Physica A 528 (2019) 121497

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Online footprints of workforce migration and economic implications

Xiaoqian Hu [a], Jichang Zhao [a,b], Hong Li [a], Junjie Wu [a,b,c,*]

[a] Department of Information Systems, School of Economics and Management, Beihang University, No.37 Xueyuan Road, Haidian District, Beijing 100191, China
[b] Beijing Advanced Innovation Center for Big Data and Brain Computing, Beihang University, No.37 Xueyuan Road, Haidian District, Beijing 100191, China
[c] Beijing Key Laboratory of Emergency Support Simulation Technologies for City Operations, No.37 Xueyuan Road, Haidian District, Beijing 100191, China

HIGHLIGHTS

- We explore workforce migration on social media during the Spring Festival.
- A prediction model is built by profiling core driving forces in migration decision.
- Mainstream patterns in inter-city workforce migration are revealed.
- Malfunction of megacities in regional development is explored.
- Agglomeration effect and filter effect are unveiled for policymaking in regional economic development.

ARTICLE INFO

Article history:
Received 8 October 2018
Received in revised form 26 January 2019
Available online 21 May 2019

Keywords:
Workforce migration
Social media
Gravity model
Regional economic development

ABSTRACT

Workforce migration plays a strong role in reallocating productive resources and provides important clues for understanding socio-economic dynamics. In this paper, we demonstrate that like a natural shock, the Spring Festival in China culturally drives workforce to travel between workplaces and hometowns, and hence the resulting trajectories in online social media open an unparallelled window to explore laws in nation-wide workforce migration. To understand the dynamics behind workforce migration between Chinese cities, a prediction model is built by profiling the tradeoff between interest articulation and cost elusion in migration decision. Interesting migration patterns are further revealed and economic implications beyond these patterns, like the filter effect of labour market and the agglomeration effect of core cities, are fully explored, which is of great help to understand different roles of Chinese cities in their own and regional economic developments. To our best knowledge, our work is among the few studies that can leverage the full data in a social media platform for comprehensive investigation.

1. Introduction

Workforce remains one of the most fundamental elements for social production in modern societies, even in the knowledge economy era. The workforce's intercity migration within a country, usually between hometowns and workplaces, essentially reflects the reallocation of production resources responsible for normal operation of national economy.

* Corresponding author at: Department of Information Systems, School of Economics and Management, Beihang University, No.37 Xueyuan Road, Haidian District, Beijing 100191, China.
E-mail address: wujj@buaa.edu.cn (J. Wu).

https://doi.org/10.1016/j.physa.2019.121497
0378-4371/

Indeed, workforce migration exerts a strong influence on socio-economic issues [1,2] including, just to name a few, the balanced development of regional economy, national transportation design, urban infrastructure planning, policymaking in household registration systems and climate change mitigation [3]. Understanding laws of workforce migration at a national scale could help solve these issues.

While individual movement is stochastic, the collective behaviour is often not random but closely linked to extensive social, economic [4,5] and even political factors [6]. Laws or patterns underlying group travel can emerge as the population size increases [7,8], suggesting that the laws of workforce migration can be probed and modelled. Along this line, tremendous efforts have been devoted to profiling and forecasting human migration in recent years [9–11]. In addition, diverse patterns of human mobility are revealed [12–14], with the driving forces extensively explored [15–18].

Most existent studies, however, assume that human migration behaviour is determined only by the attraction of cities, neglecting people's subjective will of seeking benefits and the capability requirements of different labour markets. In fact, from the economics viewpoint, the driving force behind workforce migration is a higher economic benefit or the Pareto optimality of the economy [19–21]. The development of science and technology also leads to a major labour market concern, more with technical capacity rather than with workforce size. Moreover, from the cognition perspective, sensitivity regarding influential factors varies for different individuals [22]. Hence, the cost of migration should not be always represented by the spherical distance. Richer information should be included as indicators to demonstrate different cognitive sensitivities of the workforce when making behavioural decisions.

The data adopted in previous studies also impair the ability to conduct an in-depth study of workforce migration. Specifically, survey samples from questionnaires or government censuses constitute the major data source in previous studies on the workforce [13,23,24]. However, it is well known that limited samples, biased sample selection and the uncontrollable quality of investigations might lead to unreliable survey data. Although government censuses supply valid nation-wide records, they incur an enormous expense and are very time-consuming (once per decade in China) and coarse-grained in the province level (see http://www.stats.gov.cn/tjsj/pcsj/). In summary, traditional investigations cannot offer the desired data in terms of cost, granularity and reliability.

The explosive growth of social media in recent decades, such as the Twitter-like service Weibo in China, has offered an unprecedented opportunity to understand human movements, which is further boosted by the continuous penetration of smart and mobile devices to our daily life. In order to distinguish the workforce migration behaviours from ordinary movements, we adopt the Spring Festival scenario and observe the online footprints of workforce in Weibo. Being the most important traditional custom in China, the Spring Festival culturally drives people working outside back to their hometowns to reunite with the families and celebrate the Chinese New Year, despite of long distances, large economic costs and other hardships on journey (see Appendix section *A*). It can be treated as a natural shock to workforce movement at a national scale [25,26], leading to an extremely high traffic peak called the Spring Festival Travel Rush before and after the Spring Festival Eve [27–29]. The huge volumes of footprints left in online social media thus form an ideal data source for modelling workforce migration at a national scale.

In this paper, we collect ALL Weibo users' online footprints during the 2017 Spring Festival Travel Rush and thus establish a complete picture of workforce migration in China. To explore the core driving forces of this systematic migration, we introduce economic factors and various cost measures to the well-known gravity model (GM) [30–32], and design an extension model called GM_e to make high-quality prediction of workforce flows between cities. We also investigate the diverse patterns of workforce migration by employing clustering, and explore migrant workers' subtle tradeoff between interest articulation and cost elusion in different patterns. The results indeed verify the interest-driven intrinsic motivation of the workforce and the extrinsic limitation in capability requirements of labour markets. Our study suggests that the non-negligible entanglement between social media and macroeconomic behaviours is insightful for policy making in socio-economic issues.

2. Materials and methods

2.1. Migration footprints in Weibo

Weibo is a Twitter-like service in China with nearly 400 million active users. All tweets in Weibo during the Spring Festival Travel Rush from Jan. 13 to Feb. 21, 2017 were collected (see http://zizhan.mot.gov.cn/zfxxgk) to form a complete picture of migration. A 4-tuple (m, u, t, g) was extracted from each post, where m and u are the unique identifiers of the post and the user respectively, t is the timestamp of the post, and g is the location of the user when publishing the post, including the country, province and city information. In general, 371 distinct cities were extracted from the total over 4 billion posts.

We reorganized all the 4-tuples in chronological order according to user ID, and thus obtained the complete trajectory of each user in a city level. That is, if a person travels from city A to city B and then to city C, we have a trajectory containing two flows: A to B and B to C. We then summed all users' trajectories to construct a directed inter-city migration network G, in which nodes represent cities and the edge (g_i, g_j) denotes the migration flux $f_{i,j}$ from city i to city j. An undirected migration network G_u can then be constructed with the undirected edge representing the migration flux $F_{i,j}$ transferring between cities i and j, *i.e.*, $F_{i,j} = f_{i,j} + f_{j,i}$ (see Appendix section *B*). In general, there are 61,759 edges (*i.e.*, city pairs) in undirected G_u and 120,361 edges in directed G, totalling 41,454,268 migration traces. It is worth noting that while we preserved complete trajectories, about 73% of migrant users changed their locations only between two cities during the Spring Festival Travel Rush.

2.2. The national railway line data

We also collected national railway line data from the train schedule released by China's Ministry of Railways, which contains 5,878 trains in total and 12,382 city pairs that have directed train connection. It turns out that the flows of online Weibo footprints between these 12,382 city pairs have a positive correlation with the numbers of trains going between them (see Fig. S1 in Appendix section *G*), which supports using the Weibo data for workforce migration investigation. Nevertheless, the 12,382 city pairs only account for one-fifth of city pairs extracted from the Weibo data. This reflects the unique value of online footprints in Weibo, which indeed provides a factual and much broader view of a nationwide workforce migration in China.

2.3. Demographic and economic indexes

To explore the economic driving force to workforce migration, we further collected the GDP (gross domestic product), per capita GDP, and the number of permanent residents data of all cities from the 2016 statistical yearbook in China (there is a delay in governmental statistics). Furthermore, three more indexes including the ratio of practitioners in high-technology industries, the disposable income per capita, and the investment of funds for research and development (R&D), were also collected from the 2016 statistical yearbook.

2.4. Distance measures between cities

To measure inter-city accessibility, we use three distance measures, namely the actual geographical distance, the travel distance and the travel time in route planning. The former is the distance computed by the longitude and latitude of cities on Google maps. For city i with longitude and latitude (x_i, y_i) and city j with (x_j, y_j), the distance is defined as

$$d_{ij} = D \times \arcsin\left(\sqrt{\sin^2(a) + \cos(rx_i) \times \cos(ry_j) \times \sin^2(b)}\right), \tag{1}$$

where $rx_i = \frac{1}{180}\pi x_i$, $ry_j = \frac{1}{180}\pi y_j$, $a = \frac{1}{2}(y_i - y_j)$, $b = \frac{1}{2}(x_i - x_j)$, and D is the diameter of the earth. The travel distance and travel time between cities were crawled from the travel path planning in Baidu Map API, which reflect the true land transportation costs.

2.5. The extended gravity model

We here propose an extended gravity model to explore the latent driving forces to workforce migration. In the classical gravity model (GM), human migration is assumed to be closely related to the attractiveness and accessibility of locations. The former is often estimated by the population size, while the latter by the geographical distance between two locations. Specifically, the migration flux between cities i and j is computed as $F_{ij} = a\frac{P_i \cdot P_j}{d_{ij}^{\gamma}}$, where P_i denotes the number of people residing in city i, F_{ij} represents the undirected migration flow between cities i and j, d_{ij} is the geographical distance between the two cities, and a and $\gamma > 0$ are constant parameters for adjustment. This model implies that the migration flux increases with the population size and decreases with the distance between the two locations.

The demographic factor in the classical GM model, however, might not fully capture the dynamics behind labour force migration. From the economics perspective, profit maximization, utility optimization and intuitive revenue enhancement [19–21,33] are treated as the influence factors when workers make migration decisions. From the sociology perspective, workforce migration aims at improving living conditions [16,17]. That is, immigrant areas with abundant job opportunities, higher wage levels and other preferential treatment conditions produce a "pull" force, while harsh living conditions in the original residential place would "push" human emigration. Both views imply the subjective factor in workforce migration: the pursuit of profit maximization.

As a result, we introduce economic factors and propose a general GM extension model (GM_e) as follows:

$$F_{ij} = a\frac{E_i^{\alpha} \cdot E_j^{\beta}}{C_{ij}^{\gamma}}, \tag{2}$$

where E_i stands for the economic indicator of city i, C_{ij} represents the migration cost between cities i and j, and α, β and γ are constant parameters assumed to be positive. For economic indicators, GDP is one of the most important and easily accessible indexes that reflects a region's economic situation, the income level and the required level of labour skills (see Fig. S2 in Appendix section *G*). So it is natural to select GDP as a candidate for economic indicator in GM_e, which leads to a new model: G-GM_e. Considering the demographic information included in GDP, we also select per capita GDP as another economic indicator and propose the model: avgG-GM_e. Moreover, to characterize the direction of an inter-city flux, we further introduce a variant of G-GM_e, named dirG-GM_e, in which E_i and E_j denote the GDP of the origin and destination cities respectively, and F_{ij} in Eq. (2) is changed to f_{ij}, indicating the migration flux from city i to j. Note that due to the undirected nature of G-GM_e and avgG-GM_e, we set α and β to 1. We adopted three types of migration cost C_{ij}

Table 1
The fitting results for different models. In this table, γ is the exponent for the distance measurements, and α and β are exponents for the GDP of origins and destinations, respectively. Additionally, ** stands for $p < 0.01$, and *** represents $p < 0.001$, all parameters in the models have been examined by significance tests. And ** on the model name stands for $p < 0.01$ in the F-test of the regression model. R^2 represents the R square measure. In general, the fitting is satisfactory with significant parameters and models. From the prediction view by R^2, the dirG-GM$_e$ is the best.

Distance metric	Model	γ	α	β	R^2
Geographical distance	GM**	0.314***	–	–	0.024
	avgG-GM$_e$**	0.757***	–	–	0.084
	G-GM$_e$**	0.287***	–	–	0.025
	dirG-GM$_e$**	0.307***	0.928***	0.935***	0.587
Travel distance	GM**	0.391***	–	–	0.037
	avgG-GM$_e$**	0.861***	–	–	0.108
	G-GM$_e$**	0.341***	–	–	0.035
	dirG-GM$_e$**	0.337***	0.902***	0.909***	0.598
Travel time	GM**	0.390***	–	–	0.062
	avgG-GM$_e$**	0.731***	–	–	0.133
	G-GM$_e$**	0.317***	–	–	0.051
	dirG-GM$_e$**	0.363***	0.917***	0.924***	0.591

for the above models, namely the geographical distance, the travel distance and the travel time, with the details given in Section 2.4.

The parameter estimation of the above models is straightforward by following the least square method for log-linear regressions [34] (see Appendix section *C*). In addition to the R-square measure, we also adopt the well-known measure named SSI based on Sørensen index [35] for goodness-of-fit evaluation, which is defined as $SSI = \frac{2\sum_i \sum_j \min(F_{ij}, \hat{F}_{ij})}{\sum_i \sum_j F_{ij} + \sum_i \sum_j \hat{F}_{ij}}$, where F_{ij} and $\hat{F}_{ij}$ denote the actual and predicted migrant fluxes between city i and city j, respectively. The range of SSI is in [0,1], with a larger value indicating a better predictive capability and vice versa.

3. Results

3.1. Driving forces for inter-city workforce migration

We use the Weibo data as the inter-city realistic flux (*i.e.*, F_{ij} or f_{ij}) in Eq. (2) to estimate the parameters of each model (see Appendix section *C*), and then use the learnt model to obtain the inter-city predicted flux. Table 1 shows the estimated parametric values and the R-square values of fitness, and Fig. 1 illustrates the realistic and predict fluxes when using the travel time as the distance measure (other distance measures see Fig. S3-S4 in Appendix section *G*). Fig. 2 further compares the goodness-of-fit of the four models with different distance measures in terms of the SSI measure.

As can be seen from Fig. 2, independent of distance measures, the G-GM$_e$ and dirG-GM$_e$ models demonstrate much higher predictive power than the GM model, and the avgG-GM$_e$ model performs the worst. This justifies the rationality of introducing economic factors to our GM$_e$ model for workforce migration modelling, and the GDP index seems to be a better proxy than the per capital GDP index as well as the number of permanent residents index when measuring the attractiveness of a city to migrant workers (see Appendix section *D*). Moreover, Fig. 2 also shows that all the models with travel time as the distance achieve the best performances. This implies that travel time might be the better proxy than the geographical distance in measuring the intangible cost in the minds of migrant workers.

We then focus on the estimated parametric values. As reported in Table 1, all the parameters have positive and significant values, suggesting bigger migration flows between close and well-developed cities. This well explains that economic and distance factors are indeed the important driving forces for workforce migration. Moreover, the α and β parameters are close to each other in the dirG-GM$_e$ model (0.917 and 0.924 respectively), implying that economic levels of the origin and destination cities exert an approximately equivalent influence on migrant workers. As a consequence, we can reasonably ignore the direction of migration flows, and the dirG-GM$_e$ model can therefore reduce to the undirected G-GM$_e$ model, which is the default model used for the following study.

One might expect a stronger migration when there is a larger difference in economic situations between two cities. To validate this, we also introduce the economic difference between two cities into our GM$_e$ model as an extra attractiveness factor other than the two cities' GDP, or even the only factor. The results, however, show that the new models have much worse predictive performances than the G-GM$_e$ model, implying that the economic difference is not a very critical factor for flux prediction (see Appendix section *E*). This might be explained by the fact that a city's GDP level implies the technology capability requirement of the labour market in that city (see Fig. S2 in Appendix section *G*). Workforce from a city at a much lower GDP level is less capable of meeting the technology capability requirement of a much more developed city, even though a larger economic difference might incur a greater economic attractiveness to labour force.

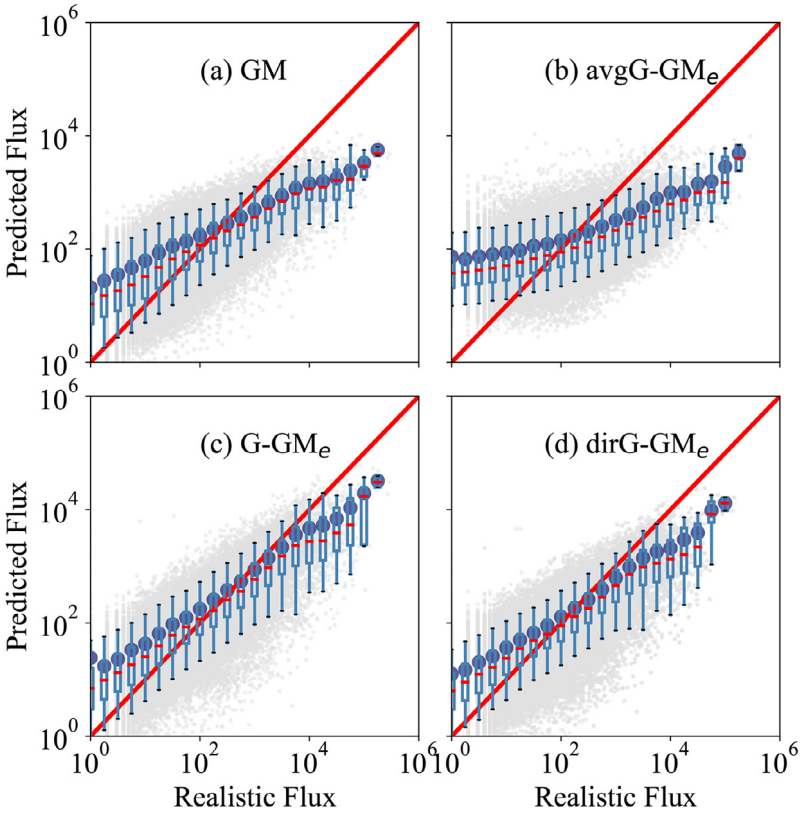

Fig. 1. Flux prediction results for four models. The distance measure is the travel time. The grey points are a scatter plot of fitness between the realistic labour migration flux and forecast flux between each pair of cities. The red lines of $y = x$ indicate the best prediction, and the red points represent the means of the predicted migration flows. Additionally, the whiskers of the box plots show the 5th and 95th percentiles, respectively.

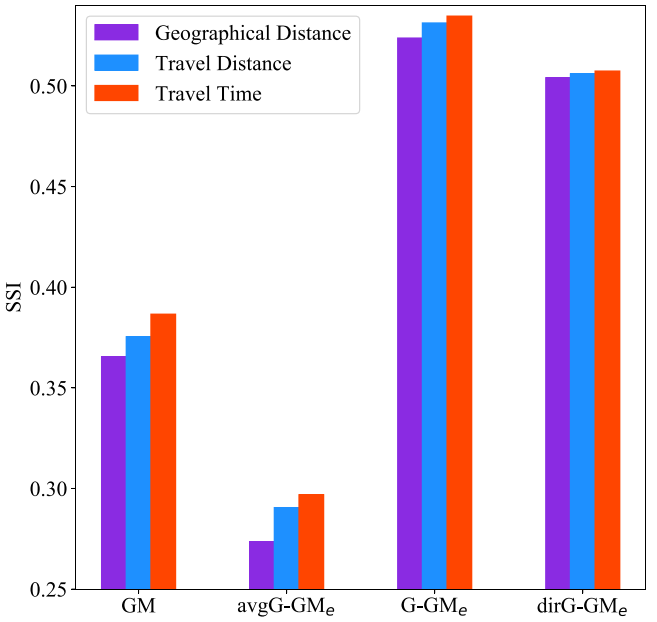

Fig. 2. Comparison of four models in terms of SSI. The bars with different colours show the predictive capability of models with different distance metrics.

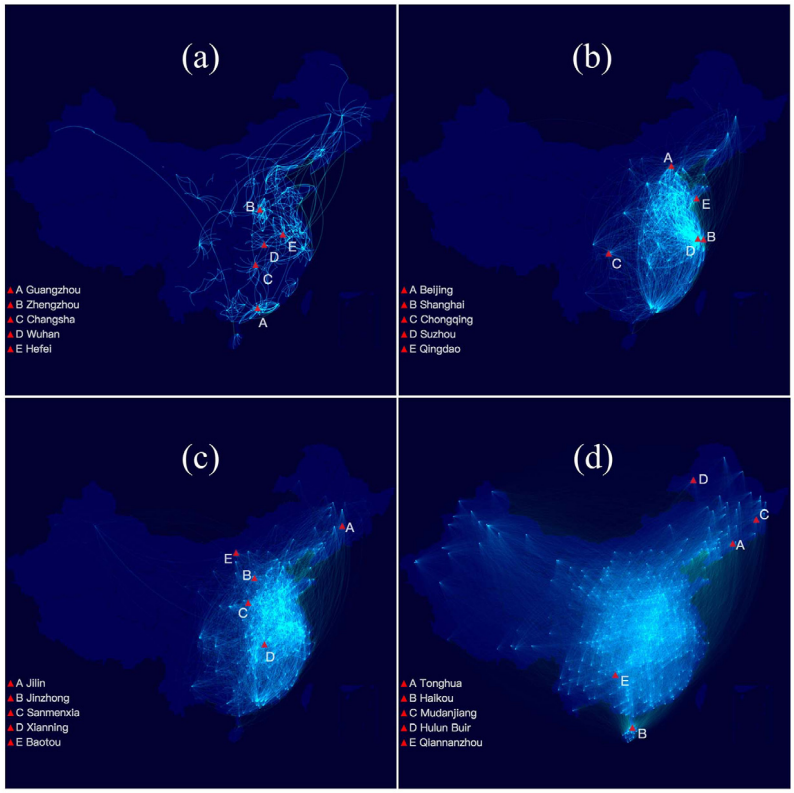

Fig. 3. Migration flows of grouped trajectories. The bright lines denote the migration flows between cities. (a) **Pattern I**, (b) **Pattern II**, (c) **Pattern III** and (d) **Pattern IV**. A brighter area denotes that there are more city pairs with migration flows. The average migration flows between city pairs is 18,522, 3182, 1045 and 195, respectively. Each group accounts for 29%, 28%, 20% and 23% of the total flux, respectively. The average GDP product of city pairs in different groups is 1.7E+07, 3.1E+07, 8.8E+06 and 1.7E+06 and the average travel time is 1.3E+04, 3.7E+04, 5.4E+04 and 1.5E+05 respectively (the specific distribution of features see Fig. S9-S11 in Appendix section *G*). The representative cities in each pattern are marked by red triangles. Specifically, we first sort cities by their migration flux and for each pattern, we only keep cities whose flux ratios of the corresponding pattern are more than a threshold (50% in Pattern I & II & IV and 40% in Pattern III). Then for each pattern, the top five cities are selected as the representative ones.

3.2. Mainstream patterns in inter-city workforce migration

Given economic benefits (GDP) and travel costs (travel time) as the significant driving forces to workforce migration, we here take a further step to recognize the mainstream patterns hidden inside these inter-city migrations, with the purpose to reveal and diagnose the economic development status of Chinese cities. To this end, the *K*-means clustering approach [36] is employed to divide city pairs into different groups, with the normalized GDP product of each city pair and the normalized inter-city travel time and migration flux as the three features, and the cosine similarity as the approximation measure (see Appendix section *F*). The silhouette coefficient [37] and the elbow method [38] are both used to determine the optimum number of groups, which is finally set to 4 for the best performance (see Fig. S5-S8 in Appendix section *G*). Fig. 3 shows the four groups (patterns) of city pairs on the map, with the detailed information of different groups given in the caption.

Pattern I: Migration between local core cities and their surroundings. As demonstrated in Fig. 3(a), the city pairs of Group I generally show divergent and star-like patterns, covering some local core cities such as Guangzhou, Zhengzhou, Changsha, Wuhan and Hefei and their surrounding cities. Most of these core cities are the capital cities of different provinces, and thus have great economic influences to their neighbourhood cities. The workforce in this group spends the least travel time but generates huge volumes of migration flux that dominates other groups. This makes it the primary economic development pattern for Chinese cities.

Pattern II: Migration towards developed cities. Fig. 3(b) shows the city pairs in Group II, where the most developed cities in China, such as Beijing, Shanghai, Chongqing and Suzhou, are the key destinations for the migration. This implies that the workforce migration in this pattern is more sensitive to cities' economic levels. While the average travel time between the city pairs in this pattern is the second-lowest among the four groups, the total volume of flux is the second-largest, implying that this is the secondary economic development pattern for Chinese cities.

Pattern III: Migration between undeveloped cities. Fig. 3(c) shows the geographical distribution of city pairs of Pattern III, which mainly contains undeveloped cities located in central China, such as Jinzhong, Sanmenxia, Xianning, *etc.* Indeed, compared with those in Pattern I and Pattern II, the city pairs in Pattern III have relatively smaller GDP products, longer travel time, and the total flux volume is also much smaller. This implies that the migrants in this pattern pursue less economic benefits but meanwhile afford higher transportation costs.

Pattern IV: Migration due to emotions. As shown in Fig. 3(d), cities in this pattern geographically cover all areas of China, particularly the peripheral provinces such as Xinjiang, Tibet, Heilongjiang, and Hainan, as well as some remote cities located on China's borders, such as Tonghua, Haikou, Mudangjiang, Hulun Buir and Qiannanzhou. City pairs in this pattern possess the lowest product of GDP but the longest travel time compared with other groups, implying that these migrations might be "emotional" rather than profit-driven. The smallest average flux volume in this pattern further reveals that this is not a mainstream pattern in workforce migration.

The four patterns are the natural result of the selective behaviour of workforce in interest articulation (measured by GDP) and cost elusion (measured by travel time). We also employ G-GM$_e$ to model the migration flux inside each group and gain significant improvements in prediction performance compared with the global model (see Fig. S12 and Fig. S13 in Appendix section *G*), which further validates the effectiveness of the four patterns. According to the total flux volume data in Fig. 3, Patterns I and II reflect the two mainstream modes in economic development of Chinese cities. The difference is that Pattern I highlights the impact of local core cities while Pattern II emphasizes the impact of national core cities.

It has been long criticized that there exist extremely unbalanced economic developments in Asian countries including China, which have forced the emergence of various megacities like Beijing, Shanghai, Tokyo, Seoul, *etc.* [39]. The workforce migration in Pattern II indeed agrees with this claim and demonstrates the great economic attraction of developed cities. Nevertheless, it is interesting to see that the economic imbalance in China is getting reversed — the local core cities in Pattern I exhibit the economic **agglomeration effect** on nearby cities, which could foster the development of regional economics and relieve the pressure of megacities. The largest flux volume but small number of city pairs in Pattern I indicate that regional economic development in China is getting more and more important, although there is still much room for improvements.

Compared with Patterns I and II, Pattern III represents a relatively backward development mode, with migrant workers bearing larger travel costs but less profits. The reason behind this seemingly irrational choice might be the **filter effect** of the technology capability requirements of labour market in high-GDP cities, which prevents not-well-educated workforce from low-GDP cities from entering local core cities or developed cities. In summary, our results suggest that the positive agglomeration effect of local core cities and the negative filter effect of labour market in developed cities are two critical factors for the governance of unbalanced economic development in China.

3.3. Malfunction of megacities in regional economic development

Regional economic development (RED) has become the focal point of the Chinese government since the drop of GDP growth rate in recent years. As mentioned above, workforce migration in Pattern I reflects the agglomeration effect of local core cites and is therefore regarded as the best mode for RED. Along this line, a natural idea is to evaluate a city's function in RED via the ratio of the city's migration flux in Pattern I. We define a city i's locality as the cumulative ratio of migration flux among top r nearest neighbourhood cities, which is given by

$$c(i,r) = \frac{\sum_{(i,j)\in T_i(r)} F_{ij}}{\sum_{(i,j)\in T_i} F_{ij}}, \tag{3}$$

where $T_i(r)$ is the set of city pairs between city i and top r cities sorted ascendingly by travel time, and T_i is the set of city pairs between city i and other cities. If $c(i,r)$ increases to 1 rapidly as r grows, city i is deemed to have good locality and thus can function well in regional economic development.

Fig. 4 shows the locality of each city with the increasing r. It is surprising that two developed cities (also titled "national center city"), *i.e.*, Beijing and Chongqing, behaviour differently from other high-GDP cities for lack of good locality. In particular, Beijing and Chongqing seem to contribute the least in regional economic development among the most important cities in China, such as Guangzhou and Shenzhen that lead the Pearl River Delta region, and Shanghai that leads the Yangtze River Delta region. Take Beijing for example. As the capital of China and circled by the Hebei province in economic backwardness, Beijing has long been expected to play a significant lead role in radiating the regional development, which however is not very sanguine as indicated by our results. That is why the Chinese government has set the economic integration of the Beijing-Tianjin-Hebei region as a national strategy since its release in 2014, hoping to boosting the regional development via policy instruments.

We further analyse the influential factors for the malfunction of Beijing and Chongqing in regional economic development. Fig. 5 compares the GDP of core cities and their surrounding cities. It is obvious that the surrounding cities of Beijing and Chongqing generally have a worse economic matching status than that of Shanghai, Shenzhen and Guangzhou. This implies that the economic disparity might be the important factor for the megacities' malfunction. That is, the workforce in low-GDP surrounding cities could not get well-trained to meet the skill requirements of the labour markets in megacities like Beijing and Chongqing, which discourages the labour mobility and eventually hurts the regional economic development. Therefore, to promote regional development in a coordinated manner, improving the technical level of workforce in undeveloped areas is a substantial requirement, in addition to the policy-oriented economic cooperation.

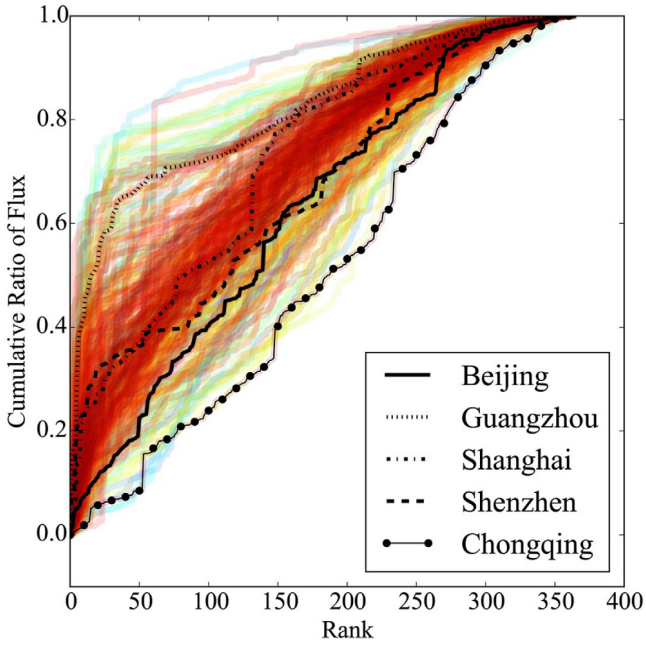

Fig. 4. Locality of cities with increasing distance. The cumulative ratio of the migration flux of the neighbouring cities from the near to the distant, defined as $c(i, r)$, is computed to show the locality characteristics of the target city. The colour of lines stands for the GDP of city i, and a deeper red colour means better economic status.

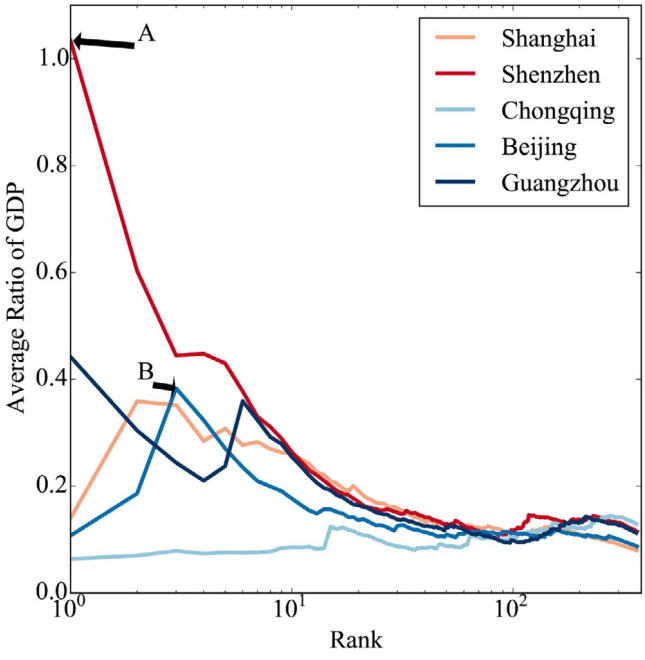

Fig. 5. Comparison between core cities and their surrounding cities. To understanding the extent of GDP matching between target cities and their neighbouring cities, the average ratio of GDP is defined as $l(i, r) = \frac{\sum_{(i,j)\in T_i(r)} E_j}{rE_i}$, where $T_i(r)$ represents the set of pairs between city i and top r cities, sorted by travel time, and E_i is the GDP of city i. Because Guangzhou has a higher GDP than Shenzhen, A is an abnormal point in the line of Shenzhen. B is the highest point in the line of Beijing, and the inflection point results from Tianjin, which is the only city with a high GDP that is similar to that of Beijing.

4. Discussion

The social media boom indeed offers an opportunity to understand human behaviours in unparallelled richness and fine granularity. To the best of our knowledge, this is the first study to systemically explore the workforce migration based on the full size of online footprints in a social media platform. Our study overcomes the limitations of traditional survey-based approaches and justifies the feasibility of probing workforce migration both collectively and individually using social media. More importantly, the present study confirms the possibility of quasi real-time policy making for social-economic issues, which is traditionally constrained by the long investigation procedure and often leads to the problem of being "out-of-sync" that could hurt the flexibility of policy formulation.

Taking the Spring Festival Travel Rush as a natural shock to investigate the workforce migration in China is interesting and reasonable. The Spring Festival is absolutely different from ordinary holidays. It has extraordinary significance to Chinese people, and evolves into a traditional cultural heritage that entrenches the emotional attachment to home in Chinese people's minds [40]. Indeed, it has been widely recognized in the literature that almost all of migrant workers would try their bests to return home to reunite with families and celebrate the Spring Festival, despite of long distances, large economic costs and other hardships on journey [41]. This provides a precious chance to observe migration fluxes between hometowns and workplaces yearly in a fine-grain granularity.

We systematically discuss the core driving forces that hide behind the nation-wide workforce migration. By introducing economic factors into the classical GM model, our G-GM$_e$ model shows much better performances than GM in inter-city fluxes modelling, which in turn justifies the power of interest articulation (measured by city GDP) and cost elusion (measured by inter-city travel time) in understanding the world's yearly largest human mobilization. Since people could have different sensitivities to economic stimulation, we further investigate the different migration patterns in terms of individual wills, and unveil some interesting phenomena closely related to the regional economic development in China, including the positive agglomeration effect of some local core cities, the negative filter effect of labour markets in megacities, and some malfunctioned megacities like Beijing and Chongqing.

It could be generally recognized from our results that the regional economic development led by local provincial big cities keeps rising in China and has absorbed the largest volume of workforces. But how to remove the barrier of technical capability requirements set naturally by the labour markets in big cities, remains the critical challenge and calls for wise solutions. This is also critically important for megacities like Beijing to exert great impact to the integration of Beijing-Tianjin-Hebei regional economic development, which has risen as the national strategy of China.

In addition to regional economic development, our study could also shed light on solving some critical social-economic issues. The prediction model of migration flows offers a quantitative perspective to evaluate and design policies on household registration systems, floating population management and inter-city transportation planning in real time. Meanwhile, the disclosed patterns of workforce migration not only have implications for the government in policy formulation and evaluation but also can inspire enterprises and individuals to solve real issues. For example, the tradeoff between the benefit and distance costs in making migration decision could offer suggestions regarding where to release recruitment information to solve the problem of labour shortage, especially in the southeast coastal areas of China.

5. Conclusions

In this study, full size of geographical tweets posted on Weibo during the Spring Festival Travel Rush in 2017 are considered as a natural experiment to understand the nation-wide workforce migration. Intrinsic motivations for migration decision, such as the pursuit of interest and cost avoidance, are discussed systematically by introducing the economic indicators into gravity laws to form a new model called G-GM$_e$ for migration flux modelling. We find that the GDP product and travel time between city pairs are excellent indicators for workforce prediction. In particular, GDP reflects cities' potential benefits as well as the technical requirements of their labour markets, which are both crucial for understanding migration decision of workforce. Under the combined effect of intrinsic motivations and external restrictions on the labour market, workforce migration in China presents four typical patterns, among which the agglomeration effect of local core cities and the filter effect of labour markets are identified and discussed from the perspective of regional economic development. As an application, Beijing's and Chongqing's lack of a leading role in regional radiation and the mismatch between labour market skills in surrounding areas and labour market requirements in these two cities, are detected from workforce migration patterns and inspire related policy formulation and evaluation.

Reference data

The data used in this research is publicly available and can be downloaded freely through: https://doi.org/10.6084/m9.figshare.5513620.v2

Declaration of competing interest

We have no competing interests.

CRediT authorship contribution statement

Xiaoqian Hu: Conceptualization, Writing-original draft, Data curation, Formal analysis, Methodology, Writing-review&editing. **Jichang Zhao:** Data curation, Formal analysis, Methodology, Writing-review&editing. **Hong Li:** Writing-review&editing. **Junjie Wu:** Conceptualization, Writing-original draft, Writing-review&editing.

Acknowledgements

W.J. was supported by the National Natural Science Foundation of China (NSFC) (71531001, 71725002, U1636210). L.H. thanks the support from NSFC (71471009). Z.J. thanks the support from NSFC (71871006).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.physa.2019.121497.

References

[1] G.J. Abel, N. Sander, Quantifying global international migration flows, Science 343 (6178) (2014) 1520–1522.
[2] J. Shen, Changing patterns and determinants of interprovincial migration in china 1985-2000, Population, Space and Place 18 (3) (2012) 384–402.
[3] S. Shayegh, Outward migration may alter population dynamics and income inequality, Nature Clim. Change (2017) advance online publication.
[4] S. Luo, F. Morone, C. Sarraute, M. Travizano, H.A. Makse, Inferring personal economic status from social network location, Nature Commun. 8 (2017) 15227.
[5] R. Li, L. Dong, J. Zhang, X. Wang, W.X. Wang, Z. Di, H.E. Stanley, Simple spatial scaling rules behind complex cities, Nature Commun. 8 (1) (2017).
[6] H. Castañeda, Migration is part of the human experience but is far from natural, Nat. Hum. Behav. 1 (2017) 0147 EP.
[7] Y. Liu, Z. Sui, C. Kang, Y. Gao, Uncovering patterns of inter-urban trip and spatial interaction from social media check-in data, PLoS One 9 (1) (2014).
[8] C. La Porta, O. Chepizhko, C. Giampietro, E. Mastrapasqua, M. Nourazar, M. Ascagni, M. Sugni, U. Fascio, L. Leggio, C. Malinverno, Bursts of activity in collective cell migration, Proc. Natl. Acad. Sci. USA 113 (41) (2016) 11408.
[9] F. Simini, M.C. Gonzalez, A. Maritan, A. Barabási, A universal model for mobility and migration patterns, Nature 484 (7392) (2012) 96–100.
[10] X. Yan, C. Zhao, Y. Fan, Z. Di, W. Wang, Universal predictability of mobility patterns in cities, J. R. Soc. Interface 11 (100) (2014) 20140834.
[11] X. Yan, W. Wang, Z. Gao, Y. Lai, Universal model of individual and population mobility on diverse spatial scales, Nature Commun. 8 (1) (2017) 1639.
[12] M. Levy, Scale-free human migration and the geography of social networks, Physica A 389 (21) (2010) 4913–4917.
[13] J.E. Blumenstock, Inferring patterns of internal migration from mobile phone call records: evidence from Rwanda, Inf. Technol. Dev. 18 (2) (2012) 107–125.
[14] A. Ghosh, V. Berg, K. Bhattacharya, D. Monsivais, J. Kertesz, K. Kaski, A. Rotkirch, Migration patterns across the life course of families: Gender differences and proximity with parents and siblings in Finland, 2017, arXiv preprint arXiv:1708.02432.
[15] C.C. Fan, Interprovincial migration, population redistribution, and regional development in China: 1990 and 2000 census comparisons, Prof. Geogr. 57 (2) (2005) 295–311.
[16] E.G. Ravenstein, The laws of migration, J. Stat. Soc. London 48 (2) (1885) 167–235.
[17] E.S. Lee, A theory of migration, Demography 3 (1) (1966) 47–57.
[18] X. Li, H. Xu, J. Chen, Q. Chen, J. Zhang, Z. Di, Characterizing the international migration barriers with a probabilistic multilateral migration model, Sci. Rep. 6 (1) (2016) 32522.
[19] J. Decressin, A. Fatas, Regional labor market dynamics in Europe, Eur. Econ. Rev. 39 (9) (1995) 1627–1655.
[20] W.J. Carrington, E. Detragiache, T. Vishwanath, Migration with endogenous moving costs, Am. Econ. Rev. 86 (4) (1996) 909–930.
[21] E.J. Taylor, The new economics of labour migration and the role of remittances in the migration process, Int. Migr. 37 (1) (1999) 63–88.
[22] M. Khasin, E. Khain, L.M. Sander, Fast migration and emergent population dynamics, Phys. Rev. Lett. 109 (24) (2012) 248102.
[23] E. Zagheni, V.R.K. Garimella, I. Weber, et al., Inferring international and internal migration patterns from twitter data, in: Proceedings of the 23rd International Conference on World Wide Web, ACM, 2014, pp. 439–444.
[24] B. Hawelka, I. Sitko, E. Beinat, S. Sobolevsky, P. Kazakopoulos, C. Ratti, Geo-located twitter as the proxy for global mobility patterns, Cartogr. Geogr. Inf. Sci. 41 (3) (2014) 260–271.
[25] J. Xu, A. Li, D. Li, Y. Liu, Y. Du, T. Pei, T. Ma, C. Zhou, Difference of urban development in China from the perspective of passenger transport around Spring Festival, Appl. Geogr. 87 (2017) 85–96.
[26] Y. Liu, X. Liu, S. Gao, L. Gong, C. Kang, Y. Zhi, G. Chi, L. Shi, Social sensing: A new approach to understanding our socioeconomic environments, Ann. Assoc. Am. Geogr. 105 (3) (2015) 512–530.
[27] J. Li, Q. Ye, X. Deng, Y. Liu, Y. Liu, Spatial-Temporal analysis on Spring Festival travel rush in China based on multisource big data, Sustainability 8 (11) (2016) 1184.
[28] L. Wang, Q. Zhang, Y. Cai, J. Zhang, Q. Ma, Simulation study of pedestrian flow in a station hall during the Spring Festival travel rush, Physica A 392 (10) (2013) 2470–2478.
[29] X. Wang, C. Liu, W. Mao, Z. Hu, L. Gu, Tracing the largest seasonal migration on earth, 2014, arXiv preprint arXiv:1411.0983.
[30] G.K. Zipf, The P_1P_2/D hypothesis: On the intercity movement of persons, Am. Sociol. Rev. 11 (6) (1946) 677–686.
[31] G. Krings, F. Calabrese, C. Ratti, V.D. Blondel, Urban gravity: a model for inter-city telecommunication flows, J. Stat. Mech. Theory Exp. 2009 (07) (2009).
[32] R. Lambiotte, V.D. Blondel, C. De Kerchove, E. Huens, C. Prieur, Z. Smoreda, P. Van Dooren, Geographical dispersal of mobile communication networks, Physica A 387 (21) (2008) 5317–5325.
[33] J. Taylor, S. Rozelle, A. De Brauw, Migration and incomes in source communities: A new economics of migration perspective from China, Econom. Dev. Cult. Chang. 52 (1) (2003) 75–101.
[34] S. Goh, K. Lee, J.S. Park, M. Choi, Modification of the gravity model and application to the metropolitan Seoul subway system, Phys. Rev. E 86 (2) (2012) 026102.
[35] M. Lenormand, S. Huet, F. Gargiulo, G. Deffuant, A universal model of commuting networks, PLoS One 7 (10) (2012) e45985.

[36] J. Macqueen, Some methods for classification and analysis of multivariate observations, in: Proc. of Berkeley Symposium on Mathematical Statistics and Probability, 1967, pp. 281–297.
[37] P.J. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, J. Comput. Appl. Math. 20 (1) (1987) 53–65.
[38] R.L. Thorndike, Who belongs in the family, Psychometrika 18 (4) (1953) 267–276.
[39] The World's Cities in 2016, United Nations, http://www.un.org/en/development/desa/population/publications/pdf/urbanization/the_worlds_cities_in_2016_data_booklet.pdf.
[40] C. Dong, C. Chen, A sociological interpretation of the "Spring Festival" transportation problem, Popul. J. 30 (2008) 31–34 (in Chinese).
[41] Y. Chen, G. Miao, Construction and resolution of the problem of the Spring Festival transportation: Sociological analysis from the perspective of social risk, Summit Publ. Manag. 2 (2008) 004 (in Chinese).