NBER WORKING PAPER SERIES

SOCIAL NETWORKS, ETHNICITY, AND ENTREPRENEURSHIP

William R. Kerr
Martin Mandorff

Working Paper 21597
http://www.nber.org/papers/w21597

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
September 2015, Revised July 2019

Comments are appreciated and can be sent to wkerr@hbs.edu. We thank Emek Basker, Gary Becker, Michel Beine, Ola Bengtsson, Gustaf Bruze, Dennis Carlton, Barry Chiswick, Rob Fairlie, Matthew Gentzkow, John Haltiwanger, Emil Iantchev, Svante Janson, Mini Kaur, Steven Lalley, Anne Le Brun, Ben Mathew, Trang Nguyen, Andriy Protsyk, Yona Rubinstein, Jesse Shapiro, Rachel Soloveichik, Chad Syverson, Catherine Thomas, Robert Topel and Nick Wormald and seminar participants for very valuable comments. We thank Meir Brooks and Rahul Gupta for excellent research support. The theory section of this paper draws heavily from Mandorff's Ph.D. Dissertation at the University of Chicago. Financial support from the Marcus Wallenberg Foundation, the Jan Wallander and Tom Hedelius Foundation, the Esther and T.W. Schultz Dissertation Fellowship, the Markovitz Dissertation Fellowship, the Kauffman Foundation, and Harvard Business School is gratefully acknowledged. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Social Networks, Ethnicity, and Entrepreneurship
William R. Kerr and Martin Mandorff
NBER Working Paper No. 21597
September 2015, Revised July 2019
JEL No. D21,D22,D85,F22,J15,L14,L26,M13

ABSTRACT

We study the relationship between ethnicity, occupational choice, and entrepreneurship. Immigrant groups in the United States cluster in specific business sectors. For example, Koreans are 34 times more concentrated in self employment for dry cleaning than other immigrant groups, and Gujarati-speaking Indians are 108 times more concentrated in managing motels. We quantify that smaller and more socially isolated ethnic groups display higher rates of entrepreneurial concentration. This is consistent with a model of social interactions where non-work relationships facilitate the acquisition of sector-specific skills and result in occupational stratification along ethnic lines via concentrated entrepreneurship.

William R. Kerr
Harvard Business School
Rock Center 212
Soldiers Field
Boston, MA 02163
and NBER
wkerr@hbs.edu

Martin Mandorff
Swedish Competition Authority
Konkurrensverket SE-103 85
Stockholm, Sweden
Martin.Mandorff@kkv.se

1 Introduction

Immigrants engage in self-employment and entrepreneurship more than natives, especially among new arrivals. Using the 2007-2011 Current Population Surveys, Fairlie and Lofstrom (2013) calculate that immigrants represent 25% of new US business owners compared to their 15% workforce share. Moreover, immigrant business owners tend to specialize in a few industries, and these industries vary across ethnic groups. Prominent US examples include Korean dry cleaners, Vietnamese nail care salons, Yemeni grocery stores, and Punjabi Indian convenience stores. Earlier ethnic specializations included Jewish merchants in medieval Europe and Chinese launderers in early twentieth century California. Despite the importance of these patterns economically—for example, *The Economist* reported that one-third of all US motels in 2016 were owned by Gujarati Indians—few studies examine the origin or consequences of this ethnic specialization for self-employment.

We focus on how small ethnic group size and isolated social interactions among group members can yield entrepreneurial specialization. We develop a simple model that considers a small industry where self-employed entrepreneurs benefit from social interactions outside of work (e.g., family gatherings, religious and cultural functions, meetings with friends). At these social events, self-employed entrepreneurs can discuss customer trends, share best practices, coordinate activities, and so on. The model describes how a small ethnic minority group that has restricted social interactions can have a comparative advantage for self-employment, similar to the account of Chung and Kalnins (2006) for better resource access through ethnic networks in the case of Gujarati hotel owners.

We analyze the model's predictions using Census Bureau data for the United States in 2000. The size of groups and their social isolation, which we measure using in-marriage rates, strongly predict industrial concentration for immigrant self-employed entrepreneurs. A 10% decline in group size raises the group's industry concentration for self-employment by 6%, and a 10% increase in group isolation boosts concentration by 5%. These results are robust under many specification variants and using instrument variable techniques outlined below.

We focus on ethnic group size and social isolation due to the exceptionally broad and pervasive nature of immigrant concentration for self-employment. While the particulars vary across ethnic groups, time periods and national settings, we consistently observe self-employed specialization among immigrant groups. Thus, we seek a general mechanism that does not revolve around the traits of any single ethnic group or setting, and our empirical analysis includes as many immigrant groups in the United States as possible. Understanding how group-level behavior can generate group-level differences is important, as we know that the higher immigrant propensity towards entrepreneurship remains after controlling for the observable traits of individuals. Moreover, group size and group social isolation manifest in many ways in the literature: risk sharing, provision of support and mentoring, sanctions for misbehavior, etc. These factors become more powerful for smaller, tighter ethnic groups. Group influences could also lead to behavioral factors prompting self employment (e.g., Åstebro et al., 2014).

While there are anecdotal and sociological accounts of how social interactions can connect to entrepreneurial activity (e.g., Fairlie and Robb, 2007), the reflection problem described by Manski (2003) makes identification of social interaction effects challenging. Econometric challenges like omitted factors and reverse causality also exist. We consider two instrument variable specifications to address this issue. One approach uses the 1980 group sizes and in-marriage rates in the United States. Our second approach instruments US ethnic group size with the predictions from a gravity model for migration to the United States and instruments US in-marriage rates with those observed for the same ethnic group in the United Kingdom. These estimations, along with many robustness checks, confirm the OLS results.

Our work connects to prior studies of immigrant entrepreneurship and self-employment behavior[1] but with a focus on industrial specialization across groups. Classic accounts of entrepreneurship focus on factors like risk taking (Kihlstrom and Laffont, 1979), business acumen (Lucas, 1978) or skill mix (Lazear, 2005), with the connection of entrepreneurship to migration being frequently noted but unexplained. We emphasize how social interactions can generate group-level effects around entrepreneurship and industry choice that go beyond the individual traits. We also relate to literatures regarding minority and immigrant group occupational specialization.[2] Our setting resembles but differs substantially from the standard theory of discrimination, as we analyze environments when groups are economically integrated but culturally isolated.[3] These important differences shape whether minority group isolation provides a comparative advantage for self-employment. We also relate to literature on the importance of social interactions for economic behavior within or outside of the workplace.[4]

[1] See Chung and Kalnins (2006), Fairlie (2008), Fairlie et al. (2010), Hunt (2011), Patel and Vella (2013), and Kerr and Kerr (2017, 2018). Fairlie and Lofstrom (2013) and Kerr (2013) provide reviews.

[2] Kuznets (1960) observes that "all minorities are characterized, at a given time, by an occupational structure distinctly narrower than that of the total population and the majority." Our theory is also related to the concept of ethnic capital (Borjas, 1992, 1995) and group assimilation (Lazear, 1999). Patel et al. (2013) provide a review. Classics from the sociology literature include Light and Bonacich (1991) and Light and Gold (2000), and our NBER working paper provides further references. Sharma (2019) provides a recent UK depiction.

[3] To illustrate how market interaction can take place without social interaction, consider a scene from Shakespeare's *The Merchant of Venice* (Act 1, Scene III) depicting the social divide between the Christians and Jews in Renaissance Europe. Following a negotiation over a large loan to a Christian man who has always scorned him, the Jewish moneylender Shylock comments: "I will buy with you, sell with you, talk with you, walk with you, and so following; but I will not eat with you, drink with you, nor pray with you."

[4] Important examples include Granovetter (1973), Glaeser et al. (1996), and Glaeser and

2 A Model of Entrepreneurial Clustering

2.1 Model Set-Up

We construct a simple model to illustrate how social isolation and small group size can generate ethnic entrepreneurial clustering when social interactions and production are complementary. To keep the model tractable and intuitive, we make several strong assumptions. Everyone has equal ability and is divided into two ethnic groups. Group A is the minority, with a continuum of individuals of mass N_A, and group B has mass $N_B > N_A$. Both groups have equal access to industries and there is no product market discrimination, but the groups are socially segregated and spend their leisure time separately. Social interactions are random within ethnic groups, such that each person interacts with a representative sample of individuals in their own group.

We analyze how these two ethnic groups sort across two industries. Industry 1 has a production structure where self-employed entrepreneurs obtain advantages through social interactions with other self-employed entrepreneurs in the same industry. When socializing during family gatherings and religious/cultural functions, entrepreneurs in this industry can mentor each other and exchange industry knowledge and professional advice. The more an entrepreneur socializes with other entrepreneurs, the more knowledge is exchanged. Industry 0, by contrast, exhibits constant returns to scale with worker productivity normalized to one. This industry can be equally comprised of individuals working in self-employment or in larger firms; the core assumption is that private social interactions do not have the same benefit in industry 0 as they do in industry 1.

More formally, define X_l for $l \in \{A, B\}$ as the fraction of the population in group l

Scheinkman (2002). Durlauf and Fafchamps (2006) and Durlauf and Ioannides (2010) provide broad reviews.

who are self-employed entrepreneurs in industry 1. Since social interaction is random within groups, a fraction X_l of the friends and family members of every individual in group l are also self-employed entrepreneurs in industry 1. For industry 1, denote individual entrepreneurial productivity in group l as $\theta\left(X_l\right)$. Our assumption that productivity increases when socializing with other entrepreneurs in industry 1 is formally stated as:

Assumption 1a *Entrepreneurial productivity in industry 1 increases in specialization:* $\theta' > 0$.

Denote aggregate output of industry 1 as Q_1, which is a function of the distribution (X_A, X_B):

$$Q_1\left(X_A, X_B\right) = X_A N_A \theta\left(X_A\right) + X_B N_B \theta\left(X_B\right). \tag{1}$$

Since social interaction plays no role for industry 0, its aggregate output is simply:

$$Q_0\left(X_A, X_B\right) = \left(1 - X_A\right) N_A + \left(1 - X_B\right) N_B. \tag{2}$$

Demand for the two industries need to be complementary enough to avoid the complications of multiple optima possibly generated by non-convexities. We simply assume them to be perfect complements via a Leontief utility function for consumers:

$$U\left(q_0, q_1\right) = \min\left(q_0, \frac{q_1}{v}\right), \tag{3}$$

where $v > 0$ is a preference parameter and q_0 and q_1 are individual consumption of each industry's output, respectively.

2.2 The Pareto Problem

We now describe the efficient outcome. Since the outputs of both industries have unitary income elasticities, distributional aspects can be ignored when characterizing the efficient outcome. The problem simplifies to choosing an industry distribution (X_A, X_B) that maximizes a representative utility function $U\left(Q_0\left(X_A, X_B\right), Q_1\left(X_A, X_B\right)\right)$. A marginal analysis is inappropriate since this is a non-convex optimization problem. We consider instead the most specialized industry distributions, where as many individuals as possible from a single group A or B are self-employed entrepreneurs in industry 1.

Figure 1 depicts the production possibilities for the two specialized distributions. Define $V\left(X_A, X_B\right) \equiv Q_1/Q_0$ as the ratio of industry outputs under the distribution (X_A, X_B). Along the curve with the kink $V\left(1, 0\right)$ in the figure, group A specializes as self-employed entrepreneurs in industry 1. Starting from a position on the far right where everyone works in industry 0, members of group A are added to the set of self-employed entrepreneurs in industry 1 as we move leftward along the x-axis. When the kink at $V\left(1, 0\right)$ is reached, all members of group A are self-employed entrepreneurs in industry 1. Thereafter, continuing leftward, members of group B are also added to industry 1 until $Q_0 = 0$. Similarly, along the curve with the kink $V\left(0, 1\right)$, group B first specializes as self-employed entrepreneurs in industry 1. Members of group B are added moving leftward along the x-axis until the kink at $V\left(0, 1\right)$, where all Bs are working in industry 1. Thereafter members of group A are also added until $Q_0 = 0$.

The curve with minority specialization is above the curve with majority specialization, so long as the need for self-employed entrepreneurs in industry 1 is sufficiently small. A large fraction of As are self-employed entrepreneurs in industry 1 when the minority specializes, allowing minority entrepreneurs to socialize mostly with other entrepreneurs in industry 1, improving productivity. The same is not true for the majority, since even if a large fraction of self-employed entrepreneurs in industry 1 are Bs, most Bs are nevertheless employed in industry 0.

The argument can be generalized to show that minority specialization is Pareto efficient so long as industry 1 is small enough. Perfect complementarity simplifies the problem of solving for the optimal allocation, since any bundle where industrial outputs are in the exact ratio v of the Leontief preferences (3) is strictly preferable to all other bundles that do not include at least as much of each industry. The Pareto optimal distribution (X_A, X_B) must therefore satisfy $v = V(X_A, X_B)$. Define the total number of entrepreneurs in the population as $M \equiv X_A N_A + X_B N_B$. It follows that:

Proposition 1 *If* $v \leq V(1,0)$*, all self-employed entrepreneurs in industry* 1 *belong to minority group* A*.*

Proof: Take the distribution $(X_A, 0)$ where X_A is such that $v = V(X_A, 0)$. This is feasible since $v \leq V(1,0)$. Assume by contradiction that it is not the uniquely efficient distribution. Then there exists an alternative distribution (X_A', X_B') with $Q_1' \geq Q_1$ and $Q_0' \geq Q_0$. Given $Q_0' \geq Q_0$, it follows that $M' \leq M$, or equivalently, $X_A' N_A + X_B' N_B \leq X_A N_A$, which implies $X_A' \leq X_A$ and $X_B' < X_A$, with $X_A' < X_A$ if $X_B' = 0$. Manipulating the expression for Q_1':

$$\begin{aligned} Q_1' &= (M' - X_B' N_B)\,\theta(X_A') + X_B' N_B \theta(X_B') \\ &< (M - X_B' N_B)\,\theta(X_A) + X_B' N_B \theta(X_A) = Q_1 \end{aligned}$$

This contradicts $Q_1' \geq Q_1$.■

The efficient outcome requires that a single group specializes as self-employed entrepreneurs in industry 1, and importantly, which group specializes is not arbitrary. Minority specialization is more efficient since the minority's social isolation enables entrepreneurs in A to socialize mostly with other entrepreneurs in their small isolated group. For $v \leq V(1,0)$, the transformation curve and the curve with minority specialization in Figure 1 coincide. Group A has absolute and comparative advantages as self-employed entrepreneurs in industry 1. If the demand for industry 1 is sufficiently great, however, then the minority is too small to satisfy demand by themselves. In the special case when $v = V(0, 1)$, the demand for industry 1 is great enough for group B to specialize completely. In this case minority involvement would dilute the majority's productivity advantage, and the Pareto efficient solution is for Bs to specialize in being self-employed entrepreneurs in industry 1.

Corollary *If* $v = V(0, 1)$*, all self-employed entrepreneurs in industry* 1 *belong to the majority,* B*.*

Thus, the relationship between group size and productivity is not monotonic, and the group with the absolute advantage is the group with a population size that most closely adheres to the size of industry 1. Other production possibilities generated by more unspecialized distributions, such as $X_A = X_B$, are not displayed in Figure 1. Our online theoretical appendix proves that a convex production function in social interactions ($\theta'' > 0$) is sufficient to ensure that at least one group specializes, in which case the efficient frontier is the outer envelope of the curves shown in Figure 1. Consequently, above a certain value of v, there is a discrete jump from minority specialization to majority specialization.

2.3 Model Discussion

This simple model provides a stark economic environment for considering how isolated social interactions impact the sorting of ethnic groups over industries. While our model considers only two industries, this simplification is not as limiting as it may first appear. The model captures a setting where a small industry of self-employed entrepreneurs can benefit through non-work interactions. Allowing the baseline industry 0 to be an aggregate of many constant-returns-to-scale industries would still lead to the efficient solution being for the small ethnic group to specialize in being the self-employed entrepreneurs if their group size matches the demand preferences for industry 1. In fact, framed this way, the baseline industry 0 would be expected to be quite large to any one industry, making it more likely that the minority group should specialize.

Another obvious simplification is that we only have two ethnic groups, whereas the world is much more diverse. Yet, a complex model allowing for several small industries and also several minority ethnic groups would lead to the same conclusions. For example, consider an economy with industries $1a$ and $1b$ that have equal demand and display the same productivity benefit for social interaction. Also allow there to be two minority groups of equal size. If the demands for industries $1a$ and $1b$ are sufficiently small, then the efficient outcome is for one minority group to specialize in being self-employed entrepreneurs in $1a$, and for the other minority group to specialize in $1b$. Which minority group specializes in which sector is arbitrary. In this multi-sector economy with sector-specific skills, otherwise-similar groups consequently specialize in different business sectors. Pushing further, if the economy has several small industries of varying sizes that benefit from these social interactions, and multiple minority ethnic groups, the efficient outcome will be characterized by minority groups specializing in specific self-employment industries as much as possible.

Our online theoretical appendix also provides several formal extensions to the model: analysis of competitive market outcomes; occupational stratification and the dynamics of group specialization; endogenous social interactions and marriage markets; and the potential formation of splinter groups. These extensions reinforce the core insight that a small and socially isolated group can have advantages for industry specialization towards self-employment.

An additional extension in the appendix considers individual heterogeneity in ability and earnings and predicts that members of an ethnic group can achieve greater earnings when entering a common self-employed industrial specialization. This is important for distinguishing this choice to specialize from discrimination against minority groups. The empirical work of Patel and Vella (2013) show a positive earning relationship for immigrant groups and common group occupational choices, and we note below some complementary evidence from our own data.[5]

3 Analysis of US Entrepreneurial Stratification

3.1 US Census of Populations Data

We analyze the 2000 Census of Populations using the Integrated Public Use Microdata Series (IPUMS). We focus on the 5% state-level sample, and we use person weights to create population-level estimates. We also use the 1980 5% sample to construct one set of instruments, while a second set of instruments uses 1991 information on the United Kingdom from IPUMS-International.

We define ethnic groups using birthplace locations and, to a lesser extent, language spoken. We merge some related birthplace locations (e.g., combining England, Scotland, Wales, and non-specific U.K. designations into a single group). We also utilize the detailed language variable to create sub-groups among some larger birthplaces (e.g., separating Gujarati and Punjabi Indians into separate groups). Our preparation develops 146 potential ethnic groups from 198 birthplace locations, although most of our empirical work focuses on 77 larger ethnic groups that have at least one industry with ten or more IPUMS observations (equivalent to about 200 workers in the industry

[5] The favorable economic outcome does not necessarily carry over to utility. Related work includes Chiswick (1978), Borjas (1987), Simon and Warner (1992), Rauch (2001), Mandorff (2007), Bayer et al. (2008), Beaman (2012), and Cadena et al. (2015).

nationally).

We assign industry classification and self-employment status through the industry and class-of-work variables. IPUMS uses a three-digit industry classification to categorize work setting and economic sector of employment. Industry is distinct from an individual's technical function or occupation, and those operating in multiple industries are assigned to the industry of greatest income or amount of time spent. We utilize the 1990 IPUMS industry delineations for temporal consistency. The class-of-work variable identifies self-employed and wage workers[6], and we define a "cluster" as an {industry, class of work} pairing. For example, a self-employed hotelier is classified differently than a wage earner in the hotel industry. The sample excludes those whose self-employment status is unknown and industries without self-employment.[7]

We retain males between 30 and 65 years old who are living in metropolitan statistical areas. We further require that immigrants to have migrated between 1968 and 1990 and to have 20 years of age or older at entry. The start year of 1968 reflects the Immigration Reform Act of 1965, and the final entry of 1990 was chosen to avoid issues related to migration for temporary employment (which is employer-sponsored in America). Our final sample contains 1,604,350 observations, representing 34,984,436 people. Of these, 143,327 observations, representing 3,141,080 people, are immigrants.

[6] In the IPUMS data, self-employment is assigned when it is the main activity of an individual (e.g., not capturing academics who consulting part-time). The definition includes both owners of employer firms and sole proprietors.

[7] Our final sample includes 200 industries. We are cautious to not rely on very aggressive definitions of industry boundaries, even if this leads us to underestimate some concentration. For example, Greek restaurateurs will sort into Greek restaurants and Chinese restaurateurs into Chinese restaurants, independent of social relationships, but we consider the restaurant industry as a whole to avoid taste-based factors. Similarly, we look at industries on a national basis, even though additional clustering happening at localized levels for some industries (e.g., taxi cabs). We use this uniform approach to be consistent over industries, versus for example defining the motel industry in a different way from taxi cabs, and because ethnic connections can provide long-distance knowledge access (e.g., Rauch, 2001; Agrawal et al., 2008).

3.2 Clustering in Entrepreneurial Activities

We design "overage" ratios to quantify for an ethnic group the heightened rate of self-employment it displays for a particular industry and also across the full range of industries. Our primary metrics focus on the specialization evident among self-employed individuals only, while robustness checks build samples combining wage earners and self-employed.[8]

We first define $OVER_{lk}$ as the ratio of an ethnic group l's concentration in an industry k to the industry k's national employment share. Thus, if an ethnic group l has N_l total workers and N_l^k workers in industry k, then $X_l^k = N_l^k/N_l$ and $OVER_{lk} = X_l^k/X^k$. This baseline metric measures the over- or under-representation of the ethnic group for a specific industry, and by definition both cases exist for an ethnic group across the full range of industries.

To aggregate these industry-level values into an overall measure of industry concentration for an ethnic group, our primary metric takes a weighted average using the share of the group's self-employment by industry as the weight:

$$OVER1_l = \sum_{k=1}^{K} OVER_{lk} X_l^k. \tag{4}$$

Our estimations ultimately use the log value of this $OVER1$ metric. We also test the following variants:

1. Weighted average over the three largest industries for ethnic group l: $OVER2_l = \sum_{k'=1}^{3} OVER_{lk'} X_l^{k'} / \sum_{k'=1}^{3} X_l^{k'}$, where $k' = k$ such that $\sum_{k'=1}^{3} N_l^{k'}$ is maximized.

[8] It may seem appealing to use wage earners instead as a counterfactual to self-employed workers. This approach is not useful, however, as ethnic entrepreneurs show a greater tendency to hire members of their own ethnic groups into their firms (e.g., Andersson et al., 2014a,b; Åslund et al., 2014; Kerr et al., 2015).

2. Weighted average over the three largest industry-level overages for ethnic group l: $OVER3_l = \sum_{k'=1}^{3} OVER_{lk'} X_l^{k'} / \sum_{k'=1}^{3} X_l^{k'}$, where $k' = k$ such that $\sum_{k'=1}^{3} OVER_{lk'}$ is maximized.

3. Maximum overage: $OVER4_l = \max_l [OVER_{lk}]$.

These calculations measure extreme values, and we need to be careful about small sample size. Our earlier requirement that ethnicities have at least one industry with ten or more IPUMS observations avoids spurious clusters that could appear in small ethnic groups and obscure industries due to very small sample size or small population size.

We investigate our entrepreneurial concentration hypotheses over the 77 primary ethnic groups. $OVER1_l$ takes the weighted sum across industries, while $OVER2_l$ considers the three largest industries for an ethnic group. In most cases, $OVER2_l$ is bigger than $OVER1_l$ as concentration is often linked to substantial numerical representation; some exceptions happen when an ethnic group is focused on bigger industries. We calculate our metrics of extreme values, captured in $OVER3_l$ and $OVER4_l$, over ethnic group-industry clusters where we have at least ten observations.

Figure 2 displays the 16 ethnic groups with the highest $OVER1_l$ metric. There is substantial entrepreneurial clustering, with Yemeni immigrants displaying the overall highest industrial concentration for entrepreneurship. Appendix Tables 1a and 1b give a detailed list of overage ratios for each ethnic group and the industries with the largest overage ratio. In most cases, the industry where the ethnic group displays the highest concentration for self-employment is the same as the industry where the ethnic group shows the highest concentration for total employment. Appendix Tables 2a and 2b document the strong correlations between the four overage metrics.

3.3 Social Isolation and In-Marriage Rates

We measure social isolation and concentrated group interactions through within-group marriage rates evident among ethnicities. This metric is a strong proxy if sorting in the marriage market is similar to sorting in other social relationships.[9] High marriage rates within an ethnic group, also termed in-marriage, suggest greater social isolation and stratification.

We calculate in-marriage rates for ethnicities using a second dataset developed from IPUMS. We focus on women and men who immigrated to the United States between the ages of 5 and 15 and who are between ages 30 and 65 in 2000, and exclude individuals already married at the time of immigration. Importantly, this sample is mutually exclusive from the earlier sample used to calculate our overage metrics.

Most immigrant groups are socially segregated with respect to marriage, some very strongly so. With random matching for marriage and equal male and female migration, in-marriage rates would roughly equal a group's fraction of the overall population. Group in-marriage rates (also shown in Appendix Table 1a) are typically much higher, almost always exceeding 50%. Pairwise correlations of 0.51 and 0.60 exist for in-marriage rates and the $OVER1_l$ and $OVER2_l$ metrics, respectively.

3.4 OLS Empirical Tests

To quantify whether smaller and more-socially isolated ethnic groups have greater industrial concentration for entrepreneurship, we use the following regression approach:

$$OVER1_l = \alpha + \beta_1 SIZE_l + \beta_2 ISOL_l + \varepsilon_l, \tag{5}$$

[9] Representative work includes Kennedy (1944), Bisin and Verdier (2000), and Bisin et al. (2004). Furtado (2010) shows how inter-marriage can provide better access to the formal labor market, and Furtado and Theodoropoulos (2011) consider shifts in likelihood of inter-marriage by when someone migrates to the US. Furtado and Trejo (2013) provide an extended review.

where $SIZE_l$ is the negative of the log value of group size and $ISOL_l$ is the log in-marriage rate of the group. We take the negative of group size so that our theoretical prediction is that β_1 and β_2 are positive. We report all coefficients in unit standard deviation terms for ease of interpretation. Our baseline regressions winsorize variables at their 10% and 90% levels to guard against outliers, weight estimations by log ethnic employment for each group, and report robust standard errors.

Column 1 of Table 1 measures that a one standard-deviation decrease in group size is correlated with a 0.63 increase in average entrepreneurial concentration across all industries. Similarly, a one standard-deviation increase in the in-marriage rate is correlated with a 0.52 standard-deviation increase in overage. The next columns consider robustness checks on our metric design. Column 2 uses the full worker sample, Column 3 calculates overages only relative to immigrant populations by excluding natives from the denominator shares, and Column 4 adds rural workers into the self-employment overage calculations.[10] The results are very robust to these adjustments. Column 5 shows that a focus on the three largest industries for an ethnic group (i.e., $OVER2_l$ discussed above) increases the relative importance of social isolation for predicting overages. Columns 6 and 7 examine extreme values using the $OVER3_l$ and $OVER4_l$ metrics defined above. These extreme values show a weaker connection to group size, placing even more prominence on group isolation.[11]

Table 2 shows additional robustness checks on the $OVER1_l$ outcomes. Columns 2 and 3 drop sample weights and winsorization steps, respectively, Column 4 introduces fixed effects for each origin continent, Column 5 uses a median regression format, and

[10] Faggio and Silva (2014) analyze differences in self-employment alignment to entrepreneurship in urban and rural areas.

[11] We obtain similar results when modifying of our overage measures with industry-level propensities for being an employer firm vis-à-vis sole proprietors using data from the Survey of Business Owners.

Column 6 bootstraps standard errors. These last two columns should be compared to Column 2 given their unweighted nature.

Columns 7 and 8 test whether smaller sample sizes for ethnic groups mechanically create concentration ratios, using Monte Carlo simulations. In one version, used for Column 7, we draw industry and self-employment status independently from each other, which means that we tend to predict the same self-employment rates across industries. In a second version used in Column 8, we jointly draw the two components such that we mimic the industry-by-industry entrepreneurship rates observed in the data. From these 1000 Monte Carlo simulations, we calculate for each ethnic group the average observed overage. Introducing these controls does not impact our estimations except that the size relationship diminishes modestly.

We further test the relationships of relative size and isolation on entrepreneurial clustering by using non-parametric regressions. We partition our size and isolation variables into terciles and create indicator variables for each combination of {smallest size, medium, largest size} and {most isolated, medium, least isolated}, and assign ethnic groups that fall into [largest size, least isolated] as the reference category.

The results continue to support the theory, as depicted in Figure 3. The [smallest size, most isolated] groups have entrepreneurial concentrations that are 2.5 standard deviations greater than the [largest size, least isolated] groups (see Table 3 for detailed coefficients). Equally important, the pattern of coefficients across the other indicator variables shows the relationships are quite regular and not due to a few outliers. For example, holding the ethnic group size constant, higher levels of social isolation strongly and significantly correspond to larger overages. Flipping it around and holding social isolation, smaller group sizes promote greater concentration within each isolation category, with the exception of the least socially isolated tercile.

3.5 IV Empirical Tests: 1980 Values

We next consider IV specifications to test against reverse causality concerns (e.g., where isolated business ownerships lead to greater social isolation or lower group sizes) or omitted variables.[12] We use two sets of instruments. The first set builds upon a dynamic version of our model where initial conditions have lasting and persistent impacts for immigrant groups; this has been demonstrated empirically by Patel and Vella (2013). We thus use the lagged 1980 values of ethnic group size and in-marriage rates in the United States to instrument for 2000 levels. The ethnic divisions in 1980 are less detailed than in 2000, and in some cases, the same 1980 value must be applied to several 2000 ethnic groups. We thus cluster standard errors around the 43 groups present in the 1980 data, with other aspects of the IV estimations being the same as OLS specifications.

The first-stage results with this instrument set are quite strong. The first two columns of Table 4 show that these instruments have very strong individual predictive power and a combined joint F-statistic of 24.[13] The exclusion restriction requires that the 1980 group sizes and in-marriage levels only impact entrepreneurship in 2000 to the extent that they shape current group size and social isolation. This seems reasonable, although some 1980 respondents may still be employed in 2000, and this may carry with it persistence that violates the exclusion restriction.

The second-stage results in Column 3 are quite similar to the OLS findings. The IV specifications suggest that a one standard-deviation decrease in ethnic group size increases overage by 0.76 standard deviations. A one standard-deviation increase in

[12] An example would be an ethnic group disproportionately located in areas with stringent employment verification procedures (e.g., Amuedo-Dorantes and Bansak, 2012; Orrenius and Zavodny, 2016), leading to more social and workplace isolation.

[13] The F-statistic comes from the Kleibergen-Paap Wald rank F-statistic used when standard errors are clustered or robust and is based off the Cragg-Donald F-test for weak instrumentation.

isolation leads to a 0.52 standard-deviation increase in entrepreneurial concentration. These results are well-measured and economically important. The size coefficient grows modestly from its OLS baseline, while the in-marriage rate coefficient declines slightly. The results are precisely enough estimated that we can reject at a 5% level the null hypothesis in Wu-Hausman tests that the instrumented regressors are exogenous. These IV results strengthen the predictions of our theory that smaller, more isolated groups are more conducive to entrepreneurial clustering.

3.6 IV Empirical Tests: Gravity Model and UK Values

Our second IV approach uses as instruments the predicted ethnic group size from a gravity model and in-marriage rates from the United Kingdom in 1991. To instrument for ethnic group size, we use a gravity model to quantify predicted ethnic size based upon worldwide migration rates to the United States. The original application of gravity models was to trade flows, where studies showed that countries closer to each other and with larger size tended to show greater trade flows, similar to the forces of planetary pull. This concept has also been applied to the migration literature, and we similarly model

$$SIZE_l = \alpha + \beta_1 DIST_l + \beta_2 POP_l + \varepsilon_l, \tag{6}$$

where $DIST_l$ is the log distance to the United States from the origin country and POP_l is the log population of the origin country. For this purpose, we estimate log ethnic group size in the United States as the dependent variable (without a negative value being taken as in earlier estimations). Unsurprisingly, lower distance ($\beta_1 = -1.56$ (s.e.=0.22)) and greater population ($\beta_2 = 0.38$ (s.e.=0.06)) are strong predictors of ethnic group size in the United States. We take the predicted values from this regression for each ethnic group as our first instrument.

For our second instrument of in-marriage rates in the United States, we calculate the in-marriage rates in the 1991 UK Census of Populations. This approach is attractive as the social isolation evident in the United Kingdom a decade before our study is only likely to be predictive of US self-employment rates to the extent that the British isolation captures a persistent trait of the ethnic group. The limitation of this instrument is that we are only able to calculate this for 24 broader ethnic divisions. We map our observations to these groups and cluster the standard errors at the UK group level.

Columns 4-5 of Table 4 again report the first-stage relationships. The instruments remain individually predictive of their corresponding endogenous regressor, and they have a joint F-statistic of 35.5. Similar to the 1980 US instruments, the minimum 2SLS relative bias that can be specified is less than 10%. This implies that we can specify a very small bias and still reject the null hypothesis that the instruments are weak. The bias level is determined by the minimum eigenvalue statistic and the Stock-Yogo 2SLS size of the nominal 5% Wald test.

The second-stage results are again comparable to our core OLS findings. The size results are a bit lower than OLS, while the social isolation effects are even stronger than OLS, with elasticities of around 0.67. We now reject at a 10% level that the instrumented regressors are exogenous.

Tables 5-7 show robustness checks with the two IV approaches. Results are very similar with simple adjustments like excluding sample weights, dropping winsorization, or using bootstrapped standard errors. The results with simulated overage controls are very similar for the first instrument set, and we can only instrument for social isolation in the second set as the predicted size relationship in the gravity model has an insufficient first stage when including the simulated metrics. Intuitively, both the instrument and predicted overage are being built upon the same data, making it hard to separate them.

Further analyses show comparable patterns with the alternative overage metric designs. The results for social isolation are robust in all specifications. Those for group size are mostly robust, with a few exceptions with the predicted size instruments. We also find very similar results when expanding the gravity equation to have a squared distance term or an indicator for Canada and Mexico as bordering countries or when using underlying components of the gravity equation as direct instruments.

In summary, the OLS and IV variants provide consistent support to the model. The strongest findings are those for social isolation, which is a very strong predictor of immigrant entrepreneurial concentration. The weight of the evidence also supports that smaller group sizes promote entrepreneurial concentration.

3.7 Earnings Estimations

Our model makes an additional prediction that members of an ethnic group can achieve greater earnings when entering a common entrepreneurial setting. In our framework, social complementarities produce a positive relationship between earnings and entrepreneurship at the group level. This prediction is in direct contrast to what would be expected if discrimination in the marketplace is the most important factor leading to segmented group self-employment. The empirical work of Patel and Vella (2013) strongly shows a positive earning relationship for immigrant groups and common group occupational choices using the 1980-2000 Census of Populations data. To close the loop for this paper, we thus provide a brief analysis of earnings and refer readers to these complementary pieces for additional evidence.

Table 8 provides individual-level estimations of the earnings relationship. The outcome variable is the log yearly income of individuals. We report three core explanatory variables. The first is whether the individual is self-employed. The second is the percentage of an individual's ethnic group who are self-employed (similar to the values reported in Table A1a), regardless of industry. Third, we measure the share of the individual's ethnic group that is employed in the industry of the focal individual. With the model developed, we anticipate both of these group measures to have positive predictive power. For natives, these latter variables are simply measured over the whole US-born population.

Our estimations also include many unreported controls for individuals that relate to earnings. We include fixed effects for PUMA geographical locations and for industries. We also control for high-school and college education, whether the individual is a native or an immigrant, whether the individual is fluent in the English language, and fixed effects for seven age categories and seven age-at-immigration categories. Regressions cluster standard errors by ethnic group and use IPUMS sample weights.

The first three columns show that all three elements are predictive of earnings. Being self-employed (a binary measure) is directly associated with a 3% increase in total earnings in the cross-section. A 1% increase in the rate of overall self-employment for an ethnic group connects to a 1% increase in total earnings. To aid interpretation, the bottom of the table also provides the standard deviation x beta coefficient for group-level variables; a one standard-deviation increase (0.0255) in group self-employment connects to 3% higher earning. Similarly, looking at ethnic group concentration for the individual in his particular industry, a 1% increase in group concentration connects to a 0.6% increase in total earnings. In standard-deviation terms, the relative effect of 5% is even larger than the 3% for group self-employment. Columns 4-6 show similar outcomes when we exclude those in professional occupations and holders of doctorate degrees.

These results support the model's structure and are consistent with a potential positive benefit from immigrant entrepreneurial concentration. It is important for future theoretical and empirical work to consider both owners and employees of firms. Empirical work can particularly target employer-employee datasets to observe more detailed hiring and wage patterns; such work can also evaluate job transitions during the assimilation of new members of ethnic groups, perhaps ultimately leading to starting their own business.

4 Conclusions

A striking feature of entrepreneurship is the degree to which immigrants of different ethnic backgrounds cluster into self-employment in different industries. These concentrations are sufficiently visible to be captured in popular culture (e.g., the Indian immigrant entrepreneur Apu who runs the convenience store in *The Simpsons*), and the cumulative magnitudes can be shocking: the Asian American Hotel Owners Association claims to be the largest hotel owners association in the world and represent half of the hotels in the United States. Yet, while noticeable, the economic implications of these tendencies are underexplored.

Our model outlines how the social interactions of small, socially isolated groups can give rise to this self-employment pattern by reducing the cost of acquiring sector-specific skills. Our online appendix explores several extensions to the basic framework, and many other avenues for future research exist. A fruitful path would be to model the intergenerational transmission of skills and to follow occupational structure and entrepreneurial persistence across generations. This interaction mechanism can also be applied to the study of the transmission of other types of skills beyond entrepreneurship.

Empirically, the Census data confirm small and, especially, socially isolated immigrant groups in the United States display heightened entrepreneurial clustering. Further quantifying these forces in employer-employee data and firm operating data are important to understand hiring patterns, career trajectories, and market power. The recent US patterns resemble many earlier observations of the economic success and social isolation of specialized minority groups throughout history. We hope this study can be replicated in settings outside of the United States given its general nature (Fairlie et al., 2010).

References

[1] Agrawal, Ajay, Devesh Kapur, and John McHale. 2008. How do spatial and social proximity influence knowledge flows? Evidence from patent data. *Journal of Urban Economics* 64: 258-269.

[2] Amuedo-Dorantes, Catalina, and Cynthia Bansak. 2012. The labor market impacts of mandated employment verification systems. *American Economic Review* 102(3): 543-548.

[3] Andersson, Fredrik, Monica Garcia-Perez, John Haltiwanger, Kristin McCue, and Seth Sanders. 2014. Workplace concentration of immigrants. *Demography* 51(6): 2281-2306.

[4] Andersson, Fredrik, Simon Burgess, and Julia Lane. 2014. Do as the neighbors do: The impact of social networks on immigrant employment. *Journal of Labor Research.*

[5] Åslund, Olof, Lena Hensvik, and Oskar Skans. 2014. Seeking similarity: How immigrants and natives manage in the labor market. *Journal of Labor Economics* 32(3): 405-442.

[6] Åstebro, Thomas, Holger Herz, Ramana Nanda, and Roberto Weber. 2014. Seeking the roots of entrepreneurship: Insights from behavioral economics. *Journal of Economic Perspectives* 28: 49-70.

[7] Bayer, Patrick, Stephen Ross, and Giorgio Topa. 2008. Place of work and place of residence: Informal hiring networks and labor market outcomes. *Journal of Political Economy* 116: 1150-1180.

[8] Beaman, Lori. 2012. Social networks and the dynamics of labor market outcomes: Evidence from refugees resettled in the US. *Review of Economic Studies* 79: 128-161.

[9] Bisin, Alberto, and Thierry Verdier. 2000. Beyond the melting pot: Cultural transmission, marriage, and the evolution of ethnic and religious traits. *Quarterly Journal of Economics* 115: 955-988.

[10] Bisin, Alberto, Giorgio Topa, and Thierry Verdier. 2004. Religious intermarriage and socialization in the United States. *Journal of Political Economy* 112: 615-664.

[11] Borjas, George. 1987. Self-selection and the earnings of immigrants. *American Economic Review* 80: 531-553.

[12] Borjas, George. 1992. Ethnic capital and intergenerational mobility. *Quarterly Journal of Economics* 107: 123-150.

[13] Borjas, George. 1995. Ethnicity, neighborhoods and human capital externalities. *American Economic Review* 85: 365-390.

[14] Cadena, Brian, Brian Duncan, and Stephen Trejo. 2015. The labor market integration and impacts of U.S. immigrants. In *Handbook of the Economics of International Migration*, edited by Barry Chiswick and Paul Miller. Amsterdam: North Holland. 1197-1259.

[15] Chiswick, Barry. 1978. The effect of Americanization on the earnings of foreign-born men. *Journal of Political Economy* 86: 897-921.

[16] Chung, Wilbur, and Arturs Kalnins. 2006. Social capital, geography, and the survival: Gujarati immigrant entrepreneurs in the U.S. lodging industry. *Management Science* 52(2): 233-247.

[17] Durlauf, Steven, and Marcel Fafchamps. 2006. Social capital. In *Handbook of Economic Growth*, edited by Philippe Aghion and Steven Durlauf. Amsterdam: North Holland.

[18] Durlauf, Steven, and Yannis Ioannides, 2010. Social interactions. *Annual Review of Economics* 2: 451-478.

[19] Faggio, Giulia, and Olmo Silva. 2014. Self-employment and entrepreneurship in urban and rural labour markets. *Journal of Urban Economics* 83(1): 67-85.

[20] Fairlie, Robert. 2008. *Estimating the Contribution of Immigrant Business Owners to the U.S. Economy.* Small Business Administration, Office of Advocacy Report.

[21] Fairlie, Robert, Harry Krashinsky, and Julie Zissimopoulos. 2010. The international Asian business success story? A comparison of Chinese, Indian and other Asian businesses in the United States, Canada and United Kingdom. In *International Differences in Entrepreneurship*, edited by Josh Lerner and Antoinette Schoar. Chicago, IL: University of Chicago Press.

[22] Fairlie, Robert, and Magnus Lofstrom. 2013. Immigration and entrepreneurship. In *The Handbook on the Economics of International Migration*, edited by Barry Chiswick and Paul Miller. Amsterdam: North-Holland Publishing.

[23] Fairlie, Robert, and Alicia Robb. 2007. Families, human capital, and small business: Evidence from the Characteristics of Business Owners Survey. *Industrial and Labor Relations Review* 60: 225-245.

[24] Furtado, Delia. 2010. Why does intermarriage increase immigrant employment? The role of networks. *B.E. Journal of Economic Analysis & Policy* 10(1): Article 101.

[25] Furtado, Delia, and Nikolaos Theodoropoulos. 2011. Interethnic marriage: A choice between ethnic and educational similarities. *Journal of Population Economics* 24(4): 1257-1279.

[26] Furtado, Delia, and Stephen Trejo. 2013. Interethnic marriages and their economic effects. In *International Handbook on the Economics of Migration*, edited by Amelie Constant and Klaus Zimmermann. Northampton, MA: Edward Elgar, 276-292.

[27] Glaeser, Edward, Bruce Sacerdote, and José Scheinkman. 1996. Crime and social interactions. *Quarterly Journal of Economics* 111: 507-548.

[28] Glaeser, Edward, and José Scheinkman. 2002. Non-market interaction. In *Advances in Economics and Econometrics: Theory and Applications, Eight World Congress*, edited by Mathias Dewatripont, Lars Peter Hansen, and Stephen Turnovsky. Cambridge, UK: Cambridge University Press.

[29] Granovetter, Mark. 1973. The strength of weak ties. *American Journal of Sociology* 78: 1360-1380.

[30] Hunt, Jennifer. (2011). Which immigrants are most innovative and entrepreneurial? Distinctions by entry visa. *Journal of Labor Economics* 29(3): 417-457.

[31] Kennedy, Ruby. 1944. Single or triple melting-pot? Intermarriage trends in New Haven, 1870-1940. *The American Journal of Sociology* 49: 331-339.

[32] Kerr, Sari Pekkala, and William Kerr. 2017. Immigrant entrepreneurship. In *Measuring Entrepreneurial Businesses: Current Knowledge and Challenges*, edited by John Haltiwanger, Erik Hurst, Javier Miranda, and Antoinette Schoar. Chicago, IL: University of Chicago Press.

[33] Kerr, Sari Pekkala, and William Kerr. 2018. Immigrant entrepreneurship in America: Evidence from the Survey of Business Owners 2007 & 2012. Working Paper no. 24494, NBER Cambridge, MA.

[34] Kerr, Sari Pekkala, William Kerr, and William Lincoln. 2015. Skilled immigration and the employment structures of U.S. firms. *Journal of Labor Economics* 33(S1): S147-S186.

[35] Kerr, William. 2013. U.S. high-skilled immigration, innovation, and entrepreneurship: Empirical approaches and evidence. Working Paper no. 19377, NBER, Cambridge, MA.

[36] Kerr, William, and Martin Mandorff. 2018. Social networks, ethnicity, and entrepreneurship. Working Paper no. 21597 (revised), NBER, Cambridge, MA.

[37] Kihlstrom, R., and Jean-Jacques Laffont. 1979. A general equilibrium entrepreneurial theory of firm formation based on risk aversion. *Journal of Political Economy* 87: 719-748.

[38] Kuznets, Simon. 1960. Economic structure and life of the Jews. In *The Minority Members: History, Culture, and Religion*, edited by Louis Finkelstein. Philadelphia, PA: Jewish Publication Society of America.

[39] Lazear, Edward. 1999. Culture and language. *Journal of Political Economy* 107: 95-126.

[40] Lazear, Edward. 2005. Entrepreneurship. *Journal of Labor Economics* 23: 649-680.

[41] Light, Ivan, and Edna Bonacich. 1991. *Immigrant Entrepreneurs: Koreans in Los Angeles, 1965-1982.* Berkeley, CA: University of California Press.

[42] Light, Ivan, and Steven Gold. 2000. *Ethnic Economies.* Bingley, UK: Emerald Group.

[43] Lucas, Robert. 1978. On the size distribution of business firms. *Bell Journal of Economics* 9: 508-523.

[44] Mandorff, Martin. 2007. Social networks, ethnicity, and occupation. University of Chicago Ph.D. Dissertation.

[45] Manski, Charles. 1993. Identification of endogenous social effects: The reflection problem. *Review of Economic Studies* 60: 531-542.

[46] Orrenius, Pia, and Madeline Zavodny. 2016. Do state work eligibility verification laws reduce unauthorized immigration? *IZA Journal of Migration* 5:5.

[47] Patel, Krishna, Yevgeniya Savchenko, and Francis Vella. 2013. Occupational sorting of ethnic groups. In *International Handbook on the Economics of Migration.* Edward Elgar Publishing.

[48] Patel, Krishna, and Francis Vella. 2013. Immigrant networks and their implications for occupational choice and wages. *Review of Economics and Statistics* 95(4): 1249-1277.

[49] Rauch, James. 2001. Business and social networks in international trade. *Journal of Economic Literature* 39: 1177-1203.

[50] Sharma, Babita. 2019. *The Corner Shop: Shopkeepers, the Sharmas and the Making of Modern Britain.* Hodder & Stoughton.

[51] Simon, Curtis, and John Warner. 1992. Matchmaker, matchmaker: The effect of old boy networks on job match quality, earnings and tenure. *Journal of Labor Economics* 10(3): 306-330.

Table 1: OLS estimations

	Log weighted average overage across all industries [OVER1]				Log weighted average overage using three largest industries for ethnic group [OVER2]	Log average of three largest overage ratios for ethnic group [OVER3]	Log largest overage ratio for ethnic group [OVER4]
	Baseline estimation	Using total worker sample	Excluding natives from denominator shares	Including rural workers			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Inverse of log ethnic group size	0.634	0.595	0.398	0.602	0.375	0.130	0.068
(small groups have larger values)	(0.069)	(0.073)	(0.080)	(0.072)	(0.080)	(0.076)	(0.076)
Log isolation of ethnic group	0.519	0.514	0.578	0.529	0.640	0.722	0.706
	(0.067)	(0.066)	(0.083)	(0.068)	(0.072)	(0.070)	(0.075)
R-Squared value	0.61	0.53	0.47	0.58	0.51	0.53	0.51

Notes: Estimations describe the OLS relationship between industry concentration for ethnic entrepreneurship and ethnic group size and in-marriage isolation. The outcome variable in Columns 1-4 is the log weighted average overage ratio across industries for each ethnic group, where the weights are levels of self employment in each industry per group. Variables are winsorized at their 10%/90% levels and transformed to have unit standard deviation for interpretation. Regressions are weighted by log ethnic group employee counts in MSAs, include 77 observations, and report robust standard errors. Column 2 considers the metric that uses all employed workers for the ethnic group, Column 3 compares industry-level overages only to rates of other immigrant groups, and Column 4 includes rural workers in the sample. Column 5 restricts the overage measure to just the three largest self-employment industries for an ethnic group. Columns 6-7 consider extreme values among industries by ethnic group. These latter overages are done without reference to industry importance in terms of ethnic group self-employment, but they require at least ten observations exist for an ethnic group - industry cluster to be included.

Table 2: Robustness checks on OLS estimations

	Baseline estimation (Table 1, Column 1)	Without sample weights	Without winsorization	Including fixed effects for origin continent	Using median regression format	Using bootstrapped standard errors	Including simulated overage control1	Including simulated overage control2
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Inverse of log ethnic group size	0.634	0.630	0.629	0.552	0.586	0.630	0.509	0.524
(small groups have larger values)	(0.069)	(0.067)	(0.062)	(0.070)	(0.092)	(0.070)	(0.188)	(0.182)
Log isolation of ethnic group	0.519	0.521	0.511	0.485	0.529	0.521	0.550	0.538
	(0.067)	(0.065)	(0.066)	(0.091)	(0.091)	(0.070)	(0.070)	(0.067)
Log predicted overage1							0.155	
							(0.195)	
Log predicted overage2								0.123
								(0.186)
R-Squared value	0.612	0.626	0.629	0.650	0.428	0.626	0.577	0.612

Notes: See Table 1. Columns 2-6 provide robustness checks on the baseline specification. Regressions in Columns 5 and 6 are unweighted and should be referenced against Column 2. Column 5 reports pseudo R-squared values. Columns 7 and 8 include control variables for predicted overage ratios based upon 1000 Monte Carlo simulations. In these simulations, pools of similarly sized ethnic groups to our true sample are formed and randomly assigned industry and entrepreneurship status according to national propensities. From these random assignments, we calculate 1000 overage metrics for each ethnic group that exactly mirror our primary data construction. The average of these simulations is entered as a control variable. In the first version included in Column 7, self-employment status and industry status are separately randomized, such that we overall predict roughly the same self-employment rate in each industry. In the second version included in Column 8, self-employment status and industry are jointly drawn such that we overall replicate observed self-employment levels across industries.

Table 3: OLS relationships with non-parametric forms

	Log weighted average overage across all industries [OVER1]	Log weighted average overage across three largest industries [OVER2]	Log average of three largest overage ratios for ethnic group [OVER3]	Log largest overage ratio for ethnic group [OVER4]
	(1)	(2)	(3)	(4)
(0,1) Indicator: ethnic size in smallest third x	2.472	2.276	1.826	1.572
(0,1) Indicator: ethnic isolation in highest third	(0.188)	(0.168)	(0.155)	(0.180)
(0,1) Indicator: ethnic size in smallest third x	1.514	0.753	0.416	0.375
(0,1) Indicator: ethnic isolation in middle third	(0.271)	(0.380)	(0.368)	(0.362)
(0,1) Indicator: ethnic size in smallest third x	1.048	0.280	-0.654	-1.002
(0,1) Indicator: ethnic isolation in lowest third	(0.280)	(0.273)	(0.243)	(0.251)
(0,1) Indicator: ethnic size in middle third x	1.581	1.211	1.127	1.044
(0,1) Indicator: ethnic isolation in highest third	(0.322)	(0.374)	(0.253)	(0.260)
(0,1) Indicator: ethnic size in middle third x	0.908	0.573	0.351	0.338
(0,1) Indicator: ethnic isolation in middle third	(0.313)	(0.314)	(0.345)	(0.362)
(0,1) Indicator: ethnic size in middle third x	0.428	-0.038	-0.443	-0.542
(0,1) Indicator: ethnic isolation in lowest third	(0.228)	(0.220)	(0.276)	(0.306)
(0,1) Indicator: ethnic size in largest third x	0.802	0.944	0.927	0.767
(0,1) Indicator: ethnic isolation in highest third	(0.369)	(0.361)	(0.309)	(0.300)
(0,1) Indicator: ethnic size in largest third x	0.126	0.279	0.329	0.294
(0,1) Indicator: ethnic isolation in middle third	(0.312)	(0.334)	(0.297)	(0.299)
(0,1) Indicator: ethnic size in largest third x (0,1) Indicator: ethnic isolation in lowest third	Excluded group			
R-Squared value	0.57	0.49	0.55	0.54

Notes: See Table 1. Effects are measured relative to largest and least isolated ethnic groups.

Table 4: IV estimations

	Instrumenting with 1980 ethnic group size and in-marriage rates in United States			Instrumenting with predicted ethnic group size from gravity model and in-marriage rates in United Kingdom		
	First stage for group size	First stage for group isolation	Second stage results	First stage for group size	First stage for group isolation	Second stage results
	(1)	(2)	(3)	(4)	(5)	(6)
Instrument for size	0.877	-0.063		0.706	-0.018	
	(0.044)	(0.055)		(0.069)	(0.115)	
Instrument for isolation	-0.075	0.721		-0.142	0.587	
	(0.043)	(0.114)		(0.109)	(0.078)	
	F stat =	Bias =		F stat =	Bias =	
	23.6	<10%		35.5	<10%	
Inverse of log ethnic group size			0.757			0.487
			(0.077)			(0.132)
Log isolation of ethnic group			0.516			0.665
			(0.099)			(0.119)
Exogeneity test p-value			0.034			0.091

Notes: See Table 1. Estimations describe the IV relationship between industry concentration for ethnic entrepreneurship and ethnic group size and in-marriage isolation. The column headers indicate the instruments used. The 2SLS relative bias reports the minimum bias that can be specified and still reject the null hypothesis that the instruments are weak. This level is determined through the minimum eigenvalue statistic and Stock and Yogo's (2005) 2SLS size of nominal 5% Wald test. The null hypothesis in Wu-Hausman exogeneity tests is that the instrumented regressors are exogenous. The test statistic used is robust to clustering of standard errors. Regressions cluster standard errors by the 43 and 24 ethnic groups in the US 1980 and UK 1990 datasets used to build the respective instruments.

Table 5: Robustness checks on IV estimations

					Isolation IV Only		Double IV	
	Baseline estimation (Table 2, Col. 1 & 6)	Without sample weights	Without winsorization	Using bootstrapped standard errors	Including simulated overage control1	Including simulated overage control2	Including simulated overage control1	Including simulated overage control2
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	A. IV results using 1980 ethnic group size and in-marriage rates in United States							
Inverse of log ethnic group size	0.757	0.748	0.689	0.748	0.519	0.547	1.254	1.220
(small groups have larger values)	(0.077)	(0.072)	(0.084)	(0.085)	(0.232)	(0.116)	(0.355)	(0.332)
Log isolation of ethnic group	0.516	0.526	0.554	0.526	0.539	0.516	0.465	0.468
	(0.099)	(0.091)	(0.145)	(0.095)	(0.122)	(0.212)	(0.133)	(0.125)
F statistic	23.6	23.4	6.9	34.6	33.1	37.5	15.4	23.0
Exogeneity test p-value	0.034	0.043	0.100	0.011	0.915	0.912	0.014	0.012
	B. IV results using predicted group sizes and UK in-marriage rates							
Inverse of log ethnic group size	0.487	0.476	0.506	0.476	0.315	0.334	Insufficient	Insufficient
(small groups have larger values)	(0.132)	(0.123)	(0.091)	(0.105)	(0.185)	(0.179)	first stage	first stage
Log isolation of ethnic group	0.665	0.639	0.464	0.639	0.772	0.751		
	(0.119)	(0.111)	(0.089)	(0.135)	(0.089)	(0.091)		
F statistic	35.5	34.1	13.5	20.0	40.7	29.8		
Exogeneity test p-value	0.091	0.084	0.160	0.061	0.137	0.166		

Notes: See Tables 1 and 4.

Table 6: IV estimations with alternative metric designs

	Log weighted average overage across all industries [OVER1]				Log weighted average overage using three largest industries for ethnic group [OVER2]	Log average of three largest overage ratios for ethnic group [OVER3]	Log largest overage ratio for ethnic group [OVER4]
	Baseline estimation (Table 2, Col. 1 & 6)	Using total worker sample	Excluding natives from denominator shares	Including rural workers			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	A. IV results using 1980 ethnic group size and in-marriage rates in United States						
Inverse of log ethnic group size	0.757	0.636	0.491	0.730	0.531	0.272	0.193
(small groups have larger values)	(0.077)	(0.063)	(0.135)	(0.086)	(0.110)	(0.126)	(0.123)
Log isolation of ethnic group	0.516	0.469	0.771	0.532	0.696	0.759	0.720
	(0.099)	(0.104)	(0.113)	(0.097)	(0.091)	(0.087)	(0.107)
F statistic	23.6	54.4	23.6	23.6	23.6	23.6	23.6
Exogeneity test p-value	0.034	0.403	0.081	0.040	0.019	0.042	0.078
	B. IV results using predicted group sizes and UK in-marriage rates						
Inverse of log ethnic group size	0.487	0.466	0.386	0.444	0.132	0.075	0.043
(small groups have larger values)	(0.132)	(0.120)	(0.141)	(0.132)	(0.109)	(0.100)	(0.090)
Log isolation of ethnic group	0.665	0.550	0.696	0.712	0.861	0.905	0.853
	(0.119)	(0.177)	(0.130)	(0.122)	(0.125)	(0.104)	(0.088)
F statistic	35.5	10.5	35.5	35.5	35.5	35.5	35.5
Exogeneity test p-value	0.091	0.107	0.687	0.055	0.022	0.239	0.464

Notes: See Tables 1 and 4.

Table 7: IV results with alternative gravity model designs for predicted size

	Baseline estimation (Table 2, Column 6)	Including border in the gravity model	Including distance squared in the gravity model	Using distance and population as instruments	Using distance, population, and border as instruments	Using distance, population, and distance squared as instruments
	(1)	(2)	(3)	(4)	(5)	(6)
Inverse of log ethnic group size	0.487	0.483	0.483	0.522	0.524	0.522
(small groups have larger values)	(0.132)	(0.131)	(0.130)	(0.149)	(0.148)	(0.150)
Log isolation of ethnic group	0.665	0.665	0.665	0.680	0.624	0.673
	(0.119)	(0.120)	(0.120)	(0.111)	(0.084)	(0.083)
F statistic	35.5	36.2	35.8	22.2	17.0	17.0
Exogeneity test p-value	0.091	0.086	0.096	0.029	0.063	0.024
Overidentification test p-value				0.174	0.283	0.394

Notes: See Tables 1 and 4.

Table 8: Estimations for log yearly income of individual

	Baseline estimation			Excluding professionals and PhDs		
	(1)	(2)	(3)	(4)	(5)	(6)
Percent of self-employed in individual's ethnic group (1)	1.145 (0.334)		1.122 (0.335)	1.091 (0.347)		1.067 (0.349)
Share of group that is working in an individual's industry (2)		0.680 (0.205)	0.615 (0.201)		0.624 (0.210)	0.562 (0.208)
Indicator for individual being self-employed	0.031 (0.002)	0.033 (0.004)	0.030 (0.002)	0.022 (0.002)	0.025 (0.004)	0.022 (0.002)
Observations	1,560,890	1,560,890	1,560,890	1,286,318	1,286,318	1,286,318
1 SD change x beta (1)	0.029		0.029	0.028		0.027
1 SD change x beta (2)		0.055	0.050		0.050	0.046

Notes: Estimations describe the OLS relationship between log yearly income of individuals and entrepreneurial activity of their ethnic group. Sample is taken from 2000 Census IPUMS. Sample includes native males and immigrant males who migrate after 1968 (effective date of the Immigration Reform Act of 1965), are aged 30-65 in 2000, and have lived in the United States for at least 10 years. Sample excludes workers whose self-employment status is unknown or not applicable, industries without self-employment, and workers living outside of metropolitan areas. Baseline estimation includes fixed effects for the following person-level traits (category counts in parentheses): PUMA geographical location (625), industry (200), native/immigrant (2), age (7), age at immigration for migrants (7), education (3), and English language fluency (2). Regressions cluster standard errors by ethnic group and use IPUMS sample weights. The bottom of the table provides the standard deviation x beta coefficient for the group-level variables (0.0255 for (1), 0.0810 for (2)). Columns 4-6 exclude workers in professional occupations and holders of doctorate degrees.

Online Appendix: Empirical

Table A1a provides our largest overage ratios ordered by $OVER1_l$. We find evidence of strong entrepreneurial clustering. For example, Gujarati Indians have an average overage ratio of 33 across the industries of their self-employment work, and an average overage ratio of 59 in their three largest industries. Their max overage is in the hotel and motel industry, which we further explore in Table A1b. The last three columns of Table A1a provide broader statistics about each ethnic group, such as its total employment (entrepreneurial and wage workers), self-employment share, and in-marriage rates.

Table A1b displays the maximum overages observed at the industry level for ethnic groups, ordered by max self-employment overage. The table displays for the ethnic groups their industry of max self-employment overage, the industry of max overage when using all workers, and the industry where the most workers for the ethnic group are occupied in terms of absolute counts. In 17 of 25 cases shown, the industry where the ethnic group displays the highest concentration for self-employment is the same as the industry where the ethnic group shows the highest concentration for total employment. In 8 of 25 cases, the industry of maximum concentration is also the industry where the ethnic group employs the most workers in an absolute sense. The industry size variable ranks industries from largest (1) to smallest (200) in terms of their overall size in the economy. Most of the maximum-concentration industries in the first two industry lists are of moderate size; industries in the third set for highest absolute count of ethnic employees tend to be larger industries.

It is noteworthy that some important factors aid group concentration are conceptually similar to but are not captured by our theoretical and empirical work. For example, we treat the taxi industry as a single industry for our empirical work, but taxi markets are often segmented by cities. Frequent travelers note the degree to which different ethnic groups appear to dominate the taxi industry on a city-by-city basis, with the most important group for each city being different. In fact, more broadly, many industries of maximum concentration (e.g., grocery stores, gas stations) are cases where geography can play an important role. This suggests we are likely under-estimating true concentration in this regard.[1]

On a related note, social interaction effects should in principle be relevant to any setting where the complementarity between social interaction and skill acquisition is strong. However, occupations and industries that require specific education and skills that are typically acquired early in life are not amenable to the forces that we model in which immigrants arrive in the United States as adults. Thus, adult immigrants

[1] Unfortunately, the data counts become very thin for segmenting by geography using IPUMS. Future work using universal linked employer-employee data can analyze these features.

find it harder to enter the medical profession, despite its significant interplay between social and professional interactions, given medicine's deep professional requirements and extensive training period. Many of the displayed entrepreneurial activities that are subject to ethnic concentration have much shorter training cycles and fewer degree or occupational licensing requirements.

Tables A2a and A2b report pairwise correlations and pairwise rank correlations for eight variants in overage ratios. All correlations exceed 0.4 and are statistically significant at a 5% level.

Appendix Table A1a: Ethnic groups displaying the greatest self-employment industrial concentration

Ethnic group, designated by country of origin or sub-groups available in IPUMS	Weighted average overage ratio over all industries	Weighted average overage ratio for three largest self-employment industries for ethnicity	Self-employment industry with max overage ratio	Total employment in sample	Share of employment classified as self-employed	In-marriage rate
Yemen	50.0	64.2	Grocery stores	2,322	26%	86%
Eritrea	35.4	45.5	Taxicab service	3,338	17%	100%
Gujarati	32.8	59.4	Hotels and motels	26,373	25%	93%
Ethiopia	27.2	43.9	Taxicab service	8,760	14%	64%
Bangladesh	20.5	27.6	Taxicab service	11,770	16%	86%
Chaldean	16.1	35.0	Grocery stores	5,429	33%	88%
Haiti	16.1	29.8	Taxicab service	58,971	8%	75%
Ghana	15.9	20.6	Taxicab service	10,975	11%	68%
Afghanistan	15.3	20.9	Taxicab service	6,432	24%	76%
Nigeria	13.6	29.5	Taxicab service	27,232	18%	64%
Tonga	12.0	14.5	Landscape and horticultural services	2,685	27%	77%
Morocco	11.3	11.2	Construction	5,346	23%	32%
Punjabi	10.5	21.8	Gasoline service stations	16,453	27%	96%
Jordan	10.0	17.6	Grocery stores	7,674	35%	68%
Laos	9.9	3.6	Agricultural production, crops	19,635	9%	77%
Pakistan	9.9	18.5	Taxicab service	35,722	22%	83%
Dominican Republic	8.7	16.6	Taxicab service	70,576	13%	62%
Cambodia	8.5	7.8	Eating and drinking places	16,245	15%	82%
Iraq	8.5	3.4	Offices and clinics of physicians	4,598	32%	60%
Turkey	8.1	3.4	Eating and drinking places	10,438	27%	60%
Korea	8.0	15.0	Laundry, cleaning, and garment services	91,928	45%	70%
Australia	7.9	2.1	Construction	4,910	23%	32%
Hungary	7.6	3.1	Construction	6,697	26%	32%
Syria	7.5	11.0	Offices and clinics of physicians	7,623	41%	57%
Sri Lanka (Ceylon)	7.3	9.1	Offices and clinics of physicians	4,010	26%	50%

Notes: Descriptive statistics from 2000 Census IPUMS. Sample includes males immigrating after 1968 (effective date of the Immigration Reform Act of 1965), aged 30-65 in 2000, and living in the United States for at least 10 years. Sample excludes workers whose self-employment status is unknown or not applicable, industries without self-employment, and workers living outside of metropolitan areas. The overage ratios and industry titles are specific to self-employment and weight industries by the number of self-employed workers for the ethnic group. Two small groups that are partially composed of residual individuals are not listed in this table but have overage values in this range (Indochina, ns 9.4; Africa, ns/nec 8.2). The employment column displays the total workforce size included in the sample for each ethnic group.

Appendix Table A1b: Maximum overage clusters and industry employment ranks by ethnic group

Ethnic group	Industry of max overage for self-employed sample	Index	Industry size	Industry of max overage for total worker sample	Industry size	Industry of max total employment	Industry size
Gujarati	Hotels and motels	108.1	31	Liquor stores	146	Hotels and motels	31
Yemen	Grocery stores	75.0	13	Grocery stores	13	Grocery stores	13
Eritrea	Taxicab service	61.0	77	Taxicab service	77	Taxicab service	77
Ethiopia	Taxicab service	52.6	77	Taxicab service	77	Taxicab service	77
Bangladesh	Taxicab service	47.1	77	Taxicab service	77	Eating and drinking places	4
Haiti	Taxicab service	42.3	77	Taxicab service	77	Construction	1
Nigeria	Taxicab service	38.1	77	Taxicab service	77	Hospitals	5
Ghana	Taxicab service	35.3	77	Taxicab service	77	Hospitals	5
Punjabi	Gasoline service stations	34.6	88	Taxicab service	77	Taxicab service	77
Korea	Laundry, cleaning, etc. services	33.5	94	Shoe repair shops	200	Laundry, cleaning, etc. services	94
Afghanistan	Taxicab service	32.5	77	Taxicab service	77	Eating and drinking places	4
Jordan	Grocery stores	28.1	13	Taxicab service	77	Grocery stores	13
Dom. Republic	Taxicab service	27.2	77	Taxicab service	77	Construction	1
Armenian	Jewelry stores	25.7	138	Jewelry stores	138	Construction	1
Pakistan	Taxicab service	25.6	77	Taxicab service	77	Taxicab service	77
Lebanon	Gasoline service stations	23.5	88	Gasoline service stations	88	Eating and drinking places	4
Chaldean	Grocery stores	20.6	13	Liquor stores	146	Grocery stores	13
Tonga	Landscape/horticultural services	18.2	25	Landscape/horticultural services	25	Construction	1
India	Hotels and motels	17.8	31	Offices and clinics of physicians	36	Computer and data processing	8
Portugal	Fishing, hunting, and trapping	16.5	170	Dyeing and finishing textiles	176	Construction	1
Ecuador	Taxicab service	15.6	77	Apparel and accessories	106	Construction	1
Iran	Apparel, fabrics, and notions	14.3	144	Apparel, fabrics, and notions	144	Eating and drinking places	4
Vietnam	Fishing, hunting, and trapping	13.4	170	Fishing, hunting, and trapping	170	Electrical machinery/equipment	14
USSR/Russia	Taxicab service	13.2	77	Taxicab service	77	Construction	1
Ukraine	Taxicab service	13.2	77	Taxicab service	77	Construction	1

Notes: See Table A1a. Table is ordered by the 25 largest self-employment overage ratios at the industry level for ethnic groups. The industry size variable ranks industries from largest (1) to smallest (200). The table also displays for each ethnic group the industry of maximum overage when considering all employed workers and the industry where the greatest number of workers are employed.

Appendix Table A2a: Pairwise correlations of various overage metrics

	Sample	Metric	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
(1)	Self-employed	Log weighted average overage ratio across all industries [OVER1]	1							
(2)		Log weighted average overage ratio in three largest industries [OVER2]	0.946	1						
(3)		Log average of three largest overage ratios for ethnic group [OVER3]	0.923	0.961	1					
(4)		Log largest overage ratio for ethnic group [OVER4]	0.859	0.927	0.966	1				
(5)	All workers	Log weighted average overage ratio across all industries [OVER1]	0.832	0.767	0.731	0.631	1			
(6)		Log weighted average overage ratio in three largest industries [OVER2]	0.835	0.796	0.785	0.685	0.948	1		
(7)		Log average of three largest overage ratios for ethnic group [OVER3]	0.555	0.627	0.640	0.630	0.541	0.632	1	
(8)		Log largest overage ratio for ethnic group [OVER4]	0.470	0.577	0.530	0.522	0.476	0.495	0.900	1

Notes: Table displays correlations between ethnic group overage measures calculated on both self-employment and industry total employment. All correlations are significant at a 5% level.

Appendix Table A2b: Pairwise rank correlations of various overage metrics

	Sample	Metric	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
(1)	Self-employed	Log weighted average overage ratio across all industries [OVER1]	1							
(2)		Log weighted average overage ratio in three largest industries [OVER2]	0.808	1						
(3)		Log average of three largest overage ratios for ethnic group [OVER3]	0.588	0.789	1					
(4)		Log largest overage ratio for ethnic group [OVER4]	0.569	0.760	0.971	1				
(5)	All workers	Log weighted average overage ratio across all industries [OVER1]	0.835	0.821	0.661	0.648	1			
(6)		Log weighted average overage ratio in three largest industries [OVER2]	0.706	0.859	0.719	0.678	0.872	1		
(7)		Log average of three largest overage ratios for ethnic group [OVER3]	0.589	0.739	0.768	0.816	0.760	0.743	1	
(8)		Log largest overage ratio for ethnic group [OVER4]	0.587	0.705	0.705	0.742	0.749	0.724	0.955	1

Notes: See Appendix Table A2a. Table displays rank correlations between ethnic group overage measures calculated on both self-employment and industry total employment. All correlations are significant at a 5% level.

Online Appendix: Theory

The theory in this paper consists of two fundamental building blocks. First, social interactions and production are complementary. Second, different social relationships are not close substitutes for one another. The former is analyzed in the main text, and this appendix begins with additional discussion. We then consider pricing equilibrium and social networks with endogenous matching. The numbering of assumptions and propositions continues from the main text.

1 Discussion of Baseline Model

1.1 Quality and Convex Productivity

In addition to the quantity of social interactions with other self-employed entrepreneurs, the quality of these interactions could also matter for productivity. Let individual productivity for self-employed entrepreneurs in industry 1 increase both in the quantity and average productivity of other entrepreneurs in the sector of the same group. Write this as

$$\theta = \phi + \delta X_l \overline{\theta}, \tag{1}$$

where $\phi > 0$ is a productivity term, $0 < \delta < 1$ is a social multiplier, X_l is the fraction of entrepreneurs in group l, and $\overline{\theta}$ is the average productivity of these entrepreneurs. Solving for equilibrium productivity by setting θ equal to $\overline{\theta}$, individual productivity in group l is a function:

$$\theta\left(X_l\right) = \frac{\phi}{1 - \delta X_l}. \tag{2}$$

Under these conditions, productivity is convex in the degree of specialization when taking both the quantity and the quality of interaction into account.[1] With this result in mind, we make the following assumption:

Assumption 1B *Productivity of self-employed entrepreneurs in industry* 1 *is convex in specialization:* $\theta'' > 0$.

Assumption 1B allows a full characterization of the efficient solution without having to resort to explicit functional form. We discuss further below. Convex productivity gives the following result:

[1] This specification highlights the differences from a standard interaction model. The standard model is generally specified so that individual productivity is a function of a group-specific term ϕ and the discounted mean of the group, $\delta\overline{\theta}$. Solving $\theta = \phi + \delta\overline{\theta}$, interaction exacerbates the difference in ϕ across groups, $\theta = \frac{\phi}{1-\delta} > \phi$, but the degree of specialization X_l has no effect on productivity.

Lemma *If productivity is convex, both groups never work in both industries.*

Proof: Assume by contradiction that an efficient distribution (X_A, X_B) exists where $0 < X_l < 1$ for $l = \{A, B\}$. Consider a marginal change ϵ in the ethnic composition of self-employed entrepreneurs in industry 1 while holding fixed the overall number of said entrepreneurs M (and therefore also the outputs of both industries). Taking the derivative of Q_1 with respect to ϵ, and evaluating it at $\epsilon = 0$:

$$\frac{\partial Q_1}{\partial \epsilon}\left(X_A + \frac{\epsilon}{N_A}, X_B - \frac{\epsilon}{N_B}\right) = \theta\left(X_A\right) + X_A\theta'\left(X_A\right) - \theta\left(X_B\right) - X_B\theta'\left(X_B\right) \qquad (3)$$

Since (X_A, X_B) is efficient, and since X_l is interior, this derivative has to be zero.[2] But with convex productivity the derivative is zero only at $X_A = X_B$, which is the global minimum. This contradicts efficiency. ■

The efficient economy aims for maximum ethnic homogeneity in self-employed entrepreneurship in industry 1. Ruling out that both groups work in both sectors implies that only the specialized distributions along the two curves depicted in Figure 1 of the main text could possibly coincide with the transformation frontier. The shape of the entire transformation frontier can therefore be deduced by tracing out the maximum of the two curves in that figure.

Proposition 2 *If productivity is convex, there is a cutoff value v^* such that for $v < v^*$, the minority group specializes as self-employed entrepreneurs in industry 1, whereas for $v > v^*$, the majority specializes.*

Proof: Direct from Proposition 1 and Lemma proofs with convexity. ■

The right panel of Figure 1 of the main text also shows how the degree of specialization varies with the size of industry 1, as governed by v, and the cutoff value v^* for majority group specialization. The greater the value of v, the greater is the demand for industry 1 and the more people work in it. As industry 1 increases in size, the interaction externality generates a characteristic discrete jump from one type of equilibrium to another. At the point v^*, where many from group B have also joined self-employed entrepreneurship in industry 1, the economy abruptly moves from minority specialization to majority specialization.

[2] If the derivative is nonzero, then the output of industry 1 could increase while keeping the output of industry 0 constant. By subsequently increasing the number of workers in industry 0 marginally, a Pareto improvement is feasible, thus contradicting efficiency.

1.2 The Case of Non-Convex Productivity

To see that convexity is needed for the Lemma on ethnic homogeneity to hold, consider a non-convex production function where a threshold fraction must work as self-employed entrepreneurs in industry 1 for interaction to have value: $\theta > 0$ if $X_l \geq b$ and zero otherwise. This specification violates the assumption that productivity is strictly increasing in the degree of specialization. Then, if the demand for industry 1 output is so great that a single group cannot satisfy it entirely, $v > V(0,1)$, and if in addition $V(b,b) < v < V(b,1)$, efficiency requires that both ethnic groups work in both industries, contradicting the Lemma.

To see why, consider what would happen if one of the groups specialized completely. In this case the non-specialized group's degree of specialization would be positive but below b, causing the self-employed industry 1 entrepreneurs in that group to have zero productivity. If, however, the industrial distribution was unspecialized instead, with $X_A = X_B$, then self-employed industry 1 entrepreneurs in both groups would be as productive as those in the most productive group were under the alternative. Clearly this would be Pareto superior, contradicting the Lemma. This special case shows how the Lemma fails for non-convex productivity, and how in this case the qualitative features of specialization will depend on specific functional form assumptions. Recall however that the results for both $v \leq V(1,0)$ and $v = V(0,1)$ are more general and apply both for convex and non-convex productivity. This condition is less important for the remaining model discussion.

2 The Price Equilibrium

The model in the main text characterizes the efficient outcome. The focus now turns to the competitive outcome. An equilibrium analysis will yield two insights into how social interaction affects distribution over industries. First, it shows how stratifying forces act to make groups more and more different, and second, how group earnings are positively related to the degree of specialization.

To see how social interaction works as a stratifying force, begin by introducing time into the analysis, with $t = 0, 1, ..., \infty$. Dynamics are built into the model by making the interaction effect work with a lag. Denote by X_l^t the degree of specialization in period t for group l, and let self-employed individual entrepreneurial productivity in industry 1 in period t be a function $\theta\left(X_l^{t-1}\right)$. This one-period lag specification for the interaction effect could easily be generalized to a distributed lag. Interaction now effectively works as a form of social capital, with the group's self-employment activities in the previous period benefiting individual productivity today. Let p_1^t and p_0^t be the prices of industry 1 output and industry 0 output respectively. Entrepreneurial earnings in industry 1 are $y_{1,l}^t = p_1^t \theta\left(X_l^{t-1}\right)$ and worker earnings in industry 0 are $y_{0,l}^t = p_0^t$. Competitive industrial choice is straightforward to derive in this setting; defining the relative price of industry 0 output to industry 1 output as $p^t = \frac{p_0^t}{p_1^t}$, an individual in group l joins industry 1 as a self-employed entrepreneur if

$$\theta\left(X_l^{t-1}\right) \geq p^t \tag{4}$$

and favors being a worker in industry 0 if $\theta\left(X_l^{t-1}\right) \leq p^t$. Since individuals have identical skills, aggregate labor supply for group l is discontinuous, with:

$$X_l^t = \begin{cases} 1 & \text{if } \theta\left(X_l^{t-1}\right) > p^t \\ [0,1] & \text{if } \theta\left(X_l^{t-1}\right) = p^t \\ 0 & \text{if } \theta\left(X_l^{t-1}\right) < p^t. \end{cases} \tag{5}$$

Avoid for now the knife-edge *unspecialized* case where $X_A^{t-1} = X_B^{t-1}$. Since there is a single price of labor, p^t, at least one of the two groups A and B must then be in a corner:

$$\left(X_A^t, X_B^t\right) = \begin{cases} \left(X_A^t = 1, 0 < X_B^t\right) \text{ or } \left(X_A^t \leq 1, X_B^t = 0\right) & \text{if } X_A^{t-1} > X_B^{t-1} \\ \left(0 < X_A^t, X_B^t = 1\right) \text{ or } \left(X_A^t = 0, X_B^t \leq 1\right) & \text{if } X_A^{t-1} < X_B^{t-1} \end{cases} \tag{6}$$

In equilibrium, supply must satisfy (6) and production must meet demand so that markets clear. Because of perfect complementarity, meeting demand reduces to satisfying $v = V\left(X_A^t, X_B^t\right)$. The resulting equilibrium distribution is unique. To see why, take the case when group l is more specialized than group l' in the previous period, with $X_l^{t-1} > X_{l'}^{t-1}$. Given that at least one of the two groups must be in a corner according to (6), the equilibrium distribution must either be of the type $\left(X_l^t, 0\right)$ or of the type $\left(1, X_{l'}^t\right)$. Since the function V is strictly increasing in both arguments, it follows that $V\left(1, X_{l'}^t\right) > V\left(X_l^t, 0\right)$. Only one distribution can consequently make V equal to v.

The equilibrium distribution is therefore uniquely determined by the distribution in the previous period. Continuing to avoid the knife-edge unspecialized case, define a function ϕ that maps every previous distribution into a new distribution:

$$\left(X_A^t, X_B^t\right) = \phi\left(X_A^{t-1}, X_B^{t-1}\right) \tag{7}$$

Next, proceed to characterize stationary equilibrium distributions. Like other equilibrium distributions, stationary distributions must satisfy (6) and must meet demand. Following the same argument as above, based on V being strictly increasing in both arguments, it follows that there is a stationary equilibrium where each of the two groups specializes. Denote the stationary distribution as $\left(X_A^A, X_B^A\right)$ when the minority specializes, and the stationary distribution as $\left(X_A^B, X_B^B\right)$ when the majority specializes.

Finally, returning for a moment to the unspecialized knife-edge case where $X_A^{t-1} = X_B^{t-1}$, this type of initial condition is of measure zero and therefore not elaborated on. Note only that since V is strictly increasing in both arguments, there can only be one such stationary unspecialized equilibrium distribution. Denote that equilibrium distribution as $\left(X_A^U, X_B^U\right)$. In the unspecialized case, although there is only one stationary equilibrium, the uniqueness of equilibria no longer applies. To summarize, there are consequently three stationary equilibrium distributions: two specialized, $\left(X_A^A, X_B^A\right)$ and $\left(X_A^B, X_B^B\right)$, and one unspecialized, $\left(X_A^U, X_B^U\right)$. Figure A1 shows the two specialized equilibria, as well as the knife-edge equilibrium, when v is less than $V(1,0)$.

2.1 Industrial Stratification

Our next analysis shows that the dynamic system in (7) converges to a stationary specialized equilibrium, so long as the interaction externality is not too strong. This analysis only examines unspecialized initial conditions, which establishes convergence on measure one. Consider what happens to the aggregate production of industry 1 when one (infinitesimal) person in group l becomes a self-employed entrepreneur in that industry. First, aggregate production increases by an amount equal to the individual productivity of that person, $\theta(X_l)$. In addition, all other self-employed entrepreneurs in industry 1 from group l benefit from the interaction externality when socializing with this new entrepreneur. Individual productivity therefore increases by $\frac{1}{N_l}\theta'(X_l)$ for all $X_l N_l$ self-employed industry 1 entrepreneurs in group l. Consequently, the internalized effect on aggregate production of one person joining the self-employed entrepreneurial sector of industry 1 is $\theta(X_l)$, and the external effect is $X_l\theta'(X_l)$. Assume that the external effect is smaller than the internal effect.[3]

Assumption 2 *The internal effect dominates:* $\theta'(X_l)X_l < \theta(X_l)$.

This condition is satisfied if productivity is concave in X_l, but it also holds for some convexity as long as $\theta(0) > 0$. To see why the assumption is needed for the system to be stable, consider the extreme case when group A has no mass at all, with $N_A = 0$. Since the derivative of V with respect to X_A^t is zero in this case, group A can be ignored altogether in the general equilibrium analysis. There is then a single stationary level of specialization for group B; denote this value as X_B^*.

Consider a perturbation in period t so that the majority starts out with too many entrepreneurs in industry 1, $X_B^t > X_B^*$, shown in Figure A2. Such a deviation boosts the interaction effect in period $t+1$ relative to the stationary equilibrium, $\theta(X_B^t) > \theta(X_B^*)$. With perfect complementarity, the outputs of both industry 0 and industry 1 must therefore increase relative to their stationary equivalents. Increasing the output of industry 0 requires an increase in the number of workers in that industry, and consequently, a decrease in the number of self-employed entrepreneurs in industry 1 to

[3] We thank Rachel Soloveichik for this interpretation of Assumption 2.

below the stationary value X_B^*. With fewer of these entrepreneurs in period $t+1$ than the stationary number, the tables turn in period $t+2$, so that the interaction effect now is reduced to below that in the stationary equilibrium. Reducing the production of industry 0 and industry 1 in period $t+2$ in response, the number of industry 0 workers in period $t+2$ has to decrease and the number of self-employed industry 1 entrepreneurs has to increase relative to the stationary equilibrium. These reversals repeat every period in cobweb-style dynamics.[4]

The question of whether the system is stable reduces to whether the number of self-employed entrepreneurs in industry 1 in period $t+2$ is less than the number of such entrepreneurs in period t, so that the degree of specialization in group B gets closer and closer to the stationary value X_B^* over time. Using the derived direction of the change in industry 1 production, $Q_1^{t+1} > Q_1^{t+2}$, this latter inequality can be equivalently expressed, after multiplying and dividing the left-hand side by X_B^t and dividing both sides by $X_B^{t+1} N_B$, as:

$$X_B^t \frac{\theta\left(X_B^t\right)}{X_B^t} > X_B^{t+2} \frac{\theta\left(X_B^{t+1}\right)}{X_B^{t+1}} \tag{8}$$

Given that productivity is not too convex, as stipulated by Assumption 2, it follows that $\frac{\theta(X_l)}{X_l}$ is strictly decreasing in X_l. Since $X_B^t > X_B^{t+1}$, equation (8) then establishes that $X_B^t > X_B^{t+2}$. This proves convergence and the stability of group B's degree of specialization around X_B^*.

Having established stability in the case of $N_A = 0$, the same example also serves to show how the stratifying force comes into play. Let group B be in its stable state, with $X_B^t = X_B^*$, and perturb the minority's industry distribution so that $X_A^t > X_B^*$. Since group B is so much greater in size than group A, the former is unaffected by the perturbation and the price continues to be locked in at $p^{t+1} = \theta\left(X_B^*\right)$. The interaction effect in period $t+1$, generated by the perturbation in period t, then results in everyone in group A becoming more productive as self-employed entrepreneurs in industry 1 than as workers in industry 0, with $\theta\left(X_A^t\right) > p^{t+1}$. Group A's degree of specialization consequently jumps from X_A^t to $X_A^{t+1} = 1$, and the distribution stays in this stratified state forever. This stratification result is extended later for the general case of any population size of the two groups, and it follows that for $l \in \{A, B\}$ and $l' \in \{A, B\}$:

Proposition 3 *Initial differences result in long-run specialization: If group l is more specialized than group l' initially, $X_l^0 > X_{l'}^0$, then group l specializes in the long run and the limiting distribution is $\left(X_A^l, X_B^l\right)$.*

[4] The flip-flopping character of the equilibrium distribution is a result of the one-period lag specification for the interaction effect. The distribution would change more gradually with a more general specification allowing for distributed lags.

Proof: Consider the equilibrium sequence of industry distributions:

$$\left(\left(X_A^1, X_B^1\right), \left(X_A^2, X_B^2\right), ...\right) \tag{9}$$

If one group l is more specialized than the other group l' initially, $X_l^0 > X_{l'}^0$, supply in (5) requires that the equilibrium sequence begins in one of the following three ways:

$$\left(\left(X_l^1, X_{l'}^1\right), \left(X_l^2, X_{l'}^2\right), ...\right) = \begin{cases} \left(\left(< 1, 0\right), ...\right) \\ \left(\left(1, \geq 0\right), \left(1, \geq 0\right), ...\right) \\ \left(\left(1, \geq 0\right), \left(< 1, 0\right), ...\right). \end{cases} \tag{10}$$

The proof proceeds by establishing that the sequence converges to $\left(X_A^l, X_B^l\right)$ in each of these three cases. Define the variable $\lambda\left(X_l\right) \equiv \frac{\theta(X_l)}{X_l}$ for $X_l > 0$. From Assumption 2 it follows that $\lambda'\left(X_l\right) < 0$. Proceed to establish convergence:

Case 1 $X_l^1 < 1$ *and* $X_{l'}^1 = 0$.

Show first that group l' stays out of entrepreneurship in industry 1 for good. By contradiction: if not, then there exists a time t where $X_{l'}^{t+1} = 0$ and $X_{l'}^{t+2} > 0$. Since supply must satisfy (6) it then follows that $X_l^{t+1} > 0$ and $X_l^{t+2} = 1$. The change in the output of industry 1 can then be written as:

$$Q_1^{t+2} - Q_1^{t+1} = N_l\left(\theta\left(X_l^{t+1}\right) - X_l^{t+1}\theta\left(X_l^t\right)\right) + X_{l'}^{t+2}N_{l'}\theta\left(X_{l'}^{t+1}\right). \tag{11}$$

This difference is strictly positive if the first term is positive. Clearly this is the case if $X_l^{t+1} \geq X_l^t$. If, instead, $X_l^{t+1} < X_l^t$, then again focusing on the first term:

$$\begin{aligned} \theta\left(X_l^{t+1}\right) - X_l^{t+1}\theta\left(X_l^t\right) &= \lambda\left(X_l^{t+1}\right)X_l^{t+1} - X_l^{t+1}\lambda\left(X_l^t\right)X_l^t \\ &= X_l^{t+1}\left(\lambda\left(X_l^{t+1}\right) - \lambda\left(X_l^t\right)X_l^t\right) > 0. \end{aligned} \tag{12}$$

This establishes that $Q_1^{t+2} > Q_1^{t+1}$. Since the output production of both industries must move in the same direction to clear the market, because of perfect complementarity, it follows that the output of industry 0 also increases from $t+1$ to $t+2$. This in turn requires that the number of workers in industry 0 increases, or equivalently, that the number of self-employed entrepreneurs in industry 1 decreases:

$$X_l^{t+2}N_l + X_{l'}^{t+2}N_{l'} < X_l^{t+1}N_l + X_{l'}^{t+1}N_{l'}. \tag{13}$$

Since $X_l^{t+2} = 1$ and $X_{l'}^{t+1} = 0$, this inequality can be simplified as $N_l + X_{l'}^{t+2}N_{l'} < X_l^{t+1}N_l$. This inequality is a contradiction and establishes that group l' stays out of self-employed entrepreneurship in industry 1 for good. The stationary equilibrium must consequently be of the form $\left(X_l^l, 0\right)$.

Assume first that $X_l^t > X^*$, in which case it is easy to show that $Q_1^{t+1} > Q_1^l > Q_1^{t+2}$ as well as $X_l^{t+1} < X_l^l < X_l^{t+2}$. Since $Q_1^{t+1} > Q_1^{t+2}$ it follows that:

$$\begin{aligned} X_l^{t+1} N_A \theta \left(X_l^t\right) &> X_l^{t+2} N_A \theta \left(X_l^{t+1}\right) \\ X_l^{t+1} \lambda \left(X_l^t\right) X_l^t &> X_l^{t+2} \lambda \left(X_l^{t+1}\right) X_l^{t+1} \\ X_l^t \lambda \left(X_l^t\right) &> X_l^{t+2} \lambda \left(X_l^{t+1}\right). \end{aligned} \tag{14}$$

The last line implies that $X_l^t > X_l^{t+2}$. The exact same argument, but with reverse inequalities, can be made for $X_l^t < X_l^l$. Therefore, having established that $X_l^t > X_l^{t+2} > X_l^l$ when $X_l^t > X_l^l$, and vice versa when $X_l^t < X_l^l$, it has been shown that X_l^t approaches the stationary equilibrium value X_l^l over time. This establishes convergence in Case 1.

Case 2 *$X_l^1 = 1$, $X_{l'}^1 \geq 0$, $X_l^2 = 1$ and $X_{l'}^2 \geq 0$.*

Show first that in this case, group l stays specialized for good. By contradiction: if not, then there exists a time t when $X_l^t = 1$, $X_l^{t+1} = 1$ and $X_l^{t+2} < 1$. Since supply must satisfy (6), it follows that $X_{l'}^{t+2} = 0$. The change in the output of industry 1 can be written as

$$Q_1^{t+2} - Q_1^{t+1} = N_l \left(X_l^{t+2} \theta (1) - \theta (1)\right) - X_{l'}^{t+1} N_{l'} \theta \left(X_l^t\right) < 0. \tag{15}$$

Since the supply of output of both industries must move in the same direction to clear the market, it follows that the output of industry 0 also decreases, which requires that the number of self-employed entrepreneurs in industry 1 increases:

$$X_l^{t+2} N_l + X_{l'}^{t+2} N_{l'} > X_l^{t+1} N_l + X_{l'}^{t+1} N_{l'}. \tag{16}$$

Since $X_{l'}^{t+2} = 0$ and $X_l^{t+1} = 1$, this inequality can be rewritten as $X_l^{t+2} N_l > N_l + X_{l'}^{t+1} N_{l'}$, which is a contradiction. This establishes that group l stays specialized in industry 1 for good. The stationary equilibrium must consequently be of the form $\left(1, X_{l'}^l\right)$. By the same argument as in Case 1, the sequence can be shown to approach the stationary equilibrium value $X_{l'}^l$ over time, both if $X_{l'}^t > X_{l'}^l$ and if $X_{l'}^t < X_{l'}^l$. This establishes convergence in Case 2.

Case 3 *$X_l^1 = 1$ and $X_{l'}^1 \geq 0$ and $X_l^2 < 1$ and $X_{l'}^2 = 0$.*

By the same argument in Case 1, it follows that group l' stays out of entrepreneurship in industry 1 permanently. Repeating the arguments in Case 1, convergence can then be established also in Case 3.

Consequently, in all three cases there is convergence. ■

This also implies that the stationary unspecialized equilibrium $\left(X_A^U, X_B^U\right)$ is unstable. If the minority group is slightly more specialized initially, then the economy converges to minority specialization $\left(X_A^A, X_B^A\right)$, and if the opposite is true, then the economy converges to majority specialization $\left(X_A^B, X_B^B\right)$. Over time, social segregation amplifies initial group differences.

2.2 Initial Conditions and Multiple Groups

Depending on the initial conditions, as is clear from Proposition 3, either of the two groups A and B can specialize as self-employed entrepreneurs in industry 1. Social interaction amplifies initial differences, but it does not explain why they are there to begin with. The difference in group size has some implications for what initial conditions to expect, however.

Consider an economy with more than two groups. As before, the group with more self-employed entrepreneurs in industry 1 initially will specialize in the long run. If the initial industrial distribution is subject to randomness, one of the smaller groups is likely to be the most specialized initially. To see why, let the initial distribution be generated by random draws, where each person becomes a self-employed entrepreneur in industry 1 with probability ρ.[5] This probability structure results in the same expected initial degree of specialization for all groups, but since the population size varies across groups, the variance in the degree of specialization also varies. The smallest groups have the largest variance, and therefore, the smallest groups are most likely to exhibit the lowest and also the greatest initial degrees of specialization. Consequently, with the smallest groups the most likely to specialize initially, as interaction amplifies initial differences over time, the smallest groups are also the most likely to specialize in the long run.

2.3 Assimilation

Our model does not feature assimilation of immigrants and their offspring and thus yields permanent social and industrial segregation. In our framework, assimilation would reduce the social isolation of an ethnic group (or some members of it) to the majority group. Our framework then predicts the industry choices of the assimilated individuals to look like those of the majority, especially if another ethnic group shows strong social isolation.

[5] These draws can be partially correlated within groups with the assumption that the correlation is the same for every group.

2.4 Heterogeneity and Earnings

Social complementarities also have implications for earnings. To examine how interaction effects would show up in earnings data, it is necessary to move away from the framework of identical skills. Returning to a static environment, endow each person i with entrepreneurial skills relevant to self-employment in industry 1, $s_1(i)$, and with another set of skills necessary for industry 0, $s_0(i)$. Self-employed entrepreneurial earnings in industry 1 are now a function of both interactions and skills. Denote the earnings of individual i in group l when she is a self- employed entrepreneur in industry 1 as $y_1(X_l, i) = p_1 \theta(X_l) s_1(i)$, and when she is a member of industry 0 as $y_0(i) = p_0 s_0(i)$. Defining the ratios $s \equiv \frac{s_1}{s_0}$, $p \equiv \frac{p_0}{p_1}$, and $q \equiv p \frac{y_1}{y_0}$, the earnings-maximizing industry choice of individual i is to consider becoming a self-employed entrepreneur in industry 1 if:

$$q(X_l, i) \geq p \tag{17}$$

and to consider working in industry 0 if $q(X_l, i) \leq p$. Here the term $q(X_l, i) = \theta(X_l) s(i)$ summarizes the individual's comparative advantage in self-employed entrepreneurship in industry 1, at parity prices, as a function of social interaction and skills.

When individuals have different skills, the character of the price equilibrium depends crucially on the marginal self-employed entrepreneur and how her comparative advantage changes as more and more untalented people also become entrepreneurs in industry 1. If the benefits of interaction are weak and the marginal entrepreneur "deteriorates" as more intrinsically untalented people enter the industry, then the economy reduces to a standard Roy model, or sorting model, with a unique unspecialized equilibrium. Only if the interaction effect is strong enough to overcome skill heterogeneity can interaction change the character of the equilibrium.

Without loss of generality, order individuals from the greatest to the smallest comparative advantage in industry 1-style entrepreneurship, so that the skill ratio is decreasing in i, $s'(i) \leq 0$. The marginal entrepreneur is then the individual indexed by $i = X_l$, and her comparative advantage is $q(X_l, X_l)$. To prevent the economy from reducing to a sorting model, assume that the interaction effect trumps heterogeneity:

Assumption 3 *Interaction dominates at the margin:* $\frac{d}{dX_l} q(X_l, X_l) > 0$.

This assumption implies that the solid line in Figure A3 is upward sloping. The equilibrium distribution (X_A, X_B) must be competitively supplied and enough output must be produced by both industries to meet demand. Using a similar line of reasoning as in the previous section, based on V being strictly increasing in both arguments, it follows from Assumption 3 that there are three equilibria: one unstratified, denoted $\left(X_A^U, X_B^U\right)$; one where the minority group A specializes, denoted $\left(X_A^A, X_B^A\right)$; and one where the majority group B specializes, denoted $\left(X_A^B, X_B^B\right)$.[6]

In the equilibrium where minority A specializes as self-employed entrepreneurs in industry 1, the mean earnings of members of group A are higher than the mean earnings of members of group B, and vice versa in the equilibrium where group B specializes. To see why, let $y = \max\left(y_0, y_1\right)$ be actual individual earnings, and denote mean group earnings as $\mu = \int_0^1 y di$.

Proposition 4 *Earnings covary with self-employed entrepreneurship in industry 1:* $\mu\left(X_l\right) > \mu\left(X_{l'}\right)$ *if* $X_l > X_{l'}$.

Proof: Since people sort into industries, mean earnings can be rewritten as

$$\mu\left(X_l\right) = \int_0^1 y_0\left(i\right) di + \int_0^{X_l} \left(y_1\left(X_l, i\right) - y_0\left(i\right)\right) di \tag{18}$$

Rearranging, the difference in mean earnings between the two groups is:

$$\mu\left(X_l\right) - \mu\left(X_{l'}\right) = \int_0^{X_{l'}} \left(y_1\left(X_l, i\right) - y_1\left(X_{l'}, i\right)\right) di + \int_{X_{l'}}^{X_l} \left(y_1\left(X_l, i\right) - y_0\left(i\right)\right) di \tag{19}$$

where both parts of the expression are positive. The first part is strictly positive due to the interaction effect, $\frac{\partial y_1\left(X_l, i\right)}{\partial X_l} > 0$, and the second part is positive because of sorting, $y_1\left(X_l, i\right) \geq y_0\left(i\right)$ for all $i \leq X_l$. ■

This unequivocal effect on mean earnings at the group level does not carry through to the industry level. Depending on the joint distribution of skills, mean earnings in either industry can increase or decrease as interaction increases self-employed entrepreneurial productivity in industry 1 and shifts people of different ability between industries. The effect of interaction on industry earnings is similar to the effect of changing skill prices, which cannot be signed for a general skill distribution (Heckman and Honore, 1990).

The difference in mean earnings, normalized in units of industry 0 output, is shown in Figure A4 for the equilibrium with minority specialization. The exact derivation is included below. The relative price of industry 0 to industry 1 outputs is always such that the marginal entrepreneur is indifferent between industries. Keeping track of whether the marginal entrepreneur is in group A or in group B depending on the industrial distribution, the equilibrium price can be expressed as:

$$p = \begin{cases} q\left(X_l, X_l\right) & \text{if } X_l > X_{l'} \text{ and } X_{l'} = 0, \text{ or } X_l < X_{l'} \text{ and } X_l > 0 \\ q\left(X_{l'}, X_{l'}\right) & \text{if } X_l > X_{l'} \text{ and } X_{l'} > 0, \text{ or } X_l < X_{l'} \text{ and } X_l = 0 \end{cases} \tag{20}$$

[6] Note that Assumptions 2 and 3, when combined, put both an upper and a lower bound on the interaction effect: $-\frac{d \ln s}{dX_l} < \frac{d \ln \theta}{dX_l} < \frac{1}{X_l}$.

When increasing the number of self-employed entrepreneurs in industry 1 in equilibrium with minority specialization, the relative price of industry 0 output to industry 1 output increases continuously as the marginal entrepreneur in group A becomes more and more productive. This increase in price continues until all As are self-employed entrepreneurs in industry 1. To expand industry 1's self-employed entrepreneurial sector further from the point where everyone in group A are entrepreneurs, the price has to drop discretely from $p = q(1, 1)$ to $q(0, 0)$, to lure the unproductive Bs into the sector as well. The earnings differential between groups A and B moves accordingly, as shown in Figure A4, increasing continuously until all As are self-employed entrepreneurs in industry 1, at which point earnings jump in response to the discontinuous drop in the relative price.

Derivation of Earnings Differential in Figure A4: Mean earnings denominated in terms of industry 0 outputs are:

$$\frac{\mu(X_l)}{p_0} = \int_0^{X_l} p^{-1}\theta(X_l)\, s_1(i)\, di + \int_{X_l}^{1} s_0(i)\, di. \tag{21}$$

Replace the relative price of industry 0 output to industry 1 output, $p = \frac{p_0}{p_1}$, with the comparative advantage of the marginal entrepreneur, q, since these two are equal in equilibrium. Denote the earnings differential as $\Delta(X_l, X_{l'}) \equiv \frac{\mu(X_l) - \mu(X_{l'})}{p_0}$. It can be expressed as:

$$\Delta(X_l, X_{l'}) = \int_0^{X_{l'}} q^{-1}\left(\theta(X_l) - \theta(X_{l'})\right) s_1(i)\, di + \int_{X_{l'}}^{X_l} \left[q^{-1}\theta(X_l)\, s_1(i) - s_0(i)\right] di. \tag{22}$$

For $X_l < 1$ and $X_{l'} = 0$, where $q = q(X_l, X_l)$, and $q(X_l, X_l) = \theta(X_l)\, s(X_l)$, differentiating with respect to X_l gives

$$\frac{\partial \Delta(X_l, 0)}{\partial X_l} = -s'(X_l)\, s(X_l)^{-2} \int_0^{X_l} s_1(i)\, di > 0. \tag{23}$$

For $X_l = 1$ and $X_{l'} = 0$, the drop in price from $q(1, 1)$ to $q(0, 0)$ results in a jump in the mean earnings differential equal to

$$\Delta(1, 0)\big|_{p=q(0,0)} - \Delta(1, 0)\big|_{p=q(1,1)} = \left(q(0, 0)^{-1} - q(1, 1)^{-1}\right) \theta(1) \int_0^1 s_1(i)\, di > 0. \tag{24}$$

For $x = 1$ and $X_{l'} > 0$, where $q = q(X_{l'}, X_{l'})$, differentiating with respect to $X_{l'}$ gives

$$\frac{\partial \Delta(1, X_{l'})}{\partial X_{l'}} = -\frac{dq}{dX_{l'}} q^{-2}\theta(1) \int_0^1 s_1(i)\, di + s'(X_{l'})\, s(X_{l'})^{-2} \int_0^{X_{l'}} s_1(i)\, di - 2s_0(X_{l'}) < 0. \tag{25}$$

■

3 Relationships in a Social Network

Since interactions have been restricted to be random, the analysis has so far abstracted from changes in the social structure that could arise in response to the productive value of interaction. The most interesting question is whether the majority will split up into smaller social groups, formed around choice of industry, to capitalize on interaction. If such splinter groups could form *costlessly*, then social interaction would no longer be able to generate industrial stratification along ethnic lines.

By developing a utility-based theory of interaction, explicitly stating social preferences and characterizing the optimal social structure, this section shows that splinter groups will not arise so long as preferences are sufficiently diverse, and so long as different social relationships are not close substitutes for one another. Under these two premises it is costly to confine social interactions to within a small group since the quality of social matches deteriorates with decreasing group size.

The theory developed in this section is constructed around a standard marriage market as in Becker (1973). In addition to spousal matching, people are also related by birth, which yields a larger social structure where individuals are interrelated not just pairwise but in a social network. Since the social network is derived as the outcome of matching, the problem analyzed here is different in nature from the problems most commonly analyzed in the social network literature, for example in Jackson and Wolinsky (1996), which focuses on strategic interaction between identical agents.

3.1 The Marriage Market

Take a very large finite population $i = 1, ..., N$, which is divided into mutually exclusive and exhaustive *families* by birth, with each family consisting of $d > 3$ individuals. Every person i independently draws a trait t_i, which could be for example beauty or intelligence, uniformly distributed between zero and one:

Assumption 4 *Individual traits t_i are independent draws.*

The independence of the draw signifies what can be thought of as maximal diversity: even within families people have different traits.

Based on realized traits, each person is assigned a spouse. To simplify, there are no gender restrictions and spouses can belong to the same family.[7] Traits are assumed to be complementary inputs in marriage. A marriage between i and j yields utility $u(t_i, t_j)$, where the function u is symmetric and strictly increasing with a positive cross-derivative:

[7] Removing gender restrictions maps this problem into a one-sided assortative matching problem. One-sided assortative matching is used in a different context in Kremer (1993).

Assumption 5 *Inputs are complementary:* $u(t_i, t_j) = u(t_j, t_i)$, $u_1 > 0$, $u_2 > 0$ *and* $u_{1,2} > 0$.

Since different relationships produce different utility, social relationships are not perfect substitutes and there is an optimal matching of spouses. Assume that utility is transferable, in which case the efficient spousal matching has to maximize aggregate utility. Labelling individuals according to rank, so that $t_1 < t_2 < \ldots$,[8] it follows that the efficient matching is positively assortative: person one marries person two, person three marries person four, ..., and person $N-1$ marries person N. To see this, let the matching function v be symmetric and the cross-derivative positive. For traits $t_1 < t_2 < t_3 < t_4$, we show that the only efficient matching is (t_1, t_2) and (t_3, t_4). As in Becker (1973), we use a property of v when the cross-derivative is positive,

$$v(a, d) + v(c, b) < v(a, b) + v(c, d) \tag{26}$$

for $a < c$ and $b < d$. Take an arbitrary efficient matching (x_1, x_2) and (x_3, x_4), which is a permutation of the traits t_1, t_2, t_3 and t_4. Without loss of generality, relabel these traits pairwise so that $x_1 < x_2$ and $x_3 < x_4$. Also without loss of generality, relabel the pairs so that $x_1 < x_3$. This implies that $x_1 < x_3 < x_4$. Using the symmetry of v, the aggregate utility from the arbitrary efficient matching can be written as $v(x_1, x_2) + v(x_4, x_3)$. Since $x_1 < x_4$ it follows from (26) that $x_2 < x_3$, otherwise aggregate utility could be increased by interchanging x_2 and x_3, just as b and d were interchanged in (26). Consequently, with $x_1 < x_2 < x_3 < x_4$, the arbitrarily chosen efficient matching (x_1, x_2) and (x_3, x_4) is identical to the efficient matching (t_1, t_2) and (t_3, t_4).

3.2 Splinter Groups

Say that two people i and j are *related* if they are married and/or belong to the same family. Define a *splinter group* as a proper subset of the population where no one in the subset is related to anyone outside of that subset. Given an efficient assignment of spouses in a very large population where traits are independently distributed, it follows that:

Proposition 5 *The probability that splinter groups exist is zero.*

Proof: Define a d-regular multigraph with loops, where every vertex corresponds to a family, and every edge corresponds to a marriage. A splinter group is equivalent to an unconnected component of this graph. Assortative marriages on independent traits generate a random configuration of vertices. A random configuration is equivalent to

[8] Since having equal-valued traits, $t_i = t_j$, is of measure zero, this possibility is ignored.

a regular random multigraph, as defined in Janson et al. (2000). A regular random multigraph is asymptotically almost surely Hamilitonian for $d > 3$ (Janson et al. 2000). Connectivity follows from Hamiltonicity, which rules out the existence of unconnected components, and consequently, the existence of splinter groups. ■

A partial explanation for this result is that if person i marries person j, then because of the independence of traits, it is unlikely that anyone else in i's family marries into j's family as well. As the population grows larger, it becomes less and less likely that there is more than one marriage between the families of i and j. This "mismatch" prevents i and j, and their families, from socially isolating themselves from the larger population. The problem is more interesting than what this partial intuition conveys, however. The likelihood of more than one marriage between two particular families decreases as the population grows larger, but on the other hand, the number of families for whom this event could occur increases. If, for example, d had been equal to two, then these two effects would have balanced, so that small splinter groups would have formed even as the population approached infinity. This proof most likely also goes through for $d \geq 3$, since it really only needs connectivity and since connectivity is closely related to cubic graphs. The fourth edge is necessary in the case of multigraphs to ensure Hamiltonicity, but Hamiltonicity is stronger than connectivity.

In addition to the above proof, we can provide a more structured intuition for no splinter groups by using a branching tree to trace out relationships in the population. Let Σ be the set of all families. Define an arbitrary family in Σ as the singleton set $\sigma(0)$. Let $\sigma(1)$ be the set of families in $\Sigma/\sigma(0)$ with at least one family member married to someone in the original family $\sigma(0)$. Define $\sigma(2)$ as the set of families in $\Sigma/(\sigma(0) \cup \sigma(1))$ with at least one family member married to someone in $\sigma(1)$. Continuing by iteration to more and more distant relations, let $\sigma(r)$ be the set of families in $\Sigma/(\sigma(r-2) \cup \sigma(r-1))$ married to someone in $\sigma(r-1)$. The variable r denotes what is sometimes called the degree of separation between the initial family $\sigma(0)$ and the families in $\sigma(r)$. The degree of separation is a measure of the social distance between individuals; compare Milgram (1967). The collection of these sets, $\cup_{q=0}^{r} \sigma(q)$, constitutes a branching tree. The sets in this collection are mutually exclusive, but if there are splinter groups, the sets are not exhaustive even as $r \to \infty$. Denote by $s(r)$ the cardinality of the set $\sigma(r)$. Since each family in $\sigma(r)$ is composed of d family members, where at least one member in each family by definition is married into $\sigma(r-1)$, the expansion of the tree $\cup_{q=0}^{r} \sigma(q)$ is bounded by

$$s(r+1) \leq s(r)(d-1). \tag{27}$$

If equation (27) holds with equality, then as r increases $s(r)$ very soon encompasses the entire population. It turns out that the equation generally holds as an inequality, however. The reason for this slowdown is threefold. First, a person in $\sigma(r)$ could marry another person in $\sigma(r)$. Second, a family in $\sigma(r)$ could have more than one family member married to someone in $\sigma(r-1)$. Thirdly, several people in $\sigma(r)$ could marry into the same family. These three types of events combine to prevent each family in $\sigma(r)$ from contributing a full $d-1$ new families to $\sigma(r+1)$, and consequently cause (27) to hold as an inequality.

Applying the branching tree $\cup_{q=0}^{r} \sigma(q)$ to the efficient assortative matching, the branching tree is overwhelmingly likely to grow to encompass the entire population in the limit. Since the branching tree only expands to include people who are directly or indirectly related, this limit result is equivalent to Proposition 5 that there are no splinter groups. To see why the entire population is included in the limit, consider what would happen if it were not true, if the branching tree died out without having reached a positive fraction of the population. If this were the case, then $\sigma(r)$ would eventually have to grow arbitrarily small relative to the remainder set $\Sigma / (\sigma(r-2) \cup \sigma(r-1))$, and therefore the likelihood that someone in $\sigma(r)$ married someone else in $\sigma(r)$ rather than in the remainder set, or that several people in $\sigma(r-1)$ married into the same family in $\sigma(r)$ rather than in the remainder set, or that several people in $\sigma(r)$ married into the same family in the remainder set, must also grow arbitrarily small. But then equation (27) should hold as an equality, implying that $s(r+1) > s(r)$, which contradicts the premise that the branching tree died out without having reached the entire population. Consequently, everyone in the population is either directly or indirectly related, and there are no splinter groups.

3.3 Implications for Productivity

The social network developed here allows more individual choice than the random interaction model analyzed earlier, since here industry choice can be made contingent on every aspect of the social structure. The main results from the random interaction model continue to hold nevertheless. A large group cannot align social relationships so as to maximize productivity in a small industry where social interaction and productivity are complementary, without incurring the cost of deteriorating social matches that comes from breaking up into smaller groups. This follows from the result that no splinter groups arise under first-best matching on social traits. Since the social choice set of ethnic minority groups is restricted anyway, these groups can limit their social interactions to a single industry at no alternative cost. Ethnic minorities are therefore well suited for social interaction-intensive industries.

A social network with the same properties could also be derived from a meeting technology where spouses meet and marry at random. The social structure derived here can therefore equally well be thought of as arising in a rigid environment where people meet randomly, as arising from efficient matching. Since randomness is likely to play a role in who marries whom, this adds additional strength to the result. Breaking up into smaller groups does not only carry a social utility cost, but also carries the cost of bypassing random marriages.

3.4 Future Model Extensions

An interesting extension for future work is to include both general and specific skills in the same framework. In such a model of spillovers between sectors, it should be possible to derive stratification in overall entrepreneurial activity as well as industry stratification between different forms of self-employed entrepreneurship at the same time. This would correspond to the current situation in the United States, where groups like the Koreans are strongly clustered in a few business sectors, while at the same time being overrepresented as self-employed owners in almost all other business activities as well.

4 Additional Appendix References:

Heckman, James and Bo Honore. 1990. The empirical content of the Roy model. *Econometrica* 58: 1121-1149.

Jackson, Matthew, and Asher Wolinsky. 1996. A strategic model of social and economic networks. *Journal of Economic Theory* 71: 44-74.

Janson, Svante, Tomasz Luczak, and Andrzej Rucinski. 2000. *Random Graphs.* New York: John Wiley.

Kremer, Michael. 1993. The O-ring theory of economic development. *Quarterly Journal of Economics* 108: 551-575.

Milgram, Stanley. 1967. The small world problem. *Psychology Today* 22: 61-67.

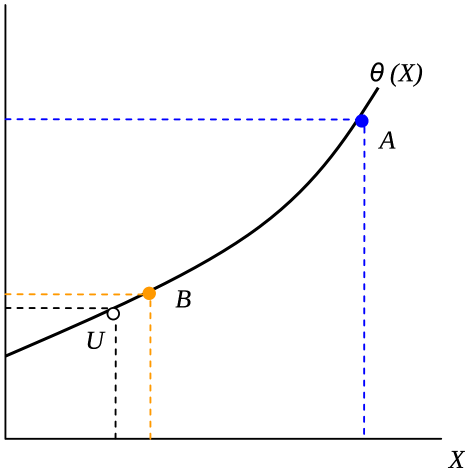

Figure A1. Individual productivity and the three stationary equilibria: one specialized equilibrium with minority specialization (A), one specialized equilibrium with majority specialization (B), and one unstratified equilibrium (U).

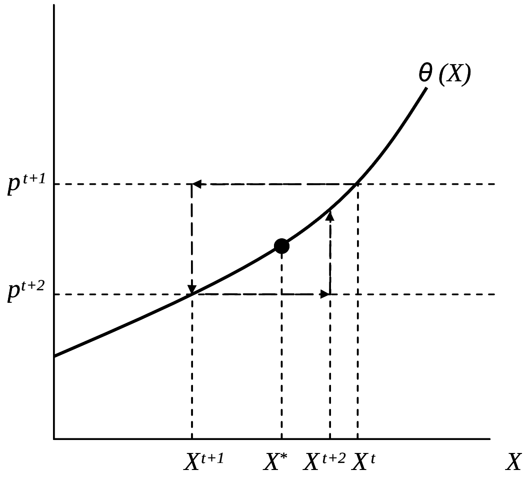

Figure A2. Stable dynamics when the internal effect dominates.

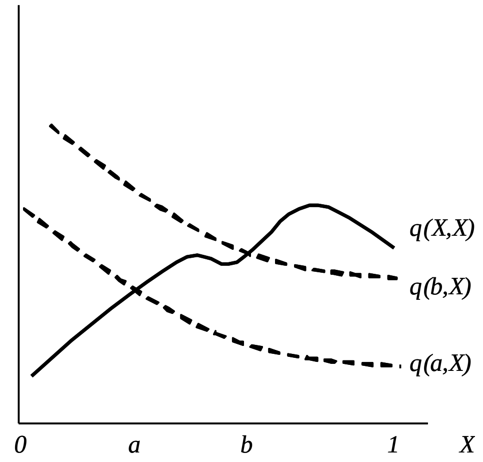

Figure A3. Sorting versus interaction effects in individual productivity. The dotted lines illustrate how the interaction effect raises productivity at all ability levels when specialization increases from *a* to *b*. The solid line shows the productivity of the marginal entrepreneur, for whom *i*=*X* at every level of *X*.

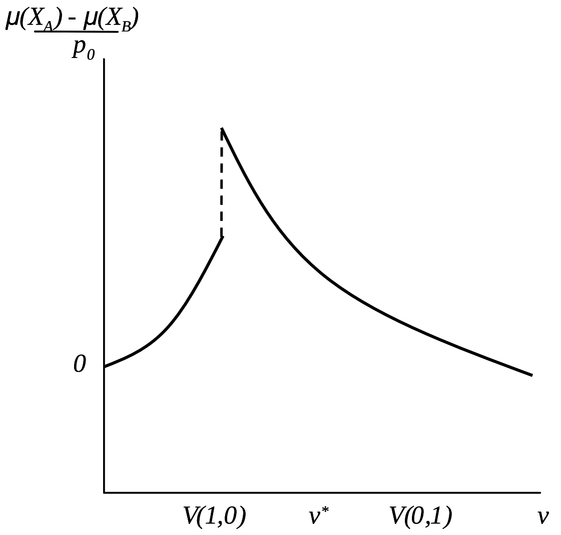

Figure A4. The difference in mean earnings between group *A* and group *B*, for different values of *v*, when minority group *A* specializes.