Entrepreneurship and Co-Villager Networks of Internal Migrants in China

Ying Deng and Xiangjun Ma*

May 9th, 2017 (Preliminary)

Abstract

Internal migrants consist of a large share of the labor market in China. However, insufficient attention has been paid to the heterogeneity of internal migrants, particularly regarding their industry choices as well as their decisions of starting a business upon migration. Employing a unique nationwide survey data, we find that internal migrants are more likely to become entrepreneurs in certain industries if their co-villagers are also entrepreneurs in these industries. Therefore, we observe industry concentration among self-employed migrants. For example, migrants from Sichuan concentrate in the construction industry and those from Zhejiang concentrate in the wholesale and retail industry. We apply the spatial autoregressive (SAR) linear probability model and the SAR logit model to investigate the co-villager network effects on migrants' self-employing decision in three large industries that cover around 65 percent of the self-employed observations, wholesale and retail, lodging and catering as well as construction. Mimicking dominates in both wholesale and retail, and construction industries with significantly positive network effects. This may due to information sharing, such as tactics of running a business, certain channels of product supply, changes of market demand, etc. However, competition dominates in lodging and catering industry with significantly negative network effects.

Keywords: *Entrepreneurship, Co-Villager Network, Spatial Autoregressive Model, Internal Migrants*

JEL Classifications: *R23; C21; L26*

*Ying Deng: University of International Business and Economics, 10 Huixin East Street, Chaoyang District, Beijing, China, 100029, ydeng.econ@gmail.com; Xiangjun Ma: University of International Business and Economics, xm2e@virginia.edu. All errors are the responsibility of the authors.

1 Introduction

Migration in China largely shapes the labor geography of China. While the total number of immigrants in the U.S are 42.4 million, there are 253 million internal migrants, which is one-third of total urban population as of 2014.[1] Although rural to urban migration has been extensively explored in the literature (Meng and Zhang (2001); Zhu (2002); Yang and Guo, 1996), relatively little attention has been paid as to how heterogeneous the migrants are, particularly regarding their labor market choice upon migration. Similar to the concentration of minority groups in specific occupations found in the U.S., we also observe industry clusters among internal migrants from the same home province in China, especially for industries where self-employment is the norm.[2] To recognize and explore such patterns are important for researchers to understand the distribution of entreprenureship in China as well as the network effects on the labor market of internal migrants.

There is large variation of the proportion of entreprenurers across migrants' origin or home provinces. According to our data in 2014, there are 32.4 percent self-employed migrants versus 67.6 percent wage workers or house helpers in the nation, while there are 62.4 percent self-employed migrants among all migrants from Zhejiang province, almost twice as of the average level of the nation (see Table 1A). Similarly, Fujian is another province popularized by entreprenureship, with a percentage of 59.4. On the other hand, there are only 7.9 percent self-employed migrants among all migrants from Yunnan province. Further, around 70 percent self-employed Zhejiang migrants are involved in the "Wholesale and Retail" industry and the other 30 percent are distributed in the other 10 industries. From the perspective of the other dimension, "Wholesale and Retail", "Lodging and Catering", and "Construction" are the 3 industries in which the ratios of self-employed migrants are highest. There are 76.3 percent, 42.2 percent, and 23.3 percent self-employed migrants in these industries, separately. Moreover, self-employed migrants from these three industries account for 43.5%, 14.46% and 7.12% of the total number of self-employed migrants, respectively. Therefore, in our empirical analysis, we focus on these three industries.

The key finding of this paper is that the probability of migrants becoming self-employed entrepreneurs in wholesale and retail industry is higher if their co-villagers are entrepreneurs in this industry. In other words, we observe clusters for self-employed migrants from the same home province in wholesale and retail industry. Similar pattern is also found in construction industry and with a larger network effect. However, we obtain significantly negative network effects in lodging and catering industry which may due to competitions. Table 1B presents the proportion of migrants for each home province given the industry among all self-employed migrants. It seems that the largest proportions of self-employed migrants in all three industries on which we focus are almost all from Anhui, Henan, and Sichuan, except for Zhejiang and Hunan migrant clusters in "Wholesale and Retail". However, it is possible that these are also the largest three provinces from which the largest amount of migrants flows out. Therefore, we also report the relative proportion, obtained by substracting the ratio of migrants for each home province (see Column 1 in Table

[1] Source for Chinese statistics: "Report on China's Migrant Population Development 2015" edited by the National Population and Family Planning Commission of China. Source of U.S statistics is the U.S. Census Bureau's 2014 American Community Survey (ACS).

[2] Kerr and Mandorff (2015) list a few examples, such as the Korean dry cleaners and the Indian motel owners in the U.S. as well as the Jewish merchants in Medieval Europe and the Chinese launderers in early twentieth century California.

1C) from the absolute proportion. In this case, the largest clusters for self-employed migrants in "Wholesale and Retail", "Lodging and Catering", and "Construction" are from Zhejiang, Anhui, and Sichuan.

There are two sources of explanation for the concentration of entreprenureship among migrants from certain origin provinces as well as industry clusters. The first one is that there are common characteristics within the origin home province that are associated with industry choice and entrepreneurship. For example, historical experiences and culture in a region could shape people's character and entreprenurership over time. Botticini and Eckstein (2005) study the transition of Jewish from farmers to merchants or other skilled occupations as a result of higher literacy. Bosacich (1973) cites liquidity, education, religious identity, culture to be factors that affect occupational clusters. Industry structure in the region would also affect the work choice of migrants who flow out of this region (Ma, Qiu, and Son, 2017).

The second source of explanation is the co-villager network, which plays an important role for self-employed migrants to acquire skills and share information in running business in an industry. This is related to social capital concept demonstrated in Borjas (1992, 1995) and Ports (1998). In tradition, Chinese culture values co-villager interactions. Co-villager interaction is one of the instrumental interactions which refer to pragmatic transactions and exchanges related to work or business, such as mutual help in finding employment, job assignment, information exchange, work place cooperation, business transactions and so forth (Chen and Chen, 2004). Kerr and Mandorff (2015) build a model to study the stratification of occupational choice and entrepreneurship, in which social interactions and skill acquisition within ethnic groups are the main mechanism. Relevant skills or knowledge may include how to start or take over a business, market conditions, how to establish supplier, customer, and employee relationships, and insights into legal and tax-related issues. Social interaction with people in the same industry can reduce the cost of skill acquisition. It is convenient to exchange industry information and professional advice if the person is strongly tied in a co-villager network with many entreprenurers in the sector.

In fact, the influence of social networks on migrants' work decisions has been documented by a large strand of literature (Calvo-Armengol and Jackson, 2004; Montgomery, 1991; Carrington, Detragiache, and Vishwanath , 1996). Studies on "Guanxi" (Chinese expression of social networks) in China have also received much attention (Zhang and Li, 2003; Lovett, Simmons and Kali, 1999). As to the co-villager network in China, Chen, Jin and Yue (2010) find that one's migration decision is influenced by her co-villagers because co-villagers help each other in moving cost and job search at the destination. However, the above research focuses on the job search and employment status of wage workers, rather than self-employed migrants.

This paper applies a spatial autoregressive (SAR) model to examine how migrant workers' decision of being an entrepreneur in certain industries is affected by their co-villager networks. The SAR model, also known as the Cliff and Ord (1973) model, has been widely used in empirical analysis of social networks.[3] Lin (2010) applies the SAR model to identify the peer effects in students' academic achievements. Baltagi and Yen (2014) allow spatial correlation among neighboring hospitals and estimate the effects of externalities generated by competition and knowledge spillovers on hospital treatment rates. In our case, the dependent variable of the SAR model is a

[3]LeSage and Pace (2009) have a detailed introduction of the SAR model. Its applications are widespread from studying spatial interactions at the macro level (countries, cities, etc.) to investigating social interactions at the micro level (households, individuals, etc.).

binary variable indicating whether a migrant choose to be self-employed in a specific industry.

We employ the Dynamic Monitoring Survey of Internal Migrants conducted by the National Population and Family Planning Commission of China from 2011 to 2014. This is a nationwide survey with 128,000 to 200,937 individual migrants residing in 326 to 337 host cities in 31 provinces of China. This includes detailed demographic and work information. In particular, we know whether the migrant is self-employed and which industry she is working in. Also, we have the home province and the host city of the migrant. In the benchmark SAR linear and logit regressions, we find that if most co-villagers (from the same home province residing in the same host city), who form the co-villager network of an individual migrant, are self-employed entrepreneurs in a certain industry, it is more likely that the individual migrant will also become an entrepreneur in that industry. [4]

The remainder of the paper is organized as follows. Section 2 presents the SAR linear probability model and the SAR logit model. Section 3 describes data sources. Section 4 discusses the empirical results. Section 5 concludes.

2 Methodology

We construct the spatial autoregressive model (SAR) for three industries "Wholesale and Retail", "Lodging and Catering", and "Construction", respectively. However, we take wholesale and retail industry as an example in the following demonstration, models for the other two industries are constructed similarly.

2.1 The SAR Linear Probability Model

The SAR linear probability model is given by

$$Self employ - retail_i = \lambda \sum_{j=1}^{n} w_{ij} \cdot Self employ - retail_j + X_i\beta + e_i, \quad i = 1,2,...,n, \qquad (1)$$

where $Self employ - retail_i$ is a binary variable of migrant i's choice on whether to become a self-employed entrepreneur in wholesale and retail industry. Specifically, $Self employ - retail_i = 1$ if migrant i is self-employed in wholesale and retail industry, otherwise, $Self employ - retail_i = 0$. Similar variables are defined for "lodging and catering" and "construction" industries. X_i is a $1 \times k$ vector of exogenous characteristics of migrant i, including gender, (rural or urban) Hukou status, marrital status, age, age-squared, years the migrant stayed in the residential city and its quadratic form, race as well as the fixed effects of the home province, the current residential city, education, and the year the survey has been conducted. e_i is an independent error with mean 0 variance $\tilde{\sigma}_i^2$. β is a $k \times 1$ vector of coefficients of these exogenous regressors. The spatial lagged dependent variable $\sum_{j=1}^{n} w_{ij} \cdot Self employ - retail_j$ is a weighted average of the decisions on whether or not to become self-employed in the wholesale and retail industry of all other migrants in the co-villager network of migrant i. We specify the spatial weight matrix W as an $n \times n$ predetermined, row-normalized, symmetric, block-diagonal, sparse matrix with diagonal elements equal to zero. A

[4] Hukou is a social identity status in China, categorized into two types: rural or urban.

typical element w_{ij} represents the weight of migrant j in i's network, hence, $w_{ii} = 0$ and $\sum_{j=1}^{n} w_{ij} = 1$ for $i = 1, 2, ..., n$. The weight matrix is constructed upon migrant i's co-villager network, which is defined as people flowing out of the same home province and residing in the same host city as migrant i. Only people within the network will make an impact on each other, as they are more likely to share information with each other. It is assumed that each migrant is equally affected by all others in the co-villager network. For example, if there are 40 migrants from Zhejiang province working in Beijing, then for each migrant, her Beijing-Zhejiang network consists of 39 people and each person is assigned a weight of 1/39 so as to satisfy the row-normalized condition of the spatial weight matrix. As a result, the spatial lagged dependent variable is the weighted average of the contracting decision of these 39 migrants in the network with the same weight of 1/39. Other migrants in the sample are excluded from the network and assigned zero weights. The coefficient λ captures the network effects on the probability of becoming self-employed in the wholesale and retail industry.

An endogeneity issue arises as the decisions of becoming self-employed in the wholesale and retail industry of individuals in the network are also influenced by the decision of the objective individual in a symmetric way. In other words, $\sum_{j=1}^{n} w_{ij} \cdot Self employ - retail_j$ is correlated with the error term u_i. To achieve consistent estimates, we follow the spatial two-stage least squares (S2SLS) estimates proposed by Kelejian and Prucha (1998). They suggest using all the exogenous variables to construct a set of instruments for the endogenous spatial lagged dependent variable. Define $X = (X_1^{'}, X_2^{'}, ..., X_n^{'})'$ as an $n \times k$ matrix of all exogenous regressors, and then the instrument set is (X, WX).[5]

2.2 The SAR Logit Model

A SAR logit model could better accommodate the binary nature of our dependent variable. However, the nonlinear transformation of a logit model adds complexity in the estimation procedure as the typical maximum likelihood estimation (MLE) often involves n integrals in the likelihood function which can be burdensome when the sample size is large. Several approaches have been proposed to produce consistent estimates for the SAR model with a limited dependent variable. McMillen (1992) suggests an expectation–maximization (EM) algorithm to estimate the coefficients of the spatial probit model. Pinkse and Slade (1998) provide conditions of employing the generalized methods of moments (GMM) estimation to the spatial probit model. LeSage (2000) proposes Bayesian simulation approaches for SAR models.[6] Yet the computation intensities of these estimators depend highly on the sample size n, since they require the inversion of an $n \times n$ matrix. In fact, our data set includes more than 230,000 individuals, which makes it difficult to apply some of the estimation approaches mentioned above, even with strong computational power.

Klier and McMillen (2008) propose a linearized GMM (LGMM) approach, which is specifically designed for large samples. The linearized logit version of the spatial GMM estimator reduces the estimation to two steps - standard logit followed by two-stage least squares. We adopt their LGMM approach and consider the SAR logit model

[5] See Lee (2003) and Kelejian, Prucha and Yuzefovich (2004) for S2SLS estimation with different sets of instruments.

[6] See LeSage and Pace (2009) and Smirnov (2010) for detailed reviews of the estimation methods for the spatial discrete choice models.

$$Selfemploy-retail_i^* = \lambda \sum_{j=1}^{n} w_{ij} \cdot Selfemploy-retail_j^* + X_i\beta + u_i, \quad i = 1,2,...,n, \qquad (2)$$

$$Selfemploy-retail_i = I(Selfemploy-retail_i^* > 0),$$

where $Selfemploy-retail_i^*$ is a latent continuous variable measuring the propensity of migrant i becoming a self-employed entrepreneur in the wholesale and retail industry. Migrant $i's$ propensity depends upon the spatially weighted average of propensities of other migrants in her network, which is expressed as $\sum_{j=1}^{n} w_{ij} \cdot Selfemploy-retail_j^*$, where w_{ij} is the $(i,j)th$ element of the spatial weight matrix W, defined in the same way as in the last section. In fact, $Selfemploy-retail_i^*$ is unobservable. Instead, we can only observe a binary variable $Selfemploy-retail_i$ which is an indicator function of $Selfemploy-retail_i^*$ taking the value of 1 when migrant i becomes a self-employed entrepreneur in the wholesale and retail industry, and 0 otherwise. Coefficient λ captures the network effects on the propensity of becoming a self-employed entrepreneur in the wholesale and retail industry. $\lambda > 0$ implies that higher propensities of becoming a self-employed entrepreneur in the wholesale and retail industry of migrants in the co-villager network increases the propensity of becoming a self-employed entrepreneur in the same industry of migrant i. $\lambda < 0$ implies the opposite. u_i follows logistic distribution with mean 0 variance σ_u^2.

2.3 The Linearized GMM Estimation

We estimate the SAR logit model using a linearized version of the generalized method of moments (LGMM) suggested by Klier and McMillen (2008). Consider the matrix form of Equation (2)

$$Selfemploy-retail^* = \lambda W * Selfemploy-retail^* + X\beta + u, \qquad (3)$$

$$Selfemploy-retail = I(Selfemploy-retail^* > 0),$$

where $Selfemploy-retail^*$, $Selfemploy-retail$, and u are the $n \times 1$ vectors of $Selfemploy-retail_i^*$, $Selfemploy-retail_i$, and u_i, respectively. X is an $n \times k$ matrix of the k exogenous regressors. The reduced form can be written as

$$Selfemploy-retail^* = (I-\lambda W)^{-1}X\beta + (I-\lambda W)^{-1}u. \qquad (4)$$

The variance-covariance matrix is proportional to $\Sigma = [(I-\lambda W)'(I-\lambda W)]^{-1}$, which implies heteroskedasticity and autocorrelation when spatial dependence exists, i.e. $\lambda \neq 0$. Denote the ith diagonal element $of\, \Sigma\, as\, \sigma_i^2$, we take into account the heteroskedasticity by defining $X_i^* = \frac{X_i}{\sigma_i}$ and $X^{**} = (I-\lambda W)^{-1}X^*$ where X^* is an $n \times k$ matrix of X_i^*.

As in Pinkse and Slade(1998), the generalized logit residual can be represented by $\varepsilon_i = Selfemploy-retail_i - P_i$, where $P_i = \frac{exp(X_i^{**}\beta)}{1+exp(X_i^{**}\beta)}$. Define the gradient terms as $G_i = (G_{\beta i}, G_{\lambda i})$, where $G_{\beta i} = \frac{\partial P_i}{\partial \beta} = P_i(1-P_i)X_i^{**}$ and $G_{\lambda i} = \frac{\partial P_i}{\partial \lambda} = P_i(1-P_i)[H_i\beta - \frac{X_i^{**}\beta}{\sigma_i^2}\Lambda_{ii}]$. H_i is the ith row of matrix $H = (I-\lambda W)^{-1}WX^{**}$, Λ_{ii} is the ith diagonal element of $\Lambda = \frac{1}{2}\Sigma[W'(I-\lambda W) + (I-\lambda W)'W]\Sigma$. When $\lambda = 0$, $(I-\lambda W)^{-1}$ degenerates to an identity matrix I. Thus, β can be consistently estimated by

a standard logit model ignoring the spatial structure. The notation can also be greatly simplified as $X_i^{**} = X_i$. Let $\Gamma = (\beta', \lambda)'$ and $\Gamma_0 = (\hat{\beta}_0', 0)'$, where $\hat{\beta}_0$ is the estimate of β in the standard logit model. The gradient terms reduce to $G_{\beta i} = P_i^0(1-P_i^0)X_i$ and $G_{\lambda i} = P_i^0(1-P_i^0)H_i^0\beta_0$ when $\lambda = 0$, where H_i^0 is the ith row of an $n \times k$ matrix $H^0 = WX$.

Linearizing ε_i around the initial estimates of parameter Γ_0, we have

$$\varepsilon_i \approx \varepsilon_i^0 - G_i(\Gamma - \Gamma_0), \tag{5}$$

where $\varepsilon_i^0 = Contract_i - P_i^0$ and $P_i^0 = \frac{exp(X_i\hat{\beta}_0)}{1+exp(X_i\hat{\beta}_0)}$. Re-organizing equation (5) we can obtain

$$\varepsilon_i^0 + G_i\Gamma_0 \approx G_i\Gamma + \varepsilon_i. \tag{6}$$

Define Z_i as the ith row of a matrix of instruments Z. Consider a theoretical moment condition $E(Z_i'\varepsilon_i) = 0$, the corresponding sample moment is thus

$$m(\beta, \lambda) = \frac{1}{n}\sum_{i=1}^{n} Z_i'\varepsilon_i. \tag{7}$$

A GMM estimator can be achieved by minimizing $\varepsilon' ZMZ'\varepsilon$ with respect to β and λ, where M is a positive definite weight matrix and ε is defined as the $n \times 1$ vectors of ε_i. Note that the computation load depends on the sample size n as it involves the inversion of an $n \times n$ matrix. Thus, the GMM estimator becomes infeasible when n is large. However, if M is set to $(Z'Z)^{-1}$, minimizing $\varepsilon' ZMZ'\varepsilon$ is equivalent to conducting a two-stage least squares (2SLS) estimation of a regression with ε as the error term and Z as the set of instruments. From Equation (6), instead of minimizing $\varepsilon' Z(Z'Z)^{-1}Z'\varepsilon$ with respect to β and λ, the GMM estimator can be achieved by performing 2SLS estimation of $\varepsilon_i^0 + G_i\Gamma_0$ on G_i, using a matrix of instruments Z.

In sum, the LGMM estimation procedure can be conducted in the following two steps:

Step 1: Estimate a standard logit model of $Self employ - retail$ with respect to all the exogenous variables X to obtain a consistent estimate of β_0, $\hat{\beta}_0$. Then calculate the residuals ε_i^0 as well as the gradient terms $G_i = (G_{\beta i}, G_{\lambda i})$.

Step 2: Denote $\varepsilon^0 + G_\beta\hat{\beta}_0$, G_β and G_λ as the matrix counterparts of $\varepsilon_i^0 + G_{\beta i}\hat{\beta}_0$, $G_{\beta i}$, and $G_{\lambda i}$. Conduct 2SLS estimation of $\varepsilon^0 + G_\beta\hat{\beta}_0$ on G_β and G_λ using Z as a set of exogenous instruments. More specifically, the 2SLS estimation involves the following two stage regressions:

- Stage 1: Regress G_β and G_λ on Z, respectively, to obtain the predicted values $\hat{G}_\beta$ and $\hat{G}_\lambda$.
- Stage 2: Regress $\varepsilon^0 + G_\beta\hat{\beta}_0$ on $\hat{G}_\beta$ and $\hat{G}_\lambda$. Thus, the corresponding coefficients of $\hat{G}_\beta$ and $\hat{G}_\lambda$ are the estimates of β and λ, respectively. In the empirical analysis, we employ $Z = (X, WX)$ as the instrument set for the LGMM estimation.

The advantage of the LGMM method is that no matrix needs to be inverted, because it requires only the standard logit and linear 2SLS estimation. The linearization significantly reduces the computation time and load as long as λ is small and the true structure is given by Equation (2). See Klier and McMillen (2008) for a detailed discussion of the finite sample properties of LGMM estimation.

3 Data and Summary Statistics

The main data source is the Dynamic Monitoring Survey of Internal Migrants conducted by the National Population and Family Planning Commission of China from 2011 to 2014. This is an annually repeated cross-sectional survey of migrant workers in various host cities, and representative at the host city level. In each year's survey, there are 128,000 to 200,937 individual migrants who are 16-59 years old and residing in 326 to 337 host cities in 32 provinces. From each host province, 2,000 - 15,000 individuals are surveyed depending on the population size of the province. The survey contains each migrant's demographic information, such as birth year, gender, ethnicity, education level, marital status, hukou (rural or urban) and origin (i.e., place of registered hukou), current residential city, how many years the migrant has stayed in the residential city, etc. The survey also provides each individual's employment status, i.e., his/her employment types and industries.

There are 15 industries as categorized in the survey including manufacturing, mining, agriculture/forestry/pasture/husbandry/fishing, construction, electric/coal/water supply, wholesale and retail, lodging and catering, social service, finance/insurance/real estate, transportation/storage/communication, health care/sport/public welfare, education/culture/radio/movie/television, R&D/technology service, government/political organizations/social groups, and other industries. Wholesale and retail, lodging and catering, and construction are the three industries with the highest proportion of self-employed migrants, except for agricultureforestry/pasture/husbandry/fishing and other industries. There are 76.3 percent, 42.2 percent, and 23.3 percent self-employed migrants in these industries, separately. Therefore, in our empirical analysis, we focus on these three industries.

We first define a co-villager network as people from the same province and currently residing in the same host city. In spite of the large population of provinces in China, the average size of the network restricted in the host city is not very big. In our sample with 230,639 individuals, the median size of networks is 81 co-villagers. The mean and standard deviation of the network size are 236 and 381, respectively. The dependent variable is a dummy variable $Self employ - retail_i$ equals to 1 if the individual migrant is self-employed in wholesale and retail, and 0 if the individual is a wage worker or not working in this industry.

4 Empirical Results

Tables 2-4 present the estimation results of the SAR linear probability and the SAR logit models for wholesale and retail industry, lodging and catering industry as well as construction industry, respectively. As specified in Section 2, we use $Z = (X, WX)$ as a set of instruments to deal with the endogeneity problem resulting from the spatial lagged dependent variable. X represents a set of exogenous variables. Specifically, Columns (1) and (3) only control for individual demographic variables including gender (a dummy equals to 1 if gender is male, and 0 otherwise), Hukou status (a dummy equals to 1 if rural, and 0 if urban), marriage status (a dummy equals to 1 if married or have married, and 0 if single), race (a dummy equals to 1 if Han, and 0 if others), age, age-squared, years the migrant has stayed in the residential city and its quadratic form. In Columns (2) and (4) we add fixed effects of the year in which data has been collected, the current residential city, home province and education level.

In wholesale and retail industry, both the SAR linear probability model (using the S2SLS estimation method) and the SAR logit model (using the LGMM estimation method) show significant

positive coefficient estimates of the spatial lagged dependent term. This implies that a migrant's probability of becoming a self-employed entrepreneur in wholesale and retail industry increases with the decisions of other migrants in her co-villager network who choose to be self-employed in the same industry. However, the inclusion of more control variables lead to a reduction of about 1/3 on the network effects, dropping from 0.3619 to 0.2168 for the SAR linear probability models. Taking Column (2) as the baseline outputs for the SAR linear probability model with controls for fixed effects, the result suggests that when the percentage of migrants in the co-villager network becoming a self-employed entrepreneur in wholesale and retail industry increases by 10%, the probability for the targeting migrant to become self-emloyed in the same industry increases by about 2%. Similar spatial pattern is found in construction industry, but with larger magnitude in the spatial dependence. According to Column (2) in Tables 2 and 4, the network effect is 0.3605 in construction industry while it is only 0.2168 in wholesale and retail industry. However, unlike these two industries, lodging and catering industry shows significantly negative spatial correlation. This implies that a migrant's decision on becoming self-employed in lodging and catering industry decreases other co-villager migrants' probabilities of becoming self-employed in the same industry. A plausible explanantion is that the requirements to start a business, including early investment in acquiring operation permission, decorating store, hiring employees, etc., are usually lower in the wholesale and retail industry or construction industry than that in the lodging and catering industry. Migrants are more likely to become self-employed in industries with lower entry barriers, especially when they can acquire skills and share information with co-villagers who are self-employed in the same industry. However, for industries with higher entry requirements, such as lodging and catering, migrants may choose to work for their self-employed co-villagers instead of starting a new one.

We also find that female migrants are more likely to start a business in wholesale and retail industry as well as lodging and catering industry, whereas male migrants are more likely to become self-employed in construction industry. This is because construction industry is usually male dominated as it involves a lot of heavy manual work comparing to the other two industries. Married migrants, in general, have higher probability to start their own business as it provides more job flexibility. Han migrants and migrants with rural Hukou are more likely to become self-employed in wholesale and retail industry as well as construction industry. But Han migrants have a negative effect and Hukou status does not have a significant effect in lodging and catering industry.

In addition, the relationship between age and the probability of becoming a self-employed entrepreneur is inverted U-shaped in most cases. That is, the probability of becoming self-employed increases as age increases at the early stage, however, the increasing rate decreases as migrants becoming more and more experienced. At some point, for example, around 23.5 years old as showed in Column (2) for migrants in wholesale and retail industry, the probability starts to fall as migrants getting older.

Lastly, in wholesale and retail industry, we observe a similar inverted U-shape relationship between the time migrants have stay in the residential city and the probability of becoming self-employed entrepreneurs. Yet in the other two industries, we find that the longer migrants have stayed in a city, the less likely they are to start a business.

5 Conclusion

This paper employs a comprehensive data set from a nationwide survey to study how co-

villager networks affect internal migrants' decision of becoming self-employed entrepreneurs in three industries, including wholesale and retail, lodging and catering as well as construction. We include a spatial lagged dependent term to capture the network effect. Besides the SAR linear probability model estimated by the S2SLS method, a linearized GMM approach is conducted to estimate the SAR logit model given the large data set. The network effect is significantly positive in wholesale and retail as well as construction industries while larger in construction industry. Mimicking dominates in these two industries probably because co-villager networks could help migrants build skills required in a sector and share information, such as tactics of running a business, certain channels of product supply, changes of market demand, etc. However, the network effect is significantly negative in lodging and catering industry which may due to intense competitions.

Appendix

Table 1: Proportion of Self-Employed Migrants

Table 1A: Proportion of Self-Employed Migrants for Each Home Province

Home Province	Proportion of self-employed workers in the home province (%)
Beijing	9.90
Tianjin	23.46
Hebei	30.94
Shanxi	32.70
Inner Mongolia	25.43
Liaoning	21.73
Jilin	25.62
Heilongjiang	28.14
Shanghai	14.29
Jiangsu	32.85
Zhejiang	**62.40**
Anhui	32.68
Fujian	**59.36**
Jiangxi	34.71
Shandong	31.33
Henan	35.19
Hubei	34.48
Hunan	36.51
Guangdong	37.90
Guangxi	23.27
Hainan	18.75
Chongqing	34.51
Sichuan	28.36
Guizhou	14.63
Yunnan	7.94
Tibet	15.38
Shaanxi	30.14
Gansu	34.64
Qinghai	46.00
Ningxia	26.33
Xinjiang	32.27

Table 1B: Proportion for Each Home Province

Home Province	Among all self-employed migrants in **wholesale and retail** industry, the proportion from each home province (%)	Among all self-employed migrants in **lodging and catering** industry, the proportion from each home province (%)	Among all self-employed migrants in **construction** industry, the proportion from each home province (%)
Beijing	0.03	0.04	0.05
Tianjin	0.17	0.11	0.21
Hebei	5.52	3.14	3.17
Shanxi	1.19	2.8	0.85
Inner Mongolia	1.01	0.94	1.32
Liaoning	0.74	0.63	0.32
Jilin	1.11	1.82	0.63
Heilongjiang	2.67	3.27	2.17
Shanghai	0.06	0	0.05
Jiangsu	3.65	1.55	3.65
Zhejiang	**8.36**	2.83	1.69
Anhui	**10.08**	**16.57**	**11.04**
Fujian	6.04	5.27	2.8
Jiangxi	6.05	6.12	5.92
Shandong	6.4	4.71	4.7
Henan	**11.1**	**10.11**	**9.51**
Hubei	6.72	5.27	5.97
Hunan	**8.77**	4.3	6.23
Guangdong	1.59	0.7	1.11
Guangxi	1.14	1.75	1.8
Hainan	0.08	0.11	0.11
Chongqing	3.41	5.27	5.92
Sichuan	**7.59**	**9.24**	**22.35**
Guizhou	1.26	1.91	2.91
Yunnan	0.31	0.25	0.58
Tibet	0.01	0	0
Shaanxi	1.91	3.05	1.53
Gansu	2.36	5.65	2.96
Qinghai	0.25	1.39	0.16
Ningxia	0.13	0.49	0.11
Xinjiang	0.26	0.65	0.21

Table 1C: Relative Proportion for Each Home Province

Home Province	Proportion of migrants of the home province (%)	Among all self-employed migrants in **wholesale and retail** industry, the **relative** proportion from each home province (%)	Among all self-employed migrants in **lodging and catering** industry, the **relative** proportion from each home province (%)	Among all self-employed mig **construction** industry, the **r** proportion from each home pro
Beijing	0.12	-0.09	-0.08	-0.07
Tianjin	0.19	-0.02	-0.08	0.02
Hebei	4.61	0.91	-1.47	-1.44
Shanxi	1.40	-0.21	1.4	-0.55
Inner Mongolia	1.61	-0.6	-0.67	-0.29
Liaoning	1.01	-0.27	-0.38	-0.69
Jilin	1.74	-0.63	0.08	-1.11
Heilongjiang	3.51	-0.84	-0.24	-1.34
Shanghai	0.08	-0.02	-0.08	-0.03
Jiangsu	3.07	0.58	-1.52	0.58
Zhejiang	2.63	**5.73**	0.2	-0.94
Anhui	**12.03**	-1.95	**4.54**	-0.99
Fujian	2.35	3.69	2.92	0.45
Jiangxi	5.46	0.59	0.66	0.46
Shandong	5.78	0.62	-1.07	-1.08
Henan	**10.90**	0.2	-0.79	-1.39
Hubei	5.66	1.06	-0.39	0.31
Hunan	6.15	2.62	-1.85	0.08
Guangdong	1.05	0.54	-0.35	0.06
Guangxi	2.52	-1.38	-0.77	-0.72
Hainan	0.13	-0.05	-0.02	-0.02
Chongqing	3.52	-0.11	1.75	2.4
Sichuan	**11.41**	-3.82	-2.17	**10.94**
Guizhou	3.86	-2.6	-1.95	-0.95
Yunnan	1.68	-1.37	-1.43	-1.1
Tibet	0.02	-0.01	-0.02	-0.02
Shaanxi	2.49	-0.58	0.56	-0.96
Gansu	3.58	-1.22	2.07	-0.62
Qinghai	0.30	-0.05	1.09	-0.14
Ningxia	0.43	-0.3	0.06	-0.32
Xinjiang	0.60	-0.34	0.05	-0.39

Table 2:Network Effects on Self-employment in Wholesale & Retail Industry

	SAR Model		SAR Logit Model	
	(1)	(2)	(3)	(4)
W*Selfemp_retail	0.3619***	0.2168***	0.0335***	0.0455***
	(0.0134)	(0.0135)	(0.0108)	(0.0063)
Sex	-0.0119***	-0.0176***	-0.0832***	-0.2047***
	(0.0014)	(0.0014)	(0.0123)	(0.0153)
Hukou Status	0.0349***	0.0104***	0.3268***	0.0727***
	(0.0021)	(0.0023)	(0.0211)	(0.0238)
Marriage Status	0.0747***	0.0679***	0.9827***	1.9335***
	(0.0023)	(0.0023)	(0.0544)	(0.1200)
Age	0.0067***	0.0047***	0.0863***	0.0934***
	(0.0006)	(0.0006)	(0.0050)	(0.0089)
Age^2	-0.0001***	-0.0001***	-0.0010***	-0.0012***
	(0.0000)	(0.0000)	(0.0001)	(0.0001)
Stay years	0.0112***	0.0121***	0.0926***	0.1770***
	(0.0004)	(0.0004)	(0.0038)	(0.0083)
Stay years^2	-0.0004***	-0.0004***	-0.0032***	-0.0061***
	(0.0000)	(0.0000)	(0.0002)	(0.0003)
Race	0.0465***	0.0184***	0.6903***	0.3736***
	(0.0031)	(0.0032)	(0.0424)	(0.0540)
Regressions Control for				
Year	no	yes	no	yes
Host City	no	yes	no	yes
Home Province	no	yes	no	yes
Education Level	no	yes	no	yes
# of Observations	230,639	230,639	230,639	230,639
R-Squared	0.1223	0.1767	0.3120	0.6484

Note: Columns (1),(2) are spatial two-stage least squares estimates with different sets of controls while columns (3),(4) are linearized GMM estimates with the corresponding sets of controls. Standard deviations are in parenthesis. ***, **, * indicate significance at the 1%, 5% and 10% level.

Table 3:Network Effects on Self-employment in Lodging & Catering Industry

	SAR Model		SAR Logit Model	
	(1)	(2)	(3)	(4)
W*Selfemp_catering	-0.0623*	-0.1015***	-0.3138***	-0.1347***
	(0.0346)	(0.0276)	(0.0235)	(0.0200)
Sex	-0.0040***	-0.0054***	-0.0931***	-0.1053***
	(0.0009)	(0.0009)	(0.0201)	(0.0239)
Hukou Status	0.0150***	-0.0009	0.4420***	-0.0460
	(0.0013)	(0.0015)	(0.0387)	(0.0404)
Marriage Status	0.0327***	0.0284***	1.3862***	0.9327***
	(0.0015)	(0.0015)	(0.1039)	(0.2470)
Age	0.0036***	0.0032***	0.1100***	0.0723***
	(0.0004)	(0.0004)	(0.0073)	(0.0159)
Age^2	-0.0000***	-0.0000***	-0.0013***	-0.0009***
	(0.0000)	(0.0000)	(0.0001)	(0.0002)
Stay years	-0.0013***	-0.0008***	-0.0284***	-0.0136**
	(0.0002)	(0.0002)	(0.0054)	(0.0059)
Stay years^2	0.0000	0.0000	0.0002	-0.0000
	(0.0000)	(0.0000)	(0.0003)	(0.0003)
Race	-0.0161***	-0.0240***	-0.3625***	-0.5085***
	(0.0020)	(0.0021)	(0.0374)	(0.0765)
Regressions Control for				
Year	no	yes	no	yes
Host City	no	yes	no	yes
Home Province	no	yes	no	yes
Education Level	no	yes	no	yes
# of Observations	230,639	230,639	230,639	230,639
R-Squared		0.0389	0.0810	0.1697

Note: Columns (1),(2) are spatial two-stage least squares estimates with different sets of controls while columns (3),(4) are linearized GMM estimates with the corresponding sets of controls. Standard deviations are in parenthesis. ***, **, * indicate significance at the 1%, 5% and 10% level.

Table4:Network Effects on Self-employment in Construction Industry

	SAR Model		SAR Logit Model	
	(1)	(2)	(3)	(4)
W*Selfemp_construction	0.5740***	0.3605***	0.1769***	0.0318***
	(0.0286)	(0.0421)	(0.0192)	(0.0117)
Sex	0.0215***	0.0214***	1.0545***	-0.0926
	(0.0006)	(0.0007)	(0.0713)	(0.3514)
Hukou Status	0.0095***	0.0051***	0.5708***	0.0546
	(0.0009)	(0.0011)	(0.0703)	(0.0899)
Marriage Status	0.0113***	0.0111***	0.6596***	-0.2494
	(0.0011)	(0.0011)	(0.0989)	(0.2975)
Age	0.0015***	0.0014***	0.0950***	-0.0125
	(0.0003)	(0.0003)	(0.0109)	(0.0291)
Age^2	-0.0000***	-0.0000***	-0.0011***	0.0002
	(0.0000)	(0.0000)	(0.0001)	(0.0004)
Stay years	-0.0004***	-0.0004**	-0.0177***	-0.0082
	(0.0002)	(0.0002)	(0.0058)	(0.0063)
Stay years^2	0.0000***	0.0000***	0.0010***	0.0005*
	(0.0000)	(0.0000)	(0.0003)	(0.0003)
Race	0.0073***	0.0074***	0.5625***	-0.0684
	(0.0014)	(0.0015)	(0.1051)	(0.2018)
Regressions Control for				
Year	no	yes	no	yes
Host City	no	yes	no	yes
Home Province	no	yes	no	yes
Education Level	no	yes	no	yes
# of Observations	230,639	230,639	230,639	230,639
R-Squared	0.0361	0.0452	0.2730	0.6271

Note: Columns (1),(2) are spatial two-stage least squares estimates with different sets of controls while columns (3),(4) are linearized GMM estimates with the corresponding sets of controls. Standard deviations are in parenthesis. ***, **, * indicate significance at the 1%, 5% and 10% level.

References

[1] Baltagi, Badi H., and Yin-Fang Yen. "Hospital treatment rates and spillover effects: Does ownership matter?." Regional Science and Urban Economics 49 (2014): 193-202.

[2] Botticini, Maristella and Zvi Eckstein.Jewish occupational selection: Education, restrictions,or minorities? The Journal of Economic History, 65.4 (2005): 922-948.

[3] Calvo-Armengol, Antoni, and Matthew O. Jackson. "The effects of social networks on employment and inequality." The American Economic Review 94.3 (2004): 426-454.

[4] Carrington, William J., Enrica Detragiache, and Tara Vishwanath. "Migration with endogenous moving costs." The American Economic Review 86.4 (1996): 909-930.

[5] Chen, Yuyu, Jin G. Zhe, and Yue Yang. "Peer migration in China." NBER Working Paper #15671 (2010).

[6] Cliff, Andrew, and Keith Ord. Spatial Autocorrelation. Pion, London, 1973.

[7] Kelejian, Harry H., and Ingmar R. Prucha. "A generalized spatial two-stage least squares procedure for estimating a spatial autoregressive model with autoregressive disturbances." The Journal of Real Estate Finance and Economics 17.1 (1998): 99-121.

[8] Kelejian, Harry H., Ingmar R. Prucha, and Yevgeny Yuzefovich. "Instrumental variable estimation of a spatial autoregressive model with autoregressive disturbances: Large and small sample results." Advances in Econometrics: Spatial and Spatio-Temporal Econometrics (2004): 163-198.

[9] Klier, Thomas, and Daniel P. McMillen. "Clustering of auto supplier plants in the United States." Journal of Business & Economic Statistics 26.4 (2008): 460–471

[10] Lee, Lung-fei. "Best spatial two-stage least squares estimators for a spatial autoregressive model with autoregressive disturbances." Econometric Reviews 22.4 (2003): 307-335.

[11] LeSage, James P. "Bayesian estimation of limited dependent variable spatial autoregressive models." Geographical Analysis 32.1 (2000): 19-35.

[12] LeSage, James, and Robert Kelley Pace. Introduction to Spatial Econometrics. CRC Press, 2009.

[13] Lin, Xu. "Identifying peer effects in student academic achievement by spatial autoregressive models with group unobservables." Journal of Labor Economics 28.4 (2010): 825-860.

[14] Lovett, Steve, Lee C. Simmons, and Raja Kali. "Guanxi versus the market: Ethics and efficiency." Journal of International Business Studies 30.2 (1999): 231-247.

[15] McMillen, Daniel P. "Probit with spatial autocorrelation." Journal of Regional Science 32.3 (1992): 335-348.

[16] Montgomery, James D. "Social networks and labor-market outcomes: Toward an economic analysis." The American economic review (1991): 1408-1418.

[17] Pinkse, Joris, and Margaret E. Slade. "Contracting in space: An application of spatial statistics to discrete-choice models." Journal of Econometrics 85.1 (1998): 125-154.

[18] Smirnov, Oleg A. "Modeling spatial discrete choice." Regional Science and Urban Economics 40.5 (2010): 292-298.

[19] Zhang, Xiaobo, and Guo Li. "Does guanxi matter to nonfarm employment?." Journal of Comparative Economics 31.2 (2003): 315-331.