Kamel, Donia; Pollacci, Laura

Working Paper

Academic Migration and Academic Networks: Evidence from Scholarly Big Data and the Iron Curtain

CESifo Working Paper, No. 10377

Provided in Cooperation with:
Ifo Institute – Leibniz Institute for Economic Research at the University of Munich

CESifo WORKING PAPERS

10377
2023

April 2023

Academic Migration and Academic Networks: Evidence from Scholarly Big Data and the Iron Curtain

Donia Kamel, Paura Pollacci

CESifo

Impressum:

CESifo Working Papers
ISSN 2364-1428 (electronic version)
Publisher and distributor: Munich Society for the Promotion of Economic Research - CESifo GmbH
The international platform of Ludwigs-Maximilians University's Center for Economic Studies and the ifo Institute
Poschingerstr. 5, 81679 Munich, Germany
Telephone +49 (0)89 2180-2740, Telefax +49 (0)89 2180-17845, email office@cesifo.de
Editor: Clemens Fuest
https://www.cesifo.org/en/wp
An electronic version of the paper may be downloaded
· from the SSRN website: www.SSRN.com
· from the RePEc website: www.RePEc.org
· from the CESifo website: https://www.cesifo.org/en/wp

Academic Migration and Academic Networks: Evidence from Scholarly Big Data and the Iron Curtain

Abstract

Iron Curtain and Big Data are two words usually used to denote completely two different eras. Yet, the context the former offers and the rich data source the latter provides, enable the causal identification of the effect of networks on migration. Academics in countries behind the Iron Curtain were strongly isolated from the rest of the world. This context poses the question of the importance of academic networks for migration post the fall of the Berlin Wall and Iron Curtain. Using Microsoft Academic Knowledge Graph, a scholarly big data source, mapping of academics' networks is possible and information about the size and quality of their co-authorships, by location is achieved. Focusing on academics from Eastern Europe (henceforth EE) from 1980-1988 and their academic networks (1980-1988), We investigate the effect of academic network characteristics, by location, on the probability to migrate post the fall of the Berlin Wall in 1989 and up to 2003, marking the year many EE countries held referendums or signed treaties to join the EU. The unique context ensures that there was no anticipation of the fall of the Eastern Bloc and together with the data that offers unique rich information, identification is achieved. Approximately 30k academics from EE were identified, from which 3% were migrants. The results could be explained by two channels, the cost and signalling channel. The cost channel is how the network characteristic reduces or increases the cost of migration and thus acting as a facilitator or a de-facilitator of migration. The signal channel on the other hand in which the network characteristic serves as a signal for the academic himself and his quality and his potential contribution and addition to the new host institution, thus also serving as a facilitator or a de-facilitator of migration. We find that mostly network size and quality results could be explained by the cost channel and signalling channel, respectively. Size of the network tends to be more important than the quality, which is a context-specific result. We find heterogeneous effects by fields of study that align with previous lines of research. Heterogeneous effects are explained by two things: threat of attention and arrest from KGB and the role of reputation, language, and network barriers.

JEL-Codes: C550, D850, F500, I200, I230, J240, N340, N440, O150.

Keywords: networks, migration, academic networks, Big Data, brain drain, Iron Curtain, Eastern Europe.

Donia Kamel
Department of Economics
Paris School of Economics / France
donia.kamel@psemail.eu

Laura Pollacci
Department of Computer Science
University of Pisa / Italy
laura.pollacci@di.unipi.it

May 2022
This is a work in progress. Future edits and additions are outlined in the last section. We would like to thank Professor Hillel Rapoport for his continuous support and insightful advice, which I'm sure will continue until this paper reaches the best version it can reach. We would also like to thank Professors Philipp Ketz and Ekaterina Zhuravskaya for their guidance. This work is supported by the European Union − Horizon 2020 Program under the scheme “INFRAIA-01-2018-2019 – Integrating Activities for Advanced Communities”, Grant Agreement n.871042, “SoBigData++: European Integrated Infrastructure for Social Mining and Big Data Analytics” (http://www.sobigdata.eu) and by the Horizon2020 European project “HumMingBird − Enhanced migration measures from a multidimensional perspective” (Grant Agreement n. 870661).

1 Introduction

The decision to migrate is one of the most important decisions an individual can make. As such, this decision is influenced and shaped by a lot of factors such as inequality levels at home and intended destination, returns to education, migration costs, employment prospects and many more. Networks play a vital role in such decision as they influence all of the above factors in a way or another. However, identifying the role of networks in migration is an empirically hard task to do.

Data on individual networks are very limited. In fact, Breza et al (2020) notes that "social network data are often prohibitively expensive to collect, limiting empirical network research", motivating the quest for alternative methods and data sources. Additionally, due to the data constraints, most of the existing empirical evidence in this strand of the literature, make a restricting implicit assumption. This literature assumes that all potential migrants benefit from the networks at destination equally (Bertoli and Ruyssen, 2018). This empirical evidence ranges from looking at share of households with a migrant at the village (McKenzie and Rapoport, 2010), size of diaspora in each destination country (Bertoli and Moraga, 2015 and Beine et al., 2011, 2015) or at the country level (Bertoli, 2010); whilst all making the implicit assumption that migrants benefit *equally* from networks. On the contrary, Blumenstock et al. (2019) abstracts from this assumption by utilizing big data in the form of call details records from mobile phone use in Rwanda which offers detailed information on the structure of individual networks. Alongside this general limitation in the literature, the data on migration is also limited, offering no information on the intention to migrate, which could be done through a close examination of one's individual networks. Combining the two in a single empirical setting not only requires good data but also good identification strategy. To overcome the data constraints and endogeneity concerns, especially reverse causality, we look at a specific historical context using a scholarly big data source. This research project examines the effect of two network characteristics, size, and quality, of *academics* in countries behind the Iron Curtain (1980-1988), by location, on their probability to migrate (1989-2003). We will come back to the contributions and importance of this research towards the end of this section, for now some theoretical motivation and parallels will be drawn.

Network theory suggests that the topological structure of the network affects the utility an individual draws from his social network, yet, there is considerable ambiguity about which *type* of social capital matters the most (Blumenstock et al., 2019). Different structures and characteristics of networks derived from network theory align with some forms of social capital. Jackson (2018) breaks down the typology of social capital into 7 fundamental forms of capital. They are all fundamental to the welfare of society and distinguishing between the different forms and understanding their distributions is important as they are a driver to inequality and immobility (Jackson, 2018).

Social network theory's notion of centrality offers interesting parallels to this research. Centrality is a micro measure, in a way, as it is at the individual level, thus allowing us to compare nodes (academics) and say something about how a given node relates to the overall network. There are 4 main groups: degree centrality (size), closeness, betweenness and neighbour's characteristics (Bonacich, 1991). These different notions capture complementary aspects of a node's position, and each measure could be seen as more appropriate for certain settings and less for others. With the nature of the data, this analysis focuses on degree centrality and neighbour's characteristics as we look into the *size* and *quality* of academics' networks (direct co-authorships). In fact, network size aligns with *degree centrality* in social network theory and *community capital* in social capital theory, and network quality which aligns with *power-, prestige-, eigenvector- centrality* in social network theory and *reputation capital* in social capital theory. The motive behind these labels and how they are synonymous to the characteristics of the networks under study in this research is presented in the following paragraphs. Additionally, they allow an easier interpretation of the results and the potential proposed channels of effects.

Degree centrality, which relates to the size of the network, could be also linked to Jackson's (2018) *community capital* highlighting the ability to be part of or sustain cooperative behaviour, which aligns with the notion of co-authoring and collaborating with other academics. In the migration literature, the prevailing view is that migrants tend to go to places where they have larger networks. This is due to the fact that such connections reduce the cost of migration (Carrington et al., 1996) and can help with improving job prospects at destination (Patel and Vella, 2013). However, degree centrality misses interesting features of a network, it does not measure how well located a node is in a network (Bonacich, 1991). For example, one node or academic might have a small network in terms of its size such that it has few links linking it to other academics or nodes, yet it might have an important and a critical place in the network or simply that these few links are of high importance and significance. Thus, given the characteristics of the data, the focus will be also on *reputation capital*, or the notion of "prestige-, power-, eigenvector- related centrality measures.

Generally, this form of (reputation) capital is a valuable input into production as such academics who are of great reputation, if for example requested for joint research and consequently becoming a co-author for a paper, attending a conference or peer reviewing a paper, adds more confidence into the production process simply due to his/her "reputation", a synonym to *quality* in this research. Additionally, this form of capital is a stock that can be acquired over time and used, which is synonymous to being or becoming a top academic and having or gaining top academics in your network. In other words, being a top academic or being connected to a top academic (through co-authorship) can be used over time to facilitate migration decisions to another country (and another academic institution). Therefore, the quality of the network is synonymous to reputation capital, one of the fundamental forms of social capital mentioned by Jackson (2018). These are more intricate measures based on the idea that a node's importance is determined by how important its neighbours are, i.e. proximity to important ones. In fact, this is central to such phenomena such as: citation ranking, google page rankings...etc., which are some proxies for the quality of an academic and his network, as will be detailed later on.

This importance of this research is multidimensional. Firstly, it overcomes data limitations inherent in the empirical literature studying networks and migration. Through Microsoft Academic Knowledge Graph, we are able to identify the academic network of each academic through co-authorships by location. Thus, we have a measure of *individual* networks and we abstract from the implicit assumption done by the literature. In fact, this assumption, that migrants benefit equally from networks at destination, contradicts the theoretical representation of social networks (Jackson, 2010) and the empirical evidence on the interactions of members of a migrant network (Comola and Mendola, 2015). MAKG offers rich information about the publications, fields, and affiliations of each academic as we will see in the data section. Through changes of affiliations spotted through publication we can identify if an academic is a migrant or not. Thus, this data source overcomes a lot of data constraints in the existing literature enable the identification of individual academic networks and migrants.

Computer scientists have studied migration of academics using MAKG before (Ricciarelli et al., 2020). However, the evidence from such papers is purely descriptive. The evidence is promising in the sense that it gives power to the use of big data in answering migration questions. However, it abstracts from causality completely. A major endogeneity concern of the effect of networks on migration is reverse causality, which is eliminated given the context of this research. As mentioned previously, we study academics from EE and their networks (1980-1988) on their post fall of Berlin Wall migration (1989-2003). The fall of the Berlin Wall and Iron Curtain was not anticipated. In fact, the Soviet Union wasn't ripe for any sudden change (Borjas and Doran, 2012). Many historians in fact note that this was a sudden effect, with many scholars overestimating the political and economic power of the Eastern Bloc (Polyak 2012, Howe 1990; Laqueur 1996). The Iron Curtain didn't just act as a political boundary but also a geographical one severely limiting migration. As such, we find that indeed there was no out of EE migration from 1980-1988. Additionally, the absence of this anticipation overcomes the reverse causality concern. As such, academics, didn't manipulate

their networks strategically in anticipation of migration and are assumed to be exogenous. This also poses the question why academic networks are specifically important to look at. The academic market of the Eastern Bloc and the Soviet Union was a bit isolated given the context. As such, the role of networks, at home, at destination and other countries, must have had a huge role in facilitating migration post the fall. It is important to understand what how different networks affected migration of academics.
Thus, in conclusion, this research is important due to the various and vast contributions it offers to different strands of the literature. It first contributes to the current wave of research on human migration through the lens and perspective of big data [1]. However, it expands on this literature by focusing on a unique historical context that offers a much closer step to achieve the causal impact of networks on migration, thus it also expands on the literature focusing on migration post the fall of the Berlin Wall, the Iron Curtain and the dissolution of the Soviet Union. By focusing on academics, it contributes to the limited literature on academic migration and what shapes and affects their migration decisions (Teichler, 2015). It contributes to the vast and extensive literature on brain drain as the focus of this research is on academic migration. This research also provides a new empirical perspective on the determinants of academic migration paying particular attention to academic networks. As such, it also contributes to the strands of literature on the empirical relationship between networks and migration, which is an empirically hard task to do as mentioned before. The alignment between network theory and social capital theory also makes this research contribute to the empirical literature on social network theory. The remaining parts of this thesis is organized as follows: section 2 presents the theoretical predictions and preview of results, section 3 presents the data, section 4 presents the identification strategy, section 5 presents the empirical strategy, section 6 presents the results, section 7 concludes, and section 8 summarizes improvements and further work to be done as this is a work in progress.

2 Summary of Theoretical Predictions and Results

In this section I present briefly the theoretical predictions expected from this analysis in parallel with the results. In this analysis, I test the assumption that the effects of network size and quality, by location, on migration decisions of academics, go through two distinct channels: cost and signaling channels. The cost channel is how the network characteristic reduces or increases the cost of migration and thus acting as a facilitator or a de-facilitator of migration. The signal channel on the other hand in which the network characteristic serves as a signal for the academic himself, his quality and his potential contribution and addition to the new host institution, thus also serving as a facilitator or a de-facilitator of migration.

Network size, synonymous to degree centrality and community capital, is defined as the sum of connections, through co-authorships, by location, of the academic from Eastern Europe, from 1980 to 1988. The theory predicts that an increase in home network size, would decrease the probability to migrate (1989-2003), the dependent variable as defined earlier. The channel at play would be the cost channel such that a greater number of connections or co-authorships at home makes migration more costly as leaving networks behind and forming new ones is costly. This could be especially magnified for outside migration than within EE migration as I distinguish between types of migration using MNL. This is because migrating out of EE in a way means complete loss of co-authorships in the region, much more compared to within-EE migration. In fact, this aligns with the results such that an increase in home network size (80-88) by one unit reduces the probability to migrate (1989-2003) by 0.1 - 0.05pp (least specified to the most specified specification). Additionally, an increase in home network size increases the chances of an academic not migrating compared to his chances of migrating to another EE. In fact, an increase by 1 unit makes the academic 1.034-1.01 times more likely to not migrate as compared to migrating to another EE. Similarly and in line with the theoretical prediction, an increase in home network size lowers the chances of migrating out of EE compared to migrating within EE, in fact, the risk or odds are 3.7% lower in the migration to EE group as compared

[1] See Sîrbu et al., 2020

to migrating outside. The alignment of the results for home network size with the predictions is explained by the cost channel and how an increase in home network size decreases the probability to migrate as it acts more as a pull factor and implies loss of connections if an academic migrates.

Similarly for destination network size, the theoretical prediction is that an increase reduces the cost of migrating as connections have already been established and as such could ease integration into the new academic institution. Distinguishing by type of migration, this would increase the chances of migrating within EE (reference group in MNL results) versus not migrating and also the chances of migrating outside of EE. In fact, an increase in destination network for migrants, on average, increases the probability to migrate post 1988 and up to 2003 by $6.8 - 7.0$pp. The results could be explained by the fact that increased connections at destination reduces the costs of migrating as connections have already been established which could ease integration.

For foreign network size, which is defined similarly but by location of academic who neither belong to home country of the academic nor destination. The effect is ambiguous as such, assuming that foreign connections are more likely to be close in terms of distance, i.e. belonging to other EE countries, then this would mean that an increase in size would increase the chances of migrating within EE as compared to not migrating following the same logic of reduces cost of migration and easing integration. However, foreign network could also be a proxy for options, openness and also quality, and thus the effect could go through the signalling channel in which this measure of openness facilitates migration. As such, if we see that an increase in foreign network size increases the chances of migrating out of EE as compared to within EE, then it's highly probably that the signalling channel is at play, given the assumption that foreign connections are close in terms of distance. In other words, if foreign connections are close, yet an increase increases the probability to migrate outside of EE as compared to within then, this means that the signalling proxy of openness, options and quality, is stronger than the cost of leaving all of connections in EE. From OLS results, I find that an increase in foreign network size on average, increases the probability to migrate by only 0.1pp and it is not significant in the most specified specification. Results from MNL confirm the assumption that foreign connections are more likely to be close in terms of distance because an increase in foreign network size by 1 unit increases the chances of an academic migrating within EE as compared to not migrating. Additionally, an increase in foreign network size by 1 unit increases the chances of migrating within EE as compared to migrating out of EE, thus going in line with that the effect was channeled through the cost channel more than the signalling channel.

Network quality, by location and from 1980-1988, is proxied by the average citation count or rank of the network. Limitations of this measure are discussed later and in the improvements/further work section. For home network quality, as quality increases it could act more as a signal of the academic himself and thus facilitate migration. Conversely, if it increases this could increase the costs of migration as it is costly leaving valuable connections at home. Thus, my theoretical prediction would be that the signalling channel outweighs the cost channel given that it would signal more the quality of the academic and not deter migration as migrating is usually done to better, more prestigious universities, on average, given also the fact that the Eastern Bloc was completely isolated. As such, a decrease in home network quality would decrease the chance of migrating outside versus migrating within EE. Indeed, the results confirm such prediction, but *marginally* as the coefficients are not economically significant. A decrease in home network quality, increases the chances of not migrating versus migrating within EE, yet the effect is not very economically significant as the relevant logit coefficient will decrease by 0.0001. Similarly, a decrease in home network quality decreases the chances of migrating outside of EE versus migrating within the EE. Thus, these results ae explained by the signalling channel.

For the destination and foreign networks quality, the signalling channel is expected to be at play. Higher quality academics at destination would increase the probability to migrate as they are a signal of the academic himself as he is able to publish with reputable academics. Higher quality academics at foreign countries would increase the probability to migrate as they are a signal of the academic himself as he is able to publish with reputable academics. It is also a signal to the host university about the openness, quality and options of the prospective academic migrant. Better quality destination and foreign networks significantly increase the probability to migrate, however, the effect is economically insignificant, being 0.0003pp and 0.0001pp. and in some specifications insignificant. This highlights that size of the network matters more than the quality, and could be explained by the special context we are in such that academics from EE were segregated from the rest of the academic community such that any additional connection would be of great help and would increase migration prospects, irrespective of the quality of that connection. Additionally, even though it seems that the effect is through the signalling channel, the MNL results from foreign network quality offers other insights. An increase in foreign network quality increases the chances of an academic migrating within EE versus migrating out of EE as the risk or odds are 0.1% lower. This highlights that a greater foreign network quality, which is assumed to be usually in other EE countries, has an effect through the cost channel as leaving the region completely means loss of these foreign connections completely, thus, this acts more as a pull factor. Thus, foreign network quality operates through both channels, it depends if we distinguish between the types of migration or not. Evidence from MNL confirms that the effect of destination and foreign network quality is economically insignificant.

Lastly, given the literature and the theory presented above, I also have some theoretical predictions on potential heterogenous effects by the broad discipline an academic belongs to. For home network size, it is not really easy to predict, potentially no heterogeneity of home network size by broad discipline. Same for home network quality, but if signalling channel outweighs the cost channel then some fields may have this effect stronger than others, i.e., fields that are more reputable or face lower network and language barriers, such as mathematicians, according to Borjas and Doran (2012). For destination and foreign network size it's a bit ambiguous but if Borjas and Doran's (2012) argument is true, such that Mathematicians were more isolated and faced more risk than other fields, then the effect would be greater for them. Increase in foreign network size might have a greater positive effect for fields like Arts and Humanities aligning with Becker et al's (2021) that academics from fields were there are large network barriers and less quality signalling then networks play a bigger role in increase the probability to migrate. For destination and foreign network quality, the direction of the effect is ambiguous.

Indeed, as you will see, the heterogeneous effects I find can be explained by two things: the role of reputation, network and language barriers and the threat of attention from KGB and arrest. Heterogeneous effects of destination network size that is significant, statistically and economically, 20.5pp, for Mathematicians, Computer scientists and Engineers. This aligns with Borjas and Doran (2012) that argue that any Soviet Mathematician that tried to communicate with scholars outside of the Soviet Union, particularly in the US, could risk the potential attention from the KGB or even arrest. Thus due to the extremely limited contact, an additional contact at destination would increase migration prospects from them, more than any other field, especially since they were of high quality/reputation. An increase in foreign network size increases the probability of academics from the Arts and Humanities to migrate significantly more (2.1pp)than academics from other fields. Fields with larger network barriers and less quality signalling, their foreign network, plays a more important role in facilitating migration aligning with Becker et al (2021) results. Additionally, an increase in home network quality has a significantly different and positive effect on the migration probability of Mathematicians, Computer Scientists and Engineers. Evidence confirms that the signalling channel outweighs the cost channel, more for them when the home network belongs to the top 25% in field and region, whereas other disciplines would need their networks to be from top 10%, aligning with Borjas and Doran (2012). Thus, how the results could be explained by the two reasons listed above.

To briefly conclude, the following figures explain very simply the research questions I intend to answer. What I am interested in is whether, for example, academic A is more likely to migrate than academic B where both are from, for example, East Germany, and both have same number of connections at home and migrated to the United States, but the difference is that A has more connections at the United States compared to B, as demonstrated by 1. Secondly, what I am also interested in is whether, A is more likely to migrate than B given the above but also given that they have the same number of connections yet A's network is of higher quality, as demonstrated by 2. The opposite is also studied such that I also interested in whether a greater network size at home would make an academic less likely to migrate or if a better quality or a more reputable network at home would make academic less or more likely to migrate. Lastly, it is important to note that other factors might be at play especially dealing with selection, omitted variable bias, and reverse causality and thus it is important to carefully describe the context of this research, and the economic intuitions and motivations behind the approach taken.

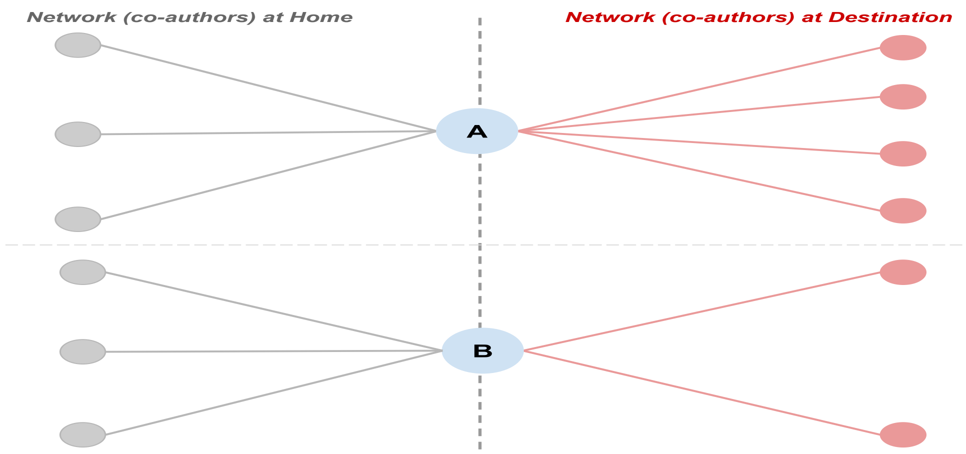


Figure 1: Illustrative Example of the Academic Network of an Academic Migrant - Size

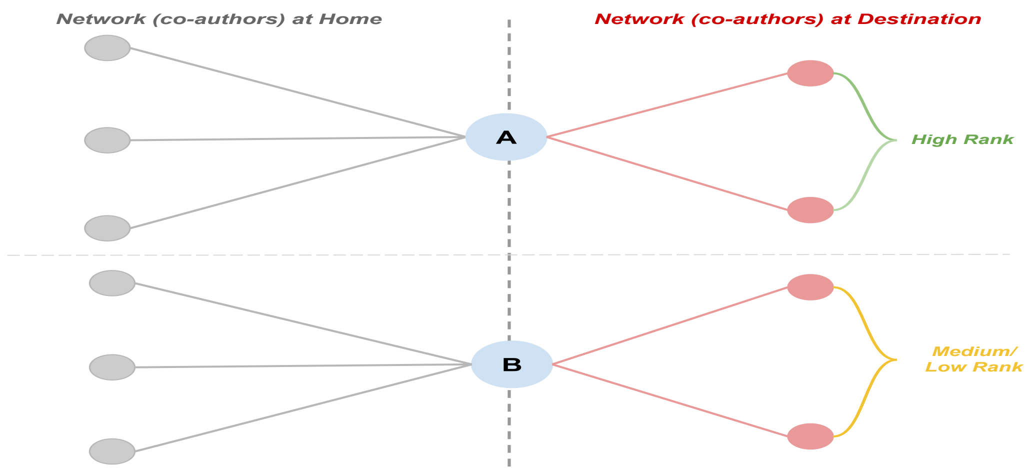


Figure 2: Illustrative Example of the Academic Network of an Academic Migrant - Quality

3 Data

Microsoft Academic Knowledge Graph includes information about academic publications with links to citations, conferences, journals, academic, fields of study and institution. As mentioned previously, there are ambiguities about the relationship between networks and migration, from an empirical viewpoint. These ambiguities arise because it has been historically difficult to differentiate between distinct sources of social capital in a single empirical setting. Traditional migration data does not link social network structures to migration decisions and collecting network data on its own is quite difficult, combining the two is even more so. All of this motivates the quest of searching for alternative data that might give more insights or help us understand and answer some important research questions. The use of big data for migration question is a research question in itself and the power of it in helping us answer questions regarding migration is yet to be confirmed. Microsoft Academic Knowledge Graph is a (scholarly) big data source as it fulfils the 4 Vs of big data. As such, the scale of the data, its different forms, its uncertainty and limitation and its constant update, coincide with volume, variety, veracity and velocity, respectively. Traditional scholarly data sources on the other hand is small, covers few entity types, covers only specific domains, usually outdated and usually cover data primarily from a single publisher. It is also important to explain why MAKG is a better scholarly big data source compared to other similar sources. In contrast with WikiData for example, it contains significantly more bibliographic information, 13x more, and 8x more references. Additionally, most of the papers in WikiData are in English in contrast with MAKG where only 60% of the papers are in English, which is particularly relevant in this context. Additionally, DBLP in RDF (Computer Science biblio) is restricted to a single discipline, Springer's Sci Graph is restricted to publications derived from one publisher, Sci Graph and lastly Open Citations models papers and their citation relations without considering other entity types, such as fields of study as provided by MAKG. [2]

Thus, MAKG is a very rich scholarly big data source that overcomes several limitations posed by traditional data sources and also other scholarly big data sources. It is currently being investigated by Computer Scientists to study the migration of academics through identifying changes in their affiliation and looking, descriptively, at their characteristics (Ricciarelli et al., 2020). The evidence from such papers is purely descriptive such that they provide figures and statistics on the number of migrants they find in their sample and study how patterns change over time. They also study how this pattern changes as the scope of the analysis changes, i.e., focusing on specific fields, years, affiliation locations...etc. The evidence is promising in the sense that it gives power to the use of big data in answering migration questions. However, it abstracts from causality completely, thus failing to identify the causes behind these patterns. Additionally, it makes the assumption that the affiliation location observed for each author is his nationality. This is a limitation of the data, yet, given the context of this research, this limitation is not of a concern as we will argue later. The following sections explain the structure of the data and specifically how the network of each researcher/academic is formed and how it is classified.

3.1 Identification of Academics and Researchers

In MAKG, there exists an ID for each paper that is in the data. For each paper, there exists a list of IDs for all co-authors, the year, the journal/conference, affiliation, field...etc. As such, picking a paper with ID x in year 1988, for example, with authors y and z, we identify two authors. These authors are then searched across the dataset in order to get their publications, field, conferences, citations, rank, and affiliation for each publication. Thus, everything is extracted from the IDs given to each paper in the dataset as shown in the Schema 6. This is then repeated for all papers such that all co-authors are extracted and the above mentioned information are given for them over the years, hinting at the procedure at which the network of each co-author is identified. From the above example, authors y and z are co-authors and thus they

[2] A comprehensive comparison between the different scholarly big data sources is provided in Paszcza (2016).

are in the network of each other in this specific year, 1988, in which paper x was published. Focusing on each specific author, the identification of his co-authors across all papers and all years are done in the same manner allowing the formation of the author's network each year and characterising the network by size and quality depending on the co-authors location, which brings us to the next point.

Sample of Affiliation Locations and the identification of Migrants:

For each published paper, there exists information on the co-authors and the affiliations of the paper. For the time period considered, 1980-1988, all of the Eastern European authors, have a single affiliation location per year, so the problem of having multiple affiliation locations given the different affiliations of the papers published by a specific author is not present here. The focus of the analysis is on Eastern European authors observed in the 9-year span prior the fall of the Berlin Wall. For each author and for each year, their location based on the *country* of affiliation is extracted and is considered their location and nationality at that year. For example, if an author's affiliation was Saint Petersburg State University in Russia for the year 1986, then his location would be Russia. One advantage of being in the context of focusing on countries behind the Iron Curtain is the fact that their affiliation location is a very good predictor of their location and nationality given the fact that migration has been halted. Therefore, sample is restricted to those having affiliation in Eastern Europe and thus belonging to countries behind the Iron Curtain. This includes East Germany, Poland, Hungary, Romania, Bulgaria, Albania, Czechoslovakia and the USSR. The data of course includes locations based on current names of these countries. Thus, the USSR includes 15 modern countries, which are: Ukraine, Georgia, Belarus, Uzbekistan, Armenia, Azerbaijan, Kazakhstan, Kyrgyzstan, Moldova, Turkmenistan, Tajikistan, Latvia, Lithuania and Estonia. Additionally, Czechoslovakia is now Czech Republic and Slovakia. Lastly, Germany is treated as one in the sample, which is important to note for later. All of this is taken into consideration when restricting the sample into only Eastern European academics, which are 29,914 unique academics in 1980-1988, out of approximately 11 million academics observed in this time period. Map 3 is the pre-1991 map as such it shows all of the USSR. It is used to show the location of the Eastern European academics identified in the data and gives support for identification as no other academics were included. Additionally, due to the cross-sectional nature of this empirical research as we will see, the dependent variable is migrating post 1988 up until 2003. Out of the 29,914 academics, 855 of them migrated since the fall of the Berlin Wall, thus accounting to almost 3%. However, not all 858 migrants migrated outside of the Eastern Bloc, 509 migrated outside of it, accounting to 1.7%. The destination locations are shown in maps 4 and 5, for all migrants and out-migrants, respectively. It is clear that the most common home country was the USSR and the most common destination is the United States. It is important to note that that the percentage of migrants is not a low % when compared to other research utilising big data to identify migrants. For example, a recent research using Twitter to identify migrants, used over 60 million tweets, were only able to identify 5k migrants (Sirbu et al., 2021). Yet, the contributions this data, MAKG, offers useful insights, even with its limitations, which we will discuss later. Additionally, the newest version of the data, which we started working on, overcomes many of the limitations of this one, this is discussed in the last section.

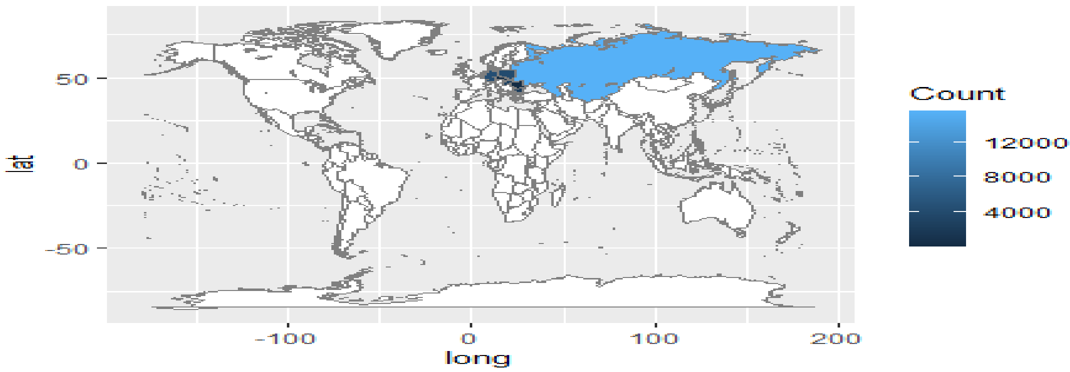
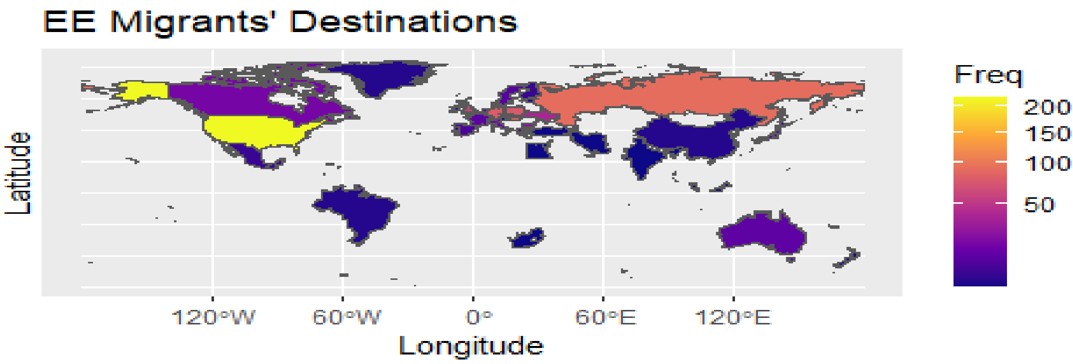


Figure 3: Location of the Eastern European Academics identified



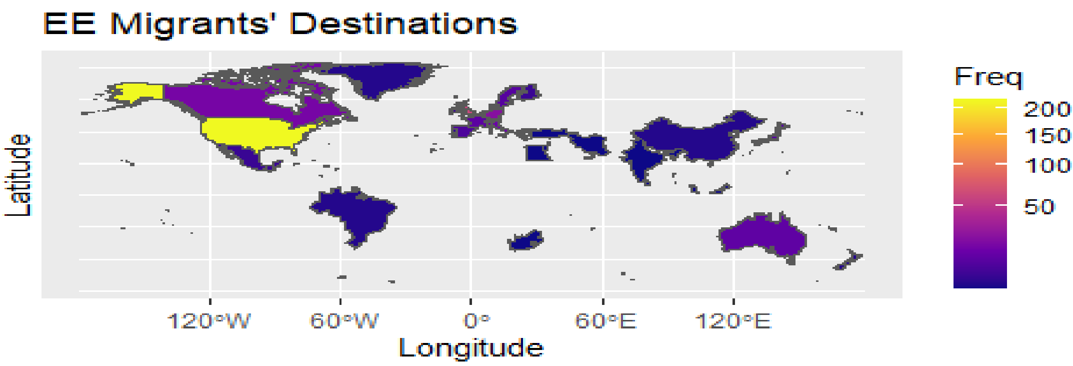


Figure 5: Outside of Eastern European Migration Post 1988

To further explain the data extraction process and the definition of a migrant, consider the following example. Suppose we observe an author ID of x who is observed firstly in 1985 through a paper ID of a and an affiliation SPBU. For each affiliation, there are information on its unique ID and its name with links to the Wikipedia page. Then, the location, by *country* is extracted from the Wikipedia page of the university or research institution. As mentioned before, here we attribute affiliation to nationality and location as argued previously because of the focus on EE academics during the time of the Iron Curtain. Back to the figure, if academic x is then observed in 1990 for example through a paper ID of i for example and his affiliation is now MIT. The location of this affiliation by country is extracted in the same manner as before and now since in 1990 his last affiliation country is not identical to his/her current affiliation country, then he is considered a Migrant in 1990.

The issue with extracting location by *country* instead of by *city* is that this exercise would treat Germany as one, which it was not before 1990, simply because Wikipedia denotes location by current names of countries. In fact, it was divided into East and West, same for Berlin. Thus, for those with German-located affiliations, the location is determined by extracting the city and allocating it to East or West Germany depending on the historical maps of Germany 16, 17. For Berlin, it was done manually as distance from the remains of the Berlin Wall with institutions belonging to the East side of it included in the sample and those belonging to the West side of it, dropped from the sample.

3.2 Network Classification

As mentioned previously, for each author, the location of his affiliation is extracted for each year. Consequently, the location of his/her network can also be identified each year. Network here are direct co-authorships. This allows for the classification of the network by location such that an academic's network could be divided into: *Home*, *Destination*, *Foreign* and *Unlocated*. The home network size by definition refers to the network an academic has in his own country of affiliation with the size being the absolute number of his co-authors per year from 1980-1988. Home is defined as the country of affiliation first observed for this academic, which is important to note when we later discuss within EE migration. The destination network also by definition has non zero values only for migrants that have contacts at destination. Destination here is defined as the country the academic migrates to post 1988 up until and including 2003. Since we don't have multiple migrations between 1988 and 2003 of migrants, this simplifies the classification of networks such that there is only 1 destination per migrant academic. The network can further be classified into foreign, which includes the number of co-authors an author has that neither are located at his/her country nor the destination country, this serves as a measure for openness, quality, and options. The unlocated network includes the authors of which their affiliation location was not identified through the process used above. The limitation and improvement will be discussed in the last section. The total size of the academic network is the simple summation of all these network classifications. Given the cross-sectional nature of this empirical analysis to investigate the effect of networks *pre* 1989 on migration decisions *post* 1988, the network size at home is defined as the sum of all co-authors at home, i.e., same affiliation country, from the period 1980-88. Similarly, the network size at destination is defined as the sum of all co-authors at destination, i.e. the host country of the academic in which he/she migrated to after 1988 and up until 2003, from the period 1980-1988. Same applies for the foreign and unlocated network. Before commenting on the academics' quality in this sample and the summary statistics from the network sizes, we will introduce the quality proxies first.

The *quality* of the network of the academic, which mirrors the *reputation capital* form of social capital and *power-eigenvector-prestige* centrality measure, can also be investigated from the data. Since the authors are extracted and known from the Paper IDs as mentioned previously, the citation count for each author can also be extracted. Additionally, and by construction, so does the paper count. Both are useful measures of an academic's quality, and the same could be applied so as to get the quality of his/her network. Another measure that is uniquely identified in Microsoft Academic Knowledge Graph is the *rank* of the academic. The

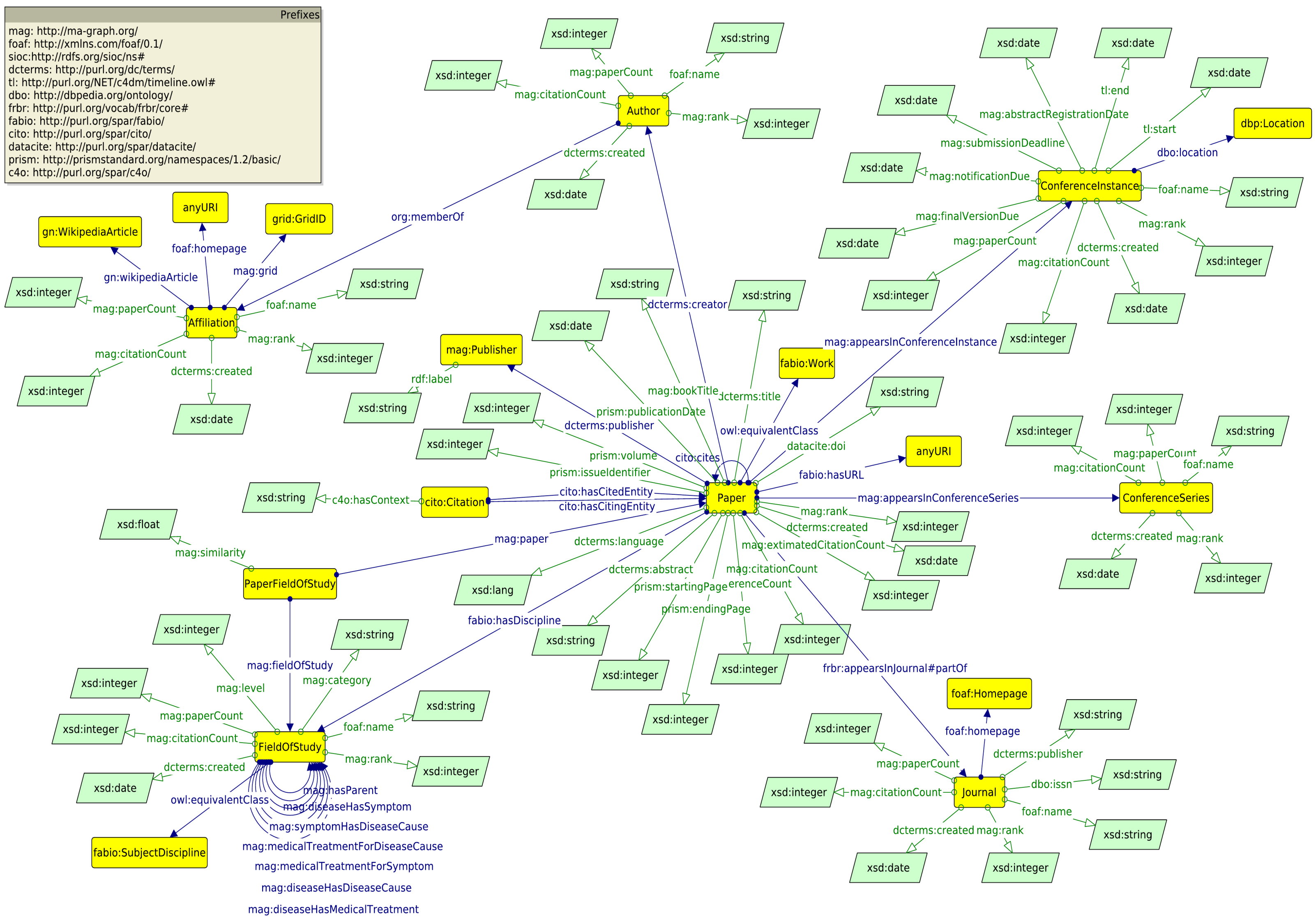


Figure 6: Microsoft Academic Knowledge Graph Schema

rank is a static value allocated to each academic reflecting the "log probability of an entity being important" with importance calculated using relationships with other academics in the data (MAKG,2021). Even though the definition provided is far from comprehensive and open to several interpretations, it hints strongly at the concept of power-prestige-eigenvector- centrality. This measure, even though is useful, is limited in the sense that it is static. It is given for each academic and does not change over time and thus does not provide any information on how the author progresses over time or how he/she was during the period of 1980-1988. However, it still serves as a good complement to the number of citations and publications, that are not constant over time. Additionally, it fits the cross-sectional nature of this analysis and due to data limitations with other measures of quality, the histograms that show the distribution of quality of networks and academics will use the rank. The equation used by MAKG to show how the rank is calculated is as follows.

$$\text{Rank} = - 1000 * \ln(\text{probability of an entity being important})$$

Similarly, the quality of the network is divided based on the location of the authors. At face value and looking at the data description, it is easy to think that the network of an academic could be divided as: number of "high quality", "medium quality", and "low quality", by home, destination, other and unlocated. However, after careful consideration of the data, there appears to be anomalies that are not intuitive. In other words, looking at all of the data we find some academics with more than 80,000 citations and paper counts. This severely restricts our ability to classify networks by quality. Additionally, it is not just that there some anomalies but the distributions are very uneven, and there are some missing information that inhibit my ability to classify academics and consequently the networks they are in. Thus, we have decided to proxy quality by year at home, destination, other and unlocated, based on the following measure. [3]

$$Quality_{ix(80-88)} = \frac{\sum_{j_x \neq i}^{n_{i(80-88)}} Quality_{j_x(80-88)}}{n_{ix}}$$

As such, for every academic i, he has a value for the proxy of quality at location x. Quality could be rank or citation count. The measure is the average of the respective measure chosen for academic i's network in x, which are j that are also in x and are not equal to i, the sum of the measure of respective quality is then divided by the total number of co-authors of academic i at x from the period 1980 to 1988. This mitigates the issue of missing information and anomalies slightly and enables me to proxy the quality of the network of an academic. This is a limitation of this research due to the data, however there are some room for improvements with the newest version of MAKG.

Now that the construction of networks and identification of authors and academics have been explained, we present the summary statistics for the relevant variables. Figure 18 in the appendix shows the distribution of the rank of academics by the sample, for migrants only and for non-migrants only. Due to the definition of the rank, a higher rank value implies a lower quality and a lower reputation. The black line in the figures denote the mean rank of the respective sample. For migrants, the average rank is 15421 compared to an average rank of 18845 for non-migrants. As such, it shows that migrants are of better quality as their rank is lower due to the definition of rank measure. The distribution is also more symmetrical implying that academics with nearly all quality levels migrate yet those in the middle of the distribution, considered medium quality, migrate the most as the frequency is largest. The distribution on the other hand for non-migrants is left tailed highlighting that non-migrants have a large share of academics who are of low quality. Thus, this highlights the importance of controlling for quality and thus we create the following measure:

$$\text{Top Academic} = 1 \textit{ if}$$
$$\text{Top 10\% Citation Count } \textbf{and} \text{ Bottom 10\% Rank} \mid \text{Broad Discipline/Field}$$

[3] Note that, for the current improvements we are working on, and the newest version of MAKG, a lot of these limitations will be mitigated as mentioned in the last section.

The *and* condition stems from the static nature of the rank and the data limitations of the citation count (the anomalies) as explained earlier. This way, we ensure that the academic is truly a top/highly reputable academic. Based on this measure, 2439 academics are considered high quality, i.e., belonging to the top 10% of citation count and bottom 10% of ranks in his/her respective field, in the sample, i.e. the region. Only 559 migrants were high quality, out of a total of 855 and only 308 were high quality out of the migrants that migrated out of EE.

Figures 19 and 20 provide descriptive statistics about the size and quality of the different network types. They offer a visualization for the network statistics in table 1. What is concluded is that, migrants tend to have larger networks, smaller networks at home, larger foreign networks, and of course by construction larger destination networks. Their home network quality through the rank has more observations towards the middle of the distribution in contrast with that of migrants that have a greater concentration towards the higher end of the rank values. Due to the definition of the rank, this implies that migrants had higher quality home networks could be a result of their own quality and thus, controlling is crucial. Their destination network quality has a high frequency of 0s due to the fact that not all migrants had connections at their destination. This is important to keep in mind when we interpret their results in the upcoming sections. Note that the anomalies mentioned earlier can be spotted from table 1 by looking at paper count and citation count values. Note that the average academic rank is considered high (i.e. low quality) given the definition of the rank. This might hint at the fact that academics identified in this sample are on average not of high quality/high reputation/high importance...etc and would potentially imply that selection out of MAKG based on being a low quality academic and thus missed out on being included is mitigated. Since the inverse of the rank is the proxy of quality, i.e. an increase in rank means an increase in quality, given the equation of the rank, then we see that on average migrants are of better quality as compared to non-migrants, especially those who migrated out of EE.

3.3 Other Information from the Data

Field:

One advantage of MAKG is that it is not restricted to one field or a single discipline, compared to DBLP in RDF, which is the Computer Science bibliography. As mentioned previously and as obvious from the schema 6, every single information is extracted from the paper published. As such, the fields that are presented are the fields of the paper and not of the academic per say. These "fields" sometimes are very specific, so an academic for a paper might consequently have a field of "economics of education" whereas a better field to distinguish him from other would be "economics". Thus, for each academic during the period 1980-1988 I identify their most frequent field, i.e., the mode, and if it is a specific "field" I allocate it to a "general" field or a broad discipline. Philosophy, history and arts are considered under Arts and Humanities. Economics, political science, sociology, psychology, geography, and business are under social sciences. Biology, physics, chemistry, environmental science, geology and material science are under natural sciences. Mathematics, computer science and engineering are under Mathematics and Engineering. The last category is medicine. This procedure contributing to understanding more the characteristics of the academics observed in the sample. Yet, as mentioned previously, since everything is extracted from the papers, there might cause some errors. Yet, by classifying them based on broad disciplines as well limits the scope of error as it would be very hard, for example, to mark a mathematician as a biologist.

Figure 7 shows the frequencies of broad disciplines and specific fields for migrants and non-migrants. The mode for both migrants and non-migrants is academics from the natural sciences followed by mathematicians and engineers. For the specific fields, the mode also for both groups appear to be from biology, chemistry, physics, medicine, and mathematics. As we can see mathematicians serve as a greater percentage of migrants as compared to non-migrants. Additionally and most importantly, they tend to cluster and migrate to the US, this potentially hints at potential heterogeneity of network effect for mathematicians as we will

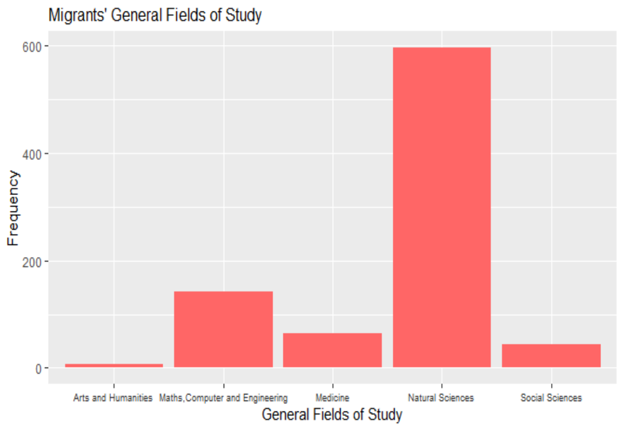
Table 1: Summary Statistics

Group	All		Non-Migrants		EE-EE Migrants		EE-Out Migrants		All Migrants	
Statistic	Mean	St. Dev.	Mean	St. Dev.	Mean	St. Dev.	Mean	St. Dev.	Mean	St. Dev.
Total Pre 1989 Citation Count	2853.319	15,593.170	2,206.514	12,984.220	19,119.080	32,757.940	30,178.820	57,591.740	24,996.990	47,891.200
Total Pre 1989 Paper Count	156.543	497.510	135.842	452.650	845.488	822.457	882. 678	1,283.288	865.253	1,091.418
Author's Rank	18,743.080	1,860.162	18,845.530	1,776.205	15,421.750	998.255	15,071.860	1,193.323	15,235.790	1,119.305
Total Pre 1989 Network Size	13.634	20.717	13.519	20.531	18.394	21.395	16.897	29.492	17.598	26.010
Total Pre 1989 Home Network Size	7.888	9.654	7.964	9.687	6.143	9.026	4.552	6.993	5.298	8.044
Total Pre 1989 Destination Network Size	0.054	0.720	0.000	0.000	2.507	4.071	1.362	3.557	1.899	3.847
Total Pre 1989 Foreign Network Size	1.119	5.316	1.035	4.832	4.091	8.444	3.892	12.854	3.986	11.002
Sum of (average) Home Network Paper Count	137.367	329.372	128.785	321.283	421.026	419.706	440.077	470.497	431.151	447.262
Sum of (average) Home Network Citation Count	2,660.209	9,297.958	2,349.420	8,474.091	11,155.350	17,468.160	15,190.350	24,737.430	13,299.830	21,719.310
Average of (average) Home Network Rank	18,213.960	1,791.563	18,259	1,783.138	16,831.901	1,183.203	16,529.260	1,449.043	16,671.060	1,338.912
Sum of (average) Destination Network Paper Count	5.996	95.728	0.000	0.000	224.262	432.628	199.808	601.538	211.266	528.998
Sum of (average) Destination Network Citation Count	235.383	5,281.980	0.000	0.000	5,982.287	16,945.290	10,410.980	38,275.250	8293.829	30,286.060
Average of (average) Destination Network Rank	201.121	1,796.753	0.000	0.000	1,609.747	4,614.100	4,860.422	7,158.474	7.086.590	8,064.044
Sum of (average) Foreign Network Citation Count	1323.5	7935.505	1092.838	7117.583	8779.526	20904.07	9501.788	20077.9	9163.043	20460.66
Average of (average) Foreign Network Rank	4282.103	7559.283	4096.382	7475.007	9113.111	7944.362	11902.45	7233.833	10594.23	7697.939
Observations	29,914		29,059		346		509		855	

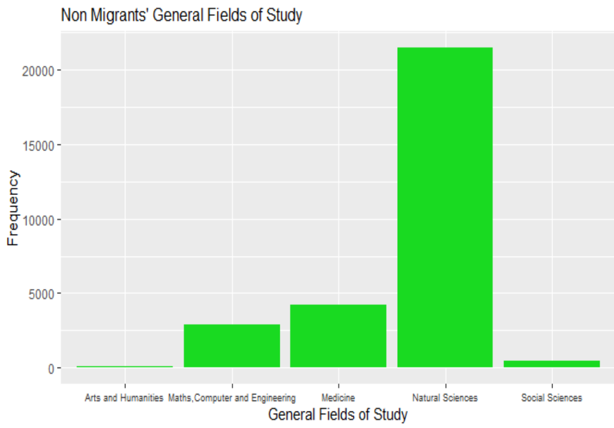
see later on. Additionally, looking at the distribution of specific fields by their home location we see that for the USSR, the second most common field was Mathematics followed by Medicine. This was also the case in Poland. Looking at the shares of specific fields in all countries in EE even those within the USSR, which demonstrates which regions of the USSR (later became specific countries) or parts of the EE that were dominant in specific fields. For example, the majority of mathematicians were from Ukraine (i.e. the USSR), same for Engineering (Ukraine and Czech Republic). Ukraine (i.e. USSR) was also dominant in most of the social sciences, except for geography where Poland had a greater share. Looking at distribution of destination countries by broad disciplines, we see that the US almost had the largest share of migrants in all broad disciplines, especially mathematicians.

Figure 7: Broad Disciplines and Specific Fields: Migrants and Non-Migrants

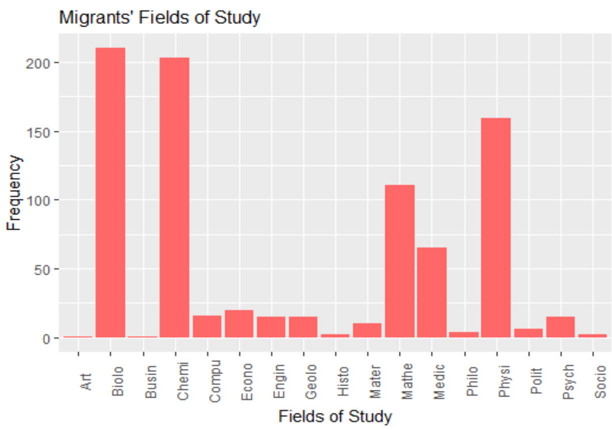
(a) Broad Disciplines: Migrants



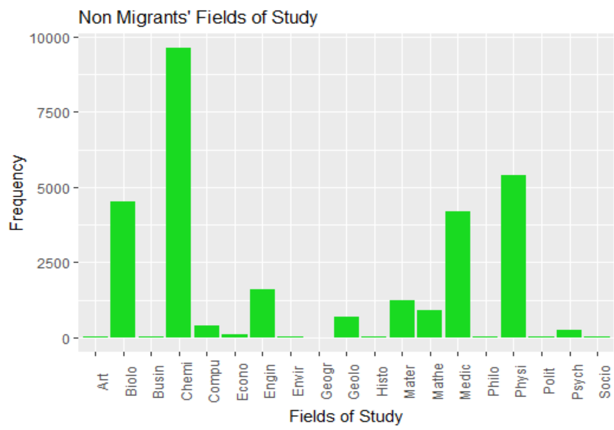
(b) Broad Disciplines: Non-Migrants



(c) Specific Fields: Migrants



(d) Specific Fields: Non-Migrants



Looking at network size and quality distribution by broad discipline offers some interesting insights. As we can see from figures 8 and 9 natural scientists, mathematicians, computer scientists and engineers have larger networks at home. This aligns with what is expected and mentioned earlier. In contrast, social scientists and humanities have lower home network sizes. This could drive heterogeneous effects of networks and thus motivate the analysis in the upcoming sections. Regarding the quality, looked at here through the average rank of the home network, in which a higher rank means lower quality or worse reputation, it appears that in all fields, the "mode" quality is low as the rank is very large. As such, this highlights that connections at

home are not necessarily highly reputable, thus, an increase in their quality might have a signalling effect that outweighs the cost effect, as we will see in the upcoming sections. Figure 21 shows histograms of destination size and quality by broad disciplines. It shows that mostly migrants did not have a lot of connections at destination, apparent from the zeros in the destination network size and rank(as there are no academics). This also hints at the fact that there is no perfect separation between migrants and non-migrants thus logit regression is valid. Finally, 22 shows the count of academics that belong to the top 10 of citation count and bottom 10 percent of rank (due to the definition of rank), by field. This highlights the importance of controlling for not just field but also by the quality of the academic as we see that a large share of migrants are of top quality, especially for Mathematicians and Natural Scientists.
Language: MAKG has several advantages over other scholarly big data sources. One of its main advantages is that it is not exclusive only to papers written in English. In fact, only 60% of its papers are in English, which is particularly useful in this context since we are focusing on Eastern European academics prior the fall of the Iron Curtain. Language of the academic is defined as his mode, i.e. the most frequent language of publication in the time span of 1980-1988. Since the language of each paper is provided by MAKG, this is an easy task. However, given that we have noticed some limitations of MAKG, it is important to re-check or re-confirm this extraction. A random sub-sample of papers has been chosen such that their abstracts are used as inputs for the Python language detect library [4]. The outcome confirms the data provided by MAKG and the consistency of the abstracts' language as detected by Python and the language noted by MAKG. Additionally, for more information and comprehensiveness, I extract the second most frequent language and consequently for each language I aggregate the article count. The most frequent language of publication is English such that 29,214 academics has English as their most frequent language of publication. This is then followed by German, 386 and French 249. Other languages include Polish and Russian. For the second most frequent language, French followed by German followed by English are the most common second language.

Age: Unfortunately, in the data, there is no direct measure for age, thus, I resort to a proxy. Since academics are identified through the papers in the dataset as explained previously and shown by the schema 6, the year of first publication can be extracted from the data. This can be used to proxy the age of the academic in reference to the year 1989. This age proxy is defined as the following and the square of it, for reasons mentioned previously, is also controlled for.

$$\text{Age} \approx 1989 \text{ - Year of First Publication}$$

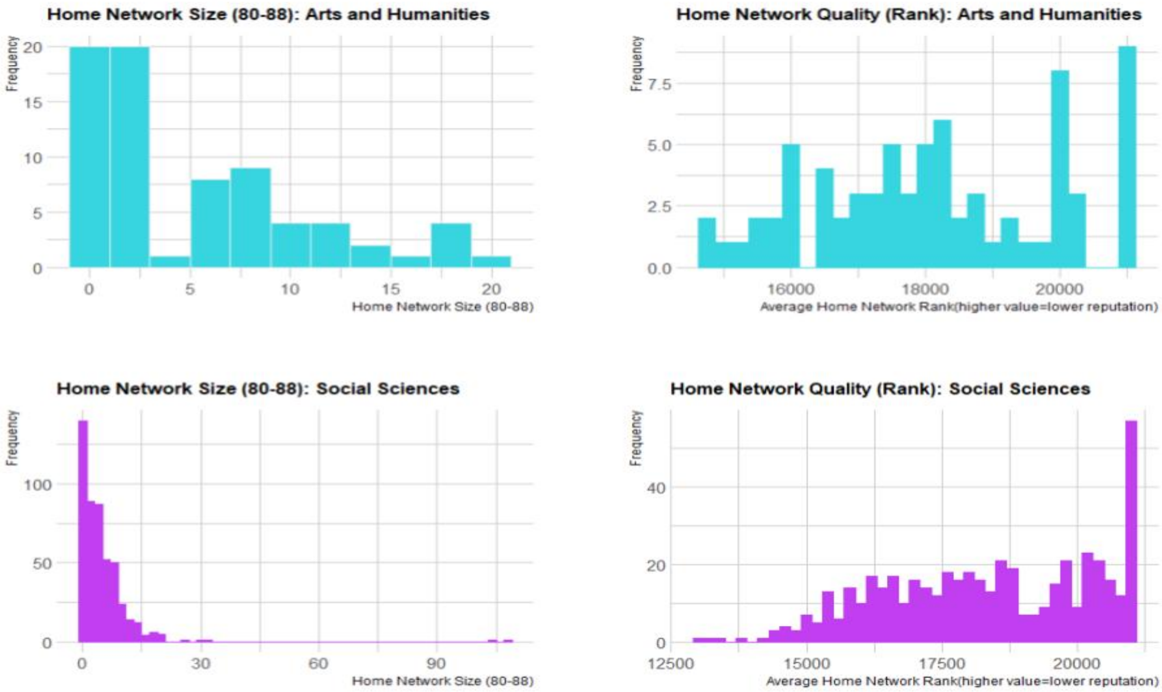
Additionally, it is one of the variables that confirms the correct identification of Eastern European academics such that there are no anomalies in their ages. In other words, it confirms that we don't have values that are ≥ 0, i.e. academics included from 1989 onwards. This further gives support for the data extraction process.

Figure 10 shows the distribution of the age proxy for migrants and non-migrants as defined previously. On average, migrants are slightly older than non-migrants with an average of 10 versus non-migrants with an average of 8. The difference between the means is statistically significant. The same differences hold when I divide by type of migration. The difference between the groups' age means highlight the importance of controlling for age as mentioned previously. However, it is not conclusive that age would increase the probability to migrate as the standard deviation for age for migrants is much larger compared to non-migrants. Thus, there are some outlier values that could be driving this high mean. It would also hint at the potentially inversely U relationship of age with migration. Thus, even though this evidence is insightful, it is not causal of course and not conclusive about the relationship between age and migration.

[4] https://pypi.org/project/langdetect/

Figure 8: Home Network Size and Quality by Broad Discipline

(a) Arts and Humanities and Social Sciences



(b) Medicine and Natural Sciences

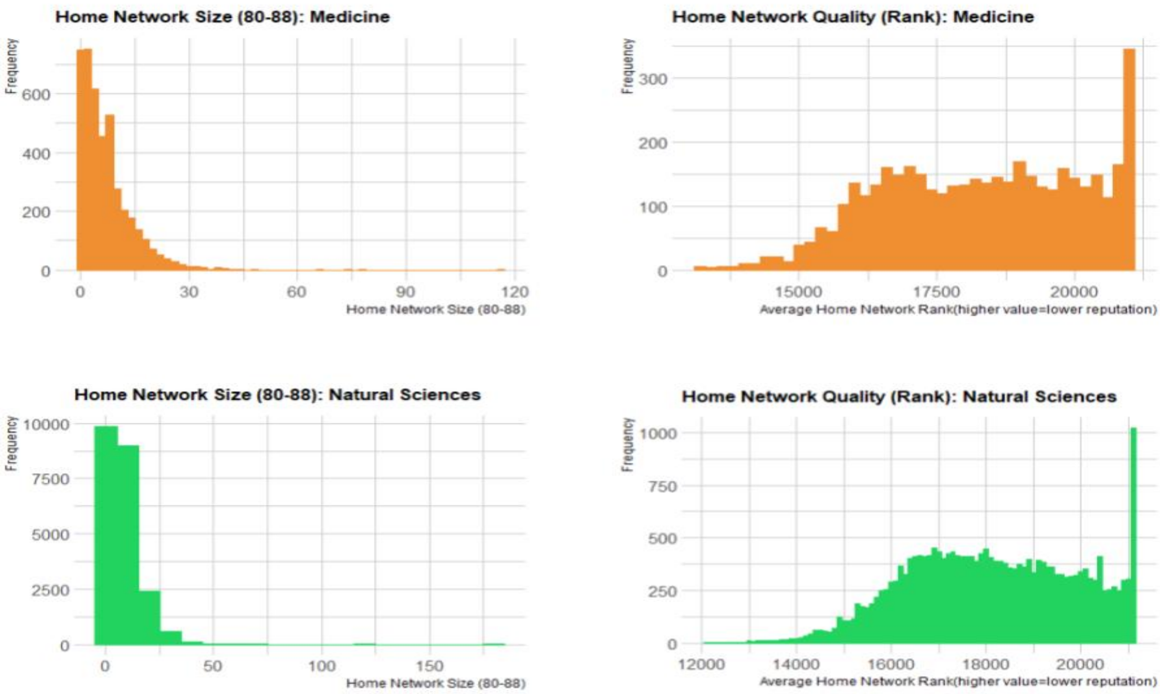


Figure 9: Network Size and Quality by Broad Discipline

(a) Maths, Computer Science and Engineering

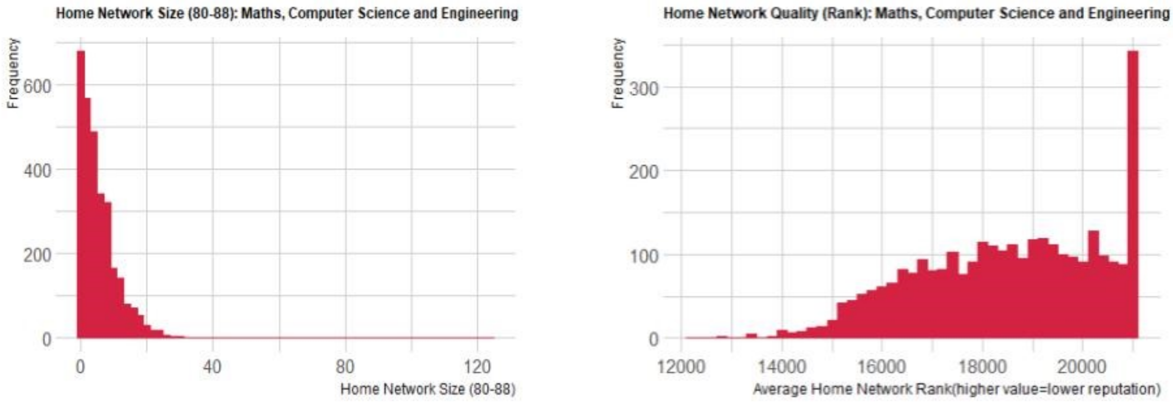
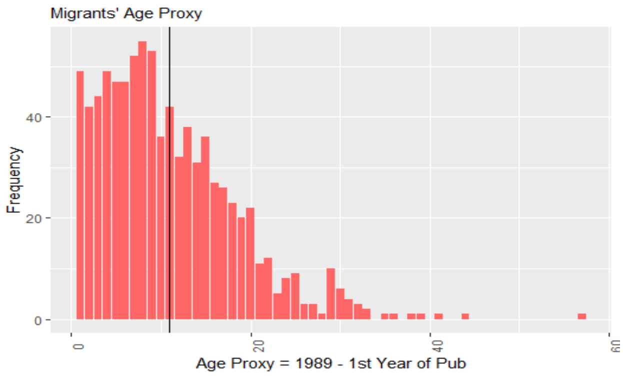
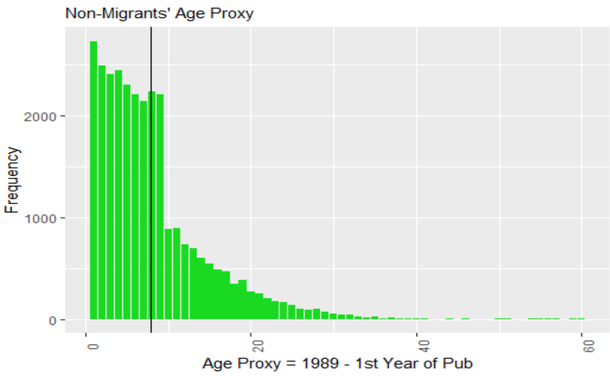


Figure 10: Age Proxy

(a) Migrants



(b) Non-Migrants



4 Identification Strategy

4.1 Context: Fall of Berlin Wall/Iron Curtain

To recap, this research focuses on Eastern European academics behind the Iron Curtain identified in the data from 1980-1988, and tracked up until 2003 and how characteristics of their pre-1989 networks, size and quality, affect their migration decisions post 1988 (i.e. from 1989 (incl) onwards). The choice of dates for this analysis is historically motivated and aligns with the identification strategy and some data restrictions[5].

[5] Note that the new version of MAKG overcomes many of the limitations and that's why the next step would be to increase the sample size by looking at academics post 1945

The focus on this context achieves two things, first, it ensures the correct identification of Eastern European academics and secondly, it ensures that there are no strategic network manipulation and thus no reverse causality concern. Before explaining how these two achievements are met, it is important to highlight and present the context in which this empirical analysis focuses on.

During this period, the Iron Curtain, a political boundary dividing Europe into two separate areas from the end of the Second World War in 1945 up until the end of the Cold War in 1991, was in place. As a result, it severely limited migration between the East and the West from 1950 up until its fall in 1991 (Van Mol and de Valk, 2016). On the other hand, Europe west of the Iron Curtain was instead a transformed continent of immigration thus leading to immigration becoming an important political issue in Western Europe (Bade, 2008).

The fall of the Iron Curtain was not anticipated (Borjas and Doran, 2012; Laqueur 1996; Polyak 2002). In fact, "most believed the system was so strong that it would never essentially change" (Walter Laqueur, 1996). Some, who had some optimistic views, foresaw a change, however over decades and generations. Laqueur (1996) notes that in the West, Sovietologists were surprised of the fall and this political shift, everyone overestimated the Soviet political power and economic performance. Borjas and Doran (2012) argue that the "political system was of the existing Soviet state was not ripe for a sudden change". This is essentially important for the reverse causality endogeneity concern.

The series of events that proceeded border openings and the collapse of the Soviet Union has led to the largest migration wave in and from eastern Europe ever since the events of refugee and forced migration of WWII (Bade, 2008). These events included not only the collapse of the Soviet Union but also tensions with its former states, civil wars in Yugoslavia, and revolutionary changes. All in all, after the opening of the Iron Curtain in November 1989 marked by the fall of the Berlin Wall, immigration from eastern Europe started and surged in all categories, and thus including migration of academics and scientists (Marshall, 2000). Thus, the collapse of the Iron Curtain induced new migration flows and enabled and facilitated the migration of academics and researchers from Eastern Europe, the focus of this analysis.

The context of this empirical analysis ensure the correction identification of EE academics. The data is presented in a way in which an academic's publication has an affiliation from which a location is extracted. In general contexts, the affiliation location is not necessarily synonymous to nationality. However, by focusing on affiliations with countries in the Eastern Bloc, we ensure that the sample is of only Eastern European academics. As such, those who are identified with Eastern European affiliations did actually belong to the Eastern Bloc. This is due to migration restrictions and the context in that region at that time. Secondly, it is important to demonstrate descriptively how the Iron Curtain was an effective barrier to migration. For the Eastern European academic identified, prior the fall of the Berlin Wall and from 1980, there has been only 263 unique migrants. All of these academics migrated to other Eastern European and there was no outside migration at all, further proving that, in this context, the Iron Curtain and the Berlin Wall effectively halted migration. Eastern European academics from 1980-1988 migrated to Poland, Ukraine, Russia, Moldova, Hungary, Czech Republic, East Germany, Bulgaria, Slovakia, Romania and Latvia, all of which were part of the Eastern Bloc. Out of these countries, the Czech Republic, Hungary, Poland, Latvia, and Slovakia later joined the European Union and 77 of these migrants migrated to these countries, which is important to control for.

4.2 Reverse Causality

There are two sources of potential bias in every empirical analysis, reverse causality and omitted variable bias. Firstly, reverse causality is when the dependent variable itself affects the explanatory variable and thus the relationship between the two is not one way. In this context, reverse causality occurs if migration decision affects and shapes the networks of academics and researchers. For example, an academic who decided to migrate to the United States in 1991 would strategically start forming connections at destination years before, in anticipation of migration. This is usually done to facilitate migration and reduce its costs and enhance integration into the new university or institution. This concern is very problematic in the migration-networks literature as it is very hard to disentangle the effect of the network on migration decision from the effect of migration itself on the network. This research contributes to the literature by focusing on the fall of the Berlin Wall and the Iron Curtain as mentioned before such that networks pre the fall were not manipulated or strategically formed simply because migration was halted and there was not any anticipation of the alternative situation. Even though this is reasonable to assume, evidence in support of the absence of reverse causality is presented by the following figures.

Firstly, figure 11 shows the evolution of the share of the co-authors of the academic at home, at destination, other and unlocated, over the time span of 1980-1988, for post 1988 migrants only. In other words, it depicts how the geographic distribution of the academic migrants' network changes over time, prior the fall of the Berlin Wall. The figure shows that both shared, home and at destination remain roughly constant throughout the period of 1980-88. Thus, it can be concluded that academic migrants appear not to publish papers with co-authors at destination *strategically* before migrating. If this was the case, it is expected to see a spike in number of co-authors at destination or a gradual increase in the years prior the fall of the Berlin Wall. A comment on the high share of unlocated academics is provided in the further work section.

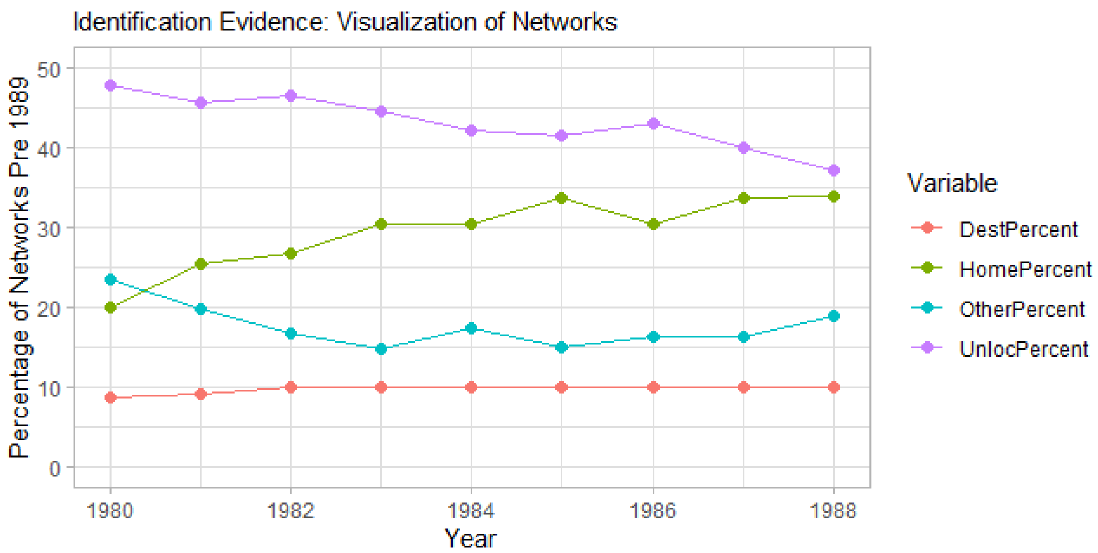


Figure 11: Geographical Networks' Share for Migrants Pre 1989

Figure 12 gives further evidence in support of using pre-89 (pre fall of Berlin Wall and Iron Curtain) networks for reasons mentioned previously. After the fall of the Berlin Wall and consequently the Iron Curtain, it is expected that academic who migrated would have a larger share of their network from their new "home" country, i.e. defined here as their destination country. As such, this is what is apparent in figure 12 such that the share of home network falls and the share of destination network increases. In fact, the two patterns intersect and switch clearly indicating migration and integration into the new universities/institutions. The share of destination network surpasses the share for the home network from 1991 onwards and the share of the other network increased and stabilizes at a "new high" it reached in figure 11. This aligns logically with what is expected in terms of the evolution of the shares of different networks and gives support for my identification strategy such that academic migrants did not strategically alter their network and thus, the concern of reverse causality is mitigated.

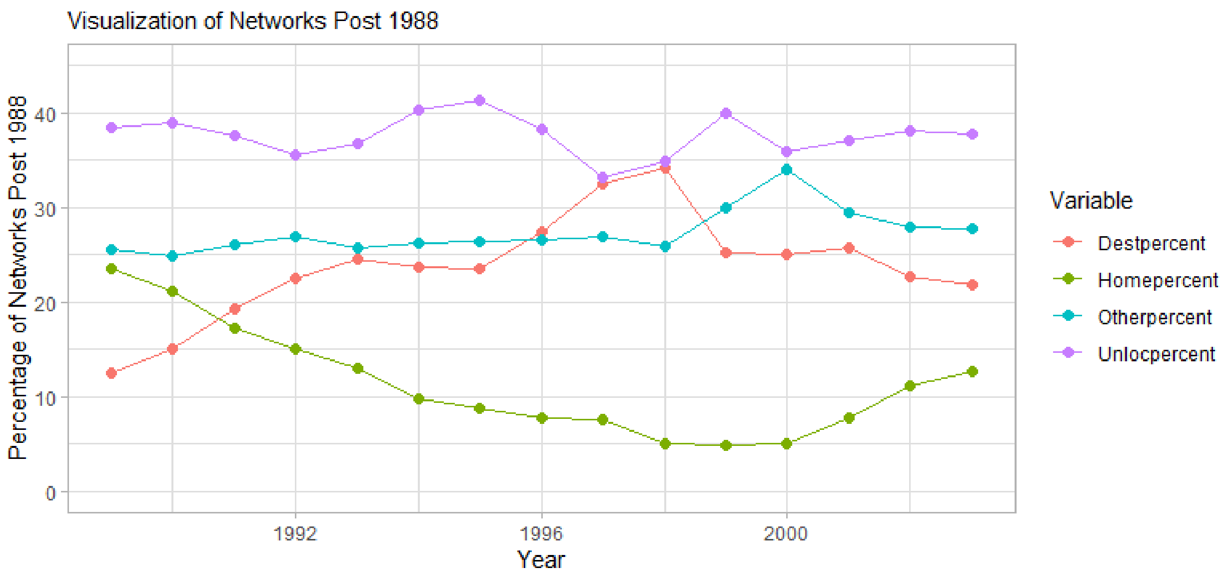


Figure 12: Geographical Networks' Share for Migrants Post 1988

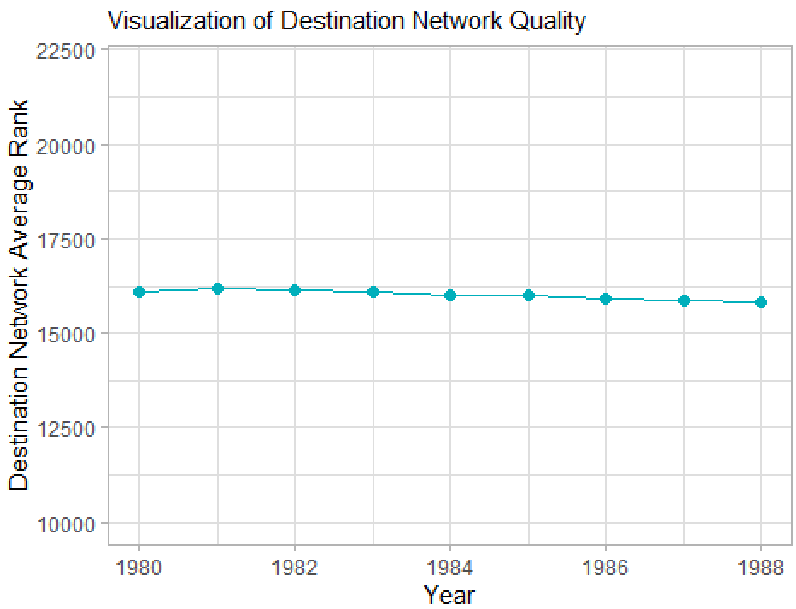


Figure 13: Pre 1989 Destination Network Proxy (Rank) Mean

One would think that migrants from an EE country to another EE country have a higher probability of strategically altering their network simply because the Iron Curtain didn't prevent migration within the Bloc, yet there is no evidence of this in terms of network manipulation as apparent from figure (a) in 23 in the Appendix. . Additionally, one would think that out migrants could specifically strategically manipulate their network more in contrast with the other migrants as they're moving outside of the Eastern Bloc, which could mean higher costs of migration, if they anticipate the fall of the Iron Curtain/Berlin Wall. As such, we find no evidence of this with the share of destination network constant throughout the period of 1980-1988, as apparent from figure (b) in 23 in the Appendix. This gives more evidence in support of the identification strategy.

Looking at the shares could hide some other types of network manipulation. For example, a low quality academic could be substituted by a high quality or a highly reputable academic at destination to further facilitate migration. If network manipulation occurs through knowing few highly known/established academics at destination, then this means that these authors, their quality would be very close to the inverse (due to the equation of how the rank is calculated) of the rank given by MAKG. In other words, if authors strategically choose to co-author with certain academics that are considered high quality, then it is certain that *they* view them as high quality *then* and as such the inverse of the rank coefficient given to them is actually a good proxy for them, given that the rank is a static number. However, given the previous limitations about the rank and other quality proxies, the quality of a destination network is the average rank for the network in each year. Thus, it is heavily influenced by the *size* of the destination network itself. Since the destination network share appears to be constant, then this should not be a concern. However, we still look at the evolution of average rank of destination network in each year. Figure supports this claim as we can see the evolution is roughly constant over the year. Even though there might be a decline apparent in 1988, this is not very economically significant as the difference between the average destination network probability of being important in 1980 compared to that of 1988 is just 0.00000141.

4.3 Omitted Variable Bias

The second, equally important, threat to identification, would be omitted variable bias. This occurs when variables that are not controlled for, i.e. in the error term, not only affect the dependent variable, the probability to migrate, but also the network structure or characteristic that we are focusing on. In other words, the characteristics of the structure of the network may be a proxy for other characteristics that are field, quality, location or individual specific. Examples of potential selection into migration, whether on characteristics of the individual or location has been given above and the literature provided the motivation and economic intuition behind the provision of the controls used in this analysis. The host of controls includes: proxy for age, language of publication, papers published percentile, quality of the academic, different types of previous inter-EE migration, field, home and destination countries and even field x home x destination countries fixed effects. The latter two control for factors that would affect all individual considering the same home-destination pair, such as "gravity" effects like specific institutions located in specific locations attracting specific nationalities from specific fields, or other factors such as amenities, wages...etc. These controls address nearly all observable aspects in which selection into migration occurs, as shown and motivated extensively by the literature. As such, we argue that the scope for omitted variable bias would be limited in this context, yet there are some remaining concerns, especially regarding the *unobservable* individual characteristics. Since the main model of analysis is a cross-section as all explanatory variables are all associated with the same single period (80-88), this inhibits the ability to include individual fixed effects that would be otherwise included if the analysis was a panel. Individual fixed effects would absorb all time invariant individual heterogeneity and would deal with the issue that some academics are just inherently more likely to migrate than others. Thus, this is important to take into consideration, and as such, think of robustness tests.

5 Empirical Strategy

This is a cross sectional analysis such that the unit of analysis is the academic i. This is because in order to answer the question of this research, the network characteristic sum, or mean, depending on the variable, before the fall for each characteristic, i.e., size and quality, is used as the explanatory variable and the dependent variable would be migration post the fall. Formally, the main economic specific is as follows:

$$\begin{aligned}\text{Migrated Post 1988}_i = &\beta \cdot \textbf{Pre1989 Network Characteristic}_{i_x} + \gamma_1 \cdot \text{Age Proxy}_i \\ &+ \gamma_2 \cdot \text{Age Proxy}_i^2 + \mu \cdot \text{High Quality}_i + \nu \cdot \text{Previous Migration}_i \\ &+ \alpha_{hd} + \pi_l + \nu_f + p_i + \epsilon_i\end{aligned}$$

The dependent variable is *Migrated Post 1988*, which is a dummy variable equal to 1 if academic i migrated from his/her Eastern European country to another country, as defined in the data section, in any year following the fall of the Berlin Wall, 1989 included, up to 2003. The main explanatory variable, as explained above, is, either the sum of the size of the network at home or at destination for the years 1980-1988 or the relevant measure for the quality at home or at destination for the same years. For quality, the sum is not as straightforward to do as the sum of the network size. This is due to the data limitations mentioned above and how the quality proxies are themselves measured, which are the average of the rank and citation count of the network by location. As mentioned previously, there are controls that are motivated intuitively and by the literature, and mitigate the presence of selection on observables.

To mitigate endogeneity concerns, especially given the empirical strategy utilised in this analysis, a wide host of controls are included that are strongly motivated by the literature. These variables, if not controlled for, would otherwise lead the estimated coefficients on the network characteristic to be biased as not only they're expected to be correlated with the network characteristics but also with the probability to migrate.

The age proxy is the difference between the first publication year of the author and 1989, and the square is also controlled for. This is motivated by Gould and Moav (2014) that argue that academics in the middle of their careers, are more likely to emigrate than the oldest or the top researchers, to countries with high residual wage inequality. Additionally, if we expect older academics to be more established, more experienced, therefore could be of higher quality/reputation, more connected, and thus potentially higher probability of migrating, then we would falsely attribute that to the effect of the network (Becker et al., 2020). On the other hand, if we expect that as an academic gets older, he/she has less time to get the return to investment to migration. Thus it's important to control for age.

For High Quality, it is dummy variable that is equal to 1 if the academic belongs to the top 10 percent of academics in his field from the EE sample. As such, if an academic belong to the top 10 of total citation count for the period 1980-1988 and belongs to the bottom 10 percent of rank values (due to how the rank is calculated) in his field and in the sample of EE academics, he is considered a high quality/highly reputable academic. The quality of the academic himself could be a potential confounder, how this influences his network and the self-selection of certain researchers, especially top researchers, into emigration. Becker et al., (2020) gives the example of certain places employing the best mathematicians, which in turn form a network, such that if the best mathematicians emigrate then falsely attributing the effect of their quality to their network might be the case. Additionally, simply, better academics or researchers, may simply have a larger network thus further facilitating migration for them. On the contrary, they could simply not need a large network to migrate simply because their reputation is a signal for their quality or due to the other factors that we will see. Top researchers might also be attracted to certain locations due to certain characteristics specific to home and destination that affect their network structure and characteristics, might speak specific languages, might be of similar age groups, which are all important to control for.

Another important control variable to account for selection into migration by certain academics is language, which is done by looking at the mode language of publication. A concern would be that proficiency in a foreign language affects the distribution of the network of an academic and the expected returns from migrating to the countries where this language is spoken (Bertoli and Ruyssen, 2016). Gould and Moav (2014) argue that there is large variation across countries, one of which is language barriers, which may affect the direction and size of selection. De la Croix et al (2020) study the academic market and the rise of universities in medieval and early modern Europe to investigate agglomeration and sorting patterns of academics and the role of positive selection into migration. In fact, De la Croix et al (2020) find that the patterns of sorting in agglomeration give evidence to a functioning academic market that is made possible by the use of a common language, which is Latin, and political fragmentation. In fact, the persistence of Latin as a "lingua franca" reduced cost and allowed scholars to teach anywhere. This hints at the importance of language and how it could serve as a barrier or a cost reducer and a facilitator for migration and thus knowledge of foreign language must be controlled for (Becker et al, 2020). In fact, Becker et al (2020), find that the two most attractive locations were the United States and the United Kingdom as a result of the lower language barrier, by studying academic networks and high skilled emigration from Nazi Germany, where a similar pattern is seen in my data, especially for the United States.

Previous migration is a categorical variable with 3 levels, one which is no previous migration between 1980-1988, the second is previous migration between 1980-1988 to an Eastern European country that did not become part of the EU and the third is previous migration to an Eastern European country that did become part of the EU after 2003. This controls for the fact that migrating once facilitates future migration as the individual becomes less attached to his home. Additionally, the inclusion of this control is historically motivated as many countries that were in Eastern Communist Bloc became part of the EU 2003 on-wards. Thus, controlling for this captures some unobserved characteristics of the individuals that might make more individuals more likely to migrate than others.

Now turning to fixed effects in this model, we use *felm* in R to fit linear models with multiple group fixed effects, which is very similar to lm whilst using dummies so as to mimic fixed effects, and it also facilitates clustering of standard errors, which will be motivated briefly. This method uses the Method of Alternating projection to sweep out multiple group effects from the normal equations before estimating the remaining coefficients with OLS. This method reduces the potential for omitted variable bias and fulfills the notion of fixed effects and clustered standard errors as we have seen that there are some indications of multiple group effects through the descriptive statistics section above (Cameron et al., 2011). We include fixed effects for the most frequently used language of publication for the academics in 1980-1988, π_l, fixed effects for broad disciplines or general fields as outline in the data section, ν_f, and paper quantile fixed effects p_i. We include home-destination country pair fixed effects, α_{hd} ,and also cluster standard errors accordingly as motivated by prior evidence. We also include field specific fixed effects.

It's really important to include field and home-destination fixed effects. Some fields simply publish more than other or have larger networks by nature of their field. For example, computer scientists tend to have 4+ co-authors for each papers versus economists who have much less on average. Additionally, some fields were more demanded in other countries post 1988 or faced different patterns and trends post the collapse of the Soviet Union and the Iron Curtain, for example the influx of Soviet Mathematicians into the US (Borjas and Doran, 2012). This brings us to the next point which is that there are some factors that are specific to the home and destination pair as Borjas and Doran's (2012) example and Becker et al's (2020) example of Gottingen employing the best mathematicians. Gould and Moav (2014) examine how levels and sources of income inequality shape how a country attracts high skilled workers. The predictions of their model, which imply that skilled workers, in contrast with less-skilled workers, are more likely to migrate to a country with a higher return to skill relative to their home country and the reverse holds (Borjas, 1987). Gould and Moav (2014) argue that a significant component of a country's inequality is the returns to skill and thus the dispersion of the wage distribution has implications for the type of selection that occurs with a concentrated compressed distribution leads to positive selection, where the most skilled leave, and a dispersed leads to negative selection.

This is binary regression model, more specifically, a linear probability model since the dependent variable for each academic, i, is either equal to 0 or 1, which depends on the above mentioned explanatory variables ($Pr(Y = 1|X = x) = x^{'}\beta$). This could easily be fitted using linear regression, OLS. We use robust standard errors as the ϵ_i is heteroskedastic in LPMs due to the fact that the variance is not constant and depends on the value of the independent variables X (Wooldridge, 2002). More specifically, We use clustered robust standard errors to account for the fact that the data is structured in clusters as such standard errors are correlated within groups of observations. As such, we avoid having spuriously low estimated standard errors that lead to overly narrow confidence intervals and low p-values thus increasing the scope for making wrong conclusions (Wooldridge, 2003; Cameron et al, 2006; Roberts, 2013).

To account for different types of migration, an estimation model that consider all types of migration as an output variable without them being ordered is preferred. As such, distinction between migrants who migrated to other Eastern European countries post 1988 from migrants who migrated out of Eastern European countries, should be done. As such, we extend this analysis by using a multinomial logit regression where the outcome variable has 3 possible types: *Not Migrating*, *Migrating to another Eastern European Country* and *Migrating outside of Eastern Europe*. The logit coefficients reported are relative to the chosen reference category and just implies how a 1 unit increase in a variable would decrease/increase the logit coefficient of the category shown relative to the reference category. Relative risk ratios allow an easier interpretation of the logit coefficients as they are the exponentiated value of the logit coefficients. These ratios give how more/less likely it is to stay at the stated category as compared to the reference category as a result of an increase in the variable of interest. This offers great insights as it allows me to compare the effect of network across all categories and highlight the potential mechanisms and channels at play.

It is important to highlight more the properties and use of the multinomial logit model. The motivation behind the use of MNL is due to the nature of the outcome variable. Additionally, MNL is currently being used extensively in the migration literature (Blumenstock, 2019; Dahl and Sorenson, 2010) and most importantly it offers sound microeconomic foundation of utility maximization with a random utility model (Blumenstock;2019, Mcfadden, 1974). Additionally, it's important that estimation meets all necessary requirements for estimation of MNL. Firstly, the Independence of Irrelevant Alternative (IIA) property which means that the probability ratio for 2 alternatives depend only on the characteristics of these two alternatives and not on other alternatives. This relies on the notion that errors are iid, which might be violated in practice especially in cases where some important variables are omitted. Having no correlation between the errors condition is mitigated by the use of clustered robust standard errors. IIA can be tested using the Hausman-McFadden test (1984), which is a variation of the Hausman (1978) test (Vijverberg, 2011; Chen, 2007). The null hypothesis is that the IIA property is met. If so, running MNL on a subset of alternatives or the full set of alternatives, provides consistent and efficient estimates. If on the other hand the IIA property does not hold then the estimated coefficients from MNL on full set of alternatives are no longer consistent but coefficients from MNL on a subset of alternatives are consistent if the subset is selected properly (Hausman and McFadden, 1984;Vijverberg, 2011). The MNL regressions done meets the IIA property as we fail to reject the null hypothesis of McFadden test. An additional requirement is that the groups of the outcome variable cannot be perfectly separated by the predictor. Which is the case in the data as previous statistics show, thus, there is no threat of having unrealistic coefficients and wrongly attributed economic significance.

6 Results

6.1 Network Size

The results for the effect of network size by location, i.e., *home*, *destination* and *foreign* are presented. As it will be shown, I find that *mostly* the results can be explained by the **cost channel** in which, depending on the location of the size of the network we are looking at, the size of the network reduces or increases the costs of migration and thus acting as a facilitator or a de-facilitator of migration, i.e. a push or a pull factor. Starting with home network size, table 2 columns 1 and 2, the latter with the most rigorous specification with controls and fixed effects; an increase in home network size (1980-1988) by 1 unit reduces the probability of an academic to migrate (1989-2003) by $0.1 - 0.05$pp. This shows that authors have a higher cost to leave their already established academic networks as their size increases thus reducing the probability to migrate. However, this decrease is not very economically significant, yet very statistically significant. Regressions from logit regression confirm the direction of the effect thus not included here for redundancy.

Turning to MNL estimation in order to distinguish between the types of migration, the dependent variable is now a categorical variable with more than 2 levels that do not follow a particular order. It has 3 levels, one level is for not migrating at all between 1988-2003, one for migrating to another EE country in this time frame and one for migrating outside of the Eastern Bloc in this time frame. However, as obvious from the above specifications, the FieldxHomexDest fixed effect specification is not included, as it is not possible or quite difficult to include the same restrictive set of fixed effects as I did for the linear regression. The logit coefficients reported are relative to the chosen reference category and just implies how a 1 unit increase in a variable would decrease/increase the logit coefficient of the category shown relative to the reference category. From them we can get the relative risk ratios which give how more/less likely it is to stay at the stated category as compared to the reference category as a result of an increase in the variable of interest. This offers great insights as it allows me to compare the effect of networks across the different categories. Note that the reference group in the following tables is those who migrated to other Eastern European countries post 1988. Thus, the interpretation of the coefficients is done in reference to this group. The estimated coefficient on home network size for those who didn't migrate relative to those who migrated to another EE country are positive and significant. In column 3, the coefficient suggests that for a 1 unit increase in home network size, the logit coefficient for those who didn't migrate relative to those who did migrate but to another EE, will go up by that amount 0.059. In other words, if this happens, the chances of an academic staying in the reference category, i.e. migrating to another EE, are lower compared to staying in this category, i.e. not migrating. The result confirm that having a greater network size at home increases the chances of staying at home relative to migrating to another EE country which is close and potentially similar on some aspects compared to the home country. The relevant risk ratios, which are the exponentiated values of logit coefficients, imply that that keeping all variables constant, if home network size increases by 1 unit, an academic is 1.071 times more likely to *not* migrate compared to the category of migrating to another EE as the risk or odds are 7.1% higher, aligning with the above interpretation. [6]

The estimated coefficient on home network size for those who migrated to a country outside of EE relative to those who migrated to another country are negative and significant, column 4 in 2. This implies that if the home network size increases by 1 unit, the logit coefficient for migrating outside of EE relative to migrating to another EE will go down by -0.037, implying that chances of migrating to another EE in the scenario of an increase in home network size are higher compared to migrating outside of EE. The risk or odds are 3.7% lower in the migration to EE group as compared to migrating outside. This result highlights the cost channel such that a greater network size at home increases the cost of migrating but the effect of the cost is asymmetric such that the cost is larger when migrating outside of Eastern Europe

[6] Note that relative risk ratios are simply the exponentiated value of the logit coefficient. If in the case they're insignificant, I state that.

as this is associated with a potential complete loss of such network, whilst migrating to a nearby Eastern European country with potentially more similar characteristics, the cost is not as big. Thus, looking at home network size effects we conclude that the results are explained through the cost channel in which it's costly to leave home when connections at home increase, however the effect is not very economically significant. Additionally, the cost channel increases the chances of not migrating significantly compared to migrating within EE and decreases the chances of migrating outside of EE compared to migrating within EE.

Looking at OLS results for destination and foreign network size, i.e columns (5) and (6) in table 2. Note that by definition, destination networks are only for migrants, even though some have none. Foreign networks are the size of network of co-authors that aren't in home country nor destination country, thus for migrants and non-migrants (from 1980-1988). In column 5, the least specified specification, implies that a unit increase in the size of destination networks for migrants, on average, increases the probability to migrate post 1988 and up to 2033 by 6.8 pp. On the other hand, an increase in foreign network size on average, increases the probability to migrate of an academic to migrate by only 0.1pp. In the more specified specifications, the effect of an increase of 1 unit in destination network increases in magnitude to 7.0pp and remains statistically significant, whilst for the estimated coefficient on foreign network size is no longer significant. The results could be explained by the fact that increased connections at destination reduces the costs of migrating as connections have already been established which could ease integration. On the other hand, the insignificant positive coefficient on the foreign network size imply that this measure of openness, quality and options does not have a significant role in migration. Yet, since it's positive, and given the assumption that foreign connections are usually close in terms of distance then, this potentially implies that the signalling channel is the one at play. It's important to look at MNL results to have better understanding of the channels in play.

The estimated logit coefficient on foreign network size for those who didn't migrate relative to those who migrated to another EE country, i.e. column (7), is negative and significant. In column 7, the coefficient suggests that for a 1 unit increase in foreign network size, the logit coefficient for those who didn't migrate relative to those who did migrate but to another EE, will go down by that amount 0.015. In other words, if this happens, the chances of an academic staying in the reference category, i.e., migrating to another EE, are higher compared to staying in this category, i.e. not migrating. The result confirms that having a greater foreign network size increases the chances of migrating to another EE relative to not migrating. For relative risk ratios, keeping all variables constant, if foreign network size increases by 1 unit, an academic is 0.985 times more likely to *not* migrate compared to the category of migrating to another EE as the risk or odds are 1.5% lower, aligning with the above interpretation. The estimated logit coefficients on foreign network size for those who migrated to country outside of EE relative to those who migrated to another country is negative and insignificant. The negative coefficients imply that for a unit increase in foreign network size, the chances of migrating out of EE are lower, yet the coefficients are not significant. Thus foreign network results from MNL confirm the assumption that foreign connections are more likely to be close in terms of distance because an increase in foreign network size increases the chances of an academic migrating within EE as compared to not migrating. Additionally, an increase in foreign network size increases the chances of migrating within EE as compared to migrating out of EE, thus going in line with that the effect was channeled through the cost channel more than the signalling channel in the sense that leaving the region completely means complete loss of these foreign connections, which are close in terms of distance to the home country.

Since the destination network size is specific to migrants, the logit coefficients across all categories are very negative and significant especially comparing those who didn't migrate to those who migrated to another EE. The estimated logit coefficient in column 7 show that an increase in destination network size for migrants implies a significant increase in chances of an academic for migrating to another EE instead of not migrating. The chances of staying at home are nearly 0 as we look at the relative risk ratios. The magnitude of effects is lower when comparing those who migrated to another EE and those who migrated outside,but statistically insignificant. However, the results indicate that for a unit increase in destination network size, the chances of an academic migrating outside of EE as compared to migrating within EE are lower. The interpretation is not as straightforward with the foreign network size as the destination network is only defined for migrants.

Table 2: Effects of Home, Destination and Foreign Network Size on Probability to Migrate Post 1988 up to 2003.

	Dependent variable:							
	Migrated 1989-2003							
	Felm		*MNL*		*Felm*		*MNL*	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
			Didn't Migrate	Migrated outside of EE			Didn't Migrate	Migrated outside of EE
Home Network Size (80-88)	-0.001^{***}	-0.005^{***}	0.059^{***}	-0.037^{***}	–	–	–	–
	(0.0002)	(0.0002)	(0.009)	(0.013)	–	–	–	–
Destination Network Size (80-88)	–	–	–	–	0.068^{***}	0.070^{***}	-14.528^{***}	-1.177
	–	–	–	–	(0.018)	(0.012)	(0.004)	(0.960)
Foreign Network Size (80-88)	–	–	–	–	0.001^{***}	0.0001	-0.015^{**}	-0.062
	–	–	–	–	(0.0003)	(0.003)	(0.006)	(0.701)
Controls	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Most frequent language of publication FE	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Paper Quantiles FE	No	Yes	Yes	Yes	No	Yes	Yes	Yes
General Field FE	No	No	Yes	Yes	No	No	Yes	Yes
Home Country FE	Yes	No	No	No	Yes	No	No	No
Home-Destination Country Pair FE	No	No	Yes	Yes	No	No	Yes	Yes
FieldxHomexDestination FE	No	Yes	No	No	No	Yes	No	No
R^2	0.002	0.626	–	–	0.610	0.697	–	–
Adjusted R^2	0.002	0.621	–	–	0.609	0.693	–	–
Residual Std. Error	0.166 (df = 29912)	0.103 (df=29496)	–	–	0.104(df=29860)	0.092 (df=29495)	–	–
Akaike Inf. Crit.	–	–	5,566.566	5,566.566	–	–	1,784.059	1,784.059

Note: Clustered robust standard errors in (1) and (5) by home-country, in (2) and (6) by field-home-destination, and in (3), (4), (7), (8) by home-destination.
Total N=29,914, Mig-Out=509, Mig-EE=346. Controls include high quality, age proxies, and previous migration
*p<0.1; **p<0.05; ***p<0.01

6.2 Network Quality

Table 3 provides the results from OLS and MNL estimation for the effect of network quality, from 1980-1988, on the probability to migrate post 1988 and up to and including 2003. The general conclusion is that the results are explained by the signalling channel, in which, a higher network quality, even at home, increases the probability to migrate, however the effects are mostly economically insignificant. This implies that size or quantity plays a more important role, which could be specific to this context as the academics from EE were completely isolated and thus the quantity mattered more than the quality.

Only average citations and not rank used for destination and foreign networks because not all academics have them and thus the rank would be equal to 0, thus misleading. For consistency, I report the results from home average citations too, which are very comparable (with opposite signs) to the the estimated coefficients on average home rank. Looking at home network quality, the results show that the signalling channel marginally outweigh the cost channel, as the coefficients are positive, yet, economically insignificant, and in the most specified equation, insignificant. This means that a higher network quality at home increases probability of migration thus it serves as a signal of quality more than a measure of cost for left connections back home. An increase in home network quality through an increase in average citation, increases the probability to migrate (0.0003pp, 0.00004pp), with the latter being statistically insignificant, and both being economically insignificant. The direction of effects are confirmed by the estimated logit coefficient, not reported here for redundancy. This economic insignificance is translated to results from MNL regressions which also confirm the signalling channel outweighs the cost channel. Looking at columns 3 and 4 we find that an increase in home network quality, decreases the chances of not migrating versus migrating within EE, yet the effect is not very economically significant as the logit coeff will only decrease by 0.0001. Similarly, an increase in home network quality increases the chances of migrating outside of EE versus migrating within the EE. As expected, the economic insignificance translates into RRR of 1.

Turning to destination and foreign network quality also implies that higher network qualities signal a higher quality for the academic himself and thus increases the probability to migrate. However, as we see from table 3, the effect is not economically significant, and sometimes statistically insignificant. This highlights that size of the network matters more than the quality, and could be explained by the special context we are in such that academics from EE were segregated from the rest of the academic community such that any additional connection would be of great help and would increase migration prospects, irrespective of the quality of that connection. Additionally, even though it seems that the effect is through the signalling channel, the MNL results from foreign network quality offers other insights.
Distinguishing between the types of migration, an increase in foreign network quality increases the chances of an academic migrating within EE versus migrating out of EE as the risk or odds are 0.1% lower. This highlights that a greater foreign network quality, which as mentioned before, is assumed to be usually in other EE countries, has an effect through the cost channel as leaving the region completely means loss of these foreign connections completely, thus, this acts more as a pull factor. Thus, foreign network quality operates through both channels, it depends if we distinguish between the types of migration or not. Additionally, evidence from MNL results confirms that the effect of destination and foreign network quality is economically insignificant. The economic insignificance and sometimes the statistical insignificance imply that the size of the network matters more than the quality of the network. This aligns with the fact that academics from EE were highly segregated from the rest of the academic community such that any additional connection would be of great help and would increase migration prospects, irrespective of the quality of that connection.

I turn to another measure of quality that is specific to the home network. An academic network is considered *high quality* if average rank of the network and citation counts belong to the top 25 percent (bottom for rank given it's definition) in the academic's field of study and in the region of Eastern Europe. A more restrictive measure is that a network is considered high quality if it belongs to the top 10 percent in citation counts (and bottom for rank given its definition) in the academic's field of study and in the region of EE. Since this quality measure is dependent on the field, I include it as part of the following heterogeneity analysis.

Table 3: Effects of Home, Destination and Foreign Network Size on Probability to Migrate Post 1988 up to 2003.

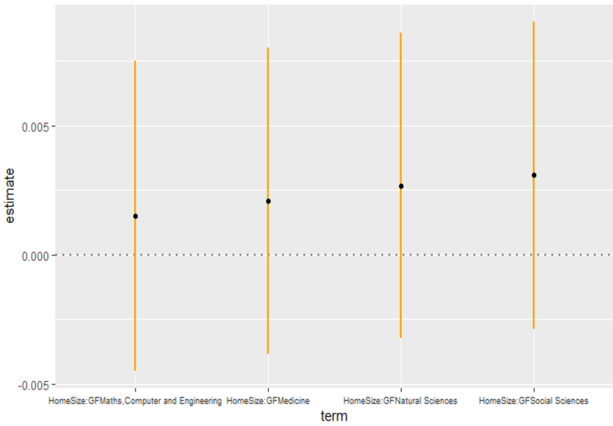
	Dependent variable:							
	Migrated 1989-2003							
	Felm		*MNL*		*Felm*		*MNL*	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
			Didn't Migrate	Migrated outside of EE			Didn't Migrate	Migrated outside of EE
Avg Home Network Citations (80-88)	0.000003***	0.0000004	−0.0001***	0.0001***	–	–	–	–
	(0.00000)	(0.00000)	(0.00001)	(0.00001)	–	–	–	–
Avg Destination Network Citations (80-88)	–	–	–	–	0.000003	0.000002	−0.036***	0.00000
	–	–	–	–	(0.00000)	(0.00000)	(0.005)	(0.001)
Avg Foreign Network Citations (80-88)	–	–	–	–	0.000001***	0.0000001	0.00000	−0.001***
	–	–	–	–	(0.00000)	(0.00000)	(0.00000)	(0.00004)
Controls	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Most frequent language of publication FE	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Paper Quantiles FE	No	Yes	Yes	Yes	No	Yes	Yes	Yes
General Field FE	No	No	Yes	Yes	No	No	Yes	Yes
Home Country FE	Yes	No	No	No	Yes	No	No	No
Home-Destination Country Pair FE	No	No	Yes	Yes	No	No	Yes	Yes
FieldxHomexDestination FE	No	Yes	No	No	No	Yes	No	No
R^2	0.039	0.628	–	–	0.540	0.637	–	–
Adjusted R^2	0.038	0.625	–	–	0.539	0.632	–	–
Residual Std. Error	0.163 (df = 29895)	0.102 (df=29664)	–	–	0.113(df=29860)	0.101 (df=29595)	–	–
Akaike Inf. Crit.	–	–	5,688.229	5,688.229	–	–	1,992.501	1,992.051

Note: Clustered robust standard errors in (1) and (5) by home-country, in (2) and (6) by field-home-destination, and in (3), (4), (7), (8) by home-destination.
Total N=29,914, Mig-Out=509, Mig-EE=346. Controls include high quality, age proxies, and previous migration
*p<0.1; **p<0.05; ***p<0.01

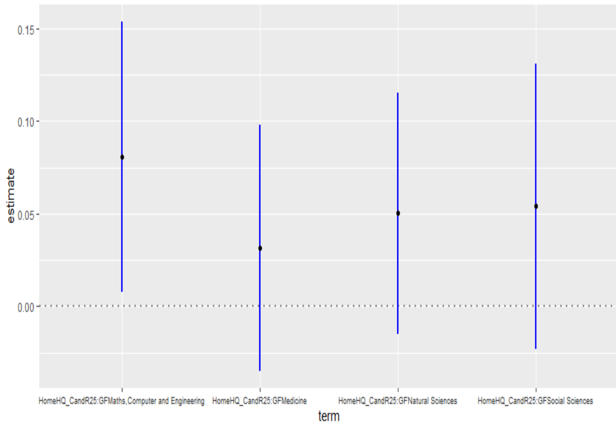
6.3 Heterogeneous Effects

It is important to look at specific characteristics of the network and the potential heterogeneous effects that might be concealed by the estimated results presented above. This is because such heterogeneous effects may show certain characteristics that either facilitate or de-facilitates migration. In this section I look at the interesting results from looking at potential heterogeneity driven by broad disciplines or general fields. This includes natural sciences, arts and humanities, social sciences, medicine and mathematics and computer science. This section summarizes and interprets the most important results and implications which lie under two key themes or conclusions. The first set of results is explained by the role of reputation, network and language barriers whilst the second set of results is explained by the threat of attention from KGB and arrest.

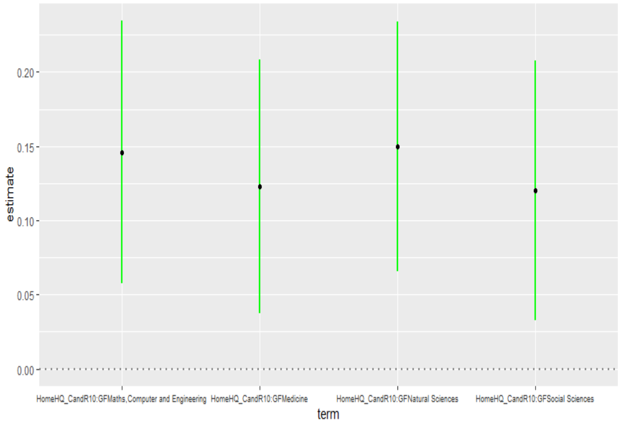
From the regression of interacting broad disciplines with home network size, and all else is the same there appears to be no heterogeneous effects on probability to migrate. This is apparent from figure (a) in 14 Looking at the same regression but for home network quality, first defined less strictly, i.e. high quality home network if the network belongs to top 25% of citations and (bottom 25%) of ranks in the field, and in the region, I find that the only significant interaction term is that for Maths, Computer and Engineering highlighting that a better home network quality increases the probability to migrate for engineers, computer scientists and mathematicians more than other academics, the relevant coefficients are depicted in figure (b) 14. Looking at the more strict definition, i.e. home network considered top quality if belonging to the top 10% in citations and (bottom) 10% of rank, , figure (c) in 14, all the relevant estimated coefficients of interactions are significant. This implies that as network quality improves more, the more it facilitates the migration of academics as it gives a stronger signalling effect. This asymmetry between figure b and c and the coefficient being only significant for mathematicians, computer scientists and engineers in figure b could potentially be explained by the fact that German and Soviet mathematicians and engineers were widely recognized as world-leading and thus a network belonging to top 25 in the region-field has a greater effect for this specific broad discipline whereas other disciplines' home network would need to belong to top 10 instead. Thus, this implies that the *signalling channel* that comes from a more qualified more reputable home network, is stronger for Mathematicians and Engineers. This evidence aligns with Borjas and Doran (2012) which states that Soviet Mathematicians were of top tier.



(a) Home Network Size Heterogeneity



(b) Home Network Quality (Top 25) Heterogeneity



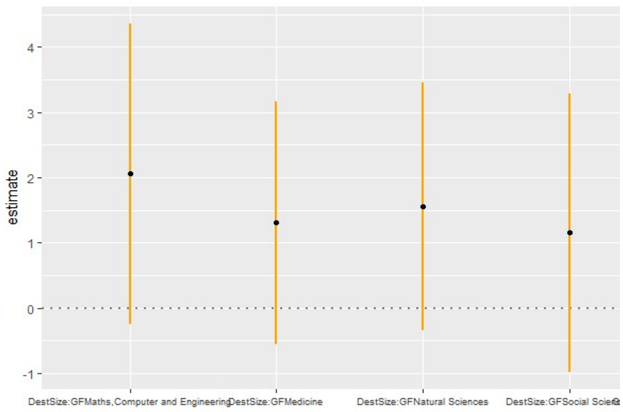
(c) Home Network Quality (Top 10) Heterogeneity

Figure 14: Home Network Characteristics Heterogeneous Effects: Estimated(OLS) Coefficients of interactions of characteristic with General Field

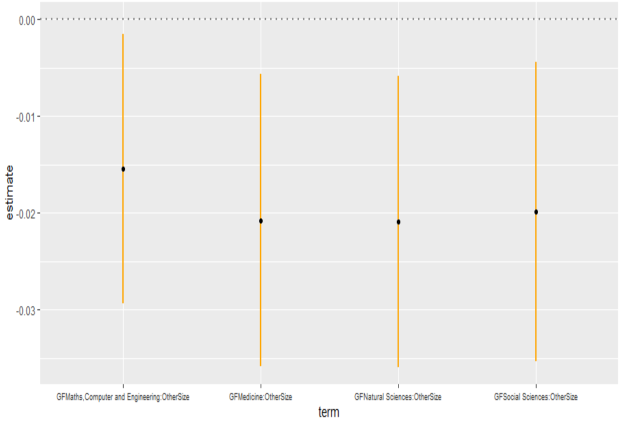
Looking at heterogeneous effects of foreign network size through broad disciplines on probability to migrate also lies under this set of results that imply and emphasize the role of reputation and network and language barriers. Note that in figure (b) in 15 Arts and Humanities broad discipline is not included due to the issue of perfect multicollinearity and that not all factors be included in the regression. However, it is easy to see the estimated coefficient if an academic belongs to the broad discipline of Arts and Humanities, which is simply the coefficient on Foreign Network Size, which is reported to be (2.1pp). This estimated coefficient is very positive and significant as compared to the other estimated coefficients for other fields. Fields with larger network barriers and less quality signalling, their foreign network, plays a more important role in facilitating migration. This aligns with Becker et al (2021) which find similar results and argue that academics from fields were there are large network barriers and less quality signalling (vs Mathematicians and Scientists) then networks play a bigger role in increase the probability to migrate. Thus, all of this evidence points towards the role of reputation and network and language barriers.

There is also heterogeneous effects of destination network size that is significant, statistically (at the 10%) and economically, as the value is 20.5pp as figure (c) in 15 shows, only for Mathematicians, Computer

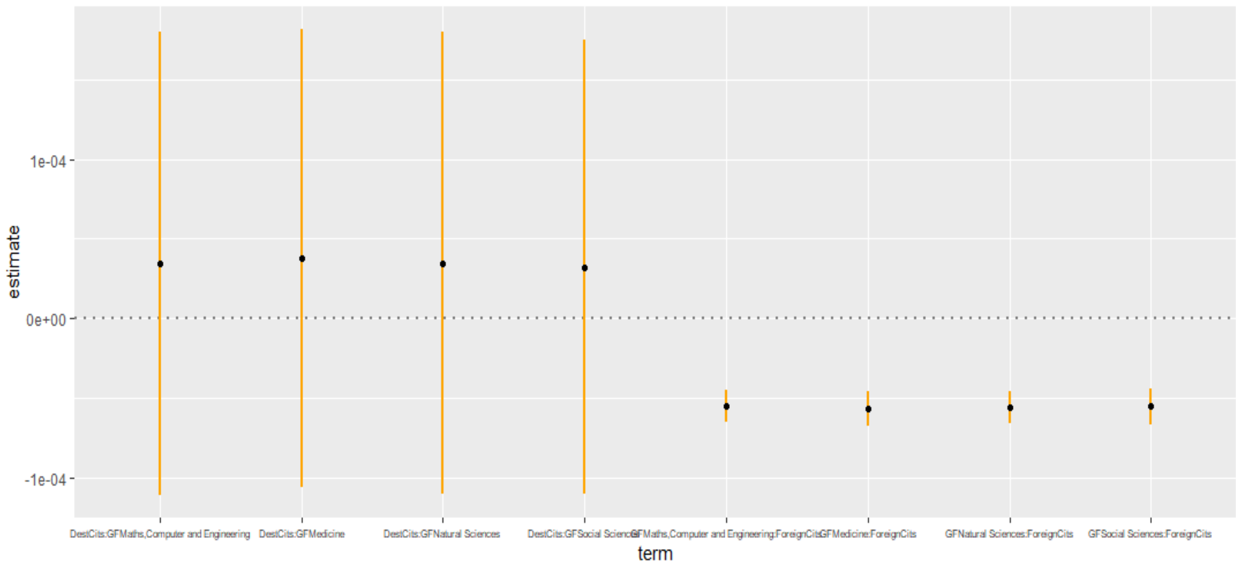
Scientists and Engineers. Note that figure 15 shows confidence intervals for the 95% level. This result is specifically explained by the threat of attention from KGB and arrest, which was specifically relevant for Mathematicians more than any other academics. This also is supported by Borjas and Doran (2012) that study the productivity of American mathematicians post the influx of Mathematicians from the Soviet Union post its fall. They state that there was almost no interaction between mathematicians in the US and the Soviet Union due to the fact that anyone who tried to communicate with a scholar in the US faced the potential risk of attention from the KGB and arrest (Borjas and Doran, 2012). In fact, they argue that this might be specific to this field and that they are not aware of any fields that might have been similarly affected by the Soviet Union's dismantling and they also noted the power and strength of Soviet Mathematicians. Thus, this specific context and particularly since the US is the most favoured destination in this sample, this is very intuitive. Since mathematicians by construct nearly had no interaction with the US, an additional connection at their potential destination has a significantly positive effect on their probability to migrate there, especially since they were of high quality and reputation.



(a) Destination Network Size Heterogeneity



(b) Foreign Network Size Heterogeneity



(c) Destination and Foreign Network Quality Heterogeneity

Figure 15: Destination and Foreign Network Size and Quality: Heterogeneous Effects by Broad Discipline

are we including this Looking at figures (d) and (e) and columns 4 and 5, we are investigating the joint effect of network size and quality. The effects are negative and significant implying that the cost channel outweighs the signalling channel, thus the negative signs.

7 Conclusion

Iron Curtain and Big Data are two words usually used to denote completely different eras. However, using a scholarly big data source, Microsoft Academic Knowledge Graph, I am able to study the effect of academic networks on migration post the fall of the Iron Curtain. Using this data, I am able to identify the academics from the Eastern Bloc such that their affiliation is their location and nationality due to the fact the Iron Curtain halted migration outside of the Bloc. Additionally, due to the fact that the fall wasn't anticipated, I find no evidence of strategic manipulation of networks prior the fall of Berlin Wall 1989, and thus focusing on this context specifically, reduces tremendously the scope for reverse causality. The rich data source also mitigates the risk of omitted variable bias. Thus, I focus on academics from EE from 1980-1989 and their academic networks (1980-1989), and investigate the effect of their network characteristics, size and quality, by location, on the probability to migrate post the fall of the Berlin Wall 1989 and up to 2003, marking the year many EE countries held referendums or signed treaties to join the EU. I test the assumption that size and quality characteristics of the network go through two distinct channels, *cost* and *signalling.*

The evidence I find, as outlined concisely in the section of summary of theoretical predictions and results, offers some interesting insights. Regarding network size, the cost channel seem to explain the results, especially resolving the ambiguity of the effect of foreign network size, which could go through cost or signalling channels, when distinguishing between the types of migration. Regarding network quality, the signalling channel seems to explain the majority of the results, except for foreign network quality. This is driven by the fact that the assumption of foreign connections being close in terms of distance, is confirmed, by the data and by the results. As such, distinguishing between types of migration, a higher foreign network quality, decreases the chances of migrating out of EE with respect to migrating within EE. I also find evidence that size of network matters more than quality, which appears to be a specificity of the context we are in. Lastly, heterogeneous effects of networks by broad disciplines can be explain by two things: threat of attention from KGB and arrest, and the role of reputation, language and network barriers. In conclusion, this research abstracts from certain limitations inherit in the empirical investigation of the relationship between networks and migration by focusing on a specific context and using novel data sources. Even though the question of using big data in migration is still a research question in itself, I think this paper contributes to various strands of literature as mentioned previously.

8 Improvements and Robustness

As mentioned previously, one of the limitations of the extraction of networks is that there is a high share of unlocated academics. This has been extracted by the UNIPI team. We have managed to find a way to decrease this number by detailing the code more (finer units than country) so as to minimize this. There is a new version of MAKG, which has less abnormal values as mentioned previously. This overcomes data contamination issues and having extremely heavy tails. If this still persists, potentially quantile regression could be used as one of the main motivation behind using it is for robustness, especially against outliers and heavy tails, which is the context here. I am investigating this further. Additionally, the new version offers more comprehensive information on journals, fields and academics. It's still under investigation. With the newest version, I am able to spot more migrants post 1989.

I am also trying to define *destination* network differently as this definition limits the interpretation of results. This is because, a non-migrant, by definition, has no destination network, as explained previously.

This might be problematic as it is not really clear what should be considered as the destination network. This is mitigated above by using *foreign network*. However, I can try to overcoming this in a different way. One way is to define an observation at the level of the individual-potential destination. As such, pre 1989, each individual has a specific number of observations (depending on the destination countries in the sample, or even equal to the number of countries in the whole world), one for each potential destination that the individual could migrate to. These observations for each individual are not i.i.d, and as such the multinomial logit model would come in handy in this context. This is because it treats the decision to migrate as a single decision with x+1 alternatives (staying at home and the x alternative options). This might have some limitations to the extent of fixed effects that could be included, so a trade-off exists.

Another work to be done is improving the definition of the quality of the network as the interpretation of the one used here is not very straightforward. I could turn to the median, which is much more robust to outliers, or create a specific index. I could also perform robustness check through the second and third most frequent language of publication as mentioned previously. I could also check the identification evidence presented above by fields of study and see if the argument still holds.

An additional aspect that I want to investigate, is the functioning and productivity of Eastern European academics ever since the 1950s up to 1989 marking the fall of Berlin Wall. To do so, I have extracted all academics from these countries and currently cleaning the data. Additionally, Professor Hillel and I wanted to look at Eastern Germany on its own, specifically if the reverse causality concern is still not an issue in that context. It would also be interesting to investigate certain research questions by focusing on Eastern Germany, potentially comparing the academics' productivity with their Western German academics.

9 Bibliography

Becker, S.O., Lindenthal, V., Mukand, S. and Waldinger, F., 2021. Persecution and Escape: Professional Networks and High-Skilled Emigration from Nazi Germany.

Beine, M., Docquier, F. and Özden, Ç., 2011. Diasporas. Journal of Development Economics, 95(1), pp.30-41.

Beine, M., Docquier, F. and Özden, Ç., 2015. Dissecting network externalities in international migration. Journal of Demographic Economics, 81(4), pp.379-408.

Belot, M. and Ederveen, S., 2012. Cultural barriers in migration between OECD countries. Journal of Population Economics, 25(3), pp.1077-1105.

Bertoli, S., 2010. Networks, sorting and self-selection of Ecuadorian migrants. Annals of Economics and Statistics/Annales d'Économie et de Statistique, pp.261-288.

Bertoli, S. and Ruyssen, I., 2018. Networks and migrants' intended destination. Journal of Economic Geography, 18(4), pp.705-728.

Bertoli, S., Moraga, J.F.H. and Ortega, F., 2013. Crossing the border: Self-selection, earnings and individual migration decisions. Journal of Development Economics, 101, pp.75-91.

Bircan, T., 2020. "Gaps in Migration Research. Review of migration theories and the quality and compatibility of migration data on the national and international level". (Deliverable n°2.1). Leuven: HumMingBird project 870661 H2020

Blumenstock, J.E., Chi, G. and Tan, X., 2019. Migration and the value of social networks.

Bonacich, P., 1991. Simultaneous group and individual centralities. Social networks, 13(2), pp.155-168.

Borjas, G.J. and Doran, K.B., 2012. The collapse of the Soviet Union and the productivity of American mathematicians. The Quarterly Journal of Economics, 127(3), pp.1143-1203.

Breza, E., Chandrasekhar, A.G., McCormick, T.H. and Pan, M., 2020. Using aggregated relational data to feasibly identify network structure without network data. American Economic Review, 110(8), pp.2454-84.

Cameron, A.C., Gelbach, J.B. and Miller, D.L., 2011. Robust inference with multiway clustering. Journal of Business Economic Statistics, 29(2), pp.238-249.

Cameron, A.C. and Miller, D.L., 2015. A practitioner's guide to cluster-robust inference. Journal of human resources, 50(2), pp.317-372.

Carrington, W.J., Detragiache, E. and Vishwanath, T., 1996. Migration with endogenous moving costs. The American Economic Review, pp.909-930.

Cheng, S. and Long, J.S., 2007. Testing for IIA in the multinomial logit model. Sociological methods research, 35(4), pp.583-600.

Comola, M. and Mendola, M., 2015. Formation of migrant networks. The Scandinavian Journal of Economics, 117(2), pp.592-618.

De la Croix, D., Docquier, F., Fabre, A. and Stelter, R., 2020. The Academic Market and the Rise of Universities in Medieval and Early Modern Europe (1000-1800).

Docquier, F. and Rapoport, H., 2012. Globalization, brain drain, and development. Journal of Economic Literature, 50(3), pp.681-730.

Giulietti, C., Wahba, J. and Zenou, Y., 2018. Strong versus weak ties in migration. European Economic Review, 104, pp.111-137.

Gould, E.D. and Moav, O., 2016. Does high inequality attract high skilled immigrants?. The Economic Journal, 126(593), pp.1055-1091.

Hanck, C., Arnold, M., Gerber, A. and Schmelzer, M., 2019. Introduction to Econometrics with R. University of Duisburg-Essen.

Hausman, J. and McFadden, D., 1984. Specification tests for the multinomial logit model, Econometrica (52, 5).

Hojman, D.A. and Szeidl, A., 2008. Core and periphery in networks. Journal of Economic Theory, 139(1), pp.295-309.

Howe, Marvine, "For Emigre Scientists, Job Hunting Is Difficult," New York Times, August 12, 1990.

Jackson, M.O., 2020. A typology of social capital and associated network measures. Social Choice and Welfare, 54(2), pp.311-336.

Jackson, M.O. and Watts, A., 2002. The evolution of social and economic networks. Journal of Economic Theory, 106(2), pp.265-295.

Jackson, M.O., 2010. Social and economic networks. Princeton university press.

James, C., Pappalardo, L., Sîrbu, A. and Simini, F., 2018. Prediction of next career moves from scientific profiles. arXiv preprint arXiv:1802.04830.

Laqueur, Walter. The Dream That Failed: Reflections on the Soviet Union. (New York: Oxford University Press, 1996).

McKenzie, D. and Rapoport, H., 2007. Network effects and the dynamics of migration and inequality: Theory and evidence from Mexico. Journal of development Economics, 84(1), pp.1-24.

McKenzie, D. and Rapoport, H., 2010. Self-selection patterns in Mexico-US migration: the role of migration networks. the Review of Economics and Statistics, 92(4), pp.811-821.

Menard, S., 2010. Logistic regression: From introductory to advanced concepts and applications. Sage.

Patel, K. and Vella, F., 2013. Immigrant networks and their implications for occupational choice and wages. Review of Economics and Statistics, 95(4), pp.1249-1277.

Polyak, B. T., “History of Mathematical Programming in the USSR: Analyzing the Phenomenon,” Mathematical Programming, 91 (2002), 401–416.

Roberts, M 2013, Robust and Clustered Standard Errors, lecture notes, Harvard University, delivered 6 March 2013.

Sîrbu, A., Andrienko, G., Andrienko, N., Boldrini, C., Conti, M., Giannotti, F., Guidotti, R., Bertoli, S., Kim, J., Muntean, C.I. and Pappalardo, L., 2020. Human migration: the big data perspective. International Journal of Data Science and Analytics.

Van Mol, C. and De Valk, H., 2016. Migration and immigrants in Europe: A historical and demographic perspective. In Integration processes and policies in Europe (pp. 31-55). Springer, Cham.

Varian, H, 2014. ”Big Data: New Tricks for Econometrics,” The Journal of Economic Perspectives, 28(2), 3-27.

Vijverberg, W.P., 2011. Testing for IIA with the Hausman-McFadden test (No. 5826). IZA Discussion Papers.

White, R. and Buehler, D., 2018. A closer look at the determinants of international migration: decomposing cultural distance. Applied Economics, 50(33), pp.3575-3595.

Wooldridge, J.M., 2002. Econometric analysis of cross section and panel data MIT press. Cambridge, MA, 108.

Wooldridge, J.M., 2003. Cluster-sample methods in applied econometrics. American Economic Review, 93(2), pp.133-138.

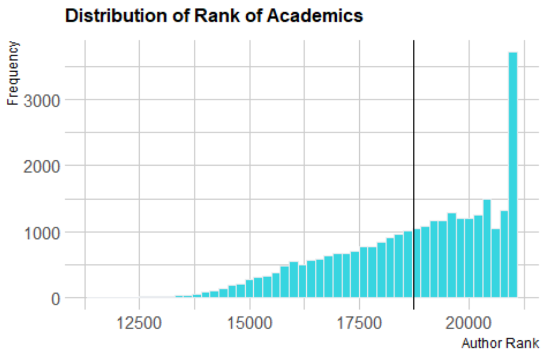
10 Appendix

Figure 16: Map of Germany Pre 1990: East and West Germany

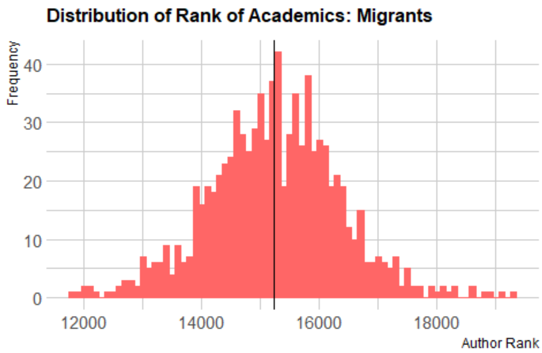
Figure 17: Map of Eastern Germany Pre 1990

Figure 18: Author Quality Distributions

(a) Rank: All of the Sample



(b) Rank: Migrants



(c) Rank: Non-Migrants

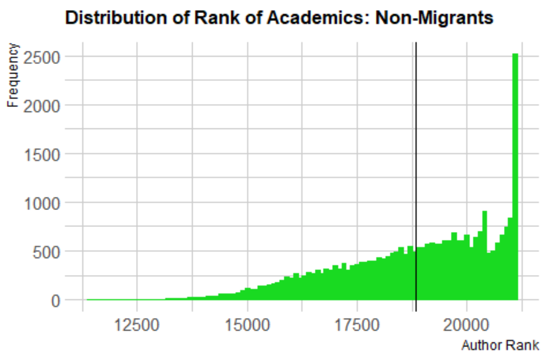
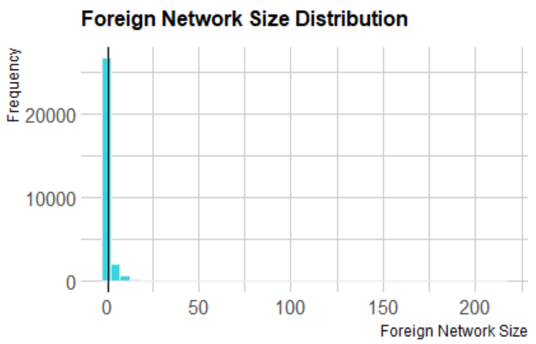
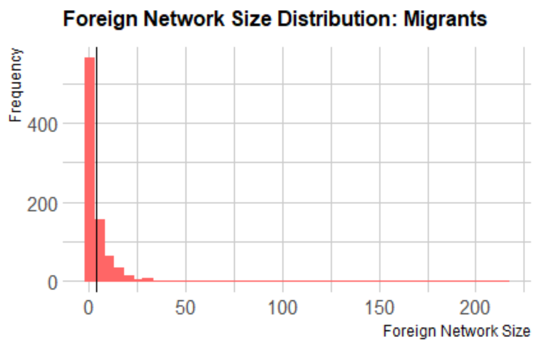


Figure 19: Network Size Distributions

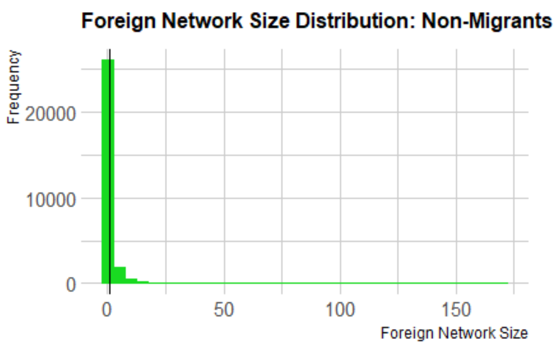
(a) Foreign Network Size: All of the Sample



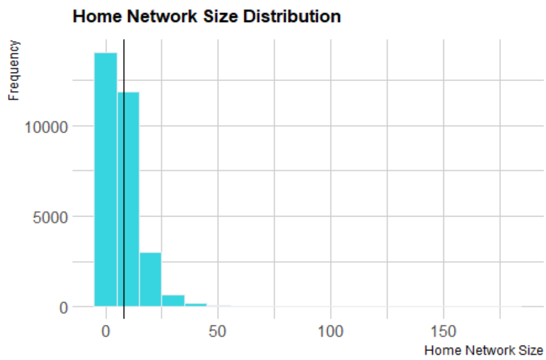
(b) Foreign Network Size: Migrants



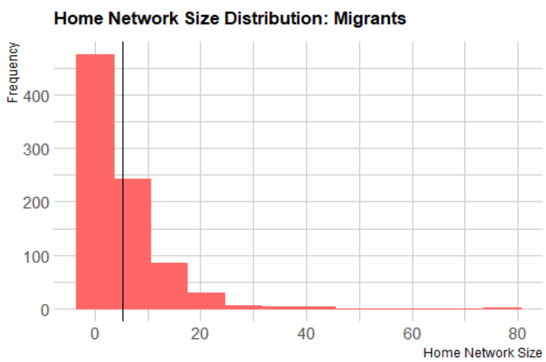
(c) Foreign Network Size: Non-Migrants



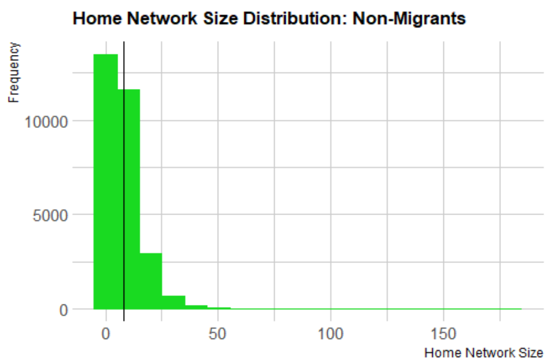
(d) Home Network Size: All of the Sample



(e) Home Network Size: Migrants



(f) Home Network Size: Non-Migrants



(g) Destination Network Size: Migrants Only

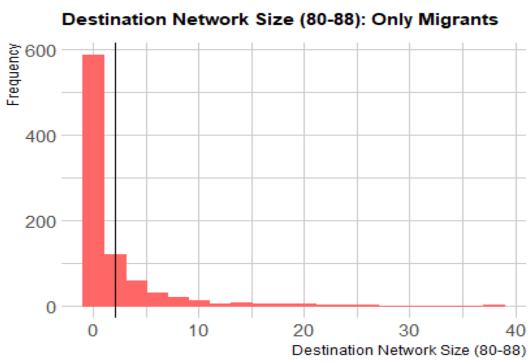
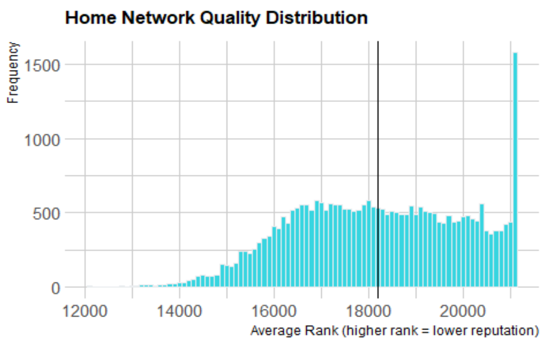
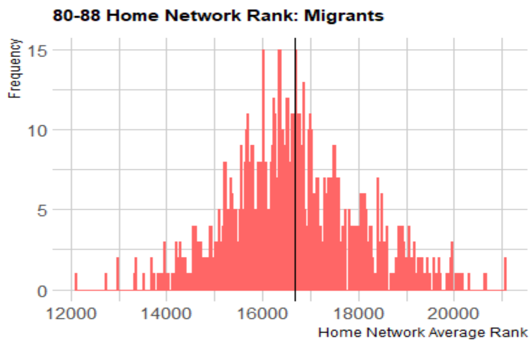


Figure 20: Network Quality Distributions

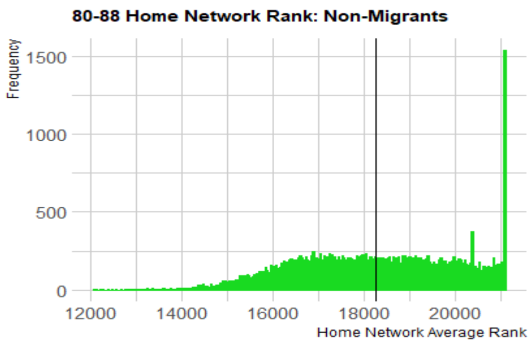
(a) Home Network Quality: All of the Sample



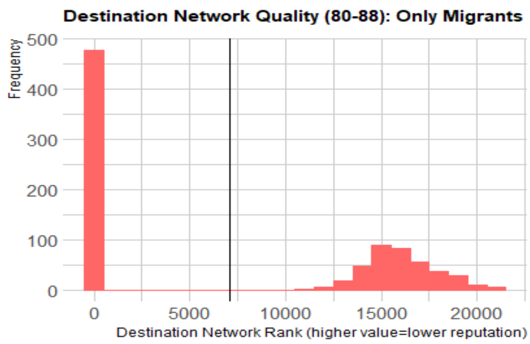
(b) Home Network Quality: Migrants



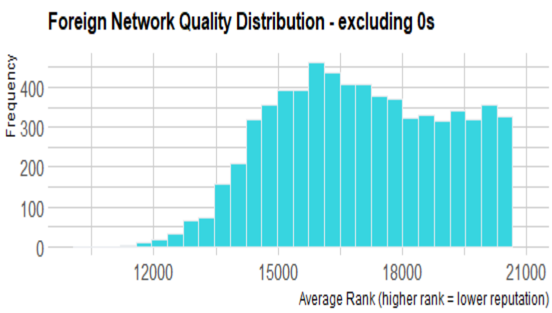
(c) Home Network Quality: Non-Migrants



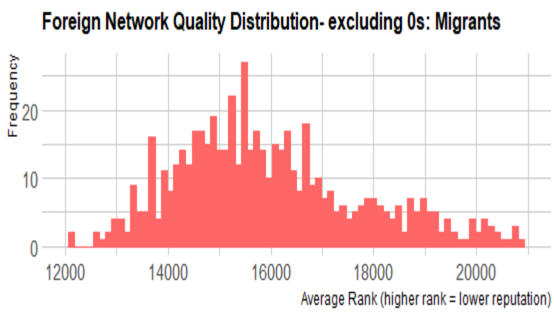
(d) Destination Network Quality: Migrants Only



(e) Foreign Network Quality: All of the Sample



(f) Foreign Network Quality: Migrants



(g) Foreign Network Quality: Non-Migrants

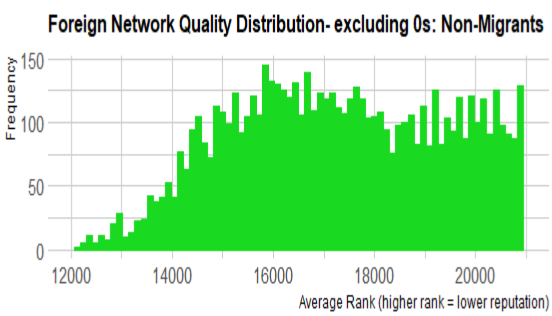
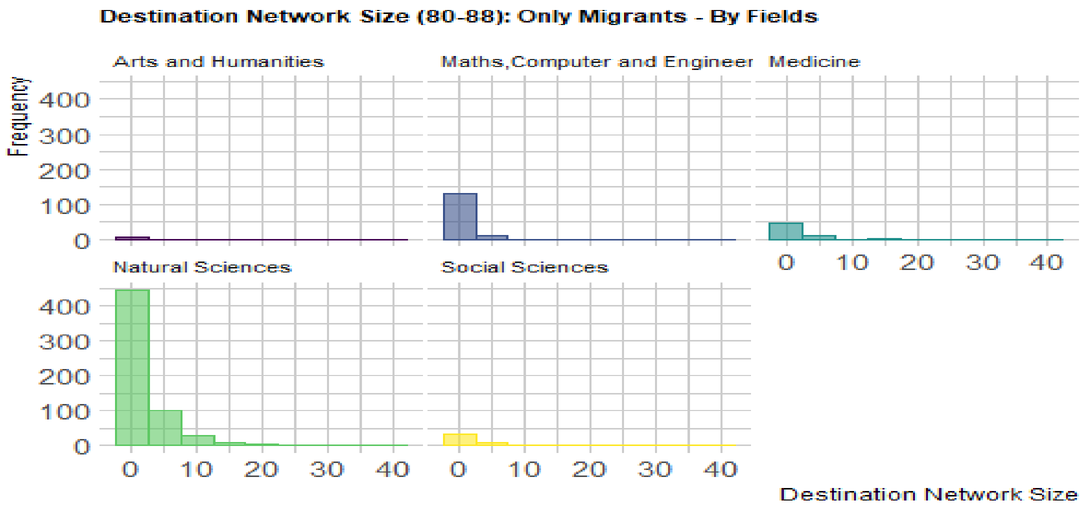


Figure 21: Destination Network Size and Quality by Broad Discipline: Migrants Only

(a) Size



(b) Quality

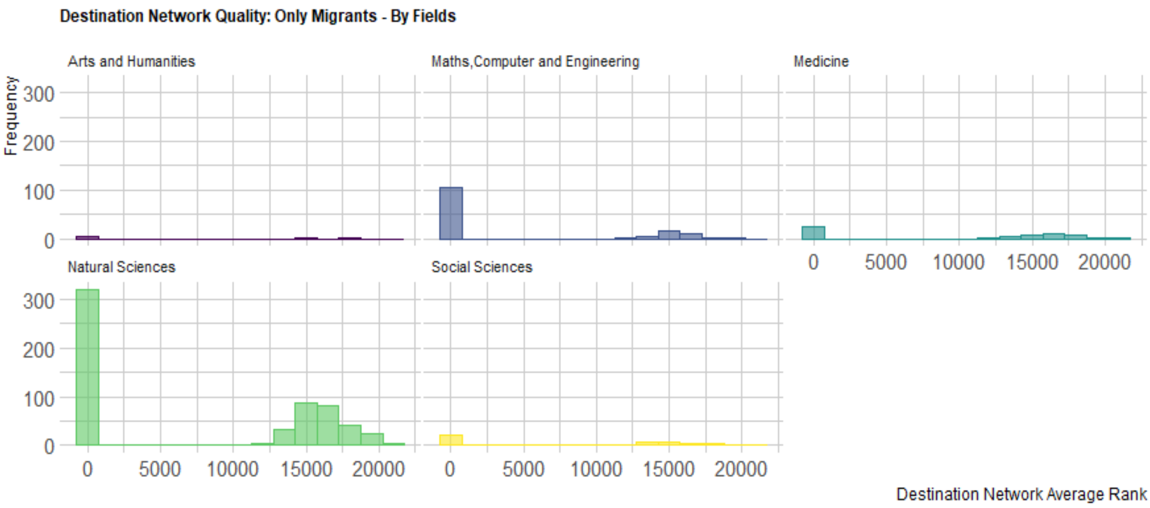
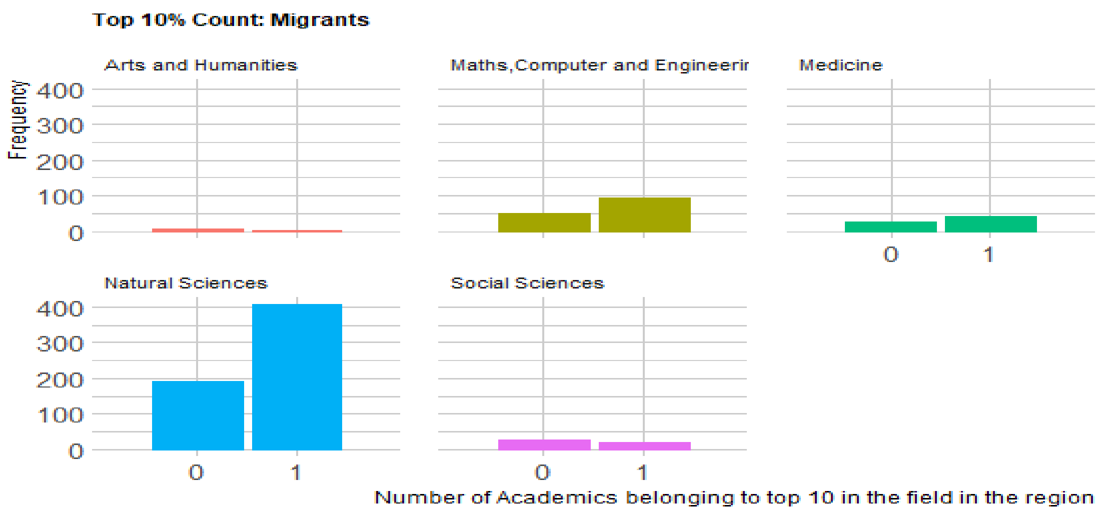
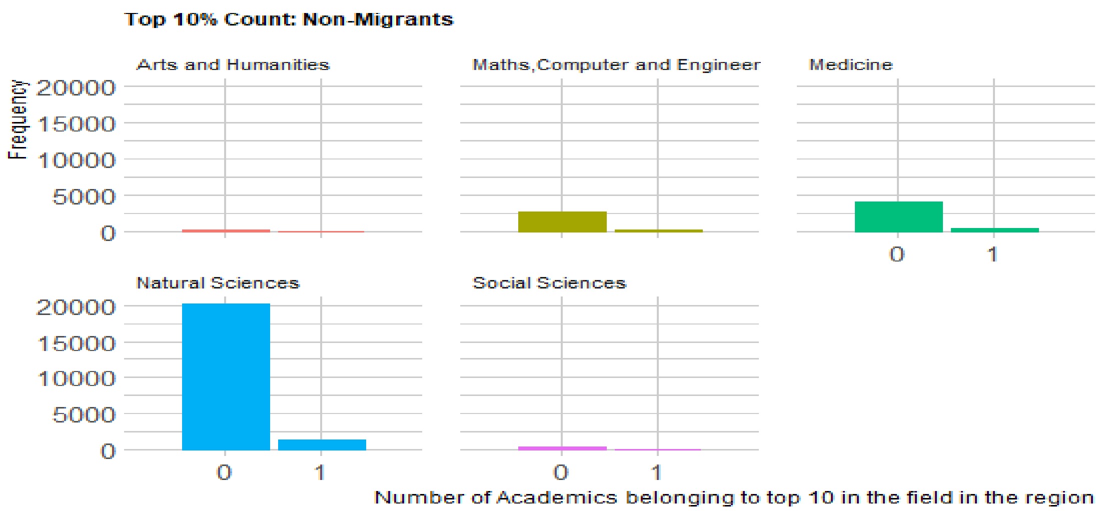


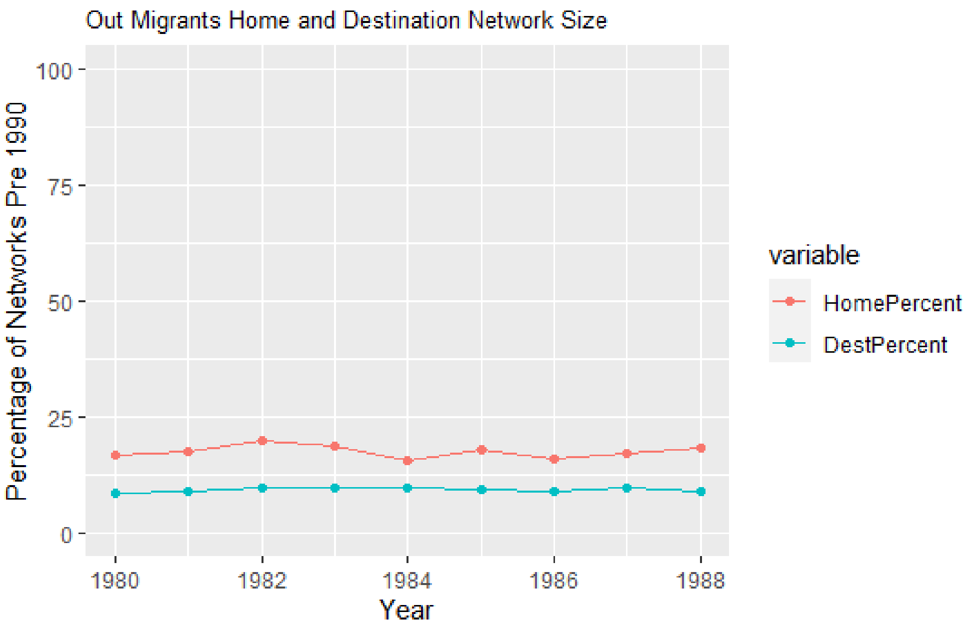
Figure 22: Top 10% Academics by Broad Disciplines

(a) Migrants

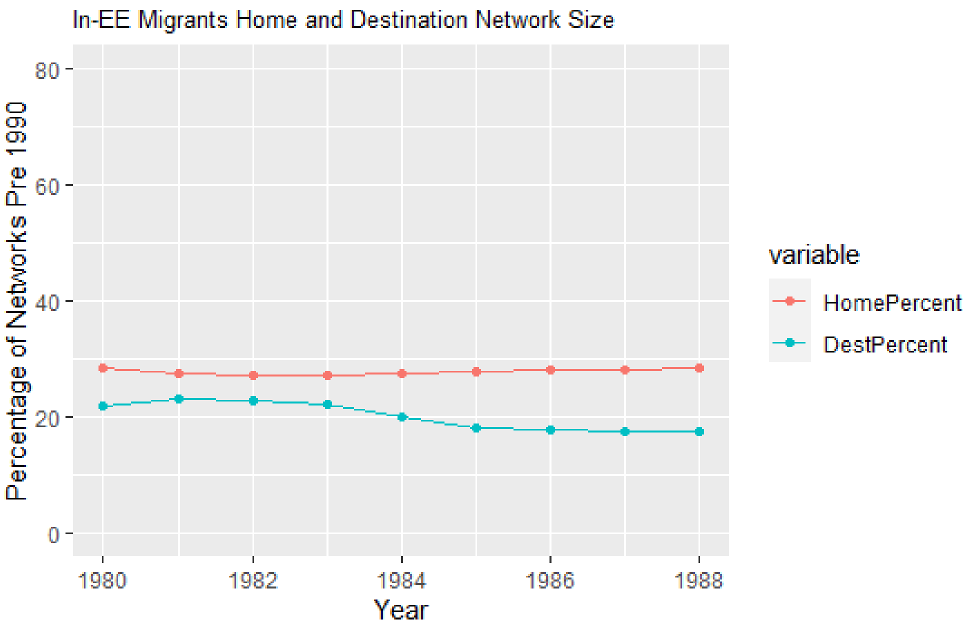


(b) Non-Migrants





(a) Out Migrants Home and Destination Network Shares



(b) Out Migrants Home and Destination Network Shares

Figure 23: Further Identification Evidence: Out and In-EE Migrants Network Shares