Article:

Elsner, Ben, Narciso, Gaia and Thijssen, Jacco Johan Jacob orcid.org/0000-0001-6207-5647 (2018) Migrant Networks and the Spread of Information. Oxford Bulletin of Economics and Statistics. pp. 659-688. ISSN 0305-9049

https://doi.org/10.1111/obes.12216

Migrant Networks and the Spread of Information*

August 4, 2017

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, that is, they migrate only if they gain. We test these predictions empirically using data on recent Mexican migrants to the United States. To instrument for the quality of networks, we exploit the settlement of immigrants who came during the Bracero program in the 1950s. The results are consistent with the model predictions, providing evidence that connections to a better integrated network lead to better outcomes after migration.

Word count: 12,010

JEL codes: F22, J15, J61

*We would like to thank the Editor, two anonymous referees, as well as 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. Margaryta Klymak and Gaspare Tortorici provided excellent research assistance. Elsner gratefully acknowledges funding from the Irish Research Council for the Humanities & Social Sciences (IRCHSS).

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.[1]

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 potential 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

[1] Throughout the paper, we use the terms "integration" and "assimilation" interchangeably.

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 integration of past immigrants, low wages of current immigrants, and overall a positive correlation between both variables. To address this endogeneity, we instrument for the integration 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 74 USD, 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 (forthcoming) 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.*, 2015). 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: Descriptive Evidence and Theory

We begin by presenting the core idea of the paper, namely that more integrated migrant networks provide information of higher quality to potential migrants in their country of origin. Using data from migrant surveys in Ireland and Germany, we provide descriptive evidence of the nature and frequency of information flows between migrant networks and their communities in the country of origin. Finally, we formulate a decision model that explains how the information received from the network affects post-migration outcomes, and how this relationship is varies by networks of different degrees of integration in the host society.

2.1 Idea: Network Integration and Information Quality

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.*, 2013), 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

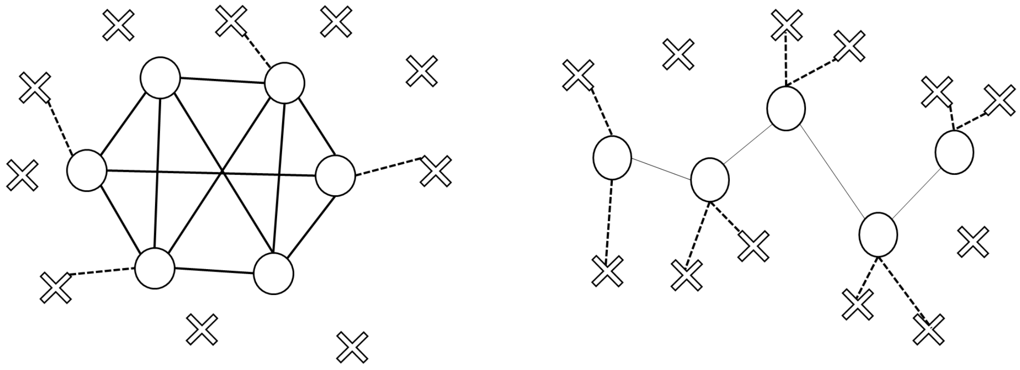

Figure 1: Ethnic enclave (left) and loosely connected network (right)

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.

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 and transmit accurate information.

2.2 Migration and information flows: suggestive evidence

A key ingredient to our theoretical reasoning is actual information flows between the network and the potential migrant before migration. While it seems natural that migrants communicate with their contacts in the destination before their departure, only few datasets comprise information on the frequency and intensity of these contacts.[2] With respect to Mexican migration to the

[2] The two notable exceptions are the IAB-SOEP Migration Sample (Brücker *et al.*, 2014) and a survey of immigrants in Dublin collected by Batista & Narciso (forthcoming). In particular, the survey conducted in Ireland comprises rich information on both the immigrants' integration in Ireland and their information flows with their countries of origin. One variable that can serve as proxy for the integration of the network in the destination is having friends among or at least regular contact with locals. Around 30% of all respondents in

US, to the best of our knowledge, the Mexican Migration Project (MMP) is the only survey that contains some information about interactions between Mexican migrants and potential migrants.

Although the amount of information available in the MMP on this topic is quite limited, the evidence provided is in line with our theoretical reasoning. According to the MMP data for the period 2000-2016, the majority of Mexican migrants (around 60%) were in contact with their home-community members during their trip to the US. Interestingly, the proportion of migrants in contact with their community of origin is higher during the last trip relative to their first trip to the US. Overall, these figures provide some suggestive evidence for the interaction between Mexican immigrants in the US and potential migrants in Mexico as well as the likely flow of information between origin and destination communities.

Additional evidence is provided in qualitative as well as quantitative field studies. Massey *et al.* (1994), who conducted a field study in 19 Mexican communities, stress the importance of networks in providing information to potential migrants. In particular, the authors stress how information about job opportunities in the US might not be available to potential migrants living in communities with few migrants in the US. Using in-depth interviews with 138 Mexican migrants and their families, Garip & Asad (2016) find strong evidence for the role of networks in the migration decision. Most interviewed migrants received information on labor market prospects from members of their home community that had previously migrated to the US. Most respondents report that networks helped them with information on job opportunities as well as with the knowledge of local amenities. Based on observational data, Winters *et al.* (2001) further examine the importance of networks for the decision of Mexicans to migrate to the US as well as the level of migration. The paper explores different types of networks — family as well as non-family networks. The results point to an important role of both types of networks in providing information about migration as well as help once the migrant is in the US. These findings are in line with our own calculations from the MMP data, according to which the vast majority of Mexican migrants (over 80%) receive support from the extended migration network upon arrival to the US.

this survey state that they have at least one Irish friend. Comparing immigrants with and without friends among locals, the survey data show that both groups are equally likely to provide information about the destination to people in their home country, although immigrants with Irish friends tend to communicate with a larger number of people in their home country.

2.3 A Model of 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 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.[3]

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.

3.1 Mexican Migration to the US

To test this relationship empirically, we use data on Mexican immigrants in the US. According to the American Community Survey, in 2011, there are over 33 million Hispanics in the US with Mexican origins, of which over 11 million were born in Mexico. Until today, there are

[3] 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.

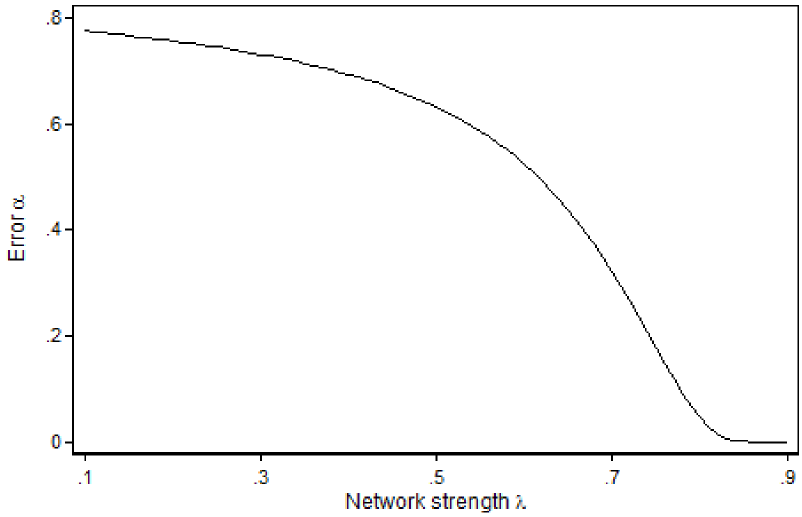

Figure 2: Network quality and the likelihood of taking a wrong decision

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

significant migration flows to the US. Between 2005 and 2010, for example, an estimated 1.1% of the Mexican population migrated internationally, mostly to the US.[4]

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. However, 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.

3.2 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

[4] Source: Census of Population and Housing 2010 (Censo de Poblacion y Vivienida).

of the Mexican census provided by IPUMS.[5] 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.[6]

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.[7] 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 (2016) and Campos-Vazquez & Lara (2012), selective out-migration of 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

[5] 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.

[6] Moreover, the census only includes people who reside in the US; it thus does not include people who visit the US on a tourist visa or any other short-term visitors (Hanson, 2006).

[7] As districts, we use CONSPUMAs (consistent 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.

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.[8] 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.[9]

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.

3.3 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

[8] 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.

[9] 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. Wages from both countries are taken from the US and Mexican censuses. Wages are self-reported. 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.[10] Initially, we only include workers with a positive income in the wage regressions. In Appendix F, we test the robustness of the wage predictions using a two-step selection model on the full sample.

3.3.1 Counterfactual wages

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

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

from which we obtain an estimate for skill prices in Mexico, $\hat{\boldsymbol{\beta}}_{\boldsymbol{MEX}}$. The vector $\boldsymbol{X_{MEX}}$ includes a set of education dummies, age, and age squared, as well as interactions of the education dummies with age and age squared. Unobservable determinants of log wages are determined by the i.i.d error term ε. The functional form — log wages regressed on education, age and other observable characteristics — represents a Mincer earnings function, although the interaction terms allow us to have a separate age-earnings gradient for each education level. Compared to a regression with wages in levels as dependent variable, the log transformation of wages typically ensure a better model fit. Moreover, the Mincer equation is firmly grounded in a theory of human capital investment (Mincer, 1974).

[10] See Appendix B for a description of the samples and the wage adjustment.

Contrary to what is done in large parts of the literature, the goal of estimating Equation (5) is *not* to obtain causal estimates for one or more parameters, but to obtain a prediction of a person's expected earnings given that person's observable characteristics. Nonetheless, it is important to assume that the error term is i.i.d (independently and identically distributed). Otherwise we would systematically over- or under-predict a person's wage. As argued by Polachek (2008), a crucial assumption of earnings functions is that the age-earnings profile is constant across education groups, which is often not the case and would potentially violate the i.i.d assumption. To alleviate this concern, and to improve the fit of the regression model, we include interaction terms of the education dummies with age and age squared.[11] The regressors included in the model explain 30.8% of the total variation in log wages. The goodness of fit could be improved by adding regressors with additional predictive power, although our choice of regressors is constrained by the joint availability in the US and Mexican censuses.

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

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

In some specifications, we will use wages in levels rather than logs, which we obtain by taking the exponential of the predicted log wage, $\widehat{\text{wage}} = exp(\boldsymbol{X_{US}}\boldsymbol{\hat{\beta}_{MEX}})$. 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.[12]

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 between zero and 500 USD per month.[13]

[11] To graphically assess the validity of the i.i.d assumption, we plotted the residuals from the wage regressions, which display a symmetric distribution centered at zero. See Appendix C.

[12] The PPP conversion implicitly assumes that migrants consume their entire income in the US.

[13] Whether workers with, say, a 200 USD difference are indeed better off in monetary terms depends on the moving costs and a person's discount rate. For a given discount rate, the longer it takes a person to recover the moving costs, the worse.

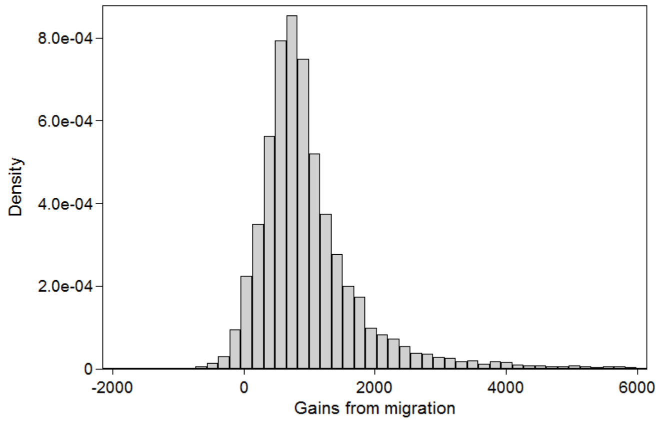

Figure 3: Gains from emigration

Note: The graph shows the distribution of the gains 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.

3.3.2 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 F, 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).

3.4 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.[14]

We calculate the assimilation index at the smallest geographic unit that is consistently available across multiple rounds of the US census, namely the so-called *CONSPUMA* . In each round, the Census Bureau defines PUMAs (Public Use Microdata Area), small geographic units with a population between 100,000 and 200,000 people. PUMAs 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. The definition of PUMAs changes with every census round. To make PUMAs comparable over time, the US Census Bureau has introduced CONSPUMAs that have the same boundaries from 1980 to 2010 and are larger than the original PUMAs.[15] 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 CONSPUMA between 1995 and 2000, we assign the assimilation index of Mexicans that lived in the same area in 1990.

[14] It is important to note here that we use assimilation as a statistical concept rather than a sociological one. According to the sociological definition, a person is assimilated if he/she has given up his cultural identity, as opposed to integration, which is defined as showing commitment to the host society while maintaining one's cultural identity (Harles, 1997). A further definition of assimilation that has been used in economics is wages (Chiswick, 1978; Borjas, 1985). In contrast, we use a statistical definition, whereby we see migrants as assimilated if we cannot statistically distinguish them from natives based on a large array of observable characteristics. This means that our measure is broader than assimilation based on wages.

[15] According to the Census Bureau, PUMAs cannot be matched across census rounds. The size of CONSPUMAs ranges between 100,000 and 4.3 million inhabitants.

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 — both men and women — in the sample and run in each metropolitan 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 for each CONSPUMA.

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.[16]

We then restrict the sample to Mexicans only, and pretend for the moment that we do not

[16] To calculate the assimilation index, we need to run a separate probit regression in each local area. We choose metropolitan areas as geographic units here, because each metropolitan area has enough Mexicans for the estimator to converge. In small CONSPUMAs, especially in areas with a low share of immigrants, the number of Mexicans is not sufficient for the estimator to converge. However, the analysis yields a separate prediction for every person, which we then aggregate at the CONSPUMA level.

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.

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

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 CONSPUMA, $\widehat{p_m}$, 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 CONSPUMA-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.

The assimilation index is based on personal characteristics as well as variables related to economic well-being, such as wages, employment, or the earnings score. In the empirical analysis to follow, it is important to show that the results are robust to the inclusion or exclusion of several variables. In addition to the baseline results, we present estimations with an assimilation index that includes self-reported language skills — a direct measure of assimilation. We also present results for an index that only includes personal characteristics.

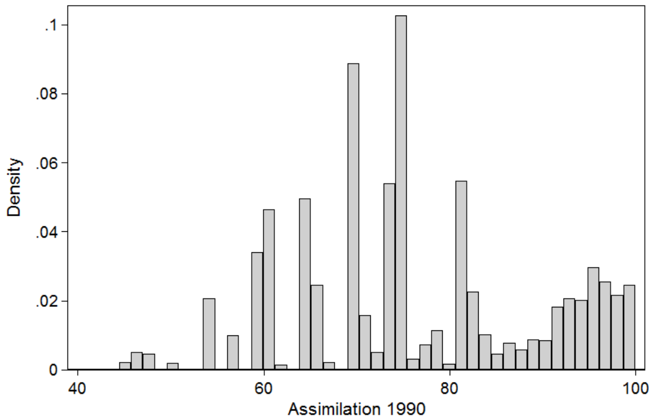

Figure 4: Assimilation index in 1990

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 CONSPUMA.

Table 1: Summary Statistics of the main variables

Variable	*N*	*Mean*	*SD*	*Min*	*Max*
A. Aggregate data					
Monthly income (USD)	202	1224.29	528.70	395	4417
Income difference US-Mexico	202	890.32	476.31	67	3558
Assimilation in 1990	202	85.99	14.10	44	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					
Monthly income (USD)	20131	1121.30	1168.79	0	14614
Income difference US-Mexico	20131	810.29	1152.19	-937	13798
Age	20131	28.62	8.72	18	64
Age at immigration	20131	26.66	8.70	18	64
Assimilation index	20131	76.06	13.05	44	100
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.

3.5 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 CONSPUMA 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).[17]

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 800 USD more than they would earn in Mexico — although there is a large degree of heterogeneity in the income difference.

4 Identification and Estimation Strategy

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

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

in which the dependent variable y_{ij} is a measure for the success of migrant i in CONSPUMA j. To align the empirical analysis with the theoretical model, in our baseline regressions, the dependent variable is the wage difference between the monthly wage in the US and the predicted counterfactual wage in Mexico in USD. We will also report regression results with US wages in

[17] The mean assimilation index differs between Table 1 and Figure 4, because in the former, every Mexican community receives equal weight, while in the latter, observations are weighted by the number of Mexicans in the community.

levels and logs as dependent variable.

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 polynomial in the number of Mexicans that have lived in the CONSPUMA in 1995. Moreover, 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 error term that captures all other factors that determine the wage difference but are not controlled for in the regression.

4.1 Inference

Statistical inference is challenging due to three factors. First, in some regressions, the dependent variable varies at the individual level whereas the regressor is a group variable and only varies across CONSPUMAs, which means that the error terms are potentially correlated within CONSPUMAs. Second, the assimilation index is a generated regressor, i.e. the result of a prediction, which potentially leads to an underestimate of the standard errors. Finally, in some specifications the dependent variable is the difference between the actual wage in the US and a predicted counterfactual wage in Mexico. Therefore, one component of the dependent variable is predicted, which may introduce heteroskedasticity and potentially inflates the variance. In light of these issues, reliable inference can only be drawn if standard errors are adjusted appropriately.

Our solution simultaneously solves the first two problems, namely clustering and generated regressors. In all unweighted regressions — which are the source of most results reported in this paper — we compute bootstrapped standard errors with 1000 replication. In regressions where an individual-level variable is regressed on a group variable, we use a block bootstrap whereby a block equals a CONSPUMA. While, in general, bootstrapped standard errors are immune to the bias resulting from including a generated regressor, bootstrapping at the block level accounts for the within-CONSPUMA correlation of the error term, thus eliminating the clustering problem (Davidson & MacKinnon, 2006). To remedy the third problem, one solution would be to report heteroskedasticity-robust or clustered standard errors.[18] However, these do not adjust for the bias from generated regressors. To assess the robustness of our inference in the presence of a generated dependent variable, we will perform a robustness check whereby the dependent variable is the US wage alone.

4.2 Instrumental Variable Strategy

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 that affects one group of migrants but not another — 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.

[18] Lewis & Linzer (2005) show that generated dependent variables lead to heteroskedasticity. Cameron & Miller (2015) explain why clustered standard errors account for heterogeneity.

4.2.1 The Bracero Progam

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. The program was initially introduced as a wartime measure to compensate for the labor shortage in agriculture, and it was subsequently expanded and extended by Congress. Over the duration of the program, from 1942 to 1964, around 4.6 million Mexican workers temporarily moved to the US,[19] mainly to work in agriculture. The number of Mexican migrants entering to the US under the Bracero program steadily increased since 1942 until its peak in 1959 with 437,643 new admissions (Calavita, 1992). The number of new admissions subsequently declined until the end of the program in 1964 (McElroy & Gavett, 1965). The recruitment process involved four parties. Federal officials informed Mexican authorities about the amount of labour requested by American agricultural businesses. Mexican authorities then selected suitable candidates before the final assessment was performed in the United States. Applicants that were suited for the job were temporarily employed before being repatriated. Although during the initial phase guest-workers hailed from, and were recruited in, Mexico City, American labor demand had increasingly been satisfied by individuals coming from rural areas, who were arguably more accustomed to agricultural occupations. Further recruitment centers were opened in cities closer to the border, such as Chihuahua, Hermosillo, and Monterrey.

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 helped create more closed-up Mexican communities, resulting in a low degree of assimilation of Mexicans living in the same places 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.[20] Clearly, a higher

[19] Most sources report estimates of the total number: the Bracero History Archive reports 4.6 million (*www.braceroarchive.org/about*, accessed on July 4, 2017), Massey & Liang (1989) 4.5 million.

[20] We compute the number of Bracero migrants in a CONSPUMA from the 2000 census based on the number of hispanic Mexican-born immigrants who immigrated between 1942 and 1964.

share of Bracero immigrants predicts a low level of assimilation. Whether this relationship is strong enough to eliminate weak instrument bias will depend on the specification. The conventional threshold for instruments not to be considered weak is an F-Statistic of the excluded instrument of $F > 10$ (Stock *et al.*, 2002). As shown by Bound *et al.* (1995) and Staiger & Stock (1997), weak instruments can introduce two biases into an estimate. First, in finite samples, weak instruments lead to a small sample bias that goes in the same direction as the OLS estimate that includes the endogenous regressor. Second, a small violation of the exclusion restriction can be severely inflated, introducing a bias of unknown sign. In addition, with weak instruments, the two-stage-least-squares estimator can understate the standard errors in the second-stage, leading to an under-rejection of the null hypothesis of no effect (Moreira, 2003).[21]

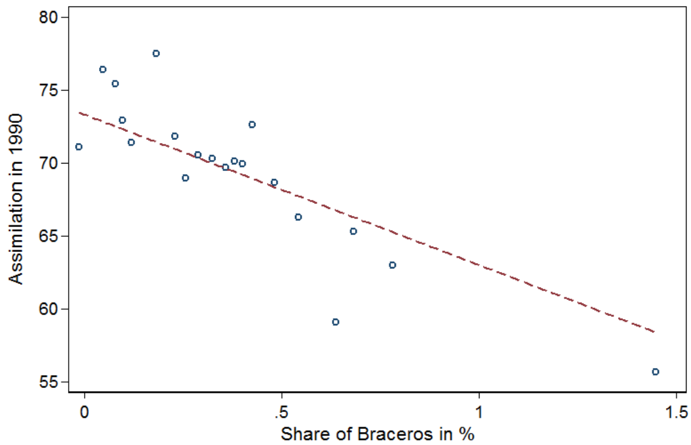

Figure 5: First stage relationship

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 fourth-order polynomial in the number of Mexicans in a CONSPUMA.

4.2.2 Instrument validity

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

[21] We assess the severity of these issues for our estimations in Appendix E.

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. In robustness checks, we will also control for several characteristics of the Mexican communities, such as average education and the employment rate.

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. Panel A in Table 2 displays the results for the following estimating equation

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

where $\boldsymbol{R_j}$ includes the average wage of US natives in 2000 in all specifications, and a fourth-order polynomial in the number of Mexicans in some.[22] Standard errors in all columns except (2) and (5) have been computed using a bootstrap with 1000 replications. For the weighted regressions in Columns (2) and (5), we report heteroskedasticity-robust standard errors.

Column (1) displays the OLS estimate for β in Equation (11). The partial correlation is positive and statistically significant at the 1%-level. An increase in the assimilation index by one point increases the monthly wage difference between US and Mexican wages by 5.8 USD. This may not sound like a large effect; but increasing the assimilation index by one standard deviation (SD=14), increases the wage difference by 81.2 USD per month, or 17.5% of a standard

[22] Controlling for a higher-order polynomial in the size of the community within a CONSPUMA allows us to account for the uneven size distribution of Mexican communities. We assess the robustness of the estimates to the functional form in Section 5.3.

Table 2: Network assimilation and the success of recent migrants

Dependent variable: wage difference USA - Mexico

A. CONSPUMA level						
	OLS	OLS	OLS	IV	IV	IV
	(1)	(2)	(3)	(4)	(5)	(6)
Assimilation index	5.815***	6.734***	4.868***	6.286***	6.893	3.555
	(1.547)	(0.980)	(1.704)	(2.362)	(4.208)	(50.708)
Weighted by size	No	Yes	No	No	Yes	No
Control: nr of Mexicans	No	No	Yes	No	No	Yes
First stage:						
Share Braceros				-16.110***	-7.835***	-2.536
				(4.037)	(2.890)	(2.466)
F-Statistic				12.53	7.35	2.49
N	202	202	202	202	202	202

B. individual level						
	OLS	IV	IV	OLS	IV	IV
	(1)	(2)	(3)	(4)	(5)	(6)
Assimilation index	5.033***	6.626***	6.204***	5.057***	9.082***	8.846***
	(0.895)	(1.351)	(1.211)	(1.145)	(3.378)	(2.993)
First stage:						
Share Braceros		-14.804***	-14.643***		-7.742**	-7.683**
		(5.402)	(5.506)		(3.577)	(3.607)
Control: nr of Mexicans	No	No	No	Yes	Yes	Yes
Exclude if wage US= 0	No	No	Yes	No	No	Yes
F-Statistic		13.94	12.84		6.59	6.08
N	20131	20131	15082	20131	20131	15082

Note: This table displays the results of OLS- and IV-regressions of the difference in monthly wages on the assimilation index. The counterfactual wages for Mexico are based on a Mincer equation, as explained in Section 3.3.1. In Panel A), the unit of observation is a CONSPUMA, while in Panel B), the unit of observation is an individual. Standard errors in Panel A have been bootstrapped with 1000 replications, with the exception of those in Columns (2) and (5), which are heteroskedasticity-robust. Standard errors in Panel B have been computed with a block bootstrap with 1000 replications. Significance levels: * p<0.1, ** p<0.05, *** p<0.01.

deviation in the wage difference.

While in Column (1) every CONSPUMA receives 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 fourth-order polynomial in the number of Mexicans. In this case, the point estimate is slightly smaller and less precisely estimated. It is statistically significant with bootstrapped standard errors, but insignificant when we use conventional standard errors. 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, there is a strong negative first-stage relationship between the share of Braceros and the assimilation index. In the unweighted regression in Column (4), the instrument is strong enough to rule out a weak instrument problem. Once we weight the regression by the size of the Mexican community, the instrument is weaker (F=7.35) and the estimates are less precise, although the magnitude of the point estimate remains in a similar range as the OLS estimates. In column (6), however, the instrument weakens (F=2.49), resulting in an imprecise estimate that is lower than all other estimates reported in this row.

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. Panel B in Table 2 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 5 USD 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 13.9. The point estimate is slightly larger than 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 fourth-order polynomial in the number of Mexicans in a CONSPUMA. The first stage now becomes weaker and the point estimates are higher. However, the difference in point estimates between Columns (3) and (6) and Columns (2) and (5) are statistically insignificant.

In sum, these estimates suggest that a one-point increase in the assimilation index increases the wage difference by 5-9 USD. 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 between 100 and 180 USD per month.

5.2 Robustness checks: controlling for characteristics of Mexicans in 1990

In a next step, we assess the robustness of our results to controls for characteristics of the Mexican community such as the size, average education, or the employment rate.

In several specifications, we control for the size of the Mexican community, which helps us to isolate the impact of the assimilation of a community from the impact of the size itself. Given the uneven size distribution of Mexican communities across CONSPUMAS, ranging from zero to over 500,000, we chose to control for the number of Mexicans per CONSPUMA with a fourth-order polynomial. Higher-order polynomials are more flexible than linear or quadratic controls, but are also more taxing on the degrees of freedom, which can decrease the precision of the estimates.

To assess the robustness of the estimates to the functional form of these controls, Colums (1)-(3) in Table 3 present the estimation results for OLS and IV regressions with varying controls for the size of the Mexican community, ranging from a linear control to a fourth-order polynomial. All other control variables are the same as those used in Section 5. With the addition of higher-order polynomials, going from left to right, the point estimates remain at a similar level. In some rows of Table 3 the estimates are larger with linear controls than in specifications with a quartic while in others the opposite is true. However, the point estimates within a row are never statistically different from each other. The statistical significance of the coefficient declines with the addition of higher-order terms, as does the strength of the excluded instruments in the first stage. Therefore, including a fourth-order polynomial leads to more conservative estimates compared to including a linear control or a quadratic.

In Columns (4)-(6), we additionally control for several characteristics of the Mexican community in a given CONSPUMA in 1990, namely the average years of schooling, the share of women, the share of people who are married and the employment rate. These controls are potentially important in OLS estimations, because community characteristics other than assimilation may directly affect an immigrants' outcome while being correlated with the assimilation index.

Once the characteristics are included in the regressions, the point estimates become significantly larger. In the OLS regressions, this is particularly so when we consider the aggregate level in Panel A, while the point estimates remain similar in the individual-level OLS regressions in Panel C. In the IV regressions, the estimates conditional on network controls are considerable larger than the those without controls. At the same time, the first stage of the instrument becomes weaker and falls below the commonly used threshold of an F-statistic of 10. The weak instruments can provide one explanation for the increase in coefficients, although this increase is similar to the one in Panel A, which is not subject to weak instrument bias.

Overall, Table 3 suggests that the overall conclusion that a more assimilated network leads to better outcomes for newly arrived immigrants holds and is robust to the inclusion of community controls. The difference in magnitude of the estimated coefficients between the aggregate- and individual-level regressions can be explained by the different weights CONSPUMAs receive in each specification. In the CONSPUMA-level regressions, each CONSPUMA receives equal weight whereas in the individual-level regressions, the weight is proportional to the size of the Mexican community in a CONSPUMA. If, for example, the marginal effect of greater assimilation is larger in smaller communities, this will lead to larger estimates in CONSPUMA-level regressions. And while not all coefficients in Table 3 are statistically significant, most of them are, and they consistently have a positive sign.

The fact that the coefficient of the assimilation index remains large and statistically significant also highlights that the assimilation index measures the relative difference in characteristics between Americans and Mexicans in an area. Control variables such as the average level of education or average employment rates of Mexicans may be correlated with the assimilation index, but the results suggest that there is significant variation in the relative difference between Mexicans and Americans.

5.3 Additional results and robustness checks

In addition to the analysis described above, we perform a series of robustness checks, which we summarize here. Further details and regression outputs can be found in Appendix D.

Results with (log) wages as dependent variable In all specifications presented so far, the dependent variable has been the wage difference between the actual wage in the US and a counterfactual wage in Mexico, which was predicted based on observable characteristics. This choice of dependent variable is informed by the theoretical model, which predicts that migrants with access to more assimilated networks are more likely to do better relative to their situation in Mexico. The wage difference approximates this wage difference between both countries.

However, the fact that the counterfactual wages are predicted potentially creates a redundancy in the econometric model. The reason for this is that the counterfactual wages have been predicted based on the same observable characteristics $\boldsymbol{X_{ij}}$ that are controlled for in the individual-level regressions. In fact, if set of variables predicting the counterfactual wage was completely contained in the set of control variables, the variation in the dependent variable would only come from US wages. In our case, the prediction is based on several variables that are not included in the regression — mainly interactions between education dummies and age — but, nonetheless, much of the variation in the counterfactual wage is absorbed by the control variables.

In Table 5 in Appendix D, we re-estimate the same models as in Table 2 but use as dependent variable the level of US wages. In most cases, the point estimates are slightly larger than those in Table 2 but are in the same ballpark. This small difference indicates that, indeed, most of the variation in the dependent variable is due to variation in US wages.

Table 3: Controlling for network characteristics

Dependent variable: wage difference USA - Mexico

	(1)	(2)	(3)	(4)	(5)	(6)
A. CONSPUMA level - OLS results						
Assimilation index	5.540***	5.467***	4.868***	11.171***	11.336***	10.805***
	(1.574)	(1.596)	(1.704)	(2.851)	(2.908)	(2.806)
B. CONSPUMA level - IV results						
Assimilation index	5.889*	5.833	3.555	13.325**	14.465*	15.744
	(3.436)	(6.402)	(50.708)	(6.433)	(8.232)	(51.119)
First stage:						
Share Braceros	-11.826***	-9.148***	-2.536	-9.437***	-7.389**	-2.293
	(3.609)	(3.414)	(2.466)	(3.453)	(2.974)	(2.202)
F-Statistic	8.70	5.85	2.49	8.36	6.56	2.50
C. individual level - OLS results						
Assimilation index	4.592***	4.709***	5.057***	5.558***	5.482***	5.617***
	(0.950)	(1.016)	(1.145)	(1.385)	(1.425)	(1.520)
D. individual level - IV results						
Assimilation index	6.968***	7.127***	9.082***	13.525	14.612	13.618*
	(2.173)	(2.619)	(3.378)	(13.216)	(59.190)	(7.603)
First stage:						
Share Braceros	-9.395***	-8.996**	-7.742**	-3.819	-3.292	-4.508*
	(3.613)	(3.688)	(3.577)	(3.557)	(2.397)	(2.382)
F-Statistic	9.61	8.67	6.59	3.10	3.35	4.29
Network controls (1990)						
Nr of Mexicans	linear	quadratic	quartic	none	linear	quartic
Average years of schooling	no	no	no	yes	yes	yes
Share women	no	no	no	yes	yes	yes
Share married	no	no	no	yes	yes	yes
Employment rate	no	no	no	yes	yes	yes

Note: This table displays the estimation results with additional controls for characteristics of the Mexican community in 1990. In all regressions, the dependent variable is the difference between the US wage and the counterfactual wage in Mexico. Panels A and B display results at the CONSPUMA-level, while Panels C and D display results at the individual level. The controls are the same as in the baseline regressions presented in Section 5. Standard errors in Panels A and B have been computed with a standard non-parametric bootstrap with 1000 replications. Standard errors have been computed with a block bootstrap with 1000 replications. Significance levels: * p<0.1, ** p<0.05, *** p<0.01.

In Table 6, we estimate the same models as before, but use as dependent variable the log US wage. The results shown here correspond to those of an augmented Mincer equation. The point estimates lie between 0.004 and 0.008, which means that a one-point increase in the assimilation index raises an immigrant's wage in the US by approximately 0.4-0.8%.

Assimilation index based on men only The calculation of the assimilation index was based on the observable characteristics of both Mexican men and women, whereas the empirical results shown in Table 2 are based on Mexican men only. It could be the case that the assimilation of men only as opposed to all Mexicans is the relevant determinant of the success of migrants. To test for this possibility, we calculate the assimilation index based on Mexican men only, and otherwise run the same regressions as before. The results, shown in Appendix D.3, are similar to those based on the assimilation of both.

The impact on women In the main analysis, we exclusively focus on men, mainly because men typically have higher labor force participation rates as well as inelastic labor supply. In Appendix D.4, we estimate the effects of network assimilation on the wage difference of women. The results are less clear-cut than for men. While we find large positive and statistically significant estimates in the IV regressions, we find very small effects once we account for the extensive margin of labor supply. This suggests that the quality of networks affects women through labor force participation and employment rather than wages.

Robustness to leaving out large Mexican communities The descriptive statistics in Table 1 show that the size of Mexican communities varies considerably, with many small and some very large communities. One concern could be that our results are driven by large Mexican communities. In Appendix D.5, we perform a robustness check in which we drop all Mexican communities larger than 200,000 from the sample. That way, 1.9% of all CONSPUMAs with Mexican communities and 19% of all individual-level observations are dropped. The results are robust to the exclusion of these communities. In the aggregate regressions, we lose statistical precision, whereas the estimates from individual-level regressions are statistically significant and very similar to the baseline results reported in Table 2.

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.

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.

The central contribution of this paper is its focus on the quality of networks, and its importance for the outcomes of migrants. While most of 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. It therefore complements earlier evidence by Massey & Denton (1985) and Hatton & Leigh (2011), among others, who suggest that immigrant groups, as they assimilate economically and culturally, become more accepted by the native population.

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**(2), 194–227.

Acemoglu, Daron, Dahleh, Munther A., Lobel, Ilan, & Ozdaglar, Asuman. 2011. Bayesian Learning in Social Networks. *Review of Economic Studies*, **78**(4), 1201–1236.

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. forthcoming. Migration and Information Flows: Evidence from a Field Experiment among Immigrants in Ireland. *World Bank Economic Review.*

Bauer, Thomas, Epstein, Gil S., & Gang, Ira N. 2005. Enclaves, Language, and the Location Choice of Migrants. *Journal of Population Economics*, **18**(4), 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. 2013. Living and Working in Ethnic Enclaves: Language Proficiency of Immigrants in U.S. Metropolitan Areas. *Papers in Regional Science*, **92**(2), 305–328.

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

Beine, Michel, Docquier, Frédéric, & Özden, Çağlar. 2015. Dissecting Network Externalities in International Migration. *Journal of Demographic Economics*, **81**, 379–408.

Biavaschi, Costanza. 2016. Recovering the Counterfactual Wage Distribution with Selective Return Migration. *Labour Economics*, **38**(1), 59–80.

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. 1985. Assimilation, Changes in Cohort Quality and the Earnings of Immigrants. *Journal of Labor Economics*, **3**(4), 463–489.

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.

Bound, John, Jaeger, David A., & Baker, Regina M. 1995. Problems with Instrumental Variables Estimation when the Correlation Between the Instruments and the Endogenous

Explanatory Variables is Weak. *Journal of the American Statistical Association*, **90**(430), 443–450.

Brücker, Herbert, Kroh, Martin, Bartsch, Simone, Goebel, Jan, Kühne, Simon, Liebau, Elisabeth, Trübswetter, Parvati, Tucci, Ingrid, & ., Jürgen Schupp. 2014. The new IAB-SOEP Migration Sample: An Introduction into the Methodology and the Contents. *SOEP Survey Papers, Series C*, **216**.

Calavita, Kitty. 1992. *Inside the State: The Bracero Program, Immigration, and the I.N.S.* Routledge.

Cameron, A. Colin, & Miller, Douglas L. 2015. A Practitioner's Guide to Cluster-Robust Inference. *Journal of Human Resources*, **50**(2), 317–372.

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), 239–281.

Chiswick, Barry R. 1978. The Effect of Americanization on the Earnings of Foreign-born Men. *The Journal of Political Economy*, **86**(5), 897–921.

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.

Davidson, Russell, & MacKinnon, James G. 2006. Bootstrap Methods in Econometrics. *Pages 812–838 of:* Mills, Terence C., & Patterson, Kerry (eds), *Palgrave Handbook of Econometrics*, vol. 1. Palgrave Macmillan.

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.

Garip, Filiz, & Asad, Asad L. 2016. Network Effects in Mexico-U.S. Migration: Disentangling the Underlying Social Mechanisms. *American Behavioral Scientist*, **60**(10), 1168–1193.

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.

Harles, John C. 1997. Integration before Assimilation: Immigration, Multiculturalism and the Canadian Polity. *Canadian Journal of Political Science*, **30**(4), 711–736.

Hatton, Timothy J., & Leigh, Andrew. 2011. Immigrants Assimilate as Groups, not just as Individuals. *Journal of Population Economics*, **24**(2), 389–419.

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.

Lewis, Jeffrey B., & Linzer, Drew A. 2005. Estimating Regression Models in Which the Dependent Variable Is Based on Estimates. *Political Analysis*, **13**(4), 345–364.

Massey, Douglas S., & Denton, Nancy A. 1985. Spatial Assimilation as a Socioeconomic Outcome. *American Sociological Review*, **50**(1), 94–106.

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.

Massey, Douglas S., Goldring, Luin, & Durand, Jorge. 1994. Continuities in Transnational Migration: An Analysis of Nineteen Mexican Communities. *American Journal of Sociology*, **99**(6), 1492–1533.

McElroy, Robert C., & Gavett, Earle E. 1965. Termination of the Bracero Program: Sofe Effects of Farm Labor and Housing Needs. *Agricultural Economic Report, US Department of Agriculture.*

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), 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*, **102**(C), 116–127.

McPherson, Miller, Smith-Lovin, Lynn, & Cook, James M. 2001. Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology*, **27**(1), 415–444.

Mincer, Jacob. 1974. *Schooling, Experience and Earnings.* NBER.

Moreira, Marcelo J. 2003. A Conditional Likelihood Ratio Test for Structural Models. *Econometrica*, **71**(4), 1027–1048.

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**(1), 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**(7), 1160–1186.

Polachek, Solomon W. 2008. Earnings Over the Life Cycle: The Mincer Earnings Function and Its Applications. *Foundations and Trends in Microeconomics*, **4**(3), 165–272.

Staiger, Douglas, & Stock, James H. 1997. Instrumental Variables Regression with Weak Instruments. *Econometrica*, **65**(3), 557–586.

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

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.

Online Appendices

(Not for publication)

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 3.3 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 less than 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).[23]

C Residuals from wage regression

A key assumption underlying the prediction of counterfactual wages is that the error terms are independently and identically distributed (i.i.d). While we cannot directly assess the validity of this assumption for the population, we can make an assessment based on the residuals of the regression. Figure 6 displays the distribution of the residuals from the wage regression (Equation (5)), along with a normal distribution as benchmark. While a Shapiro-Wilk test for normality of the residuals rejects normality at all conventional significance levels, the residuals appear to be symmetric, with a higher number close to zero than would be predicted by a normal distribution. We see this as evidence in support of the i.i.d. assumption.

[23] 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).

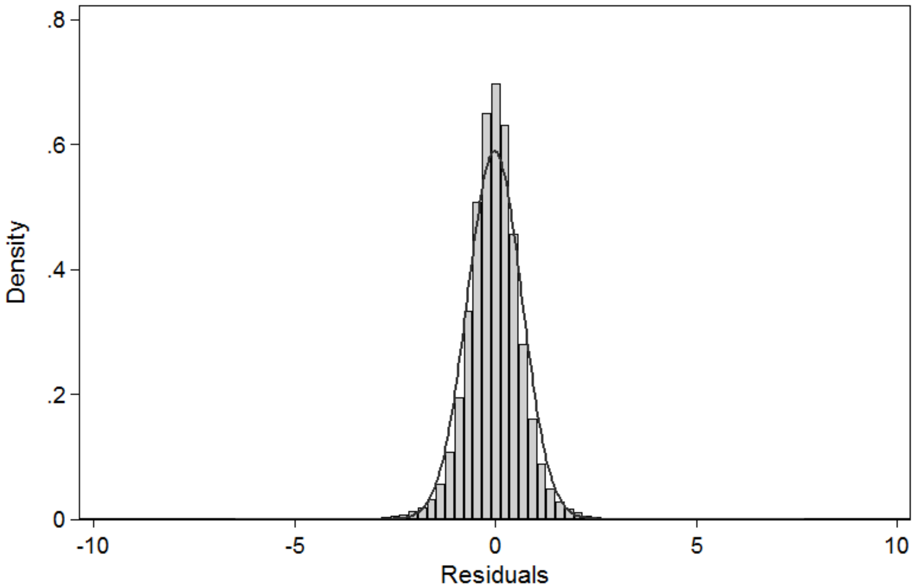

Figure 6: Residuals from a Mincer wage regression in Mexico, 2000 census

Note: This graph displays the residuals from a Mincer wage regression on education dummies, age, age squared, as well as interactions of the education dummies with age and age squared for all Mexicans with a monthly wage greater than zero. The solid line superimposes a normal distribution.

D Robustness checks

D.1 Regressions with US wages as dependent variable

Table 5 displays the results from regressions with US wages as dependent variable. As explained in Section 5.3, using the level of US wages instead of the wage difference between the US and Mexico, because Mexican wages have been predicted based on similar characteristics as those included in the vector of controls in the regression. For an interpretation of the results, please see Section 5.3.

Table 5: Robustness check: US wages as dependent variable

Dependent variable: monthly wage in the US (in USD)						
A. CONSPUMA level						
	OLS	OLS	OLS	IV	IV	IV
	(1)	(2)	(3)	(4)	(5)	(6)
Assimilation index	7.001***	7.666***	5.881***	6.009**	4.879	-3.555
	(1.677)	(0.953)	(1.775)	(2.624)	(5.457)	(34.094)
Weighted by size	No	Yes	No	No	Yes	No
Control: nr of Mexicans	No	No	Yes	No	No	Yes
First stage:						
Share Braceros				-16.110***	-7.835***	-2.536
				(4.037)	(2.890)	(2.466)
F-Statistic				12.53	7.35	2.49
N	202	202	202	202	202	202
B. individual level						
	OLS	IV	IV	OLS	IV	IV
	(1)	(2)	(3)	(4)	(5)	(6)
Assimilation index	5.014***	6.498***	6.044***	5.023***	8.790***	8.529***
	(0.899)	(1.363)	(1.222)	(1.144)	(3.401)	(3.020)
First stage:						
Share Braceros		-14.804***	-14.643***		-7.742**	-7.683**
		(5.402)	(5.506)		(3.577)	(3.607)
Control: nr of Mexicans	No	No	No	Yes	Yes	Yes
Exclude if wage US= 0	No	No	Yes	No	No	Yes
F-Statistic		13.94	12.84		6.59	6.08
N	20131	20131	15082	20131	20131	15082

Note: This table displays regression results with US wages as dependent variable. Controls are the same as in the baseline regressions reported in Section 5. Standard errors in Panel A have been bootstrapped with 1000 replications, with the exception of those in Columns (2) and (5), which are heteroskedasticity-robust. Standard errors in Panel B have been computed with a block bootstrap with 1000 replications. Significance levels: * p<0.1, ** p<0.05, *** p<0.01.

While in Table 5 the dependent variable is the US wage in levels, in Table 6 the dependent variable is the US wage in logs. Therefore, the underlying regression corresponds to a Mincer equation augmented by the network variables. For further interpretation, see Section 5.3.

Table 6: Robustness check: log US wages as dependent variable

Dependent variable: log monthly wage in the US (in USD)

A. CONSPUMA level						
	OLS	OLS	OLS	IV	IV	IV
	(1)	(2)	(3)	(4)	(5)	(6)
Assimilation index	0.004***	0.007***	0.005***	0.004*	0.004	0.007
	(0.001)	(0.001)	(0.002)	(0.002)	(0.005)	(0.056)
Weighted by size	No	Yes	No	No	Yes	No
Control: nr of Mexicans	No	No	Yes	No	No	Yes
First stage:						
Share Braceros				-16.110***	-7.835***	-2.536
				(4.037)	(2.890)	(2.466)
F-Statistic				12.53	7.35	2.49
N	202	202	202	202	202	202

B. individual level						
	OLS	IV	IV	OLS	IV	IV
	(1)	(2)	(3)	(4)	(5)	(6)
Assimilation index	0.006**	0.014***	0.005***	0.003	0.019*	0.008***
	(0.002)	(0.004)	(0.001)	(0.003)	(0.010)	(0.002)
First stage:						
Share Braceros		-14.804***	-14.643***		-7.742**	-7.683**
		(5.402)	(5.506)		(3.577)	(3.607)
Control: nr of Mexicans	No	No	No	Yes	Yes	Yes
Exclude if wage US= 0	No	No	Yes	No	No	Yes
F-Statistic		13.94	12.84		6.59	6.08
N	20131	20131	15082	20131	20131	15082

Note: This table displays regression results with log US wages as dependent variable. Controls are the same as in the baseline regressions reported in Section 5. Standard errors in Panel A have been bootstrapped with 1000 replications, with the exception of those in Columns (2) and (5), which are heteroskedasticity-robust. Standard errors in Panel B have been computed with a block bootstrap with 1000 replications. Significance levels: * p<0.1, ** p<0.05, *** p<0.01.

D.2 Different definitions of the assimilation index

The assimilation index is a statistical index that measures the similarity between Mexicans and Americans in an area with respect to several variables. The set of variables includes personal characteristics such as gender or age as well as variables related to economic well-being, such as wages or employment. Due to the limited information in the census and the nature of the index, it does not include variables that would be seen as strong indicators for assimilation in disciplines other than economics. Examples are English language skills, or the extent to which Mexicans share American values. Such information makes the construction of a statistical index more difficult because the index is based on the predictability of a person being Mexican. If one randomly picks people and one cannot tell based on observable characteristics whether the person is Mexican or American, this means that both groups are similar. If, on the other hand, one includes a variable such as English skills, this variable alone predicts very well if someone is Mexican or American and the index would indicate a low degree of assimilation.

These caveats notwithstanding, we re-calculate the assimilation index and include English language skills. The results are displayed in Table 7. The estimates in Columns (1)-(3), i.e. those without controls for the number of Mexicans, are similar to the baseline estimates. Once we control for the number of Mexicans in Columns (4)-(6), we obtain larger estimates in the IV regressions, which is most likely due to the weak first stage correlation.

In Table 8, we present the results of a further robustness check, whereby the assimilation index is purely based on personal characteristics while measures of economic well-being such as wages, employment, or the occupational earnings score, are omitted. The estimates are similar to the baseline estimates in Table 2. This could either suggest that the variation in the index is largely driven by personal characteristics, or that within CONSPUMAs, Americans and Mexicans are as similar in terms of personal characteristics as in terms of variables that indicate economic well-being.

Table 7: Robustness check: assimilation index includes English ability

Dependent variable: wage difference US wage - wage in Mexico

A. CONSPUMA level						
	OLS (1)	OLS (2)	OLS (3)	IV (4)	IV (5)	IV (6)
Assimilation index	3.604**	6.509***	1.754	7.733***	10.705*	9.345
	(1.784)	(1.177)	(2.554)	(2.945)	(6.214)	(526.427)
Weighted by size	No	Yes	No	No	Yes	No
Control: nr of Mexicans	No	No	Yes	No	No	Yes
First stage:						
Share Braceros				-13.423***	-5.045**	0.745
				(3.802)	(2.241)	(2.193)
F-Statistic				9.56	5.07	0.46
N	202	202	202	202	202	202

B. individual level						
	OLS (1)	IV (2)	IV (3)	OLS (4)	IV (5)	IV (6)
Assimilation index	4.920***	8.781***	8.163***	4.487***	14.333*	13.794**
	(0.993)	(1.853)	(1.604)	(1.194)	(8.028)	(6.225)
First stage:						
Share Braceros		-10.172**	-10.132**		-3.366	-3.429
		(4.259)	(4.340)		(2.616)	(2.641)
Control: nr of Mexicans	No	No	No	Yes	Yes	Yes
Exclude if wage US= 0	No	No	Yes	No	No	Yes
F-Statistic		11.23	10.53		4.02	3.82
N	20131	20131	15082	20131	20131	15082

Note: This table displays regression results with log US wages as dependent variable. Controls are the same as in the baseline regressions reported in Section 5. Standard errors in Panel A have been bootstrapped with 1000 replications, with the exception of those in Columns (2) and (5), which are heteroskedasticity-robust. Standard errors in Panel B have been computed with a block bootstrap with 1000 replications. Significance levels: * $p<0.1$, ** $p<0.05$, *** $p<0.01$.

Table 8: Robustness check: assimilation index excludes economic variables

Dependent variable: wage difference US wage - wage in Mexico

A. CONSPUMA level						
	OLS	OLS	OLS	IV	IV	IV
	(1)	(2)	(3)	(4)	(5)	(6)
Assimilation index	4.425***	5.767***	3.156	5.299***	5.159	2.638
	(1.422)	(1.010)	(2.139)	(2.026)	(3.520)	(8.067)
Weighted by size	No	Yes	No	No	Yes	No
Control: nr of Mexicans	No	No	Yes	No	No	Yes
First stage:						
Share Braceros				-19.309***	-10.468***	-4.779**
				(4.006)	(2.964)	(2.354)
F-Statistic				18.20	12.47	5.41
N	202	202	202	202	202	202

B. individual level						
	OLS	IV	IV	OLS	IV	IV
	(1)	(2)	(3)	(4)	(5)	(6)
Assimilation index	4.464***	5.864***	5.482***	4.814***	8.720**	8.506***
	(0.808)	(1.247)	(1.138)	(1.162)	(3.455)	(3.095)
First stage:						
Share Braceros		-16.852***	-16.755***		-7.871***	-7.877***
		(5.442)	(5.542)		(2.811)	(2.831)
Control: nr of Mexicans	No	No	No	Yes	Yes	Yes
Exclude if wage US= 0	No	No	Yes	No	No	Yes
F-Statistic		18.64	17.18		11.02	10.19
N	20131	20131	15082	20131	20131	15082

Note: This table displays regression results with log US wages as dependent variable. Controls are the same as in the baseline regressions reported in Section 5. Standard errors in Panel A have been bootstrapped with 1000 replications, with the exception of those in Columns (2) and (5), which are heteroskedasticity-robust. Standard errors in Panel B have been computed with a block bootstrap with 1000 replications. Significance levels: * p<0.1, ** p<0.05, *** p<0.01.

D.3 Assimilation index for men only

For our main analysis, we calculate the assimilation index based on the similarity of Mexicans and Americans in an area, with both groups comprising both men and women. But given that we only look at the impact of network quality on the outcomes of men, it might be that the assimilation of Mexican men may be more important, especially so because Mexican men have high labor force participation rates and may pass on job information to male newcomers. To test whether the assimilation of men is more relevant than that of men and women combined, we compute the index based on men only, and run the same regressions as before. The correlation between both indices is very high ($\rho = 0.988$), but we are not able to compute the index for nine small CONSPUMAs in which the number of observations is too small to estimate a probit model. The results, displayed in Table 9 are similar to the baseline results.

Table 9: Regression results based on assimilation of men only

Dependent variable: wage difference USA - Mexico

A. CONSPUMA level						
	OLS	OLS	OLS	IV	IV	IV
	(1)	(2)	(3)	(4)	(5)	(6)
Assimilation index	4.40**	5.30***	3.67	5.71*	6.49**	7.56
	(1.80)	(0.89)	(2.44)	(3.46)	(2.96)	(13.26)
Weighted by size	No	Yes	No	No	Yes	No
Control: nr of Mexicans	No	No	Yes	No	No	Yes
First stage:						
Share Braceros				-19.80***	-9.71***	-6.68**
				(2.37)	(3.51)	(2.65)
F-Statistic				69.74	7.67	6.37
N	193	193	193	193	193	193

B. individual level						
	OLS	IV	IV	OLS	IV	IV
	(1)	(2)	(3)	(4)	(5)	(6)
Assimilation index	4.14***	5.48***	5.16***	4.35***	7.94***	7.76***
	(0.76)	(1.07)	(0.97)	(0.93)	(2.36)	(2.14)
First stage:						
Share Braceros		-20.67***	-20.81***		-11.57***	-11.60**
		(5.44)	(5.70)		(4.38)	(4.57)
Control: nr of Mexicans	No	No	No	Yes	Yes	Yes
Exclude if wage US= 0	No	No	Yes	No	No	Yes
F-Statistic		14.47	13.33		6.99	6.43
N	20098	20098	15057	20098	20098	15057

Note: This table displays the results of OLS- and IV-regressions of difference in monthly wages on the assimilation index. The assimilation index in this robustness check has been computed for Mexican men only. In Panel A), the unit of observation is a CONSPUMA. Standard errors in parentheses are robust to heteroskedasticity. In Panel B), the unit of observation is an individual. All regressions control for the average wage of US natives. Those in Panel B also include individual-level controls. Standard errors in Panel A have been bootstrapped with 1000 replications, with the exception of those in Columns (2) and (5), which are heteroskedasticity-robust. Standard errors in Panel B have been computed with a block bootstrap with 1000 replications. Significance levels: * p<0.1, ** p<0.05, *** p<0.01.

D.4 Results for women

In our baseline analysis, we focus on men only, because men's labor supply is typically inelastic, such that the estimate is not confounded by selection into employment. To complete the picture, we provide the results for women in Table 10 based on individual-level regressions. While we find fairly large effects on the wage difference when zero wages are included, the effects diminish once we only focus on women with non-zero wages. This suggests that for women, networks mainly work through the extensive margin of the labor market. Women with access to better networks are more likely to work, be it because they have a higher labor supply, or because they find employment more easily through the network.

Table 10: Effects for women, individual-level regressions

Dependent variable: wage difference USA - Mexico						
	OLS	IV	IV	OLS	IV	IV
	(1)	(2)	(3)	(4)	(5)	(6)
Assimilation index	3.596***	9.415***	3.062	0.205	9.954*	2.344
	(0.992)	(1.887)	(2.300)	(0.907)	(5.427)	(4.770)
First stage:						
Share Braceros		-14.026***	-16.396**		-8.956**	-9.488*
		(4.258)	(7.491)		(3.711)	(5.693)
Control: nr of Mexicans	No	No	No	Yes	Yes	Yes
Exclude if wage US= 0	No	No	Yes	No	No	Yes
F-Statistic		9.66	9.66		5.28	4.95
N	14031	14031	5175	14031	14031	5175

Note: This table displays OLS- and IV-regression results at the individual level for recent immigrant women only. 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 fourth-order polynomial in the number of Mexicans. Bootstrapped standard errors, clustered at the CONSPUMA-level, with 1000 replications, are displayed in parentheses. Significance levels: * p<0.1, ** p<0.05, *** p<0.01.

D.5 Leave out large CONSPUMAs

As shown in Table 1, the size of Mexican communities varies considerably across CONSPUMAs from less than 500 to over 500,000. One concern may be that the estimates are driven by a small number of very large communities. In a robustness check, we drop from the sample the four largest communities, which have over 200,000 inhabitants and are clear outliers in the distribution of community sizes. The results, displayed in Table 11, show that the baseline estimates are not driven by those outliers.

Table 11: Results when large CONSPUMAs are left out

Dependent variable: wage difference USA - Mexico						
A. CONSPUMA level						
	OLS	OLS	OLS	IV	IV	IV
	(1)	(2)	(3)	(4)	(5)	(6)
Assimilation index	5.071***	4.835***	3.740**	6.409**	6.701	8.125
	(1.481)	(1.149)	(1.875)	(2.495)	(5.019)	(10.592)
Weighted by size	No	Yes	No	No	Yes	No
Control: nr of Mexicans	No	No	Yes	No	No	Yes
First stage:						
Share Braceros				-12.842***	-6.225**	-5.299***
				(2.705)	(2.519)	(1.716)
F-Statistic				12.07	6.11	6.25
N	198	198	198	198	198	198
B. individual level						
	OLS	IV	IV	OLS	IV	IV
	(1)	(2)	(3)	(4)	(5)	(6)
Assimilation index	4.176***	6.426***	5.957***	4.966***	10.333***	10.991***
	(0.995)	(1.801)	(1.571)	(1.235)	(3.682)	(3.156)
First stage:						
Share Braceros		-12.436***	-12.578***		-7.822**	-7.983**
		(4.444)	(4.562)		(3.247)	(3.329)
Control: nr of Mexicans	No	No	No	Yes	Yes	Yes
Exclude if wage US= 0	No	No	Yes	No	No	Yes
F-Statistic		10.28	9.55		8.19	7.72
N	16142	16142	12165	16142	16142	12165

Note: This table displays the results of OLS- and IV-regressions of difference in monthly wages on the assimilation index. CONSPUMAs with more than 200,000 Mexicans have been dropped from the sample. In Panel A), the unit of observation is a CONSPUMA. Standard errors in parentheses are robust to heteroskedasticity. In Panel B), the unit of observation is an individual. Standard errors in Panel A have been bootstrapped with 1000 replications, with the exception of those in Columns (2) and (5), which are heteroskedasticity-robust. Standard errors in Panel B have been computed with a block bootstrap with 1000 replications. Significance levels: * p<0.1, ** p<0.05, *** p<0.01.

E Inference with weak instruments

Among the results displayed in Section 5, several instrumental variable estimations are based on instrumental variables with a weak first stage. Weak instruments pose two problems for estimation and inference. First, estimates obtained through a two-stage-least-squares (TSLS) method are biased towards the OLS estimates, mainly due to the unfavorable small sample properties of TSLS (Staiger & Stock, 1997). Second, the TSLS method under-estimates the standard errors of the second-stage coefficient of interest, leading to an under-rejection of the null hypothesis of a zero effect.

To assess the severity of the small sample bias, we apply a limited information maximum likelihood estimator (LIML), which has more favorable asymptotic properties and is less prone to bias compared to a TSLS estimator. Columns (1)-(3) of Table 12 display the results for IV regressions with various controls. It appears that the small sample bias of TSLS is problematic in the aggregate regressions (Panel A) once we include controls for network characteristics. While the coefficient is precisely estimated in a regression without network controls, the estimates become imprecise once the controls are added. The individual-level regressions (Panel B), in contrast, do not seem prone to small sample bias. The point estimates are similar to those in Table 2 and the standard errors are, if anything, smaller than in the TSLS estimations, despite being clustered at the CONSPUMA level.

In Columns (4)-(6) we use a conventional TSLS estimator, but construct 95%-confidence intervals based on a conditional likelihood ratio (CLR) test (Moreira, 2003), which removes the downward bias in the estimated standard errors. In the aggregate regressions, the wide 95%-confidence intervals based on the CLR test point to imprecise estimates. Matters are different when we consider the individual-level regression, where the 95%-confidence intervals are wider than those around the estimates in Table 2, but all three coefficients are statistically significant at the 1%-level.

These results warrant some caution for the interpretation of the estimates based on aggregate regressions. Weak instruments introduce a considerable small sample bias and lead to significant under-rejection when conventional t-tests are used. However, the results in Panel B suggest that weak instruments are less of a reason for concern in the individual-level regressions. The small sample bias appears minor and the standard errors — while larger than before — are small enough for us to conclude that the estimates are statistically significant.

Table 12: IV estimation results with weak instrument corrections

	Dependent variable: wage difference USA - Mexico					
A. CONSPUMA level						
	LIML	LIML	LIML	CLR	CLR	CLR
	(1)	(2)	(3)	(4)	(5)	(6)
Assimilation index	6.286***	3.555	15.744	6.286	3.555	15.744
	[1.71, 10.86]	[-17.40, 24.52]	[-13.80, 45.29]	[-3.47, 16.12]	[-67.21, 68.42]	[-54.30, 118.81]
N	202	202	202	202	202	202
B. individual level						
Assimilation index	6.626***	9.082***	13.618**	6.966***	9.082***	13.618***
	[4.08, 9.17]	[3.57, 14.60]	[2.43, 24.80]	[4.81, 9.12]	[4.83, 13.85]	[5.64, 21.60]
N	20,131	20,131	20,131	20,131	20,131	20,131
Network controls (1990)						
Nr of Mexicans	no	quartic	quartic	no	quartic	quartic
Average years of schooling	no	no	yes	no	no	yes
Share women	no	no	yes	no	no	yes
Share married	no	no	yes	no	no	yes
Employment rate	no	no	yes	no	no	yes

Note: This table displays estimates based on instrumental variable regressions. Columns (1)-(3) report point estimates and 95%-confidence intervals based on a LIML estimator. In the individual-level regressions, standard errors are clustered at the CONSMPUMA-level. Columns (4)-(6) report point estimates from a TSLS estimator and confidence intervals based on a CLR Test (Moreira, 2003). Significance levels: * p<0.1, ** p<0.05, *** p<0.01. In Columns (4)-(6), the significance stars refer to the p-values of the CLR test.

F Counterfactual Wages

For our baseline analysis, we calculate the counterfactual wages of Mexican immigrants in the U.S. — the wages Mexican migrants would have if they were living in Mexico — based on the full Mexican census, i.e. based on all Mexicans that have not migrated. There are concerns, though, that emigrants are a self-selected along many dimensions, such that their actual counterfactual wages would be different from those predicted based on the entire Mexican census.

As an alternative, we propose to predict counterfactual wages based on a subsample of the population that is as similar as possible to the migrants. To find a more similar comparison group, we use two approaches: first, based on a wide range of observable characteristics, we perform a propensity score matching that provides us with a sample of Mexicans that are currently living in Mexico and have the same characteristics as Mexicans in the US. Second, we use a sample of internal migrants, which are presumably more similar to international migrants compared to people who stay in their local area. Moreover, to account for selection into employment, we predict wages based on a Heckman two-step selection model.

Table 13 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 13: Counterfactual Wages: Correlations

	2000			
	Baseline	PSM	Internal	Heck
Baseline	1			
PSM	0.98	1		
Internal	0.90	0.92	1	
Heckman	0.95	0.94	0.80	1

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