Published in final edited form as:

Physica A. 2018 January 15; 490: 1061–1075. doi:10.1016/j.physa.2017.08.145.

Agent-Based Modeling of China's Rural-Urban Migration and Social Network Structure

Zhaohao Fu^{a,*} and Lingxin Hao^b

^aDepartment of Civil Engineering, Johns Hopkins University

^bDepartment of Sociology, Johns Hopkins University

Abstract

We analyze China's rural-urban migration and endogenous social network structures using agent-based modeling. The agents from census micro data are located in their rural origin with an empirical-estimated prior propensity to move. The population-scale social network is a hybrid one, combining observed family ties and locations of the origin with a parameter space calibrated from census, survey and aggregate data and sampled using a stepwise Latin Hypercube Sampling method. At monthly intervals, some agents migrate and these migratory acts change the social network by turning within-nonmigrant connections to between-migrant-nonmigrant connections, turning local connections to nonlocal connections, and adding among-migrant connections. In turn, the changing social network structure updates migratory propensities of those well-connected nonmigrants who become more likely to move. These two processes iterate over time. Using a core-periphery method developed from the k -core decomposition method, we identify and quantify the network structural changes and map these changes with the migration acceleration patterns. We conclude that network structural changes are essential for explaining migration acceleration observed in China during the 1995–2000 period.

Keywords

Agent-based model; k -core; rural-urban migration; China

1. Introduction

Rural-urban migration is a persistent phenomenon in developing countries starting from the middle of the last century [1]. From the 1980s to 2010, China's rural-urban migration has grown rapidly to become the largest single migration in human history [2]. Like other developing countries, China's rural migration results from decades of urban-biased development policies that make urban destinations attractive to the rural population [3]. China's rural-urban migration, however, differs from other developing countries in a number of distinctive ways [4]. First, China received the largest amount of foreign-direct investment (FDI) in the 1990s and this exogenous shock was a necessary condition for its unprecedented rural-urban migration. Second, China's extremely low arable land to labor

*Corresponding author. zfu7@jhu.edu (Zhaohao Fu).

ratio created a surplus of rural labor that was increased by the further loss of arable land to the FDI-supported expansion of industry and the environmental degradation of land by industry [5]. This paper seeks to contribute to the literature by modeling China's rural-urban migration using agent-based modeling (ABM) to better understand generalized migration patterns.

Generalized migration patterns are based on individual behaviors influenced by social networks. These social networks are in turn shaped by individuals' migration. This interdependent relationship between individual behavior and social networks is posited in the theory of cumulative causation of migration, which states that social networks connecting migrants to nonmigrants make migration self-perpetuating [6]. Modeling the co-evolution of social networks and migration perpetuation is our central theme. Computational models capable explaining the co-evolution of migration and social networks have been proposed and applied to the case of Brazil [7].

We consider migration as a complex system where individual decision making, social interactions, and local economies interact in complex ways. Yet, we know very little about how network structure changes during the migration process. In particular, there is little known about how the sparse structure of the initial social network transforms into the cohesive structure of the mature social network that promotes further migration. In this paper, we examine a 5-year period (1995–2000) when Chinese rural-urban migration accelerated in order to understand whether and how the endogenous social network develops and boosts migration.

In our study, the endogenous social network may be a driving force of migration acceleration at a quadratic rate. We use a hybrid model that combines observational data and a computational approach to build and simulate the population-scale social network. Empirically we draw micro data from the census on the family ties and spatial locations of agents as well as the distributions of local families. Computationally we optimize empirically calibrated parameters with respect to common social network properties using Latin Hypercube Sampling method [8]. The resulting probabilities are used to connect agents between families and villages. We incorporate recent studies on agent-based economics, which provide a perspective to study two-way feedback between micro-structure and macro-structure [9]. We use agent based modeling (ABM) as a complementary approach to the statistical approach widely used in the migration literature because it allows us to better understand social phenomena from interactions among actors [10]. ABM and social interaction have been applied to studies on rural-urban migration in industrializing countries [11, 12], but they have not incorporated empirical-based social networks.

In this paper we are particularly interested in how the endogenous social network helps explain the accelerating migrant stock during the 1995–2000 period. To this end, we compare results from three ABM models. First the base model involves no social networks but specifies migration decisions as being influenced by the economy and individual demographic characteristics (*Model 1*). The second model introduces the implicit social network using the migration prevalence of the origin (*Model 2*). Our third model embeds individual migratory decisions in explicit social networks with the evolution of social

interactions (*Model 3*). *Model 3* maps the social network structure changes onto the migration acceleration patterns. Through the model comparisons we are able to show whether and how social network structures drive migration acceleration and, at the same time, how migratory acts transform the initially sparse social network into a cohesive social network. To our knowledge, this paper is the first attempt to use k -core decomposition and the further-developed core-periphery structure to understand migration.

The paper is organized as follows. Section 2 introduces the mathematical formulation of our three ABM models as well as the procedure to generate the hybrid social networks, accompanied by pseudo codes and flowcharts. Section 3 introduces our optimization approach to parameter space sampling by specifying the multiple objectives in the stepwise Latin Hypercube Sampling (LHS method). Section 4 introduces our network structure measure. The analysis of the scenarios generated by our three models is shown in Section 5. Concluding remarks are offered in Section 6.

2. Model Formulation

Our model features both empirical grounding and agent-based simulation based on important assumptions and probability theory. It contains the essential information about household membership and demographic information, while retaining the randomness of connection and migratory propensity. The empirical data comes from census and national surveys¹. Section 2.2 shows how the social network is built with both the actual family ties from the census data and randomly generated social connections according to social norms governing the family, kinship, and friendship. Our simulation-LHS iterative procedure, detailed in Section 3, calibrates network parameters while sampling sets of parameters. Our model incorporates uncertainty of parameters by using distributions of parameters. Examples include the uniform distribution of migration propensity of groups of individuals defined by demographic characteristics and the rules for decisions and moves calculated based on substantive theories (causes of migration) and probability theories, which are described in Section 2.3. To begin, we lay out the basic ABM setup in Section 2.1.

2.1. ABM Setup

There are two types of acting agents: individuals and provinces. Individual agents include nonmigrants and migrants in 1995. The provinces are divided into rural and urban areas: the rural area provides the geographic identification for the origin of individual agents; the urban area gives the identification for the destination of migrants.

Individual Agents—The model reads in 313,342 individual records with the actual age (11–44 years old), sex (male/female), years of schooling (4–14 years), family ID, the original province, as well as the destination province if migrated.

¹Our empirical data on China's rural-origin population were drawn from China 2000 Census' 1/1000 micro data (the first stage was at the household level and the final stage is to take all individuals in the sampled households). The census collected individual-level data on the current province of residence, the province of household registration (the origin province, or "hukou"), migrant status and migration history in the past five years, demographic characteristics, and household members. Individuals who out-migrated without an accompanying family member by 2000 are randomly assigned to a non-migrant family of the same origin province such that the distribution of family sizes of each origin province is constant.

Provinces—The model reads in 31 provinces in China (not including Taiwan, Hong Kong and Macau). The local economy of provincial urban areas is measured using the amount of foreign direct investment (FDI). The local economy of provincial rural areas is measured using the amount of per capita arable land. Because the rapidly growing FDI has created demand for low-cost labor [5], we consider urban FDI and rural land scarcity as the two major economic forces behind China's rural-urban migration.

2.2. Hybrid Social Network Setup

The actual population-scale social network is not directly observed. Our social network setup is a hybrid combining observational data and probability theory. We exploit the available observational information from census and national surveys as our empirical sources. In *Model 3*, we build the explicit social network through the family ties in 1995, derived from the 2000 Census, and construct social networks beyond the family ties based on probability theory and the empirical within-province family-size distributions and village-size distributions. We also set up a social network before rural-urban migration was allowed, a time we label pre-1995.

The social network G contains vertices V and edges E :

$$G=(V, E) \quad (1)$$

$$V=\{v_i, i=1, 2, \dots, 313342\} \quad (2)$$

$$E=E_1 \cup E_2 \cup E_3 \cup E_4 \quad (3)$$

where v_i stands for the vertex corresponding to agent i , and E_k are subsets of edges where the index k represents one of four layers of edges. Starting from the unconnected smallest component (family ties E_1), we sequentially add network edges. Beyond the family individuals are sequentially connected to the larger network: within villages E_2 , within provinces E_3 , and between migrants of the same destination province E_4 . For the different layers of edges, we use the probability of connection to keep the desired distribution of the degree (the number of connections for an individual).

Family ties E_1 —Nonmigrants are connected within the same origin province based on the observed family ties from the census micro data. Families are treated as complete graphs (each family member is linked with all others in the family).

Within villages E_2 —Village sizes are assigned based on the actual distribution of labor force counts of villages from community surveys and the estimated province-specific distribution of village sizes. Families in the same village are linked in two ways: (a) relatives and friends within villages: for each family we randomly assigned κ_2 relatives and friends,

who are randomly chosen from other families. For each family in a village, the probability of connecting to any other families is

$$P(E_2) = \frac{\kappa_2}{N_2^{\eta_2}} \quad (4)$$

where N_2 is the number of families in this village and $\eta_2 < 1$ is a scale factor to reduce the variation in the average degree across villages of different sizes. Without the scale factor, the larger the village is, the smaller $P(E_2)$ will be. (b) Transitivity: a probability ν is given to each of two connected vertices v_a and v_b to enclose a randomly chosen open triple, i.e., to connect to a third vertex v_c such that the complete graph of v_a , v_b and v_c is formed.

Between villages E_3 —The between-village network is built at the province level. The average degree between villages is strictly fewer than that within villages. Social ties between villages of a same province are, again, enabled in two ways: (a) relatives and friends in different villages: For each family in a village, the probability of connecting to any families in different villages within provinces is

$$P(E_3) = \frac{\kappa_3}{N_3^{\eta_3}} \quad (5)$$

where N_3 is the number of total families in this province and $\eta_3 < 1$ is a scale factor. (b) Transitivity: a probability of ν is applied to enclose a randomly chosen open triple for each of two connected vertices.

Between migrants E_4 —Edges are built to connect acquaintances among migrants in the same destination province from different origin provinces. Two ways are used to create these edges: (a) new acquaintances: with a certain probability a new migrant is connected to other migrants. For any agent migrated to a destination province, the probability of a new migrant acquainting with the earlier-migrated agents is

$$P(E_4) = \frac{\kappa_4}{N_4^{\eta_4}} \quad (6)$$

where N_4 is the number of migrants in the urban area of a province and $\eta_4 < 1$ is a scale factor. (b) Transitivity: those nonmigrant family members of a migrant are linked with his/her new acquaintances. If two migrants are connected, then each one has a ν probability to connect to one of his/her paired agent's family members.

The algorithm for building the network is attached here with a subgraph of hybrid social network shown in Figure 1. The subgraph contains two migrants in Guangdong from Sichuan and Hunan respectively. This graph shows how the four layers of edges connect the whole country.

```

Read in  $V$  and  $E_1$ . Initialize  $E_2, E_3, E_4 \leftarrow \emptyset$ ;
for  $i \leftarrow 1$  to  $N_1$  in  $E_1$  do
    if  $V_i$  is an single migrant then
        Set to a family from the same origin province by the family sizes
        distribution;
        Update  $E_1$ ;
    end
end
for  $i \leftarrow 1$  to  $N_2$  do
    for  $j \leftarrow i + 1$  to  $N_2$  do
        Connect Family  $i$  and Family  $j$  by probability  $P(E_2)$ ;
        if Family  $i$  and Family  $j$  are connected then
            Each family chosen an agent to have a  $\nu$  probability to
            connect to one of his/her paired agent's family members;
        end
    end
end
for  $i \leftarrow 1$  to  $N_3$  do
    for  $j \leftarrow i + 1$  to  $N_3$  do
        if  $i$  and  $j$  are not in the same village then
            Connect Family  $i$  and Family  $j$  by probability  $P(E_3)$ ;
            if Family  $i$  and Family  $j$  are connected then
                Each family chosen an agent to have a  $\nu$  probability to
                connect to one of his/her paired agent's family members;
            end
        end
    end
end

```

```

for all the provinces do
  for  $i \leftarrow 1$  to  $N_4$  in the same provinces do
    for  $j \leftarrow i + 1$  to  $N_4$  in the same provinces do
      Connect Migrant  $i$  and Migrant  $j$  by probability  $P(E_4)$ ;
      if Migrant  $i$  and Migrant  $j$  are connected then
        Each migrant chosen an agent to have a  $\nu$  probability to
        connect to one of his/her paired agent's family members;
      end
    end
  end
end

 $G(V, E) \leftarrow G(V, E_1 \cup E_2 \cup E_3 \cup E_4);$ 

```

2.3. Simulation Process

We simulated migratory decisions and actions as well as the social network growth and decay in monthly time ticks for 60 months from January 1996 to December 2000. Below we first lay out four decision-making rules regarding agents' migratory decisions and then the rules for network growth and decay. The first rule for migration defines the monthly migration threshold for *Model 1*, *Model 2* and *Model 3*. We use the migration proportion in

month t , $\frac{B_t}{A_0}$, as the constraints, where A_0 is the total rural agents in the initial month 0, B_t is the total migrant counts (stock) in month t . The total migrant stock trend is represented using a quadratic growth function. Then the migratory threshold at time t is:

$$\tau_t = 1 - \frac{B_t}{A_0} = 1 - \frac{B_0 + \delta_1 t + \delta_2 t^2}{A_0} \quad (7)$$

where the parameters δ_1 and δ_2 are fitted by annual total migrant stock observed in the census with the constraints of initial and ending total rural labor.

The second rule concerns individual migratory propensities. Agents are classified in three migratory propensity groups according to their demographic attributes. Individual agents' propensities to migrate are randomly assigned within the three migratory propensity groups. The three groups are given different uniform distributions of migration propensities. The high-propensity group (Group 1) is younger (11–30), male, and higher educated (9–14). The low-propensity group (Group 3) is older (31–44), female, and lower educated (4–8). All the rest are classified to a middle-propensity group (Group 2). Each nonmigrant i is assigned an initial migratory propensity score *propensity_i*, randomly within the migratory propensity group he or she belongs.

$$\text{propensity}_i \sim U[l_k, 1], v_i \in \text{Group-} k, k=1, 2, 3 \quad (8)$$

where $U[l_k, 1]$ is a uniform distribution with a common upper bound of 1 and a specific lower bound l_1, l_2, l_3 , which are calibrated so that the total migrant stock in 2000 is the same as observed in the census. The initial migratory propensity will be updated according to the next two rules.

The third rule regards the two competing influences of social networks, denoted by F_1 of non-migrants, termed local-coherent influence, and F_2 of migrants, termed migration-promoting influence. *Model 1* has no social network influences on migratory decision making: $F_1(i, t) = F_2(i, t) = 0$. In *Model 2*, at a time tick each agent simply counts the migration stock in all destinations from the same origins. So

$$F_1(i, t) = -\alpha_2 \quad (9)$$

$$F_2(i, t) = \beta_2 \sum_{j \in M_i(t)} M_{ij} \quad (10)$$

where the $F_1(i, t)$ is a constant α_2 so calibrated in order to make the total migration stock grows as the census observed.

In *Model 3*, let $d_u(i, j, t)$ be the geodesic distance from a nonmigrant i to a migrant j everywhere, $j \in M(t)$, where $M(t)$ denotes the migrants set at time t ; $d_r(i, j, t)$ be the geodesic distance from a nonmigrant i to another nonmigrant j from the same origin of $i, j \in N_r$ where $N(t)$ denotes the non-migrants set in the same rural province with agent i at time t . Then the local-coherent influence at time t $F_1(i, t)$ is defined as:

$$F_1(i, t) = \alpha_3 \sum_{j \in N(t)} \frac{1}{2^{d_r(i, j, t) - 1}} \quad (11)$$

The migration-promoting influence at time t $F_2(i, t)$ is defined as:

$$F_2(i, t) = \beta_3 \sum_{j \in M_i(t)} \frac{1}{2^{d_u(i, j, t) - 1}} \quad (12)$$

where α_3 and β_3 are the influence factors respectively. We compute the influence from one vertex to all the others using a breadth-first search (BFS). Figure 2 is a conceptual visualization how BFS works for a part of network in Figure 1. Starting from vertex 1 in the graph, the agent reaches all others agents following the tree-structure paths. Due to

computational constraint for this population-scale network, we limit the search range within $d = 3$.

The fourth rule defines the two-step migration decision. Individual i 's propensity is updated by the influence of arable land per capita $land_{pt}$ in the rural area of the origin province p at time t . The influence factor is

$$F_3(i, t) = (1 - (1 - \varphi \frac{land_{pt}}{\sum_q land_{qt}})) \quad (13)$$

where φ reflects how the economic rationality of rural people alters their migratory propensity given land scarcity. In addition, the current social network influences (in *Model 2* and *Model 3*) update a nonmigrant's migratory propensity. Taken together, for each nonmigrant at month t during the simulation, the migratory propensity score is updated:

$$P(i) = propensity_i (1 - F_1(i, t) + F_2(i, t)) (1 + F_3(i, t)) \quad (14)$$

If the updated migratory propensity is greater than the migratory threshold τ_t then the agent decides to migrate out of the origin. We assign a probability ρ to determine whether individuals would have full information on the Foreign Direct Investment (FDI) in destinations. In *Model 1*, $ER(p, t) = 0$, where $ER(p, t)$ is the total migration effect of rural province p . This means that individuals will choose the destination purely based on FDI information when full information on FDI is available, otherwise they will randomly choose a destination. In *Model 2* the network influence is measured using origin migration prevalence, $ER(p, t) = \sum_{j \in M_p(t)}$ and we use the effect factor ω_2 . In *Model 3*, for those agents who have the FDI information, the destination decision is to maximize a function of relative FDI amounts and migrant network influences to solve for the p^* -th province as the destination:

$$EF(p, t) = \sum_{j \in M_p(t)} \frac{1}{2^{d_u(i, j, t) - 1}} \quad (15)$$

$$p^* = \arg \max (1 + \omega_3 \times EF(p, t)) \frac{FDI_{p, t}}{\sum_q FDI_{q, t}} \quad (16)$$

Where $FDI_{p, t}$ is the FDI amount for urban area p in the previous year, ω_3 is the network influence factor. For those who do not have full FDI information, the destination decision is to maximize a function of migrant network influences to solve for the p^* -th province as the destination,

$$p^* = \arg \max (1 + \omega_3 \times \text{EF}(p, t)) \quad (17)$$

The social networks will grow correspondingly to the increasing migration over time through their acquaintances with earlier migrants. The process follows the setup of between migrants E_4 , through which new migrants acquaint with earlier migrants. In the meanwhile, edges in the network G will decay if two connected agents are not in the same rural province. The decay happens whenever adding new edges through E_4 . The decay occurs mainly between migrants and non-migrants if they were linked in the origins through E_2 and E_3 . Migrants should have more urban connections than rural connections. We apply a probability θ to drop a randomly picked existing non-family connection initially built in the rural province after the agent migrated.

Our simulation of the process of social network evolution follows the road map shown in Figure 3. The algorithm for the whole simulation process (*Model 3*) is shown here:

```

for  $i \leftarrow 1$  to  $N$  do
    Set to Group  $m$ ;
     $propensity_i \leftarrow U[l_m, 1]$ 
end

Build up the social network (Only for Model 3);

while  $Step \leq 60$  do
     $\tau_t = 1 - \frac{B_t}{A_0} = 1 - \frac{B_0 + \delta_1 t + \delta_2 t^2}{A_0}$ ;
    for  $i \leftarrow 1$  to  $N$  do
         $F_1(i, t) = \alpha_3 \sum_{j \in N(i)} \frac{1}{2^{d_r(i, j, t)} - 1}$ ;
         $F_2(i, t) = \beta_3 \sum_{j \in M_i(t)} \frac{1}{2^{d_u(i, j, t)} - 1}$ ;
         $F_3(i, t) = (1 - (1 - \varphi \frac{land_{pt}}{\sum_q land_{qt}}))$ ;
         $P(i) = propensity_i (1 - F_1(i, t) + F_2(i, t))(1 + F_3(i, t))$ ;
        Assign rationality by probability  $\rho$ ;
        if  $P(i) > \tau_t$  then
            if rational then
                 $EF(p, t) = \sum_{j \in M_p(t)} \frac{1}{2^{d_r(i, j, t)} - 1}$ 
                 $p^* = \arg \max (1 + \omega_3 \times EF(p, t)) \frac{FDI_{p, t}}{\sum_q FDI_{q, t}}$ 
            else
                 $p^* = \arg \max (1 + \omega_3 \times EF(p, t))$ 
            end
            Connect to other migrants by probability  $P(E_2)$ ;
            if Two migrants are connected then
                Have  $\nu$  probability to connect to one of his/her paired
                agent's family members;
                Cut a randomly picked existing non-family connection at  $\theta$ ;
            end
        else
            Stay as non-migrant;
        end
    end
end

```

3. Parameter Space Optimization

Our social network as an observational-computational hybrid is built by reading in the census household data and applying probability theory to initial parameter bounds calibrated using census data, survey data and aggregate statistics. Beyond observed family connections E_1 , we build up E_2 , E_3 , E_4 with parameters space optimization in a manner such that the

whole network possesses the stable common properties of small-world social networks [13], which are taken as the objectives in the parameter space optimization and sampling process.

With the objective of maintaining the common social network properties, we sampled parameters associated with the influence of social network, FDI, and land use. To obtain the optimal parameter space that can both satisfy the social networks properties and simulate the co-processes of migration and network evolution, we use the stepwise Latin Hypercube Sampling (LHS) method. LHS is a statistical method for generating samples for multi-dimensional problems [8]. According to McKay et al. [14], by applying LHS method, we can reduce the variance of the models' results more than the Monte Carlo Method (MC). A stepwise procedure is used to optimize the multi-objectives of the small world properties. In each step we picked subspace of the parameters that satisfies all the objectives and sampled from the subspace in the next step.

The benefit of using stepwise LHS is to explore the complexity of the parameter sample space and the targeted common properties of small-world social networks efficiently. The relations between model results and the parameters are nonlinear, where parameters in some regions are sensitive to the model's results while changes in parameters in other regions are insignificant to the results. The LHS provides us the opportunity to explore all possible highly sensitive regions of the parameter space which MC could not. In MC, the parameters are sampled separately, making it possible for us to neglect some regions of the parameter space. Also, by using the stepwise method, we avoid the parameter space where the model's results are meaningless. We efficiently shrink the size of the space and make the final sample space more acceptable to the small-world social networks.

3.1. Small-world Social Network Properties as objectives

Degree distribution—In small-world social networks, the degree distribution of social networks should follow a power-law-like distribution, which is long tailed and scale free [15]. The frequency of degree approximates a power-law function, simplified to

$$f(k) = Ak^{-\gamma} \quad (18)$$

where k is the degree. Toivonen et al. (2006) suggested an exponent $\gamma > 3$ for social networks of size $N \sim 10^6$ [16]. Graphically we use the degree distribution in log-log scale, where the curve should be downward.

Clustering coefficients—Clustering is measured using a vector of the average clustering coefficient for each degree (the clustering coefficient for degree $k = 2$ is undefined because of the triangle units used in the clustering measure). Vertex v_i with degree k has its clustering coefficient defined as $C(v_i) = 2\lambda_G(v_i)/(k(k-1))$, where $\lambda_G(v_i)$ is triangle v_i

closed in G and the possible triangle v_i can close is $\frac{k(k-1)}{2}$ [15]. In social networks, clustering coefficients should be relatively high and have a negative relationship with the degrees.

Average geodesic distance—Geodesic distance is the length of the shortest path, which is the minimal length of path between two vertexes. The average geodesic distance considers all pairs of vertices. In small world social networks, the average geodesic distance (degree of separation) is approximately 6 [17].

3.2. Latin Hypercube Sampling

We denote the parameter set as X , the number of parameters in the set as N_{dim} and we want to obtain M samples from LHS. We first define the permutation matrix $P = p_{ik}$ to be an $N_{dim} \times M$ matrix, where each row of P is a random permutation of $\{1, 2, \dots, M\}$. Let $\xi_{ik}(i = 1, 2, \dots, N_{dim}, k = 1, 2, \dots, M)$ be $N_{dim}K$ independent and identically uniformly distributed on $[0,1]$. Then the sampled parameter $x_{ik}(x_{ik} \in X)$ is defined by

$$x_{ik} = F_i^{-1}((p_{ik} - 1 + \xi_{ik})/N_{dim}) \quad (19)$$

Where $F_i^{-1}()$ is the inverse cumulative distribution function of parameter x_i .

To scale down the number of samples that are required to get the optimal parameter sets without losing accuracy of the simulation, we use the stepwise optimization method. The optimization is divided into progressive steps where the space of the latter step is a subspace of the previous one. In each step, we choose one parameter to divide into two ranges and therefore the previous space is divided into two subspaces. Then we apply a T -test to determine whether the model's social networks properties using parameters sampled from the two subspaces fall in the small-world properties range – the γ and average distance.

We choose a proper boundary to divide the previous space to make sure that we can eliminate one of the subspaces in each step. In the end, if we shrink each parameter's initial

bound by $s_i, i = 1, 2, \dots, N_{dim}$ and $0 < s_i < 1$, the final space will be $S = \prod_i^{N_{dim}} s_i$. In this case, we shrink our initial space to 0.1% of the initial space.

The optimized results for the parameters are shown in Table 1. The process of the stepwise sampling is shown in Figure 4. In (a), for an exponent $\gamma > 3$ for social networks of size $N \sim 10^6$, we abandon all the samples with $\gamma < 3$. In (b), a middle example step of sampling is shown. The blue circles refer to the samples that satisfy the criteria. For example, we use an estimate to give the optimal region for κ_4 , which, in this case, is $4 - 5$.

In our model design, *Model 1* and *Model 2* do not involve social networks. So *Model 1* and *Model 2* follow the initial range for parameters we calibrated using empirical data and we do not perform stepwise optimization. We apply LHS to reduce the variance in *Model 3*.

We validate whether model results with the optimized parameters show the stability of the above properties in Figure 5. We generate the degree distribution in the log-log axes in Figure 5a, the curves are downward-sloped, which indicates that the degree approximates a power-law like distribution. The dash lines are the boundaries for the γ , which in our case is between 3 and 4, appropriate for our network size of 313,342. Figure 5b shows a relatively

stable downward slope of the average clustering coefficient. Figure 5c shows the average geodesic distance for pairs of all vertices in the whole network. While the trend is downward, the average geodesic distance is around 6, ranging from 5.55 to 6.55. This pattern confirms the relative stability of the property of six degree of separation. In sum, these stable network properties provide a baseline for us to simulate the social network process in conjunction with the migration process.

4. Measuring Social Network Structure

One important task of our ABM models is to identify and quantify changes in the social network structure. Based on the k -core decomposition method, we have developed the core-periphery measure to capture social network structural changes over the 60 monthly intervals.

k-core decomposition

The k -core decomposition of Graph $G = \{V, E\}$ defines that a subgraph $H = (C, E|C)$ induced by the set $C \subseteq V$ is a k -core if and only if $\forall v \in C: \text{degree}_H(v) \geq k$, and H is the maximum subgraph with this property. To obtain the subgraph of k -cores in a network, we remove from it any vertices that have degree less than k iteratively. We repeat this cutting process until the degree of all vertices in the subgraph is k or over.

Core-periphery structure

After identifying a core using k -core decomposition, we place this core back into the social network to identify its peripheral spheres which are made of vertices with geodesic distance $d=1,2,3$ with any core vertices. Our simulation over the 60 monthly ticks helps portray the changes in the core-peripheral structure. This core-periphery structure includes the core in the center with high connectivity within and the periphery to which the core radiates its influence in a manner shown in Figure 2. By analyzing these structural changes over time, we see a rising cohesive structure where migrants are more densely connected to form a core and this core radiates its influence to nonmigrants in the peripheral spheres around the core.

5. Results

The presented results are based on the average of outcomes from 20 simulations for *Model 1*, *Model 2* and *Model 3*, respectively. We organize the result section in the order of network outcomes, migration outcomes, and the interdependence of network outcomes and migration outcomes.

5.1. Network Results

Although the social network as a whole sustains the stable properties of small-world social networks, there are significant changes in the subgroups of migrants and nonmigrants as well as the network structure. In particular, the cumulative causation of migration suggests that network structure change is a key driver of migration dynamics.

In this paper, we have selected $k = 12$ as an appropriate value to define the network structure according to three criteria. The first criterion regards the migrant share of core vertices. A

larger migrant share of the core is consistent with the cumulative causation theory of migration that highlights the influence of migrants on nonmigrants' future moves. We wish to find a k such that it is as large as possible while the migrant share is maximized. As shown in Figure 6a, $k=12$ satisfies the first criterion. A second criterion states that the percentage of migrants in the core is sufficiently large. Figure 6b shows that if $k > 12$, the percentage of migrants in the core falls to nearly 0. Therefore k should not be greater than 12. A third criterion is that the core size should be reasonably substantial. Figure 6c shows that the 12-core covers about 5% of the total population. It is worthy noting that the k -decomposition ensures that higher k -cores are inside the lower k -cores. In our case, cores with $k > 12$ should be inside the 12-core. Based on the three criteria, the 12-core and periphery structure is used to study the relationship between social network structural change and migration acceleration.

We visualize the core-periphery structure through the connectivity of vertices of different distances to the core shown in Figure 7 in selected years: 1995 when the number of migrants was small, 1997 when migration started to accelerate, and 2000 when migrant stock accounted for nearly 10% of the population. For 1995, the number of vertices in the core-periphery structure is small. In 1997, there are more vertices. In 2000, in addition to more connections between the core and periphery, some previous peripheral vertices with $d=1$ became core vertices; and some vertices previously distant from the core became close to the core.

We capture the core-periphery size and distance trends in Figure 8 to show that the sparsely structured social network has transitioned to a cohesively structured social network. The core size does not show a substantial increase until 1998. The $d=1$ section grows slowly at the beginning and accelerates from 1997; the $d=2$ section peaks in 1999; and the $d=3$ section grows until 1997 and then declines. We clearly see that the size of the core and the $d=1$ peripheral sphere increases over time.

With these structural shifts documented, one may wonder how the whole network retains stable properties while experiencing substantial structural changes. We compare and contrast the network measures for core and non-core vertices in Figure 9. The trends show huge gaps between core and non-core vertices. In Figure 9a, the average degree among core vertices is much greater than that among non-core vertices, which is logical because the core vertices have at least degree of k . Figure 9b shows the clustering coefficient of core and non-core vertices; both follow a downward pattern as degree increases. The clustering coefficients of core vertices are nearly 1 at degree = 12 and decrease to 0.4 at degree = 18. In contrast, the clustering coefficients for non-core vertices start from 0.12 at degree = 12 and decrease to 0.05 at degree = 18. In Figure 9c, the geodesic distance for non-core vertices is above 6 with a declining trend, while the geodesic distance for core members increases from 1.3 to 2.8. The vertices in the core easily reach others and non-core vertices reach one another through core vertices. This within-network heterogeneity in the core-periphery structure does not negate the stability of the basic properties of the whole network.

5.2. Migration Results

Our agent-based model simulates individual migratory decisions and actions under the influence of social networks and local economic forces. We empirically validate the generated aggregate outcomes from *Model 1*, *Model 2*, and *Model 3* against the aggregate data from the census². Because our model incorporates probability theory to assign the migratory propensity, we want to validate the model by the expected aggregate outcomes instead of individuals' outcomes. First, we examine the comparisons of destination-specific migrant stock on China's map as shown in the three maps in Figure 10. This figure gives a straightforward sense as to how each model performs compared to the reality. The hot side (red) of the color scheme indicates under-fit; the cold side (blue) indicates over-fit; and the middle (white) indicates good-fit. *Model 1* leaves out social networks completely and shows the poorest fit. *Model 2* uses an implicit social network measure (the origin migratory prevalence), which improves the model fit to some degree. *Model 3* incorporates the explicit social network and shows the best fit among the three models as the map is nearly white. A closer look at the most popular destination - Guangdong in South China - shows bright red (very under-fit) in *Model 1*, less red in *Model 2*, and slight red in *Model 3*. The same under-fit pattern can be found in Fujian, Jiangsu, and Zhejiang, all of which are major destinations. In contrast, an over-fit pattern is found in major sending provinces, such as Hunan. The fit, however, improves from *Model 1*, to *Model 2*, and to *Model 3*.

Second, we validate the cell counts of migrants of the origin-destination matrix. The difference between the model's matrix $M_{\text{model}}^{31 \times 31}$ and census data's matrix $M_{\text{data}}^{31 \times 31}$ represents the relative accuracy of the designed model, where the destinations are listed vertically and origins horizontally. The difference matrix $\Delta M^{31 \times 31} = M_{\text{model}}^{31 \times 31} - M_{\text{data}}^{31 \times 31}$, with δ_{ij} is the corresponding item in matrix D . Three orders of the norm of this difference matrix are used in the evaluation: the first order norm evaluates the maximum difference in migrant counts by origins; the infinite order norm evaluates the maximum difference in migrant counts by destination; the Frobenius norm evaluates the overall difference for migrant counts by origin-destination cells. Also, we calculate the Pearson correlation of all the 31×31 values of δ_{ij} . The results are shown in Table 2. The results in the First Order Norm column are large without clear patterns, which suggests that none of the three models does a good job of explaining migration out of origins. In the Infinite order norm column, there appears to be evidence that *Model 3* performs better than *Model 2*, then *Model 1*. This suggests that the models do a relatively good job in predicting migrant stock in specific destinations. The Frobenius Norm column shows a pattern mainly reflecting the combined result. Also, the Pearson correlation confirms that *Model 3* provides the better fit, then *Model 2*, and then *Model 1*.

5.3. Interdependent Social Network Process and Migration Process

So far we have examined the results for network outcomes and migration outcomes separately. Guided by the cumulative causation of migration theory, our ABM incorporates the social network process with the migration process, simulating the interdependent

²We use 1995 year data as the input for the simulation and year 1996–2000 data to validate the models' outcomes.

processes simultaneously. In this section we discuss how the social network structure change and the migration acceleration correspond to each other.

Figure 11 shows the corresponding dynamics of the core-periphery structure and the nonlocal connections of migrant and nonmigrant groups. Figure 11a shows a dramatic increase in the proportion of core vertices' neighbors that are in the core, an increase from about 10% in pre-1995 to a peak of about 45% in 1998, indicating the rapid increase in connectivity within the core. Also, non-core vertices' neighbors in the core increase, particularly from 1998. These two curves suggest that the core-periphery structure becomes increasingly cohesive as the connectivity both within-core and between-core-periphery strengthens. These trends correspond to those of nonlocal neighbors in Figure 11b. The percentage of migrants' neighbors who are nonlocal increases to about 45% in 1997 and then remains stable afterwards. At the same time, the percentage of nonmigrants' neighbors who are nonlocal increases more slowly after 1997. The similar trends in Figure 11a and Figure 11b suggest the interrelationship between the migration acceleration and the social network structural change as postulated by the cumulative causation of migration theory.

We see the matches between the core-periphery structure and the migrant-nonmigrant interaction not only at the national level but also at the origin-destination level. Figure 12 represents core and migrants size from origin to destination. The dash line is the fitted linear regression line. From the graph we can see that there is a correspondence between the core size and the migrants size in the migrant stock matrix. Here we highlight some well-known migration subsystems inside China, e.g., Sichuan to Shanghai and Hunan to Guangdong. The corresponding core size and the migrants stock are well-mapped.

These mapped patterns between the significant change of the social network structure and the timing of migration acceleration emerging from the bottomup model are not the simple sums of micro dynamics. At the micro level, new migrants join the core and influence the future moves of more nonmigrants in the periphery. At the macro level, the core-periphery expands and migrant stock accelerates. This micro-macro feedback loop suggests that the social network structure is a major driver of migration perpetuation, and at the same time, migration transforms a sparsely structured social network into a cohesively structured social network.

6. Conclusions

This paper incorporates the agent-based model of migratory decisions and an observational-computational hybrid of the social network to explain migration perpetuation and social network structural changes in a migration era. Incorporating endogenous social networks into the agent-based modeling of migratory decisions improves model performance. Our model provides better subnational migration system outcomes and quantifies social network structural changes. At the micro level, traditional social interactions have given way to migration-purpose interactions. These micro-level dynamics contribute to structural changes in the population-scale social network. The emerging macro state suggests that the increasingly cohesive social network through the expanding core-periphery structure is a

driving force of migration acceleration. The core-periphery structure of social networks gives us a new prospective to explain the rapid increase in migration stock in China.

While the social network in 1995 was sparsely connected, the social network structure after 1997 became a key promoter of migration. The comparison of the explicit social model to both the non-network model and the implicit social networks model show the need to use the explicit social network in order to capture the migration-promoting effect of the social network. This modeling strategy produces convincing results showing that the social network influence is a catalyst in guiding rural-urban migration to certain destinations, above and beyond economic and demographic influences.

Acknowledgments

This research was supported by an NICHD Grant R21HD078808. The model is performed using Java language with the use of some MASON libraries by George Mason University.

References

1. Fields GS. Rural-urban migration, urban unemployment and underemployment, and job-search activity in Idcs. *Journal of development economics*. 1975; 2(2):165–187. [PubMed: 12265914]
2. C. N. B. of Statistics. *China Statistical Yearbook*. 2013
3. Harris JR, Todaro MP. Migration, unemployment and development: a two-sector analysis. *The American economic review*. 1970; 60(1):126–142.
4. Cai N, Ma H-Y, Khan MJ. Agent-based model for rural–urban migration: A dynamic consideration. *Physica A: Statistical Mechanics and its Applications*. 2015; 436:806–813.
5. Hao L. Cumulative causation of rural migration and initial peri-urbanization in china. *Chinese sociological review*. 2012; 44(3):6–33. [PubMed: 25525564]
6. Massey DS, Zenteno RM. The dynamics of mass migration. *Proceedings of the National Academy of Sciences*. 1999; 96(9):5328–5335.
7. Barbosa Filho, HS., de Lima Neto, FB., Fusco, W. Intelligent Agent (IA), 2011 IEEE Symposium on. IEEE; 2011. Migration and social networks an explanatory multi-evolutionary agent-based model; p. 1-7.
8. Stein M. Large sample properties of simulations using latin hypercube sampling. *Technometrics*. 1987; 29(2):143–151.
9. Tesfatsion, L., Judd, KL. *Agent-based computational economics*. Vol. 2. Elsevier; Amsterdam: 2006.
10. Macy MW, Willer R. From factors to actors: Computational sociology and agent-based modeling. *Annual review of sociology*. 2002:143–166.
11. Espíndola AL, Penna T, Silveira JJ. Rural-urban migration in d-dimensional lattices. *International Journal of Modern Physics C*. 2005; 16(12):1831–1840.
12. Silveira JJ, Espíndola AL, Penna T. Agent-based model to rural–urban migration analysis. *Physica A: Statistical Mechanics and its Applications*. 2006; 364:445–456.
13. Watts DJ, Strogatz SH. Collective dynamics of small-world networks. *nature*. 1998; 393(6684):440–442. [PubMed: 9623998]
14. McKay MD, Beckman RJ, Conover WJ. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*. 2000; 42(1):55–61.
15. Newman ME. The structure and function of complex networks. *SIAM review*. 2003; 45(2):167–256.
16. Toivonen R, Onnela J-P, Saramäki J, Hyvönen J, Kaski K. A model for social networks. *Physica A: Statistical Mechanics and its Applications*. 2006; 371(2):851–860.
17. Guare J. *Six degrees of separation: A play*. Vintage. 1990

Figure 1.

A part of hybrid social network: This subgraph contains two migrants in Guangdong originally from Sichuan and Hunan respectively in 2000. This graphs shows how the 4 layers of networks connect the whole country.

(a)

(b)

Figure 2.

(2a) shows the same subgraph from Figure 1 with labeled vertices. (2b) portrays how the network spread the influence starting from Agent 1 of different geodesic distances.

Figure 3.

Flowchart: This graph shows the simulation process in *Model 3*. For *Model 1* and *Model 2*, Step 2 are missing.

(a) Power law like exponent

(b) κ_2 , κ_3 , κ_4 , ϕ of samples that satisfy the criterion**Figure 4.**

In (4a), For an exponent $\gamma > 3$ for social networks of size $N \sim 10^6$, we abandon all the samples with $\gamma \leq 3$. In (4b), the blue circle refer to the samples that satisfy the criterion. For example we can use an approximate estimate to give the optimal region for κ_4 , which in this case, will be 4 – 5.

Figure 5.

Basic Network Properties of Computational Model in China, 1995–2000: (5a): the curves are approximately downward-slope lines, which indicates that the degree follows the power-law like distribution. The dash lines are the boundaries for the γ , which in our case is between 3 and 4 appropriate for a network size of 313,355. Figure 5b shows a relatively stable average clustering coefficient. Figure 5c shows the average geodesic distance for pairs of all vertices in the whole network. While the trend is downward, the average geodesic distance is around 6, ranging from 5.55 to 6.55.

(a) % of core Vertices are Migrants (b) % of migrants located in the Core (c) % of core size to the network

Figure 6.

6a: Percentage of Core Vertices are Migrants: The percentage peaks at $k=12$. 6b:

Percentage of Migrants located in the Core: The percentage declines to 0 when $k > 12$. 6c:

Core size as the percentage the whole network.

(a) Year 1995

(b) Year 1997

(c) Year 2000

Figure 7.

Heuristic Network Visualization of the Core-Periphery Dynamics: The dots with different colors represent vertices in the core-periphery, blue for the core vertices, red for vertices with $d=1$ to the core, orange for vertices with $d=2$ to the core, and purple for vertices with $d=3$ to the core. The solid lines represent the social connections between vertices.

Figure 8. Core and periphery percentage: The core and $d=1$ section increase slowly at first and accelerate later; the $d=2$ section peaks 1999 and then decline later; the $d=3$ section increase until 1997 and then decline.

(a) Degree of actors

(b) Cluster coefficients

(c) Geodesic distance

Figure 9.

Basic Network Measures of Core and Non-core Vertices: (9a) shows the average degree among core vertices is much greater than that among non-core vertices (9b) shows the clustering coefficient of core and non-core both follow a downward pattern as degree increases. Figure (9c) shows the geodesic distance for core vertices and non-core vertices.

Figure 10.
Empirical Validation of Computational Migrant Stock by Destinations: The red is over-fit and blue is under-fit.

(a) % of Neighbors in the core

(b) % Non-local neighbors

Figure 11.

(11a) shows how dramatic is the increase in the proportion for neighbors that are in the core, which increase connections from around 10% in pre-1995 to around 45%–48% in 1998, which remains stable afterward. (11b) shows the increase in the proportion for non-local degree for migrants, which increases to around 45% in 1997, and remains stable afterward.

Figure 12.

Core and Migrants Size: the scatters in the plot represent core and migrants size from origin to destination. The dash line is fitted linear regression.

Table 1Samples result by Latin Hypercube method for *Model 3*

Sampled parameter	Initial space	Final space
κ_2	[2,5]	[4,5]
κ_3	[0,2]	[1,2]
κ_4	[0,5]	[4,5]
ν	[0.5,1]	[0.7,0.9]
β_2	[0.00002,0.00003]	[0.00002,0.00003]
β_3	[0.02,0.03]	[0.02,0.03]
θ	[0,1]	[0.5,0.8]
ρ	[0.7,0.9]	[0.7,0.8]
ω_2	[0,10]	[0,10]
ω_3	[0,10]	[5,10]

Fixed Parameter	Value
η_2	0.9
η_3	0.9
η_4	0.95
α_2	0.02
α_3	0.0005
φ	1

Table 2

Statistics for Empirical-Computational Difference in Origin-Destination Matrix of Migrant Stock

	First Order Norm	Infinite Order Norm	Frobenius Norm	Pearson Correlation
<i>Model 1</i> vs Data	11286	7547	8126	0.074
<i>Model 2</i> vs Data	10771	5394	6431	0.417
<i>Model 3</i> vs Data	10725	3040	3022	0.632