an open access journal

Citation: Akbaritabar, A. (2021). A quantitative view of the structure of institutional scientific collaborations using the example of Berlin. *Quantitative Science Studies*, *2*(2), 753–777. https://doi.org/10.1162/qss_a_00131

DOI:
https://doi.org/10.1162/qss_a_00131

Peer Review:
https://publons.com/publon/10.1162/qss_a_00131

Received: 2 September 2020
Accepted: 2 April 2021

Corresponding Author:
Aliakbar Akbaritabar
akbaritabar@demogr.mpg.de

Handling Editor:
Ludo Waltman

The MIT Press

RESEARCH ARTICLE

A quantitative view of the structure of institutional scientific collaborations using the example of Berlin

Aliakbar Akbaritabar[1,2]

[1]Max Planck Institute for Demographic Research (MPIDR),
Laboratory of Digital and Computational Demography, Rostock, Germany
[2]German Centre for Higher Education Research and Science Studies (DZHW), Berlin, Germany

Keywords: Berlin, Berlin University Alliance, bipartite community detection, coauthorship network analysis, disambiguation, internationalization

ABSTRACT

This paper examines the structure of scientific collaborations in Berlin as a specific case with a unique history of division and reunification. It aims to identify strategic organizational coalitions in a context with high sectoral diversity. We use publications data with at least one organization located in Berlin from 1996–2017 and their collaborators worldwide. We further investigate four members of the Berlin University Alliance (BUA), as a formerly established coalition in the region, through their self-represented research profiles compared with empirical results. Using a bipartite network modeling framework, we move beyond the uncontested trend towards team science and increasing internationalization. Our results show that BUA members shape the structure of scientific collaborations in the region. However, they are not collaborating cohesively in all fields and there are many smaller scientific actors involved in more internationalized collaborations in the region. Larger divides exist in some fields. Only Medical and Health Sciences have cohesive intraregional collaborations, which signals the success of the regional cooperation established in 2003. We explain possible underlying factors shaping the intraregional groupings and potential implications for regions worldwide. A major methodological contribution of this paper is evaluating the coverage and accuracy of different organization name disambiguation techniques.

1. INTRODUCTION

Researchers work for academic and nonacademic organizations and firms and use the resources from these organizations to carry out scientific work and form scientific collaborations. Coalitions and strategic ties between scientific organizations can be a *cause* and/or an *effect* of the way scientists affiliated to them communicate with each other. An example of the former is the top-down regional, national, or organizational policies that support specific types of collaborations (e.g., the COST[1] initiative to foster scientific networking in Europe). The latter is driven more by the individual motivations of scientists to start bottom-up research projects and obtain funding through interorganizational collaborations with researchers of other (inter) national organizations (e.g., ERC[2] starting, consolidator, or advanced grants).

[1] https://www.cost.eu/
[2] https://erc.europa.eu/

We aim to look at the outcome of scientific collaborations, in the form of scientific publications, that are produced through either the former or latter process. By understanding the structure of scientific collaborations between organizations, we aim to find a proxy to identify possible strategic coalitions among them that in turn could have been inspired by individual researchers. This dichotomy is nothing but a simplification to serve the purpose of the current article. We are aware that large multiorganizational collaborations are not simply agreements of individual researchers (Shrum, Genuth et al., 2007, p. 117).

Strategic coalitions could take different forms and lead to differing set of outputs (Katz & Martin, 1997; Laudel, 2002). Here we are focused on coauthorship as one of the main forms of collaboration and scientific publications as the expected output. We are aware that coauthorship offers only a reductionist view of collaboration, but nevertheless it is one of the frequently used measures of scientific collaboration (Shrum et al., 2007, pp. 7–8).

Moreover, strategic coalitions can be affected by linguistic (Avdeev, 2019), geographical (Katz, 1994), and regional proximities (Luukkonen, Persson, & Sivertsen, 1992). In an in-depth review, Small and Adler (2019) presented a diverse array of literature that emphasized the effect of *space* in the formation of social ties. Scientific organizations are populated by scientists and science is a social enterprise (Fox, 1983). Thus, it is not counterintuitive to consider scientific collaborations as a form of social tie oriented towards shared objectives (Shrum et al., 2007, p. viii). The formation of these ties is *facilitated* or *hindered* by the contextual (Akbaritabar, Casnici, & Squazzoni, 2018; Small, 2017, p. 154; Sonnenwald, 2007), social (Akbaritabar & Squazzoni, 2020; Smith-Doerr, Alegria, & Sacco, 2017), and epistemic preferences of researchers and they can result in denser or instead sparser scientific communities (Akbaritabar, Traag et al., 2020).

In addition, the increasing trend towards more collaborative work and *team science* is well known (Araújo, Araújo et al., 2017; Wuchty, Jones, & Uzzi, 2007). It is claimed that scientific fields, even social sciences, which tend to be nationally oriented, are moving towards more intense collaborations and more internationalization. For hard sciences (e.g., particle physics), scientific discovery has major reliance on multiorganizational scientific collaborations (Shrum et al., 2007, pp. 3, 7). However, studies have highlighted the differences in national or disciplinary contexts in the rate of internationalization (Babchuk, Keith, & Peters, 1999; Moed, De Bruin et al., 1991) or differing rates of benefits, in terms of impact, obtained from internationalized collaborations (Glänzel, Schubert, & Czerwon, 1999).

We intend to explore the interplay between different contextual variables and geographical space to investigate the structure of scientific collaborations in the Berlin metropolitan region. We focus on Berlin as a regional hub, with high geographical proximity, that can inspire specific organizational arrangements. Therefore, we investigate the following research questions:

- **RQ1**: How *collaborative* and *internationalized* is the scientific landscape of the Berlin metropolitan region?
- **RQ2**: Are there *field* differences in the rate of collaborative and internationalized scientific work?
- **RQ3**: How *sector* oriented is scientific collaboration in the Berlin metropolitan region?
- **RQ4**: Is there evidence of strategic coalitions, or field, regional, or organizational *agreements* in the structure of scientific collaborations in the Berlin metropolitan region?
- **RQ5**: Are there specific field, sectoral, national, or continental *cohesive subgroups* driving the scientific collaborations in the Berlin metropolitan region?

The contribution of the current paper is fourfold: (a) We focus on the scientific output of the Berlin metropolitan region and trace the share of collaborative works and identify the share of international collaborations. We separate Berlin, Germany, Europe, and continental regions worldwide to investigate possible groupings and we intend to move beyond the descriptive and macro view, which advocates for increasing internationalization. (b) We cover all major OECD scientific fields and provide a comparative view of the specificities of these fields. We also include a sectoral view based on the type of organizations. (c) We develop and use multiple organization name disambiguation techniques and compare their efficiency, coverage, and accuracy, and (d) we employ a bipartite network modeling and community detection approach and present how it can be useful in coauthorship network analysis and identification of denser groups collaborating preferentially among themselves.

The structure of the paper is as follows: Section 2 presents the prior studies. Section 3 presents our data sources and modeling strategy. Section 4 presents our findings, followed by discussion and limitations in Section 5 and conclusions in Section 6.

2. LITERATURE REVIEW

Balland, Jara-Figueroa et al. (2020) argued that scientific and complex economic activities are concentrated in urban and metropolitan areas. In a large-scale study of 353 U.S. metropolitan areas, they found that disproportionate spatial concentration increases with complexity of productive activities. Using the average number of authors in scientific publications as a proxy for complexity (due to the higher coordination cost of larger scientific teams), they found that scientific fields with higher complexities tend to have more urban concentration.

In the case of Europe, policies and initiatives are developed with the aim of building an "integrated European Research Area." Hoekman, Frenken, and Tijssen (2010) tested whether this objective has been achieved. They concluded that Europe leans towards more integration in subjects and fields that were previously national endeavors. Nevertheless, they reported that geographically localized coauthorship was prevalent, with tendencies towards high degrees of difference among fields in the regional, national, or European contexts. They found that some fields (e.g., physical sciences and life sciences) were in a more advanced stage of "Europeanization," while other fields (e.g., medicine, engineering, social science, and humanities) present a more nationally oriented profile of scientific collaboration.

Specific national contexts can present a higher or lower degree of scientific production and internationalized coauthorship. It is important to take the national context into account along with the continental and regional views. Stahlschmidt, Stephen, and Hinze (2019) and Stephen, Stahlschmidt, and Hinze (2020) found that Germany has a stable rate of growth in the number of scientific publications, similar to that of OECD countries with more established science systems (e.g., the United States, the United Kingdom, and France). Germany is moving towards higher rates of international collaborations in most scientific fields (from 46% internationalized coauthorships in 2007 to 55% in 2017 in Scopus and from 47% in 2007 to 59% in 2017 in Clarivate's Web of Science (WoS)). The United States, the United Kingdom, France, Switzerland, Italy, the Netherlands, China, Spain, Austria, and Australia are the 10 countries with the highest shares of coauthorship with Germany in Scopus. In addition, Aman (2016) presented evidence of increasing internationalization and also higher rates of citations for interorganizational and international coauthorship for the German science system in WoS from 2007 to 2012.

There are also studies specifically focused on the Berlin metropolitan region. Rammer, Kinne, and Blind (2020) found a form of selective spatial proximity between knowledge-producing institutions (e.g., universities) and knowledge-demanding institutions (e.g., innovative companies and firms). They reported a microgeographic scope where innovative firms were surrounded by same-sector firms and located closer to universities and research institutes.

Abbasiharofteh and Broekel (2020) explored the biotechnology field in the Berlin metropolitan region. They concluded that Eastern and Western organizations within Berlin are still not cohesively collaborating with each other. The “shadow” of the Berlin wall still influences the scientific collaborations of the region.

Organizational, regional, and continental coalitions are being developed to support higher rates of scientific collaboration among the actors in these contexts. A specific example is the Berlin University Alliance (BUA)[3]. The BUA was founded in February 2018 between the three main universities and one university hospital located in the Berlin metropolitan region (i.e., Freie Universität Berlin (FU), Humboldt-Universität zu Berlin (HU), Technische Universität Berlin (TU), and Charité – Universitätsmedizin Berlin (CH)) (Berlin University Alliance, Humboldt-Universität zu Berlin et al., 2018, 2019). The BUA claims to have been established based on a longstanding record of intraregional collaborations between these institutions. The interaction between these institutions began following an era of institutional isolation after the fall of the Berlin Wall during which these institutes needed to define and empower their unique identities. Afterwards, the first forms of cooperation between these institutes emerged, which led in 2003 to the establishment of a shared medical faculty between HU and FU to be located in CH’s facilities. There are examples of competition, mutual definition of exclusive research areas, and graduate programs versus close cooperation among BUA members in the past three decades. These are highlighted in the BUA’s proposal as dimensions of the unique history and strengths of the region. The BUA aims at fostering previous collaboration experiences in a new organizational form. One of our minor goals is to control whether these four institutions have a distinctive position in the structure of scientific collaborations formed in the Berlin metropolitan region.

Network analysis can be used to identify the presence of communities in coauthorship networks (Akbaritabar et al., 2020; Leone Sciabolazza et al., 2017; Palla, Barabási, & Vicsek, 2007). Quantitative models are used to examine whether collaboration patterns persist *between* or *within* denser areas of the network and in form of specific communities. Looking at the composition of these communities and identifying potential factors contributing to their cohesion helps to explain groupings in scientific collaborations.

In levels lower than continental, national, or regional frameworks, scientific organizations themselves can have strategic plans to define their overarching identities and main research foci. This might inspire researchers in a certain organization to prioritize research in specific fields and areas (Blume, Bunders et al., 1987) to show allegiance with the organization’s designated identity, which in turn could penalize researchers’ selection of innovative research themes (Rijcke, Wouters et al., 2016). Goals set out by funding agencies could affect collaborations (Nederhof, 2006; Wagner, Park, & Leydesdorff, 2015). Furthermore, the type of organization (i.e., sector) partially determines the type of research that an organization conducts and its expected outcomes.

In addition to the themes discussed above, the type of data employed to answer the research questions could have a large effect on the identified trends (Huang, Neylon et al., 2020). Bibliometric databases are not perfect and any given one could be prone to specific errors. In terms of coverage, different databases have certain policies to define what should be indexed (Huang et al., 2020). This affects the results of macro studies depending on the database employed, subset of scientific publications used, document types analyzed, and level of aggregation and normalization applied (Stahlschmidt et al., 2019; Stephen et al., 2020). In terms of cleanness of the data, there is a strong need for disambiguation of scientific entity names

[3] https://www.berlin-university-alliance.de/en/about/index.html

(e.g., authors, organizations) which might bias the quality of results (Aman, 2018; D'Angelo & van Eck, 2020; Donner, Rimmert, & van Eck, 2019). Thus, one of our methodological goals is to introduce organization name disambiguation techniques to match scientific organization names with publicly available databases (e.g., Wikidata and Global Research Identifier Database (GRID)) and evaluate the reliability of the results in comparison to established techniques.

3. DATA AND METHODS

We use Scopus 2018 data from the German Competence Centre for Bibliometrics (KB)[4]. We extract *article*, *review*, and *conference proceedings* documents published from the beginning of the database in 1996 until the end of 2017. To delineate the *Berlin* metropolitan region and to identify the scientific collaborations that occurred in the region, we select only publications that have at least one authoring organization located in Germany and Berlin. Thus, coauthorship here includes Berlin organizations and their collaborators worldwide. Our level of analysis is *scientific organizations* (i.e., each affiliation address mentioned in a publication that can be academic or nonacademic organizations or firms with which researchers are affiliated) and we do not investigate lower levels (e.g., authors).

Our data include different metadata for each publication, such as *publication year*, *title*, *affiliation addresses*, *scientific field*, *journal name*, and *document type*. We include *conference proceedings* in addition to articles and reviews as there are technical universities in the sample for which this type of document is considered influential. We use a mapping of publications to OECD scientific fields based on Scopus ASJC[5], which reduces the number of subject categories from 33 to a more interpretable six categories (OECD, 2007). We compare the aggregate data for trends between the different OECD scientific fields (i.e., *Agricultural Sciences* (AS), *Engineering Technology* (ET), *Natural Sciences* (NS), *Medical and Health Sciences* (MHS), *Humanities* (H), and *Social Sciences* (SS)). Note that some publications are assigned to multiple fields. In the aggregate analysis, we use the first field to which a publication was assigned, but in a single field view, we take publications with any assignment in the given field; thus, interdisciplinary publications are covered separately in all their assigned fields.

As described earlier, scientific organizations set goals and define strategic paths to ensure a unique research profile and identity. To have a better understanding of how BUA members introduce their own research goals and main areas, we use their self-representations in Berlin University Alliance et al. (2018, 2019). We expect to observe the prevailing roles of these institutions in the structure of scientific collaborations of the fields closer to their primary areas of focus. For **FU** this includes: "Biomedical Foundations," "Complex Systems," "Cultural Dynamics," "Educational Processes and Results," "Health and Quality of Life," "Human-Environmental Interactions," "In-Security and Security Research," "Materials Research," and "Transregional Relations." For **HU**: "Application-Oriented Mathematics," "Image Sciences," "Integrative Life Sciences," "Integrative Natural Sciences," "Research on Law and Society," "Study of Ancient Civilizations," and "Sustainability Research." For **CH**: "Cardiovascular Research & Metabolism," "Infection, Immunology & Inflammation," "Neuroscience," "Oncology," "Rare Disease & Genetics," and "Regenerative Therapies." For **TU**: "Materials, Design and Manufacturing," "Digital Transformation," "Energy Systems, Mobility and Sustainable Resources," "Urban and Environmental Systems," "Optic and Photonic Systems," and "Education and Human Health." Aside from CH, which has a focus on MHS and NS, the other three institutions are active in areas close to major OECD fields.

[4] Kompetenzzentrum Bibliometrie (KB), https://bibliometrie.info

[5] All Science Journal Classification

3.1. Organization Name Disambiguation

The data delivered by Scopus are not perfect. They are prone to error and there is a strong need for *disambiguation* of organization names (Donner et al., 2019). Without disambiguation, the coauthorship networks constructed will have multiple representations of the same actor and an artificially higher level of (dis)connectivity.

We developed two disambiguation techniques (i.e., *OrgNameString* and *OrgNameFuzzy* matching) and compared their results with a previously established technique (Research Organization Registry (ROR)[6]), as depicted in Figure 1. ROR uses data from the Global Research Identifier Database (GRID)[7] prepared by Digital Science[8], ISNI[9], Crossref, and Wikidata[10].

In *OrgNameString* matching (shown in the gray shaded area on the left of Figure 1), we standardize organization names and perform a match with GRID (snapshot of February 17, 2019). We use only the largest entity that the KB extracts from Scopus using the first part of the affiliation string before the first comma. To match, we used *string* comparison methods in Python (for simplicity we call it *OrgNameString*) which matches whole and subsets of the text strings but does not account for changes in the order of words in organization names. To remove the effect of the order of organization name parts, we split the names based on space (i.e., words) and reorder them alphabetically for both Scopus and GRID entities. We then add country names to the end of strings to allow higher precision of matching and reduce the effect of organizational homonyms[11]. For those still nonmatched, we perform another OrgNameString match with scientific organizations in Wikidata and for the still missing ones, with Wikidata entities that have geographical coordinates. We limit the results to those most promising based on a Jaro Winkler distance of more than 0.85 between the two matched names. We do this after the OrgNameString match is done, as a control for reliability. We chose this threshold based on manual evaluation of match results to have the highest accuracy. Finally, to complement the results of the OrgNameString procedure, we search organization names in an in-house database that was previously developed by Rimmert (2018)[12] by comparing organization names to Wikidata entities.

In parallel, we compare organization names with GRID by *OrgNameFuzzy* matching the names. This method takes differing word order and subsets of the name into account (we standardize the names as before and add country). For OrgNameFuzzy text matching, we use the *FuzzyWuzzy*[13] library in Python. Using *fuzz.ratio* as the scorer, we set a threshold of 80%, which was chosen based on empirical evaluation on some exemplar cases and proved to give a reliable accuracy (gray shaded area in center of Figure 1).

In a third attempt, instead of the main organization names used in previous procedures, we used the complete string of affiliation addresses delivered by Scopus to disambiguate it with the Research Organization Registry (ROR) API (see footnote 11 for an example). We obtain further information (i.e., country, geographical coordinates (longitude and latitude)) of the main address and type of

[6] https://ror.org/about

[7] https://www.grid.ac/pages/policies

[8] https://www.digital-science.com/

[9] International Standard Name Identifier, https://isni.org/

[10] https://www.wikidata.org

[11] As an example, from this address string delivered by Scopus, “Freie Universität Berlin, Department of education and psychology, DEU” KB extracts “Freie Universität Berlin” as the first part, which we use for OrgNameString and OrgNameFuzzy matching processes and after removing alphanumeric symbols, lowercasing and reordering alphabetically, we add the 3-digit ISO country code to the end (e.g., “berlin freie universität DEU”).

[12] This in-house data is only accessible through KB infrastructure.

[13] https://github.com/seatgeek/fuzzywuzzy

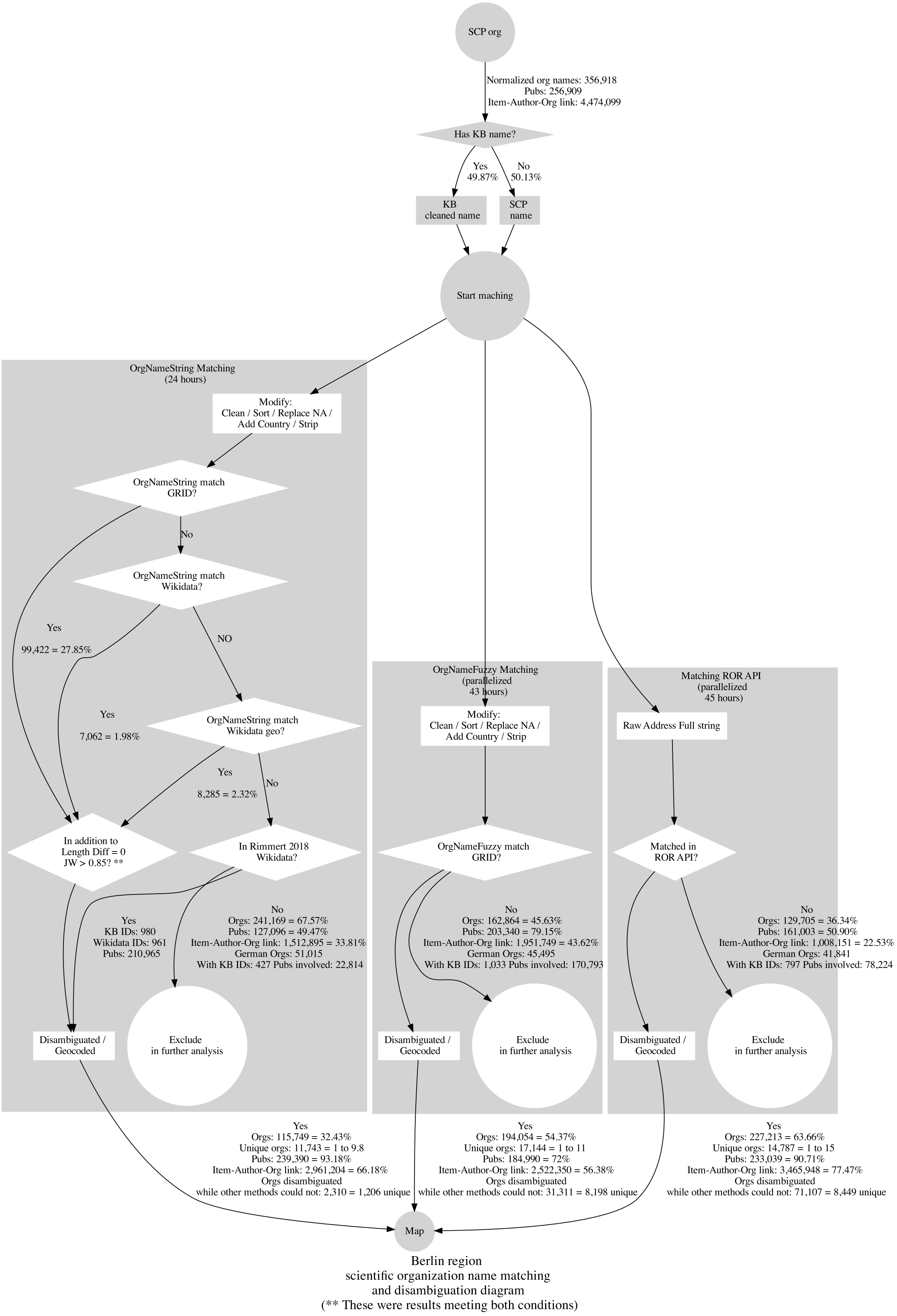

Figure 1. Organization name disambiguation techniques and comparison of coverage and accuracy.

organization as *education*, *non-profit*[14], *company*, *government*, *health-care*, *facility*[15], *archive*,[16] and others from ROR). We used the ROR snapshot from November 7, 2019. This disambiguation takes different name spellings and misspelled words, acronyms, and multiple languages into account. In order not to face API request limits, we have set up a local instance of the ROR API (see gray shaded area on the right of Figure 1).

To evaluate the quality of ROR disambiguation results, we chose two random samples of 100 organization names from Scopus data. A research assistant searched online for them and the disambiguated name from ROR to find the original address and compared the two names to determine whether the match was correct. Through this process, we identified false matches in 4% of one sample and 8% of the other sample, indicating that 96% and 92% of cases were reliably matched.

To highlight the importance and effect of the disambiguation, we define different scenarios based on nondisambiguated and disambiguated data. We present the results and implications of each scenario in construction of the coauthorship networks. However, on the basis of the accuracy and coverage results, we use the third disambiguation technique described above (i.e., ROR) for our field, geographical, and sector analysis. It is important to note that in disambiguated versions of the data, we include only publications for which all contributing organizations are disambiguated and we exclude those publications with one or more nondisambiguated coauthoring organizations. Note that due to the shared medical faculty of HU and FU from 2003, affiliations from Charité in most cases included either HU or FU (or both). We searched each individual affiliation string; if Charité was mentioned, we assigned these to Charité. ROR favored the affiliation appearing first and Charité was not always the first mentioned, thus this manual correction was necessary to identify Charité's publications.

3.2. Bipartite Network Modeling

We construct bipartite coauthorship networks (Breiger, 1974) using ties between publications and *organizations* (Katz & Martin, 1997). We treat each single publication as an event where organizations interact to produce an academic text (Biancani & McFarland, 2013). Studies on coauthorship networks usually use a one-mode projection of these bipartite networks (Newman, 2001a, 2001b). The problem with this projection is twofold. Different structures in two-mode networks are projected to the same one-mode structure, which causes an information loss about the underlying structure. Second, the one-mode projection can present an artificially higher density and connectivity due to publications with a high number of authors which project to maximally connected cliques. By adopting methods specifically developed for *bipartite networks* we are able to resolve the shortcomings. However, these methods are scarce.

To identify possible geographical, field and/or sector based coalitions between scientific organizations, we extract the largest connected component of the network, i.e., giant component, and investigate it further. Our aim is to see if there are cohesive subgroups of organizations preferentially collaborating *among* themselves. We investigate the potential underlying factors behind these groupings.

[14] Organizations that use their surplus revenue to achieve their goals. They include charities and other non-government research funding bodies. Example, the Max Planck Society (grid.4372.2)

[15] A building or facility dedicated to research of a specific area, usually containing specialized equipment. Includes telescopes, observatories and particle accelerators. Example: member institutes of the Max Planck Society (e.g., Max Planck Institute for Demographic Research, grid.419511.9)

[16] Repository of documents, artifacts, or specimens. Includes libraries and museums that are not part of a university. Example, New York Public Library (grid.429888.7)

To identify communities of coauthorship, we use *bipartite community detection* by *Constant Potts model* (CPM). CPM is a specific version of the Potts model (Reichardt & Bornholdt, 2004) proposed by Traag, Van Dooren, and Nesterov (2011) as a *resolution-limit-free* method. It resolves the resolution limit problem in modularity (Newman, 2004) which can obstruct detection of small communities in large networks (Traag, Waltman, & van Eck, 2019). We use the implementation in the *Leidenalg*[17] library in Python. Community detection emphasizes the importance of links *within* communities rather than those *between* them. CPM uses a resolution parameter γ (i.e., "*constant*" in the name), leading to communities such that the link density between the communities (external density) is lower than γ and the link density within communities (internal density) is more than γ. We set the resolution parameter for all networks (e.g., aggregate data (ROR) and scientific fields) to 6×10^{-3}. To ensure replicability of results, we use a seed. We chose this γ after exploration of the number of communities detected in contrast to the number of organizations and publications included in each bipartite community to arrive at a rather consistent distribution.

4. RESULTS

4.1. Implications of Organization Name Disambiguation

Figure 1 presents the different disambiguation techniques used and the coverage of publication-author-organization links (*OrgNameString* 66.18%, *OrgNameFuzzy* 56.38%, and *ROR* 77.47%). We present author-level counts of links here, as different authors from the same institution might mention different affiliation addresses by including department names or they might report erroneous addresses. However, in building organization level coauthorship networks, we exclude repeated organization-publication links.

Each technique successfully disambiguates a set of unique organization names that other techniques are unable to disambiguate (OrgNameString 1,206, OrgNameFuzzy 8,198 and ROR 8,449). Note that not all organizations involved in authoring a publication are successfully disambiguated: OrgNameString identified 115,749 (32.43%) organizations from 239,390 (93.18%) publications, OrgNameFuzzy identified 194,054 (54.37%) organizations from 184,990 (72%) publications, and ROR identified 227,213 (63.66%) organizations from 233,039 (90.71%) publications. We only include publications for which all contributing organizations are successfully disambiguated and this decreases our coverage to 129,813 (51%) publications in OrgNameString, 53,569 (21%) in OrgNameFuzzy, and 126,130 (49%) in ROR in favor of higher accuracy of results.

Table 1 compares the networks constructed using nondisambiguated data with the output of different disambiguation techniques. We observed a high rate of disconnectivity in the nondisambiguated network (10,269 components, including many organizations) while this was extremely reduced through disambiguation techniques (i.e., to 66 in OrgNameString, 159 in OrgNameFuzzy, and 100 in ROR). The share of nodes in the giant component, which was initially high in the nondisambiguated network (95%), further increased and covered close to 99% in all cases (numbers in the table are rounded up). In all these cases, disambiguation shows that many unique organization names delivered by Scopus need to be merged due to spelling errors and name order changes, which can affect the networks constructed to a high degree (see De Stefano, Fuccella et al. (2013) for a discussion of possible effects). In OrgNameString, the ratio of disambiguated to nondisambiguated unique organizations was 1 to 9.8 (in OrgNameFuzzy 1 to 11 and in ROR 1 to 15). This proves the high influence that disambiguation has on the results.

Table 2 presents the networks in different OECD scientific fields using ROR results. Note that the results which follow are based on the 49% of publications for which all contributing

[17] https://github.com/vtraag/leidenalg

Table 1. Berlin organizations' coauthorship networks using nondisambiguated and disambiguated data (G = giant component)

Metrics	Non disambiguated	OrgNameString	OrgNameFuzzy	ROR
No. of connected components	10,269	66	159	100
No. of biparitite nodes	613,827	135,057	58,547	133,387
No. of biparitite edges	1,083,775	246,704	89,199	246,472
% of biparitite nodes in G	95	100	99	100
% of biparitite edges in G	98	100	100	100
No. of organizations	**356,918**	**5,244**	**4,978**	**7,257**
No. of organizations in G	337,755	5,176	4,809	7,153
No. of publications (%)	**256,909**	**129,813 (51%)**	**53,569 (21%)**	**126,130 (49%)**
No. of publications in G	245,203	129,657	53,248	125,949

organization names were successfully disambiguated by the ROR technique. Each of the fields presented in Table 2 covers a different share of connected components observed in the aggregate data, ranging from 13 components in AS to 68 in Social Sciences (SS). NS has the highest number of both publications and organizations, while Humanities (H) has the smallest number of publications and organizations.

4.2. Macro View of Scientific Output of the Berlin Metropolitan Region

Figure 2 presents the raw and fractional count of publications among different OECD fields. Note that it is based on publications that have at least one collaborator from the Berlin metropolitan region and for organizations that were successfully disambiguated with the ROR technique. Nevertheless, the trends are in line with what Stephen et al. (2020) and Stahlschmidt et al. (2019) reported for Germany. Some fields show higher rates of collaborative work (e.g., see the case of NS, blue lines, first and second from top) which is evident in the gap between the lines presenting their raw and fractional counts and is in line with Shrum et al. (2007, p. 3)'s report of

Table 2. Berlin organizations' coauthorship networks in different OECD scientific fields (G = giant component, ROR organization name disambiguation)

Metrics	AS	ET	H	MHS	NS	SS
No. of connected components	13	56	34	48	55	68
No. of biparitite nodes	8,528	33,930	4,835	46,842	89,032	16,045
No. of biparitite edges	13,822	57,671	6,195	84,945	170,334	25,336
% of biparitite nodes in G	100	100	98	100	100	99
% of biparitite edges in G	100	100	99	100	100	100
No. of organizations	**1,687**	**2,991**	**798**	**3,843**	**5,970**	**2,091**
No. of organizations in G	1,668	2,933	763	3,792	5,910	2,012
No. of publications	**6,841**	**30,939**	**4,037**	**42,999**	**83,062**	**13,954**
No. of publications in G	6,828	30,851	3,988	42,913	82,989	13,849

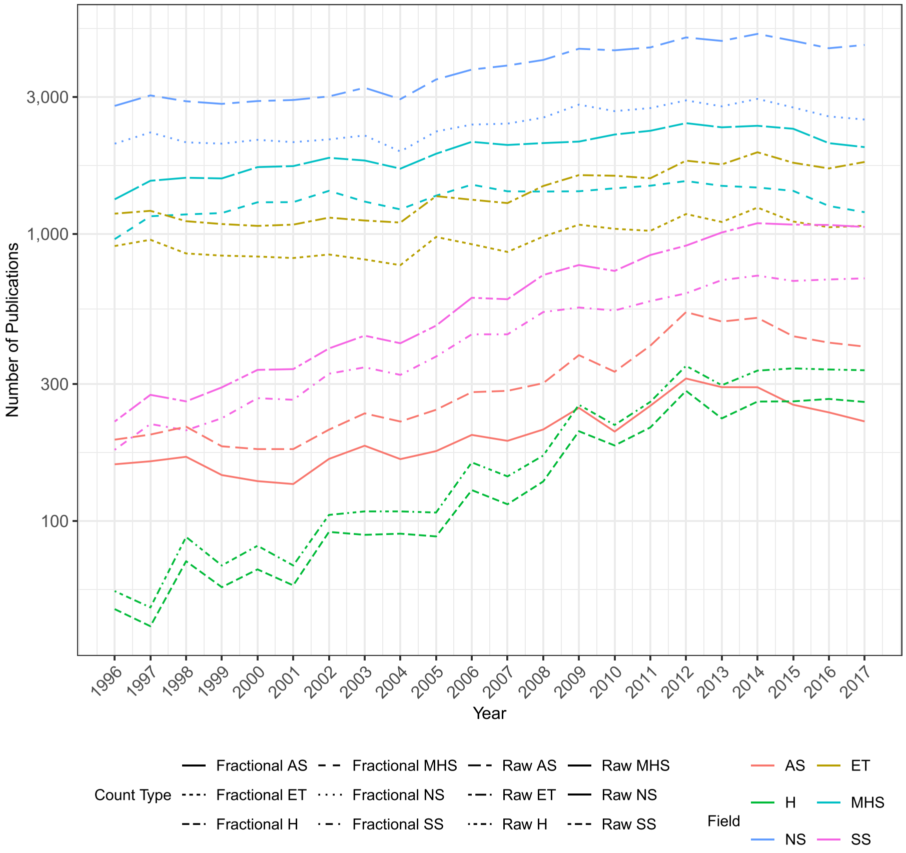

Figure 2. Raw and fractional count of Berlin publications by OECD fields (1996–2017, Scopus, fractional count based on organizations, *y*-axis on log scale).

physical sciences. In contrast, some fields that are traditionally known to be less collaborative (Leahey, 2016) present a smaller gap on the plot (e.g., Humanities). As the *y*-axis is on a log 10 scale, the figure shows that the growth over time in the raw and fractional count of publications has tripled in the case of Humanities and Social Sciences. However, this could be influenced by the higher coverage of Scopus in recent years and not merely an increase in publications (see Stahlschmidt et al. (2019) for a discussion). To investigate the internationalization of publications, Figure 3 presents the *single* (intra-DEU) versus *multiple* country publications. The trends observed are in line with the case of the German science system reported in Stahlschmidt et al. (2019). It is clear that some fields have already reached close to 50% of their publications involving *internationalization* (i.e., NS), driving the aggregate trend of increasing internationalization observed in the top panel of the figure. However, there are other fields with still lower than 25% of publications involving international collaboration (i.e., H and SS). In all fields except H and SS, an increasing trend towards further internationalization is evident (see the increasing length of black bars in the figure), while H and SS do not present a clear increasing trend and in some years the rate of internationalization has decreased. This answers our **RQ1** and **RQ2**, signaling large differences between fields in the rate of collaborative work and internationalization.

To investigate our **RQ3**, we focus on organization sectors. In total, out of 7,257 unique organizations, which includes Berlin organizations and their worldwide collaborators, there were 2,844 organizations from the education sector, 1,667 facility, 860 health-care, 587 company, 436 nonprofit, 429 government, 282 other, 124 archive, and 28 with missing sectors.

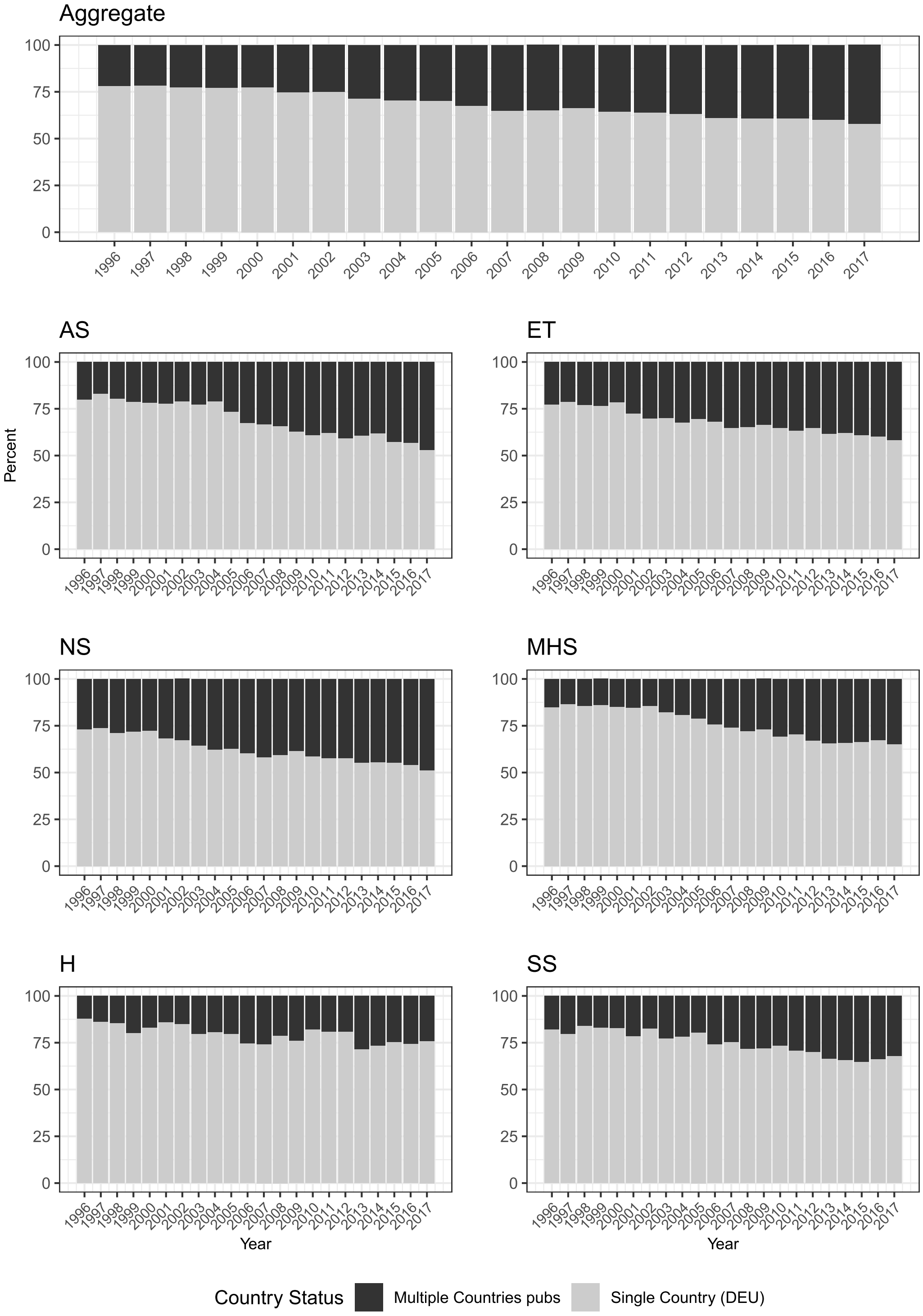

Figure 3. Share of intra-Germany versus multiple country coauthorship: (Top) aggregate; (Bottom) different fields (1996–2017, Scopus).

Table 3 presents the distribution of organizations in different sectors in the five countries with the highest numbers of organizations (i.e., China, Germany, France, the United Kingdom, and the United States, in alphabetical order of ISO codes). While education is the sector with the highest number of organizations in four countries, facility has the highest number of

Table 3. Five countries with the highest number of organizations by sector (GRID data based on Berlin sample 1996–2017, percentages calculated for each country)

Country code	Organization sector	Count	%
CHN	Education	193	56
	Facility	77	22
	Healthcare	39	11
	Government	18	5
	Company	7	2
	Nonprofit	5	1
	Other	3	1
	Archive	1	0
DEU	Facility	319	27
	Education	215	18
	Company	205	17
	Healthcare	115	10
	Nonprofit	113	10
	Other	113	10
	Government	66	6
	Archive	35	3
FRA	Facility	261	51
	Education	114	22
	Healthcare	45	9
	Government	35	7
	Company	34	7
	Other	15	3
	Nonprofit	8	2
	Archive	4	1
GBR	Education	114	32
	Healthcare	91	25
	Facility	46	13
	Company	34	9
	Nonprofit	26	7
	Government	25	7
	Other	16	4
	Archive	7	2

Table 3. (*continued*)

Country code	Organization sector	Count	%
USA	Education	423	42
	Healthcare	157	15
	Company	129	13
	Facility	110	11
	Nonprofit	95	9
	Government	44	4
	Other	34	3
	Archive	23	2

organizations in Germany, which could be an artifact of the disambiguation and exclusion of publications with nondisambiguated organizations. Figure 4 presents the geographical distribution of organizations worldwide separated by sectors and aggregated in countries. Darker colors show higher numbers of organizations in a given country. It is clear that most countries have organizations in the education sector. Another evident pattern is that more developed countries (e.g., in Western Europe, North America, and Oceania) have representatives in all sectors, which signals the higher sectoral diversity of the science systems of these countries. However, China and India are two specific cases outside the previously mentioned regions with representation in many sectors. The distribution of companies is another interesting observation, where many countries do not have any representatives, in contrast to the education sector. This answers our **RQ3** that there is a high diveresity of sectors collaborating with the Berlin region, with variation depending on the specific country.

4.3. Structure of Institutional Scientific Collaborations in Berlin

We focus now on communities of coauthorship identified from the giant component using bipartite community detection (**RQ4** and **RQ5**). This enables us to go beyond the macro descriptive view presented thus far and investigate the structure of collaborations at the individual publication level. Note that these are communities detected from the giant component, which is connected in itself; however, these communities are the denser areas of the collaboration network. We are interested to know what could be the underlying factors behind these higher densities leading to these cohesive subgroups.

Figure 5 presents the distribution of communities based on the number of organizations in each community and the aggregate number of publications of all organizations in a given community (each community is represented by one dot in the figure). The clearest observation in this figure is the field-based collaboration patterns among BUA members indicated by the shape and color of dots, where green triangles show the presence of one or more BUA member(s) in a given community.

In most cases, there is a divide between BUA members. In the aggregate view in the top panel, and in the NS, H, and SS fields, we see three BUA members populating the most prolific community (FU, HU, and CH) and one BUA member located in the second most prolific community (TU). In AS, HU and TU are in the most prolific community, i.e., 0, and CH and FU are in

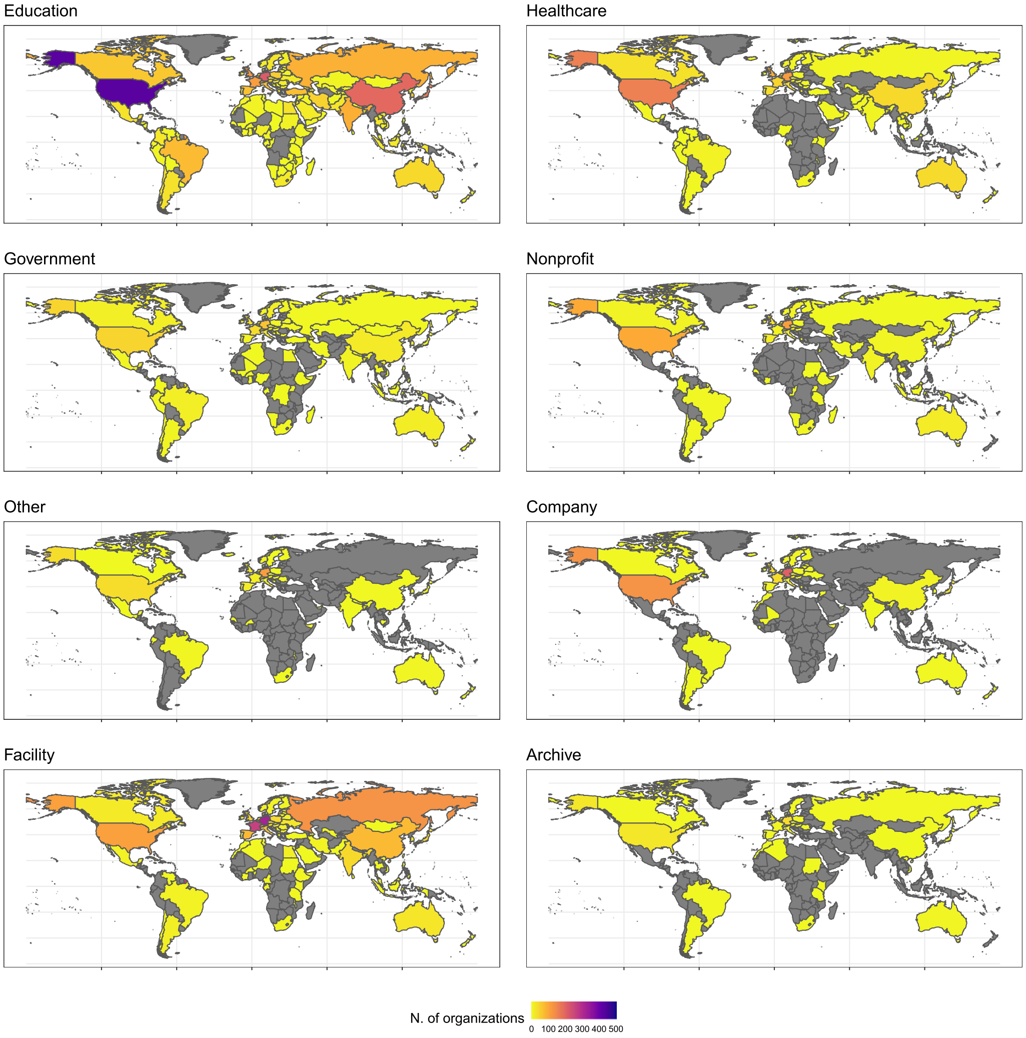

Figure 4. Countries worldwide collaborating with Berlin region by sector (color: N. of organizations (ranges from 1 to 423). If a country does not have presence in a sector, it is shown with gray color.

community 1. In the case of ET, TU is located in the most prolific community (community 0) and HU, CH, and FU are in the second most prolific community (i.e., 1). This could be due to the fact that TU, being a technical university, pursues more technical and application-oriented research.

MHS is the only case where all four BUA members are present in *a single community*, which is perhaps due to the closer cooperation among them that was formed through a shared faculty by HU and FU located in Charité (CH) in 2003. They have been successful in integrating TU into the collaboration structure of MHS.

Overall, BUA needs to integrate TU further into the structure of scientific collaborations in the region through shared projects or organizational forms. Furthermore, it is clear from aggregate and field views that not all communities are populated with the most prolific organizations (in terms of number of publications). There are communities of different sizes consisting of organizations with different levels of productivity (e.g., see the gray circles). More detail on the sectoral and geographical composition of the communities is presented in Table A1 in the Appendix.

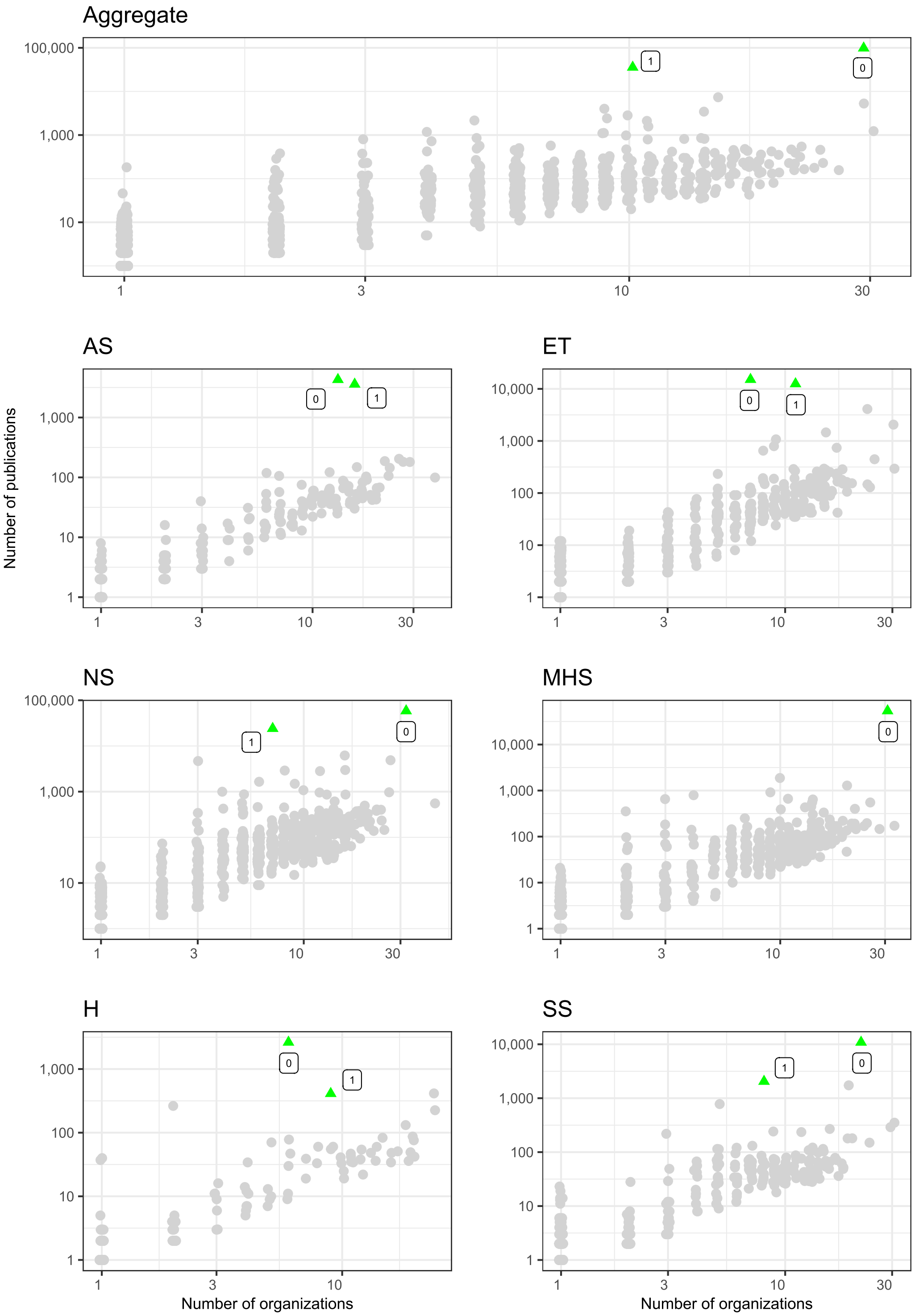

Figure 5. Organizations in communities of the giant component vs. publication (label: name of community; green triangle: includes BUA member(s); gray circle: other communities; *x* and *y* axes on log scale).

5. DISCUSSION

At first sight, and based only on descriptive analysis, we observed a highly collaborative science landscape in the Berlin region. Some fields present a high degree of difference between the *raw* and *fractional* counts of publications. Despite the prevalence of collaborative works and increasing trend towards internationalization in aggregate view, we observed that some fields (e.g., NS and AS) have more internationalized collaborations while other fields (e.g., MHS, Humanities, and Social Sciences) are less internationalized or they did not present a steady upward trend which is in line with observations by Moed et al. (1991) and Babchuk et al. (1999).

Berlin presents a specific case. It is similar to a science hub with a diverse sectoral composition of organizations, which is in line with Balland et al. (2020)'s observation in metropolitan regions in the United States and Rammer et al. (2020)'s observation of the Berlin metropolitan region. However, this could be due to our data gathering strategy, where only publications with at least one organization located in Berlin are included. Thus, there could be other collaborations between the partners, excluding Berlin organizations, that we do not cover here. Some countries present a highly diverse science system consisting of a wide range of sectors among those collaborating with Berlin organizations. However, in most countries, *education* is the prevailing sector where scientific publications are produced, which is not counterintuitive.

We modeled the scientific collaborations through *bipartite* coauthorship networks, which treat each scientific publication as an event where organizations interact in producing scientific texts. Our bipartite community detection configuration was helpful in detecting the diverse composition of organizational teams contributing to scientific publications, which could be overlooked if the network is projected to one mode due to artificially high cliquish behavior. We observed that in most fields, with the exception of Humanities and Agricultural Sciences, the most prolific communities were comprised of organizations located in the Berlin metropolitan region and they were collaborating either within Berlin or with other German organizations or exclusively with European organizations. There were of course internationalized communities in all fields, but they were not highly prolific.

Looking at the members of the BUA and their position in these cohesive subgroups presented interesting findings. Only in the MHS, which was dominated by the high productivity of Berlin and Germany, were the four BUA members located in *one* community and collaborated densely. This is likely an outcome from the efforts of the HU and FU in 2003 to jointly establish an MHS faculty, located in Charité's facilities, and our findings show that this strategic coalition has been successful in integrating other organizations from Berlin, such as TU.

Furthermore, some of the observed field division between BUA members could be remnants of the east-west division in Germany and the reorganization of research profiles and mutually exclusive definitions of areas of focus to reduce parallel work and competition that happened after the reunification. This divide was recently observed in the biotechnology field by Abbasiharofteh and Broekel (2020). TU presents a specific case and in most cases it is member of a separate community. In ET, TU's collaboration network is dominated by other Berlin and German organizations. It shows that the BUA needs to develop further strategic cooperations among the members to ensure higher integration, similar to the case of MHS. However, this might be due to the fact that we included *conference proceedings*, which is a specific publication type preferred more by the technical universities. As TU is the main technical university in our sample, the collaboration structure reflected in this document type may have affected the observed results and overinflated the divide between TU and other BUA members.

5.1. Limitations

Our paper has certain limitations. When we construct the coauthorship network at the organization level, we naturally overlook the changes that happen in the composition of researchers affiliated to those organizations. The same organization could have a highly different composition of members over time that affects the type of research carried out and collaboration ties formed. In addition, the number of available algorithms for bipartite community detection limits our possiblities to provide comparative results. Nevertheless, there are recent developments in that direction (Rossetti, Milli, & Cazabet, 2019; Taguchi, Murata, & Liu, 2020).

We use only Scopus as the main database, and although it covers documents written in German, it is dominated by English-language publications. Bibliometric databases, including Scopus, are regularly updated which can affect the temporal trends we observe here. In addition, each bibliometric database covers a specific set of scientific publications (see Stahlschmidt et al. (2019) for a comparison between WoS and Scopus), despite the similarities, there are differences in philosophies and approaches to what should be indexed (Huang et al., 2020). Furthermore, we were unable to disambiguate all the organizations in our sample, which led to excluding 51% of the publications that had one or more nondisambiguated organizations. Thus, while our results follow the general trends observed in the German scientific system (see Stahlschmidt et al. (2019), Stephen et al. (2020), and Aman (2016)), the specificities observed in the structure of scientific collaborations among BUA members and the international collaborations could be highly affected if we had higher coverage in the disambiguation techniques.

Another limitation of our data, and of research in general at the organization level, is the *superstar* researchers with multiple affiliations. We assume that these researchers have received resources from each of these multiple organizations. Thus, we consider these researchers as bridges between these organizations; however, in the networks constructed, these cases might appear as an international collaboration when really it is a single author affiliated with multiple countries. High-quality data with disambiguated records of publications at the author level would allow a more complete investigation of such cases. Another limitation of our study could be that our disambiguation techniques penalize countries using languages other than English or lesser known organizations that are usually less prolific. These organizations can be more prevalent among those we excluded from our analysis because our techniques likely did not accurately disambiguate them. In addition, different disambiguation techniques are more or less effective in identifying differing sets of organizations and any choice of technique would have implications for a subset of the organizations while penalizing another subset.

We do not have any insight into the background of individual researchers affiliated with these scientific organizations. We do not know about the motivations that form and drive the scientific collaborations and observed trends (Katz & Martin, 1997; Shrum et al., 2007, pp. 7–8; Subramanyam, 1983, 202, 209). In addition, we assume that scientific collaborations are positive interactions among collaborators that successfullly led to one or more scientific publications, which simply is not always the case and conflict arises between partners that needs to be resolved for a collaboration to proceed (Shrum et al., 2007, pp. 197–198, 202). As an example, for all fields, we observed small communities that were leaning more towards internationalized collaborations. These might be groups mainly consisting of migrant scientists who collaborate with their former scientific organizations or they may play a "boundary spanning role" among regional, national, and continental contexts. We cannot investigate these type of questions at the organization level.

Furthermore, our definition of the Berlin metropolitan region was based on the affiliation addresses, while the literature on science geography presents a diverse array of definitions (e.g., Abbasiharofteh & Broekel, 2020; Cottineau, Finance et al., 2019), from NUTS level to areas covering multiple cities, which are overlooked in our data gathering strategy.

6. CONCLUSION

We provide a quantitative, exploratory, and macro view of the structure of scientific collaborations in the Berlin metropolitan region. We chose Berlin because of its history of division and reunification. This region has undergone eras of organizational self-isolation to mutually exclusive definition of organizational research profiles and has a specific position in the German science system. Our main level of analysis was scientific *organization* (which can be academic or nonacademic organization or firm) and we investigated the share of collaborative work and internationalized work versus single country collaborations. We covered all OECD scientific fields and presented a comparative view of the similarities and differences of collaborations in these fields. By adopting a global, regional, and organization based approach, we tried to put the empirical results into different contexts.

In methodological terms, we developed two organization name disambiguation techniques (i.e., OrgNameString and OrgNameFuzzy matching) and compared their performance and coverage with an established technique (i.e., Research Organization Registry [ROR]). We presented the high impact that organization name disambiguation could have on the constructed collaborations networks and how it can bias measures and trends. We had to exclude 51% of the publications that had one or more nondisambiguated organizations to limit our analysis to successfully disambiguated cases. Considering bibliometric databases as ground truth could have implications for the results of similar studies of scientific collaborations. Future research needs to carefully consider disambiguation and provide transparent details about the disambiguation procedures followed and accuracy obtained.

We conclude that mixing macro and global views while keeping regional, national, and continental granularity can help in describing the observed quantitative trends. It is necessary to move beyond the macro descriptive view presented based on yearly publication counts or increasing trends of team science. As our investigation proved, not all members of the community are moving towards internationalization and some parts of the community, which are normally those that are the most prolific, prevail and distort the aggregate images.

To provide suggestions for the Berlin region and other regions worldwide, we borrow Shrum et al. (2007)'s conceptualization of *bureaucracy* and *technology*. To present further implications for the structure of scientific collaborations, let us here return to the specific case of Medical and Health Sciences (MHS) in the Berlin metropolitan region. The case of MHS evidently showed that a new organizational form (i.e., a shared faculty with defined bureaucratic procedures) can help to not only bridge the divide between coalition members but also form an expanded collaboration network worldwide. BUA members (as one example studied here that can be extended to other strategic coalitions formed worldwide) seem to be trying to cooperate while preserving their distinct organizational identities, which is similar to the case of geophysics discussed in Shrum et al. (2007, p. 200), where a highly formal structure prevented the collaboration becoming an extension to any of the member organizations. It is also similar to the case of materials sciences (Shrum et al., 2007, p. 201), where a more bureaucratic structure allowed higher autonomy for the collaboration partners from diverse sectors. Therefore, clear bureaucratic procedures can even enable brokered collaboration among previous competitors (Shrum et al., 2007, p. 192). Although scientific collaboration cannot become a goal in itself (e.g., Shrum et al., 2007, p. 202), BUA and similar coalitions can use this conceptualization of bureaucracy to develop clear objectives for scientific collaboration and, while preserving the distinctive profiles of the members, benchmark from the successful case of MHS to foster further intraregional and global collaborations. The resources of a single organization are limited and multiorganizational collaborations can enable higher achievements (e.g., Shrum et al., 2007, pp. 20, 119; Aman, 2016) and increase the diversity of researchers, which in turn can facilitate better science (Nielsen, Alegria et al., 2017). Nevertheless, multiorganizational scientific collaborations are complex and need a careful design to be successful.

ACKNOWLEDGMENTS

We would like to thank Sybille Hinze, Martin Reinhart, Paul Donner, Melike Janßen, and Dimity Stephen for comments and suggestions on earlier versions of this paper.

FUNDING INFORMATION

This research was done in the DEKiF project supported by the Federal Ministry for Education and Research (BMBF), Germany, grant number M527600. Data are obtained from Kompetenzzentrum Bibliometrie (Competence Center for Bibliometrics), Germany, which is funded by BMBF, grant number 01PQ17001.

DATA AVAILABILITY

The data cannot be made publicly available due to the licensing and contract terms of the original data. Python scripts to replicate the organization name disambiguation techniques and bipartite community detection may be publicly accessed at https://doi.org/10.5281/zenodo.4657325.

COMPETING INTERESTS

The author has no competing interests.

REFERENCES

Abbasiharofteh, M., & Broekel, T. (2020). Still in the shadow of the wall? The case of the Berlin biotechnology cluster. *Environment and Planning A: Economy and Space*, *53*(1), 73–94. https://doi.org/10.1177/0308518X20933904

Akbaritabar, A., Casnici, N., & Squazzoni, F. (2018). The conundrum of research productivity: A study on sociologists in Italy. *Scientometrics*, *114*(3), 859–882. https://doi.org/10.1007/s11192-017-2606-5

Akbaritabar, A., & Squazzoni, F. (2020). Gender patterns of publication in top sociological journals. *Science, Technology, & Human Values*, *46*(3), 555–576. https://doi.org/10.1177/0162243920941588

Akbaritabar, A., Traag, V. A., Caimo, A., & Squazzoni, F. (2020). Italian sociologists: A community of disconnected groups. *Scientometrics*, *124*, 2361–2382. https://doi.org/10.1007/s11192-020-03555-w

Aman, V. (2016). How collaboration impacts citation flows within the German science system. *Scientometrics*, *109*(3), 2195–2216. https://doi.org/10.1007/s11192-016-2092-1

Aman, V. (2018). Does the Scopus author ID suffice to track scientific international mobility? A case study based on Leibniz laureates. *Scientometrics*, *117*(2), 705–720. https://doi.org/10.1007/s11192-018-2895-3

Araújo, E. B., Araújo, N. A. M., Moreira, A. A., Herrmann, H. J., & Andrade, J. S. (2017). Gender differences in scientific collaborations: Women are more egalitarian than men. *PLOS ONE*, *12*(5), e0176791. https://doi.org/10.1371/journal.pone.0176791, PubMed: 28489872

Avdeev, S. (2019). International collaboration in higher education research: A gravity model approach. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3505886

Babchuk, N., Keith, B., & Peters, G. (1999). Collaboration in sociology and other scientific disciplines: A comparative trend analysis of scholarship in the social, physical, and mathematical sciences. *The American Sociologist*, *30*(3), 5–21. https://doi.org/10.1007/s12108-999-1007-5

Balland, P.-A., Jara-Figueroa, C., Petralia, S. G., Steijn, M. P. A., Rigby, D. L., & Hidalgo, C. A. (2020). Complex economic activities concentrate in large cities. *Nature Human Behaviour*, *4*(3), 248–254. https://doi.org/10.1038/s41562-019-0803-3, PubMed: 31932688

Berlin University Alliance, Humboldt-Universität zu Berlin, Technische Universität Berlin, & Charité Universitätsmedizin Berlin. (2018, February). Gemeinsam im Verbund (Together as a group). Berlin University Alliance. https://www.berlin-university-alliance.de/excellence-strategy/universities-of-excellence/index.html.

Berlin University Alliance, Humboldt-Universität zu Berlin, Technische Universität Berlin, & Charité Universitätsmedizin Berlin. (2019). *Berlin University Alliance Proposal Crossing Boundaries toward an Integrated Research Environment*. Berlin: Berlin University Alliance.

Biancani, S., & McFarland, D. A. (2013). Social networks research in higher education. In *Higher education: Handbook of theory and research* (pp. 151–215). Berlin: Springer. https://doi.org/10.1007/978-94-007-5836-0_4

Blume, S., Bunders, J., Leydesdorff, L., & Whitley, R. (Eds.). (1987). *The social direction of the public sciences*. Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-009-3755-0

Breiger, R. L. (1974). The duality of persons and groups. *Social Forces*, *53*(2), 181–190. https://doi.org/10.1093/sf/53.2.181

Cottineau, C., Finance, O., Hatna, E., Arcaute, E., & Batty, M. (2019). Defining urban agglomerations to detect agglomeration economies. *Environment and Planning B: Urban Analytics and City Science*, *46*(9), 1611–1626. https://doi.org/10.1177/2399808318755146

D'Angelo, C. A., & van Eck, N. J. (2020). Collecting large-scale publication data at the level of individual researchers: A practical proposal for author name disambiguation. *Scientometrics*, *123*(2), 883–907. https://doi.org/10.1007/s11192-020-03410-y

De Stefano, D., Fuccella, V., Vitale, M. P., & Zaccarin, S. (2013). The use of different data sources in the analysis of coauthorship networks and scientific performance. *Social Networks*, *35*(3), 370–381. https://doi.org/10.1016/j.socnet.2013.04.004

Donner, P., Rimmert, C., & van Eck, N. J. (2019). Comparing institutional-level bibliometric research performance indicator values based on different affiliation disambiguation systems. *Quantitative Science Studies*, *1*(1), 150–170. https://doi.org/10.1162/qss_a_00013

Fox, M. F. (1983). Publication productivity among scientists: A critical review. *Social Studies of Science*, *13*(2), 285–305. https://doi.org/10.1177/030631283013002005

Glänzel, W., Schubert, A., & Czerwon, H.-J. (1999). A bibliometric analysis of international scientific cooperation of the European Union (1985). *Scientometrics*, *45*(2), 185–202. https://doi.org/10.1007/BF02458432

Hoekman, J., Frenken, K., & Tijssen, R. J. W. (2010). Research collaboration at a distance: Changing spatial patterns of scientific collaboration within Europe. *Research Policy*, *39*(5), 662–673. https://doi.org/10.1016/j.respol.2010.01.012

Huang, C.-K., Neylon, C., Brookes-Kenworthy, C., Hosking, R., Montgomery, L., … Ozaygen, A. (2020). Comparison of bibliographic data sources: Implications for the robustness of university rankings. *Quantitative Science Studies*, *1*(2), 445–478. https://doi.org/10.1162/qss_a_00031

Katz, J. S. (1994). Geographical proximity and scientific collaboration. *Scientometrics*, *31*(1), 31–43. https://doi.org/10.1007/BF02018100

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

Laudel, G. (2002). What do we measure by coauthorships? *Research Evaluation*, *11*(1), 3–15. https://doi.org/10.3152/147154402781776961

Leahey, E. (2016). From sole investigator to team scientist: Trends in the practice and study of research collaboration. *Annual Review of Sociology*, *42*(1), 81–100. https://doi.org/10.1146/annurev-soc-081715-074219

Leone Sciabolazza, V., Vacca, R., Kennelly Okraku, T., & McCarty, C. (2017). Detecting and analyzing research communities in longitudinal scientific networks. *PLOS ONE*, *12*(8), e0182516. https://doi.org/10.1371/journal.pone.0182516, PubMed: 28797047

Luukkonen, T., Persson, O., & Sivertsen, G. (1992). Understanding patterns of international scientific collaboration. *Science, Technology, & Human Values*, *17*(1), 101–126. https://doi.org/10.1177/016224399201700106

Moed, H. F., De Bruin, R. E., Nederhof, A. J., & Tijssen, R. J. W. (1991). International scientific cooperation and awareness within the European community: Problems and perspectives. *Scientometrics*, *21*(3), 291–311. https://doi.org/10.1007/BF02093972

Nederhof, A. J. (2006). Bibliometric monitoring of research performance in the Social Sciences and the Humanities: A review. *Scientometrics*, *66*(1), 81–100. https://doi.org/10.1007/s11192-006-0007-2

Newman, M. E. J. (2001a). Scientific collaboration networks. I. Network construction and fundamental results. *Physical Review E*, *64*(1), 016131. https://doi.org/10.1103/PhysRevE.64.016131, PubMed: 11461355

Newman, M. E. J. (2001b). Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. *Physical Review E*, *64*(1), 016132. https://doi.org/10.1103/PhysRevE.64.016132, PubMed: 11461356

Newman, M. E. J. (2004). Detecting community structure in networks. *The European Physical Journal B – Condensed Matter*, *38*(2), 321–330. https://doi.org/10.1140/epjb/e2004-00124-y

Nielsen, M. W., Alegria, S., Börjeson, L., Etzkowitz, H., Falk-Krzesinski, H. J., … Schiebinger, L. (2017). Opinion: Gender diversity leads to better science. *Proceedings of the National Academy of Sciences*, *114*(8), 1740–1742. https://doi.org/10.1073/pnas.1700616114, PubMed: 28228604

OECD. (2007). Revised Field of Science and Technology (FOS) classification in the Frascati Manual (Classification, Field of science and technology classification, FOS, Frascati, Methodology, Research and development). https://www.oecd.org/science/inno/38235147.pdf.

Palla, G., Barabási, A.-L., & Vicsek, T. (2007). Quantifying social group evolution. *Nature*, *446*(7136), 664–667. https://doi.org/10.1038/nature05670, PubMed: 17410175

Rammer, C., Kinne, J., & Blind, K. (2020). Knowledge proximity and firm innovation: A microgeographic analysis for Berlin. *Urban Studies*, *57*(5), 996–1014. https://doi.org/10.1177/0042098018820241

Reichardt, J., & Bornholdt, S. (2004). Detecting fuzzy community structures in complex networks with a Potts model. *Physical Review Letters*, *93*(21), 218701. https://doi.org/10.1103/PhysRevLett.93.218701, PubMed: 15601068

Rijcke, S. de, Wouters, P. F., Rushforth, A. D., Franssen, T. P., & Hammarfelt, B. (2016). Evaluation practices and effects of indicator usea literature review. *Research Evaluation*, *25*(2), 161–169. https://doi.org/10.1093/reseval/rvv038

Rimmert, C. (2018). *Institutional disambiguation for further countries – an exploration with extensive use of Wikidata (project report)*. (Report). Bielefeld: Bielefeld University, Institute for Interdisciplinary Studies of Science (I^{2}SoS).

Rossetti, G., Milli, L., & Cazabet, R. (2019). CDLIB: A python library to extract, compare and evaluate communities from complex networks. *Applied Network Science*, *4*(1), 52. https://doi.org/10.1007/s41109-019-0165-9

Shrum, W., Genuth, J., Carlson, W. B., Chompalov, I., & Bijker, W. E. (2007). *Structures of scientific collaboration*. Cambridge, MA: MIT Press. https://doi.org/10.7551/mitpress/7461.001.0001

Small, M. L. (2017). *Someone to talk to*. Oxford: Oxford University Press. https://doi.org/10.1093/oso/9780190661427.001.0001

Small, M. L., & Adler, L. (2019). The role of space in the formation of social ties. *Annual Review of Sociology*, *45*(1), 111–132. https://doi.org/10.1146/annurev-soc-073018-022707

Smith-Doerr, L., Alegria, S. N., & Sacco, T. (2017). How diversity matters in the US science and engineering workforce: A critical review considering integration in teams, fields, and organizational contexts. *Engaging Science, Technology, and Society*, *3*, 139. https://doi.org/10.17351/ests2017.142

Sonnenwald, D. H. (2007). Scientific collaboration. *Annual Review of Information Science and Technology*, *41*(1), 643–681. https://doi.org/10.1002/aris.2007.1440410121

Stahlschmidt, S., Stephen, D., & Hinze, S. (2019). *Performance and structures of the German science system* (p. 91). Studien zum deutschen Innovationssystem. https://www.e-fi.de/fileadmin/Assets/Studien/2019/StuDIS_05_2019.pdf

Stephen, D., Stahlschmidt, S., & Hinze, S. (2020). *Performance and structures of the German science system 2020*. Studien zum deutschen Innovationssystem. https://www.e-fi.de/fileadmin/Assets/Studien/2020/StuDIS_05_2020.pdf

Subramanyam, K. (1983). Bibliometric studies of research collaboration: A review. *Journal of Information Science*, *6*(1), 33–38. https://doi.org/10.1177/016555158300600105

Taguchi, H., Murata, T., & Liu, X. (2020). BiMLPA: Community detection in bipartite networks by multi-label propagation. In N. Masuda, K.-I. Goh, T. Jia, J. Yamanoi, & H. Sayama (Eds.), *Proceedings of NetSci-X 2020: Sixth International Winter School and Conference on Network Science* (pp. 17–31). Cham: Springer. https://doi.org/10.1007/978-3-030-38965-9_2

Traag, V. A., Van Dooren, P., & Nesterov, Y. (2011). Narrow scope for resolution-limit-free community detection. *Physical Review E*, *84*(1), 016114. https://doi.org/10.1103/PhysRevE.84.016114, PubMed: 21867264

Traag, V. A., Waltman, L., & van Eck, N. J. (2019). From Louvain to Leiden: Guaranteeing well-connected communities. *Scientific Reports*, *9*(1), 5233. https://doi.org/10.1038/s41598-019-41695-z, PubMed: 30914743

Wagner, C. S., Park, H. W., & Leydesdorff, L. (2015). The continuing growth of global cooperation networks in research: A conundrum for national governments. *PLOS ONE*, *10*(7), e0131816. https://doi.org/10.1371/journal.pone.0131816, PubMed: 26196296

Wuchty, S., Jones, B. F., & Uzzi, B. (2007). The increasing dominance of teams in production of knowledge. *Science*, *316*(5827), 1036–1039. https://doi.org/10.1126/science.1136099, PubMed: 17431139

APPENDIX

Table A1 provides detail about the regional and sectoral composition of the most prolific or largest communities. We separate *Berlin* and *Germany (DEU)* from the rest of Europe to provide a better comparison of *intra*/*inter*-regional collaborations.

Except in two cases where we observe a more internationalized mixture of members i.e., AS (6%, 12%, and 6% from Americas, Asia, and Oceania, respectively) and H (33% from Americas), in all other cases, communities including BUA members (e.g., rows with bold font and gray background) have no members from outside Berlin and Germany and only a small share of European members (maximum is NS with 14%). This shows the national and regionally oriented structure of collaborations among the most prolific actors. However, in aggregate and all fields we observe smaller and less prolific communities with an internationalized share of members (see percentages of members from regions outside Europe on the left side of the table).

In terms of sectoral composition of the communities, education and facility prevail and have the highest shares of members in most of the communities.

Table A1. Composition of the largest and most prolific communities of the giant component by region and sector in aggregate and separate fields (N = community size, P = aggregate publications, B = number of BUA member(s))

					Region (%)								Sector (%)								
Data	cluster	N	P	B	Africa	Americas	Asia	Berlin	DEU	Europe	Oceania	No region	Archive	Company	Education	Facility	Government	Healthcare	Nonprofit	Other	No sector
	0	**29**	**98,408**	**3**				**21**	**76**	**3**				**3**	**62**	**10**	**3**	**21**			
	1	**10**	**35,734**	**1**				**60**	**30**	**10**					**30**	**60**	**10**				
ROR	2	15	7,356			20	7	7	20	47					33	40	20		7		
	3	9	4,002			11	11	22	56						33	56			11		
	4	29	5,285			10	21	3	10	55				7	62	21	3		3	3	
	5	9	2,414					11		89					67	33					
	6	10	2,822			20	30	10		40				10	80	10					
	7	5	2,157					20	40	40					40	40		20			
	8	14	3,441			14		7	14	64					79	14		7			
	9	11	2,110			9	18		18	55					18	45	18			18	
	10	11	1,571					18	45	36					18		55	9	9	9	
	23	31	1,230		3	3	61	3	16	10	3			3	68	10	3	3	10	3	
	0	**13**	**4,308**	**2**				**31**	**62**	**8**					**69**		**15**	**15**			
	1	**16**	**3,582**	**2**		**6**	**12**	**12**	**56**	**6**	**6**		**6**		**56**	**19**		**19**			
AS	4	26	204		8	8	15	4	54	12				4	50	31	8	8			
	9	27	182			15	4	4	7	67	4		4	4	59	22		7		4	
	10	29	180		3	17	34	7	10	28			3		55	28	7			3	3
	12	38	99			13	8			79				8	37	16	13	13	8	5	

Table A1. *(continued)*

					Region (%)								Sector (%)								
Data	cluster	N	P	B	Africa	Americas	Asia	Berlin	DEU	Europe	Oceania	No region	Archive	Company	Education	Facility	Government	Healthcare	Nonprofit	Other	No sector
	0	**7**	**15,116**	**1**				**71**	**29**						**29**	**71**					
	1	**11**	**12,457**	**3**				**36**	**55**	**9**					**45**	**18**	**9**	**27**			
ET	2	23	4,092			9	26	4	22	39					39	39	13	4	4		
	3	15	1,460			20	40	13		27				7	60	33					
	4	30	2,059			3	23	3	7	63				3	60	20	7		3	7	
	5	9	1,072			11			33	56				11	22	44	11			11	
	20	31	294			6	6		10	74	3			6	42	23	3	6	13	6	
	0	**32**	**58,586**	**3**				**19**	**78**	**3**				**3**	**62**	**12**	**3**	**19**			
	1	**7**	**24,045**	**1**				**57**	**29**	**14**					**29**	**57**	**14**				
NS	2	3	4,676					33	67						67	33					
	3	16	6,186			6	12	6	19	56					31	44	19		6		
	4	27	4,869			7	22	4	7	59				7	63	22	4			4	
	5	8	2,897			12		25	38	25					12	62			25		
	6	12	2,816			33	25	8		33				8	75	17					
	7	16	2,976			6	6	6		69	12				75	19	6				
	8	6	1,643					17	17	67					67	17		17			
	18	26	968			8	62	12	4	12	4				77	4	4		8	8	
	71	45	553		2	7	4			84	2				20	42	7	13	9	9	
	0	**31**	**53,817**	**4**				**19**	**81**					**3**	**61**	**10**	**3**	**23**			
MHS	1	10	1,875			20	20	20	30	10					40	40			20		
	4	20	1,286			45		5	5	45					90	10					
	19	26	552			23			4	73					46	12	15	27			
	44	28	146			7	11		4	79				4	36	36	4	18	4		
	70	33	172		64	12	6			18			3	6	33	12	6	21	9	9	

	0	**6**	**2,628**	**3**				**50**	**50**				**17**	**67**			**17**		
	1	**9**	**411**	**1**		**33**		**11**	**22**	**33**				**100**					
H	2	2	264			50		50							50			50	
	3	24	416			17	4	4	17	46	12		4	88	8				
	4	24	225		4	42		4	8	42				83	17				
	5	18	131			22	28	6	11	33				61	28	11			
	0	**22**	**10,895**	**3**				**14**	**77**	**9**			**5**	**64**	**9**		**23**		
	1	**8**	**2,046**	**1**				**50**	**38**	**12**				**38**	**50**				**12**
SS	2	19	1,733			32		5	21	37		5		89	11				
	6	29	291		3	14	14	3	3	62			3	55	31	3	3		3
	8	31	353			58		6		35			6	77	6			6	3
	47	24	150		4	12	8		8	67			4	67	8	17	4		