The World Migration Network: rankings, groups and gravity models

Quentin Cappart
Adrien Thonet
Université catholique de Louvain
Place de l'Université 1, 1348, Belgium
E-mail: quentin.cappart@uclouvain.be

***Abstract*—Human migrations, ever accelerated in a globalised world, is a growing topic of scientific investigation that calls for novel analytic methodologies able to disintricate the dynamic migratory flows across the world. This contribution studies several decades of global migration data from a network perspective. Tools developed in various fields such as PageRank, community detection or gravity models are analysed and applied to diverse aspects of migrations.**

I. Introduction

Human migrations have been demonstrated to exert a large influence on the economy [1], education [2], health [3] or labour market [4] of a country. As such it constitutes a growing field of research [5]. Network science, now a common tool in sociology [6], mobility [7], trade [8], etc. has not been used for an extensive study of the migration phenomenon, except for Fagiolo and Mastrorillo [9]. Migration data however adopt a natural network structure, where nodes represent countries and edges, migratory fluxes, weighted with a number of recorded migrants over a considered period. A network perspective is therefore natural in this context.

In this paper we focus more particularly on three topics of interests:

1) **Ranking**: each country attracts different numbers of migrants. An empirical measure of attractiveness is simply the weighted in-degree, i.e. the total number of immigrants into a country. However this measure is noisy and highly varying with time. A measure of attractiveness can also be composed from different possible explaining factors ranging from wealth to education level, or living conditions in the destination country [10]. But such a measure requires extra data and clarification of assumptions. In this work we propose a robust ranking based solely on migration data itself. The method used to obtain this ranking is called the *PageRank*.

2) **Group detection**: one shortcoming of the ranking analysis is not considering the relations that a country can have with others. The group detection analysis overcomes this lack by detecting groups of countries having strong relations together. The goal is to gather countries having important migration flows together into the same group. Group, or *community*, detection is a mature field of network science. As we will see, applying such a method can highlight unexpected relations between countries.

3) **Prediction of future migrations**: a highly studied topic in the migration literature consists in elaborating *gravity models* describing and predicting the flow of migrants between two countries, from explaining factors such as distance or the living conditions of both countries. In this paper, we propose new formulas based on the ranking measures and the community partition mentioned above, and achieve better accuracy than the state of the art models.

Before performing such analysis there is a need to have reliable data about migrations. The work presented here is based on the data of the World Bank [11] which contain about 17 million migrations that has been made between 1990 and 2000. This represents about 740 000 migrations by country, including islands and dependent states. Other data from the World Bank like population, gross domestic product, etc. are also used.

This paper is organized as follows. In Section II, the PageRank is described for the task of ranking countries instead of simple methods which present some shortcomings. Section III discusses about the group detection. Finally, Section IV presents our different gravity models.

II. Ranking analysis

A simple way to rank countries is to consider only the number of immigrants. Applied with the migration data, this method gives the ranking presented on Table I.

Table I: Most attractive countries according to the number of immigrants

Ranking	Country		
1	United States	6	Canada
2	Russian Federation	7	Ukraine
3	Germany	8	Saudi Arabia
4	France	9	United Kingdom
5	India	10	Australia

978-1-4799-8569-2/15/$31.00 ©2015 IEEE

We can observe that populated countries occupy the top of the ranking which highlights a big shortcoming of the method: the populated countries are too advantaged compared to lowly populated. It occurs because the more populated is a country, the more will be the probability to have migrants.

A clever method is thereby required to address this issue. A solution is to divide the number of immigrants by the population of their home country.

Table II: Most attractive countries according to the number of immigrants divided by the population

Ranking	Country		
1	Kuwait	6	Not. Mar. Islands
2	Qatar	7	Cayman Islands
3	United Arab Emirates	8	Macao SAR, China
4	Monaco	9	Falkland Islands
5	Andorra	10	Virgin Islands (US)

Table II presents the ranking obtained from this idea. However, these results show that the lowly populated countries are now far more advantaged. One way to get rid of the population in the computation is to elaborate a ranking based on ratio of immigrations and emigrations. Mathematically, it is expressed like this:

$$\text{ratio} = \frac{\text{weighted in-degree}}{\text{weighted out-degree}}.$$

Table III: Most attractive countries according to the $\frac{\text{weighted in-degree}}{\text{weighted out-degree}}$ ratio

Ranking	Country		
1	Qatar	6	United States
2	Mayotte	7	Cayman Islands
3	United Arab Emirates	8	Gabon
4	Saudi Arabia	9	French Guiana
5	Djibouti	10	Andorra

Although the results of Table III do not seem to advantage countries with their population, a last limitation remains: the importance of countries is not taken into account. Importance of a country can be defined in a recursive way. If a country receives migrants from an important country it will gain more importance than for a less important country. Not considering this limitation leads to have good local attractors (Mayotte, French Guiana, Gabon or Djibouti) on the top of the ranking at the expense of global attractors. However, the objective pursed is to have a ranking of global attractors. To do so, Importance of the countries must be taken into account. It directly leads to a more sophisticated method, the PageRank.

The idea of the PageRank [12] is related to the random walk. The concept of the random walk captures the behaviour of a random migrant who moves out from country to country following the migratory flow and sometimes decides to go to any country randomly. Given that weighted flow are considered, if the number of migrants going from a country to another is high, the probability that this migrant will follow this flow will be also high. By assuming that he moves out an infinity of times, the PageRank of a country is defined as the proportion of times this random migrant has been in this country. This is also referred in the literature as the stationary probability.

The PageRank has thereby the idea that all the connections are not equal, it captures instead the importance of the countries: if a country receives migrants from an 'important' country, it will gain more importance than if the origin country is less 'important'. Furthermore, it is not the population that is taken into account but only the proportion of migrants of the different countries.

To obtain the PageRank, we need to formalise it into a computable expression. First of all, let us define A_{ij} the migration adjacency matrix where are stored migrations from country i to country j. In the same way, we define $H_{ij} = A_{ij}/\sum_{j=1}^{N} A_{ij}$, The migration stochastic matrix containing the proportion of migrants for each country.

Besides, we need to add the probability to go to any countries randomly. We obtain the expression

$$\mathbf{G} = \theta\ \mathbf{H} + (1-\theta)\frac{1}{N}\mathbf{1_{N\times N}} \tag{1}$$

where N is the size of the matrix (i.e. the number of countries), $\theta \in [0,1]$ the proportion of times the random walker follows the flow and $(1-\theta)$ the proportion where the random walker moves out randomly to any country. The purpose of the term $\frac{1}{N}\mathbf{1_{N\times N}}$ is to have a matrix linking every country to all other countries with the same probability. Therefore, the larger is θ, the larger will be the probability to follow the flow of migrants.

Finally, the PageRank corresponds to the left eigenvector related to the highest eigenvalue of $\mathbf{G}$. Using Perron-Frobenius theorem [13], it can be shown that this eigenvalue is equal to 1 and is unique. According to this theorem, the PageRank always exists and as mentioned above, it is equal to the stationary probability. This leads us to the computation of the system of equations

$$\mathbf{G}^T\ \pi = \pi \tag{2}$$

where π is the PageRank. This system can be solved with iterative methods as Power method [14]. It is the method we used to compute the PageRank of migrations.

Table IV: Most attractive countries according to their PageRank

Ranking	Country		
1	United States	6	France
2	Canada	7	West Bank and Gaza
3	United Kingdom	8	Mexico
4	Germany	9	Puerto Rico
5	Australia	10	Saudi Arabia

Table IV shows the results obtained. In order to be sure about the relevance of this ranking, we identified for each country reasons explaining their presence on this top. We obtained several reasons:

- A high Human Development Index (HDI): Australia (2^{nd}), United States (3^{rd}), Canada (6^{th}), Germany (12^{th}), France (18^{th}).
- A High Gross Domestic Product (GDP): United States (1^{st}), Germany (3^{rd}), United Kingdom (4^{th}), France (5^{th}), Canada (8^{th}), Mexico (9^{th}), Australia (14^{th}).
- Tax haven countries: Puerto Rico.
- Oil producing countries: Saudi Arabia.
- Special event involving countries: West Bank and Gaza [15] [16].
- Major emigrations of the United States countries: Mexico (16%), Canada (12.4%), Puerto Rico (11.2%), United Kingdom (7.6%), Germany (5.3%), France (3.8%).

Concerning the last reason, as the United States is by far the most attractive country according to the PageRank evaluation, the major emigrations coming from it give to the destination country a good place in the ranking. Nevertheless, other attractive countries rise in top of the ranking for other reasons which supports the idea that the PageRank of migrations gives a good indicator of attractiveness.

Besides, the PageRank idea can be extended in order to obtain more information about countries. For instance, by reversing all the edges of the migration graph (i.e. immigrants for a country become emigrants and vice versa) and applying the PageRank we obtain a ranking of repulsiveness. The two PageRanks can also be combined by taking their ratio in order represent both rankings together. This result is presented on Figure 1 where green countries represent attractive countries and red repulsive.

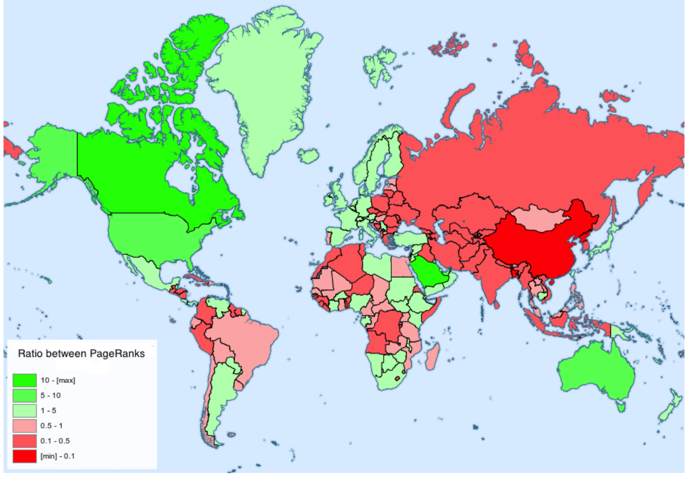

Figure 1: Map of ratio of PageRanks

III. Group detection

On this section, we will analyse how countries can be gathered using the concept of communities. To understand it, let us first define what a graph clustering is.

Definition III.1 (Graph clustering)**.** *A graph clustering is a classification of the nodes of the graph into groups where the repartition of the nodes tends to optimise a configuration such as:*

- *The connections between nodes within the same group are strong.*
- *The connections between nodes of different groups are weak.*

Following this definition, communities correspond to the groups of vertices found with a graph clustering method. The challenge behind the communities detection is to efficiently form clusters. A solution proposed by Newman and Girvan [17] considers the community problem with another point of view. They designed a metric, called the modularity (Q), having the purpose of measuring the quality of a graph partition into communities. The intuition behind this measure is to compare the density of the connections within a same community with the expected density for a same community partition [18]. The expected density means that we consider a randomised graph having the same number of nodes where every node keeps the same degree but where the edges are placed randomly.

The higher is the modularity Q, the better is the partitioning. The problem of finding the best partition turns thereby to maximise the modularity. Firstly, we need to have a computable expression for Q. For a directed weighted graph such as the migration network, an expression of modularity can be obtained with the formula [19]:

$$Q = \frac{1}{m}\sum_{i,j}\left[A_{ij} - \frac{k_i^{out}k_j^{in}}{m}\right]\delta(c_i, c_j) \qquad (3)$$

where

- A_{ij} is the weight of the edge from the node i to the node j.
- $m = \sum_{i,j} A_{ij}$ is the sum of all the weighted edges.
- k_i^{in} is the weighted in-degree of the node i.
- k_i^{out} is the weighted out-degree of the node i.
- c_i is the community of the node i.
- $\delta(c_i, c_j)$ equals 1 if i and j are in the same community, 0 otherwise.
- $(k_i^{out}k_j^{in})/m$ corresponds to the probability to have an edge from the node i to j in a random graph having the same configuration than ours.

The task is now to find the partitioning producing the highest modularity Q. A naive solution is to consider all the partitions and to select the one having the highest Q. However, the problem of finding an optimal partitioning is known to be NP-complete. For this reason, the solutions requiring to enumerate all the partitions are infeasible in practice.

To deal with this issue, Blondel et al. proposed a greedy algorithm, called the Louvain method [18], aiming to optimise the modularity. Using this algorithm, we obtain the map presented on Figure 2.

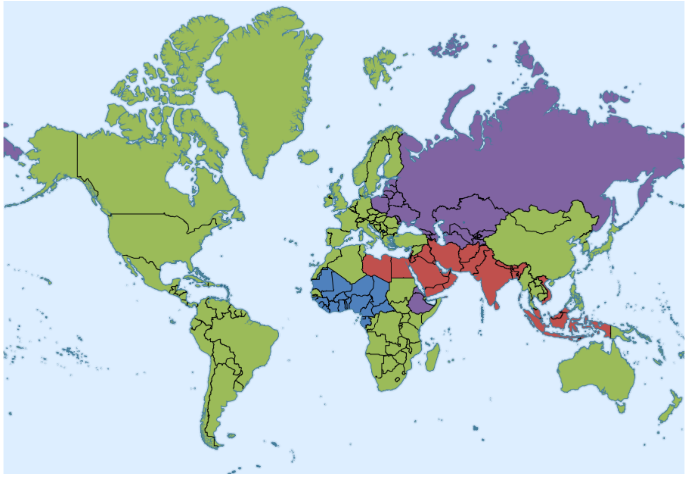

Figure 2: Communities map obtained with the Louvain method

Four groups of countries are exhibited on this map:

- The West Africa.
- A portion of the Middle East, the Indian subcontinent and the south of the Far East.
- Countries belonging mainly to the former USSR.
- A last community taking over the rest of the world.

One may ask why Ethiopia is in the same community than the former USSR countries (in purple). The reason can be explained by the historic fact stating that Ethiopia has long standing relations since the 17^{th} century with Russia [20].

The partition of countries into communities can thereby be used to discover not expected facts about countries. Furthermore, as we will see in the next section, communities can also be used to design gravity models.

IV. Gravity model

This section deals with the issue of obtaining unknown migration data using characteristics of countries. The interest of such a study is threefold: predicting future migrations, finding actual missing migration data, and identifying the main factors that explain the migratory flows.

The most used method in the literature to perform such a prediction is the gravity model [21]. The intuitive idea is to design a mathematical expression describing how countries attract people. A parallel can be done with the Newton's law of gravitation which describes the attraction of bodies. There are two families of gravity models. On the one hand, the bilateral gravity models (e.g. [22]) only considering the origin and the attractiveness of the destination, and on the other hand the multilateral gravity models (e.g. [23]) also considering the attractiveness of the alternative destinations.

This paper focus on bilateral gravity models. To define the attractiveness between countries, state of the art models [22], [24], [25] use such kinds of factors:

- **Geographic factors**: distance between countries, common boundaries, etc.
- **Linguistic factors**: common language, English spoken, etc.
- **Socio-economic factors**: population, GDP, HDI, etc.
- **Historic factors**: old migrations, former colony, etc.
- **Specific factors**: polygamy, fertility rate, etc.

Our motivation is to build a gravity model which is general and easy to explain. For example, taking parameters such as the polygamy or the fertility rate into account seems to be more arbitrary than considering the GDP or the population. Furthermore, these factors can be gathered into three forces: the attractiveness of a country (high GDP, high HDI, etc.), its repulsiveness (low GDP, low HDI, etc.), and particular relations between two countries (distance between them, sharing a common language, etc.).

We obtained in the previous sections mathematical expressions describing these forces. The idea is to use them to build gravity models. Concretely, we introduced three mathematical parameters:

- The PageRank of the destination country which is a measure of the attractiveness of the destination.
- The inverted PageRank of the origin country which is a measure of the repulsiveness of the origin.
- The community of countries which is a measure of the intensity of connections between countries.

To the best of our knowledge, none of these parameters have already been used to build gravity models. We will now explain how a computable expression for a gravity model can be obtained.

$$F_{ij} = \alpha_0 A_{1_{ij}}^{\alpha_1} A_{2_{ij}}^{\alpha_2} A_{3_{ij}}^{\alpha_3} \ldots A_{N_{ij}}^{\alpha_N} E_{ij} \tag{4}$$

Equation (4) shows the canonical form of a gravity model where

- F_{ij} is the migration flow from country i to country j.
- $A_{(1..N)_{ij}}$ are the values of the parameters used to explain the flow from i to j.
- $\alpha_{(1..N)}$ are the coefficients related to the parameters.
- E_{ij} is a term of error with an expectation of 1 ($\mathbb{E}(E_{ij}|A_{1_{ij}}, \ldots, A_{N_{ij}}) = 1$).

The principle is to determine the coefficients α from known values of flows and parameters through a regression. The resulting equation can be used thereafter to determine unknown flows if the parameters are known.

Silva and Tenreyro [26] propose to use a Poisson Pseudo-Maximum Likelihood (PPML) regression which is robust against the zero value problem and the heteroscedasticity. For this reason, this is the method we used. The quality of such a regression can be determined using the McFadden's pseudo R-squared metric [27].

To obtain the most accurate model, we built a particular model for each possible combination of parameters, and then we selected the one providing the highest R-squared value. The parameters forming our best model are the following:

1) The past migrations between two countries (0.839).
2) Sharing a common border (0.180).
3) The PageRank of the destination country (0.128).
4) The GDP ratio between the two countries (0.094).
5) The distance between the two countries (-0.092).
6) The population of the origin country (0.086).
7) The population of the destination country (0.045).

The coefficients α obtained for each parameter is indicated in parenthesis. Furthermore, it is important to make the distinction between the migration flow F and the previous migrations parameter. Past migrations parameter is commonly seen as a proxy parameter to model diaspora which are known to have a major impact on the future migrations [28]. If we build a model based on migrations occurring between 1900 and 2000, the flow F is related to the migrations that have occurred during this period while the diaspora is related to migrations which have occurred before. Concerning the mathematical parameters, they are computed from the data of flow F which must also be known to build the model. In other words, the main idea of the mathematical parameters is to give a compact description of the flow F that reveals its underlying structure.

Table V shows the performance of our model and compares it with the state of the art gravity model.

Table V: Comparison between our model and Artuc, Docquier et al. model with past migrations

	# parameters	R-squared
Artuc, Docquier et al. [22]	15	0.898
Our model	7	0.9565

With a R-squared of 0.9565 and 7 parameters, we can see that our model outperforms the other one with less parameters. However, when analysing the coefficients α obtained, we can observe that the past migrations parameter is much more significant (0.839) than the others (0.180 for the second one). This observation is accentuated when comparing the loss of the R-squared value if a particular parameter is withdrawn from the model. For the past migrations, a loss of 0.3154 is recorded whereas the second most important loss is of only 0.0029. The accuracy of the model is thereby almost determined only by this single parameter. Artuc, Docquier et al. model [22] was actually mainly outperformed by the fact than our model contained a better approximation for the diaspora.

However, this parameter is far more costly to obtain than the other parameters. While other parameters only need a piece of information for each country, modelling a diaspora requires information for each pair of countries and for each year period. For instance, past migrations require 192 080 ($196^2 \times 5$) data if 5 periods and 196 countries are considered where classical parameters only require 196 data each. At the other extreme, the mathematical parameters that we introduced do not require any new information to be obtained, they only rely on the data of the flow F.

Several gravity models in the literature [24], [25] are built without resorting to past migrations. Following the same idea, we obtained a new model with the following parameters:

1) Sharing a common border (2.060).
2) Belonging to a same community (1.835).
3) Sharing a common language (0.797).
4) The PageRank of the destination country (0.746).
5) The inverted PageRank of the origin country (0.686).
6) Speaking English in the origin country (-0.484).
7) The population of the origin country (0.320).
8) The population of the destination country (0.286).
9) The GDP ratio between the two countries (0.166).
10) The distance between the two countries (-0.053).

Table VI recaps the performance obtained and compares it with competitive models.

Table VI: Comparison between our model and existing models without past migrations

	# parameters	R-squared
Lewer and Van den Berg [24]	10	0.663
Ramos and Surinach [25]	13	0.634
Our model	10	0.7457

With a R-squared of 0.7457, our model outperforms others. Besides, unlike the model using past migrations, Table VII shows that now the accuracy of the model is not determined by a single parameter.

Table VII: The five most important R-squared value loss for the model without the past migrations

Parameter	R-squared value loss
4)	-0.0835
1)	-0.0752
2)	-0.0545
5)	-0.0257
7)	-0.0169

Among the five most representative parameters, three of them have been developed in this paper. This confirms the validity of the previous chapters where the new parameters were introduced. To the best of our knowledge, even with more specific database and with more parameters than what we are using, no model of international migrations competes with ours.

V. Conclusion

This paper used and modified several well-known mathematical concepts of network science to analyse the migration flows and proved their validity by applying them to econometric models which gave excellent results according to the metrics commonly used in this research field. Three major points in the field of migrations were developed: ranking countries, grouping them into consistent groups and elaborating gravity models.

Our contribution is twofold. So far, methods from network science has not been much used to analyse migrations. We showed however that such methods lead to consistent results. Network science can thereby be a new way to analyse migrations. Furthermore, by using these results we presented innovative gravity models having better performances than the state of the art models.

VI. Acknowledgments

We would like to thank the people who directly or indirectly contributed to this work. Firstly, our supervisor, Jean-Charles Delvenne, for his continuous support. And then, Jean-François Carpantier and Yves Deville for their reading of this work.

References

[1] The Levin Institute - The State University of New York, “Globalization101 : Economic effects of migration,” 2014, http://www.globalization101.org/economic-effects-of-migration/.

[2] C. Dustmann and A. Glitz, “Migration and education,” Handbook of the Economics of Education, vol. 4, pp. 327–439, 2011.

[3] M. Kristiansen, A. Mygind, and A. Krasnik, “Health effects of migration.” Danish medical bulletin, vol. 54, no. 1, pp. 46–47, 2007.

[4] G. J. Borjas, R. B. Freeman, and L. F. Katz, “On the labor market effects of immigration and trade,” pp. 213–244, 1992.

[5] OECD UNDESA, “World Migration in Figures,” no. October, pp. 1–6, 2013.

[6] S. P. Borgatti, “2-mode concepts in social network analysis,” Encyclopedia of complexity and system science, pp. 8279–8291, 2009.

[7] A. De Montis, M. Barthélemy, A. Chessa, and A. Vespignani, “The structure of interurban traffic: a weighted network analysis,” Environment and Planning B: Planning and Design, vol. 34, no. 5, pp. 905–924, 2007, http://www.envplan.com/abstract.cgi?id=b32128.

[8] L. De Benedictis and L. Tajoli, “The world trade network,” The World Economy, vol. 34, no. 8, pp. 1417–1454, 2011.

[9] G. Fagiolo and M. Mastrorillo, “International migration network: Topology and modeling,” Physical Review E, vol. 88, no. 1, p. 012812, Jul. 2013, http://link.aps.org/doi/10.1103/PhysRevE.88.012812.

[10] O. Bakewell, “South-south migration and human development: Reflections on african experiences,” 2009.

[11] World Bank Group, C. Ozden, C. R. Parsons, M. Schiff, and T. L. Walmsley, “World Bank Economic Review : Where on Earth is Everybody? The Evolution of Global Bilateral Migration 1960-2000,” Tech. Rep., http://data.worldbank.org/data-catalog/global-bilateral-migration-database.

[12] M. Chiang, “Networked life: 20 questions and answers,” 2012.

[13] S. U. Pillai, T. Suel, and S. Cha, “The perron-frobenius theorem: some of its applications,” pp. 62–75, 2005.

[14] L. Eldén, “A Note on the Eigenvalues of the Google Matrix,” pp. 1–3, 2003.

[15] T. M. Rempel, “Palestinian Refugees in the West Bank and the Gaza Strip,” 2006, http://www.forcedmigration.org/research-resources/expert-guides/palestinian-refugees-in-the-west-bank-and-the-gaza/alldocuments.

[16] A. Di Bartolomeo, T. Jaulin, and D. Perrin, “Palestine,” no. July, 2011.

[17] M. E. J. Newman, “The mathematics of networks,” pp. 1–12.

[18] V. Blondel, J.-l. Guillaume, R. Lambiotte, and E. Lefebvre, “Fast unfolding of communities in large networks,” J. Stat. Mech. (2008) P10008.

[19] Y. Liu, Q. Liu, and Z. Qin, “Community Detecting and Feature Analysis in Real Directed Weighted Social Networks,” Journal of Networks, vol. 8, no. 6, pp. 1432–1439, Jun. 2013, http://ojs.academypublisher.com/index.php/jnw/article/view/10238.

[20] Ministry of Foreign Affairs of Ethiopia, “Ethiopia-Russia relations,” http://www.mfa.gov.et/BilateralMore.php?pg=25.

[21] J. E. Anderson, “The Gravity Model,” Annual Review of Economics, vol. 3, no. 1, pp. 133–160, Sep. 2011, http://www.annualreviews.org/doi/abs/10.1146/annurev-economics-111809-125114.

[22] E. Artuc, F. Docquier, C. Ozden, and C. R. Parsons, “A Global Assessment of Human Capital Mobility : the Role of non-OECD Destinations,” http://perso.uclouvain.be/frederic.docquier/oxlight.htm.

[23] S. Bertoli and J. F.-H. Moraga, “Multilateral resistance to migration,” Journal of Development Economics, vol. 102, pp. 79–100, 2013.

[24] J. J. Lewer and H. Van den Berg, “A gravity model of immigration,” Economics Letters, vol. 99, no. 1, pp. 164–167, Apr. 2008, http://linkinghub.elsevier.com/retrieve/pii/S0165176507002455.

[25] R. Ramos and J. Suriñach, “A gravity model of migration between ENC and EU,” 2013.

[26] J. S. Silva and S. Tenreyro, “The log of gravity,” The Review of Economics and statistics, vol. 88, no. 4, pp. 641–658, 2006.

[27] A. C. Cameron and F. A. Windmeijer, “An r-squared measure of goodness of fit for some common nonlinear regression models,” Journal of Econometrics, vol. 77, no. 2, pp. 329–342, 1997.

[28] M. Beine, F. Docquier, and Ç. Özden, “Diasporas,” Journal of Development Economics, vol. 95, no. 1, pp. 30–41, 2011.