UCL CENTRE FOR COMPARATIVE STUDIES
OF EMERGING ECONOMIES (CCSEE)

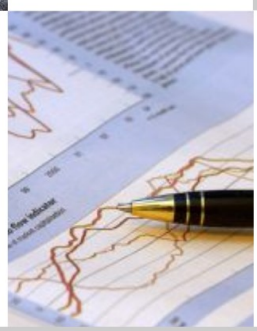
Working Paper Series

2017/4

Diversity and Collaboration in Economics

Sultan Orazbayev

CCSEE



UCL School of Slavonic and East European Studies
16 Taviton Street, London WC1H 0BW

www.ucl.ac.uk/ssees/comparative-studies-emerging-economies

Diversity and collaboration in Economics*

Sultan Orazbayev†

September 27, 2017

Abstract

Papers written by coauthors from different countries, on average, are published in better journals, have higher citations counts, and are evaluated more positively by peers. Similar 'diversity premia' exist for inter-ethnic and inter-gender collaborations. Using data on collaborations among 34 thousand economists, this paper considers possible explanations for the positive quality-diversity correlation. After controlling for a range of relevant factors, the authors' position in the global research network plays an important role in explaining variation in the quality of collaboration, proxied by citation counts and simple impact factor of the journal in which the article is published. Access to non-redundant social ties in the global research network is associated with greater quality of the collaboration. Geographic, gender and ethnic diversity premia on collaboration quality disappear after controlling for the authors' global network position, suggesting that diversity is important only to the extent that it correlates with non-redundancy of social ties.

Keywords: structural holes; diversity; collaboration networks.

JEL codes: A14, Z1.

*I would like to thank Miriam Manchin for continued support throughout this project, Douglas Nelson for constructive criticism and an extended discussion on the topic that lead to a change in the formulation of the research question, and Elodie Douarin and Francesco Fasani for further support and detailed feedback. I also thank Karuna Krishnaswamy, Jascha Grübel, Frank Neffke as well as seminar participants at Galbino, CEA Ottawa and Barcelona GSE Summer Forum for feedback. Paul Longley kindly provided the name-ethnicity matches used at the early stage of the calculations, the RePEc team (in alphabetical order: Jose Manuel Barrueco, Sune Karlsson, Thomas Krichel and Christian Zimmermann) provided help with access to the data and explanations on technical issues.

†contact@econpoint.com.

1 Introduction

Multiple empirical studies have shown that collaboration and the diversity of collaborators among academics are increasing (e.g. Gazni et al. 2012; Wuchty et al. 2007). These trends are not uniform across all fields of science, but collaboration among economists is also increasing (Hamermesh 2013; Card and DellaVigna 2013). A salient feature of the data on academic collaborations are the premia on international (e.g. Franceschet and Costantini 2010) and inter-ethnic (e.g. Freeman and Huang 2015) collaborations, reflected in higher citation counts and higher impact factor of the journal in which the joint work was published. This paper shows that such diversity premia are also observed in a sample of publications by approximately 34 thousand economists, including a premium for collaborations between authors of different gender. A social network-based explanation is used to explain all three diversity premia.

Factors affecting the quality of academic collaboration can be grouped into three broad categories: (1) gains from the division of labour (and economies of scale/scope) due to collaboration, (2) human capital of the collaborators and (3) network capital of the collaborators.[1]

1. Collaboration allows coauthors to exploit potential gains from the division of labour, hence reducing the cost of academic 'production' and allowing collaborators to focus on improving the quality of their product. The combination of diminishing returns from an additional author and increasing cost of coordination will put an upper limit on the efficient team size (Furman and Gaule 2013).

2. Human capital reflects characteristics of the collaborators, for example their skills and knowledge. As the stock of acquired knowledge accumulates, a single author may find it difficult to acquire expertise in all relevant areas, so authors with different specialisations can complement each other through collaboration (Jones 2009).

3. Network capital will reflect the role that collaborators play in providing access to the knowledge and resources embedded in their networks[2], which

[1] The social network of an author could be thought of as 'network capital' (i.e. as an asset that can be accumulated and maintained at some cost, this asset can generate returns and depreciates over time). A related term, social capital, is already used in the literature to represent trust, cohesion and related concepts in a network.

[2] Another effect of a larger, more influential network is improved diffusion of knowledge, which will also attract a larger number of citations. In this paper the relevant measure of network capital is independent of the size of the network, with some caveats. See further discussion in Section **??**.

can take various forms, including tacit knowledge and access to particular datasets/funding (for example, bulk of 2007–2013 European Framework Programme's budget was allocated for international collaboration projects). Sociologists have shown that social network properties play an important role in creativity and quality of ideas, both at the individual and team levels (Uzzi and Spiro 2005; Burt 2004).

These broad categories do not explicitly include diversity measures, for example ethnicity, gender, and geographic locations, because this will require a mechanism that links diversity and collaboration quality. Consider an example from (Freeman and Huang 2015), suppose that authors have a preference towards collaboration with someone of the same ethnicity, and are willing to forgo collaboration with someone whose skills and knowledge are more relevant to a project, but who happens to belong to a different ethnicity. Such preferences can explain why inter-ethnic collaborations are on average more productive, but to explain the premia on international and inter-gender collaborations the preferences would have to become very discriminatory. Another explanation could be that more productive authors select into international collaboration (analogous to the selection of more productive firms into exporting), perhaps because only very productive authors can afford the communication and coordination costs of collaboration over great distance, resulting in an international collaboration premium. However, this mechanism would not be able to explain the other two premia.

The social network approach provides a single explanation for all three premia. Every scientist acquires a set of knowledge through training, research and collaborations. This knowledge would be captured by the scientist's human capital. Also, the scientist would have access to the knowledge embedded in their ego network, for example close friends and colleagues, whom the scientist can contact quickly ('strong ties'). Collaborators whose ego networks overlap are then more likely to have access to very similar knowledge and resources embedded in their networks. In the extreme hypothetical case, collaboration with one's clone is unlikely to provide access to new ideas or knowledge (there would still be gains from division of labour).[3] On the other hand, if collaborators' networks are non-overlapping, then as a result of collaboration they each get access to new, non-redundant resources/information embedded in their coauthor's networks.

Clustering of authors along some measure of diversity (e.g. ethnic collaboration clusters) implies that members of a cluster are likely to share access to the same knowledge, same resources, and there will be low value-added from collaboration between very similar authors. On the other hand, collaborations with authors that are diverse, that belong to a different cluster, will result in access to different

[3]There could still be gains due to the division of labour, but this would not explain the diversity premia.

resources, different knowledge, hence it is more likely to stimulate creative, innovative work, for example through knowledge recombination. Thus, diversity is not important per se, but only to the extent that it is correlated with non-redundancy of social ties. This argument combines the 'strength of weak ties' (Granovetter 1973) with 'structural holes' (Burt 2004). Weak ties serve as channels through which non-redundant information flows between collaborators. Collaborators that have greater access to structural holes in the network are more likely to generate a higher quality publication. The advantage of the explanation based on social network is that it is not contingent on diversity, authors with same ethicity, gender and location can still benefit from collaboration if their social networks are non-redundant.

This paper uses a dataset of publications by 34 thousand economists, combining it with information on authors' location, most likely ethnicity and gender, to show that diversity premia exist for collaborations among economists. Then, by examining properties of the collaborators' networks, I show that controlling for the collaborators' access to non-redundant knowledge can explain (eliminate) the diversity premia. The dataset also includes information on working papers, including those that never transformed into a journal publication. This allows examining whether the same approach can be used to explain patterns of 'unsuccessful' collaboration (defined as collaborations that do not result in a journal publication).

The dataset is based on information collected by RePEc (*Research papers in Economics*) and related services. RePEc data contains name-disambiguation, allowing clear identification of an author's work on RePEc (see Zimmermann 2013), including working papers that remain unpublished for an extended period of time. The data also allows calculation of specialised citation counts, for example excluding self-citations or citations from the same country.[4]

The rest of the paper is organised as follows. Section 2 reviews related literature and explains the contribution to the literature. The description of data is provided in Section 3. Empirical framework and results are presented in Sections 4 and 5. The results are discussed in Section 6, which also contains the conclusions and suggestions for further research.

[4]Thelwall and Maflahi (2015) report an own-country bias in readership of scientific articles on a popular academic resource, Mendeley. If their finding is also true for the general population of academic readers, then this could explain the international collaboration premium, but not the other two premia. For robustness, this paper includes a citation measure that excludes own-country citations. See Section **??** for further details.

2 Related literature

The two relevant themes in the literature are: the link between collaboration and quality of the joint product, and the role of networks in the production and diffusion of knowledge.

Quality of an academic work is a broad concept and there are various measures used in the literature: citation counts (e.g. Hamermesh 2017), impact factors of the journals in which the collaborated publication is published (e.g. Rigby and Edler 2005) and, where available, peer-assigned ratings (e.g. Franceschet and Costantini 2010; Coupé 2013).

Each quality measure has its limitations. Citation counts can be influenced by factors unrelated to the paper content, for example the paper's sequential order in an electronic distribution list (Feenberg et al. 2017) or within a journal (Coupé et al. 2010). Publication in a prestigious journal may be affected by personal ties with the editor, though Brogaard et al. (2014) show that the editors do not seem to abuse their power. Peer assessment is not widely available as a quantitative measure and opinion of selected peers can differ significantly from perception by the broad scientific community, at least as reflected in gross citation counts (Coupé 2013). Lacking a single measure of quality, the literature on the effect of collaboration on quality typically uses multiple measures of quality.

Franceschet and Costantini (2010) use information from a research assessment exercise in Italian universities, which covered publications between 2001 and 2003. Every article included in the dataset received a rating from at least two peers and most of the articles were indexed in Thomson Reuters' Web of Science (WoS) database. Franceschet and Costantini (2010) interpret peer rating as a measure of quality and citation count as a measure of impact, however both peer rating and citation count can also be seen as measures of quality. The analysis shows that collaborated work is perceived to have higher quality and impact, a finding that is also reported by other researchers (e.g. Hamermesh 2013). However, what is the causal mechanism linking collaboration and quality?

The causality from collaboration to joint product's quality is likely to go via several channels. Jones (2009) convincingly argues that as the stock of existing knowledge increases, it becomes harder (costlier) for an individual to become an expert with both deep and broad knowledge. Acquiring deep and broad knowledge will leave less time for generating new knowledge, while choosing to specialise requires collaboration in teams. This explanation can be categorised as a 'human capital'-based approach, because the key explanatory role is played by authors' individual characteristics (albeit in relation to the changing macro-level properties).

Another causal channel is via the 'peer review', which is inherent in productive collaboration. By examining the research networks formed by collaborators within 22 R&D projects, Rigby and Edler (2005) show that more intense collabora-

tions reduce variability of quality. The intensity of collaboration is proxied by the density of the collaboration network within the project team, while the quality of collaboration for each project team is measured by the number of publications that had citation counts higher than the journal average in which they were published. This is an example of social network-based approach, where the key explanatory power is played by the authors' integration into the collaboration networks.

There is a large literature in sociology that examines the role of networks in the knowledge production process. Burt (2004) explores the hypothesis that a good idea is more likely to come to those individuals that are connected across different groups. Through empirical analysis of data on 673 managers at one of the largest electronics companies in the United States, Burt (2004) shows that managers who made and utilised connections between different groups were more likely to express an idea and discuss it with colleagues, these ideas were more likely to be engaged and judged valuable by senior management at the company. The effect of connection across groups on creativity is not that of "genius; it is creativity as an import-export business. An idea mundane in one group can be a valuable insight in another.", Burt (2004, page 388).

The link between network connectedness and creativity is explored further by Uzzi and Spiro (2005), they examine data on artistic and financial success of Broadway musicals from 1945 to 1989. By constructing a measure of the 'smallness' of the small world of artists, the authors are able to check to what extent the varying connectedness affects the success of the final product. A more connected artistic network stimulates diffusion of good ideas and practices between different clusters of artists, fostering creativity and fresh thinking, however once the network becomes too connected then the pool of ideas and practices becomes homogenous across groups, resulting in spread of common knowledge rather than novel ideas. Uzzi and Spiro (2005) find that the small world network effect on creativity is non-linear, with small increases in connectedness leading to better musical productions (both artistically and financially) and the effect turning negative after passing a certain a threshold.

The social network of economists can be described as a small world, a network with relatively low density in which it takes only several steps to connect any two authors (Goyal et al. 2006). Most economists today identify themselves as a part of an international profession, "being an economist means inhabiting not only a country-specific field, populated by fellow nationals, but also an international field", Fourcade (2009, p. 243). At the same time, individuals are influenced by country-level institutions and culture (Fourcade 2009; Montecinos and Markoff 2010). National borders through a mixture of constraints, from institution to funding, limit the diffusion of knowledge, resulting in knowledge which is embedded into national networks. In this case, strong ties are not likely to be a source of new

information Granovetter (1973). Individuals that are highly embedded only within a national network will also be more 'constrained' (i.e. exposed to redundant information) with access to the knowledge and resources embedded only in their network. By finding collaborators from foreign networks, an author can increase access to new, non-redundant information and, using Burt (2004)'s terminology, will be "at higher risk of having good ideas".

Access to new knowledge and diffusion of existing knowledge are greatly influenced by social networks. Kerr (2008) shows the role of U.S. ethnic networks in diffusion of knowledge to their home countries, a 'core to periphery' knowledge flow. However, networks also play a role in knowledge that flows in the other direction, from 'periphery to core'. Helgadóttir (2015) uses a case study of a specific theory, "expansionary austerity", diffused via a group of Italian economists, the "Bocconi boys", to argue that the transmission of ideas is not only from core to periphery, as often described in the literature, but ideas can also be diffused via networks in the other direction. Thus, generation of knowledge can occur in the peripheral states with diffusion to the more central countries. Networks, however, are a key ingredient in this process (Helgadóttir 2015). Another support for the role of networks comes from the analysis of how foreign-born US-educated engineers are helping the development of relevant industries in the sending countries by recombining resources embedded in the receiving and sending countries, Saxenian (2005).

Support for the role of networks also comes from the work on migration, especially academic migrants. Franzoni et al. (2014) use detailed information on migration history for a sample of more than 14000 scientists from 16 countries. Their analysis shows that migrant scientists outperform domestic scientists, even after correcting for potential selection bias of better scientists being more likely to migrate. Their results are consistent with theory of knowledge recombination, suggesting that it is the act of moving per se that enhances the performance of a migrant scientist. This evidence can also be seen in the light of the brokerage theory advanced in the present paper: the act of moving will result in a better integration into the destination-country's research network, hence allowing the migrant scientist to benefit from the arbitrage opportunities that may exist between the destination and origin research networks. A related finding is that return migrant scientists help to connect the domestic and the global research networks (Gibson and McKenzie 2014; Jonkers and Cruz-Castro 2013) and that return migrant scientists perform better than stayers (Baruffaldi and Landoni 2012). Movement of scientists allows expansion of collaboration over greater distances. Head et al. (2015) show that once personal ties are controlled for, the effect of distance on knowledge flows is greatly reduced or disappears in some specifications. Yet, distance does influence formation of personal ties, so being located within a particular

national space will influence an author's ties.

This paper makes a connection between these broad themes, collaboration-quality nexus and the role of networks in production/diffusion of knowledge, to explore what role the networks play in the quality of collaboration. The direction of knowledge flows is not examined here, and the underlying implicit assumption is that quality of collaboration is improved with combination of resources and knowledge from different networks, not exclusively from core to periphery. The key social network measure calculated for collaborators is the "network constraint" (Burt 2004), a measure of access to diverse, non-redundant information, with additional controls for author characteristics (human capital), as well as the number of collaborators (to capture potential non-linearities due to the economies of scale/scope).

The paper adds to the literature on the strength of weak ties and the importance of collaborators' social network by showing that access to non-redundant information greatly improves the quality of academic collaboration. It thus complements the studies that have confirmed this relationship in other areas (Burt 2004; Uzzi and Spiro 2005; Mohnen 2016; Burt 2015). The main contribution is the social network-based explanation of diversity premia, which reflect gains from collaboration across weak ties. The weak ties allow overcoming constraints of the author's homophilic network. Once the global social network information is taken into account, the diversity premia disappear.

3 Data description

3.1 Authors and works

The data used in this chapter comes from a popular Internet resource — *Research Papers in Economics* (RePEc) and its related services (CitEc, CollEc and EconPapers).[5] The dataset contains rich author-level information, including the list of publications, citation counts, co-authors, institutional affiliations and more. Table 1 shows information on the number of registered users and their works.

Several additional data sources were combined with RePEc data. First, author-provided affiliation was used to identify their geographic location, specifically country of affiliation, which was then used for calculation of physical (Mayer and Zignago 2011) and linguistic/cultural/genetic (Spolaore and Wacziarg 2015) distances between collaborators using aggregate, country-level information. Author names were used to identify their most likely ethnic group using name-based ethnicity

[5]Krichel and Zimmermann (2009) describe the challenges of keeping accurate bibliographic records and also provide a brief historical background for RePEc.

Table 1: Number of authors and collaborators among RePEc users.

	Registered users		
	total	as percentage of	
	count	authors	all users
Registered users	61,097		100.0
Authored 1+ works	47,129	100.0	77.1
Authored 10+ works	22,703	48.2	37.2
Authored 100+ works	1,102	2.3	1.8
Collaborated on 1+ works	37,796	80.2	61.9
Collaborated on 10+ works	13,599	28.9	22.3
Collaborated on 100+ works	256	0.5	0.4

Source: own calculations based on RePEc data. Notes: the number of authored works is adjusted for related works (see Appendix A); not all works are claimed by all of their authors, so the number of collaborations is likely to be underestimated.

matching software (Mateos et al. 2007). Further details on data processing are provided in Appendix A.

A useful feature of the data is that it contains information on journal publications and working papers,[6] and also allows to distinguish related works (e.g. working paper version of a journal publication). This distinction is important for obtaining accurate measures of citation counts (removing self-citations and citations from multiple versions of a single work) and publication patterns.

The information on collaborations allows constructing the social network of economists. The added advantage of using RePEc data is that it contains information on working papers which were not published - such collaborations are important in constructing a more complete picture of the connections among economists, yet they would be missed by approaches that use only information on publications. Table 2 shows descriptive statistics for the global research network constructed using information on both journal publications and working papers. The calculated measures are similar to those found in Goyal et al. (2006). There is a possible indication of self-selection of authors that register on RePEc reflected in the lower share of isolates (authors that do not collaborate with anyone else in the network) and lower average distance (the number of intermediate links connecting any two economists in the network). However, the magnitude of the differences is relatively small, keeping in mind that the calculations rely on different datasets, cover different time periods, use different identification of authors.[7]

[6] As argued in Zimmermann (2013), working papers constitute an important channel of scientific communication in Economics. Using only journal publications to reconstruct the social network of economists will bias the sample towards fruitful collaborations.

[7] In Goyal et al. (2006) authors are identified based on the last name and all the initials of their first names.

Table 2: The global research network.

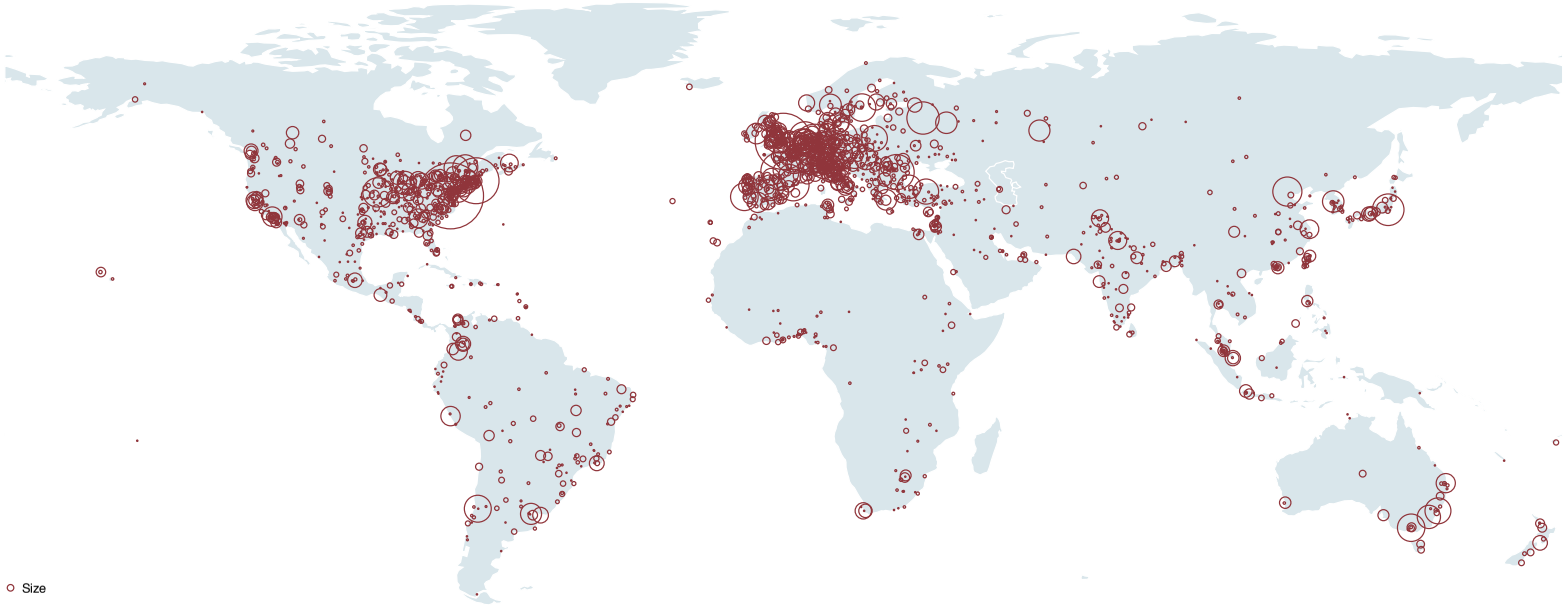
	Research networks based on journal publications	working papers	all works
Unique authors	28299	30380	34449
Giant component	26108	28494	32537
As share of total	0.92	0.94	0.94
Isolates	2191	1886	1912
As share of total	0.08	0.06	0.06
Degree:			
Median	3.00	3.00	4.00
Mean	4.88	5.56	6.20
St. dev.	5.53	6.33	7.21
Max	82	77	95
Distance:			
Median	6.19	5.93	5.70
Mean	6.38	6.07	5.87
St. dev.	0.98	0.81	0.86
Min	4.45	4.28	4.13
Max	14.57	13.84	13.77
Clustering coefficient (overall)	0.20	0.20	0.19
Clustering coefficient (average)	0.24	0.27	0.26

Source: own calculations based on RePEc data using *netsis* program in Stata. Note: the paper counts were not adjusted for related works; single-author publications are not included in the sample.

3.2 Collaboration networks and geography

Affiliation information allows identifying the location of an author, calculating the geographic diversity of the collaborators, as well as distinguish the domestic and global networks. Table 3 shows that there is a significant heterogeneity in network properties of the domestic networks. Some countries have a large 'giant component', defined as the largest set of interconnected authors, for example the United States, Netherlands, Austria. There are also countries where the network is represented by isolated groups of collaborators - for example Russia, India and China. Within the giant component, some countries are characterised by highly interconnected authors (large clustering), for example Austria, Russia and Poland, while others are relatively sparse - e.g. the United States, UK and Canada. Country networks also differ in patterns of average numbers of collaborators and how many 'handshakes' it takes on average to connect any two authors in the network (i.e. average path length, 'distance'). The differences across countries can reflect heterogeneity in the availability of resources (e.g. research funding may be allocated to areas that require or encourage strong collaboration), but also heterogeneity in network capital at macro- and micro-levels.

Figure 1: Geographic distribution of the economists.



Source: own calculations based on RePEc data. Notes: the diameter of each circle is proportional to the number of authors registered on RePEc at a given location; the location coordinates were rounded to the first decimal point precision and then aggregated for simpler visual representation.

Table 3: Domestic network statistics for the top 30 countries by number of collaborators.

		Percentage of	Degree				Distance					Clustering coeff.	
Country	N	isolates	Median	Mean	St. dev.	Max	Median	Mean	St. dev.	Min	Max	Average	Overall
United States	7316	0.03	4	5.53	5.49	56.00	5.24	5.36	0.71	3.77	10.12	0.19	0.13
Germany	2179	0.07	3	4.92	5.95	63.00	4.67	4.75	0.75	3.02	9.01	0.27	0.25
France	2171	0.08	3	4.38	4.43	42.00	5.01	5.19	0.89	3.51	10.34	0.23	0.19
United Kingdom	2054	0.10	2	3.77	3.87	37.00	5.16	5.30	0.83	3.69	9.17	0.20	0.17
Italy	2000	0.08	3	4.38	4.03	36.00	5.07	5.23	0.88	3.61	11.50	0.26	0.21
Spain	1406	0.15	3	3.45	2.81	22.00	5.60	5.73	0.94	3.89	10.61	0.27	0.25
Canada	927	0.12	2	3.14	2.65	20.00	5.36	5.49	0.87	3.81	9.50	0.16	0.17
Australia	891	0.07	3	3.79	3.43	24.00	5.17	5.27	0.81	3.77	8.54	0.23	0.20
Netherlands	756	0.05	3	4.22	4.03	34.00	4.56	4.67	0.69	3.31	7.65	0.26	0.18
Romania	659	0.37	2	2.79	2.29	22.00	4.58	4.76	1.04	3.06	8.20	0.30	0.33
Japan	562	0.10	2	3.35	3.19	18.00	4.88	5.00	0.82	3.39	9.23	0.20	0.21
Switzerland	487	0.18	2	2.91	2.48	14.00	5.78	6.02	1.14	4.02	10.30	0.23	0.27
Belgium	465	0.10	2	3.55	3.46	21.00	4.51	4.63	0.81	3.07	8.29	0.31	0.23
Sweden	442	0.12	3	3.57	3.03	18.00	4.64	4.76	0.85	3.17	9.05	0.24	0.20
Russia	441	0.54	1	2.86	3.80	30.00	2.71	2.93	0.81	1.77	5.16	0.27	0.42
Turkey	434	0.33	2	3.22	3.11	21.00	3.23	3.39	0.67	2.27	5.94	0.24	0.24
Brazil	409	0.12	2	3.14	2.72	19.00	4.61	4.76	0.82	3.29	8.24	0.19	0.18
Portugal	396	0.18	2	2.73	2.25	13.00	5.19	5.42	0.99	3.51	8.62	0.22	0.25
Colombia	330	0.11	2.5	3.78	3.53	28.00	4.12	4.22	0.81	2.65	6.90	0.27	0.20
Greece	308	0.22	2	2.84	2.36	16.00	4.17	4.49	1.10	2.68	9.88	0.25	0.21
China	272	0.56	1	2.14	2.10	18.00	3.04	3.01	0.64	1.87	5.07	0.23	0.23
India	267	0.66	1	2.06	1.68	15.00	2.10	2.18	0.58	1.32	4.26	0.16	0.20
Norway	261	0.11	3	4.02	3.24	22.00	4.05	4.28	1.00	2.78	8.20	0.27	0.23
Poland	260	0.40	2	2.93	2.47	16.00	3.14	3.23	0.70	1.95	5.29	0.35	0.38
Chile	247	0.06	3	3.49	2.79	16.00	4.78	5.30	1.62	3.51	12.44	0.19	0.17
Austria	246	0.07	3.5	6.05	6.03	27.00	3.53	3.73	0.82	2.46	8.13	0.34	0.45
Denmark	235	0.11	2	3.22	2.90	17.00	4.44	4.58	0.83	3.15	7.72	0.23	0.22
Czech Republic	215	0.29	2	3.39	3.47	24.00	2.69	2.82	0.58	1.82	4.58	0.24	0.25
Pakistan	204	0.26	2	3.17	3.00	20.00	2.92	3.06	0.56	1.98	4.93	0.32	0.25
South Africa	159	0.13	2	3.32	3.37	27.00	3.66	3.78	0.66	2.60	5.42	0.30	0.21

Source: own calculations based on RePEc data using *netsis* program in Stata. Note: both journal publication and working papers were used to determine collaborations; all calculations are unweighed; single-author publications are not included in the sample.

3.3 Diversity premia

The result of a collaboration, on average, receives more citations than a single-author work. However, not all collaborations are alike. Collaborations in which authors differ along some measure of diversity receive greater number of citations. The citation premium can be observed for collaboration across countries (Table 4), gender (Table 5) and ethnicity (Table 6). The international collaboration premium is about 15%, the mixed-gender premium is 25% compared to female only teams[8], and the inter-ethnic premium is 30%.[9]

Table 4: Summary statistics for citations by type of collaboration.

	Net citation count					
	Median	Mean	St. dev.	Min	Max	N
Journal publications						
Domestic	2	11.11	45.50	0	2748	62383
International	3	12.75	44.12	0	3485	39279
Working papers						
Domestic	0	2.32	9.22	0	484	51281
International	0	2.99	9.92	0	376	33093

Source: own calculations based on RePEc data. Note: the citation counts were adjusted for self-citations and related works.

Finally, there is a premium on collaboration between authors with different gender, see Table 5.

Table 5: Summary statistics for citations by inter-gender collaboration.

	Median	Mean	St. dev.	Min	Max	N
Journal publications						
All female	1	7.22	37.05	0	1170	2634
All male	2	12.86	47.29	0	3485	67648
Inter-gender	2	9.08	32.28	0	1964	19065
Working papers						
All female	0	1.89	5.94	0	124	2302
All male	0	2.74	10.05	0	484	53062
Inter-gender	0	2.32	8.27	0	282	18824

Source: own calculations based on RePEc data. Note: the citation counts were adjusted for self-citations and related works.

There is also a premium for inter-ethnic collaborations, see Table 6.

[8]Putting male-only and female-only collaborations into an 'own-gender' category will result in a diversity discount rather than premium. The table is decomposed to show the differences between the two own-gender collaborations. See further discussion of possible explanations for this pattern in Section **??**.

[9]Further descriptive statistics can be found in Orazbayev (2016).

Table 6: Summary statistics for citations by inter-ethnic collaboration.

	Median	Mean	St. dev.	Min	Max	N
Journal publications						
Own-ethnic	2	10.02	37.12	0	2748	35415
Inter-ethnic	3	13.11	48.33	0	3485	53943
Working papers						
Own-ethnic	0	2.30	8.71	0	377	28147
Inter-ethnic	0	2.79	9.99	0	484	46044

Source: own calculations based on RePEc data. Note: the citation counts were adjusted for self-citations and related works.

4 Empirical framework

4.1 Collaboration and quality

A stylised model of academic collaboration is described first. During the academic production process, several authors contribute their knowledge and resources and their joint work leads to a finished product (journal publication) or a semi-finished product (working paper). The quality of the product will depend on characteristics of the individual authors (human capital), their networks (network capital), as well as the size of the team (economies of scale).

Quality of the product will be measured by the net citation count. Net citation count is calculated as the gross citation count adjust for the related works (i.e. citations from multiple versions of the same work) *and* self-citations.[10]

An alternative proxy for quality is the journal (or working paper series) impact factor. The advantage of this approach is that it proxies for quality as perceived by the editor/referees at the time of publication, the disadvantage is that it is a very noisy proxy, e.g. Seglen (1997) shows that the distribution of citations to articles within a journal is very uneven, with 15% of the articles accounting for 50% of the citations and 50% of the most cited articles account for 90% of the citations. Figure 2 summarises the difference between these measures of quality. The results based on this measure are not reported, they are similar qualitatively to the results based on citation measure, but the diversity premia remain after controlling for the global network constraint.

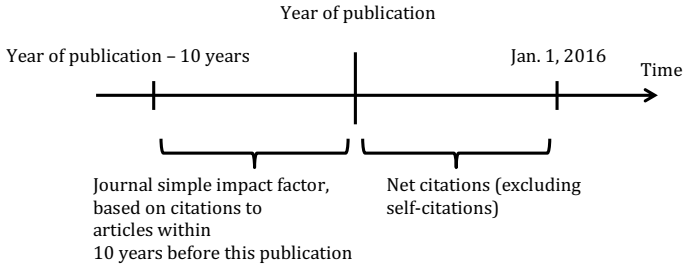
The simplified production function for academic work p can be summarised as:

$$Q_p = F(H_p, N_p, L_p), \tag{1}$$

where Q_p is quality of paper p, H_p represents human capital of the authors (for example, average research age or average quality of their prior publications), N_p

[10] A citation is considered to be a self-citation if at least one of the authors of the citing authors was an author of the cited work.

Figure 2: Measures of quality: impact factor and net citations.



Note: the diagram uses 10 years as an example of time period for calculation of simple impact factors.

represents network capital of the authors (for example, their access to diverse, non-redundant information), and L_p is the number of authors. The marginal impact of all variables is expected to be positive, with potential non-linearities in N_p (e.g. as in Uzzi and Spiro 2005) and L_p (due to increasing communication/coordination costs).

The quality of a paper is modelled to depend only on the information at the time of production, so it should be independent of the observation year t.[11] However, citation counts will vary with observation year t, so this measure of quality must be adjusted for potential influence of awareness/diffusion factors (compare with Head et al. 2015, where awareness is modelled at bilateral, citing-cited paper level). In this chapter, citations are aggregated, so awareness is a non-decreasing measure, though it can grow slower over time.). The influence of awareness/diffusion factors will be captured by the age of publication t_p and its squared value to capture potential non-linearity where awareness/diffusion slows down with the age of publication. Ideally, JEL code dummies would also be included, however JEL code information was readily available only for a relatively small sub-sample.[12]

Human capital will be measured by the research age of the authors at the time of publication, their prior citations and the number of prior publications.[13] Research age should capture authors' experience, which should have a positive (or at least non-decreasing) effect on quality. Prior citations and publications should capture author's ability/productivity as reflected in their prior work.

Network capital will be measured using "network constraint", a measure that summarises a variety of information about an author's network to reflect the extent

[11] Generally, the publication cannot be modified once published. See Mongeon and Larivière (2015) and Jin et al. (2013) for an analysis of the rare cases — article retractions.

[12] See Table B.1 in Appendix B.

[13] Additional specifications that used prior average citations per paper and average citations per paper-year were checked and the main results continue to hold.

to which an author's network provides the author with diverse, non-redundant information. This measure and its calculation for domestic and global networks are explained in the following subsection.

The variable of interest, dummy for international collaboration, is calculated using information on the author's affiliation at the time of observation t. This is not necessarily representative of the author's location at the time of publication, however at the present moment this information is not readily available.[14] Using this information, it is possible to classify collaborations to be international or domestic. Dummy variable I_p is set to equal to 1 if there are at least two different countries among collaborators that worked on paper p.

The information from RePEc is collapsed to publication level using average value of the variables for human and network capital.[15] The regression equation used to estimate the stylised model is:

$$\text{net citation}_{p,t} = exp(\alpha H_p + \beta N_p + \gamma L_p + \omega I_p + \delta_1 t_p + \delta_2 t_p^2 + \epsilon_p). \quad (2)$$

The equation is estimated for the sample of journal publications[16] using Poisson for the citation count, all specifications use robust standard errors.

The random term ϵ_p capture citations due to unexplained factors, something that is not captured by the variables used in the equation.

4.2 Domestic/global networks, structural holes and network constraint

The country of each author allows distinction between the global network, the network formed by examining collaborations among all authors regardless of their location, and the domestic networks, which are formed using only collaborations of authors located in a particular country. Figure 3 shows the distinction between the global and domestic networks on an example.

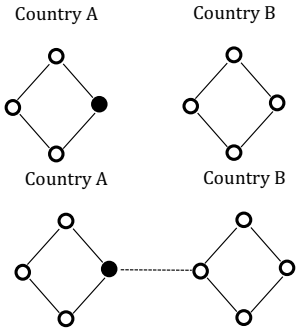
Using information on the date of publication, it is possible to construct two versions of a network - dynamic and persistent. The persistent network assumes that collaboration links between coauthors are permanent, while the dynamic network assumes that links between collaborators last only a specific period of time. Journal publications are a relatively rare phenomenon for the median author, they occur once in several years. As a result, constructing dynamic network using only actual publication year might result in networks that are very fragmented, overstating the number of isolates.

[14] The author is working on obtaining author's location at the time of publication.

[15] Alternative calculations, where the values were collapsed using maximum values, were also conducted. The main results continue to hold with that approach.

[16] The results for working papers are similar.

Figure 3: The domestic and global networks: an example.



Note: the top diagram shows domestic networks; the bottom diagram shows the global network; the black dot is an author that becomes gains access to additional, non-redundant information in the global network as a result of international collaboration (dashed lines).

Authors that collaborated in a particular year, but not in the subsequent year, are likely to have an access to their collaborator for a number of years. Moreover, the long publication lag in Economics might lead to a situation where actual publication date reflects work conducted several years ago. In a similar context, Uzzi and Spiro (2005) apply a 5-year window, so that each link lasts for 5 years since the last interaction. To take into account these shortcomings of using the publication year, every collaboration is assumed to start 3 years prior to the publication year and to last 3 years past the collaboration year, where number 3 was chosen as a rough estimate of the time from the first submission to a journal to eventual publication.

The networks can further be distinguished by the level of analysis - ego or whole network. In ego analysis, the network measures are calculated for each author, *ego*, using information on ego's collaborators that are one step apart (i.e. direct collaborators), two steps apart (i.e. including authors with whom ego's collaborators published joint work) or more.

In order to determine the most important nodes in a network, the literature typically uses the following measures of network 'centrality': betweenness, degree and closeness. Betweenness is a measure of how often an author appears on the shortest path between any two authors in the network. Degree centrality is simply a measure of the size of an authors' network. For example, if author A collaborated with B, C and D, then that author's degree is 3. Closeness is the average distance of an author to all other authors in the network. These measures have limitations when it comes to measuring access to diverse, non-redundant information. Betweenness can often result in extremely high values that are many orders of magnitude above the median level in a network. This can be especially

problematic when comparing domestic networks of different size.[17] Closeness and degree centrality measures do not distinguish between redundant (i.e. if an author collaborators with authors are also linked to each other) and non-redundant collaboration links. Non-redundant ties are important because they will be a potential source of new information.

Non-redundancy in ties implies that some of the authors in the network are not connected with each other. This absence of a link suggests a potential opportunity due to a gap in information flow, a "structural hole". Burt (2004) calculates an ego-level measure, "network constraint", which measures the extent to which an authors' network lacks structural holes. The network constraint summarises size, density and hierarchy of an individual in the social network. High values of network constraint indicate that an individual's network is dense (in this chapter - collaborators are well-connected with each other) or hierarchical (in this chapter - collaborators share information via a central contact). In a network with high constraint, there are fewer structural holes, since an individual's connections already have ways of communication with each other. Individuals with a low network constraint, on the other hand, will have greater access to structural holes in the network.

Following Burt (2004), the bilateral network constraint c_{ij} for every author i and their collaborator j was calculated as:

$$c_{ij} = (p_{ij} + \sum_{q \neq i,j} p_{iq} p_{qj})^2,$$

where p_{ij} is a measure of the strength of connection between i and j and was assigned value of 0 if i and j did not collaborate, and $\frac{1}{N_i}$ if i and j collaborated, where N_i is the number of i's collaborators.

The values were then summed across all collaborators j, giving constraint index for each author i:

$$C_i = \sum_j c_{ij}.$$

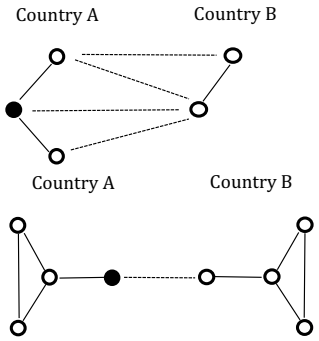
The constraint measure was calculated separately for domestic and the global networks.[18] International collaboration allows an author to connect to a different network, which will generally lead to a lower value in the author's global network

[17] This can be mitigated somewhat by normalisation that brings betweenness measure to $[0, 1]$ range, however the relative values of betweenness for central authors remain much higher than the median value. Another possibility is to calculate ego-betweenness, however this still results in very large differences among authors.

[18] As in Burt (2004), network constraint was normalised to [0,100] range and domestic isolates, i.e. authors that collaborate only internationally, were assigned domestic network constraint of 100. The results are also not sensitive to discarding domestic isolates. See Appendix C for examples.

constraint. Global constraint for authors that collaborate internationally is on average 17, while for authors that collaborate domestically it is about 25.

Figure 4: Network constraint in the domestic and global networks.



Note: the top diagram shows an example of an author (black dot) that is more constrained globally, than domestically; the bottom diagram shows an example of an author (black dot) that is less constrained globally, than domestically; dashed lines indicate international collaboration.

An author's position in the domestic and the global networks is highly correlated (0.71), however a high domestic constraint does not necessarily imply a high global constraint. Figure 4 shows two examples of the changes to the measure of constraint when considering domestic and global networks. If an author is linked to coauthors that collaborate internationally, then the author can end up being marginal in the global network, even if they are less constrained in the domestic network. On the other hand, if an author is weakly integrated into the domestic network, but collaborates extensively internationally, then the author can end up being central within the global network.

4.3 Identification strategy

The objective of this chapter is to test whether the proposed theoretical framework can explain the diversity premia associated with international, inter-ethnic and inter-gender collaborations. The estimations will use actual, observed data, and so the estimations must address potential endogeneity issues. Endogeneity arises due to non-random pairing of collaborators. One argument could be that authors' with high productivity could attract more collaborators and produce output of better quality (or research on more challenging topics). To address this concern, the estimated equations include a measure of authors' productivity at the time of collaboration (i.e. based on authors' historical publications up to the year of collaboration). Also, the network capital measures are calculated at the time of publication, so they exploit dynamic variation in collaborators' social ties.

It's also important to control for factors that can affect diffusion of knowledge. For example, a higher citation count may reflect greater awareness of an academic work rather than its intrinsic quality. Collaboration between authors who have diverse reading audiences can result in better diffusion of the joint product, raising its 'measured' quality. To address this additional controls are used: for the age of publication, number of collaborators, but also the dependent variable is modified to exclude self-citations or citations from the collaborators' countries.

Also, the information in the dataset allows computing author characteristics at the time of publication. This provides an accurate measure of 'human capital' variables, such as research age, number of prior publications and prior citations. For example, year of the first publication by an author is defined as a 'research birth year', b_a , and so the research age of author a at the time of publication p which was published in year t_p is calculated as: $\text{age}_{a,p} = t_p - b_a$. Analogous procedure is used to calculate citations and publications by author a prior to publication p, for example: prior citation $= \sum_{t=0}^{t_p-1} \text{citations}_{a,t}$.

4.4 Diversity and collaboration quality

As an extension of the main calculations, and to gain a better understanding of the source of the differences between networks, further information on collaboration characteristics is added to the estimated equation. Specifically, several measures of distance are added - physical, cultural/linguistic/genetic. Greater distance, esp. physical, will raise the cost of collaboration (though this can be mitigated by existence of personal ties). Similarly, greater cultural/linguistic distance will result in greater communication costs. However, there also can be gains from having coauthors that come from a different country or culture.

National borders limit the diffusion of knowledge. Also, national-level policies on funding can affect availability of resources. So networks from different countries can accumulate different pools of knowledge.

The estimated equations are modified to include additional terms for physical, cultural, linguistic and genetic distances between collaborators. For papers with more than two collaborators the paper level distance was calculated as either the mean of bilateral distances or the minimum. This approach is similar to Singh (2005) and Head et al. (2015), and assumes a perfect information flow within a collaboration team.

5 Results

The main purpose of the regression analysis in this section is to examine the role of "network constraint" (as a proxy for the *lack* of network capital) in collabora-

tion quality, as well as to examine how international collaboration affects access to diverse, non-redundant information. In specifications that use global network constraint, which is calculated without regard to national borders, we would expect international collaboration dummy to be insignificant, because this would be consistent with international collaboration being merely a way of expanding access to non-redundant information. However, in specifications that use the domestic network constraint, which is bounded by the national borders, we would expect the value of international collaboration dummy to be positive, consistent with the literature that finds international collaborations to be of better quality. The magnitude of the gain from international collaboration is expected to be larger for authors that are more constrained domestically, because such authors are exposed to primarily redundant information, so the marginal impact of non-redundant information should be larger.

For the remaining variables, it is expected that age of publication will be positively correlated with net citations with a possible slow-down due to saturation of awareness/diffusion processes (i.e. negative coefficient on the squared age of publication). There could be non-linearities in the role of the number of collaborators, but generally the coefficients are expected to be increasing with the number of collaborators until gains from the economies of scale are exhausted. Measures of human capital are expected to have positive coefficients, indicating benefits of accumulating experience as reflected in research age, prior publications and prior citations.

The results using net citation count as a proxy for quality are presented in Table 7. Starting with the role of international collaboration, it can be seen that international collaboration significantly improves net citation count (column 1), but this significance disappears after the global network constraint (i.e. access to non-redundant information without regard for national borders) is included in the regressions (column 3). This suggests that the role of international collaboration is in expanding access to non-redundant information.

The marginal benefit of access to non-redundant information was expected to be larger to individuals that are more constrained domestically, and column 2 of Table 7 reports this pattern. This pattern can also be seen from the predictive margins for domestic and international collaboration, see Figures 5 and 6. It can be seen from Figure 5 that the marginal impact of international collaboration is positive and greater for authors that are constrained domestically, while Figure 6 shows that the effect of international collaboration disappears after controlling for position of the authors in the global collaboration network. Together the figures imply that the role of international collaboration is in improving access to non-redundant information.

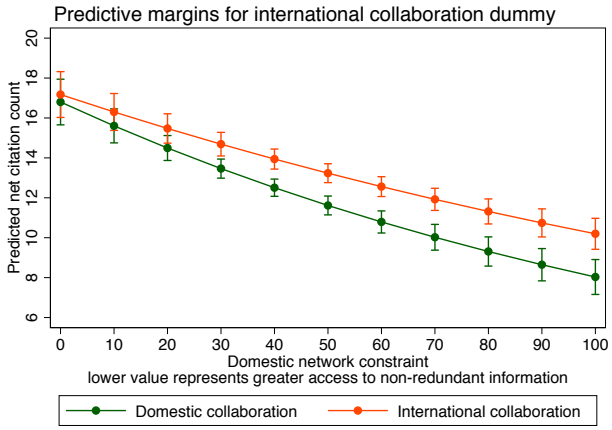
The coefficients on the team size (i.e. number of authors) suggest that as the

number of authors increases, the marginal impact on the quality is diminishing, with no significant advantage from the fifth author and beyond. This is consistent with results reported in the literature (Hamermesh 2017).

Role of awareness/diffusion was controlled through the age of publication and its squared value. The coefficients on these terms suggest that initially the impact of awareness/diffusion is large, but as the publication gets older the marginal impact of an additional year diminishes.

Contrary to the expectation, the coefficients on research age and prior publications of the authors are negative. They were expected to be positive, reflecting advantage from accumulated experience. The main reason for this result is that prior citations seem to capture a large part of the variation due to experience. Holding prior citations constant, being older or publishing more papers (i.e. on average prior papers received fewer citations) is associated with lower quality. Similar findings have been reported in the literature (Oster and Hamermesh 1998; Hamermesh 2015).[19]

Figure 5: Margins for international collaboration - domestic constraint (net citation count).



Source: own calculations based on RePEc data. Notes: paper counts are adjusted for related works; citations are adjusted for self-citations; network constraint was calculated separately for the dynamic global and domestic networks; lower values of network constraint indicate access to a more diverse, non-redundant information; journal dummies are included in all specifications; the equations are estimated using Poisson with robust standard errors.

[19]Somewhat related result from Burt (2004) is that, holding everything else constant, older managers were less likely to receive positive evaluation of their performance.

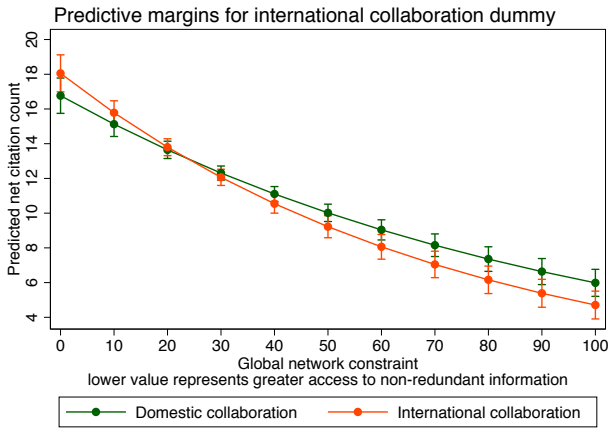
Table 7: Importance of network constraint for the dynamic network (net citation count).

	Net citation count			
Three authors	0.216	0.216	0.217	0.216
	(0.034)***	(0.034)***	(0.034)***	(0.034)***
Four authors	0.389	0.395	0.409	0.403
	(0.096)***	(0.095)***	(0.094)***	(0.094)***
Five+ authors	0.137	0.138	0.132	0.127
	(0.125)	(0.126)	(0.124)	(0.123)
Years since publication	0.256	0.256	0.256	0.256
	(0.008)***	(0.008)***	(0.008)***	(0.008)***
Years since publication, sq.	-0.007	-0.007	-0.007	-0.007
	(0.000)***	(0.000)***	(0.000)***	(0.000)***
Authors - age at the time of publication	-0.009	-0.009	-0.008	-0.008
	(0.002)***	(0.002)***	(0.002)***	(0.002)***
Authors - prior publications	0.002	0.002	0.002	0.002
	(0.001)***	(0.001)***	(0.001)***	(0.001)***
Authors - prior average citations	0.003	0.003	0.003	0.003
	(0.000)***	(0.000)***	(0.000)***	(0.000)***
International collaboration	0.103	0.022	0.008	0.074
	(0.026)***	(0.048)	(0.026)	(0.043)*
Domestic constraint	-0.006	-0.007		
	(0.001)***	(0.001)***		
Domestic constraint x International collaboration		0.002		
		(0.001)**		
Global constraint			-0.011	-0.010
			(0.001)***	(0.001)***
Global constraint x International collaboration				-0.003
				(0.001)**
Pseudo R2	0.49	0.49	0.49	0.49
N	56,432	56,432	56,432	56,432
Journal dummy	Yes	Yes	Yes	Yes

* $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

Source: own calculations based on RePEc data. Notes: paper counts are adjusted for related works; citations are adjusted for self-citations; network constraint was calculated separately for the dynamic global and domestic networks; lower values of network constraint indicate access to a more diverse, non-redundant information; journal dummies are included in all specifications; the equations are estimated using Poisson with robust standard errors.

Figure 6: Margins for international collaboration - global constraint (net citation count).



Source: own calculations based on RePEc data. Notes: paper counts were adjusted for related works and self-citations were removed; journal dummies are included in all specifications, the results with journal simple impact factor are qualitatively the same (see Figure **??**); the equations are estimated using Poisson with robust standard errors.

5.1 Robustness

The robustness of conclusions about the importance of network constraint for quality and the role of international collaboration as a mechanism of overcoming domestic network constraint will be checked in this section. The results will be checked by using different measures of the collaboration network: the persistent network, which tracks links between all collaborations regardless of when they took place; the recent network, which is based only on information for the recent years; and the ethnic networks which are based on the ethnicity of collaborators.

The persistent networks

A possible concern could be that during collaboration authors share knowledge that remains with the authors after the collaboration ends and can be used by them in future work independently (without going back to the original collaborators). In this case, the assumption of dynamic network will result in attribution of some of the 'network capital' to 'human capital' of the authors. For example, suppose an author begins a new collaboration, but makes use of knowledge acquired in an earlier collaboration (with other coauthors). Holding everything else constant, if the link to prior collaborators is no longer in the data, then the contribution of that knowledge will be given to author's characteristics ('human capital'). To

address this concern, the global and domestic persistent networks are created.

The global persistent network is constructed using information on all collaborations, regardless of the national borders and the year of collaboration. The domestic persistent networks uses information on all collaborations between authors residing in a country.[20]

Calculations that use the persistent networks are qualitatively the same.

Academic migration

One shortcoming of the dataset used in this chapter is that only the latest affiliation of an author is used when determining their domestic network.[21] However, an author could migrate between different countries during the years in the sample, when the dataset is constructed. Suppose an author spends early part of their career abroad and returns to home country towards the end of the sample. From the dataset we would observe the author's last location, home, but would not know about the migration episode. When the author's domestic network constraint is calculated for publications in their early years (when they were abroad), the author will have a high domestic network constraint or even appear as an outlier in the domestic network, because they were collaborating with authors abroad.

The inability to track an author's location at the time of collaboration will lead to two potential problems: one is mistaken classification of some domestic collaborations as international and of some international collaborations as domestic, the second problem is that the measured value of domestic network constraint may be inaccurate (the global network constraint is calculated without regard for national borders, so it is not affected by movement of authors).

Assuming that on average international collaboration is of greater quality (conditional on network capital), the average international collaboration as measured in the dataset will have lower citation count (because it is mixed with some domestic collaborations which are mistakingly classified as international) and domestic collaborations, on average, will have higher citation count (because some of the actual international collaborations are classified as domestic in the data). So, the results obtained in this chapter should be even larger when used with data that accurately tracks the authors' locations.

The second potential problem caused by cross-border movement of authors is mis-measurement of domestic network constraint. Without knowing accurate movements of all authors it's not possible to know whether the domestic network constraint is over- or underestimated in the earlier calculations. Assuming that

[20] Based on the last known affiliation, see discussion on academic migration below.

[21] This is not a criticism of RePEc data, which is of very high quality. Rather extraction of publication-level affiliation raises additional challenges. A dataset that includes affiliation of authors at the time of publication is under construction.

there is no consistent over- or underestimation, an 'error-in-variables' argument would suggest that the estimated coefficient on domestic network constraint is likely to be biased towards zero (in the context of this chapter implying a lower importance of network capital).

To reduce the influence of migration, calculations were also performed for papers that are less than 10 years old. The argument is that individual authors are more likely to have been in the same location for the more recent sample.

Qualitatively the results are very similar - the marginal effect of international collaboration is positive when controlling for domestic network constraint, but disappears once the global network constraint is used. There are several exceptions: the results for 1-year sample are not statistically significant; the marginal impact of international collaboration dummy remains positive and statistically significant for low global network constraint in the sample of papers less than 9 years old. The lack of significance for 1-year sample is not alarming, since there wasn't sufficient time for papers to be cited. The persistent significance of international collaboration premium for low values of network constraint (i.e. high access to non-redundant knowledge) could imply that migration of highly-connected researchers reduces the international collaboration premium. An alternative explanation is that authors with a low global network constraint benefit from a faster diffusion of their joint product, initially attracting citations faster than papers produced through domestic collaboration.[22]

Ethnic networks

Recent literature shows that collaborations between authors of different ethnicities receive more citations (e.g. Freeman and Huang 2015). Freeman and Huang (2015) perform analysis of a much larger sample of scientific papers, and find that controlling for prior publication history does not eliminate the premium of inter-ethnic collaboration.[23]

What could explain this result? The underlying mechanism could be related to discriminatory preferences, so that researchers discriminate towards own ethnicity, or, as some studies controversially suggest, there might be cognitive differences among different ethnic groups (e.g. Nisbett 2003). However, there is a more practical reason for why inter-ethnic collaboration can be beneficial. Ethnicity can act as a proxy for the type of knowledge acquired early on, possible migration

[22]This can be checked by calculating a citation count measure that includes citations only within the first N years. If the international collaboration premium can be found for N less than 10 using the full sample, then this would be indicative of the speed of diffusion argument rather than impact of migration. These calculations are in progress.

[23]Their paper is formulated in terms of homophily, so they find that own-ethnicity collaboration leads to a work with lower citations and impact.

history (even if more than one generation ago), and other experiences that can result in accumulation of different knowledge and access to different networks. Hence, own-ethnic collaborators are likely to share access to similar information and resources.

It's highly unlikely that ethnicity *in itself* plays a role in determining collaboration partners, rather that ethnicity can proxy for differences in network capital. Hence, categorising collaborations by ethnicity will show that inter-ethnic collaborations are, on average, more productive. To examine if this is true for the dataset on economists, the names of collaborators were matched to their most likely ethnic group using Onomap software, see the Appendix for details. This is not a 100% accurate identification of true ethnicity, however studies show that the accuracy of name-ethnicity matching is generally high (see references in Nathan 2015).

Table 6 shows that on average inter-ethnic collaboration is cited more often, though the result is very small for the sample of working papers.

Using each author's most likely ethnic group, it is possible to construct 'ethnic networks', which show collaborations among authors with the same ethnicity. This is analogous to the domestic networks used in the main calculations. If the argument from the previous section, that international collaboration allows collaborators to access non-redundant information, is valid, then the same pattern would be expected for inter-ethnic collaboration.

The results for the marginal impact of inter-ethnic collaboration show similar pattern to the one found for international collaboration. The marginal impact of inter-ethnic collaboration is higher for those that have high network constraint in the ethnic network, but once the global network constraint is taken into account the marginal effect of inter-ethnic collaboration becomes negligible.

Unlike geographical location, ethnicity is given and cannot be changed.[24] This means that migration or individual's action are unlikely to change membership in a particular ethnic network, and the ethnic network constraint can be a good proxy of access to the resources embedded in the ethnic network.

Selection effect

Another possible explanations of the results could be that there is a selection effect in which authors that have higher ability collaborate internationally. While the regression specifications include a measure of the authors' 'human capital', it is possible that there are other author-specific factors that are not included in the regression, but which *cause* both a higher quality and international collaboration. Ideally, author-specific fixed effects (or author dummies) would be included to

[24] A person's *name* can change, which might lead to a different recoding of ethnicity. Life events, e.g. marriage, could result in a mismatched ethnicity, hence transferring the effect of own-ethnic collaboration to inter-ethnic collaboration.

control for the unobserved heterogeneity, but the number of dummy variables exceeds variable limit in Stata.

Freeman and Huang (2015) faced a similar problem when examining possible selection effect based on ethnicity of authors. Their conclusion is that controlling for prior publication history of the authors reduces, but does not eliminate, the negative correlation between impact/citation percentile and own-ethnicity collaboration. Their specifications include dummies for location (specific US states), publication year, subfields and interactions between publication year and subfields. The results provide some support for selection effect, but the selection effect does not completely explain the inter-ethnic premium.

Table 8 shows marginal effects based on calculations that are similar to Freeman and Huang (2015). Consistent with Freeman and Huang (2015), excluding human capital measures raises the magnitude of coefficients on the international collaboration dummy and network constraint. This indicates some positive selection into international collaboration, but the effect is rather small. The results for the inter-ethnic collaboration are qualitatively the same.

In the main calculations, after controlling for the global network constraint, the international collaboration premium disappears. This gives further support to the role of international collaboration in providing access to non-redundant information.

Table 8: Exploring selection effects (net citations).

	Net citation count			
Three authors	3.078	3.052	3.082	3.037
	(0.495)***	(0.496)***	(0.446)***	(0.445)***
Four authors	7.330	9.009	8.521	9.837
	(1.753)***	(1.886)***	(1.816)***	(1.982)***
Five+ authors	2.237	1.394	3.052	2.192
	(1.753)	(1.712)	(1.856)	(1.810)
Years since publication	1.182	1.216	1.092	1.117
	(0.026)***	(0.027)***	(0.024)***	(0.023)***
Authors - age at the time of publication	-0.137		-0.149	
	(0.024)***		(0.024)***	
Authors - prior publications	-0.020		-0.013	
	(0.010)**		(0.010)	
Authors - prior citations	0.003		0.003	
	(0.000)***		(0.000)***	
Domestic network constraint	-0.082	-0.090		
	(0.008)***	(0.008)***		
International collaboration	1.440	1.460		
	(0.337)***	(0.341)***		
Ethnic network constraint			-0.059	-0.061
			(0.007)***	(0.007)***
Inter-ethnic collaboration			1.279	1.315
			(0.319)***	(0.316)***
Pseudo R2	0.49	0.48	0.46	0.46
N	57,549	57,549	55,061	55,061
Journal dummy	Yes	Yes	Yes	Yes

* $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

Source: own calculations based on RePEc data. Notes: paper counts are adjusted for related works; citations are adjusted for self-citations.

5.2 The link between diversity and quality

Results in the previous section are consistent with the role of international collaboration in overcoming the domestic network constraint and providing access to non-redundant information. However, why would information from different countries be necessarily non-redundant?

One explanation is that institutions in different countries might provide different incentives and impose different constraints on individuals, shaping the pool of knowledge around country-specific goals. For example, Fourcade (2009) provides a comparative overview of the development of Economics as a profession in three countries — France, UK and USA, showing that country-level institutions influence what it means to be an 'economist' in each country. This can explain, for example, the historical interest of British economists in welfare questions, the mathematical approach of the French economists (with a special focus on industrial organisation) and the quantitative approach, due to competition with 'hard' sciences for NSF funding, in the United States.

However, assuming heterogeneity only at the country-level would not be consistent with the empirical results. For example, it can be seen on Figure 5 that international collaboration premium disappears for authors that are very central domestically. This implies that heterogeneity occurs at the individual author level, rather than exclusively at the country-level. However, individuals can also internalise some of the country-specific 'network capital', and hence continue to contribute 'diversity' even within what may appear to be a purely domestic collaboration.

A similar finding has been reported by Ingersoll et al. (2014), who calculate the effect of cultural diversity on performance of soccer teams. While the argument of access to embedded knowledge in geographically-separated networks is not directly applicable in that case, the soccer players to an extent embody some of the knowledge and skills that were acquired in the origin country.

Cultural differences can also affect reception of novel ideas. Wang (2015) shows that return migrant's experience and networks abroad can be perceived negatively in xenophobic countries, reducing chances of successful transfer of knowledge. In the context of the present chapter, work of collaborators from very diverse cultures might be perceived as lower quality due to bias against very different ideas or because of the coordination/communication costs at the collaboration team level.

Both of these effects, the positive effects of recombining knowledge from different networks and the negative effect of very different cultures, imply that there will be a concave relationship between collaboration quality and cultural distance. This pattern is indeed observed in the data, as shown in Figure 7.

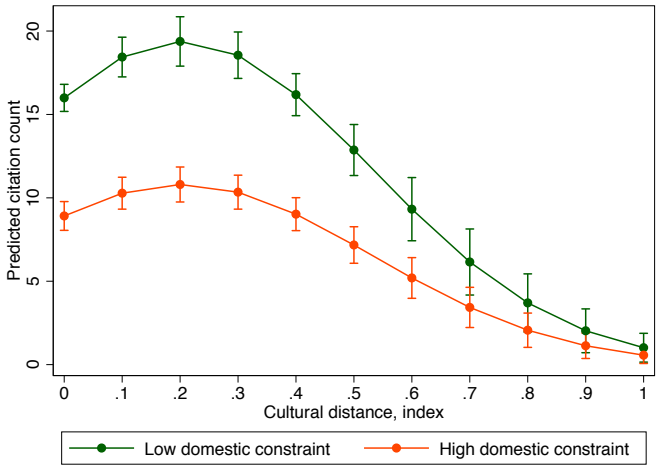
Table 9: The importance of culture (net citations).

	Net citation count	
	Journal publications	Working papers
Three authors	0.210	0.230
	(0.034)***	(0.044)***
Four authors	0.520	0.474
	(0.098)***	(0.085)***
Five+ authors	-0.114	0.761
	(0.144)	(0.258)***
Years since publication	0.303	0.269
	(0.008)***	(0.011)***
Years since publication, sq.	-0.006	-0.005
	(0.000)***	(0.001)***
Authors - age at the time of publication	-0.006	-0.004
	(0.002)***	(0.003)
Authors - prior publications	-0.004	-0.004
	(0.001)***	(0.001)***
Authors - prior citations	0.000	0.000
	(0.000)***	(0.000)***
Journal impact factor	0.070	0.241
	(0.002)***	(0.008)***
Cultural distance	0.905	0.334
	(0.290)***	(0.410)
Cultural distance, sq.	-2.592	-2.043
	(0.607)***	(0.829)**
Cultural distance x Global constraint	-0.002	0.017
	(0.004)	(0.007)***
Global constraint	-0.012	-0.016
	(0.001)***	(0.001)***
Adj. R2	0.37	0.27
N	53,371	24,155

* $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

Source: own calculations based on RePEc data. Notes: paper counts are adjusted for related works; citations are adjusted for self-citations.

Figure 7: The effect of cultural distance.



Source: own calculations based on RePEc data. Notes: paper counts are adjusted for related works; citations are adjusted for self-citations; bars represent 95% confidence intervals.

6 Discussion and conclusion

The results of this research provide an explanation for the diversity premia on international, interethnic and intergender collaboration — these collaborations provide access to the knowledge embedded in different research networks. Such collaboration could be done by migrant scientists (cf. Franzoni et al. 2014), but also through establishing links with an existing research-active diaspora or through specific programs that support international collaboration. It should be noted, though, that this theory suggests that the source (sending) country will benefit from a migrant scientist only if a research link between the migrant scientist and the domestic research network is maintained. If the migrant scientists return and maintain research links with other countries, however, the original sending country should derive benefits. For example, Gibson and McKenzie (2014) found that return migrant scientists acted as the main source of research knowledge transfer between the global and domestic research networks.

The importance of the position of the authors in the global research network can also help understand the unexpected similarity of academic productivity of "elite" stayers and movers (Hunter et al. 2009). Authors occupying a central position in the domestic network will be able to derive benefits from access to the structural holes within the domestic network and will not need to migrate or to collaborate internationally, which would act as a substitute for physical relocation. An empirical investigation of this argument can be found in Wang (2015), who uses

survey data on highly-skilled return migrants that left the United States to their home country to show that the extent of the return migrants' embeddedness in the host and origin country networks positively affects their transfer of knowledge. The survey data also allowed Wang (2015) to assess the importance of cultural differences for knowledge transfer - in xenophobic countries the foreign embeddedness of the returnee was acting as a liability rather than advantage, diminishing the chance of successful knowledge transfer.

More broadly, this paper adds to the literature on diversity and economic performance. Studies of ethnic diversity and economic performance show that the relationships are quite complex (e.g. Alesina and La Ferrara 2005) and depend on the underlying production function. The traditional approaches focused on complementarity/substitutability of skills, cultural differences and costs of collaboration, while the results of the present paper support including measures of social network access.

Also, even though the direction of knowledge flows is not examined in the present chapter, the results do imply that the direction of knowledge flows is not exclusively from 'core' to 'periphery' — international collaboration can act as a way of improving centrality within the *global* research network, including both the 'core' and the 'periphery'.

This paper provides an explanation for the mechanism underlying positive effect of geographic diversity on the quality of collaboration. The diversity premia observed in the literature disappear after controlling for the social network properties of the authors' global collaboration networks, specifically in terms of non-redundant ties. Collaboration between geographically and ethnically diverse economists is a way of accessing non-redundant information and resources (covering "structural holes") embedded in their co-authors' domestic research networks.

This provides support for policy measures that encourage international collaboration, because such collaboration is likely to increase access to new, non-redundant information, thereby raising the quality of the product compared to domestic-only collaboration. At the same time, the benefits of international collaboration are the largest for authors with high domestic network constraint, for example junior researchers or those researchers without history of academic mobility to other countries. As the authors become less constrained domestically, the benefits of international collaboration begin to diminish. Hence a policy that can effectively target domestically-constrained authors will be most effective.

Further extensions of this research are possible. One extension is to obtain a better measure of authors' location at the time of collaboration. This will allow identifying migration episodes and comparing migrant scientists with non-migrants or returnees. The present paper also didn't explore the diffusion of information, so it is interesting to see to what extent the geographic and ethnic diversity affects the

diffusion of knowledge, both in terms of physical distance and speed of diffusion. The dataset on economists collected by RePEc and related services contains high quality data and can be used to answer many interesting research questions in areas of networks, sociology of economics, team science and more.

References

Alesina, Alberto and Eliana La Ferrara (2005). "Ethnic Diversity and Economic Performance". In: *Journal of Economic Literature*, pp. 762–800.

Baruffaldi, Stefano H. and Paolo Landoni (2012). "Return mobility and scientific productivity of researchers working abroad: The role of home country linkages". In: *Research Policy* 41.9, pp. 1655–1665.

Brogaard, Jonathan, Joseph Engelberg, and Christopher A. Parsons (2014). "Networks and productivity: Causal evidence from editor rotations". In: *Journal of Financial Economics* 111.1, pp. 251–270.

Burt, Ronald S. (2004). "Structural holes and good ideas". In: *American journal of sociology* 110.2, pp. 349–399.

Burt, Ronald S. (2015). "Reinforced structural holes". In: *Social Networks* 43, pp. 149–161.

Card, David and Stefano DellaVigna (2013). "Nine Facts about Top Journals in Economics". In: *Journal of Economic Literature* 51.1, pp. 144–61.

Coupé, Tom (2013). "Peer review versus citations – An analysis of best paper prizes". In: *Research Policy* 42.1, pp. 295–301.

Coupé, Tom, Victor Ginsburgh, and Abdul Noury (2010). "Are leading papers of better quality? Evidence from a natural experiment". In: *Oxford Economic Papers* 62.1, pp. 1–11.

Desmet, Klaus, Ignacio Ortuño-Ortn, and Romain Wacziarg (2015). *Culture, Ethnicity and Diversity*. Tech. rep. National Bureau of Economic Research.

Feenberg, Daniel R, Ina Ganguli, Patrick Gaule, and Jonathan Gruber (2017). "It's Good to be First: Order Bias in Reading and Citing NBER Working Papers". In: *Review of Economics and Statistics* 99.1, pp. 32–39.

Fourcade, Marion (2009). *Economists and Societies: Discipline and Profession in the United States, Britain, and France, 1890s to 1990s*. Princeton University Press.

Franceschet, Massimo and Antonio Costantini (2010). "The effect of scholar collaboration on impact and quality of academic papers". In: *Journal of informetrics* 4.4, pp. 540–553.

Franzoni, Chiara, Giuseppe Scellato, and Paula Stephan (2014). "The mover's advantage: The superior performance of migrant scientists". In: *Economics Letters* 122.1, pp. 89–93.

Freeman, Richard B. and Wei Huang (2015). "Collaborating with People Like Me: Ethnic Coauthorship within the United States". English. In: *Journal of Labor Economics* 33.S1, pp. 289–318.

Furman, Jeff and Patrick Gaule (2013). "A review of economic perspectives on collaboration in science". Paper prepared for the Workshop on Institutional & Organizational Supports for Team Science.

Gazni, Ali, Cassidy R Sugimoto, and Fereshteh Didegah (2012). "Mapping world scientific collaboration: Authors, institutions, and countries". In: *Journal of the American Society for Information Science and Technology* 63.2, pp. 323–335.

Gibson, John and David McKenzie (2014). "Scientific mobility and knowledge networks in high emigration countries: Evidence from the Pacific". In: *Research Policy* 43.9, pp. 1486–1495.

Goyal, Sanjeev, Marco J. van der Leij, and José Luis Moraga-González (2006). "Economics: An emerging small world". In: *Journal of Political Economy* 114.2, pp. 403–412.

Granovetter, Mark S (1973). "The strength of weak ties". In: *American journal of sociology*, pp. 1360–1380.

Hamermesh, Daniel S. (2013). "Six Decades of Top Economics Publishing: Who and How?" In: *Journal of Economic Literature* 51.1, pp. 162–72.

Hamermesh, Daniel S. (2015). *Age, Cohort and Co-Authorship*. Tech. rep. National Bureau of Economic Research.

Hamermesh, Daniel S. (2017). "Citations in Economics: Measurement, Uses and Impacts". In: *Journal of Economic Literature*.

Head, Keith, Yao Amber Li, and Asier Minondo (2015). *Geography, Ties, and Knowledge Flows: Evidence from Citations in Mathematics*. Tech. rep. HKUST Institute for Emerging Market Studies.

Helgadóttir, Oddný (2015). "The Bocconi boys go to Brussels: Italian economic ideas, professional networks and European austerity". In: *Journal of European Public Policy*, pp. 1–18.

Hunter, Rosalind S, Andrew J Oswald, and Bruce G Charlton (2009). "The Elite Brain Drain". In: *The Economic Journal* 119.538, F231–F251.

Ingersoll, Keith, Edmund J Malesky, and Sebastian M Saiegh (2014). "Diversity and Group Performance: Evidence from the World's Top Soccer League". In: *APSA 2014 Annual Meeting Paper*.

Jin, Ginger Zhe, Benjamin Jones, Susan Feng Lu, and Brian Uzzi (2013). *The reverse Matthew effect: catastrophe and consequence in scientific teams*. Tech. rep. National Bureau of Economic Research.

Jones, Benjamin F (2009). "The burden of knowledge and the "death of the renaissance man": Is innovation getting harder?" In: *The Review of Economic Studies* 76.1, pp. 283–317.

Jonkers, Koen and Laura Cruz-Castro (2013). "Research upon return: The effect of international mobility on scientific ties, production and impact". In: *Research Policy* 42.8, pp. 1366–1377.

Kerr, William R. (2008). "Ethnic Scientific Communities and International Technology Diffusion". In: *Review of Economics and Statistics* 90.3, pp. 518–537.

Kosnik, Lea-Rachel D (2015). "JEL Codes: What Are They Really Telling Us?" Mimeo.

Krichel, Thomas and Christian Zimmermann (2009). "The Economics of Open Bibliographic Data Provision". In: *Economic Analysis and Policy* 39.1, pp. 143–152.

Mateos, Pablo, Richard Webber, and Paul Longley (2007). "The cultural, ethnic and linguistic classification of populations and neighbourhoods using personal names". Paper 116.

Mayer, Thierry and Soledad Zignago (2011). *Notes on CEPII's distances measures: The GeoDist database*. Tech. rep. Working paper 2011-25.

Miura, Hirotaka (2012). "Stata graph library for network analysis". In: *Stata Journal* 12.1, pp. 94–129.

Mohnen, Myra (2016). "Stars and brokers: peer effects among medical scientists". Job market paper.

Mongeon, Philippe and Vincent Larivière (2015). "Costly collaborations: The impact of scientific fraud on co-authors' careers". In: *Journal of the Association for Information Science and Technology*.

Montecinos, Verónica and John Markoff (2010). *Economists in the Americas*. Edward Elgar Publishing.

Nathan, Max (2015). "Same difference? Minority ethnic inventors, diversity and innovation in the UK". In: *Journal of Economic Geography* 15.1, pp. 129–168.

Nisbett, Richard E (2003). *The geography of thought*. New York: Free Press.

Orazbayev, Sultan (2016). "Exploring the world of Economics through RePEc data". mimeo.

Oster, Sharon M. and Daniel S. Hamermesh (1998). "Aging and productivity among economists". In: *Review of Economics and Statistics* 80.1, pp. 154–156.

Rigby, John and Jakob Edler (2005). "Peering inside research networks: Some observations on the effect of the intensity of collaboration on the variability of research quality". In: *Research Policy* 34.6, pp. 784–794.

Saxenian, AnnaLee (2005). "From brain drain to brain circulation: Transnational communities and regional upgrading in India and China". In: *Studies in comparative international development* 40.2, pp. 35–61.

Seglen, Per O (1997). "Why the impact factor of journals should not be used for evaluating research." In: *BMJ: British Medical Journal* 314.7079, p. 498.

Singh, Jasjit (2005). "Collaborative networks as determinants of knowledge diffusion patterns". In: *Management science* 51.5, pp. 756–770.

Spolaore, Enrico and Romain Wacziarg (2015). *Ancestry, Language and Culture*. Tech. rep. National Bureau of Economic Research.

Tange, Ole (2011). "GNU Parallel-the command-line power tool". In: *The USENIX Magazine* 36.1, pp. 42–47.

Thelwall, Mike and Nabeil Maflahi (2015). "Are scholarly articles disproportionately read in their own country? An analysis of Mendeley readers". In: *Journal of the Association for Information Science and Technology* 66.6, pp. 1124–1135.

Uzzi, Brian and Jarrett Spiro (2005). "Collaboration and creativity: the small world problem". In: *American journal of sociology* 111.2, pp. 447–504.

Wang, Dan (2015). "Activating Cross-border Brokerage: Interorganizational Knowledge Transfer through Skilled Return Migration." In: *Administrative Science Quarterly* 60.1, pp. 133–176.

Wuchty, Stefan, Benjamin F Jones, and Brian Uzzi (2007). "The increasing dominance of teams in production of knowledge". In: *Science* 316.5827, pp. 1036–1039.

Zimmermann, Christian (2013). "Academic Rankings with RePEc". In: *Econometrics* 1.3, pp. 249–280.

A Data preparation

This section describes how the data was prepared for the analysis. Part of the data was processed with the aid of a *parallel* tool, see Tange (2011). Some of the network calculations were performed using *netsis*, see Miura (2012).

A.1 RePEc data

The data used in this chapter is collected by Research Papers in Economics (RePEc) and a set of related services[25] that are of interest to the community of economists. These services include hosting space for working papers, software components, citation statistics, directory of economists, genealogy of academic economists, ranking of individual authors and departments, and many other services/databases. Zimmermann (2013) provides a good overview of RePEc data and the methodology used to rank journals, departments and individual authors. The following rest if this section contains a brief description of the data used in this chapter, for a more thorough exploration of RePEc data see Orazbayev (2016).

Some of the data is contributed by the registered economists, e.g. affiliations and publications, while the rest of the data — citation counts, citating authors, registered colleagues at the same institution, h-index and such — is computed by RePEc based on the data on publications and information contributed by other authors.

A.1.1 Author affiliation

Authors that are registered on RePEc have the opportunity to list the institutions with which they are affiliated, Zimmermann (2013). Any author can have multiple affiliations, in which case authors can attach to each affiliation a weight that represents roughly the share of time that the author spends at the institution. Many authors have not specified weights, but listed multiple affiliations. This presents a challenge because if those affiliations are in different countries, then it's not clear in which country the author spends most of their time.

To resolve this issue the following procedure is used:

- if all affiliations of an author are located within one country, then that country is used as the main location of the author;
- if affiliations are in different countries and the author provided weights, then the country with the largest weight (summed by all institutions per country) is used as the main location of the author;
- if affiliations are in different countries and the author did not specify weights, but one of the affiilations reflects belonging to an international research network (e.g. NBER, CEP), which does not require physical presence then such affiliations are dropped;

[25] For convenience RePEc will be used to refer to all of these services, even though each service has individual maintainers.

- if remaining affiliations are in different countries and the author did not specify weights, then the affiliation that the author provided first is used as the main location of the author.

A.1.2 Related papers

A journal publication or a current working paper might have been circulated previously as a working paper. In this case, there will be multiple entries of the same paper in the dataset, which will bias the sample towards papers that are frequently updated and/or disseminated via multiple channels. RePEc uses a procedure based on article title and author names to determine if two papers are 'related' (i.e. different versions of the same paper). RePEc does use this procedure to aggregate citations on different versions of the same work and also to determine the latest version of the paper.

For the purposes of calculations in this chapter, a custom procedure was used to determine the latest version of the paper. Given a set of related papers, each paper was assigned a score based on the following scheme: 1 point if the paper has the highest year, 2 points if it has the highest citation count (this includes citations on earlier versions), 3 points if it was published. From each set of related papers, the paper with the highest score is retained in the sample, while the others are removed from the sample.

A.1.3 Paper-level variables and collapsing of the data

A number of variables is available at the paper-level, for example the publication year, title of the journal or working paper series, number of authors, JEL codes. Some of the variables however are available at the author level, for example main country of affiliation, publications and citations prior to this paper, research age at the time of this publication. To get a paper-level value for these variables either the mean or the maximum of the values for individual coauthors is used. To calculate this, for each pair all possible pairwise combinations of coauthors are formed and author-level values are assigned, including physical and cultural distance measures. Then the dataset is collapsed to the paper-level by using either the min/max or the average approach (see Section 4.1).

A.1.4 Self citations

The raw data file provided by CollEc contains information on all papers that cite a particular paper. From this information it is possible to calculate gross citations (i.e. a simple count of all papers citing a particular paper) and net citations, where self-citations by any of the initial authors are removed. Calculated net citations are highly correlated with the gross citations (see Table below), however

the correlation is weaker for papers with fewer than 10 citations. The outlier in terms of absolute self-citations is a paper that was self-cited 251 times (this is partly due to the large number of coauthors, three, but seems to be mainly due to the specific dataset used in the paper), the outlier in terms of relative self-citations is a paper that was cited exclusively by the authors 51 times. Similar analysis can be performed to exclude citations from the same country. See Figures 10 and 9 for a diagram showing the effects of adjustment for self citations and citations from different countries.

A.1.5 Journal simple impact factors

Using the data from CitEc, it is possible to calculate the simple impact factor for each journal and working paper series. The simple impact factor of a journal/working paper series is measured as the ratio of citations received by publications within a specific time period (e.g. over 10 previous years) to the number of publications during that period.

RePEc provides journal/working paper series rankings based on the latest data, but does not provide historical values of the simple impact factor. The simple impact factors for 3-, 5- and 10-year periods were reconstructed for journals and working paper series by using citation information from CitEc. For each cited paper the citations from the same journal/working paper series are removed, then only citations from the relevant time period are counted. For example, when calculating 10-year simple impact factor for a journal in year 2010, the number of publication is the number of papers published starting from 2000 up to and including 2009, while the number of citations is the number of citations to these papers by other papers which were also published during the same time-period. This approach ensures that the simple impact factor depends only on information prior to the measurement year. As in the RePEc rankings, the simple impact factor is calculated only if there were at least 50 papers published within the time period.

A.1.6 JEL codes

A number of papers provided JEL codes in the abstract and/or keywords section of the data, and so they had to be extracted from these fields using a regular expression match procedure. The extracted codes were verified for consistency. Also, some authors provided JEL codes at limited resolution, e.g. only 2 symbols. In specifications that use the full 3-symbol JEL codes the truncated codes were padded with zeroes to keep them in the dataset. For example, if a paper contained only symbol A1, then for 3-symbol specifications this would be changed to A10.

The specification that includes a set of dummies representing JEL codes of the

paper is included in the Appendix. Moreover, there could be misallocation of JEL codes in the data, Kosnik (2015) shows that there can be significant variation in allocation of JEL codes - in the sample of papers published in American Economic Review the author- and editor-assigned JEL codes differ 42% of the time.

A.2 Genetic, cultural and linguistic distance

A growing body of literature on economic development and growth links studies the importance of cultural differences across countries (see Spolaore and Wacziarg (2015) for a good overview of relevant papers). The main measures used in the literature are based on genetics, language, religion, answers to questions on social norms, values and attitudes. One measure of genetic distance, F_{ST}, is calculated based on the frequency with which certain genes appear in different population, see Spolaore and Wacziarg (2015) for details and references to other measures. Larger genetic distance is correlated with more distant culture (e.g. as reflected in answers to poll questions on social norms and values).

The dataset[26] on genetic, linguistic and cultural distances doesn't contain information on distance for self-pairs (e.g. from Kazakhstan to Kazakhstan), however internal collaborations are an important part of the sample and as an extra assumption the self-distance (or internal distance) was set to zero for linguistic and genetic (F_{ST}) distances. A similar assumption has been made for cultural distance, even though research suggests that culture within a country is far from homogenous, see Desmet et al. (2015). In the absence of internal, region-level data it is not possible to calculate a measure of internal cultural distance, so the assumption is strong, but can be revised later once additional data is available. Since, the general conclusion of Spolaore and Wacziarg (2015) is that the genetic distance can act as a summary measure for a wide range of cultural traits, this measure will be used as the main proxy for country-specific barriers to interaction and communication between authors.

A.3 Physical distance

The data on physical distance between countries comes from CEPII, see Mayer and Zignago (2011).

A.4 Name-ethnicity matching

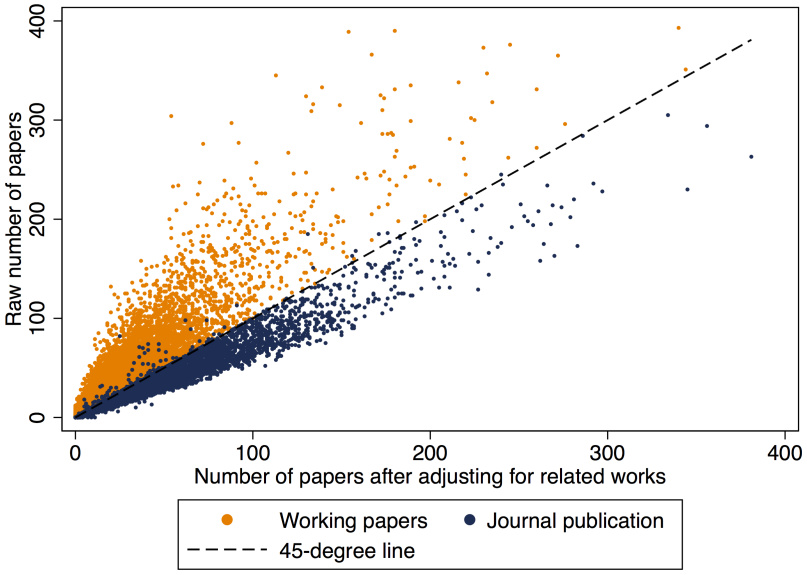
Name-ethnicity matching is performed by OnoMAP software. Further details can be found at http://www.onomap.org/FAQ.aspx and Mateos et al. (2007). Nathan

[26]The dataset is available at: http://sites.tufts.edu/enricospolaore/

(2015, page 162) provides a good description of advantages and disadvantages of OnoMAP, with further references to papers that test OnoMAP's matching precision.

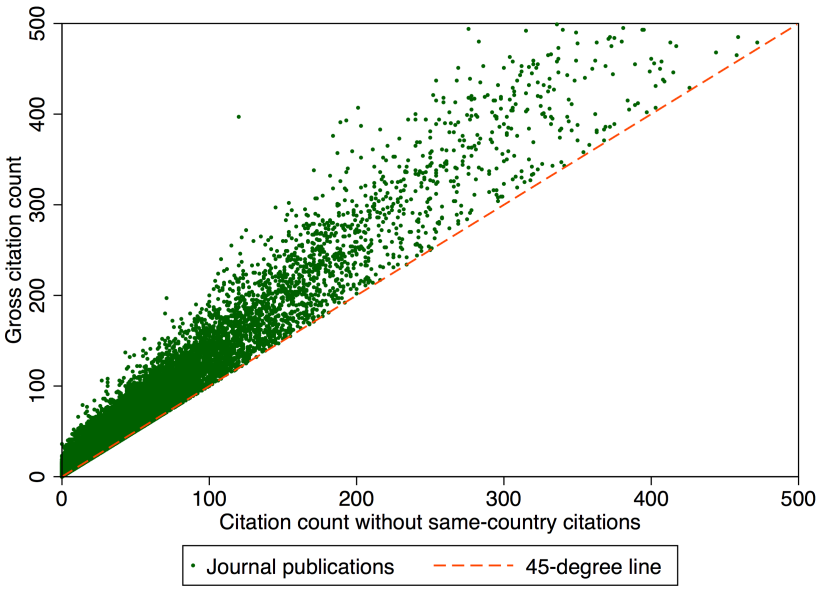
B Supporting diagrams and tables

Figure 8: Effect of correction for related works.



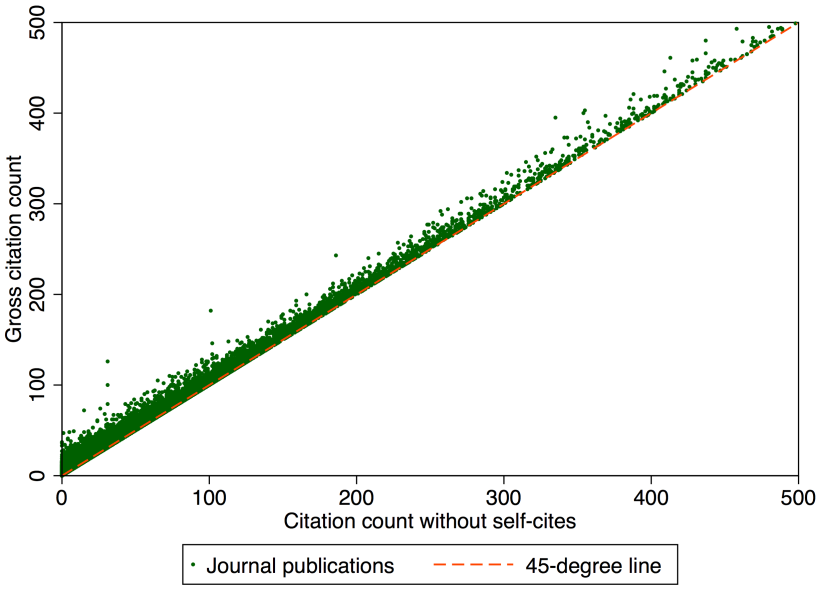
Source: own calculations based on RePEc data. Notes: paper counts are adjusted for related works; total number of publications exceeds unique papers due to co-authorships.

Figure 9: The effect of adjusting same-country citations.



Source: own calculations based on RePEc data. Notes: paper counts are adjusted for related works.

Figure 10: The effect of adjusting self-citations.



Source: own calculations based on RePEc data. Notes: paper counts are adjusted for related works.

Table B.1: Importance of centrality for the dynamic network (net citations) - inclusion of JEL dummies.

	Net citation count			
Three authors	0.268	0.269	0.257	0.255
	(0.081)***	(0.081)***	(0.079)***	(0.080)***
Four authors	0.373	0.378	0.350	0.343
	(0.141)***	(0.141)***	(0.140)**	(0.141)**
Five+ authors	-0.083	-0.070	-0.100	-0.109
	(0.198)	(0.197)	(0.195)	(0.196)
Years since publication	0.234	0.234	0.227	0.227
	(0.018)***	(0.018)***	(0.017)***	(0.017)***
Years since publication, sq.	-0.002	-0.002	-0.001	-0.001
	(0.001)*	(0.001)	(0.001)	(0.001)
Authors - age at the time of publication	0.002	0.002	0.002	0.002
	(0.004)	(0.004)	(0.004)	(0.004)
Authors - prior publications	-0.003	-0.003	-0.004	-0.004
	(0.002)*	(0.002)*	(0.002)**	(0.002)**
Authors - prior citations	0.000	0.000	0.000	0.000
	(0.000)***	(0.000)***	(0.000)***	(0.000)***
Journal impact factor	0.065	0.065	0.064	0.064
	(0.003)***	(0.003)***	(0.003)***	(0.003)***
Domestic constraint	-0.004	-0.006		
	(0.001)***	(0.002)**		
International collaboration	0.126	-0.032	0.042	0.129
	(0.068)*	(0.125)	(0.066)	(0.106)
Domestic constraint x International collaboration		0.004		
		(0.003)		
Global constraint			-0.011	-0.010
			(0.002)***	(0.002)***
Global constraint x International collaboration				-0.003
				(0.003)
Pseudo R2	0.42	0.42	0.42	0.42
N	14,712	14,712	15,093	15,093
JEL dummies	Yes	Yes	Yes	Yes

* $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

Source: own calculations based on RePEc data. Notes: paper counts are adjusted for related works; citations are adjusted for self-citations.

C Calculating structural holes: an example

Let's calculate structural holes for the nodes in the top panel of Figure 4. Consider first the domestic network in country A. Naming the black node as X and the nodes above and below it as Y and Z, respectively, we can calculate the network constraint for X,Y and Z as:

$$C_X = \sum_{j=Y,Z} c_{Xj} = (\frac{1}{2} + 0)^2 + (\frac{1}{2} + 0)^2 = \frac{1}{2},$$

$$C_Y = \sum_{j=X} c_{Yj} = (1)^2 = 1,$$

$$C_Z = \sum_{j=X} c_{Yj} = (1)^2 = 1.$$

Now examining the 'global network' for countries A and B, let the top and bottom nodes in country B network be known as V and W, respectively. For simplicity, let's focus on the network constraint for node X:

$$C_X = \sum_{j=W,Y,Z} c_{Xj}.$$

Calculation of the constraint between X and W must take into account that they have two common collaborators Y and Z:

$$c_{XW} = (p_{XW} + \sum_{q \neq X,W} p_{Xq} p_{qW})^2 = (\frac{1}{3} + \frac{1}{3} \times \frac{1}{3} + \frac{1}{3} \times \frac{1}{2})^2 \approx .37,$$

where the third expression is based on $q = Y, Z$.

Calculating the remaining bilateral constraints for X gives:

$$c_{XY} = (p_{XY} + \sum_{q \neq X,Y} p_{Xq} p_{qY})^2 = (\frac{1}{3} + \frac{1}{3} \times \frac{1}{3})^2 \approx .20,$$

$$c_{XZ} = (p_{XZ} + \sum_{q \neq X,Z} p_{Xq} p_{qZ})^2 = (\frac{1}{3} + \frac{1}{3} \times \frac{1}{3})^2 \approx .20.$$

Combining the bilateral constraints gives the overall constraint for X as:

$$C_X = \sum_{j=W,Y,Z} c_{Xj} \approx .77.$$