Clara Boothby
Qualifying exams, Fall 2020
Question 3

Prompt

What methods have been developed in quantitative studies in the science of science to study division of labor and collaboration in academic science when the primary unit of study is co-authorship on peer-reviewed publications? How are these techniques being adapted, complicated, and improved by new strategies and data sources?

Introduction: Observing Collaboration Behind the Work

Collaboration is becoming more frequent across all fields, and the average team size is growing dramatically (Wuchty et al., 2007). Into this environment, collaboration is not only central to modern research practices, but also a major area of study in the science of science. Collaboration itself is frequently understood to be one of the main drivers behind new scientific ideas and strong research practices (Bozeman & Youtie, 2017). But scientific collaboration is not solely a modern condition. The idea of scientists advancing knowledge by working together in quiet networks of collaboration was present well before the time of the Royal Society, such that knowing the major participants in one's own field through letter-writing was an indicator of one's active membership in scientific research (Kronick, 2001). However, one of the central limitations to studying co-authorship is a small amount of large-scale empirical data about what happens within collaborations. Derek de Solla Price's discussion of these networks, which he termed invisible colleges, he pointed out that despite the critical importance of invisible colleges and collaboration write large within science, the personal connections between scientists are rarely explicitly articulated or recorded, and that any studies of collaboration may need to rely on proxies (De Solla Price & Beaver, 1966; Price & Price, 1986). However, despite these challenges studying collaboration is a crucially important way to study the confluence of new ideas during collaboration, and the types of most productive collaboration between researchers, both within discipline and in interdisciplinary collaborations. Studying collaboration can serve as an entry-point into questions about field development, power, and hierarchy in scientific work.

There have been social and ethnographic works studying the lab conditions and collaboration between researchers working together on the same projects. Latour and Woolgar's *Laboratory Life* and Diana Forsythe's *Studying Those Who Study Us*, Sharon Traweek's *Beamtimes and Lifetimes*, and more recently Janet Vertesi's *Seeing like a Rover* are a few emblematic examples of qualitative studies of collaboration that are based on years of deep fieldwork, case studies, interviews and direct observation. They provide an unparalleled depth of attention to interpersonal dynamics of scientific collaboration, and the work and communication that occurs during the collaboration as researchers develop research findings prior to publication. On the other hand, the quantitative fields of science of science can use publication metadata to study collaborative work in science on a larger scale than from within the often singular working environments of one lab. As metadata is data that describes other data, in the context of the science of science, metadata describe the basic information about the scientific paper, like the title, authors, publishing journal, keywords, author affiliation and so on (Sugimoto & Larivière, 2018). However, quantitative studies that use published papers as the data source can only observe collaboration in retrospect, inferring from the finished artefact of the co-authored paper that a collaboration to produce the paper must have occurred.

In *Science In Action* Latour describes closed vs open objects, where open objects are unresolved issues where the full complexities of the situation are still on display and open for negotiation, while closed objects are cases where the object of study has been formalized and solidified so as to give the impression of completeness and straightforwardness (Latour, 2003). Latour identifies the scientific paper is a closed object, within which there had been a great number of decisions and scientific negotiations that were behind its creation, but that after publication, all those issues present the façade of having been completely solved, and now the complexities are completely hidden, appearing only as traces (Latour, 2003). When Latour was talking about this, he meant the scientific findings in the papers, but in this case, the concept is still apt and important for the working conditions and the contributions of the various collaborators who created the paper. In the case of quantitative science of science, researchers trying to study collaboration, we're given the closed object of the scientific paper, and we need to find ways to reopen that object to make appropriate inferences about the collaborations between the papers contributors. That question is, how do we look back infer things about the complexities of the scientific team that happened before the publication of the paper when all we have are completed the scientific papers and the other completed artefacts.

However, Once we, as science of science researchers, are able to pull some understanding of the collaboration and the working conditions of the researchers from behind the publication metadata, we are able to aggregate these insights into larger scale studies using the whole universe of bibliometric techniques, and look at collaboration on the level of full research careers, institutions, fields, countries, and many other units of analysis. In this way, the unique benefit of quantitative studies of collaboration is that the scale of quantitative studies may encompass multiple collaborations between a larger network of collaborators outside of the individual labs. The scale of quantitative collaboration studies can be as broad as entire fields, entire countries, and even international studies across the globe. The main wrinkle to using publication metadata to study collaboration is that collaboration happens behind-the-scenes of the resulting papers that comprise the major datasets for quantitative science of science. And it is this issue of looking behind the scenes of the research paper and inferring how the collaboration happened that this essay will be centered around.

The primary unit for studying collaboration from bibliometric perspective is co-authorship in peer-reviewed publication, which indicates who made significant contributions to the research and writing of that particular publication. While this might seem limited at first, the quantitative fields of the science of science have developed many well-known methods for studying collaboration, as complex and often invisible as it is, using only paper metadata. Here, I will overview the typical techniques and assumptions of the methodologies for studying collaboration via co-authorship, and I will discuss some recent developments that extend into new terrain for studying co-authorship outside the co-author line only. In fact, there are increasing calls and additional data sources for other information about the contributors and collaborators. These new methods will contribute a greater depth of understanding about the implicit elements of collaboration that had previously not been captured by the byline. However, these bibliometric methods also require a deep understanding on the part of the researchers of the subtle dynamics behind the decisions about who appears on the authorship line. I will also discuss these kinds of contingencies that mean that any bibliometric studies of collaboration require a thorough understanding of the ways that authorship is counted, particularly between different fields.

Section 1: Methods for studying collaboration based in on co-authorship

Within the metadata of scientific papers, the list of all the paper's co-authors records the people who contributed to developing the results and in the writing of the paper itself. This is already a very basic

unit, since knowing who was in the co-author line could be considered a simple binary variable that indicates that any researcher appearing in the co-author line contributed to the paper, while the any researchers who did not appear did not contribute. In addition, the forms of collaboration captured in the study of co-authorship is limited to the types of collaboration that produce a peer-reviewed scientific paper; in other words, collaboration that leaves behind a scientific paper as an artifact. In addition, because this kind of collaboration is tied specifically to scientific papers, looking on the paper level atomizes collaboration into a number of discrete events, particular papers that the collaborators contributed to together. Of course, papers may also be aggregated to study collaboration on the level of the author, of institutions, of countries, or many other levels , by grouping papers based on these criteria (Katz & Martin, 1997). In this way, the granular collection of coauthored papers that add up to a collaborative relationship may make a way of measuring, the duration and intensity of these collaborative relationships. Even with its simplicity, co-authorship provides the first, ever-important glimpse into the workings of the collaboration behind the scientific paper. As I will describe, these methodologies for studying collaboration through the co-author rely heavily on an understanding of the social and structural conventions within scientific collaboration about which researchers appear in the co-author line and why.

Number of coauthors

One of the major characteristics of the scientific working environment is team size, particularly in the era of modern science (Gazni et al., 2012). One of the simplest strategies for studying collaboration between authors is to count the number of authors listed in the co-author line. However, the simplicity of counting co-authors as an indicator of team size belies its importance as a research method because team-size can reveal a lot about the collaboration environment that produced the paper. For papers to have more than one author in fact creates one of the most basic motivations for studies of collaboration in science, and the insight that fewer papers are single authored in the recent past (Greene, 2007; Wuchty et al., 2007) justifies collaboration as an important research area in the science of science. Furthermore, the studies of the sizes of scientific teams are an important research area, as several studies have used the increases in numbers of co-authors per paper to demonstrate that the team size within science has been growing over time (Larivière et al., 2015; Milojević, 2014; Wuchty et al., 2007). Most studies of team size aggregate the papers in particular fields to identify these increases in team size across science, and to make informed generalizations and comparisons about the working structures between scientific fields. Although the intuition that resource-intensive research, or “hard science”, fields have a greater need for team-based science is borne out by larger team sizes in these fields, in a surprising twist, increases in team size in social science have also been observed, which demonstrates the need for aggregated comparative studies (Wuchty et al., 2007).

However, even with the apparently straightforward nature of the authorship line, there are remaining questions about the visibility of less prestigious roles within the scientific paper. For example, while technicians contribute materially to scientific research, particular in resource intensive fields like lab biology and experimental physics where they are responsible for equipment, experimental workflow, and experimentation itself, technicians appear as paper authors much less frequently than other collaborators (Barley & Bechky, 1994; Hagstrom, 1964). In this way, technicians are examples of active collaborators whose presence would be left out of studies of team-size that use the authorship line. The threshold for inclusion in the author line has been constantly under negotiation, and there is also evidence that the internal rules in science regarding which collaborators will be included as paper coauthors change not only between fields, but also over time. As different fields have different

conventions for crediting contributors, team size may be affected if instead of authorship, contributors in certain roles are given credit in a way that doesn't mean authorship, such as inclusion in the acknowledgements section (Paul-Hus et al., 2017).

Different fields also tend to have different cultures with different standards for inclusion as co-author. Where some fields opt to include all the researchers and technicians who had anything to do with the experiment, and everyone who touched the draft of the paper as co-authors, other fields reserve author credit only for those who had a substantial role in the design, experiment, and writing of the paper, shunting everyone else's contributions into the acknowledgement section. These differing conventions raise questions about the actual differences in team sizes. When current research studying collaboration finds different numbers of collaborators on papers in different fields, are the apparently smaller teams in social sciences accurate depictions that social science has smaller teams, or are these smaller teams an artifact of the different co-authorship conventions in social science, which may have a higher threshold for the contribution of authorship than in other fields (Paul-Hus et al., 2017)? However, studying the number of authors included in the co-author line is a useful and important first measure of the most basic fact about the scientific research team behind the paper and their collaboration—how many people worked together to produce that paper. The names on the co-authorship line provide the most basic binary information about the collaborators on the paper: who was a part of the scientific collaboration that produced the paper, and who was not. In the next sections, we will add additional nuance to this, showing how it's possible to glean more detailed information about division of labor.

Credit Allocation based on co-authorship

The correct apportioning of credit for a particular research finding also sheds light on the division of labor within the team that produce the scientific finding. The rise of multi-authored and interdisciplinary papers and the increasing rarity of the single-author paper also raises issues about the credit structures for scientific authorship (Flier, 2019; Greene, 2007). Oftentimes, metrics evaluating the frequency, impact, or strength for scientific work is used not only in hiring and promotion decisions for individual scientists, but for national funding decisions for research departments and institutions (Hicks, 2012). The pursuit of accurate credit is important precisely because the allocation of credit has material consequences for researchers (Latour & Woolgar, 1986; P. E. Stephan, 2012). Another dimension in the study of scientific credit is the measurement of individuals' and institutions' scientific excellence is interested in credit allocation so as to more precisely evaluate the productivity and impact of individual authors, a particularly active area in scientometrics research. Even with a purely scientific interest in accounting for the accurate division of labor on the unit of the scientific paper is important for getting a clearer picture into the inner workings of the research lab. Because the collaboration information contained in the authorship line is situated on the paper level, our desires as science of science researchers to peer into the work behind the scientific paper will necessarily involve some amount of inference. However, there are still robust and varied modern strategies for inferring division of labor from paper co-author lines when approached with an appropriate understanding of the implicit conventions affecting them.

While it's a straightforward enough intuition that a single author publication would be counted as the full work of its sole author and count as 1 paper on the CV of its author, when a multi-authored paper represents the cumulative work of multiple researchers, it becomes challenging to accurately infer how much credit each of the authors should have for the paper. Because of these challenges, many traditional measures of author productivity simply count each papers on which the researcher was listed

as author as 1 full paper, regardless of the nature of their contribution—a strategy known as full counting (Waltman, 2016). Notably, the Hirsch index (h-index) (Hirsch, 2005), which remains in widespread use for faculty performance evaluation and by granting agencies (Hicks, 2012; P. Stephan et al., 2017), classifies any paper that the author appear on as a publication in this manner. Because full counting disregards both the author role and their level of involvement, using full counting in measures of research productivity often inflates the productivity of authors who are highly collaborative and appear on many papers, and the productivity of the research fields that are highly collaborative (Greene, 2007; Vavryčuk, 2018). In this case, instead of considering full counting as a measure of authorial productivity, a better definition of what full counting actually measures could be that it represents the *participation* of authors in varied research findings rather than *contribution* (Waltman, 2016).

Introducing additional nuance into credit allocation is done with fractional credit allocation to co-authors on team papers (Shen & Barabási, 2014; Waltman, 2016). This idea is based on the understanding that if a paper represents one finding, then each author actually only completed a fraction of the work for the paper. In this way, the credit due to each researcher for team efforts will often be adjusted to reflect these fractional contributions to the papers that they are on. For this reason, fractional authorship approaches to credit allocation also often divide the impact of each paper (the number of citations) by the number of authors carrying forward the assumption that each author is fractionally responsible for the success of the paper as well as its existence (Shen & Barabási, 2014; Waltman, 2016).

However, situations where each author truly contributes exactly equally to each paper are rare and challenging to verify, even between members of the same research team. Incorporating author-order into analyses also has the capacity to reveal more tacit information about the division of labor in the research lab because author order rests on a powerful and relatively consistent convention in the sciences. The order of the co-authors in the author lines of papers is usually a deliberate indication of the relative importance of the roles each author played within the research team (Tscharntke et al., 2007). The first author most likely played a central role in conceiving of the experiment and writing the paper, and they often also contributed to the experimentation and analyses (Larivière et al., 2016). The last author, who would likely be the principal investigator (PI) in charge of a research group or senior researcher, and they would also have helped write the paper and develop the research questions and research design (Larivière et al., 2016). Finally, any middle authors are likely to play support roles during the completion of the paper, such as carrying out experimentation, data analysis, and technical work. In addition, with the understanding that co-authors who are listed earlier in the paper byline contributed more to the project, this understanding can be combined with fractional counting methods. A variety of methodologies for inferring credit use both fractional credit and author order, such that the first author would be given more fractional credit for the paper, than their following co-authors (Donner, 2020; Waltman, 2016).

The biggest limitation for methodologies for attributing credit is that these methods are based on an understanding of specific conventions and not every field practices these conventions. The un-critical use of author order as a direct proxy for relative contribution, while powerful, is not exact or a guaranteed expression of credit. In particular, while the last author is taken as highly likely to be the PI of the research group, “the practice is still unofficial in most fields,” such that “some last authors mistakenly benefit from the assumption, even when they are not officially PIs (Tscharntke et al., 2007). Another notable example of a breakdown in the convention that author order matters at all is that in most Physics papers, particularly papers with a large coauthor team, all the authors are listed in alphabetical order (Waltman, 2012), which tells those who are not familiar with the contributors, very

little about the nature of each co-authors' contribution. However, the number of papers using alphabetic ordering is on the decline, with less than 4% of publications listing authors in alphabetical order; the use of alphabetical order seems to be the exception more than the rule, and is used only in the specific contexts of math, physics, and finance fields, and even then, it seems to be limited to special cases of particularly large or small author lists (Waltman, 2012). In these cases, the exceedingly long author lists suggest that these fields also follow much more expansive conventions for including contributors in the author line than other fields with fewer authors on their papers. Nevertheless, researchers hoping to use the author line to infer differential credit between the co-authors must remain mindful of these exceptions and include opportunities to detect them during their analyses.

Network science approaches

By creating network structures from bibliometric data, the methodologies and theories of network science can be brought to bear on quantitative research in the science of science. There are several types of network that can be built from bibliometric data, and with the input data of the co-author line, a collaboration network can be created by treating each publishing author as a node and linking together authors whenever they have appeared together on a paper's co-authorship line (Newman, 2001). Technically, the co-author network is a transformation of a bipartite network between the authors and the papers that they appear on. before the co-authorship network can be generated, authors will be connected to each of the papers they appeared on. The co-authorship network is broadly interpreted as capturing the large scale dynamics of who collaborates with whom (Newman, 2001). Network science is very fruitful for taking the simple data unit of the co-authorship information in the article byline and compiling these small units into a very informative whole network.

The appealing thing about network science analyses of co-authorship is that because network science is a field in itself, new theories and methodologies from network science can easily be imported to draw out new insights on the collaboration network. Because of the sheer number of network science approaches, I'm only going to highlight a few central questions about collaboration that network science methods are frequently used to study on the co-authorship network. In general, network science approaches to studying the co-authorship network can reveal large-scale collaboration patterns, including: the percolation of scientific ideas between collaborators (Morgan et al., 2018), field formation and strength (Barabási et al., 2002; Newman, 2004; Uddin et al., 2012), community structure (Girvan & Newman, 2002), and centrality and hierarchy of researchers in the collaboration network (Leydesdorff & Wagner, 2009; Uddin et al., 2013). For example, Barabasi et al. look at the evolution of the field of physics through the development of its collaboration network, finding a scale-free relationship between the authors (Barabási et al., 2002). The scale free nature of the collaboration network also suggests that there is inequality at play in the collaboration structures of science. Position within the collaboration network and access to good collaborators is an important elements in researchers' ability to do more prestigious work (Li et al., 2013).

While large-scale collaboration networks of all the possible co-authors are useful, ego networks for researchers and their collaborators, which are small networks containing only the nodes connected to an individual author, are connected with a broad array of individual-centric strategies. These approaches using ego networks place individual researchers into the context of their collaboration network, and as a result they're useful for homing in on different individual patterns of collaboration. For example, Jadidi et al. compare the ego networks for male and female researchers to study potential gender disparities the makeup of researchers collaborators (Jadidi et al., 2017). However, one thing to remember when

considering network approaches is that, as with any method, network science's methods often need to be tweaked to fit with the full context of the scientific collaboration.

General conclusions and limitations about using only the coauthor line:

Section 1 has been an overview of the methods that use only the author line in scientific papers to infer information about the collaborative relationships of the authors who completed the research paper. From the co-authorship line as basic unit of analysis, there are many ways of maneuvering what is essentially only a list of author names into different configurations in order to study collaboration. However, at the same time, it is important to fully acknowledge that the foundation of collaboration studies using co-authorship rests on the assumption that the authorship line is an accurate reflection of the people who collaborated on the underlying project, but there are some limitations to this assumption. For example, it is debatable whether there is a one-to-one correspondence between scientific papers, projects, and collaborations because within the lived social practices of collaboration, there is sometimes little clear division between particular projects, let along papers (Lane et al., 2015). For example, while a collaboration between people might last years, and might involve multiple emails a day, the unit of co-authorship will still represent this relationship only in terms of the papers that have resulted from the relationship (Ponomariov & Boardman, 2016). Not every research collaboration creates a paper and oftentimes research collaborations often produce more than one paper.

Further complicating the co-author line as a direct account of the collaborators on the research team is that there is a certain amount of social negotiation involved with who is included as a formal co-author for a particular paper and who is not. For example, as scientific papers are conceptualized as depicting the findings of particular research projects, technicians whose work focuses on the general maintenance of the research lab may not be included as co-authors (Hagstrom, 1964). Because there are social conditions feeding into any decisions about co-authorship, many in the bibliometric field have recommended caution in using co-authorship as a proxy for collaboration, up to the point of questioning its validity altogether (Katz & Martin, 1997; Ponomariov & Boardman, 2016). Furthermore, there is considerable tacit disagreement about the necessary levels of participation for authors to qualify for co-author status on papers; so much so that professional organizations are beginning to issue guidelines about who should be included as co-authors, and there are frequent calls from researchers for journals and funding agencies to adopt more consistent standards for authorship (International Committee of Medical Journal Editors, 2015; McNutt et al., 2018).

However, if the exclusion of collaborators is one complicating issue to collaboration studies that use the co-author line, the inclusion of people who did not substantially contribute to the paper is another concern, commonly known as "hyperauthorship" (Cronin, 2001), "gift authorship" (Tscharntke et al., 2007) or "honorary authorship" (Bozeman & Youtie, 2017; Greenland & Fontanarosa, 2012). The practice arises more frequently when someone is listed as author on a paper to which they have not made a significant contribution. There are various reasons for this practice, but most often gift or honorary authorship is applied to situations where the honorary author is a senior researcher, or a well-known or otherwise prestigious member of the scientific community (Bozeman & Youtie, 2017; Greenland & Fontanarosa, 2012). In some cases the presence of the honorary contributor might result in a citation boost, in which case there's an element of self- interest for the true contributors as well (Bozeman & Youtie, 2017), and the practice is strongly discouraged as it has the potential to dilute the credit attributed to true co-authors and undermine the central tenet of authorship—that each author is responsible for the research described in the paper (Greenland & Fontanarosa, 2012). However, clear-

cut cases of honorary authorship is nonetheless challenging to catch because an explicit delineation of the contribution of each author is unfortunately not universal. Also, the threshold for co-authorship is still defined by convention and not with set standards, so it is often unclear how much a co-author does have to contribute to meet this threshold (Tscharntke et al., 2007).

However, these limitations do not invalidate the co-author line as long as we maintain a strong understanding of the conventions governing authorship. In general, most of the additional meaning that can be inferred from the co-author line is a mobilization of tacit knowledge about various social conventions surrounding authorship, including conventions about who tends to qualify for authorship, what kinds of work are understood as co-author worthy, and how author order is determined. This raises the point that even at the larger scale, the quantitative studies of collaboration rely on a knowledge of scientific conventions that have been discerned through the work of the observational set. Careful monitoring of the current conventions about co-authorship are necessary and important for the continued accuracy of these methods. No data source is perfect, and because the co-author line appears on every scientific paper with vanishingly few exceptions, the co-author line has the coverage necessary for large-scale bibliometric studies.

Section 2: Beyond the co-authorship byline techniques and developments

The byline of the scientific paper listing the co-authors is an invaluable resource because it is a feature that nearly all papers have. However, there are other features of scientific papers that carry rich information about collaboration and the roles that paper authors have fulfilled during their collaborations. These other sources of information about the collaborating relationship that produced the papers are less standard, so they may not appear on everything, but they are still useful and from the basis for a great deal of groundbreaking work. However, before discussing the less common sources of collaboration information, I will start with the most common feature first.

Departmental and institutional affiliations

An explicit statement of each author's affiliation with their research institution at the time of the paper's publication is also a nearly universal feature of papers in academic journals. The information about each author's research institution serves to provide contact information for the paper co-authors, but also because the authors' employing universities benefit from being acknowledged. One of the most basic facts about the working conditions within any particular collaboration is whether the collaborators are working in physical proximity or not, and the affiliation information provides a crucial and easy to understand indication for this essential fact. Several studies of distance collaboration using author affiliation have shown that collaborations where there is significant geographic distance between collaborators is growing more common (Gazni et al., 2012), and many of these studies credit the newly found ease of online collaboration or cheaper air travel as facilitating conditions that could be behind the increased potential for collaboration at a distance (Catalini et al., 2016; Gazni et al., 2012; Wagner et al., 2015). The presence of institutional affiliation during studies of collaboration have also allowed for a highly active, robust research topic on national and international collaboration (Gazni et al., 2012; Leydesdorff & Wagner, 2009; Wagner et al., 2015). One limitation of using author affiliation is the fact that authors sometimes have multiple affiliations and different reasons for including or excluding them from a paper. This may introduce some noise into studies of author affiliation that would need to be dealt with on a case-by-case basis.

Explicit Author-submitted Statements on Contributorship

The studies looking at author order rely on several implicit conventions, but there are also efforts underway to encourage researchers to explicitly declare a full accounting for the research tasks that each author performed for the papers they submit. Several academic journals, including Nature, PNAS and PLoS, require author statements about the division of labor (McNutt et al., 2018; "Policy on Papers' Contributors," 1999). Many major journals are starting to require breakdowns of the division of labor for the articles; these publications accomplish this by requiring authors to complete questionnaires during paper submission or by requiring a written statement on division of labor that has been approved by all the authors. This information is then included in the metadata of the articles. These efforts are often framed as a way to better allocate credit to authors and may also help to combat hyperauthorship, as these disclosures are accompanied by ethical agreements not to lie (Cronin, 2001; Greenland & Fontanarosa, 2012). Amidst the competing strategies for fractional authorship, weighting methods, etc. several researchers have specifically argued that only contributor submitted accounts of each contributors role are fully sufficient, perhaps even ideal ways to attribute author credit (Tscharntke et al., 2007; Vavryčuk, 2018).

However, while efforts to include author statements are generally approved of, the uptake of this convention has been relatively slow and implemented only in a piecemeal fashion at a handful of academic journals. Additionally, the conventions for reporting on division of labor are non-standard among journals that have them. Furthermore, some questions about the reliability of the division of labor statements have been raised, as this data about who did what is usually self-reported by the corresponding author. At the present time, there are no ways for journals to conclusively verify that the division of labor during the paper collaboration was indeed as reported. In many cases, during these statements, some authors simply check all the boxes in the submission form that everyone on the paper did everything.

Honorary mentions: Acknowledgement sections and payroll data

We have reached the point where I'd like to make few miscellaneous notes about a couple remaining data sources and methods that aren't in broad use, but which have the potential to lend additional insight into collaboration. Unlike the co-author line or an explicit statement regarding author contributions, the acknowledgement section may be the place where contributions that did not qualify for co-authorship status are reported (Cronin, 2004; Paul-Hus et al., 2017). While several fields have traditionally been understood to have smaller research teams, like social science, when the acknowledgments of papers in these fields are counted as members of the research team, the team size for social science isn't as comparatively smaller (Paul-Hus et al., 2017). Additionally, even in writing-centric fields that are commonly understood to be solitary, the extensive acknowledgement sections demonstrate the kinds of community support and collaboration, distributed cognition to use Cronin's term, necessary for writing-based endeavors, (Cronin, 2004). However, the limitation to including the acknowledgement section is that there is still no standard form for acknowledgement section nor is it consistently indexed by bibliometric databases (Paul-Hus et al., 2017). Furthermore, just as there are differing field conventions for inclusion in the authorship line, there certainly will be field differences in the writing of an acknowledgment section.

In addition, these future directions include new ideas that base studies of collaboration through different avenues, like looking at research funding, are less common, but still fascinating possibilities.

While studies of grants and outcomes of receiving grants, take as a given that project-based science endeavors are more commonly carried out within research teams of collaborators, some concerns have been raised that studying science from the level of projects is too coarse-grained because within the actual working environment of the research team, the work for a particular project is near impossible to definitively separate out from the rest of the work (Lane et al., 2015). There have also been efforts to use payroll or funding data to get more fine-grained, and specific accounts of the specific roles of the researchers working on projects, such as through the UMETRICS database project (Lane et al., 2015). The UMETRICS database is pretty new, and many of the other data sources, aren't in wide standard usage and as the dataset can only include data from universities and institutions that opt in, and specifically send in their funding data, this dataset, and others like it require a great deal of bespoke data gathering and cleaning. In addition, due to the sensitivity of payroll data, the UMETRICS dataset is also accessible to only a small group of researchers specifically associated with the project and with security clearance.

Disambiguation and ORCID

One overarching issue that must be discussed is the identification and disambiguation of the researchers in the co-author line because this touches nearly all areas of collaboration research. The disambiguation of author names is a major headache since, in order to study authorship, it's often important to be able to distinguish all the researchers publishing papers from each other. Since most analyses begin with only the names of the authors, in order to gather together any particular individuals' papers, we have to clean the text of the names and match the incidences of each of the names, but this matching process is often beset with difficulties with alternate spellings or typography errors, and duplicate names (Smalheiser & Torvik, 2009; Tang & Walsh, 2010). This issue is made more important by the fact that current disambiguation methods underperform particularly badly for authors with East Asian names, which are subject to a higher error rate as they are transcribed into English databases, or as multiple variants of family names are assigned the same English spelling (Tang & Walsh, 2010). Many bibliometric databases have implemented some form of disambiguation for authors: Leiden has one for the Web of Science (Caron & Van Eck, 2014), and Microsoft Academic Graph has one, to name a couple. Leiden's method, for example is able to improve their matching rates by taking into account a number of other factors, like field of study and affiliation, that are likely to distinguish authors from each other. Other methods use the knowledge base of the articles that might be attributed to particular authors into account when matching papers to their authors (Tang & Walsh, 2010). However, the accuracy of each one is challenging to assess because there is no ground truth dataset.

However, many hopes for alleviating the issues with author disambiguation rest on ORCID, the Open Researcher and Contributor ID. ORCID is an infrastructure to give a stable identifier to each researcher so that their contributions will be more readily and clearly identifiable and distinguishable from other researchers with similar names (orcid_about, 2012). Academic journals including several of the masthead publications like PNAS and particularly open access journals such as PLoS, are beginning to specifically encourage their authors, editors and reviewers to register for an ORCID (McNutt et al., 2018; "ORCID," n.d.). ORCID is definitely a step in the right direction because they make it easier to compile accurate lists of author publications in a transparent fashion (Haak et al., 2018), but they'll only be available for current and future research, not historical research. Even in the present day, not everyone has an ORCID or uses it reliably yet, particularly outside the US and European contexts (Youtie et al., 2017). Unique author id databases will also be face many of the same challenges of other large scale databases, including the tenuousness of data saving capabilities (Smalheiser & Torvik, 2009). I sure hope

we're able to use these cross-platform unique identifiers more consistently, because disambiguation is such a big sticking point, not only for bibliometric-based research, but also for everyday metric gathering on the individual level for the purposes of evaluation.

General conclusions and limitations for using more than the coauthor line

While there are many ways and scales for studying collaboration in the sciences, I have focused on the rather complex problem of examining inter-team dynamics in research collaboration. In this case, although the paper is the major unit of analysis, the dynamics I'm interested in actually take place below the level of the paper, within the composition of the team and the agglomeration of their labor to generate research findings and write the paper. Methods of inferring below the level of the paper to identify the actual author roles and division of labor can often greatly benefit from the addition of specialized data, because any additional information can help peel back that curtain into the actual collaborative structures within science. In fact, so many of the methods that we have for studying collaboration within the paper-producing scientific team are various ways of inferring information specifically contained in the data sources "beyond the co-author line" that I've discussed in this section.

In particular, many of the methods using the co-author line are ways of trying to infer the information that an explicit division of labor statement would answer directly. As such, the explicit author contributor statements are considered the gold standard for collaboration data because they provide some verification that all listed authors actually made substantial contributions, and they provide clarity about the author roles (McNutt et al., 2018; Tscharntke et al., 2007; Vavryčuk, 2018). Some form of author contributor statements are being required by an increasing number of journals, the fact that authorship statements as yet are being require, implemented, and aggregated by individual journals is actually the crux of the issue about contributor statements, as well as many of the other sources of beyond the authorship line data. There is so much variation in these outside methods, and many of them are not as in wide use as the methods for studying collaboration using the co-authorship line.

Although the pure authorship line does not provide anywhere close to the nuance and potential for validation that outside co-author line methods like contributor statements do, co-author line based methods still have the advantage of their easier availability and greater uniformity. At the present time, the infrastructure is still missing for gathering, standardizing, and making data sources like contributor statements or acknowledgement sections available on a large scale. However, since these initiatives like these, particularly contributor statements, have active initiatives to expand their usage and availability, hopefully coverage will become more consistent in the near future.

Conclusion

This summary highlights an ample toolkit of methods available to study collaboration within the scientific workforce, both from the co-author line in bibliometric databases, and smaller, more piecemeal bits of data from publications. The use of quantitative methods allow researchers of scientific collaboration to scale up research questions by aggregating data across larger scales, like entire disciplines, countries, or across all of science. However, the unit that we want to aggregate from is also what defines the object of study, whether we are looking at papers, journals, research teams, or individual authors. Within publication databases the scientific paper is the smallest unit of analysis. Since a paper is a finished product of the scientific research, it may seem that quantitative research can only go as deep as the paper without looking behind it. However, I argue the contrary. The aim of the

methods I have discussed is to look inside the unit of the scientific paper to make inferences about the roles and work contributions of the researchers who produced the paper.

Writing on the management of team science, Bozeman and Youtie find that within the culture of scientific research collaborations, there is a strong tendency among researchers to assume that the standards for collaboration are already clear and that their research partners already think the way they do (Bozeman & Youtie, 2017). This is not true, and as a result Bozeman and Youtie argue that explicit decisions about research roles within team science will make the strength of collaborations and resulting science more successful, not to mention more pleasant. It's interesting if not surprising that the tendency to only hint at the division of labor in the research team also seems to extend into the papers, where the most frequent, widely used methods for understanding the strength of the research collaboration and each authors roles is to hint at the relative contributions of the co-authors by putting the co-authors names in order of importance.

Many of the methods summarized above rely on a nuanced understanding of the culture of scientific work, since these methods of inferring information about collaboration from paper data, rely on small details about the tacit conventions surrounding authorship that only insiders to scientific can interpret. In particular, we must understand and use the conventions about the work requirements for being included as an author, which are not often explicitly articulated, and we have to understand the conventions around author order. We also have to make distinctions and allowances for the sometimes radically different cultures and work structures of different scientific fields. For these reasons, even as quantitative researchers in the science of science, our work must be informed by and in conversation with the kinds of deep qualitative studies of scientific work practices that can gather and articulate the kinds of unspoken conventions that undergird the material artefacts of the work, like the co-author line. In other words, in order to have an accurate and up-to-date ability to interpret the data given as number of contributors, the author order, or the acknowledgment sections, we must pay attention to the qualitative, work of our colleagues in the qualitative science of science.

Regardless, given this tacit understanding of conventions and only a small piece of information, the co-author line, for scientific publications, researchers have done an outstanding job of inferring patterns in the scientific workforce. I look forward to using and expanding on these tools in my own research.

Works Cited

Barabási, 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), 590–614. https://doi.org/10.1016/S0378-4371(02)00736-7

Barley, S. R., & Bechky, B. A. (1994). In the Backrooms of Science: The Work of Technicians in Science Labs. *Work and Occupations*, *21*(1), 85–126. https://doi.org/10.1177/0730888494021001004

Bozeman, B., & Youtie, J. (2017). *The Strength in Numbers: The New Science of Team Science*. Princeton University Press.

Caron, E., & Van Eck, N. J. (2014). Large scale author name disambiguation using rule-based scoring and clustering. *Proceedings of the 19th International Conference on Science and Technology Indicators*, 79–86.

Catalini, C., Fons-Rosen, C., & Gaulé, P. (2016). *Did cheaper flights change the direction of science?* (Issue 1520). Department of Economics and Business, Universitat Pompeu Fabra. https://ideas.repec.org/p/upf/upfgen/1520.html

Cronin, B. (2001). Hyperauthorship: A postmodern perversion or evidence of a structural shift in scholarly communication practices? *Journal of the American Society for Information Science and Technology*, *52*(7), 558–569. https://doi.org/10.1002/asi.1097

Cronin, B. (2004). Bowling alone together: Academic writing as distributed cognition. *Journal of the American Society for Information Science and Technology*, *55*(6), 557–560. https://doi.org/10.1002/asi.10406

De Solla Price, D. J., & Beaver, D. (1966). Collaboration in an invisible college. *American Psychologist*, *21*(11), 1011–1018. https://doi.org/10.1037/h0024051

Donner, P. (2020). A validation of coauthorship credit models with empirical data from the contributions of PhD candidates. *Quantitative Science Studies*, *1*(2), 551–564. https://doi.org/10.1162/qss_a_00048

Flier, J. S. (2019). Credit and Priority in Scientific Discovery: A Scientist's Perspective. *Perspectives in Biology and Medicine*, *62*(2), 189–215.

Gazni, A., Sugimoto, C. R., & Didegah, F. (2012). Mapping world scientific collaboration: Authors, institutions, and countries. *Journal of the American Society for Information Science and Technology*, *63*(2), 323–335. https://doi.org/10.1002/asi.21688

Girvan, M., & Newman, M. E. J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, *99*(12), 7821. https://doi.org/10.1073/pnas.122653799

Greene, M. (2007). The demise of the lone author. *Nature*, *450*(7173), 1165–1165. https://doi.org/10.1038/4501165a

Greenland, P., & Fontanarosa, P. B. (2012). Ending Honorary Authorship. *Science*, *337*(6098), 1019. https://doi.org/10.1126/science.1224988

Haak, L. L., Meadows, A., & Brown, J. (2018). Using ORCID, DOI, and Other Open Identifiers in Research Evaluation. *Frontiers in Research Metrics and Analytics*, *3*, 28. https://doi.org/10.3389/frma.2018.00028

Hagstrom, W. O. (1964). Traditional and Modern Forms of Scientific Teamwork. *Administrative Science Quarterly*, *9*(3), 241–263. JSTOR. https://doi.org/10.2307/2391440

Hicks, D. (2012). Performance-based university research funding systems. *Research Policy*, *41*(2), 251–261. https://doi.org/10.1016/j.respol.2011.09.007

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(46), 16569. https://doi.org/10.1073/pnas.0507655102

International Committee of Medical Journal Editors. (2015). *Defining the Role of Authors and Contributors*. http://www.icmje.org/recommendations/browse/roles-and-responsibilities/defining-the-role-of-authors-and-contributors.html

Jadidi, M., Karimi, F., Lietz, H., & Wagner, C. (2017). Gender Disparities in Science? Dropout, Productivity, Collaborations and Success of Female and Male Computer Scientists. *Advances in Complex Systems*, *21*(03n04), 1750011. https://doi.org/10.1142/S0219525917500114

Katz, J. S., & Martin, B. R. (1997). What is research collaboration? *Research Policy*, *26*(1), 1–18. https://doi.org/10.1016/S0048-7333(96)00917-1

Kronick, D. A. (2001). The Commerce of Letters: Networks and "Invisible Colleges" in Seventeenth- and Eighteenth-Century Europe. *The Library Quarterly: Information, Community, Policy*, *71*(1), 28–43. JSTOR.

Lane, J. I., Owen-Smith, J., Rosen, R. F., & Weinberg, B. A. (2015). New linked data on research investments: Scientific workforce, productivity, and public value. *The New Data Frontier*, *44*(9), 1659–1671. https://doi.org/10.1016/j.respol.2014.12.013

Larivière, V., Desrochers, N., Macaluso, B., Mongeon, P., Paul-Hus, A., & Sugimoto, C. R. (2016). Contributorship and division of labor in knowledge production. *Social Studies of Science*, *46*(3), 417–435. https://doi.org/10.1177/0306312716650046

Larivière, V., Gingras, Y., Sugimoto, C. R., & Tsou, A. (2015). Team size matters: Collaboration and scientific impact since 1900. *Journal of the Association for Information Science and Technology*, *66*(7), 1323–1332. https://doi.org/10.1002/asi.23266

Latour, B. (2003). *Science in action: How to follow scientists and engineers through society* (11. print). Harvard Univ. Press.

Latour, B., & Woolgar, S. (1986). Cycles of Credit. In *Laboratory life: The construction of scientific facts*. Princeton University Press.

Leydesdorff, L., & Wagner, C. (2009). International collaboration in science and the formation of a core group. *Journal of Informetrics*, *2*(4), 317–325. https://doi.org/10.1016/j.joi.2008.07.003

Li, E. Y., Liao, C. H., & Yen, H. R. (2013). Co-authorship networks and research impact: A social capital perspective. *Research Policy*, *42*(9), 1515–1530. https://doi.org/10.1016/j.respol.2013.06.012

McNutt, M. K., Bradford, M., Drazen, J. M., Hanson, B., Howard, B., Jamieson, K. H., Kiermer, V., Marcus, E., Pope, B. K., Schekman, R., Swaminathan, S., Stang, P. J., & Verma, I. M. (2018). Transparency in authors' contributions and responsibilities to promote integrity in scientific publication. *Proceedings of the National Academy of Sciences*, *115*(11), 2557. https://doi.org/10.1073/pnas.1715374115

Milojević, S. (2014). Principles of scientific research team formation and evolution. *Proceedings of the National Academy of Sciences*, *111*(11), 3984. https://doi.org/10.1073/pnas.1309723111

Morgan, A. C., Economou, D. J., Way, S. F., & Clauset, A. (2018). Prestige drives epistemic inequality in the diffusion of scientific ideas. *EPJ Data Science*, *7*(1), 40. https://doi.org/10.1140/epjds/s13688-018-0166-4

Newman, M. E. J. (2001). Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. *Physical Review E*, *64*(1), 016132. https://doi.org/10.1103/PhysRevE.64.016132

Newman, M. E. J. (2004). Coauthorship networks and patterns of scientific collaboration. *Proceedings of the National Academy of Sciences*, *101*(suppl 1), 5200. https://doi.org/10.1073/pnas.0307545100

ORCID. (n.d.). *PLOS*. Retrieved September 7, 2020, from https://plos.org/open-science/orcid/

orcid_about. (2012, August 17). *Our Mission* [Text]. https://orcid.org/about/what-is-orcid/mission

Paul-Hus, A., Mongeon, P., Sainte-Marie, M., & Lariviere, V. (2017). The sum of it all: Revealing collaboration patterns by combining authorship and acknowledgements. *Journal of Informetrics*, *11*(1), 80–87. https://doi.org/10.1016/j.joi.2016.11.005

Policy on papers' contributors. (1999). *Nature*, *399*(6735), 393–393. https://doi.org/10.1038/20743

Ponomariov, B., & Boardman, C. (2016). What is co-authorship? *Scientometrics*, *109*(3), 1939–1963. https://doi.org/10.1007/s11192-016-2127-7

Price, D. J. de S., & Price, D. J. de S. (1986). *Little science, big science—And beyond*. Columbia University Press.

Shen, H.-W., & Barabási, A.-L. (2014). Collective credit allocation in science. *Proceedings of the National Academy of Sciences*, *111*(34), 12325. https://doi.org/10.1073/pnas.1401992111

Smalheiser, N. R., & Torvik, V. I. (2009). Author name disambiguation. *Annual Review of Information Science and Technology*, *43*(1), 1–43. https://doi.org/10.1002/aris.2009.1440430113

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

Stephan, P., Veugelers, R., & Wang, J. (2017). Reviewers are blinkered by bibliometrics. *Nature News*, *544*(7651), 411. https://doi.org/10.1038/544411a

Sugimoto, C. R., & Larivière, V. (2018). *Measuring research: What everyone needs to know*. Oxford University Press.

Tang, L., & Walsh, J. P. (2010). Bibliometric fingerprints: Name disambiguation based on approximate structure equivalence of cognitive maps. *Scientometrics*, *84*(3), 763–784. https://doi.org/10.1007/s11192-010-0196-6

Tscharntke, T., Hochberg, M. E., Rand, T. A., Resh, V. H., & Krauss, J. (2007). Author Sequence and Credit for Contributions in Multiauthored Publications. *PLOS Biology*, *5*(1), e18. https://doi.org/10.1371/journal.pbio.0050018

Uddin, S., Hossain, L., Abbasi, A., & Rasmussen, K. (2012). Trend and efficiency analysis of co-authorship network. *Scientometrics*, *90*(2), 687–699. https://doi.org/10.1007/s11192-011-0511-x

Uddin, S., Hossain, L., & Rasmussen, K. (2013). Network Effects on Scientific Collaborations. *PLOS ONE*, *8*(2), e57546. https://doi.org/10.1371/journal.pone.0057546

Vavryčuk, V. (2018). Fair ranking of researchers and research teams. *PLOS ONE*, *13*(4), e0195509. https://doi.org/10.1371/journal.pone.0195509

Wagner, C., Park, H., & Leydesdorff, L. (2015). *The Continuing Growth of Global Cooperation Networks in Research: A Conundrum for National Governments* (Vol. 10). https://doi.org/10.1371/journal.pone.0131816

Waltman, L. (2012). An empirical analysis of the use of alphabetical authorship in scientific publishing. *Journal of Informetrics*, *6*(4), 700–711. https://doi.org/10.1016/j.joi.2012.07.008

Waltman, L. (2016). A review of the literature on citation impact indicators. *Journal of Informetrics*, *10*(2), 365–391. https://doi.org/10.1016/j.joi.2016.02.007

Wuchty, S., Jones, B. F., & Uzzi, B. (2007). The Increasing Dominance of Teams in Production of Knowledge. *Science*, *316*(5827), 1036. https://doi.org/10.1126/science.1136099

Youtie, J., Carley, S., Porter, A. L., & Shapira, P. (2017). Tracking researchers and their outputs: New insights from ORCIDs. *Scientometrics*, *113*(1), 437–453. https://doi.org/10.1007/s11192-017-2473-0