



Homophily and its effects on collaborations and repeated collaborations: a study across scientific fields

João M. Santos¹ · Hugo Horta²  · Shihui Feng³

Received: 27 June 2023 / Accepted: 18 January 2024 / Published online: 16 February 2024
© The Author(s) 2024

Abstract

This study examines the influence of homophily on research collaborations across all scientific fields, noting the role of ascribed, acquired, geographical, current career, and educational and career history attributes. Our analysis builds on previous studies by adopting a comparative approach across scientific fields and examining the effect of homophily on repeated collaborations. Our findings reveal physical proximity as a universal driver of collaboration across all scientific fields, both for overall collaborations and as a maintainer of collaborative endeavors. We also note that most attributes that are responsible for overall collaborations are similar to the attributes that play a role in maintaining them, with a few exceptions, notably when it comes to institutional characteristics. Our results also show the complex role of career history attributes, such as job variety and international mobility, in influencing the likelihood of collaboration. This study sheds light on the intricate interplay of different attributes in shaping scientific collaborations, underscoring the need for a multidisciplinary approach in future homophily studies and nuanced strategies for fostering collaborations across different scientific fields.

Keywords Homophily · Research collaborations · Science studies · Scientometrics

Introduction

Modern-day science has become largely collaborative in nature, due not only to the inherent efficiency gain obtained through working as a team but also to the increasingly complex nature of the challenges yet to be solved (Katz & Martin, 1997; Uddin et al., 2013). Because science is not monolithic, the dynamics of scientific collaboration differ substantially by field. For example, in the hard sciences, large collaborative endeavors tend to be

✉ Hugo Horta
horta@hku.hk

¹ Centro de Investigação e Estudos de Sociologia (Cies_Iscte), Iscte–Instituto Universitário de Lisboa, Lisbon, Portugal

² Social Contexts and Policies of Education, Faculty of Education, The University of Hong Kong, Hong Kong SAR, China

³ Mathematics, Science and Technology, Faculty of Hong Kong, The University of Hong Kong, Hong Kong SAR, China

the norm, particularly in the more experimental sciences that are rooted in the laboratory as the workplace (Lauto & Valentin, 2013). Researchers in the social sciences and humanities can, in many cases, conduct the bulk of their work by themselves, often even from home (Henriksen, 2018). Thus, it comes as no surprise that collaborations in the social sciences and humanities have fewer authors, and co-authorships in these fields have only recently been rising due to the growing internationalization of science and academia and the establishment of research and career incentives that push for more collaboration (Kwiek, 2018). The existence of incentives to conduct work collaboratively means that cooperation between researchers is likely to further increase for the foreseeable future (Xu, 2020). Thus, it becomes critical to understand both the processes leading to such collaborations and the continued sustainability of these collaborations.

A critical construct underlining collaborative relationships is homophily—a sociological principle that suggests that individuals have an inherent tendency to bond with others who exhibit similar attributes (Lazarsfeld & Merton, 1954). The essence of homophily extends beyond social connections to infuse professional relationships, shaping the way in which researchers in a social system such as science choose and engage with collaborators. This propensity to align and be engaged with similar individuals manifests in various forms, such as shared characteristics, interests, and beliefs, with notable implications for scientific collaboration and co-authorship. Indeed, homophily creates an additional layer of complexity in the study of collaborations. Whereas most studies have focused on the propensity to collaborate (Jeong et al., 2011), patterns of collaboration (Kwiek, 2021), and other aspects relating to the collaborative process, such as the ordering of authors and related ethical issues (Youtie & Bozeman, 2014), homophily is distinct in that the unit of analysis becomes the dyad of researchers and the shared attributes that trigger their collaboration.

The influence of homophily on collaborative research relationships can be traced to a myriad of attributes—from ascribed to acquired, geographical to cultural, and prestige to resources. Lazarsfeld and Merton's (1954) original article about homophily in social relations essentially considered two attributes: *ascribed*, which are attributes inherent to the individual, and *acquired*, which are attributes that result from real-world experiences such as education and work. The weight of ascribed attributes, such as gender, race, and age, in the formation of research collaborations has long been acknowledged. For example, same-gender researchers are more likely to collaborate than researchers of different genders (González Brambila & Olivares-Vázquez, 2021; Holman & Morandin, 2019). The role of acquired attributes, such as an individual's professional expertise (Hunter & Leahey, 2008), strategic research preferences (Evans et al., 2011), and even personality traits (Horta et al., 2022), is also becoming evident. *Geographical* and cultural attributes are particularly deserving of attention. Despite advances in digital communication, physical proximity continues to significantly influence collaborative choices. Indeed, geographical co-location has been noted as the primary attribute leading to collaboration, stressing the role of homophily in collaborative research processes (Evans et al., 2011; Horta et al., 2022). Institutional and societal identities further amplify this homophilic tendency, shedding light on how deeply entrenched social attributes can sway professional decisions in academia, including those related to collaborative endeavors (Tavares et al., 2022).

However, it is worth noting that homophily does not consistently prevail in all cases; the context and specific objectives of a research project may engender heterophily—the inclination toward dissimilarity (i.e., heterophily). For example, complementarity of skills in collaborations can lead to heterophily in network formation (Xie et al., 2016). Other rationales and mechanisms, such as the role of positional goods in research collaboration, add another layer of complexity. Traditionally, prestige has been a major determinant of

collaborations, with less renowned researchers naturally gravitating toward those with more prestige (Ebadi & Schiffauerova, 2015). The complexity of modern research increasingly necessitates collaborations that prioritize expertise and utilitarian associations rather than similarities (Feng & Kirkley, 2020). The availability of resources, such as funding and time, also dictates the course of collaborations, reinforcing the practical aspects of academic work rather than similarities between co-authors (Ubfal & Maffioli, 2011). All of these elements are associated with heterophily—engaging with others who have different attributes. Therefore, the current research collaboration arena seems to be increasingly complex and includes both heterophilic and homophilic drivers that deserve further study for researchers, research managers, and policymakers to be aware of, better understand, and act upon (as argued by Huang, 2014).

Despite some understanding of the influence of homophily (and heterophily) on research collaborations, there are still at least two known unknowns on the topic. One is the extent to which homophily affects research collaboration in different scientific fields: current studies have only focused on a single scientific discipline or field (Horta et al., 2022) or compared a few disciplines, mostly from the natural sciences (Zeng et al., 2016). The other relates to the effects that homophily may have on established research collaborations, that is, research collaborations that include the same co-authors. To the best of our knowledge, no study has performed such an analysis. Therefore, the research questions guiding this study are as follows:

RQ1a: Which homophily factors among researchers contribute to research collaborations across all fields of science?

RQ1b: What is the relative importance of the various types of attributes regarding research collaborations?

RQ1c: Are there scientific field differences in the effects of homophily on research collaborations?

RQ2a: Which homophily factors among researchers contribute to repeated collaborations across all fields of science?

RQ2b: What is the relative importance of the various attributes regarding repeated collaborations?

RQ2c: Are there scientific field differences in the effects of homophily on repeated collaborations?

Our study contributes to the literature in two ways. First, we explore homophily and research collaborations encompassing researchers from all scientific fields. Second, we assess the effect of homophily on repeat collaborations, that is, the shared attributes that lead to researchers collaborating more than once with one another.

Method

Participants

This study used a large dataset that was first collected in 2017 as part of a multi-study project. In this section, we detail how this primary data source was collected and how the working dataset was generated.

In the first step, we began by identifying all corresponding authors who published in all fields of science between 2010 and 2016. They were subsequently invited to complete an online survey consisting of the Multidimensional Research Agendas Inventory-Revised

(MDRAI-R; Horta & Santos, 2020)—an instrument to evaluate Strategic Research Agendas (SRA)—along with several career-level, demographic, and educational questions. The questions in the survey cover variables that are identified in the literature as being determinants of research homophily and can be broadly grouped into various categories of attributes (these are described further ahead). Some variables used in the survey were already tested in a previous study on homophily (Horta et al., 2022), but new variables were included as a means to expand the categories of attributes; by doing so the current study tests new attributes, adding further novelty and strengthening the study's contribution to the literature.

The participants were required to complete an informed consent form before they could proceed to the survey itself. Although the full dataset of participants who clicked the invitation link was composed of 21,016 participants, in this exercise we used only 4,855 participants. Three hundred and one participants did not complete the informed consent form and were unable to advance to the survey and an additional 1,953 quit the survey on the first page. Of the remaining 18,762 participants, only 9,162 reached the end of the survey, which was expected due to its length (roughly 30 min to complete). However, for the working sample, this number was further reduced to 4,855 participants; this was due to the presence of non-imputable missing data, notably in the career section, which was placed at the end of the survey and made optional due to privacy concerns.

In the second step, we retrieved each participant's publication records and updated bibliometric data using the Scopus API. Because the original dataset already contained the ScopusID for each participant, nominal ambiguity was not an issue. These additional data were used to produce the final, working dataset, which is detailed below.

The final working sample was roughly composed of two-thirds males ($N=3222$; 66.4%) and one-third females ($N=1633$; 33.6%), with an average age of 52 years ($M=51.64$, $SD=11.81$). Most of the participants hailed from the Medical and Health Sciences ($N=1591$; 32.8%), followed by the Natural Sciences ($N=1328$; 27.4%), Social Sciences ($N=879$; 18.1%), Engineering and Technology ($N=809$; 16.7%), Agricultural Sciences ($N=202$; 4.2%), and Humanities ($N=43$; 0.9%). Finally, in terms of geographical distribution, the participants originated from a variety of countries worldwide. The most represented countries were the United States ($N=916$; 18.9%), Italy ($N=439$; 9.0%), France ($N=272$; 5.6%), Spain ($N=254$; 5.2%), and Australia ($N=251$; 5.2%), with the remaining participants distributed over a myriad of other countries.

Data processing

Using the 4,855 participants' publication records and ScopusIDs as identifiers, we created a dyadic matrix with all possible co-author combinations among them. For each dyad, the following two variables were created: a *Collaboration* variable, with the value of 1 if the dyad had collaborated in publishing a scientific article and 0 otherwise; and a *Repeated Collaboration* variable, with the value of 1 if the dyad collaborated more than once and 0 otherwise. After removing the main diagonal of the matrix and the redundant lower half, this resulted in a dataset of 11,783,085 unique dyads ($(4855 \cdot (4855 - 1)) / 2$). For each dyad, we computed similarity and dissimilarity measures based on the original variables. Categorical variables were coded as 1 if they were identical (e.g., same gender) and 0 otherwise. The quantitative variables were computed as the absolute difference of the variable for each member of the dyad, thus creating a measure of dissimilarity (e.g., the larger the value, the larger the difference between the two members of the dyad).

Variables

This section describes the variables that were included in our models. Note that as described above, the variables were not used in their raw form but were transformed into measures of similarity (homophily) or dissimilarity (heterophily). However, for ease of presentation, here we describe them using their natural conceptual meaning. Because our analysis considers five groups of variables—*ascribed attributes*, *geographical attributes*, *present career attributes*, *educational and career history attributes*, and *acquired attributes*—we present them in this manner.

First, it should be noted that for control purposes, *Field of Science* (FOS) was included in every model. As each case in the dataset represented a dyad of researchers and not an individual researcher, there were some subtleties in our implementation and interpretation of this variable. When the members of the dyad belonged to different scientific fields, the variable took the value of 1, representing *Multidisciplinary*. This was the reference category. All other levels of the variable indicated that the members of the dyad belonged to the same scientific field. In addition, Humanities and Social Sciences were merged due to the relatively low sample size for Humanities dyads.

For *ascribed attributes*, we included *gender* and *age*, which are both self-explanatory. For the *geographical attributes*, we included the participant's *country*, *university*, and *city*. In terms of *present career*, we included *Top-Ranked University*, which indicated whether or not the participant worked at a Top 500 ARWU ranked university; *Academic Sector*, indicating whether or not the participant was working in the academic sector; and *Educational Inbreeding*, which measured educational immobility because it indicated whether the participant had pursued their entire educational path at the same university. The *educational and career history* batch of variables included *Job Changes*, which was the count of the number of job changes that occurred throughout the participant's career, including changing jobs within the same sector and among sectors of activity; *Job Country Changes*, which was the number of times that the participant engaged in international employment mobility; *Publications* and *Citations*, which are self-explanatory and were extracted from Scopus as part of the data-gathering process; *Percentage of Research Funding*, which indicated the percentage of the participant's career with access to research funding; and finally, *Percentage of Teaching*, representing the percentage of time dedicated to teaching graduates.

The final batch of variables, *acquired attributes*, consisted of the various low-level dimensions of the Multidimensional Research Agendas Inventory—Revised (MDRAI-R) (Horta & Santos, 2020). As these dimensions relate to various concepts across a wide body of literature, they are summarized in Table 1 to facilitate interpretation.

Procedure

As both the “Collaboration” and “Repeated Collaboration” dependent variables were binary, by default, logistic regression was the optimal choice (Hair et al., 2014). However, the collaboration matrix was extremely sparse; only 5,582 collaborations occurred out of 11,783,085 dyads (0.05%), and this sparsity was compounded with Repeated Collaborations (2,418 collaborations out of 11,783,085 dyads, representing 0.02%). Rare events in logistic regressions can be a potential source of bias (King & Zeng, 2001). Although some methods are robust to rare event bias, they are computationally

Table 1 Dimensions of the MDRAI-R (acquired attributes)

<i>Dimension</i>	<i>Definition</i>
Scientific ambition	Prestige. The desire to acquire recognition and academic prestige in a given field Drive to publish. Being motivated and driven by the publication of scientific articles
Divergence	Branching out. The desire to expand into other fields of study or topics Multidisciplinary. The preference for working in multidisciplinary research ventures
Discovery	The preference for working in fields or topics with the potential to lead to scientific discovery
Tolerance for low funding	The willingness to work on fields or topics for which research funding is scarce
Collaboration	Willing to collaborate. The desire to engage in collaborative scientific ventures Invited to collaborate. Having the opportunity and receiving invitations to participate in collaborative scientific ventures
Mentor influence	The researcher's mentor (PhD or otherwise) holds a degree of influence over their work
Academia driven	Field oriented. The extent to which the research agenda is influenced by scientific priorities that the field community determines by consensus Institution oriented. The propensity of the researcher to align their research agenda with the research strategic targets of their institution
Society driven	Society oriented. The incidence of society-related challenges in a research agenda Non-academic oriented. The influence and participation of laymen and non-experts in the design of the research agenda

intensive, an issue further compounded by our very large dataset. Due to computational limits, it became evident that we would be unable to conduct the entire analysis using a robust method; thus, we estimated a single model using a robust method and compared it to the model estimated by a simple logistic regression to determine whether this approach influenced the results (the logic being that consistent results would indicate the absence of a rare event bias). To this end, we employed a penalized likelihood method, or a Firth regression (Firth, 1993), using R's *logistf* library. We conducted a Firth regression using *Model V* (described further below), the most complex model and the one with the smallest number of events on the dependent variable, for the "Repeated Collaboration" variable. Comparing the results of the Firth regression and the logistic regression revealed negligible differences, notably centesimal or millesimal differences in the coefficients and no changes to the significance values. Thus, we proceeded with the logistic regressions for the remainder of the models.

The categorical variables were specified as fixed factors in the model, in which the reference category was "different" (thus, the value indicated in the table is the estimator for homophilic dyads). Four sets of analyses are presented in the subsequent sections.

Analysis 1 involved a series of hierarchical models in which "Collaboration" was the dependent variable. The variables were entered into the model in the following manner: *Model I* included ascribed attributes; *Model II* added geographical attributes; *Model III* added present career attributes; *Model IV* added educational and career history attributes; and *Model V* added acquired attributes. Each model also included the previous models' variables, as is standard practice with hierarchical regressions. This hierarchical

approach allows us to determine which types of attributes are the most important when determining the odds of collaboration. This analysis addresses RQ1a and RQ1b.

Analysis 2 took *Model V* from the previous analysis, removed the *Field of Science* (FOS) variables from the list of predictors, and instead split the model into five separate models, based on the disciplinary orientation of the members of the dyad, to examine the differences among disciplines regarding the influence of homophily factors. In the *Multi-disciplinary* model, we included only dyads in which each member belonged to a different scientific field. For the remainder of the models, each named after a specific field, we included only dyads in which both members were part of that same field. This approach allowed us to better understand how the specific dynamics of homophily operated within each field. This analysis addresses RQ1c. Finally, Analysis 3 and Analysis 4 were similar to the previous two analyses except for the fact that they used the “Repeated Collaboration” variable, allowing us to understand whether the mechanics leading to repeating collaborations differed from one-off collaborations. Analysis 3 addresses RQ2a and RQ2b, while Analysis 4 addresses RQ2c.

Results

Analysis 1 – RQ1a and RQ1b

The first analysis is a hierarchical regression on the propensity of the members of a dyad to have collaborated. The results of this analysis are shown in Table 2.

In Model I, which begins by introducing the ascribed attributes, it is possible to identify the effects of homophily in research collaborations. Researchers of the same gender are 24.14% more likely to engage in collaboration ($B = 0.216$, $OR = 1.241$, $p < 0.01$) than those of different genders. Additionally, increasing age differentials reduce the odds of collaboration at the rate of 2.17% per year of age difference ($B = -0.021$, $OR = 0.978$, $p < 0.01$), meaning that researchers of similar ages are more likely to collaborate than researchers of dissimilar ages. Field of Science also has significant effects, underlining the power of disciplinary fields, traditions, and values and norms in science and in shaping collaborations. This result is aligned with findings that show that even in multidisciplinary research, disciplinary homophily plays a key role (Feng & Kirkley, 2020).

Model II includes the geographical attributes. The new variables reveal that geographical proximity is an important predictor of collaboration. Researchers in the same country are roughly 6 times more likely to collaborate than those in different countries ($B = 1.797$, $OR = 6.030$, $p < 0.01$), whereas those in the same university are roughly 5 times more likely to collaborate ($B = 1.642$, $OR = 5.167$, $p < 0.01$), and those in the same city are 5.5 times more likely to collaborate ($B = 1.720$, $OR = 5.583$, $p < 0.01$). The variables from Model I maintain their significant effects. Proceeding to Model III, which adds current career attributes, we find two additional effects. Researchers who work in a top-ranked university yield a 17.54% increase in the odds of collaboration ($B = 0.162$, $OR = 1.175$, $p < 0.01$), and researchers working in the academic sector yield a 12.7% increase ($B = 0.120$, $OR = 1.127$, $p < 0.05$). Educational inbreeding has no effects on the odds of collaboration.

Model IV includes educational and career history attributes, and several more interesting effects emerge. Differences in job changes decrease the likelihood of collaboration ($B = -0.034$, $OR = 0.966$, $p < 0.05$), but differences in terms of international mobility actually increase the odds of collaboration ($B = 0.061$, $OR = 1.063$, $p < 0.01$). Changing jobs

Table 2 Hierarchical logistic regression for Collaboration

Variables	Model I	Model II	Model III	Model IV	Model V
Gender (identical)	0.216*** (0.028)	0.213*** (0.029)	0.211*** (0.042)	0.203*** (0.043)	0.220*** (0.050)
Age delta	−0.022*** (0.001)	−0.020*** (0.002)	−0.020*** (0.002)	−0.021*** (0.002)	−0.018*** (0.003)
FOS (Agricultural sciences)	3.194*** (0.116)	3.007*** (0.129)	2.951*** (0.169)	2.955*** (0.174)	2.885*** (0.205)
FOS (Engineering & Technology)	2.109*** (0.055)	2.020*** (0.056)	1.929*** (0.087)	1.937*** (0.090)	1.967*** (0.101)
FOS (Medical and Health sciences)	2.210*** (0.036)	2.107*** (0.037)	1.994*** (0.055)	1.963*** (0.057)	1.899*** (0.067)
FOS (Natural sciences)	2.228*** (0.039)	2.227*** (0.040)	2.212*** (0.055)	2.137*** (0.058)	2.153*** (0.066)
FOS (Social sciences)	2.179*** (0.051)	2.014*** (0.053)	2.022*** (0.076)	2.084*** (0.079)	2.150*** (0.088)
Country (identical)		1.797*** (0.033)	1.745*** (0.048)	1.714*** (0.051)	1.728*** (0.058)
University (identical)		1.642*** (0.079)	1.428*** (0.115)	1.397*** (0.119)	1.421*** (0.141)
City (identical)		1.720*** (0.071)	1.806*** (0.101)	1.837*** (0.105)	1.799*** (0.124)
Top-Ranked University (identical)			0.162*** (0.045)	0.156*** (0.047)	0.140*** (0.054)
Academic Sector (identical)			0.120** (0.046)	0.085* (0.048)	0.044 (0.056)
Educational Inbreeding (Identical)			0.053 (0.040)	0.050 (0.042)	0.096** (0.048)
Job Changes delta				−0.034** (0.015)	−0.042** (0.018)
Job Country Changes delta				0.061*** (0.021)	0.078*** (0.024)
Publications delta				0.001** (0.000)	0.001** (0.000)
Citations delta				0.000*** (0.000)	0.000*** (0.000)
Percentage Research Funding delta				−0.006*** (0.001)	−0.006*** (0.001)
Percentage Teaching delta				−0.003*** (0.001)	−0.003*** (0.001)
Prestige delta					−0.004 (0.027)
Drive to Publish delta					−0.014 (0.026)
Branching Out delta					0.019 (0.027)

Table 2 (continued)

Variables	Model I	Model II	Model III	Model IV	Model V
Multidisciplinarity delta					-0.044 (0.027)
Willingness to Collaborate delta					0.006 (0.034)
Invited to Collaborate delta					-0.064** (0.030)
Field Orientation delta					-0.010 (0.026)
Institution Orientation delta					-0.004 (0.023)
Society Orientation delta					-0.108*** (0.023)
Non-academic Orientation delta					-0.086*** (0.026)
Mentor Influence delta					-0.010 (0.021)
Tolerance to Low Funding delta					-0.027 (0.022)
Discovery delta					-0.058* (0.033)
Pseudo R ²	0.068	0.150	0.145	0.147	0.147
Observations	11,768,526	11,155,726	5,172,936	4,853,170	3,821,230

might mean changing to different sectors of activity in which one’s habits, work goals, or mentality change while retaining little in common. Less internationally mobile researchers may want to collaborate with more internationally mobile researchers to tap into global knowledge and resources (e.g., Ryan, 2015), and internationally mobile researchers may benefit in collaborating with less internationally mobile researchers, who may facilitate access to local data, knowledge, and resources. Asymmetry in both publications ($B = 0.001$, $OR = 1.000$, $p < 0.05$) and citations ($B = 0.000$, $OR = 1.000$, $p < 0.01$) increases the odds of research collaborations, probably due to the fact that less established researchers naturally gravitate towards those with a more prolific track record. However, the coefficient is close to zero and, as such, this is not a very noticeable effect. Finally, differences in the percentage of the researcher’s career with research funding decreases the odds of collaboration ($B = -0.006$, $OR = 0.994$, $p < 0.01$), and this is also the case for differences in the percentage of time allotted to teaching graduates ($B = -0.003$, $OR = 0.997$, $p < 0.01$), highlighting the differences in the dynamics between research and teaching track careers and their corresponding resources (see also Kwiek, 2018, 2020).

Finally, Model V includes acquired attributes in terms of the homophily of the dimensions of the strategic research agendas of the researchers. Of these, only a few have notable effects. As expected, asymmetries in being invited to collaborate reduce the odds of collaboration ($B = -0.064$, $OR = 0.937$, $p < 0.05$) because either one is invited to collaborate or one is not. Additionally, the larger the gap between researchers in terms of society

orientation ($B = -0.108$, $OR = 0.897$, $p < 0.01$) and non-academic consultation ($B = -0.086$, $OR = 0.917$, $p < 0.01$), the less likely they are to collaborate. Finally, differences in the researchers' orientations toward discovery-driven agendas also reduce their propensity to collaborate ($B = -0.058$, $OR = 0.943$, $p < 0.01$), because the scientific goals and working methods of researchers working on the frontier of science and those preferring to contribute through incremental advances to science is known to be radically different (Santos & Horta, 2018).

In terms of hierarchy, by evaluating the Pseudo R^2 it becomes apparent that the model that most substantially improves this measure of fit is Model II, which includes geographical attributes. This result highlights physical geographical proximity as the most important aspect in increasing the odds of collaboration. Although to some extent this is an unsurprising finding, the strength of the effect is somewhat unexpected given the current working environment and the era, in which there is an abundance of tools for engaging in remote collaborative ventures.

Analysis 2 – RQ1c

Analysis 2 takes Model V from the previous analysis and splits the model by field of science (thus removing FOS as a control variable). Each FOS considers only dyads where both members are part of the same field; otherwise, if the members are from different fields, they are considered under the “Multidisciplinary” model. Given the comparative nature of this analysis, we focus more on the effects that differ across fields (as Analysis 1 already covered global effects). This analysis is summarized in Table 3.

The first notable effect is that the homophilic effect concerning gender is not homogeneous across scientific fields. It is absent in the Agricultural sciences, Engineering & Technological sciences, and Medical & Health sciences. However, it is present in similar magnitudes in the Natural sciences and in Multidisciplinary collaborations (researchers of the same gender are approximately 25% more likely to collaborate than researchers of different genders). The magnitude is stronger in the social sciences, where same-gender researchers are 69% more likely to collaborate than different-gender researchers ($B = 0.525$, $OR = 1.690$, $p < 0.01$). These trends may relate to the overpopulation of male researchers in the natural sciences who collaborate more with other male researchers, whereas in the social sciences females researchers tend to be the majority and thus collaborate more with female researchers, but other explanations may be possible because even in the social sciences, empirical studies have found that female researchers are less likely to publish and when they do publish, they often collaborate with male researchers (Feinberg et al., 2011). Therefore, even in the field of social sciences, the observed homophily effect may be one of a minority of male researchers intensively collaborating with one another. The same may be true of Natural sciences and Multidisciplinary collaborations (Ozel et al., 2014).

Age and geographical attributes behave in a consistent manner across scientific fields and have essentially the same effect described in the global analysis. The only exception is Agricultural sciences, in which researchers collaborate more within the same country than elsewhere, but not with those in the same university or city. This may relate to the fact that much of the research (and teaching) in agricultural sciences involves fieldwork often conducted away from colleagues in the same university or city (Parr et al., 2007), thus highlighting the relevance of the principle of physical proximity to promoting homophilic collaborations. Educational Inbreeding, which did not exhibit significant effects until Model V in the previous analysis, now reveals effects that are only present in multidisciplinary

Table 3 Logistic regression for Collaboration, by Field of Science

Variables	Multidisc	Agricultural	Engineering	Medical	Natural	Social
Gender (identical)	0.248*** (0.096)	0.159 (0.460)	-0.124 (0.186)	0.153 (0.094)	0.231** (0.099)	0.525*** (0.155)
Age delta	-0.015*** (0.005)	-0.094*** (0.029)	-0.016* (0.009)	-0.025*** (0.005)	-0.009** (0.005)	-0.029*** (0.008)
Country (identical)	1.761*** (0.116)	2.907*** (0.522)	2.463*** (0.208)	1.837*** (0.112)	1.502*** (0.111)	1.521*** (0.176)
University (identical)	1.213*** (0.254)	-0.757 (1.455)	1.549*** (0.483)	1.655*** (0.249)	1.320*** (0.330)	1.616*** (0.439)
City (identical)	2.284*** (0.219)	1.976 (1.229)	1.462*** (0.464)	1.584*** (0.221)	1.325*** (0.281)	2.195*** (0.399)
Top Ranked U. (identical)	0.158 (0.104)	-0.261 (0.556)	0.003 (0.228)	0.062 (0.104)	0.230** (0.101)	-0.026 (0.160)
Academic Sector (identical)	0.027 (0.109)	-0.354 (0.424)	0.032 (0.232)	0.058 (0.103)	0.086 (0.103)	0.146 (0.229)
Educ. Inbreeding (identical)	0.166* (0.093)	0.244 (0.402)	-0.042 (0.182)	0.020 (0.094)	0.068 (0.088)	0.250* (0.151)
Job Changes delta	-0.024 (0.033)	-0.155 (0.150)	-0.105 (0.080)	-0.028 (0.033)	-0.035 (0.033)	-0.091 (0.059)
Job Country Changes delta	-0.025 (0.048)	0.544*** (0.183)	-0.045 (0.111)	0.152*** (0.046)	0.086** (0.042)	0.049 (0.077)
Publications delta	0.001* (0.000)	0.001 (0.004)	0.002*** (0.001)	0.001*** (0.000)	-0.001 (0.000)	-0.003* (0.002)
Citations delta	0.000*** (0.000)	0.000 (0.000)	-0.000 (0.000)	-0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
Percentage Research Funding delta	-0.009*** (0.002)	0.003 (0.009)	-0.000 (0.004)	-0.007*** (0.002)	-0.005** (0.002)	-0.003 (0.003)
Percentage Teaching delta	0.000 (0.002)	-0.025*** (0.009)	-0.004 (0.003)	-0.001 (0.002)	-0.004** (0.001)	-0.009*** (0.003)
Prestige delta	0.019 (0.052)	0.571*** (0.175)	-0.050 (0.106)	-0.064 (0.055)	-0.019 (0.050)	0.058 (0.083)
Drive to Publish delta	-0.015 (0.049)	-0.285 (0.262)	-0.011 (0.101)	-0.078 (0.054)	0.055 (0.048)	-0.052 (0.071)
Branching Out delta	-0.036 (0.053)	-0.016 (0.254)	0.124 (0.115)	-0.003 (0.051)	0.051 (0.054)	0.045 (0.073)
Multidisciplinarity delta	-0.040 (0.052)	0.465** (0.223)	-0.143 (0.097)	-0.001 (0.055)	-0.084* (0.050)	-0.016 (0.082)
Willingness to Collaborate delta	0.114* (0.061)	-0.233 (0.291)	-0.048 (0.131)	-0.097 (0.071)	0.035 (0.067)	-0.028 (0.092)
Invited to Collaborate delta	-0.200*** (0.060)	-0.103 (0.325)	0.090 (0.099)	-0.040 (0.066)	-0.034 (0.058)	-0.004 (0.076)
Field Orientation delta	0.046 (0.049)	-0.048 (0.236)	-0.049 (0.107)	0.015 (0.053)	-0.087* (0.051)	0.045 (0.077)
Institution Orientation delta	-0.018 (0.046)	0.064 (0.158)	0.092 (0.091)	-0.009 (0.047)	-0.016 (0.041)	-0.025 (0.067)

Table 3 (continued)

Variables	Multidisc	Agricultural	Engineering	Medical	Natural	Social
Society Orientation delta	-0.136*** (0.045)	-0.201 (0.235)	0.004 (0.094)	-0.080 (0.050)	-0.132*** (0.040)	-0.063 (0.065)
Non-academic Orientation delta	-0.071 (0.049)	-0.207 (0.212)	-0.131 (0.112)	-0.078 (0.050)	-0.090* (0.050)	-0.095 (0.079)
Mentor Influence delta	-0.037 (0.045)	-0.074 (0.209)	-0.129* (0.067)	0.003 (0.041)	0.017 (0.038)	0.038 (0.061)
Tolerance to Low Funding delta	0.001 (0.040)	0.087 (0.159)	0.025 (0.083)	-0.108** (0.044)	-0.036 (0.044)	0.043 (0.062)
Discovery delta	-0.013 (0.062)	-0.103 (0.269)	-0.082 (0.124)	-0.114* (0.067)	-0.091 (0.061)	0.098 (0.097)
Pseudo R2	0.097	0.305	0.168	0.113	0.052	0.120
Observations	2,888,297	7,750	98,346	369,370	334,971	122,496

dyads ($B=0.166$, $OR=1.180$, $p<0.1$) and in the Social Sciences ($B=0.250$, $OR=1.284$, $p<0.1$).

Job Changes does not have statistically significant effects in any field of science. Job Country Changes, which indicates international employment mobility, shows differentiated effects. In the Natural sciences and the Medical sciences, the effect of Job Changes is aligned with that shown in the global analysis. However, it is not present in Multidisciplinary dyads or in dyads from Engineering and the Social Sciences. Interestingly, this effect is present to a much greater degree in the Agricultural Sciences—each unit of difference in terms of international employment mobility increases the odds of collaboration by roughly 72% ($B=0.544$, $OR=1.723$, $p<0.01$). Publications has a largely consistent effect, and it is only non-significant for the Agricultural Sciences and the Natural Sciences. Citations, however, only maintain a significant effect for the Multidisciplinary dyads, even though the size of the effect is still marginal at best ($B=0.000$, $OR=1.000$, $p<0.01$). Differences in the Percentage of Research Funding are also not consistent, with this effect only present for Multidisciplinary dyads and for dyads from the Medical Sciences and Natural Sciences. Likewise, differences in the Percentage of Teaching are only significant for the Agricultural Sciences, the Natural Sciences, and the Social Sciences.

In terms of acquired attributes, Prestige, which did not exhibit an effect in the global model, reveals a very significant effect exclusively in the Agricultural sciences ($B=0.571$, $OR=1.770$, $p<0.01$), with each unit of difference between members of the dyad increasing the odds of collaboration by 77%. Agricultural sciences also experience a strong effect from Multidisciplinary, with each unit of difference increasing the odds of collaboration by 59.2% ($B=0.465$, $OR=1.592$, $p<0.01$). Multidisciplinary also has an effect in the Natural sciences, where it decreases the odds of collaboration ($B=-0.084$, $OR=0.919$, $p<0.1$). Invited to Collaborate, which previously exhibited a global effect, is now shown to be exclusive to multidisciplinary dyads ($B=-0.200$, $OR=0.819$, $p<0.01$), with each unit in difference reducing the odds of collaboration by 19%. Willingness to Collaborate has an effect exclusive to multidisciplinary clusters that was not present in the global analysis: each unit of difference increases the odds of collaboration by 12% ($B=0.114$, $OR=1.120$, $p<0.1$). Field orientation, which also was not significant globally, now exhibits a negative effect, which is significant only for the Natural sciences ($B=-0.087$, $OR=0.917$, $p<0.1$).

Society orientation, which was previously significant at a global level, shows some nuances; this effect is only present in Multidisciplinary dyads and dyads from the Natural sciences and in both cases reduces the odds of collaboration, which is consistent with the global analysis. Finally, Mentor Influence has a negative effect on Engineering dyads, whereas Tolerance to Low Funding and Discovery both reduce the odds of collaboration for Medical sciences dyads.

Taken as a whole, these findings reveal a mix of homophilic effects by scientific field and stress the need to understand homophily in research collaborations as part of the values, traditions, and routines of fields of science. They also stress the need to consider how knowledge stocks and flows, along with other field-specific dynamics, are made sense of and navigated by researchers in each field in a changing body of global science (see Mutz et al., 2015).

Analysis 3 – RQ2a and RQ2b

The third analysis is a hierarchical regression on the propensity for the members of a dyad to have collaborated more than once. This differs from the previous analysis because it excludes collaborations that might have occurred as a one-off partnership and are likely to reflect a continuation of efforts and the establishment of a more consolidated collaboration. The base assumption is that an unsuccessful collaboration (even one resulting in a paper) is unlikely to be repeated. Thus, these analyses are intended to represent ongoing partnerships where it is likely that a relationship of trust, skillset complementarity, mutual interests, and understanding, among other relevant characteristics, was established by the collaborating researchers (see Parker & Kingori, 2016). The results of this analysis are shown in Table 4.

The hierarchical nature of the model is identical to that in Analysis 1. Beginning with the ascribed attributes in Model I, we observe the same effects for Gender and Age as those observed for overall collaborations. The difference is that the effect of Gender is much stronger here: being of the same gender increases the odds of repeated collaboration by 36.5% ($B=0.311$, $OR=1.365$, $p<0.01$). In Model II, which focuses on geographical attributes, the effects are quite similar to those exhibited for overall collaborations, but the effects are stronger and reinforce the relevance of ascribed and geographical attributes to research collaboration homophily. Substantial differences begin to emerge in Model III. Whereas being in a Top-Ranked University increased the odds of engaging in collaborations in Table 2, Model III, it is not significant for repeated collaborations. It may be the case that researchers at top-ranked universities collaborate from time to time with researchers from the same type of university (for sake of reputation, selectivity, or to maintain research possibilities for future collaboration) but establishing longstanding collaborations with researchers at research universities could limit their research potential, opportunities to establish new collaborations, and opportunities to lead research directions (Naik et al., 2023). For these researchers, establishing stable research collaborations outside of the scope of the research universities may provide competitive advantages through collaborations of their own and the creation of networks that they can eventually dominate (Oleksiyenko & Sá, 2010) This effect may be even stronger if these collaborations are with former PhD graduates who are researchers at less reputed universities (see Celis & Kim, 2018). It is also important to remember that researchers at top-ranked universities are frequently requested to collaborate by others (not necessarily from research universities), so finding collaboration opportunities is not a problem, while collaborations bring the opportunity to face new challenges, make outstanding contributions, and reap the benefits of having a

Table 4 Hierarchical logistic regression for Repeated Collaborations

Variables	Model I	Model II	Model III	Model IV	Model V
Gender (identical)	0.311*** (0.043)	0.317*** (0.044)	0.306*** (0.065)	0.292*** (0.067)	0.344*** (0.079)
Age delta	-0.022*** (0.002)	-0.020*** (0.002)	-0.022*** (0.003)	-0.024*** (0.004)	-0.018*** (0.004)
FOS (Agricultural sciences)	3.521*** (0.164)	3.260*** (0.183)	3.189*** (0.245)	3.268*** (0.247)	3.263*** (0.286)
FOS (Engineering & Technology)	2.377*** (0.082)	2.262*** (0.084)	2.254*** (0.129)	2.264*** (0.133)	2.325*** (0.147)
FOS (Medical and Health sciences)	2.451*** (0.056)	2.299*** (0.058)	2.260*** (0.087)	2.225*** (0.090)	2.144*** (0.106)
FOS (Natural sciences)	2.383*** (0.061)	2.379*** (0.062)	2.365*** (0.090)	2.249*** (0.095)	2.259*** (0.110)
FOS (Social sciences)	2.453*** (0.077)	2.282*** (0.079)	2.363*** (0.114)	2.457*** (0.118)	2.532*** (0.131)
Country (identical)		1.876*** (0.050)	1.912*** (0.075)	1.875*** (0.079)	1.891*** (0.090)
University (identical)		1.717*** (0.107)	1.592*** (0.161)	1.595*** (0.169)	1.589*** (0.206)
City (identical)		1.925*** (0.100)	1.937*** (0.148)	1.928*** (0.155)	1.866*** (0.188)
Top-Ranked University (identical)			0.087 (0.071)	0.087 (0.073)	0.049 (0.083)
Academic Sector (identical)			0.099 (0.072)	0.054 (0.075)	-0.025 (0.086)
Educational Inbreeding (identical)			0.020 (0.062)	0.018 (0.064)	0.053 (0.074)
Job Changes delta				-0.018 (0.023)	-0.016 (0.026)
Job Country Changes delta				0.070** (0.032)	0.080** (0.036)
Publications delta				0.001*** (0.000)	0.001** (0.000)
Citations delta				0.000** (0.000)	0.000* (0.000)
Percentage Research Funding delta				-0.007*** (0.001)	-0.007*** (0.001)
Percentage Teaching delta				-0.004*** (0.001)	-0.004*** (0.001)
Prestige delta					-0.015 (0.043)
Drive to Publish delta					-0.038 (0.040)
Branching Out delta					0.055 (0.041)

Table 4 (continued)

Variables	Model I	Model II	Model III	Model IV	Model V
Multidisciplinarity delta					-0.072* (0.042)
Willingness to Collaborate delta					-0.000 (0.050)
Invited to Collaborate delta					-0.038 (0.045)
Field Orientation delta					-0.025 (0.041)
Institution Orientation delta					-0.031 (0.035)
Society Orientation delta					-0.118*** (0.035)
Non-Academic Orientation delta					-0.091** (0.040)
Mentor Influence delta					-0.004 (0.033)
Tolerance to Low Funding delta					-0.052 (0.034)
Discovery delta					0.032 (0.050)
Pseudo R ²	0.072	0.174	0.176	0.180	0.175
Observations	11,768,526	11,155,726	5,172,936	4,853,170	3,821,230

leading position in such collaborative networks (Pfothenauer et al., 2013). The prevalence of focusing research collaborations within the academic sector is also not statistically significant (as it was concerning collaborations overall). This result may suggest the greater involvement of academics in research projects with researchers and non-researchers based in other sectors of activity, as opposed to only collaborating with other academics, possibly reflecting the outcomes of policy rhetoric toward more engaged universities, university research evaluations focused on research impact, and incentives (and associated funding) to set and maintain research collaborations with other sectors of activity to foster knowledge production and exchange (Horta, 2022).

The findings regarding the variables in Model IV are similar to those obtained for overall collaborations; the only difference of note is that Job Changes are no longer significant, indicating that mobility might be important for initially engaging in collaborations, but it is not necessarily relevant to sustaining them. All other variables have effects that are identical to those described in Analysis 1. In Model V, which introduces acquired attributes, several differences of note also emerge. First, being Invited to Collaborate and pursuing Discovery-driven Agendas, both variables that decrease the odds of collaboration, are no longer relevant when we consider repeated collaborations. Second, Society Orientation and Non-academic Orientation maintain their previous effects. Third, a previously non-significant variable, Multidisciplinary, now plays a role, decreasing the odds for repeated collaborations by 7% per unit increase in difference ($B = -0.072$, $OR = 0.930$, $p < 0.1$). The

explanation for this may lie in the fact that researchers may try a multidisciplinary research collaboration once (despite one or several having a multidisciplinary research agenda, whereas other researchers do not), but the effort may be completely expended in that collaboration, and from then on, the differentials in research agenda concerning multidisciplinary stances make further collaborations less likely.

Overall, in terms of hierarchy, geographical attributes are again the most important aspect, this time in predicting the odds of repeated collaborations. The inclusion of the other models does not improve the explanatory power significantly, even though many of their variables are statistically significant and thus help explain the phenomenon of research collaboration homophily.

Analysis 4 – RQ2c

In this final analysis, we compare Model V from the previous analysis of repeated collaborations across the various disciplinary dyads. Similar to Analysis 2, we focus more on the differences across fields to avoid repetition of the previous analysis. The results of this exercise are summarized in Table 5.

Similar to the findings concerning overall collaborations, the effect of Gender shows differentiated effects for Multidisciplinary dyads and for dyads from the Natural and Social Sciences. Medical Sciences also emerges as significant in this analysis, with same-gender researchers in this field being 30% more likely to repeat collaborative ventures ($B=0.265$, $OR=1.303$, $p<0.01$) than different-gender researchers. The remaining effects of Gender are similar to those previously observed. Age, likewise, has a similar effect to that previously observed, but with a small difference: for repeated collaborations, the age differential is no longer significant for Engineering and the Natural Sciences. The geographical attributes again reveal their relevance and consistency, and they are the most important predictors of research collaboration homophily for all fields of science.

Educational Inbreeding has a differentiated effect not seen in the first analysis. Inbreeding at the educational level reduces the odds of collaboration by 22% in the Medical sciences ($B=-0.259$, $OR=0.772$, $p<0.1$), but actually increases the odds of collaboration by 31% in the Natural sciences ($B=0.276$, $OR=1.317$, $p<0.1$). Job Country Changes has a differentiated field effect that is similar to that observed in the overall collaboration analysis, but for repeated collaborations we also now see a new and negative effect—for Engineering, asymmetry in international mobility reduces the odds of collaboration by 26% per unit of difference ($B=0.290$, $OR=0.748$, $p<0.1$). Publications and citations also matter differently by field, but even when they are significant, they maintain their rather modest effects.

Percentage of Research Funding also exhibits a differentiated effect similar to global collaborations; in this case, the key difference is that whereas asymmetry previously reduced the odds of collaboration in the Natural sciences, for repeated collaboration asymmetry at this level no longer matters. Likewise, Percentage of Teaching, which previously reduced the odds of collaboration for Agricultural Sciences in terms of global collaborations, is no longer significant for repeated collaborations.

Finally, we consider acquired attributes. We begin by observing the effect of Prestige in Agricultural Sciences, an effect that was present for global collaborations, and so we will not revisit it. However, Drive to Publish also now has a negative effect. Asymmetry in Drive to Publish reduces the odds of repeated collaboration by 55% per unit of difference, exclusively in the Agricultural Sciences ($B=-0.801$, $OR=0.448$, $p<0.05$). However,

Table 5 Logistic regression for Repeated Collaboration, by Field of Science

Variables	Multidisc	Agricultural	Engineering	Medical	Natural	Social
Gender (identical)	0.477*** (0.169)	-0.506 (0.544)	0.040 (0.266)	0.265* (0.140)	0.276* (0.160)	0.799*** (0.228)
Age delta	-0.015* (0.008)	-0.106** (0.045)	-0.020 (0.013)	-0.022*** (0.008)	-0.006 (0.007)	-0.034*** (0.013)
Country (identical)	1.849*** (0.209)	3.971*** (0.735)	2.381*** (0.319)	1.904*** (0.167)	1.815*** (0.169)	1.749*** (0.244)
University (identical)	1.561*** (0.442)	-1.950 (1.292)	1.247** (0.585)	1.837*** (0.340)	1.660*** (0.485)	1.331** (0.561)
City (identical)	2.366*** (0.411)	2.656** (1.045)	2.057*** (0.590)	1.810*** (0.318)	1.042** (0.430)	2.342*** (0.505)
Top Ranked U. (identical)	0.170 (0.183)	0.216 (0.734)	0.485 (0.382)	-0.109 (0.156)	-0.013 (0.156)	-0.070 (0.220)
Academic Sector (identical)	0.048 (0.189)	-0.179 (0.513)	0.017 (0.332)	0.027 (0.154)	-0.005 (0.159)	-0.268 (0.277)
Educ. Inbreeding (identical)	0.141 (0.159)	0.536 (0.531)	-0.034 (0.254)	-0.259* (0.139)	0.276* (0.143)	0.065 (0.212)
Job Changes delta	0.012 (0.057)	-0.255 (0.200)	-0.088 (0.101)	0.017 (0.046)	-0.051 (0.052)	-0.021 (0.079)
Job Country Changes delta	-0.059 (0.086)	0.539** (0.219)	-0.290* (0.169)	0.157** (0.064)	0.131* (0.068)	0.051 (0.102)
Publications delta	0.001 (0.001)	0.007 (0.005)	0.004*** (0.001)	0.002*** (0.001)	-0.001 (0.001)	-0.002 (0.003)
Citations delta	0.000*** (0.000)	0.000 (0.000)	-0.000 (0.000)	-0.000 (0.000)	0.000 (0.000)	-0.000 (0.000)
Percentage Research Funding delta	-0.009*** (0.003)	-0.000 (0.012)	-0.005 (0.006)	-0.010*** (0.003)	-0.004 (0.003)	-0.005 (0.004)
Percentage Teaching delta	0.001 (0.003)	-0.013 (0.008)	-0.004 (0.004)	-0.001 (0.002)	-0.007*** (0.002)	-0.015*** (0.004)
Prestige delta	0.110 (0.084)	0.643*** (0.219)	-0.071 (0.157)	-0.127 (0.083)	-0.130 (0.086)	0.131 (0.116)
Drive to Publish delta	-0.061 (0.084)	-0.801** (0.330)	0.016 (0.137)	-0.090 (0.075)	0.069 (0.084)	-0.086 (0.097)
Branching Out delta	0.116 (0.086)	0.225 (0.267)	-0.023 (0.161)	0.008 (0.073)	0.063 (0.092)	0.052 (0.096)
Multidisciplinarity delta	-0.069 (0.093)	0.173 (0.340)	-0.174 (0.135)	0.043 (0.081)	-0.236*** (0.083)	0.079 (0.108)
Willingness to Collaborate delta	0.114 (0.095)	-0.070 (0.356)	-0.052 (0.176)	-0.019 (0.093)	0.039 (0.105)	-0.146 (0.140)
Invited to Collaborate delta	-0.284*** (0.097)	0.414 (0.385)	0.051 (0.143)	-0.058 (0.098)	0.025 (0.089)	0.027 (0.096)
Field Orientation delta	-0.015 (0.086)	-0.474* (0.249)	-0.069 (0.155)	0.061 (0.076)	-0.045 (0.083)	-0.053 (0.114)
Institution Orientation delta	-0.071 (0.083)	0.217 (0.194)	0.077 (0.121)	-0.129* (0.068)	0.022 (0.064)	-0.027 (0.095)

Table 5 (continued)

Variables	Multidisc	Agricultural	Engineering	Medical	Natural	Social
Society Orientation delta	-0.013 (0.066)	0.097 (0.267)	-0.002 (0.135)	-0.129* (0.074)	-0.232*** (0.065)	-0.116 (0.098)
Non-academic Orientation delta	-0.163* (0.085)	-0.116 (0.203)	-0.077 (0.164)	-0.015 (0.073)	-0.079 (0.072)	-0.126 (0.116)
Mentor Influence delta	-0.008 (0.081)	0.120 (0.271)	-0.088 (0.093)	-0.053 (0.062)	-0.009 (0.062)	0.111 (0.082)
Tolerance to Low Funding delta	-0.072 (0.070)	0.151 (0.174)	0.075 (0.123)	-0.135** (0.062)	-0.062 (0.072)	0.052 (0.086)
Discovery delta	0.003 (0.107)	-0.075 (0.409)	0.209 (0.165)	0.017 (0.093)	-0.072 (0.105)	0.263* (0.136)
Pseudo R ²	0.125	0.373	0.207	0.146	0.071	0.144
Observations	2,888,297	7,750	98,346	369,370	334,971	122,496

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Regression coefficients are shown with robust standard errors in parentheses

Multidisciplinary, which was also an important predictor of overall collaborations for Agricultural Sciences, is no longer significant when it comes to repeated collaborations. Multidisciplinary has the same effect as previously observed for the Natural Sciences. Invited to Collaborate essentially maintains the same role as in overall collaborations, with a negative effect exclusive to the multidisciplinary dyads. Field Orientation has an effect exclusive to Agricultural Sciences, reducing the odds of collaboration by 38% per unit of difference ($B = -0.474$, $OR = 0.622$, $p < 0.1$). Society Orientation is no longer relevant for this same dyad. It has a negative effect for the Natural sciences, but it also reduces the odds of repeated collaboration in the Medical sciences ($B = 0.129$, $OR = 0.878$, $p < 0.1$). Finally, Tolerance to Low Funding has the same effect as observed for overall collaborations when it comes to the Natural sciences.

Overall, these findings underline different dynamics and variations across scientific fields concerning the determinants of homophily in research collaborations, thus highlighting the need for analyses that include both global perspectives on science and observations of separate fields of science. The complexity of our findings between overall collaborations and repeated collaborations in each field of science further underlines the need to understand the specific dynamics, values, and working traditions of each field to be able to ascertain more definitive interpretations of these findings.

Conclusions

This study contributes to the knowledge of homophily as an important aspect of scientific collaboration by expanding on previous analyses and the literature, along with bringing new perspectives and findings. The first noteworthy finding is that geographical attributes, specifically taking the form of physical proximity at any level (university, city, and country in our analysis), are the most important driver of collaboration in any field of science. This finding has been noted in previous studies (e.g., Akbaritabar & Barbato, 2021; Bergé, 2017; Evans et al., 2011), but this is the first study to identify geographical attributes as a

phenomenon that is not bound by scientific disciplines. It is the most influential factor for research collaboration homophily across the board, independent of values, norms, routines, or any other specific dynamics associated with different fields of science. Indeed, this is one of the few effects that is truly consistent across disciplines. Even gender homophily, which is consistently documented as one of the key aspects of homophily in scientific collaboration (e.g., Abramo et al., 2013), is shown here to be specific to multidisciplinary collaborations and to collaborations within the natural sciences and the social sciences (being particularly notable in the latter). Gender is also a key homophilic determinant of research collaborations in the medical sciences when it comes to repeated collaborations. The other ascribed trait, Age, is also a relatively universal determinant of homophilic research collaborations across the board, except for Engineering and only regarding repeated collaborations. In all other cases, age differentials decrease the odds of collaboration, indicating that researchers tend to gravitate toward collaborating with those within their own age range. This finding may seem to contradict two standing narratives. The first narrative is that there is a growing number of research collaborations between PhD students (likely to be younger) and their supervisors (likely to be older) (Abbasi et al., 2012). The second narrative is that lesser-known researchers tend to gravitate toward those with more prestige (who are likely to be older) and collaborate with them to raise their research profile even when facing entry into selective research collaboration networks (e.g., Wagner et al., 2015). Some tendencies of these narratives are observed in the presence of asymmetry in publications and citations contributing to increased chances of collaboration, possibly reflecting a tendency among less-established researchers (including PhD students) to seek partnerships with more prolific counterparts, but the effect of these variables is negligible.

Another notable finding is that the attributes that significantly predict overall collaborations are mostly the same as those that predict repeated collaborations, suggesting that the mechanism that initiates the partnership is similar to the one that maintains it. There are, of course, notable exceptions. For example, educational inbreeding tends to contribute to overall collaborations in multidisciplinary and social sciences dyads but has no effect in maintaining them. For repeated collaborations, educational inbreeding actually hinders collaborations in the medical sciences, despite playing a role in maintaining collaborations in the natural sciences. Being in a top-ranked university is also a strong predictor of collaborations in the natural sciences, but not in maintaining those collaborations. This suggests that educational inbreeding and working at a top-ranked university might be useful for bringing researchers together, but they play no active role in maintaining frequent research collaborations for the reasons explained in the discussion of the findings of Analysis 3, although some others that are more related to social interaction or practical issues pertaining to the research process may be of importance.

Differences in career history attributes, such as the number of jobs held and international job mobility, also impact collaboration likelihood in complex ways. While job variety reduces the odds of collaboration, international job mobility appears to enhance them. Those who are internationally mobile tend to collaborate more with the less internationally mobile. This can have two possible explanations. First, this may relate to diaspora networks whereas internationally mobile researchers establish collaborations with researchers in their home country to contribute to the scientific capacity of or to maintain scientific links with the home country (Langa et al., 2018). Second, expat researchers may have an interest in collaborating with researchers of the host country to have an easier access to national and local funding, other resources, knowledge and data, or to be better integrated into the hosting country national scientific community; it is likely that researchers in the hosting country also benefit from collaborating with

these expat researchers to better tap into global flows of knowledge, resources and international scientific communities and networks (Wang et al., 2019). However, additional granularity is found when we compare these effects across fields of science. Notably, international job changes enhance collaboration within agricultural, medical, and natural sciences, but deter repeated collaborations within engineering (while maintaining their effect on the aforementioned fields). This finding suggests that international career mobility can simultaneously open new collaboration opportunities and interrupt existing collaborative partnerships, which is not surprising given the great importance of geographical attributes, as noted above. The literature also seems to give credence to this pattern and our interpretation of these findings (Wang et al., 2019).

Overall, this study underscores the complex ways in which ascribed, geographical, career, educational, and acquired attributes shape collaborations. Some attributes have consistent effects for both overall and repeated research collaborations, whereas many others display discipline-specific effects, highlighting the context-dependent nature of research collaborations and the role homophily plays in establishing research collaborations in different environments and contexts. These findings shed light on the need to tackle future studies on homophily from both multidisciplinary and disciplinary perspectives and to consider distinguishing between three categories of homophilic research collaboration attributes (two of which are assessed in this study): 1) overall (referring to the existence of a research collaboration independently from the frequency of collaborations); 2) initiators (one-time research collaboration only); and 3) maintainers (sustainable research collaborations, defined as those including two or more collaborations between the same researchers). This study focuses only on the overall collaborators and maintainers, but future research may also include initiators, because from a practical perspective, this study also highlights the necessity of nuanced strategies to foster and sustain collaborations across different scientific fields.

To conclude, some considerations should be made regarding possible limitations that the analysis may have. First, in terms of the effects of geographical proximity, it may be possible that due to the recent pandemic, many researchers have shifted their way of collaboration and adopted remote collaborative tools more frequently to collaborate. However, there are studies suggesting that the increased use of remote collaborative tools for research during the pandemic may lessen in the near future or even revert to pre-pandemic levels (Ziamba & Eisenhardt, 2022). The pandemic shift was not manifested in our data, but future studies aiming to assess the effects of the pandemic on post-pandemic research collaborations can shed further light on how the adoption of remote collaboration tools may have changed the importance of geographical proximity. Second, our study does not cover all possible variables which can potentially influence the initiation or the continuation of a previous collaboration—human behavior is complex and there are likely many other factors at play, perhaps more than can realistically be stochastically modeled. To understand the reasonings that led to a research collaboration, the participants of a survey would need to be asked about the motivations for collaborating in every collaboration in the dataset, which would not be realistic in terms of feasibility.

Acknowledgements This work was supported by the Fundação para a Ciência e a Tecnologia (FCT) under Project 2020.03729.CEECIND and by the Research Grants Council, University Grants Committee, Hong Kong SAR (Grant Number 17604119).

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long

as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Abbasi, A., Hossain, L., & Leydesdorff, L. (2012). Betweenness centrality as a driver of preferential attachment in the evolution of research collaboration networks. *Journal of Informetrics*, 6(3), 403–412.
- Abramo, G., D'Angelo, C. A., & Murgia, G. (2013). Gender differences in research collaboration. *Journal of Informetrics*, 7(4), 811–822.
- Akbaritabar, A., & Barbato, G. (2021). An internationalized Europe and regionally focused Americas: A network analysis of higher education studies. *European Journal of Education*, 56(2), 219–234.
- Bergé, L. R. (2017). Network proximity in the geography of research collaboration. *Papers in Regional Science*, 96(4), 785–815.
- Celis, S., & Kim, J. (2018). The making of homophilic networks in international research collaborations: A global perspective from Chilean and Korean engineering. *Research Policy*, 47(3), 573–582.
- Ebadi, A., & Schiffauerova, A. (2015). How to receive more funding for your research? Get connected to the right people! *PLoS ONE*, 10(7), e0133061.
- Evans, T. S., Lambiotte, R., & Panzarasa, P. (2011). Community structure and patterns of scientific collaboration in business and management. *Scientometrics*, 89(1), 381–396. <https://doi.org/10.1007/s11192-011-0439-1>
- Feinberg, G., Watnick, B., & Sacks, A. (2011). Solo vs. collaborative research in the social sciences and higher education: unraveling the realities of male-female research publication patterns in the context of gender politics and social justice issues. *Journal of Multidisciplinary Research*, 3(3), 47–63.
- Feng, S., & Kirkley, A. (2020). Mixing patterns in interdisciplinary co-authorship networks at multiple scales. *Scientific Reports*. <https://doi.org/10.1038/s41598-020-64351-3>
- Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, 80(1), 27–38. <https://doi.org/10.1093/biomet/80.1.27>
- González Brambila, C. N., & Olivares-Vázquez, J. L. (2021). Patterns and evolution of publication and co-authorship in Social Sciences in Mexico. *Scientometrics*, 126(3), 2595–2626. <https://doi.org/10.1007/s11192-020-03644-w>
- Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2014). Multivariate data analysis. *Pearson Education Limited*. <https://doi.org/10.1108/RAUSP-05-2019-0098>
- Henriksen, D. (2018). What factors are associated with increasing co-authorship in the social sciences? A case study of danish economics and political science. *Scientometrics*, 114(3), 1395–1421. <https://doi.org/10.1007/s11192-017-2635-0>
- Holman, L., & Morandin, C. (2019). Researchers collaborate with same-gendered colleagues more often than expected across the life sciences. *PLoS ONE*, 14(4), e0216128. <https://doi.org/10.1371/journal.pone.0216128>
- Horta, H. (2022). Trust and incentives in academic research and the position of universities within innovation systems. *Higher Education*, 84, 1343–1363.
- Horta, H., Feng, S., & Santos, J. M. (2022). Homophily in higher education research: A perspective based on co-authorships. *Scientometrics*, 127(1), 523–543. <https://doi.org/10.1007/s11192-021-04227-z>
- Horta, H., & Santos, J. M. (2020). The Multidimensional research agendas inventory—revised (MDRAI-R): Factors shaping researchers' research agendas in all fields of knowledge. *Quantitative Science Studies*, 1(1), 60–93. https://doi.org/10.1162/qss_a_00017
- Huang, J. S. (2014). Building research collaboration networks—an interpersonal perspective for research capacity building. *Journal of Research Administration*, 45(2), 89–112.
- Hunter, L., & Leahey, E. (2008). Collaborative research in sociology: Trends and contributing factors. *The American Sociologist*, 39(4), 290–306. <https://doi.org/10.1007/s12108-008-9042-1>
- Jeong, S., Choi, J. Y., & Kim, J. (2011). The determinants of research collaboration modes: Exploring the effects of research and researcher characteristics on co-authorship. *Scientometrics*, 89(3), 967–983. <https://doi.org/10.1007/s11192-011-0474-y>
- Katz, J. S., & Martin, B. R. (1997). What is research collaboration? *Research Policy*, 26(1), 1–18.

- King, G., & Zeng, L. (2001). Logistic regression in rare events data. *Political Analysis*, 9(2), 137–163. <https://doi.org/10.1093/oxfordjournals.pan.a004868>
- Kwiek, M. (2018). International research collaboration and international research orientation: Comparative findings about European academics. *Journal of Studies in International Education*, 22(2), 136–160.
- Kwiek, M. (2020). Internationalists and locals: International research collaboration in a resource-poor system. *Scientometrics*, 124, 57–105.
- Kwiek, M. (2021). What large-scale publication and citation data tell us about international research collaboration in Europe: Changing national patterns in global contexts. *Studies in Higher Education*, 46(12), 2629–2649.
- Langa, P. V., Swanzy, P., & Uetela, P. (2018). Advancing collaboration between african diaspora and africa-based scholars: Extracts of interviews with selected african diaspora scholars. *Journal of Higher Education in Africa*, 16(1/2), 135–155.
- Lauto, G., & Valentin, F. (2013). How large-scale research facilities connect to global research. *Review of Policy Research*, 30(4), 381–408. <https://doi.org/10.1111/ropr.12027>
- Lazarsfeld, P. F., & Merton, R. K. (1954). Friendship as a social process: A substantive and methodological analysis. *Freedom and Control in Modern Society*, 18(1), 18–66.
- Mutz, R., Bornmann, L., & Daniel, H. D. (2015). Cross-disciplinary research: What configurations of fields of science are found in grant proposals today? *Research Evaluation*, 24(1), 30–36.
- Niak, C., Sugimoto, C. R., Larivière, V., Leng, C., & Guo, W. (2023). Impact of geographic diversity on citation of collaborative research. *Quantitative Science Studies*. https://doi.org/10.1162/qss_a_00248
- Oleksiyenko, A., & Sá, C. M. (2010). Resource asymmetries and cumulative advantages: Canadian and US research universities and the field of global health. *Higher Education*, 59, 367–385.
- Ozel, B., Kretschmer, H., & Kretschmer, T. (2014). Co-authorship pair distribution patterns by gender. *Scientometrics*, 98, 703–723.
- Parker, M., & Kingori, P. (2016). Good and bad research collaborations: Researchers' views on science and ethics in global health research. *PLoS ONE*, 11(10), e0163579. <https://doi.org/10.1371/journal.pone.0163579>
- Parr, D. M., Texler, C. J., Khanna, N. R., & Battisti, B. T. (2007). Designing sustainable agriculture education: Academics' suggestions for an undergraduate curriculum at a land grant university. *Agriculture and Human Values*, 24, 523–533.
- Pfotenhauer, S. M., Jacobs, J. S., Pertuze, J. A., Newman, D. J., & Roos, D. T. (2013). Seeding change through international university partnerships: The MIT-Portugal program as a driver of internationalization, networking and innovation. *Higher Education Policy*, 26, 217–242.
- Ryan, L. (2015). Friendship-making: Exploring network formations through the narratives of Irish highly qualified migrants in Britain. *Journal of Ethnic and Migration Studies*, 41(10), 1664–1683.
- Santos, J. M., & Horta, H. (2018). The research agenda setting of higher education researchers. *Higher Education*, 76, 649–668.
- Tavares, O., Sin, C., Sá, C., Bugla, S., & Amaral, A. (2022). Inbreeding and research collaborations in Portuguese higher education. *Higher Education Quarterly*, 76(1), 102–115. <https://doi.org/10.1111/hequ.12301>
- Ubfal, D., & Maffioli, A. (2011). The impact of funding on research collaboration: Evidence from a developing country. *Research Policy*, 40(9), 1269–1279.
- Uddin, S., Hossain, L., & Rasmussen, K. (2013). Network effects on scientific collaborations. *PLoS ONE*, 8, 2.
- Wagner, C. S., Horlings, E., Whetsell, T. A., Mattsson, P., & Nordqvist, K. (2015). Do Nobel laureates create prize-winning networks? An analysis of collaborative research in physiology or medicine. *PLoS ONE*, 10(8), e0136478. <https://doi.org/10.1371/journal.pone.0136478>
- Wang, J., Hooi, R., Li, A. X., & Chou, M. (2019). Collaboration patterns of mobile academics: The impact of international mobility. *Science and Public Policy*, 46(3), 450–462.
- Xie, W. J., Li, M. X., Jiang, Z. Q., Tan, Q. Z., Podobnik, B., Zhou, W. X., & Stanley, H. E. (2016). Skill complementarity enhances heterophily in collaboration networks. *Scientific Reports*, 6, 18727.
- Xu, X. (2020). China 'goes out' in a centre-periphery world: Incentivizing international publications in the humanities and social sciences. *Higher Education*, 80(1), 157–172. <https://doi.org/10.1007/s10734-019-00470-9>
- Youtie, J., & Bozeman, B. (2014). Social dynamics of research collaboration: Norms, practices, and ethical issues in determining co-authorship rights. *Scientometrics*, 101, 953–962.
- Zeng, X. H. T., Duch, J., Sales-Pardo, M., Moreira, J. A. G., Radicchi, F., Ribeiro, H. V., Woodruff, T. K., & Amaral, L. A. N. (2016). Differences in collaboration patterns across discipline, career stage, and gender. *PLOS Biology*, 14(11), e1002573. <https://doi.org/10.1371/journal.pbio.1002573>

Ziembra, E. W., & Eisenhardt, M. (2022). The effect of Covid-19 pandemic on ICT usage by academics. *Journal of Computer Information Systems*, 62(6), 1154–1168.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.