Migrant Networks and the Spread of Misinformation*

Benjamin Elsner† Gaia Narciso‡ Jacco J. J. Thijssen§

January 26, 2016

Abstract

Diaspora networks provide information to future migrants, which affects their success in the host country. While the existing literature explains the effect of networks on the outcomes of migrants through the size of the migrant community, we show that the quality of the network is an equally important determinant. We argue that networks that are more integrated in the society of the host country can provide more accurate information to future migrants about job prospects. In a decision model with imperfect signalling, we show that migrants with access to a better network are more likely to make the right decision, — namely, they migrate only if they gain. We test these predictions empirically using data on recent Mexican migrants to the US. To instrument for the quality of networks, we exploit the settlement of immigrants who came during the Bracero program in the 1950s. The results provide strong evidence that connections to a better-integrated network lead to better outcomes after migration.

JEL codes: F22, J15, J61

*We would like to thank Simone Bertoli, Herbert Brücker, Joanna Clifton-Sprigg, Tommaso Colussi, Margherita Comola, Christian Danne, Rachel Griffith, Joachim Jarreau, Julia Matz, Imran Rasul, Bas ter Weel, Mathis Wagner, and seminar participants at IZA, University College Dublin, the University of Southern Denmark, as well as the conferences ISNE, TEMPO, IEA, ESEM, OECD immigration workshop, RES, NORFACE, ESPE, EALE, IZA/SOLE for helpful comments. Elsner gratefully acknowledges funding from the Irish Research Council for the Humanities & Social Sciences (IRCHSS).

†Institute for the Study of Labor (IZA). Address: Schaumburg-Lippe-Str. 5-9, 53113 Bonn, Germany. elsner@iza.org.

‡Corresponding Author. Trinity College Dublin and CReAM. narcisog@tcd.ie.

§University of York, and The York Management School. jacco.thijssen@york.ac.uk.

1 Introduction

Prior to moving, migrants face significant uncertainty about their job prospects abroad, which is why they often seek advice from existing diaspora networks. A large amount of literature has shown that diaspora networks indeed influence the decision to migrate and affect migrants' success in the host country (Beaman, 2012; Pedersen *et al.*, 2008; Edin *et al.*, 2003). Throughout this literature, the size of the network has been identified as the main determinant. In this paper, we provide a different perspective on the role of diaspora networks by showing that the quality of these networks — measured by their degree of integration in the host society — has an equally important impact on the decisions and success of future migrants.

We argue that the integration of migrant networks in the host country determines both the decision to migrate and the outcomes after migration. Because existing networks differ in their degree of integration, some networks are able to provide more accurate information than others concerning job prospects. Well-integrated networks that have a great deal of interaction with the world surrounding them have better knowledge of local labor markets than enclaves, whose members typically have little social interaction outside the network. Potential migrants with access to a better-integrated network can base their decision on more accurate information, which in turn makes them more likely to make a correct decision: they migrate if they can expect to secure a job that makes them better off, whereas they stay if they can expect a job that makes them worse off.

To illustrate the underlying mechanism, we explore the link between information flows and the success of migrants in a simple two-period decision model. Initially, the migrant has some knowledge about her expected income abroad, albeit not enough to convince her that migration will be beneficial. She then receives information from the network and updates her beliefs about expected income from migration. To the extent that a more integrated network provides a more truthful signal and spreads less misinformation, a migrant who receives this information is more likely to make the right decision given her true income prospects in the receiving country.

We test this prediction using data on recent Mexican immigrants in the US. Mexican communities are spread out all across the US, allowing us to exploit a significant degree of variation in the characteristics of these communities. Communities in traditional destinations such as Los Angeles and Houston are typically more enclaved than those in newer destinations. Key to the empirical analysis is measuring both the quality of the network and the success of immigrants. For the quality of the network, we compute an assimilation index that measures the degree of similarity between Mexicans and Americans in an area with respect to a wide range of characteristics. As the social networks literature has shown, people with similar characteristics have more interaction, which leads to a more efficient aggregation of information (McPherson *et al.*, 2001; Acemoglu *et al.*, 2011), and ultimately to more accurate information on job prospects that can be passed on to future migrants. To measure the success of migrants, we take the difference between the wages of Mexicans in the US and Mexico. As the data do not allow us to observe Mexicans in both countries at the same time, we predict counterfactual wages in Mexico based on a large set of observable characteristics, and interpret a larger difference between income in the US and Mexico as a lower likelihood that the migrant has made a mistake in her decision to migrate.

Identification is threatened by the presence of unobserved factors that may induce a spurious relationship between the characteristics of the established network and the outcomes of newly arrived migrants. For example, a local industry may have attracted a lot of low-skilled migrants in the past, and does so until today, resulting in a low degree of assimilation of past immigrants, low wages of current immigrants, and overall a positive correlation between both variables. To address this endogeneity, we instrument for the assimilation of the network in 1990 with the settlement of Mexican migrants who arrived during the Bracero program between 1942 and 1964. The Bracero program was a guest worker program that mainly attracted low-skilled Mexicans who worked in agriculture and construction. Arriving initially as temporary migrants, these workers had little incentive to integrate in American society, casting a long shadow on the integration of Mexican communities today. Areas with a high share of Bracero immigrants have significantly less assimilated Mexican communities in the 1990s. At the same time, after controlling for network size and vintage, the settlement of workers in the 1950s should affect outcomes of newly arrived migrants in 2000 only through the characteristics of the network.

The results confirm the prediction of the model: migrants with access to better integrated networks are significantly more likely to be better-off in the US. An increase in the assimilation index by one standard deviation increases the monthly income difference between the US and Mexico by 90USD, or 16% of a standard deviation.

The previous literature has highlighted the importance of information in migration decisions. In particular, it has been shown that migrants generally may have incorrect beliefs about their prospective income abroad. McKenzie *et al.* (2013), for example, interviewed Tongan migrants before moving to New Zealand, and find that they significantly under-estimate their income in New Zealand. The discrepancy between the predicted and the realized income is mainly explained by the negative experiences of previous migrants. On the contrary, the work of Farré & Fasani (2013) shows that potential migrants can also over-estimate the gains from migration. They exploit exogenous variation in the availability of TV signals in Indonesia, and show that areas that receive more information about other regions of the country have lower emigration rates. However, not all information flows between migrant networks and their home country are equally accurate. Batista & Narciso (2013) stress the importance of the quality and frequency of information flows for the flow of remittances. They use a randomized control trial to increase the communication flows between immigrants and their networks abroad, showing that increased communication flows have a positive impact on the value of remittances, due to better control over remittance use and increased trust. Our paper contributes to the literature on information and migration by developing a straightforward theoretical link of the quality of information to the integration of migrant networks in the host society, and by testing how much the integration of the networks matters for migrant outcomes after migration.

By focusing on the quality of migrant networks, this paper provides a new perspective within the literature on network effects in international migration. Generally speaking, the literature defines a network as the number of previous migrants in a given destination and studies how existing networks affect the decisions and outcomes of future migrants. One strand of this literature documents that migration is path-dependent, with new migrants moving to places where they find an established community from their home countries (Pedersen *et al.*, 2008; Beine *et al.*, 2010). Growing migrant communities also affect the skill selection of subsequent migrants through a reduction in moving costs, and an increase available low-skilled jobs within the community (Carrington *et al.*, 1996; Winters *et al.*, 2001; Munshi, 2003; McKenzie & Rapoport, 2010; Beine *et al.*, 2011). In terms of outcomes after migration, larger migrant communities need not necessarily benefit newly arrived migrants. As shown by Beaman (2012), existing networks can provide information about jobs, but once the networks become larger, there is also an increased competition among the recipients of this information. Using data on resettled refugees in the US, she shows that a growing network hurts the current arrival cohort, but increases the employment and income prospects of future cohorts. Our paper introduces the quality of the network as an additional determinant of the success of newly arrived migrants. The social structure of migrant networks affects earnings on top of the scale effect found in previous papers.

Finally, the paper extends the literature concerning the impact of ethnic enclaves on the labor market outcomes of immigrants. Borjas (1995) shows that enclaves create human capital externalities that persist over generations. Children in ethnic enclaves grow up in a homogeneous, 'closed' environment, which often leads to a persistence in skill differentials compared to people outside the enclave. Nonetheless, enclaves can also have a positive impact on the earnings of newly arrived immigrants (Edin *et al.*, 2003) as well as the likelihood of finding employment in the destination (Andersson *et al.*, 2009). While these papers document the impact of networks on the outcomes of immigrants that have already emigrated, our paper shows that networks can even have an impact on migration decisions *before* emigration. Not only do migrant networks provide help in finding a job once a migrant has arrived, they also provide information to potential migrants in their home country, thereby affecting the beliefs of the potential migrant, and ultimately her success in the receiving country.

2 Migrant Networks, Information Flows, and Migrant Outcomes: Theory

2.1 The Integration of Networks and the Quality of Information

Our basic argument is that migrant communities that are more integrated in the society of their host country are able to give better information to future migrants. Members of a more integrated community have a better knowledge of the labor market and can give future migrants more accurate information about job prospects. This argument is consistent with the *strength-of-weak-ties* hypothesis (Granovetter, 1973, 2005), which states that in many situations, acquaintances – weak ties – are able to provide more important information than close family and friends – strong ties – because any two acquaintances have fewer social ties in common and receive information from a larger number of sources outside one's own network. In contrast, close friends and family are more likely to have the same contacts and information sources; thus, information easily becomes redundant.

Figure 1 illustrates two examples of migrant networks with different degrees of integration. The figure on the left describes an ethnic enclave, whose members, represented by the circles, have close connections within the network but very few connections to the outside world, represented by the crosses. An enclave is a typical example of a network with a high degree of closedness. This is a pervasive pattern in social networks, which the literature often refers to as inbreeding homophily — the fact that individuals with similar characteristics form close ties among one another (McPherson *et al.*, 2001; Currarini *et al.*, 2009). The graph on the right, in contrast, represents a well-integrated network whose members have weak connections among each other but strong connections to the outside world.

There are at least two reasons why a potential migrant would receive better information from a well-integrated network than from an enclave. First, the well-integrated network has more connections to the outside world. Its members receive more information and therefore have better knowledge about job perspectives in the receiving country. By contrast, members of an enclave typically have little knowledge of the language of the host country (Lazear, 1999; Bauer *et al.*, 2005; Beckhusen *et al.*, 2012), which makes interactions with natives difficult. While an enclave might offer job opportunities within the migrant community, it has very limited information on the labor market outside the enclave.

Second, members of the well-integrated network only have weak ties among one another; therefore, misinformation — false beliefs about the world outside the network — is unlikely to persist. The members of an enclave, on the other hand, interact mostly with other members of the enclave; thus, each member updates her beliefs based solely on interactions with other members. As shown in a series of theoretical papers , misinformation is more likely to persist in such closely connected communities (Acemoglu *et al.*, 2010; Golub & Jackson, 2010, 2012; Bikchandani *et al.*, 1992).

While the two network formations in Figure 1 represent polar cases that illustrate the differences between migrant networks, in reality, most networks will lie somewhere in between. In the theoretical analysis, we therefore introduce a parameter $\lambda \in (0, 1]$ that describes the ability of the network to aggregate accurate information.

2.2 Migrant networks, information, and migration decisions

To formalize the basic underlying mechanism, we consider the decision problem of a potential migrant who has imperfect information about his expected earnings abroad. His network, that is, people he knows in the destination, can reduce this uncertainty by providing him with more information about earnings abroad. We model the potential migrant's decision as a Bayesian decision problem with imperfect signaling, in which the migrant updates his prior beliefs after

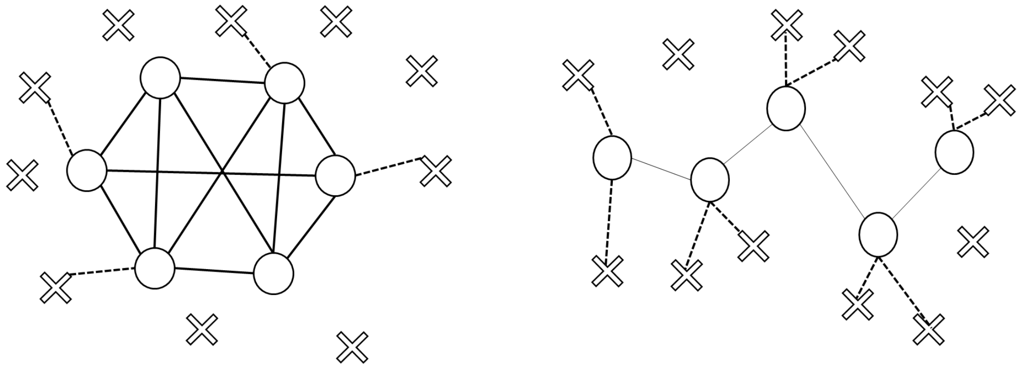

Figure 1: Ethnic enclave (left) and loosely connected network (right)

Note: These two panels depict models of migrant networks. The circles represent the migrant network; the crosses represent information sources outside the network. The network on the left is an ethnic enclave, with strong connections within the network but weak connections to the outside world. The network on the right is a loosely connected migrant network, with strong connections to the outside world and weak connections within the network.

receiving a signal from the network.

The network knows more about the labor market in the destination than the migrant himself, but does not have perfect knowledge. The quality of the network, described by $\lambda \in (0, 1]$, is larger the more integrated a network is in the society of the destination. If a potential migrant decides to move, he has to pay a sunk cost $M > 0$. We assume that a migrant is risk neutral, and maximizes expected income. He moves as soon as the expected wage differential between at home and abroad, w, is greater than the sunk cost. We view w as the realization of a random variable $\tilde{w}$.

The migrant has a prior about his expected earnings abroad, given by

$$\tilde{w} \sim N(\mu_0, \sigma_0^2). \tag{1}$$

We assume that $\mu_0 < M$, such that *a priori* migration is not beneficial. To get better information about expected earnings, the migrant receives a signal, θ, from the network, which has a conditional distribution

$$\theta | w \sim N\left(w, \frac{1-\lambda}{\lambda}\sigma^2\right). \tag{2}$$

If the network has perfect knowledge of the labor market, $\lambda = 1$, then the signal is perfect, whereas if the network knows nothing about job prospects, that is, if $\lambda \to 0$, then the signal is pure noise.

After receiving the signal, the migrant updates his beliefs. Applying Bayes' rule, the posterior distribution of $\tilde{w}$ is

$$\tilde{w} | \theta \sim N(\mu_1(\theta), \sigma_1^2), \tag{3}$$

where

$$\frac{1}{\sigma_1^2} = \frac{1}{\sigma_0^2} + \frac{1}{\sigma^2}\frac{\lambda}{1-\lambda}, \text{ and } \mu_1(\theta) = \sigma_1^2\left(\frac{\mu_0}{\sigma_0^2} + \frac{\theta}{\sigma^2}\frac{\lambda}{1-\lambda}\right).$$

The migrant moves if $\mu_1 > M$. A migrant makes an error in his migration decision if he migrates although it would have been beneficial to stay at home. This can be the case if he received a positive signal from his network, migrated based on the belief that he will be better off abroad, while he learned after moving that migration was not beneficial, i.e. $w < M$.

The probability of making an ex-post error in the migration decision can be expressed as a function of the signal, which is in turn a function of the network quality λ,

$$\alpha(\theta) = P(\tilde{w} < M|\theta) = \Phi\left(\frac{M - \mu_1(\theta)}{\sigma_1}\right). \tag{4}$$

Figure 2 provides a numerical example that illustrates the negative relationship between the network quality and the probability of making an error in the migration decision.[1]

3 Data and Measurement

The theory predicts a reduced-form relationship between the integration of migrant networks and the likelihood that migrants make a mistake in their decision to migrate. The more integrated the network is in the host country, the more likely it is that a migrant has ex post a higher income than in the home country, and the less likely it is that he made an error in his decision to migrate.

To test this relationship empirically, we use data on Mexican immigrants in the US. Focusing on Mexicans allows us to exploit a significant degree of variation in the characteristics of Mexican communities across the entire country. Mexicans have had a long tradition of emigration to the US, leading to well-established Mexican communities in many US cities. Nonetheless, the settlement pattern changed in the 1990s. While until the 1980s most Mexicans went to California, Texas, and Chicago, many Mexicans in the 1990s settled in areas that had no significant pre-existing Mexican community, such as Atlanta, Denver, Raleigh-Durham, Seattle, or Washington, D.C. (Card & Lewis, 2007). This gradual diffusion of Mexicans across the US led to a great deal of heterogeneity across Mexican communities, both in terms of size and integration. Another advantage of looking at one nationality is that it reduces unobserved heterogeneity because the network characteristics and the success of migrants differ less within a nationality than between different nationalities.

3.1 Main dataset

The core dataset is the 2000 US census, supplemented with information from the 1990 US census and the 2000 Mexican census. We use the 5%-sample of the US census, and the 10%-sample

[1] This error is analogous to a type-I-error. The potential migrant tests the hypothesis that his income is higher in the US than at home, based on the observation of the signal.

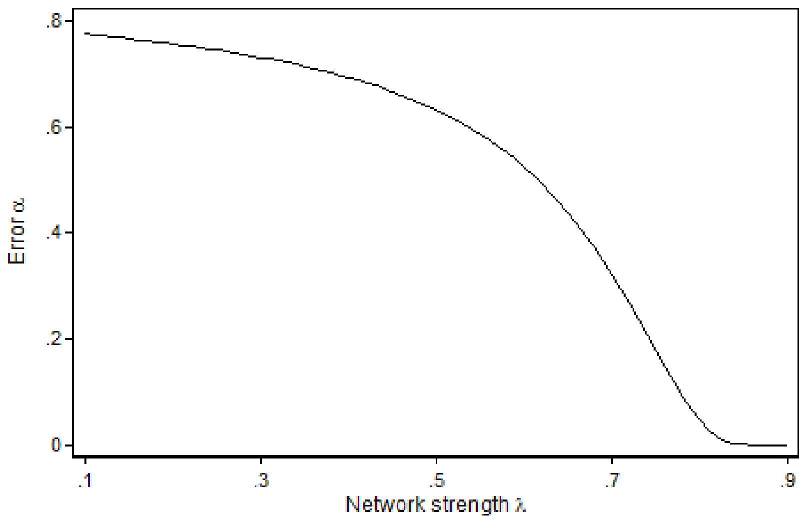

Figure 2: Network quality and the likelihood to take a wrong decision

Note: Numerical example based on $M = 0.8$, $\sigma_0 = \sigma = 1$, $\theta = 0.5$.

of the Mexican census provided by IPUMS.[2] The US census is representative at the individual and household level, and includes both legal and illegal migrants, but without containing an identifier for illegal migrant status. Moreover, the census only includes people who stay in the US long-term; it does not include people who are on a tourist visa or any other short-term visitors (Hanson, 2006).

Our sample consists of Mexican immigrant men who arrived in the US between 1995 and 2000. We define immigrants as Mexican citizens who were born in Mexico and report in the census that they were residing in Mexico 5 years ago. The sample is restricted to Mexicans aged 18-64 who were at least 18 years old when they moved to the US and who moved to a district with at least 20 Mexicans.[3] An outline of further restrictions to the sample can be found in Appendix B.2.

The restriction of the sample to recent migrants is the result of a trade-off between having a measure of lifetime success on the one hand, and accurate information on the network, as well as a less selective sample on the other. The gold standard for measuring the success of migrants would be to compare their lifetime earnings in the US with counterfactual lifetime earnings in Mexico. Unfortunately, detailed data on the entire earnings history of migrants is not available. If we used information from a single census round on migrants who have been in the US for a long time, we would not be able to reconstruct a migrant's network at the time of arrival. Moreover, as shown by Biavaschi (2012) and Campos-Vazquez & Lara (2012), selective out-migration of

[2] Ipums: Steven Ruggles, J. Trent Alexander, Katie Genadek, Ronald Goeken, Matthew B. Schroeder, and Matthew Sobek. Integrated Public Use Microdata Series: Version 5.0 [Machine-readable database]. Minneapolis: University of Minnesota, 2010.

[3] As districts, we use consistent PUMAs (public use microdata area). A cutoff of 20 is necessary for our measure of integration. As this measure is based on a probit model at the CONSPUMA-level, a minimum number of observations is required for convergence.

more successful migrants would lead to an under-estimation of the success of migrants. With the focus on recent migrants, we can only measure their short-term success, although this enables us to obtain a more precise measure of their network, and base the estimation on a less selective sample.

For our analysis, the census offers two advantages. First, it is the only dataset that is large enough to cover Mexican communities across the whole of the US, allowing us to exploit a large degree of variation in terms of network quality, size, and vintage across the US *within* one nationality. Second, the census contains rich information on individual and household characteristics, such as the age at the time of immigration, birth place, current employment, education and family situation.

Besides these advantages, the US census has two limitations: it has no direct information on the network of the migrant or the information flows between the network and the migrant prior to migration.[4] A further limitation is that it contains no information on wages prior to migration, which would be helpful to compare the migrants' situation in Mexico and the US. Recently available longitudinal datasets, such as ENET or the Mexican Family and Life Survey, contain this information, but they have limited information on outcomes after migration.[5]

A further concern with data on Mexicans in the US is the undercounting of illegal migrants. The majority of Mexicans in the United States arrive as illegal immigrants and only receive their residence permit at a later stage (Massey & Malone, 2002; Hanson, 2006). While the census does not ask respondents about their legal status, some illegal migrants might fear negative consequences and choose not to take part in the survey or might not be available for some other reason. The undercounting of illegal migrants can lead to selection bias if the least-skilled migrants are more likely to be excluded. While we are aware that undercounting might bias the results, it is important to note that the extent of undercounting has decreased significantly over the last census rounds: from a 40% undercount rate in 1980 (Borjas *et al.*, 1991) and 15-20% in the 1990s (Bean *et al.*, 2001; Costanzo *et al.*, 2002) to around 10% in the 2000 survey (Card & Lewis, 2007). Moreover, Chiquiar & Hanson (2005) show that undercounting only causes minor changes to the wage distribution of Mexicans in the US, which means that there is no systematic undercount of a particular skill level.

3.2 Measuring the Success of Migrants

Next, we turn to the construction of the dependent variable. To be in line with the theory, we require a measure for an error in the migration decision — that is, a variable that indicates whether a Mexican in the US would have been better off staying in Mexico rather than incurring the fixed moving cost and earning an income in the US. The error in a migration decision could then be measured by a binary variable that takes value one if the earnings in Mexico are larger

[4] While other datasets, such as the Mexican Migration Project, contain some information on the assistance of friends and family members in the migration decision, they do not contain information on the broader network that goes beyond family and friends, and they have limited variation in networks across destinations in the US.

[5] See Appendix A for a discussion of other datasets on Mexicans in the US.

than the earnings in the US minus moving costs. Given that we cannot observe moving costs, it is difficult to construct this measure without introducing a great deal of measurement error.

To proxy for the success or failure of migrants, we use the difference between wages in the US and Mexico. The larger the value of this difference, the higher the wage in the US relative to Mexico, and the less likely it is that an immigrant has made an error in her decision to migrate. We calculate the wage difference as the difference between the actual monthly wage in the US, and a counterfactual wage of workers in Mexico with the same observable characteristics. As Mexicans in the US and Mexico might differ with respect to the number of working hours, we adjust wages by the number of working hours in a typical work week and the number of weeks worked in a typical year. In addition, we convert Mexican wages into US dollars and account for differences in price levels using a PPP factor.[6] Initially, we only include workers with a positive income in the wage regressions. In Appendix C.1, we test the robustness of the wage predictions using a two-step selection model on the full sample.

3.2.1 Counterfactual wages

To predict the counterfactual wages, we first use the 2000 Mexican census to regress monthly wages on a vector of personal characteristics

$$\text{wage} = \boldsymbol{X_{MEX}}\boldsymbol{\beta_{MEX}} + \varepsilon, \tag{5}$$

from which we obtain an estimate for skill prices in Mexico, $\hat{\boldsymbol{\beta}}_{\boldsymbol{MEX}}$. $\boldsymbol{X_{MEX}}$ includes a set of education dummies, a dummy for marital status, age, and age squared, as well as interactions of the education dummies with the dummy for marital status, age, and age squared. ε is an i.i.d error term that captures unobservable determinants of wages. The interaction terms allow us to have a separate age-earnings gradient for each education level.

Using the same characteristics for Mexicans in the US, $\boldsymbol{X_{US}}$, we then predict the counterfactual wages as

$$\widehat{\text{wage}} = \boldsymbol{X_{US}}\hat{\boldsymbol{\beta}}_{\boldsymbol{MEX}}. \tag{6}$$

To make both wages comparable, we convert the counterfactual wages into US dollars and adjust for differences in price levels using PPP data from the Penn World Tables.[7]

The difference between the actual and the counterfactual wage yields the gains from emigration. Figure 3 shows the distribution of the gains for Mexicans with a positive wage income in the US. As can be seen, most Mexican workers in the US are financially better off than in Mexico. The average Mexican in 2000, conditional on working, earns around 700 USD per month more in the US. Around 5% of the distribution would be better off in Mexico, while around 25% have a wage difference of less than 500 USD per month.

[6] See Appendix B for a description of the samples and the wage adjustment.
[7] The PPP conversion implicitly assumes that migrants consume their entire income in the US.

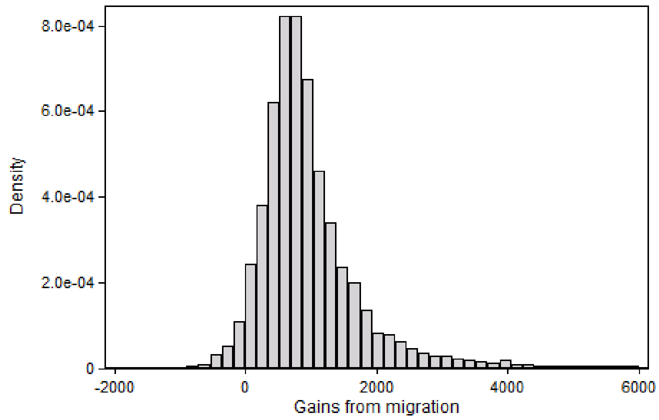

Figure 3: Gains from emigration

Note: The graph shows the distribution of the losses from emigration in 2000, which is measured as the difference between the actual and counterfactual monthly income. The graphs only include workers with a positive income in the US.

3.2.2 Counterfactual wages and self-selection

The prediction of counterfactual wages in Equation (6) assigns to every Mexican in the US the average wage of a worker in Mexico with the same observable characteristics. But this measure could be biased if migrants and non-migrants differ with respect to unobservable characteristics, which is very likely given that education, age, gender and marital status only represent some of the factors that determine wages. Unobserved factors such as IQ, confidence, motivation or self-selection into a certain type of firm or industry potentially have a large impact on wages and can explain wage differentials between workers with identical observable characteristics.

The literature provides ample evidence that emigrants from Mexico are not a random sample of the entire Mexican population. While the earlier literature based its analysis on observable characteristics and found that Mexican emigrants were mainly selected from the center of the Mexican income distribution (Chiquiar & Hanson, 2005; Orrenius & Zavodny, 2005), more recent studies have shown that Mexican emigrants are negatively selected on unobservable characteristics. Using longitudinal data that tracks Mexican workers across the border, Ibarraran & Lubotsky (2007), Fernández-Huertas Moraga (2011), and Ambrosini & Peri (2012), find that pre-migration earnings were on average lower for emigrants than for stayers.

Due to negative self-selection, the counterfactual wages are upward-biased because we assign to every Mexican emigrant a higher income than he would actually have. Consequently, the dependent variable — the difference between the US wage and the counterfactual wage — is downward-biased. While our cross-sectional data does not allow us to directly analyze the magnitude of the selection bias, we can get an idea of its importance by using different samples to predict the counterfactual wages. If we cannot directly observe counterfactual wages, the second best way is to predict them based on Mexicans who are as similar as possible to Mexicans in the US. Two obvious candidates are internal migrants and return migrants, because both groups are by definition more mobile than never-migrants, and should be more comparable to migrants to the US. To be sure, the different migration decisions — migration to the US, return to Mexico, migration within Mexico — may be driven by different selection patterns (Borjas, 1987; Borjas & Bratsberg, 1996; Bartolucci *et al.*, 2013). However, as we show in Appendix C.1, the predicted Mexican wages are similar regardless of the method, suggesting that selection bias is negligible.

The wage difference between Mexico and the US measures the success of migrants based on their economic situation in the first five years after migration. While we believe that it represents a suitable measure, it should be noted that wage differences might not be the only indicator for the success of migrants, with local amenities, available housing and other location-specific factors possibly contributing to the utility of a destination. If migrants maximize utility rather than income in their location choice, then we should not be surprised if a considerable share has wage differentials close to zero. While non-monetary factors might play a role in location choice, recent literature has shown that a model of income maximization can explain most of the variation in location choices of both internal and international migrants (Kennan & Walker, 2011; Grogger & Hanson, 2011).

3.3 Measuring the Integration of Migrant Communities

A further key ingredient to the empirical analysis is the integration of migrant communities. As outlined in Section 2.1, there are good reasons to believe that better integrated networks have better knowledge about the labor markets in a given area because they have more interaction with the world outside the network. Thus, incorrect beliefs would not easily spread in such a community. As it is most likely that migrants received some information from the network they eventually moved to, we measure the network variable for each migrant using characteristics of Mexicans that already lived in the same local area in the US.

The question is how to measure whether a migrant community is well-integrated in the area. The literature on social networks suggests statistics that measure the degree of homophily — the likelihood that a person only interacts with people of the same group (McPherson *et al.*, 2001). An enclave would have a high degree of homophily, as its members interact mostly with each other but not with people outside the enclave. A direct measure of homophily requires very detailed data on the connections within a community.

Given that we cannot measure direct links between members of Mexican communities, we proxy the network quality with an assimilation index that measures the similarity between Mexicans and Americans in a given area with respect to a large set of observable characteristics. If Mexicans and Americans are similar with respect to variables such as age, education, fertility, occupation, and home ownership, they most likely have more interaction with Americans, and hence the network is well-integrated and has access to more accurate knowledge about the labor market. On the contrary, if Mexicans and Americans in an area are very different in their characteristics, there is probably little interaction between the two groups.

We calculate the assimilation index at the smallest geographic unit available in the US census, the *consistent PUMA* (*conspuma*). PUMAs (Public Use Microdata Area) are small geographic units in the US census, with a population between 100,000 and 200,000 people. They do not cross state borders and their boundaries are re-drawn with every census so that the size of each PUMA never exceeds 200,000 people. To make PUMAs comparable over time, the US Census Bureau has introduced *consistent* PUMAs that have the same boundaries from 1980 to 2010 and are larger than the original PUMAs.[8] As we want to calculate the assimilation index of the communities before the most recent migrants arrived, we use conspumas. To every migrant who moved to a given consistent PUMA between 1995 and 2000, we assign the assimilation index of Mexicans that lived in the same area in 1990.

Following Vigdor (2008), we calculate the assimilation index as a statistical measure of similarity between Mexicans and Americans in an area. The assimilation index is low if we randomly draw people from a given area, and their observable characteristics clearly identify them as Mexicans or Americans. On the contrary, if we cannot tell both groups apart based on observable characteristics, Mexicans and Americans are very similar, which is reflected in a high assimilation index.

We proceed in three steps. First, we use all Mexicans and Americans in the sample and run in each local area a separate probit regression of a binary variable (1 if Mexican, 0 if US citizen) on a large set of observable characteristics. We then restrict the sample to all Mexicans in the area, and use the probit estimates to predict the probability of being Mexican based on their observable characteristics. A failure to predict that someone with given characteristics is Mexican means that the person is very similar to US natives in the same area. Finally, we use the predicted probabilities of all Mexicans in an area to compute the assimilation index.

We first run the following probit regression:

$$P(\text{Mexican } |\boldsymbol{X}) = \Phi(Z) = \Phi(\boldsymbol{X\beta}), \tag{7}$$

where $\Phi()$ is the cumulative density function (CDF) of the standard normal distribution. $\boldsymbol{X}$ contains the following variables: marital status, gender, education (4 categories, see Appendix B.1), employment status, number of children, age, and home ownership. We also include the median income of the person's occupation in 1990 (variable ERSCOR90) to see whether migrants work in similar occupations compared to Americans. The sample for the calculation of the assimilation index is more restrictive than the sample used in the regressions in the next section. It consists of all Mexicans between 25 and 64 years who live in a metropolitan area with at least 20 Mexicans. To increase statistical power, we estimate Equation 7 at the level of metropolitan area, and use the estimates to compute a separate assimilation index for each conspuma.

We then restrict the sample to Mexicans only, and pretend for the moment that we do not know if a person is Mexican or American. Using the estimated coefficients $\hat{\beta}$, we predict for every person i in the sample the probability that the person actually is a Mexican.

$$\hat{p_i} = \Phi(\hat{Z}) = \Phi(\boldsymbol{X\hat{\beta}}), \tag{8}$$

[8] The size of CONSPUMAs ranges between 100,000 and 4.3 million inhabitants.

where Φ is the cumulative distribution function of the joint normal distribution. The higher this probability, the more different is the person from the US citizens living around her. If the observable characteristics perfectly predict that a person is Mexican, then this implies that the person has a low degree of assimilation in her local area, whereas if the person was highly assimilated, we would not be able to statistically distinguish her from an American.

To obtain the assimilation index for an entire Mexican community in a conspuma, we take the average predicted probability for each PUMA, $\widehat{p_m}$, and calculate the estimate of the assimilation index as

$$\widehat{\text{index}_m} = 100(1 - \widehat{p_m}). \tag{9}$$

Figure 4 shows the distribution of the assimilation index in 1990. The density was calculated based on PUMA-level data weighted by the number of Mexicans in a conspuma such that each bar reflects the number of Mexicans living in an area with a given assimilation index. As the figure shows, there is considerable heterogeneity in the degree of assimilation across conspumas. The largest number of Mexicans live in areas with an assimilation index between 40 and 80. Networks with assimilation indices above 80 are mostly small, although there are also a number of smaller networks that have an assimilation index lower than 80.

3.4 Descriptive Statistics

Table 1 displays the descriptive statistics for the US census in 2000. Panel A shows the aggregate statistics at the conspuma-level, while panel B shows the individual-level statistics of the sample. In the regressions to follow, we will use both aggregate and individual data.

The aggregate variables in panel A are computed conditional on at least one Mexican living there. The distribution of Mexicans across the US is heavily skewed, with a large number of small communities and a small number of large communities. The median share of Mexicans in a PUMA is 0.9% and the median number is 1,700, while the largest number of Mexicans in a conspuma is more than 500,000 (a conspuma within Los Angeles). The area with the largest concentration has 35% Mexicans (McAllen-Edinburg-Pharr-Mission, TX).

Panel B displays the characteristics of immigrants who recently arrived in the US. Most immigrants come to the US in their early 20s. The vast majority has a lower secondary education or less, while there are very few Mexicans immigrants with a college education. The median Mexican moved to a community with an assimilation index of 70. For most immigrants, migration pays off, with Mexicans in the US earning on average around 700 USD more than they would earn in Mexico — although there is a large degree of heterogeneity in the income difference.

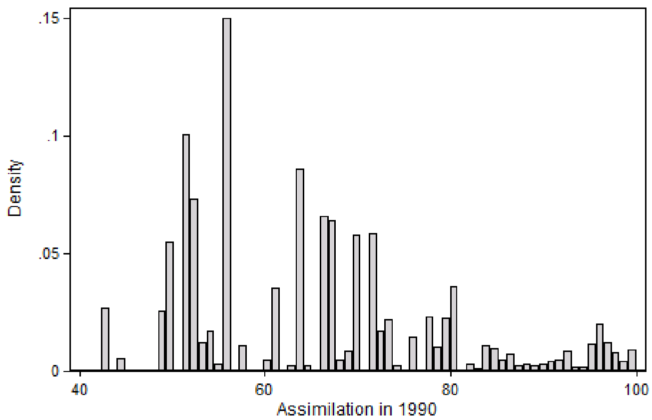

Figure 4: Assimilation index in 1990

Note: The graph shows the distribution of the assimilation index in 1990. It is based on conspuma-level data, weighted by the number of Mexicans per PUMA

Table 1: Summary Statistics of the main variables

Variable	*N*	*Mean*	*SD*	*Min*	*Max*
A. Aggregate data					
Income difference US-Mexico	202	764.15	452.93	-84	3174
Assimilation in 1990	202	84.78	16.00	42	100
Share of Braceros (in %)	202	0.22	0.45	0	4
Share of Mexicans (in %)	202	4.50	6.57	0	35
Nr of Mexicans (in 1000)	202	20.95	56.39	0	555
Mean wage of US natives (monthly)	202	2493.44	504.30	1477	3906
B. Individual-level data					
Income difference US-Mexico	20131	695.09	1148.72	-1425	13558
Age	20131	28.62	8.72	18	64
Age at immigration	20131	26.66	8.70	18	64
Married	20131	0.48	0.50	0	1
High-school dropout	20131	0.14	0.35	0	1
Lower secondary school	20131	0.49	0.50	0	1
Upper secondary school	20131	0.33	0.47	0	1
College Degree	20131	0.04	0.21	0	1

Notes: Aggregate statistics are computed at the conspuma-level, conditional on at least one Mexican living in the area. The share of Braceros is the share of Mexicans in the population of a conspuma that immigrated between 1942 and 1964, during the time of the Bracero guest worker program. Individual-level data as well as the income difference in Panel A is based on men only.

4 Identification and Estimation Strategy

To estimate the effect of network quality on the success of migrants, we fit the following regression:

$$y_{ij}^{2000} = \alpha + \beta \text{ assim}_j^{1990} + \boldsymbol{R_j'}\boldsymbol{\gamma} + \boldsymbol{X_{ij}'}\boldsymbol{\delta} + \varepsilon_{ij}, \tag{10}$$

in which y_{ij} is the difference between wages in the US and Mexico of newly arrived migrants in the 2000 census. Index i refers to a Mexican immigrant who lives in conspuma j. The regressor of interest is the assimilation index for all Mexicans that lived in conspuma j in 1990, $assim_j^{1990}$. Given the differences in the characteristics of conspumas with respect to economic performance and the size of the existing Mexican community, we control for a vector of conspuma characteristics, $\boldsymbol{R_j}$, which includes the average income of US natives, as well as a quartic in the number of Mexicans that have lived in the conspuma in 1995. Moreover, the characteristics of Mexicans may differ across conspumas. To make Mexican immigrants comparable across the US, we control for a vector of observable characteristics, $\boldsymbol{X_{ij}}$, which includes dummies for four education levels (high school dropouts, high school degree, some college, completed college), a dummy for being married, and a quadratic in age. Finally, ε_{ij} is an i.i.d. error term that captures all other factors that determine the wage difference but are not controlled for in the regression. Given that we regress an individual variable on a group variable, we cluster the standard errors at the conspuma level.

To estimate the causal effect of network quality on the success of migrants, one would ideally want to randomly assign new immigrants to different types of networks and observe the differences in the outcome of interest after they have migrated. Given that such an experiment is not available for Mexicans in the US, an alternative approach would be to find exogenous variation in the quality of networks that is unrelated to other factors that might affect the outcome of interest. In the absence of a clean quasi-experiment — for example, a change in migration policies — we rely on instrumental variables that affect the assimilation of local Mexican communities but have no direct effect on the success of migrants.

The assimilation index is potentially endogenous in this regression, in which case the estimate for β could not be interpreted as a causal effect. Endogeneity could arise because migrants self-select into areas based on local amenities, such as existing migrant networks, employment opportunities or public services. This concern is particularly important in our estimation of Equation (10) which regresses the success of current migrants on the assimilation of previous migrants, which in turn could be seen as a proxy for the success of previous migrants. If current migrants with a higher earnings potential move to areas with a high degree of assimilation, we would observe a positive correlation, which would be spurious and purely due to the self-selection of immigrants into areas. The control variables in Equation (10), which include a person's education level and age, only capture the observable component of a person's earnings potential, whereas the selection can also be based on unobservable characteristics such as motivation, language skills, or the ability to adapt to a new environment.

To address the endogeneity problem, we use the settlement of Mexicans in the US during the Bracero program as an instrumental variable. The Bracero Program was a temporary migration program that allowed Mexicans to take up temporary agricultural work in the US. Over the duration of the program, from 1942 to 1964, around 4.5 million Mexican workers came to the

US to work in agriculture, and for the railroad system. As shown by Massey & Liang (1989), many of these workers took repeated trips to the US before eventually settling there. Most Bracero workers were low-skilled, and the temporary nature of the program gave them little incentive to integrate into US society after arrival. The low degree of integration of the Braceros seemingly created more closed-up Mexican groups, resulting in a low degree of assimilation of Mexicans in 1990. Figure 5 displays the first-stage relationship between the share of Braceros and the assimilation in 1990, controlling for conspuma and individual characteristics. Clearly, a higher share of Bracero immigrants predicts a low level of assimilation.

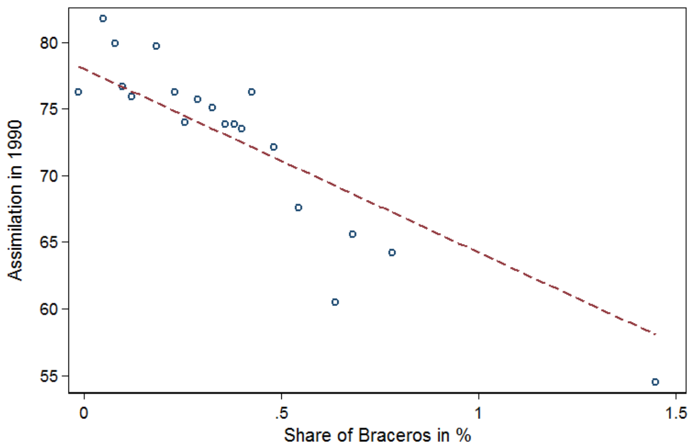

Figure 5: First stage regression

Note: This graph displays a bin scatter of the first-stage relationship between the share of Bracero immigrants in a conspuma, and the assimilation index in 1990. The dashed line shows the coefficient of the first stage regression of the assimilation index on the share of Braceros, individual control variables, as well as controls for average US wages, and a quartic in the number of Mexicans in a conspuma.

The identifying assumption behind this instrument is that the share of Bracero immigrants affects the success of current migrants only through the assimilation of the network. We believe that this assumption holds because the share of Braceros is very small compared to the total population in a conspuma. As shown in Table 1, the share of Braceros in the total population of a conspuma is 0.22%. Therefore, it is unlikely that the average share of Braceros had an effect on the broader economy of an area, and that this effect would be noticeable almost 40 years later.

One potential violation of the exclusion restriction, however, is the impact of Braceros on the size of the network, which in turn may affect both the assimilation of a community and the wages of recent immigrants. As shown by Beaman (2012), the size of the network directly affects the performance of immigrants, positively through a higher number of jobs within the network, and negatively through greater competition for these jobs. In order to control for this potential transmission channel, we include a polynomial of the network size in the regression.

5 Results

5.1 Results at the aggregate level

We first explore the relationship between network quality and the success of migrants at the conspuma-level. Table 2 displays the results for the following estimating equation

$$y_j^{2000} = \alpha + \beta \text{ assim}_j^{1990} + \boldsymbol{R_j'}\gamma + \varepsilon_j, \tag{11}$$

where $\boldsymbol{R_j}$ includes the average wage of US natives in all specifications, and a quartic in the number of Mexicans in some. Conventional standard errors are displayed in parentheses. Because the assimilation index has been estimated and therefore contains estimation error, we also report bootstrapped standard errors with 500 replications in brackets for unweighted regressions.

Table 2: Regression results at the conspuma-level

Dependent variable: wage difference USA - Mexico

	OLS (1)	OLS (2)	OLS (3)	IV (4)	IV (5)	IV (6)
Assimilation index	4.91**	5.46***	4.51	4.90	4.94*	3.98
	(2.05)	(0.85)	(2.88)	(3.43)	(2.54)	(9.80)
	[1.33]		[1.67]	[1.87]		[5.74]
Weighted by size	No	Yes	No	No	Yes	No
Control: nr of Mexicans	No	No	Yes	No	No	Yes
First stage:						
Share Braceros				-21.14***	-12.74***	-9.52***
				(2.04)	(3.52)	(2.26)
F-Statistic				107.55	13.10	17.79
N	202	202	202	202	202	202

Note: The table presents the results of OLS- and IV-regressions of difference in monthly wages on the assimilation index. Unit of observation is conspuma. All regressions include a control for the average wage of US natives. In Columns (2) and (5) the regressions are weighted by the number of Mexicans in a conspuma. Columns (3) and (6) include controls for a quartic in the size of the Mexican community. Conventional standard errors are displayed in parentheses, bootstrapped standard errors with 500 replications in brackets. Bootstrapped standard errors are not reported for weighted regressions. Significance stars refer to the conventional standard errors: * $p<0.1$, ** $p<0.05$, *** $p<0.01$.

Column (1) displays the plain OLS estimate for β in Equation (11). The partial correlation is positive and statistically significant at the 5%-level.[9]. An increase in the assimilation index by one point increases the monthly wage difference between US and Mexican wages by around 5USD. This may not sound like a large effect; but increasing the assimilation index by one

[9] Significance levels refer to the conventional standard errors

standard deviation, increases the wage difference by 80USD per month, or 16% of a standard deviation in the wage difference.

While in Column (1) every conspuma received equal weight, regardless of the size of the Mexican community, in Column (2) we weight the regression with the number of Mexicans in a conspuma, giving higher weight to areas with a larger Mexican community. The estimate in this specification is larger and more precise, indicating that the effect is more pronounced in larger Mexican communities. In Column (3) we directly control for a quartic in the number of Mexicans. In this case, the point estimate is of a similar magnitude as before, but less precise. In sum, accounting for size, either through weighting or through controls, does not change the estimates dramatically.

In Columns 4-6, we estimate the same specifications as in Columns 1-3, but instrument for the assimilation index with the share of Braceros in a conspuma. As shown in Figure 5, the first stage relationship is negative and strong enough to rule out a weak instrument problem. Compared to the OLS results, the point estimates are less precise, but are of a similar magnitude. In Column (6), when we control for the size of the network, is the effect around 20% smaller, but the difference is statistically insignificant.

.

5.2 Individual-level results

The aggregate results confirm our hypothesis that a more integrated network leads to better outcomes for migrants. We now turn to the estimation of Equation (10) with individual-level data. This enables us to control for more observable characteristics, and gives greater weight to areas with a large number of recent immigrants

Table 3 displays the estimates for β based on individual-level regressions as outlined in Equation 10. Columns (1)-(3) present the results without controls for network size. All regressions include individual-level controls, as well as a control for the average wage of US natives in a conspuma. As before, we report both conventional and bootstrapped standard errors, which are both clustered at the conspuma-level.

The result from an OLS regression in Column (1) is similar in magnitude to the estimates at the aggregate level. An increase in the assimilation index by one point is associated with a 4.4USD increase in the monthly wage difference. In Column (2), we instrument the assimilation index with the share of Braceros. Again, the first stage is negative and sufficiently strong, with an F-statistic of 19.7. The point estimate is slightly larger than in the OLS coefficient.

One problem with the census data is that around one quarter of the sample have zero wages in the US. So far, we have taken a wage of zero at face value, but we cannot be sure whether the person actually earns zero, or whether his wage was coded as zero and is actually unknown. To assess whether the estimates are affected by zero wages, we re-estimate the IV-estimation, dropping all observations with zero income. As shown in Column (3), the zero wages do not significantly affect the results.

In Columns (4)-(6), we estimate the same specifications, but in addition control for a quartic

Table 3: Regression results at the individual level

Dependent variable: wage difference USA - Mexico						
	OLS	IV	IV	OLS	IV	IV
	(1)	(2)	(3)	(4)	(5)	(6)
Assimilation index	4.43***	5.01***	4.72***	5.02***	6.65***	6.52***
	(0.73)	(1.00)	(0.92)	(0.96)	(2.03)	(1.89)
	[0.74]	[1.07]	[0.92]	[1.12]	[2.48]	[2.19]
First stage:						
Share Braceros		-22.56***	-22.62***		-13.85***	-13.80***
		(5.08)	(5.33)		(3.89)	(4.05)
Control: nr of Mexicans	No	No	No	Yes	Yes	Yes
Exclude if wage US= 0	No	No	Yes	No	No	Yes
F-Statistic		19.71	18.04		12.67	11.59
N	20131	20131	15082	20131	20131	15082

Note: This table displays OLS- and IV-regression results at the individual level. All regressions include individual-level controls, as well as a control for the average wage of US natives. Columns (4)-(6) also control for a quartic in the number of Mexicans. Standard errors, clustered at the conspuma-level, are displayed in parentheses. Bootstrapped standard errors, clustered at the conspuma-level, with 500 replications, are displayed in brackets. Significance stars refer to the clustered standard errors: * p<0.1, ** p<0.05, *** p<0.01.

in the number of Mexicans in a conspuma. The first stage now becomes weaker, but the F-statistic is still above the conventional level of 10 that would indicate a weak instrument problem. The estimated effects are now larger. A one-point increase in the assimilation index increases the wage difference by 6.5USD. Put differently, an increase in the assimilation index by one standard deviation increases the wage difference by 20% of a standard deviation.

In sum, these results show that the quality of pre-existing networks has a significant impact on the success of migrants. Moving from the 25th percentile of the assimilation index to the 75th percentile, or going from Waco, TX, to Amarillo, TX, results in an increase in the gains from migration by 143 USD per month.

6 Conclusion

Migrant communities around the world differ not only in their size but also in their degree of integration in the host society. In this paper, we study how the integration of existing migrant communities affects the migration decisions and economic outcomes of future migrants. Following the literature on social networks, we argue that more integrated networks have a better knowledge of the labor market in that destination and therefore give more accurate information to future migrants about job opportunities. We first explore this mechanism in a decision model with imperfect signalling, which predicts that migrants who receive information from better-integrated networks make fewer errors in their migration decisions, and they migrate earlier.

Using data on recent Mexican immigrants in the US, we test these predictions empirically. The focus on Mexico allows us to exploit a significant variation in the size and social structure of migrant communities across the United States. We measure the two variables of interest — the likelihood of making an error and the quality of the migrant network — using the wage difference between the US and Mexico and an assimilation index that measures the similarity of Mexicans and Americans in an area with respect to a large number of observable characteristics. To overcome omitted variable bias, we instrument the assimilation index with past changes in the diffusion of Mexicans across the US and with past settlement patterns of low-skilled Mexicans who came to the US during the Bracero program. Our results confirm our hypothesis, namely that migrants with access to a better-integrated network had a significantly larger wage differential between the US and Mexico and, hence, were less likely to make an error in their migration decision.

With its focus on the quality of networks, this paper offers a new perspective on the role of networks in international migration. While the previous literature has proxied the strength of migrant networks through their size, we show, both theoretically and empirically, that the quality of networks has a sizable impact on the economic outcomes of migrants.

In addition, the theoretical model and empirical findings offer new insights for the study of social networks in general. Most of the empirical literature focuses on the impact of the architecture of social networks on individual members of the network. Our paper shows that the social structure of networks also affects people outside the network — in our case, potential migrants who still live in the country of origin — through the network's ability to aggregate information. If more integrated communities have better knowledge and are able to provide more accurate information, this benefits the recipients of the information.

References

Acemoglu, Daron, Ozdaglar, Asuman, & Gheibi, Ali Parandeh. 2010. Spread of (Mis)Information in Social Networks. *Games and Economic Behavior*, **70**, 194–227.

Acemoglu, Daron, Dahleh, Munther A., Lobel, Ilan, & Ozdaglar, Asuman. 2011. Bayesian Learning in Social Networks. *Review of Economic Studies*, **78**(4).

Ambrosini, J. William, & Peri, Giovanni. 2012. The Determinants and the Selection of Mexico-US Migrants. *The World Economy*, **35**(2), 111–151.

Amuedo-Dorantes, Catalina, & Mundra, Kusum. 2007. Social Networks and their Impact on the Earnings of Mexican Migrants. *Demography*, **44**(4), 849–863.

Andersson, Fredrik, Burgess, Simon, & Lane, Julia. 2009. Do as the Neighbors Do: The Impact of Social Networks on Immigrant Employment. *IZA Discussion Paper*, **4423**.

Bartolucci, Cristian, Villosio, Claudia, & Wagner, Mathis. 2013. Who Migrates and Why? *Collegio Carlo Alberto, mimeo.*

Batista, Catia, & Narciso, Gaia. 2013. Migration and Information Flows: Evidence from a Field Experiment among Immigrants in Ireland. *CReAM Discussion Paper Series, Centre for Research and Analysis of Migration (CReAM), Department of Economics, University College London.*, **1331**.

Bauer, Thomas, Epstein, Gil S., & Gang, Ira N. 2005. Enclaves, Language, and the Location Choice of Migrants. *Journal of Population Economics*, **18**, 649–662.

Bauer, Thomas, Epstein, Gil S., & Gang, Ira N. 2007. Herd Effects or Migration Networks? The Location Choice of Mexican Immigrants in the U.S. *Research in Labor Economics*, **26**, 199–229.

Beaman, Lori. 2012. Social Networks and the Dynamics of Labor Market Outcomes: Evidence from Refugees Resettled in the U.S. *Review of Economic Studies*, **79**(1), 128–161.

Bean, Frank D., Corona, Rodolfo, Tuiran, Rodolfo, Woodrow-Lafield, Karen A., & van Hook, Jennifer. 2001. Circular, Invisible, and Ambiguous Migrants: Components of Difference in Estimates of the Number of Unauthorized Mexican Migrants in the United States. *Demography*, **38**(3), 411–422.

Beckhusen, Julia, Florax, Raymond J.G.M., de Graaff, Thomas, Poot, Jacques, & Waldorf, Brigitte. 2012. Living and Working in Ethnic Enclaves: Language Proficiency of Immigrants in U.S. Metropolitan Areas. *IZA Discussion Paper*, **6363**.

Beine, Michel, Docquier, Frédéric, & Özden, Çağlar. 2010. Diasporas. *Journal of Development Economics*, **95**(1), 30–41.

Beine, Michel, Docquier, Frédéric, & Özden, Çağlar. 2011. Dissecting Network Externalities in International Migration. *CESifo Discussion Paper*, **3333**.

Biavaschi, Costanza. 2012. Recovering the Counterfactual Wage Distribution with Selective Return Migration. *IZA Discussion Paper*, **6795**.

Bikchandani, Sushil, Hirshleifer, David, & Welch, Ivo. 1992. A Theory of Fads, Fashion, Custom and Cultural Change as Informational Cascades. *Journal of Political Economy*, **100**(5), 992–1027.

Borjas, George J. 1987. Self-Selection and the Earnings of Immigrants. *The American Economic Review*, **77**(4), 531–553.

Borjas, George J. 1995. Ethnicity, Neighborhoods and Human-Capital Externalities. *The American Economic Review*, **85**(3), 365–390.

Borjas, George J., & Bratsberg, Bernt. 1996. Who Leaves? The Outmigration of the Foreign-Born. *The Review of Economics and Statistics*, **78**(1), 165–176.

Borjas, George J., Freeman, Richard B., & Lang, Kevin. 1991. Undocumented Mexican-born Workers in the United States: How Many, How Permanent? *Chap. 2, pages 77–100 of:* Abowd, John M., & Freeman, Richard B. (eds), *Immigration, Trade and the Labor Market*. University of Chicago Press.

Campos-Vazquez, Raymundo M, & Lara, Jaime. 2012. Self-selection Patterns among Return Migrants: Mexico 1990-2010. *IZA Journal of Migration*, **1**(8).

Card, David, & Lewis, Ethan G. 2007. The Diffusion of Mexican Immigrants During the 1990s: Explanations and Impacts. *Chap. 6, pages 193–228 of:* Borjas, George J. (ed), *Mexican Immigration to the United States*. University Of Chicago Press.

Carrington, William J., Detragiache, Enrica, & Vishwanath, Tara. 1996. Migration with Endogenous Moving Costs. *The American Economic Review*, **86**(4), 909–930.

Chiquiar, Daniel, & Hanson, Gordon H. 2005. International Migration, Self-Selection and the Distribution of Wages: Evidence from Mexico and the United States. *Journal of Political Economy*, **113**(2).

Costanzo, Joseph M., Davis, Cynthia, Irazi, Caribert, Goodkind, Daniel, & Ramirez, Roberto. 2002. Evaluating Components of International Migration: The Residual Foreign Born. *US Census Bureau Population Division Working Paper*, **61**.

Currarini, Sergio, Jackson, Matthew O., & Pin, Paolo. 2009. An Economic Model of Friendship: Homophily, Minorities and Segregation. *Econometrica*, **77**(4), 1003–1045.

Edin, Per-Anders, Frederiksson, Peter, & Aslund, Olof. 2003. Ethnic Enclaves and the Economic Success if Immigrants - Evidence from a Natural Experiment. *The Quarterly Journal of Economics*, **118**(1), 329–357.

Farré, Lidia, & Fasani, Francesco. 2013. Media Exposure and Internal Migration: Evidence from Indonesia. *Journal of Development Economics*, **102**(C), 48–61.

Fernández-Huertas Moraga, Jesús. 2011. New Evidence on Emigrant Selection. *The Review of Economics and Statistics*, **93**(1), 72–96.

Golub, Benjamin, & Jackson, Matthew O. 2010. Na ive Learning in Social Networks and the Wisdom of Crowds. *American Economic Journal: Microeconomics*, **2**(1), 112–149.

Golub, Benjamin, & Jackson, Matthew O. 2012. How Homophily Affects the Speed of Learning and Best-Response Dynamics. *The Quarterly Journal of Economics*, **127**(3), 1287–1338.

Granovetter, Mark. 2005. The Impact of Social Structure on Economic Outcomes. *Journal of Economic Perspectives*, **19**(1), 33–50.

Granovetter, Mark S. 1973. The Strengh of Weak Ties. *American Journal of Sociology*, **78**(6), 1360–1380.

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

Hanson, Gordon H. 2006. Illegal Migration from Mexico to the United States. *Journal of Economic Literature*, **44**(4), 869–924.

Ibarraran, Pablo, & Lubotsky, Darren. 2007. Mexican Immigration and Self-Selection: New Evidence from the 2000 Mexican Census. *Chap. 5, pages 159 – 192 of:* Borjas, George J. (ed), *Mexican Immigration to the United States*. University Of Chicago Press.

Kennan, John, & Walker, James R. 2011. The Effect of Expected Income on Individual Migration Decisions. *Econometrica*, **79**(1), 211–251.

Lazear, Edward P. 1999. Culture and Language. *The Journal of Political Economy*, **107**(6), S95–S126.

Massey, Douglas S., & Liang, Zai. 1989. The Long-Term Consequences of a Temporary Worker Program: The US Bracero Experience. *Population Research and Policy Review*, **8**(3), 199–226.

Massey, Douglas S., & Malone, Nolan. 2002. Pathways to Legal Immigration. *Population Research and Policy Review*, **21**(6), 473–504.

Massey, Douglas S., & Zenteno, René. 2000. A Validation of the Ethnosurvey: The Case of Mexico-U.S. Migration. *International Migration Review*, **34**(3), 766–793.

McKenzie, David, & Rapoport, Hillel. 2007. Network Effects and the Dynamics of Migration and Inequality: Theory and Evidence from Mexico. *Journal of Development Economics*, **84**, 1–24.

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

McKenzie, David, Gibson, John, & Stillman, Steven. 2013. A Land of Milk and Honey with Streets Paved with Gold: Do Emigrants Have Over-Optimistic Expectations about Incomes Abroad? *Journal of Development Economics*, **forthcoming**.

McPherson, Miller, Smith-Lovin, Lynn, & Cook, James M. 2001. Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology*, **27**, 415–444.

Munshi, Kaivan. 2003. Networks in the Modern Economy: Mexican Migrants in the U.S. Labor Market. *The Quarterly Journal of Economics*, **118**(2), 549–597.

Orrenius, Pia M., & Zavodny, Madeline. 2005. Self-Selection among Undocumented Immigrants from Mexico. *Journal of Development Economics*, **78**, 215–240.

Pedersen, Peder J., Pytlikova, Mariola, & Smith, Nina. 2008. Selection or Network Effects? Migration Flows into 27 OECD Countries, 1990-2000. *European Economic Review*, **52**, 1160–1186.

Vigdor, Jacob L. 2008. Measuring Immigrant Assimilation in the United States. *Center for Civic Innovation, Civic Report*, **53**.

Winters, Paul, de Janvry, Alain, & Sadoulet, Elisabeth. 2001. Family and Community Networks in Mexico-U.S. Migration. *Journal of Human Resources*, **36**(1), 159–184.

A Other datasets

Given the available data on Mexican migration in the US, a researcher faces the trade-off between using a large representative dataset with little direct information on networks and without a longitudinal dimension on the one hand, and small datasets that can offer this additional dimension yet cannot provide the variation in network characteristics that we would need on the other. Using the census, we opted for sample size, which we consider as a necessary condition to say anything about diaspora networks.

Other datasets on Mexicans in the US, unfortunately, are too small for our analysis. The household surveys ENET (Encuesta Nacional de Empleo Trimestral), ENADID (Encuesta Nacional de la dinámica demográfica), and the Mexican Family Life Survey (MxFLS) are conducted in Mexico, and have little information on Mexicans that already reside in the US. The Mexican Migration Project (MMP), a survey of Mexican migrants that contains both migrants and non-migrants, has some information on family and friends in the US, and on the help of these networks in crossing the border and finding a job. Numerous studies use the MMP to analyze the effect of networks on migration decisions (Munshi, 2003; Bauer *et al.*, 2005; Amuedo-Dorantes & Mundra, 2007; McKenzie & Rapoport, 2007; Bauer *et al.*, 2007). The MMP is representative of migration flows to the US (Massey & Zenteno, 2000), but it is not representative of the stocks. Additionally, it does not have any information on the characteristics of friends and family networks in the US, which is what our analysis requires.

B Data appendix

B.1 Education Groups

For the prediction of the counterfactual wages in Section 3.2 and for the regressions in Section 5 we use four broad education groups. Clustering the workers into broad education groups makes the interpretation of the estimates easier and allows us to match the Mexican and the US data. Table 4 shows the education groups for the Mexican and the US census. For the Mexican census we take the variable *years of schooling* (YRSCHL). The US census distinguishes between 11 education groups (variable EDUC).

Table 4: Education groups in the Mexican and US census

Nr	Education group	Mexican census	US census
1	High-school dropouts	less than 5 years of schooling	education group 1
2	Lower secondary education	5-9 years of schooling	education groups 2-4
3	Upper secondary education	10-12 years of schooling	education groups 5-7
4	Third-level education	13 or more years of schooling	education groups 8-11

B.2 Data Cleaning US census

In the US census we exclude the following observations:

- younger than 18 and older than 64 years,
- younger than 18 at the time of immigration,
- if still enrolled in education (SCHOOL=2),
- self-employed people,
- with an annual wage income (INCWAGE) higher than 200,000 USD, as these were clear outliers,
- living in Hawaii and Alaska,
- if born to American parents in Mexico (CITIZEN=1),
- with unknown income,
- who work less than 7 hours a week (UHRSWORK) or less than 8 weeks a year (WKSWORK1, not available for 1980), or if any of these is missing,
- if they live in group quarters (hospitals, prisons, etc; GQ=3 or GQ=4)
- if they moved to a district (CONSPUMA) with at least 20 Mexicans.

To make wages comparable between the US and Mexico, we use monthly wages. We obtain monthly wages by dividing the annual wages by 12. Given that not all Mexicans work throughout the entire year and work full-time, we adjust the income by weeks worked per year (WKSWORK1) and by hours worked in a typical workweek (UHRSWORK). In the 1980 census, we obtain the adjusted monthly income by multiplying the nominal monthly income by 40 (the full time equivalent) and divide it by the actual hours worked. From 1990 onwards, we also have information on the average weeks per year, and thus the adjusted income is calculated as

$$\text{adjusted income} = \text{ nominal income } \frac{52 * 40}{\text{weeks worked } * \text{ hours worked}}. \tag{12}$$

In the ACS, the number of weeks worked comes in six categories, and we use the midpoints for each category (7; 20; 33; 43.5; 48.5; 51). In some rare cases, the denominator in Equation (12) is very small — if the person has worked few hours and few weeks — and we drop every observation that yields an adjusted wage income of more than 15,000 USD per month.

B.3 Mexican census

We use the 10% files of the Mexican census in 1990, 2000, and 2010 for the estimation of counterfactual wages. The following observation are excluded:

- younger than 18 and older than 64 years
- more than 100 or less than 10 hours of work per week (HRSWORK1)
- self-employed

Monthly income is taken from the variable INCEARN. As with the US census, we adjust monthly income by hours of work by multiplying it with 40 and dividing it by the usual hours of work per week (HRSWORK1). To convert the monthly wage into PPP dollars, we divide the adjusted wage by a PPP factor (price level Mexico over Price level US) and the exchange rate (pesos per dollar).[10]

[10] The PPP factor is the amount of goods in return for one dollar in the US over the amount of goods in return for one dollar in Mexico. The PPP factor was 0.48 in 1990, 0.63 in 2000, and 0.68 in 2010. The exchange rates were 2.83 pesos per dollar in 1990, 9.2845 in 2000, and 12.6287. Sources: Penn World Tables (PPP) Mexican Central Bank (Exchange Rate).

C Robustness checks

C.1 Counterfactual Wages

We predict counterfactual wages using three approaches: a sample based on propensity score matching, a sample consisting of internal migrants, and a Heckman selection model that accounts for selection into employment. Table 5 shows the correlation coefficients for the counterfactual wages on the entire sample of Mexicans in the US. The correlation coefficients are remarkably large, which gives us confidence that the straightforward prediction of Mexican wages does not suffer from severe selection bias.

Table 5: Counterfactual Wages: Correlations

	2000				**2010**			
	Baseline	PSM	Internal	Heck	Baseline	PSM	Internal	Heck
Baseline	1				1			
PSM	0.98	1			0.99	1		
Internal	0.90	0.92	1		0.92	0.93	1	
Heckman	0.95	0.94	0.80	1	0.98	0.98	0.89	1

Note: The table displays the correlations between different predictions of counterfactual wages of Mexicans in the US.