

Language and geopolitical distance across multiple margins of international scientific collaboration

Yizheng He

Chengdu University of Information Technology

Ming Li

53489641@qq.com

Chengdu University of Information Technology

Article

Keywords: international scientific collaboration, language distance, geopolitical distance, collaboration networks, extensive margin, intensive margin

Posted Date: October 26th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11113197/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Language and geopolitical distance across multiple margins of international scientific collaboration

Yizheng He¹, Ming Li^{1*}

Short title: Language and geopolitical distance in scientific collaboration

¹ School of Statistics, Chengdu University of Information Technology

No. 10 Xingfu Road, Longquanyi District, Chengdu 610103, Sichuan, China

* Corresponding author: Ming Li

Email: 53489641@qq.com

Yizheng He: 1801511029@qq.com | ORCID: <https://orcid.org/0009-0008-2790-281X>

Abstract

International scientific collaboration is associated with both linguistic and geopolitical distance, while bilateral collaboration rates alone obscure whether networks adjust through relationship formation, maintenance, termination, scale, or intensity. Using English-language DOI journal articles indexed in OpenAlex from 2018 to 2025, we construct a bilateral country-pair panel spanning economics, mathematics, and medicine, with 33 focal countries and 173 partner countries. Six outcomes capture relative partner allocation, absolute bilateral coauthored output, active collaboration, entry, exit, and intensity among active ties. Standardized linguistic distance and pre-period geopolitical distance enter jointly in trajectory and state-specific transition models, with direct coefficient-comparison tests and robustness checks. Linguistic-distance associations follow the expected direction on all six margins and are generally stable. Geopolitical-distance evidence is concentrated in relative partner allocation, collaboration entry, and active-tie intensity, with greater sensitivity to network construction and data coverage. Direct tests show substantial overlap: collaboration exit is the only margin with a coefficient difference at the 5% level. The multi-margin framework reveals how the two distances are associated with multiple forms of network adjustment and provides more informative indicators for science and technology policy monitoring.

Keywords: international scientific collaboration; language distance; geopolitical distance; collaboration networks; extensive margin; intensive margin

1. Introduction

International scientific collaboration has become an important organizational form of knowledge production and scientific innovation. As research activities increasingly cross national borders, linguistic differences, geographic distance, political relations, and institutional environments among partners are also increasingly likely to affect the formation and maintenance of research teams. Existing studies have documented the long-run expansion of international scientific collaboration networks and have shown that geography, culture, political relations, sanctions, and institutional conditions are all closely associated with cross-national scientific collaboration (Frame and Carpenter, 1979; Wagner and Leydesdorff, 2005; Jones et al., 2008; Hoekman et al., 2010; Adams, 2012; Freeman et al., 2015; Hou et al., 2021; Makkonen and Mitze, 2023; Charpin et al., 2024; Wu et al., 2024; Gadgin Matha et al., 2026; Zhao, 2026). At the same time, language barriers are also considered to increase the costs of communication, mutual understanding, and coordination within cross-national research teams and to be associated with participation in collaboration and research performance (Hwang, 2013; Cao et al., 2024). Against the backdrop of deepening global scientific collaboration and an increasingly complex external environment, understanding the relationship of linguistic distance and geopolitical distance with international scientific collaboration has become an important question for analyzing changes in the organization of global knowledge production.

Bilateral collaboration rates do not reveal how scientific collaboration networks adjust. The proportion of a country's research output coauthored with a partner can fall because fewer relationships form, existing relationships end, bilateral publication volume contracts, activity weakens within continuing ties, or the focal country's total output changes. These processes have different empirical meanings, so the analysis separates relationship formation and termination from changes in absolute scale, relative allocation, and collaboration intensity.

Comparing linguistic and geopolitical distance requires attention to how each may relate to several collaboration margins. Linguistic distance proxies communication and coordination frictions in activities such as discussing research questions, communicating technical concepts, allocating tasks,

and drafting and revising manuscripts (Hwang, 2013; Cao et al., 2024). Geopolitical distance captures differences in the political and institutional positions surrounding cross-border research relationships (Hou et al., 2021; Charpin et al., 2024; Gadgin Matha et al., 2026). Neither measure maps to a fixed mechanism: both may be associated with relationship formation, continuation, termination, and intensity. A unified bilateral country-pair framework makes their outcome profiles directly comparable.

An empirical comparison needs outcome measures that separate network margins, continuous bilateral measures of language and geopolitics, and joint estimation on the same sample. Separate statistical significance does not establish a difference between coefficients because their uncertainty is correlated within the joint model. The comparison therefore uses direct coefficient-difference tests based on the estimated covariance.

These distinctions also matter for science and technology policy and research monitoring. Tracking entry, exit, bilateral volume, and active-tie intensity alongside relative partner allocation identifies where network adjustment occurs and makes collaboration indicators more informative.

We use English-language DOI papers indexed in OpenAlex from 2018 to 2025 to construct a focal-country–partner-country–discipline–year panel. The sample covers economics, mathematics, and medicine and includes 33 focal countries, 173 partner countries, and 136,224 panel observations. The baseline network is directed by the first author's country. Six outcomes describe relative partner allocation, absolute bilateral coauthored output, active collaboration, entry, exit, and intensity among active ties.

The models include standardized bilateral linguistic distance and pre-period geopolitical distance jointly. Linguistic distance is constructed as one minus the equally weighted mean of DICT's *cnl* and *lpn* measures, and geopolitical distance is the pre-period difference in United Nations General Assembly ideal points. Partner collaboration rate, bilateral coauthored papers, active collaboration, and active-tie intensity are estimated with country-pair trajectory models that include focal-country–partner-country–discipline, focal-country–discipline–year, and partner-country–discipline–year fixed effects. State-specific transition models estimate entry and exit within risk sets defined by the previous period's collaboration status. Within-model covariance provides direct tests of coefficient differences. Robustness checks cover alternative network constructions, full and fractional counting, hyperauthorship exclusion, finite-cluster inference, leave-one-focal-country-out analysis, observable-coverage restrictions, and supplementary PPML sensitivity analysis.

Linguistic distance has a broad and consistent outcome profile across all six collaboration margins. Greater distance is associated with a more negative trajectory in partner collaboration rates, fewer bilateral coauthored papers, lower active-collaboration status and entry probabilities, higher exit probabilities, and weaker intensity among active ties. Geopolitical-distance evidence is concentrated in relative partner allocation, collaboration entry, and active-tie intensity, and some estimates are more sensitive to network construction and observable-quality restrictions. Direct coefficient-difference tests show substantial overlap across most margins. Only collaboration exit rejects equality of the two coefficients at the 5% level, yielding an exit-specific difference rather than an overall ranking.

The central contribution is a unified multi-margin comparison of linguistic and geopolitical distance in international scientific collaboration. The framework separates relative allocation, absolute scale, active status, entry, exit, and intensity among active ties, while estimating both continuous distance measures in the same country-pair models. Coefficient differences are evaluated directly from the covariance in the joint model. Alternative network constructions, counting methods, sample restrictions, and inference procedures distinguish stable patterns from specification-sensitive results.

The observational design estimates conditional associations. Fixed effects, pre-period distance measures, and control-variable trends absorb stable differences and common changes, while time-varying unobserved bilateral factors may remain. The coefficients therefore describe how linguistic and geopolitical distance are associated with collaboration trajectories and transition probabilities; they are not causal effects identified from exogenous variation. The analysis asks which margins carry each association and where the two coefficients differ statistically.

2. Literature Review

2.1 International Scientific Collaboration and Network Structure

International scientific collaboration is an important organizational form through which knowledge production crosses national and institutional boundaries. Early research focused mainly on the expansion of international scientific collaboration and the evolution of its network structure. Frame and Carpenter (1979) conducted an early systematic study of international scientific collaboration; Wagner and Leydesdorff (2005) analyzed its growth from the perspectives of network structure and self-organization; and Jones et al. (2008), Adams (2012), and Freeman et al. (2015) further discussed trends in scientific collaboration from the perspectives of research teams, research networks, and the organization of collaboration. Collectively, these studies show that research activities increasingly depend on collaborative networks spanning institutions, regions, and national borders.

As international scientific collaboration has deepened, research has gradually shifted from the overall scale of collaboration toward specific factors shaping the formation and evolution of collaboration networks. Hoekman et al. (2010) found that, despite continued advances in remote communication and collaboration technologies, spatial distance remains closely associated with scientific collaboration and that this relationship varies over time and across disciplines. Liu et al. (2023), focusing on the structure of collaborative relationships, discussed the distinct roles of strong and weak ties in international scientific collaboration networks. These findings indicate that examining only the overall scale or degree of internationalization cannot fully capture structural changes within scientific collaboration networks.

At the country-pair level, scientific collaboration can adjust through relationship formation or termination and through changes in output or intensity within continuing ties. Relative partner allocation, absolute bilateral volume, active status, entry, exit, and active-tie intensity therefore describe different features of the network.

2.2 Linguistic Distance and International Scientific Collaboration

Language is an important medium for knowledge exchange, task coordination, and manuscript writing in cross-national scientific collaboration; linguistic differences may therefore constitute communication and coordination frictions. During collaboration, researchers need to communicate continuously about research questions, technical concepts, task allocation, and manuscript drafting and revision, and linguistic differences may increase the costs of understanding and coordination in this process.

Existing studies have examined the relationship between linguistic factors and cross-border collaboration from multiple perspectives. Hwang (2013) linked language barriers to participation, organizational integrity, collaborative performance, and interorganizational relationships in international scientific collaboration. Cao et al. (2024) further used bibliometric data to study linguistic distance and found a significant association between distance from English and research performance. The importance of linguistic factors is not confined to scientific research. In the context

of cross-border economic activity, Melitz and Toubal (2014) showed that common native language, common spoken language, and translation capacity all influence the organization of cross-border exchange. These studies indicate that language is not merely a background characteristic of international interaction but may be closely associated with the communication costs faced in cross-border exchange.

In recent years, the measurement of linguistic differences has also expanded from discrete indicators of whether countries share a common language to continuous measures of linguistic connectivity across countries. Gurevich et al. (2025) constructed data on linguistic connectivity within and between countries, providing a new measurement basis for characterizing linguistic differences at the country-pair level. Compared with a common-language dummy variable, continuous measures can capture variation in linguistic connectivity across country pairs and provide a more detailed basis for comparing the relationship between linguistic distance and international scientific collaboration.

Country-pair linguistic distance is examined here as a proxy for average communication and coordination frictions across six observable margins, extending prior work focused on participation, performance, or aggregate collaboration.

2.3 Geopolitical Factors and International Scientific Collaboration

In addition to linguistic factors, political and institutional relations between countries are an important topic in research on international scientific collaboration. Although scientific collaboration primarily serves knowledge production and academic exchange, cross-national collaboration still occurs within specific international political and institutional environments. Political relations, sanctions, and geopolitical conflict may therefore be associated with changes in scientific collaboration networks.

Existing studies have examined this issue from different perspectives. Hou et al. (2021) comprehensively analyzed the relationships of scientific, economic, geopolitical, and cultural factors with international scientific collaboration, showing that geopolitical conditions are an important dimension for understanding such collaboration. Makkonen and Mitze (2023), in the context of collaboration between the European Union and Russia, studied the relationship of geopolitical conflict and sanctions with international knowledge flows. Charpin et al. (2024) further examined the relationship between geopolitical tensions and changes in international scientific collaboration. Gadgin Matha et al. (2026) studied the relationship between geopolitical distance and research and innovation collaboration at the global level. Thus, the association between geopolitics and scientific collaboration has received substantial attention in the literature.

Geopolitical associations may appear in relationship formation, exit, or changes within continuing ties. Margin-specific outcomes make these adjustment processes visible.

Geopolitical relations encompass United Nations General Assembly voting positions, bilateral sanctions, alliance relationships, export controls, security reviews, and discipline-specific regulation. The analysis uses a political-position proxy and interprets its association at the country-pair level.

2.4 Review of the Literature

The literature establishes that scientific collaboration has a network structure and is associated with linguistic, political, and institutional conditions. A remaining empirical gap is the lack of a joint comparison of linguistic and geopolitical distance across distinct margins of collaboration adjustment.

The analysis distinguishes six outcomes: relative allocation, absolute scale, active collaboration status, collaboration entry, collaboration exit, and intensity among active ties.

Linguistic and geopolitical distance come from different research traditions and need not vary together. A country pair may be linguistically close but politically distant, or politically close but linguistically distant. Joint models separate their associations under a common sample and control structure.

3. Data, Variables, and Measurement

3.1 Data Sources and Sample Construction

The publication records come from OpenAlex, which indexes scholarly works and information on their authors, publication venues, institutions, and concepts (Priem et al., 2022). The sample is restricted to English-language DOI journal articles in economics, mathematics, and medicine from 2018 to 2025. After matching countries, disciplines, and years, the final focal-country–partner-country–discipline–year panel contains 136,224 observations, covering 33 focal countries, 173 partner countries, 5,676 focal-country–partner-country combinations, and 17,028 country-pair–discipline groups. Table 1 reports the sample structure. The three disciplines and the English-language DOI criterion define the study population.

The baseline analysis constructs a directed international collaboration network. A paper is assigned to a focal country from the observable institutional-affiliation country of the first listed author, and links are formed with the other observable institutional-affiliation countries on the paper. This rule consistently defines focal-country output and partner relationships. Corresponding-author status, project leadership, individual contribution shares, and co-first authorship are unavailable in the data, so these roles are not used to assign network direction.

Table 1. Sample composition and descriptive statistics**Panel A. Sample**

Statistic	Value
Observations	136,224
Focal countries	33
Partner countries	173
Fields	3
Years	8
Unique focal-partner dyads	5,676
Unique focal-partner-field groups	17,028
Zero collaboration cell share	74.18%
2018 weighted mean collaboration rate	0.002539

Panel B. Descriptive statistics

Variable	N	Mean	SD	P25	Median	P75	Min	Max
Partner collaboration rate	136,224	0.003	0.011197	0	0	0.001431	0	0.462366
Pair papers	136,224	1.279459	4.870391	0	0	1	0	215
Focal papers	136,224	457.665404	172.221481	327	483.5	588.25	71	774
Geopolitical distance (standardized)	5,676	0.001486	1.007122	-0.967477	0.054341	0.7695	-1.456244	4.307213
Linguistic distance (standardized)	5,676	0.022075	0.988744	-0.221106	0.487858	0.52149	-5.675524	0.522185
Geopolitical distance (raw ideal-point gap)	5,676	1.008102	0.694652	0.339769	1.044558	1.537832	0.002647	3.977934
Linguistic distance (unstandardized)	5,676	0.960064	0.078956	0.940645	0.997259	0.999944	0.505083	1

Note. The baseline is a directed first-author network covering 2018–2025 English-language DOI articles in Economics, Mathematics, and Medicine. Linguistic distance is constructed from DICL *cnl* and *lpn* and standardized over the 6,405-dyad pre-control panel; benchmark dyads retain this scaling.

3.2 Collaboration Outcomes

For a given focal country, partner country, discipline, and year, the number of bilateral coauthored papers (PairPapers) is the number of papers assigned to the focal country that contain at least one observable institutional affiliation from the partner country; the number of focal-country papers (FocalPapers) is the focal country's total number of papers in the discipline and year. Based on these measures, the partner collaboration rate is first constructed as follows:

$$\text{PartnerCollabRate}_{ijft} = \frac{\text{PairPapers}_{ijft}}{\text{FocalPapers}_{ift}} \quad (1)$$

A multinational paper may connect the focal country to multiple partner countries simultaneously and therefore enter multiple country pairs; the same FocalPapers denominator for a focal-country–discipline–year is also repeated across different partners. Accordingly, the partner collaboration rate (PartnerCollabRate) measures the share of the focal country's total papers in the corresponding discipline and year that are coauthored with a given partner.

Five additional outcomes separate the other network adjustment processes:

Number of bilateral coauthored papers: PairPapers, the absolute number of collaborative papers between the focal country and partner country in the corresponding discipline and year.

Active collaboration status: equals 1 when PairPapers is greater than 0 and 0 otherwise.

Collaboration entry: defined only for country pairs with zero bilateral coauthored papers in the previous year; equals 1 when the current collaboration relationship is positive, i.e., $\Pr(y_t > 0 | y_{t-1} = 0)$

Collaboration exit: defined only for country pairs with a positive number of bilateral coauthored papers in the previous year; equals 1 when the current collaboration relationship is zero, i.e., $\Pr(y_t = 0 | y_{t-1} > 0)$

Intensity of active collaborative relationships: the natural logarithm of PairPapers conditional on PairPapers being greater than 0.

These variables correspond to different observable margins of international scientific collaboration. The partner collaboration rate captures relative partner allocation; PairPapers measures the absolute scale of bilateral collaboration; active collaboration status captures whether a relationship exists in the annual network; collaboration entry and exit capture relationship formation and termination, respectively; and the intensity of active collaborative relationships describes collaboration scale conditional on the relationship being present.

3.3 Network construction and counting rules

The baseline analysis uses full counting: for a paper involving multiple countries, each observable focal-country–partner-country link is counted as one bilateral paper link. As an alternative, under fractional counting, if a focal-country paper contains K observable partner countries, each partner receives a weight of $1/K$. The former records all bilateral links formed by a paper, whereas the latter ensures that the partner-link weights generated by one paper sum to one.

The study also reconstructs an all-institutional-affiliations network over the same fixed focal-country–partner-country–discipline universe. A paper is retained whenever the focal country appears in the observable set of institutional-affiliation countries, and all other observable countries are treated as partners without using the first author to determine direction. Reconstructed counts are merged into the common analysis panel, so the alternative remains defined on the fixed analytical dyad universe.

3.4 Distance Measures and Control Variables

Linguistic distance is constructed from two fields in the Dataset on International Connectivity for Language (DICL; Gurevich et al., 2025): *cnl*, which captures common-native-language connectivity, and *lpn*, which captures native-language linguistic proximity. We define bilateral language connectivity as $CL_{ij} = 0.5(cnl_{ij} + lpn_{ij})$ and linguistic distance as $L_{ij} = 1 - CL_{ij}$. The resulting directed-dyad measure is standardized over the 6,405 focal-partner dyads in the broader pre-control competition panel, and the 5,676-dyad benchmark sample retains this scaling. At the country-pair level, DICL proxies average communication and coordination frictions; it does not identify a team's working language or an individual researcher's language proficiency.

Geopolitical distance is constructed from final-passage United Nations General Assembly ideal points following Bailey et al. (2017). For each directed focal-partner dyad, we calculate the absolute bilateral ideal-point difference in each year from 2015 to 2017 and then take its mean across the three baseline years. The resulting measure is standardized across the 6,405 directed focal-partner dyads in the broader pre-control country-pair panel, and the benchmark sample retains this scaling. The measure captures one dimension of political position. Bilateral sanctions, export controls, security reviews, alliance relationships, and discipline-specific regulation lie outside its scope.

The baseline bilateral controls draw geographic and historical measures from the CEPII GeoDist database (Mayer and Zignago, 2011): the log of population-weighted geographic distance (*distw*), indicators for contiguity (*contig*), a colonial relationship (*colony*), a common colonizer (*comcol*), and historical membership in the same country (*smctry*). Pre-period trade intensity and regional-trade-agreement exposure come from the CEPII Gravity Database (Conte et al., 2022). For each dyad-year from 2015 to 2017, bilateral trade is constructed as the sum of the two directed *trade_flow_baci* flows.

Annual trade intensity is then defined as $\log\left(1 + \frac{\text{bilateraltrade}}{\sqrt{GDP_{it}GDP_{jt}}}\right)$, averaged over the available baseline years, and standardized. Regional-trade-agreement exposure is based on *fta_wto*: directed records are aggregated within each dyad-year, converted to an annual indicator equal to one when the observed value is positive, and then averaged over 2015–2017 and standardized. In the trajectory models, each baseline control is interacted with a linear trend defined as *year* minus 2018; the transition models use the corresponding dyad-level controls.

3.5 Empirical Models

3.5.1 Dyadic Trajectory Models

For the partner collaboration rate, bilateral coauthored paper count, active collaboration status, and log active-tie intensity, the benchmark specification is

$$Y_{ijft} = \beta_L(L_{ij} \times T_t) + \beta_G(G_{ij} \times T_t) + \gamma'(X_{ij} \times T_t) + \alpha_{ijf} + \lambda_{ift} + \mu_{jft} + \varepsilon_{ijft} \quad (2)$$

where L_{ij} is standardized language distance, G_{ij} is standardized pre-period geopolitical distance, T_t is a linear year trend beginning in 2018, and X_{ij} denotes the vector of baseline bilateral controls described in Section 3.4. Focal-partner-field fixed effects α_{ijf} absorb all time-invariant pair-field differences. Focal-field-year fixed effects λ_{ift} absorb changes common to a focal country's field in each year, including its publication scale and broad national conditions. Partner-field-year fixed effects μ_{jft} absorb partner-specific field-year conditions.

The coefficient β^L is the difference in the annual outcome trajectory associated with a one-standard-deviation larger language distance, conditional on the included fixed effects and control trends. β^G has the analogous interpretation for geopolitical distance. For the active-collaboration-status outcome, the

model is a weighted linear probability model. For active-tie intensity, the sample contains only positive pair-paper cells and the coefficient is a conditional log-intensity trajectory.

All models enter language and geopolitical distance jointly. Observations are weighted by FocalPapers, giving more weight to focal–field–years with more underlying articles. Standard errors are clustered two ways, by focal country and partner country. The trajectory, active-collaboration-status, entry, and active-tie-intensity models contain 33 focal and 173 partner clusters; the exit risk set contains 33 focal and 172 partner clusters.

3.5.2 Entry and Exit Risk-set Models

Entry and exit are estimated on state-specific risk sets from 2019 to 2025. The entry sample contains dyad-field observations with $y_{t-1} = 0$; its outcome is one if $y_t > 0$. The exit sample contains observations with $y_{t-1} > 0$; its outcome is one if $y_t = 0$. These weighted linear probability models include focal–field–year and partner–field–year fixed effects, the two distance measures, and the baseline bilateral controls. Pair-field fixed effects are not included because the transition specification estimates associations with time-invariant distance within the conditional risk sets.

The entry coefficients describe probability differences among previously inactive dyads, and the exit coefficients describe probability differences among previously active dyads. Each estimate is specific to its risk-set population.

3.5.3 Direct Coefficient-comparison Tests

For each margin, we test $H_0: \beta^L = \beta^G$ using the estimated within-model covariance:

$$\text{Var}(\hat{\beta}_L - \hat{\beta}_G) = \text{Var}(\hat{\beta}_L) + \text{Var}(\hat{\beta}_G) - 2\text{Cov}(\hat{\beta}_L, \hat{\beta}_G) \quad (3)$$

Statements about differences between the distance profiles rely on this test, which incorporates the estimated within-model covariance. Separate p-values do not test coefficient equality.

The design estimates conditional observational associations in dyadic trajectories and transition probabilities. Language and geopolitical distance are predetermined rather than exogenously varied, and time-varying bilateral conditions not absorbed by the fixed effects and control trends may remain correlated with the outcomes. The coefficients therefore do not identify causal effects.

4. Results

4.1 Partner Collaboration Rate

The benchmark partner-collaboration-rate model shows negative annual gradients for both distance measures (Table 2). The language-distance coefficient is -0.000082 (95% confidence interval, -0.000127 to -0.000038). In percentage-point units, a one-standard-deviation larger language distance is associated with a 0.0082-percentage-point steeper annual decline in the focal country's collaboration rate with that partner. The geopolitical-distance coefficient is -0.000063 (-0.000122 to -0.000004), equivalent to a 0.0063-percentage-point steeper annual rate trajectory.

These estimates describe relative partner allocation. Because FocalPapers is repeated across partners and a multinational article can enter multiple dyads, the rates are not mutually exclusive partner shares. A lower trajectory means that a partner accounts for a declining fraction of the focal country's field-specific output under this dyadic definition, while total international collaboration may move differently.

Figures 1 and 2 trace the timing of these associations. With 2018 as the reference, the language-distance coefficients become increasingly negative after 2019 and reach -0.000518 in 2025

(-0.000828 to -0.000207). The geopolitical path is less monotonic and reaches -0.000498 in 2025 (-0.000945 to -0.000051). The joint test across post-2018 language coefficients has a p-value of 0.038809; the corresponding geopolitical test has a p-value of 0.067295. The paths describe divergence from the 2018 coefficient in a descriptive dynamic specification.

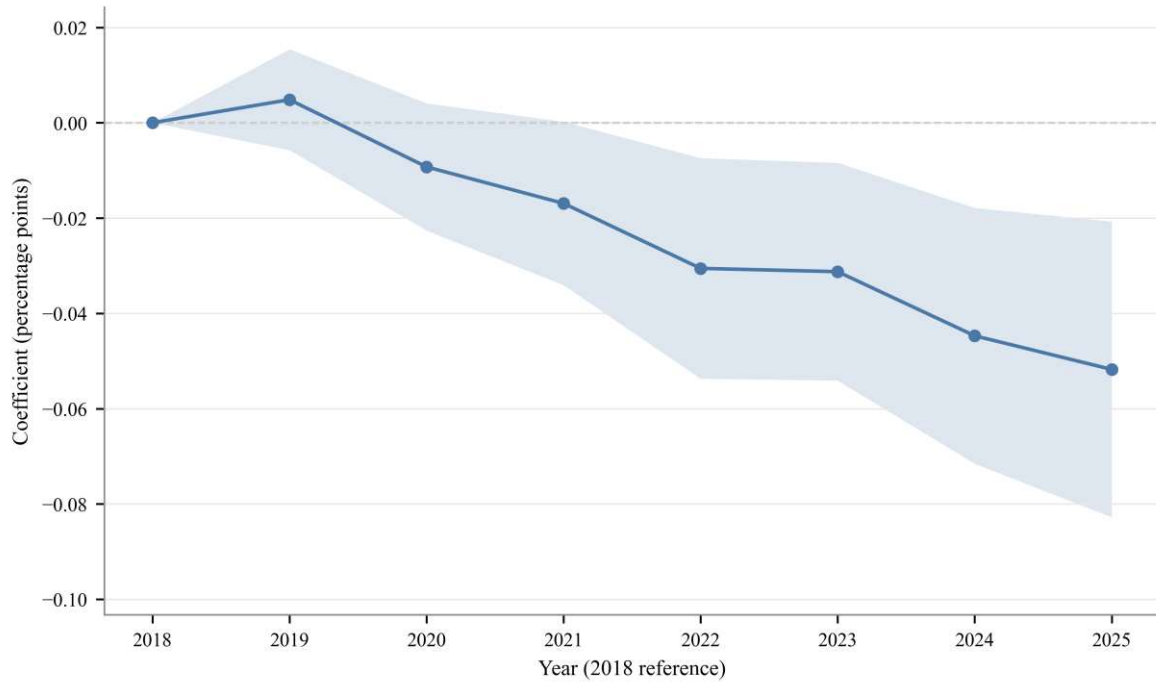


Figure 1. Dynamic estimates for language distance in the partner collaboration rate. Points show annual coefficients and the shaded area shows 95% confidence intervals relative to the 2018 reference.

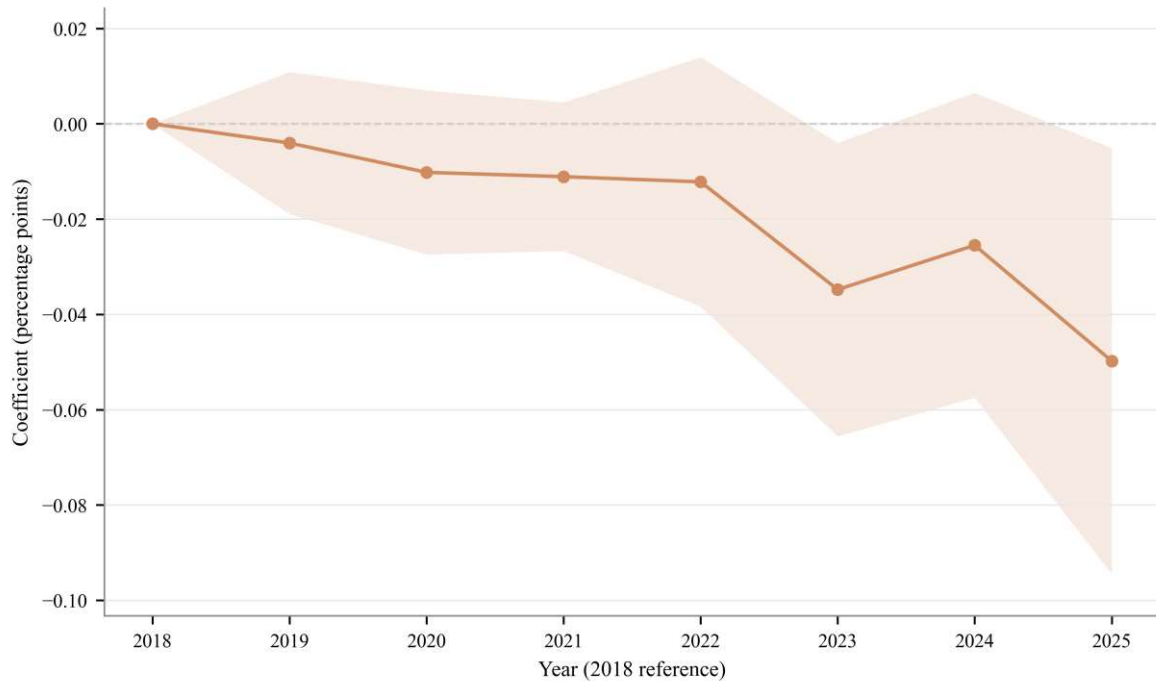


Figure 2. Dynamic estimates for geopolitical distance in the partner collaboration rate. Points show annual coefficients and the shaded area shows 95% confidence intervals relative to the 2018 reference.

4.2 Absolute Bilateral Collaboration Volume

The absolute-count model separates changes in a partner's rate from changes in joint publication volume. For language distance, the pair-count coefficient is -0.029551 papers per year for a one-standard-deviation difference (95% confidence interval, -0.045139 to -0.013963). The direction and interval show that the language-rate gradient is accompanied by a negative trajectory in absolute pair-paper counts within the same fixed-effect and control framework.

The geopolitical pair-count coefficient is -0.016477 , but its confidence interval spans -0.039968 to 0.007013 . The point estimate is consistent with lower absolute volume, but precision is insufficient to distinguish it from zero. The geopolitical pattern is therefore clearer for relative partner reallocation than for absolute collaboration volume.

Table 2. Language and geopolitical distance across partner collaboration rate and bilateral publication volume

Margin	Distance	Coefficient	SE	95% CI	p	N	Risk set
Partner collaboration rate	Language	-8.22e-05	2.18e-05	[-0.000127, -3.79e-05]	0.000648	136,224	all balanced cells
Partner collaboration rate	Geopolitical	-6.33e-05	2.90e-05	[-0.000122, -4.16e-06]	0.036702	136,224	all balanced cells
Pair paper count	Language	-0.029551	0.007653	[-0.045139, -0.013963]	0.000516	136,224	all balanced cells
Pair paper count	Geopolitical	-0.016477	0.011532	[-0.039968, 0.007013]	0.162754	136,224	all balanced cells

Note. Coefficients are annual distance-by-trend gradients. Models jointly include standardized linguistic and geopolitical distance together with trends in baseline geographic, historical, trade, and regional-trade-agreement controls. Observations are focal-paper weighted; standard errors are clustered by focal and partner country.

4.3 Active Collaboration Status and Collaboration Entry

Active collaboration status indicates whether any focal–partner paper is observed in a field-year (Table 3). Language distance is associated with a -0.001873 annual active-collaboration-status gradient (95% confidence interval, -0.003371 to -0.000374), or -0.1873 percentage points per year for a one-standard-deviation difference. The geopolitical coefficient is -0.000323 (-0.001405 to 0.000759). Its interval includes zero, leaving the annual active-collaboration-status trajectory imprecisely estimated.

The entry model isolates the formation margin. Among 88,829 lagged-zero observations, a one-standard-deviation larger language distance is associated with a 0.031796 lower probability of entering an active state (95% confidence interval, -0.043370 to -0.020222). This is a 3.18-percentage-point probability difference within the entry risk set. The geopolitical association is also negative: -0.021368 (-0.030930 to -0.011805), or 2.14 percentage points. Both distance measures are therefore associated with lower formation among previously inactive dyads.

4.4 Collaboration Exit

Among 30,367 lagged-positive observations, language distance is associated with a higher probability of exit. The coefficient is 0.055251 (95% confidence interval, 0.037621 to 0.072881), equivalent to a 5.53-percentage-point difference for a one-standard-deviation larger language distance. The geopolitical coefficient is 0.016747 (-0.001389 to 0.034882). Its positive point estimate is in the expected direction, but the conventional two-way-clustered interval includes zero.

Exit provides the clearest formal difference between the two proxies. The covariance-based language-minus-geopolitical contrast is 0.038505 (95% confidence interval, 0.012204 to 0.064806 ; $p=0.005438$). Within the lagged-positive risk set, the language-distance exit association is therefore 3.85 percentage points more positive than the geopolitical-distance association.

4.5 Intensity among Active Ties

Conditional on positive pair-paper cells, both distance measures have negative intensity trajectories. The language-distance coefficient is -0.006815 log points per year (95% confidence interval, -0.010994 to -0.002636) across 35,171 observations. The geopolitical coefficient is -0.007615 (-0.013303 to -0.001928). Active ties thus display negative conditional-intensity gradients alongside the entry and exit patterns.

The two intensity coefficients are close. Their language-minus-geopolitical contrast is 0.000801 (95% confidence interval, -0.006445 to 0.008046 ; $p=0.823366$). The data do not support a difference between them on this margin.

Table 3. Active collaboration status, entry, exit, and intensity among active ties

Margin	Distance	Coefficient	SE	95% CI	p	N	Risk set
Active collaboration status	Language	-0.001873	0.000736	[-0.003371, -0.000374]	0.015949	136,224	all balanced cells
Active collaboration status	Geopolitical	-0.000323	0.000531	[-0.001405, 0.000759]	0.54766	136,224	all balanced cells
Entry	Language	-0.031796	0.005682	[-0.04337, -0.020222]	3.51e-06	88,829	lagged-zero risk set, 2019-2025
Entry	Geopolitical	-0.021368	0.004695	[-0.03093, -0.011805]	7.28e-05	88,829	lagged-zero risk set, 2019-2025
Exit	Language	0.055251	0.008655	[0.037621, 0.072881]	3.59e-07	30,367	lagged-positive risk set, 2019-2025
Exit	Geopolitical	0.016747	0.008903	[-0.001389, 0.034882]	0.069111	30,367	lagged-positive risk set, 2019-2025
Active-tie intensity	Language	-0.006815	0.002051	[-0.010994, -0.002636]	0.002244	35,171	positive cells only
Active-tie intensity	Geopolitical	-0.007615	0.002792	[-0.013303, -0.001928]	0.01028	35,171	positive cells only

Note. Entry conditions on lagged-zero cells; exit conditions on lagged-positive cells; active-tie intensity conditions on positive current cells. Models are focal-paper weighted and use two-way focal and partner clustering. Transition models include focal-field-year and partner-field-year fixed effects; trajectory models also include focal-partner-field fixed effects.

4.6 Comparing Language and Geopolitical Distance across Margins

Figure 3 summarizes the six outcome-specific estimates without placing unlike outcome units on a common magnitude scale. Language distance has the expected adjustment sign on all six margins: negative for the rate, count, activity, entry, and conditional intensity, and positive for exit. All six conventional two-way-clustered estimates are statistically distinguishable from zero at the 5% level. The geopolitical-distance profile has precise conventional evidence for the partner collaboration rate, entry, and active-tie intensity. Its pair-count and active-collaboration-status intervals include zero, as does its conventional exit interval.

The difference in the breadth of individually precise estimates is descriptive. Table 4 and Figure 4 show that the formal contrast intervals include zero for the partner collaboration rate, pair-paper counts, active collaboration status, entry, and active-tie intensity. Only exit rejects equality at 5%. The result is an exit-specific difference within otherwise overlapping coefficient profiles.

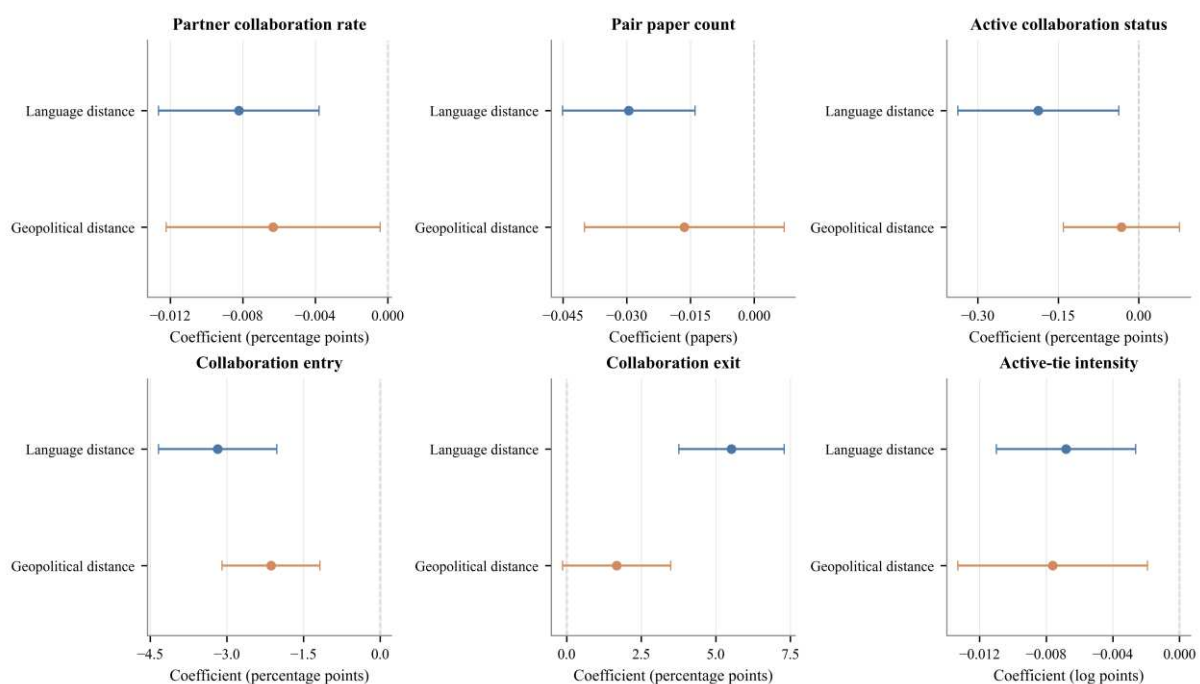


Figure 3. Estimated language- and geopolitical-distance gradients across collaboration margins. Points show coefficients and horizontal bars show 95% confidence intervals from the jointly specified models.

Table 4. Direct coefficient contrasts between language and geopolitical distance

Margin	Language	Geopolitical	Language - geopolitical	SE	95% CI	p	N
Partner collaboration rate	-8.22e-05	-6.33e-05	-1.90e-05	2.68e-05	[-7.35e-05, 3.55e-05]	0.483856	136,224
Pair paper count	-0.029551	-0.016477	-0.013073	0.010823	[-0.03512, 0.008973]	0.235943	136,224
Active collaboration status	-0.001873	-0.000323	-0.00155	0.000848	[-0.003276, 0.000177]	0.076884	136,224
Entry	-0.031796	-0.021368	-0.010428	0.007312	[-0.025323, 0.004467]	0.163529	88,829
Exit	0.055251	0.016747	0.038505	0.012912	[0.012204, 0.064806]	0.005438	30,367
Active-tie intensity	-0.006815	-0.007615	0.000801	0.003557	[-0.006445, 0.008046]	0.823366	35,171

Note. Contrasts use the full within-model two-way-clustered covariance matrix. Exit is the only margin for which equality is rejected at 5%.

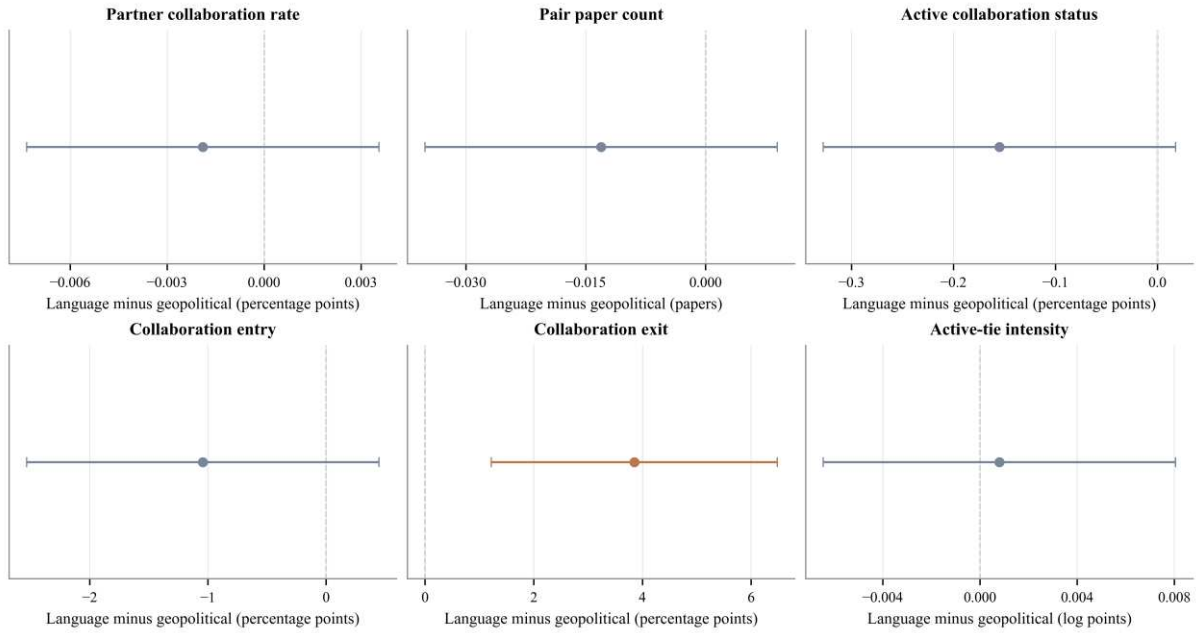


Figure 4. Direct language-versus-geopolitical coefficient contrasts. Points show the within-model language coefficient minus the geopolitical coefficient; horizontal bars show covariance-based 95% confidence intervals and the dashed vertical line marks zero. Exit is the only interval that excludes zero.

5. Robustness and Validity

5.1 Alternative Network Construction

The baseline directed first-author outcomes are compared with all-affiliation outcomes reconstructed on the frozen analytical dyad universe. Both are evaluated under full and fractional counting, with and without papers flagged at the hyperauthorship cap, producing eight constructions (Table 5). The directed construction assigns a work to a focal country through its first author. The all-affiliation construction retains a work when the focal country appears anywhere in its observed affiliation-country union and counts each other observed country as a partner. Full counting gives every observed partner link one unit, whereas fractional counting allocates one unit across the observed partners. Denominators and regression weights are rebuilt for each construction before the resulting counts are merged onto the common panel.

Language-distance signs are stable across all eight constructions and all six margins. The coefficient ranges are -0.000152 to -0.000041 for the partner collaboration rate, -0.107233 to -0.015030 for pair-paper counts, -0.001932 to -0.001272 for active collaboration status, -0.032154 to -0.031522 for entry, 0.030647 to 0.055567 for exit, and -0.010024 to -0.005017 for active-tie intensity. The ranges differ because outcome construction changes the scale, denominators, weights, and transition risk sets within the common dyad universe; their common sign provides the relevant comparison.

The geopolitical profile is more sensitive. Five margins retain the expected sign in all eight constructions, but active collaboration status does so in four. The active-collaboration-status coefficient ranges from -0.000323 to 0.000773 . Other ranges are -0.000174 to -0.000030 for the partner collaboration rate, -0.114703 to -0.006499 for pair-paper counts, -0.027607 to -0.020903 for entry, 0.013113 to 0.016747 for exit, and -0.009091 to -0.002163 for active-tie intensity. Together, the checks show stable language-distance signs and greater construction sensitivity in the geopolitical estimates.

5.2 Full Versus Fractional Counting

Full and fractional counting answer different measurement questions. Full counting records every observed bilateral link created by a paper, matching the network interpretation of a collaboration connection. Fractional counting gives less weight to each partner link on papers spanning many countries and prevents a multinational paper from contributing more than one unit across the focal country's partner set. The stable language signs across both rules show that its profile is not confined to the allocation of large multinational teams. Several geopolitical magnitudes shift more visibly, consistent with their broader construction sensitivity.

5.3 Hyperauthorship Exclusion

The directed baseline contains 357,317 retained paper records and 357,208 after excluding observations at the hyperauthorship flag; the corresponding all-affiliation counts are 474,323 and 473,480. Removing capped records leaves the language sign profile unchanged. The exercise addresses observed cap risk; author identities and affiliations beyond OpenAlex's 100-authorship cap remain unavailable.

Table 5. Alternative network constructions, counting rules, and hyperauthorship exclusions

Construction	Margin	N	Risk set	Lang. coef.	Lang. SE	Lang. p	Geo. coef.	Geo. SE	Geo. p
Directed first author, full	Partner collaboration rate	136,224	all balanced cells	-8.22e-05	2.18e-05	0.000648	-6.33e-05	2.90e-05	0.036702
Directed first author, full	Pair paper count	136,224	all balanced cells	-0.029551	0.007653	0.000516	-0.016477	0.011532	0.162754
Directed first author, full	Active collaboration status	136,224	all balanced cells	-0.001873	0.000736	0.015949	-0.000323	0.000531	0.54766
Directed first author, full	Entry	88,829	lagged-zero risk set, 2019-2025	-0.031796	0.005682	3.51e-06	-0.021368	0.004695	7.28e-05
Directed first author, full	Exit	30,367	lagged-positive risk set, 2019-2025	0.055251	0.008655	3.59e-07	0.016747	0.008903	0.069111
Directed first author, full	Active-tie intensity	35,171	positive cells only	-0.006815	0.002051	0.002244	-0.007615	0.002792	0.01028
Directed first author, full, no hyperauthorship	Partner collaboration rate	136,224	all balanced cells	-8.13e-05	2.15e-05	0.000644	-6.01e-05	2.86e-05	0.043309
Directed first author, full, no hyperauthorship	Pair paper count	136,224	all balanced cells	-0.029022	0.007541	0.000535	-0.015014	0.011334	0.19465
Directed first author, full, no hyperauthorship	Active collaboration status	136,224	all balanced cells	-0.001932	0.000742	0.013853	-0.000208	0.000537	0.701154
Directed first author, full, no hyperauthorship	Entry	88,867	lagged-zero risk set, 2019-2025	-0.031973	0.005654	2.95e-06	-0.020903	0.00466	8.79e-05
Directed first author, full, no hyperauthorship	Exit	30,329	lagged-positive risk set, 2019-2025	0.055567	0.008606	2.91e-07	0.016699	0.008896	0.069635
Directed first author, full, no hyperauthorship	Active-tie intensity	35,101	positive cells only	-0.006778	0.002066	0.002504	-0.007491	0.002782	0.011195
Directed first author, fractional	Partner collaboration rate	136,224	all balanced cells	-4.12e-05	1.03e-05	0.000356	-3.02e-05	1.83e-05	0.10885
Directed first author, fractional	Pair paper count	136,224	all balanced cells	-0.015172	0.004377	0.001526	-0.006574	0.007661	0.397212
Directed first author, fractional	Active collaboration status	136,224	all balanced cells	-0.001873	0.000736	0.015949	-0.000323	0.000531	0.54766
Directed first author, fractional	Entry	88,829	lagged-zero risk set, 2019-2025	-0.031796	0.005682	3.51e-06	-0.021368	0.004695	7.28e-05
Directed first author, fractional	Exit	30,367	lagged-positive risk set, 2019-2025	0.055251	0.008655	3.59e-07	0.016747	0.008903	0.069111
Directed first author, fractional	Active-tie intensity	35,171	positive cells only	-0.010024	0.002371	0.000184	-0.009091	0.003089	0.005999
Directed first author, fractional, no hyperauthorship	Partner collaboration rate	136,224	all balanced cells	-4.11e-05	1.03e-05	0.00035	-3.01e-05	1.83e-05	0.109861
Directed first author, fractional, no hyperauthorship	Pair paper count	136,224	all balanced cells	-0.01503	0.004345	0.001555	-0.006499	0.00765	0.40189
Directed first author, fractional, no hyperauthorship	Active collaboration status	136,224	all balanced cells	-0.001932	0.000742	0.013853	-0.000208	0.000537	0.701154
Directed first author, fractional, no hyperauthorship	Entry	88,867	lagged-zero risk set, 2019-2025	-0.031973	0.005654	2.95e-06	-0.020903	0.00466	8.79e-05
Directed first author, fractional, no hyperauthorship	Exit	30,329	lagged-positive risk set, 2019-2025	0.055567	0.008606	2.91e-07	0.016699	0.008896	0.069635
Directed first author, fractional, no hyperauthorship	Active-tie intensity	35,101	positive cells only	-0.009593	0.002378	0.000318	-0.008865	0.003004	0.005878
All affiliations, full	Partner collaboration rate	136,224	all balanced cells	-0.000152	3.76e-05	0.000318	-0.000174	5.29e-05	0.002506
All affiliations, full	Pair paper count	136,224	all balanced cells	-0.107233	0.027352	0.000438	-0.114703	0.035693	0.002988
All affiliations, full	Active collaboration status	136,224	all balanced cells	-0.001272	0.000774	0.110193	0.000773	0.000464	0.105331
All affiliations, full	Entry	68,903	lagged-zero risk set, 2019-2025	-0.031522	0.007747	0.000288	-0.025421	0.006697	0.00062

Construction	Margin	N	Risk set	Lang. coef.	Lang. SE	Lang. p	Geo. coef.	Geo. SE	Geo. p
All affiliations, full	Exit	50,293	lagged-positive risk set, 2019-2025	0.030647	0.005944	1.26e-05	0.01407	0.005437	0.014407
All affiliations, full	Active-tie intensity	58,317	positive cells only	-0.005319	0.002072	0.015149	-0.002442	0.001943	0.217873
All affiliations, full, no hyperauthorship	Partner collaboration rate	136,224	all balanced cells	-0.000147	3.65e-05	0.000329	-0.000165	5.18e-05	0.003278
All affiliations, full, no hyperauthorship	Pair paper count	136,224	all balanced cells	-0.102428	0.026091	0.000431	-0.106656	0.034373	0.003987
All affiliations, full, no hyperauthorship	Active collaboration status	136,224	all balanced cells	-0.001296	0.000757	0.096512	0.000517	0.000462	0.271916
All affiliations, full, no hyperauthorship	Entry	69,748	lagged-zero risk set, 2019-2025	-0.032154	0.007675	0.000205	-0.027607	0.006782	0.000287
All affiliations, full, no hyperauthorship	Exit	49,448	lagged-positive risk set, 2019-2025	0.031763	0.006195	1.37e-05	0.013113	0.005858	0.032283
All affiliations, full, no hyperauthorship	Active-tie intensity	57,348	positive cells only	-0.005017	0.002086	0.022143	-0.002163	0.00188	0.258455
All affiliations, fractional	Partner collaboration rate	136,224	all balanced cells	-4.96e-05	1.32e-05	0.000685	-4.34e-05	2.33e-05	0.071016
All affiliations, fractional	Pair paper count	136,224	all balanced cells	-0.035731	0.009252	0.000515	-0.021919	0.014665	0.144812
All affiliations, fractional	Active collaboration status	136,224	all balanced cells	-0.001272	0.000774	0.110193	0.000773	0.000464	0.105331
All affiliations, fractional	Entry	68,903	lagged-zero risk set, 2019-2025	-0.031522	0.007747	0.000288	-0.025421	0.006697	0.00062
All affiliations, fractional	Exit	50,293	lagged-positive risk set, 2019-2025	0.030647	0.005944	1.26e-05	0.01407	0.005437	0.014407
All affiliations, fractional	Active-tie intensity	58,317	positive cells only	-0.007832	0.00262	0.005334	-0.006037	0.002721	0.033701
All affiliations, fractional, no hyperauthorship	Partner collaboration rate	136,224	all balanced cells	-4.97e-05	1.32e-05	0.000694	-4.32e-05	2.33e-05	0.073442
All affiliations, fractional, no hyperauthorship	Pair paper count	136,224	all balanced cells	-0.035131	0.00914	0.000542	-0.021256	0.014582	0.154654
All affiliations, fractional, no hyperauthorship	Active collaboration status	136,224	all balanced cells	-0.001296	0.000757	0.096512	0.000517	0.000462	0.271916
All affiliations, fractional, no hyperauthorship	Entry	69,748	lagged-zero risk set, 2019-2025	-0.032154	0.007675	0.000205	-0.027607	0.006782	0.000287
All affiliations, fractional, no hyperauthorship	Exit	49,448	lagged-positive risk set, 2019-2025	0.031763	0.006195	1.37e-05	0.013113	0.005858	0.032283
All affiliations, fractional, no hyperauthorship	Active-tie intensity	57,348	positive cells only	-0.007277	0.002799	0.014011	-0.00564	0.002529	0.032865

Note. Counts, denominators, weights, and transition risk sets are reconstructed for each network rule on the frozen analytical dyad universe before merging to the common focal-partner-field panel.

5.4 Finite-cluster Inference and Leave-one-focal-country-out Checks

The focal dimension contains 33 clusters, making finite-cluster diagnostics important. Wild-cluster inference uses 999 bootstrap draws at the focal-country level for five margins: the partner collaboration rate, pair-paper counts, entry, exit, and active-tie intensity. Active collaboration status is not included in this diagnostic. The wild-cluster p-values for the language coefficients are 0.001 on each of the five reported margins. For geopolitical distance they are 0.011 for the partner collaboration rate, 0.156 for pair counts, 0.001 for entry, 0.023 for exit, and 0.012 for active-tie intensity. For geopolitical exit, the conventional two-way interval includes zero, whereas the focal-level wild-bootstrap p-value is below 0.05.

Leave-one-focal-country-out estimates provide a coefficient-stability diagnostic. Among the ten audited coefficients, none changes sign and none moves by more than 50% in magnitude after removing one focal country. The language ranges are -0.000090 to -0.000076 for the partner collaboration rate, -0.032175 to -0.025956 for pair counts, -0.034212 to -0.029521 for entry, 0.051096 to 0.058005 for exit, and -0.007302 to -0.006072 for active-tie intensity. The corresponding geopolitical ranges are -0.000077 to -0.000050 , -0.022450 to -0.009151 , -0.023947 to -0.019720 , 0.013751 to 0.019527 , and -0.008833 to -0.006416 .

The largest relative movement is 0.444644 for the geopolitical pair-count coefficient. Conventional 5% significance classifications change in some focal-country deletions for the geopolitical partner-collaboration-rate and exit estimates. Coefficient directions remain stable, while precision changes for selected geopolitical margins. Bangladesh has the highest recorded leverage for most audited coefficients, and Vietnam has the highest leverage in the entry models; deletion leaves the coefficient directions unchanged.

5.5 Coverage Sensitivity

OpenAlex coverage is assessed using restrictions recoverable from the frozen files: high first-author affiliation observability, low authorship missingness, low cap risk, their intersection, and a strict observed complete-enumeration rule. Across the benchmark and five restricted samples, every language-distance coefficient retains the expected sign on all six margins. Under the strict complete-enumeration restriction, selected geopolitical coefficients change sign: pair counts become 0.014144, active collaboration status 0.000274, and active-tie intensity 0.007005. Their p-values are 0.145858, 0.811773, and 0.115505, respectively, so these reversals are themselves imprecise, but they show that the geopolitical profile is more sensitive to observable quality restrictions.

Observed-work completeness among positive records is high, but it measures only records present in OpenAlex. Completeness is undefined for zero-work dyads, and the current files do not recover partner-specific differential affiliation retention. Whether missing partner affiliations covary with language or geopolitical distance remains unknown. Coverage is therefore a central validity limitation.

5.6 Supplementary PPML Sensitivity Analysis

As a supplementary count-model sensitivity analysis, Supplementary Table S3 reports a Poisson pseudo-maximum likelihood model with high-dimensional fixed effects (PPML-HDFE) estimated on the broader pre-control language–geopolitics panel rather than on the 136,224-observation benchmark panel. The broader panel contains 153,720 observations, 35 focal countries, and 184 partner countries. Separation arising from fixed-effect groups with all-zero outcomes removes 80,951 observations, leaving 72,769 estimable rows; all 35 focal and 184 partner countries remain represented. The benchmark panel has a narrower country universe because imposing the CEPII GeoDist and Gravity controls removes two focal countries and 11 partner countries.

The geopolitical coefficient is -0.004900 (95% confidence interval, -0.011119 to 0.001318 ; $p=0.118517$), and the linguistic-distance coefficient is -0.001667 (95% confidence interval, -0.004823 to 0.001489 ; $p=0.290718$). Both estimates retain negative signs but are statistically imprecise.

Re-estimating the linear collaboration-rate model on the same 72,769 rows retained by the separation screen yields -0.000095 for geopolitical distance (95% confidence interval, -0.000155 to -0.000036) and -0.000129 for linguistic distance (95% confidence interval, -0.000182 to -0.000076). The difference in precision is therefore not explained solely by sample size. The linear-rate and PPML specifications use the same retained row universe but differ in outcome scale and estimand, so the PPML estimates provide a complementary specification sensitivity check.

Detailed inference, coverage, and specification sensitivity results are reported in Supplementary Table S3.

6. Further Analysis: Retrospective Heterogeneity in Collaboration Persistence

As a supplementary heterogeneity analysis, focal–partner–field relationships are classified according to their 2018–2025 collaboration histories. The baseline definition classifies a relationship as persistent when collaboration is positive in all eight years; alternative thresholds require at least seven or at least six positive years. Relationships with at least one positive year that do not meet the relevant persistence threshold are classified as switchers, while never-active relationships are excluded from this analysis. Because the full-sample classification uses outcomes from future years to define relationship status, it is inherently retrospective. On positive collaboration cells, we estimate focal-paper-weighted linear models of log partner collaboration rate with separate language-distance $\times$ linear-trend and geopolitical-distance $\times$ linear-trend slopes for persistent and switching relationships. The models include status-interacted baseline bilateral control trends, focal–field–year-by-status and partner–field–year-by-status fixed effects, and focal–partner–field fixed effects. The linear trend is defined as year minus 2018, and standard errors are two-way clustered by focal and partner country. For each distance measure, the persistence-gradient contrast is the difference between the persistent and switching slopes.

$$Y_{ijft} = \sum_{s \in \{0,1\}} \mathbf{1}\{P_{ijf} = s\} [\beta_{L,s}(L_{ij}\tau_t) + \beta_{G,s}(G_{ij}\tau_t) + \gamma'_s(\mathbf{X}_{ij}\tau_t) + \alpha_{ift,s} + \delta_{jft,s}] + \mu_{ijf} + \varepsilon_{ijft} \quad (4)$$

$$\Delta_L = \beta_{L,1} - \beta_{L,0} \quad (5)$$

$$\Delta_G = \beta_{G,1} - \beta_{G,0} \quad (6)$$

$$C = \Delta_G - \Delta_L \quad (7)$$

Here, $P_{ijf} = 1$ denotes a persistent relationship and $P_{ijf} = 0$ denotes a switching relationship; never-active relationships are excluded. Y_{ijft} is the log partner collaboration rate on positive collaboration cells, and $\tau_t = t - 2018$.

Across alternative full-sample thresholds, the geopolitical persistence-gradient contrast minus the linguistic persistence-gradient contrast is consistently negative (Supplementary Table S1). Requiring positive collaboration in all eight years yields an estimate of -0.019928 ($p=0.002570$); requiring positive collaboration in at least seven of the eight years yields -0.019653 ($p=0.001066$); and requiring positive collaboration in at least six of the eight years yields -0.018163 ($p=0.022159$). The pattern is consistent across thresholds: the linguistic-distance gradient is more negative among relationships classified ex post as switching, whereas the geopolitical-distance gradient is more negative among relationships classified ex post as persistent.

Forward-looking rolling classifications yield weaker evidence (Supplementary Table S2). When classification is based on 2018–2020 information and evaluated over 2021–2025, the geopolitical-minus-linguistic persistence-gradient contrast is -0.024150 , with a 95% confidence interval of -0.053123 to 0.004823 and $p=0.099248$. When classification is based on 2019–2021 information and evaluated over 2022–2025, the same contrast is -0.016576 , with a 95% confidence interval of -0.059489 to 0.026336 and $p=0.437177$, and the persistent-versus-switching pattern for geopolitical distance reverses sign in the second window. The prospective estimates do not support a robust heterogeneity claim; persistence heterogeneity remains suggestive retrospective evidence.

7. Conclusions and Policy Implications

Using English-language DOI journal articles indexed in OpenAlex from 2018 to 2025, the analysis covers economics, mathematics, and medicine across 33 focal countries, 173 partner countries, and 136,224 panel observations. Six margins describe relative partner allocation, absolute bilateral coauthored output, active collaboration, entry, exit, and intensity among active ties. Country-pair trajectory models and state-specific risk-set models estimate the corresponding associations.

Standardized linguistic distance is constructed from the Dataset on International Connectivity for Language (DICL; Gurevich et al., 2025), and pre-period geopolitical distance follows the ideal-point framework of Bailey et al. (2017). Both measures enter jointly with trends in the baseline bilateral controls. The geopolitical measure is fixed before the analysis period. Direct covariance-based tests compare the two coefficients within each model.

Linguistic distance has a broad and consistent profile across the six collaboration margins. Greater distance is associated with lower annual trajectories in partner collaboration rates and bilateral coauthored papers, a lower probability that collaboration remains active, and lower entry among previously inactive country pairs. It is also associated with higher exit and weaker intensity among active ties. This sign profile is stable across alternative network direction, full and fractional counting, hyperauthorship exclusion, observable-quality restrictions, and leave-one-focal-country-out estimation.

Geopolitical-distance evidence is concentrated in partner collaboration rate, collaboration entry, and intensity among active ties. The negative point estimate for bilateral coauthored papers is imprecise, as are the conventional estimates for active status and exit. Some geopolitical coefficients change sign under the strictest observable-coverage restriction, and their precision is more sensitive to leave-one-focal-country-out estimation and network construction.

Direct covariance-based tests show overlapping profiles on five margins: partner collaboration rate, the number of bilateral coauthored papers, active collaboration status, collaboration entry, and intensity among active ties. Collaboration exit is the only margin on which equality of the two coefficients is rejected at the 5% level. The statistically distinguishable difference is therefore specific to relationship termination; the estimates do not support a general ranking.

The observational design estimates conditional associations rather than causal effects identified from exogenous variation.

DICL is a country-pair proxy for average communication and coordination frictions, not a measure of team working language or individual language proficiency. Using pre-period ideal points reduces contemporaneous overlap with outcomes from 2018 to 2025. United Nations General Assembly ideal-point distance captures one dimension of political and institutional position; sanctions, export controls, security reviews, alliances, and discipline-specific regulation lie outside its scope.

OpenAlex coverage remains incomplete, and the available data cannot identify partner-specific affiliation retention or determine whether missingness is related to the two distance measures. PPML and linear models use different outcome scales and estimands; the supplementary PPML specification yields weaker statistical precision.

International collaboration monitoring should combine relative partner allocation with bilateral volume, active status, entry, exit, and active-tie intensity. Reporting the denominator, risk set, and counting rule for each indicator helps research management authorities and institutions identify the stage at which a network is changing.

Better data would connect country-pair measures to team-level processes. Language-related monitoring could include working language, researchers' language proficiency, multilingual practices, and translation support. Geopolitical monitoring could incorporate bilateral sanctions, export controls, security reviews, alliance relationships, and discipline-specific regulation. More complete author and institutional-affiliation data, together with team size, repeated collaboration, project funding, and collaboration duration, would clarify the processes behind formation, termination, and changes in intensity.

Monitoring indicators should be used in proportion to the evidence. Linguistic and geopolitical distance can identify country-pair patterns in collaboration networks, but they are not measures of a specific team's collaborative capacity and do not justify restrictive measures. The exit difference warrants follow-up across countries, disciplines, and institutional contexts. On margins where coefficient equality is not rejected, separate significance levels do not provide a basis for ranking the two distances. Policies involving cross-language research services, international project management, or geopolitical risk can be evaluated through pilots, process documentation, and more detailed data.

CRedit authorship contribution statement

Yizheng He: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Visualization, Writing – original draft, Writing – review & editing.

Ming Li: Writing – review & editing, Supervision, Resources, Project administration, Methodology, Funding acquisition, Conceptualization.

Funding

The author(s) declared that financial support was received for this work and its publication. This work was supported by the Major Program of the National Social Science Fund of China (Grant No.21&ZD153) and the Major Project of the 2023 Open Research Fund of the Key Open Laboratory of Statistical Information Technology and Data Mining, National Bureau of Statistics of China, and the China Government Statistical Research Center (Grant No. SDL202301).

Declaration of competing interest

The authors declare that there are no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work, the authors used OpenAI's ChatGPT to translate and polish the article.

Acknowledgments

We want to thank all those who contributed to this study.

Persistence heterogeneity tables are reported in Supplementary Tables S1 and S2.

References

- Adams, J. (2012). The rise of research networks. *Nature*, 490(7420), 335–336. <https://doi.org/10.1038/490335a>
- Bailey, M. A., Strezhnev, A., & Voeten, E. (2017). Estimating dynamic state preferences from United Nations voting data. *Journal of Conflict Resolution*, 61(2), 430–456. <https://doi.org/10.1177/0022002715595700>
- Cao, Y., Sickles, R. C., Triebs, T. P., & Tumlinson, J. (2024). Linguistic distance to English impedes research performance. *Research Policy*, 53(4), 104971. <https://doi.org/10.1016/j.respol.2024.104971>
- Charpin, R., London, J., & Vincent, N. (2024). The effect of geopolitical tensions on international research collaborations and its implications for global operations management. *International Journal of Production Economics*, 268, 109120. <https://doi.org/10.1016/j.ijpe.2023.109120>
- Conte, M., Cotterlaz, P., & Mayer, T. (2022). *The CEPII Gravity Database*. CEPII Working Paper No. 2022-05. <https://www.cepii.fr/CEPII/en/publications/wp/abstract.asp?NoDoc=13432>
- Frame, J. D., & Carpenter, M. P. (1979). International research collaboration. *Social Studies of Science*, 9(4), 481–497. <https://doi.org/10.1177/030631277900900405>
- Freeman, R. B., Ganguli, I., & Murciano-Goroff, R. (2015). Why and wherefore of increased scientific collaboration. In A. B. Jaffe & B. F. Jones (Eds.), *The changing frontier: Rethinking science and innovation policy* (pp. 17–48). University of Chicago Press. <https://doi.org/10.7208/chicago/9780226286860.003.0002>
- Gadgin Matha, S., Hartog, M., Chacua, C., Balland, P.-A., Hausmann, R., & Yildirim, M. A. (2026). Burning bridges or bridging divides: Geopolitics and collaboration in research and innovation. *Research Policy*, 55(9), 105580. <https://doi.org/10.1016/j.respol.2026.105580>
- Gurevich, T., Herman, P. R., Toubal, F., & Yotov, Y. V. (2025). A dataset on linguistic connectivity across and within countries. *Scientific Data*, 12(1), 542. <https://doi.org/10.1038/s41597-025-04692-8>
- 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>
- Hou, L., Pan, Y., & Zhu, J. J. H. (2021). Impact of scientific, economic, geopolitical, and cultural factors on international research collaboration. *Journal of Informetrics*, 15(3), 101194. <https://doi.org/10.1016/j.joi.2021.101194>
- Hwang, K. (2013). Effects of the language barrier on processes and performance of international scientific collaboration, collaborators' participation, organizational integrity, and interorganizational relationships. *Science Communication*, 35(1), 3–31. <https://doi.org/10.1177/1075547012437442>
- Jones, B. F., Wuchty, S., & Uzzi, B. (2008). Multi-university research teams: Shifting impact, geography, and stratification in science. *Science*, 322(5905), 1259–1262. <https://doi.org/10.1126/science.1158357>
- Liu, J., Guo, X., Xu, S., & Zhang, Y. (2023). Quantifying the impact of strong ties in international scientific research collaboration. *PLOS ONE*, 18(1), e0280521. <https://doi.org/10.1371/journal.pone.0280521>
- Makkonen, T., & Mitze, T. (2023). Geopolitical conflicts, sanctions and international knowledge flows: EU–Russia collaboration during the Ukraine crisis. *The World Economy*, 46(10), 2926–2949. <https://doi.org/10.1111/twec.13421>
- Mayer, T., & Zignago, S. (2011). *Notes on CEPII's distances measures: The GeoDist database*. CEPII Working Paper No. 2011-25. <https://www.cepii.fr/CEPII/en/publications/wp/abstract.asp?NoDoc=3877>
- Melitz, J., & Toubal, F. (2014). Native language, spoken language, translation and trade. *Journal of International Economics*, 93(2), 351–363. <https://doi.org/10.1016/j.jinteco.2014.04.004>
- Priem, J., Piwowar, H., & Orr, R. (2022). *OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts*. arXiv:2205.01833. <https://doi.org/10.48550/arXiv.2205.01833>
- Wagner, C. S., & Leydesdorff, L. (2005). Network structure, self-organization, and the growth of international collaboration in science. *Research Policy*, 34(10), 1608–1618. <https://doi.org/10.1016/j.respol.2005.08.002>
- Wu, L., Yi, F., Bu, Y., Lu, W., & Huang, Y. (2024). Toward scientific collaboration: A cost-benefit perspective. *Research Policy*, 53(2), 104943. <https://doi.org/10.1016/j.respol.2023.104943>
- Zhao, Y. (2026). Does distance influence international collaboration among scientists? Evidence from library and information science. *Humanities and Social Sciences Communications*, 13(1), 462. <https://doi.org/10.1057/s41599-026-06767-2>

Data availability

The data and supporting materials underlying this study are provided with this revision for editorial and peer-review scrutiny. The shared materials include the retained source data and structured OpenAlex extracts, processed datasets, the final analytical country-pair panel, variable definitions and transformation documentation, the recovered data-processing and statistical-analysis scripts, locked result files, and the retained table and figure code. The study draws on OpenAlex, the Dataset on International Connectivity for Language (DICT), United Nations General Assembly ideal-point data, and the CEPII GeoDist and Gravity databases. The accompanying documentation identifies the available source versions and retrieval information, source-specific redistribution information, and the retained processing scripts. Historical OpenAlex API response JSON was not retained, and the exact final G2 panel-construction and PPML estimation scripts were not recovered; these limitations and the available reconstruction evidence are documented in the package.

Ethical Approval

This article does not contain any studies with human participants performed by any of the authors.

Informed Consent

This article does not contain any studies with human participants performed by any of the authors. Therefore, informed consent was not applicable.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryInformation.docx](#)
- [HSSCReplicationPackage.zip](#)