Preliminary draft
Comments welcome

Social-Family Network and Self-Employment: Evidence from Temporary Rural-Urban Migrants in China

Junfu Zhang and Zhong Zhao*

Abstract

We hypothesize that individuals with a larger social-family network are more likely to choose self-employment. We test this hypothesis using data on temporary rural-urban migrants in China. The size of a migrant's social-family network is measured by the number of relatives and friends this migrant greeted during the past Spring Festival. Our empirical analysis faces two challenges. First, there is an endogeneity problem in that a migrant may want to develop and maintain a large social-family network exactly because he is self-employed. For this reason, a simple correlation between the probability of being self-employed and the size of the migrant's social-family network cannot be interpreted as causal. Second, the size of social-family network is measured using survey data, which is subject to measurement errors. To overcome these problems, we take an instrumental variable (IV) approach. More specifically, we examine how faraway an individual migrated when he first moved to a city and use this variable to instrument for the current size of social-family network. We establish the credibility of the IV by emphasizing the unique institutional context of rural-urban migration in China and focusing on the sample of migrants who originally started as wage workers in urban areas and currently are not on their first jobs. Our IV results indeed show that a rural-urban migrant with a larger social-family network is more likely to be self-employed in the city. This finding is robust to alternative model specifications and various restrictions on the sample used in estimation.

Keywords: Social-family network, self-employment, rural-urban migrants.

JEL classifications: J23, J61, D85.

---

* Zhang is an assistant professor of economics at Clark University and Zhao a professor of economics at Renmin University of China. Both are research fellows at the IZA. We thank Randy Akee, Evan Due, Shihe Fu, Delia Furtado, Wayne Gray, Jiang Qian, and Xin Meng for stimulating discussions on this topic. We are grateful for comments and suggestions from participants at the RUMiCI conference in Yogyakarta, Indonesia and seminars at Clark University and Renmin University of China. Collection of the Rural Urban Migration in China (RUMIC) data used in this paper is financed by IZA, ARC/AusAid, the Ford Foundation, and the Ministry of Labor and Social Security of China. Zhong Zhao would like to acknowledge financial support from the Mingde Scholar Program at Renmin University of China.

1. Introduction

In this paper, we examine how having access to a larger social-family network affects an individual's choice of self-employment.[1]

Self-employment and especially creating and running a business is often a much more challenging task than finding a job and working for an employer. Consider a cook who wants to work at a restaurant. He only needs to find a job that fits his qualifications and interests, and then routinely provides his services at the restaurant day after day. What if he wants to have his own restaurant? Then he has to find a location where a new restaurant can possibly survive; he needs to rent a place for the operation; he needs to secure some money as start-up capital; he needs to deal with local government bureaucracies to obtain permits and licenses; he may need some helpers even if he still works as the cook himself.

Getting the restaurant open is only a start. The owner then has to continuously think about how to attract more customers, where to find dependable suppliers of raw food, and how to cut costs; he needs to figure out the demand fluctuations during a day, throughout a week, and over a year, and respond to them accordingly; he needs to know at least some elementary accounting to keep books in order. Additionally, there are all kinds of random events to deal with, some of which have little to do with the normal business of preparing and serving food. For example, two customers have a heated argument in the restaurant that needs intervention; some local rogues demand a protection fee; a customer gets sick after a meal at the restaurant and threatens to sue; and so on. Clearly, a self-employed person has a lot more to manage than a typical employee.

At each stage of this endeavor of self-employment, a social and family network is often the most reliable source of assistance. This is particularly true in a society like China, on which our empirical analysis will focus. For example, when one needs financial capital for initial investment, one turns to family members, relatives, and close friends. Similarly, when a self-employed individual needs to find customers, the word-of-mouth advertising by friends and relatives is often more effective than advertising

[1] Self-employment is the status of working for oneself instead of for an employer. A self-employed individual may work alone or own and run a business that also hires other people. Following the general practice in the literature (see, e.g., Evans and Jovanovic, 1989; Evans and Leighton, 1989; Blanchflower and Oswald, 1998; Fairlie, 1999; and Parker, 2004), we sometimes also refer to a self-employed person as an entrepreneur and the choice of self-employment as entrepreneurship.

through formal channels; when a small business owner needs to hire a helper, he or she also asks friends and relatives for recommendations and referrals. More importantly, in a developing country where the institutional environment is full of uncertainty and hidden rules, a self-employed individual constantly needs personal connections to facilitate the navigation in such a system (Yueh, 2009). Indeed, an extensive literature has documented that the self-employed rely heavily on the assistance of friends and family members.[2]

Given that a well-developed social-family network can greatly increase the feasibility of self-employment and enhance the chance of success for the self-employed, one would naturally hypothesize that individuals with a larger social-family network are in a better position to choose self-employment. In this paper, we empirically investigate whether this is indeed true.

Our study brings together two strands of literature. One concerns how personal networks affect an individual's labor market outcome and the other regards the various factors that influence a person's decision to become self-employed or engage in entrepreneurship.

There is a vast and growing literature on networks and labor market outcomes, focusing mainly on how social and family connections increase one's employment opportunities and earnings.[3] The bulk of this literature is motivated by the idea that a large social-family network facilitates job search because social contacts, relatives, and family members can provide job openings information as well as referrals. This line of research does not distinguish between wage workers and the self-employed. We want to emphasize here that a well-developed network is more important for the self-employed than for wage workers. Without a supportive network, it is still possible to find a job but will be extremely difficult to survive as a self-employed. Moreover, while "weak ties" and "informal networks" are generally good enough to be helpful when one looks for a wage-earning job (Granovetter, 1973; Bayer et al. 2008), strong ties are often necessary for the kind of assistance needed during self-employment. For example, an acquaintance

[2] See, for example, Birley (1985), Burt (1997), Brüderl and Preisendörfer (1997), Allen (2000), and Greve and Salaff (2003).

[3] See Montgomery (1991), Ioannides and Loury (2004), Jackson (2008) for excellent reviews of this literature. Some recent work has paid careful attention to the problem of endogenous network formation (see, e.g., Munshi, 2003; Luke and Munshi, 2006; Beaman, 2009; and Laschever, 2009). Bian (1994) and Zhang and Li (2003) study networks and labor market outcomes in the context of China, but neither investigates the relationship between social-family network and the choice of self-employment.

in your neighborhood may provide you some useful information about job openings at his company, but it is unlikely that he will lend you money when you are in need of capital as a small business owner. The latter type of help almost always comes from family members, relatives, or close friends. For these reasons, we expect that a large social-family network is crucial for self-employment and having access to such a network puts one in a better position to be self-employed.

The existing literature on the choice of self-employment or entrepreneurship has mostly focused on factors such as liquidity constraint, human capital, and family background. There has been considerable evidence that higher household wealth increases the probability of entrepreneurship, perhaps by relaxing capital market constraints.[4] A person's human capital matters too. For example, Lazear (2004, 2005) shows that individuals with more balanced skills, acquired through formal education or work experience, are more likely to become entrepreneurs. Others find that family and social backgrounds, such as having a self-employed parent or residing in highly entrepreneurial neighborhoods, also have an effect on the choice of self-employment.[5] However, the size of social-family network, as a potential determining factor in self-employment decisions, is very much under-researched.

We have been able to discover only two empirical studies related to ours, each examining the simple correlation between the size of a person's social-family network and the choice of self-employment. Using survey data on 595 residents in the U.S. state of Wisconsin, Allen (2000) finds that the probability of self-employment is positively correlated with the size of family network, although not correlated with the number of friends. Using survey data on some 9,000 working-age adults in 13 cities in China, Yueh (2009) finds that an individual is more likely to be self-employed when the size of social network is larger. Her estimates indicate that having one more person in the social

[4] Evans and Jovanovic (1989), Evans and Leighton (1989), Holtz-Eakin, Joulfaian, and Rosen (1994a, 1994b), Blanchflower and Oswald (1998), and Fairlie (1999) all offer some supportive evidence, although Hurst and Lusardi (2004) cast doubt on some of these findings. In the context of China, Wang (2008) finds that the relaxation of constraints on capital (as well as job mobility), as a result of a wealth shock created by a housing reform, has increased self-employment.

[5] See, for example, Lentz and Laband (1990), Dunn and Holtz-Eakin (2000), Hout and Rosen (2000), and Giannetti and Simonov (2009). Djankov et al. (2006) report that in China immediate and extended family members of entrepreneurs are nearly three times more likely to be entrepreneurs themselves than family members of non-entrepreneurs.

network is associated with a 0.03-percentage-point increase in the probability of self-employment, a very small though statistically significant effect.

There are two problems with empirical analyses based on simple multivariate regressions using survey data, such as those reported by Allen (2000) and Yueh (2009). One is the endogeneity issue. That is, not only may a large social-family network affect the probability of choosing self-employment, the self-employed also have incentive to develop a large network. As a result, a simple correlation between self-employment status and network size does not necessarily imply a causal effect of network size on the choice of self-employment. Indeed, Yueh (2009) has recognized this very concern and cautioned against interpreting her estimated coefficient as a causal effect. The other problem stems from the measurement of social-family network size. Because it is difficult to directly count a person's family members and friends, researchers have to rely on self-reported numbers in survey data to measure the size of social-family network. Both Allen (2000) and Yueh (2009), as well as our study here, use such self-reported data. Due to respondents' imperfect recall and their tendency to report rounded numbers, we suspect that there are serious measurement errors in network-size variables constructed using survey data. Such errors, if not taken into account, will also bias the estimates from simple multivariate regressions.

In this paper, we empirically examine whether individuals with a larger social-family network are more likely to choose self-employment, paying close attention to the issues of endogeneity and measurement errors. We use a survey database that was recently constructed in China. Our data contain detailed information about rural-urban migrants in China, including many variables on their personal characteristics as well as their social and family networks.

We take the standard instrumental variable (IV) approach to overcome the endogeneity and measurement-errors problems. In particular, we use *the distance from home province when a migrant first moved to the urban area* as an instrument for social-family network size today. As will be shown, the migrants who originally migrated far away from home tend to have a smaller social-family network today, because the networks of rural people tend to be local and long-distance migration disrupts their previously established networks. Using this distance as an instrument, we are assuming

that it does not directly affect a migrant's self-employment decision through any other uncontrolled channels.

We believe this assumption is plausible for several reasons. First, as we will emphasize below, the unique institutional context of rural-urban migration in China has determined that the first-time migrants face a great deal of uncertainty and almost always consider the move temporary. The decision where to migrate in the first time is particularly uninformed and largely random, depending on where some early movers they knew had gone. So the distance of the first-time migration is arguably exogenous. Second, our analysis focuses on a sample of migrants who all started as wage workers in urban sectors and all have changed jobs over time. Since none of them was self-employed originally and none of them is on the first job any more, it is plausible that whether they are self-employed today is not directly affected by the distance of their first migration. And third, we control for home province fixed effects. By comparing migrants from the same province, we think it is more reasonable to consider the distance of first migration as exogenous to today's employment status.

We find that migrants with a smaller social-family network, as a result of a longer-distance migration in the past, are less likely to be self-employed today. This finding holds true for various network size measures and it is robust to different model specifications and sample restrictions. We consider these results convincing evidence that the size of social-family network has a positive effect on the choice of self-employment.

The rest of the paper is organized as follows. Section 2 discusses the unique institutional context in China. Section 3 describes the data and our empirical strategies. Section 4 presents empirical results. Section 5 concludes with some remarks.

2. Institutional Setting

2.1 Temporary rural-urban migration in China

Along with its fast economic growth, China has experienced a rapid urbanization during the past three decades. The share of urban population in China has risen from 18 percent in 1978 to 46 percent in 2008. This fast urban growth is achieved primarily through a massive migration from rural areas to cities (Zhang and Song, 2003).

According to the National Bureau of Statistics, by the end of 2008, there was a total of 225 million rural-urban migrants in China.[6]

This recent wave of rural-urban migration in China occurred in a unique institutional context. On the one hand, there is a long-standing residence registration (*hukou*) system in China, designed to control the movement of people within the country (Chan and Zhang, 1999). Each individual is issued a residence permit, a so-called *hukou*, which gives the person the right to live in a jurisdiction and access local public goods such as public education and health care. If a person with a rural *hukou* wants to move to a city and work in the urban sectors, he or she has to apply through the relevant bureaucracies. Since the mid-1980s, this system has been gradually relaxed and the controls have been weakened, primarily in response to the rapid expansion of the urban economy and the increased demand for cheap labor in urban sectors. However, although people with a rural *hukou* are now generally allowed to find work in urban areas, jobs in certain urban sectors are still reserved only for residents with the local urban *hukou* and the migrants from rural areas have very limited access to urban public goods.[7]

On the other hand, a household responsibility system was implemented in the late 1970s in countryside, which was a key component of economic reform in China (Lin, 1992). In rural areas, land ownership belongs to local economic collectives. Under the household responsibility system, land use right is contracted to households, with the size of the land for each household determined by the number of household members who have a *hukou* in the village. As long as farmers fulfill grain procurement obligations, they can retain the surplus for their own use or sell it on the market. Over the years, the central government removed most of the procurement obligations; in 2006, China also repealed all agricultural taxes to lift the burden on farmers.[8] Thus a farmer who does not want to seek employment in urban areas can make a basic living by farming on his family's land. Similarly, a migrant who has difficulty in finding a job in urban areas, due to a slowdown

[6] In China, these migrants are commonly referred to as *nongmingong*, meaning "farmers-turned workers."
[7] For example, jobs in government agencies and state-owned enterprises are generally inaccessible to rural-urban migrants without an urban *hukou*. Migrant workers are overrepresented in blue-collar occupations (Meng and Zhang, 2001). Also, rural-urban migrants are not entitled to housing, medical, and educational subsidies available to urban residents. For example, if these migrants want to have their children enrolled in public schools in the city, they have to pay an extra "temporary student fee" that is many times higher than the tuition paid by regular local students.
[8] Occasionally, farmers still have to pay head taxes and fees to fund local public works.

of the urban economy or any other reasons, can always return to his village and resume farm work on his family's land.

Because of this institutional arrangement, rural-urban migration in China gives the impression of being "temporary." The migrants, even having lived and worked in a city for many years, tend to consider themselves as outsiders and are reluctant to make an effort to assimilate into the city. They also tend to be footloose and move from one city to another to chase jobs. Partly because they feel unwelcome in the city and partly because they have access to a piece of land back in their villages, rural-urban migrants tend to consider their villages as homes and many of them leave their children in their villages together with grandparents. These migrants regularly send money back to pay for their children's education, build houses, or make other investments (Wei, 2008).

2.2 China as a "guanxi society"

In Chinese, *guanxi* means connections. China is a "*guanxi* society" where connections really matter and personal relationships are central in every aspect of the society. Despite a comprehensive economic reform aimed to establish institutions compatible with a modern market economy, doing business in China, to a great extent, is still about managing interpersonal relationships rather than faceless transactions (Xin and Pearce, 1996; Luo, 2007).

Consider an aspiring entrepreneur who needs to borrow some money from a bank in China. His most important task is not to craft a sound business plan or put up enough collateral. Rather, he will have to find out whether he can get to know one of the loan officers in person through a friend or a relative. Such a personal connection is often more helpful than a good business plan.

The same is true for the self-employed; their business opportunities often come through personal connections. In Xu and Qian (2009), there is a revealing story about a rural-urban migrant who makes a living by sharpening scissors for others. He is very good at his job, but he earns far less money than a local competitor. The local person, not necessarily a better scissor-sharpener, knows the owner of an apparel factory that has thousands of scissors and uses his service every other week. Similarly, while an ordinary scrap metal collector has to dig around at junk yards, a person whose relative is managing a state-owned steel factory can regularly pick up some waste metal at several plants.

This importance of personal connections in China has two implications. First, if the size of social-family network indeed affects one's choice of self-employment, we should expect to see this effect in China more than most other societies. Second, and perhaps more importantly, this usefulness of personal connections in China implies that the self-employed will intentionally build and maintain a large network. Thus it is absolutely necessary to develop an identification strategy to solve the reverse-causation problem.

3. Data and Empirical Strategies

This study uses a unique survey database on Rural-Urban Migration in China and Indonesia (RUMiCI). The RUMiCI database has been constructed by a team of researchers from Australia, China, and Indonesia. They secured funding to conduct a five-year longitudinal survey in China and Indonesia, with the goal of studying issues such as the effect of rural-urban migration on income mobility and poverty alleviation, the state of education and health of children in migrant families, and the assimilation of migrant workers into the city.

The first wave of the survey was conducted in 2008 and the data became available in 2009. In China, three representative samples of households were surveyed, including a sample of 8,000 rural households, a sample of 5,000 rural-urban migrant households, and a sample of 5,000 urban households. In this paper, our empirical analyses use information mainly from the migrant sample.

The migrants surveyed are randomly chosen from fifteen cities that are the top rural-urban migration destinations in China (see Figure 1). Eight of these cities are in coastal regions (Shanghai, Nanjing, Wuxi, Hangzhou, Ningbo, Guangzhou, Shenzhen, and Dongguan); five of them are in central inland regions (Zhengzhou, Luoyang, Hefei, Bengbu, and Wuhan); and two of them are in the west (Chengdu and Chongqing). A sampling procedure is very carefully designed to ensure that migrants in the database constitute a representative sample of all the migrants in the fifteen cities.[9]

The migrant survey was designed to collect information about every household member. It asks detailed questions about the respondent's personal characteristics,

[9] See the RUMiCI Project's homepage (http://rumici.anu.edu.au/joomla/) for detailed documentation of the sampling method.

educational background, employment situation, health status, children's education, social and family relationship, major life events, income and expenditure, housing and living conditions, etc. The resultant database contains more than 700 variables. In terms of basic information of a household member, we know the person's age, gender, education level, current address, home address before migration, etc. Related to the person's employment experience, we know whether the person is self-employed or a wage worker, occupation, monthly income, how he/she found the current job, what was his/her first job, how he/she found the first job, etc. For the self-employed, we know why they chose self-employment, the amount and sources of money they borrowed for initial investment, the number of workers they currently hire, etc. Particularly useful for our study, the survey also asked about the migrant's social and family network. We know who the migrant's important social contacts are and whether they live in the same city, whether the migrant's parents and siblings also live in the same city, how many people the migrant greeted during the past Spring Festival, etc.

In our regression analysis, the dependent variable is whether an individual is self-employed or not today. Among all of the migrant household heads in the database, 19.6 percent are self-employed.[10] These individuals can be restaurant owners, convenient store owners, scrap metal collectors, street vendors who sell fruits, snacks, cigarettes, clothing, souvenirs, etc. or provide services such as shining shoes and repairing bicycles and electronics.[11] A large proportion of these self-employed migrants simply work alone; only a quarter of them (25.4 percent) also hire other people. Among those who hire other people, the average number of employees is 3.5.

Our key independent variable is the size of a person's social-family network. To measure this size, we use the number of friends one greeted during the past Spring Festival, the number of relatives one greeted during the past Spring Festival, or the sum of these two numbers.

[10] This proportion increases to 22.9 percent if we consider the whole sample, including both the heads and other working members of the households.

[11] In the version of the database used here, a migrant's occupation and industry are in the form of verbal descriptions that directly record the respondent's answer to the survey question and are not numerically coded. Whereas this allows us to see exactly what kind of work each migrant does, it is impossible to tabulate these occupations and industries in a straightforward way.

Spring Festival is the most important traditional holiday in China, which starts from the first day and ends on the fifteenth day of the first month, according to the Chinese lunar calendar. There are many traditional activities during the festival, which vary widely across different regions in the country. But one tradition is followed throughout the country. That is, during the festival, people will greet family members, relatives, and friends, wishing them a happy, healthy, and wealthy new year. We therefore use the self-reported number of friends and relatives an individual greeted during the festival to measure the size of this migrant's social-family network. It is worth noting that although traditionally greetings were mostly sent through personal visits, in recent years greetings by phone, post, or even email have also become common. Therefore, the persons greeted are not necessary local people.

This network size measure is a behaviorally revealed one that is more relevant for our purpose in this study. For example, a person may have a first cousin who is by definition one of his relatives. However, if they have had a bad relationship and have not been on speaking terms, or if they have lived far away from each other and have lost contact, then the cousin is in effect out of this person's network of relatives. It is important to count the cousin out for our purpose because it is unlikely the cousin will provide any help when this person needs assistance during self-employment. Our measure will achieve this because if a relative is effectively outside a person's network, this person will not have greeted him during the Spring Festival. Similarly, we believe that only a friend greeted is a friend indeed.

A network size measure like ours also has its drawbacks. For example, if a person has already chosen self-employment, he may have incentive to greet more friends and relatives simply because he has used or will likely seek their assistance during self-employment. For this reason, a simple correlation between self-employment status and network size cannot be interpreted as a causal effect of network size on the choice of self-employment. It may be a result of reverse causation. That is, self-employment may have caused one to develop and maintain a large social-family network, an effect that is also interesting in itself but not exactly what we intend to study here.

Another issue with the network size measure is the concern of measurement errors, a problem that is common to survey data. During the survey, a respondent has to recall

how many friends and relatives he greeted. Due to imperfect memory or lack of effort to do an accurate count, a respondent tends to report a number that appears to be a best guess. As we can see in Figure 2, most surveyed individuals reported salient numbers, numbers that are multiples of five or ten.[12] There is no reason to believe, for example, that a person is so much more likely to have actually greeted twenty than nineteen friends or relatives. Thus the spiky distributions in Figure 2 are almost surely a result of rounding or misreporting. As well known, measurement errors in the independent variable will bias the OLS coefficient toward zero. Therefore, even if a larger social-family network indeed increases the probability of self-employment, a simple OLS regression may fail to identify a statistically significant effect because of random errors in the measurement of network size.

The standard technique to overcome these reverse-causation and measurement-errors problems is to instrument for the independent variable, which is the approach we take here. That is, we will use an instrumental variable that is correlated with the network size measure but does not affect the choice of self-employment through any other unaccounted channels. The particular IV we will use is the distance from home province when a migrant first left his village to work in the urban sector.

More specifically, we construct a distance variable using information about a migrant's home address and the province he migrated to when he first left his village.[13] Since this first migration typically occurred a few years ago (with a median of six years ago) and the RUMiCI project focuses on the migrant's current situation, the survey did not ask about the exact destination of the first migration at the sub-provincial level. So we can only construct a distance variable at the province level. For each migrant, we calculate the log railway distance between the capital of the home province and the capital of the first destination province.[14] If the home province is the same as the first destination province, we set the log distance equal to zero.

---

[12] This tendency to report salient numbers seems to be a common issue in survey data rather than an idiosyncratic feature of our data here. For example, working with U.S. firm level data, Neumark et al. (2007) report a similar problem with a firm size variable measured by self-reported number of employees.

[13] We use the word "province" to refer to all provincial level jurisdictions in China, including 23 provinces, five autonomous regions, and four direct-control municipalities.

[14] Only one province, Hainan (which is on an island), is not connected with other provinces through railway. There are only two migrants from Hainan in the database, so we simply dropped those two observations.

We expect, and the data have confirmed, that the distance of the first migration is correlated with the number of friends and relatives greeted during the past Spring Festival. The reason is simple. For people who grew up in rural China, their social and family networks are highly local, because they usually interact with and marry with other people in the same or nearby villages. A person who migrated far away would have been disconnected with many individuals in his original network for a considerable period of time. This is true even if the migrant later moved back to a city closer to the home village. Because of this disruption, he tends to lose contact with some friends and relatives in his network. In the meantime, because he moved far away from home, he tended to know few locals and thus had difficulty in developing a new network.

Our key identifying assumption is that the distance of the first migration does not affect today's choice of self-employment through any other channels that are not controlled for in our regressions. We cannot test this assumption but believe it is plausible given the specific context of rural-urban migration in China and the particular samples of migrants used for estimation.

In recent years, as rural-urban migration has become an increasingly prominent social phenomenon in China, many field studies have been conducted to document the life experiences of these migrants.[15] We have therefore learned a great deal about the decision process during these migrations, from both anecdotal and statistical evidence. The key fact to keep in mind is that a typical villager in China had no chance to travel to many places and had very limited information about how the urban economy is organized in different cities. It is clear that the migration is usually triggered by a need or an urge to improve one's individual or family economic conditions. But where to migrate in the first time is mostly an accidental choice not based on an informed calculation of feasibility and potential returns in different locations.

A migrant chose a particular city in the first time almost always because he knew somebody who had already been there. It could be a relative, a neighbor, a friend, or simply an acquaintance who had already migrated to that city and demonstrated that it

[15] See, for example, Lü (2009), Wei (2008), and Xu and Qian (2009).

might be feasible for this person to do the same thing (Zhao, 1999, 2003).[16] Also, because the migration is not meant to be permanent, the first-timers tend to have a trial-and-error attitude: Let me give it a shot and see what happens. For this reason, when looking at a random sample of migrants, it seems reasonable to think of their first migration distance as random, especially after controlling for home province fixed effects. That is, given two first-time migrants from the same province, whether one went farther away than the other is likely to be exogenous, driven mostly by whether one knew somebody who had migrated far away. Note that we do not need this distance to be completely random; we only need that it is exogenous to the choice of self-employment today.

The most serious threat to the credibility of our identification strategy is that the first migration destination and the type of the first job in urban sectors (whether self-employed or not) may be jointly determined. If this is true, it is problematic to think of the distance of first migration as exogenous to a migrant's self-employment decision, especially for those who are still on their first jobs in cities today. To overcome this problem, in our empirical analysis below we will focus on the sample of migrants who did not start as self-employed and who are not on their first jobs today. In other words, we will examine the sample of migrants who all moved to urban areas to work for some employers and all changed their jobs over time. Some of them would change from wage workers to self-employment and others would remain as wage workers but have moved to different employers. We then ask the following empirical question: Among all these rural-urban migrants who started as wage workers and later changed their jobs, who are more likely to have chosen self-employment today? Because all the migrants in this sample started as wage workers in urban sectors, it is much more plausible to assume that their first migration destinations were not chosen for the purpose of self-employment. It is thus reasonable to exclude the distance of the first migration from the main equation that explains a migrant's self-employment status today.

Another threat to the credibility of our identification strategy is the possibility that the distance of first migration is correlated with some unobserved characteristics of the

[16] Our survey asked the first-time migrants the question "who provided you the information for job hunting in the urban sector." Relatives (52.7 percent) and other migrants from the same village (28.4 percent) overwhelmingly top the list of answers.

migrant that in turn are correlated with the migrant's choice of self-employment. In that case, the distance is not a valid instrumental variable. A most plausible scenario is perhaps that the more adventurous individuals are more likely to migrate far away from home and those people are also more willing to take risks and therefore more likely to choose self-employment. As it turns out, we find that individuals who migrated far away in the first time tend to have a smaller social-family network today and are less likely to be self-employed today. Therefore, this concern about unobserved attitude toward risks actually works against our findings. In particular, if it is indeed true that the less risk-averse individuals tend to migrate a longer distance and are more likely to choose self-employment, then the true effect of network size will be even higher than what we find. That is, our IV estimate can be thought of as a lower bound of the true effect.[17]

Our main estimating equation is as follows:

$$y_{ji} = \alpha + \beta s_{ji} + X_{ji}\gamma + HP_j + \varepsilon_{ji} \quad (1)$$

where the outcome variable y_{ji} is a dummy variable taking value 1 if migrant *i* from province *j* is self-employed; s_{ji} is the key independent variable that measures the size of social-family network for this individual; X_{ji} is a vector of control variables including the migrant's age, gender, years of schooling, and marital status;[18] HP_j is a home-province fixed effect that captures the effect of all unobserved factors common to migrants from province *j*; and ε_{ji} is the error term.

When using the IV strategy, we estimate two-stage least squares (2SLS) regressions with the following first-stage equation:

$$s_{ji} = \kappa + \phi d_{ji} + X_{ji}\lambda + HP_j + \mu_{ji} \quad (2)$$

[17] In general, if we estimate the slope of the equation $y = \alpha + \beta x + \varepsilon$ using a variable *z* to instrument for *x*, then $\beta_{IV} = \beta + Cov(z, \varepsilon)/Cov(z, x)$. In our case, if indeed people migrating faraway are more willing to take risks, then we have $Cov(z, \varepsilon) > 0$. We know that migrating faraway is negatively correlated with network size, i.e., $Cov(z, x) < 0$, so $\beta_{IV} < \beta$. That is, our IV coefficient underestimates the true effect.

[18] Given the literature on the entrepreneur's liquidity constraint, it seems necessary to control for income or wealth in the regression. However, we only observe a household's current income, which is a result of the choice between self-employment and wage work instead of its causes. We have included years of schooling, which should have picked up some of the income or wealth effects.

where d_{ji} is the log-distance between the home and destination provinces when individual *i* from province *j* first migrated to a city. Predicted s_{ji} from this first-stage regression are then used for estimating equation (1) in the second stage.

Note that our dependent variable is a dichotomous variable. In ordinary situations, it is natural to use a logit or probit specification. However, when one needs to instrument for an endogenous independent variable, a linear probability model is a preferred setup (Angrist and Krueger, 2001). Thus we will focus on this linear model in our empirical analysis. Although not presented here, we have also run parallel regressions with an IV probit specification; the results are qualitatively similar.

4. Empirical Results

We present our empirical results in this section.

4.1 Descriptive statistics

The survey of rural-urban migrants was conducted at the household level. Some migrants are married; their spouses, and sometimes their grown-up children, may stay in the same household and also work in the city. In our empirical analysis, we focus on the household heads only. We also exclude the household heads aged below 16 or above 70. And finally, we drop any observations with a missing dependent, independent, control, or instrumental variable. This procedure leaves us with 4,505 observations, for which the descriptive statistics are shown in the left four columns of Table 1.

Twenty percent of the household heads are self-employed; 69 percent are male; and 54 percent are married. Their average age is 30.4 and average years of schooling is 9.3. When they first migrated out of rural areas, 48 percent went to a city in the same province and 52 percent migrated to a different province. The average log distance between the home and destination provinces during the first migration is 3.153, which translates to 23.4 kilometers.[19] The distribution of this log distance is highly skewed, with a maximum of 8.313 (equal to 4,077 kilometers).

On average, a household head greeted 34 people during the past Spring Festival, 18 of them are identified as friends and 13 of them relatives. So relatives and friends sum up to 32. They also greeted a couple of other people who are neither friends nor relatives.

[19] This average is so small partly because we have forced the log distance of all within-province migrations to be zero.

These are most likely neighbors or coworkers a migrant regularly bumps into, feels compelled to say Happy New Year out of politeness, but does not consider as friends. We will refer to these social contacts as acquaintances. We think that acquaintances have only weak ties to the migrant. It is unlikely that they will provide substantial assistance to the migrant when needed. We therefore do not expect them to affect the migrant's choice of self-employment.

Notice that in the fourth column of Table 1 these network size measures have very large maximum values. For example, the maximum number of people greeted is 9,999.[20] The maximum number of friends and relatives greeted is 1,996, with half of them being friends and the other half relatives. It is hard to believe that these extreme values contain any real information; it seems they are either carelessly made-up numbers coming from the interviewees or recording errors made by the data collectors.

Given the small average network size, we are concerned that these extreme values could seriously bias our estimation. We examine the distribution of the total number of people greeted in more details and find its 99th percentile to be 200. We therefore decide to use this as a cutoff point and drop all observations with this total number higher than 200. Since some outliers have missing values of other variables and they will be dropped from our analysis anyway, this rule of deleting outliers only excluded 32 other observations, reducing our sample size from 4,505 to 4,473. Our regression analysis below starts with this sample of 4,473 observations.

In the last two columns of Table 1, we also present the descriptive statistics of the sample excluding outliers. Naturally, all the network size measures now have smaller means and standard deviations. Notice that the means and standard deviations of all other variables are virtually unchanged. We conduct t-tests to compare the means between the whole sample and the sample excluding outliers, and find that the difference in the mean is never statistically significant for any non-network-size variable. This suggests that the outliers are essentially a random subset of the whole sample, and therefore dropping them will unlikely introduce serious sample selection biases.

[20] This is not a code for missing values. Only one migrant reported a total of 9,999 people greeted; the next largest total reported is 3,000.

To be cautious, we have also run parallel regressions using the whole sample, including all the outliers. The results are qualitatively similar, although the estimation of the network size coefficient is generally less precise with a slightly lower t-value.

4.2 Regression results

We now present regression results. We use the number of friends, the number of relatives, and the number of friends and relatives as alternative measures of the size of an individual's social-family network. We run OLS and 2SLS regressions to estimate equation (1). For each set of regressions, we try four different samples defined as follows:

- Sample A — all household heads aged between 16 and 70 years, excluding outliers whose total number of contacts is above the 99th percentile.
- Sample B — all household heads in sample A who are not on their first jobs in urban sectors.
- Sample C — all household heads in sample A whose first jobs in urban sectors were not self-employment and who are not on their first jobs in urban sectors.
- Sample D — all household heads in sample A whose first jobs in urban sectors were not self-employment and who are not on their first jobs in urban sectors, excluding those who chose to become self-employed at some point only because they could not find any wage-earning jobs.

Going from sample A to B imposes a most stringent restriction, reducing the number of observations by 44 percent, from 4,473 to 2,489. This implies that many of the migrants are still on their first jobs after they migrated out of rural areas. Among those who have left their first jobs, some have stayed in the same city, yet others moved to different cities or even different provinces. For example, there are 434 household heads who originally moved out of their home provinces but currently work in cities within the home provinces. There are also 225 household heads who originally migrated to cities within their home provinces but currently work in places outside the home provinces. Presumably these individuals moved to different provinces because over time they found better job opportunities in other provinces.

Among the 2,489 household heads who have changed jobs, there are also moves into and out of self-employment. There are 416 individuals who started as wage workers in urban areas but have now become self-employed. There are 41 household heads who

were initially self-employed and have now become wage workers. There are also 55 household heads who started and have remained as self-employed.

In sample C, we further exclude all the 96 household heads who started with self-employment when they first moved to cities. As discussed above, this sample restriction makes it more reasonable to think that the distance in the first migration does not directly affect the choice of self-employment today. To be precise, regressions using Sample C answer the following question: Among all the individuals who originally migrated to cities only to take wage-earning jobs, who are more likely to have moved into self-employment today?

For descriptive purposes, we divide sample C into two groups. One group initially migrated to cities within the home province and the other initially to cities outside the home province. The first group, migrants who originally stayed within the home province, have an average of 34.05 friends and relatives today; the other group, those who originally moved out of the home province, have an average of 28.57 friends and relatives today. We show in Figure 3 that in the first group, 826 migrants changed jobs over time but remained as wage workers, and 215 migrants (or 20.65 percent of this group) moved from the wage-worker status to self-employment. In contrast, in the second group that initially moved outside of the home province, 1,151 migrants changed jobs but remained as wage workers and only 201 (or 14.87 percent of this group) switched from wage work to self-employment. That is, those who originally moved far away from home—and therefore have smaller social-family networks today—are less likely to become self-employed. These differences between the two groups are the key variations in the data that help us identify the effect of network size on the choice of self-employment.

Sample D adds one more restriction on sample C by dropping the household heads who moved from wage work to self-employment at some point because they could not find wage-earning jobs. The idea is that some migrants may stay as self-employed for the time being primarily to avoid unemployment. These people did not mean to be entrepreneurs and their choices may not be determined by the same factors as other self-

employed individuals.[21] We consider results from samples C and D more convincing, although we also present results from samples A and B for comparison purposes.

We first look at how the number of friends affects the choice of self-employment and the results are in Table 2. Columns (1)-(4) show the results from OLS regressions, using samples A-D; columns (5)-(8) are corresponding results from 2SLS regressions. In every regression, we include the same set of control variables, a constant, and home province fixed effects.

The number of friends has small positive coefficients in OLS regressions, some of which are statistically significant and others not. Consider column (4), which uses sample D and gives the largest coefficient among all OLS regressions. It suggests that one more friend is associated with a 0.045-percentage-point increase in the probability of being self-employed. It takes 22 friends—close to one standard deviation, which is 26.8 for sample D—to increase the probability of self-employment by one percentage point. Therefore, even if one believes this effect is true, its magnitude is too small to be of much economic significance.

In columns (5)-(8), the number of friends still has positive coefficients in 2SLS regressions. But in contrast, these IV coefficients are all substantially higher and all statistically significant. Consider results in column (8), again estimated using sample D. The coefficient suggests that one more friend leads to a 1.15-percentage-point increase in the probability of self-employment, 25 times higher than the corresponding OLS coefficient. Although the magnitude of the IV coefficient changes across different samples, they are more or less of the same order.

Our discussion above suggested that the OLS coefficient of network size may be biased for two reasons. One is reverse causation, which biases the coefficient upward; the other is measurement errors in the key independent variable, which biases the coefficient

[21] Among all of the self-employed migrants (in our sample A), only a small fraction (12 percent) ended up being self-employed because they cannot find wage work. Most of them choose self-employment because it brings a higher income (38 percent), it gives more flexibility and freedom (29 percent), or it allows one to be one's own boss (19 percent).

toward zero. Given that our IV estimates are so much larger, it seems that biases from measurement errors are dominant in the OLS regressions.[22]

The coefficients of control variables show rather consistent patterns across different samples. Age and being male generally have small coefficients and are statistically insignificant except in one case. The coefficients of schooling and marital status, in contrast, are always statistically significant. The IV results suggest that one more year of schooling decreases the probability of self-employment by about three percentage points, which is a sizeable effect. One possible explanation is that employers prefer to hire the more educated and consequently such individuals have better alternatives on the job market. It is also possible that better educated people have become more risk averse and do not want to face the higher uncertainty associated with self-employment. Yet another possible reason is that self-employment opportunities are highly concentrated in low-status services and the more educated may want to stay away from those occupations either because they have higher aspirations or because of social pressure.[23] Married migrants are 20-percentage-point more likely to be self-employed, perhaps because the married couple have complementary skills that make self-employment more feasible, or because a spouse provides an extra source of income and serves as a sort of insurance for the self-employed household head.

Home province fixed effects are included in all regressions. This is important because heterogeneity across home provinces may affect both the dependent variable and the endogenous independent variable. On the one hand, migrants may have different numbers of friends simply because social customs and population densities differ across home provinces. For this reason, home province fixed effects should be included in the first-stage regression. On the other hand, self-employment rate can also vary across migrants from different home provinces due to unobserved factors. For example, Sichuan

[22] In the context of studying the rate of returns to education, Ashenfelter and Krueger (1994) also find that measurement errors in self-reported schooling cause serious biases in their estimates but omitted ability variables do not.

[23] There has been a long history of occupational stratification in the Chinese society. Different social statuses are attached to different occupations and one of the motivations to obtain more education is to enter a higher-status and well-respected occupation. It is a general expectation that the more educated should not enter a low-status occupation. In 2003, a young man who graduated from the prestigious Beijing University was found to work as a self-employed butcher and make a living by selling pork. That became big news in China and generated a lot of discussion in the media and on the Internet.

cuisine is very popular in many urban areas in China and as a result migrants from Sichuan may disproportionally concentrate in the food services industry and work as self-employed restaurant owners. Therefore, we should also control for home province fixed effects in the second-stage regression.

For all 2SLS regressions, we present the first stage F statistics to show the correlation between our instrumental variable and the endogenous independent variable. The statistic is generally large enough to alleviate serious concerns over potentially weak instruments.

Table 3 presents results from a similar set of regressions, only that now we use the number of relatives as the independent variable. The results are qualitatively similar. In OLS regressions, the coefficient of number of relatives is never statistically significant and always very small. In contrast, the coefficient in 2SLS regressions is always statistically significant and always much larger. For example, the IV coefficient from sample D is 40 times as large as the OLS coefficient estimated from the same sample. It suggests that one more relative increases the probability of self-employment by three percentage points.[24]

A casual comparison of the results in Tables 2 and 3 suggests that the effect of an extra relative is larger than that of an extra friend. This makes sense. In China, it is generally believed that "blood is thicker than water," meaning that kinship is more important than friendship. Therefore, it is hardly surprising to find that an extra relative has more influence than an extra friend.

We should point out that the specifications in Tables 2 and 3, including either the number of friends or the number of relatives as an independent variable but not both, are not ideal. As expected, these two variables are positively correlated, with a correlation coefficient of 0.55. Therefore, if both variables have positive effects on the choice of self-employment, then including only one of them in the regression will overestimate the coefficient because it will capture part of the positive effect of the other variable.

Ideally, we want to include both as independent variables in our regression and separately identify the effect of each variable. However, we have only one plausible IV,

[24] We must note here that because we are estimating a linear probability model and because the IV coefficient is better understood as a local average treatment effect, it is inappropriate to extrapolate this estimate too far away from the sample mean.

which does not allow us to deal with two endogenous independent variables simultaneously. As a compromise, we use the sum of these two variables as an alternative measure of network size. This imposes the assumption that the effect of an extra friend is the same as the effect of an extra relative, which may not be true given our discussion above. Nonetheless, this seems to be the most reasonable way to construct a network size measure that incorporates the effects of both friends and relatives. The regression results using this measure are shown in Table 4.

We see again that in OLS regressions the number of friends and relatives always has a very small and positive coefficient. It is statistically significant in three out of four regressions. The IV coefficients are again all statistically significant and substantially larger than the OLS coefficients. Also, although the IV coefficients vary across different samples, they are more or less of the same order. Using sample D, the IV coefficient is 22 times as large as the OLS coefficient, again implying that biases from measurement errors dominate endogeneity biases in the OLS regressions. The IV results suggest that an extra friend or relative increases the probability of self-employment by 0.8 percentage points. This is indeed smaller than the effects found in either Table 2 or 3, confirming the suspicion that using the number of friends or relatives only in the regression will overestimate the effect.

We pause here to give a quick summary of the results. Using different network size measures and different data samples, we find consistent evidence that more friends or relatives lead to a higher probability of self-employment. Using our preferred samples C and D, estimates from our preferred specification (in Table 4) show that an extra friend or relative increases the probability of self-employment by 0.8-1 percentage point. We also find that naïve OLS regressions greatly underestimate the effect, most likely because of measurement errors in the explanatory variables.[25]

4.3 Additional and sensitivity analysis

4.3.1 Choice vs. duration of self-employment

Up to this point, we have always been speaking about the *choice* of self-employment and how it is affected by the size of a person's social-family network.

[25] Yueh's (2009) estimates are more or less of the same order as our OLS estimates here, suggesting that there may be serious measurement-errors biases in her estimates.

However, the independent variable in our regression analyses does not exactly measure entry into self-employment. Rather, it indicates the state of being self-employed at a particular point in time, which is determined not only by the choice but also the duration of self-employment. Consider the following extreme case. At any point in time, suppose that all individuals enter self-employment with exactly the same probability. Further assume that self-employed individuals with a larger social-family network will be more successful and thus stay longer in self-employment, and others will soon move out of self-employment. This scenario would be equally consistent with our empirical findings, but it is really about the effect of network size on the duration rather than the choice of self-employment.

This duration-effect interpretation implies that conditional on being self-employed at present, those with a larger social-family network should have been in the self-employment status longer. Our survey indeed asked when each person started the current job. So we can check whether a larger network leads to a longer self-employment spell. Focusing on the self-employed individuals in samples C and D, we rerun all the 2SLS regressions, using the duration of self-employment as the dependent variable and keeping all the right-hand side variables exactly the same. We find that the size of social-family network never significantly affects the duration of self-employment, no matter which network-size measure is used. Not only is the coefficient of network size not statistically significant, but also that it is always negative. That is, it has a "wrong" sign that contradicts the duration-effect interpretation. Therefore, our findings are indeed about the choice of self-employment and we will continue to interpret the results this way.

4.3.2 Effects of other contacts

As mentioned in the data section, there are often some people a migrant greeted during the Spring Festival but did not identify as friends or relatives. We call them acquaintances and believe that they are unlikely to have an effect on the migrant's choice of self-employment. To verify this, we run exactly the same set of regressions as in Tables 2-4, but instead use the number of acquaintances as the explanatory variable. The results are in Table 5. This change of the explanatory variable makes rather striking differences. Now the coefficient of the number of acquaintances is never statistically significant, whether in the OLS or 2SLS regressions. The first stage F statistic in the

2SLS regressions is always very small, implying that one would have some acquaintances no matter where one lives, regardless of the distance of the first migration. This exercise confirms our expectation that acquaintances are not as helpful as friends and relatives and therefore do not affect one's self-employment decision. More importantly, these results suggest that the statistically significant IV coefficients shown in Tables 2-4 represent real effects rather than some artifacts in this particular database.

4.3.3 Excluding "roaming" migrants

We further explore the data to gain a deeper understanding of the mechanisms behind our empirical findings. As indicated earlier, there are a large number of migrants who originally moved out of (stayed in) their home provinces but are currently working inside (outside) their home provinces. Maybe it is these migrants, who moved back and forth, that really suffered a loss in terms of their family and social connections. Perhaps moving far away just one time does not cause so much disruption to a migrant's social-family network; after all, one can start developing a new network after settling down in the destination city. We therefore examine whether our main results are driven by the "roaming" migrants who did not settle down after their first moves.

Starting with sample D, we construct two samples by excluding those who moved across provinces after they first arrived in a city. They are as follows:

- Sample E — all household heads in sample D who first migrated to cities within (or outside) the home province and are still in cities within (or outside) the home province.
- Sample F — all household heads in sample D who are still in the same province as the original destination province.

Clearly sample F is a subset of sample E.

Using these two samples, we run 2SLS regressions with exactly the same specifications as reported in Tables 2-4, and the results are in Table 6. Column pairs 1-2, 3-4, and 5-6 show results using number of friends, number of relatives, and number of friends and relatives as independent variables, respectively. Comparing these results with those in Tables 2-4, we find that the coefficients of the network size variables are slightly smaller, but of the same order. They are more precisely estimated and their statistical significance tends to be higher. The first-stage F statistics are all much higher, suggesting

that excluding the roaming migrants increases the correlation between first-migration distance and network size. Although not presented in Table 6, in each case we have also run a companion OLS regression and the results are qualitatively identical to those in Tables 2-4. That is, OLS coefficients are positive but much smaller than the IV coefficients. Overall, this analysis suggests that our main results are not driven by those roaming migrants.

4.3.4 Controlling for city characteristics

Up to this point, we have controlled for individual characteristics and home province fixed effects but not any destination-city characteristics. We experimented with the idea of adding city fixed effects. However, because our distance variable is crudely measured at the province level and because there are 27 home province dummies and 15 city dummies, these fixed effects tend to explain away most of the variations in our instrumental variable, rendering our IV strategy ineffective.

As we examine the data at a more detailed level, we find that migrants in cities such as Guangzhou, Dongguan, Shenzhen, Ningbo tend to come from faraway rural areas and have low self-employment rates; and migrants in cities such as Hefei, Bengbu, Zhengzhou, and Luoyang tend to come from nearby rural areas and have high self-employment rates. These between-city variations seem to be crucial for identifying the effects of network size. Including city fixed effects will simply dump all these variations and this is why such specifications do not produce any precise estimates.

However, this source of identification (i.e., between-city variations) indeed causes concerns. We find that cities with low self-employment rates among migrants are mostly in coastal areas and cities with high self-employment rates are mostly in inland areas. Simply due to geographic constraints, coastal cities would tend to draw migrants from further region than inland cities. At the same time, coastal cities also tend to have many manufacturing plants in exporting industries that hire a large number of migrant workers. So we wonder whether our empirical analysis has simply picked up this effect that reflects structural differences between coastal and inland cities but has little to do with social-family networks.

To address this concern, we have developed two strategies. First, we construct three structural-characteristics variables for each city and use them as control variables.

The first one is the share of the labor force working in the private sector. From the *Urban Statistical Yearbook of China*,[26] we know the number of employees in government and state- or collectively-owned enterprises, the number of workers in self-employment or in privately-owned businesses, and the number of registered unemployed workers in each city. We use these three numbers to calculate the share of the labor force in the private sector. If this share is higher in a city, we expect its residents to choose self-employment with a higher probability. The other two variables are the share of tertiary-sector employment in total city employment and the share of tertiary-sector GDP in total city GDP, both directly available from the yearbook. Because the tertiary sector includes service industries and the self-employed tend to concentrate in those industries, we expect a higher self-employment rate if these two shares are higher.

We use sample D for this analysis and focus on the number of friends and relatives as the network size measure. We run 2SLS regressions using the same first-stage specification as in Table 4, i.e., controlling for individual characteristics and home province fixed effects. The only difference is with the second stage, in which we now add the city level controls. Since the three city characteristics are correlated, we add them into the second-stage regression in every possible combination. The results are in columns (1)-(7) of Table 7.

We want to compare the coefficient of friend and relatives with that in the last column of Table 4, which is 0.0083. The coefficients in columns (1)-(7) of Table 7 are all smaller, but only slightly, ranging from 0.0061 to 0.0081. They are still statistically significant. These results suggest that one more friend or relative will increase the probability of self-employment by 0.6-0.8 percentage point, still a sizable effect. Each of the three city characteristics is statistically significant in at least one specification, and the coefficients of tertiary-sector employment share and tertiary-sector GDP share always have the expected sign.

Second, we directly control for self-employment rate in each city in the second-stage regression. The idea is that if some city characteristics make it easier or more desirable for any city resident to be self-employed, then those characteristics should be reflected in the overall self-employment rate among city residents. So directly including

[26] We use the 2008 edition of the yearbook, which publishes the 2007 data on all cities in China.

the self-employment rate in the second-stage regression is largely equivalent to controlling for all relevant city characteristics.

As mentioned in the data section, the RUMiCI research project team also interviewed 5,000 randomly selected urban households in eighteen cities, including all of the fifteen cities covered by the rural-urban migrant surveys. They asked city residents questions similar to those in the migrant surveys, including whether they are working and whether they are self-employed. Using these data, we calculate the self-employment rate for each city, and add it to the second-stage regression as a control variable. We calculate the self-employment rate for each city in two ways, one is the share of self-employed workers in all the working household heads aged between 16 and 70, and the other is the share of self-employed workers in all the household heads (whether working or not) aged between 16 and 70.

Results from the regressions using each of these extra control variables are in Columns (8)-(9) of Table 7. The coefficient of the relatives and friends variable is 0.0078 and 0.0073, again very similar to the baseline estimate. It turns out that city self-employment rate has a negative coefficient. In cities where more local residents are self-employed, rural-urban migrants are less likely to be self-employed. That is, self-employed migrants and self-employed city residents are substitutes. This is actually consistent with the notion that rural-urban migrants tend to take the jobs shunned by local residents.

Overall, these results in Table 7 show that our findings are robust.

4.3.5 Alternative IV for number of relatives

Finally, we experiment with an alternative IV for the number of relatives.[27] It is natural that people in large families tend to have more relatives. The RUMiCI survey asked each household head many questions about his or her parents. One particular question is how many children in total a parent had. We use the answer to this question to construct a "number of siblings" variable. If this variable is missing for a household head's father, we use the mother's number of children when it is available.[28] When a household head is married, the same question was also asked about his or her parents-in-

[27] We thank Wayne Gray for suggesting this idea.
[28] Given the low divorce rate among the older generations in China, a mother's number of children is almost always identical to the father's.

law. However, whereas the parents' number of children is usually not an individual's own choice, the number of children in the spouse's family is the individual's choice (by association) and thus is less exogenous. So we do not count the children of the parents-in-law and will simply control for marital status as before.

We use this alternative IV in two ways: (1) use the number of siblings alone to instrument for the number of relatives; and (2) use the number of siblings, its interaction with age, and its interaction with age squared to instrument for the number of relatives. The first specification is straightforward; we expect an individual to have more relatives if he or she grew up in a larger family. The second specification further takes into account the possibility that the number of additional relatives associated with each extra sibling varies nonlinearly with age.[29]

Results from these exercises are in Table 8. Here our alternative IVs should work regardless whether or not a migrant started as wage worker and whether or not he is still on the first job. Therefore, we should trust the estimates from the whole sample. Nonetheless, for comparison purposes, we have again estimated the model using samples A-D.[30] The left four columns correspond to the first specification and the right four columns the second specification. In all these IV regressions, the coefficient of the number of relatives cannot be precisely estimated. It is not statistically significant except in one case. However, the size of the coefficient is of similar order to the estimates obtained in our baseline regressions. It implies that one extra relative increases the probability of self-employment by 0.61-1.97 percentage points. Although not presented in Table 8, we have also estimated the corresponding OLS coefficients. Similar to the baseline results, the OLS coefficients (ranging between 0.0006 and 0.0013) are still much smaller than the IV coefficients. Given that these regressions use completely different IVs, we find it rather reassuring that these results, although not statistically significant, at least point to the same direction as the baseline estimates.

[29] When a person is young, one more sibling does not necessarily imply many more relatives because the sibling is probably not married yet. For a middle-aged person, one more sibling implies a much larger number of relatives because the sibling most likely has married. When a person is old, some siblings may be dead and so the effect may be smaller again.

[30] These samples are constructed in exactly the same way as the samples used in Tables 2-4. However, because here we are using different IVs and dropping all observations with any missing IVs, we end up with four samples that are slightly different from those used in Tables 2-4.

We run some parallel regressions using the number of friends or friends and relatives as the independent variable. However, the siblings variable is only very weakly correlated with these independent variables, failing the requirement of a good IV. This of course is not surprising given that people make friends in many different contexts and siblings do not necessarily bring more friends.[31]

5. Conclusion

A large body of existing literature suggests that the self-employed rely heavily on family members, relatives, and friends for informational, financial, and operational assistance. Their success often hinges on access to a well developed social-family network. We therefore hypothesize that individuals connected with a larger social-family network are in a better position to choose self-employment. We test this hypothesis using a newly constructed database on rural-urban migrants in China. The migrants in the database reported the number of people they greeted during the past Spring Festival; they also identified how many of these contacts are their friends and how many are relatives. We use this information to measure the size of a migrant's social-family network.

We recognize two potential problems with using naïve multivariate regressions to identify the effect of network size on the choice of self-employment. One is the reverse causation that leads to upward biases in OLS estimates; the other is measurement errors (particularly in our data here and perhaps also in other survey data sources) in the network size variable that cause downward biases in OLS estimates.

We take the standard IV approach to overcome these problems. In particular, we use the distance from home province when a migrant first moved to the urban area to

[31] We have also explored another idea about a possible alternative IV. The IV used in our baseline regressions is based on the fact that individuals originally migrating farther from home tend to end up with a smaller social-family network today. This migration distance is obviously related to where one's home province is and where job opportunities turn up in urban sectors. Ideally, we would want to construct an expected migration distance for each individual based on where he lived and the overall migration patterns—similar to the idea employed by Card (2001)—and use this expected distance to instrument for network size. For example, if a farmer who grew up in Jiangsu province first moved out of his village in 2000, and if in that year one third of the migrants from Jiangsu went to Guangdong and two thirds went to Shanghai, then we can calculate the expected migration distance based on this aggregate migration pattern. If this kind of information is available for multiple years, migrants from the same province will have different expected migration distances depending when they first moved out, so it still allows us to control for province fixed effects. The advantage of using this expected distance is that it is primarily determined by exogenous factors (e.g., where one was born and which region in the country was booming in a particular time period). Unfortunately, we could not find any reliable data on detailed rural-urban migration patterns in China, even at the province level, and thus could not implement this idea.

instrument for network size today. We believe the exclusion condition is likely to be satisfied in the particular institutional context of rural-urban migration in China and especially for the sample of migrants who first started with wage-earning jobs in urban sectors and have moved on to different jobs over time. We find that the migrants who initially moved further away—and therefore have fewer friends and relatives today—are less likely to shift from wage work to self-employment. We consider this result rather convincing evidence that the size of social-family network affects one's self-employment decision. Our IV estimates are substantially larger than OLS estimates, suggesting that measurement-error biases dominate endogeneity biases in OLS regressions.

References

Aldrich, Howard E. and Jennifer E. Cliff (2003). "The Pervasive Effects of Family on Entrepreneurship: Toward a Family Embeddedness Perspective," *Journal of Business Venturing* 18, 573-596.

Aldrich, Howard E. and Cathrine Zimmer (1986). "Entrepreneurship through Social Networks," in D. Sexton & R. Smiler (eds.), *The Art and Science of Entrepreneurship*, New York: Ballinger, 3-23.

Allen, W. David (2000). "Social Networks and Self-Employment," *Journal of Socio-Economics* 29, 487-501.

Angrist, Joshua D. and Alan B. Krueger (2001). "Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments," *Journal of Economic Perspectives* 15(4), 69-85.

Ashenfelter, Orley and Alan B. Krueger (1994). "Estimates of the Economic Returns to Schooling from a New Sample of Twins," *American Economic Review* 84, 1157-1173.

Bayer, Patrick, Stephen L. Ross, Giorgio Topa (2008). "Place of Work and Place of Residence: Informal Hiring Networks and Labor Market Outcomes," *Journal of Political Economy* 116, 1150-1196.

Beaman, Lori A. (2009). "Social Networks and the Dynamics of Labor Market Outcomes: Evidence from Refugees Resettled in the U.S." mimeo, Department of Economics, Northwestern University.

Bian, Yanjie (1994). "Guanxi and the Allocation of Urban Jobs in China," *China Quarterly* 140, 971-999.

Birley, Sue (1985). "The Role of Networks in the Entrepreneurial Process," *Journal of Business Venturing* 1, 107-117.

Blanchflower, David and Andrew Oswald (1998). "What Makes an Entrepreneur? Evidence on Inheritance and Capital Constraints," *Journal of Labor Economics* 16, 26–60.

Brüderl, Josef and Peter Preisendörfer (1997). "Network Support and the Success of Newly Founded Business," *Small Business Economics* 10, 213-225.

Burt, Ronald S. (1997). "The Contingent Value of Social Capital," *Administrative Science Quarterly* 42, 339–365.

Card, David (2001). "Immigrant Inflows, Native Outflows, and the Local Labor Market Impacts of Higher Immigration," *Journal of Labor Economics* 19, 22-64.

Chan, Kam Wing and Li Zhang (1999). "The Hukou System and Rural-Urban Migration in China: Processes and Changes," *China Quarterly* 160, 818-855.

Davidsson, Per and Benson Honig (2003). "The Role of Social and Human Capital among Nascent Entrepreneurs," *Journal of Business Venturing* 18, 301-331.

Djankov, Simeon, Yingyi Qian, Gérard Roland, and Ekaterina Zhuravskaya (2006). "Who Are China's Entrepreneurs?" *American Economic Review Papers and Proceedings* 96(2), 348-352.

Dunn, Thomas and Douglas Holtz-Eakin (2000). "Financial Capital, Human Capital, and the Transition to Self-Employment: Evidence from Intergenerational Links," *Journal of Labor Economics*, 2000, vol. 18, 282-305.

Evans, David S. and Boyan Jovanovic (1989). "An Estimated Model of Entrepreneurial Choice under Liquidity Constraints," *Journal of Political Economy* 97, 808–827.

Evans, David S. and Linda Leighton (1989). "Some Empirical Aspects of Entrepreneurship," *American Economic Review* 79, 519–535.

Fairlie, Robert W. (1999). "The Absence of the African-American Owned Business: An Analysis of the Dynamics of Self-Employment." *Journal of Labor Economics* 17, 80–108.

Giannetti, Mariassunta and Andrei Simonov (2009). "Social Interactions and Entrepreneurial Activity," *Journal of Economics & Management Strategy* 18, 665–709.

Granovetter, Mark (1973). "The Strength of Weak Ties," *American Journal of Sociology* 78, 1360-1380.

Greve, Arent and Janet W. Salaff (2003). "Social Networks and Entrepreneurship," *Entrepreneurship: Theory and Practice* 28, 1-22.

Hoang, Ha and Bostjan Antoncic (2003). "Network-Based Research in Entrepreneurship: A Critical Review ," *Journal of Business Venturing* 18, 165-187.

Holtz-Eakin, Douglas, David Joulfaian, and Harvey S. Rosen (1994a). "Sticking It Out: Entrepreneurial Survival and Liquidity Constraints," *Journal of Political Economy* 102, 53–75.

Holtz-Eakin, Douglas, David Joulfaian, and Harvey S. Rosen (1994b). "Entrepreneurial Decisions and Liquidity Constraints," *Rand Journal of Economics* 23, 334–347.

Hout, Michael and Harvey S. Rosen (2000). "Self-Employment, Family Background, and Race," *Journal of Human Resources* 35, 671–694.

Hurst, Erik andb Annamaria Lusardi (2004). "Liquidity Constraints, Household Wealth, and Entrepreneurship," *Journal of Political Economy* 112, 319–347.

Ioannides, Yannis M. and Linda Datcher Loury (2004). "Job Information Networks, Neighborhood Effects, and Inequality," *Journal of Economic Literature* 42, 1056-1093.

Jackson, Matthew O. (2008). *Social and Economic Networks*, Princeton, NJ: Princeton University Press.

Laschever, Ron (2009). "The Doughboys Network: Social Interactions and the Employment of World War I Veterans," mimeo, Department of Economics, University of Illinois at Urbana-Champaign.

Lazear, Edward P. (2004). " Balanced Skills and Entrepreneurship," *American Economic Review Papers and Proceedings* 94(2), 208-211.

Lazear, Edward P. (2005). "Entrepreneurship," *Journal of Labor Economics* 23, 649-80.

Lentz, Bernard F. and David N. Laband (1990). "Entrepreneurial Success and Occupational Inheritance among Proprietors," *Canadian Journal of Economics* 23, 563–579.

Lin, Justin Yifu (1992). "Rural Reforms and Agricultural Growth in China," *American Economic Review* 82, 34-51.

Luke, Nancy and Kaivan Munshi (2006). "New Roles for Marriage in Urban Africa: Kinship Networks and the Labor Market in Kenya," *Review of Economics and Statistics* 88. 264-282.

Luo, Yadong (2007). *Guanxi and Business*, 2nd edition, Singapore: World Scientific Publishing Co.

Lü, Guoguang (2009), ed. *Oral History of Farmers Turned Workers*, Wuhan: Hubei People's Press.

Meng, Xin, and Junsen Zhang, (2001). "The Two-Tier Labor Market in Urban China: Occupational Segregation and Wage Differentials between Urban Residents and Rural Migrants in Shanghai," *Journal of Comparative Economics* 29, 485-504.

Montgomery, James (1991). "Social Networks and Labor-Market Outcomes: Toward an Economic Analysis," *American Economic Review* 81, 1408-1418.

Munshi, Kaivan (2003). "Networks in the Modern Economy: Mexican Migrants in the U.S. Labor Market," *Quarterly Journal of Economics* 118, 549–599.

Neumark, David, Junfu Zhang, and Brandon Wall (2007). "Employment Dynamics and Business Relocation: New Evidence from the National Establishment Time Series," *Research in Labor Economics* 26, 39-83.

Parker, Simon C. (2004). *The Economics of Self-Employment and Entrepreneurship*, New York: Cambridge University Press.

Wang, Shing-Yi (2008). "Credit Constraints, Job Mobility and Entrepreneurship: Evidence from a Property Reform in China," mimeo, Department of Economics, New York University.

Wei, Cheng (2008). *An Investigation of Farmers Turned Workers in China*, Beijing: Law Press.

Xin, Katherine R. and Jone L. Pearce (1996). "Guanxi: Connections as Substitutes for Formal Institutional Support," *The Academy of Management Journal* 39, 1641-1658.

Xu, Xuchu and Wenrong Qian (2009), eds. *Survival Story—Interviews with 50 Farmers Turned Workers*, Hangzhou: Zhejiang University Press.

Yueh, Linda (2009). "Self-Employment in Urban China: Networking in a Transition Economy," *China Economic Review* 20, 471-484.

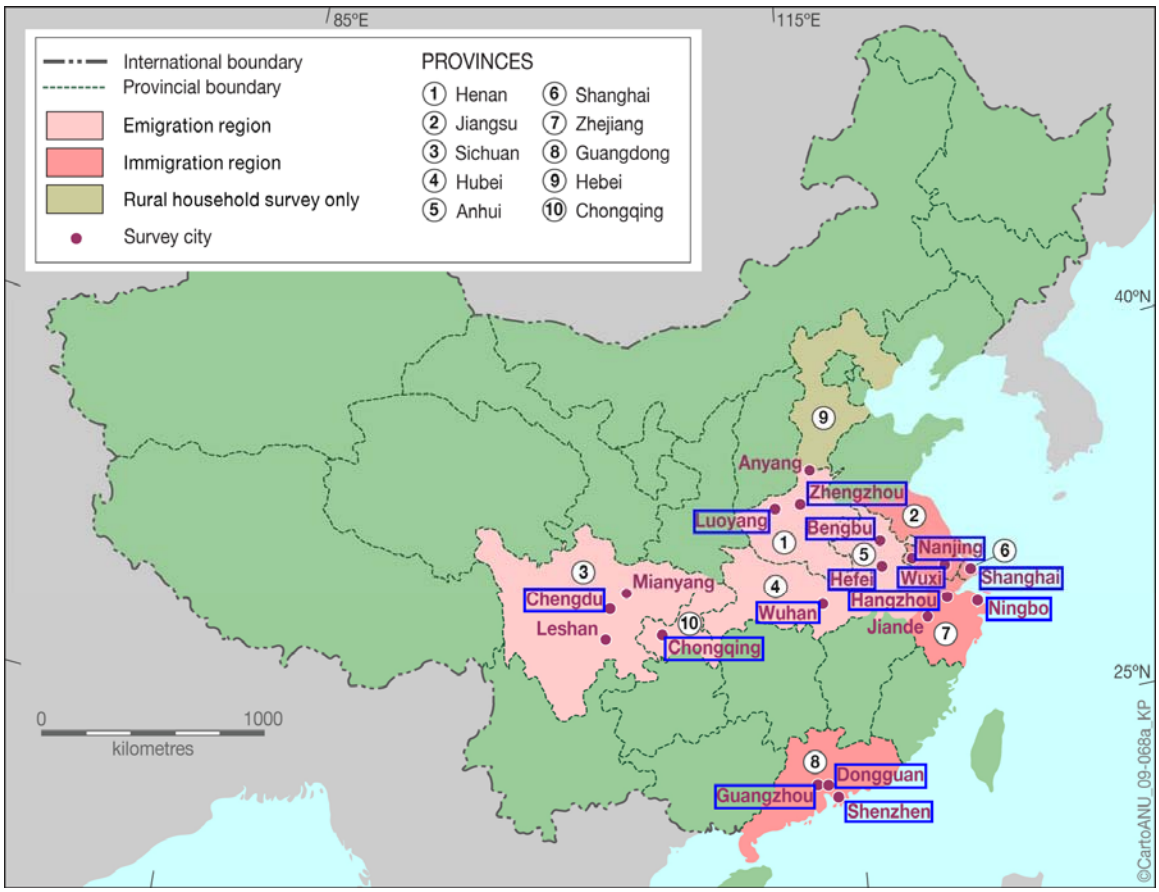
Zhang, Kevin Honglin and Shunfeng Song (2003). "Rural–urban migration and urbanization in China: Evidence from time-series and cross-section analyses," *China Economic Review* 14, 386-400.

Zhang, Xiaobo and Guo Li (2003). "Does *Guanxi* Matter to Nonfarm Employment?" *Journal of Comparative Economics* 31, 315-331.

Zhao, Yaohui (1999). "Leaving the Countryside: Rural-To-Urban Migration Decisions in China," *American Economic Review Papers and Proceedings* 89, 281 -286.

Zhao, Yaohui (2003). "The Role of Migrant Networks in Labor Migration: The Case of China," *Contemporary Economic Policy* 21, 500-511.

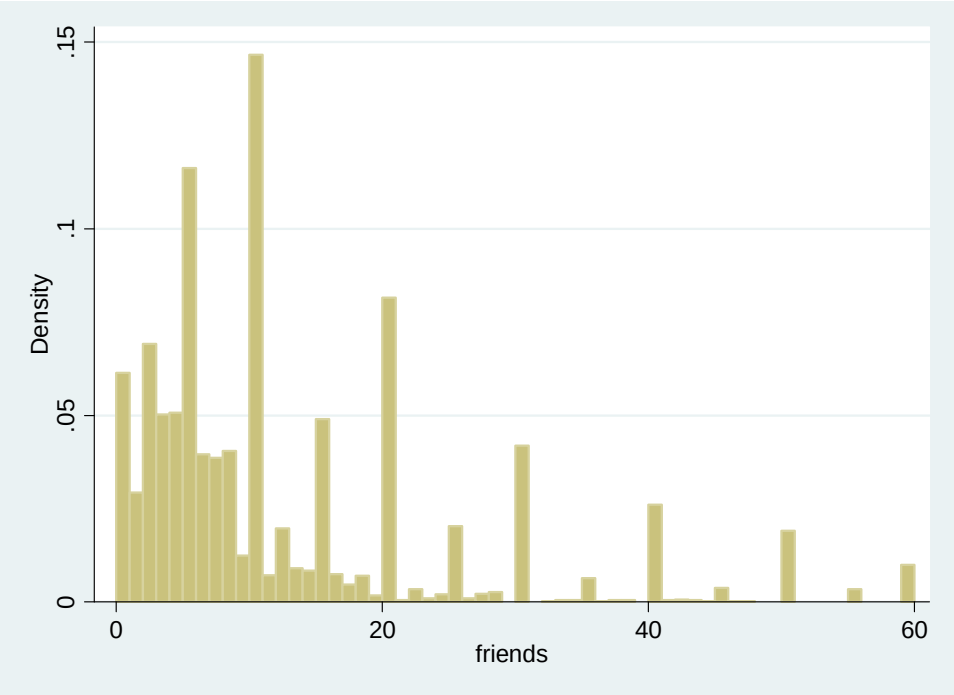
Figure 1: The Fifteen Cities Where Rural-Urban Migrants Are Surveyed



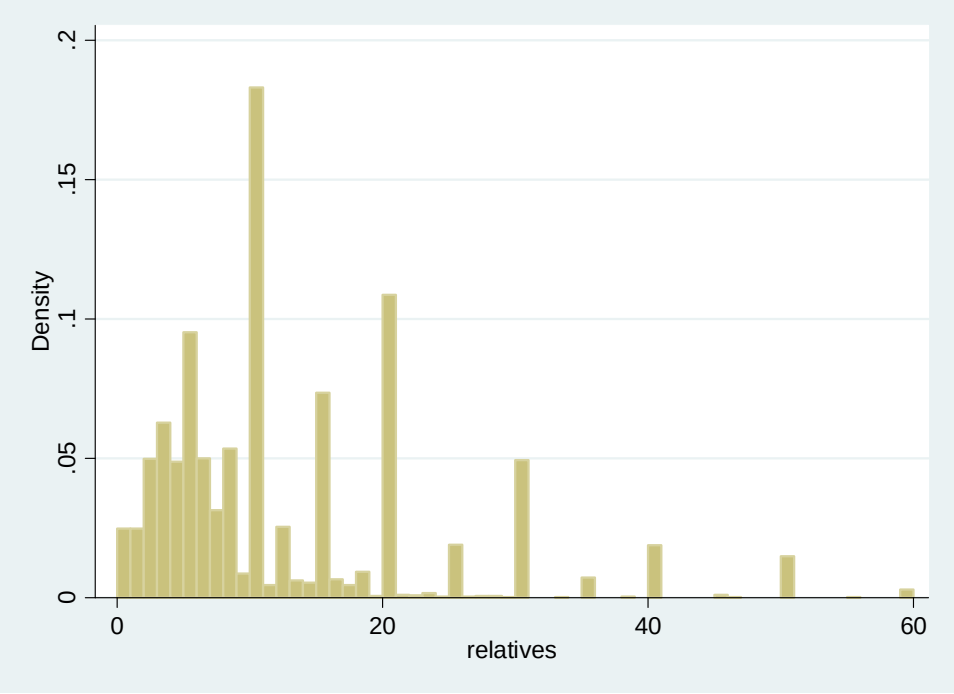
Source: The Rural-Urban Migration in China and Indonesia Project Website (http://rumici.anu.edu.au/joomla/index.php?option=com_content&task=view&id=49&Itemid=52), with modifications.

The rural-urban migrants are surveyed in the 15 cities that are highlighted with blue rectangles. Urban households are surveyed in all the 18 cities on this map.

Figure 2: Potential measurement errors in self-reported measures of network size



(a) Histogram of number of friends greeted



(b) Histogram of number of relatives greeted

Figure 3: Distance of first migration and the choice of self-employment

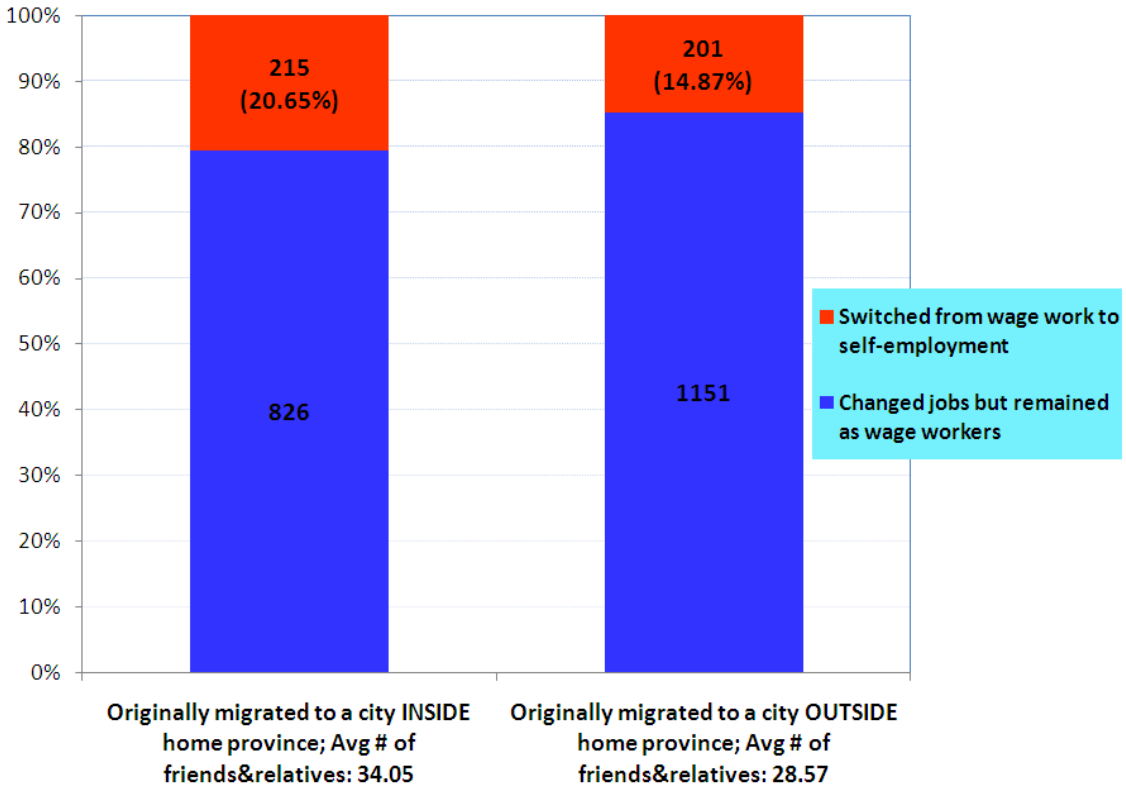


Table 1: Descriptive statistics

Variables	A. Summary Statistics (whole sample)				B. Summary Statistics (excluding outliers)	
	Mean	**Std. Dev.**	**Min**	**Max**	**Mean**	**Std. Dev.**
Whether currently self-employed	0.197	0.398	0	1	0.196	0.397
Number of friends	18.36	40.27	0	998	16.15	24.34
Number of relatives	13.23	21.52	0	998	12.53	12.43
Number of friends & relatives	31.58	56.18	0	1996	28.68	32.32
Number of other acquaintances	3.70	157.3	0	9974	0.672	5.582
Total number of people greeted	34.11	162.7	0	9999	28.50	30.95
Sex	0.691	0.462	0	1	0.690	0.462
Age	30.37	10.22	16	69	30.38	10.23
Married	0.535	0.499	0	1	0.535	0.499
Years of schooling	9.302	2.386	1	20	9.288	2.380
Log distance from home province when first migrated	3.153	3.430	0	8.313	3.148	3.429
Whether stayed within home province when first migrated	0.482	0.500	0	1	0.482	0.500

The statistics in column A are based on the whole sample that includes all household heads aged between 16 and 70.
The statistics in column B are based on the truncated sample that excludes all household heads with a number of total contacts above the 99^{th} percentile (larger than 200).
T-tests are conducted to compare means between the two samples. No statistically significant differences are detected for any of the variables other than the five network-size variables

Table 2: Number of friends and choice of self-employment

(Dependent Variable: whether self-employed or not)

	OLS Regressions				**2SLS Regressions (IV: Log distance when first migrated)**			
	(1) Sample A	**(2) Sample B**	**(3) Sample C**	**(4) Sample D**	**(5) Sample A**	**(6) Sample B**	**(7) Sample C**	**(8) Sample D**
Number of friends	.00027	.00045	.00043*	.00045**	.0110**	.0147**	.0134**	.0115**
	(.00025)	(.00026)	(.00022)	(.00022)	(.0047)	(.0063)	(.0056)	(.0048)
Age	.0015	-.0002	-.0014	-.0023	.0032*	.0026	.0014	.0001
	(.0013)	(.0015)	(.0017)	(.0017)	(.0017)	(.0022)	(.0021)	(.0019)
Sex	-.0202	.0060	.0060	.0050	-.0547*	-.0439	-.0382	-.0320
	(.0173)	(.0215)	(.0193)	(.0191)	(.0331)	(.0428)	(.0364)	(.0332)
Years of schooling	-.0144***	-.0130***	-.0123***	-.0109***	-.0303***	-.0363***	-.0326***	-.0285***
	(.0027)	(.0029)	(.0031)	(.0031)	(.0069)	(.0113)	(.0099)	(.0092)
Married	.2139***	.2019***	.1991***	.1946***	.2107***	.2069***	.2040***	.1983***
	(.0343)	(.0388)	(.0393)	(.0403)	(.0319)	(.0388)	(.0407)	(.0404)
Constant	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Home province fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
First-stage F statistic	----	----	----	----	14.56	9.05	9.20	9.46
No. of observations	4,473	2,489	2,393	2,349	4,475	2,491	2,395	2,351

Outliers above the 99th percentile (with more 200 total contacts) are excluded from regression. Sample A includes all household heads aged between 16 and 70 years. Sample B includes all household heads in sample A who are not on their first jobs in urban sectors. Sample C includes all household heads in sample A whose first jobs were not self-employment and who are not on their first jobs in urban sectors. Sample D includes all household heads in sample A whose first jobs were not self-employment and who are not on their first jobs in urban sectors, excluding those who choose self-employment because they cannot find other jobs.

Standard errors in parentheses are robust to heteroskedasticity and clustered at the city level.

*** statistically significant at the 1 percent level; * statistically significant at the 5 percent level; * statistically significant at the 10 percent level.

Table 3: Number of relatives and choice of self-employment

(Dependent Variable: whether self-employed or not)

	OLS Regressions				2SLS Regressions (IV: Log distance when first migrated)			
	(1) Sample A	**(2) Sample B**	**(3) Sample C**	**(4) Sample D**	**(5) Sample A**	**(6) Sample B**	**(7) Sample C**	**(8) Sample D**
Number of relatives	.00008	.00074	.00079	.00075	.0268**	.0422**	.0369*	.0300**
	(.00047)	(.00061)	(.00059)	(.00058)	(.0107)	(.0206)	(.0192)	(.0128)
Age	.0015	-.0002	-.0015	-.0023	.0028*	.0011	-.0006	-.0016
	(.0013)	(.0015)	(.0017)	(.0017)	(.0016)	(.0021)	(.0021)	(.0019)
Sex	-.0194	.0073	.0071	.0063	-.0405	-.0113	-.0091	-.0041
	(.0170)	(.0214)	(.0192)	(.0189)	(.0338)	(.0461)	(.0407)	(.0348)
Years of schooling	-.0140***	-.0127***	-.0120***	-.0105***	-.0272***	-.0313***	-.0282***	-.0238***
	(.0026)	(.0028)	(.0030)	(.0029)	(.0051)	(.0086)	(.0082)	(.0065)
Married	.2138***	.1997***	.1968***	.1925***	.1372***	.0855	.0995**	.1159***
	(.0348)	(.0386)	(.0391)	(.0403)	(.0381)	(.0530)	(.0455)	(.0333)
Constant	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Home province fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
First-stage F statistic	----	----	----	----	14.04	8.97	7.45	12.49
No. of observations	4,473	2,489	2,393	2,349	4,475	2,491	2,395	2,351

Outliers above the 99th percentile (with more 200 total contacts) are excluded from regression. Sample A includes all household heads aged between 16 and 70 years. Sample B includes all household heads in sample A who are not on their first jobs in urban sectors. Sample C includes all household heads in sample A whose first jobs were not self-employment and who are not on their first jobs in urban sectors. Sample D includes all household heads in sample A whose first jobs were not self-employment and who are not on their first jobs in urban sectors, excluding those who choose self-employment because they cannot find other jobs.

Standard errors in parentheses are robust to heteroskedasticity and clustered at the city level.

*** statistically significant at the 1 percent level; * statistically significant at the 5 percent level; * statistically significant at the 10 percent level.

Table 4: Number of friends and relatives and choice of self-employment

(Dependent Variable: whether self-employed or not)

	OLS Regressions				2SLS Regressions (IV: Log distance when first migrated)			
	(1) Sample A	**(2) Sample B**	**(3) Sample C**	**(4) Sample D**	**(5) Sample A**	**(6) Sample B**	**(7) Sample C**	**(8) Sample D**
Number of friends & relatives	.00016	.00037*	.00037**	.00037**	.0078**	.0109**	.0098**	.0083***
	(.00019)	(.00021)	(.00018)	(.00021)	(.0032)	(.0045)	(.0041)	(.0032)
Age	.0015	-.0002	-.0014	-.0023	.0031*	.0022	.0008	-.0004
	(.0013)	(.0015)	(.0017)	(.0017)	(.0016)	(.0020)	(.0020)	(.0018)
Sex	-.0200	.0061	.0060	.0052	-.0505	-.0355	-.0304	-.0243
	(.0172)	(.0217)	(.0194)	(.0192)	(.0327)	(.0429)	(.0371)	(.0330)
Years of schooling	-.0143***	-.0130***	-.0124***	-.0110***	-.0294***	-.0351***	-.0314***	-.0272***
	(.0027)	(.0029)	(.0031)	(.0030)	(.0060)	(.0100)	(.0089)	(.0080)
Married	.2135***	.2009***	.1981***	.1936***	.1893***	.1755***	.1761***	.1754***
	(.0343)	(.0386)	(.0391)	(.0402)	(.0322)	(.0340)	(.0348)	(.0352)
Constant	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Home province fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
First-stage F statistic	----	----	----	----	18.12	11.90	11.66	13.14
No. of observations	4,473	2,489	2,393	2,349	4,475	2,491	2,395	2,351

Outliers above the 99th percentile (with more 200 total contacts) are excluded from regression. Sample A includes all household heads aged between 16 and 70 years. Sample B includes all household heads in sample A who are not on their first jobs in urban sectors. Sample C includes all household heads in sample A whose first jobs were not self-employment and who are not on their first jobs in urban sectors. Sample D includes all household heads in sample A whose first jobs were not self-employment and who are not on their first jobs in urban sectors, excluding those who choose self-employment because they cannot find other jobs.

Standard errors in parentheses are robust to heteroskedasticity and clustered at the city level.

*** statistically significant at the 1 percent level; * statistically significant at the 5 percent level; * statistically significant at the 10 percent level.

Table 5: Number of acquaintances and choice of self-employment

(Dependent Variable: whether self-employed or not)

	OLS Regressions				2SLS Regressions (IV: Log distance when first migrated)			
	(1) Sample A	(2) Sample B	(3) Sample C	(4) Sample D	(5) Sample A	(6) Sample B	(7) Sample C	(8) Sample D
Number of acquaintances	.00037	.00164	.00034	.00008	.5500	.3769	.3281	.2711
	(.00105)	(.00235)	(.00151)	(.00130)	(.8912)	(.3182)	(.3114)	(.2525)
Age	.0014	-.0001	-.0013	-.0021	-.0032	-.0052	-.0002	-.0011
	(.0014)	(.0015)	(.0018)	(.0018)	(.0079)	(.0058)	(.0017)	(.0014)
Sex	-.0205	.0068	.0070	.0068	.0241	.0031	-.0290	-.0254
	(.0178)	(.0221)	(.0201)	(.0198)	(.0966)	(.0652)	(.0495)	(.0420)
Years of schooling	-.0145***	-.0123***	-.0116***	-.0100***	-.0538	-.0389**	-.0261**	-.0233**
	(.0028)	(.0029)	(.0031)	(.0031)	(.0592)	(.0197)	(.0112)	(.0096)
Married	.2135***	.1977***	.1941***	.1904***	.2398*	.1898***	.1609***	.1626***
	(.0355)	(.0408)	(.0414)	(.0426)	(.1246)	(.0489)	(.0461)	(.0373)
Constant	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Home province fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
First-stage F statistic	----	----	----	----	0.41	1.12	0.88	0.93
No. of observations	4,329	2,422	2,329	2,286	4,331	2,424	2,331	2,288

Outliers above the 99th percentile (with more 200 total contacts) are excluded from regression. Sample A includes all household heads aged between 16 and 70 years. Sample B includes all household heads in sample A who are not on their first jobs in urban sectors. Sample C includes all household heads in sample A whose first jobs were not self-employment and who are not on their first jobs in urban sectors. Sample D includes all household heads in sample A whose first jobs were not self-employment and who are not on their first jobs in urban sectors, excluding those who choose self-employment because they cannot find other jobs.

Standard errors in parentheses are robust to heteroskedasticity and clustered at the city level.

*** statistically significant at the 1 percent level; * statistically significant at the 5 percent level; * statistically significant at the 10 percent level.

Table 6: Sensitivity analysis excluding "roaming" migrants
(Dependent Variable: whether self-employed or not)

	2SLS Regressions (IV: Log distance when first migrated)					
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
	Sample E	**Sample F**	**Sample E**	**Sample F**	**Sample E**	**Sample F**
Number of friends	.0121***	.0105***				
	(.0030)	(.0033)				
Number of relatives			.0363***	.0271***		
			(.0107)	(.0096)		
Number of friends & relatives					.0091***	.0076***
					(.0022)	(.0024)
Age	.0009	.0012	-.0007	-.0014	.0005	.0005
	(.0021)	(.0024)	(.0024)	(.0019)	(.0022)	(.0022)
Sex	-.0496*	-.0371	-.0355	-.0094	-.0461*	-.0293
	(.0272)	(.0281)	(.0319)	(.0289)	(.0276)	(.0279)
Years of schooling	-.0226***	-.0223***	-.0211***	-.0157***	-.0222***	-.0205***
	(.0065)	(.0065)	(.0043)	(.0053)	(.0054)	(.0057)
Married	.1877***	.1600***	.0848*	.1047**	.1620***	.1446***
	(.0383)	(.0485)	(.0479)	(.0436)	(.0368)	(.0460)
Constant	Yes	Yes	Yes	Yes	Yes	Yes
Home province fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
First-stage F statistic	30.32	30.76	11.66	19.46	26.72	31.60
No. of observations	1,729	1,567	1,729	1,567	1,729	1,567

Outliers above the 99th percentile (with more 200 total contacts) are excluded from regression. Sample E includes all household heads in sample D that have always stayed within or outside the home province. Sample F includes all household heads in sample D that have always stayed in the same province. Sample D is defined as in Tables 2-5; it includes all household heads in sample A whose first jobs were not self-employment and who are not on their first jobs in urban sectors, excluding those who choose self-employment because they cannot find other jobs.
Standard errors in parentheses are robust to heteroskedasticity and clustered at the city level.
*** statistically significant at the 1 percent level; * statistically significant at the 5 percent level; * statistically significant at the 10 percent level.

Table 7: Sensitivity analysis with city-level controls
(Dependent Variable: whether self-employed or not)

	2SLS Regressions using Sample D (IV: Log distance when first migrated)								
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**	**(9)**
Number of friends and relatives	.0077***	.0081***	.0061**	.0070***	.0062**	.0063**	.0062**	.0078***	.0073***
	(.0026)	(.0029)	(.0031)	(.0024)	(.0031)	(.0030)	(.0029)	(.0027)	(.0026)
Share of private-sector employment in labor force in the city	-.0014			-.0023*	.0007		-.0007		
	(.0016)			(.0013)	(.0017)		(.0017)		
Share of tertiary-sector employment in total city employment		.0046		.0057**		.0033	.0039		
		(.0029)		(.0027)		(.0025)	(.0027)		
Share of tertiary-sector GDP in total city GDP			.0061***		.0068*	.0052**	.0043		
			(.0024)		(.0035)	(.0026)	(.0036)		
Share of self-employed in working household heads in the city								-.4207	
								(.2807)	
Share of self-employed in all household heads in the city									-.6036**
									(.2433)
Constant and controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Home province fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
No. of observations	2,349	2,349	2,349	2,349	2,349	2,349	2,349	2,349	2,349

Outliers above the 99th percentile (with more 200 total contacts) are excluded from regression. All regressions use Sample D, which is the same sample D as used in Tables 2-5. Sample D includes all household heads in sample A whose first jobs were not self-employment and who are not on their first jobs in urban sectors, excluding those who choose self-employment because they cannot find other jobs.
First-stage regressions are the same as in Table 4, column 8, using sample D.
In Columns (1)-(7), additional control variables are constructed using data from *China Urban Statistical Yearbook*. In Columns (8)-(9), additional control variables are constructed using data from RUMiCI urban household surveys.
In each regression, age, sex, marital status, and years of schooling are included as controls.
Standard errors in parentheses are robust to heteroskedasticity and clustered at the city level.
*** statistically significant at the 1 percent level; * statistically significant at the 5 percent level; * statistically significant at the 10 percent level.

Table 8: Number of relatives and choice of self-employment, results from alternative IVs

(Dependent Variable: whether self-employed or not)

	2SLS Regressions (IV: Number of siblings)				2SLS Regressions (IVs: Number of siblings, age*siblings, age*age*siblings)			
	(1) Sample A	(2) Sample B	(3) Sample C	(4) Sample D	(5) Sample A	(6) Sample B	(7) Sample C	(8) Sample D
Number of relatives	0.0090	0.0093	0.0065	0.0066	0.0197*	0.0102	0.0071	0.0061
	(.0080)	(.0089)	(.0085)	(.0085)	(.0115)	(.0082)	(.0074)	(.0080)
Age	.0041	.0030*	.0015	.0004	.0047	.0030*	.0016	.0004
	(.0017)	(.0018)	(.0020)	(.0020)	(.0019)	(.0018)	(.0020)	(.0020)
Sex	-.0130	.0155	.0171	.0155	-.0212	.0153	.0169	.0156
	(.0195)	(.0223)	(.0202)	(.0199)	(.0272)	(.0230)	(.0206)	(.0205)
Years of schooling	-.0157***	-.0116***	-.0109**	-.0096**	-.0212***	-.0119***	-.0111***	-.0094**
	(.0048)	(.0043)	(.0043)	(.0042)	(.0067)	(.0045)	(.0043)	(.0043)
Married	.1673***	.1586***	.1636***	.1607***	.1358***	.1561***	.1620***	.1618***
	(.0370)	(.0343)	(.0359)	(.0406)	(.0352)	(.0341)	(.0365)	(.0432)
Constant	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Home province fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
First-stage F statistic	9.04	9.93	10.56	8.06	3.07	3.70	3.69	2.91
No. of observations	4,416	2,607	2,518	2,470	4,416	2,607	2,518	2,470

Outliers above the 99th percentile (with more 200 total contacts) are excluded from regression. Sample A includes all household heads aged between 16 and 70 years. Sample B includes all household heads in sample A who are not on their first jobs in urban sectors. Sample C includes all household heads in sample A whose first jobs were not self-employment and who are not on their first jobs in urban sectors. Sample D includes all household heads in sample A whose first jobs were not self-employment and who are not on their first jobs in urban sectors, excluding those who choose self-employment because they cannot find other jobs. Notice that the sample sizes here are different from those in Tables 2-4. This is because we are using different IVs here and dropping different sets of observations due to missing IVs.

Standard errors in parentheses are robust to heteroskedasticity and clustered at the city level.

*** statistically significant at the 1 percent level; * statistically significant at the 5 percent level; * statistically significant at the 10 percent level.