ANNUAL REVIEWS

Annual Review of Economics

Social Networks and Migration

Kaivan Munshi

Department of Economics, Yale University, New Haven, Connecticut 06520, USA;
email: kaivan.munshi@yale.edu

Annu. Rev. Econ. 2020. 12:503–24

First published as a Review in Advance on May 8, 2020

The *Annual Review of Economics* is online at economics.annualreviews.org

https://doi.org/10.1146/annurev-economics-082019-031419

JEL codes: D02, J61, J62, L14, O12, O17

Keywords

community networks, migration, Roy model, assimilation, mutual insurance

Abstract

The frictions that restrict migration are among the largest sources of inefficiency in the global economy. The first step in designing policies to address these frictions is to understand the fundamental forces that drive migration. However, the Roy model—the workhorse model of migration in economics—does a poor job of explaining many important features of this phenomenon. This limitation can be rectified by adding migrant networks to the Roy model. A rich qualitative literature in the social sciences has documented the role played by social networks in supporting migrants in their new locations. Economists have advanced this literature by identifying and quantifying the contribution of these networks to migration. Although much progress has been made over the past two decades, important gaps in the literature remain: Migrant assimilation has received little theoretical or empirical attention, and a richer characterization of the social interactions that support these networks is needed to tie research on migration to the economic literature on networks.

1. INTRODUCTION

Average earnings in a given occupation differ dramatically across countries. For example, a carpenter's monthly wage in 1995 was $42 in India, $125 in Mexico, and $2,299 in the United States (Rosenzweig 2010). Large wage gaps persist even after adjusting for differences in skills, and a natural question to ask is why workers do not move to close these gaps. Some of the frictions that restrict the mobility of workers are legal or political in nature. However, even in environments where individuals can seemingly move at will, they often fail to take advantage of what appear to be large economic opportunities. The failure of entrepreneurs and workers to move to locations where they are most productive is arguably one of the largest sources of inefficiency in the global economy. There is evidently a role for policy in reducing these inefficiencies, but to design such policies it is necessary to understand the fundamental forces that drive migration.

The canonical characterization of migration in economics is based on the Roy model, which specifies that the migration decision is based on economic payoffs at the origin and destination, together with moving costs. While the Roy model has the advantage of simplicity, it cannot explain a number of stylized facts that are associated with migration: (*a*) Migrants often fail to move to destinations where they would appear to realize the largest economic gain, and (*b*) they move, instead, to locations (and sectors) where others from their origin (i.e., birthplace) moved before. The Roy model also does a poor job of explaining migrant selection on ability, which is typically measured by education, both in the cross-section across origins and over time from a given origin. It is possible to augment the Roy model along multiple dimensions to explain the stylized facts. As shown below, a more parsimonious approach that can also explain these facts adds destination networks to the Roy model.

A rich social science literature, summarized by Munshi (2014), describes how networks organized around predetermined social groups, defined by kinship or origin location, have historically supported and continue to support the movement of their members. In recent years, economists have advanced this literature by identifying and quantifying the role played by social networks in determining migration. Early studies in economics (e.g., Munshi 2003, Beaman 2012) exploited special research settings to identify network effects. While this was an important first step, there are many settings in which the quasi-random variation in the migrant networks that is needed to establish their causal effect on migration is unavailable. The central thesis of this article is that it is nevertheless possible to credibly establish that migrant networks are active using a three-step approach. First, a theoretical model of migration that incorporates networks must be developed; a convenient way to do this is to add destination networks to the Roy model. Second, an exogenous source of variation in origin characteristics must be identified. The model can then be used to derive predictions for migrant flows, migrant selection, and destination outcomes with respect to this origin characteristic, both in the cross-section and over time. If the predictions are sufficiently rich, they can then be used to validate the model by ruling out alternative non-network explanations. Third, evidence that is directly indicative of the underlying network-based mechanism must be provided. I describe two examples of this approach, one applied to Indian data (Munshi 2011) and the other to Chinese data (Dai et al. 2019), below. Incorporating networks in analyses of migration is only justified if they contribute substantially to payoffs at the destination and, hence, to migration flows. The advantage of the approach that I am advocating is that once the model has been validated, its structural parameters can be estimated. Counterfactual simulations that shut down the network can subsequently be used to quantify its impact on migration.

The emerging literature on networks and migration in economics has largely focused on destination networks. Recently, however, there has been some research that explores the connection between origin networks, specifically providing mutual insurance to their members, and migration (Munshi & Rosenzweig 2016, Morten 2019). Once again, the Roy model, suitably augmented to incorporate networks at the origin, can be used to analyze this phenomenon. And once again, the validated model can be used for counterfactual simulations and to quantify the contribution of the (origin-based) social networks to migration. The received evidence thus far is based on a small number of studies, and much more work is needed on the quantification dimension. That said, all of the evidence indicates that social networks, either at the origin or the destination, are important determinants of migration. Moreover, a recurring message from counterfactual policy analyses is that interventions that attempt to boost migration without taking account of the underlying networks can have substantial negative consequences.

Much progress has been made by economists over the past two decades in understanding how social networks affect migration. Nevertheless, major gaps in the literature remain. One important area that remains open for future research is migrant assimilation or, conversely, the failure to assimilate. The Roy model, suitably augmented to incorporate a network component, serves as the starting point for analyses of assimilation as well. However, a complete analysis of assimilation would need to go further and model the community identity that can sometimes lock migrants into particular occupations and locations. It would also need to account for the interactions between migrant communities and for those between specific communities and the native population; the network can no longer be treated as operating independently in these analyses. A major challenge with empirical analyses of migrant assimilation is that this process can extend over multiple generations. With the increased availability of administrative data for research, however, this challenge may not be unsurmountable.

A second potential direction for future research would be to enrich the social interactions that provide information and support cooperation in migrant networks. The implicit assumption in most existing models of networks and migration is that individuals match randomly within their communities and that interactions within a prespecified social group alone are relevant for economic outcomes. The latter assumption must be relaxed once we allow for interactions between networks, as discussed above. Moreover, interactions may not be random even within networks, which implies that statistics such as network centrality will also be relevant in determining outcomes (and accompanying migration). Current analyses, based on network size or on statistics that are broadly associated with social connectedness in the population from which the networks are drawn, must work hard to establish a causal role for the networks. The additional intra-network heterogeneity is even more difficult to identify, and innovative research designs will thus need to be developed for this purpose; without validating the augmented model, any attempts to quantify its contribution to migration will have limited value.

2. THE ROY MODEL

The workhorse model of migration in economics, first developed by Roy (1951) and subsequently popularized by Borjas (1987), specifies that the individual's location choice depends on the payoff at the origin, the payoff at the destination, and the cost of moving. These will, in turn, depend on the individual's ability or, equivalently, their education.

Denote the individual's ability by ω. Let ω^σ be the payoff at the origin and $A_0\omega$ the payoff at the destination. Once ability heterogeneity is introduced in the payoffs, moving costs can be ignored when deriving implications for selection on ability. In particular, if $\sigma < 1$ (or $\sigma > 1$), there will be an ability threshold above (or below) which individuals choose to migrate. Moreover, an increase in the payoff gap between origin and destination, generated by an exogenous improvement in the destination economy (measured by A_0), will shift down the ability threshold and unambiguously increase migration.

Although the Roy model has the virtue of simplicity, it does a poor job of explaining observed patterns of migration:

1. A large number of studies, typically using OECD data, have tested the prediction that bilateral migration flows should be increasing in the origin–destination wage gap (e.g., Beine et al. 2011, Bertoli & Fernandez-Huertas Moraga 2012, Docquier et al. 2014). Although the wage gap is positively associated with migrant flows between countries, it accounts for only a small part of the variation in these flows. What matters more for the origin–destination flows is the existing stock of migrants from the origin at the destination.
2. The nature of migrant selection from Mexico to the United States, a topic of substantial policy and research interest, remains unresolved. Depending on the context, migrants are found to be drawn from very different parts of the ability/education distribution (e.g., Chiquiar & Hanson 2005; Cuecuecha 2005; Orrenius & Zavodny 2005; Ibarran & Lubotsky 2007; Mishra 2007; Fernandez-Huertas Moraga 2011, 2013). If the payoff function is the same for all origin locations, as specified above, then this finding cannot be reconciled with the Roy model.
3. McKenzie & Rapoport (2007) document an initial positive selection on education in the Mexican origin communities that they study, which is replaced by negative selection later in time. Historical and contemporary evidence shows that the nature of migrant selection to the United States (across all countries, not just Mexico) has also changed over time (Abramitzky & Boustan 2017). The Roy model, as specified above, cannot explain these dynamic patterns of migrant selection.

It is possible to augment the benchmark Roy model to reconcile it with the empirical facts listed above. For example, if moving costs are relevant, then migrants from a given origin will locate disproportionately at specific proximate destinations. The stock of migrants at a given destination will proxy for the moving cost in that case, if this variable is not fully accounted for in the estimating equation. Moreover, if destination conditions, measured by A_0, are improving over time, then initial positive selection on ability can subsequently be replaced by negative selection. If there is variation in the onset of this improvement across destinations favored by migrants from different origins, then the heterogeneity in migrant self-selection that has been documented in the cross-section can also be explained.

The benchmark Roy model must be extended along multiple dimensions to explain the empirical facts listed above. Alternatively, we will see that a single extension—the incorporation of destination networks—is sufficient to explain all of these facts. Apart from parsimony, the additional advantage of the network-based approach is that it is supported by a wealth of anecdotal evidence that origin community–based networks have historically supported, and continue to support, the occupational and spatial mobility of their members (see Munshi 2014 for a review of the literature).

3. THE ROY MODEL WITH DESTINATION NETWORKS

Migrants drawn from the same origin provide many forms of mutual economic support to each other at the destination. Being newcomers to the destination labor market, they often begin their careers in sectors of the economy that are characterized by short-term contracts and frequent job turnover. Without established credentials, it is difficult for employers to assess the ability of potential employees. These employees, in turn, must somehow learn about the jobs that become available. Under these circumstances, members of the network that are employed at a given point in time can provide referrals for unemployed members of their network. The incumbent workers have a reputation to maintain with their firms and thus have an incentive to refer able workers. Workers, once hired through the network, have an incentive to work diligently to avoid the social sanctions that they would face if they shirked instead.

Migration is associated with spatial mobility and, typically, with accompanying occupational mobility. Given that opportunities in the formal sector are limited, international migrants often end up being self-employed. Entrepreneurs in developing countries are also, for the most part, first- or second-generation migrants who move to centers of production to establish and then subsequently grow their businesses. These entrepreneurs, especially the first-generation entrepreneurs, will rely on each other for capital, connections to buyers and suppliers, and information about new technologies and markets. The provision of such help without immediate compensation requires a high level of commitment, which can once again be supported by the threat of social sanctions. It is thus not surprising that businesses in developing countries, and migrant businesses in advanced economies, are concentrated among a few communities based on common origin or kinship ties (Munshi 2014).

To add destination networks to the Roy model, we follow Dai et al. (2019) and let the payoff of a migrant with ability ω in period t be $A_0(1+h)^{n_{t-1}}\omega$, where n_{t-1} measures the number of individuals from the migrant's network who are active at the destination as of period $t-1$, and h is the help that each of them provides. Notice that the only difference between this specification of the destination payoff and the corresponding specification in the benchmark Roy model is the inclusion of the $(1+h)^{n_{t-1}}$ term. Letting θ' denote $\log(1+h)$, the preceding expression can be rewritten as

$$A_0 \exp(\theta' n_{t-1})\omega. \qquad 1.$$

This specification of the destination payoff, which applies to both migrant workers and entrepreneurs, has many features in common with the standard formulation of payoffs in endogenous growth and agglomeration models. The important difference is that the complementarities in help provided, and the accompanying increasing returns to network size, are restricted to individuals from the same origin. In agglomeration and endogenous growth models, n_{t-1} would be the total number of individuals or firms operating at a destination, regardless of their social origin. Ciccone & Hall (1996), for example, use the number of workers per square kilometer as a proxy for agglomeration effects in a given location. We will see below that this distinction can be used to disentangle network effects from agglomeration effects and, more generally, any common destination effect.

To derive migrant flows with the augmented Roy model, we retain the previous specification of the origin payoff, ω^σ, except that we now place the restriction that $\sigma \in (0, 1)$. Assuming that $\log\omega \sim U[0, 1]$, there exists a threshold ability $\underline{\omega}_t$ in each period t for which the individual with that ability level is indifferent between moving and staying:

$$\log \underline{\omega}_t = \frac{-1}{1-\sigma}\left[\log A_0 + \theta' n_{t-1}\right]. \qquad 2.$$

In each period, individuals with ability above the threshold migrate, while the remainder stay at the origin. If the ability distribution is the same in all cohorts and individuals are infinitely lived, then this implies that entry (and the stock of migrants) will be increasing over time. This cannot go on forever, of course, and hence the model only applies to the initial phase of the migration process for a given origin.

Munshi (2011) describes how the process of spatial (and occupational) migration commences; in general, a fortuitous confluence of circumstances is required to jump-start the network. Thus, different origin communities will start sending migrants at different points in time even when the underlying fundamentals that determine the payoffs at the origin and the destination are the same. The resulting heterogeneity in network dynamics can explain (*a*) why the stock of migrants from a given origin at a given destination, rather than the average wage differential between the two locations, is a stronger predictor of migrant flows; (*b*) why there is no clear pattern in the cross-section across origin communities with regard to migrant selection on ability; and (*c*) why positive selection is replaced by negative selection over time among migrants drawn from the same origin community. While the Roy model with destination networks can simultaneously explain the three stylized facts on migration listed above, non-network explanations incorporating origin and destination heterogeneity are also available. The discussion that follows describes how these potentially coexisting mechanisms can be disentangled empirically and how network effects can be identified more generally.

4. IDENTIFYING NETWORKS

The identification of network effects is a challenging statistical problem. To see why this is the case in the context of migration, we derive the relationship between migrant flows from a given origin to a given destination, e_t, and the lagged stock of migrants from that origin at that destination, n_{t-1}, from Equation 2. We obtain

$$e_t = L + \theta n_{t-1}, \qquad 3.$$

where $L \equiv 1 + \frac{\log A_0}{1-\sigma}$, $\theta \equiv \frac{\theta'}{1-\sigma}$.

The implicit assumption in the preceding equation is that conditions at the destination, measured by the A_0 parameter, are fixed over time. The obvious threat to the identification of network effects with this specification is that unobserved serially correlated shocks in the destination economy could generate a spurious correlation between e_t and n_{t-1}. For example, the shock in period t, C_t, will be positively associated with e_t. The corresponding shock in period $t-1$, C_{t-1}, will be positively associated with e_{t-1} and, hence, n_{t-1}. If these shocks are excluded from the estimating equation, then the estimated θ coefficient could be positive and significant even when destination networks are absent.

Munshi (2003) uses lagged rainfall shocks at the origin, which determine migration flows but are orthogonal to destination shocks, as instruments for n_{t-1}. In most settings, however, valid instruments are not available. McKenzie & Rapoport (2007, 2012) and Woodruff & Zenteno (2007) use historical variation in access to railroads among Mexican origin communities to estimate network effects for migrants from those communities in the United States. Fixed origin characteristics cannot be used as statistical instruments for the current migrant stock in this way, even if they are accidentally determined, because they could, in turn, have shaped other origin characteristics, such as investments in human capital, that have persistent effects on migration. As shown below, heterogeneity in origin characteristics can nevertheless be used in tandem with dynamic predictions from models of migration to provide credible evidence that networks are active. Heterogeneity within the network with respect to tenure at the destination can also be used to identify network effects. I discuss each approach in turn.

4.1. Heterogeneity Within the Network

The specification of the network effect thus far is based on the implicit assumption that all incumbent members are equally effective in increasing the payoffs of new entrants. In practice, we would expect established migrants to be better positioned to provide different forms of help to new arrivals. Beaman (2012) formalizes this intuition by developing a model of job referrals that is based, in turn, on work by Calvo-Armengol & Jackson (2004). There is no ability heterogeneity in this model, and the level of entry in each period, or cohort, is exogenously determined.

Each entering cohort, c, has N_c agents. Each agent works for S periods. The employment rate within the cohort in period $t \in [c, c + S - 1]$ is denoted by S_c^t. The probability that an employed individual will exogenously lose their job at the beginning of any period is b. Each agent (employed or unemployed) hears directly about a job opening with probability a in each period. If the individual is unemployed, they fill the position. If they are employed, they pass on the information to a random unemployed network member. Given this decision rule, the dynamics of employment for a given cohort, c, can be described as follows:

$$S_c^t = a + r^t \quad if \quad t = c, \tag{4.}$$

$$S_c^t = (1 - b)S_c^{t-1} + \left[1 - (1 - b)S_c^{t-1}\right](a + r^t) \quad if \quad t \in (c, c + S - 1), \tag{5.}$$

where r^t is the probability of receiving a referral from the network in period t, conditional on being unemployed. This probability is the ratio of two terms: (*a*) the number of employed individuals who receive redundant job information directly and thus pass it on to the network, and (*b*) the number of unemployed individuals in the network (prior to the receipt of the referrals). While r^t will, in general, vary over time, depending on the size of incoming cohorts, we assume that it is constant on average. To simplify the analysis that follows, let r^t be a constant parameter $\overline{r}$, with the property that $a + \overline{r} < 1$. Collecting terms, Equation 5 can be rewritten as

$$S_c^t = (1 - b)[1 - (a + \overline{r})]S_c^{t-1} + (a + \overline{r}). \tag{6.}$$

Equation 6 describes a one-dimensional, first-order, dynamical system with the property that S_c^t is increasing over time. While this tells us how employment will vary for a given cohort with experience, the model also has dynamic implications for the effect of a given entering cohort's size on employment in other cohorts. First, an increase in the size of the entering cohort, N_t, instantaneously decreases employment for all cohorts, S_c^t. This result is obtained because all members of the entering cohort start the period unemployed. Hence, they do not provide any referrals to other members of the network, while at the same time lowering everyone's probability of receiving a referral in that period. The resulting decline in r^t unambiguously lowers employment for all cohorts from Equations 4 and 5. Second, the impact of an increase in N_t on initial employment in the cohorts that follow is monotonically increasing over time. This result is obtained because the employment rate is increasing over time within any cohort. Each cohort thus starts off as a net consumer of referrals and later becomes a net supplier. Cohorts that enter sufficiently close to t are thus negatively affected by an increase in N_t, but the relationship subsequently switches.

An individual's employment probability could increase with experience, independently of the network, if there is learning on the job. Similarly, if individuals from each origin occupy distinct narrow niches in the destination labor market, then an influx of migrants could lower employment (and wages) for earlier arrivals from the same origin. The prediction of the model that is perhaps most difficult to explain without networks is the one that links employment rates in entering cohorts to the size of preceding cohorts. In general, it is easier to rule out non-network explanations when multiple predictions from the network-based model are available. Beaman (2012) thus proceeds to derive additional predictions for wage dynamics when networks are active, basing her analysis on work by Calvo-Armengol & Jackson (2007). Suppose that individual i receives job information that has an attached wage offer w^o_{ict} with probability a. If unemployed, they accept the offer as before. If employed, they take the job if $w^o_{ict} > w_{ict}$, their current wage. Otherwise, the offer is passed on to a randomly selected member of the network who is unemployed or has a lower wage. Given this decision rule, the optimal policy is to accept any offer if unemployed and to accept a higher offer if employed.

In this labor market, wages are (weakly) increasing with experience, because more draws are received over time and lower wage offers can be discarded. Moreover, the distribution of wages through the direct channel first-order stochastically dominates the distribution of referred wages because the network only offers jobs that were rejected by the initial recipient. This insight generates the following implications: First, an increase in the size of entering cohort N_t (weakly) increases the initial wages of its members, while simultaneously lowering their probability of employment. This is because a greater fraction of these entrants receive their jobs through the direct channel (as r_t has declined). Second, an increase in entering cohort size N_t results in a monotonic decline in the entering wage of subsequent cohorts. This is because the fraction of initial entrants employed through the indirect channel due to cohort t is increasing over time.

Beaman tests these rich dynamic predictions for employment and wages with data on non-family-reunification refugees who were resettled by one agency, the International Refugee Committee (IRC), in the United States between 2001 and 2005. Initial employment and initial wages are measured three months after arrival. The longer-term labor market trajectory for a given individual, however, is unavailable. The network is defined by national origin; the implicit assumption is that individuals from the same country can cooperate at the destination even if they do not have preexisting social ties. To allow for the lagged effect of previous cohorts on the outcomes of subsequent entrants, Beaman uses aggregate data on IRC placements by national origin and location in the United States. The estimated coefficient on current entry flows in the employment equation is negative and significant, as predicted by the model. This coefficient is also increasing across successive lags, from t to $t-2$, as predicted by the model. However, the coefficients on $t-3$ and $t-4$ are less well behaved. The results with wages as the outcome fare even worse: Higher current entry does not increase initial wages for that cohort, nor does it (increasingly) reduce wages for subsequent cohorts.

One possible reason for the failure of the model, particularly the wage component, is that the wage offer process and the accompanying specification of the job acceptance process do not accurately reflect the functioning of the labor market. Beaman does allow for frictions in the referral process, but while this provides one possible explanation for why the data do not match the model, it does not generate additional testable predictions. A second reason for the failure of the model is that the network is not measured accurately. A rich social science literature describes how preexisting communities support the movement of their members, and it is possible that the domain of the network is more restrictive than assumed by Beaman in her analysis. Moreover, there is no reason to assume that networks are restricted to individuals placed by the IRC. Finally, a third reason for the failure of the model is that the assumption that entry flows across successive cohorts are exogenously determined and independent is at odds with reality. The IRC would certainly be aware of these networks, if they did exist, and would consequently tailor its refugee placement to take account of them. One approach to deal with this endogeneity problem would be to model the placement, just as selection into the network was modeled above. Another approach would be to instrument for the entry flows; Beaman's strategy of incorporating a variety of fixed effects may not be sufficient. Indeed, once the endogeneity in entry flows is properly accounted for, the coefficient on lagged entry flows in the employment equation is monotonically increasing across successive lags, as documented by Munshi (2003).

4.2. Heterogeneity Across Origins

While Beaman's (2012) innovative analysis exploits heterogeneity in the experience of agents within the network, an alternative approach exploits heterogeneity across origins to identify destination networks. The advantage of this approach is that exogenous origin characteristics can be utilized for the analysis. As noted, these characteristics could potentially be (accidentally) correlated with independent determinants of migration and outcomes at the destination. Thus, the challenge is to derive predictions that can be used to rule out alternative non-network mechanisms.

Returning to the Roy model with destination networks, as specified in Equation 3, migrant flows from a given origin to a given destination in period t can be derived as a function of lagged flows,

$$e_t = L + \theta \sum_{\tau=0}^{t-1} e_\tau, \tag{7.}$$

where e_τ is entry in period τ. Solving Equation 7 recursively, we obtain

$$e_t = (L + e_0)(1 + \theta)^{t-1}. \tag{8.}$$

Assuming that initial entry, e_0, is exogenously determined, it follows from Equation 8 that e_t is (*a*) increasing in θ at each point in time, (*b*) increasing over time for any θ, and (*c*) increasing more steeply in θ over time. Given that $e_t = 1 - \log \underline{\omega}_t$, corresponding predictions (with the sign reversed) can also be derived for the ability of the marginal entrepreneur. Recall that $\theta \equiv \frac{\theta'}{1-\sigma}$, where θ' measures destination network quality, which in turn is determined by origin characteristics (as discussed below), and σ measures the returns to ability at the origin. The predictions of the model derived with respect to θ above can thus be tested by exploiting variation in θ' or σ. However, origin characteristics that determine θ could also be correlated with other characteristics that are positively and independently associated with migrant flows, matching the first prediction without requiring networks to be active. Similarly, the second prediction could be generated, without networks, if economic conditions at the destination are improving over time relative to conditions at the origin. It is the third prediction, which is similar to the interaction effect in difference-in-difference analyses but which exploits the dynamic implications of the network-based model, that is most effective in ruling out alternative explanations.

As noted, the predictions of the network model can be tested by exploiting variation in either network quality, which will in general be determined by social connectedness in the origin population or in the returns to ability at the origin. Munshi (2011) exploits the latter source of variation in his analysis of network dynamics in the Indian diamond industry. Although India is the world's largest producer of polished diamonds, it does not produce rough diamonds. These diamonds are imported, for the most part from the Antwerp market, and then cut and polished before being sold to foreign buyers. Two traditional business castes—the Marwaris and the Palanpuri Jains—have controlled the business end of the industry from its inception in the late 1960s, leaving the cutting and polishing to the Kanbi Patels, a community of lower-caste agricultural laborers known informally in the industry as the Kathiawaris (Engelshoven 2002). The industry structure changed in the late 1970s with the discovery of massive diamond deposits in Australia's Argyle mines. This supply shock jump-started the Kathiawari business network, and today all three communities account for a substantial share of the industry.

Most exporters visit the Antwerp market for a few days, each month or every other month, to source rough diamonds. They tend to specialize in stones of a particular size, and while each exporter has a small number of regular suppliers, they also want to buy stones from other (different) suppliers from one trip to the next. Given the high value of the diamonds, most exporters must rely on supplier credit. What the network does is to allow its members to receive credit from suppliers with whom they do not have long-term connections; other members of the community who do have established relations with those suppliers stand as guarantors for them, with the threat of social sanctions ensuring that the recipients of the referrals do not renege on their commitments. The use of social sanctions to support economic cooperation has been examined theoretically by Kandori (1992) and more recently, with explicit attention to the network architecture, by Jackson et al. (2012). The institutional mechanism based on community sanctions that is used to support cooperation in the modern diamond industry is also remarkably similar to the mechanism described by Greif (1993) in his analysis of the Maghribi traders a thousand years ago.

Although the origin and the destination in the Roy model have thus far been defined in space, they could, equivalently, be defined by occupation. In the discussion that follows, we will thus think of the origin as the traditional occupation for a given community and the destination as the diamond business (which is the same for all communities). The origin occupation (i.e., the outside option) for the Kathiawaris is farming or industrial labor, neither of which is particularly remunerative, whereas the Marwaris and Palanpuris have many attractive opportunities outside the diamond industry. In the context of our model, this implies that σ should be lower for the Kathiawaris. Once the business networks have formed at the destination (i.e., in the diamond industry), Equation 8 thus predicts that the Kathiawari network will strengthen relatively rapidly. As noted, a rapidly strengthening network is accompanied by a decline in the ability of the marginal entrant. Munshi (2011) tests the model's dynamic predictions for selection on ability with retrospective data obtained from a survey of diamond export firms conducted in 2004–2005. He measures ability by the firm's owner's (or senior partner's) education and business experience (i.e., whether or not they are a first-generation businessman). The son of a businessman in a developing economy inherits his parent's wealth and connections, which improve payoffs in the same way as individual ability. Using either measure, Munshi documents that ability is declining more steeply across successive entering cohorts (with establishment year) among the Kathiawaris. While 30% of entering Kathiawaris in the late 1970s were first-generation businessmen, this statistic had increased to 80% by 2004. This intergenerational occupational mobility was accompanied by spatial movement: While the entrepreneurs were based in Mumbai, all the first-generation businessmen were born in rural areas (if their parents were farmers) or in smaller towns (if their parents were engaged in industrial labor).

Our interpretation of the weakening ability of the entering entrepreneurs over time, especially the Kathiawari entrepreneurs, is that they are being supported by a strengthening caste-specific network. To establish that networks are active, however, it is necessary to rule out alternative explanations. This can be done by systematically relaxing different assumptions in the model. In the Roy model, the returns to ability at the origin σ (i.e., the outside options) vary across communities but not over time. Suppose, instead, that σ was declining over time, relatively rapidly for the Kathiawaris. This then would explain the observed dynamics of selection on ability across communities without requiring networks to be active.

To rule out this alternative explanation, additional information on payoffs at the destination is needed. If weaker Kathiawaris are moving into the diamond business (the destination) simply because their outside options are worsening over time, then they should fare relatively poorly at the destination. In contrast, if networks are active, then two opposing forces are in play: the negative selection on ability, with negative consequences for payoffs at the destination, and the positive effect of the network. Munshi (2011) uses panel data over a 10-year period (1995–2004), or for the latest available period for more recent entrants, to examine destination payoffs across communities and over time. The Kathiawaris keep pace with their Marwari and Palanpuri rivals, despite the decline in the relative ability of entrants from that community over time. Indeed, once this compositional change is accounted for with firm fixed effects (which can be included in the estimating equation because this is an unbalanced panel), the Kathiawari export trajectory is significantly steeper than are the corresponding trajectories for the other two communities or, for that matter, its own trajectory without fixed effects.

The preceding result indicates that there is an underlying force that is giving a relative boost to Kathiawari firms in the diamond industry. In the Roy model with networks, external conditions at the destination (measured by the A_0 parameter) are the same for all firms and constant over time. What varies endogenously across communities and over time is the network effect, θn_{t-1}. Suppose, instead, that networks are absent ($\theta = 0$) but A_0 is increasing over time, relatively rapidly for the Kathiawaris. This would be the case if communities occupy distinct niches in the polished diamond (destination) market and the Kathiawari niche grows faster for exogenous reasons. This explanation is more difficult to rule out because it essentially mirrors the endogenous evolution of the community network. The Kathiawaris do, in fact, tend to specialize in small stones. However, once this specialization is accounted for in the empirical analysis, the estimated cross-community differences grow even larger.

Having systematically ruled out alternative explanations, the final step in the analysis is to provide direct support for the network mechanism. Munshi (2011) does this in two ways: First, he shows that the fraction of marriages within the caste and within the industry, which are especially useful in supporting cooperation, are increasing relatively rapidly over time for the Kathiawaris. Second, he shows that particular firm-specific organizational structures that are associated with participation in the network are disproportionately (and increasingly) favored by the Kathiawaris.

The preceding discussion provides a template to test the Roy model with destination networks. First, a source of exogenous variation in origin characteristics must be identified. This can subsequently be used to test the predictions of the model. Second, alternative explanations, which are generated by relaxing different assumptions of the model, must be ruled out. Third, direct support for the network mechanism must be provided. A recent working paper by Dai et al. (2019) incorporates each of these components, as does Munshi (2011), but exploits variation in network quality, θ', rather than outside options, σ, for the analysis.

The setting for Dai et al.'s (2019) research is the Chinese economy. The Chinese economy has grown at an unprecedented rate over the past 30 years, despite the fact that the conventional ingredients for successful economic development—effective legal systems and well-functioning market institutions—are absent (Allen et al. 2005). It has been argued that informal mechanisms based on reputation and trust substituted for the missing formal institutions (Peng 2004, Song et al. 2011, Greif & Tabellini 2017), and case studies of Chinese production clusters do indicate that long-established relationships among relatives and neighbors from the rural origin substitute for legally enforced contracts between firms (Fleisher et al. 2010, Nee & Opper 2012). Dai and colleagues advance this line of research by utilizing comprehensive data covering the universe of registered firms over many years to identify and quantify the role played by informal community-based networks in the growth of private enterprise in China. Starting from almost no private firms in 1990, there were 15 million registered private firms in 2014, most of which were owned by migrant entrepreneurs who were born in rural areas. Firms need workers, who will also typically move from rural areas to centers of production, and thus the growth of private enterprise in China has been directly or indirectly associated with the largest internal migration in history.

The first step in Dai et al.'s (2019) analysis is to identify a determinant of network quality, θ'. They argue that population density in rural counties, which is mechanically correlated with spatial proximity, is positively associated with social connectedness. Spatial proximity results in more frequent social interactions that, under plausible assumptions on the matching process, give rise to more interconnected social networks (Coleman 1988, Jackson et al. 2012). Interconnected networks sustain greater economic cooperation via norms based on community enforcement (Greif 1993, 1994; Greif & Tabellini 2017). The authors obtain evidence supporting the preceding argument from the China Family Panel Survey, covering a nationally representative sample of households, which shows that the frequency of local social interactions and trust in neighbors are both increasing with population density. However, this result is obtained in rural counties but not cities, possibly because social homogeneity, which independently determines cooperation, is found to be increasing with population density in counties and decreasing with population density in cities. The analysis linking population density to entrepreneurship thus focuses on county-born businessmen; their firms account for two-thirds of all registered private firms (and a comparable share of total registered capital) in China.

The majority of county-born entrepreneurs establish their firms outside their birth counties, so occupational mobility and spatial mobility track together in this economy. The key assumption, which is consistent with the sociological literature on the role of the hometown in supporting migration in China (e.g., Honig 1996, Goodman 1995) and the literature on temporary migration in economics (e.g., Morten 2019), is that these entrepreneurs remain connected to, and can be sanctioned by, their origin communities. The second step in Dai et al.'s (2019) analysis is to construct an augmented version of the Roy model with destination networks, which can be used to validate the preceding hypothesis and to identify a role for business networks. Entrepreneurs now choose between multiple destinations (defined by sectors and locations), and the postentry mutual help at the destination is now reinforced by a preentry referral process, which increasingly channels firms from a given origin (birth county) into an initially favored destination. The interaction between the two types of spillovers generates dynamic increasing returns to network size in any given destination, as well as increased sectoral and spatial within-sector concentration over time. Entry and concentration are also shown to be increasing in θ', which is measured by origin population density, at each point in time, with an increase in this slope over time (at early stages of the industrialization process). The well-documented clustering by migrants in particular sectors and locations has previously been attributed to the presence of underlying networks (Carrington et al. 1996, Munshi 2003); what the augmented model does is to place additional testable structure on the nature of this clustering.

An additional feature of Dai et al.'s (2019) model is that it incorporates capital investment. The destination payoff is now specified as $A_0 \exp(\theta n_{it-1}) \omega^{1-\alpha} K^{\alpha}$, where n_{it-1} is the lagged stock of firms from the origin in destination i and the parameter $\alpha \in (0, 1)$. All firms face an interest rate r, and capital K is chosen by the entrepreneur (given their ability ω) at each point in time to maximize the firm's profit. Denote $A_0 \exp(\theta n_{it-1})$ by community total factor productivity (CTFP). The network-based spillovers that raise CTFP over time have two conflicting effects on the marginal entrant's initial capital. The direct effect, for a given level of ability ω, is to increase firm size by raising TFP, but as discussed above, an increase in CTFP brings in less able firms at the margin to lower TFP. The latter effect is shown to dominate generically; the marginal entering firm from a higher–population density county will be unambiguously smaller, with this negative relationship strengthening over time as networks get larger. In contrast, postentry growth in firm size, which is determined by changes in CTFP alone, is increasing in population density.

The third step in the analysis tests the predictions discussed above using administrative data obtained from the State Administration of Industry and Commerce (SAIC) over the 1990–2009 period. A unique feature of these data, which cover the universe of registered private firms in China, is that the citizenship ID is available for each firm's legal representative who is designated as the entrepreneur in the analysis. The birth county can be recovered from the citizenship ID. Because firms from many different birth counties end up in the same sectors and locations, sector–time period and location–time period fixed effects can be included in the estimating equation. This controls flexibly for time-varying resources provided by local governments and for the destination-based productivity spillovers emphasized in the endogenous growth and agglomeration literatures. The model generates dynamic predictions for the relationship between birth county population density and multiple outcomes—firm entry and concentration, the marginal entrepreneur's ability, initial firm size, and firm growth—and the data match each of them. These results are used to rule out non-network explanations, which are obtained by systematically relaxing the assumptions of the model. The explanation that is most difficult to rule out is one in which entrepreneurs from higher–population density counties have preferential access to sectors and locations that exogenously grow faster (this just mimics the effect of the endogenously evolving network). The seemingly contradictory result that firms from higher–population density counties start smaller but then grow faster within the same destination (once sector and location fixed effects are included in the estimating equation) is most useful in ruling out this explanation.

The final step in validating the network-based model is to test the underlying mechanism, which is that initial entry in a given destination (sector and location) should generate subsequent entry in that destination (due to increased postentry productivity and the preentry referral effect). Consistent with the dynamic multiplier effect that is implied by the model, one additional entrant from the birth county in a particular sector and location in the initial 1990–1994 period is associated with seven additional entrants over the 2000–2004 period and nine additional entrants over the 2005–2009 period. Moreover, the initial entry effect is increasing with birth county population density. While these results provide direct support for the presence of business networks organized around the birth county, Dai et al. (2019) find that initial entry from other origins in a given sector and location has no consequences for subsequent entry. The absence of a (negative) cross-community effect goes against the view that the members of the network are simply colluding to extract rents (Brooks et al. 2016) or that networks are competing with each other for subsidized credit (Bai et al. 2019).

5. THE ROY MODEL WITH ORIGIN NETWORKS

Given the information and enforcement problems that migrants face in the destination economies, it is not surprising that the literature in economics and other social sciences has largely focused on the role of destination networks in supporting migration. More recently, however, a literature has emerged in economics that examines how networks providing mutual insurance to their members at the origin can shape migration decisions. Once again, the Roy model, suitably augmented to incorporate a network component, can be used to analyze this relationship.

The discussion that follows is based on work by Munshi & Rosenzweig (2016). We now suppress heterogeneity in individual ability. Income at the rural origin is exogenously determined and noisy; risk-averse households thus benefit from a mutual insurance arrangement that smooths their consumption over time. To begin with, assume that all households have the same average income at the origin. However, opportunities at the urban destination, which become available to one or more members of the household, vary exogenously across households. The key assumption is that households with members who migrated permanently, rather than temporarily, will have reduced access to the origin-based insurance network. This is because, first, permanent migrants cannot be as easily punished by the insurance network, and their family back home in the origin now has superior outside options (in the event that the household is excluded from the network). It follows that households with such migrants cannot credibly commit to honoring their future obligations at the same level as households without migrants. Second, an information problem arises if the migrant's income cannot be observed (which is more likely with permanent migration). If the household is treated as a collective unit by the network, it always has an incentive to underreport its destination income so that transfers flow in its direction. A household must thus weigh the gain in (average) income from migration against the loss in insurance when deciding whether or not to send its members to the destination. Note that the implicit assumption underlying this tradeoff, as suggested by Banerjee & Newman (1998), is that formal insurance, which could be used to substitute for the network, is unavailable to migrants.

If we assume, as Munshi & Rosenzweig (2016) do, that full risk sharing can be supported by the network, then each household has two choices: It can keep all its members at the origin and enjoy the benefits of the network, or it can send one or more members to the destination but then lose access to network insurance entirely. Let the total population be P and let the number of households that choose to participate in the network be n. A larger network does a better job of smoothing consumption, and thus the payoff for those households that choose to keep all their members at the origin, $W^O(n)$, is increasing in n. All households have the same origin income; the only source of heterogeneity in the model is the income boost from migration, ϵ, which is private information. It is straightforward to verify that there exists a threshold $\underline{\epsilon}$ for which the associated household is indifferent between moving and staying, and this pins down n in equilibrium.

In general, a better functioning network will be associated with a higher threshold (it takes more to leave) and, hence, lower migration. This is the argument that Munshi and Rosenzweig (2016) make to explain why permanent migration is so low in India, on account of the fact that caste-based insurance networks in that country are exceptionally well functioning. One way to test this hypothesis, following Dai et al. (2019), would be to exploit variation in network quality generated by exogenous caste characteristics to assess whether superior networks lead to lower migration. Munshi and Rosenzweig are unable to identify caste characteristics that are suitable for this exercise. What they do instead, as does Beaman (2012), is to exploit heterogeneity within the caste: First, they theoretically identify which households benefit less (or more) from the origin network. Next, they test whether it is precisely those households that are more (or less) likely to send members to the destination.

To implement the preceding exercise, we now introduce heterogeneity in average origin income across households. With diminishing marginal utility, the total surplus generated by the insurance arrangement can be increased by redistributing income so that relatively poor households consume more than they earn on average. This gain from redistribution must be weighed against the cost to the members of the network from the accompanying decline in its size, since relatively wealthy households will now be more likely to leave, and smaller networks are less able to smooth consumption. Solving at the same time the income-sharing rule, which determines redistribution, and the participation decision, which determines the level of migration, is a challenging exercise. Previous research (e.g., Genicot & Ray 2003, Abramitzky 2008) has solved for either the income-sharing rule or the level of participation, holding the other constant. Munshi & Rosenzweig (2016) are able to show, under reasonable conditions, that the income-sharing rule will be set so that there is some amount of redistribution in equilibrium, even after accounting for its consequences for migration (and consumption smoothing). This implies that relatively wealthy households benefit less from the network, everything else being equal, and so will be more likely to have migrant members.

Munshi and Rosenzweig's second theoretical prediction exploits heterogeneity in origin (rural) income risk. With full risk sharing, all households face the same variance in consumption ex post, regardless of the ex ante variance in their income. This implies that households facing greater income risk benefit more from the network and, therefore, should be less likely to have migrant members. This is precisely the opposite of what we would expect if insurance networks were absent: Households facing greater income risk at the origin would then be more likely to send members to the destination as a way of diversifying their income (Lucas & Stark 1985). Alternative explanations are available for the positive association between relative income and migration. For example, this association could be observed if destination networks are active and migrants are positively selected on ability, which, in turn, is correlated with origin income. However, the negative association between origin income risk and migration is less easy to explain without a role for insurance networks at the origin.

Munshi & Rosenzweig (2016) test these predictions of the model with data from the ICRISAT surveys and the most recent round (for year 2006) of the Rural Economic Development Survey (REDS), a nationally representative survey of rural Indian households that has been administered at multiple points in time over the past four decades. Their analysis establishes that (*a*) there is substantial redistribution of income within castes, (*b*) relatively wealthy households within their caste are more likely to report that one or more adult male members have permanently left the village, and (*c*) households with a higher coefficient of variation in their rural income are less likely to have migrant members. Recall that the key reason origin-based insurance networks dampen permanent migration is that they are associated with a loss in network services. Munshi and Rosenzweig use multiple rounds of the REDS to document that relatively wealthy households in their caste are (*a*) less likely to migrate, as above; (*b*) less likely to give or receive transfers within their caste; and (*c*) less likely to marry within their caste (which is a precondition for access to the network).

Although permanent migration may be low in India, temporary (seasonal) migration has been increasing over time (Morten 2019). The commitment and information problems that arise with permanent migration are less relevant for temporary migration; migrants can be punished directly once they return to the origin if they renege on their obligations and information about their destination income is also readily available to the origin community (because they must return soon enough with it). Temporary migration thus may not result in a loss in origin-network services.

Continuing to focus on insurance networks at the origin, the canonical model of mutual insurance with full risk sharing derives the consumption of each network member in each state of the world as the solution to a social planner's welfare maximization problem. For example, with logarithmic preferences, each household receives a fixed share of the total income in each period (state), where the share is exogenously determined. To incorporate temporary migration, we follow Morten (2019) and assume that in each period the destination experiences an event q_t that follows an independent and identically distributed (i.i.d.) process with probabilities $\pi^q(q_t)$; this assumption differs from the one proposed by Munshi & Rosenzweig (2016), who assume that destination opportunities are household specific and deterministic. If we continue to assume full risk sharing, then the migration decision can also be included in the social planner's problem; the optimal number of temporary migrants would be selected in each period to maximize welfare (although we would now need to introduce some heterogeneity, perhaps in moving costs, to pin down the level of migration).

What Morten (2019) does in her analysis, instead, is to build on the more complicated model of mutual insurance with dynamic limited commitment developed by Ligon et al. (2002). This model is based on the idea that in some states of nature, the sender could prefer to default (and forego future insurance) rather than provide the large transfers that are required to satisfy full risk sharing. Optimal transfers in this model will continue to maximize social welfare, but they will now be subject to a participation constraint in each state. When migration is added to the model, the participation constraint must be satisfied at two points in time in each period: (*a*) once origin incomes are realized but prior to temporary migration, and (*b*) after temporary migration (and resulting income realizations).

Morten estimates the parameters of the model using ICRISAT data, which cover a small number of villages in one region of India but provide detailed information over time on consumption, income, transfers, and temporary migration for sampled households. Although the estimated model is useful for counterfactual policy simulations (as discussed below), the analysis does not satisfy the conditions we have specified for a credible identification of network effects. The first limitation is that the model does not generate predictions that are specific enough to rule out alternative explanations. Morten states that improved risk sharing at the origin has two effects on migration: (*a*) It reduces the incentive to migrate temporarily in response to negative income shocks at the origin, and (*b*) it increases migration by smoothing the effect of income uncertainty at the destination. However, these (relatively general) implications are not derived explicitly from the model. A possibly more promising way forward would be to assume full risk sharing. The resulting model would be more analytically tractable and could also potentially deliver more specific predictions. This simplification, moreover, has empirical justification. The testable implication of full risk sharing is that the household's consumption should be unaffected by contemporaneous income shocks. Studies from across the world, including India, have documented close-to-full risk sharing, and Morten's own analysis estimates the coefficient on the income shock to be similar in magnitude to the point estimates obtained in those studies.

A second limitation of Morten's analysis is that it fails to exploit exogenous variation at the origin and the destination, which is needed to disentangle the competing mechanisms through which origin networks shape migration. Without reduced-form tests of these mechanisms, network effects cannot be credibly identified. Moreover, the village is treated as the domain of the network. Although this assumption has been made for convenience in a number of recent studies on networks (not related to migration), it is at odds with a well-established social science literature and an emerging literature within economics (e.g., Munshi & Rosenzweig 2006, 2016; Munshi 2011; Mazzocco & Saini 2012) that documents that the endogamous caste, spanning many villages, is the social unit around which networks supporting different activities, including internal migration (Dhillon et al. 2013), are organized.

6. DIRECTIONS FOR FUTURE RESEARCH

6.1. Quantification

Early analyses of networks and migration (e.g., Munshi 2003, 2011; Beaman 2012) focused on the identification of network effects. That task is largely complete, and the idea that social networks support the occupational and spatial mobility of their members is now well established in economics. However, the argument that networks should be incorporated in theoretical and empirical analyses of migration is only justified if the magnitude of their role is substantial, and much less effort has been devoted to this question.

The studies listed above use special research settings to identify network effects. However, such settings are less useful for quantifying the magnitude of the network effects. A well-designed quantification exercise would begin by deriving predictions from a model of migration that incorporates networks and then testing them with comprehensive (ideally, economy-wide) data. If an exogenous source of variation is available and the model generates sufficiently precise predictions, then alternative explanations can be ruled out and network effects can be credibly identified, as discussed above. The next step in the quantification exercise would be to estimate the structural parameters of the model and then predict the level of migration (and associated outcomes) with and without networks. This is the approach taken by Munshi & Rosenzweig (2016) and Dai et al. (2019). Munshi and Rosenzweig, using nationally representative data from India, estimate that a 50% improvement in formal insurance, which substitutes for network insurance for households with migrant members, would more than double the migration rate, from 4% to 9%. Dai and colleagues, using data that cover the universe of registered private firms in China, estimate that rural hometown networks increase the number of firms established between 1995 and 2004 by 11% and the capital stock in 2004 by 12.5% for the economy as a whole.

While the preceding results indicate that social networks have played an important role in shaping migration in major developing economies, their welfare consequences must also be evaluated. One obvious inefficiency that arises when networks are active is that information and cooperation do not cross community lines, resulting in a misallocation of resources. For example, Banerjee & Munshi (2004) compare capital investment and export trajectories between local firms and outsiders in Tirupur's production cluster. They document that local businessmen hold more capital stock than do the outsiders on average, at all levels of experience. Nevertheless, production grows faster for the outsiders at all levels of experience: They start with lower levels of output but outstrip the locals after five years of experience. Based on a model of production in which entrepreneurial ability and capital are complementary inputs, these empirical facts can only be observed simultaneously if outsiders have higher ability on average and locals face a lower cost of capital. These intercommunity differences are not surprising: Local entrepreneurs in Tirupur belong to a wealthy agricultural caste with few alternative uses for its capital, whereas migrant entrepreneurs are drawn from castes and communities with many generations of business experience. The fact that cheap capital fails to move from the local community to the more competent outsiders is, nevertheless, indicative of an allocative inefficiency. This misallocation should not be observed within communities, where we expect that well-functioning networks will ensure that all entrepreneurs face the same interest rate; indeed, Banerjee and Munshi find that firms holding more capital stock do grow faster (and have higher levels of production) within communities. It is only across communities that the negative correlation between capital stock and production is obtained, presumably because the outsiders cannot credibly commit to repaying the locals for the credit they receive.

Banerjee & Munshi's (2004) analysis is based on a single production cluster (one industry and one location). Thus, although it identifies a particular inefficiency that arises when networks are active, it is less useful for quantifying the efficiency and equity consequences of this inefficiency. A properly specified and validated model, estimated with comprehensive (economy-wide) data, could be used for this purpose (as discussed above). With the increasing availability of administrative and nationally representative data, quantification exercises of this nature are becoming increasingly feasible, and this suggests a fruitful direction for future research. Given the inefficiencies that arise when networks are active, and the spillovers that these institutions generate, there is an obvious role for policy. Once structural models of migration with networks have been estimated, they can be readily applied to evaluate counterfactual policies, as done by Munshi & Rosenzweig (2016), Dai et al. (2019), and Morten (2019). A recurring message from these analyses is that well-intentioned programs that attempt to boost migration in specific (target) households can have unintended negative consequences for other members of the network.

6.2. Assimilation

The discussion thus far has focused on the formation of migrant networks and their subsequent (short-run) dynamics. While these networks provide different forms of support for their members, they will ultimately decay; in developing countries, this will be because they are no longer needed once markets are functioning efficiently, whereas with international migration it will be because their members have assimilated in the host economy. Either way, the dynamics will mirror (in reverse) the growth in networks described above.

In the context of the Roy model, the origin payoff is now the payoff in the migrant network, and the destination payoff is the payoff obtained if the individual chooses to work independently in the modern (host) economy. Following Munshi & Rosenzweig (2006), the origin payoff, $W^O(n_{t-1})$, depends on the number of community members from the previous cohort or generation who selected into the network (and who provide referrals and support for those who follow). The destination payoff, $W^D(\omega)$, is increasing with the individual's ability. The process starts with some individuals being exogenously shifted out of the network. This results in an interior ability threshold above which individuals endogenously select out of the network in the cohort that follows. Once the dynamics have been initiated, the threshold becomes progressively lower from one cohort to the next, until the network collapses and all its members are absorbed (i.e., assimilated) in the modern (host) economy.

Munshi & Rosenzweig (2006) examine this process of network decay in Mumbai. This city emerged as an industrial center in the middle of the nineteenth century. With industrialization came the movement of workers from rural areas to the city. Social historians who have studied this process (e.g., Morris 1965, Chandavarkar 1994) describe how caste networks supported this internal migration, consistent with the discussion above. The point of departure for Munshi and Rosenzweig's analysis is an economic shock that occurred more than a hundred years later, around 1990, by which time the caste networks had been in place for multiple generations. There was an exogenous restructuring of the Indian economy, which increased wages in the corporate and service sectors where the caste networks were irrelevant. As predicted, Munshi and Rosenzweig document a movement out of the traditional networked occupations across successive cohorts, with higher-ability individuals leading the way. However, they also uncover substantial frictions that slow down the process of occupational mobility.

The explanation that Munshi & Rosenzweig (2006) propose for these frictions is based on an endogenously determined community identity. Individuals who select out of the traditional occupation often migrate (permanently) elsewhere. The conventional punishment mechanisms that maintain cooperation within communities will then no longer be effective. An alternative strategy to maintain cooperation under these circumstances would be to instill a strong sense of community identity in childhood, which would ensure that individuals remain tied to their community (and select into the traditional occupation) in adulthood. Self-interested individuals do not internalize the cost to their community, on account of a weaker network, when they exit the traditional occupation. While community identity may thus be welfare enhancing when it is first put in place, it can result in a dynamic inefficiency if it persists in subsequent cohorts (or generations) past the point where it is optimal for high-ability individuals to remain in the traditional occupation.

Cultural norms are persistent by design, which explains why heavily networked blue-collar communities often appear to stubbornly resist change, despite the fact that these same communities were extremely dynamic in the distant past when their networks formed. The same argument would apply to migrant assimilation. The idea that identity, and values more generally, are purposefully instilled is in line with previous theoretical work on this topic in economics (e.g., Bisin & Verdier 2000, Tabellini 2008). To model migrant assimilation, however, it is not enough to analyze community identity and the resulting tension between individual and group mobility at each point in time. The dynamic process through which migrant networks form, become established over multiple generations, and then serve as the point of departure for further (individual or group) mobility must also be characterized. Such analyses require data over multiple generations, and administrative data satisfying these requirements are starting to become available. Migrant assimilation has received little theoretical or empirical attention in economics, despite its importance, and it would thus appear to be a fruitful area for future research.

6.3. Social Interactions

Our characterization of networks thus far has assumed, implicitly or explicitly, that individuals match randomly within their communities and that interactions within the community alone are relevant for cooperation. Moreover, networks are assumed to operate independently. While these assumptions may be realistic in many settings, there will clearly be environments in which a small number of networks (with market power) are competing with each other or where interactions both inside and outside the social group are relevant.

Esteban & Ray (1994), using an axiomatic approach, describe how the group-size distribution in a population can determine conflict or, conversely, cooperation. They argue that polarization, in which there is (*a*) a high degree of homogeneity within each group, (*b*) a high degree of heterogeneity across groups, and (*c*) a small number of significantly sized groups, is more relevant for conflict than fractionalization (concentration). To test this hypothesis, Bazzi et al. (2019) exploit the exogenous (quasi-random) variation in ethnic group composition across recently settled Indonesian villages that was generated by the government's trans-migration program over the 1979–1988 period. They find that national identity, measured by the language spoken at home, intermarriage, and children's name choice is increasing in ethnic fractionalization and decreasing in polarization, as predicted. They also obtain complementary results with local public good provision and conflict as outcomes. While these results are novel and informative, they are obtained in a nonrepresentative setting that is less useful for quantification. Moreover, the analysis is missing a dynamic component, which would also help with identification (as argued above). As with assimilation, the analysis of the interactions between migrant groups and between specific migrant groups and the native population has received little theoretical or empirical attention in the economics literature and is an open area for research.

The model developed by Bazzi et al. (2019) assumes that the ethnic composition of the population is exogenously determined and that individuals are matched randomly to each other (without regard to their ethnicity). In practice, the matching process may be biased toward in-group interactions, and this can have consequences for the pattern of relationships that forms (Currarini et al. 2009). While the preceding discussion focused on the interactions between networks, the interactions within networks (i.e., the network architecture) will also have consequences for economic outcomes, including migration. The obvious challenge here is that the network architecture is typically not observed in its entirety, and even if it is, it is endogenously determined. In an interesting recent paper, Blumenstock et al. (2019) make progress on both dimensions by utilizing data covering the universe of mobile phone activity in Rwanda over a five-year period. Their objective is to disentangle two alternative mechanisms—cooperation and information—through which social networks could potentially support migration. Particular network architectures are associated with each mechanism: Information capital is associated with expansive networks, whereas cooperation capital is associated with interconnected networks. Network statistics corresponding to each mechanism can be constructed at each potential destination location, for a given migrant, based on their history of phone calls at each location—that is, their direct network links and, more importantly, their contacts' history of phone calls at those locations (i.e., their network links). The idea is to see whether network statistics associated with information or network statistics associated with cooperation, as constructed from the phone links, are more influential in determining migration.

An innovative feature of Blumenstock et al.'s (2019) analysis, which substantially reduces the potential for omitted variable bias, is that the network statistics are constructed using the direct contact's contacts. While this approach, and the idea of using cell phone data, has many appealing features, it also has limitations. First, the motivation for the horse race between information and cooperation is unclear. Both mechanisms are evidently relevant in general, with their importance depending on the circumstances. Second, the cell phone data do not provide information on individual characteristics, the nature of the relationship between connected individuals, or the individual's origin (birth) location. The absence of the latter is particularly problematic for analyses of migration, because movements from the origin to the destination cannot be distinguished from movements in the opposite direction, and neither can networks in the two (very different) types of location. Research on networks and migration using cell phone data is still at a nascent stage and much more progress needs to be made, but with the increasing availability of such data for research there is also much promise.

DISCLOSURE STATEMENT

The author is not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

Research support from the National Science Foundation through grants SES-0431827 and SES-0617847, Cambridge-INET (Institute for New Economic Thinking), the Keynes Fund and the Newton Trust at the University of Cambridge, and the Agence Nationale de la Recherche (ANR) under the EUR Project ANR-17-EURE-0010 is gratefully acknowledged. I am responsible for any errors that may remain.

LITERATURE CITED

Abramitzky R. 2008. The limits of equality: insights from the Israeli kibbutz. *Q. J. Econ.* 123(3):1111–59

Abramitzky R, Boustan L. 2017. Immigration in American economic history. *J. Econ. Lit.* 55(4):1311–45

Allen F, Qian J, Qian M. 2005. Law, finance, and economic growth in China. *J. Financ. Econ.* 77(1):57–116

Bai C-E, Hsieh C-T, Song M. 2019. *Special deals with Chinese characteristics*. NBER Work. Pap. 25839

Banerjee A, Munshi K. 2004. How efficiently is capital allocated? Evidence from the knitted garment industry in Tirupur. *Rev. Econ. Stud.* 71(1):19–42

Banerjee A, Newman AF. 1998. Information, the dual economy, and development. *Rev. Econ. Stud.* 65(4):631–53

Bazzi S, Gaduh A, Rothenberg AD, Wong M. 2019. Unity in diversity? How intergroup contact can foster nation building. *Am. Econ. Rev.* 109(11):3978–4025

Beaman LA. 2012. Social networks and the dynamics of labour market outcomes: evidence from refugees resettled in the US. *Rev. Econ. Stud.* 79(1):128–61

Beine M, Docquier F, Ozden C. 2011. Diasporas. *J. Dev. Econ.* 95:30–41

Bertoli S, Fernandez-Huertas Moraga J. 2012. *Visa policies, networks and the cliff at the border*. IZA Discuss. Pap. 7094, Inst. Labor Econ., Bonn, Ger.

Bisin A, Verdier T. 2000. "Beyond the melting pot": cultural transmission, marriage, and the evolution of ethnic and religious traits. *Q. J. Econ.* 115:955–88

Blumenstock JE, Chi G, Tan X. 2019. *Migration and the value of social networks*. Work. Pap., Univ. Calif., Berkeley

Borjas G. 1987. Self-selection and the earnings of immigrants. *Am. Econ. Rev.* 77(4):531–53

Brooks WJ, Kaboski JP, Li YA. 2016. *Growth policy, agglomeration, and (the lack of) competition*. NBER Work. Pap. 22947

Calvo-Armengol A, Jackson MO. 2004. The effects of social networks on employment and inequality. *Am. Econ. Rev.* 94(3):426–54

Calvo-Armengol A, Jackson MO. 2007. Networks in labor markets: wage and employment dynamics and inequality. *J. Econ. Theory* 132(1):27–46

Carrington WJ, Detragiache E, Vishwanath T. 1996. Migration with endogenous moving costs. *Am. Econ. Rev.* 86(4):909–30

Chandavarkar R. 1994. *The Origins of Industrial Capitalism in India: Business Strategies and the Working Classes in Bombay, 1900–1940*. Cambridge, UK: Cambridge Univ. Press

Chiquiar D, Hanson GH. 2005. International migration, self-selection, and the distribution of wages: evidence from Mexico and the United States. *J. Political Econ.* 113(2):239–81

Ciccone A, Hall RE. 1996. Productivity and the density of economic activity. *Am. Econ. Rev.* 86(1):54–70

Coleman JS. 1988. Social capital in the creation of human capital. *Am. J. Sociol.* 94:S95–120

Cuecuecha A. 2005. *The immigration of educated Mexicans: the role of informal social insurance and migration costs*. Work. Pap., Inst. Tecnol. Auton. Mex., Mexico City

Currarini S, Jackson MO, Pin P. 2009. An economic model of friendship: homophily, minorities, and segregation. *Econometrica* 77(4):1003–45

Dai R, Mookherjee D, Munshi K, Zhang X. 2019. *The community origins of private enterprise in China*. Work. Pap., Peking Univ., Beijing

Dhillon A, Iversen V, Torsvik G. 2013. *Employee referral, social proximity and worker discipline: theory and evidence from India*. CESifo Work. Pap. 4309, CESifo, Munich, Ger.

Docquier F, Peri G, Ruyssen I. 2014. The cross-country determinants of potential and actual migration. *Int. Migr. Rev.* 48(Suppl. 1):S37–99

Engelshoven M. 2002. Rural to urban migration and the significance of caste: the case of the Saurashtra Patels of Surat. In *Development and Deprivation in Gujarat*, ed. G Shah, M Rutten, H Streefkerk, pp. 294–313. New Delhi: Sage

Esteban J-M, Ray D. 1994. On the measurement of polarization. *Econometrica* 62(4):819–51

Fernandez-Huertas Moraga J. 2011. New evidence on emigrant selection. *Rev. Econ. Stat.* 93(1):72–96

Fernandez-Huertas Moraga J. 2013. Understanding different migrant selection patterns in rural and urban Mexico. *J. Dev. Econ.* 103:182–201

Fleisher B, Hu D, McGuire W, Zhang X. 2010. The evolution of an industrial cluster in China. *China Econ. Rev.* 21(3):456–69

Genicot G, Ray D. 2003. Group formation in risk-sharing arrangements. *Rev. Econ. Stud.* 70(1):87–113

Goodman B. 1995. *Native Place, City, and Nation: Regional Networks and Identities in Shanghai, 1853–1937*. Berkeley: Univ. Calif. Press

Greif A. 1993. Contract enforceability and economic institutions in early trade: the Maghribi traders' coalition. *Am. Econ. Rev.* 83(3):525–48

Greif A. 1994. Cultural beliefs and the organization of society: a historical and theoretical reflection on collectivist and individualist societies. *J. Political Econ.* 102(5):912–50

Greif A, Tabellini G. 2017. The clan and the corporation: sustaining cooperation in China and Europe. *J. Comp. Econ.* 45(1):1–35

Honig E. 1996. Regional identity, labor, and ethnicity in contemporary China. In *Putting Class in Its Place: Worker Identities in East Asia*, ed. EJ Perry, pp. 225–43. Berkeley, CA: Inst. East Asian Stud.

Ibarran P, Lubotsky D. 2007. Mexican immigration and self-selection: new evidence from the 2000 Mexican census. In *Mexican Immigration in the United States*, ed. G Borjas, pp. 159–92. Chicago: Univ. Chicago Press

Jackson MO, Rodriguez-Barraquer T, Tan X. 2012. Social capital and social quilts: network patterns of favor exchange. *Am. Econ. Rev.* 102(5):1857–97

Kandori M. 1992. Social norms and community enforcement. *Rev. Econ. Stud.* 59(1):63–80

Ligon E, Thomas JP, Worrall T. 2002. Informal insurance arrangements with limited commitment: theory and evidence from village economies. *Rev. Econ. Stud.* 69:209–44

Lucas REB, Stark O. 1985. Motivations to remit: evidence from Botswana. *J. Political Econ.* 93(5):901–18

Mazzocco M, Saini S. 2012. Testing efficient risk-sharing with heterogeneous risk preferences. *Am. Econ. Rev.* 102(1):428–68

McKenzie D, Rapoport H. 2007. Network effects and the dynamics of migration and inequality: theory and evidence from Mexico. *J. Dev. Econ.* 84:1–24

McKenzie D, Rapoport H. 2012. Self-selection patterns in Mexico-U.S. migration: the role of migration networks. *Rev. Econ. Stat.* 92(4):811–21

Mishra P. 2007. Emigration and wages in source countries: evidence from Mexico. *J. Dev. Econ.* 82(1):180–99

Morris MD. 1965. *The Emergence of an Industrial Labor Force in India: A Study of the Bombay Cotton Mills, 1854–1947*. Berkeley: Univ. Calif. Press

Morten M. 2019. Temporary migration and endogenous risk sharing in village India. *J. Political Econ.* 127(1):1–46

Munshi K. 2003. Networks in the modern economy: Mexican migrants in the U.S. labor market. *Q. J. Econ.* 118(2):549–97

Munshi K. 2011. Strength in numbers: networks as a solution to occupational traps. *Rev. Econ. Stud.* 78:1069–101

Munshi K. 2014. Community networks and the process of development. *J. Econ. Perspect.* 28(4):49–76

Munshi K, Rosenzweig M. 2006. Traditional institutions meet the modern world: caste, gender and schooling choice in a globalizing economy. *Am. Econ. Rev.* 96(4):1225–52

Munshi K, Rosenzweig M. 2016. Networks and misallocation: insurance, migration, and the rural-urban wage gap. *Am. Econ. Rev.* 106(1):46–98

Nee V, Opper S. 2012. *Capitalism from Below: Markets and Institutional Change in China*. Cambridge, MA: Harvard Univ. Press

Orrenius PM, Zavodny M. 2005. Self-selection among undocumented immigrants from Mexico. *J. Dev. Econ.* 78(1):215–40

Peng Y. 2004. Kinship networks and entrepreneurs in China's transitional economy. *Am. J. Sociol.* 109(5):1045–74

Rosenzweig M. 2010. *Global wage inequality and the international flow of migrants*. Work. Pap., Yale Univ., New Haven, CT

Roy AD. 1951. Some thoughts on the distribution of earnings. *Oxford Econ. Pap.* 3(2):135–46

Song Z, Storesletten K, Zilibotti F. 2011. Growing like China. *Am. Econ. Rev.* 101(1):196–233

Tabellini G. 2008. The scope of cooperation: values and incentives. *Q. J. Econ.* 123(3):905–50

Woodruff C, Zenteno R. 2007. Migration networks and microenterprises in Mexico. *J. Dev. Econ.* 82(2):509–28

Annual Review of Economics

Volume 12, 2020

Contents