Peer Migration in China*

Yuyu Chen
Peking University
Ginger Zhe Jin
University of Maryland & NBER
Yang Yue
Peking University

January 1, 2010

Abstract

This paper highlights the role of social networks in the internal migration of China. With over 130 million rural labors migrating to the city each year, China is experiencing the largest internal migration in the human history. Using the 2006 China Agricultural Census, we show that individual migration decision is not only dependent on individual attributes (such as age, gender, education and distance) but also highly clustered: migration pattern varies greatly across close-by villages; but migrants from the same village tend to go to the same destination for the same occupation. After using China's one-child policy as instruments for neighbors' migration decision in the same village, we conclude that the clustered migration is most likely driven by same-origin villagers helping each other in moving cost and job search at the destination.

* Contact information: Yuyu Chen, Guanghua School of Management, Peking University. Email: chenyuyu@gsm.pku.edu.cn. Ginger Jin, Department of Economics, University of Maryland, College Park, MD 20742. Email: jin@econ.umd.edu. Yang Yue, Guanghua School of Management, Peking University. Email: shananyueyang@gmail.com. This project is a collaborative effort with a local government of China. We would like to thank Hongbin Cai, Wei Li, Brian Viard, Roger Betancourt, Loren Brandt, Judy Hellerstein, John Ham and Matthew Chesnes for helpful comments. All errors are our own.

I. Introduction

In the past 20 years, China has witnessed an explosive growth of labor migration. Cai (1996) estimates that 34.1 million workers had left their rural home for urban jobs in 1990. This number increased to 67 million in 1999 (Huang and Pieke 2003) and 134.8 million in 2005 (Sheng 2006). Given the large income gap between the East, the Middle, and the West (Lin, Wang and Zhao 2006), it is not surprising that over one-fourth of the migration is across province (Cai 1996 & National Bureau of Statistics 2006).

Rural-to-urban migration is one of the most important drivers for economic growth. According to Young (2003), a rising labor participation rate, most of which is attributable to the transfer of labor out of agriculture, accounts for nearly one-ninth of the 7.8% annual GDP growth of China.[1] However, due to residence permit and other institutional barriers in China, most migrating workers do not migrate permanently to the city (Zhao 1999a & 1999b). Most of them leave their families at the origin and travel between rural home and urban jobs every year. This leads to a number of social issues including traffic congestion, lack of labor protection, child development problems, and a link of macro risks between origin and destination.[2]

The most fundamental questions underlying the massive Chinese migration are who migrate, why they migrate, and where they migrate to.[3] Classical theories argue that an individual will migrate if the discounted present value of income gains exceeds the direct cost of migration (Sjaastad 1962 and Becker 1975). These models emphasize geographic attributes (e.g. distance to destination), individual characteristics (e.g. age and education) and market factors (e.g. wage gap and land ownership).[4] Recently, researchers have explored the role of social interactions in migration, both theoretically (Carrington et al. 1996) and empirically (Munshi 2003, Mckenzie and Rapoport 2007, Woodruff and Zenteno 2007). A separate literature examines the effects of social networks in job search, stressing that social networks may explain a large number of empirical facts including employment heterogeneity, persistent unemployment, and income inequality (see Ioannides and Loury 2004 for a detailed review).

---

[1] Young (2003) calculates the annual 7.8% GDP growth based on published data from 1978 to 1998. He shows that the deflated annual growth is 6.1%, of which 0.9% can be attributable to the increase of labor participation. 0.9% is approximately one-nine of 7.8%.

[2] For example, the Chinese railway system has accommodated 192 million passenger-trips during the 40-day rush of 2009. A significant part of the traffic is driven by rural migrants going home before the Spring Festival and returning to urban work after the festival (http://news.xinhuanet.com/newscenter/2009-02/19/content_10849579.htm).

[3] Another important question is the impact of migration on the economy of both origin and destination. The literature has documented the effect of remittance of mine workers on agricultural productivity (Lucas 1987), the effect of migration on income inequality (Mckenzie and Rapoport 2007), the effect of migration networks on microenterprises in Mexico (Woodruff and Zenteno 2007), and the effect of migration on child health (Mckenzie and Hildebrandt 2005).

[4] See surveys of migration in Rosenzweig (1988), Borjas (1994) and Lucas (1997).

The goal of this paper is documenting the role of social networks in migration while controlling for individual factors such as age, gender, education, land pressure, and geographic distance to destination. Based on 5.9 million individual observations from the 2006 China Agriculture Census, we presumably observe every one that has a residential permit (hukou) in a continuous rural area, including those who have migrated for remote jobs. Because many agricultural, governmental and social activities are organized at the village level, Chinese village is a natural host of social networks. When a villager migrates to the city and comes back for holiday or family visit, the information he has about the destination spreads out fast to others in the same village. Given the fact that many Chinese cities impose barrier to entry for rural migrants, having an acquaintance at the destination also implies a substantial reduction of moving cost. Both "social-network" arguments imply that migrants should cluster by village. Consistently, we observe migration rate varies greatly across nearby villages but migrants from the same village cluster tend to cluster at the same destination for the same occupation.

As argued in Manski (1993), a cluster of economic actions does not necessarily imply that one's action affects the action of other people in the same network. Clustered migration may be driven by villagers having similar individual characteristics or facing similar institutional environments. Even if we rule out these correlated effects, the migration decision of one villager may add peer pressure on non-migrants or create general equilibrium effects within the village (through land redistribution and increased demand for agricultural labor), both of which could influence the migration of other villagers directly. How to distinguish these alternative explanations from the role of social networks in information and cost sharing is the main challenge of our empirical analysis.

One way to circumvent the correlated effects is finding an instrument variable that affects A's migration decision but not that of B directly except for the social interactions between A and B. By definition, such an instrumental variable must be individual-specific. Village-level instruments such as rainfall in the sending community (as used in Munshi 2003) may help determine the overall migration rate in that village, but could affect both peer and self migration and therefore does not help identify how peer migration affects an individual's own migration decision.[5]

The evolution of China's one-child policy offers a unique opportunity to construct individual-level instruments. For cultural reasons, the rural areas of China have a strong preference for boys. To accommodate this preference, the central government of China issued "Document 7" on April 13, 1984, which effectively allows rural households to have a second baby if the first child is a girl. As shown later in the paper, not only does this policy minimize sex selection on the firstborns[6], it also implies that rural

[5] As acknowledged by Munshi (2003), the time lag between early and later migrants does not circumvent the problem because rainfall of the past may have a long lasting effect on the local economy of the origin and therefore directly affect current migration.

[6] Sex selection may take several forms ranging from selective abortion, abandon of newborns, to infanticide.

households with a girl firstborn are more likely to have a second child. In other words, the gender-specific family planning policy plus the nature-determined sex of the firstborn generate exogenous variations in family size and gender composition of children, both of which are important factors to be considered in one's migration decision. Under the assumption that one household's fertility outcomes do not directly affect the migration decision of its neighbors, they are valid instruments. Later on we report a number of robustness checks to support this assumption.

After constructing instruments based on the sex of peers' firstborns and whether peers' first and second births are multiples, we find that one percentage point increase in the percent of peers migrating out of a village will increase one's own migration probability by 0.727 percentage points. If we ignore other long run considerations (say aging), counterfactual simulations suggest that a village starting with a 1% of migration in the first year will reach a migration rate of 6% by the fifth year and over 60% by the eleventh year.

The instrumental variable approach allows us to distinguish social interactions from the correlated effects, but it does not identify whether the social interactions arise because people of the same village use the village-wide social network to reduce moving cost and obtain job information at the destination, or because of peer pressure and general equilibrium effects at the origin. We argue these two explanations are separable because only the former implies that (1) migrants from the same village cluster at the same destination for the same occupation, and (2) the strength of the social interactions is greater for the origins that are more difficult to travel from. The strong cluster by destination and occupation, plus an incremental effects of peer migration by distance to the provincial capital, leads us to conclude that social network effects are the dominant force underlying the migration clusters. Towards the end of the paper, we argue that these social network effects could have profound implications on a number of socio-economic issues in China.

The rest of the paper is organized as follows. Section 2 provides a brief literature review on migration, social networks, and the use of fertility history as instrumental variables. Section 3 describes the background and data. Section 4 lays out a basic specification, examines the validity of instruments, and reports the instrumental variable results that distinguish social interactions from correlated effects. Section 5 examines three types of social interactions, namely the social network effects in cost reduction and information sharing, peer pressure at the origin, and the general equilibrium effects in land use. Section 6 simulates the snowball effect of peer migration and discusses other implications that peer migration may have on rural and urban development. A brief conclusion is offered in Section 7.

2. Literature Review

The existing literature has stressed the importance of social networks in both migration and job search, but empirical evidence still lags behind theory. In job search, the model of Calvo-Armengol and Jackson (2004) shows that job information sharing within a social network can explain why employment rate varies across networks, why unemployment rate persists in some networks, and why inequality across networks can be long lasting. Their model implies that a public policy that provides incentives to reduce initial labor market dropout could have a positive and persistent effect on future employment. In a similar spirit, Carrington et al. (1996) establish a dynamic model of labor migration in which earlier migrants help later migrants to reduce the moving costs at the same destination. In their model, migration occurs gradually but develops momentum over time. It explains why migration tends to cluster in geography and why migratory flows may increase even as wage differentials narrow.

In comparison, numerous empirical facts are consistent with the social network theory, but causal links are difficult to establish. For example, on the decision of migrating or not, having friends or relatives in Manila or Hawaii is positively correlated with whether an adult Philippine moves to these two destinations (Caces et al. 1985), having kin at a destination increases the probability of Mexico rural residents to migrate out of Mexico (Taylor 1986), and living in a village that has more early migrants tends to encourage one to migrate within China but this correlation disappears if the early migrants return to the origin village permanently (Zhao 2003). On life after migration, US immigrants are shown to be more geographically concentrated than natives of the same age and ethnicity and often employed together (Bartel 1989, LaLonde and Topel 1991). All these findings are suggestive that peer migrants may help improve information and reduce moving costs. But they are also consistent with the alternative explanation that kins, friends, neighbors and people from the same origin share common preferences, have lived in similar areas, and therefore make similar migration decisions.

Researchers have used three ways to identify social network effects from confounding factors: one is controlling for a large number of group fixed effects (say census block group as in Bayer et al. 2008) and then exploring employment cluster by a smaller unit (say census block) within the controlled group. The underlying assumption is that there is no unit-level correlation in unobserved individual attributes after taking into account the broader group.[7] This method is unlikely to succeed in our context, as one could argue that individuals from the same village may have similar unobserved attributes and these attributes differ across villages.

The second approach hinges on random assignment of peers. Duflo and Saez (2003) design a randomized experiment to study social interactions among college employees regarding participation in a Tax Deferred Account. Another example is the Moving to Opportunity (MTO) program, which provides

[7] Similar identification strategy has been used in Aizer and Currie (2004) and Bertrand et al. (2000).

housing vouchers to a randomly selected group of poor families in five US cities. Studies have documented the effects of the MTO program on adolescent behavior and adult outcomes (see e.g. Kling, Liebman and Katz 2007). While the interpretation of these findings is subject to debate[8], the social network effects to be studied in this paper are different from most MTO evaluations: instead of examining whether a change of neighborhood affects the behavior and economic outcomes of the treated families, we focus on closely-knit, long-established networks (village) and examines how an exogenous shock to some members of a network result in behavior of others in the same network.

The third identification approach is using instrumental variables (IV). For example, Angrist and Lang (2004) examine whether reassigning Boston school students to more affluent suburbs under the Metropolitan Council for Educational Opportunity (Metco) program has any impact on the performance of non-Metco students, using the predicted assignment as an IV for the actual fraction of Metco students in the class. Maurin and Moschion (2009) study a French mother's labor market participation in association with neighbors' participation, using the sex composition of neighbor's eldest siblings as IV.

The instrumental variables we propose to use are similar to that of Maurin and Moschion (2009). As detailed in Section 4, we argue that whether one has a girl firstborn or multiples in the first and second births are related to one's own migration decision, but do not affect neighbors directly. Similar identification strategy has been pursued in settings other than migration and social network effects. For instance, Rosenzweig and Wolpin (1980) use twins as an exogenous shock to study of the quantity-quality tradeoff in family fertility; Angrist and Evans (1998) use the sex composition of the two eldest siblings as an instrument to identify the effect of family size on mother's labor market participation. In a recent study that evaluates the effect of family size on school enrollment, Qian (2009) makes the same use of China's one-child policy as we do, and instruments family size by the interaction of an individual's sex, date of birth and region of birth.

We believe our instrument is more suitable for identifying the social network effects of migration than several community-level instruments used in the recent migration literature. For example, Munshi (2003) uses rainfall in the sending community as an instrument for the prevalence of Mexico migrants from that community in the US, and finds that the more established migrants there are, the better the employment status is for a new migrant from the same village. He attributes this finding to the positive role that migrant networks play in locating jobs and reducing migration costs in the US. However, as Munshi (2003) acknowledges, lagged rainfall also directly determines current employment outcomes at the origin and hence the individual's migration decision. This is why he focuses on the effect of a

[8] Clampet-Lundquist and Massey (2008) claims that MTO was a weak intervention and therefore uninformative about neighborhood effects. Ludwig et al. (2008) make a counter-argument.

migration network on employment at the destination conditional on a person has migrated to the US, not the migration decision itself.

Mckenzie and Rapport (2007) use historic migration rates as instruments for the stock of migration in the sending community and study how migration prevalence affects an individual's current migration decision and the income inequality within a community. Since historic migration rate is a community variable, it helps explain the current migration rate at the community level. But at the individual level, historic migration rate could directly affect both the extent to which village residents have ever migrated in the past and one's concurrent migration decision, especially if the current employment opportunities in that village depend on historic migration. This is the same caveat as the rainfall instrument. We overcome this problem because our instruments are at the individual instead of community level.

3. Background and Data

The land-population pressure is more acute in China than in other countries. According to the World Development Indicators constructed by the World Bank, China's rural population per square kilometer of arable land was 592 in 2000. Although this number has declined to 542 in 2005 (probably due to migration and fertility control), it is still higher than that of US (33), Mexico (98), and India (489). The high population density implies that the rural areas of China potentially have a large amount of agricultural labor that could be more productive in other activities.

The transfer from agricultural labor to non-agricultural activities takes two forms in China. One is working for local Township-Village-Enterprises. These enterprises often locate in the same village or same town, allowing workers to commute between home and work every day. The other form is migrating to a far-away city, working there, and coming back to the rural home occasionally for holidays, family visits, or agricultural seasons. As shown in Figure 1, the percentage of rural population that engages in local non-agriculture work has increased from roughly 16% in 1985 to over 25% in 2005. In comparison, the growth in the percentage of people migrating for city work is much faster, from 2.2% in 1985 to nearly 20% in 2005.

A large fraction of the rural-to-urban migration moves across province, mainly from the less developed inland provinces to the more developed coastal cities in the East. Lin, Wang and Zhao (2006) show that, in the year of 2000, the urban per capita income is 142% higher than that of rural areas. The average income gap between the West, Middle and East is not as large, but per capita income of 12 coastal provinces on the East is still 65% higher than that of inland areas. These income gaps are one of the most fundamental reasons of why inland rural laborers migrate to coastal provinces. Based on the 2000 China Population Census, Cai and Wang (2003) show that rural-to-urban migration accounts for

52% of intra-province moves but 78% of inter-province moves. Within inter-province migration, 75% of migrants move from the West and Mid-west to the East (Wang, Wu and Cai 2003). In terms of occupation, rural migrants are concentrated in manufacturing (32%), construction (22%), services (12%) and retail (5%) (Sheng 2008). Probably due to skill, language, or other differences, most western migrants are construction workers while eastern migrants are concentrated in manufacturing.

Most rural migrants cannot obtain a permanent residential permit in their working cities. Before 1984, most individual activities, including employment, schooling and social benefits, were closely tied to an individual's residence permit (*Hukou*). The enforcement of the *Hukou* system was relaxed over time, partly because more and more enterprises are not province-owned and do not require local *Hukou* for employment, partly because Chinese government has adopted a series of policies that allow people to live in cities without local *Hukou* (Chen 2006).

For example, if a rural-to-urban migrant wants to work in a city for more than one month, the city will issue a temporary residential permit (TRP) condition on the migrant's employer proving his or her employment status. Alternatively, the migrant can apply for a TRP by staying in the house of a local resident (or a hotel) but that resident (or hotel owner) must show a valid residential permit of the city to the local police. If a migrant is caught without *Hukou* and TRP, s/he is subject to fine and could be sent back home. Like *Hukou*, TRP constitutes a barrier to entry into urban areas and is quite controversial. Some cities tried to eliminate TRP, but many of them end up reactivating it because local residents prefer to have it due to safety reasons.[9] By the time of our sample period (2006), most destinations observed in our sample still issue and enforce TRP.

In short, although the relaxed *Hukou* system and the introduction of TRP have fostered rural-to-urban migration, they do not facilitate permanent migration to the city. Without urban *Hukou*, rural migrants have to keep their families at home, work alone in the city, and tolerate discrimination in schooling, housing, health insurance, work protection, and retirement benefits. Knight, Song and Jia (1999) document that migrants on average spend 6.8 months away from home in 1993. Using newer data, Du (2000) finds that the away time is on average 8 months per year for migrants from Sichuan (a western province) and 7 months for migrants from Anhui (a middle province).

Despite the inconvenience of long-distance travel between home and work, most migrants work away from home for years and do not return permanently to their rural origin. Using data from six provinces, Zhao (2002) estimates that 8.3% of those that migrated in 1998 returned to the origin area and remained there from the end of 1998 to August 1999. Based on more recent data, Sheng (2008) finds that

[9] As documented in Wan, Wang and Li (2004), temporary residential permit has reduced urban crime because (1) it facilitates the management of rural-to-urban migrants, (2) it helps the police to target crimes committed by or towards migrants, and (3) the police can educate migrants in laws and law enforcement.

among all the rural people who migrated in 2003, only 7.1% returned home and did not migrate again during 2004. The return percentage is the lowest for those in the 16-25 age group (6.2%) and the highest for age 45 and above (16.7%).

Labor market As summarized in Zhao (2005) and Cai, Park and Zhao (forthcoming), the labor market within China is segregated due to institutional barriers and social discrimination. Despite the large inflow of rural labor into urban areas, rural and urban workers are not close substitutes. Most rural-to-urban migrants are concentrated in dangerous, dirty or low-pay jobs that urban workers prefer not to undertake. Since most migrants are unskilled and do not have families in the city, being able to endure hardship and being willing to accept strict management are their two main assets.

The labor market for rural-to-urban migrants is also plagued by the lack of information. After surveying 439 rural migrants in the city of Chang Sha in Spring 2004, Chen (2005) finds that most migrants found the job via informal channels: 57.2% relied on the introduction of relatives, friends, or migrants from the same origin; 13.2% contacted potential employers directly; 6.1% responded to employer recruitment (excluding mass media ads); 1.9% were self-employed; and only 1.4% found a job via advertisements on TV, newspapers or billboards. The fraction of government-organized migration is even smaller (0.5%). When asked how easy it is to find a job in the city, 44.5% answered difficult or very difficult. For the biggest hurdle of job search, 38.3% mentioned the lack of a social network and 25.1% mentioned the lack of job information. Instead of surveying people that have migrated to the city, Du, Park and Wang (2005) asked 582 rural households in four western counties to list the most important factors that affected their migration decision in 2000. The lack of information and social networks is the third most mentioned factor determining men's migration, only lagging behind agricultural labor demand and low education.[10]

Local governments play a limited role in matching rural migrants and urban employers. For example, destination governments often organize job markets for rural-to-urban migrants, but these markets are held in a conference center within the city. Rural migrants must go to the city first before attending these job markets. At the other end, origin governments could contact far-away employers directly and organize a group of rural residents to migrate to a destination that already offer jobs for the migrants. But according to the 2003 rural survey conducted by the National Bureau of Statistics, only 3.3% of the 113.9 million rural migrants were employed via government-organized migration (Jian 2005).

[10] The four most important factors for men are (1) agricultural labor demand (25.9 percent), lack of education or skills (25.3 percent), lack of information and social networks (18.3 percent), and inability to finance transportation and search costs (14.1 percent).. The three most important factors for women are (1) unwillingness to be separated from children (46.2 percent), agricultural labor demand (21.0 percent), and lack of education or skills (12.7 percent).

The rest relied on friends and relatives (41.3%) or self search (55.4%). These numbers, together with the individual surveys cited above, suggests that social networks and other informal channels play a dominant role in determining whether, when and where to migrate.

Probably because of the information problem (and lack of labor protection for migrants), 2004 witnessed a significant shortage of migrating labor. According to the estimates from the Department of Labor in September 2004, roughly 10% of positions were left unfilled in the Pearl River triangle area and the labor shortage was concentrated in low-pay, skillful, and labor-intensive jobs such as toy and electronic assembly. This event suggests that many employers in coastal provinces did not foresee the short supply of migrant labor. In response to the problems exposed in the 2004 labor shortage, the central government issued a new policy in 2006 aiming to improve job information and training opportunities for migrating labor. In Section 5, we will discuss whether government-organized migration is a valid explanation for the migration patterns observed in our sample.

One-child policy China started fertility control in early 1970s. The crude birth rate, measured by the number of childbirth per 1000 people per year, has declined sharply from 33‰ of early 1970s to 21‰ in 1980s and 14.03‰ in 2000 (Men and Zeng 2003). Similarly, total fertility rate (TFR), which captures the average number of children that would be born per woman throughout all her fertility years, also decreases from 5.8 in 1949-1969 to 5.44 in 1970 and 2.24 in 1980s (Zhang 1998). According to the World Development Indicators published by the World Bank, China's TFR is 1.72 in 2007, which is much lower than the population replacement rate (2) and the TFR of US (2.10), UK (1.90) or India (2.68).

The most common fertility control methods used in early 1970s were educating young couples to reduce fertility and increase birth space, giving away birth control materials, and offering paid leave after free and voluntary sterilization. These tools were the most effective in urban areas, but less successful in rural areas mostly because rural households have no retirement benefits and often rely on sons for elderly care (Liang and Lee 2006). The strong boy preference made it difficult to implement the one-child policy, especially if the firstborn is a girl. The aggressive implementation of the "one-child" policy in early 1980s and the subsequent conflicts in rural areas motivated the central government to relax the one-child policy in 1984. The most important change in the 1984 policy is allowing a rural family to have a second child if the firstborn is female. With an intention to reduce infanticide of firstborn girls, the percentage of rural households receiving a second child permit has increased from 5% in 1982 to 50% in 1986 (White 1992).

The actual implementation for the permit of a second child is up to local governments. The province of our data area stipulates that a household with both parents having rural *hukous* is eligible to apply for the permit of a second child if (1) the first born is a girl, or (2) at least one of the parents belongs to a minority ethnicity, or (3) the mother of the child is a single child herself and the father of the child lives with the parents of the child's mother. However, the government won't issue the permit if the

mother was less than 30 year old at the time of the first birth and the birth space between the two births is less than four years. If a rural household has a second birth without the permit, the household is subject to monetary fines.

The family planning policy that allows a second child after a girl firstborn has several implications. First, there should be little gender selection in the firstborns if every family with a girl firstborn is allowed to have more children. Second, conditional on a girl firstborn, households with a strong boy preference will increase family size and try to get a boy in the second birth. Both implications are confirmed in Ebenstein (2009). He shows that the sex of firstborn is balanced (51% being boy) and has changed little between 1982 and 2000, hence the imbalance between male and female as observed in Sen (1990) is mostly driven by gender selection for the second and later-borns. As detailed in Section 4, the gender-specific family planning policy plus the nature-determined sex of firstborns generate exogenous variations in family size and gender composition of children, which are important factors to be considered in one's migration decision but not that of neighbors. This allows us to use the sex of peers' firstborns and whether peers' first births are multiples as instruments for peer migration.

As documented in Li, Zhang, and Zhu (2005), China's one-child policy is only applicable to the *Hans*, which represent 92% of the Chinese population. Hence the above instrumental variable logic does not work for ethnic minorities. Unfortunately, our data do not contain information as to whether an individual is a minority or not. The inability to distinguish *Hans* and minorities will weaken the power of our instruments but not to a great extent. Since the data tell us whether a village is a gathering place for minorities, we will present results with and without minority villages as a robustness check.

Data Description The National Bureau of Statistics of China has organized local governments to conduct two rounds of the China Agricultural Census (CAC) in 1996 and 2006 respectively. Aiming to produce reliable statistics for rural population and activities, the CAC is designed to cover every individual that resided or had registered residence in every village at the time of interview. The exhaustive nature of CAC, and the administrative nature of village, allows us to have a clear boundary of social networks by village. It also allows us to test if a boarder definition of social network will yield different results.

Drawn from the 2006 CAC, our data cover all the rural residents in a poor area of China as of December 31, 2006. This project is a collaborative effort with the local government, in order to better understand rural population and agricultural activities in this particular area. We are not allowed to reveal the geographic location, but we can assure readers that the studied area belongs to an inland province whose per capita income is significantly lower than the national average. In total, we observe 5.9 million individuals in 1.4 million households and 3,986 villages. These villages belong to 250 townships and spread across 8 counties. The size of the whole census area is roughly 16,000 km^2 total, with on average

area of 4 km^2 per village. Compared to other migration data that often contain a limited sample of households from a small number of communities (e.g. the Mexican Migration Project used in both Munshi 2003 and Mckenzie and Rapoport 2007 surveys 57 rural communities and 200 households per community), we are able to define who and who are in the same network, the demographics of network members, and the migration decision of each member.

The main part of the data was collected at the household level. The household head was asked to enter information for every family member. If a resident was away from home at the time of interview, his/her information was still collected from the household. By this design, we observe detailed household information including how many individuals reside in the household, their relationship to the household head, their age and gender composition, the amount of contract land, the amount of land in use, ownership of housing, the self-estimated value of house(s), ownership of durable goods, the availability of electricity, water and other amenities, the number of household members that receive government subsidies, and engagement in various agricultural activities.

Individual level data are limited to age, sex, education, employment, occupation, and the number of months away from home for out-of-township employment in 2006. Since a child in the studied area may get married as early as 17 and daughters often leave their own home after marriage, we restrict our child definition to age 0-16. One complication of the data is that we do not always have a clear definition of spouse- or parent-child relationship because one only reports his/her relationship to the household head. This problem is more likely to occur in households with three or more generations (9.99%), than in households with two (82.03%) or one (7.98%) generations. Out of the 1.19 million households that have at least two generations, 0.856 million (or 71.69%) has children under age 16. For the 0.661 million households that we can clearly identify parent-child relationship for all family members, we know the age and sex of each child. We use this information to infer the birth order of each child. For the other households, each adult's own fertility information is missing. However, we know the number of people by age groups at the household level. This is why our examination of individual migration decisions control for a long list of household attributes and a short list of individual attributes.

Although we cannot track every one's fertility history, we can still calculate the average percentage of multiple birth or the average gender of firstborns at the village level, as long as the real values of these variables are uncorrelated with whether the household relationship is too complicated to infer who is whose child. To justify this assumption, we will present results including only two-adult families. Another implication from the lack of fertility history in the "complicated" families is our village-average calculation may introduce large measurement errors if the village is small. To address this concern, we drop from the study sample all the 21 villages that have less than 100 adults between age 17 and 60.

Supplemental data were collected at the village level including the size of the village in both arable land and registered population, whether the village is a place for minority gathering, the distance to the nearest bus/rail/dock station, access to water, electricity and other amenities, and whether the village has a national poverty status (as designated by the Central government). Due to measurement errors in the registered population (e.g. some children may not register at birth), we calculate the number of adults per village from our study sample and use it to proxy village population. The data also include several township level variables, including the number and nature of township-village-enterprises, the distance between the township and the county center, whether there is a highway exit within the boundary of the township, and registered population of the township.

Above all, the 2006 CAC data is especially suitable for the study of social network effects in migration because we observe one's own migration as well as the migration decision of almost all the other adults in the same village. One shortcoming is that we only observe the migration decision at the data collection time and cannot identify who have migrated long before 2006 and who just started to migrate in 2006. For this reason, the social network effects identified in this study only reflect a correlation between self and peer migration, which could be driven by a group of adults migrating together or some migrants migrating early and then helping others to migrate afterwards. Similarly, we only observe where a migrant migrates to as of December 31, 2006; we don't know whether s/he has moved directly from the village to the destination, or stepwise from the village to some intermediary location first and then from the intermediary location to the reported destination.

Our sample construction involves several steps. For the purpose of migration, we focus on adults between age 17 and 60. Individuals that have non-rural *Hukou* and students that are currently enrolled in school are dropped from the sample. After dropping villages that have less than 100 adults (to reduce measurement error in peer migration), we end up with a final sample of 3.3 million adults in 3950 villages. These villages belong to eight counties, which allow us to examine village-by-village differences within each county. In theory, we can also impose township fixed effects and focus on village variations within the same township. However, since on average we have only 15 villages per township and these villages are geographically adjacent to each other, township fixed effects will absorb a large amount of heterogeneity across villages. In light of this, we control for township variables in the main specification (with county fixed effects) but apply township fixed effects for a robustness check.

The CAC questionnaire asks explicitly how many months a respondent has been away from the residential village for *out-of-township employment* during 2006. One month away from home is defined as being away for more than 15 days in that month. Based on this question, we define an adult as a "migrant" if s/he has been away for at least one month in 2006. This definition yields 17.08% of adults in our sample being migrants in 2006. As shown in Figure 2, the majority of migrants report that they stay

away from home for at least 10 months a year. This suggests that most migrants live and work in a far-away place and only come back home for short visits. An alternative definition of migration as six-or-more months away from home renders very similar results. Each migrant is also required to report the migration destination and occupation. Destination is reported by whether the migrant works in or out of the studied area if it is within the same province, and by province if the destination is out of the studied province. Occupation is reported in the category of manufacturing, construction, services, or other.

Data Summary Table 1 reports summary statistics for major migration destinations. In addition to the percent of migrants going to each destination, we report the number of bus/railway hours needed to transport to each destination from the center of the studied area[11], as well as the relative income across destinations. Scaling the 2006 per (rural) capita income of the studied area to one, Table 1 shows that almost all the destinations have significantly higher income than the studied area; some are even eight or ten times higher.[12] Consistent with the literature, the most attractive destinations are either high-income or within a short distance. However, income gap and distance do not explain everything. For example, destination F has the highest income per capita in the list. The next highest-income destination (A) is almost the same distance from the sampled area as F, but the percent of migrants to A (27.86%) is much higher than to F (1.92%). Apparently other forces are at work when people decide where to migrate.

Table 2 reports summary statistics for the individual, household, village, and township-level variables to be used in our study. For comparison, we first report the whole sample and then by migrants and non-migrants separately. Consistent with the literature, migrants are on average 10 years younger, have one more year of schooling, and are more likely to be male and the head of household. To be more specific, Figure 3 reports the percent of migration by age and gender. It is clear that young adults aged 20-25 are most likely to migrate. Migration tendency declines sharply after age 30. The percent of migration is similar for men and women before age 22, but men are significantly more likely to migrate after 22, probably because married women stay home for childbearing, child care and elderly care.

In terms of family structure, Table 2 shows that migrants are more likely to come from a household that has fewer children under age 16. Interestingly, the probability of having any boy is 41% for migrating households, which is much lower than that of non-migrants (51%). This difference suggests that migration may be related to the boy preference. For instance, parents that prefer boys may want to give better child care to boys, or parents with boys may feel less necessary to work and save for

[11] Since there is no railway station in the studied area, we first compute the bus hours from the area center to the province capital and add that to the number of railway hours from the province capital to other provinces.
[12] The comparison is based on China National Statistical Book, so the income difference may reflect differences in observable attributes. For example, a rural migrant to A may not expect to earn the average income in A because he is less educated and does not have full access to all the job opportunities of his education level due to hukou requirement in some city jobs. In this sense, Table A is only suggestive.

themselves because their sons will provide elderly care in the future. Table 5 explores these channels in more details.

As expected, both capital ownership and ease to transport differ between migrants and non-migrants. Migrants are more likely to have a lower house value and some outstanding loans, but their contracted land (at the household level) is no less than that of non-migrants. The latter is masked by the difference in the number of adults within a household. At the village level, migrants do have less land per adult. As we would expect, migrants have less land in use than non-migrants because they spend the most time away from home. In terms of transportation, migrants are 13% closer[13] to the nearest bus, rail or dock station, and they are more likely to live in a village with more access to drivable road.

Migration clusters All the social network theories predict heterogeneity across networks. Accordingly, Figure 4 plots the histogram of migration percentage per village and shows that this percent can be as low as 0% and as high as over 50%. For the same reason, the last panel of Table 2 reports the percent of same-village adults that migrate in 2006 excluding adults in own household, conditional on whether the individual under study is a migrant or not. Comparing columns (2) and (3), it is clear that the migrants are more likely to come from high-migration villages. This could be driven by social interactions among villagers, or omitted attributes that are similar for every one in the same village.

Our first attempt to separate social interactions from omitted variables is taking a village as the unit of observation and regressing migration percentage per village on village level variables including registered village population, whether the village is a poverty village, whether the village is a minority gathering place, average house value, average people per household, average age, average gender, land per household, average education of adults, distance to the nearest bus/rail/dock station, and township fixed effects. This regression has an R-square of 0.578, which suggests that village-level observables only explain 57.8% of the cross-village variation in migration. Figure 5 plots the histogram of residuals from this regression. The comparable dispersion of Figures 4 and 5 confirms the impression that a large fraction of across-village migration variations are driven by something else other than fundamental socio-economic difference across villages.

The cluster pattern of migration is more striking if we examine the distribution of destination and occupation within each village. For example, the first row of Table 3 shows that, if we single out the most common occupation within each village, 75.1% of same-village migrants go to such a destination. This number is much higher than what we would get if we redo the exercise by township (51.93%), county (46.12%) or the whole area (46.48%). Similarly, the percent of migrants to the most common destination is more concentrated by village (63.8%) than by township (39.82%), county (31.86%), or the whole area (27.86%). The rest of Table 3 shows that same-village migrants are more clustered by the combination of

[13] This percentage is computed by 1- (avg distance of migrants)/(avg distance of non-migrants) = 1-5.35/6.16.

destination, occupation and surname than migrants from the same township or the same county. All these statistics support the conjecture that each village is a closely-knit social network and people interact with each other much more within the village than across villages. Given the facts that the average area per village is only 4 km^2 and villages in the same township are adjacent by definition, the migration clusters shown here is similar to the employment clusters documented in Bayer et al. (2008). In Bayer et al. (2008), workers residing in the same census block tend to work in the same census block, as compared to residents of nearby census blocks. However, unlike Bayer et al. (2008), we use instrumental variables to further control for potential omitted variables at the village level.

The last panel of Table 2 reports summary statistics for the percent of same-village adults that have a girl firstborn (conditional on the first birth is not a multiple), and the percent of same-village adults whose first birth is a multiple, both excluding adults in own household. Although the mean of these two variables are only different in the third decimal point across migrants and non-migrants, the t-statistics for the test of equal mean is very large (777.1 and 52.4) thanks to the large sample. As shown in Table 5, the correlation between one's migration decision and whether this individual has a girl firstborn is highly significant, once we control for county fixed effects. More evidence for the validity of instruments and tests of weak instruments are presented in Section 4.

4 Basic Specification with Instruments

For an individual *i* in household *h*, village *v*, township *t* and county *k*, the basic specification is:

$$y_i = \alpha_k + \beta x_i + \gamma x_h + \delta x_v + \psi x_t + \lambda \bar{y}_{-i|v} + \varepsilon_i \qquad (1)$$

where y_i is a binary variable indicating whether individual *i* is a migrant in 2006; α_k denotes county fixed effects; x_i denotes *i*'s individual attributes such as age, gender, year of schooling, whether the firstborn singleton is a girl, whether the first birth is multiple, whether the second birth is multiple, as well as the minimum and maximum ages of own children. As discussed before, the variables on own children have missing values because some individuals do not have first or second birth, some individuals' family relationship is too complicated to determine parent-child relationship, and if the maximum age is above 16 (and therefore may not live at home due to marriage) we cannot identify the gender of the firstborn. Accordingly, we include a missing dummy for each of the self-child variables. To address the potential concern of sample selection, we will present a robustness check conditional on the sample of two-adult families so that they all have valid information on the self-child variables.

We control for a long list of household attributes in x_h, partly because most of our demographic variables are collected at the household level, and partly because migration decisions may be made by the household as a whole instead of by each individual separately. Within x_h, the key variables are the

number of family members by age group (0-7, 7-16, 17- 23, 24-44, 45-59, 60+), whether there is at least one boy (aged 0-16) in the household, the amount of contract land, the debt status of the household, and the prevalence of the household head's surname in the village. The last one captures the household's political status and extent of social networks within the village. We don't control for the amount of land in use by household because this could be a result of migration. In section 5, we will examine how land in use of non-migrants correlates with the degree of peer migration. The most important village level variables (x_v) includes the distance to the nearest bus/rail/dock station, access to drivable roads, the total adult population, and the total acreage of arable land. The latter two attempt to capture the degree of land-population pressure in the village. Township level variables (x_t) include the number of township-village-enterprises, the presence of highway exit(s) in the township, and the registered population.

The center of interest is the coefficient (λ) on the degree of peer migration in the same village ($\overline{y}_{-i|v}$), where $\overline{y}_{-i|v}$ is measured by the percent of same-village adults that migrate in 2006 (exclude all adults in household *h*). Equation (1) is estimated by a linear probability model. Errors are clustered by village (*v*) and adjusted for heteroscadasticity. To the extent that the omitted variables in the error term capture similar socioeconomic status, common preferences, or common environment , we expect $\lambda_{OLS} > \lambda_{2SLS}$.

Validity of Instruments We propose three instruments for peer migration ($\overline{y}_{-i|v}$). They are: (1) the percent of same-village adults whose first birth involves two (or more) children; (2) the percent of same-village female labors that reside in the households with a girl firstborn; and (3) the percent of same-village male labors that reside in the households with a girl firstborn. All three instruments are conditional on the households for which we can clearly define the oldest child, excluding household *h*. We focus on firstborns only, because births of higher order are more likely subject to sex selection (Ebenstein 2009). Both instruments (2) and (3) capture the presence of girl firstborns, but we construct them separately because, as shown below, having a firstborn girl tends to *encourage* male adults of that household to migrate for work but *discourage* females from migration. Capturing this differential effect will enhance the strength of our instruments.

More specifically, consider a village of four households: A, B, C and D. Household A has three adults: a husband, a wife, and the husband's younger sister. Assume A's firstborn is a girl. In comparison, household B is a nuclear family with a husband, a wife and a firstborn of boy. Household C has a husband, a wife, the husband's father (below age 60), and a firstborn of boy. From D's point of view, there are 4 female labor in neighboring households, 2 of which are in the household (A) with a girl firstborn. So the

percent of female labor in the households with girl firstborns is 2/4. By the same logic, the percent of male labor in the households with girl firstborns is 1/4. These two percentages are different because households A and C contain adult members other than the immediate parents of the firstborn. In our data, among households that have children (age 0-16) and adult labor (age 17-60), 0.90% involve adult siblings like A, 11.85% involve grandparents like C. This suggests that most variations between instruments (2) and (3) come from three-generation households.

The validity of the instruments relies on two assumptions: first, the gender of -*i*'s firstborn and whether -*i* has multiples in the first birth must be correlated with -*i*'s own migration decision; second, these variables must be uncorrelated with the other households' migration decision in the same village. In the absence of sex selection[14], the occurrence of twins, triplets, or a girl in the firstborn should be out of the control of a household. However, this does not automatically imply the second assumption holds because we encounter several measurement errors and the gender composition of adults within a household could be an endogenous choice conditional on the fertility outcome.

The most primary measurement error lies in the definition of firstborn. Since our data capture a one-time snapshot, the oldest child in our sample may not be the first birth if some elderly sibling(s) has grown out of age 16 or died before the data collection time. This may introduce some sex selection in favor of boys in the observed oldest child, even if the actual firstborns are balanced in sex. While we cannot rule out such sex selection, it is comforting to find that the percent of singleton girl in the observed first births (48.60%, versus 50.70% for singleton boy) is close to the natural ratio (James 1987 and Cai and Lavely 2005). Consistent with Ebenstein (2009), we also find significant gender differences in second and later-children. As shown in Table 4, households with a girl firstborn are more likely to have a second child (71%) than those with a boy firstborn (68%). Moreover, the second and later births are more likely to have (at least one) boy if the firstborn is a girl (87%) than otherwise (64%). Put it another way, the probability of having (at least one) boy and having (at least one) girl after firstborn is very close if the firstborn is a boy (64% vs. 64%), but far away if the firstborn is a girl (87% vs. 52%). All these numbers are larger than 50% because they include children born after the second births. In particular, 16.37% of all children with clear parent-child definition are third-born, and 5.51% are fourth or above. Combined, households with a girl firstborn tend to have more children (2.28) than those of a boy firstborn (2.02). This confirms the conjecture that the gender of firstborn affects family size and the gender composition of later-borns. Another way to get around the mis-definition of firstborn is only computing our instruments

[14] In a rural area as poor as our sample, there is no fertility treatment services.

conditional on the neighboring households that have all adults aged at or below 35. As shown below, our main results are robust to this alternative definition.

The second measurement error is that we may miscount two close-by births as twins because our data only report age in years instead of months or days. This data problem may lead to (1) an over-estimate on the percent of multiple births, and (2) a higher-than-natural rate of mixed gender in these multiples. The latter could occur if a girl first-born in January motivates the birth of a subsequent boy in November or December. To check these concerns, we find that among all the first births the likelihood of having two or more children at the same age is 0.70%, which is consistent with the natural probability of multiple births in both the international literature (James 1987) and the period of time in China before the implementation of one-child policy (Cai and Lavely 2005). Regarding gender mix, the percent of multiple firstborns with mixed gender (0.27%) is slightly higher than that of all boys (0.24%) and all girls (0.19%). We propose two robustness checks to address the potential measurement issue: one is not using the percent of neighboring households having multiples as an instrument; the other is restricting the calculation of the instrument to the households whose oldest same-age children are all boys. As shown below, our main results are robust to both alternatives.

Even if the fertility outcome of firstborns is exogenous, one may argue that the number of female or male labors in a household is correlated with the fertility outcome or adult composition in another household. This could happen if the two household heads are close relatives. For example, consider two middle-age brothers who have a mother of 55 years old. If one brother has a girl firstborn, he may invite the mother to live in and take care of the baby so that he can migrate out for work. This change of living arrangement may leave the other brother more (or less) likely to migrate. This story generates a potential correlation between one brother's error term and his corresponding instruments. Unfortunately, our data do not indicate whether two households are relatives or how close their blood is. However, as a robustness check, we can restrict the instruments to neighboring households that have different surnames as the studied household. While this solution is imperfect, a finding robust to this alternative definition of instrument help reduces the econometric concern.

A related concern is that the fertility outcome of two households may be correlated directly either because genetic links or peer effects in fertility. This does not generate a problem for our instrumental variable strategy if we control for the number of children and gender mix in one's own household and such control is sufficient to capture all the correlations between one's own migration decision and fertility outcome. However, an econometric bias could arise if the actual relationship of migration and fertility is non-linear. Restricting instruments to neighboring households with different surnames may limit the genetic link of fertility; and running the same specification without controlling for self fertility may provide a robustness check. Our results are robust to both.

The discussion above focuses on a potential correlation between the instruments and the error term. The other assumption for the validity of instruments is that individual *–i's* fertility outcome must be correlated with *-i*'s own migration decision. As a first test, we regress *-i*'s migration status (y_{-i}) on the gender of *-i*'s firstborn, controlling for nothing else but county fixed effects. As shown in Table 5 Column 1, this regression suggests a significant, positive correlation between migration and having a girl firstborn. This correlation hardly changes if we add *-i*'s age, education and distance to the nearest bus/rail/dock station in the regression (Column 2). The third column of Table 5 includes an interaction of *-i* being female and having a girl firstborn. Results suggest that having a girl firstborn tends to affect men and women in opposite ways: men (fathers, uncles and grandfathers) are more likely to migrate, probably because they need to earn more income for an increased family size; women (mothers, aunties and grandmothers) are less likely to migrate, probably because they have to stay home to bear the second child and take care of the children when they are young. Note that the separate correlation for men and women are 4-6 times larger than the pooling effect, which explains why our instrument variables exploit the adult gender composition in the households with a girl firstborn.

Intuitively, we suspect the correlation between migration and having a girl firstborn may arise via two channels. First, because the government allows rural residents to have a second child if the firstborn is a girl, the gender of firstborn is closely correlated with the number of children in the household. To the extent that the need for parental time increases with the number of children, especially young children, we would expect that having a girl firstborn will hinder migration. Alternatively, one can make a counter argument that raising more children requires more resources and migration is an important means to get resources. The second channel of influence is that households with boys may have different preferences for income and time. For example, a household with strong boy preference may be more willing to take close care of boys, which implies a smaller tendency to work away from home. Another possibility is that households with a boy may want to accumulate more wealth so that they can build a house when the boy is ready to marry; and migration is one way to achieve the goal. Alternatively, one may argue that because parents rely on their sons for elderly care, they may have fewer incentives to work away from home and save for themselves if they have sons. Tan (2003) suggests somewhat the opposite: while parents may still have such perception in their mind (which justifies the boy preference in fertility), parents' actual economic return from sons is no higher than that of daughters. The main reasons are (1) adult sons tend to give less percentage of their own income to the parents, (2) more and more adult sons do not live with their parents after marriage, and (3) daughters also offer elderly care to the parents, especially if the parents have no sons.

As shown in Columns 4-6 of Table 5, a household that has a girl firstborn is likely to have a larger family size, a greater number of children, and a smaller likelihood of having any boy. To the extent

that the family size and gender mix of children affect -*i*'s migration decision directly, -*i*'s migration decision will be correlated with whether -*i*'s first singleton birth is a girl and whether -*i* has had a multiple birth(s). We will recheck these correlations in the 2SLS results as we control for one's own family size and gender mix of children directly in Specification (1).

Conditional on the validity of instruments, one may suspect a weak instrument problem as the percent of girl firstborns should be close to 50%, especially for large villages.[15] As a first check, we note that the percent of girl firstborns ranges from 17.7% to 78.5% at the village level, the median number of adult labors per village is roughly 800, and the distribution of village size is 30.1% (100-1000 adults), 48.8% (1000-2000), 15.88% (2000-3000) and 5.19% (3000+). This suggests that we will have fair amount of variation in our instruments. To be sure, the instrumental variables results reported below are accompanied with a conditional likelihood ratio (LR) test for weak instruments. We adopt conditional LR because it is more robust than Anderson-Rubin and score tests (Andrews and Stock 2007).

Key results The key results of specification (1) are presented in Table 6. In addition to the OLS results in Column (1), we present three columns of instrumental variable estimates: the first using the percent of same-village adults having multiples in the first birth as the only instrument for the percent of peer migration in the same village; the second using the percentages of female and male labors residing in the households with a girl firstborn as instruments; and the third using all three instruments. As shown in Panel A of Table 6, all three instruments are highly significant and have expected signs in the first stage. In particular, peers having multiple firstborns increase the propensity of peer migration, while more female (male) labors residing in the girl-firstborn families have a negative (positive) effect on peer migration.

Panel B of Table 6 shows the 2SLS estimates. One consistent finding across the four columns is that migrants are younger, more educated, and have more access to drivable roads. They are also more likely to be male, and have less house value and less contract land. These patterns are consistent with the existing literature on both international migration (Rosenzweig 1988, Lucas 1997) and internal migration within China (Zhao 1999a & 1999b). The number of household members in any age group has a positive effect on migration, suggesting that the need for more resources to support a large family dominates the need to take care of family members. The coefficient of having a boy in the household is negative, suggesting that boy preference may hinder migration because parents (and grandparents) want to spend more times with boys or because they can rely on their sons (and grandsons) to provide elderly care and therefore have fewer incentives to work and save for themselves.

The key 2SLS coefficient for the effects of peer migration, λ, is 0.631 in column (2) and increases moderately to 0.768 (column 2) and 0.727 (column 3) if we change instruments. All three

[15] Similar argument applies to the percent of multiple birth.

estimates are lower in magnitude than the OLS estimate (0.930) reported column 1), which suggests that λ_{OLS} tends to over-estimate the actual magnitude of social interactions because omitted individual or community variables tend to affect self and peer migrations in similar ways. All three 2SLS estimates pass the conditional likelihood ratio test for weak instruments, with tight intervals of λ well above zero. While not reported, the reduced form regressions corresponding to columns (2) to (4) show similar statistical significance, which suggest that one's own migration decision is indeed correlated with neighbors' fertility outcome and family composition.[16] An over-identification test for the three instruments yields an F-statistics of 147.9 with p-value less than 1%. One explanation is that the IVs affect different parts of peers and therefore imply different types and magnitudes of social interactions.

Taking Table 6 Column (4) as the preferred specification, the 2SLS estimate suggests that every one percentage point increase in the percent of same-village adults migrating away will increase one's own migration probability by 0.727 percentage point. Two factors may explain this seemingly large effect of social interactions: First, most Chinese rural-to-urban migrants leave families at the origin and therefore have plenty of opportunities to communicate with people in the same village. Second, due to the lack of job information via formal channels, potential migrants must rely on friends, relatives, and other social networks. Given the geographic sparseness of rural areas, current migrants in the same village is likely the most important source of job information in remote destinations.

Robustness Checks We perform a number of robustness checks to ensure that the reported effects of social interactions are not driven by sample selection, variable construction, or invalid instruments.

To address the concern that working away from home for 1-2 months is not migration, we redefine migration as working away from home for at least 6 months. Column (2) of Table 7 shows that the 2SLS coefficient (using all three instruments) is similar (0.674 vs. 0.727). The concern of omitted variables at the township level leads us to replace county fixed effects with township fixed effects in the specification with all three instruments. The key coefficient of peer migration is similar (0.789) and remains significant at 1% level. Column (4) excludes the fertility outcomes of own household from the right hand side, out of the concern that there may be direct peer effects in fertility and the linear specification is not sufficient to control for the impact of self fertility on self migration. This concern is ungrounded, as the key coefficient is very similar to the main results (0.741 vs. 0.727).

Regarding the validity of instruments, Table 3 Column (5) reports that conditioning the percent of neighbors' multiple birth on all-boy twins yields a similar coefficient of peer migration (0.681) as compared to 0.631 in the Column 2 of Table 6. To address the measurement issue of firstborns, Column

[16] For example, in the reduced form regression corresponding to column (4), the coefficients for three instruments are 0.300 (standard error 0.152), 0.582 (0.125), and -0.502 (0.118). All are significant at the 1% level.

(6) of Table 7 restricts the percent of peers having a girl firstborn to the households that have all adults aged at or below 35. The magnitude of the key 2SLS result is lower than before (0.622 versus 0.768) but remains significant. Column (7) of Table 7 limits the percent of peers having twins or a girl firstborn to the households that have a different surname as the one under study. Obtaining similar results (0.810) in this alternative specification suggests that the observed correlation between self and peer migration is not driven by households being close relatives. In all these robustness checks, the key coefficient has a larger standard error than the main results, as the alternative instruments utilize fewer variations in the data.

Additional robustness checks consider two extra groups of peers. The first group is the adults that live in the same township but not in the same village. Since township covers a set of adjacent villages, same-township adults may communicate across villages. Column (8) of Table 7 report the 2SLS estimates including both the percent of peer migration in the same village and the percent of peer migration in the same township but different villages. Both are instrumented by all three IVs constructed for the relevant peers. Results suggest that peers from same township but different villages have a positive peer effect on one's own migration decision, but its magnitude (0.0108) is much smaller than that of same-village peers (0.782) and statistically close to zero. This confirms our grouping of social network by village.

In Column (9), we add information about the second peer group, namely the percent of migration for the adults that live in the same household. In theory, migration within a household may be positively correlated due to social interactions or unobserved household factors. The correlation could also be negative if the household makes individual migration decisions jointly (for example, insurance concern may motivate the household to diversify in agricultural and non-agricultural activities), or if there are unobserved individual factors that are different across family members. Unfortunately, our instruments – the gender of firstborns and the occurrence of multiple births – are applicable to both parents, hence we cannot use them for the percent of same-household adults that migrate. For this reason, the regression reported in Column (8) includes same-household migration, but do not use any instrument for this variable. Although we still use instruments for the peers outside the household, the coefficient on same-household peers does not necessarily identify the causal effect within a household. Keeping this in mind, Column (8) suggests that there is some positive correlation in the migration decisions of self and other household members, but its magnitude is much lower than the effects from other people in the same village. The relatively smaller coefficient on the percent of same household migration indicates that insurance concern may be one non-trivial factor in the migration decisions within a household.

To address concerns on potential sample selection, Column (10) of Table 7 condition the analysis sample on the households that have only two adults. These households have a simple relationship among family members, which allow us to clearly define fertility history. As shown in Table 7, the effect of peer migration is slightly lower for this sub-group of population (0.604 vs. 0.727 for full sample), probably

because two-adult families have a hard time finding live-in help for child care which is much needed if parents stay away from home for a long time. The last column of Table 7 excludes minority (non-Han) gathering villages from the sample because minorities are not subject to the one-child policy and minorities are much less likely to migrate than the Hans. The 2SLS effect of peer migration changes little (0.710 vs. 0.727) by this new sample definition.

In an unreported table, we also explore whether the 2SLS effect of peer migration is non-linear. In particular, we first regress self migration on all the control variables in Specification 1, the residual from this regression is refereed to as resid1. We then regress each instrument on the same control variables and name its residual as resid2, resid3, and resid4. In the third step, we regress resid1 on a quadratic function of resid2, resid3, and resid4 respectively. The results suggest that all three instruments affect one's migration decision non-linearly. The effect from the percent of neighbors having multiple birth in the firstborn is concave and its positive sign disappears when the percent of multiples reaches 3.4%. This is very high considering the average is 0.7% in our sample. The effects from the percentages of female and male labors residing in the households with a girl firstborn are always positive and slightly convex, starting from nearly 0.8.

5. Mechanisms of Social Interactions

The instrumental variable results presented above help identify social interactions from omitted individual or village-level variables, but they do not identify the mechanisms underlying these social interactions. This goal of this section is empirically identifying three types of social interactions.

The first is the social network effects as argued in Calvo-Armengol and Jackson (2004) and Carrington et al. (1996): migrants that belong to the same social network may help each other reduce migration cost and identify job opportunities at the destination. For example, if a young female migrant performs well in a manufacturing factory, the employer may ask her to help find new employees like her. Since most rural migrants have limited social networks in the city, most likely she will pass this information to her peers at the rural home. Even if the factory that the migrant is working for does not need new employees, she may hear about job opportunities in similar factories in the same city. Such information is useful for peers to consider when they decide whether and where to migrate. If peers migrate together to one city, they often share housing, which makes it easy to exchange job information during job search.

Social network theories imply that migrants from the same village should be more likely to cluster in the same destination and same occupation. This statement also implies that the village level cluster must differ substantially across villages, even though these villages are observationally similar. Table 3 already presents evidence in support of this prediction. To the extent that one is most familiar

with job opportunities that are specific to one's own age, gender and education group, or one is more willing to share job information with relatives, the social network effects should be the strongest among villagers of same age, gender, education and surname. The strength of the social network effects also depends on alternative channels of information, which predicts that they should be higher in the remote villages that have limited access to road or mass media.

The second type of social interactions is peer pressure at the origin. Suppose inside a village, migration is viewed as a positive signal of ability. Observing more peers migrating out of the village could make a non-migrant feel that he is inferior. If such peer pressure exists, we should observe stronger peer effects within similar age, gender and education, but not necessarily within the same destination for the same occupation.[17]

Unlike the above two mechanisms which both predict a positive correlation between self and peer migration, the third type of social interaction could suggests a negative correlation: for example, the more people migrate out of the village, the more agricultural resources and opportunities will be left for those that stay and therefore increase the opportunity cost of migration for these people. The positive λ_{2SLS} found in our main specification suggests that this general equilibrium effect does not dominate the social network effects or peer pressure at the origin. That being said, we can test the general equilibrium effects directly by looking at whether a non-migrant household uses more land if it is located in a high-migration village. Like peer pressure at the origin, general equilibrium effects focus on omitted variables at the origin and therefore do not predict migration cluster by destination and occupation.

Empirical Evidence Our empirical detection of social interaction mechanisms starts with two revised specifications. First, suppose we classify adults into nine age groups (17-20, 21-25, 26-30, 31-35, 36-40, 41-45, 46-50, 51-55, 56-60) and individual *i* belongs to age group *a*. The following specification examines how the percent of migration of age group 1 to 9 affects individual *i*'s own migration decision conditional on the sample of age group *a*:

$$y_{i|a} = \alpha_k + \beta x_i + \gamma x_h + \delta x_v + \lambda_1 \overline{y}_{-i,a_1|v} + \lambda_2 \overline{y}_{-i,a_2|v} + ... + \lambda_9 \overline{y}_{-i,a_9|v} + \varepsilon_i. \quad (2)$$

The percent of peer migration in group *a* ($\overline{y}_{-i,a|v}$) is instrumented based on the percent of female firstborns and the likelihood of multiple birth within the adults of group *a* in the same village.

Following Specification (2), Tables 8, 9, 10 and 11 report the 2SLS regression results for adults grouped by age, gender, surname and education. A general patterns standing out of these tables is that the effects of peer migration are the strongest for the adults within similar age, gender, education and

[17] The peer pressure could be related to destination or occupation if going to a specific destination, say Beijing, has a positive signaling value in the eyes of peers. In the data, destination- or occupation-specific peer pressure is not distinguishable from the social network effects, not only because they are observational equivalent, but also because such peer pressure is likely to rely on the help from earlier migrants to result in clustered migration.

surname. This finding is consistent with social network effects or peer pressure but not with the general equilibrium effects. Social interactions across different demographic groups are occasionally significant and asymmetric: for example, males have significantly positive peer effects on females and villagers aged 36-40 have significantly positive effects on those of 26-35, but the reverse effects are close to zero. These findings could reflect some strength of social network effects, as males (or the old) are more able to help females (the young) in a stranger community than vice versa. One exception is that migrants between 17 and 20 have some positive influence on older migrants and the effect declines with the age gap. This is probably because 17-20 is the prime age of migration and this group is more acceptable to new information.

Specification 2 is applicable to the sub-samples grouped by age, gender, education and surname, but not by destination and occupation. This is because destination and occupation are choices made by migrants conditional on migration. To detect social interactions by destination and occupation, we return to the full sample. Suppose there are *n* potential destinations and y_i^d is denoted one if individual *i* migrates to destination *d,* 0 otherwise. We then regress y_i^d on the percent of same-village adults migrating to destinations 1, 2, … n. This specification can be written as:

$$y_i^d = \alpha_k + \beta x_i + \gamma x_h + \delta x_v + \lambda_1 \overline{y}_{-i|v}^{d_1} + \lambda_2 \overline{y}_{-i|v}^{d_2} + ... + \lambda_n \overline{y}_{-i|v}^{d_n} + \varepsilon_i. \qquad (3)$$

The coefficients, $\{\lambda_1, \lambda_2, ...\lambda_n\}$, capture the correlation between self destination and peer destination, which could be driven by social interactions, or a destination-specific omitted variable that affects both self and peers. Unfortunately, our instruments are only relevant for whether peers migrate or not, not where to migrate or what to do after migration. To identify social interactions by destination or occupation, we must find additional instruments for each destination or each occupation.

We compute the travel distance from village *v* to destination d_j by summing up the distance of the village to the nearest bus/rail/dock station, the distance from the station to the township it belongs to, and the distance from the township to the destination. For destinations within the sampled province, we define the distance as distance from village *v* to the provincial capital. For destinations that fall in the residual category of "others", we compute the distance from village *v* to the biggest city of an adjacent province. Based on the distance variables, we define the instruments for $\overline{y}_{-i|v}^{d_j}$ as the distance from *v* to d_j times the three instruments used in Specification 1. Because individual *i*'s decision to migrate to d_j will take into account the distance to all alternative destinations, the 2SLS regression of $y_{i|v}^{d_j}$ on $\left\{\overline{y}_{-i|v}^{d_1}, ...\overline{y}_{-i|v}^{d_j}, ...\overline{y}_{-i|v}^{d_J}\right\}$ also controls for the distances from *v* to $\left\{d_1, ..., d_j, ..., d_J\right\}$ directly. Since we run a regression of $y_{i|v}^{d_j}$ for

each destination separately with county fixed effects, any unobserved correlation between the origin county and the destination is already accounted for.

It is more difficult to construct occupation-specific instruments because we know nothing about the employers of migrants. However, there are natural demographic differences across occupation: most construction workers are male, most service industry workers are female, and manufacturing jobs usually requires more skills than construction and service jobs. All these jobs prefer young to old. In light of these variations, we first compute the percent female, the percent of each age group (16-22,23-29, 30-39, 40+), and the percent of each education group (6 years of schooling, 7-9, 10-12, 13+) for each village. We then interact them with the three instruments used before as the IVs for $\overline{y}^{o_1}_{-i|v}$(manufacturing), $\overline{y}^{o_1}_{-i|v}$(service), $\overline{y}^{o_1}_{-i|v}$ (construction), and $\overline{y}^{o_1}_{-i|v}$(other). Like the destination regressions, the 2SLS regression specific to each occupation controls for the percent of female, percent of age groups, and percent of education groups at the village level.

Tables 12 and 13 report the 2SLS regression of Specification 3 for migration choice of destination and occupation. It is apparent that migrants from the same village are highly clustered by destination and occupation. All the coefficients on the diagonal (indicating the same destination or the same occupation) are positive, significant, and close to one. Two of them are even slightly bigger than one (destinations A and B), but t-test suggests that neither of them is statistically different from one. In contrast, most off-diagonal coefficients (indicating peer effects across destination and occupation) are insignificant, many are even significantly negative. Overall, the within-village cluster by occupation is most consistent with sharing job information at or about the destination. The cluster by destination is consistent with both the reduction of moving cost and the sharing of job information.

To address whether migrants leaves more agricultural resources and opportunities for the non-migrants in the same village, Table 14 regresses land in use of each non-migrant household (i.e. no adult migrates in the household) on the percent of same-village adults that migrate in 2006. We use the same instruments (all three) for peer migration as in Specification 1. The OLS results confirm a positive correlation between peer migration and the land use of non-migrants, but this correlation is no longer significant once we use instruments to control for omitted individual or village-level characteristics. In unreported tables, we try the same specification on other agricultural activities such as short-run employment for agricultural labor, fertilizer use, and the adoption of agricultural technology. Results are similar to that of land use: most show significant correlations with peer migration in OLS, but the significance disappears when we use instruments. These results suggest that eve if the general equilibrium effects exist, they are likely to reflect omitted characteristics, or being absorbed by remaining members of

the migrating households and therefore do not cause significant spillovers on other households in the same village.

Heterogeneous effects of peer migration So far the strong cluster of migrants by destination and occupation suggests that the most likely mechanism is the social network effects: peer migrants may help each other reduce moving costs and locate job opportunities at the same destination. Recall that the social network theory also implies heterogeneous strength of networks: the origins that are more difficult to travel from or have less information about the outside world should rely more on social networks. Table 15 includes the interactions of peer migration and distance from the respondent's residential village to the nearest bus/rail/dock station, the center of the county, the center of the studied area, and the provincial capital. In the last column we also interact peer migration with whether the respondent resides in a village that has access to TV signals. All regressions use the same three instruments for peer migration as in Specification 1. Results suggest that longer distance to the nearest station and county center does not imply larger effects of peer migration. However, longer distances to the area center and provincial center do have a positive and statistically significant effect on the strength of peer influence. This result is sensible because the area we study is not large and provincial capital is the most important stop if one wants to take railway or bus to other provinces. In comparison, the strength of peer influence is not sensitive to TV access, which confirms the fact that there is little job information via formal channels.

Alternative explanation The strong clustering by age, gender, education, destination and occupation could also be driven by local governments organizing group migration to a specific destination or by far-away employers recruiting a large number of workers from the same origin. While we cannot rule out these organized efforts, they are unlikely the driving force in our data for the following reasons.

First of all, both the 2004 shortage of migrating labor and the 2006 nationwide promotion of migrant labor markets, suggest that most organized efforts, if they exist, took place after 2004. In fact, almost all the major events we can find in the government documents of the data area in terms of organized migration happened in October 2006, as a response to the central government policy. According to the local newspaper, 65% of the migrants that worked away from home in 2006 have migrated out of the area by 2004, and 88.5% have migrated by 2005. Among the people that started to migrate in 2006, only 9.8% were organized by the Department of Labor in the local governments. These numbers suggest that the majority of the migrants observed in our 2006 cross-sectional data did not migrate because of government-organized migration or recruiting.

Second, for the organized migration or group recruiting to explain our key empirical findings, it must be organized at the village level because we already control for county fixed effects and results are robust if we further control for county fixed effects. All the anecdotes we can find in the mass media regarding organized migration or organized recruiting cover administrative units at or above the township

level. For example[18] in 2006, a vocational high school of a sampled county has signed a two-year contract with a city of destination A to train 1,500 adults per year. Graduates of the training program are guaranteed to work for an electronics factory or a shoe factory in destination B. In 2007, the Chairman of a large employer visited the sampled area and negotiated with the area government for a group labor contract of 2,000 migrants. A large town in the sampled area has actively searched for job opportunities since 1984 and the total number of migrants from this particular town has exceeded 21,000 by the end of 2008. All these activities, if equally effective for the whole administrative unit (area, county, or town), should already be controlled for in our fixed effects.

The remaining question is whether organized migration or recruiting occurs at the village level. If such village-level activities are correlated with whether a village has more firstborns being girls or more births being multiples, their effects will survive the instruments. In theory, this possibility is not zero: for example, if a village has more girl firstborns and therefore has a larger family size on average, the village head may face greater land-population pressure hence is more motivated to search for migration opportunities. Since all the firstborns computed in our sample are under age 16 at the time of the survey, the above story will only hold in reality if the village leader is sophisticated and forward-looking enough to predict the land-population pressure in the future.

This argument leads to two empirical tests: in the first test, we control for the demographics (gender, education, and military experience) of village cadres[19], which hopefully capture some unobserved village-level activities. As shown in Table 15, adding cadre demographics generates no change in our main results: the 2SLS coefficient for same-village peer migration only changes slight (0.735 versus 0.727) and remains highly significant. In the mean time, all the cadre demographics are statistically insignificant from zero. In the second test, we reconstruct all the IVs based on children of age 0-12 instead of 0-16. Again, results barely change: the 2SLS coefficient for same-village peer migration is 0.635, which is similar to the key results cited above (0.727).

To summarize, both anecdotes and empirical analysis lead us to believe that organized migration or organized recruiting is *not* the main reason driving the clustered pattern of migration. The most likely explanation is people of the same village share job information and help each other reduce moving costs at the destination. This conclusion is consistent with the importance of social networks as cited in a number of individual surveys conducted within China (Section 3).

6. Potential Implications of Peer Migration

[18] To protect data confidentially, we cannot provide precise citations for these anecdotes. They are available upon referee request.

[19] Village cadres refer to the village head and the Communist Party leader of the village.

What does peer migration imply for the migration rate in the long run? Typically, social network effects are assumed symmetric between self and peers. This symmetric assumption is unlikely to hold in the context of migration: almost all the empirical studies on migrant networks emphasize that previous migrants can influence non-migrants to move away; but once a migrant has moved, his or her future migration decision is unlikely to be affected by those that stay at the origin, unless the migrant returns permanently and reconsiders migration next year. Given the fact that only 7-8% of Chinese migrants return home the year after moving (Sheng 2008, Zhao 2002), we suspect the social network effects identified in our study is mostly driven by previous migrants affecting new migrants but not vice versa.

Unfortunately, we cannot test this conjecture because our data is only cross-sectional. Rather, we assume that the social network effects found in the study is restricted to the single direction from previous migrants to non-migrants. Under this assumption, social network effects imply that supporting a few rural residents to migrate to the city could lead their neighbors to do the same thing, every round of migration embodies a larger number of new migrants, these new migrants will influence the migration of the remaining villagers in the next round.

To illustrate the strength of this snowball effect, we consider a village of population one. For simplicity, the population is assumed to be homogenous (say young adults subject to the risk of migration) and never ages. Suppose the government subsidizes 1% of the population to migrate to the city in year one. Our peer effect estimate suggests that every 1% increase in peer migration will increase the remaining population's probability of migration by 0.727%. Assuming this peer influence does not change over time, we simulate the percentage of migrating population for the next thirty year in three scenarios: first, we assume every year 8% of the existing migrants will return to the rural area and these people will be subject to the risk of migration as much as the non-migrant population. In the second scenario, the likelihood of returning is still 8% but only half of the returning migrants will be subject to the risk of migration next year. In comparison, the third scenario assumes all the returning migrants will never migrate again.

Figure 6 shows the simulated migration path for all three scenarios. As we expect, scenario #1 will converge to a steady province with 88.6% of migrants, at which time the percentage of new migrants is equal to the percentage of returning migrants. In the other two scenarios, the returning migrants' inertia to migrate in the future becomes a greater countervailing force against the social multiplier effect. As a result, the percentage of migrating population reaches the maximum of 76% in year 14 for scenario #2 (68% in year 13 for scenario 3) and gradually declines afterwards. While each of these three scenarios are counterfactual (and inconsistent with the reality because we do not account for aging, marriage, childbearing and other life events), they suggest that a small initiation of migration could lead to a large wave of migration in the next few decades.

In addition to the snowball effect, peer migration could have large, persistent, and sometimes alarming implications for other socio-economic issues. For example, the railway traffic in the 40-day rush around the Spring Festival sets a new record every year: the 2009 traffic is 10.6% higher than 2008, and 2008 is 11% higher than 2007.[20] Although the central government has invested a lot to enhance the supply of railway service, it is difficult to catch up with the soaring demand. Railway tickets are extremely hard to get during the rush days, and many railway stations, especially those in large cities, have 24-hour police in order to reduce crime and accidents. The clustered pattern of migration found in our data also implies that the traffic between a migration origin and a migration destination is likely clustered, which could create serious congestion on specific routes even if the overall market is not crowded.

Another implication of peer migration is significant demographic change in high-migration villages. Since migrants are more likely to be male, young and better-educated, and the peer effects are the strongest among similar age and gender, the remaining population of high migration villages is likely to be concentrated in children, women and the elderly. To support this argument, we group the 3,950 villages in our sample according to whether its migrants-to-adults ratio falls in the brackets: 0-10%, 10-20%, 20-30%, or 30% and above. For each of the four village groups, we plot the headcount of migrants and non-migrants separately at every integer age (Figures 7-11). In all four figures, the age distribution has small spikes every 2-3 years. One potential reason is some rural households tend to calculate age by lunar calendar[21], but this is unlikely the main driver because the birth rate recorded by the local government by the regular calendar varies greatly from year to year.[22]

Comparing the four figures, we conclude that the overall population structure is similar across the four groups but because of migration, the villages that have the highest migration percentage (>30%) are significantly short of young and middle age adults. The difference in the percent of female is less striking: in the highest-migration villages, the percent of female adults in the remaining non-migrant population is 51%, as compared to 47.4% in the lowest migration villages. These migration-generated demographic changes, especially those in age composition, could have profound impact on agricultural productivity, child development, education, and elderly care.

Lastly, there is no doubt that peer migration clustered by destination and occupation establishes a strong employment link between origin and destination. This link could affect the vulnerability of the macro economy. Take the on-going economic recession as an example. Since a large fraction of rural migrants are concentrated in export-oriented manufacturing, the reduced international demand in 2008

[20] 2008 data: http://news.xinhuanet.com/video/2008-03/03/content_7704253.htm. 2007 data: http://info.cec-ceda.org.cn/jtwl/pages/20070207_41568_7_3.html。

[21] Some rural households count age as the number of lunar years that a person's life has covered regardless of the month of birth. For example, if a person was born in lunar December, she could be two year old in the lunar count even before she reaches her first birthday.

[22] The actual birth rate in the surveyed area ranges from 1.28% (1996, 2006) to 2.00% (2003) and 2.23% (2001).

and the subsequent return of unemployed labor has created serious problems for inland provinces.[23] This problem could be worsened by clustered migration because it reduces the origin area's ability to diversify the macroeconomic risk.

7. Conclusion

The unprecedented labor migration in China provides an excellent opportunity to deepen our understanding of whether, why and where laborers migrate. Using whether peers have a girl firstborn and whether peers have multiple birth(s) as instruments, we find a large, positive, and significant effects of social interactions within a village. The cluster pattern of migration suggests that peer migrants help each other reduce moving costs and locate job opportunities at the same destination.

The social network effects found in our study implies that a policy that subsidizes a small fraction of the rural population to migrate could have a large and persistent effect on subsequent migrations from the same village. Interestingly, not only does the snowball effect help transfer agricultural labor to non-agricultural activities, it could also create a number of socio-economic implications including a dramatic demographic change in the high migration villages, transportation congestion, and increased vulnerability to potential macroeconomic shocks. Evaluating the impact of migration on these social-economic issues will be a promising direction for future research.

[23] http://news.xinhuanet.com/newscenter/2009-03/02/content_10928532.htm.

References

Aizer, Anna and Janet Currie (2004) "Networks Or Neighborhoods? Correlations In The Use Of Publicly-Funded Maternity Care In California," *Journal of Public Economics* 88(12): 2573-2585.

Andrews, Donald W.K. and James H. Stock (2007) "Inference with Weak Instruments," *Advances in Economics and Econometrics, Theory and Applications: Ninth World Congress of the Econometric Society,* Vol. III, ed. by R. Blundell, W. K. Newey, and T. Persson. Cambridge, UK: Cambridge University Press, 2007.

Angrist, Joshua D and Evans William N (1998) "Children and Their Parents' Labour Supply: Evidence from Exogenous Variation in Family Size" *American Economic Review*, vol. 88(3) : 450-77.

Angrist, Joshua D. and Kevin Lang (2004) "Does School Integration Generate Peer Effects? Evidence from Boston's Metco Program" *American Economic Review*, Vol. 94, No. 5 (Dec., 2004), pp. 1613-1634

Bartel, Ann P. (1989) "Where Do the New U.S. Immigrants Live?" Journal of Labor Economics 7(4): 371-391.

Bayer, Patrick; Stephen L. Ross and Giorgio Topa (2008) "Place of Work and Place of Residence: Informal Hiring Networks and Labor Market Outcomes," *Journal of Political Economy,* 116(6), pages 1150-1196, December.

Becker, Gary (1975) *Human Capital: A Theoretical and Empirical Analysis* (2nd Edition), New York: National Bureau of Economic Research.

Bertrand, Marianne; Erzo F. P. Luttmer and Sendhil Mullainathan (2000) "Network Effects and Welfare Cultures," *Quarterly Journal of Economics* 115:3, pp. 1019-55.

Borjas, George J. (1994) "The Economics of Immigration," *Journal of Economic Literature* 32(4): 1667-1717.

Caces, F; F Arnold; J.T. Fawcett and R.W.Gardner(1985) "Shadow Household and Competing Auspices: Migration Behavior in Philippines*" Journal of Development Economics*, 17:5-25.

Cai, Fang (1996) "An economic analysis for labor migration and mobility" *Social Sciences in China* (Spring): 120–35.

Cai, Fang; Albert Park; and Yaohui Zhao "The Chinese Labor Market in the Reform Era" forthcoming in Loren Brandt and Thomas Rawski, eds., *China's Great Economic Transformation* (Cambridge University Press).

Cai, Fang and Wang Dewen (2003) "Migration As Marketization: What Can We Learn from China's 2000Census Data?" *The China Review*, Vol. 3, No. 2 : 73–93.

Cai, Yong and William Lavely (2003) "China's Missing Girls: Numerical Estimates and Effects on Population Growth" *The China Review*, Vol. 3, No. 2: 13–29

Calvo-Armengol, Tony and Matthew Jackson (2004) "The Effects of Social Networks on Employment and Inequality" *American Economic Review* 94(3): 426-454.

Carrington, W.J; E. Detragiache and T. Vishwanath (1996) “Migration with Endogenous Moving Costs” American Economic Review, 86(4): 909-30.

Chen, Jinyong (2006) “Reform of Hukou Policy and Rural-urban Migration in China” in Fang Cai and Zansheng Bai *edited Labor Migration in Transition China* 2006, Social Science Academy Press (China).

Chen, Jun(2005)” On the Problem of Information Shortage Encountered by Peasant-workers in Job Hunting”, *Hunan Social Sciences*, 2005(5): 83-85.

Du, Yang, Albert Park, and Sangui Wang(2005) “Migration and rural poverty in China”, *Journal of Comparative Economics*, 33:4 (2005): 688-709.

Duflo, Esther and Emmanuel Saez (2002): “The Role of Information and Social Interactions in Retirement Plan Decisions: Evidence from a Randomized Experiment,” *The Quarterly Journal of Economics*, Vol. 118, No. 3 (Aug., 2003), pp. 815-842

Du, Yin (2000) “Rural Labor Migration in Contemporary China: An Analysis of Its Features and the Macro context” in West, Loraine & Zhao, Yao Hui edited *Rural Labor Flows in China*, Institute of East Asian Studies, University of California.

Ebenstein, Avraham (2009) ”Estimating a Dynamic Model of Sex Selection in China” March 2009, job market paper.

Fei, John C. H and Ranis Gustav (1964) *Development of the Labor Surplus Economy: Theory and Policy*, New Haven, CT: Yale University Press.

Huang, Ping and Frank N. Pieke (2003) “China Migration Country Study” Paper presented at the Regional Conference on Migration, Development and Pro-Poor Policy Choices in Asia, Dhaka, June 21–24, 2003.

Ioannides, Yannis and Linda Datcher Loury (2004) "Job Information Networks, Neighborhood Effects and Inequality" *Journal of Economic Literature* 42(4): 1056-1093).

James,W.H (1987) “The Human Sex ratio, Part1: A Review of the Literature” *Human Biology* 59:721–752.

Jian, Yulan(2005)” China's rural labor force transfer of the status quo and countermeasures”, *The Rural Economics*, 2005(5): 110-112.

Kling, Jeffrey R; Jeffrey B. Liebman and Lawrence F. Katz (2007) “Experimental Analysis of Neighborhood Effects” *Econometrica* 75(1) : 83–119.

Knight, John; Song Lina and Jia Huaibin (1999) “Chinese Rural Migrants in Enterprises: Three Perspectives” *Journal of Development Studies* 35: 73-104

LaLonde, R. J., and R. H. Topel (1991) "Labor Market Adjustments to Increased Immigration," in J. M. Abowd and R. B. Freeman, eds., *Immigration, Trade, and the Labor Market*, Chicago: University of Chicago Press.

Lee, Everett S (1966) “A theory of migration” *Demography* 3: 47-57.

Lewis, W. Arthur (1954) “Economic Development with Unlimited Supplies of Labour “ *Manchester School of Economic and Social Studies* 22:139-91.

Li, Hongbin; Junsen Zhang and Yi Zhu (2006) “The Effect of the One-Child Policy on Fertility in China: Identification Based on Differences-in-Differences” *Working Paper*, The Chinese University of Hong Kong.

Liang, Qiusheng and Che-Fu Lee (2006) “Fertility and Population Policy: An Overview” in *Fertility, Family Planning, and Population Policy in China*, edited by D.L. Poston, Jr; Che-Fu Lee; Chiung-Fang Chang; Sherry L. McKibben and Carol S. Walther, published by Routledge Taylor & Francis Group.

Lin, Justin Yifu; Gewei Wang and Yaohui Zhao “Regional Inequality and Regional Transfers in China” in Cai Fang and Bai Nansheng edited *Labor Migration in Transition China* 2006, Social Science Academy Press (China).

Lowry, I (1966) *Migration and metropolitan growth: Two Analytical Models.* San Francisco: Chandler.

Lucas, Robert E.B. (1987) “Emigration to South Africa’s mines” *American Economic Review* 77: 313-330.

Lucas, Robert E.B. (1997) “Internal Migration in Developing Countries” Chapter 13, *Handbook of Population and Family Economics*, edited by M.R. Rosenzweig and O. Stark 1997, Elsevier.

Manski, Charles F (1993) “Identification of Endogenous Social Effects: The Reflection Problem “ *Review of Economic Studies*, 60:531-542.

Maurin, Eric; Hulie Moschion (2009) “ The Social Multipiler and Labor Market Participation of Mother” *American Economic Journal: Applied Economics* 2009 1(1): 251-272.

Mckenzie, David and Hillel Rapoport (2007) “Network Effects and the Dynamics of Migration and Inequality: Theory and Evidence from Mexico” *Journal of Development Economics* 84: 1-24.

Mckenzie, David and Nicole Hilderbrandt (2005) “The Effects of Migration on Child Health in Mexico” *Economia* 6(1): 257-289.

Men, Kepei and Wei Zeng (2003 “Prediction of China Population Over the Next 50 Years”, *The Journal of Quantitative & Technical Economics (in Chinese)*: F224.

Munshi, Kaivan (2003) “Networks in the Modern Economy: Mexican Migrants in the U.S. Labor Market” *Quarterly Journal of Economics*, 118(2): 549-600.

Qian, Nancy (2009) “Quantity-Quality and the One Child Policy: the Only-Child Disadvantage in School Enrollment in Rural China” NBER *working paper* #14973.

Rosenzweig, Mark R and Kenneth L Wolpin (1980) “Testing the Quantity-Quality Fertility Model: The Use of Twins as a Natural Experiment” *Econometrica*, January 1980a, 48(1) : 227-40.

Rosenzweig, Mark (1988) “Labor Markets in Low-income Countries” Chapter 15 in the *Handbook of Deveopment Economics*, Volume 1, edited by H. Chenery and T. N. Srinivasan, Elsevier.

Sen, Amartya (1990) “More than 100 million women are missing” *New York Review of Books* 37(20): 61-66.

Sheng, Laiyun (2008) *Flows or Migration: the Economic Analysis China’s Rural Labor Flow*. Published by Shanghai Far East Publishers(China).

Sjaastad , Larry A (1962) “The costs and returns of human migration” *The Journal of Political Economy*, 1962, vol. 70, no. S5: 80-93.

Tan, Kejian (2003) “Value Difference of Sons and Daughters to Their Families in Poverty-stricken Areas”, *Population Research (in Chinese)*: 2003 27(2).

Taylor, J. Edward (1986) “Differential migration, networks, information and risk” in *Migration, Human Capital and Development*, edited by O. Stark, published by JAI Press, Greenwich, CT.

Wan, Chuan; Dezhu Wang and Bo Li (2004) “Debates on Temporary Residence Permit System and Necessity for Adopting It” *Chinese Journal of Population Science (in Chinese)*: C29.

Wang, Dewen; Wu Yaowu; and Cai Fang (2003) “Migration, unemployment, and urban labor market segregation in China’s economic transition” working paper, Beijing: Institute of Population and Labor Economics, Chinese Academy of Social Sciences.

White, Tyrene (1992) “Birth Planning between Plan and Market: The Impact of Reform on China’s One-Child Policy”, *China’s Economic Dilemmas in the 1990’s: The Problems of Reforms, Modernization, and Interdependence. Studies in Contemporary China.* Armonk, N.Y.U.S.C.J.E. Committee and London, Sharpe, 1992: 252-69.

Woodruff, Christopher and Rene Zenteno (2007) “Migration Networks and Microenterprises in Mexico” *Journal of Development Economics* 82: 509-528.

Young, Alwyn (2003) “Gold into Base Metals: Productivity Growth in the People’s Republic of China during the Reform Period” *Journal of Political Economy*, Vol 111, no. 6.

Zhang, Weiqing (1998) *Introduction to Family Planning in China (in Chinese)*, China Population Publishing House.

Zhao, Yaohui (1999a) “Leaving the Countryside: Rural-to-Urban Migration Decision in China” *American Economic Review Papers and Proceedings, May 1999,* 89(2): 281-286.

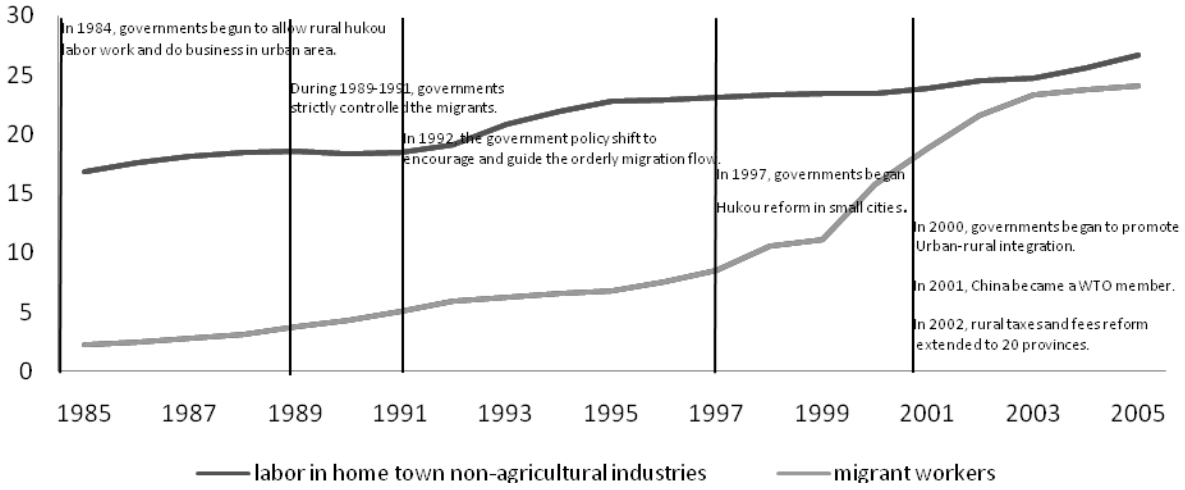
Zhao, Yaohui (1999b) “Labor Migration and Earnings Differences: the Case of Rural China” *Economic Development and Cultural Change* 47(4): 767-782.

Zhao, Yaohui (2002) “Cause and Consequences of Return Migration: Recent Evidence from China” Journal of Comparative Economics 30, 376-394.

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

Zhao, Zhong (2005) “Migration, labor market flexibility, and wage determination China: a review”, *The Developing Economies*, 43, 285-312.

Figure 1: Transfer of rural labor to non-agriculture activities, 1985-2004, all China



Data source：China Rural Statistical Yearbook 2006

Figure 2: Distribution of migrants by # of months away from home in 2006, study sample

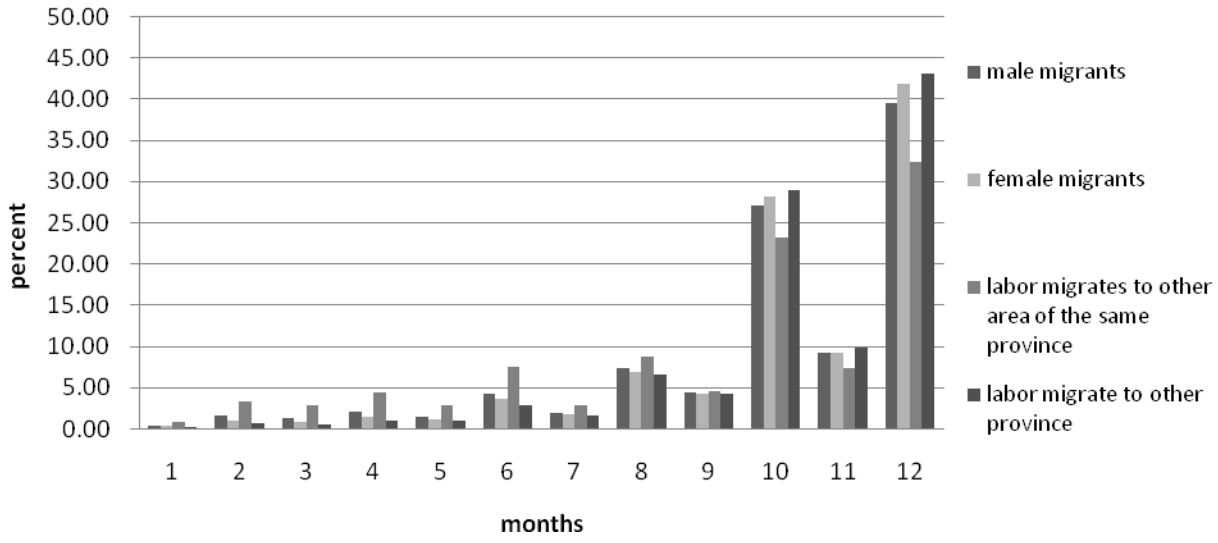


Figure 3: Percent of adults that migrate in 2006, by age and gender, study sample

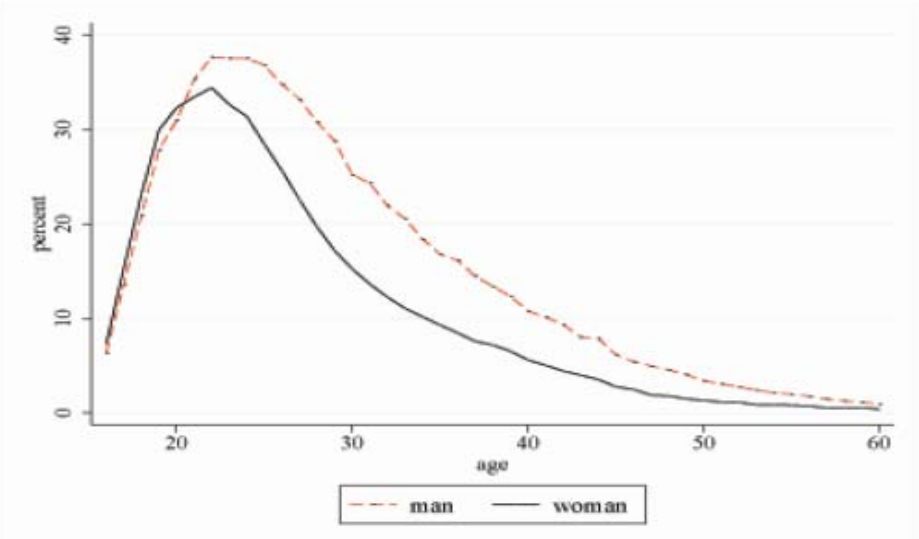


Figure 4: Histogram of migration percentage per village, raw data, study sample

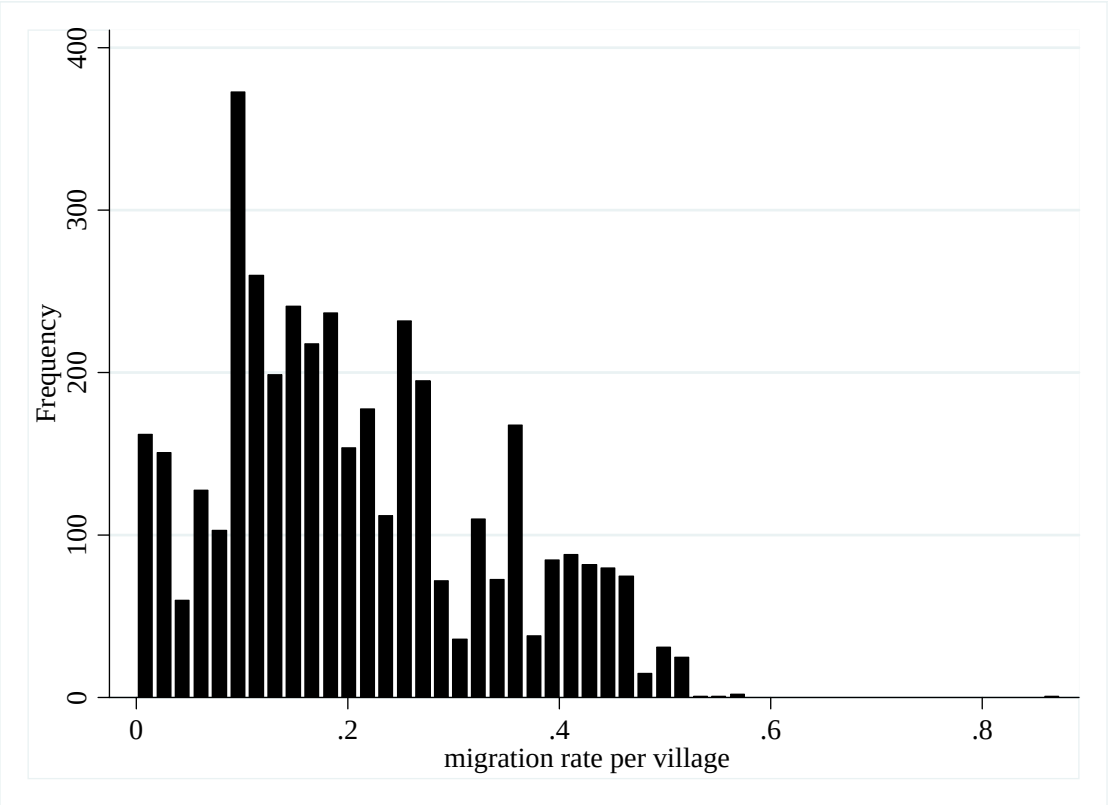


Figure 5: Histogram of unexplained migration percentage per village, after village-level regression, study sample

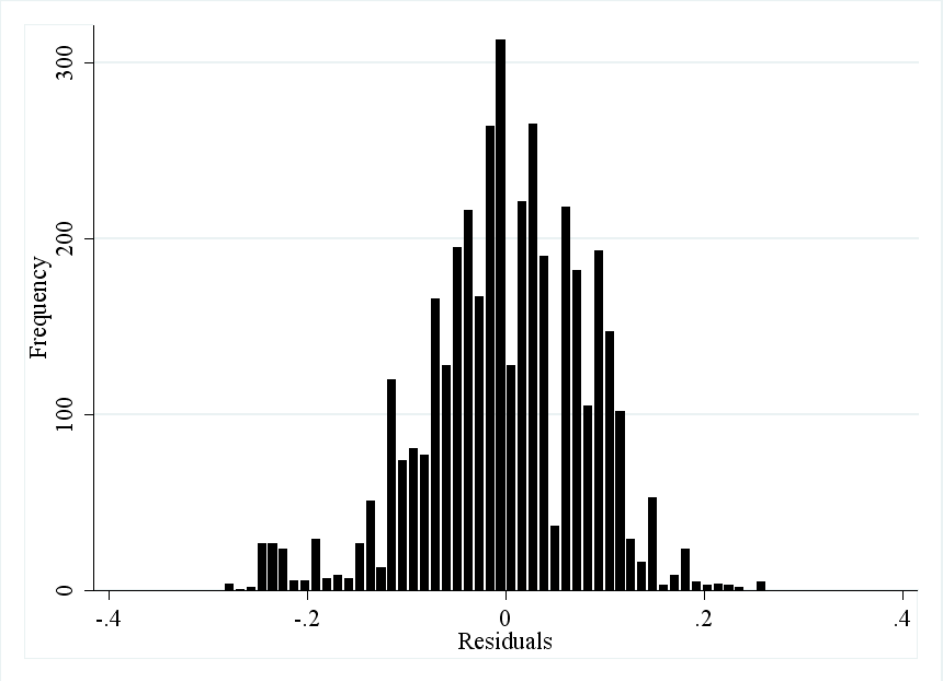


Figure 6: Simulated migration

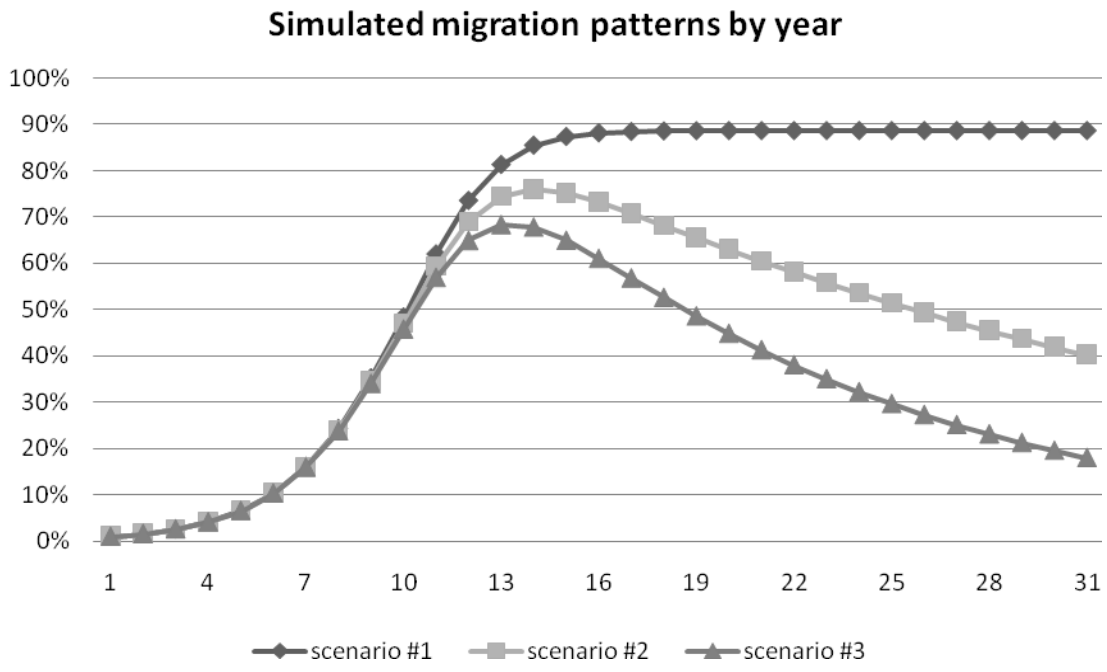


Figure 7: Age distribution in villages where 0-10% adults migrate

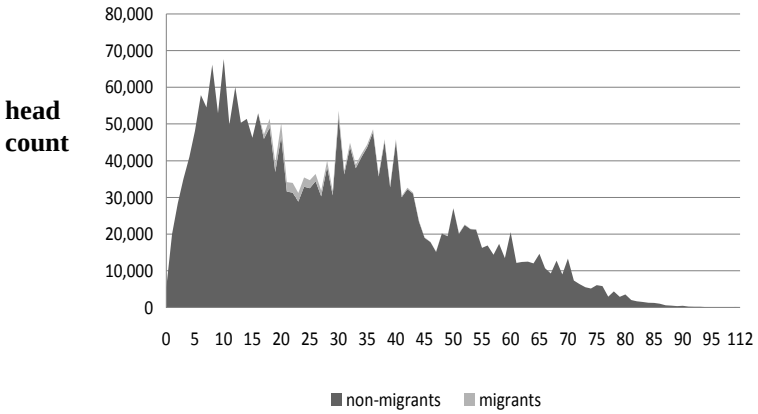


Figure 8: Age distribution in villages where 10-20% adults migrate

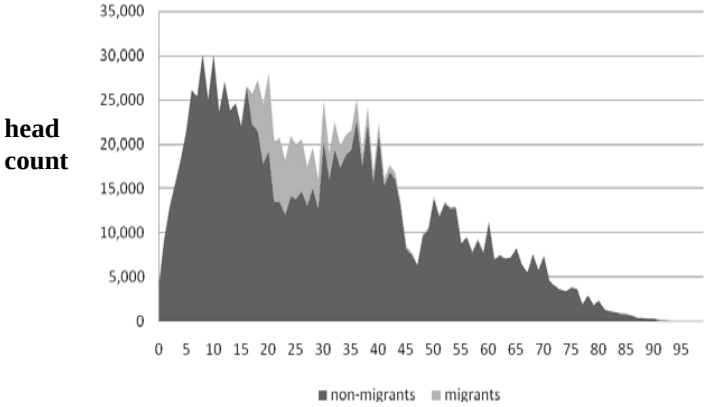


Figure 9: Age distribution in villages where 20-30% adults migrate

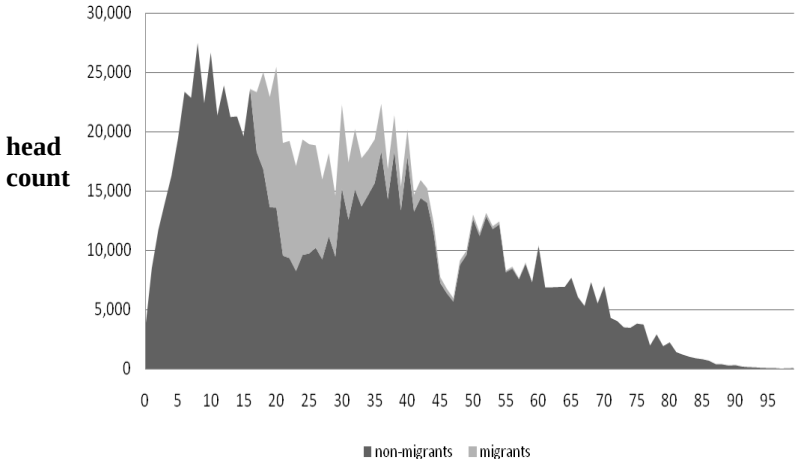


Figure 10: Age distribution in villages where 30+% adults migrate

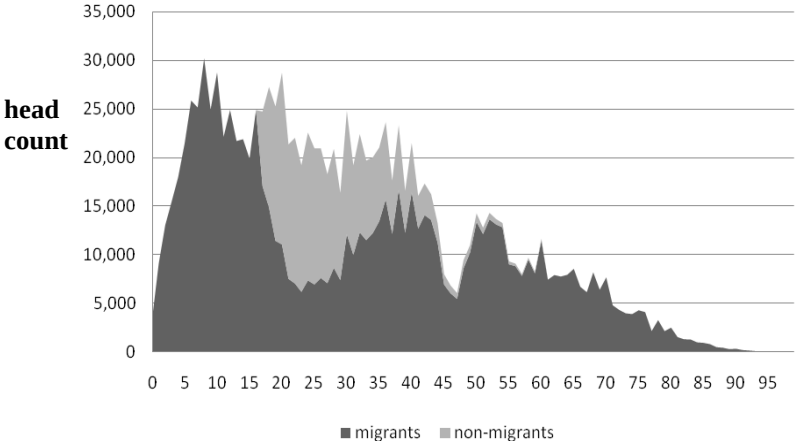


Table 1: Distribution of migrants by destination, study sample

Destination	% migrants	Per capita income of 2006 (relative to rural of the sampled area)		Railway hours to destination
		rural	urban	
Within province				
within the sampled area	10.20%	1.00	3.57	0
outside the sampled area	16.05%	0.97	4.56	3-4
Across province				
A	27.86%	3.59	8.40	26
B	20.18%	2.49	7.57	20
C	9.68%	2.37	6.71	38
D	4.85%	1.10	5.07	10.5
E	3.53%	2.85	6.83	36
F	1.92%	4.47	10.98	26.5
G	0.20%	1.41	5.57	6.5
H	0.63%	1.47	4.51	13

Table 2: Summary statistics by migration status, study sample

	All Adults age 17-60		Non-Migrants		Migrants	
	(1)		(2)		(3)	
	mean	std dev	mean	std dev	mean	std dev
Panel A: Individual Attributes						
age	35.63	(11.66)	37.34	(11.61)	27.36	(7.70)
years of schooling	6.44	(2.87)	6.15	(2.93)	7.84	(2.04)
female	0.47	(0.50)	0.48	(0.50)	0.38	(0.48)
being household head	0.35	(0.48)	0.38	(0.49)	0.17	(0.38)
whether my first single birth is girl[1]	0.49	(0.50)	0.49	(0.50)	0.49	(0.50)
whether my first birth is multiples[2]	0.01	(0.09)	0.01	(0.08)	0.01	(0.11)
minimum age of my own children under age 16[3]	7.58	(4.74)	7.61	(4.71)	7.39	(4.90)
maximal age of my own children under age 16[3]	10.29	(4.74)	10.40	(4.69)	9.65	(4.94)
Panel B: Household attributes						
household head age	45.39	(11.16)	44.70	(10.95)	48.73	(11.56)
household head being female	0.05	(0.22)	0.05	(0.22)	0.07	(0.26)
household head yr of schooling	6.26	(2.83)	6.27	(2.82)	6.24	(2.88)
# of HH members age 0-6	0.40	(0.71)	0.41	(0.71)	0.38	(0.69)
# of HH members age 7-16	0.87	(1.11)	0.91	(1.12)	0.66	(1.00)
# of HH members age 17-23	0.83	(1.02)	0.78	(0.99)	1.08	(1.11)
# of HH members age 24-44	1.59	(0.90)	1.56	(0.88)	1.72	(0.99)
# of HH members age 45-59	0.75	(0.88)	0.73	(0.88)	0.85	(0.89)
# of HH members age 60+	0.23	(0.54)	0.21	(0.51)	0.35	(0.65)
# of girls in the HH age 0-16	0.58	(0.81)	0.61	(0.82)	0.48	(0.76)
# of boys in the HH age 0-16	0.68	(0.81)	0.71	(0.82)	0.55	(0.77)
Has any boy in the HH? (age 0-16)	0.49	(0.50)	0.51	(0.50)	0.41	(0.49)
estimated house value (10,000 yuan)	2.36	(2.80)	2.37	(2.87)	2.28	(2.43)
have any outstanding loans	0.14	(0.34)	0.13	(0.34)	0.15	(0.36)
contract land (mu)[5]	3.45	(2.54)	3.44	(2.55)	3.48	(2.45)
land in use (mu)[5]	4.13	(3.04)	4.15	(3.04)	3.99	(3.06)
Prevalence of HH head's surname in the village	0.10	(0.12)	0.10	(0.12)	0.10	(0.11)
HH head's surname is the largest surname in the village	0.21	(0.41)	0.21	(0.41)	0.21	(0.41)
HH head's surname is the second largest in the village	0.13	(0.34)	0.13	(0.34)	0.13	(0.33)
HH head's surname is the third largest in the village	0.10	(0.29)	0.10	(0.29)	0.10	(0.29)
HH head's surname is below the third largest in the village	0.56	(0.50)	0.56	(0.50)	0.57	(0.50)

(to be continued)

Table 2 (continued)

	All Adults age 17-60		Non-Migrants		Migrants	
	(1)		(2)		(3)	
	mean	std dev	mean	std dev	mean	std dev
Panel C: Village attributes						
distance to the nearest bus, rail, or dock station (kilometer)	6.02	(8.23)	6.16	(8.37)	5.35	(7.45)
whether the village is a minority gathering	0.30	(0.46)	0.31	(0.46)	0.26	(0.44)
the village has a national poverty status	0.51	(0.50)	0.51	(0.50)	0.49	(0.50)
# of adults 17-60 in the village	1083.00	(545.67)	1083.73	(543.30)	1079.43	(556.97)
average land per adult (mu)[4]	1.63	(0.70)	1.64	(0.71)	1.59	(0.64)
whether regular water use is guaranteed	0.37	(0.48)	0.36	(0.48)	0.43	(0.49)
have organized production and sale of agriculture products	0.09	(0.28)	0.08	(0.28)	0.10	(0.29)
% of natural gathering groups that have access to electricity	0.96	(0.16)	0.96	(0.16)	0.97	(0.13)
% of natural gathering groups that have access to telephone	0.78	(0.34)	0.77	(0.35)	0.83	(0.30)
% of natural gathering groups that have access to TV signal	0.91	(0.26)	0.90	(0.27)	0.92	(0.24)
% of natural gathering groups that have access to drivable road	0.66	(0.31)	0.66	(0.32)	0.68	(0.30)
Panel D: attributes of adults in the same village						
% of adults in the village that are migrants (exclude all adults in self HH)	0.17	(0.14)	0.15	(0.13)	0.29	(0.12)
% of first-born children being girl (exclude self, single birth only)	0.49	(0.06)	0.49	(0.06)	0.49	(0.05)
% of first-born children being multiples (exclude self)	0.01	(0.01)	0.01	(0.01)	0.01	(0.01)
Observations	3327996		2759453		568543	

Notes: An individual is defined as migrant if s/he has been away from the village for the reason of work for more than 15 days in 2006. For data reasons as described in Section 2, (1) "whether own first child is girl" has 1834855 missing observations (55.1%), (2) "whether own first birth is multiples" has 1823197 missing observations (54.78%), (3) minimum and maximum child age has 1257900 missing observations each (37.8%). (4) One Chinese mu is equal to 666.7 square meters or 0.1647 acres.

Table 3: Within-village cluster of migrants, by destination, occupation and surname, (conditional on migrants only)

	Average per village	Average per township	Average per county	Whole area
% of migrants in the most common occupation	75.10%	51.93%	46.21%	46.68%
	(0.68%)	(1.00%)	(1.71%)	
% of migrants in the most common destination	63.80%	39.82%	31.66%	27.86%
	(0.76%)	(0.84%)	(4.54%)	
% in the most common occupation, conditional on migrants in the most common destination	83.80%	59.07%	54.27%	59.05%
	(0.59%)	(1.12%)	(5.90%)	
% in the most common occupation of each destination	95.20%	58.33%	49.83%	47.63%
	(0.34%)	(0.81%)	(3.26%)	
% in the most common occupation, conditional on migrants with the most common surname	82.60%	52.80%	45.77%	43.97%
	(0.60%)	(0.85%)	(3.26%)	
% in the most common destination, conditional on migrants with the most common surname	74.50%	41.31%	31.73%	26.80%
	(0.69%)	(1.02 %)	(4.77%)	
% in the most common occupation of each surname	90.10%	61.99%	48.95%	47.03%
	(0.47%)	(0.76 %)	(1.52 %)	
% in the most common destination of each surname	85.80%	51.73%	34.90%	28.67%
	(0.56%)	(0.75 %)	(3.57 %)	
N of observations	3950	250	8	1

Standard error in parentheses. The percentages are computed as follows: suppose 11 migrants of a village went to two destinations (A and B) for two occupations (X1 and X2). If 5 went to destination A with 3 in X1 and 2 in X2, and the other 6 went to B with 1 in X1 and 5 in X2, the % in the most common occupation is 7/11, the % in the most common destination is 6/11 , the % in most common occupation conditional on the most common destination is 5/6, and the % in the most common occupation of each destination is (3+5)/(5+6). Percentages by surname are computed similarly.

Table 4: Summary of adults by the number and gender of child birth, study sample

	# of observations		migrate or not	have second birth or not	have boy in second or later birth1	have girl in second or later birth1	Number of girls under age 16	Number of boys under age 16
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Panel A: All adults aged 17-60							
All adults	3327996		0.17				0.58	0.68
			(0.38)				(0.80)	(0.81)
Adults with kids	2081151		0.15				0.92	1.08
			(0.36)				(0.85)	(0.78)
Adults with clear firstborn definition	1504799		0.14	0.69	0.76	0.58	1.00	1.17
			(0.35)	(0.50)	(0.43)	(0.49)	(0.86)	(0.80)
	Panel B: Conditional on the first-born child being single birth							
firstborn is boy	758616	50.41%	0.14	0.68	0.64	0.64	0.55	1.52
			(0.35)	(0.46)	(0.48)	(0.48)	(0.70)	(0.67)
firstborn is girl	734495	48.81%	0.14	0.71	0.87	0.52	1.48	0.80
			(0.35)	(0.45)	(0.33)	(0.50)	(0.73)	(0.74)
	Panel C: Conditional on the first-born children being multiple birth							
firstborns are all girls	3123	0.21%	0.21	0.70	0.89	0.40	2.35	0.78
			(0.41)	(0.46)	(0.31)	(0.49)	(0.64)	(0.70)
firstborns are all boys	3989	0.27%	0.20	0.39	0.55	0.63	0.28	2.24
			(2.60)	(0.49)	(0.50)	(0.48)	(0.55)	(0.51)
firstborns have mixed gender	4576	0.31%	0.22	0.48	0.68	0.57	1.32	1.38
			(0.42)	(0.50)	(0.46)	(0.49)	(0.59)	(0.56)

Notes: Unit of analysis is adult labor aged 17-60 as defined in the study sample. (1) conditional on the families that have second birth.

Table 5 Test for the validity of instruments

	Dependent Variable					
	migrate or not			# of kids	Family size	Having at least one boy
	(1)	(2)	(3)	(4)	(5)	(6)
having a girl firstborn	0.00392***	0.00331***	0.0132***	0.248***	0.256***	-0.318***
	(0.000824)	(0.000823)	(0.000994)	(0.00388)	(0.00399)	(0.00164)
female* girl firstborn			-0.0201***			
			(0.000940)			
1 if firstborn info missing	0.0437***	0.0371***	0.0373***	-1.444***	-0.771***	-0.680***
	(0.00142)	(0.00136)	(0.00136)	(0.00450)	(0.00729)	(0.00153)
female		-0.0325***	-0.0281***	-0.133***	-0.785***	-0.0367***
		(0.000947)	(0.000929)	(0.00596)	(0.0100)	(0.00197)
age		-0.0212***	-0.0213***	0.247***	0.390***	0.0832***
		(0.000349)	(0.000349)	(0.00174)	(0.00233)	(0.000485)
age square		0.000159***	0.000160***	-0.00310***	-0.00430***	-0.00106***
		(3.53e-06)	(3.53e-06)	(2.06e-05)	(2.81e-05)	(5.67e-06)
distance to nearest station		0.00879***	0.00877***	-0.00256***	0.0338***	0.00191***
		(0.000273)	(0.000273)	(0.000693)	(0.00107)	(0.000180)
years of schooling		9.70e-05	9.81e-05	0.000858*	0.00145**	0.000160**
		(0.000225)	(0.000225)	(0.000442)	(0.000619)	(7.74e-05)
County dummy	control	control	control	control	control	control
Numbers of Observations	3,327,996	3,327,996	3,327,996	1,153,804	1,153,804	1,153,804
Level of Observations	individual	individual	individual	household	household	household
R square	0.0879	0.2016	0.2017	0.5298	0.1986	0.4667

Notes: Significance at 10% (*), 5% (**), 1% (***). Robust standard errors in parentheses. The error terms are clustered by village.

Table 6 OLS and 2SLS regressions on Specification 1, study sample

Panel A: First stage				
	Dependent Variables: peer migration ($y_{-i,v}$)			
	(1) OLS	(2) 2SLS	(3) 2SLS	(4) 2SLS
% of same-village adults having multiples in the first birth		0.346**		0.300**
		(0.152)		(0.152)
% of same-village female labors residing in households with a girl firstborn			0.600***	0.582***
			(0.125)	(0.125)
% of same-village males residing in households with a girl firstborn			-0.518***	-0.502***
			(0.118)	(0.118)
Number of observations		3327996	3327996	3327996
R square		0.196	0.196	0.197
Panel B: Second Stage				
	Dependent Variable: self migration ($y_{i,v}$)			
% of same-village adults migrating	0.930***	0.631***	0.768***	0.727***
	(0.00326)	(0.106)	(0.0407)	(0.0438)
Conditional LR test for weak IV		[0.571, 0.672]	[0.721,0.787]	[0.685,0.743]
female	-0.0485***	-0.0484***	-0.0484***	-0.0484***
	(0.00117)	(0.00119)	(0.00118)	(0.00118)
age	-0.0116***	-0.0117***	-0.0117***	-0.0117***
	(0.000269)	(0.000278)	(0.000272)	(0.000272)
age square	2.74e-05***	3.02e-05***	2.89e-05***	2.93e-05***
	(2.94e-06)	(3.10e-06)	(2.97e-06)	(2.97e-06)
years of schooling	0.00852***	0.00871***	0.00862***	0.00865***
	(0.000233)	(0.000242)	(0.000232)	(0.000232)
# of HH members aged 0-6	0.00642***	0.00764***	0.00708***	0.00725***
	(0.000944)	(0.00103)	(0.000957)	(0.000957)
# of HH members aged 7-16	0.00743***	0.00857***	0.00805***	0.00821***
	(0.000610)	(0.000697)	(0.000618)	(0.000614)
# of HH members aged 17-23	0.00374***	0.00565***	0.00478***	0.00504***
	(0.000397)	(0.000756)	(0.000466)	(0.000469)
# of HH members aged 24-44	0.0256***	0.0282***	0.0270***	0.0274***
	(0.000610)	(0.00101)	(0.000705)	(0.000688)
# of HH members aged 45-59	0.0325***	0.0347***	0.0337***	0.0340***
	(0.000670)	(0.000997)	(0.000739)	(0.000737)
# of HH members aged 60+	0.0170***	0.0206***	0.0189***	0.0194***
	(0.000716)	(0.00141)	(0.000878)	(0.000880)
having at least one boy in HH	-0.00598***	-0.00575***	-0.00585***	-0.00582***
	(0.000868)	(0.000896)	(0.000877)	(0.000881)
minimum age of own child	0.00309***	0.00310***	0.00309***	0.00310***
	(0.000128)	(0.000131)	(0.000129)	(0.000129)
having a girl firstborn	0.00218***	0.00191***	0.00203***	0.00200***

	(0.000710)	(0.000726)	(0.000716)	(0.000717)
first birth is multiples	0.00510	0.0124***	-0.0316***	-0.0305***
	()	(0.00467)	(0.00475)	(0.00457)
second birth is multiples	-0.0128***	-0.0110***	-0.0118***	-0.0116***
	(0.00351)	(0.00363)	(0.00355)	(0.00357)
Is household head	-0.0353***	-0.0353***	-0.0353***	-0.0353***
	(0.00121)	(0.00121)	(0.00121)	(0.00121)
age of household head	0.00341***	0.00360***	0.00351***	0.00354***
	(6.85e-05)	(9.31e-05)	(7.38e-05)	(7.34e-05)
household head is female	0.0226***	0.0244***	0.0236***	0.0238***
	(0.00145)	(0.00166)	(0.00151)	(0.00152)
years of schooling for household head	-0.00253***	-0.00191***	-0.00220***	-0.00211***
	(0.000189)	(0.000304)	(0.000217)	(0.000222)
estimated house value	-0.00281***	-0.00355***	-0.00321***	-0.00331***
	(0.000159)	(0.000329)	(0.000196)	(0.000207)
contract land (mu)	-0.00150***	-0.00262***	-0.00211***	-0.00226***
	(0.000182)	(0.000436)	(0.000244)	(0.000251)
Prevalence of HH head's surname in village	-0.00102	0.00148	0.000340	0.000681
	(0.00290)	(0.00440)	(0.00343)	(0.00366)
distance to nearest bus/rail/dock station	-7.96e-05**	-1.95e-05	-4.70e-05	-3.88e-05
	(3.18e-05)	(8.12e-05)	(5.19e-05)	(5.95e-05)
village is a minority gathering	0.00122*	-0.00414	-0.00169	-0.00242*
	(0.000639)	(0.00255)	(0.00127)	(0.00145)
village has national poverty status	-0.00157**	-0.00200	-0.00180*	-0.00186
	(0.000633)	(0.00157)	(0.00103)	(0.00118)
# of adults in the village	-0.00722***	-0.0124***	-0.0100***	-0.0107***
	(0.00104)	(0.00310)	(0.00186)	(0.00209)
arable land in the village (mu)	0.00322***	0.00677***	0.00515***	0.00564***
	(0.000794)	(0.00218)	(0.00133)	(0.00149)
regular water use is guaranteed	-0.00131**	-0.00356**	-0.00253**	-0.00284**
	(0.000607)	(0.00179)	(0.00104)	(0.00120)
village has organized production and sale of agricultural products	0.00241**	0.00308	0.00278*	0.00287
	(0.000985)	(0.00250)	(0.00165)	(0.00189)
% of natural gathering groups within the village having access to electricity	-0.00311*	0.00314	0.000281	0.00113
	(0.00176)	(0.00458)	(0.00287)	(0.00323)
% of natural gathering groups within the village having access to telephone	0.000538	0.0122***	0.00685***	0.00843***
	(0.000908)	(0.00455)	(0.00208)	(0.00228)
% of natural gathering groups within the village having access to TV signals	0.00247*	-0.00206	1.03e-05	-0.000608
	(0.00130)	(0.00324)	(0.00196)	(0.00222)
% of natural gathering groups within the village having access to drivable roads	0.00373***	0.00383*	0.00378**	0.00380**
	(0.000932)	(0.00225)	(0.00148)	(0.00170)
Observations	3327996	3327996	3327996	3327996
R-squared	0.276	0.267	0.273	0.272

Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. All regressions control for county fixed effects, registered township population, presence of highway exit, and # of township-village-enterprises in township. Errors are clustered by village.

Table 7 Robust checks on Specification 1

	Dependent variables: migration										
Sample	**Main Sample**									**Family with only two adults**	**Exclude minority gathering villages**
Specification	**Results from Table 6 Column 4 2SLS (1)**	**Redefine migration as 6+ months away 2SLS (2)**	**Use township FE instead of county FE 2SLS (3)**	**Exclude own household fertility outcomes from the right hand side 2SLS (4)**	**IV conditional on all-boy multiples only 2SLS (5)**	**IV conditional on girl firstborns with adult age at or below 35 2SLS (6)**	**IV conditional on same-village households with different surnames 2SLS (7)**	**include migration of other adults in different villages but same township 2SLS (8)**	**include migration of other adults in the same household 2SLS (9)**	**2SLS (10)**	**2SLS (11)**
% of same-village adults that migrate (exclude own household)	0.727***	0.674***	0.789***	0.741***	0.681***	0.622***	0.810***	0.782***	0.482***	0.604***	0.710***
	(0.0438)	(0.116)	(0.0757)	(0.0407)	(0.148)	(0.0825)	(0.0458)	(0.104)	(0.0396)	(0.0186)	(0.0217)
% of same-township adults that migrate (exclude own village)								0.0108			
								(0.103)			
% of same-household adults that migrate (exclude self)									0.144***		
									(0.002)		
Control Other Variables	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes		Yes	Yes
Numbers of Observations	3,327,996	3,327,996	3,327,996	3,327,996	3,327,996	3,327,996	3,327,996	3,327,996	3,327,996	2,138,948	2,330,227
R-squared	0.272	0.268	0.275	0.260	0.270	0.267	0.259	0.274	0.333	0.226	0.273

Notes: Significance at 10% (*), 5% (**), 1% (***). Robust standard errors in parentheses. All regressions control for county fixed effects (except for column 3) and all the other variables used in Table 6. Errors are clustered by village.

Table 8 2SLS regression by age cohort

	Dependent Variable: Migration					
sample	**17-20**	**21-25**	**26-30**	**31-35**	**36-40**	**41-45**
	(1)	(2)	(3)	(4)	(5)	(6)
% of same-village adults aged 17-20 that migrate	1.142***	0.697***	0.577***	0.287**	0.0989*	0.0366
	(0.0563)	(0.145)	(0.175)	(0.117)	(0.0532)	(0.0271)
% of same-village adults aged 21-25 that migrate	-0.224***	0.339**	-0.0586	0.0810	0.0422	0.0288
	(0.0513)	(0.143)	(0.170)	(0.113)	(0.0503)	(0.0258)
% of same-village adults aged 26-30 that migrate	-0.0572	-0.320**	-0.0311	-0.497***	-0.222***	-0.122***
	(0.0573)	(0.163)	(0.206)	(0.144)	(0.0626)	(0.0328)
% of same-village adults aged 31-35 that migrate	0.108	0.153	-0.245	0.423**	-0.141	0.0133
	(0.0832)	(0.229)	(0.274)	(0.207)	(0.0938)	(0.0442)
% of same-village adults aged 36-40 that migrate	0.238	0.417	0.993**	0.753**	1.305***	0.170**
	(0.148)	(0.389)	(0.481)	(0.339)	(0.150)	(0.0723)
% of same-village adults aged 41-45 that migrate	-0.213	-0.339	0.0118	0.168	-0.112	0.742***
	(0.187)	(0.496)	(0.556)	(0.382)	(0.179)	(0.0947)
Observations	341307	442035	472301	498009	506443	339076
R-squared	0.275	0.299	0.290	0.226	0.149	0.096

Notes: Significance at 10% (*), 5% (**), 1% (***). Robust standard errors in parentheses. All regressions include county fixed effects, all the control variables used in Table 6, and the percentages of same village adults aged 46-50, 51-55, 56-60 that migrate. Errors are clustered by village. Shaded cells refer to influence from peers of the same age group.

Table 9 2SLS regression by gender

	Dependent variable: migration	
Sample	**male**	**female**
	(1)	**(2)**
% of same-village male adults that migrate	0.766***	0.237***
	(13.34)	(3.52)
% of same-village female adults that migrate	0.097	0.437***
	（1.91）	(7.36)
Numbers of Observations	1,780,353	1,547,643
R-Squared	0.29	0.27

Notes: Significance at 10% (*), 5% (**), 1% (***). Robust standard errors in parentheses. All regressions include county fixed effects and all the control variables used in Table 6. Errors are clustered by village. Shaded cells refer to influence from peers of the same gender.

Table 10 2SLS regression by surname groups

	Dependent Variable: migration			
	(1)	(2)	(3)	(4)
sample	**Adults with most dominant surname**	**Adults with second most dominant surname**	**Adults with third most dominant surname**	**Adults with other surnames**
% of migrants among adults that have the most dominant surname in the village	0.976***	0.0948	0.154	0.231*
	(0.150)	(0.213)	(0.230)	(0.122)
% of migrants among adults that have the second most dominant surname in the village	0.259**	1.239***	0.378*	0.300**
	(0.101)	(0.172)	(0.213)	(0.145)
% of migrants among adults that have the third most dominant surname in the village	-0.118	-0.321*	0.435*	-0.00815
	(0.111)	(0.179)	(0.225)	(0.147)
% of migrants among adults that have other surnames in the village	-0.258	-0.231	-0.208	0.302*
	(0.210)	(0.310)	(0.301)	(0.175)
Observations	754602	457514	337638	1973942
R-squared	0.279	0.263	0.274	0.268

Notes: Significance at 10% (*), 5% (**), 1% (***). Robust standard errors in parentheses. All regressions include county fixed effects and all the control variables used in Table 6. Errors are clustered by village. Shaded cells refer to influence from peers of the same surname group.

Table 11 2SLS regression by education group

	Dependent Variable: migration			
Sample	**Edu: 0-5**	**Edu: 6-9**	**Edu: 10-12**	**Edu>12**
	(1)	**(2)**	**(3)**	**(4)**
% of migrants among adults whose years of schooling is less than 7	0.917***	0.661***	0.858***	0.706
	(7.90)	(2.66)	(3.22)	(1.03)
% of migrants among adults whose years of schooling is between 7 and 9	-0.11	0.304	-0.076	-0.178
	(-1.51)	(-1.85)	(-0.41)	(-0.35)
% of migrants among adults whose years of schooling is between 10 and 13	0.112**	0.339***	0.753***	0.361
	(2.55)	(2.97)	(7.12)	(1.1)
% of migrants among adults whose years of schooling is above 12	-0.009	0.007	-0.042	0.527***
	(-0.59)	(0.14)	(-0.83)	(3.39)
Observations	1358,858	834404	49331	5579
R-squared	0.2	0.26	0.23	0.26

Notes: Significance at 10% (*), 5% (**), 1% (***). Robust t-statistics in parentheses. All regressions include county fixed effects and all the control variables used in Table 6. Errors are clustered by village. Shaded cells refer to influence from peers of the same education group.

Table 12 2SLS regression by destination, study sample

	Dependent Variable: Migration to				
	Within province	A	B	C	D
	(1)	(2)	(3)	(4)	(5)
% of same village adults migrating to same province	0.996***	-0.00512	-0.00255	-0.00197	0.00125
	(0.0170)	(0.0283)	(0.0279)	(0.00826)	(0.00419)
% of same village adults migrating to A	0.0501**	1.103***	0.109***	0.0341***	0.00942*
	(0.0207)	(0.0364)	(0.0369)	(0.0107)	(0.00506)
% of same village adults migrating to B	-0.0597	0.0130	1.009***	0.000879	-0.0240**
	(0.0384)	(0.0632)	(0.0626)	(0.0185)	(0.00992)
% of same village adults migrating to C	-0.0327	-0.138	-0.132	0.964***	0.000261
	(0.0548)	(0.0981)	(0.0984)	(0.0282)	(0.0135)
% of same village adults migrating to D	-0.0928	-0.153*	-0.125	-0.0648**	0.967***
	(0.0568)	(0.0848)	(0.0790)	(0.0274)	(0.0160)
Observations	3327996	3327996	3327996	3327996	3327996
R-squared	0.102	0.106	0.097	0.068	0.043

Notes: Significance at 10% (*), 5% (**), 1% (***). Robust standard errors in parentheses. All regressions include county fixed effects, all the control variables used in Table 6, and percent of same village adults migrating to other destinations. Errors are clustered by village. Shaded cells refer to influence from peers migrating to the same destination.

Table 13 2SLS regression by occupation, study sample

	Dependent Variable: Migration for			
	(1) manufacturing	**(2) service**	**(3) construction**	**(4) other jobs**
% of same-village adults migrating for manufacturing jobs	0.927***	-0.00439	-0.0120	-0.0258
	(0.0743)	(0.0164)	(0.0133)	(0.0373)
% of same-village adults migrating for service jobs	-0.454*	0.896***	-0.0405	-0.205*
	(0.243)	(0.0540)	(0.0430)	(0.122)
% of same-village adults migrating for construction jobs	0.0414	0.0115	0.937***	-0.0197
	(0.163)	(0.0361)	(0.0304)	(0.0822)
% of same-village adults migrating for other jobs	-0.236***	-0.0531***	-0.0276*	0.879***
	(0.0759)	(0.0167)	(0.0142)	(0.0386)
Observations	3327996	3327996	3327996	3327996
R-squared	0.190	0.065	0.071	0.132

Notes: Significance at 10% (*), 5% (**), 1% (***). Robust standard errors in parentheses. All regressions include county fixed effects, and all the control variables used in Table 6. Errors are clustered by village. Shaded cells refer to influence from peers migrating for the same occupation.

Table 14: Impact of peer migration on the land in use by non-migrating households

Sample	Dependent Variable: land	
	Exclude migration family	
	OLS	2SLS
	(1)	(2)
% of same-village adults that migrate	1.393***	-0.131
	(0.089)	(0.794)
Numbers of Observations (household)	984,110	984,110
R-squared	0.737	0.733

Notes: Significance at 10% (*), 5% (**), 1% (***). Robust standard errors in parentheses. All regressions include county fixed effects, own contract land, and all the control variables used in Table 6. Errors are clustered by village.

Table 15: Heterogeneous effects of peer migration, study sample

	Dependent Variable: Self migration, 2SLS				
	(1)	(2)	(3)	(4)	(5)
% of same-village adults that migrate	0.796***	0.776***	0.825***	0.538***	0.781***
	(0.0346)	(0.039)	(0.0397)	(0.101)	(0.0397)
% of same-village adults that migrate * distance to the nearest bus/rail/dock station	-0.00085				
	(0.00776)				
% of same-village adults that migrate * distance to county center		0.00016			
		(0.00014)			
% of same-village adults that migrate * distance to the center of the studied area			0.00054*		
			(0.000282)		
% of same-village adults that migrate * distance to the provincial capital				0.00131**	
				(0.000357)	
% of same-village adults that migrate * have TV access					0.0182
					(0.0274)
Numbers of Observations (household)	3,327,996	3,327,996	3,327,996	3,327,996	3,327,996
R-squared	0.274	0.274	0.275	0.274	0.274

Notes: Significance at 10% (*), 5% (**), 1% (***). Robust standard errors in parentheses. All regressions include county fixed effects and all the control variables used in Table 6. IVs are the same as in Column (4) of Table 6. Errors are clustered by village.

Table 16: Empirical tests for organized migration or organized recruiting

	Dependent Variable: Migration , all 2SLS	
	Include village cadre demographics	**Redefine IVs based on children age 0-12 only**
	(1) 2SLS	**(2) 2SLS**
% of same-village adults migrating	0.735***	0.672***
	(0.0417)	(0.0477)
Numbers of Observations	3,327,996	3,327,996
R-Squared	0.272	0.269

Notes: Significance at 10% (*), 5% (**), 1% (***). Robust standard errors in parentheses. All regressions include county fixed effects, all the control variables used in Table 6, and village cadre demographics (gender, education and veteran status for village head and communist party leader.). Errors are clustered by village.