An Empirical Evaluation of the Impact of Regional Diasporas in North Rhine-Westphalia on Future Migration Patterns

Mémoire réalisé par
Yannik Schenk

Promoteurs
Frédéric Docquier
Florian Mayneris

Lecteur
Gani Aldashev

Année académique 2016 - 2017

En vue de l'obtention du titre académique de
Master en sciences économiques, finalité spécialisée

Ecole d'économie de Louvain/UCL • Place Montesquieu 3 • 1348 Louvain-la-Neuve
Département des Sciences économiques/UNamur • Rempart de la Vierge 8 • 5000 Namur

Avant-Propos

First of all, I would like to thank my two promoters Frédéric Docquier and Florian Mayneris for accompanying this thesis with their expertise, many valuable ideas and patience. Moreover, I would like to thank Gani Aldashev, who pointed me more than once in the right direction while I was still in search of a topic that truly interests me for this thesis. Last but defenitly not least, I would like to thank my parents and friends for their unconditional support through the last month and years. Especially, I want to express my honest gratitude to João and Jan, without whom this work clearly would not have been possible.

An Empirical Evaluation of the Impact of Regional Diasporas in North Rhine-Westphalia on Future Migration Patterns

Master Thesis

Submitted to:

Supervisor: Prof. Frédéric Docquier
Co-Supervisor: Prof. Florian Mayneris
Reader: Prof. Gani Aldashev

Economics School of Louvain
Université Catholique de Louvain

Louvain-La-Neuve, August 25, 2017

Submitted by:

Yannik Schenk
Student ID: 72621500
Degree Program: Master en sciences économiques, finalité spécialisée
Université Catholique de Louvain

Address: Parvis Saint-Gilles 24, 1060, Saint-Gilles, Belgique
E-Mail: yannik.schenk@student.uclouvain.be

Contents

List of Tables

List of Figures

1 Introduction

A considerable stream of empirical and theoretical work has investigated the impact of regional networks of migrants (or diasporas) on future migrations streams. Understanding the role of the diasporas for migration patterns is key for many policy questions. On the one hand, diasporas have the potential to strongly shape future migration patterns (henceforth referred to as network effect). On the other hand, they are related to many questions regarding integration and assimilation of migrants within a country. Especially in the context of the recent refugee crisis, those questions clearly regained importance: according to numbers of the German Federal Office for Statistics (DESTATIS), Germany accommodated around 1.2 million asylum seekers between 2012 and 2015. Over the next years, the management of this quickly growing stock of migrants will become a key challenge for the German society, economy and political leaders alike. To address this issue properly, it is indispensable to gather deeper understanding of related dynamics. The goal of the present work is to contribute to this understanding.

Explicitly, I shed some light on the questions regarding how regional migrant networks can be expected to influence future migration flows. I use official German government data on the stocks of foreigners of specific nationalities in the 53 administrative municipalities in North Rhine-Westphalia (NRW) over ten years to construct yearly migration flows into municipalities by nationality (origin). Then I estimate the effect of regional diasporas on these flows by means of ordinary least squares (OLS) and Pseudo-Poisson maximum likelihood (PPML) estimation, using a full panel specification with multiple origins, multiple destinations and several years. This allows me to control carefully for unobserved factors and decreases the risk of misspecification and endogeneity. This reflects an improvement regarding existing literature which mainly uses cross-sectional data (Beine et al. (2011); Grogger and Hanson (2011)) or data including only one destination (Clark et al. (2007)). As such, the method of this work is most closely related to the approach of Ortega and Peri (2013), who use a similar panel specification to estimate the impact of migration policies. However, they do not explicitly control for regional networks, which is the main focus of this work.

The contribution of this thesis is threefold: Firstly, I present different econometric approaches when dealing with imperfect data and econometric obstacles in the frame of gravity equations of migration. Secondly, the work exploits a large and highly disaggregated panel dataset of the recent development in German cities, which allows for the implementation of fully-fledged panel methods to estimate the network effect. Thirdly, I extend the basic framework by further factors to assess whether and how they impact upon the strength of the network effect: I analyze the difference in localization choice patterns between refugees and ”traditional” migrants. I analyze in how far the level of economic development at the origin country determines whether a migrant relies on a regional network in the destination city. Finally, I present evidence for the relative importance of the assimilation channel (direct support through local diasporas) and the policy channel (by qualifying for family reunion programs) as carrier for network effects.

The remainder of this thesis is structured in the following way: Section 2 recaps typical drivers of migration and discusses typical channels that underlie the network effect. Section

3 theoretically develops the econometric specification that is brought to the data, based on a Random Utility Maximization Model of migration. Section 4 describes the dataset used, its caveats and suited econometric strategies to deal with them. Section 5 presents baseline estimation strategy and results, and various robustness checks. Section 6 presents three different extensions of the baseline model. Section 7 concludes.

2 Determinants of Migration

Economic theory explains international migration, like its internal counterpart, by geographic differences in the supply and demand of labour. It generally conceives the migration movements as aggregation of individual income maximizing decisions. Countries with large endowment of labour, relative to capital, have low equilibrium market wages. On the other side, developed economies that are relatively capital abundant have higher equilibrium market wages. The resulting wage differentials trigger migration of workers from low to high wage economies (Harris and Todaro (1970); Lewis (1954) Ranis and Fei (1961); Michael P Todaro (1976)). The heterogeneity of workers, for example in terms of human capital, must be recognized in this framework. The return to human capital will be higher in a relatively human capital scarce environment. Hence the flow of human capital must not be aligned with the general flow of labour and people (Borjas (1987); Massey et al. (1993)).

Taking a more detailed lens, a microeconomic point of view regards migration as the utility maximizing choice of an individual across different destinations compared to the situation in the origin country, given her characteristics. Individuals build expectations about their future utility in different scenarios, and expectations related to the costs of migration. Higher wage and higher return to a specific form of human capital incentivize migration. On the other side are costs that the individual migrant would need to bear: the material costs of travelling, legal costs, the costs of maintenance while resettling and looking for work, the effort in adapting to a new culture (e. g. learning a language) and the psychological costs of leaving family, friends and well-known environment behind and engaging in the risky endeavour to create something new. If this cost-benefit calculation leads the individual to expect a positive net return to migration, she will migrate to the most favourable destination (Borjas (1990); Massey et al. (1993); Sjaastad (1962); Michael P. Todaro (1969)).

In their seminal work, Massey et al. (1993) draw a conceptual distinction between the conditions for the initiation of international migration and the perpetuation of international migration. While several ex-ante factors like wage differentials, job opportunities, general quality of life and different sorts of migration costs clearly play an important role in shaping migration flows, migration itself reshapes the conditions in the receiving country. Hence the conditions that initiate migration in the first place can differ from those that perpetuate migration over time and space. Migrant networks have received the highest attention from researchers as main driver of this process of cumulative causation (Massey (1990); Myrdal (1957)). Networks of migrants at the destination can alleviate migration related costs and risks. The first migrant who settles at a new destination has no network to draw support from. Subsequently, every migrant who settles at a destination is inevitably connected to potential future migrants due to kinship and friendship ties. Hence each migrant reduces migration costs of someone else, and induces a change in the cost-benefit calculation of a group of related non-migrants. In this sense, migration leads to higher migrant stocks and higher migrant stocks in turn promote further migration. The migration process can be self-accelerating, even when underlying push- and pull-factors remain unchanged (Beine et al. (2015)).

Beine et al. (2015) emphasize two different main channels through which migrant networks influence migration patterns. The policy channel is related to the overcoming of legal entry barriers imposed by the destination country. The EU Family Reunification Directive (2003/86/EC) has created an EU-wide framework for the family reunification of third-country nationals, or nationals of the relevant Member State, with third-country nationals. All national provisions concerning the family reunification of Germans or third-country nationals with other third-country nationals in Germany are set out in Sections 27 – 36 of the German Residence Act.

Under certain criteria, the presence of a relative can hence facilitate the administrative procedure, offset legal barriers and decrease policy-related migration costs. This has naturally a direct impact upon subsequent migration. The larger the number of people with ties to a third country, the more people are potential candidates to benefit from the family reunification program. Eligible under the German regulation are manly spouses, registered partners and first degree relatives, but also more distant relatives under an additional set of conditions. EU citizens generally benefit from freedom of movement as set out in the Act on the General Freedom of Movement for EU citizens and are not regulated by the Residence Act.

The second channel through which migrant networks shape migration patterns was termed assimilation channel. Much research has addressed the question how regional networks help new arriving migrants and promote further migration. According to Massey et al. (1993) network connections reflect some sort of social capital that can fulfill several functions for an arriving migrant. Giulietti et al. (2013)) phrase the value of an ethnic network for migrants in the following words: " *[...] networks are known to facilitate the economic and social integration into the host country by providing social support and useful information towards finding employment. In fact, social networks play a dual role as a channel for information and a substitute for specific skills needed in the host country – such as the language or the knowledge of institutions – which immigrants, and especially those who have recently arrived, might lack.*"

Various contributions have proven those network effects (Beine et al. (2011); Beine and Parsons (2015); Beine et al. (2011) Bertoli and Fernández-Huertas Moraga (2012) Munshi (2003)) and discussed operating channels. Munshi (2003) shows that social networks at the destination in the U.S are related to higher employment rates and higher paid jobs among Mexican migrants who share the same origin community. Further, the networks provide financial support and housing assistance for its members. Using survey data from Sri Lankan residents in Milan, Comola and Mendola (2015) find additional evidence for the assisting function of established networks for new-coming migrants. Based on administrative social media data, Mayer and Puller (2008) find that common race is amongst the strongest predictors for friendship between two college students, controlling for various other personal and background dimensions. This research shows that a common origin can facilitate the creation of new personal ties, for a multitude of reasons.

The creation of institutions and infrastructure is closely related to the assimilation channel. This can for example be in the form of an adequate religious infrastructure, enabling migrants to maintain their religious habits; the creation of firms and job opportunities that facilitates labour market entry of the newcomers; or other forms of ethnically and culturally relevant constructs. One can wonder whether those institutions or the migrant networks themselves are important for further migration. Naturally, any kind of ethnical institutions would not develop, and can hardly persist in the absence of regional diasporas. Hence the underlying definition of network effects should include a broad range of those factors, which are due to regional diasporas.

Summing up, many reasons have been discussed why migrants might be more likely to choose a destination where a diaspora of fellow countrymen already exists. The bigger the diaspora, the higher is the chance that a random immigrant has ex-ante personal ties to someone, which would then impact upon his decision to migrate to a certain destination. Both, the policy channel through the relaxation of legal entry barriers and the assimilation channel that eases integration after the crossing of the boarder are relevant in understanding the impact of networks.

3 Micro-foundation: A Random Utility Maximization Model of Migration

This section presents a model of migration choice across multiple destinations and derives the micro-foundation for the econometric specification that is brought to the data in the next section. The approach follows the canonical Random Utility Maximization model of migration (RUM), that has been widely applied as micro-foundation for gravity equations [see Beine et al. (2016) and Ramos (2016) for an overview]. The prospective migrant compares its utility at home j with the utility at destination k, which is part of the set of potential destinations $k \in D$. If the expected utility gain at any destination is bigger than the migration costs, an individual will migrate to the most beneficial option.

In line with the neoclassical model of international migration discussed in section 2, we can describe the utility of an individual i, born in country j, who does not migrate in year t with the following equation:

$$U_{jj}^{it} = w_j^t + A_j^t + A_j + \epsilon_{jj}^{it}. \tag{1}$$

where w_j^t represents the part of expected utility that is due to monetary benefits in the native country j, depending on the labour market situation and cost of living. I assume a common trend in earnings at origin across all residents. A (amenities) reflects the country or city specific non-pecuniary part of utility. It is useful to divide this component into time-persistent factors A_j (e.g. geographic location, fixed amenities, climate) and time-varying factors A_j^t (e.g. push factors, such as civil wars or environmental catastrophes in the country of origin).

ϵ^{it}_{jj} represents the stochastic individual-specific utility component, introducing heterogeneity into the model, which gives rise to distinct decisions within the pool of potential migrants in the origin country. Analogously, we can define i's utility if he migrates to city k by:

$$U^{it}_{jk} = w^t_k + A^t_k + A_k - C^t_{jk} + \epsilon^{it}_{jk}, \tag{2}$$

where C^t_{jk} additionally encompasses the costs of migrating from origin j to destination city k in year t. A^t_k includes the time varying part of utility (e.g. attitudes towards migration, public expenditure, institutions, security). Additionally, A^t_k and A_k the share of expected utility that is common to all destinations in Germany. w^t_k represents then the city specific monetary benefit, related to regional labor markets and costs of living.

The net utility of an individual i that decides to migrate from origin country j to the German city k thus can be expressed by

$$U^i_{jkt} - U^i_{jjt} = [w^t_k - w^t_j] + [(A^t_k + A_k) - (A^t_j + A_j)] - C^t_{jk} + [\epsilon^{it}_{jk} - \epsilon^{it}_{jj}]. \tag{3}$$

It is helpful to note that (3) is consistent with the classical categorization of push-, pull- and intervening factors that is for example elaborated in Black et al. (2011). Better conditions at a potential destination in economic terms (w^t_k) and non-pecuniary terms (A^t_k and A_k) will foster migration, while improving conditions at origin (w^t_j, A^t_j, A_j) will have a decelerating impact (and vice versa). Intervening factors (C^t_{jk}), or the costs of migration, are those that facilitate or restrict migration and create a sort of spatial disequilibrium, e. g. the limited mobility of people. If an individual is willing to migrate, the net utility gain of migrating must at least be large enough to recompense the migration costs, so that $U^{it}_{jk} - U^{it}_{jj} \geq 0$.

The distributional assumptions on the stochastic term ϵ^{it}_{jk} determine the expected probability that migrating to destination k, or staying at j, represents the utility maximizing choice of any individual i. In line with the literature, the random term is assumed to be independently and identically Extreme Value Type-1 (0,1) distributed. Following McFadden (1974) on discrete choice problems, we then obtain the probabilities that individual i, born in country j moves to location k in t as

$$P[U^{it}_{jk} = \max_{l \in D} U^{it}_{jl}] = \frac{exp[w^t_k + A^t_k + A_k - C^t_{jk}]}{\sum_{l \in D} exp[w^t_l + A^t_l + A_l - C^t_{jl}]}; \tag{4}$$

and, similarly, the probability of not migrating as

$$P[U^{it}_{jj} = \max_{l \in D} U^{it}_{jl}] = \frac{exp[w^t_l + A^t_j + A_j]}{\sum_{l \in D} exp[w^t_l + A^t_l + A_l - C^t_{jl}]}. \tag{5}$$

In (4) and (5) it becomes evident that a decrease in the attractiveness of alternative destinations (decrease of the denominator) leads to an increase of migration from j to k. This has often been illustrated by the thought experiment that Krugman (1995) made regarding trade: if we would move two random countries to Mars while keeping their attractiveness and accessibility unchanged, the migration flows between the two countries would increase because migration to other destinations would become very costly. This phenomena has been termed multilateral resistance to migration (MRM) in the migration literature (Bertoli and Fernández-Huertas Moraga (2013)).

By the law of large numbers, the ratio of migrants from j to k to the number of stayers can be expressed as

$$\frac{P[U^{it}_{jk} = \max_{l \in D} U^{it}_{jl}]}{P[U^{it}_{jj} = \max_{l \in D} U^{it}_{jl}]} = \frac{m_{jkt}}{m_{jjt}} = \frac{exp[w^t_k + A^t_k + A_k - C^t_{jk}]}{exp[w^t_j + A^t_j + A_j]}, \tag{6}$$

or in logs

$$lnm_{jkt} = [w^t_k - w^t_j] + [(A^t_k + A_k) - (A^t_j + A_j)] - C^t_{jk} + lnm^t_{jj}. \tag{7}$$

As a direct result of the distributional assumption on ϵ^{it}_{jk}, the bilateral migration flow can be described as a function of origin and destination characteristics only, where the denominators of (4) and (5) have cancelled out. The intuition behind this is, that the variation in attractiveness of alternative destinations induces an equal, relative change of the number of migrants from j to k and the number of stayers. This property is well-known as the independence of irrelevant alternatives. McFadden (1984) argues that the assumptions allow for the estimation of a choice model when only a subset of alternative choices is known to the econometrician. This property is useful in this context, since the dataset contains only data over a very specific subset of destinations: cities in NRW.

Nevertheless, it can be argued that the underlying assumptions are strong and might jeopardize the chances that the model is correctly specified. If the assumption à la McFadden fails, any econometric specification that does not control for MRM faces an endogeneity problem. A simple example can illustrate this point: Consider the choice of European destinations for potential migrants in Morocco. If the European economy faces a decline as a whole, this has two impacts on the migration flow from Morocco to Germany. Firstly, migration decreases because the attractiveness of Germany declines due to worsened labour market conditions. Secondly, the relative attractiveness of Germany increases since alternative destinations, such as France, Italy, Spain become less attractive. This would then lead to the underestimation of the importance of economic development in Germany. This consideration shows that it would be necessary to introduce a measure for all possible destinations to the estimation. Recent contributions suggested different ways to deal with multilateral resistance in the framework of a gravity model of international migration. Mayda (2010) addresses this issue by the inclusion of weighted average income per capita in the other destinations as a control for their aggregated time-varying attractiveness. Ortega and Peri (2013) derive an extended version of usual RUM model. Based on a nested logit model (Berry (1994) they add a correction term that accounts for changes in alternative destinations. This is consistent with an econometric specification with origin-year fixed effect. Beine et al. (2014) use destination-year dummies, which allows them to partly control for the dynamic resistance terms. In this work, I account for MRM on binational level by the inclusion of an origin-year dummy, which absorbs all

Germany-origin country specific factors. A city-year dummy will account for the arguably less important MRM across German cities.

I proceed by decomposing the bilateral cost component C_{jk}^t in the following way:

$$C_{jk}^t = d_j^t + T_j^t - \gamma ln(1 + M_{jk}^t) \tag{8}$$

where d_j^t reflects all binational, persistent and time-varying factors. Usual candidates are geographical distance between origin and destination, reflecting transportation costs, the presence of a common boarder, and measures for cultural distance, colonial ties and common language, reflecting non-monetary costs. The variable T_j^t stands for the tightness of entry laws and varies by year and origin country, but not by city.[1] That reflects general migration policy changes in Germany, and origin country specific agreements at the same time (e.g. EU east-enlargement, visa regulation or recognition of refugee status). I exploit the fact, that policies and other migration cost determinants are common to all cities in NRW, while the value of networks is region-specific. This comes in handy to account for all relevant cost factors by the inclusion of an origin-year dummy into the econometric specification, if one is to estimate the coefficient of the regional diaspora variable. The main variable of interest, M_{jk}^t reflects the diaspora size of a certain nationality in a city in t. A larger diaspora is expected to reduce migration costs. I allow for non-linearity by taking the logarithm of the network M_{jk}^t . As the size of the diaspora expands, the marginal impact of an additional migrant declines. To guarantee finite moving costs along with the log-definition, one unit is added to the diaspora size.[2]

Plugging the migration costs (8) in (7) leads to

$$lnm_{jk}^t = w_k^t - w_j^t + A_k^t + A_k - A_j^t - A_j - d_j^t - T_j^t + \gamma ln(1 + M)_{jk}^t) + lnm_{jj}^t. \tag{9}$$

The bilateral gross migration flow hence increases in the size of the income differential, as with the differential of non-pecuniary benefits. It decreases with bilateral migration costs and with tighter migration laws, whereas the presence of a diaspora at destination can decrease these costs and foster migration. The coefficient $\gamma$is henceforth the focus of interest in this work.

In the following, I derive the econometric basic specification, consistent with the theoretical result in (9). As the main interest here falls upon the diaspora variable, I firstly adopt a somewhat parsimonious specification in terms of observable controls. I circumvent the usual problems of misspecification by adding different dummy variables to the model specification and therefore fully exploit size and panel structure of the dataset. Firstly, I introduce the fixed effect $\mu_j^t = ln\ m_{jjt} - (A_j^t + A_j) - d_j^t - T_j^t$, which essentially captures all, time-varying and persistent, binational factors influencing migration from the country j to Germany, thus to all

[1]Municipal governments have only little impact on laws targeting migration and no leeway regarding distribution of refugees.

[2]This approach has largely been implemented in related literature. Robustness checks show that estimates are not impacted beyond the effect due to extension of the sample. See for example Bertoli and Fernández-Huertas Moraga (2015), Beine et al. (2011) or Beine and Parsons (2015).

German cities alike.[3] Further, I include a second fixed effect $\mu_k^t = w_k^t + A_k^t + A_k$ that accounts for all city specific pull forces. Time varying components are related to labour markets, level of rents, social benefits or the attitude towards migration in city k. Time persistent factors are for example location, climate, size and type of the city (large metropolis or small township). All those determinants are usually difficult to grasp with adequate observables, even when much data is available. Introducing the fixed effects μ_j^t and μ_k^t, the first econometric specification becomes

$$lnm_{jk}^t = \alpha_0 + \gamma ln(1 + M_{jk}^t) + \mu_j^t + \mu_k^t + \eta_{jk}^t \tag{10}$$

where η_{jk}^t represents a well-behaved error term with $\mathbb{E}(\eta_{jk}^t) = 0$. The implemented fixed effects significantly decrease the risk of misspecification and biased estimates. The resulting specification (10) is suited to analyse the network effect in the most precise way and reflects the baseline approach of this work. Later, I will partially relax the fixed effect structure to be able to estimate the impact of other origin-year varying factors. Accounting for inertia in location choices, all explanatory variables considered are lagged by one year. In line with the intuition, this specification delivers the best fit.

4 Dataset, Limitations and Econometric Strategy

4.1 Data

The data exploited in this paper originates from the Landesdatenbank NRW, a database that offers official statistical data of the federal state North-Rhine Westfalia and is directly associated to the German Federal Statistical Office. The main-dataset includes data on the size of stocks of foreigners, covering 206 different nationalities in the 53 municipalities of NRW over a horizon of ten years from 2006 to 2015.[4] The data is originally gathered by the central register for foreigners (CRF). Inter alia, this register collects data on nationality, place of birth and place of residence in Germany of foreigners, who do not hold a German passport and that stay at least three month in Germany.[5] The registration at the CRF is obligatory and a necessary condition, when dealing with any kind of official authorities and the omission can be sanctioned.[6] The data can therefore be assumed to be relatively accurate. The data on local networks is expanded by municipal data on GDP per capita, unemployment rate, population, naturalization rates and living-space, which are likewise extracted from the Landesdatenbank NRW. Further I use data on the total size of diasporas in Germany from DESTATIS, the German federal office for statistics.

[3] In principle lnm_{jj}^t should be treated as endogenous variable, as it is directly concerned by the amount of emigrants $\sum_{l \in D} lnm_{jl}^t$ As in Beine et al. (2011), I disregard this problem by assuming that the outflow of migrants is small relative to the number of residents m_{jj}^t.

[4] By constructing migration flows as the difference in stocks over two subsequent years, I lose the first year of observations.

[5] Individuals who hold a German and a foreign passport are not considered.

[6] According to Gesetz über das Ausländerzentralregister (Law over the central registration system for foreigners)

The exploited dataset brings along strength and weaknesses that emphasize the complementary value of this examination to existing studies. Firstly, I draw on yearly administrative data, as opposed to many studies which use census data, collected only in longer time intervals (Beine et al. (2014); Beine et al. (2015)) or cross-sectional approaches (Grogger and Hanson (2011)). The longitudinal dimension of the dataset allows for the implementation of fully-fledged panel methods, which are able to control much more carefully for usually unobserved determinants of migration, coping with many of the econometrical challenges that typically arise in this context (Ortega and Peri (2013)). Moreover, Beine et al. (2016) emphasize that longer time periods are at odds with the assumption of the RUM model, that the deterministic component of utility does not vary within the period. On the down side, the short time-intervals between the observations could hamper the identification of effects, given the large inertia in migration movements. The factors that mattered in the preceding years could still matter in the current year.

Existing literature has shown that the benefits of networks decrease with geographic distance between network member and newcomer (Comola and Mendola (2015); Massey and Aysa-Lastra (2005); Patacchini and Zenou (2012)). Much research on ethnic network effects has used data on international migration flows (Beine et al. (2011)), or looked at the development in large metropolitan regions (Beine et al. (2015)). The disaggregated nature of the dataset on municipality level should thus be of advantage for the detection of network effects, as for many of the channels through which networks operate, locational proximity is a necessary condition. It could be argued that regional diasporas exert externalities on other cities which are in proximity. I rely on the assumption that the main effect of diaspora unfolds within the regarding city, and that regional spillovers are limited.

4.2 Measurement Error

On the down side, the structure of the data poses econometric challenges that must be dealt with. The dependent variable in (10), the equation that is to be estimated based on the derived RUM model, is the gross flow of migrants: the absolute number of individuals moving from one place to the other in a particular direction. In the absence of data on direct bilateral gross migration flows, these are approximated by the difference in stocks between two subsequent years as $\text{migrants\_in}_t \approx stock_t - stock_{t-1}$. This approach fails to account for death, naturalization and out migration (return or internal migration) and can thus introduce measurement error into the estimation. The following equation demonstrates this:

$$\text{stock}_t - \text{stock}_{t-1} = \text{migrants\_in}_t - \text{migrants\_out}_t - \text{death}_t - \text{naturalisation}_t \tag{11}$$

$$\text{migrants\_in}_t = \text{stock}_t - \text{stock}_{t-1} + \text{migrants\_out}_t + \text{death} + \text{naturalisation}_t \tag{12}$$

While (11) is the observed measure for the main part of this analysis, (12) would be the theoretically justified concept. Hence, to derive the correct measure of migration from the observed data on stocks, data on out migration, death rates and naturalization would be necessary. Further, when it comes to the migrant stocks as explanatory variable, the relevant measure should consist of people with a migratory background: residents that have kinship or friendship ties to others, who are located abroad. By considering only those that do not hold a German passport, we fail to capture a part of this group, e.g. those that hold a double nationality, or those that have already been naturalized. Due to data constraints, I can only account for naturalization for a subsample. Sufficient data to adopt a concept of migratory background that would account for individuals naturalized before 2006, and for individuals holding a double nationality is not available. I provide results based on an subsample adjusted for naturalization for robustnes purposes in section 5.1.

When it comes to death rates, appropriate data is not available neither. The results rest on the assumption that death rates are relatively insignificant, compared to the number of migrants, and do not pose major problems to the estimation. Birth does not pose a problem as children of foreign born parents automatically hold both nationalities and hence do not appear as foreigners in the data. At the same time, children younger than ten are unlikely to exercise any network externalities.

Lastly, it is not possible to control for return migration, internal migration or migration from Germany to third countries. In the presence of migration costs, the decision to move to a particular country and to stay in that country are not the same (Ramos (2016)). Nevertheless, gross flows, net flows and migrant stocks are closely related: Forces which attract migrants to a certain destination (gross flow) might also increase their propensity to stay at that destination (and hence increase stock and net flow). Subsequently, the constructed dependent variable reflects both, the pull-forces and the stay-forces. Additionally, regarding internal migration movements, the networks are also likely to play a role when it comes to the redistribution of migrants who are already in Germany. The relevant concept, when measuring the network effect of regional diasporas, becomes thus a question of definition. Both, in migration and out migration should be affected by networks at the city level, and the development of stocks contains relevant information about both mechanisms.

All in all, the measure for migration adopted here can be expected to be smaller than the true gross migration flows. Nevertheless, the measure obtained represents a reasonable approximation for the concept at stake and should allow to derive robust conclusions.

4.3 Zeros, Negative Values and Heteroscedasticity

Another problem, resulting from the approximation of gross flows by variations in stocks, is the occurrence of negative flow values. Due to return migration, death and naturalization the obtained measure is smaller than zero for some city-country combinations in some years. By definition, the gross migrant flow is non-negative and the usual log-log specification of the log-linearized pseudo gravity model does not allow for negative values. The descriptive stats in table 1 reveal that the share of observations with negative flows lies over one fourth of the total dataset and is thus relatively important. Former research has either excluded those observations from the sample (Beine et al. (2011)), set them to zero (Bertoli and Fernández-Huertas Moraga (2012)), or added them to the proxy for the flow in the opposite direction, assuming unanimous return migration (Beine and Parsons (2015)). All those approaches are merely imperfect solutions in response to limited data availability. This analysis applies the usual approach of a truncated regression, dropping observations with negative flows. This is likely to lead to some sort of sample selection bias: it must be assumed that there is a correlation between the covariates and the probability that the bilateral flow is smaller than zero. I carry out robustness checks to address this issue in section 5.

A related problem, which has received broader attention in trade and migration literature, is related to the occurrence of the many zero values, both for flows and stocks in the dataset. Firstly, it must be noted that, while posing technical problems to the estimation, this observation is fully consistent with the theory. Migration costs might be restrictively high for some origin-destination combinations. Consequently, we obtain quiet coherent series with zero migration and stocks for some city-country combinations across time, where the populations are rather low or/and migration costs high (e.g. Andorra or Samoa). The usual OLS regression will drop those observations, as the log specification of the dependent variable leads to undefined values in those cases. Dropping zero observations in the way that OLS does it, potentially leads to sample selection bias. The goal of this work is to explain migration streams based on the regressors in the model. The probability that a migration flow is smaller than zero or zero depends on the covariates, hence the usual OLS estimators that drops those values is likely to be biased.[7] In fact, with 45563 nearly half of the observations of the dataset show zero flows. Instead of throwing away these observations, other methods have been discussed in the trade literature, such as adding one unit to the dependent variable, or replacing the zero observations by a small arbitrary value [see Frankel et al. (1997) for a survey of the procedures]. However these methods will generally lead to inconsistent estimates of the parameter of interest, while the extent depends on data and model structure (Frankel et al. (1997); Santos Silva and Tenreyro (2006).

Next to the sample selection bias, Santos Silva and Tenreyro (2006) emphasize a second issue resulting from the log-linearization of (6). They make the point that a well-behaved error term with $\mathbb{E}[\eta_{jkt}] = 1$ for estimating the multiplicative gravity model in (6) does not imply that $\mathbb{E}[ln(\eta_{jkt})] = 0$ after the log-linearization in (7). The expected value of $ln(\eta_{jkt})$ depends on higher moments of the distribution of η_{jkt}, including its variance. Thus, in the presence of heteroscedasticity, it will be correlated with some of the covariates. This in turn invalidates one important assumption of consistency of OLS estimates. It is important to note that this kind of heteroscedasticity cannot be dealt with by simply applying robust standard errors along with the OLS estimation, since it affects the parameter estimates

[7] Similar to the issue related to negative values discussed above.

themselves. Consequently, they show that, even controlling for fixed effects, the presence of heteroscedasticity can generate strikingly different estimates when a gravity equation is log-linearized, rather than estimated in levels. The results which are presented later generally support that finding in this context.

Santos Silva and Tenreyro (2006) propose to estimate the original multiplicative form of the equation in levels, using a Poisson pseudo maximum likelihood (PPML) estimator. They show that the PPML estimator produces unbiased estimators in presence of heteroscedasticity, even when the data does not follow a Poisson distribution. Besides it allows to circumvent sample bias by the inclusion of observations with zero values of the dependent variable, here the migration flows, since there is no need to take its logarithmic form. Evidence based on Monte Carlo simulations shows that PPML performs strongly in datasets with a large number of zeros and heteroscedasticity (Santos Silva et al. (2015), Santos Silva and Tenreyro (2011)). At the same time, it allows for the same intuitive elasticity interpretation of coefficients as of the conventional OLS estimation with the log of the dependent variable. Therefore, in the subsequent section I will use the PPML estimator as main estimation strategy.

On a more technical side the PPML estimator, together with the specification with many fixed effects can pose estimation problems to statistical software. Several numerical issues can avoid convergence when calculating maximum likelihood estimators [see Santos Silva and Tenreyro (2010); Santos Silva and Tenreyro (2015) for a detailed discussion]. This work is conducted with Stata 12 SE. In fact, the usual Stata commands — poisson — and its panel version — xtpoisson — do not achieve convergence in the given set-up. For the following PPML estimations, I use the user-written command -ppml - (Santos Silva and Tenreyro (2015)). This command applies the routine proposed by Santos Silva and Tenreyro (2010), while it is able to deal with several high-dimensional fixed effects. Explicitly, the routine guarantees the existence of estimates by dropping problematic regressors. If any of the regressors that are dropped is a dummy variable, by default — ppml — will drop the observations with the less frequent value of the excluded dummy. This is done to prevent the observations identified by the dropped dummy from being treated as part of the base category. The aforementioned explains the slightly differing sample sizes for alternative specifications. Using the same command with a less demanding structure of fixed effects delivered the same results as the integrated commands — poisson — and — xtpoisson —, and can be regarded as a reasonable approach to the issue at hand. However, the not random shift in considered sample could potentially introduce limited sampling bias to the estimates.

Table 1: Descriptive Statistics

		Flows							*Stocks*					
	Number	Flows>0	Flows=0	Flows<0	Mean	Std. Err.	max	min	Stocks=0	Stocks & Flows = 0	Mean	Std. Err.	max	min
full dataset														
Observations	98262	30646	45563	22053	4.62	53.45	3167	-1341	31080	29775	173.89	1133.27	66175	0
Origin Countries	206													
Cities	53													
Years	9													
subsample adj. for naturalization														
Observations	42453	24991	7379	10083	16.15	79.70	3175	-892	2844	2737	427.58	1839.72	67794	0
Origin Countries	89													
Cities	53													
Years	9													

Note: The table shows descriptive statistics for the full dataset where naturalization rates are not considered, and the naturalization augmented subsample.

5 Baseline Regression Results and Robustness Checks

5.1 PPML, OLS and Fixed Effects

Table 2 presents estimation results of the basline model, that focuses solely on the role of networks. In column 1 I estimate the network effect with one-dimensional city, origin country and year fixed effects. Column 2 reveals that adding the multiplicative city-year fixed effect, controlling for city specific, time varying pull-factors does not impact much upon the coefficient. This is likely due to the fact that change on city level is rather slow and much of the difference in attractiveness between cities is absorbed by the time persistent city fixed effect, while the year fixed effects absorbs general development attractiveness of German cities as migration destination. In column 3, I add a multiplicative origin-year fixed effect that absorbs all time varying forces that cause migration in the origin country, the so-called push-factors. The result indicates that controlling for those origin-year varying factors is more important in this context: The estimated coefficient of regional networks decreases from around 0.77, in the first two specifications, to 0.7 when controlling for those factors. Column 4 presents the preferred, and theoretically justified specification with, both city-year and origin-year fixed effects, hence controlling for all relevant push- and pull-factors. I omit the one-dimensional fixed effects, since they are entirely absorbed by their two-dimensional counterpart in this specification. The resulting coefficient indicates that a 10 percent increase in the diaspora size is associated with a 7 percent increase in migration in the subsequent year. The model explains 85% of the observed variance in migration flows, expressed by Efron's pseudo R^2 for maximum likelihood estimations. This result is reassuringly consistent with findings of other papers which estimated this network effect based on structural gravity models (Beine et al. (2011); Beine and Parsons (2015); Bertoli and Fernández-Huertas Moraga (2012)). In their survey of related work, Beine et al. (2016)) state that this elasticity typically lies around 0.7 for migration to OECD countries.

A further option for the econometric specification would be the additional inclusion of a city-origin fixed effect, controlling for any dyadic time-persistent factors that affect migration flows. As argued in section 3, I control for binational factors by the inclusion of μ_j^t, which should be largely more important in this context. Additionally, the estimation of such a model delivers counter-intuitive results, indicating that the aforementioned specification would be too demanding, given the variability in the data at hand. In fact, the descriptive statistices in table 1 demonstrate that the major part of variability in the data is due to variation across stocks. I report the results in the appendix.

Next, I address the concern related to potential measurement error when calculating the migrant flows based on the observed stocks of migrants in the cities. As discussed in section 4, the biggest concern when measuring the development of migrant population in cities are arguably the high and heterogeneous naturalization rates across different diasporas. Intuitively, the naturalization rates for the relatively old Turkish diasporas are higher than for diasporas which developed in recent years, e.g. Syrians in Germany. The failure to correct for different naturalization patterns could bias the estimates of the network effect. I dispose

Table 2: Baseline Regressions and Robustness Checks

	1	2	3	4	5	6	7
dependent var.	flow	flow	flow	flow	flow	flow>0	ln flow
Lagged ln diaspora	.7648*** (.0203)	.7745*** (.0169)	.6963*** (.0167)	.7040*** (.0131)	.7107*** (.0137)	.6719*** (.0133)	.3532*** (.0082)
City FE	yes	yes	yes	no	no	no	no
Origin FE	yes	yes	yes	no	no	no	no
Year FE	yes	yes	yes	no	no	no	no
Year*Origin FE	no	no	yes	yes	yes	yes	yes
Year*City FE	no	yes	no	yes	yes	yes	yes
(pseudo) R^2	.8394	.8523	.7799	.8498	.8343	.8552	.7414
Nb. obs.	76209	76209	64879	64879	30958	30506	30646
Dataset	full	full	full	full	nat. adjusted	full	full
Method	PPML	PPML	PPML	PPML	PPML	PPML	OLS

Note: Standard errors are in parenthesis. ***, **, * significant at 1%, 5%, 10% level. Standard errors calculated as Efron's pseudo R^2 for maximum likelihood estimations. Flows are calculated as the difference between the regional diasporas in two subsequent years.

of data on observed naturalization rates for a subset of 89 nationalities. I augment the yearly observed diasporas by the aggregated sum of naturalized citizens for each nationality. It is not possible to observe whether naturalized individuals stay in the city where they have been naturalized in subsequent years. I assume that people which qualify for naturalization are relatively immobile, given that they spent already several years before in Germany and likely created relatively stable living conditions. Further, those that leave a city might be counterbalanced by incoming individuals that have been naturalized elsewhere and hence are not identified as migrants in the dataset. Comparing column 4 and 5 reveals consistent coefficient estimates for both datasets, while the number of observations drastically shrinks due to data constraints regarding naturalization rates. I will henceforth draw on the full dataset for the following estimations.

Lastly, comparing column 6 and 7 points out another interesting result. Section 4.3 advocated the PPML estimator for two different reasons: Firstly, it is less prone to sampling bias since it includes observations with zero flow into the estimation, which are naturally excluded by the OLS approach, due to the log-specification of the dependent variable. Secondly, it delivers consistent estimates in the presence of heteroscedasticity in the framework of gravity models. I estimate the same specification respectively with OLS and PPML, and limit the sample for the PPML estimation to the same subsample of observations with positive flows.[8] The estimated coefficient of networks is nearly twice as large when estimated with PPML. This finding suggests that the second, heteroscedasticity-related reason might be the more

[8]The sample sizes slightly deviate because the algorithm used for the PPML estimation drops observations that prevent the estimation from converging.

important one in this context. The result is in line with the literature that revealed similar results based on simulations (Santos Silva and Tenreyro (2006), (2015)).

5.2 Rolling Regressions with Different Flow and Stock Thresholds

Table 3 presents further robustness checks of the baseline model. One concern is that the estimation might suffer a sampling bias, and consequently be inconsistent. As discussed in section 4, former research has essentially excluded observations with negative flows from the estimations. This could introduce correlation between the covariates and the error term. To shed some light on how sensitive the estimated network coefficient reacts to this issue, I estimate the basic model with origin-year and city-year fixed effects, excluding observations with small flows at different thresholds (flow>-1; flow>0; flow>5; flow>10). I exploit the fact that the transition from positive to negative flows is gradual and negative flows are small in average, with a mean of -10.85. Therefore, I argue that observations with small positive flows are likely to resemble those with negative flows in other relevant characteristics. Column 1 to 4 show the results, starting with the largest possible sample, considering all non-negative flows, and gradually augmenting this threshold. While the sample size quickly shrinks with increasing threshold, from 76209 observations at flow>-1 to only 8673 observations for which the migrant flow is bigger than 10, the estimated coefficient of regional diasporas remains relatively stable at around 0.7. This result is reassuring regarding the potential sample selection bias, as it indicates that the estimated coefficient reacts not very sensitive to sample selection based on the size of flows. The consistent downward trend of the coefficient with augmenting threshold, however indicates that the network coefficient estimates could be slightly upwards biased. Yet, this represents only an indication since observations with negative flows could be substantially different.

Another source of concern is related to the frequent occurrence of pairs of zero observations in the dataset. For a large part of the dataset, stocks and flows in a given year are zero for some nationality-city combinations. This can either be due to small destination cities in terms of population, or due to remote or small origin countries that generally don't send many migrants to Germany. In both cases migration costs become prohibitively high. Consequently, these observations qualify only to a limit degree to draw conclusions about networks. The descriptive statistics in table 1 reveal that roughly half of observations in the dataset have zero sized diasporas, and again for nearly 96% of those, the subsequent migration flow is zero. Spurious correlation between these two variables could be the consequence. To address this concern, I follow the same strategy as before. I re-estimate the baseline model with different thresholds for the minimum network size in a city. Again, demonstrated in columns 5 to 9 of table 3, excluding zero stocks, or dropping small stocks does alter the estimated coefficient very little, while the sample size decreases significantly. Further, it can be conjectured that the network effect is driven equally by small and large networks, as the estimated coefficient changes little, even when limiting the estimation to networks with over 500 individuals.

Table 3: Rolling Regressions with augmenting Thresholds for Flows and Stocks

	1	2	3	4	5	6	7	8	9
Threshold	flow>-1	flow>0	flow>5	flow>10	L.stock0	L.stock>0	L.stock>20	L.stock>100	L.stock>500
Lagged									
ln diaspora	.7040***	.6720***	.6526***	.6376***	.7040***	.7019***	.7110***	.7238***	.7463***
	(.0131)	(.0133)	(.0145)	(.0156)	(.0131)	(.0132)	(.0138)	(.0145)	(.0167)
Year*Origin	yes	yes	yes	yes	yes	yes	yes	yes	yes
Year*Country	yes	yes	yes	yes	yes	yes	yes	yes	yes
pseudo R^2	.8498	.8457	.8395	.8381	.8498	.8480	.8460	.8510	.8725
Nb. obs.	64879	30646	12788	8673	64879	44934	27747	21226	16490
Dataset	full	full	full	full	full	full	full	full	full
Method	PPML	PPML	PPML	PPML	PPML	PPML	PPML	PPML	PPML

Note: Standard errors are in parenthesis. ***, **, * significant at 1%, 5%, 10% level. Standard errors calculated as Efron's pseudo R^2 for maximum likelihood estimations. Flows are calculated as the difference between the regional diasporas in two subsequent years.

5.3 Endogeneity and Shift-Share Instrumentalization

Short of a randomized experiment in which regional stocks of migrants are varied across cities and time exogenously and randomly, I cannot rest assured that the regressions ran, necessarily reflect causality between regional networks and subsequent migration. This is especially a concern when estimating network effects as the presence of a network is just the consequence of aggregated past migration. Unobserved factors could have triggered past migration streams and future migration streams alike. The result would be the correlation of the error term of the estimated model with the diaspora size, thus endogeneity that leads to an inconsistent (and likely upwards biased) estimate of the network coefficient. Additionally Beine et al. (2016) point out that endogeneity could also arise from some kind of reverse causality when flows are computed from stocks. The usual way forward to tackle this issue would be the implementation of an instrumental variable strategy: Replacing the critical network variable with a variable that is correlated with stocks but arguably uncorrelated with future migration flows.

Several instruments have been used in the related literature to dispel those endogeneity concerns. Beine et al. (2011) and Beine et al. (2014) used the presence of guest-worker programs to instrument the present network sizes. The idea is that a guest-worker programs in the past are related to larger present networks, while not having an impact on future migration beyond. Munshi (2003) uses the rainfall in the origin-communities as instrument for the level of migration. They argue that rainfall is negatively correlated with out-migration at origin, while being unrelated to US labor market shocks.

Due to the panel nature of the dataset, the identification of an appropriate external instrument becomes very difficult. The applied approaches are only suitable instruments when dealing with a cross-section of data, as they cannot instrument stocks that vary over time. Hence it would be necessary to find an adequate instrument for each year of the panel which excludes typical exogenous shocks, such as guest-worker programs.

One alternative that has been largely applied in regional economics is the so-called shift-share instrumentalization (Ottaviano and Peri (2006); Ruist et al. (2017)). To implement this strategy, I use data on the total national diasporas in Germany by origin and year, originating from DESTATIS.[9] By multiplying the regional stock of migrants from origin j in city k in 2006, the first panel year observed, with the growth rate of the total diaspora of origin j in Germany between 2006 and t, I obtain an internal instrument for regional diasporas for each year. I exploit the fact that I have data on the disaggregated stocks in NRW, as well as on the development of total stocks by origin in Germany. This instrumentalization procedure thus generates local and temporal variation by exploiting variation in national inflows, which are arguably less endogenous, regarding local dyadic factors. At the same time, the national development is naturally a good predictor of the local migration dynamics, and hence computed stocks are correlated with the observed stocks. In the absence of suitable instruments this presents a merely imperfect solution to the issue at hand. Nevertheless, it

[9]The Federal Statistical Office of Germany

decreases the potential for endogeneity and represents a further robustness check.

Firstly, the growth rates of national diasporas for each year and origin country with respect to its stock in 2006 are computed according to the following formula:

$$g_j^{t=6;t} = \frac{M_j^t - M_j^{t=6}}{M_j^{t=6}} \tag{13}$$

Then I proceed by computing the instrumented diasporas for each year t by multiplying the observed diaspora in 2006 $M_{jk}^{(t=6)}$ with 1 plus the growth rate of the national diaspora during that period

$$\tilde{M}_{jk}^t = M_{jk}^{t=6} * (1 + g_j^{t=6;t}) \tag{14}$$

I then estimate (10) instrumenting the observed local diaspora by the shift-share instrument. The downside of this approach is related to the relatively short panel (9 years). The shift-share procedure can be problematic, especially given the inertia of location choice in the context of migration. Additionally, the change in migrant stocks is naturally small relative to their size, and the main share of variability in the data stems from variation over diasporas; only little is due to the variation of diasporas over time. The growth rates of regional diasporas and those of national diasporas show only a correlation of 20%. Thus, while the national development clearly matters for regional diasporas, many other factors at the city level play likewise. By its construction as a combination of observed regional diasporas, and national growth rates, the shift-share IV closely related to the potentially endogenous network variable. In fact, the correlation between observed diasporas and computed diasporas lies at 99.29%. This in turn poses a problem regarding the exclusion restriction of the instrument, as it is likely to suffer the same problems since the instrumented variable. It is therefore unlikely that the instrument will entirely solve the endogeneity problem.

Table 4 presents the familiar PPML result and the related PPML-IV estimation. For technical reasons, I estimate the specification with separate city, origin and year fixed effects.[10] As shown in section 5.1, this has only a minor impact on the estimated network coefficient. The instrumental strategy produces a significantly smaller estimate of the regional network effect. This indicates a correlation of the observed network variable with the error term of the baseline model. Relevant unobserved drivers of migration seem in average to be positively correlated with the network variable. This finding lends some support to the ex-ante expectation: The estimated coefficient could be upwardly biased and overestimate the importance of networks in this context.

6 Model Extensions

In the following section, the baseline specification will be extended to address different empirical questions that will shed further light on the functioning and importance of migrant networks. While the data structure allows to consider broader aspects, the main focus of

[10]The preferred specification with two-dimensional fixed effects leads to convergence issues when estimating the PPML-IV model.

Table 4: Shift-Share Instrumentalization

	1	2
dependent var.	flow	flow
Lagged ln diaspora	0.765***	.4403***
	(0.0019)	(.0237)
City FE	yes	yes
Origin FE	yes	yes
Year FE	yes	yes
Nb. obs.	76209	66725
Dataset	full	full
Method	PPML	IV-PPML

Note: Standard errors are in parenthesis. ***, **, * significant at 1%, 5%, 10% level.
Flows are calculated as the difference between the regional diasporas in two subsequent years.

this analysis will remain on the network effect of regional diasporas and their impact on subsequent migration patterns. Firstly, I investigate whether potential refugees show different migration and agglomeration behaviour than conventional, economic migrants. Secondly, I will add data on HDI and GDP per capita in the country of origin. In this way, I analyse how far migrants originating from more or less developed, or wealthy countries (based on HDI/ GDP) differ in their reliance on diasporas. Thirdly, following the approach of Beine et al. (2015), the importance of assimilation and policy channel (especially family reunification programs) is disentangled, exploiting data on regional and national diasporas.

6.1 Introducing Origin-specific Control Variables

In all three cases I add a variable that varies with origin country and year to the econometric baseline specification in (10). When estimating the coefficient of such a variable, it is indispensable to exclude the origin-year varying fixed effect from the specification, since it would entirely absorb coefficients of all other origin-year varying covariates. This comes at the cost that not for all relevant origin-year varying factors might be accounted anymore, leading to misspecification of the model. Those factors include push forces in the origin country and time varying, binational determinants. An estimation that falls short in accounting for these determinants is likely to suffer endogeneity and to produce biased estimates. The panel structure of the dataset allows me to replace the origin-year fixed effect by a one-dimensional origin dummy. This dummy will account for all factors that are persistent during the 9-year period of consideration, and will hence deal with most issues related to migration costs in physical and political dimensions, which tend to stay relatively stable. Nevertheless, if one is to estimate the impact of origin-year varying factors in this context (e.g the impact of war or HDI on the network effect), it becomes key to introduce adequate control variables which account for the remaining time-varying component.

With this in mind, I introduce a set of observed control variables. First, I add the logarithm of GDP per capita at origin, its quadratic term to account for credit constraints at origin ($\text{GDP}_j^t; \text{GDP}_j^{2t}$) (Belot and Hatton (2012); Mayda (2010)), and (HDI_j^t).[11] Next, I create a dummy variable that takes the value of one, if a country is part of Schengen and zero otherwise, reflecting the related decrease in legal migration costs (schengen_j^t). I use data from the latest version of the PRIO armed conflict database (Gleditsch et al. (2002)) that contains all armed conflicts in which at least one party is the government of a state. A war dummy takes the value one if an armed conflict that caused more than 999 deaths within one year takes place in the respective country (war_j^t). Finally, I retrieve data from the Freedom House database on political rights and civil freedom. I take the mean of both dimensions that ranges from one (not free) to seven (entirely free) (free_j^t).[12] While serving as control variables in all extensions, I will draw on the same data to investigate potential moderating effects. The econometric specification hence becomes:

$$\begin{aligned} ln\ m_{jk}^t = \alpha + \gamma\ ln(1 + M_{jk}^t) + \beta_1\ \text{GDP}_j^t - \beta_2\ \text{GDP}_j^{2t} - \beta_3\ \text{HDI}_j^t + \beta_4\ \text{war}_j^t \\ + \beta_5\ \text{schengen}_j^t - \beta_6\ \text{free}_j^t + \mu_k^t + \eta_{jk}^t. \quad (15) \end{aligned}$$

Hereafter, I will refer to the whole set of observed origin-year varying variables as the vector O_j^t to shorten notation. It has to be noted that (15) deviates in a second point slightly from the theoretically justified specification (10), since it does not properly account for the share of residents $ln\ m_j j^t$. This will be partially resolved by the inclusion of origin country fixed effect μ_j.

[11]International data on GDP per capita and HDI retrieved from the World Bank.

[12]Freedom House. 2016. Freedom in the World, 1973-2016.

Table 5: Extensions

	1	2	3	4	5	6	7
Destination							
L.ln diaspora	.7040*** (.0130)	.7745*** (.0169)	.7621*** (.0174)	.7757*** (.0169)	.0332 (.1132)	.3651*** (.0665)	.7264*** (.0214)
L. national diaspora							-.0016 (.0835)
Origin							
war			.7403*** (.0817	1.4848*** (.2389)	.7331*** (.0808)	.7437*** (.0788)	.6601***
(.0807)							
war* L.ln diaspora				-.1302*** (.0424)			
ln GDP/c			11.8129*** (1.1926)	11.8417*** (1.1942)	12.8256*** (1.2334)	12.9954*** (1.2426)	10.0269*** (1.2646)
(ln GDP/c)^2			-.6148*** (.0613)	-.6144*** (.0615)	-.6930*** (.0642)	-.6798*** (.0634)	-.5322*** (.0664)
ln GDP/c * L.ln diaspora					.0752*** (.0115)		
hdi			-14.4764*** (2.2626)	-14.5175*** (2.2820)	-14.3187*** (2.2552)	-17.2062*** (2.3047)	-11.9371*** (2.6240)
hdi * L.ln diaspora						.5228 *** (.0832)	
Schengen			.2141*** (.0819)	.2107*** (.0818)	.1836** (.0811)	.1979** (.0813)	
Political Rights/ Civil Freedom			-.1026*** (.0406)	-.1011*** (.0403)	-.1005** (.0403)	-.0989** (.0401)	-.21491***
(.0386)							
Origin FE	no	no	yes	yes	yes	yes	yes
Year* City	yes	yes	yes	yes	yes	yes	yes
Year* Origin	yes	no	no	no	no	no	no
pseudo R^2	.8498	.7188	.7576	.7585	.7578	.7582	.7427
Nb. obs.	64879	74312	53704	53704	53704	53704	41193
Dataset	full	full	full	full	full	full	full
Method	PPML	PPML	PPML	PPML	PPML	PPML	PPML

Note: Standard errors are in parenthesis. ***, **, * significant at 1%, 5%, 10% level. Standard errors calculated as Efron's pseudo R^2 for maximum likelihood estimations. Flows are calculated as the difference between the regional diasporas in two subsequent years. All covariates are lagged by one year to account for inertia in location choices.

In the first three columns of table 5, I test whether the introduced origin country specific variables adequately control for push-factors. Comparing the first two columns shows that controlling for country-related push-factors is important: Dropping the origin-year dummy biases the network coefficient upwards. The respective share of observed variability that is explained by the model, measured by Efron's pseudo R^2 for maximum likelihood estimations, decreases by around 13 percentage points, compared to the full specification. Column 3 presents the first set of results. I introduce the full set of origin-year varying observables. Compared to column 1, we see that this specification naturally falls short in accounting for all factors formerly covered by the fixed effect μ_j^t. The R^2 remains clearly under the value, but does increase by around 4 percentage points compared to the model without any controls. Similarly, the coefficient on networks moves slightly towards the value estimated by the full specification in column 1. This result gives rise to some concerns and indicates that the subsequent estimations could be biased due to omitted variables. That highlights the typical difficulty to account for all relevant push-factors in the origin country by observable control variables.

Nevertheless, it is reassuring that the relative change in coefficient between columns 1 and 3 remains moderate and all coefficients are significant at the 0.01 level, with their signs going into the expected directions. Column 3 can hence be considered as the next set of results: a war in the country of origin leads to more migration from this country in the subsequent year. This positive coefficient absorbs both, push factors that force people out of the country, and a decrease in policy related entry costs as the asylum procedure decreases usual policy costs, e.g. the application for visas. Secondly, the non-linear effect of GDP per capita at origin is in line with the theory of credit constraints to migration. Many people can simply not afford to migrate to Germany as the related monetary costs are too high. Hence the initial increase in economic welfare empowers people to migrate. Eventually the decreasing push force, related to improving economic condition in the origin country, starts to impact negatively on out-migration. The large coefficient of HDI can be explained by the definition of HDI between 0 and 1. Improving conditions in the origin country clearly decrease emigration, once accounted for potential credit constraints. The positive estimated coefficient of the Schengen dummy highlights the migration fostering effect when policy costs decrease. Finally, the variable measuring the degree political and civil freedom shows as expected a negative coefficient. Freedom can be seen as a part of country specific utility that renders a country more attractive and hence "pushes" fewer out.

Next, I will precede by extending (15) in three different ways and present the empirical results.

6.2 Location Choice of Refugees

The refugee crisis that slowly developed with the riots in Syria in 2011 and peaked in 2015 has governed the political landscape in Europe and worldwide ever since. In 2008 Eurostat reported 225 150 people who sought asylum in a country of the European Union. Till 2011 this number rose to 309 040 and peaked in 2015 with 1 322 825 asylum applications in Europe. While France was the preferred destination till the end of 2011, Germany outstripped it largely in the subsequent years, driven by a comparatively open refugee policy. In 2015 Germany received nearly one third of all asylum applications to European countries, roughly six times as many as France in the same year (Eurostat, 2017). The structure of migrant inflow in Germany shifted thus from predominantly economic migrants, seeking for income opportunities, towards a larger share of refugees in recent years. In this context, it is clearly policy relevant to understand whether one can expect the localization patterns of refugees to differ from those of "conventional" migrants. Differences could stem from either different preferences of refugees on the one hand, or from policy induced legal burdens (like a mandatory reallocation mechanism of refugees) on the other.

Several factors can give reason to expect different localization patterns. For once, the preceding conditions of people who leave their country of origin during war time clearly differ from those of other migrants. War is generally unpredictable and concerned individuals are often left with only little time to prepare their escape. They must leave financial resources behind and leave with insufficient information regarding travel and destination. A priori, there is no clear expectation evident about the sign of the effect. Refugees could rely more on friendship and kinship ties of regional networks, as they reflect the most credible and accessible way to gather upfront, destination-specific information. On the other hand, the destination of a refugee might be more determined by random factors and sudden opportunities, since rational decision-making could suffer from difficult surrounding conditions.

On the policy side, refugees are generally not free to self-select their city of residence after their arrival in Germany. According to the "Königsteiner Schlüssel", which is the basis for the distribution mechanism of refugees in Germany since 1993, they shall be distributed over the federal states to two third relative to their tax revenue and to one third to their population. At the point of first contact they will be registered and subsequently allocated to their final destination, according to the regarding quotas. Only parents and their children have the right to be reunified at the same location. After the assignment, the concerned must remain at the refugee camp till the asylum proceedings are finalized, which can take up to 6 month. Hence it can be expected that the network effect is smaller when the freedom of location choice is limited both in the short term (direct obligation) and in the long-term (region specific adjustment cost already borne at determined location).[13]

The dataset allows me to investigate these questions at hand with recent data from German cities. I use the introduced dummy war_j^t as an indicator for whether a migrant qualifies as refugee. It can be expected that an individual migrant who comes from a country that recently hosted a war has higher average chances to be granted the refugee status in Germany. Hence the war dummy can be considered as a good proxy for the migrant status of a given immigrant. I extend equation (15) with an interaction term of this war dummy with the diaspora size. Estimating the coefficient of this multiplicative term allows to analyze whether the network effect differs for potential refugees. The econometric specification becomes then the following:

[13] See the website of the German Ministry for Refugees and Migration offers more detailed information on the distribution system: http://www.bamf.de/DE/Fluechtlingsschutz/AblaufAsylv/Erstverteilung/erstverteilung-node.html .

$$ln\ m_{jk}^t = \alpha + \gamma\ ln(1 + M_{jk}^t) + \lambda_1\ ln(1 + M_{jk}^t) * \text{war}_j^t + \beta\ O_j^t + \mu_j + \mu_k^t + \eta_{jk}^t. \qquad (16)$$

In table 5 column 4, I present the results for the first extension. As discussed in section 6.1 the results should be considered with caution, as it is likely to suffer some degree of omitted variable bias. This remark weighs similarly for the following 2 extensions. Further, it must be noted that the interpretation of interaction effects in non-linear models is not straight forward. They are not stable and vary with covariates. Especially when the estimated coefficients are relatively small this can lead to fallacious conclusions. Presented results should hence only be considered as an indication for the underlying effect (Ai and Norton (2003); Greene (2010)). I find that the network effect of regional diasporas decreases for migrants originating from a country in conflict. With the given data, it is impossible to disentangle which share of this effect is policy induced (limited self-selection in location decisions) and which part is due to heterogeneous preferences of refugees. Further I cannot evaluate whether, laying the negative impact of policy interventions by side, refugees tend to agglomerate more or less than other migrants. One promising way forward could be to consider different refugee crisis, where a stronger or weaker reallocation mechanism was in place. In this way one could disentangle the policy channel and the true preferences of incoming migrants. Another interesting path could be to analyze micro data on the mobility patterns after the period of obligation once residence permit is granted. This direction of research could answer whether the allocation mechanism has only temporary effects or does shape the distribution of migrants that came as refugees in the long term.

6.3 Interaction with HDI/GDP in Origin-country

Related literature has dedicated considerable effort to investigate the sorting function of networks. McKenzie and Rapoport (2010) found evidence for a higher value of networks for low-skilled migrant. In line with that, Beine et al. (2011) identify patterns of sorting. Explicitly the network effect is found to be stronger for less educated migrants. Giulietti et al. (2013) find that the less educated are more likely to obtain employment through personal contacts within a network. The general line of argument that lies behind can be sketched in this way: In average, less educated and less wealthy migrants face higher migration costs. Local networks decrease migration costs. In consequence, unskilled and/or poor migrants tend to rely more on those networks. Potential drivers can for example be credit constraints or language deficits.

Subsequently, I extend equation (15) in turn by two interaction terms to analyze whether the dataset at hand grants confirmation to preceding findings. Firstly, I interact the regional diaspora size with the log of GDP in the country of origin. Secondly, I interact the regional diaspora size with the HDI value of the origin country. Ex-ante, I expect for the coefficient of both interaction terms to have a negative sign, since a higher state of development should increase average skill-level amongst migrants and decrease reliance on networks. This leads to the following equations:

$$ln\ m_{jk}^t = \alpha + \gamma\ ln(1 + M_{jk}^t) - \lambda_2\ ln(1 + M_{jk}^t) * \text{GDP}_j^t + \beta O_j^t + \mu_j + \mu_k^t + \eta_{jk}^t \tag{17}$$

$$ln\ m_{jk}^t = \alpha + \gamma\ ln(1 + M_{jk}^t) - \lambda_3\ ln(1 + M_{jk}^t) * \text{HDI}_j^t + \beta O_j^t + \mu_j + \mu_k^t + \eta_{jk}^t \tag{18}$$

It is obvious that both measures capture many different effects. Therefore, it is difficult to break down any results into clear causal channels. GDP per capita has been used to proxy for a broad array of factors. For once, wealth generally only plays a role when it is devoted to a cause. Secondly, it is closely related to an array of other factors, such as domestic political conditions and the development of financial markets. For decades, the GDP has been the main tool to measure the state of development of a nation, in all relevant dimensions, including education and health. The later created, more inclusive human development index adds health in terms of life expectation at birth, and education, as the average years of schooling at the age of 25, to the purely economic dimension. By controlling for the domestic political situation through the war and freedom measures we can purge most political effects out of the estimated coefficients. Additionally, when estimating the GDP interaction, HDI will control for the health and educational dimension. Vice versa GDP will control for the economic dimension of the HDI. However, the estimated models remain rather descriptive in nature and largely cover similar constructs of determinants, only differing in their weights.

A second issue relates to the heterogeneity within a country. GDP and HDI are by construction equal for all nationals. In the presence of some sorting mechanism, those that migrate could systematically differ from those that do not. Both constructs would then embody only little information about migrants. I assume that both measures are reasonable approximations for the average level among migrants.

Table 5 column 5 shows the results for the estimation with GDP-diaspora interaction. I find that the network effect tends to increase with higher GDP at origin. At the same time the network coefficient becomes very small and insignificant. Emphazising the discussed underlying endogeneity issues.

Table 5 column 6 presents similar results for the HDI variable. The effect of a larger diaspora increases with higher HDI at origin. Additionally, the coefficients of the war dummy and GDP at origin are not estimated statistically significant any more, while their size remains stable. This is likely due to collinearity of the different origin-year varying measures, as discussed above. Especially when omitted variables are likely to play a role, it is difficult to empirically disentangle the effects of the various determinants, since they are highly correlated.

Disregarding the doubts on the consistency of the estimates, I find some evidence for an increasing network effect for migrants originating from higher developed countries based on both, GDP and HDI. This finding is to some degree at odds with the literature. McKenzie and Rapoport (2010), Beine et al. (2011) and Giulietti et al. (2013) find evidence for an increasing importance of a network for low-skilled migrants. It can be argued that the applied measures here, GDP and HDI, do not sufficiently grasp the skill dimension of migrants and hence differ from former estimated models. Additionally, the outcome could be the result of the failure of the homogeneity assumption within the origin country. If increasing development reinforces the sorting of migrants, migrants originating from more developed countries could in fact be less skilled, and therefore rely more on networks.

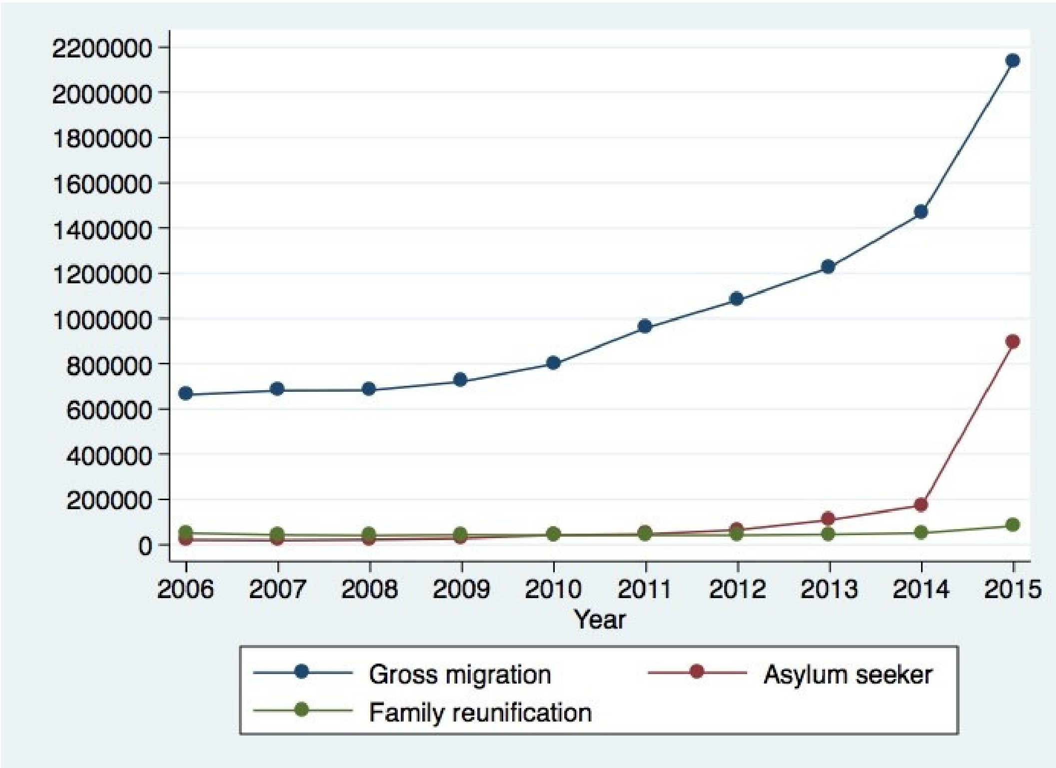
6.4 Disentangling Network Effects

As discussed in section 2, one can fundamentally separate the mechanism through which migrant networks benefit future migrants into two channels. First, the assimilation channel contains all private costs such as social and cultural adjustment costs, finding employment and housing and create a new livelihood. Secondly, the policy channel refers to the overcoming of legal barriers that migrants face before crossing the frontier. Especially, in the frame of family reunification programs, migrants that already possess a German residence permit or citizenship are eligible to sponsor further family members. In this section I implement an identification strategy that is able to disentangle these two effects, using macro data on migration flows and diasporas. The applied approach is in large parts inspired by the identification strategy in (Beine et al. (2015)). Using US census data for metropolitan regions, they find that the assimilation channel (the policy channel) is associated with 73% (27%) of the total network effect. It must be noted that these findings largely depend on the structure of bilateral migration laws, and hence are likely to differ in the German context. Especially, given the less restrictive migration law in Germany relative to the US, it can be expected that the policy channel is less relevant. Figure 2 illustrates the relative importance of the family reunification channel in Germany. The share of migrants entered through family reunification programs decreased gradually from 7.6% in 2006 to only 3.9% in 2015. To put that number in relation, in the reference work Beine et al. (2011) state that about 70% of legal migrants were admitted via family reunification programs in the year 2000 in the US.

I exploit the fact that network externalities for migrants are driven by different entities for assimilation and policy channel. The probability for a migrant to obtain legal entry and residence permit through family reunification programs depends on the total size of the network already present in Germany, not on the distribution within Germany across different cities. Oppositely, the assimilation channel is mainly regionally operating. Much of the assistance and information transferred, for example regarding housing or jobs, are just of value in the regional context [see for example (Comola and Mendola (2015); Giulietti et al. (2013); Mayer and Puller (2008)). On the statistical grounds, the two channels could not be identified if the geographic distribution of new immigrants or network members was either randomly distributed across cities, or, at the other extreme, highly concentrated. While the beneficiaries are likely to reside with their sponsors, a considerable part (especiall eligible second degree relatives) will settle in other cities, following for example housing or job market dynamics. This condition should therefore clearly be fulfilled.

Assuming the same logarithmic functional form of network externalities for each origin j and for assimilation and policy channel alike, I can estimate their relative relevance. I decompose the migration costs as in equation (8), but allow now for separate effects of the

Figure 1: Development of Gross Migration, Asylum Seekers and Family Reunification in Germany



Note: Underlying numbers are retrieved from the German Federal Office for Statistics. Y-axis: Number of Migrants for each category. X-axis: Years according to the timeframe considered in this work from 2006 till 2015 regional diaspora M^t_{jk} and the national diaspora M^t_j in the following way:[14]

$$C^t_{jk} = d^t_j - \gamma_1\, ln(1 + M^t_{jk}) + T^t_j - \gamma_2 ln(1 + M^t_j). \tag{19}$$

Plugging (13) into (7) and applying the usual transformation to the econometric specification leads to

$$lnm^t_{jk} = \alpha + \gamma_1\, ln(1 + M^t_{jk}) + \gamma_2\, ln(1 + M^t_j) + \beta O^t_j + \mu_j + \mu^t_k + \eta^t_{jk} \tag{20}$$

where γ_1 and γ_2 are respectively the coefficients of assimilation and policy channel. The intuition of this specification is the following: increasing the size of the regional diaspora, while holding the total diaspora size fixed will only influence migration through the assimilation channel. Vice versa, variation of the total national diaspora, while regional diasporas remain stable only impacts upon migration through the policy channel. While hypothesizing that the assimilation channel is mostly operating regionally, it is likely that some part information and assistance could be nationally valuable. For example, information regarding general legislative procedure in Germany if one is to apply for a residence permit do not need to come from a friend at the local level. The estimated coefficient of the policy channel should hence be seen as upper bound. I limit the analysis to the relevant subsample of origin countries which are not part of Schengen, and thus do face significant policy related migration costs.

Table 5 column 7 presents the estimation of the latter specification. The coefficient for the regional diasporas remains stable at around 0.73, while the total national diaspora is not found to have a significant impact on migration. While it is surprising that the policy channel is found to be irrelevant with this identification strategy, it is to some degree consistent with the expectation of significantly lower impact than revealed in cited literature. Generally, the results should be considered with the highest degree of caution. As highlighted in section 6.1, presented estimation is likely to suffer omitted variable bias. Other, unobserved origin-year varying factors could be correlated with the covariates and the error term in (14). Hence, any clear-cut conclusions could be fallacious.

Additionally, it must be noted that migration streams due to family reunification are marked by a high degree of inertia. Figure 1 illustrates that the main growth in gross migration was driven by incoming asylum seekers in recent years. Credible estimations on how much additional migrants will follow, exploiting the possibility of joining one's family in Germany, are not available yet. Many still unknown factors will influence these numbers. However, it seems clear that the family reunification rates will substantially increase in the following years. It could be an interesting exercise to repeat this analysis at a later point and compare the results.

[14] Note that M^t_j represents the sum of all regional diasporas in Germany, not just the sum of diasporas in NRW.

7 Conclusion

The present work explores the importance of regional diasporas in explaining future migration pattern in the German federal state North-Rhine Westphalia. I first develop a simple theoretical framework that emphasizes the enforcing effect of regional migrant networks through a reduction in migration costs. Bringing the theoretical results to the data, I implement pseudo maximum likelihood estimator to quantify this network effect. The PPML strategy is well suited to deal with the specific structure of the dataset, including many zero observations and a high degree of heteroscedasticity. Strikingly different results between the OLS and PPML estimates confirm previous evidence, suggesting that OLS can produce strongly biased estimates in this context.

Estimating the theoretical justified specification with origin-year and city-year fixed effects, I find that a 10% larger regional network is related to 7% larger migration flow in the subsequent year. This results clearly highlights that any policy targeting migration must recognize the importants of regional networks.

I proceed by extending the baseline model with three different additional aspects. First, I replace the origin-year fixed effect by a set observed of origin-year control variables. This adjustment is necessary to estimate the impact of other origin-year varying factors on the network mechanism. The results emphasize the difficulty to control for all relevant push-factors by means of observables. All following findings should hence be interpreted with caution. Nevertheless, the results lend support to former empirical evidence on determinants of migration: the membership to the Schengen area increases migration, since it decreases policy related migration costs. A lower degree of political and private freedom, and the occurrence of wars in the origin country increases migration. A higher level of development decreases the incentive to migrate, once accounted for potential credit constraints.

The first extension reveals that the network effect is weaker for potential refugees. It remains open to further research to which degree this finding is driven through policies that impede the freedom of location choice, or through heterogeneous preferences of refugees. A more detailed understanding is clearly policy relevant when dealing with the recent refugee crisis in the long-run. Secondly, I find that migrants from higher developed countries rely more on network. This finding is partially at odds with the literature. It remains unclear whether this is the result of a bias due to omitted variables, or because the implemented measures do not grasp a similar concept. Thirdly, I find evidence that the policy channel is largely less relevant than the assimilation in the given context. This reflects relatively low policy related migration costs. It remains interesting to observe this development in future years: the recently accommodated refugees are likely to trigger a second wave of migrants that benefit of family reunification procedures.

References

Ai, Chunrong, and Edward Norton. 2003. "Interaction terms in logit and probit models." *Economics Letters* 80 (1): 123–129.

Beine, Michel, Simone Bertoli, and Jesús Fernández-Huertas Moraga. 2016. "A Practitioners' Guide to Gravity Models of International Migration." *The World Economy* 39 (4): 496–512.

Beine, Michel, Frédéric Docquier, and Çağlar Özden. 2011. "Diasporas." *Journal of Development Economics* 95 (1): 30–41.

———. 2015. "Dissecting Network Externalities in International Migration." *JODE - Journal of Demographic Economics* 81 (4): 379–408.

Beine, Michel, Romain Noël, and Lionel Ragot. 2014. "Determinants of the international mobility of students." *Economics of Education Review* 41 (C): 40–54.

Beine, Michel, and Christopher Parsons. 2015. "Climatic Factors as Determinants of International Migration." *Scandinavian Journal of Economics* 117 (2): 723–767.

Belot, Michèle V. K., and Timothy J. Hatton. 2012. "Immigrant Selection in the OECD." *The Scandinavian Journal of Economics* 114 (4): 1105–1128. ISSN: 1467-9442.

Berry, Steven T. 1994. "Estimating Discrete-Choice Models of Product Differentiation." *The RAND Journal of Economics* 25 (2): 242–262.

Bertoli, Simone, and Jesús Fernández-Huertas Moraga. 2012. *Visa Policies, Networks and the Cliff at the Border.* Working Papers 2012-12. FEDEA.

———. 2013. "Multilateral resistance to migration." *Journal of Development Economics* 102 (C): 79–100.

———. 2015. "The size of the cliff at the border." *Regional Science and Urban Economics* 51 (C): 1–6.

Black, Richard, Dominic Kniveton, and Kerstin Schmidt-Verkerk. 2011. "Migration and Climate Change: Towards an Integrated Assessment of Sensitivity." *Environment and Planning A* 43 (2): 431–450.

Borjas, George J. 1987. *Self-Selection and the Earnings of Immigrants.* NBER Working Papers 2248. National Bureau of Economic Research, Inc.

———. 1990. "Friends or strangers: the impact of immigrants on the US economy."

Clark, Ximena, Timothy J. Hatton, and Jeffrey G. Williamson. 2007. "Explaining U.S. Immigration, 1971 - 1998." *The Review of Economics and Statistics* 89 (2): 359–373.

Comola, Margherita, and M. Mendola. 2015. *The Formation of Migrant Networks.* PSE - Labex "OSE-Ouvrir la Science Economique". HAL.

Frankel, Jeffrey A, Ernesto Stein, and Shang-Jin Wei. 1997. *Regional trading blocs in the world economic system.* Peterson Institute.

Giulietti, Corrado, Christian Schluter, and Jackline Wahba. 2013. "With a Lot of Help from my Friends: Social Networks and Immigrants in the UK." *Population, Space and Place* 19 (6): 657–670.

Gleditsch, Nils Petter, Peter Wallensteen, Mikael Eriksson, Margareta Sollenberg, and Håvard Strand. 2002. "Armed Conflict 1946-2001: A New Dataset." *Journal of Peace Research* 39 (5): 615–637.

Greene, William. 2010. "Testing hypotheses about interaction terms in nonlinear models." *Economics Letters* 107 (2): 291–296.

Grogger, Jeffrey, and Gordon Hanson. 2011. "Income maximization and the selection and sorting of international migrants." *Journal of Development Economics* 95 (1): 42–57.

Harris, John R., and Michael P. Todaro. 1970. "Migration, Unemployment and Development: A Two-Sector Analysis." *The American Economic Review* 60 (1): 126–142.

Krugman, Paul. 1995. "Increasing returns, imperfect competition and the positive theory of international trade." Chap. 24 in *Handbook of International Economics,* 1st ed., edited by G. M. Grossman and K. Rogoff, 3:1243–1277. Elsevier.

Lewis, W. Arthur. 1954. "Economic Development with Unlimited Supplies of Labour." *The Manchester School* 22 (2): 139–191.

Massey, Douglas S. 1990. "Social Structure, Household Strategies, and the Cumulative Causation of Migration." *Population Index* 56 (1): 3–26.

Massey, Douglas S, Joaquin Arango, Graeme Hugo, Ali Kouaouci, Adela Pellegrino, and J Edward Taylor. 1993. "Theories of international migration: A review and appraisal." *Population and development review:* 431–466.

Massey, Douglas S., and Maria Aysa-Lastra. 2005. "Social capital and international migration from Latin America." In *Paper presented at the Expert group meeting on international migration and development in Latin America and the Caribbean, United Nations Secretariat.*

Mayda, Anna Maria. 2010. "International migration: a panel data analysis of the determinants of bilateral flows." *Journal of Population Economics* 23 (4): 1249–1274.

Mayer, Adalbert, and Steven Puller. 2008. "The old boy (and girl) network: Social network formation on university campuses." *Journal of Public Economics* 92 (1-2): 329–347.

McFadden, Daniel. 1974. "Conditional logit analysis of qualitative choice behavior." *Frontiers in Econometrics.*

———. 1984. "Econometric analysis of qualitative response models." Chap. 24 in *Handbook of Econometrics,* 1st ed., edited by Z. Griliches and M. D. Intriligator, 2:1395–1457. Elsevier.

McKenzie, David, and Hillel Rapoport. 2010. "Self-Selection Patterns in Mexico-U.S. Migration: The Role of Migration Networks." *The Review of Economics and Statistics* 92 (4): 811–821.

Munshi, Kaivan. 2003. "Networks in the Modern Economy: Mexican Migrants in the U. S. Labor Market." *The Quarterly Journal of Economics* 118 (2): 549–599.

Myrdal, Gunnar. 1957. *Economic theory and under-developed regions.* Technical report. Methuen.

Ortega, Francesc, and Giovanni Peri. 2013. "The effect of income and immigration policies on international migration." *Migration Studies* 1 (1): 47–74.

Ottaviano, Gianmarco I.P., and Giovanni Peri. 2006. "The economic value of cultural diversity: evidence from US cities." *Journal of Economic Geography* 6 (1): 9–44.

Patacchini, Eleonora, and Yves Zenou. 2012. "Ethnic networks and employment outcomes." *Regional Science and Urban Economics* 42 (6): 938–949.

Ramos, Raul. 2016. "Gravity models: A tool for migration analysis." *IZA World of Labor* (February): 239–239.

Ranis, Gustav, and John C. H. Fei. 1961. "A Theory of Economic Development." *The American Economic Review* 51 (4): 533–565.

Ruist, Joakim, Jan Stuhler, and David A Jaeger. 2017. "Shift-Share Instruments and the Impact of Immigration–preliminary–."

Santos Silva, João, and Silvana Tenreyro. 2006. "The Log of Gravity." *The Review of Economics and Statistics* 88 (4): 641–658.

———. 2010. "On the existence of the maximum likelihood estimates in Poisson regression." *Economics Letters* 107 (2): 310–312.

———. 2011. "Further simulation evidence on the performance of the Poisson pseudo-maximum likelihood estimator." *Economics Letters* 112 (2): 220–222.

———. 2015. *PPML: Stata module to perform Poisson pseudo-maximum likelihood estimation.*

Santos Silva, João, Silvana Tenreyro, and Frank Windmeijer. 2015. "Testing Competing Models for Non-negative Data with Many Zeros." *Journal of Econometric Methods* 4 (1).

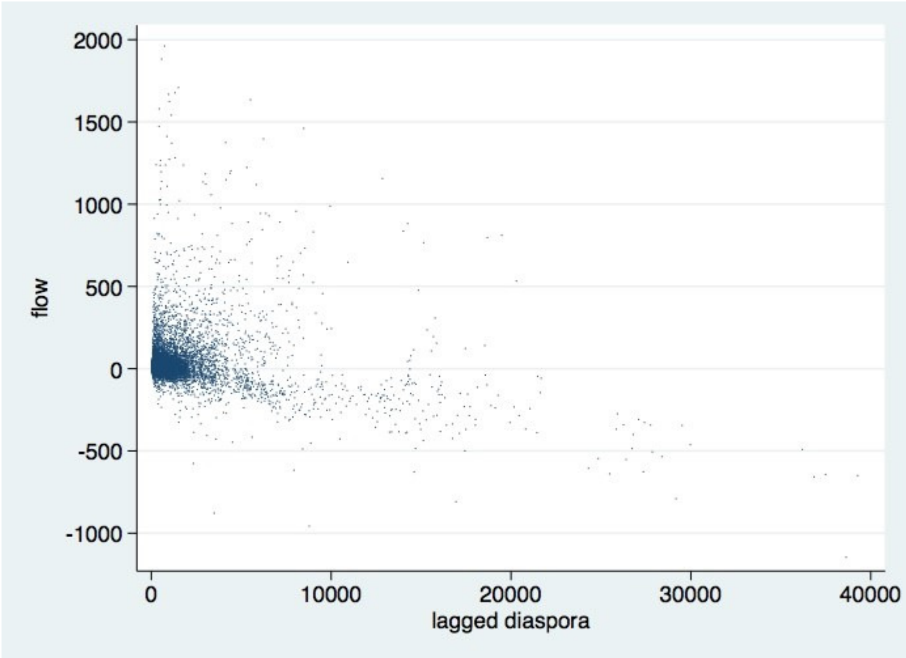
Sjaastad, Larry A. 1962. "The Costs and Returns of Human Migration." *Journal of Political Economy* 70 (5): 80–93.

Todaro, Michael P. 1969. "A Model of Labor Migration and Urban Unemployment in Less Developed Countries." *The American Economic Review* 59 (1): 138–148.

Todaro, Michael P. 1976. *Internal migration in developing countries.* International Labour Office.

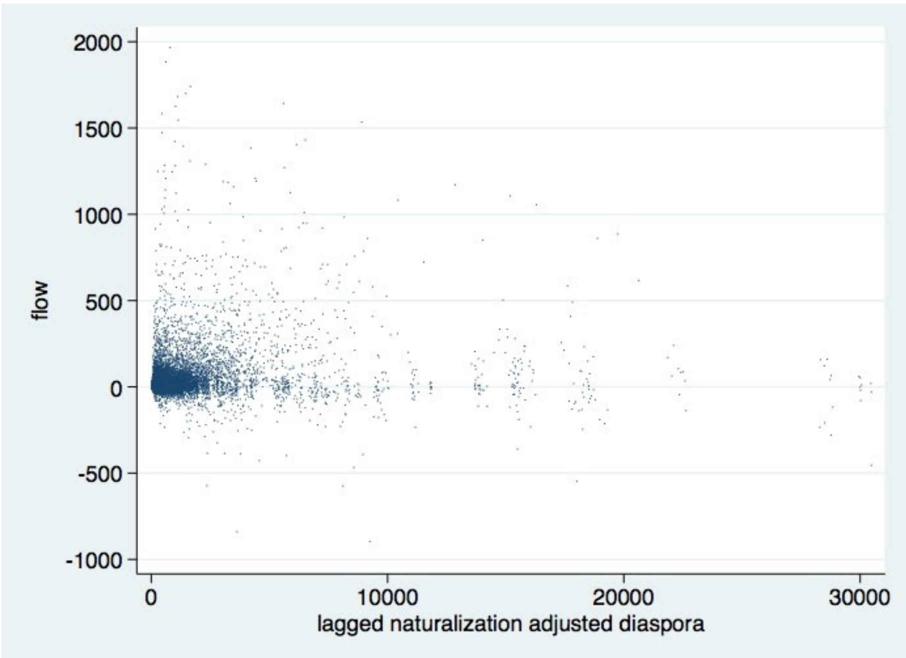
Appendix

Figure 2: Scatter Plot of Flows and Regional Diasporas for the Full Dataset not adjusted for Naturalization



Note: Y-axis: Flows calculated as the difference in stocks between two subsequent years. X-axis: Regional stocks of migrants of the same nationality.

Figure 3: Scatter Plot of Flows and Regional Diasporas for the Subsample Naturalization adjusted Diasporas



Note: Y-axis: Flows calculated as the difference in naturalization augmented stocks between two subsequent years. X-axis: Regional naturalization augmented stocks of migrants of the same nationality.

Table 6: Introducing Dyadic Fixed Effects

	1	2	3	4	5
dependent var.	flow	flow	flow	flow	lnflow
Lagged ln diaspora	.6275*** (.0131)	.7239*** (.0470)	-.2569*** (.0633)	-.2643*** (.0426)	-.2369*** (.0215)
City FE	yes	no	no	no	no
Origin FE	yes	no	no	no	no
Year FE	yes	no	no	no	no
Year*Origin FE	no	no	yes	yes	yes
Year*City FE	no	yes	no	yes	yes
City*Origin FE	yes	yes	yes	yes	yes
R^2				.9340	.8250
Nb. obs.	48927	48927	47849	47912	29253
Dataset	full	full	full	full	full
Method	PPML	PPML	PPML	PPML	OLS

Note: Standard errors are in parenthesis. ***, **, * significant at 1%, 5%, 10% level. Standard errors calculated as Efron's pseudo R^2 for maximum likelihood estimations. Flows are calculated as the difference between the regional diasporas in two subsequent years.