

# Strong versus Weak Ties in Migration\*

Corrado Giulietti<sup>†</sup>      Jackline Wahba<sup>‡</sup>      Yves Zenou<sup>§</sup>

(forthcoming in the European Economic Review)

Accepted version: February 2018

## Abstract

This paper studies the role of strong versus weak ties in rural-to-urban migration decisions in China. We develop a network model that puts forward the different roles of weak and strong ties in helping workers to migrate to the city. We use unique longitudinal data that allow us to test our model by focusing on first-time migration. We address the endogeneity of weak ties using an instrumental variable procedure. Our results indicate that weak and strong ties provide different type of help and hence act as complements in the migration decision, with the interactive effect being particularly strong above a certain threshold of weak ties.

**Keywords:** social networks, internal migration, China.

**JEL Classification:** O15, J61.

---

\*The Longitudinal Survey on Rural Urban Migration in China (RUMiC) comprises three parts: the Urban Household Survey, the Rural Household Survey and the Migrant Household Survey. It was initiated by a group of researchers at the Australian National University, the University of Queensland and the Beijing Normal University and was supported by the Institute for the Study of Labor (IZA), which provides the Scientific Use Files. The financial support for RUMiC was obtained from the Australian Research Council, the Australian Agency for International Development (AusAID), the Ford Foundation, IZA and the Chinese Foundation of Social Sciences. We are grateful to the editor and one of the associate editors, as well as to two anonymous referees. We thank participants at the 5<sup>th</sup> IZA/CIER Annual Workshop on Research in Labor Economics, the 7<sup>th</sup> Migration & Development Conference, the 1<sup>st</sup> PhD Workshop on the Economics of Migration, the 1<sup>st</sup> World Congress of Comparative Economics, the 6<sup>th</sup> SEBA-GATE Annual Workshop, seminar participants at IZA, Reading University, Università Cattolica di Milano, and to Lorenzo Cappellari, Benjamin Elsner, Mark R. Rosenzweig and Ulf Zölitz for useful comments.

<sup>†</sup>University of Southampton, UK. Email: c.giulietti@soton.ac.uk.

<sup>‡</sup>University of Southampton and CPC, UK. E-mail: J.Wahba@soton.ac.uk.

<sup>§</sup>Monash University, Australia, and University of Southampton, UK. E-mail: yves.zenou@monash.edu.

# 1 Introduction

Social interactions, whether regular or occasional, influence individual decisions and behaviors. The effects of social networks on economic activity have been well documented (see Jackson, 2008, Ioannides, 2013, Jackson and Zenou, 2013, Jackson et al., 2017, for recent surveys), particularly in the labor market, where social networks play an important role in transmitting information about jobs (Ioannides and Loury, 2004, Bayer et al., 2008, Topa, 2001, 2011). Social networks are also widely recognized as very influential in migration decisions (Munshi, 2003, McKenzie and Rapoport, 2007, 2010, Beine et al., 2011a,b, Dolfin and Genicot, 2010, Bertoli and Moraga, 2015). Nonetheless, little is known about the mechanisms through which networks exert such effects. The aim of this paper is to investigate the role of networks in depth by disentangling the effect of strong and weak ties in migration decisions.

A large body of literature exists concerning the role played by the different types of network in the labor market. In particular, Granovetter (1973, 1974, 1983) shows that weak ties are superior to strong ties in terms of providing support in getting a job.<sup>1</sup> Indeed, in a close network where everyone knows each other, information is shared and thus potential sources of information are quickly shaken down, whereby the network rapidly becomes redundant in terms of access to new information. By contrast, Granovetter stresses the *strength of weak ties* involving a secondary ring of acquaintances who have contacts with networks outside the ego's network and therefore offer new sources of information about job opportunities.

In the present paper, we investigate whether this is also true for migration decisions. Accordingly, we first derive a theoretical model to illustrate the different channels through which social networks may affect migration decisions. More precisely, we consider a dynamic model in which individuals belong to different dyads and dyad members do not change over time. As a result, two individuals belonging to the same dyad hold a *strong tie* with each other. However, each dyad partner can meet other individuals outside the dyad partnership, referred to as *weak ties* or random encounters. By definition, weak ties are transitory and only last for one period. Weak ties only provide information about job, while strong ties, besides job information, provide more concrete help, such as financial support.

Individuals can be in two different states, namely having migrated or not. Accordingly, there will be three different types of dyads: both members have migrated, one member has migrated and the other has not, or both members have not migrated. In this model, only workers who have migrated in the city can provide information about jobs to rural workers.

---

<sup>1</sup>Granovetter (1973, 1974, 1983) defines *weak ties* in terms of a lack of overlap in personal networks between any two agents, i.e., weak ties refer to a network of acquaintances who are less likely to be socially involved with one another. Formally, two agents A and B have a weak tie if there is little or no overlap between their respective personal networks. Vice versa, the tie is *strong* if most of A's contacts also appear in B's network.

Strong ties provide more reliable information than weak ties. Hence, information about jobs is essentially obtained through strong and weak ties and thus social networks.

We show that a unique steady-state equilibrium exists and explicitly determine the migration rate in the economy. We also show that the probability of migrating increases with the social interactions with strong ties, the social interactions with weak ties and the job arrival rate, and decreases with the job-destruction rate in the city (or the emigration rate).

We subsequently test these theoretical results using a unique longitudinal dataset in China (RUMiC), where we observe the individuals and their networks prior to migration. As in the theoretical model, we define different types of networks based on the strength of the social interaction. First, we define strong ties based on the five closest contacts of the household head. Strong ties refer to contacts who could be relatives, friends or neighbors but are not members of the household. We subsequently estimate the impact of the strong ties who have migrated on the individual’s subsequent migration decision. Second, we define weak ties as the share of previous migrants from the village. We measure the network at a point in time that precedes migration by one year.

There are several challenges when attempting to estimate the effects of networks on migration. First, endogeneity could arise because unobservable factors could affect both the network characteristics and the migration decision. Selectivity could also arise to the extent that only individuals of a certain type self-select into certain networks. Second, there may be common local shocks simultaneously triggering individual and village migration. Third, measurement issues could affect the way in which social networks are measured and reported in survey data. Our identification strategy addresses all these issues. First, in our regression model we include several characteristics of both strong and weak ties that are often unobserved in other surveys. Second, we note that endogeneity issues related to strong ties are annihilated by the fact that strong ties refer to contacts of the household head rather than household members, which represent the majority of our sample. Using a pre-determined network makes it unlikely that choices between the individual and the strong ties are co-determined. Third, in order to address issues related to unobservable shocks at the village level and potential measurement error, we implement an instrumental variable procedure. We construct an instrument that exploits two sources of exogenous variations: weather shocks and village land. Using data from local weather stations, we construct rainfall shocks for the years preceding the survey. We subsequently derive a variable that measures the size of land pertaining to the “remaining village members” (i.e., excluding the individual’s land). The instrument is the interaction between the lagged rainfall shocks and the village land, with the rationale being that the larger the productivity shock accruing to the agricultural production, the higher the probability that villagers will migrate to the city.

Our results indicate that both weak and strong ties matter in the migration decision process. We also find that weak and strong ties act as complements in the migration decision,

showing that the probability of migration is largest when at least one of the individual's strong ties has migrated and the fraction of migrants in the village is beyond a certain threshold. We finally provide evidence of the existence of several channels behind the complementarity. We show that individuals are more likely to migrate when the strong tie has provided financial help or is self-employed, or when the individual found the job through the social network. The role of these channels is stronger when the size of the weak ties is larger, emphasizing that different types of ties provide different type of help.

The remainder of the paper unfolds as follows. In the next section, we discuss the contribution of our paper compared to the literature on social networks and migration. In Section 3, we present the background context by describing the migration and the role played by networks in China. The theoretical model is developed in Section 4, before the data and identification strategy are discussed in Section 5. In Section 6, we present our main empirical results. We discuss and address endogeneity issues and perform several robustness checks in Section 7 and subsequently explore the channels behind the complementarity between strong and weak ties in Section 8. Finally, Section 9 concludes.

## 2 Related literature

There is a growing body of literature concerning social networks and migration, especially in developing countries. One important question concerns the extent to which the influence of networks is significant in addition to the role of the traditional factors (such as the wage differential between the origin and the destination country, the bilateral distance between the two countries, etc.). The empirical literature based on structural gravity models (Beine et al., 2011a,b, Bertoli and Moraga, 2015, Beine and Parsons, 2015) finds an elasticity of about 0.4, which means that a ten percent increase in the bilateral migration stock will lead, on average, to a four percent increase in the bilateral migration flow over the next ten years.

At the microeconomic level, it is important to understand the exact role of networks in the migration decision. As noted by Dolfin and Genicot (2010), migrant networks can facilitate migration in three different ways: providing information about the migration process itself, providing information about jobs at the destination and aiding integration after arrival, and helping to finance the costs of migration.

Recent work provides support for the role of networks in finding jobs at migrants' destinations. Using Mexican rainfall as an instrument for the size of migrants' US networks, Munshi (2003) finds that larger networks substantially improve Mexican immigrants' likelihood of US employment. McKenzie and Rapoport (2007, 2010) investigate the role of networks in alleviating migration costs, finding evidence that community networks tend to lower costs, especially for the less educated. Orrenius and Zavodny (2005) find that having a father or

brother who has migrated to the US increases the likelihood of migration for males.<sup>2</sup>

In most of these papers, networks are measured by taking the share of migrants in the destination country from the same village of origin (see e.g., Munshi, 2003, McKenzie and Rapoport, 2007, 2010). However, this is clearly a very rough measure of social networks and one needs to open this black box to better understand the role of networks on migration. In the present paper, we measure weak ties in the same way but also include strong ties, defined as the closest contacts nominated by the head of the household. We show that one underestimates the effect of social networks on migration by not taking into account the strong ties in the mobility process. Moreover, we are able to show that there are strong complementarities between these two types of ties, especially above a certain threshold of the size of the weak ties. This is important because it means that both types of social interactions are key to understanding migration decision and that weak and strong ties reinforce each other.

Finally, we believe that our results shed light on the mechanisms through which networks encourage migration, because weak and strong ties provide different types of help in the migration process. Weak ties only provide information about jobs at the destination. Strong ties, besides giving job information, also provide more concrete help to facilitate migration, such as financial support and job opportunities. Because weak and strong ties have different roles in encouraging migration, we believe that disentangling between these two types of ties is of paramount importance.

## 3 Migration and social networks in China

### 3.1 Migration in China

Internal migration in China is important in terms of its magnitude and consequences. China is experiencing mass rural-urban migration, triggered by the economic reform that started at the end of the 1970s. Prior to that period, the combination of the household registration system (hukou) and the imposed quotas for per capita consumption considerably limited internal mobility. Agricultural productivity increased with the beginning of the economic restructuring, yielding both an excess rural labor force and a more stable supply of food. Furthermore, these changes were accompanied by a rise in the inflow of foreign investment in urban areas, which itself created a high demand for low-priced labor force. The combination of these vicissitudes progressively generated the largest movement of labor in human history. Recent estimates reveal that 168 million migrant workers moved from their rural residence to urban areas (National Bureau of Statistics of China, 2013).

While partially reformed, the hukou system persists and continues to influence the size

---

<sup>2</sup>See also Wahba and Zenou (2005).

and composition of the rural-to-urban migrant flows (Zenou, 2012). Hukou regulations imply that migrants are only allowed to live in cities for a few years and exclusively for working reasons. Furthermore, migrants often lack access to better paid jobs, social security or good schools for their children. In such a context in which most migrations are precarious in nature yet are also frequent and comprise a large share of rural households, the role of social networks becomes crucial. For example, the network can provide information about job opportunities in the city, as well as effective help to facilitate the move. In the next subsection, we provide a description of the main aspects of the social networks in China and how its role in influencing economic outcomes has been analyzed in the literature to date.

## 3.2 Social networks in China

Social connections – also known as *guanxi* – permeate many aspects of the Chinese culture and are frequently used to achieve the most disparate tasks in daily life, ranging from getting a job to providing favors. Given such pervasiveness, a growing number of economists have started to explore how *guanxi* affects economic decisions and outcomes. A first set of studies looks at the role of social networks in the context of the labor market. For example, Zhang and Zhao (2015) analyze the role of social networks in promoting the self-employment of migrants in Chinese cities. One feature of their study is that they take into account the endogenous formation of migrants’ networks after migration. Long et al. (2017) examine the impact of networks on migrants’ wages using the proportion of labor migrants in the home village as a proxy for the village social network.

The role of networks in the migration decision has previously been analyzed by Zhao (2003), who found that previous rural-to-urban migration – represented by the network of earlier migrants – positively influences subsequent migrations. However, the actual network is not observed in her study. Instead, she relies on approximating it with the proportion of migrants from the same village who migrated in a given year. A similar network measure is used by Chen et al. (2010), who examine the role of networks in determining the cluster of migrants in certain destinations. Our paper contributes to the migration literature by studying the role played by different types of *guanxi* – namely weak and strong ties – where these networks are observed.

# 4 Theoretical model

## 4.1 Assumptions, notations and definitions

Consider a population of individuals of size one.

**Dyads** We assume that individuals belong to mutually-exclusive two-person groups, referred to as *dyads*. We say that two individuals belonging to the same dyad hold a *strong tie* to each other. We assume that dyad members do not change over time. A strong tie is created once and forever and can never be broken. Thus, we can consider strong ties as links between members of the same family or between very close friends.

Individuals can be in either of two different states: having migrated (state 1) or not having migrated (state 0). Dyads – which comprise paired individuals – can thus be in three different states, as follows:

- (i) both members have migrated – we denote the number of such dyads by  $d_2$ ;
- (ii) one member has migrated and the other has not migrated ( $d_1$ );
- (iii) both members have *not* migrated ( $d_0$ ).

For example, a  $d_2$  dyad means that both workers in the dyad (i.e., who have a strong tie relationship with each other) have migrated in the city, while a  $d_1$  dyad means that one person in the dyad has migrated while the other has not.<sup>3</sup>

**Aggregate state** By denoting the migration rate and the non-migration rate at time  $t$  by  $m(t)$  and  $n(t)$ , where  $m(t), n(t) \in [0, 1]$ , we have:

$$\begin{cases} m(t) = 2d_2(t) + d_1(t) \\ n(t) = 2d_0(t) + d_1(t) \end{cases} \quad (1)$$

The population normalization condition can then be written as

$$m(t) + n(t) = 1 \quad (2)$$

or, alternatively,

$$d_2(t) + d_1(t) + d_0(t) = \frac{1}{2} \quad (3)$$

**Social interactions** Time is continuous and individuals live forever. We assume repeated random pairwise meetings over time. Matching can take place between dyad partners or not. At time  $t$ , each individual can meet a weak tie with probability  $\omega(t)$  and her strong tie partner with probability  $\alpha(t)$ . We assume that these probabilities are constant and exogenous and do not vary over time, and thus they can be written as  $\omega$  and  $\alpha$ . We do not assume anything about  $\omega$  and  $\alpha$  being complements or substitutes; rather, we simply assume that individuals

---

<sup>3</sup>The inner ordering of dyad members does not matter.

spend some time with their weak ties (captured by  $\omega$ ) and some time with their strong ties (captured by  $\alpha$ ).

We refer to matching within the dyad partnership as *strong ties*, and to matching outside the dyad partnership as *weak ties* or random encounters. Information is exchanged within each matched pair as explained below.

**Information transmission** Workers migrate from the rural area to the urban area. We denote workers who have migrated (to the urban area) by workers of type  $m$  or  $m$ -workers or *migrants* and workers who have not migrated (and thus live in the rural area) by workers of type  $n$  or  $n$ -workers or *non-migrants*. In order to migrate, a rural worker needs to have some information about a job opportunity in the urban area. When the rural worker obtains this information about a job in the city, she may migrate to the city. Each piece of information about an opportunity is taken to arrive only to the  $m$ -workers, who can subsequently direct it to one of their contacts who have not migrated (through either strong or weak ties). In other words, *only individuals who have already migrated and live in the city can help a rural individual to migrate to the city by providing her with some information about an opportunity (about a job) in the urban area.*

To be more precise, a worker of type  $m$  (a migrant) hears of a job opportunity in the city at the exogenous rate  $\lambda$ . The migrant will transmit this job opportunity to a non-migrant depending on whether the latter has a weak tie or a strong relationship with the migrant. Quite naturally, we assume that the quality of the job information is better when it comes from a strong tie rather than from a weak tie. To be more precise, a non-migrant will always migrate whenever she receives job information from her strong tie, while she will migrate with probability  $0 < p < 1$  when the job information comes from a weak tie. This is because a strong tie – who has a long-term relationship with the non-migrant – always provides reliable information about jobs and can even provide financial help or housing and help with other aspects of urban life for the migrant. On the other hand, a weak tie – who is a random encounter and has a short relationship with the non-migrant – can provide less reliable information about jobs.

Another interpretation is the following. Strong ties provide two “services” to their dyad friend. First, they provide job information. Second, they provide concrete support for finding a job, such as financial help. Weak ties only provide job information. An individual can migrate with only job information (i.e., no financial help) but her probability to migrate is lower than if she receives both job information and financial help. As a result, a potential migrant receives from her weak tie only information about a job, and her probability to migrate is  $\lambda p$ . On the other hand, from her strong tie she will receive both information about a job and financial help, which increases her probability to migrate to  $\lambda$ , which is greater than  $\lambda p$ .

We also assume that there is an exogenous rate  $\delta$  that a  $m$ -worker goes back to the rural area. This is because migrants in the city can lose their jobs (at rate  $\delta$ ) and subsequently need to return to the rural area. This is particularly relevant in the case of China, since migrants without urban hukou are not allowed to stay in the city without working. Similarly, access to welfare programs is usually restricted to urban hukou holders. In other words,  $\delta$  is the exogenous rate at which urban migrants return to the rural area. We call this the emigration rate or the (city) job destruction rate.<sup>4</sup>

Therefore, migrants who hear about an opportunity in the city pass on this information to their current matched partner, who can be a strong or weak tie. Thus, information about opportunities in the city is essentially obtained through social networks. This information transmission protocol defines a Markov process. The state variable is the relative size of each type of dyad. Transitions depend on labor market turnover and the nature of social interactions as captured by  $\omega$  and  $\alpha$ . Owing to the continuous time Markov process, the probability of a two-state change is zero (small order) during a small interval of time  $t$  and  $t + dt$ . This means that both members of a dyad cannot change their status at the same time. For example, two rural workers cannot migrate to the city at the same time, i.e., during  $t$  and  $t + dt$ , the probability assigned to a transition from a  $d_0$ -dyad to a  $d_2$ -dyad is zero. They can eventually both migrate, although it will take some time.

**Flows of dyads between states** It is readily checked that the net flow of dyads from each state between  $t$  and  $t + dt$  is given by:

$$\begin{cases} \dot{d}_2(t) = (\alpha + \omega m(t)p)\lambda d_1(t) - 2\delta d_2(t) \\ \dot{d}_1(t) = 2\omega m(t)\lambda p d_0(t) - (\delta + \alpha\lambda + \omega m(t)\lambda p)d_1(t) + 2\delta d_2(t) \\ \dot{d}_0(t) = \delta d_1(t) - 2\omega m(t)\lambda p d_0(t) \end{cases} \quad (4)$$

Let us explain these equations in further detail, starting with the first one. Subsequently, the variation of dyads comprising two  $m$ -workers ( $\dot{d}_2(t)$ ) is equal to the number of  $d_1$ -dyads in which the  $n$ -worker has migrated to the city (through either her strong tie with probability

---

<sup>4</sup>In our model, return migrants do not convey information about jobs in cities, since once migrants lose their jobs and return home, their information about jobs is obsolete. This seems a plausible assumption for China, considering the speed at which vacancies for “migrant jobs” get filled (Knight and Yueh, 2004) and that the majority of migrants do not have access to unemployment insurance since they do not own an urban hukou (Fleischer and Yang, 2003). There is also direct evidence that return migrants do not convey useful information about jobs. For example, Zhao (2003), investigating the role of migrant networks in the context of China, distinguishes the migration network into “experienced” migrants (i.e., individuals who migrated before but they are still migrating) and return migrants (i.e., individuals who migrated before but returned for good in the home village). She finds that experienced migrants have a positive effect on migration of others, but return migrants do not.

$\alpha\lambda$  or her weak tie with probability  $\omega m(t)\lambda p$ ) minus the number of  $d_2$ -dyads in which one of the two migrants has returned to the rural area. In the second equation, the variation of dyads comprising one migrant and one non-migrant ( $\dot{d}_1(t)$ ) is equal to the number of  $d_0$ -dyads in which one of the non-migrants has migrated to the city (only through her weak tie with probability  $\omega m(t)\lambda p$ , since her strong tie also lives in the rural area and therefore cannot transmit any opportunity in the city) minus the number of  $d_1$ -dyads in which either the migrant worker has returned to the rural area (with probability  $\delta$ ) or the rural worker has migrated to the city thanks to her strong or weak tie (with probability  $[\alpha + \omega m(t)p]\lambda$ ) plus the number of  $d_2$ -dyads in which one of the two migrants has returned to the rural area. Finally, in the last equation, the variation of dyads comprising two rural workers ( $\dot{d}_0(t)$ ) is equal to the number of  $d_1$ -dyads in which the migrant worker has returned to the rural area minus the number of  $d_0$ -dyads in which one of the rural workers has migrated to the city (only through her weak tie, with probability  $\omega m(t)\lambda p$ ). These dynamic equations reflect the flows across dyads, as in the graph below.

[Insert Figure 1 here]

## 4.2 Steady-state equilibrium

In a steady state  $(d_2^*, d_1^*, d_0^*)$ , each of the net flows in (4) is equal to zero. Setting these net flows equal to zero leads to the following relationships:

$$d_2^* = \frac{(\alpha + \omega m^* p)\lambda}{2\delta} d_1^* \quad (5)$$

$$d_1^* = \frac{2\omega m^* \lambda p}{\delta} d_0^* \quad (6)$$

where

$$d_0^* = \frac{1}{2} - d_2^* - d_1^* \quad (7)$$

$$m^* = 2d_2^* + d_1^* \quad (8)$$

$$n^* = 1 - m^* \quad (9)$$

**Definition 1** *A steady-state labor market equilibrium is a five-tuple  $(d_2^*, d_1^*, d_0^*, m^*, n^*)$  such that equations (5), (6), (7), (8) and (9) are satisfied.*

We have the following result:

### Proposition 1

(i) There always exists a steady-state equilibrium  $\mathcal{N}$  where all rural individuals do not migrate to the city whereby only  $d_0$ -dyads exist, namely  $d_2^* = d_1^* = m^* = 0$ ,  $d_0^* = 1/2$  and  $n^* = 1$ .

(ii) If

$$\frac{\delta}{\lambda} < \frac{\omega p + \sqrt{\omega p(4\alpha + \omega p)}}{2} \quad (10)$$

there exists an interior steady-state equilibrium  $\mathcal{I}$  where  $0 < m^* < 1$  is defined by

$$m^* = \frac{\sqrt{\lambda(4\delta\alpha + \alpha^2\lambda + 2\alpha\omega\lambda p + \omega^2\lambda p^2)} - 2\delta - \alpha\lambda + \omega\lambda p}{2\omega\lambda p} \quad (11)$$

$0 < n^* < 1$  by (9), and  $0 < d_0^* < 1/2$  is given by:

$$d_0^* = \frac{\delta^2}{\omega\lambda^2 p(\alpha + \omega p) + \omega\lambda p\sqrt{\lambda(4\delta\alpha + \alpha^2\lambda + 2\alpha\omega\lambda p + \omega^2\lambda p^2)}} \quad (12)$$

Furthermore, the other dyads are given by:

$$d_1^* = \frac{2m^*\omega\lambda p}{\delta}d_0^* \quad (13)$$

$$d_2^* = \frac{\omega\lambda^2 p(\alpha + \omega m^*)}{\delta^2}m^*d_0^* \quad (14)$$

If condition (10) holds, then an interior equilibrium always exists. Indeed, if the rate at which migrants return to the rural area  $\delta$  is sufficiently low and/or the information-contact rate  $\lambda$  is sufficiently high, then an interior equilibrium exists. Otherwise, all workers will not migrate and stay in the rural area and the steady-state equilibrium  $\mathcal{N}$  will prevail. Interestingly, in the interior steady-state equilibrium  $\mathcal{I}$ , we know  $m^*$ , the fraction of workers who have migrated to the city, as well as those among them who have a strong tie migrating to the city ( $d_2^*$  dyads), those who have a strong tie living in the rural area and those who have not migrated ( $d_1^*$  dyads).

To better understand this result, let us analyze the steady-state equilibrium when (i) there are no weak ties and (ii) when there are no strong ties.

**Proposition 2** *Assume (10).*

(i) *When there are no weak ties, i.e.,  $\omega = 0$ , then the equilibrium fraction of workers who have migrated to the city (i.e., the migration rate) is equal to:  $m_{\omega=0}^* = 0$ .*

(ii) When there are no strong ties, i.e.,  $\alpha = 0$ , then the equilibrium fraction of workers who have migrated to the city (i.e., the migration rate) is equal to:<sup>5</sup>  $m_{\alpha=0}^* = 1 - \frac{\delta}{\omega\lambda p}$ .

(iii) If we compare the different migration rates, we have:

$$m_{\omega=0}^* < m_{\alpha=0}^* < m^* \quad (15)$$

where  $m^*$  is defined by (11).

The proof of this proposition is straightforward and is thus omitted. First, this proposition shows that weak ties are more valuable than strong ties in terms of migration decision. This is because there is an asymmetry in the role of strong versus weak ties because weak ties help migration in any dyad while strong ties have no impact on migration in the  $d_0$  dyad. This also explains why, when  $\omega = 0$ , individuals get stuck in the  $d_0$  dyad since there is no possibility for them to leave this dyad by migrating.

Second, expression (15) shows that, for the migration decision, it is always better to have both weak and strong ties than only one of them. This is because there is some complementarity effect between weak and strong ties in the migration decision. Indeed, even if each of them enters in an additive way in the migration decision, they complement each other in the equilibrium decision of migrating.

### 4.3 Comparative statics

The key variable of our model is  $m^*$ , which is given by (11). Since we have a Markov process,  $m^*$  can have two equivalent interpretations: it is either the fraction of time each worker has spent migrating over his/her lifetime or *the unconditional probability of migrating for any worker in the steady state*. For the purpose of our empirical analysis, we use the second interpretation. As can be seen in (11),  $m^*$  is a function of the social interaction with strong ties ( $\alpha$ ), the social interaction with weak ties ( $\omega$ ), the rate at which migrants hear about a job ( $\lambda$ ) and the job destruction rate in the city ( $\delta$ , or the emigration rate). Let us now analyze the impact of these variables on the probability of migrating  $m^*$ . We have the following result:

**Proposition 3** *Assume that condition (10) holds. Accordingly, the probability of migrating increases with the social interactions with strong ties  $\alpha$ , the social interactions with weak ties  $\omega$  and the job arrival rate  $\lambda$ , and decreases with the job destruction rate  $\delta$ , namely:*

$$\frac{\partial m^*}{\partial \alpha} > 0, \quad \frac{\partial m^*}{\partial \omega} > 0, \quad \frac{\partial m^*}{\partial \lambda} > 0, \quad \frac{\partial m^*}{\partial \delta} < 0$$

---

<sup>5</sup>Observe that (10) reduces to  $\delta < \omega\lambda p$  when  $\alpha = 0$ , which guarantees that  $0 < m_{\alpha=0}^* < 1$ .

Furthermore, the cross effect of weak and strong ties on the probability of migrating is undetermined, namely  $\partial^2 m^* / \partial \alpha \partial \omega$  has an ambiguous sign.

This is an interesting result, showing that both weak and strong ties can help a rural worker to migrate to the city. It also shows how the state of the economy in the city (captured by  $\delta$  and  $\lambda$ ) can affect the migration of workers. Finally, it shows that there are important and complex cross effects of the impact of both weak and strong ties on the probability of migration.

#### 4.4 Econometric equation

We would like to test Proposition 3, in other words, evaluating the role of weak and strong ties in the migration decision of rural workers in China. Observe that  $m^*$  is defined by (11) as well as by  $m^* = 2d_2^* + d_1^*$ . Thus, using (5) and (6), we have:

$$m^* = \left[ \frac{(\alpha m^*) \lambda \omega + (\omega m^*)^2 p \lambda + \delta (\omega m^*)}{\delta} \right] \frac{2\omega \lambda p}{\delta} d_0^* \quad (16)$$

where  $d_0^*$  is defined by (12). In (16), we see that  $m^*$ , the probability of migration is a function of  $\omega m^*$ , the probability of migration for weak ties and  $\alpha m^*$ , the probability of migration for strong ties, i.e.,  $m^* = f(\alpha m^*, \omega m^*, \lambda, \delta)$ . This is what we want to test. Therefore, a (linear) reduced form of (16) can be written as:<sup>6</sup>

$$m_i = \beta_0 + \beta_1 S_i^m + \beta_2 W_v^m + \beta_3 (S_i^m \times W_v^m) + \beta_4 \delta_v + \beta_5 \lambda_d + \mathbf{X}'_i \theta + \varepsilon_i \quad (17)$$

where  $m_i$  is the (unconditional) probability of migrating for individual  $i$ ,  $S_i^m$  is the fraction of  $i$ 's strong ties who have migrated,  $W_v^m$  is the fraction of  $i$ 's weak ties who have migrated,  $S_i^m \times W_v^m$  is the interaction term between weak and strong ties,  $\delta_v$  measures the return migration (which varies at the village level),  $\lambda_d$  measures the job arrival rate in urban areas (which varies at the county level),  $\mathbf{X}'_i$  are individual and household attributes of  $i$ , along with indicators for the provinces of residence and additional observable characteristics of the strong and weak ties. We will describe more in detail the variables of the econometric model in the next Section.

---

<sup>6</sup>In robustness checks, we will also test for a non-linear effect of  $S_i^m$  and  $W_v^m$ .

## 5 Data and econometric strategy

### 5.1 Data sources and descriptive evidence

Our analysis is based on the Rural-Urban Migration in China (RUMiC) data collected as part of a large scale project conducted in China and comprising a rural household survey (RHS), an urban household survey and a migrant household survey (MHS). For our purposes, we extract data from the first and second waves of the RHS (see Akgüç et al., 2014 for a technical description of the RUMiC panel dataset). The RHS covers the main migrant sending provinces and was conducted using the random samples from the annual China household income and expenditure surveys carried out in rural villages.<sup>7</sup> The first wave of data was collected in early 2008 (wave I) and contains data that mostly refer to the situation of individuals and households in 2007. Similarly, the second wave was collected in early 2009 (wave II) and contains information about 2008. The survey has detailed information about household members – including those who are currently migrating – and comprises socio-demographic characteristics, labor market outcomes, migration history and the family situation prior to leaving the hometown. We also draw information from two surveys that accompany the RHS: a village survey and a household expenditure survey. From the latter, we obtain information such as household income, land and housing. From the village survey (which is administered to the village cadre), we extract information about the village, including the number of individuals who are currently migrating, as well as a wealth of other socio-economic indicators. Having village information has various advantages, as it allows us to control for many confounding factors associated with outmigration from the village (our measure of weak ties). Nonetheless, our measure of weak ties might also have some measurement issues, which we will discuss in detail below.

We measure  $m_i$  by the probability that the individual will be observed to migrate in the following year, *conditioning on not having ever migrated before*. In other words,  $m_i$  is an indicator that takes the value of 1 if the individual will be observed to migrate for the first time during 2008 (i.e., wave II of the RHS) and 0 otherwise. One of the important features of our analysis is that we model “future migration” as a function of present covariates and particularly network characteristics. The key explanatory covariates are based on two dimensions of the network: the strength of social interactions (weak and strong ties) and their migration status. Strong ties refer to the closest contacts of the family. They can be relatives, friends, neighbors or other acquaintances but are not household members, i.e., not immediate family members. Strong ties are nominated in the survey by the head of the household or – in his/her absence – by the spouse. The head of the household could be individual  $i$  or someone closely related to him/her (the wife or children). Most of our

---

<sup>7</sup>The provinces covered by the RHS are: Anhui, Chongqing, Guangdong, Hebei, Henan, Hubei, Jiangsu, Sichuan, and Zhejiang.

sample comprises individuals who are not household heads. Hence, it is plausible to assume that individual  $i$  does not select his/her strong ties network. We will elaborate on this point below. We measure  $S_i^m$  with a variable ranging from 0 to 1 that captures individual  $i$ 's share of strong ties who migrated. The number of strong ties nominated in the survey varies across individuals, with some indicating only one and some indicating five contacts. However, 80% of respondents nominate at least three contacts. Hence, in our analysis, we use information on all available strong ties nominated by the head of the household, and we also check the robustness of our results for different definitions (e.g., strong ties defined by the closest contact only). As for weak ties, we measure  $W_v^m$  with the share of individuals who have migrated out of the village (where  $i$  resides) and moved temporarily to urban areas. In other words, weak ties represent the entire social space, which we approximate with the village.

We also have information concerning two additional parameters stemming from the theoretical model: the emigration rate ( $\delta$ ) and the rate at which job information reaches the network ( $\lambda$ ). For  $\delta$ , we construct a measure of return migration at the village level using the RHS data. In particular, we use a subsample of individuals that are not employed in the main analysis, and combine information on return migration and migration intentions. We construct an indicator that takes the value of 1 if the individual is a return migrant and he/she does not intend to migrate again and 0 otherwise. This allows us to capture effective return migration and hence exclude those who plan to migrate again (e.g., circular migrants). We then aggregate the return migration indicators and obtain a measure at the village level (i.e., the share of return migrants).<sup>8</sup>

For  $\lambda$ , we retrieve information from the Migrant Household Survey (MHS), which collects information on a sample of rural migrants living in 15 cities.<sup>9</sup> We obtain a proxy for social network use that varies at the county level in the following manner. From the MHS, we extract information about the individual's *first job after migration*. The question asks retrospectively about who provided information about the city job. We construct an indicator that takes the value of 1 if the migrant obtained information from migrant workers from the same village and 0 otherwise (on average, 34% of migrants obtained information through the village network). This variable captures the use of network by migrants and can be considered to proxy for the "speed" at which migrants hear about a job opportunity. To "match" this variable with the RHS in a credible manner, we combine information on the province of origin of the migrant in the MHS with information on the county of residence of individuals

---

<sup>8</sup>One might wonder whether return migration is correlated with the weak tie variable. We have checked this and found that the correlation is practically inexistent (0.005, s.e. 0.819).

<sup>9</sup>The 15 cities are Guangzhou, Dongguan, Shenzhen, Zhengzhou, Luoyang, Hefei, Bengbu, Chongqing, Shanghai, Nanjing, Wuxi, Hangzhou, Ningbo, Wuhan and Chengdu. Note that 86% of migrants in the MHS come from the nine provinces surveyed by the RHS, making the former a complementary source of information about migration.

in the RHS. As a first step, we calculated the Euclidean distance between counties of a given province and the city where the migrant from that province lives. Within each province of origin, we normalize the distance between counties and the cities whereby the longest distance takes the value of 0 and the shortest the value of 1. This allows us to “re-weight” the value of  $\lambda$  by giving more importance to migrants from nearby counties (which arguably send more migrants due to shorter distance) and less to migrants from faraway counties. Importantly, this weighting implies that the value of  $\lambda$  varies at the county level, thereby still permitting the identification of province fixed effects in our regression model.

Other covariates include additional characteristics of the networks. For the weak ties, we include several variables capturing socio-economic characteristics of the village: the village income and public expenditure, whether the village has made investment in health or education in the previous year, the amount of grain subsidy, whether the village is located in one of the poverty alleviation counties, the presence of primary schools, the share of workforce involved in agricultural activities, and whether the closest train station is located at a distance above 5 kilometers. For the strong ties, we use information concerning the relationship status (being a relative vis-à-vis a neighbor or friend), the marital status and whether the strong tie has above 8 years of schooling. Similar to the migration status, we average the values of these variables over all available strong ties.

Finally, we include characteristics at the individual (e.g., gender, age, marital status, number of children, education, employment status, self-reported health status and body mass index) and household level (e.g., income, land, housing) as well as an indicator for each of the provinces in which the individual lives.

For the purposes of our study, we restrict the sample to individuals aged 16 to 35. The reason is that 75% of first-time migrants belong to this age window (the median age being 30 years) and we want to focus on individuals who move exclusively for labor-related reasons. This yields a sample size of 2,073.<sup>10</sup>

In Table 1, we present individual and household characteristics, as well as the network’s attributes.

*[Insert Table 1 here]*

The first entry in the table refers to our response variable, i.e., whether the individual migrates for the first time in the following year. The value of 0.084 indicates that 8.4% of our initial sample of individuals for whom we have information in wave I of the RHS are observed as migrants in wave II of the survey. The percentage of males is below 50%, reflecting the fact that males of this age are more likely to have already migrated (and hence are excluded from our sample). Individuals in the sample are relatively young on average (27 years). This is unsurprising given that we limit the sample to those below 35 years. About 60% are married

---

<sup>10</sup>To corroborate our results, we perform several sensitivity checks, for example including individuals up to 64 years old.

and have more than 8 years of schooling. The large majority belong to the Han ethnicity. About half of individuals in the sample are farmers. The table also contains information on household characteristics. The average household size is above four. In each household, there is less than one elderly person aged above 64. The average household income is about 7.8 ln RMB.

In terms of network characteristics, on average, 5.4% of the strong ties reported being a migrant in wave I, whereas the percentage for weak ties is more than double (14%). These figures reflect the magnitude of the migration phenomenon in rural China. The table also reports other characteristics of both weak and strong ties. About one-third of the strong ties are a relative of the household head; about 90% of them are married; half report having more than 8 years of schooling. In terms of weak ties' characteristics, almost half of the village workforce is employed in agricultural activities. About 50% of the villages have primary schools. About one-sixth of villages are located in economically depressed areas (i.e., located in one of the key counties where the national poverty reduction program was active).<sup>11</sup>

In Table 2, we provide evidence about dyads. We classify strong ties as “migrated” when at least one strong tie has migrated and “not migrated” when none of the strong ties has migrated. In the theoretical model, we normalized the total population to 1. In the real world, this is not the case and we denote by  $T$  the total population of individuals in our sample who were surveyed in wave I ( $T = 2,073$ ). This does not include the strong ties. Since each person has one strong tie, the total population is  $2T = 4,146$ .<sup>12</sup> Furthermore, in the theoretical model, we assume that  $d_{10} = d_{01} = d_1$ . This is also true in the data, with a small discrepancy. Indeed,  $d_{10}$  is the number of dyads in which the person interviewed has migrated while her strong tie has not. Similarly,  $d_{01}$  is the number of dyads where the person interviewed has not migrated while her strong tie has. Note that each dyad has two persons, whereby – for example – the number of persons interviewed who have migrated with a strong tie who has also migrated is  $d_2$ , while the total number of migrants who have migrant strong ties is  $2d_2$ . In Table 2, we see that  $d_0 = 1,729$ , which means that there are 1,729 interviewed persons who have not migrated and whose strong ties have not migrated. We can also see that  $d_{10} \neq d_{01}$ , i.e., there are 144 interviewed persons who have migrated but their strong ties have not ( $d_{10}$ ) while there are 171 interviewed individuals who have not migrated but their strong ties have ( $d_{01}$ ). Finally, there are 29 interviewed persons who have migrated and whose strong tie has migrated ( $d_2$ ). If we want to calculate the *unconditional*

---

<sup>11</sup>In 2001, 592 counties were designated as “key counties for national poverty alleviation and development”. These areas are targeted by governmental programs that took place between 2001 and 2010 and were aimed at reducing poverty and fostering economic development.

<sup>12</sup>To be consistent with the theoretical model, we assume here that each individual has only one strong tie. To check the migration status of the strong tie, we only consider the first strong tie nominated by the interviewed person. However, in the empirical analysis, we will consider all possible strong ties that a person nominates (up to five).

probability of migrating for an interviewed person, it is given by  $(d_{10} + d_2) / 2,073 = 0.084$ , which is the percentage given in Table 1. However, if we want to calculate  $m^*$ , as defined in the theoretical model by (11), we obtain:

$$m^* = d_{10}^* + d_{01}^* + 2d_2^* = \frac{d_{10}^*}{4,146} + \frac{d_{01}^*}{4,146} + \frac{2d_2^*}{4,146} = 0.09$$

As a result, the *unconditional* probability of migrating for any person (including both interviewed individuals and their strong ties) is 9%, which is higher than 8.4%, the *unconditional* probability of migrating for an interviewed individual. This is because there is more migration from strong ties with non-migrant interviewed individuals than from interviewed persons with non-migrant strong ties ( $d_{01}^* = 171 > d_{10}^* = 144$ ).

[Insert Table 2 here]

Figures 2, 3 and 4 provide information about some characteristics of the strong ties. In Figure 2, we classify strong ties depending on the type of help that they provide, distinguishing between financial help, psychological help, help with daily affairs and no help. This figure shows that strong ties who have migrated are more likely to provide financial help than those who have not migrated, while the opposite is true for psychological help. This indicates that the strong ties who have migrated are relatively more prone to supplying material and financial support (and not only information) to the potential migrants.

[Insert Figure 2 here]

In Figure 3, we depict the frequency of contacts with strong ties. There are stark differences depending on the migration status of the strong tie. Those who have migrated have somewhat less frequent contact with individuals in rural areas than those who have not migrated. This is certainly due to the fact that the geographical distance between strong ties is less for those who have not migrated than for the migrants.

[Insert Figure 3 here]

Finally, we report some statistics about money and gifts exchange in Figure 4. While the amount of money received is generally larger for both strong ties who did not migrate and those who did, the difference is more marked for the latter group. This complements the information about financial help reported in Figure 2.

[Insert Figure 4 here]

## 5.2 Identification

There are several challenges when trying to estimate network effects, mostly related with the endogeneity of the network.

First, *network endogeneity of strong ties* might arise because there could be unobservable factors affecting both friendship formation and migration decision.<sup>13</sup> Moreover, since individuals only name up to five persons as their strong ties and we do not observe the whole network (i.e., strong ties have not been interviewed), the problem of selectivity into networks is similar to the endogeneity of choosing one’s strong tie. To mitigate this issue, we perform several checks. First, in one of our sensitivity checks, we narrow down our sample to only strong ties who are known to the household head since age 16. In this case, it is difficult to believe that the reason why two 16 year-olds become friends is because they anticipate that they will help each other to migrate many years later. Second, when checking the covariates’ balance between individuals who have strong ties who have migrated and individuals who have strong ties who have not migrated, we find that most characteristics are very similar between the two groups, suggesting that endogeneity of migration is not a major issue in our sample.

*Endogeneity of weak ties* might be a more serious issue. In our settings, major problems do not arise from network formation. Since we approximate weak ties with the village and given that rural individuals typically do not change their village residence, the choice of network does not seem to be an issue (i.e., it is plausible to assume that network membership is “conditionally” exogenous). On the other hand, approximating weak ties with the village might create two additional sources of endogeneity bias related to omitted variables and measurement error. Omitted variable bias can arise due to unobservable shocks that might simultaneously affect the village and the individual’s migration. Part of this endogeneity issue is mitigated by the fact that we measure the individual’s migration one year after the weak ties are measured. However, this strategy will only be able to annihilate the role of transient shocks. On the other hand, more persistent shocks – i.e., which could affect migrations for more than one year – still have the potential to generate endogeneity bias. The sign of the bias is a priori unknown. For example, if shocks induce more migration and if they do so for both the individual and the rest of the village, the OLS estimator will yield upwardly biased estimates.

Measurement issues could also be at work. Our measure of weak ties has some advantages with respect to those used in the literature (see Munshi, 2003, McKenzie and Rapoport, 2007,

---

<sup>13</sup>Observe that the reflection problem (Manski, 1993) does not arise here because the reference group is different from one individual to the other. The closest contacts nominated by the head of household are usually *not* the same across different households, especially when people nominate at least three closest contacts. This is a similar approach to the network literature where the presence of intransitive triads in the network solves the reflection problem (Lee, 2007, Bramoullé et al., 2009, Calvó-Armengol et al., 2009, Blume et al., 2011)

2010, Dolfin and Genicot, 2010). Indeed, since the weak ties variable comes from village data, we can control for a wealth of village characteristics that are usually unobserved in other surveys. However, one drawback of approximating the social space of the individual with the village is that weak ties could be measured with error. In our settings, two individuals from the same village have the same weak ties, which might not be the case. Measurement error is common to all studies using micro data and where weak ties are measured using some form of spatial approximation (e.g., village, cities, regions, etc.). Surprisingly, measurement issues are often ignored, although this is a serious source of endogeneity that typically biases the OLS estimator towards zero.

To address the endogeneity of weak ties, we implement an instrumental variable strategy. The literature has shown that weather conditions are a valid instrument for the migration of weak ties in the context of mobility from rural areas. This is the case because weather events – such as rainfalls – have a direct effect on agricultural production in rural areas, thereby influencing the incentives to migrate to other (most likely urban) areas. For example, Munshi (2003) shows that rainfall in the origin community in Mexico is an exogenous predictor for the network size in the U.S.. Giles and Yoo (2007) use rainfall shocks to instrument for the size of the migrant network in rural China. Using weather shocks as an instrument also seems feasible in our case, although our settings are slightly more complicated. While the outcome variable in the aforementioned cases does not involve migration (being employment rates of migrants in the U.S. for Munshi (2003) and household consumption of rural residents in the case of Giles and Yoo, 2007), in our case the response variable is the individual’s migration decision. This means that any instrument that affects the migration of the *entire* village will also have a direct effect on the migration of the individual, thereby violating the exclusion restriction. Hence, we need a slight modification of the instrumental variable strategy. For each individual, we construct a shock measure that pertains to the “remaining village members”, i.e., excluding the individual him-/herself. In order to obtain variation across individuals, we interact the rainfall shock – as described below – with the land size of the remaining village members. This interaction term constitutes our instrumental variable.

We use land size to proxy for agricultural productivity since we do not have direct measures of agricultural productivity at the village level. Using household-level data from the RHS, we have checked that indeed land size positively correlates with income from agriculture (in absolute level) and with the share of income from agriculture (relative to total income).<sup>14</sup> Identification hinges on the assumption that the instrument only affects the individual’s

---

<sup>14</sup>We estimated two regression models where the dependent variables are the log income from agriculture and the share of income from agriculture. The key explanatory variable is land size, and the specification includes all individual, household and village level covariates included in our regressions in Table 3 column 5. The coefficient for land size when the dependent variable is log agricultural income is 0.096 (s.e. 0.021) and when the dependent variable is the share of agricultural income is 0.015 (s.e. 0.004). These correlations are robust to the inclusion/exclusion of income in the regression.

migration through weak ties' migration. It is important to note that in the instrumental variable regressions, we include the main effect for rainfall shock and the main effect for the land pertaining to the remaining village members as control variables. We further saturate the model by including an interaction term between the individual's land and the rainfall shock. The rationale is that after controlling for the shock that is common to everyone in the village, the size of the land of remaining village members and the individual's own shock, rainfall that affects remaining village members' agricultural production will influence weak ties' migration but will not directly influence the migration of the individual.

Formally, the second stage is defined as follows:

$$m_i = \beta_0 + \beta_1 S_i^m + \beta_2 W_v^m + \beta_3 (S_i^m \times W_v^m) + \beta_4 \delta_v + \beta_5 \lambda_d + \beta_6 R_d + \beta_7 L_{v*} + \beta_8 (R_d \times L_i) + \mathbf{X}_i' \theta + \varepsilon_i \quad (18)$$

where  $R_d$  indicates the rainfall shock in the county and  $L_{v*}$  is the land of the "remaining" village members (which by construction varies across households within the same village).

The first stage for the main effect model is:

$$W_v^m = \gamma_0 + \gamma_1 R_d + \gamma_2 (R_d \times L_{v*}) + \gamma_3 L_{v*} + \gamma_4 (R_d \times L_i) + \gamma_5 S_i^m + \gamma_6 \delta_v + \gamma_7 \lambda_d + \mathbf{X}_i' \tau + \nu_i \quad (18a)$$

Note that for the interaction model, there are two first stages (one for the main effect of weak ties and one for the interaction between weak and strong ties). The first stages for the interaction model are:

$$W_v^m = \chi_0 + \chi_1 R_d + \chi_2 (R_d \times L_{v*}) + \chi_3 (S_i^m \times R_d \times L_{v*}) + \chi_4 L_{v*} + \chi_5 (R_d \times L_i) + \chi_6 S_i^m + \chi_7 \delta_v + \chi_8 \lambda_d + \mathbf{X}_i' \tau + \nu_i \quad (18b)$$

$$S_i^m \times W_v^m = \pi_0 + \pi_1 R_d + \pi_2 (S_i^m \times R_d \times L_{v*}) + \pi_3 (R_d \times L_{v*}) + \pi_4 L_{v*} + \pi_5 (R_d \times L_i) + \pi_6 S_i^m + \pi_7 \delta_v + \pi_8 \lambda_d + \mathbf{X}_i' \tau + \mu_i \quad (18c)$$

The instrument for equation (18a) is  $R_d \times L_{v*}$  and for equations (18b) and (18c) are  $S_i^m \times R_d \times L_{v*}$  and  $R_d \times L_{v*}$ . Note that both the first and second stages include the main effect for the rainfall shock ( $R_d$ ), the main effect for the land of other village members ( $L_{v*}$ ), and a term for the interaction between the rainfall shock and the land of the individual

$(R_d \times L_i)$ .<sup>15</sup> Importantly, these are control variables and not exclusion restrictions. In particular, the term  $R_d \times L_i$  controls for how the household agricultural production shock directly influences the individual’s decision to migrate. In summary, identification comes from the extent to which rainfall shocks, through their effect on village land’s productivity (excluding the individual’s), affect villagers’ decision of migration.<sup>16</sup>

To obtain  $R_d$ , we follow a procedure similar to Giles and Yoo (2007). In particular, we obtained rainfall data for the 1980-2007 period from the China weather stations (there are about 400 of them in the whole country). We select only stations inside or within a close distance from the counties where villages in our sample are located. For this purpose, we calculate the Euclidean distance between the centroid of the county and the location of the weather station and select only those stations within a radius of 1. This procedure gives us 200 weather stations.<sup>17</sup> For each weather station, we calculate the average amount of rainfall of each year between 2000 and 2006. We subsequently calculate the absolute difference between the log rainfall in each year and the log average rainfall for the period 1980-1999. The difference between yearly rainfall and the long run average rainfall determines the rainfall “shock” for the seven years before individuals in our sample are observed to migrate. The larger the shock, the higher the migration to urban areas that is expected to occur. This is the case because both scarcity and excess of rainfall are detrimental to agricultural production, thereby inducing affected individuals to devise strategies to cope with yield loss, such as migration to urban areas. The “cumulative” effect of these shocks over the years is an exogenous predictor of network size at the time of the survey. As argued by Munshi (2003) and Giles and Yoo (2007), not all the rainfall shocks are expected to affect network size. Reflecting the cumulative effect that weather conditions have on networks, shocks that occurred in a more distant time are better predictors than those that occurred in a time closer to when the network is measured. After some experiments, we found the fourth to seventh lags being the relevant ones.

## 6 Baseline results

The aim of our empirical analysis is to examine the effects of both weak and strong ties in the individual migration decisions, as well as their potential interactive effects. Accordingly, we estimate equation (17).

---

<sup>15</sup>The main effect  $L_i$  is already included in  $\mathbf{X}$  in both stages.

<sup>16</sup>Alternative identification strategies have been implemented in the literature to model migration choices. For example, Kinnan et al. (2018) use the hukou reforms and labor demand shocks in destination areas as exogenous “pull” factors for migration. In our case using pull factors as instrumental variables is more problematic, since, e.g., a shock in destination areas is likely to simultaneously affect both individual migration (our response variable) and the village migration (weak ties).

<sup>17</sup>Note that by following this procedure, a few stations can fall within more than one county.

Table 3 presents the baseline case of the OLS linear probability model’s estimates.<sup>18</sup> The columns represent alternative models in which we vary the set of regressors. In order to account for heterogeneous characteristics of the sending regions, note that all models include province fixed effects. Our preferred estimation is column 5, where we have the full set of controls. The table reports only the parameter estimates of the strong and weak ties’ variables. Estimates of additional covariates are reported in Table A1 in the Appendix.<sup>19</sup> Since the values of  $\lambda$  and  $\delta$  are estimated, we calculated bootstrap robust standard errors.<sup>20</sup>

Let us now focus on the role of weak and strong ties in migration. The estimate of  $S_i^m$  in column 1 is positive and highly statistically significant. The magnitude implies that a one standard deviation increase in the migration of the strong ties is associated with a 0.071 standard deviation increase in the probability of the individual’s migration. The estimates in column 2 also reveal a similar positive association between weak ties and the individual’s migration. One standard deviation increase in  $W_v^m$  is associated with a 0.086 standard deviation increase in individual migration. While the difference is not large, the comparison of the estimates in the first two columns suggests that the predictions from expression (15) (i.e.,  $m_{\omega=0}^* < m_{\alpha=0}^*$ ) are empirically corroborated. The pattern of results does not change when we jointly estimate weak and strong ties. In column 4, we further add the parameters  $\delta$  and  $\lambda$ . The estimates match the theoretical model’s predictions: Higher return migration has a detrimental effect on first-time migration, whereas a more frequent use of social networks in the urban areas is positively associated with migration.

In columns 1 to 4, equation (17) is estimated with the restriction  $\beta_5 = 0$ , i.e., no cross effect of weak and strong ties. The pattern of results becomes more interesting in column 5, where we interact the strong and weak tie migration variables ( $S_i^m \times W_v^m$ ), i.e., we relax the assumption that  $\beta_3 = 0$ . In this model, we center the values of  $S_i^m$  and  $W_v^m$  around their means. Hence, for example, one reads the parameter estimate for  $S_i^m$  as the partial effect of strong ties on migration when the level of weak ties is equal to the mean. The estimates of  $\beta_1$  and  $\beta_2$  are virtually unchanged. Importantly, however, the interaction effect is positive and statistically significant. Its magnitude is also sizeable when compared to the main effects of  $S_i^m$  and  $W_v^m$ : a one standard deviation increase in  $S_i^m \times W_v^m$  implies a 0.076 standard deviation increase in the probability of migrating.

We have estimated additional models in order to check the robustness of the results to the inclusion of control variables, the measurement of strong and weak ties and the sample

---

<sup>18</sup>We also estimated probit models, finding remarkably similar marginal effects. We privileged estimation with a linear model to ease the interpretation of the interaction term.

<sup>19</sup>In terms of estimates of standard individual and household covariates, we find that being young has a positive impact on migration. Males are more likely to migrate, although the estimate is not statistically significant. Being a farmer makes individuals more likely to migrate. Similarly, larger land size is a factor favoring migration.

<sup>20</sup>We choose the sample size (2,073) as the draw sample size and set 1,000 as the number of replications of the bootstrap procedure.

selection. First, we have ascertained that the estimates of our preferred specification in Table 5 are not sensitive to the exclusion of all controls for weak and strong ties. Second, we have found that including individuals of all ages produces a similar pattern of results of our baseline, albeit with somewhat smaller magnitudes. Third, we explore the presence of non-linear effects of social networks, finding that the quadratic terms of weak and strong ties are both insignificant. Last, we corroborated that measuring weak and strong ties using variables in levels instead of shares produces similar results. We include the results of these additional robustness checks in Table A2 in the Appendix.

[Insert Table 3 here]

Table 3 reveals two important results. First, if we were to measure the social networks effect only through weak ties – which is the standard way in which researchers have tested so far (e.g., Munshi, 2003, McKenzie and Rapoport, 2007, 2010) – we would underestimate the true role of the network, given the importance of strong ties. Second, the positive and statistically significant estimate of  $\beta_3 = \partial^2 m_i / \partial W_v^m \partial S_i^m$  means that *weak and strong ties act as complements in the migration decision*. In other words, the higher the fraction of villagers who migrate, the greater the effect of migrants’ strong ties on the individual’s migration decision.

In order to quantify the impact of weak versus strong ties on the migration decision, we report in Table 4 the predicted probabilities of migration based on the estimates of Table 3. Let us start with the estimations of column 4, where the interaction term is not included. First, we observe that when nobody has migrated in the village ( $W_v^m = 0$ ), the effect of strong ties on migration is important. Indeed, about 9.5% of individuals migrate when their strong ties have also migrated, while only 4.5% will migrate if their strong ties have not migrated. More importantly, if we compare weak and strong ties, one needs to reside in a village where more than 20% of people have migrated (corresponding to the eighth decile) to obtain a migration rate of about 9.4% when the strong tie has not migrated. In other words, the impact of a strong tie who has migrated on a person’s own migration decision is “equivalent” to the impact of 20% of the people from the village who have migrated. The table also shows that the difference in migration probability between a strong tie who has migrated and one who has not is relatively constant when the share of weak ties who migrate increases. This is clearly not the case when there are interaction effects, as reported in the second part of Table 4. The complementarity between the two types of ties is evident. If only 10% of weak ties have migrated (fourth decile), the individual probability of migrating does not significantly change whether or not the strong ties have migrated (7.2 vs 8.3%). By contrast, if 20% of weak ties have migrated (eighth decile), the individual’s probability of migration increases substantially (from 10.2% to 17.3%) when strong ties have migrated compared to the case in which they have not. This means that weak and strong ties are

complements; in particular, the interactive effect between weak and strong ties is largest in the last deciles of the weak ties migration distribution. One potential alternative explanation – besides complementarity – is that the two type of ties are strongly correlated. If this were the case, the interaction term could partially capture the fact that higher levels of weak ties migration correspond to higher levels of strong tie migration. However, this does not seem to be the case. As we show in Table A3, the level of weak ties migration is fairly similar irrespective of strong ties migration.

[Insert Table 4 here]

## 7 Endogeneity issues

### 7.1 Weak ties

In this section, we perform additional analysis to address main potential endogeneity issues. We first investigate the instrumental variable approach for weak ties described in the previous section, before performing several tests to check the potential endogeneity of strong ties.

Table 5 presents the instrumental variable (IV) estimates, where weak ties migration is instrumented with the interaction term between rainfall shock and the land size pertaining to remaining village members. We report the instrumental variable estimates that correspond to the OLS models in columns 4 and 5 of Table 3. Our preferred estimates in Table 5 are in columns 1 and 2, although we also report results from alternative models. As discussed, more distant shocks are expected to be stronger predictors of the weak ties migration. For example, Munshi (2003) finds that rainfall lags between four and six years preceding migration are stronger predictors of network size. Similarly, Giles and Yoo (2007) found that the fourth, fifth and sixth lags of rainfall shocks are strong predictors for weak ties migration. After some investigation, we concluded that rainfall lags between four and seven years before weak ties are observed (wave I) yield the strongest first-stage estimates.

In Table 5, together with second stage estimates, we report the first-stage results and several statistics pertaining to the relevance of the instrument and endogeneity tests. In column 1, it can be seen that the instruments are a strong predictor of the endogenous variable (note, the average value of  $L_{v*}$  – land of other village members – is 3,402 mu, while the average value of  $R_d$  – rainfall shock in the county – is 0.282 for  $t - 4$ , 0.181 for  $t - 5$ , 0.253 for  $t - 6$  and 0.193 for  $t - 7$ ). The F-stat in column 1 of Table 5 is 33, above the value of the F-statistic for the 5% maximal IV size (16.85) as set by Stock and Yogo (2005). The result of the Durbin-Wu-Hausman test confirms the presence of endogeneity (under the assumption of instrument validity) at the 1% significance level. In terms of the second stage, the estimates of weak ties are much larger than the OLS, albeit one should interpret

the IV estimates with caution, considering that the 5% confidence intervals of the OLS and IV estimates partially overlap. The larger estimate signals the presence of downward bias in the OLS, possibly related to measurement error in the weak ties migration variable. The estimate of strong ties migration is slightly smaller than in column 4 of Table 3. In column 2, we report the instrumental variable estimates of the interaction model. Note that there are eight instruments for the first stage of  $W_v^m$  and eight for the first stage of  $S_i^m \times W_v^m$ . In the table, we only report the estimates of the four instruments which are important for each stage (i.e.,  $\chi_2$  from equation (18b) and  $\pi_2$  from equation (18c)). The first-stage estimates reveal that the instruments are relevant and endogeneity is present. In this case, the F-statistic for the first stage is 16.35, slightly below the 5% maximal IV relative bias (17.70), and above the threshold set for the 10% maximal IV relative bias for the case of two endogenous regressors and eight instruments (10.22). Overall, the results from the second stage mimic the OLS results in the sense that we find evidence of complementarity, as captured by the positive and statistically significant interaction term. For both models, the results of the Hansen J-statistic indicate that one cannot reject the null hypothesis that jointly all the instruments are uncorrelated with the error term and can hence be excluded from the second stage equation.<sup>21</sup>

In columns 3 and 4, we estimate an alternative model to further corroborate the validity of our instrumental variable procedure. We exclude villages where the land was redistributed in the previous year (about 7% of the villages). Since land in rural villages might be temporarily redistributed across households depending of the current number of residents, one might worry that migration in the past – which is not observable to us but might be correlated with current village migration – could affect the exogeneity of the instrument (which is constructed using the land of remaining households). We think that it is quite unlikely that past migration invalidates our instrumental variable strategy. One simple reason is that, on average, the land size at the village level will not change. While some households might obtain more land as a consequence of the redistribution and others might obtain less land, such redistribution is not related to the migration behavior of the single household. The estimates in columns 3 and 4 confirm that the pattern of results is very similar when we focus our attention on the subset of villages where land was not redistributed.

In columns 5 and 6 we finally check the sensitivity of the IV estimates to the exclusion of village controls. The rationale of this check is the following: If the rainfall shocks were correlated with current village factors and if this correlation would substantially bias the IV estimates, then the second stage estimates should be sensitive to the presence/absence of village observables. However, the comparison of the estimates in columns 5 and 6 with those

---

<sup>21</sup>In unreported analysis, we have also explored an alternative model using only the sixth and seventh lags of rainfall shocks. The results are qualitatively similar to those reported in column 1 and 2 of Table 5, with the exception of the p-values for the Durbin-Wu-Hausman test, which were somewhat larger.

in columns 1 and 2 show no appreciable difference.

[Insert Table 5 here]

## 7.2 Strong ties

In this subsection, we perform several checks for our definition of strong ties. Besides weak ties, endogeneity issues related to *network formation and migration decision* could also affect strong ties. For example, individuals may form relationships with friends that could subsequently help them to migrate. Another possibility is that the migration decisions are co-determined between the individual and the strong ties (despite observing them at two different points in time). In order to address the endogeneity of network formation, we estimate alternative OLS and instrumental variable models to those presented thus far. In the first test of Table 6, we consider only strong ties who are known to the household head since the age of 16. This robustness check aims to exclude the potential endogeneity in the choice of the strong tie. As can be seen, the pattern of the results is unchanged. In the second test, we exclude from the sample all individuals who report having at least one relative among their strong ties. This robustness check aims to understand the extent to which our results concerning strong ties and its interactive effect with weak ties are due to the presence of relatives within the close network. Once again, the estimates are only mildly sensitive to this change in definition (despite the substantial sample size reduction).

In the third test, we check the sensitivity of our results to a different characterization of the strong ties. In particular, we change the definition of strong ties and include only the “closest contact” (or strongest tie). This is because it is possible that selectivity or endogeneity issues could be at work in the choice (and reporting) of the number of strong ties. However, this does not seem to be an issue in our sample as the results are qualitatively the same. Finally, in the last test, we attempt to address the endogeneity in the migration decision of the strong ties. For example, if a few years before the survey the strong tie and the individual decided together to migrate, the coefficient of the strong ties migration would not capture a fully exogenous effect, but rather a joint decision, and thus it would be biased. However, if our estimates were driven by this type of bias, we should only observe our results for those individuals who planned to migrate (and decided to do so jointly with the strong ties). On the other hand, our results should not hold for those not planning to migrate in the near future. To investigate this, we reduce the sample to individuals who report – in wave I – that they have not planned to migrate in the near future (i.e., up to one year). The pattern of estimates is once again remarkably similar to the baseline IV case.

To further corroborate that the endogeneity of strong ties migration is not a major issue, in Table A3 in the Appendix we check the balance of covariates depending on the migration status of strong ties. Indeed, we note that a large and statistically significant difference arises

in the migration status of the individual. The probability of migration is 14.5% for individuals reporting that at least one tie has migrated, but only 7.7% for those whose strong ties have not migrated (this difference is remarkably similar to the OLS estimate of  $\beta_2$ ). However, in terms of other covariates, we note a remarkable similarity. There are only a few differences in terms of characteristics of the weak and strong ties (arguably these characteristics are also determinants of the migration status of the strong tie) and a few differences in the regional distribution (consistently with what reported in Table 3 for the individual migration, strong ties are also more likely to come from less developed villages). While this is not a formal test for the exogeneity of the migration status, it ensures that the results concerning the strong ties and its interactive effect with weak ties are not substantially influenced by endogeneity or selection issues. This should not come as a surprise considering that strong ties are distant 2 from the individual and hence it is unlikely that the individual selects his/her close network and that the migration decision are simultaneously or endogenously taken.

[Insert Table 6 here]

## 8 The complementarity between strong and weak ties

In our theoretical model, we argue that weak ties only provide information about jobs while strong ties, besides job information, also provide complementary help such as financial and moral support, and possibly job opportunities. In this section, we empirically explore the existence of these potential channels.

That weak and strong ties provide different types of information in job search is an argument already embedded in seminal work on social networks (Granovetter, 1973, 1983). In the case of migration – we conjecture – strong ties can provide concrete support that goes beyond information. For example, strong ties can provide financial help or a direct connection with the employer in urban areas. Furthermore, if strong ties own a small business or are self-employed and seek to hire workers, they can offer a job opportunity to the perspective migrant. Tangible help from close connections is especially relevant in the context of China, since the majority of migrants have a rural hukou and hence cannot rely on institutional support such as unemployment benefits. This help would take place within the system of gifts and favor exchange that characterizes social interactions in China.

To empirically explore the channels behind the complementarity of strong and weak ties, we accessed ancillary information on how the current job of individuals is obtained, the type of help received from the strong ties and the occupation of the strong ties.<sup>22</sup> This information

---

<sup>22</sup>Information on how the job is obtained refers to time  $t + 1$  (when the individual makes the migration choice), while information about occupation and type of help from the strong tie refers to time  $t$  (when characteristics of the strong tie are observed).

is available both for individuals who migrates and those who do not. For the three channels above, we constructed the following: a) one indicator which takes the value of 1 if the job is obtained through the social network (family members, relatives, friends or acquaintances) and 0 otherwise (e.g., through direct application or employment agencies); b) one indicator which takes the value of 1 if at least one strong ties provides financial help and 0 otherwise; c) one indicator which takes the value of 1 if at least one strong ties is self-employed and 0 otherwise.

In the first step, we cross-tab each of the three indicators with the indicator for whether at least one of the strong ties has migrated. For each cell, we computed the predicted probabilities based on column 2 of Table 5. We report the predicted probabilities in the first two columns of Table 7, which represent the migration status of the weak ties. The rows of this table represent the indicators for job-finding methods, type of help and employment status of the strong ties. We found that the probability of migration is highest when we observe at the same time that the strong ties have migrated and the job was obtained through the social network (0.152). This probability is twice as large than when the individual found the job through other channels and the strong ties did not migrate (0.076). It is also higher than when the strong ties migrate but the individual found a job through other channels (0.119). Similarly, we found that the probability of migration is the highest when the strong ties have migrated and provided financial help (0.135) and when the strong ties have migrated and are self-employed (0.142). The observed gaps in predicted probabilities show that different types of help received by close connections are important for the migration decision, and that such effect is greater when the strong ties have migrated.

In the second step, we add the size of the weak ties as an additional dimension. The aim is to explore whether and how weak ties – who provide only job information – complement with the strong ties’ characteristics. We compute the predicted probability for the subsamples of individuals who report relatively small values of weak ties (below or at the median of the weak ties distribution) and of individuals who report large values of weak ties (above the median). We report the predicted probabilities in the last four columns of Table 7. Here we are interested in comparing the probability of migration, e.g., between individuals who report finding a job through other channels, strong ties who have not migrated, and small values of weak ties (0.053) with individuals who report finding a job through the network, strong ties who have migrated and large values of weak ties (0.216). Now that we consider the role of weak ties, the gap in probabilities is much larger. Similar large values of predicted probabilities emerge when considering the case of individuals who report – besides having a large network of weak ties – that their strong ties have migrated and provided financial help (0.182) or that their strong ties have migrated and are self-employed (0.186).<sup>23</sup>

---

<sup>23</sup>In unreported results, we investigated different definitions for the explored channels. We constructed and indicator for whether all strong ties migrated, one for whether all strong ties provided financial help and

In summary, Table 7 provides evidence that the effect of personal connections on the migration decision is greatest when the individual’s strong ties provide financial support, help in finding a job or job prospects, and when at the same time the individual possesses a relatively large network of weak ties, who are able to provide diverse job information.

## 9 Conclusion

It is well established that individual migration is affected by the individual’s family and community networks. However, little research exists investigating the mechanisms by which networks exert such effects.

We study these mechanisms by first developing a theoretical model that illustrates the different channels through which networks may affect migration decisions. Using unique data from China, we subsequently estimate the role played by social networks in the migration decision using networks prior to migration, and distinguishing between strong and weak ties. Strong ties are measured by the closest contacts (excluding household members) of the household head, while weak ties are determined by the fraction of migrants from the village in which the individual resides.

Our results indicate that both weak and strong ties matter in the migration decision process. We also show that one underestimates the effect of social networks on migration by not taking into account the strong ties in the mobility process. We finally find that weak and strong ties act as complements in the migration decision, with the interactive effect between weak and strong ties being particularly strong above a certain threshold of the size of weak ties’ migration.

We believe that our results shed light on the mechanisms by which networks encourage migration, because weak and strong ties provide different types of help in the migration process. Weak ties usually provide some information about jobs at the destination, whereas strong ties usually provide information about jobs plus concrete support to migrate. Given that weak and strong ties play different roles in encouraging migration and because they complement each other, we believe that it holds paramount importance to disentangle between these two types of ties. Thus, our paper provides a first attempt at answering a very important question in relation to both social networks and migration, although clearly more research is needed to better understand the relationship between the two.

---

one for whether all strong ties are self-employed. We found that with this more restrictive definition, the gaps in terms of predicted probabilities are much larger. We reached a similar conclusion when considering a more restrictive definition for the size of the weak ties, i.e., below the first quartile and above the fourth quartile of the weak ties distribution.

## Bibliography

- Akgüç, M., C. Giulletti, and K. F. Zimmermann (2014). The RUMiC Longitudinal Survey: Fostering Research on Labor Markets in China. *IZA Journal of Labor & Development* 3(1), 1–14.
- Bayer, P., S. L. Ross, and G. Topa (2008). Place of Work and Place of Residence: Informal Hiring Networks and Labor Market Outcomes. *Journal of Political Economy* 116(6), 1150–1196.
- Beine, M., F. Docquier, and Ç. Özden (2011a). Diasporas. *Journal of Development Economics* 95(1), 30–41.
- Beine, M., F. Docquier, and Ç. Özden (2011b). Dissecting Network Externalities in International Migration. *CESifo Working Paper Series 3333*.
- Beine, M. and C. Parsons (2015). Climatic Factors as Determinants of International Migration. *The Scandinavian Journal of Economics* 117(2), 723–767.
- Bertoli, S. and J. F.-H. Moraga (2015). The Size of the Cliff at the Border. *Regional Science and Urban Economics* 51, 1–6.
- Blume, L. E., W. A. Brock, S. N. Durlauf, and Y. M. Ioannides (2011). Identification of Social Interactions. In J. Benhabib, A. Bisin, and M. O. Jackson (Eds.), *Handbook of Social Economics, Vol. 1B*, pp. 853–964. Amsterdam: Elsevier.
- Bramoullé, Y., H. Djebbari, and B. Fortin (2009). Identification of Peer Effects Through Social Networks. *Journal of Econometrics* 150(1), 41–55.
- Calvó-Armengol, A., E. Patacchini, and Y. Zenou (2009). Peer Effects and Social Networks in Education. *The Review of Economic Studies* 76(4), 1239–1267.
- Chen, Y., G. Z. Jin, and Y. Yue (2010). Peer Migration in China. *NBER Discussion Paper 15671*.
- Dolfin, S. and G. Genicot (2010). What do Networks do? The Role of Networks on Migration and “Coyote” Use. *Review of Development Economics* 14(2), 343–359.
- Fleischer, B. M. and D. Yang (2003). Labor laws and regulations in china. *China Economic Review* 14(4), 426–433.
- Giles, J. and K. Yoo (2007). Precautionary Behavior, Migrant Networks, and Household Consumption Decisions: An Empirical Analysis Using Household Panel Data from Rural China. *The Review of Economics and Statistics* 89(3), 534–551.

- Granovetter, M. S. (1973). The Strength of Weak Ties. *American journal of sociology*, 1360–1380.
- Granovetter, M. S. (1974). *Getting a Job: A study of Contacts and Careers*. Cambridge, MA: Harvard University Press.
- Granovetter, M. S. (1983). The Strength of Weak Ties: A Network Theory Revisited. *Sociological Theory* 1, 201–233.
- Ioannides, Y. M. (2013). *From Neighborhoods to Nations: The Economics of Social Interactions*. Princeton: Princeton University Press.
- Ioannides, Y. M. and L. D. Loury (2004). Job Information Networks, Neighborhood Effects, and Inequality. *Journal of Economic Literature* 42(4), 1056–1093.
- Jackson, M. O. (2008). *Social and Economic Networks*. Princeton: Princeton University Press.
- Jackson, M. O., B. W. Rogers, and Y. Zenou (2017). The Impact of Social Networks on Economic Behavior. *Journal of Economic Literature* 55, 49–95.
- Jackson, M. O. and Y. Zenou (2013). *Economic Analyses of Social Networks, The International Library of Critical Writings in Economics*. London: Edward Elgar Publishing.
- Kinnan, C., S.-Y. Wang, and Y. Wang (2018). Access to Migration for Rural Households. *American Economic Journal: Applied Economics*, forthcoming.
- Knight, J. and L. Yueh (2004). Job mobility of residents and migrants in urban china. *Journal of comparative economics* 32(4), 637–660.
- Lee, L. (2007). Identification and Estimation of Econometric Models with Group Interactions, Contextual Factors and Fixed Effects. *Journal of Econometrics* 140(2), 333–374.
- Long, W., S. Appleton, and L. Song (2017). The Impact of Job Contact Networks on Wages of Rural–Urban Migrants in China: A Switching Regression Approach. *Journal of Chinese eConomiC and Business studies* 15(1), 81–101.
- Manski, C. F. (1993). Identification of Endogenous Social Effects: The Reflection Problem. *The Review of Economic Studies* 60(3), 531–542.
- McKenzie, D. and H. Rapoport (2007). Network Effects and the Dynamics of Migration and Inequality: Theory and Evidence from Mexico. *Journal of Development Economics* 84(1), 1–24.

- McKenzie, D. and H. Rapoport (2010). Self-selection Patterns in Mexico-U.S. Migration: The Role of Migration Networks. *The Review of Economics and Statistics* 92(4), 811–821.
- Munshi, K. (2003). Networks in the Modern Economy: Mexican Migrants in the U.S. Labor Market. *The Quarterly Journal of Economics* 118, 549–599.
- National Bureau of Statistics of China (2013). *China Statistical Yearbook 2010*. China Statistics Press.
- Orrenius, P. M. and M. Zavodny (2005). Self-selection among Undocumented Immigrants from Mexico. *Journal of Development Economics* 78(1), 215–240.
- Stock, J. H. and M. Yogo (2005). Testing for Weak Instruments in Linear IV Regression. In J. H. S. Donald W. K. Andrews (Ed.), *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*, pp. 80. Cambridge: Cambridge University Press.
- Topa, G. (2001). Social Interactions, Local Spillovers and Unemployment. *The Review of Economic Studies* 68(2), 261–295.
- Topa, G. (2011). Labor Markets and Referrals. In J. Benhabib, A. Bisin, and M. Jackson (Eds.), *Handbook of Social Economics, Vol. 1B*, pp. 1193–1221. Amsterdam: Elsevier.
- Wahba, J. and Y. Zenou (2005). Density, Social Networks and Job Search Methods: Theory and Application to Egypt. *Journal of Development Economics* 78(2), 443–473.
- Zenou, Y. (2012). Housing Policies in China: Issues and Options. *Regional Science Policy & Practice* 4(4), 393–417.
- Zhang, J. and Z. Zhao (2015). Social-family Network and Self-employment: Evidence from Temporary Rural–urban Migrants in China. *IZA Journal of Labor & Development* 4(1), 1–21.
- Zhao, Y. (2003). The Role of Migrant Networks in Labor Migration: The Case of China. *Contemporary Economic Policy* 21(4), 500–511.

# Tables and Figures

Table 1: Summary statistics

Individual characteristics		Network characteristics	
Migrates in 2008	0.084 (0.277)	$S_i^m$ : strong ties migration <sup>+</sup>	0.054 (0.187)
Male*	0.432 (0.496)	$S_i^c$ : relative <sup>+</sup>	0.316 (0.402)
Age	27.180 (5.518)	$S_i^c$ : married <sup>+</sup>	0.887 (0.277)
Birth order	1.924 (1.163)	$S_i^c$ : high education (> 8 years) <sup>+</sup>	0.534 (0.424)
Married*	0.621 (0.485)	$W_v^m$ : weak ties migration <sup>+</sup>	0.140 (0.088)
Number of children	0.782 (0.831)	$W_v^c$ : village income (RMB/100000)	0.832 (2.914)
Han ethnicity*	0.992 (0.090)	$W_v^c$ : village expenditure (RMB/100000)	0.519 (1.469)
Urban hukou*	0.083 (0.275)	$W_v^c$ : investment in health/education*	0.494 (0.500)
Above 8 years of education*	0.580 (0.494)	$W_v^c$ : located in poverty alleviation county*	0.135 (0.341)
Work training*	0.148 (0.355)	$W_v^c$ : grain subsidy (RMB)	19.553 (14.269)
Farmer*	0.535 (0.499)	$W_v^c$ : presence of primary school*	0.512 (0.500)
Good or excellent health*	0.890 (0.313)	$W_v^c$ : workforce in agriculture <sup>+</sup>	0.490 (0.246)
BMI ( $\times 100$ )	0.216 (0.026)	$W_v^c$ : distance from train station >5Km*	0.306 (0.461)
Household size	4.539 (1.277)		
Number of elderly in household (age > 64)	0.177 (0.452)	$\delta_v$ : return migration <sup>+</sup>	0.066 (0.110)
Household land (Mu)	4.220 (4.466)		
Housing size (m <sup>2</sup> /100)	1.486 (0.816)	$\lambda_d$ : migrants using networks <sup>+</sup>	0.191 (0.049)
Household labor income (ln RMB)	7.804 (3.062)		
<b>Provinces</b>			
Hebei*	0.066 (0.248)	Hubei*	0.092 (0.289)
Jiangsu*	0.141 (0.349)	Guangdong*	0.144 (0.351)
Zhejiang*	0.135 (0.342)	Chongqing*	0.053 (0.223)
Anhui*	0.085 (0.279)	Sichuan*	0.108 (0.310)
Henan*	0.177 (0.382)		

Observations: 2,073. Source: RUMiC data, RHS waves I and II. All variables refer to wave I except “Migrates in 2008” which refers to wave II. Standard deviations in parentheses.

(\*) refers to 0/1 variables. (+) refers to shares.

$S_i^m$  /  $W_v^m$  refers to strong / weak ties who migrated.  $S_i$  /  $W_v$  refers to other characteristics of the strong / weak ties. Strong ties refer to the five closest persons in the network; Weak ties refer to village.

Table 2: Dyads

		Strong ties		
		Not migrated	Migrated	Total
<b>Individual</b>	Does not migrate	1729	171	1900
	Migrates	144	29	173
	Total	1873	200	2073
	$d_0$	= 1729/2073	0.8341	
	$d_{10}$	= 144/2073	0.0695	
	$d_{01}$	= 171/2073	0.0825	
	$d_2$	= 29/2073	0.0140	

Source: RUMiC data, RHS waves I and II. Strong ties refers to migration status of the strong ties in wave I; Individual refers to migration status of the individual in wave II.

Entries in the table refers to cross-tabulations of the dependent variable with a 0/1 variable for whether at least one strong ties migrated (Migrated) or none of the strong ties migrated (Not Migrated).

Table 3: Weak and strong ties in migration

	1	2	3	4	5
$S_i^m$ : strong ties migration	0.1048*** (0.0403)		0.1014** (0.0394)	0.1008** (0.0398)	0.1049*** (0.0392)
$W_v^m$ : weak ties migration		0.2730*** (0.0762)	0.2640*** (0.0751)	0.2558*** (0.0739)	0.2903*** (0.0771)
$S_i^m \times W_v^m$					1.1910*** (0.4600)
$S_i$ : friend	-0.0128 (0.0144)		-0.0089 (0.0142)	-0.0112 (0.0142)	-0.0097 (0.0143)
$S_i$ : married	0.0323 (0.0231)		0.0382 (0.0234)	0.0418* (0.0235)	0.0457** (0.0226)
$S_i$ : high education	-0.0162 (0.0143)		-0.0133 (0.0143)	-0.0044 (0.0142)	-0.0033 (0.0143)
$W_v$ : village income		0.0014 (0.0023)	0.0012 (0.0024)	-0.0007 (0.0024)	-0.0006 (0.0024)
$W_v$ : village expenditure		-0.0020 (0.0054)	-0.0007 (0.0055)	0.0047 (0.0055)	0.0044 (0.0054)
$W_v$ : investment in health/education		0.0103 (0.0138)	0.0140 (0.0139)	0.0080 (0.0137)	0.0097 (0.0136)
$W_v$ : in poverty alleviation county		0.0402 (0.0247)	0.0410* (0.0245)	0.0328 (0.0241)	0.0346 (0.0240)
$W_v$ : grain subsidy		0.0001 (0.0004)	0.0001 (0.0005)	0.0004 (0.0004)	0.0004 (0.0004)
$W_v$ : presence of primary school		-0.0263** (0.0132)	-0.0277** (0.0133)	-0.0296** (0.0132)	-0.0327** (0.0133)
$W_v$ : workforce in agriculture		0.0383 (0.0322)	0.0454 (0.0325)	0.0451 (0.0321)	0.0500 (0.0317)
$W_v$ : distance from train station >5km		0.0135 (0.0144)	0.0096 (0.0144)	0.0090 (0.0143)	0.0059 (0.0141)
$\delta$ : return migration rate				-0.0990** (0.0415)	-0.0996** (0.0414)
$\lambda$ : migrants using networks				1.7495*** (0.4236)	1.7875*** (0.4230)
Individual and household covariates	Y	Y	Y	Y	Y
Province fixed effects	Y	Y	Y	Y	Y
$R^2$	0.09	0.10	0.10	0.11	0.12
N	2073	2073	2073	2073	2073

Source: RUMiC data, RHS waves I and II. The dependent variable is from wave II; all control variables are from wave I. Sample is composed by individuals aged 16 to 35 who have never migrated before.

Dependent variable is a binary variable indicating whether the individual migrates in 2008. Regressions estimated with linear probability models. Robust bootstrapped standard errors in parentheses. \*/\*\*/\*\* indicate significance at the 0.1/0.05/0.01 level.  $S_i^m / W_v^m$  is the share of strong / weak ties who migrated.  $S_i / W_v$  refer to other characteristics of the strong / weak ties. See Table A1 for estimates of individual and household covariates.

Table 4: Predicted probabilities

Deciles	Values of $W_v^m$	Main effect model				Interaction model			
		$S_i^m = 0$		$S_i^m = 1$		$S_i^m = 0$		$S_i^m = 1$	
		Prob.	St. Err.	Prob.	St. Err.	Prob.	St. Err.	Prob.	St. Err.
1	0.019	0.047	0.010	0.095	0.020	0.049	0.010	0.016	0.034
2	0.041	0.053	0.008	0.101	0.020	0.056	0.009	0.035	0.030
3	0.072	0.061	0.007	0.109	0.019	0.065	0.007	0.062	0.025
4	0.097	0.067	0.006	0.115	0.019	0.072	0.006	0.083	0.022
5	0.124	0.074	0.006	0.122	0.019	0.080	0.006	0.106	0.020
6	0.148	0.080	0.006	0.128	0.019	0.087	0.006	0.127	0.020
7	0.171	0.086	0.006	0.134	0.019	0.093	0.007	0.147	0.022
8	0.202	0.094	0.008	0.142	0.020	0.102	0.008	0.173	0.026
9	0.225	0.100	0.009	0.148	0.020	0.109	0.009	0.193	0.030
10	0.287	0.116	0.013	0.164	0.022	0.127	0.013	0.246	0.042

Source: RUMiC data, RHS waves I and II. Entries are predicted probabilities calculated at different values of  $S_i^m$  and  $W_v^m$  and obtained from estimating the regression model in column 4 (Main effects) and in column 5 (Interaction model) of Table 3 and keeping other predictors at the mean values.  $S_i^m / W_i^m$  refers to strong / weak ties who migrated.  $S_i^m = 0$  indicates that no strong tie migrated;  $S_i^m = 1$  indicates that at least one strong tie has migrated.

Table 5: Instrumental variable estimates

	Full sample		No reallocated land		No village controls	
	Second stage					
$S_i^m$ : strong ties migration	0.0830** (0.0395)	0.0935** (0.0435)	0.0926** (0.0412)	0.1149** (0.0480)	0.0791** (0.0396)	0.0887** (0.0435)
$W_v^m$ : weak ties migration	1.2128*** (0.4486)	1.2149*** (0.4509)	1.1674*** (0.4019)	1.2155*** (0.3954)	1.0785** (0.4372)	1.0515** (0.4452)
$S_i^m \times W_v^m$		2.2131** (1.1013)		2.8150** (1.1572)		2.2312** (1.0999)
	First stage					
	$W_v^m$					
Village land $\times$ Shock t-4	0.0380*** (0.0071)	0.0365*** (0.0074)	0.0403*** (0.0075)	0.0385*** (0.0079)	0.0427*** (0.0073)	0.0415*** (0.0075)
Village land $\times$ Shock t-5	-0.0041 (0.0075)	-0.0035 (0.0075)	-0.0032 (0.0066)	-0.0028 (0.0069)	-0.0018 (0.0078)	-0.0015 (0.0077)
Village land $\times$ Shock t-6	0.0491*** (0.0058)	0.0489*** (0.0059)	0.0534*** (0.0070)	0.0531*** (0.0071)	0.0482*** (0.0058)	0.0480*** (0.0058)
Village land $\times$ Shock t-7	0.0289*** (0.0091)	0.0292*** (0.0093)	0.0222** (0.0104)	0.0220** (0.0107)	0.0280*** (0.0093)	0.0284*** (0.0094)
	$S_i^m \times W_v^m$					
$S_i^m \times$ Village land $\times$ Shock t-4		0.0020 (0.0167)		0.0014 (0.0160)		0.0030 (0.0165)
$S_i^m \times$ Village land $\times$ Shock t-5		0.0122 (0.0230)		0.0075 (0.0245)		0.0100 (0.0232)
$S_i^m \times$ Village land $\times$ Shock t-6		0.0065 (0.0118)		0.0083 (0.0124)		0.0067 (0.0118)
$S_i^m \times$ Village land $\times$ Shock t-7		0.0485** (0.0202)		0.0483** (0.0190)		0.0483** (0.0205)
Kleibergen-Paap Wald F rk statistic	32.998	16.355	32.829	16.613	32.096	15.783
Durbin-Wu-Hausman test (p-value)	0.010	0.039	0.016	0.029	0.039	0.158
Hansen J statistics (p-value)	0.836	0.422	0.878	0.488	0.515	0.219
N	2073	2073	1928	1928	2073	2073

Source: RUMiC data, RHS waves I and II. Rainfall data are from the China Meteorological Data Service Center. All regressions include the covariates of Table 3. Robust bootstrapped standard errors in parentheses. \*/\*\*/\*\* indicate significance at the 0.1/0.05/0.01 level.

The instrumental variable is the interaction between village land and the lagged rainfall shock. Village land refers to the size of the village land excluding that of the individual. Rainfall shocks are absolute differences between the log rainfall in each weather station in a given year and the log average rainfall between 1980 and 1999 for each weather station. Weather stations are matched with the county of residence of individuals.

Table 6: Strong ties - sensitivity analysis

	Knows $S_i^m$ before age 16				Excluding relatives			
	OLS		Instr. Var.		OLS		Instr. Var.	
$W_v^m$	0.2583*** (0.0809)	0.2978*** (0.0835)	1.2496*** (0.4656)	1.3324*** (0.4545)	0.3358*** (0.0915)	0.3878*** (0.0969)	1.0575** (0.5355)	0.9937* (0.5389)
$S_i^m$	0.1190*** (0.0438)	0.1241*** (0.0442)	0.1018** (0.0450)	0.1131** (0.0504)	0.1653*** (0.0632)	0.1342** (0.0597)	0.1507** (0.0638)	0.1117* (0.0663)
$S_i^m \times W_v^m$		1.3884*** (0.4947)		2.4285** (1.1457)		1.3689** (0.5556)		1.9796* (1.0176)
N	1787	1787	1787	1787	1181	1181	1181	1181
	1 <sup>st</sup> friend only				Did not intend to migrate in 2008			
	OLS		Instr. Var.		OLS		Instr. Var.	
$W_v^m$	0.2620*** (0.0743)	0.1955*** (0.0729)	1.1949*** (0.4469)	0.9471* (0.4921)	0.2979*** (0.0704)	0.3303*** (0.0733)	1.1580*** (0.4444)	1.1860*** (0.4337)
$S_i^m$	0.0801** (0.0328)	0.0920*** (0.0338)	0.0723** (0.0328)	0.0982** (0.0390)	0.0997** (0.0422)	0.1087** (0.0430)	0.0906** (0.0408)	0.1059** (0.0457)
$S_i^m \times W_v^m$		0.9886*** (0.3698)		1.9665* (1.1215)		1.2488*** (0.4679)		1.8892* (1.0466)
N	2073	2073	2073	2073	1890	1890	1890	1890

Source: RUMiC data, RHS waves I and II. All regressions include the covariates of Table 3. Robust bootstrapped standard errors in parentheses. \*/\*\*/\*\* indicate significance at the 0.1/0.05/0.01 level.  
 Knows  $S_i^m$  before age 16: sample includes only individuals who know the strong ties since age 16 or before. Excluding relatives: sample includes only individuals whose strong ties are not relatives. 1<sup>st</sup> friend only: strong ties correspond to first contact nominated by the household head. Did not intend to migrate in 2008: only includes individuals who in wave I declared they did not want to migrate.

Table 7: Strong and weak ties complementarity

	Full sample		$W_v^m \leq \text{median}$		$W_v^m > \text{median}$	
	$S_i^m = 0$	$S_i^m = 1$	$S_i^m = 0$	$S_i^m = 1$	$S_i^m = 0$	$S_i^m = 1$
Job through other channels	0.076 (0.099)	0.119 (0.136)	0.053 (0.069)	0.076 (0.086)	0.099 (0.117)	0.157 (0.159)
Job through network	0.081 (0.122)	0.152 (0.167)	0.046 (0.071)	0.086 (0.090)	0.121 (0.151)	0.216 (0.199)
	Full sample		$W_v^m \leq \text{median}$		$W_v^m > \text{median}$	
	$S_i^m = 0$	$S_i^m = 1$	$S_i^m = 0$	$S_i^m = 1$	$S_i^m = 0$	$S_i^m = 1$
Strong tie: No financial help	0.077 (0.104)	0.120 (0.155)	0.052 (0.074)	0.056 (0.067)	0.101 (0.122)	0.166 (0.182)
Strong tie: Financial help	0.080 (0.111)	0.135 (0.144)	0.049 (0.065)	0.089 (0.093)	0.112 (0.138)	0.182 (0.170)
	Full sample		$W_v^m \leq \text{median}$		$W_v^m > \text{median}$	
	$S_i^m = 0$	$S_i^m = 1$	$S_i^m = 0$	$S_i^m = 1$	$S_i^m = 0$	$S_i^m = 1$
Strong tie: Other occupation	0.085 (0.113)	0.118 (0.137)	0.056 (0.074)	0.080 (0.088)	0.114 (0.136)	0.163 (0.168)
Strong tie: Self employed	0.062 (0.091)	0.142 (0.157)	0.039 (0.057)	0.079 (0.087)	0.086 (0.113)	0.186 (0.179)

Source: RUMiC data, RHS waves I and II.  
 Predicted probabilities calculated using values of  $S_i^m$  and  $W_v^m$  from column 1 of Table 5 (Main effects) and column 2 of Table 5 (Interaction model) estimated using probit model and keeping other predictors at the mean values.  $S_i^m / W_v^m$  refers to strong / weak ties who migrated.  $S_i^m = 0$  indicates that none of the strong ties migrated;  $S_i^m = 1$  indicates that at least one strong tie has migrated. Strong ties refer to the five closest persons in the network; Weak ties refer to village.

Figure 1: Flows between rural and urban areas

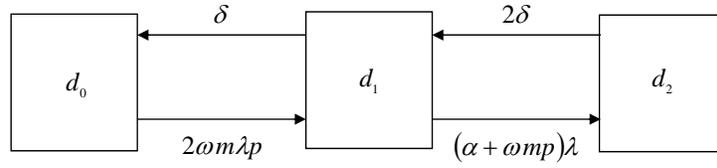


Figure 2: Strong ties: Type of help

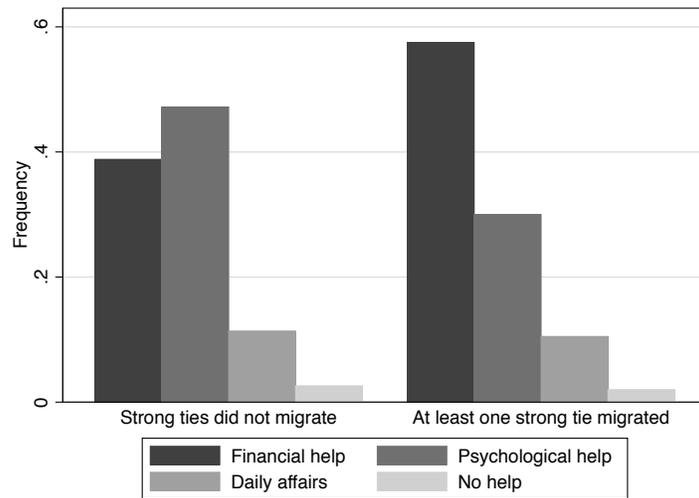


Figure 3: Strong ties: Frequency of contacts

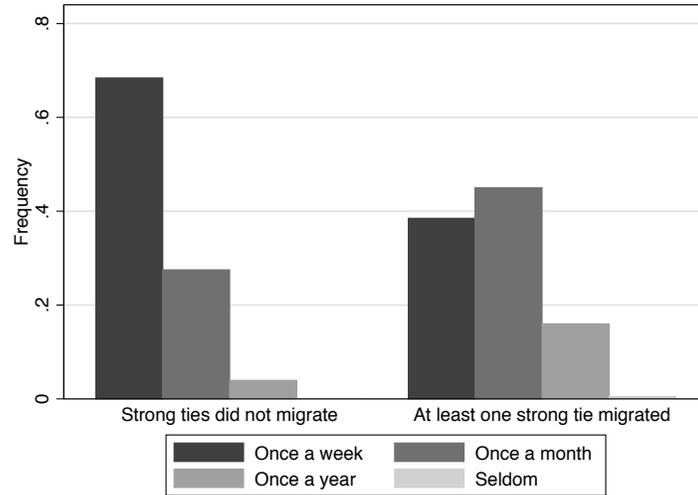
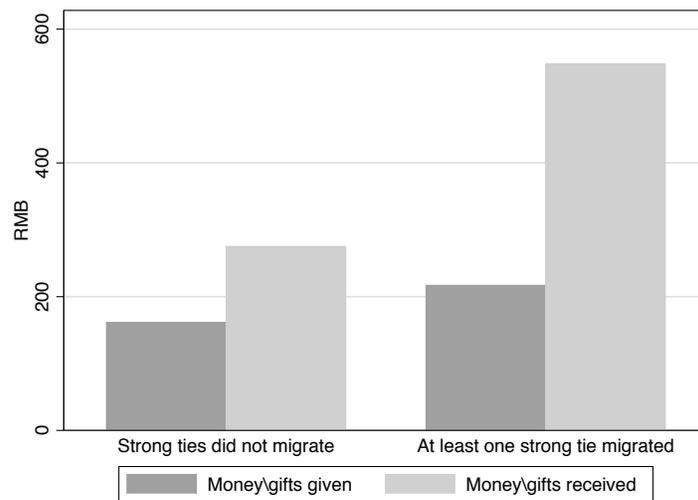


Figure 4: Strong ties: Money/gifts exchange



## Appendix

### Theoretical Model

**Proof of Proposition 1:** We establish the proof in two steps. First, Lemma 1 characterizes all steady-state dyad flows. Lemma 2 then provides conditions for their existence.

**Lemma 1** *There exists at most two different steady-state equilibria: (i) a non-migration equilibrium  $\mathcal{N}$  such that  $m^* = 0$  and  $n^* = 1$ , (ii) an interior equilibrium  $\mathcal{I}$  such that  $0 < m^* < 1$  and  $0 < n^* < 1$ .*

**Proof:** By combining (5) to (8), we easily obtain:

$$m^* = [(\alpha + \omega m^* p)\lambda + \delta] \frac{2\omega m^* \lambda p}{\delta^2} d_0^* \quad (18)$$

We consider two different cases.

(i) If  $m^* = 0$ , then equation (18) is satisfied. Furthermore, using (5) and (6), this implies that  $d_1^* = d_2^* = 0$  and, using (7) and (9), we have  $d_0^* = 1/2$  and  $n^* = 1$ . This is referred to as steady-state  $\mathcal{N}$ .

(ii) If  $m^* > 0$ , then solving equation (18) yields:

$$m^* = \frac{1}{\lambda \omega p} \left[ \frac{\delta^2}{2\omega \lambda p d_0^*} - \delta \right] - \frac{\alpha}{\omega p}$$

Define  $Z = \alpha/(\omega p)$ ,  $B = \delta/(\omega \lambda p)$ . This equation can now be written as:

$$m^* = \frac{B^2}{2d_0^*} - B - Z > 0 \quad (19)$$

Moreover, by using (5) and (6), we obtain:

$$d_1^* = \frac{2m^*}{B} d_0^*, \quad d_2^* = \frac{(Z + m^*) m^*}{B^2} d_0^* \quad (20)$$

- Let us first focus on the case where  $m^* = 1$ . In that case, it has to be that only  $d_2$ -dyads exist and thus  $d_0^* = d_1^* = 0$ , which, using (20) implies that:  $d_2^* = 0$ . So this case is not possible.

- Let us now thus focus on the case:  $0 < m^* < 1$  (which implies that  $0 < n^* < 1$ ).

By plugging (19) and (20) into (7) and after some algebra, we obtain that  $d_0^*$  solves  $\Phi(d_0^*) = 0$  where  $\Phi(d_0^*)$  is the following second-order polynomial:

$$\Phi(d_0^*) = -\frac{Z}{B}d_0^{*2} - \frac{(1+Z)}{2}d_0^* + \left(\frac{B}{2}\right)^2 = 0 \quad (21)$$

This completes the proof of the lemma. ■

## Lemma 2

(i) *The steady-state equilibrium  $\mathcal{N}$  always exists.*

(iv) *The steady-state equilibrium  $\mathcal{I}$  exists when*

$$\frac{\delta}{\lambda} < \frac{\omega p + \sqrt{\omega p(4\alpha + \omega p)}}{2}$$

### Proof.

(i) In this equilibrium  $m^* = 0$ . There are only  $d_0$ -dyads. So when a  $d_0$ -dyad is formed it is never destroyed and thus this equilibrium is always sustainable.

(ii) We know from Lemma 1 that a steady-state  $\mathcal{I}$  exists and that  $m^* \neq 1$ . We now have to check that  $m^* > 0$  and  $0 < d_0^* < 1/2$ . Let us thus verify whether there exists some  $0 < d_0^* < 1/2$  such that  $\Phi(d_0^*) = 0$ , where  $\Phi(\cdot)$  is given by (21). We have  $\Phi(0) = (B/2)^2 > 0$  and  $\Phi'(0) = -(1+Z)/2 < 0$ . Therefore, (21) has a unique positive root smaller than  $1/2$  if and only if

$$\Phi(1/2) = \frac{1}{4} \left[ B^2 - (1+Z) - \frac{Z}{B} \right] = \frac{1}{4} \left(1 + \frac{1}{B}\right) (B^2 - B - Z) < 0.$$

The unique positive solution to  $B^2 - B - Z = 0$  is  $[1 + \sqrt{1 + 4Z}]/2$ , which, using the value of  $Z$ , is equal to:  $\left[1 + \sqrt{1 + 4\alpha/(\omega p)}\right]/2$ . Then,  $d_0^* < 1/2$  if and only if  $B < \left[1 + \sqrt{1 + 4\alpha/(\omega p)}\right]/2$ , equivalent to:

$$\frac{\delta}{\lambda} < \frac{\omega p + \sqrt{\omega p(4\alpha + \omega p)}}{2}$$

Observe that  $d_0^* < 1/2$  guarantees that  $m^* > 0$ .

Finally, to obtain (11) and (12), we proceed as follows. First, we plug the values of  $d_2^*$  and  $d_1^*$  from (5) and (6) into  $m^* = 2d_2^* + d_1^*$  to obtain (18) or equivalently:

$$m^* = \left[ \frac{(\alpha + \omega m^* p)\lambda}{\delta} + 1 \right] \frac{2\omega m^* \lambda p}{\delta} d_0^* \quad (22)$$

Then, we plug the values of  $d_2^*$  and  $d_1^*$  from (5) and (6) into  $d_0^* = \frac{1}{2} - d_2^* - d_1^*$  to obtain:

$$d_0^* = \frac{\delta^2}{2\delta^2 + [(\alpha + \omega m^* p)\lambda + 2\delta]\omega m^* p \lambda} \quad (23)$$

By solving simultaneously equations (22) and (23), we get (11) and (12). ■

**Proof of Proposition 3:** By differentiating (11), we obtain:

$$2\omega\lambda p \frac{\partial m^*}{\partial \alpha} = \frac{2\delta\lambda + \alpha\lambda^2 + \omega\lambda^2 p}{\sqrt{\lambda(4\delta\alpha + \alpha^2\lambda + 2\alpha\omega\lambda p + \omega^2\lambda p^2)}} - \lambda \quad (24)$$

Thus

$$\frac{\partial m^*}{\partial \alpha} > 0 \Leftrightarrow 2\delta + \alpha\lambda + \omega\lambda p > \sqrt{\lambda(4\delta\alpha + \alpha^2\lambda + 2\alpha\omega\lambda p + \omega^2\lambda p^2)}$$

It is easily verified that this inequality is equivalent to  $\delta + \omega\lambda p > 0$ , which is always true.

By differentiating (11), we obtain:

$$2\lambda p \omega^2 \frac{\partial m^*}{\partial \omega} = \left[ \frac{\lambda^2 p (\alpha + \omega p)}{\sqrt{\lambda(4\delta\alpha + \alpha^2\lambda + 2\alpha\omega\lambda p + \omega^2\lambda p^2)}} + \lambda p \right] \omega - \left[ \sqrt{\lambda(4\delta\alpha + \alpha^2\lambda + 2\alpha\omega\lambda p + \omega^2\lambda p^2)} - 2\delta - \alpha\lambda + \omega\lambda p \right]$$

Thus

$$\begin{aligned} \frac{\partial m^*}{\partial \omega} &> 0 \Leftrightarrow \left[ \frac{\lambda^2 p (\alpha + \omega p)}{\sqrt{\lambda(4\delta\alpha + \alpha^2\lambda + 2\alpha\omega\lambda p + \omega^2\lambda p^2)}} + \lambda p \right] \omega \\ &> \sqrt{\lambda(4\delta\alpha + \alpha^2\lambda + 2\alpha\omega\lambda p + \omega^2\lambda p^2)} - 2\delta - \alpha\lambda + \omega\lambda p \end{aligned}$$

which is equivalent to:

$$(2\delta + \alpha\lambda) \sqrt{\lambda(4\delta\alpha + \alpha^2\lambda + 2\alpha\omega\lambda p + \omega^2\lambda p^2)} > 4\delta\alpha\lambda + \alpha^2\lambda^2 + \alpha\omega\lambda^2 p$$

That is

$$(2\delta + \alpha\lambda)^2 \lambda (4\delta\alpha + \alpha^2\lambda + 2\alpha\omega\lambda p + \omega^2\lambda p^2) > (4\delta\alpha\lambda + \alpha^2\lambda^2 + \alpha\omega\lambda^2 p)^2$$

It is easily verified that this inequality is always true.

By differentiating (11), we obtain

$$2\omega p\lambda^2 \frac{\partial m^*}{\partial \lambda} = \frac{2\delta\alpha\lambda + \alpha^2\lambda^2 + 2\alpha\omega\lambda^2 p + \omega^2\lambda^2 p^2}{\sqrt{\lambda(4\delta\alpha + \alpha^2\lambda + 2\alpha\omega\lambda p + \omega^2\lambda p^2)}} - \sqrt{\lambda(4\delta\alpha + \alpha^2\lambda + 2\alpha\omega\lambda p + \omega^2\lambda p^2)} + 2\delta$$

Thus

$$\begin{aligned} & \frac{\partial m^*}{\partial \lambda} > 0 \\ \Leftrightarrow & \frac{2\delta\alpha\lambda + \alpha^2\lambda^2 + 2\alpha\omega\lambda^2 p + \omega^2\lambda^2 p^2}{\sqrt{\lambda(4\delta\alpha + \alpha^2\lambda + 2\alpha\omega\lambda p + \omega^2\lambda p^2)}} + 2\delta > \sqrt{\lambda(4\delta\alpha + \alpha^2\lambda + 2\alpha\omega\lambda p + \omega^2\lambda p^2)} \end{aligned}$$

This is equivalent to

$$\sqrt{\lambda(4\delta\alpha + \alpha^2\lambda + 2\alpha\omega\lambda p + \omega^2\lambda p^2)} > \alpha\lambda$$

which are clearly always true .

Moreover, by differentiating (11), we obtain

$$2\omega\lambda p \frac{\partial m^*}{\partial \delta} = \frac{2\alpha\lambda}{\sqrt{\lambda(4\delta\alpha + \alpha^2\lambda + 2\alpha\omega\lambda p + \omega^2\lambda p^2)}} - 2\delta$$

Thus

$$\frac{\partial m^*}{\partial \delta} < 0 \Leftrightarrow \frac{\alpha\lambda}{\sqrt{\lambda(4\delta\alpha + \alpha^2\lambda + 2\alpha\omega\lambda p + \omega^2\lambda p^2)}} < \delta$$

It is easily verified that this inequality is always true. ■

Finally, let us now calculate  $\frac{\partial^2 m^*}{\partial \alpha \partial \omega}$ . By differentiating (24), we obtain:

$$\begin{aligned} & 2\lambda p \frac{\partial^2 m^*}{\partial \alpha \partial \omega} = \\ = & \frac{\omega\lambda^2 p \sqrt{\lambda(4\delta\alpha + \alpha^2\lambda + 2\alpha\omega\lambda p + \omega^2\lambda p^2)} - \frac{(2\delta\lambda + \alpha\lambda^2 + \omega\lambda^2 p)[\lambda(4\delta\alpha + \alpha^2\lambda + 2\alpha\omega\lambda p + \omega^2\lambda p^2) + \omega\lambda^2 p(\alpha + \omega p)]}{\sqrt{\lambda(4\delta\alpha + \alpha^2\lambda + 2\alpha\omega\lambda p + \omega^2\lambda p^2)}}}{\omega^2\lambda(4\delta\alpha + \alpha^2\lambda + 2\alpha\omega\lambda p + \omega^2\lambda p^2)} + \frac{\lambda}{\omega^2} \end{aligned}$$

This implies that

$$\begin{aligned} & \frac{\partial^2 m^*}{\partial \alpha \partial \omega} \geq 0 \\ \Leftrightarrow & \frac{\omega\lambda^2 p \sqrt{\lambda(4\delta\alpha + \alpha^2\lambda + 2\alpha\omega\lambda p + \omega^2\lambda p^2)} - \frac{(2\delta\lambda + \alpha\lambda^2 + \omega\lambda^2 p)[\lambda(4\delta\alpha + \alpha^2\lambda + 2\alpha\omega\lambda p + \omega^2\lambda p^2) + \omega\lambda^2 p(\alpha + \omega p)]}{\sqrt{\lambda(4\delta\alpha + \alpha^2\lambda + 2\alpha\omega\lambda p + \omega^2\lambda p^2)}}}{\omega^2\lambda(4\delta\alpha + \alpha^2\lambda + 2\alpha\omega\lambda p + \omega^2\lambda p^2)} \end{aligned}$$

$$+\frac{\lambda}{\omega^2} \stackrel{\approx}{\geq} 0$$

This is equivalent to:

$$\begin{aligned} & \frac{\omega\lambda p\sqrt{\lambda(4\delta\alpha + \alpha^2\lambda + 2\alpha\omega\lambda p + \omega^2\lambda p^2)} + \lambda(4\delta\alpha + \alpha^2\lambda + 2\alpha\omega\lambda p + \omega^2\lambda p^2)}{\omega^2(4\delta\alpha + \alpha^2\lambda + 2\alpha\omega\lambda p + \omega^2\lambda p^2)} \\ \stackrel{\approx}{\geq} & \frac{(2\delta\lambda + \alpha\lambda^2 + \omega\lambda^2 p) [\lambda(4\delta\alpha + \alpha^2\lambda + 2\alpha\omega\lambda p + \omega^2\lambda p^2) + \omega\lambda^2 p(\alpha + \omega p)]}{\omega^2\lambda(4\delta\alpha + \alpha^2\lambda + 2\alpha\omega\lambda p + \omega^2\lambda p^2)\sqrt{\lambda(4\delta\alpha + \alpha^2\lambda + 2\alpha\omega\lambda p + \omega^2\lambda p^2)}} \end{aligned}$$

or equivalently:

$$\begin{aligned} & \omega\lambda^2 p(4\delta\alpha + \alpha^2\lambda + 2\alpha\omega\lambda p + \omega^2\lambda p^2) \\ & + \lambda(4\delta\alpha + \alpha^2\lambda + 2\alpha\omega\lambda p + \omega^2\lambda p^2)\sqrt{\lambda(4\delta\alpha + \alpha^2\lambda + 2\alpha\omega\lambda p + \omega^2\lambda p^2)} \\ \stackrel{\approx}{\geq} & (4\delta\alpha + \alpha^2\lambda + 2\alpha\omega\lambda p + \omega^2\lambda p^2)(2\delta\lambda + \alpha\lambda^2 + \omega\lambda^2 p) \\ & + \omega\lambda p(2\delta\lambda + \alpha\lambda^2 + \omega\lambda^2 p)(\alpha + \omega p) \end{aligned}$$

It is clearly impossible to sign this inequality and thus  $\frac{\partial^2 m^*}{\partial \alpha \partial \omega}$  has an ambiguous sign.

## Further Statistics and Empirical Results

Table A1: Weak and strong ties in migration - full estimates

	1	2	3	4	5
$S_i^m$ : strong ties migration	0.1048*** (0.0403)		0.1014** (0.0394)	0.1008** (0.0398)	0.1049*** (0.0392)
$W_v^m$ : weak ties migration		0.2730*** (0.0762)	0.2640*** (0.0751)	0.2558*** (0.0739)	0.2903*** (0.0771)
$S_i^m \times W_v^m$					1.1910*** (0.4600)
$S_i$ : friend	-0.0128 (0.0144)		-0.0089 (0.0142)	-0.0112 (0.0142)	-0.0097 (0.0143)
$S_i$ : married	0.0323 (0.0231)		0.0382 (0.0234)	0.0418* (0.0235)	0.0457** (0.0226)
$S_i$ : high education	-0.0162 (0.0143)		-0.0133 (0.0143)	-0.0044 (0.0142)	-0.0033 (0.0143)
$W_v$ : village income		0.0014 (0.0023)	0.0012 (0.0024)	-0.0007 (0.0024)	-0.0006 (0.0024)
$W_v$ : village expenditure		-0.0020 (0.0054)	-0.0007 (0.0055)	0.0047 (0.0055)	0.0044 (0.0054)
$W_v$ : investment in health/education		0.0103 (0.0138)	0.0140 (0.0139)	0.0080 (0.0137)	0.0097 (0.0136)
$W_v$ : in poverty alleviation county		0.0402 (0.0247)	0.0410* (0.0245)	0.0328 (0.0241)	0.0346 (0.0240)
$W_v$ : grain subsidy		0.0001 (0.0004)	0.0001 (0.0005)	0.0004 (0.0004)	0.0004 (0.0004)
$W_v$ : presence of primary school		-0.0263** (0.0132)	-0.0277** (0.0133)	-0.0296** (0.0132)	-0.0327** (0.0133)
$W_v$ : workforce in agriculture		0.0383 (0.0322)	0.0454 (0.0325)	0.0451 (0.0321)	0.0500 (0.0317)
$W_v$ : distance from train station >5km		0.0135 (0.0144)	0.0096 (0.0144)	0.0090 (0.0143)	0.0059 (0.0141)
$\delta$ : return migration rate				-0.0990** (0.0415)	-0.0996** (0.0414)
$\lambda$ : migrants using networks				1.7495*** (0.4236)	1.7875*** (0.4230)
Male	0.0103 (0.0126)	0.0107 (0.0127)	0.0114 (0.0126)	0.0109 (0.0126)	0.0118 (0.0126)
Age	-0.0085*** (0.0017)	-0.0086*** (0.0016)	-0.0085*** (0.0016)	-0.0084*** (0.0016)	-0.0084*** (0.0016)
Birth order	0.0072 (0.0053)	0.0062 (0.0053)	0.0064 (0.0053)	0.0063 (0.0053)	0.0065 (0.0053)
Married	0.0051 (0.0200)	0.0072 (0.0205)	0.0072 (0.0205)	0.0042 (0.0205)	0.0051 (0.0205)
Number of children	-0.0132 (0.0088)	-0.0123 (0.0091)	-0.0127 (0.0092)	-0.0128 (0.0091)	-0.0121 (0.0091)
Han ethnicity	-0.0047 (0.0651)	0.0071 (0.0635)	0.0157 (0.0655)	-0.0928 (0.0740)	-0.0893 (0.0760)
Has urban hukou	-0.0287* (0.0162)	-0.0226 (0.0160)	-0.0216 (0.0162)	-0.0208 (0.0158)	-0.0176 (0.0156)
Above 8 years of schooling	0.0114 (0.0134)	0.0080 (0.0134)	0.0115 (0.0136)	0.0099 (0.0137)	0.0105 (0.0137)
Has work training	0.0085 (0.0163)	0.0087 (0.0161)	0.0067 (0.0162)	0.0067 (0.0161)	0.0054 (0.0160)
Farmer	0.0499*** (0.0145)	0.0398*** (0.0152)	0.0394*** (0.0153)	0.0406*** (0.0150)	0.0386*** (0.0150)
Health above average level	0.0030 (0.0169)	0.0078 (0.0174)	0.0110 (0.0171)	0.0118 (0.0170)	0.0134 (0.0171)
BMI	0.2840 (0.2162)	0.2661 (0.2147)	0.2685 (0.2167)	0.3270 (0.2174)	0.2956 (0.2166)
Household size	0.0081 (0.0054)	0.0063 (0.0056)	0.0063 (0.0056)	0.0056 (0.0055)	0.0054 (0.0055)
Number of elderly in the household	0.0093 (0.0156)	0.0067 (0.0151)	0.0071 (0.0152)	0.0076 (0.0151)	0.0085 (0.0150)
Household land	0.0038** (0.0019)	0.0041** (0.0019)	0.0040** (0.0019)	0.0035* (0.0019)	0.0036* (0.0019)
Housing size	-0.0025 (0.0073)	0.0010 (0.0073)	0.0005 (0.0074)	-0.0014 (0.0076)	-0.0030 (0.0074)
Household labor income	0.0018 (0.0020)	0.0024 (0.0020)	0.0025 (0.0020)	0.0024 (0.0020)	0.0024 (0.0020)
Province fixed effects	Y	Y	Y	Y	Y
$R^2$	0.09	0.10	0.10	0.11	0.12
N	2073	2073	2073	2073	2073

Source: RUMiC data, RHS waves I and II. The dependent variable is from wave II; all control variables are from wave I. Sample is composed by individuals aged 16 to 35 who have never migrated before. Dependent variable is a binary variable indicating whether the individual migrates in 2008. Regressions estimated with linear probability models. Robust bootstrapped standard errors in parentheses. \*/\*\*/\*\*\*/ indicate significance at the 0.1/0.05/0.01 level.  $S_i^m / W_v^m$  is the share of strong / weak ties who migrated.  $S_i / W_v$  refer to other characteristics of the strong / weak ties.

Table A2: Robustness checks

	Without $S_i$ and $W_v$	All ages	Quadratic $S_i^m$ and $W_v^m$	$S_i^m$ and $W_v^m$ in level
$S_i^m$ : strong ties migration	0.0931** (0.0381)	0.0133* (0.0068)	0.1779 (0.1323)	0.0589*** (0.0162)
$W_v^m$ : weak ties migration	0.2877*** (0.0723)	0.0669*** (0.0179)	0.2619* (0.1440)	0.0603*** (0.0160)
$S_i^m \times W_v^m$	1.1049** (0.4388)	0.2576** (0.1111)	1.1933** (0.4636)	0.1171*** (0.0429)
$(S_i^m)^2$			-0.0872 (0.1497)	
$(W_v^m)^2$			0.0816 (0.3733)	
$R^2$	0.11	0.06	0.12	0.12
N	2073	11333	2073	2073

Source: RUMiC data, RHS waves I and II. All regressions except those in the first column include the covariates of Table 3. Robust bootstrapped standard errors in parentheses. \*/\*\*/\*\* indicate significance at the 0.1/0.05/0.01 level.

Table A3: Summary statistics - by migration status of strong ties

	Individual characteristics		Network characteristics		
	$S_i^m = 0$	$S_i^m = 1$	$S_i^m = 0$	$S_i^m = 1$	
Migrates in 2008	0.077 (0.267)	0.145# (0.353)	$S_i^m$ : strong ties migration <sup>+</sup>	0.000 (0.000)	0.556# (0.292)
Male*	0.438 (0.496)	0.380 (0.487)	$S_i^c$ : relative <sup>+</sup>	0.311 (0.402)	0.361 (0.402)
Age	27.194 (5.507)	27.055 (5.638)	$S_i^c$ : married <sup>+</sup>	0.891 (0.274)	0.850 (0.300)
Birth order	1.920 (1.149)	1.960 (1.287)	$S_i^c$ : high education (> 8 years) <sup>+</sup>	0.522 (0.428)	0.646# (0.368)
Married*	0.620 (0.486)	0.630 (0.484)	$W_v^m$ : weak ties migration <sup>+</sup>	0.139 (0.088)	0.154# (0.090)
Number of children	0.776 (0.836)	0.830 (0.784)	$W_v^c$ : village income (RMB/100000)	0.868 (3.037)	0.492# (1.226)
Han ethnicity*	0.993 (0.083)	0.980 (0.140)	$W_v^c$ : village expenditure (RMB/100000)	0.535 (1.526)	0.369# (0.742)
Urban hukou*	0.080 (0.272)	0.105 (0.307)	$W_v^c$ : investment in health/education*	0.502 (0.500)	0.420# (0.495)
Above 8 years of education*	0.582 (0.493)	0.560 (0.498)	$W_v^c$ : located in poverty alleviation county*	0.133 (0.340)	0.150 (0.358)
Work training*	0.143 (0.350)	0.195 (0.397)	$W_v^c$ : grain subsidy (RMB)	19.862 (14.453)	16.655# (12.077)
Farmer*	0.539 (0.499)	0.500 (0.501)	$W_v^c$ : presence of primary school*	0.514 (0.500)	0.490 (0.501)
Good or excellent health*	0.894 (0.308)	0.850 (0.358)	$W_v^c$ : workforce in agriculture <sup>+</sup>	0.491 (0.247)	0.474 (0.237)
BMI (×100)	0.216 (0.026)	0.217 (0.024)	$W_v^c$ : distance from train station >5Km*	0.300 (0.458)	0.365 (0.483)
Household size	4.544 (1.284)	4.495 (1.216)			
Number of elderly in household (age > 64)	0.178 (0.453)	0.170 (0.450)	$\delta_v$ : return migration <sup>+</sup>	0.063 (0.108)	0.092# (0.120)
Household land (Mu)	4.170 (4.383)	4.693 (5.171)			
Housing size (m <sup>2</sup> /100)	1.480 (0.765)	1.540 (1.189)	$\lambda_d$ : migrants using networks <sup>+</sup>	0.191 (0.049)	0.190 (0.050)
Household labor income (ln RMB)	7.785 (3.089)	7.983 (2.790)			
<b>Provinces</b>					
Hebei*	0.064 (0.244)	0.085 (0.280)	Hubei*	0.093 (0.290)	0.085 (0.280)
Jiangsu*	0.146 (0.353)	0.100# (0.301)	Guangdong*	0.141 (0.348)	0.170 (0.377)
Zhejiang*	0.130 (0.336)	0.185 (0.389)	Chongqing*	0.056 (0.230)	0.020# (0.140)
Anhui*	0.081 (0.272)	0.125 (0.332)	Sichuan*	0.105 (0.306)	0.135 (0.343)
Henan*	0.186 (0.389)	0.095# (0.294)			

Observations: 1,873 for  $S_i^m = 0$  and 200 for  $S_i^m = 1$ . Source: RUMiC data, RHS waves I and II. All variables refer to wave I except "Migrates in 2008" which refers to wave II. Standard deviations in parentheses.

(\*) refers to 0/1 variables. (+) refers to shares.

$S_i^m / W_v^m$  refers to strong / weak ties who migrated.  $S_i / W_v$  refers to other characteristics of the strong / weak ties. Strong ties refer to the five closest persons in the network; Weak ties refer to village. # indicates that a t-test for the difference of two means is significant at the 5% level.