can operationalize the key aspects to indirectly reflect the impact of enablers. We group these metrics into two categories: data production and data-to-knowledge measures as shown in Figure 2.

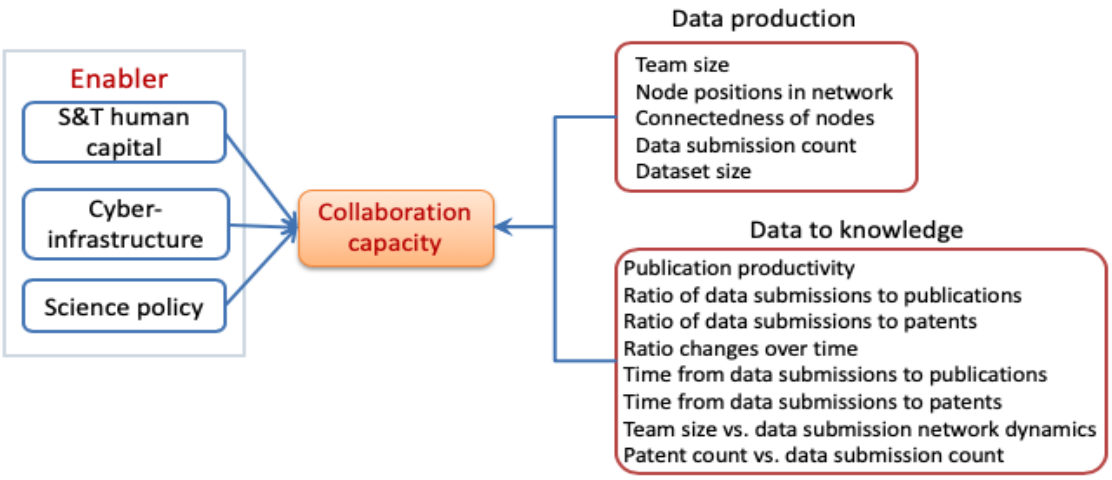

Figure 2. The collaboration capacity framework with enabling components and operationalized measures

While we can't verify the entire framework in this paper, we can verify some connections. Specifically, we noted above that collaboration has been found to be positively related to productivity (Lee & Bozeman, 2005; Qin, 1995), and so we would expect that collaboration capacity is positively related to productivity in our data. We also noted that S&T human capital focuses on the career trajectories of scientists and their ongoing ability to enhance their capabilities and make contributions (Bozeman et al., 2001). Thus, we also expect that researchers that become more (or less) productive over their careers, will gain (lose) collaboration capacity with the change in their productivity. As such, we offer the following hypotheses that we will test:

H1) Authors who started off being low performers but became high performers will tend to exhibit higher collaboration capacity, as measured by how many new authors they tend to work with.

H2) Authors who are consistently high performers will tend to exhibit higher collaboration capacity

H3) Authors who are consistently low performers will tend to have lower collaboration capacity and thus will tend to collaborate with fewer new co-authors

H4) Authors who start high, but become low performers over time will tend to collaborate with fewer new co-authors

H5) Those who work with more new co-authors will tend to publish more, holding all else equal.

The collaboration capacity framework above, along with our discussion of networks, also suggests that an actor's position in both the publication and data submission networks also probably plays a role in collaboration capacity. As such, we will explore the following two research questions:

Q1) Is an author's position in the publication co-author network, as measured by typical centrality measures, related to them having higher collaboration capacity, as measured by average number of new co-authors?

Q2) Is an author's position in the data submission co-author network, as measured by typical centrality measures, related to them having higher collaboration capacity, as measured by average number of new co-authors?

METHODS

The GenBank metadata contains descriptions about the molecular sequences in annotation records, which include authors for publications and data submissions, and in some cases, for patents if the sequences have been filed for patent applications. We chose GenBank metadata as the primary data for two main reasons: first, molecular sequences play a critical role in modern biomedical research, and GenBank from its inception till now spans several decades, offering unprecedented time series data to study collaboration networks and capacity; and second, data submission metadata (pre-publication collaboration), publication metadata, and metadata for patent applications (post-publication) supply trace data for the whole research lifecycle in molecular sequencing, which makes a perfect case for testing collaboration capacity and the impact of its enablers.

Data

Data for this work was downloaded from the GenBank FTP server as compressed semi-structured text files. We downloaded all annotation records from 1982 to 2018. These records include the metadata of sequence data that

researchers uploaded to GenBank. The metadata includes the authorship, associated references, release date of a sequence and other data related to the data submission. We parsed and loaded the metadata into a relational database, which gave us 227,905,057 annotation records, and then ran author disambiguation (Chin et al., 2013) on the names. We cross-checked the results using Microsoft Academic graph, SCOPUS and Web of Science, accomplishing 89% accuracy. We are left with 877,134 unique authors after disambiguation. We note that some records include journal publications, and some contain other information, such as sequence data submission information.

For each year between 1992 and 2018 we construct a co-author network of the journal publication data and data submission data. However, between 1982 and 1991 the data is sparse. It isn't until 1992 that significant numbers of entries show up in the database. Thus, we have two networks, pub and sub, for each year of data.

Since we are intent on understanding how certain productivity measures and network attributes are related to collaboration capacity, we need to look at a set of actors who show up consistently over time. We don't require them to necessarily publish (or submit data) every year, but to be present in the data repeatedly over time. To satisfy this, we break our yearly networks into seven windows, spanning three years each as represented in figure 3. To be included an author must have published at least once in each window over the seven windows. Note that seven windows at three years each gives us 21 years, which is a reasonable approximation of a researcher's career and gives us enough time to understand collaboration capacity. We start with 1997 and end with 2018. In total, we end up with 6,503 authors who are present in the data from start to finish. For each of these authors we derive network statistics, such as centrality measures.

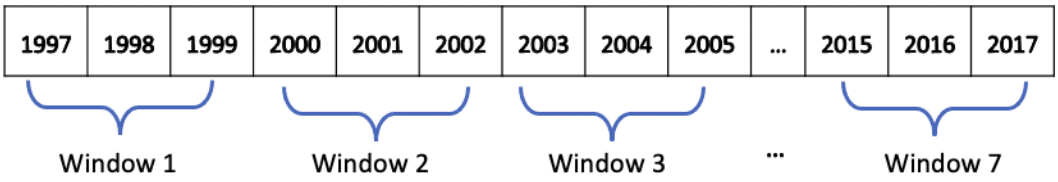

Figure 3. Scheme for 3-year windows

We do recognize that many scholars have careers longer than 21 years. Indeed, if a student researcher goes on to become faculty, and if they continue to do research until they retire, they could have a 40-year career. Thus, our 21-year span could be capturing researchers at the beginning, middle or end of their career. We believe careers this long would be the exception, not the norm, and so with over 6,000 observations we think it is probably that are capturing most of the majority of researchers careers. Still, we do see this as a limitation of the work.

As a means of measuring collaboration capacity, we find the average number of new co-authors for each of the 6,503 authors. This is calculated by finding the number of new co-authors someone has worked with in each of the 3-year windows and averaging that. As an example of this, if in window 4 (starting in year 10 to going to year 12), an author worked with a new co-author, they would count as one new co-author in that window, but if that co-author showed up again in the 5th window, they would then not be considered new. Once we count the number of new co-authors in each window, we find the average for each author across all windows. The assumption being that those with more collaboration capacity will tend to find more new co-authors.

We also make three seven-year windows (each being 1/3 of the total timespan), which we use to categorize authors into one of five groups based on their publication productivity. These groups are, 1) "LowToHigh", which means an author's average number of publications started low, compared to the others, but increased over time; 2) "ConsistentlyLowPerf" is for authors who produced a low number of publications in each of the three windows; 3) "HighToLow" is the category for authors who started out as high performers, but then tended to have lower numbers in later windows; 4) "ConsistentlyHighPerf" are those authors who consistently had higher numbers of publications across all three windows; and finally, 5) "typical" for all other authors. This last group are authors who tended more toward the average number of publications, or otherwise don't fit into one of the other groups.

Regression analysis

To test our hypothesis, we use a standard ordinary least squares regression model. The strength of such a model is in its simplicity and flexibility (Faraway, 2004). We can easily include a number of variables, which the model will hold constant while finding the variance explained by the other variables. Such models are also flexible enough to include constant and dichotomous variables. We use the regression in this study to test which factors, and find out by how much, affect the collaboration capacity of an author. To make the coefficients comparable, we normalize all of the numeric variables (Faraway, 2004). As is normal, we report the model's performance along with the estimates

for each of the coefficients. We note that models with very high numbers of observations, like ours (6,503) can make interpretation of p-values less reliable (Faraway, 2004) and so also report the confidence interval.

Since we are interested in how both network measures and performance measures are related to collaboration capacity, we build a model where our dependent variable is the average number of new co-authors, which is described in the previous section. The data here is skewed and so as is often done (Faraway, 2004), we log-transform the variable to make it comply with the assumption of a normal distribution. For this model, all of our independent variables are drawn from window-2, which covers the years from 2000-2002. However, the coefficients' significance and R-squared where fairly constant regardless of which window was used. The following are our independent variables:

LowToHigh: this is a 1 if we categorized the author as initially a low performer, but they became a high performer by the final 7-year window. The base state by which this variable is compared against it "typical". A total of 1,769 authors fell into this group, or nearly 27%.

HighToLow: authors that started out as high performers, but became low performers over the timeframe, are coded as 1, otherwise the variable is zero. The base state by which this variable is compared against it "typical". Slightly more than 15% (948) of our authors fell into this group.

ConsistentlyLowPerf: The variable is 1 for authors who consistently produced a low number of publications. The base state by which this variable is compared against it "typical". Only 5%, or 309 authors are in this group.

ConsistentlyHighPerf: Authors who consistently produced high numbers of publications are coded as 1, otherwise the variable is zero. As before, the base state by which this variable is compared against it "typical". These actors make up the smallest group with only 261 authors, or 4%.

P_3YrAvg2: This is the number of publications for the author during the second three-year window. All remaining variables were drawn from the same window. This tests the idea that those who find more new co-authors (our measure for collaboration capacity) just tend to publish more.

The following variables were calculated from either the publication co-author network or the submission co-author network.

D_3YrAvg2: This variable captures the author's degree centrality, or number of links, in the publication network.

C_3YrAvg2: This is the authors closeness, which measures a nodes proximity to other nodes in the network, in the publication network.

B_3YrAvg2: This variable is the betweenness centrality of the author within the publication network. Betweenness measures how many paths between other nodes the author is on.

E_3YrAvg2: Eigenvector centrality, which is a measure of influence in a network, is captured in this variable

Sub_D_3YrAvg2: The author's degree centrality in the submission network.

Sub_C_3YrAvg2: The author's closeness centrality in the submission network.

Sub_B_3YrAvg2: The author's betweenness centrality in the submission network.

Sub_E_3YrAvg2: The author's Eigenvector centrality in the submission network.

RESULTS

The model preforms reasonably well and, as we can see from the r-squared values, explains 52% of the variance in the data. We calculated the Variance Inflation Factor (VIF) for the independent variables to ensure that the model didn't suffer from multicollinearity. All values were below 4, a standard cutoff for the test (Faraway, 2004), which indicates the model doesn't suffer from multicollinearity. We also made diagnostic plots to verify the remaining model assumptions. Recall that the dependent variable for the model is the average number of new co-authors a given author has over the three-year windows.

	Estimate	Std.Error	t.val	p.val	ci.2.5	ci.97.5
(Intercept)	1.334***	0.005	293.241	0.000	1.325	1.343
LowToHigh	0.225***	0.008	28.996	0.000	0.21	0.24
ConsistentlyLowPerf	-0.26***	0.016	-16.382	0.000	-0.291	-0.229
HighToLow	-0.151***	0.01	-14.514	0.000	-0.171	-0.131
ConsistentlyHighPerf	0.114***	0.022	5.239	0.000	0.072	0.157
P_3YrAvg2	0.046***	0.006	7.442	0.000	0.034	0.058
D_3YrAvg2	0.14***	0.007	20.558	0.000	0.127	0.153
C_3YrAvg2	0.069***	0.004	15.537	0.000	0.06	0.078
B_3YrAvg2	-0.006	0.006	-1.135	0.256	-0.017	0.005
E_3YrAvg2	-0.041***	0.005	-8.174	0.000	-0.051	-0.031
Sub_D_3YrAvg2	0.02***	0.006	3.472	0.001	0.009	0.031
Sub_C_3YrAvg2	0.049***	0.004	11.479	0.000	0.041	0.057
Sub_B_3YrAvg2	0.015***	0.005	3.206	0.001	0.006	0.025
Sub_E_3YrAvg2	0.015***	0.005	3.216	0.001	0.006	0.024
R^2: 0.518, F-stat: 536.865, df: (14, 6489)						
P-value significance codes: *** <= 0.001, ** <= 0.01, * <= 0.05, . <= 0.10						

Table 1. Regression analysis results

We note that both LowToHigh and ConsistentlyHighPerf are significant and positive, which suggests that authors in these categories tend to have higher collaboration capacity than typical authors. Interestingly, LowToHigh has a higher magnitude estimate than ConsistentlyHighPerf, implying that those who start out disadvantaged in some ways, but are good networkers, tend to end up with higher collaboration capacity than those who tend to be consistent high performers.

ConsistentlyLowPerf and HighToLow are both significant and negative, suggesting that these authors tend to have lower collaboration capacity than typical authors. Since the estimate for ConsistentlyLowPerf is of a larger magnitude than for HighToLow, we might assume that lower performance tends to be related to lower collaboration capacity overall.

The number of publications an author has (P_3YrAvg2 in the model) is positively related to the average number of new co-authors a given author will tend to have. Obviously, with this analysis we cannot determine if more publications attracts more new co-authors or if a byproduct of more co-authors just tends to lead to more publications.

Note that the publication network measures D_3YrAvg2, C_3YrAvg2 and E_3YrAvg2 are all significant, though E_3YrAvg2 has a negative estimate while the others are positive. The largest magnitude estimate in this group is D_3YrAvg2 at 0.14, which suggests that a unit increase in standard deviation of degree, or how many links an author has in the publication author network, causes a 0.14 standard deviation increase in the average number of new co-authors a given author will tend to have. The magnitude for closeness centrality appears low, but the significance suggests that those on short paths to many others in the network will tend to find more new collaborators, holding all else equal. More generally, position in the network matters for collaboration capacity. Interestingly, eigenvector centrality is significant and negative with a small effect size. Since eigenvector centrality measures influence in the network, where those with higher scores tend to be connected to other highly influential nodes, we might presume that some actors at the top get comfortable collaborating with others at the top and tend to find new collaborators less frequently. Note that the effect size is small, and due to the nature of regression, other things are held constant, but we did note above that authors in the ConsistentlyHighPerf category tended to find fewer new authors to collaborate with than those in the LowToHigh category, and these very top, consistently high performers, may be best connected in the network.

Next, we look at the submission network. All of the variables were positive and significant, though the effect sizes tended to be very small. But overall, we can glean that position in the submission network plays a role in finding new collaborators later. Recall that, the data submission represents a stage in a research lifecycle earlier than

publication. Thus, the significance of these variables suggests that those who were well positioned in the data submission network, tended to gain more publication co-authors later in their careers, but given the effect size, other factors probably matter more.

Finally, the model explains 52% of the variance, suggesting that developing new co-author relationships, or a researcher's collaboration capacity, either has a great deal of randomness to the process or that our data lacks some key significant predictors. We believe both are probably true and that finding additional data to link to our current data could be a fruitful new direction.

DISCUSSION

The collaboration capacity framework discussed in this paper offers a new way to examine collaboration networks from theoretical and methodological perspectives. Collaboration networks are a phenomenon where the structures and positions of researchers in these networks are closely tied to the effective use and support of S&T human capital, science policy, and cyberinfrastructure. Our framework views collaboration capacity as an indicator of how effectively S&T human capital is used by a researcher, and how effectively science policy and cyberinfrastructure support science. While authors are a typical measure for S&T human capital, measures for evaluating the effectiveness and impact of science policy and cyberinfrastructure have not been straightforward in the literature to date. An example is research funding. How much funding should be allocated to which research field is generally considered as part of science policy, yet whether such funding may be obtained by a given researcher requires an effective orchestration of all three enablers. That is, they need S&T human capital for creating a compelling grant proposal, the science policy must make funding opportunities, and to ensure success, they need resources and tools (cyberinfrastructure). In this sense, the collaboration capacity framework offers a way to operationalize the metrics for assessing the effectiveness and impact of S&T human capital, science policy, and cyberinfrastructure.

The empirical evidence from this study shows that as new authors are added to an author's collaborator network, we see an increase in their overall capacity to collaborate, but the strength of the relationship differs among different performance groups. When researchers' career trajectory maintains an upward trend, i.e., their publication performance goes from low to high, they appear to be better connected in the network and have stronger collaboration capacity than other groups. This may imply that from a science policy perspective, the highest returns may come from investing in mid-career researchers who are showing an upward trend in productivity. The differing relationships between collaboration capacity and performance levels raise questions for future work based on the enablers of collaboration capacity in Figure 2. For example, what are the structural patterns of the teams for consistently high and/or from-low-to-high performance groups? How does the ability to secure funding associate with the performance level? Is there any correlation between collaboration capacity and impactful breakthroughs? Answering these questions will require data beyond what can be found in the GenBank metadata, and we have already started collecting and linking this data. For example, we have selected sample authors from the four performance groups to collect more data regarding their affiliations, status, and positions for tracking their career trajectories and other research outputs. We have also collected NIH funding data to link with the authors in our dataset. The triangulation with these new data will allow us to gain more insights into the phenomenon we have identified from current study, as well as further validate the collaboration capacity framework.

The GenBank metadata used in this study included both publications and data submissions. The data submissions data adds a new angle for research looking at collaboration networks. In GenBank metadata records, not all publication authors are data submission authors or vice versa. Our analysis shows that authors' positions in data submission network contribute to their eventual collaboration capacity, which is positively related to career trajectory (consistently high, from lor to high, from high to low, and consistently low performance). The inclusion of the data submission metadata in our dataset gave us an opportunity to examine an earlier stage of collaboration in the research lifecycle (pre-publication) that could be used for modeling researchers' career trajectories and the relationships between career trajectories and collaboration capacity enablers listed in the collaboration capacity framework.

In this work our dependent variable was the number of new publication authors and so we note that it might also be interesting to do similar work looking at data submission productivity. In other words, does collaborative capacity also lead to higher productivity in data submissions to GenBank? Since we assume that data submissions are more likely to occur earlier in a career, such a finding might reasonably suggest that some actors just have more collaborative capacity which then might lead to more opportunities to make data submissions. This could be as simple as some people having a more outgoing personality. Our data are too limited for us to untangle such effects, but this could be a future direction.

Due to limited the space of this paper, not all of the variables in Figure 2 were tested or discussed in this work. Still, the empirical findings illustrate the entangled relations between researchers' performance and network positions and has raised new questions for further investigation. Thus, our work makes a contribution to science of science studies

by providing partial support for the collaboration capacity framework. The work also makes a contribution by offering a new way to measure collaboration capacity over the span of researchers' careers.

CONCLUSION

This paper introduced a collaboration capacity framework and showed the results from analyzing the GenBank metadata that include both publications and data submissions. The results from our work support our hypothesis that authors who are better connected in the network have higher collaboration capacity, as represented by the number of new authors added during a period. The empirical evidence raises several questions for further investigation that will need other data sources to address. Metadata from scientific data repositories are massive in volume and complex in structures and relations. The use of such data requires intensive computational processing before they are ready for analysis. This type of metadata as a new data source for quantitative study of science has great potential to be explored. The complexity and methodological challenge in this data source cannot be underestimated.

REFERENCES

Barabási, A.-L., & Albert, R. (1999). Emergence of scaling in random networks. *Science*, *286*(5439), 509–512.

Barabâsi, A.-L., Jeong, H., Néda, Z., Ravasz, E., Schubert, A., & Vicsek, T. (2002). Evolution of the social network of scientific collaborations. *Physica A: Statistical Mechanics and Its Applications*, *311*(3–4), 590–614.

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

Bozeman, B., & Corley, E. (2004). Scientists' collaboration strategies: Implications for scientific and technical human capital. *Research Policy*, *33*(4), 599–616. https://doi.org/10.1016/j.respol.2004.01.008

Bozeman, B., Dietz, J. S., & Gaughan, M. (2001). Scientific and technical human capital: An alternative model for research evaluation. *International Journal of Technology Management*, *22*(7–8), 716–740.

Braun, T., Glänzel, W., & Schubert, A. (2001). Publication and cooperation patterns of the authors of neuroscience journals. *Scientometrics*, *51*(3), 499–510.

Chin, W.-S., Juan, Y.-C., Zhuang, Y., Wu, F., Tung, H.-Y., Yu, T., Wang, J.-P., Chang, C.-X., Yang, C.-P., & Chang, W.-C. (2013). Effective string processing and matching for author disambiguation. *Proceedings of the 2013 KDD Cup 2013 Workshop*, 1–9.

Corley, E. A., Bozeman, B., Zhang, X., & Tsai, C.-C. (2019). The expanded scientific and technical human capital model: The addition of a cultural dimension. *The Journal of Technology Transfer*, *44*(3), 681–699. https://doi.org/10.1007/s10961-017-9611-y

Costa, M. R. (2014). The dynamics of social capital in scientific collaboration networks. *Proceedings of the American Society for Information Science and Technology*, *51*(1), 1–4.

Costa, M. R., 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. https://doi.org/10.1007/s11192-016-1954-x

De Solla Price, D., & Gursey, S. (1976). Studies in scientometrics, part I: transience and continuance in scientific authorship. *International Forum on Information and Documentation*, *1*(2), 17–24.

Dietz, J. S., & Bozeman, B. (2005). Academic careers, patents, and productivity: Industry experience as scientific and technical human capital. *Research Policy*, *34*(3), 349–367.

Faraway, J. J. (2004). *Linear models with R* (Vol. 63). Chapman and Hall/CRC.

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

Hara, N., Solomon, P., Kim, S.-L., & 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.

Howison, J., Wiggins, A., & Crowston, K. (2011). Validity issues in the use of social network analysis with digital trace data. *Journal of the Association for Information Systems*, *12*(12), 2.

Lee, S., & Bozeman, B. (2005). The impact of research collaboration on scientific productivity. *Social Studies of Science*, *35*(5), 673–702.

Melin, G. (2000). Pragmatism and self-organization: Research collaboration on the individual level. *Research Policy*, *29*(1), 31–40.

Nagaoka, S., Motohashi, K., & Goto, A. (2010). Patent statistics as an innovation indicator. In *Handbook of the Economics of Innovation* (Vol. 2, pp. 1083–1127). Elsevier.

Nahapiet, J., & Ghoshal, S. (1998). Social capital, intellectual capital, and the organizational advantage. *Academy of Management Review*, *23*(2), 242–266.

NCBI. (2019). *GenBank Overview*. https://www.ncbi.nlm.nih.gov/genbank/

Newman, M. E. (2001a). Scientific collaboration networks. I. Network construction and fundamental results. *Physical Review E*, *64*(1), 016131.

Newman, M. E. (2001b). Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. *Physical Review E*, *64*(1), 016132.

Newman, M. E. (2003). The structure and function of complex networks. *SIAM Review*, *45*(2), 167–256.

Qin, J. (1995). Collaboration and publication productivity: An experiment with a new variable in Lotka's law. *ISSI'95 (Proceedings of the Fifth Biennial International Conference of the International Society for Scientometrics and Infometrics)*, 445–454.

Qin, J. (2010, February 3). Empirically assessing impact of scholarly research. *Proceedings of the IConference*. https://www.ideals.illinois.edu/bitstream/handle/2142/14924/qin.pdf?sequence=2&origin=publication_detail

Qin, J., Hemsley, J., & Bratt, S. (2018). Collaboration capacity: Measuring the impact of cyberinfrastructure-enabled collaboration networks. *Science of Team Science (SciTS)*. Science of Team Science (SciTS), Galveston, TX.

Qin, J., Hemsley, J., & Bratt, S. (2021). The Role of Cyberinfrastructure-Enabled Collaboration Networks in Supporting Collaboration Capacity. *Available at SSRN*. https://doi.org/10.2139/ssrn.3887529

Singh, J. (2005). Collaborative networks as determinants of knowledge diffusion patterns. *Management Science*, *51*(5), 756–770.

Stephan, P. (2012). *How economics shapes science*. Harvard University Press.

Velden, T., Haque, A., & Lagoze, C. (2010). A new approach to analyzing patterns of collaboration in co-authorship networks: Mesoscopic analysis and interpretation. *Scientometrics*, *85*(1), 219–242.

Wuchty, S., Jones, B. F., & Uzzi, B. (2007). The increasing dominance of teams in production of knowledge. *Science*, *316*(5827), 1036–1039.