Economics Working Papers

2015-03

Networks and Selection in International Migration to Spain

Nina Neubecker, Marcel Smolka and Anne Steinbacher

AARHUS
UNIVERSITY
BUSINESS AND SOCIAL SCIENCES
DEPARTMENT OF ECONOMICS AND BUSINESS

Networks and Selection in International Migration to Spain*

Nina Neubecker‡ Marcel Smolka§ Anne Steinbacher¶

DIW Berlin Aarhus University

This version: January 21, 2015; first version: May 2012

Abstract

This paper provides new evidence on migrant networks as determinants of the scale and skill structure of migration, using aggregate data from a recent migration boom to Spain. We develop a three-level nested multinomial logit migration model. Our model accommodates varying degrees of similarity of destinations located in the same region (or the same country), allowing for a rich structure of substitutability across alternative destinations. We find strong positive network effects on the scale of migration and a strong negative effect on the ratio of high-skilled to low-skilled migrants. Simplifying restrictions on substitutability across destinations are rejected by the data.

JEL Codes: F22, J61

Keywords: international migration · migrant networks · nested multinomial logit model · skill structure of migration · Spain

*This paper is a revised version of DIW Discussion Paper No. 1306 (published in May 2013). The first version of this paper was published in May 2012 as a University of Tübingen Working Paper in Economics and Finance (No. 35). We have benefitted from valuable suggestions by Wilhelm Kohler, Udo Kreickemeier, Peter Eppinger, Gordon Hanson, Johannes Pfeifer, Melissa Siegel, two anonymous referees, as well as seminar and conference participants in Copenhagen, Göttingen, Tübingen, Granada, Beijing, Stuttgart-Hohenheim, and Munich. Research assistants at the University of Tübingen and the University of Aarhus have provided excellent support. Marcel Smolka gratefully acknowledges financial support from the Tuborg Foundation, as well as from the Volkswagen Foundation under the project "Europe's Global Linkages and the Impact of the Financial Crisis". Part of the work on this paper was done while Marcel Smolka was a visiting PhD student at University College London (UCL). The hospitality of the Department of Economics and the Centre for Research and Analysis of Migration (CReAM) at UCL is gratefully acknowledged.

‡German Institute for Economic Research, DIW Berlin, Mohrenstraße 58, 10117 Berlin, Germany.

§**Corresponding author:** Department of Economics and Business, Aarhus University, Fuglesangs Allé 4, Building 2632, 8210 Aarhus V, Denmark. E-mail: msmolka@econ.au.dk, Phone: +4 8716 4974.

¶No longer affiliated with an academic institution.

1 Introduction

An established body of literature argues that already settled migrants, often simply called a migrant network, facilitate migration for prospective newcomers, for example through informal job referrals among co-national peers (Munshi, 2003).[1] In this paper, we provide new evidence on migrant networks as determinants of the total size (scale) and skill structure of migration, drawing on aggregate data from a recent migration boom to Spain. Spain is an interesting case to look at. The country developed into one of the world's most attractive destinations for migrants due to its strong economic growth ahead of the Global Financial Crisis. From 1997 to 2009, Spain received roughly six million new migrants.[2] The foreign-born share among the total population has increased dramatically over the past few years, starting out from 4.9% in 2000 and approaching 14.1% in 2008 (OECD, 2010, 240).

The model we develop to identify network effects in migration is a three-level nested multinomial logit (NMNL) model along the lines of McFadden (1984, 1422-1428). The main feature of our model is the rich structure of permissible substitution patterns across alternative migration destinations. The basic idea is that migration destinations located in the same country or region are easier to substitute for one another because they are "similar" (to varying degrees, which we model through heterogeneous similarity parameters[3]): they share the same legal and political framework; they have a common cultural background; they engage in similar economic activities and so on. The model we use is thus more general than the standard multinomial logit (MNL) model (McFadden, 1984, 1411-1415) which features a uniform degree of cross-destination substitutability.

More specifically, our model allows for unobserved heterogeneity at the individual-level (such as productivity, language skills, or taste idiosyncrasies) to flexibly interact with the characteristics of alternative migration destinations, and to do so at different hierarchical levels: countries, regions, and provinces. Take, for instance, individuals that differ by age (assumed unobservable). Young professionals will find those provinces attractive that have low income tax rates. Middle-aged people with young children, in contrast, will care more about the quality of public schools. And elderly people will perhaps look for good climatic conditions for their retirement home. These important variables are not randomly distributed across provinces, but rather, they vary (sharply or roughly) by countries and regions.[4] This is especially true for all policy variables. Not only must these variables, therefore,

[1] Massey (1988, 396) defines migrant networks as "[...] sets of interpersonal ties that link migrants, former migrants, and nonmigrants in origin and destination areas through the bonds of kinship, friendship, and shared community origin."

[2] Of these migrants, 13.6% are Romanians, followed by Moroccans (11.1%), Ecuadorians (8.2%), Colombians (6.1%), Britons (5.3%), and Bolivians (4.7%). Unless stated otherwise, all migration figures in this paper are own calculations based on data from the Spanish Instituto Nacional de Estadística (INE).

[3] To the best of our knowledge, no other random utility model that could be estimated with our data would allow us to do likewise. For example, the generalized nested logit (GNL) model by Wen & Koppelman (2001) could be used to closely approximate our three-level NMNL, but its estimation is not feasible with our data.

[4] Other margins of variation are possible, too. For example, some individuals might have strong preferences for mega-

be controlled for in the estimation, but their interactions with the (unobserved) individual-specific heterogeneity must also be taken into account. We show that failing to do so risks introducing an omitted variables bias in the estimation of the determinants of migration based on aggregate migration data.

The Spanish case is well-suited for an empirical study of network effects in migration. The data available from the Spanish Instituto Nacional de Estadística (INE) are of exceptional quality and coverage. They allow us to exploit variation in migration across a large number of countries of origin, as well as across all regions and provinces of destination in Spain. Although we focus on a single county of destination, the identifying variation is thus of a bilateral nature, and we can use a rich structure of fixed effects in order to control for confounding factors (such as bilateral migration policies, which are often difficult to observe and measure), and provide estimates that are consistent with the rich structure of cross-destination substitutability featured in our NMNL model.

Obtaining consistent and unbiased estimates of network effects in migration is not trivial. The main endogeneity concern is the two-way relationship between migration costs and migrant networks, defined as the number of migrants from a certain nationality that are already settled in a certain destination. On the one hand, the migrant network appears as an argument in the migration cost function determining future migration. On the other hand, the migrant network is the result of past migration, and is thus itself influenced by migration costs. As our data distinguish among different countries of origin as well as among different provinces of destination in Spain, we can go beyond the existing literature in the way we control for unobserved heterogeneity in migration costs through fixed effects. For example, if people from Latin America (independently of the exact country of origin) are generally more welcome in, say, Barcelona than in other provinces in Spain, then we can control for this in the estimation. To further strengthen our analysis, we instrument migrant networks by historical internal migration flows in Spain.

Our estimates reveal robustly positive network effects on the scale of migration. The effects are of considerable size, and overall similar to those reported in the received literature. Since individual migration moves are independent of the effect they have on the migration decisions of others, our results have important policy implications. In a dynamic model of labor migration, network effects indicate a welfare loss in the laissez-faire transition path equilibrium (Carrington et al., 1996; Chau, 1997). From the perspective of a social planner who wants to maximize world welfare, they call for migration subsidies that accelerate the speed of migration. Our estimates also attest to strong negative

cities such as London or New York. Testing for these alternative structures is beyond the scope of our paper. However, we believe that the structure we impose in terms of *territorial entities* (countries, regions, and provinces) is a natural and plausible choice, for the reasons given above.

effects of migrant networks on the skill structure of migration, defined as the ratio of high-skilled to low-skilled migrants. This finding accords with the idea that high-skilled individuals have lower effective migration costs than low-skilled individuals (Chiswick, 1999). Intuitively, migrant networks are more important for low-skilled than for high-skilled individuals, and thus bias the skill structure of migration toward the low-skilled ones.

Our estimates strongly reject a uniform degree of substitutability across alternative destinations, working against the standard MNL model in our application to the Spanish case. We find pronounced heterogeneity in the estimated network coefficients (reflecting heterogeneous similarity parameters across regions), an observation that has (to the best of our knowledge) received no attention so far in the literature. We use the structural interpretation of our network coefficients in order to exploit this heterogeneity and compute elasticity values for the network effect. The estimated elasticity is lowest for the destinations located in the region of Extremadura, slightly exceeding a value of 0.1. It is highest for the destinations located in the region of Cataluña, lying in the vicinity of 0.55. We conclude from our results that the ease with which one destination can be substituted for another one is highest in the region of Cataluña, arguably the region with the highest degree of political and cultural autonomy in Spain.

Our paper is related to recent estimates of network effects based on aggregate migration data. Beine et al. (2011) investigate the determinants of the scale and skill structure of migration between the years 1990 and 2000 to 30 OECD countries. They find that economies that already host migrants from a given country attract both a larger number of new migrants as well as a larger fraction of low-skilled migrants from that country.[5] Similar results are obtained by Beine & Salomone (2013) who study potential gender differences in network effects. The paper by Beine et al. (2012) disentangles what the authors call local and national network externalities, saying that local migrant networks facilitate the assimilation of migrants in the host society, while nation-wide migrant networks help overcome the legal entry barriers to migration. However, all of these papers derive the estimated migration functions from a standard MNL model that assumes a uniform degree of cross-destination substitutability.[6]

Our paper is also related to a number of macro-level studies that are more generally concerned

[5]See also Grogger & Hanson (2011, 53) for complementary evidence. McKenzie & Rapoport (2010) find positive self-selection on education from Mexican migrants to the U.S. to be more likely, the larger the number of return migrants in the origin community. Bertoli (2010) finds a positive interaction between the number of migrants abroad and the extent of negative self-selection, using individual-level data on Ecuadorian emigrants. González & Ortega (2011, 2013) as well as Farré et al. (2011) show that historical networks in Spain can be used to instrument for recent migration flows.

[6]While revising this paper, we became aware of research by Bertoli & Fernández-Huertas Moraga (2012) (published as Bertoli & Fernández-Huertas Moraga, 2015). They use the same migration data as Beine et al. (2011) in order to estimate network effects in migration, relaxing the assumption of a uniform degree of substitutability across alternative destinations. The most general version of their estimated model reduces to a two-level NMNL model with a homogeneous similarity parameter for all "nests" (countries and regions in our paper); see our online Addendum A for details.

with the determinants of international migration.[7] In this literature, migrant networks robustly rank among the most important factors shaping migration, but the estimated migration functions often lack an explicit micro-foundation (Clark et al., 2007; Lewer & Van den Berg, 2008; Pedersen et al., 2008; Mayda, 2010). Two recent papers, Bertoli & Fernández-Huertas Moraga (2013) and Ortega & Peri (2013), develop micro-founded random utility migration models in order to estimate the determinants of migration. In both papers, the standard MNL assumption of a uniform degree of cross-destination substitutability is relaxed. Bertoli & Fernández-Huertas Moraga (2013) use the same Spanish data source as we do in this paper. They argue that the Common Correlated Effects (CCE) estimator, a panel estimator proposed by Pesaran (2006), yields consistent estimates of the migration function under arbitrary specifications of the cross-nested logit (CNL) model due to Vovsha (1997). The CNL model allocates a "portion" of each destination to a set of "nests" (countries and regions in our paper), assuming, contrary to our model, that there is a single similarity parameter for all nests.[8] Ortega & Peri (2013) investigate the impact of income and immigration policies on migration to OECD countries, using panel data detailed by country of origin and country of destination.[9] Their model, best understood as a two-level NMNL model with a single similarity parameter for all nests, allows for a higher degree of substitutability across destinations that are located outside the individual's country of origin. However, neither Bertoli & Fernández-Huertas Moraga (2013) nor Ortega & Peri (2013) identify the effects of migrant networks on the scale and skill structure of migration, as we do in this paper.

The remainder of this paper is organized as follows. Section 2 characterizes individual decision making in a three-level NMNL model. This model allows us to derive estimable equations for the scale and skill structure of migration. In Section 3 we present our estimation strategy and we introduce the data we employ in our econometric analysis. In Section 4 we present our estimation results and we provide a structural interpretation of these results in terms of our NMNL migration model. Section 5 concludes.

[7] For the location choice of migrants within borders, see Bartel (1989), Zavodny (1997, 1999), Chiswick & Miller (2004), Card & Lewis (2007), and Jayet et al. (2010). Selected survey-based studies on migration decisions at the micro-level include Åslund (2005), Baghdadi (2005), Bauer et al. (2005, 2009), and Dolfin & Genicot (2010).

[8] The CNL model is a special case of the GNL model. Unlike the GNL model, the CNL model cannot be used to approximate our three-level NMNL model; see Wen & Koppelman (2001). Bertoli et al. (2013) employ the CNL model in order to study the effect of the recent economic crisis in Europe on migration to Germany.

[9] In Ortega & Peri (2009), a previous version of Ortega & Peri (2013), the authors also study the effects of migration on employment, investment, and productivity.

2 The Model

In this section we develop a multi-country random utility framework with many countries of origin and many provinces of destination within countries.

2.1 Basic Setup

We assume that the decision making process leading to migration follows a hierarchical structure in which provinces (the final migration destinations) are grouped into higher-level territorial entities (nests). Individuals "eliminate" nests until a single province remains. Decision making can be described in a hierarchical manner[10]: first to which country to migrate (including the country of origin); second which region to move to within the chosen country; and third which province to pick within the preferred region.[11] We index the countries of origin by $i = 1, \ldots, I$; the countries of destination (the primary nests) by z or $y = 1, \ldots, Z$; the regions of destination (the secondary nests) by r or $\ell = 1, \ldots, R$; and the provinces of destination by j or $k = 1, \ldots, J$.[12] Let the country of origin i be one element in each of the sets $\{1, \ldots, Z\}$, $\{1, \ldots, R\}$, and $\{1, \ldots, J\}$, thus representing a degenerate nest with a single final migration destination. Define A_{zr} as the set of provinces in region r of country z, and A_z as the set of regions in country z.

We write the utility of individual o who migrates from country i to province j as:

$$U_{ij}^{o} = Y_j - C_{ij} + e_{ij}^{o}, \tag{1}$$

where the index $o = 1, \ldots, m_i$, identifies individuals originating from country i, the terms Y_j and C_{ij} are sub-utility functions for moving from country i to province j, and the term e_{ij}^{o} is a stochastic (random) utility variable with idiosyncratic realizations for each province $j = 1, \ldots, J$. This variable reflects any type of unobserved individual-specific heterogeneity that influences an individual's decision to migrate (age, productivity, family status, occupation etc.). The function Y_j summarizes characteristics of province j such as the wage rate, the state of the housing market, or the climate. It is assumed independent of the individual's country of origin. The function C_{ij} captures the costs of moving and assimilation, henceforth called migration costs. Similar to Beine et al. (2011, 33-34), we hypothesize that these costs are a decreasing and globally convex function of the migrant network, M_{ij}, defined

[10] We assume that each decision in this hierarchy is made conditional on both the fixed preceding decisions and the optimal succeeding decisions. Hence, individuals decide on all aspects of their migration moves simultaneously (cf. Domencich & McFadden, 1975, 33-46).

[11] In Ortega & Peri (2013), the first decision is between migrating abroad or staying at home. Our estimation is compatible with this additional structure.

[12] Strictly speaking, the provinces j and the nests r and z are i-specific. We omit this index in order to avoid notational clutter.

as the number of co-national migrants already settled in province j. A convenient specification of migration costs that incorporates the idea of positive but diminishing returns to the migrant network uses the log of M_{ij}:

$$C_{ij} = c_{iz} + c_{ir} + c_{ij} - \theta \ln(1 + M_{ij}), \quad j \in A_{zr}, r \in A_z, \tag{2}$$

where the parameter $\theta > 0$ is a measure for the strength of the network effect, and where we add one to the variable M_{ij} before taking logs in order to abstract from infinitely large migration costs. The other cost components not related to the migrant network will be described in more detail below. Suffice it to say here that, for a given country of origin i, they vary across either countries of destination (c_{iz}), regions of destination (c_{ir}), or provinces of destination (c_{ij}). For expositional convenience, we define $U_{ij} \equiv U^o_{ij} - e^o_{ij} = Y_j - C_{ij}$ and $\xi_{ij} \equiv Y_j - c_{ij} + \theta \ln(1 + M_{ij})$.

Individuals are assumed to choose from the set of provinces the alternative from which they derive the highest utility:

$$j^o = \text{argmax}(U^o_{i1}, \ldots, U^o_{iJ}), \quad j^o \in \{1, \ldots, J\}. \tag{3}$$

The probability that individual o from country i migrates to province j is equal to the probability that this individual associates the largest utility with moving to province j:

$$\begin{aligned} P^o_i(j^o = j) &= \Pr(U^o_{ij} > U^o_{ik} \ \forall\, k \in \{1, \ldots, J\} : k \neq j) \\ &= \Pr(e^o_{ik} - e^o_{ij} < U_{ij} - U_{ik}; \\ &\quad \forall\, k \in \{1, \ldots, J\} : k \neq j). \end{aligned} \tag{4}$$

By the laws of conditional probability, we can express this probability as a product of transition probabilities:

$$P^o_i(j^o = j) = P^o_i(j^o = j | j^o \in A_{zr}) P^o_i(j^o \in A_{zr} | r \in A_z) P^o_i(r \in A_z), \quad j \in A_{zr}, r \in A_z. \tag{5}$$

These probabilities depend on the distribution assumed for the random utility variables, $e^o_{i1}, \ldots, e^o_{iJ}$. Let $\mathbf{g}_i = (g_{i1}, \ldots, g_{iJ})$ be a $(1 \times J)$ row vector with non-negative entries, and let H_i be a non-negative function of $\mathbf{g}_i$ with:

$$\lim_{g_{ij} \to \infty} H_i(\mathbf{g}_i) = +\infty \quad \text{for } j = 1, \ldots, J. \tag{6}$$

Furthermore, assume that H_i is homogeneous of degree one in $\mathbf{g}_i$, and let H_i have mixed partial derivatives of all orders, with non-positive even and non-negative odd mixed derivatives. It can be

shown that the function

$$F_i(e^o_{i1},\ldots,e^o_{iJ}) = \exp\left[-H_i\left(\exp[-e^o_{i1}],\ldots,\exp[-e^o_{1J}]\right)\right] \tag{7}$$

is a multivariate extreme value distribution function, and that, if $(e^o_{i1},\ldots,e^o_{iJ})$ is distributed F_i, (4) can be written as:

$$\begin{aligned} P^o_i(j^o=j) &= \frac{\exp[U_{ij}]}{H_i(\exp[U_{i1}],\ldots,\exp[U_{iJ}])}\frac{\partial H_i(\exp[U_{i1}],\ldots,\exp[U_{iJ}])}{\partial \exp[U_{ij}]} \\ &= \frac{\partial \ln H_i(\exp[U_{i1}],\ldots,\exp[U_{iJ}])}{\partial U_{ij}}; \end{aligned} \tag{8}$$

see McFadden (1978, 80-81; 1981, 226-230).[13]

We depart from the received literature in that we introduce a function H_i that generates the response probabilities of a three-level NMNL model. It allows for the random utilities associated with provinces in the same region (or the same country) to be mutually correlated, whereas the random utilities associated with provinces in different countries are independent. This means that an individual that has strong preferences for a certain destination j is likely to also have stronger preferences for other destinations in the same region (or country) as destination j. The strength of this effect depends on how "homogeneous" the region/country is (i.e. how similar the provinces are that belong to this region/country).

Define on the half-open unit interval two parameters, λ_z and κ_r $(0<\kappa_r,\lambda_z\leq 1)$, measuring the similarity of the provinces in country z and region r, respectively. These two parameters govern the degree of substitutability across alternative migration destinations. High parameter values indicate little similarity among provinces (and weak correlations among the random utilities), low parameter values indicate much similarity (and strong correlations). We thus assume:

$$\begin{aligned} H_i(\exp[U_{i1}],\ldots,\exp[U_{iJ}]) &= \sum_z\left(\sum_{r\in A_z}\left(\sum_{j\in A_{zr}}\exp[U_{ij}/(\kappa_r\lambda_z)]\right)^{\kappa_r}\right)^{\lambda_z} \\ &= \sum_z \exp[-c_{iz}]\left(\sum_{r\in A_z}\exp[-c_{ir}/\lambda_z]\left(\sum_{j\in A_{zr}}\exp[\xi_{ij}/(\kappa_r\lambda_z)]\right)^{\kappa_r}\right)^{\lambda_z}. \end{aligned} \tag{9}$$

It is instructive to note that the function $H_i(\cdot)$ nests the generating function for the response probabilities of the standard MNL model as a special case with $\kappa_r=\lambda_z=1\ \forall r,z$. This rules out any correlation among the random utilities. We will return to this in more detail below. From equations

[13] We show in our online Addendum B how to derive (8).

(8) and (9) it follows that each transition probability in equation (5) has a closed-form analytical solution[14]:

$$P_i^o(r \in A_z) = \exp[\Omega_{iz}\lambda_z - c_{iz} - \Psi_i], \tag{10}$$

$$P_i^o(j^o \in A_{zr} | r \in A_z) = \exp[\Phi_{ir}\kappa_r - c_{ir}/\lambda_z - \Omega_{iz}], \tag{11}$$

$$P_i^o(j^o = j | j^o \in A_{zr}) = \exp[\xi_{ij}/(\lambda_z\kappa_r) - \Phi_{ir}], \tag{12}$$

where Φ_{ir}, Ω_{iz}, and Ψ_i are "inclusive values" defined as:

$$\Phi_{ir} \equiv \ln \sum_{k \in A_{zr}} \exp[\xi_{ik}/(\lambda_z\kappa_r)], \tag{13}$$

$$\Omega_{iz} \equiv \ln \sum_{\ell \in A_z} \exp[\Phi_{i\ell}\kappa_\ell - c_{i\ell}/\lambda_z], \tag{14}$$

$$\Psi_i \equiv \ln \sum_z \exp[\Omega_{iz}\lambda_z - c_{iz}]. \tag{15}$$

The inclusive values Φ_{ir}, Ω_{iz}, and Ψ_i summarize the characteristics of all provinces in region r, all provinces in country z, and all provinces in the complete set of final migration destinations, respectively. Using equation (5) along with equations (10) to (15) and aggregating over all individuals from country i, we can write the expected rate of migration from country i to province j as:

$$\frac{m_{ij}}{m_i} = \frac{\exp[\xi_{ij}/(\lambda_z\kappa_r) - c_{ir}/\lambda_z - c_{iz}]}{\exp[\Psi_i + (1-\kappa_r)\Phi_{ir} + (1-\lambda_z)\Omega_{iz}]}, \tag{16}$$

where m_{ij} is the number of individuals migrating from i to j, and m_i is the initial population in country i. This ij-specific migration rate depends on the inclusive values Φ_{ir}, Ω_{iz}, and Ψ_i. It is therefore responsive to the attractiveness of all provinces $k = 1, \ldots, J$, whether in the same region r (or the same country z) as province j or not.[15] For example, consider the elasticity of the ij-specific migration rate, m_{ij}/m_i, with respect to Y_k, the characteristics of province k, where $j \in A_{zr}, r \in A_z$,

[14]For example, in order to derive $P_i^o(r \in A_z)$, one has to compute $\partial \ln H_i(\cdot)/\partial(-c_{iz})$, and similarly for the other transitional probabilities. We show in our online Addendum C how to compute $P_i^o(j^o = j) = \partial \ln H_i(\cdot)/\partial U_{ij}$.

[15]One might refer to the inclusive values as "multilateral resistance" terms; see Bertoli & Fernández-Huertas Moraga for a discussion of multilateral resistance to migration, and Anderson & van Wincoop (2003) for multilateral resistance in the gravity equation for international trade flows. Mayda (2010) speaks of "multilateral pull" effects. Anderson (2011) sketches a general equilibrium migration model with multilateral resistance. See also Hanson (2010, 4373-4375) for a discussion.

and $k \in A_{y\ell}, \ell \in A_y$. Straightforward though cumbersome differentiation yields[16]:

$$\begin{aligned}\frac{\partial \ln(m_{ij}/m_i)}{\partial \ln(Y_k)} &= Y_k \left[\frac{I(j,k)}{\lambda_z \kappa_r} - \left(\frac{m_{ik}}{m_i} \right) \right. \\ &\quad \left. - \frac{I(\ell,r)}{\lambda_z \kappa_r}(1-\kappa_r) \left(\frac{m_{ik}}{m_{ir}} \right) - \frac{I(y,z)}{\lambda_z}(1-\lambda_z) \left(\frac{m_{ik}}{m_{iz}} \right) \right], \end{aligned} \tag{17}$$

where $m_{ir} = \sum_{j \in A_{zr}} m_{ij}$, $m_{iz} = \sum_{r \in A_z} m_{ir}$, and $I(a,b) = 1$ if $a = b$ and zero otherwise.[17] Given that $0 < \kappa_r, \lambda_z \leq 1$, this elasticity is positive for $k = j$ and negative for all other provinces $k \neq j$.

Any change in the conditions in some province $k \neq j$ induces *non-uniform* effects on the ij-specific migration rate, depending on whether this province belongs to the same country or region as province j. In particular, the elasticity in (17) is largest (in absolute value) for any change in the conditions in other provinces in the same region, $I(\ell,r) = I(y,z) = 1$. The fact that these substitution effects are strongest within regions and weakest across countries is due to the similarity of provinces in the same region (and in the same country).

In the standard MNL model with $\lambda_z = \kappa_r = 1 \; \forall r, z$, the pattern of cross-elasticities is much simpler: for $k \neq j$, (17) collapses to $\partial \ln(m_{ij}/m_i)/\partial \ln(Y_k) = -Y_k m_{ik}/m_i$ (independently of whether or not the provinces j and k share the same region or country). The corresponding ij-specific migration rate in (16) equals:

$$\left. \frac{m_{ij}}{m_i} \right|_{\lambda_z, \kappa_r = 1} = \frac{\exp[\xi_{ij} - c_{ir} - c_{iz}]}{\exp[\Psi_i]} = \frac{\exp[U_{ij}]}{\sum_k \exp[U_{ik}]}, \tag{18}$$

where the inclusive values Φ_{ir} and Ω_{iz} (but not Ψ_i) disappear. One approach to get rid of Ψ_i is to compute the relative odds, i.e., the ij-specific migration rate (namely, the fraction of the population in i who migrate to j) relative to the i-specific stay rate (namely, the fraction of non-migrants of the population in i):

$$\frac{m_{ij}}{m_{ii}} = \exp[U_{ij} - U_{ii}], \tag{19}$$

which is independent of the number and characteristics of other provinces $k \neq i, j$, a property known as the independence of irrelevant alternatives (IIA) assumption (McFadden 1974, 1978).[18] Thus, estimating a log-linearized version of (19) (rather than of (18)) has the advantage that no attention needs to be paid to the inclusive values, provided that the IIA assumption is not violated. In our more general NMNL modeling framework, the relative odds become:

$$\frac{m_{ij}}{m_{ii}} = \frac{\exp[\xi_{ij}/(\lambda_z \kappa_r) - \xi_{ii} - c_{ir}/\lambda_z + c_{i\ell} - c_{iz} + c_{iy}]}{\exp[(1-\kappa_r)\Phi_{ir} + (1-\lambda_z)\Omega_{iz}]}, \tag{20}$$

[16] We show in our online Addendum D how to compute this elasticity.

[17] Notice that $I(j,k) = 1$ implies $I(\ell,r) = I(y,z) = 1$, but not the other way around.

[18] Strictly speaking, the standard MNL model as such does not imply the IIA property. The IIA property would indeed be absent in the standard MNL model if U_{ij} was a function of any of the characteristics of province $k \neq i, j$.

where $j \in A_{zr}, r \in A_z$ and $i \in A_{y\ell}, \ell \in A_y$, and where we have used the fact that the country of origin i represents a single final migration destination. It is thus easy to verify that the odds ratio between any two provinces located in different regions is not independent of the number and characteristics of other provinces. This involves a partial relaxation of the IIA assumption, which needs to be addressed explicitly in the estimation, whether we use a log-linearized version of (16) or of (20).[19] Given that the variable m_i in (16) is exogenous, while the variable m_{ii} in (20) is endogenous and potentially difficult to observe, we use the ij-specific migration rate in (16) for our econometric implementation.

2.2 Scale of Migration

Substituting ξ_{ij} in (16), taking logs, and rearranging terms thus yields the following migration function for $j \in A_{zr}$,$r \in A_z$[20]:

$$\begin{aligned} \ln(m_{ij}) &= \frac{\theta}{\lambda_z \kappa_r} \ln(1 + M_{ij}) + \ln(m_i) + \frac{1}{\lambda_z \kappa_r} Y_j - c_{iz} - \frac{1}{\lambda_z} c_{ir} - \frac{1}{\lambda_z \kappa_r} c_{ij}, \\ &\quad -\Psi_i - (1 - \lambda_z)\Omega_{iz} - (1 - \kappa_r)\Phi_{ir}. \end{aligned} \tag{21}$$

Identification of the network effect is thus complicated by the presence of both the different cost components and the inclusive values. Moreover, the network coefficient (defined as $\eta_{zr} \equiv \eta(\lambda_z, \kappa_r) = \frac{\theta}{\lambda \kappa_r}$) is a decreasing function of λ_z and κ_r, which means that a greater similarity among provinces is associated with larger network coefficients. For low values of λ_z and κ_r, it is easy to substitute one province in Spain for another one, especially if they are located in the same region. Hence, a small increase in the migrant network in province $k \in A_{zr}$, will lead a large number of individuals to substitute another province $j \in A_{zr}$ by province k, other things held constant.

2.3 Skill Structure of Migration

We now distinguish between high-skilled and low-skilled individuals, denoted by h and l, respectively. We augment the utility function by a parameter $\gamma^s > 0, s \in \{h, l\}$, representing the ease with which individuals are able to cope with migration costs (decreasing with higher values):

$$U_{ij}^o = Y_j - \gamma^s C_{ij} + e_{ij}^o, \tag{22}$$

where $s = h$ if individual o is high-skilled and $s = l$ otherwise. We assume $\gamma^h < \gamma^l$, so that high-skilled individuals have lower effective migration costs than low-skilled individuals. This assumption

[19] The same applies to the CNL migration model estimated in Bertoli & Fernández-Huertas Moraga (2013).

[20] Notice that we keep the subscript z throughout the paper (although we have just one country of destination in the estimation), as this will make it easier to understand the computation of the network elasticity.

is in line with Chiswick (1999), who argues that the high-skilled can handle their migration process more efficiently than the low-skilled. We can thus derive one migration function for each skill group by complete analogy to equation (21). Subtracting the equation for low-skilled migrants from the same equation for high-skilled migrants, we obtain:

$$\begin{aligned} \ln\left(\frac{m_{ij}^h}{m_{ij}^l}\right) &= \frac{\theta\gamma^*}{\lambda_z\kappa_r}\ln(1+M_{ij}) + \ln\left(\frac{m_i^h}{m_i^l}\right) - \gamma^* c_{iz} - \frac{\gamma^*}{\lambda_z} c_{ir} - \frac{\gamma^*}{\lambda_z\kappa_r} c_{ij} \\ &\quad -\Psi_i^* - (1-\lambda)\Omega_{iz}^* - (1-\kappa_r)\Phi_{ir}^*, \end{aligned} \tag{23}$$

where the variables with an asterisk (*) are differences between the corresponding parameters (or variables) for high-skilled and low-skilled individuals. Since $\gamma^* < 0$, the ratio of new high-skilled to new low-skilled migrants is a decreasing function of the migrant network. This result is due to the fact that individuals differ in their effective costs of migration, and that this difference is less important for low levels of migration costs. Hence, it is the low-skilled individuals who benefit the most from a reduction in migration costs through a larger migrant network.[21]

3 Estimation Strategy and Data

In this section we describe our estimation strategy and the different variables we use in the estimation. The model for the scale of migration is estimated at the level of provinces in Spain, whereas due to reasons of data availability the model for the skill structure of migration is estimated at the level of regions.[22] The baseline sample we use to estimate the scale model comprises the 55 most important countries of origin listed in Table A.1 in Appendix A. These are all countries with at least 630 migrants in Spain in the year 1996. We choose this sample to cover those countries that are responsible for the lion's share of Spanish migration[23], and to make sure we have sufficient cross-sectional variation that we can exploit for identification purposes. The baseline sample we use to estimate the model for the skill structure of migration derives, in principle, from the same set of 55 countries we use for the scale model. However, the actual estimation is carried out on 28 countries due to insufficient data for the dependent variable (the skill structure of migration); see Table A.1 in Appendix A for a list of these countries. All migration data we use in this paper come from the Spanish Instituto Nacional de Estadística (INE). The corresponding internet sources are given in Table A.2 in Appendix A.

[21] This is reflected in the following inequality: $\partial U_{ij}(\gamma^l)/\partial M_{ij} > \partial U_{ij}(\gamma^h)/\partial M_{ij}$. In this respect, our modeling approach is akin to the one in Beine et al. (2011).

[22] Spain consists of 52 provinces and 19 regions. We exclude the provinces (enclaves) of Ceuta and Melilla due to their specific geographical location, and thus end up with 50 provinces nested in 17 regions. See `http://www.ine.es/daco/daco42/codmun/cod_provincia.htm` and `http://www.ine.es/daco/daco42/codmun/cod_ccaa.htm` (both accessed on 04/17/2012) for a list of provinces and regions, respectively.

[23] See chapter 4 in Neubecker (2013) for trends in aggregate migration to Spain over the period considered.

3.1 Statistical Inference: Cross-sectional vs. Time-series Variation

Our NMNL migration model explains both the scale and the skill structure of migration by the migrant network at destination. It is designed to address whether destinations can expect to receive more migrants as well as less-skilled migrants from those countries that have larger migrant networks to start with. Hence, the variation our model attempts to explain is cross-sectional: *across* different countries of origin, as well as *across* different provinces/regions of destination. Statistical inference is therefore based on cross-sectional estimates in the following. This approach is in line with virtually all of the recent literature that tries to identify network effects in aggregate migration data; see Beine et al. (2011, 2012), Grogger & Hanson (2011), Beine & Salomone (2013), Neubecker & Smolka (2013), and, most recently, Bertoli & Fernández-Huertas Moraga (2015).

Some recent papers on the determinants of migration estimate similar models using panel methods to exploit time-series variation in aggregate migration data; see e.g. Ortega & Peri (2013), Bertoli & Fernández-Huertas Moraga (2013), Bertoli et al. (2013), and Beine & Parsons (forthcoming). The advantage of this approach is that it allows to control for additional unobserved heterogeneity in the estimation (e.g. for pairs of countries of origin and destination). However, this practice hinges on the assumption that the random utility parameters $(e^o_{i1}, \ldots, e^o_{iJ})$ individuals draw in one period are replaced by other, independent draws in the next period. In other words, there is zero autocorrelation in unobserved heterogeneity at the individual-level, and individuals have erratic preferences through time. This is a problematic assumption, because "ability, productivity, health status, taste idiosyncrasies, and many other unobservables are likely to be persistent over time" (Norets, 2009, 1665).[24] Ignoring this persistence results in significantly biased estimates of migration costs; see Bayer & Jüssen (2012). We therefore believe that our model should be estimated on cross-sectional data (rather than time-series data), and that we need an alternative estimator in order to tackle the type of unobserved heterogeneity the above-mentioned studies control for through panel estimations. We develop and apply such an alternative estimator below.

3.2 Scale of Migration

The dependent variable is the log of the migration flow to provinces in Spain, obtained from the Spanish Residential Variation Statistics and aggregated from the beginning of 1997 until the end of 2006.[25] We use a period of ten years in order to make our estimates comparable to those obtained in the received literature (Beine et al., 2011, 2012; Grogger & Hanson, 2011; Beine & Salomone,

[24] This is all the more true for high-frequency (i.e. quarterly or monthly) data, such as those used in Bertoli & Fernández-Huertas Moraga (2013) and Bertoli et al. (2013).

[25] Migrants are defined as individuals whose last country of residence (other than Spain) corresponds to their country of birth and nationality.

2013; Bertoli & Fernández-Huertas Moraga, 2015; and Beine & Parsons, forthcoming). The period we choose covers Spain's unprecedented migration boom, which came to a sudden stop in the wake of the global financial and economic crisis in 2007/08.[26] The migrant network, M_{ij}, is given by the number of already settled migrants in 1996, as reported by the Spanish Municipal Register. We rely on population figures disaggregated by nationalities and provinces as of May 1, 1996.

From the year 2000 onwards, our migration data are likely to include both documented and undocumented migrants due to the incentives deriving from the *"Law on the Rights and Freedoms of Aliens in Spain and their Social Integration"* (*Ley Orgánica 4/2000, artículo 12*). This law became effective in 2000 and entitled all registered foreigners to free medical care under the same conditions as Spanish nationals, irrespective of their legal status.[27] Each registrant must provide his or her name, surname, sex, usual domicile, nationality, passport number, as well as the place and date of birth.[28] Since this information is confidential and must not be communicated to other administrative units, the probability of forced repatriation is plausibly independent of registration.

We identify the model from the within-cluster variation across provinces in the data. We begin with a parsimonious fixed effects (FE) specification where clusters are defined in terms of countries of origin. Hence, we compute all variables in equation (21) as deviations from their country means (within-transformation).[29] This approach wipes out, first, all terms with subscript i, and thus controls for the initial population size in the country of origin as well as for the inclusive value Ψ_i; and, secondly, it wipes out all terms with subscript iz, because our migration data refer to a single country of destination z. By eliminating c_{iz}, it thus controls, for example, for the impact of country-specific migration policies and the geographical and cultural distance between the country of origin and Spain, as well as for the inclusive value Ω_{iz}.

In more demanding specifications of our FE model, we define clusters in terms of pairs of countries of origin and regions of destination. Hence, we compute all variables as deviations from their country-and-region means. In addition to the above-described country effects, this approach wipes out all terms with subscript ir. These terms include, first, the multilateral resistance term Φ_{ir}, so that this approach is fully compatible with our three-level NMNL model; and secondly, they include the cost term c_{ir}

[26] See Bertoli & Fernández-Huertas Moraga (2013) and chapter 4 in Neubecker (2013) for detailed descriptives on aggregate migration flows to Spain during that period.

[27] As part of its austerity measures in 2012, the Spanish government has restricted this access to health care for undocumented migrants from September 2012 onwards. Exceptions are made for pregnant women and minors, as well as for cases of emergency care. (`http://www.presseurop.eu/en/content/news-brief/2614611-no-more-free-treatment-undocumented-migrants` based on `http://elpais.com/elpais/2012/08/29/opinion/1346265472_538020.html`, accessed on 08/31/2012).

[28] See INE at `http://www.ine.es/en/metodologia/t20/t203024566_en.htm`, accessed on 08/19/2011.

[29] When zero values inflate the dependent variable, the FE estimator delivers inconsistent estimates. In our baseline sample we observe only a modest number of zero migration flows (5.75% of all country-province pairs) and therefore apply the FE estimator. We will use an estimator that can handle zero migration flows in our robustness analysis.

representing the geographical and cultural distance between the country of origin and the region of destination. Important elements of this distance derive from a cultural, political, and historical context. For example, the different regions in Spain feature considerable heterogeneity in terms of native languages; the Basque Autonomous Community and Navarre both have strong cultural ties with the Northern Basque Country, which is part of French national territory[30]; the region of Galicia has long been suffering from a chronic growth weakness leading to mass emigration in the 19th and 20th century, in particular to Latin American countries.

All other migration costs are summarized in the term c_{ij}. Some of these costs, for example those related to the attitudes of the native population toward migrants, may be specific to the province of destination j but independent of the country of origin i. We control for these province-specific migration costs by including a set of province fixed effects in the estimation. The province fixed effects also absorb the impact of province-specific pull factors summarized in the term Y_j. Some other migration costs may be specific to both the province of destination and the world region of origin (grouping countries of origin). An example would be that individuals from Ecuador feel attracted not only by a network of co-national migrants (i.e., migrants from Ecuador) but also by a network of migrants from other Latin American countries; see Neubecker & Smolka (2013). This additional effect, a "cross-national" network externality, would lower the migration costs for potential migrants from Ecuador, leading to a higher incidence of migration. In more demanding specifications of our model, we therefore control for these other migration costs with a set of world region-and-province fixed effects.[31]

As further control variables, we include bilateral trade and capital flows where possible. Both variables could be part of the cost term c_{ij}. Trade is not only facilitated by, but is also conducive to a good infrastructure for traveling and transportation. Capital invested by foreign firms could create demand for specific types of labor, especially foreign labor. Data on both trade and foreign direct investment (FDI) are provided by the Spanish Ministry of Industry, Tourism and Trade. We measure ij-specific trade flows by the sum of exports and imports (in Euros) in the year 1996. These information are taken from DataComex Statistics on Spanish Foreign Trade. Ideally, we would like to use FDI stocks to measure inward investment but we only have information on gross FDI inflows (in Euros). These are available from DataInvex Statistics on Foreign Investments in Spain and detailed

[30]The Basque Autonomous Community and Navarre form the Spanish part of the Basque Country (*País Vasco* in Spanish; *Euskal Herria* in Basque language).

[31]In terms of world regions, we distinguish between East Asia & Pacific; Eastern Europe & Central Asia; Latin America & Caribbean; Middle East & North Africa; North America, Australia & New Zealand; South & South-East Asia; Sub-Saharan Africa; and Western Europe. For a similar classification used by the IMF, see `http://www.imf.org/external/datamapper/region.htm`, accessed on 07/25/2012.

by the country of the last owner and by the region of destination in Spain.[32] Due to limited data availability, we have to use FDI flows for the year 1997. However, we think that endogeneity is unlikely, because the decision to engage in FDI is often made some time before the actual investments are carried out.

In case we omit ij-specific variables that are correlated with both m_{ij} and M_{ij}, the migrant network is endogenous to the subsequent migrant flow. In view of our extended FE specification, it is difficult to think of any such omitted variable. However, suppose there is a province-specific labor demand for workers from a certain nationality, such as the demand for German engineers in SEAT's car production in Barcelona. Then, the FE model may produce biased and inconsistent estimates. Consistent estimation would call for an instrument that is uncorrelated with the structural error term but correlated with the endogenous regressor. We instrument country i's migrant network in province j with historical internal migration flows in Spain, defined as the log of the number of people holding country i's nationality and migrating from province j to any other province $k \neq j$ in Spain in 1988 (henceforth simply called internal migration).[33]

Because it indicates a large historical network, internal migration can be expected to correlate positively with the migrant network in 1996.[34] Our first-stage regressions attest to a statistically significant positive (partial) correlation. Its significance is also reflected in relatively high values for the first-stage F statistics. For internal migration to be a valid instrument, it must be uncorrelated with the structural error term.[35] This assumption could be violated if a large internal migration observed for a certain province reflects and signals a poor matching quality (for example in terms of jobs) between this province and the corresponding migrants, thus leading to a lower incidence of migration today. However, this signaling effect does not necessarily render our instruments endogenous. One reason is that most, if not all, of the variation in the matching quality across countries and provinces is absorbed into our fixed effects. Another, probably more important, reason is that the signaling effect should be captured by the (observable) migrant network itself, given that this network is a function of all past migration flows. We use internal migration in 1989 as a second excluded instrument. This allows us to perform tests on overidentifying restrictions and check for instrument exogeneity.

[32] Hence, the effect of FDI on migration is not identified in the model controlling for country-and-region fixed effects.

[33] The year 1988 is the first year for which these information are available. It is well before the start of the Spanish migration boom. We add one to the number of people before taking logs in order to keep observations with zero migration flows.

[34] It follows from its definition, however, that internal migration also reduces the size of the historical network.

[35] Therefore, the focus on *internal* migration is on purpose because it excludes return migrants who could shape future migration in one way or the other.

3.3 Skill Structure of Migration

Aggregate migration data with reliable information about the skill structure of migration can only be constructed at the level of regions rather than provinces. We deal with this issue in two different ways. First, we simplify the structure of our model to a two-level NMNL model in which the regions of destination are the final migration destinations within the primary nest of Spain. This approach is straightforward, and is therefore the one we describe in the following. Secondly, we develop an estimation strategy at the regional level that is fully consistent with the three-level NMNL model presented above. This approach is offered in our robustness analysis.

Defining regions of destinations (indexed here by j) as the final migration destination, we re-write Equation (23) as:

$$\ln\left(\frac{m_{ij}^h}{m_{ij}^l}\right) = \frac{\theta\gamma^*}{\lambda_z}\ln(1+M_{ij}) + \ln\left(\frac{m_i^h}{m_i^l}\right) - \gamma^* c_{iz} - \frac{\gamma^*}{\lambda_z}c_{ij} - \Psi_i^* - (1-\lambda_z)\Omega_{iz}^*. \tag{24}$$

The dependent variable measures the skill structure of migration. Skill-specific migration flows are obtained from the National Immigrant Survey 2007 (NIS). The survey gathers unique information on a total of 15,465 migrants through field interviews conducted between November 2006 and February 2007; see Reher & Requena (2009, 255-261) for this and the following information.[36] Migrants report, *inter alia*, their year of arrival in Spain, their first destination in Spain, as well as their highest level of education they completed before migrating. They are defined as individuals aged 16 years or older who were born abroad and have lived in Spain for more than a year, or at least intended to stay for more than a year at the time the survey was conducted.[37] Importantly, this definition is independent of the individual's legal status, so the data again include documented and undocumented migrants. We aggregate the number of migrants by country of birth and region of destination, distinguishing between individuals with completed tertiary education before migrating (high-skilled) and all other individuals (low-skilled) and applying the provided population weights. Although the data can be considered representative of migrants who arrived shortly before the survey was taken, the numbers for earlier cohorts are less reliable due to the lack of information on migrants who died, returned, or migrated onward. We deal with the trade-off between a large number of individuals and data representativeness in that we consider only migrants who arrived in Spain between January 1, 2002, and December 31, 2006.

[36] The sample was obtained through a relatively complex three-stage sampling scheme designed to offer reliable and representative data to policy makers and researchers. More detailed information on the sampling can be found in Reher & Requena (2009) as well as in INE (2007).

[37] Foreign-born individuals with Spanish nationality from birth who migrated to Spain within two years after birth are not considered as migrants.

The migrant network, M_{ij}, is measured by the number of settled migrants as of January 1, 2002. These data, detailed by country of origin and region of destination, are taken from the Spanish Municipal Register. The sum of import and export values in 2001 is collected at the level of regions. Investment stocks as of 2001 are approximated by gross FDI inflows from the beginning of 1998 until the end of 2001. Country-specific fixed effects, absorbing, among other things, the inclusive values Ψ_i^* and Ω_{iz}^*, are wiped out by applying the corresponding within-transformation to the data. Hence, cross-regional differences in the migrant network of a given country of origin are used as identifying variation so that we cannot control for country-and-region fixed effects. We instead augment the model by observable variables that are likely to influence the migration costs. We control for the geographical distance between the country of origin i and the region of destination j, using the STATA module GEODIST by Picard (2010) in combining geographical data on the countries of origin from Mayer & Zignago (2006) and on the regions of destination from the Spanish Wikipedia/GeoHack webpage. We control for a common language through an indicator variable that is equal to one if at least 80% of the region's total population are native speakers of a language spoken by at least 20% of the people living in the country of origin, and zero otherwise. The information on native languages in Spain are taken from a number of recent survey studies.[38] Language information on the countries of origin come from Mayer & Zignago (2006). The influence of all terms indexed j is absorbed by a set of dummy variables for the different regions of destination. The complete specification of our model furthermore controls for world region-and-region fixed effects.

We also apply the instrumental variables approach to this model, by analogy to the model for the scale of migration. In particular, we instrument the migrant network in 2002, M_{ij}, with the log of the number of people holding country i's nationality and migrating from region j in Spain to any other region $k \neq j$ in Spain in 1988. As before, we use the corresponding migration flow in 1989 as a second excluded instrument.

4 Estimation Results

In this section we present and discuss our estimation results. We start with a descriptive look at the relationship between migrant networks and the scale and skill structure of migration to different destinations in Spain. Figure 1(a) is a scatter plot for migration between 1997 and 2006 versus migrant networks in 1996, where each dot represents a different pair of country of origin and province of destination. We observe a positive correlation between the two variables. Figure 1(b) is a scatter plot for the skill structure of migration between 2002 and 2006 versus migrant networks at the beginning

[38] See Table A.2 in Appendix A for a list of surveys.

of 2002, where now each dot represents a different pair of country of origin and region of destination. The figure suggests a weak negative correlation between the two variables. In what follows, we test whether these correlations reflect a causal relationship running from migrant networks to the scale and skill structure of migration, and we provide a structural interpretation of our estimation results in terms of our NMNL migration model. We also discuss the results of several robustness checks.

<<Figures 1(a) and 1(b) about here>>

4.1 Results for the Scale of Migration

In this subsection we present the estimation results of the model for the scale of migration as specified in equation (21). We first estimate an *average* network coefficient that abstracts from potential differences in the parameter κ_r across regions. Tables 1 and 2 show the results from the FE model and the two stage least squares (2SLS) FE model, respectively. In columns (a) and (b) of both tables, we eliminate country fixed effects via an adequate within-transformation of the data. The number of observations is equal to 2,593, which is the result of having 55 countries of origin, 50 provinces of destination, and 157 undefined values for the dependent variable due to zero migrant flows ($55 \times 50 - 157 = 2,593$). In columns (c) to (f), we eliminate country-and-region fixed effects by modifying the within-transformation accordingly. This excludes all country-and-region pairs that have no within variation (for example due to regions that consist of one province), and thus reduces the number of observations to 2,200.

In the most parsimonious specification of the FE model in column (a) of Table 1, the estimated network coefficient is equal to 0.689.[39] The coefficient is statistically significant at the 1% level and estimated with very high precision (heteroskedasticity-robust standard error, clustered by countries of origin, equal to 0.029). When we augment the model by FDI and trade flows in column (b), we find a positive and statistically significant coefficient of the FDI variable. Yet, the point estimate of this coefficient is equal to 0.012 and thus implies a moderate quantitative importance only. Trade relations, instead, do not seem to have a significant impact on the scale of migration. More importantly, the estimates of the network coefficient are virtually unchanged in this version of the model. However, once we control for country-and-region fixed effects in columns (c) and (d), we see a drop in the estimated network coefficient down to 0.54, which corresponds to a decrease by roughly 20%. We see a further reduction by more than 10% once we take out the variation that is constant for each pair of world regions of origin and provinces of destination via dummy variables.

[39] This estimate of the average network coefficient is virtually identical to the local network externality estimated by Beine et al. (2012).

Unobserved heterogeneity in our model has two sources: first, the inclusive values, and secondly, the different cost components. Failing to account for the inclusive values leads to downward-biased estimates of the network coefficient due to a positive covariance between the migrant network and the terms Ψ_i, Ω_{iz}, and Φ_{ir}, respectively. Failing to account for the different cost components, in turn, leads to upward-biased estimates of the network coefficient due to a negative covariance between the migrant network and the terms c_{iz}, c_{ir}, and c_{ij}, respectively. Given that our estimation results point towards a sizeable upward bias in the estimation of the network coefficient in specifications (a)-(d), the second source of unobserved heterogeneity clearly "dominates" the first one.

<<Tables 1 and 2 about here>>

The 2SLS FE estimations in Table 2 strengthen our interpretation of a quantitatively important causal effect of migrant networks on the scale of migration.[40] They suggest a somewhat larger role for the network effect, with a coefficient ranging between 0.718 and 0.955. The difference between the FE estimates and the 2SLS FE estimates could be due to stochastic measurement errors in the migrant network, which would result in downward-biased estimates of the network coefficient when applying the FE estimator; see Hausman (2001). As in the FE estimations, the network coefficient is smallest when controlling for country-and-region effects as well as for world region-and-province effects. The loss in precision from using the 2SLS FE approach is fairly small if interpreted relative to the FE model. The effects of both trade and FDI on the scale of migration are essentially zero. Our next specification allows for cross-regional differences in the similarity parameter κ_r, which implies region-specific network coefficients, η_{zr}. The specification employed is equivalent to the one reported in column (f) of Table 1, except for the fact that we now interact the migrant network with dummy variables for the different regions of destination. Table 3 reveals substantial heterogeneity in the estimated network coefficient across regions. It is largest for the region of Cataluña (0.795) and smallest for the region of Extremadura (0.155).[41] Hence, individuals seem to consider the provinces in the region of Cataluña (Barcelona, Girona, Lleida, and Tarragona) to be very similar to each other, relative to the provinces in the region of Extremadura (Badajoz and Cáceres). This result accords with the pronounced autonomy of Cataluña in terms of its political and cultural life. It is not surprising either that two other regions with a second official language, Comunitat Valenciana and Galicia, rank next to Cataluña in terms of the size of the estimated network coefficient. For the Basque country,

[40] The first-stage F statistic for the joint significance of the excluded instruments is relatively high and thus points to the relevance and strength of the instruments. In all the specifications employed, it exceeds the critical value of 10, which is required for reliable inference in the case of a single endogenous regressor (Stock et al., 2002, 522).

[41] In the estimation, the region of Cataluña serves as the reference region. The differences between the network coefficients estimated for Cataluña and for either of the other regions (except for the regions of Comunitat Valenciana and Canarias) are statistically significant at least at the 10% level according to t-tests.

however, we find a surprisingly low network coefficient (equal to 0.287), which indicates that migrants view the provinces in this region as rather heterogeneous. At any rate, the large and significant cross-regional differences in the estimated network coefficient show that the assumption of a uniform degree of cross-destination substitutability featured in the standard MNL model is too restrictive to be plausible in the Spanish case.[42]

<<Table 3 about here>>

The estimated network coefficients can be used to compute both the network elasticity of migration as well as the cross-elasticities of the network defined as:

$$\begin{aligned}\frac{\partial \ln(m_{ij})}{\partial \ln(1+M_{ik})} &= \theta\left[\frac{I(j,k)}{\lambda_z\kappa_r} - \left(\frac{m_{ik}}{m_i}\right)\right. \\ &\left. -\frac{I(\ell,r)}{\lambda_z\kappa_r}(1-\kappa_r)\left(\frac{m_{ik}}{m_{ir}}\right) - \frac{I(y,z)}{\lambda_z}(1-\lambda_z)\left(\frac{m_{ik}}{m_{iz}}\right)\right]. \qquad (25)\end{aligned}$$

The network elasticity ($j = k$) is a function of the network parameter θ, the similarity parameters κ_r and λ_z, and the relative attractiveness of the province of destination j (reflected by the shares m_{ij}/m_i, m_{ij}/m_{ir}, and m_{ij}/m_{iz}). Neither κ_r nor λ_z can be estimated directly due to the use of aggregate migration data. This implies an uncertainty about the true network *elasticity*, which would prevail even if the true network *coefficient*, η_{zr}, was known with certainty.[43] However, we can compute estimates of the upper and lower bounds for this elasticity, separately for each region of destination. For this purpose, we use the results reported in Table 3 in order to compute estimates of the ratio $\kappa_r/\kappa_\ell = \eta_{z\ell}/\eta_{zr}, \forall r, \ell \in A_z$. Since the region of Extremadura features the lowest estimated network coefficient, its similarity parameter κ_r can take on any value between zero and one, while the similarity parameters for all other regions $\kappa_\ell, \ell \neq r$, must be strictly lower than one. For example, the range of permissable similarity parameter values for the region of Cataluña runs from zero to 0.195 (= 0.155/0.795).

Figure 2(a) shows counterfactual network elasticities by region of destination as a function of the similarity parameter of the region of Extremadura, κ_r. The exact value of κ_r is unknown, but fixing this parameter also fixes the similarity parameters of all other regions. In order to focus on the heterogeneity in the network elasticity that is due to differences in the similarity parameters across

[42] An alternative interpretation of the heterogeneity in the estimated network coefficient is that local labor markets differ across regions, so that networks are more effective in some regions than in others. However, we believe that the differences are too substantial to be attributed to differences in local labor markets alone.

[43] Schmidheiny & Brülhart (2011) discuss a related type of uncertainty in a two-level NMNL model. They show that the Poisson model and the standard MNL model are the polar cases of a two-level NMNL model with two nests, one being a degenerate nest with a single alternative, and the other one featuring many alternatives with a single similarity parameter $\lambda \in (0,1)$. When λ is unknown, the elasticities of the Poisson model and of the standard MNL model can thus serve as boundary values for the true elasticities.

regions, we have imposed the following assumptions: first, there are 200 countries of destination outside the country of origin i; second, each of these countries consists of 51 provinces that are uniformly distributed across 17 regions; and third, all provinces abroad are equally attractive destinations, with an overall fraction of migrants in the total population equal to three percent, $\sum_{j \neq i} m_{ij}/m_i = 0.03$. These assumptions imply: $m_{ij}/m_i = 1/340,000$, $m_{ij}/m_{ir} = 1/3$, and $m_{ij}/m_{iz} = 1/51$. For the provinces in the region of Extremadura, we find a network elasticity that slightly exceeds a value of 0.1. For the provinces in the region of Cataluña, the same elasticity lies in the vicinity of 0.55. These are quite large differences. For any given region, the difference between the upper and the lower bound (i.e., the permissable range) of the network elasticity is roughly equal to 0.05, so the uncertainty about the network elasticity is a minor issue here. Importantly, the figure also incorporates the uncertainty about the country-specific similarity parameter λ_z, which can take on any value between zero and one. This uncertainty, which turns out to be almost irrelevant for the computation of the network elasticity, is reflected in the thickness of the upward-sloping lines.[44]

<<Figures 2(a) and 2(b) about here>>

We have also computed the cross-elasticities of the network based on (25), by analogy to the network elasticity. Cross-elasticities for two provinces belonging to one of the regions listed in Table 3 are depicted in Figure 2(b). For the provinces in the region of Extremadura, we find an extremely low cross-elasticity that ranges between 0.0 and -0.05. For the provinces in the region of Cataluña, the same cross-elasticity lies between -0.22 and -0.27. In Figures B.1(a) and B.1(b) in Appendix B, we also depict the cross-elasticities when the two provinces j and k are located in different regions of the same country and when they are located in different countries, respectively. These cross-elasticities are not specific to any region of destination in Spain, they are lower (in absolute value) than the cross-elasticities depicted in Figure 2(b), and they are characterized by a higher uncertainty about their true values.

Robustness Analysis

We conduct a series of robustness checks and document the corresponding regression results in Appendix C. A first issue has to do with the fact that the migration rate m_{ij}/m_i from which we derive the estimating equation (21) is an *expected* migration rate, with the *actual* migration rate depending on the realizations of the random utility parameters for all individuals $(e^o_{i1}, \ldots, e^o_{iJ})$. In the presence of heteroskedasticity in the stochastic deviations of the expected migration rate from the true migration

[44] Individual lines are upward-sloping because, for a given similarity parameter λ_z and a given estimate of the network coefficient η_{zr}, a larger similarity parameter κ_r for the region of Extremadura is only compatible with a larger network parameter θ.

rate, the FE estimator on the log-linearized migration function in (21) yields inconsistent estimates. This problem has been discussed in the context of the gravity equation in international trade by Santos Silva & Tenreyro (2006). The solution that these authors propose in order to handle this problem is to estimate the equation in levels, using the Poisson pseudo-maximum-likelihood (PPML) estimator. The estimator is consistent even in the presence of heteroskedastic errors in the level equation. A further advantage of using this estimator in the gravity context is that observations with zero migration flows can be included in the estimation (precisely because the estimation is in levels rather than in logs). This issue has also been discussed in Santos Silva & Tenreyro (2006).[45] Table C.1, which presents results for the scale model of migration based on the PPML estimator, shows that our baseline estimations lead to a certain underestimation of the average network coefficient. This corroborates our results from the 2SLS FE estimator and indicates that the FE estimates should be interpreted as a lower bound for the true size of the network effect.

We obtain further evidence in this direction by considering different time periods in our analysis. As argued above, we have chosen to aggregate the migration flow over a ten-year period in our baseline estimations, in order to make our estimates comparable to those provided in the literature. The underlying assumption is that changes in the relative attractiveness of migration destinations over the period considered are not material for the estimation. This assumption is shared by Beine et al. (2011, 2012), Grogger & Hanson (2011), Beine & Salomone (2013), Neubecker & Smolka (2013), and Bertoli & Fernández-Huertas Moraga (2015). What happens if we shorten or extend the period of aggregation? Intuitively, the initial allocation of already settled migrants should be most relevant for the early movers, and gradually fade out as we move on in time. Table C.2 shows that this is what we find in the data. We start out from a model that includes the migration flow in the year 1997 as the dependent variable, and incrementally extend the period of aggregation by one year as we move from one column to the next. The largest point estimates for the network coefficient are in fact found for the periods that aggregate migration flows over one or two years (standing at close to 0.9).[46] Overall, the results we obtain indicate an almost monotonic decline in the estimated network coefficient up to the year 2007 as we move away in time from the initial network allocation in 1997. Interestingly, from 2007 onwards up to the year 2013 (the last year for which data are available) the estimates vary indistinguishable (in a statistical sense) between 0.621 and 0.633. One possible explanation for this is that the global financial and economic crisis in 2007/08 led to a sharp reduction in new migration so

[45] For the same reasons the PPML estimator is also used in Beine et al. (2011, 2012) and Bertoli & Fernández-Huertas Moraga (2015).

[46] In the interest of space we only report the estimated network coefficient for the extended FE specification (estimated with the PPML estimator) corresponding to column (f) in Table C.1.

that the aggregate migration flows did not change much afterwards.[47]

In a further robustness check, we apply alternative sample selection criteria in order to see whether our results suffer from endogenous sample selection. In particular, we consider all observations (country-province pairs) with a migrant network of more than either 10, 20, or 50 migrants in the year 1996.[48] Applying these criteria results in unbalanced samples of 98, 90, or 74 countries, respectively. Again, the results we obtain (not reported) indicate a slightly larger average network coefficient than do our baseline estimates.

A further concern might be a potential estimation bias due to non-stochastic measurement errors in our migration data. The migration data we consider in our baseline estimations covers the period 1997-2006. To the extent that undocumented migrants arrived in or before 1996 and registered in later years (especially due to the *Ley Orgánica 4/2000* in 2000), we understate the true size of the migrant network in 1996 and overstate the true size of the migrant flow over the period 1997-2006. We show in Appendix D that our extended FE specification is entirely immune to both types of measurement errors under a relatively mild assumption, namely that the ratio of "mismeasured" to observed migrants is constant within clusters.

4.2 Results for the Skill Structure of Migration

Table 4 reports the results from FE estimations of our model for the skill structure of migration as specified in equation (24). The full data matrix would contain 935 pairs of 55 countries of origin and 17 regions of destination. However, the migrant skill ratio (the dependent variable) is missing for a substantial share of observations due to the small sample size of the NIS. Moreover, due to the inclusion of fixed effects for the countries of origin, parameter identification requires that each country of origin has at least two regions with non-missing values for the dependent variable. Therefore, the total number of observations in the baseline estimations is 234. Figure A.1(a) in Appendix A shows that most of the variation in the data comes from countries in South America: Argentina, Colombia, Bolivia, Ecuador, Cuba, Brazil, Venezuela, and Peru all have at least 11 regions of destination for which data are available. Other important countries are, for example, Romania (13 regions), Poland (9), and Morocco (9). Not surprisingly, these are also the countries that rank high in the overall incidence of migration to Spain over the period considered. Figure A.1(b) shows that the two most important regions of destination with sufficient data are Cataluña (27 countries of origin) and Madrid (21). Overall, we believe, therefore, that the cross-sectional coverage of our data (in terms of both

[47] Simple correlations between the migration flows for different periods of aggregation support this explanation. The correlation between the flow from 1997-2006 and the flow from 1997 is 0.524. In contrast, the correlation between the flow from 1997-2006 and the flow from 1997-2013 is 0.977.

[48] Sample selection based on explanatory variables is a type of exogenous sample selection; see Wooldridge (2009, 323).

countries and regions) is rich enough to be meaningfully exploited in the type of regression analysis we consider here.

In all the specifications employed in Table 4, we find a robustly significant negative impact of migrant networks on the skill structure of migration, as suggested by theory. The estimated coefficient varies between -0.506 and -0.637, so the differences across specifications are rather small. Neither the trade variable nor the FDI variable turns out to be statistically significant. This finding accords with the poorly suggestive evidence in favor of a positive effect of trade or FDI on the scale of migration. Maybe surprisingly, the effects of a common language and geographical proximity are often estimated to be zero and have an unexpected sign, but one should keep in mind here that identification comes only from within-cluster variation.

<<Tables 4 and 5 about here>>

Table 5 reports the results from the 2SLS FE estimations. They suggest a causal interpretation of the effects of migrant networks.[49] In all the specifications considered, the estimated coefficient of the migrant network is negative and statistically significant at the 5% level. The point estimates range between -0.534 and -1.105 and are thus found to be smaller (in absolute value) than those obtained from the FE estimations. In the full specification of the model in column (f), the migrant network is the only structural explanatory variable that has a statistically significant effect.

In order to interpret our results in terms of elasticities, we compute:

$$\frac{\partial \ln(m_{ij}^h/m_{ij}^l)}{\partial \ln(1+M_{ij})} = \theta\gamma^* \left[\frac{1}{\lambda_z} - \left(\frac{m_{ij}}{m_i}\right) - \frac{1-\lambda_z}{\lambda_z}\left(\frac{m_{ij}}{m_{iz}}\right)\right], \tag{26}$$

where we have assumed, for simplicity, that $m_{ij}/m_i = m_{ij}^h/m_i^h = m_{ij}^l/m_i^l$ and $m_{ij}/m_{iz} = m_{ij}^h/m_{iz}^h = m_{ij}^l/m_{iz}^l$. We assume, as before, that there are 200 countries of destination outside the country of origin i; that each of these countries consists of 17 regions; and that all regions abroad are equally attractive destinations, with an overall fraction of migrants in the total population equal to three percent.[50] Then, because the similarity parameter λ_z can take on any value between zero and one, an estimated coefficient of the migrant network equal to -0.621 (as in column (f) of Table 4) implies that the corresponding elasticity lies between -0.621 and -0.584.

[49] The first-stage F test suggests that our instruments are relevant in specifications (a), (b), and (c), but that they might be weak in specifications (d), (e), and (f).

[50] This implies that $m_{ij}/m_i = 3/340,000$ and $m_{ij}/m_{iz} = 1/17$.

Robustness Analysis

We have checked the robustness of these results and the validity of some underlying assumptions in various ways. A first concern is measurement error in the dependent variable due to the small sample size of the underlying survey data source. Figure A.2 in Appendix A shows that many observations in our baseline sample come from country-region pairs for which the survey records few respondents. Of course, the more respondents we have for a given country-region pair, the more reliable is the skill ratio that we compute from the survey and that we use in the estimation.[51] We must therefore expect the skill ratio to be mismeasured for a sizeable fraction of observations. However, while stochastic measurement error in the dependent variable leads to less precisely estimated coefficients, it does not lead to inconsistent estimates; see Hausman (2001). Hence, we believe that the skewness of the distribution in Figure A.2 toward small numbers does not invalidate our estimates.

This interpretation finds support in additional regressions that we carry out on restricted estimation samples. For example, when we restrict the sample to observations for which the skill ratio is constructed on the basis of at least 15 migrants in the survey, the estimation sample reduces to 75 observations, but the estimates still reflect a highly significant negative effect of the migrant network on the skill structure of migration. In this particular example, we obtain an estimated coefficient of -0.519 (with a 95% confidence interval of $[-.232; -.772]$) for a specification that resembles column (f) in Table 4. This estimate is not distinguishable (in a statistical sense) from the estimate we find based on the unrestricted sample. We obtain similar results when we employ a threshold of 10 migrants in the survey (rather than 15 migrants), which results in an estimation sample of 105 observations.

A related issue is a potential sample selection bias that could be due to the large number of missing values for the migrant skill ratio. In order to investigate this issue, we develop a Heckman (1976)-style procedure similar to the one proposed by Wooldridge (1995, 123-124).[52] We describe this procedure in detail in Appendix F. We find no evidence for sample selection bias in our analysis.

Next, we estimate the model with the PPML estimator rather than the FE estimator. The estimation results we obtain based on this estimator are reported in Table E.1 in Appendix E. We find strong evidence for a negative and significant skill effect of migrant networks also with this alternative estimator.

Moreover, following the methodology proposed by Grogger & Hanson (2011, 53-54), we have excluded the possibility that individuals group regions of destination into nests at the sub-country

[51] This is an implication of the law of large numbers.

[52] Technically, the two-step Heckman procedure for testing and correcting for sample selection bias could be applied if the country fixed effects were not differenced out but, rather, if they were estimated by including a set of country dummy variables. However, this approach would result in inconsistent estimates due to the incidental parameters problem described in Neyman & Scott (1948).

level. To do so, we have repeatedly estimated the scale model as given by equation (21), using regional data instead of provincial data and each time excluding the observations for one region. The estimated network coefficient is very stable across regressions, ranging from 0.665 to 0.719.

Finally, we have estimated a migration function that describes migration into regions of destination but derives from the three-level NMNL model featuring provinces as the final migration destinations. The starting point is to use equations (10) and (11) in order to compute the probability $P_i^o(j^o \in A_{zr}) = P_i^o(j^o \in A_{zr}|r \in A_z)P_i^o(r \in A_z)$, separately for each skill group. It is easy to show that this alternative migration function depends, among other things, on the number of provinces in each regional nest and on the within-nest distribution of migrant networks across provinces. This last argument is part of a highly non-linear term, which collapses to zero if we look at regions that consist of a single province. Hence, we have estimated the model excluding all regions that consist of more than one province. In spite of the reduced number of observations, our estimates continue to reflect a negative and statistically significant impact of migrant networks on the skill structure of migration.[53]

5 Conclusion

In this paper, we have documented strong positive network effects on the scale of migration and a strong negative effect on the ratio of high-skilled to low-skilled migrants. Both types of effects are robust across alternative estimators, estimation samples, and sets of control variables. Our identification strategy is based on a three-level NMNL model that allows for varying degrees of substitutability across alternative migration destinations. The ease with which one destination in Spain can be substituted by another one depends on whether the two destinations are located in the same region or not; in case they are, it also depends on the degree of homogeneity of that region. Our approach is corroborated by the significant degree of heterogeneity in the estimated network elasticity across regions.

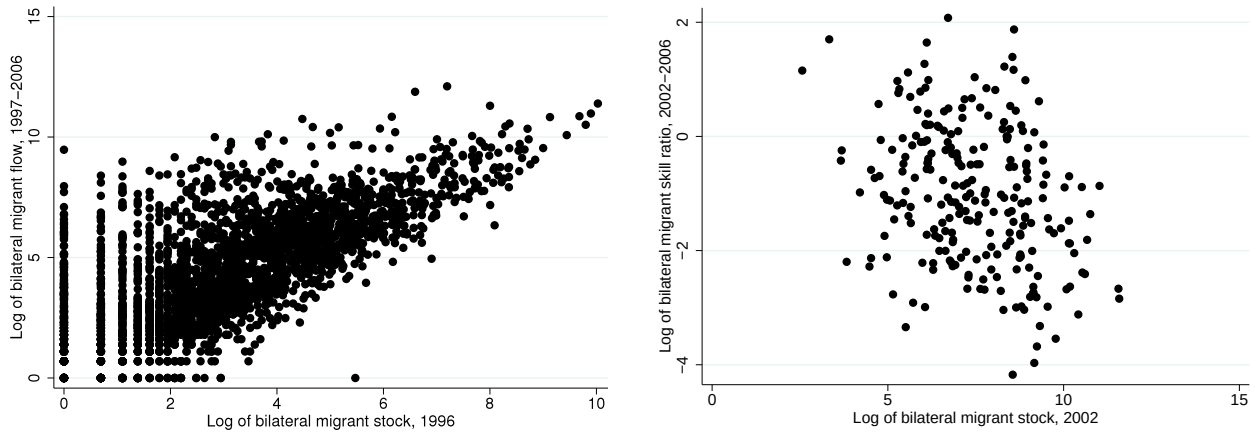
The results we document in this paper can inform the current debate about immigration in Europe. Frattini & Dustmann (2014), for example, show that recent migrants from countries that joined the EU in 2004 made a particularly strong positive fiscal contribution to the UK. This finding seems in line with the idea, supported by our results, that destinations with small migrant networks attract relatively more high-skilled migrants (that can be expected to perform well in the labor market, and to claim less benefits). However, our estimates imply that, other things equal, the average skill of the people that are yet to migrate from these countries to the UK will decline, whereas the overall number of migrants from these countries might remain high.

[53] We have also experimented with two alternative estimation approaches following Quigley (1976) and Lerman (1976). Both include the full set of regions in Spain and are summarized in McFadden (1978, 91-94). Again, we have obtained a robustly significant, negative impact of migrant networks on the skill structure of migration.

Our findings also add to the understanding of the recent migration phenomenon in Spain. This migration has gained momentum through Spain's strong economic growth in the years before the Global Financial Crisis. It has changed the size and composition of the country's population and labor supply, with potentially important effects on a number of key macroeconomic variables such as wages, unemployment, and production, as well as on the national welfare state. The recent economic recession in Spain is reflected in a sharp decline in new migration and a significant amount of return migration in the very short run. The analysis of the structural relationships among past migration, future migration, and the labor market outcomes involves highly complex dynamics. Attempts to study these dynamics are a challenging yet promising avenue for future research.

Figures and Tables

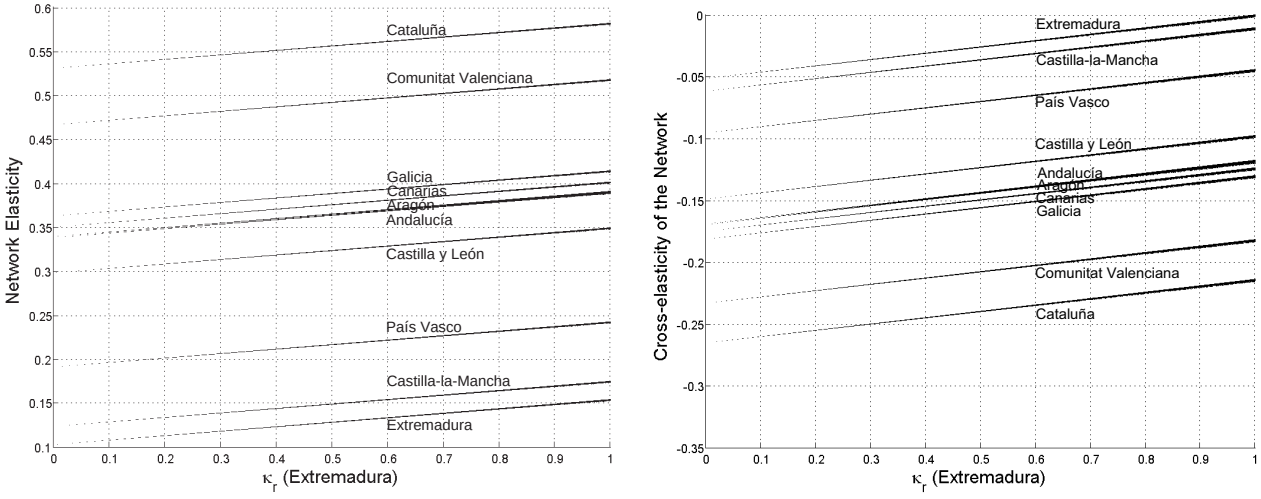
Figure 1: Migrant Networks and the Scale and Skill Structure of Migration



(a) $\ln(m_{ij})$ plotted against $\ln(1+M_{ij})$, provincial level

(b) $\ln(m_{ij}^h/m_{ij}^l)$ plotted against $\ln(1+M_{ij})$, regional level

Figure 2: Counterfactual Network Elasticities and Cross-elasticities



(a) Network Elasticities

(b) Cross-elasticities for $j, k \in A_{zr}$

Table 1: Scale of Migration – FE Model†

	Dependent Variable: Migration Flow (Province-Level 1997-2006)					
	(a)	(b)	(c)	(d)	(e)	(f)
Stock of Migrants	0.689***	0.683***	0.540***	0.541***	0.470***	0.470***
(Province-Level 1996)	(0.029)	(0.029)	(0.029)	(0.029)	(0.033)	(0.033)
FDI Flow		0.012**				
(Region-Level 1997)		(0.005)				
Trade Flow		0.005		0.003		0.008
(Province-Level 1996)		(0.007)		(0.007)		(0.007)
Province Effects	Yes	Yes	Yes	Yes	Nested	Nested
Country Effects	Yes	Yes	Nested	Nested	Nested	Nested
Country-and-Region Effects	No	No	Yes	Yes	Yes	Yes
World Region-and-Province Effects	No	No	No	No	Yes	Yes
Observations	2,593	2,593	2,200	2,200	2,200	2,200
Within R2	0.791	0.792	0.669	0.669	0.763	0.764

†All variables are in natural logs. Heteroskedasticity-robust standard errors (clustered by countries of origin or pairs of countries of origin and regions of destination) are given in parentheses. *,**,*** denote significance at the 10%, 5%, 1% levels, respectively. The regressions include all countries of origin with at least 630 nationals residing in Spain in 1996 (55 countries of origin). See Section 3 for a detailed description of all variables.

Table 2: Scale of Migration – 2SLS FE Model†

	Dependent Variable: Migration Flow (Province-Level 1997-2006)					
	(a)	(b)	(c)	(d)	(e)	(f)
Stock of Migrants	0.955***	0.952***	0.823***	0.825***	0.718***	0.721***
(Province-Level 1996)	(0.069)	(0.070)	(0.080)	(0.080)	(0.101)	(0.101)
FDI Flow		0.004				
(Region-Level 1997)		(0.005)				
Trade Flow		0.004		0.005		0.008
(Province-Level 1996)		(0.007)		(0.008)		(0.007)
Province Effects	Yes	Yes	Yes	Yes	Nested	Nested
Country Effects	Yes	Yes	Nested	Nested	Nested	Nested
Country-and-Region Effects	No	No	Yes	Yes	Yes	Yes
World Region-and-Province Effects	No	No	No	No	Yes	Yes
Observations	2,593	2,593	2,200	2,200	2,200	2,200
Within R2	0.770	0.770	0.635	0.635	0.745	0.745
Robust first stage F	31.44	30.73	18.57	18.51	12.82	12.79
Hansen J test	0.0128	0.0194	0.420	0.384		
Hansen J test p-value	0.910	0.889	0.517	0.535		
Endogeneity test	14.52	14.21	10.86	10.93		
Endogeneity test p-value	0.000	0.000	0.001	0.001		
Kleibergen-Paap LM test	20.23	20.15	24.32	24.30	20.35	20.33
Kleibergen-Paap LM p-value	0.000	0.000	0.000	0.000	0.000	0.000

†All variables are in natural logs. Heteroskedasticity-robust standard errors (clustered by countries of origin or pairs of countries of origin and regions of destination) are given in parentheses. *,**,*** denote significance at the 10%, 5%, 1% levels, respectively. The regressions include all countries of origin with at least 630 nationals residing in Spain in 1996 (55 countries of origin). The (log) stock of migrants in 1996 is instrumented with the (log) migration flows of foreign nationals within Spain in 1988 and in 1989. See Section 3 for a detailed description of all variables.

Table 3: Estimated Network Coefficients, by Region of Destination†

Region r	Estimate of η_{zr}	Region r	Estimate of η_{zr}
Cataluña	0.795	Andalucía	0.507
Comunitat Valenciana	0.699	Castilla y León	0.447
Galicia	0.544	País Vasco	0.287
Canarias	0.525	Castilla-La Mancha	0.186
Aragón	0.509	Extremadura	0.155

†This table reports region-specific estimates of the network coefficient, η_r. The specification employed is equivalent to the one reported in column (f) of Table 1, except that we interact the migrant network with dummy variables for the different regions of destination. F tests reveal that each of the above-reported network coefficients – with the exception of the one for Extremadura – is significant at least at the 5% level. The number of observations is 2,200 and the within R^2 is 0.771.

Table 4: Skill Structure of Migration – FE Model†

	Dependent Variable: Migrant Skill Ratio (Region-Level 2002-2006)					
	(a)	(b)	(c)	(d)	(e)	(f)
Stock of Migrants	-0.513***	-0.510***	-0.506***	-0.626***	-0.637***	-0.621***
(Region-Level 2002)	(0.085)	(0.084)	(0.087)	(0.094)	(0.090)	(0.098)
FDI Flow			-0.006			-0.012
(Region-Level 1998-2001)			(0.019)			(0.015)
Trade Flow			-0.001			0.080
(Region-Level 2001)			(0.079)			(0.095)
Language		0.248	0.246		0.463***	0.559***
(Region-Level)		(0.209)	(0.210)		(0.149)	(0.131)
Distance		-0.636*	-0.657*		-1.450	-1.388
(Region-Level)		(0.373)	(0.369)		(1.159)	(1.148)
Region Effects	Yes	Yes	Yes	Nested	Nested	Nested
Country Effects	Yes	Yes	Yes	Yes	Yes	Yes
World Region-and-Region Effects	No	No	No	Yes	Yes	Yes
Observations	234	234	234	234	234	234
Within R2	0.245	0.261	0.261	0.466	0.477	0.481

†All variables except for the language dummy are in natural logs. Heteroskedasticity-robust standard errors (clustered by countries of origin) are given in parentheses. *,**,*** denote significance at the 10%, 5%, 1% levels, respectively. See Section 3 for a detailed description of all variables.

Table 5: Skill Structure of Migration – 2SLS FE Model†

	Dependent Variable: Migrant Skill Ratio (Region-Level 2002-2006)					
	(a)	(b)	(c)	(d)	(e)	(f)
Stock of Migrants	-0.534**	-0.550***	-0.549***	-1.002***	-1.089***	-1.105***
(Region-Level 2002)	(0.210)	(0.210)	(0.210)	(0.380)	(0.404)	(0.422)
FDI Flow			-0.005			0.009
(Region-Level 1998-2001)			(0.021)			(0.029)
Trade Flow			0.004			0.077
(Region-Level 2001)			(0.080)			(0.101)
Language		0.244	0.243		0.325*	0.344
(Region-Level)		(0.205)	(0.206)		(0.178)	(0.215)
Distance		-0.637*	-0.649*		-1.877	-1.795
(Region-Level)		(0.371)	(0.366)		(1.192)	(1.167)
Region Effects	Yes	Yes	Yes	Nested	Nested	Nested
Country Effects	Yes	Yes	Yes	Yes	Yes	Yes
World Region-and-Region Effects	No	No	No	Yes	Yes	Yes
Observations	234	234	234	234	234	234
Within R2	0.245	0.260	0.260	0.419	0.411	0.409
Robust first stage F	13.93	12.40	11.62	5.961	6.119	5.210
Hansen J test	0.863	0.613	0.673			
Hansen J test p-value	0.353	0.434	0.412			
Endogeneity test	0.110	0.146	0.177			
Endogeneity test p-value	0.740	0.702	0.674			
Kleibergen-Paap LM test	11.41	10.85	10.42	8.258	8.520	7.860
Kleibergen-Paap LM p-value	0.003	0.004	0.005	0.016	0.014	0.020

†All variables except for the language dummy are in natural logs. Heteroskedasticity-robust standard errors (clustered by countries of origin) are given in parentheses. *,**,*** denote significance at the 10%, 5%, 1% levels, respectively. The (log) stock of migrants in 2002 is instrumented with the (log) migration flows of foreign nationals within Spain in 1988 and in 1989. See Section 3 for a detailed description of all variables.

References

[1] Anderson, James E., "The Gravity Model," *Annual Review of Economics*, 3 (2011), 133–160.

[2] Anderson, James E., and Eric van Wincoop, "Gravity with Gravitas: A Solution to the Border Puzzle," *American Economic Review*, 93:1 (2003), 170–192.

[3] Åslund, Olof, "Now and forever? Initial and Subsequent Location Choices of Immigrants," *Regional Science and Urban Economics*, 35:2 (2005), 141–165.

[4] Baghdadi, Leila, "Mexico-U.S. Migration: Do Spatial Networks Matter?," Universite Paris I. mimeograph (2005).

[5] Bartel, Ann P., "Where Do the New U.S. Immigrants Live?," *Journal of Labor Economics*, 7:4 (1989), 371–391.

[6] Bauer, Thomas, Gil S. Epstein, and Ira N. Gang, "Enclaves, Language, and the Location Choice of Migrants," *Journal of Population Economics*, 18:4 (2005), 649–662.

[7] Bauer, Thomas, Gil S. Epstein, and Ira N. Gang, "Measuring Ethnic Linkages among Migrants," *International Journal of Manpower*, 30:1/2 (2009), 56-69.

[8] Bayer, Christian, and Falko Jüssen, "On the Dynamics of Interstate Migration: Migration Costs and Self-Selection," *Review of Economic Dynamics*, 15:3 (2012), 377-401.

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

[10] Beine, Michel, Frédéric Docquier, and Çağlar Özden, "Dissecting Network Externalities in International Migration," University of Luxembourg mimeograph (2012).

[11] Beine, Michel, and Christopher Parsons, "Climatic Factors as Determinants of International Migration," *Scandinavian Journal of Economics* (forthcoming).

[12] Beine, Michel, and Sara Salomone, "Network Effects in International Migration: Education versus Gender," *Scandinavian Journal of Economics*, 115:2 (2013), 354–380.

[13] Bertoli, Simone, "Networks, Sorting and Self-selection of Ecuadorian Migrants," *Annals of Economics and Statistics*, 2010:97/98 (2010), 261–288.

[14] Bertoli, Simone, and Jesús Fernández-Huertas Moraga, "Visa Policies, Networks and the Cliff at the Border," IZA Discussion Paper No. 7094 (2012).

[15] Bertoli, Simone, and Jesús Fernández-Huertas Moraga, "Multilateral Resistance to Migration," *Journal of Development Economics*, 102 (2013), 79–100.

[16] Bertoli, Simone, and Jesús Fernández-Huertas Moraga, "The Size of the Cliff at the Border," *Regional Science and Urban Economics*, 51 (2015), 1-6.

[17] Bertoli, Simone, Herbert Brücker, and Jesús Fernández-Huertas Moraga, "The European Crisis and Migration to Germany: Expectations and the Diversion of Migration Flows," IZA Discussion Paper No. 7170 (2013).

[18] Card, David and Ethan G. Lewis, "The Diffusion of Mexican Immigrants During the 1990s: Explanations and Impacts," in George J. Borjas (Ed.), *Mexican Immigration to the United States* (Chicago: University of Chicago Press, 2007), chapter 6, 193–227.

[19] Carrington, William J., Enrica Detragiache, and Tara Vishwanath, "Migration with Endogenous Moving Costs," *American Economic Review*, 86:4 (1996), 909–930.

[20] Chau, Nancy, "The Pattern of Migration with Variable Migration Cost," *Journal of Regional Science*, 37:1 (1997), 35–54.

[21] Chiswick, Barry R., "Are Immigrants Favorably Self-Selected?," *American Economic Review: Papers and Proceedings*, 89:2 (1999), 181–185.

[22] Chiswick, Barry R., and Paul W. Miller, "Where Immigrants Settle in the United States," *Journal of Comparative Policy Analysis*, 6:2 (2004), 185–197.

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

[24] Dolfin, Sarah, and Garance Genicot, "What Do Networks Do? The Role of Networks on Migration and 'Coyote' Use," *Review of Development Economics*, 14:2 (2010), 343–359.

[25] Domencich, Thomas A., and Daniel L. McFadden, *Urban Travel Demand: A Behavioral Analysis* (North Holland: Amsterdam, 1975. Reprinted by The Blackstone Company: Mount Pleasant, MI, 1996.)

[26] Frattini, Tommaso, and Christian Dustmann, "The Fiscal Effects of Immigration to the UK," *Economic Journal*, 125:580 (2014), F593–F643.

[27] Farré, Lídia, Libertad González, and Francesc Ortega, "Immigration, Family Responsibilities and the Labor Supply of Skilled Native Women," *The B.E. Journal of Economic Analysis & Policy*, 11:1 (2011).

[28] González, Libertad, and Francesc Ortega, "How Do Very Open Economies Absorb Large Immigration Flows? Evidence from Spanish Regions," *Labour Economics*, 18 (2011), 57-70.

[29] González, Libertad, and Francesc Ortega, "Immigration and Housing Booms: Evidence from Spain," *Journal of Regional Science*, 53:1 (2013), 37–59.

[30] Grogger, Jeffrey, and Gordon H. Hanson, "Income Maximization and the Selection and Sorting of International Migrants," *Journal of Development Economics*, 95:1 (2011), 42–57.

[31] Hanson, Gordon H., "International Migration and the Developing World," in D. Rodrik and M. Rosenzweig (Eds.), *Handbook of Development Economics, Vol. 5* (Amsterdam: North-Holland, 2010), 4363–4414.

[32] Hausman, Jerry, "Mismeasured Variables in Econometric Analysis: Problems from the Right and Problems from the Left," *Journal of Economic Perspectives*, 15:4 (2001), 57–67.

[33] Heckman, James J., "The Common Structure of Statistical Models of Truncation, Sample Selection, and Limited Dependent Variables and a Simple Estimator for Such Models," *Annals of Economic and Social Measurement*, 5:4 (1976), 475–492.

[34] INE, "National Immigrant Survey 2007. Methodology," (2007).

[35] Jayet, Hubert, Nadiya Ukrayinchuk, and Giuseppe De Arcangelis, "The Location of Immigrants in Italy: Disentangling Networks and Local Effects," *Annals of Economics and Statistics*, 2010:97/98 (2010), 329–350.

[36] Lerman, Steven R., "Location, Housing, Automobile Ownership, and Mode to Work: A Joint Choice Model," *Transportation Research Record*, 610 (1976), 6–11.

[37] Lewer, Joshua J., and Hendrik Van den Berg, "A Gravity Model of Immigration," *Economics Letters*, 99:1 (2008), 164–167.

[38] Massey, Douglas S., "Economic Development and International Migration in Comparative Perspective," *Population and Development Review*. 14:3 (1988), 383-413.

[39] Mayda, Anna M., "International Migration: A Panel Data Analysis of the Determinants of Bilateral Flows," *Journal of Population Economics*, 23:4 (2010), 1249–1274.

[40] Mayer, Thierry, and Soledad Zignago, "Notes on CEPII's Distances Measures," (2006).

[41] McFadden, Daniel L., "Conditional Logit Analysis of Qualitative Choice Behavior," in Paul Zarembka (Ed.), *Frontiers in Econometrics* (New York: Academic Press, 1974), 105–142.

[42] McFadden, Daniel L., "Modelling the Choice of Residential Location," in Anders Karlqvist, Lars Lundqvist, Folke Snickars, and Jörgen W. Weibull (Eds.), *Spatial Interaction Theory and Planning Models* (Amsterdam: North-Holland, 1978), 75–96.

[43] McFadden, Daniel L., "Econometric Model of Probabilistic Choice," in Charles F. Manski and Daniel L. McFadden (Eds.), *Structural Analysis of Discrete Data with Econometric Applications* (Cambridge, MA: MIT Press, 1981), chapter 5, 198-272.

[44] McFadden, Daniel L., "Econometric Analysis of Qualitative Response Models," in Zvi Griliches and Michael D. Intriligator (Eds.), *Handbook of Econometrics, Vol. II* (Amsterdam: Elsevier Science Publishers, 1984), chapter 24, 1396–1457.

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

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

[47] Neubecker, Nina, *Essays on the Migration of Heterogeneous Individuals,* PhD thesis, University of Tübingen (2013).

[48] Neubecker, Nina, and Marcel Smolka, "Co-national and Cross-national Pulls in International Migration to Spain," *International Review of Economics & Finance*, 28 (2013), 51–61.

[49] Norets, Andriy, "Inference in Dynamic Discrete Choice Models with Serially Correlated Unobserved State Variables," *Econometrica*, 77:5 (2009), 1665–1682.

[50] Neyman, Jerzy, and Elizabeth L. Scott, "Consistent Estimates Based on Partially Consistent Observations," *Econometrica*, 16:1 (1948), 1–32.

[51] OECD, *International Migration Outlook: SOPEMI 2010,* OECD (2010), Paris.

[52] Ortega, Francesc, and Giovanni Peri, "The Causes and Effects of International Migrations: Evidence from OECD Countries 1980-2005," NBER Working Paper No. 14833 (2009).

[53] Ortega, Francesc, and Giovanni Peri, "The Effect of Income and Immigration Policies on International Migration," *Migration Studies*, 1:1 (2013), 47-74.

[54] Pedersen, Peder J., Mariola Pytlikova, and Nina Smith, "Selection and Network Effects – Migration Flows into OECD Countries 1990-2000," *European Economic Review*, 52:7 (2008), 1160–1186.

[55] Pesaran, M. Hashem, "Estimation and Inference in Large Heterogeneous Panels with a Multifactor Error Structure," *Econometrica*, 74:4 (2006), 967–1012.

[56] Picard, Robert, "GEODIST: Stata Module to Compute Geodetic Distances," Statistical Software Components, Boston College Department of Economics (2010).

[57] Quigley, John M., "Housing Demand in the Short Run: An Analysis of Polytomous Choice," in NBER (Ed.), *Explorations in Economic Research, Vol. 3, No. 1* (NBER, 1976), chapter 3, 76–102.

[58] Reher, David, and Miguel Requena, "The National Immigrant Survey of Spain: A New Data Source for Migration Studies," *Demographic Research*, 20:12 (2009), 253–278.

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

[60] Schmidheiny, Kurt, and Marius Brülhart, "On the Equivalence of Location Choice Models: Conditional Logit, Nested Logit and Poisson," *Journal of Urban Economics*, 69:2 (2011), 214–222.

[61] Stock, James H., Jonathan H. Wright, and Motohiro Yogo, "A Survey of Weak Instruments and Weak Identification in Generalized Method of Moments," *Journal of Business and Economic Statistics*, 20:4 (2002), 518–529.

[62] Vovsha, Peter, "Application of Cross-nested Logit Model to Mode Choice in Tel Aviv, Israel, Metropolitan Area," *Transportation Research Record*, 1607 (1997), 6–15.

[63] Wen, Chieh-Hua, and Frank S. Koppelman, "The Generalized Nested Logit Model," *Transportation Research Part B*, 35:7 (2001), 627–641.

[64] Wooldridge, Jeffrey M., "Selection Corrections for Panel Data Models Under Conditional Mean Independence Assumptions," *Journal of Econometrics*, 68:1 (1995), 115–132.

[65] Wooldridge, Jeffrey M., *Introductory Econometrics. A Modern Approach,* Fourth Edition, International Student Edition (South-Western Cengage Learning, 2009).

[66] Zavodny, Madeline, "Welfare and the Locational Choices of New Immigrants," *Federal Reserve Bank of Dallas Economic Review*, (1997), 2–10.

[67] Zavodny, Madeline, "Determinants of Recent Immigrants' Locational Choices," *International Migration Review*, 33:4 (1999), 1014–1030.

A Data Appendix

Table A.1: List of Countries Considered in the Empirical Analysis, by World Region†

EAST ASIA & PACIFIC	Cuba*	NORTH AMERICA,	WESTERN EUROPE
China*	Dominican Republic*	AUSTRALIA	Austria
Japan	Ecuador*	& NEW ZEALAND	Belgium*
Korea	El Salvador	Australia	Denmark
Philippines	Honduras	Canada	Finland
	Mexico*	United States*	France*
EASTERN EUROPE	Peru*		Germany*
& CENTRAL ASIA	Uruguay*	SOUTH	Ireland
Bosnia and Herzegowina	Venezuela*	& SOUTHEAST ASIA	Italy*
Bulgaria*		India	Netherlands*
Poland*	MIDDLE EAST	Pakistan	Norway*
Romania*	& NORTH AFRICA		Portugal*
Russia*	Algeria*	SUB-SAHARAN	Sweden
	Egypt	AFRICA	Switzerland
LATIN AMERICA	Iran	Angola	United Kingdom*
& CARIBBEAN	Lebanon	Cape Verde	
Argentina*	Morocco*	Equatorial Guinea	
Bolivia*	Syria	Gambia	
Brazil*		Guinea	
Chile*		Mauritania	
Colombia*		Senegal	

† The baseline estimation sample for the scale model includes all countries of origin with at least 630 migrants residing in Spain in the year 1996. These are the 55 countries listed above. The corresponding sample we use for the model for the skill structure of migration includes all of the above countries that have sufficient data for the dependent variable (i.e. the skill structure of migration). These are the 28 countries marked with an asterisk.

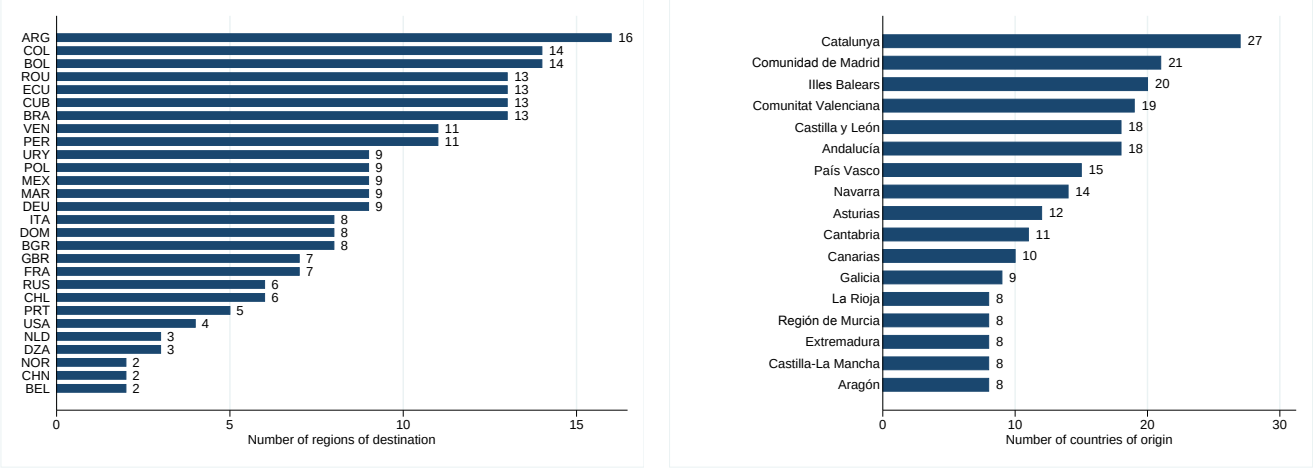
Table A.2: Data Sources

Variable	Definition	Data Sources
Migrant Flow m_{ij}	Migrants who registered at municipalities in Spain between January 1, 1997, and December 31, 2006 (or other years depending on the regression), by province of destination (or region of destination) and by country of origin. Migrants are defined as individuals whose last country of residence (other than Spain) corresponds to their country of birth and nationality.	Spanish Residential Variation Statistics, INE, http://www.ine.es/en/prodyser/micro_varires_en.htm, accessed on 10/05/2010 (as well as one 11/24/2014 for the revision)
Migrant Skill Ratio m^h_{ij}/m^l_{ij}	Ratio of new high-skilled migrants over new low-skilled migrants, aggregated from 2002 to 2006, by region of destination in Spain and by country of birth. Migrants are individuals aged 16 years or older who were born abroad and have lived in Spain for more than a year, or at least intended to stay for more than a year at the time the survey was conducted.	National Immigrant Survey 2007, INE, http://www.ine.es/prodyser/micro_inmigra.htm, accessed on 10/05/2010
Migrant Network M_{ij}	Number of settled migrants as of May 1, 1996, by province of destination (or region of destination) in Spain and by nationality.	Population by Nationality, Autonomous Communities and Provinces, Sex and Year, Municipal Register, Main Population Series since 1998, INE, http://www.ine.es/jaxi/menu.do?type=pcaxis&path=%2Ft20%2Fe245&file=inebase&L=0, accessed on 10/07/2010
Trade Flow	Sum of exports and imports, by province (or region) in Spain and by country of destination/origin.	DataComex Statistics on Spanish Foreign Trade, Spanish Government, Ministry of Industry, Tourism and Trade, http://datacomex.comercio.es/principal_comex_es.aspx, accessed on 10/20/2010
FDI Flow	Gross FDI flow in Euros, by region in Spain and by country of the last owner.	DataInvex Statistics on Foreign Investments in Spain, Spanish Government, Ministry of Industry, Tourism and Trade, http://datainvex.comercio.es/principal_invex.aspx, accessed on 10/20/2010
Historical Internal Migrant Flow	People moving from one province (or region) to another province (or region) in Spain in 1988 and 1989, by province (or region) in Spain and by nationality.	Spanish Residential Variation Statistics, INE, http://www.ine.es/en/prodyser/micro_varires_en.htm, accessed on 10/05/2010
Geographical Distance	Distances are constructed on the basis of latitudinal and longitudinal data for regions in Spain and countries of origin and using the STATA module GEODIST by Picard (2010).	SpanishWikipedia/GeoHack, http://es.wikipedia.org, accessed on 09/05/2011; Mayer & Zignago (2006)

Table A.2 *continued*

Variable	Definition	Data Sources
Indicator for Common Lanuage	This variable is equal to one if at least 80% of a region's population in Spain are native speakers of a language spoken by at least 20% of the people in the country of origin; it is zero otherwise.	*Cataluña*: Generalitat de Catalunya, Institut d'Estadística de Catalunya (2008). Enquesta d'usos lingüístics de la població 2008. *Comunidad Foral de Navarra*: Instituto de Estadística de Navarra (2001). Censo 2001 de Población y Viviendas en Navarra. *Comunitat Valenciana*: Universidad de Salamanca (2007). Estudio CIS No. 2.667. La identitad nacional en España. *Galicia*: Instituto Galego de Estatística (2008). Enquisa de condicións de vida das familias. Coñecemento e uso do galego. Edición 2008. *Illes Balears*: Villaverde i Vidal, J. A. (2003). L'Enquesta Sociolingüística 2003. Principals Resultats. *País Vasco*: Universidad de Salamanca (2007). Estudio CIS No. 2.667. La identitad nacional en España. *Countries of origin*: Mayer & Zignago (2006).

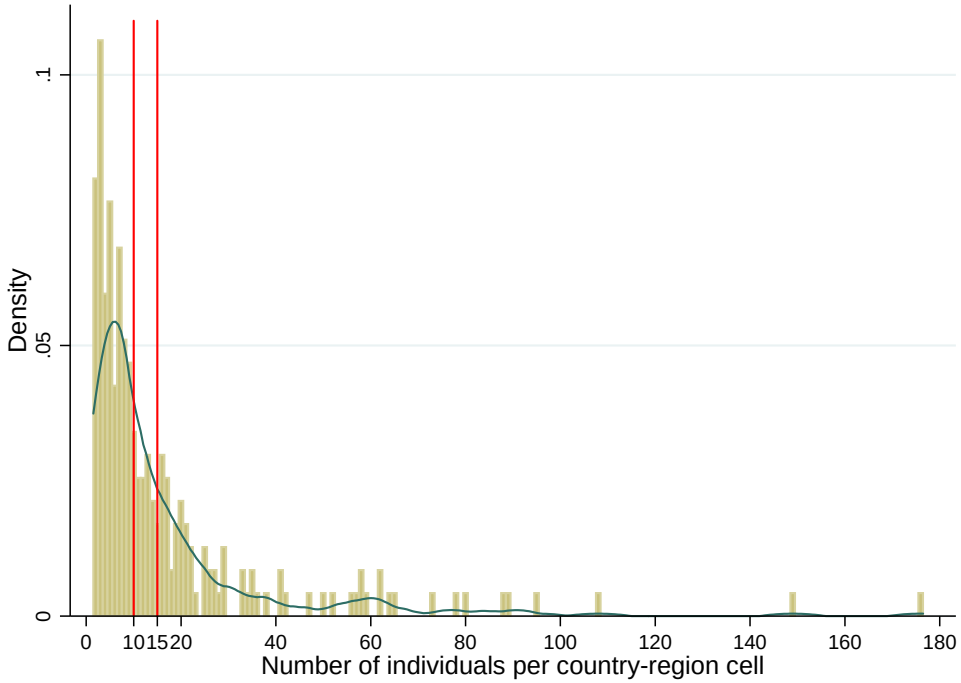
Figure A.1: Data Distribution for the Skill Structure of Migration



(a) Countries of origin (ISO 3166) in the baseline sample

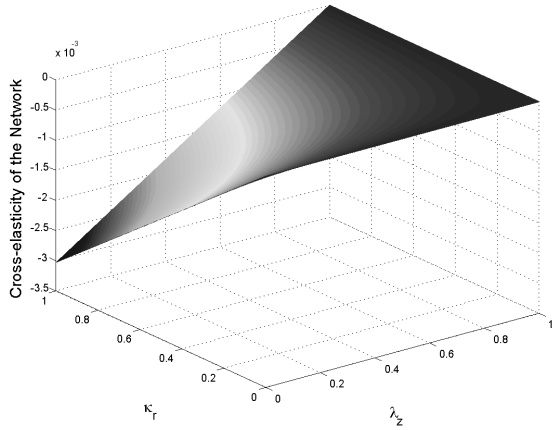
(b) Regions of destination in the baseline sample

Figure A.2: Distribution of Individuals in Survey Data

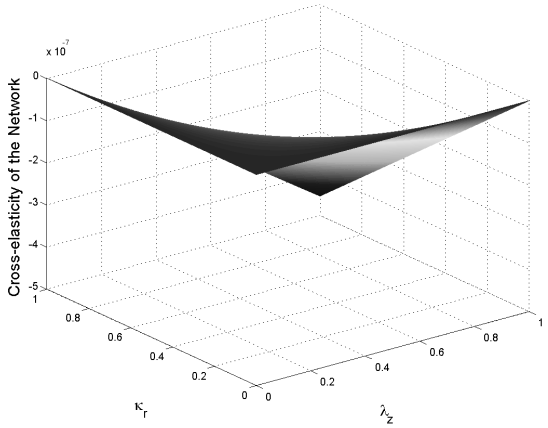


B Counterfactual Cross-elasticities of the Migrant Network

Figure B.1: Counterfactual Cross-elasticities of the Migrant Network



(a) Cross-elasticities for $j \in A_{zr}$ and $k \in A_{z\ell}$, $r \neq \ell$



(b) Cross-elasticities for $j \in A_{zr}$ and $k \in A_{y\ell}$, $z \neq y$

C Further Regressions for the Scale of Migration

Table C.1: Scale of Migration – Poisson Model (PPML)†

	Dependent Variable: Migration Flow (Province-Level 1997-2006)					
	(a)	(b)	(c)	(d)	(e)	(f)
Stock of Migrants	0.814***	0.805***	0.681***	0.681***	0.638***	0.638***
(Province-Level 1996)	(0.036)	(0.041)	(0.045)	(0.045)	(0.063)	(0.061)
FDI Flow		0.008				
(Region-Level 1997)		(0.014)				
Trade Flow		0.010		0.005		0.023
(Province-Level 1996)		(0.012)		(0.014)		(0.015)
Province Effects	Yes	Yes	Yes	Yes	Nested	Nested
Country Effects	Yes	Yes	Nested	Nested	Nested	Nested
Country-and-Region Effects	No	No	Yes	Yes	Yes	Yes
World Region-and-Province Effects	No	No	No	No	Yes	Yes
Observations	2,750	2,750	2,365	2,365	2,365	2,365

†Explanatory variables are in natural logs. Heteroskedasticity-robust standard errors (clustered by countries of origin or pairs of countries of origin and regions of destination) are given in parentheses. *,**,*** denote significance at the 10%, 5%, 1% levels, respectively. The regressions include all countries of origin with at least 630 nationals residing in Spain in 1996 (55 countries of origin). See Section 3 for a detailed description of all variables.

Table C.2: Scale of Migration – Poisson Model (PPML) for Different Periods of Aggregation†

	Dependent Variable: Migration Flow (Province-Level 1997-X)								
	X=1997	*X=1998*	*X=1999*	*X=2000*	*X=2001*	*X=2002*	*X=2003*	*X=2004*	*X=2005*
Stock of Migrants	0.880***	0.889***	0.868***	0.819***	0.759***	0.722***	0.701***	0.669***	0.646***
(Province-Level 1996)	(0.055)	(0.049)	(0.047)	(0.054)	(0.058)	(0.062)	(0.063)	(0.062)	(0.062)
Observations	1,811	2,125	2,221	2,318	2,333	2,345	2,351	2,365	2,365
	X=2006	*X=2007*	*X=2008*	*X=2009*	*X=2010*	*X=2011*	*X=2012*	*X=2013*	
Stock of Migrants	0.638***	0.621***	0.622***	0.624***	0.627***	0.630***	0.632***	0.633***	
(Province-Level 1996)	(0.061)	(0.061)	(0.059)	(0.058)	(0.058)	(0.057)	(0.057)	(0.056)	
Observations	2,365	2,365	2,365	2,365	2,365	2,365	2,365	2,365	

†Explanatory variables are in natural logs. All regressions include the trade flow at the province-level (in 1997), country-and-region fixed effects, as well as world region-and-province fixed effects. Heteroskedasticity-robust standard errors (clustered by pairs of countries of origin and regions of destination) are given in parentheses. *,**,*** denote significance at the 10%, 5%, 1% levels, respectively. The regressions include all countries of origin with at least 630 nationals residing in Spain in 1996 (55 countries of origin). Missing observations to 2,365 are due to zero variation in the migration flow within clusters. See Section 3 for a detailed description of all variables.

D Measurement Error

We argue that the potential non-stochastic measurement errors discussed at the end of Section 4.1 are unlikely to result in biased estimates. Let $\tilde{m}_{ij} < m_{ij}$ and $\tilde{M}_{ij} > M_{ij}$ denote the unobserved true size of the migrant flow and the migrant network, respectively. Let the relationship between the migrant flow and the migrant network be given by the following equation:

$$\ln(\tilde{m}_{ij}) = \eta_{zr} \ln(\tilde{M}_{ij}). \tag{D.1}$$

Let y_{ij} denote the ratio of unobserved (i.e. "excess") migrants to observed migrants in the flow, and let x_{ij} denote the ratio of unobserved (i.e. unregistered) migrants to observed migrants in the network. Hence, $\tilde{m}_{ij} = (1 - y_{ij})m_{ij}$ and $\tilde{M}_{ij} = (1 + x_{ij})M_{ij}$ and thus:

$$\ln((1 - y_{ij})m_{ij}) = \eta_{zr} \ln((1 + x_{ij})M_{ij}), \tag{D.2}$$

which can be rewritten as:

$$\ln(m_{ij}) = \eta_{zr} \ln(M_{ij}) + \eta_{zr} \ln(1 + x_{ij}) - \ln(1 - y_{ij}). \tag{D.3}$$

The last two terms in equation (D.3), if not controlled for, may introduce a bias in the estimation of the network coefficient η_{zr}. Obviously, a sufficient condition for our FE model controlling for country-and-region fixed effects to deliver unbiased estimates is:

$$v_{ij} = v_{ir}, \quad v = \{x, y\}. \tag{D.4}$$

Hence, the type of mismeasurement potentially present in our migration data is not a problem *per se* for the estimation. For example, suppose that migrants are possibly measured with error, so that $x_{ij} \leq 0$ and $y_{ij} \leq 0$ for all provinces in Spain. Furthermore suppose that these errors are large for some regions of destination but small for others, and that they are large for some countries of origin but small for others. Then, a mild but sufficient condition for our estimates to be unbiased is: $x_{ij} = x_{ik}$ and $y_{ij} = y_{ik}$, where $j \neq k$ and $j, k \in A_{zr}$.

E Further Regressions for the Skill Structure of Migration

Table E.1: Skill Structure of Migration – Poisson Model (PPML)[†]

	Dependent Variable: Migrant Skill Ratio (Region-Level 2002-2006)					
	(a)	(b)	(c)	(d)	(e)	(f)
Stock of Migrants	-0.392**	-0.403***	-0.390***	-0.635***	-0.672***	-0.647***
(Region-Level 2002)	(0.155)	(0.155)	(0.149)	(0.143)	(0.126)	(0.132)
FDI Flow			-0.025			-0.020
(Region-Level 1998-2001)			(0.021)			(0.015)
Trade Flow			0.012			0.026
(Region-Level 2001)			(0.098)			(0.149)
Language		0.392	0.408		0.010	0.088
(Region-Level)		(0.351)	(0.334)		(0.224)	(0.189)
Distance		-0.264	-0.350		-1.179	-1.164
(Region-Level)		(0.476)	(0.443)		(1.797)	(1.796)
Region Effects	Yes	Yes	Yes	Nested	Nested	Nested
Country Effects	Yes	Yes	Yes	Yes	Yes	Yes
World Region-and-Region Effects	No	No	No	Yes	Yes	Yes
Observations	234	234	234	234	234	234

[†]All explanatory variables except for the language dummy are in natural logs. Heteroskedasticity-robust standard errors (clustered by countries of origin) are given in parentheses. *,**,*** denote significance at the 10%, 5%, 1% levels, respectively. See Section 3 for a detailed description of all variables.

F Testing for Sample Selection Bias

We briefly present our procedure for identifying a potential sample selection bias in the model for the skill structure of migration. It is a slight modification of Wooldridge (1995, 123-124), who proposes a method for testing for sample selection bias in panel data. It will become evident below that we impose very strong assumptions on the selection equation and the mechanism governing selection. These assumptions would often be inappropriate if we were to derive *corrections* for a sample selection bias in models with fixed effects. It turns out, however, that they do not pose a threat to the correct *testing* for a sample selection bias. For further details on this, the reader is referred to Wooldridge (1995).

We start by rewriting the model for the skill structure of migration as:

$$y_{ij} = \mu_i + \mathbf{x}_{ij}\boldsymbol{\beta} + u_{ij}, \qquad j = 1, \ldots, J, \tag{F.1}$$

where y_{ij} is the ij-specific log of the ratio of high-skilled migrations to low-skilled migrants, μ_i is an unobserved country fixed effect, $\mathbf{x}_{ij}$ is a $1 \times K$ vector of explanatory variables (including region dummies and interactions between region dummies and world region dummies), $\boldsymbol{\beta}$ is a $K \times 1$ vector of parameters to be estimated, and u_{ij} is an independent and identically distributed error term. We explicitly allow for $E(\mu_i|\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iJ}) \neq E(\mu_i)$. Since J is fixed, the asymptotic analysis is valid for $I \to \infty$. Now suppose that $(y_{ij}, \mathbf{x}_{ij})$ is sometimes unobserved, and that $\mathbf{s}_{ij} = (s_{i1}, \ldots, s_{iJ})'$ is a vector of selection indicators with $s_{ij} = 1$ if $(y_{ij}, \mathbf{x}_{ij})$ is observed and zero otherwise. Define $\mathbf{x}_i \equiv (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iJ})$ and $\mathbf{s}_i \equiv (\mathbf{s}_{i1}, \ldots, \mathbf{s}_{iJ})$ and suppose that $E(u_{ij}|\mu_i, \mathbf{x}_i, \mathbf{s}_i) = 0 \ \forall j$, which implies that the selection process is strictly exogenous conditional on μ_i and $\mathbf{x}_i$. Then, our FE estimator employed in the main text is consistent and asymptotically normal even when selection arbitrarily depends on $(\mu_i, \mathbf{x}_i)$ (Wooldridge 1995, 118).

In our application, the explanatory variables $\mathbf{x}_{ij}$ are observed for all regions $j = 1, \ldots, J$. The variable y_{ij} is observed if $s_{ij} = 1$, but not otherwise. For each $j = 1, \ldots, J$, define an unobserved latent variable

$$h_{ij}^* = \boldsymbol{\delta}_{j0} + \mathbf{x}_{i1}\boldsymbol{\delta}_{j1} + \cdots + \mathbf{x}_{iJ}\boldsymbol{\delta}_{jJ} + v_{ij}, \tag{F.2}$$

where v_{ij} is a stochastic term independent of $(\mu_i, \mathbf{x}_i)$, and $\boldsymbol{\delta}_{jp}$ is a $(K+1) \times 1$ vector of unknown parameters, $p = 1, 2, \ldots, J$.[54] The binary selection indicator is defined as $s_{ij} \equiv 1[h_{ij}^* > 0]$. Since $\mathbf{s}_i$ is a function of $(\mathbf{x}_i, \mathbf{v}_i)$, where $\mathbf{v}_i \equiv (v_{i1}, \ldots, v_{iJ})'$, a sufficient condition for the selection process to be strictly exogenous conditional on μ_i and $\mathbf{x}_i$ is:

$$E(u_{ij}|\mu_i, \mathbf{x}_i, \mathbf{v}_i) = 0, \qquad j = 1, \ldots, J. \tag{F.3}$$

Under (F.3), there is no sample selection bias. An alternative that implies sample selection bias is:

$$E(u_{ij}|\mu_i, \mathbf{x}_i, \mathbf{v}_i) = E(u_{ij}|v_{ij}) = \rho v_{ij}, \qquad j = 1, \ldots, J, \tag{F.4}$$

[54] In the following, $\mathbf{x}_{ij}$ includes one element more than in equation (F.1), despite the fact that we use the same notation for convenience. We thus assume that there is exactly one exclusion restriction in equation (F.1). In the estimation, we use the log of the number of people holding country i's nationality and migrating from region j in Spain to any other region $k \neq j$ within or outside Spain over the period from January 1, 2006, to December 31, 2007, as an exclusion restriction.

where $\rho \neq 0$ is some unknown scalar. Under the alternative (F.4) we have:

$$E(y_{ij}|\mu_i, \mathbf{x}_i, \mathbf{s}_i) = \mu_i + \mathbf{x}_{ij}\boldsymbol{\beta} + \rho E(v_{ij}|\mu_i, \mathbf{x}_i, \mathbf{s}_i) = \mu_i + \mathbf{x}_{ij}\boldsymbol{\beta} + \rho E(v_{ij}|\mathbf{x}_i, \mathbf{s}_i). \tag{F.5}$$

Let $E(v_{ij}|\mathbf{x}_i, \mathbf{s}_i) = E(v_{ij}|\mathbf{x}_i, s_{ij})$ and assume a standard uniform distribution for v_{ij}. Then,

$$E(v_{ij}|\mathbf{x}_i, s_{ij} = 1) = E(v_{ij}|\mathbf{x}_i, v_{ij} > -\mathbf{x}_i\boldsymbol{\delta}_j) = (1 + \mathbf{x}_i\boldsymbol{\delta}_j)/2. \tag{F.6}$$

and

$$E(y_{ij}|\mu_i, \mathbf{x}_i, s_{ij} = 1) = \rho^* + \mu_i + \mathbf{x}_{ij}\boldsymbol{\beta} + \rho^*\mathbf{x}_i\boldsymbol{\delta}_j, \tag{F.7}$$

where $\rho^* \equiv \rho/2$ and $\mathbf{x}_i$ now includes unity as its first element. The procedure to test for sample selection bias is as follows. We first obtain estimates of $\mathbf{x}_i\boldsymbol{\delta}_j$ by estimating region-specific selection equations (where s_{ij} is the dependent variable) derived from equation (F.2), using linear probability models for the full data matrix. We then estimate equation (F.7) in an FE framework (within-transformed data), using only observations with $s_{ij} = 1$. We finally test $H_0 : \rho = 0$, using the t-statistic for ρ^*.

Economics Working Papers

2014-15: Bo Sandemann Rasmussen: An Interpretation of the Gini Coefficient in a Stiglitz Two-Type Optimal Tax Problem

2014-16: A. R. Lamorgese, A. Linarello and Frederic Warzynski: Free Trade Agreements and Firm-Product Markups in Chilean Manufacturing

2014-17: Kristine Vasiljeva: On the importance of macroeconomic factors for the foreign student's decision to stay in the host country

2014-18: Ritwik Banerjee: On the Interpretation of Bribery in a Laboratory Corruption Game: Moral Frames and Social Norms

2014-19: Ritwik Banerjee and Nabanita Datta Gupta: Awareness programs and change in taste-based caste prejudice

2014-20: Jos Jansen and Andreas Pollak: Strategic Disclosure of Demand Information by Duopolists: Theory and Experiment

2014-21: Wenjing Wang: Do specialists exit the firm outsourcing its R&D?

2014-22: Jannie H. G. Kristoffersen, Morten Visby Krægpøth, Helena Skyt Nielsen and Marianne Simonsen: Disruptive School Peers and Student Outcomes

2014-23: Erik Strøjer Madsen and Yanqing Wu: Globalization of Brewing and Economies of Scale

2014-24: Niels-Hugo Blunch and Nabanita Datta Gupta: Social Networks and Health Knowledge in India: Who You Know or Who You Are?

2014-25: Louise Voldby Beuchert and Anne Brink Nandrup: The Danish National Tests – A Practical Guide

2015-01: Ritwik Banerjee, Tushi Baul, and Tanya Rosenblat: On Self Selection of the Corrupt into the Public Sector

2015-02: Torben M. Andersen: The Nordic welfare model and welfare services - Can we maintain acceptable standards?

2015-03: Nina Neubecker, Marcel Smolka and Anne Steinbacher: Networks and Selection in International Migration to Spain