

Urban Dreams of Migrants

A Case Study of Migrant Integration in Shanghai

*Work in progress

Yang Yang
Zhejiang University
yangya@zju.edu.cn

Chenhao Tan
University of Washington
chenhao@chenhaot.com

Zongtao Liu
Zhejiang University
tomstream@zju.edu.cn

Fei Wu
Zhejiang University
wufei@cs.zju.edu.cn

Yueting Zhuang
Zhejiang University
yzhuang@zju.edu.cn

ABSTRACT

An unprecedented human mobility has driven the rapid urbanization around the world. In China, the fraction of population dwelling in cities increased from 17.9% to 52.6% between 1978 and 2012. Such large-scale migration poses both significant challenges for policymakers and important questions for researchers.

To investigate the process of migrant integration, we employ a one-month complete dataset of telecommunication metadata in Shanghai with 54 million users and 698 million call logs. We find systematic differences between locals and migrants in their *mobile communication networks* and *geographical locations*. For instance, migrants have more diverse contacts and move around the city with a larger radius than locals after they settle down. By distinguishing new migrants (who recently moved to Shanghai) from settled migrants (who have been in Shanghai for a while), we demonstrate the integration process of new migrants in their first three weeks. Moreover, we formulate classification problems to predict whether a person is a migrant. Our classifier is able to achieve an F1-score of 0.82 when distinguishing settled migrants from locals, but it remains challenging to identify new migrants because of class imbalance. Yet we show that an increasing fraction of new migrants is classified as locals week by week.

KEYWORDS

urban migrants, migrant integration, mobile communication networks, geographical locations

1 INTRODUCTION

In a big city like L.A. you can spend a lot of time surrounded by hundreds of people yet you feel like an alien or a ghost or something.
— Morley

Millions of people migrate to cities to realize their urban dreams, ranging from pursuing potential job opportunities to embracing an open dynamic culture [28]. These migrants contribute to the prosperity of cities by constituting a substantial part of the workforce in the cities, strengthening the political and economic status of the cities, and bringing diverse cultures to the cities.

Despite the great benefits brought by migration, policymakers and scholars have well recognized that the fast rate of migration

poses great challenges [3, 28]. Segregation and social inequality have become significant issues in the migration process. For instance, migrants may settle in slums with health hazards [6]; they tend to be overworked but underpaid [39]; their children may be excluded from schools [19]. These problems might be even more salient in China, a developing country with an unprecedented speed of urbanization [3]. It is thus an important research question to understand how migrants integrate into a city.

Furthermore, the life of migrants is in constant flux. Some migrants may eventually settle down in the city, while others may go back to their hometown or drift to another city. This process resembles users' behavior in online communities [2, 12, 13, 27, 33, 45], but presents more complex dynamics because moving offline requires considerably more efforts. Accurate modeling and prediction of migrant integration can potentially inform urban policymaking.

In this work, we are interested in two central components of migrant integration: the locations where a migrant lives and moves around, and the people that a migrant interacts with and befriends. First, because cities are divided into neighborhoods with varying characteristics, there may exist systematic differences between locals and migrants in where they live. For example, Figure 1 shows the geographical distributions of locals and migrants compared to the overall average in Shanghai. Somewhat surprisingly, locals are more active in the periphery of the city, whereas migrants relatively concentrate in the center of Shanghai. This observation echoes previous findings that existing residents flee from central cities, known as "white flight" [17]. It yet remains an open question how migrants' active areas evolve as they integrate into the city.

Another important aspect of migrant integration is how migrants build their personal networks. As humans are social animals, whether a migrant can successfully develop a personal network is crucial in her integration process [20]. In particular, Yue et al. [53] show that migrant-resident ties are significantly associated with migrant integration. However, it remains unclear how a migrant makes initial friends and then slowly build a personal network in a new city. It is also unknown what characteristics differentiate the social networks of migrants from those of locals.

Organization and highlights of this work. In order to investigate the above two aspects, we conduct a case study of Shanghai, one of the biggest cities in China, and present the first large-scale quantitative exploration of migrant integration. We employ a *one-month complete* dataset of telecommunication metadata from China

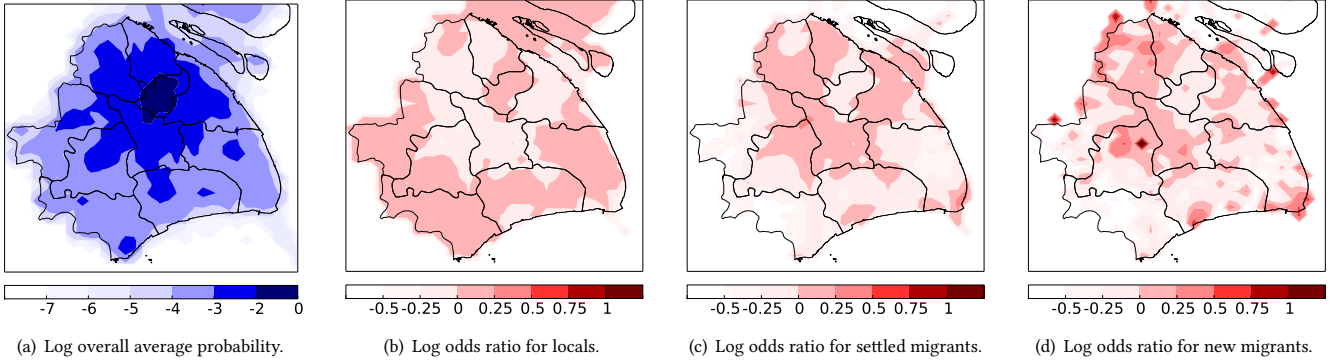


Figure 1: Geographical distributions of locals, settled migrants (who have been in a new city for a while), and new migrants (who recently moved to a new city) in Shanghai. Each person is represented by the center of their active areas (formal definitions will be presented in Section 3). Figure 1(a) shows the log probability of all users in each region and this probability constitutes a comparison point for the other figures. Each of the right three figures shows the log odds ratio of each group compared to the overall average in Figure 1(a), i.e., $\log P_{\text{group}} - \log \bar{P}$, where \bar{P} is the overall average in Figure 1(a) and P_{group} , $\text{group} \in \{\text{locals, settled migrants, new migrants}\}$ is the probability to fall in a region within that particular group of people. Intuitively, a red region in the right three figures suggests that this group of people are disproportionately frequent in that region. Settled migrants tend to be in the central part of the city, while locals are in the periphery. New migrants are similar to settled migrants by and large, but have a few dark areas on the periphery. The darkest point in Figure 1(d) correspond to Songjiang University Town, a hub of universities.

Telecom,¹ which contains 698 million call logs between 54 million mobile users in Shanghai. To identify a comparison point of migrants, we define *locals* as the persons that were born in Shanghai, the counterpart of migrants. As migrants may undergo different stages in their integration, we further differentiate migrants that have been in a new city for a while, *settled migrants*, from migrants that recently moved to a new city, *new migrants*. We present details of the dataset in Section 2.

First, we explore how locals, settled migrants, and new migrants differ in their mobile communication networks and geographical locations in Section 3. We find interesting differences between these three groups. For instance, in terms of communication networks, a substantial fraction of new migrants’ contacts are fellow townsmen, people who were born in the same province. This pattern suggests that townsmen are essential for new migrants to build their initial personal networks in a new city. Surprisingly, settled migrants have an even higher fraction of townsmen contacts, indicating that they may have grown their townsmen network as they stay longer in Shanghai. In terms of locations, in addition to the differences in Figure 1, we find that settled migrants tend to have a larger radius.

Second, we use the calling logs over different time periods to give a brief dynamic view of the integration process in Section 4. Despite the short time span, we observe that new migrants become increasingly similar to settled migrants in most characteristics, while features of settled migrants and locals tend to be stable over time. This contrast suggests that the features that we employ can indeed reflect the integration process to some extent. Meanwhile, we observe that the integration slows down in the final week. One possible explanation is that not all new migrants eventually become

settled migrants and the slow integration is due to the ones that encounter difficulty fitting into the city. This hypothesis is worth further investigation.

Finally, we formulate prediction tasks to distinguish migrants from locals in Section 5. Using the features that we propose, we are able to build a classifier that significantly outperforms the baselines and achieve an F1-score of 0.82 on settled migrants, indicating that it is not a difficult prediction task to separate settled migrants from locals. We also observe that if we apply this classifier to new migrants, an increasing fraction of new migrants is classified as locals over time. However, it remains challenging to identify new migrants because the number of new migrants is very small compared to settled migrants and locals.

Our work is a first step towards understanding migrant integration and informing urban policymakers. This challenging problem necessarily involves efforts from a wide range of disciplines, including anthropology, computer science, economics, sociology and urban planning. We thus provide an overview of related work on this issue in Section 6 and offer some concluding discussions in Section 7.

2 EXPERIMENTAL SETUP

In this section, we introduce our dataset and the framework that we use to study mobile communication networks and geographical information of locals and migrants.

2.1 Dataset

We use a dataset that contains *complete* telecommunication records between mobile users using China Telecom in Shanghai, spanning a month from September 3rd, 2016, to September 30th, 2016 (four

¹China Telecom Corporation is a Chinese state-owned telecommunication company and the third largest mobile service providers in China.

weeks). The data is provided by China Telecom, the third largest mobile service provider in China. Our dataset consists of about 54 million users and 698 million call logs between them. A call log was recorded as long as it was made in Shanghai and either the caller or the callee was a user of China Telecom (some of the 54 million users use other mobile services). Each call log contains the caller’s number, the callee’s number, the starting time, and the ending time. Since personal identification is required to obtain a mobile number, we are able to retrieve personal attributes, including age, sex, and birthplace, for users of China Telecom that opened their accounts in Shanghai.² In addition, we have the GPS location of the corresponding telecommunications tower used during the call for users of China Telecom, which roughly approximates the locations of them. Our dataset was anonymized by China Telecom for privacy concerns. Throughout the paper, we report only average statistics without revealing any identifiable information of individuals.

2.2 Framework

We categorize users in our dataset into three groups based on their birthplaces and this categorization constitutes the basis for our computational framework. We refer to people that were born in Shanghai as *locals*. The rest people who were not born in Shanghai are migrants. To assess different stages of migrant integration, we separate migrants that have no call logs in the first week (*new migrants*), from migrants that have at least one call log in the first week (*settled migrants*). We further require each local and settled migrant to have call logs at every week, and each new migrant to have call logs at each of the last three weeks, to make sure that these users lived in Shanghai during our four-week span. We filtered around 15,000 users that have abnormally high degrees, who likely corresponded to fraudsters, delivery persons, or customer services according to a user type list provided by China Telecom. In the end, we have 1.7M *locals*, 1.0M *settled migrants*, and 22K *new migrants*.

Mobile communication networks. One core component of our study is a weekly mobile communication network based on the call logs. Grouping by weeks allows us to account for variations between weekdays and weekends. Formally, we build a directed graph $G_t = (V_t, E_t)$ for each week t ($t \in \{1, 2, 3, 4\}$), where V_t is the set of users, and each directed edge $e_{ij} \in E$ indicates that v_i calls v_j ($v_i, v_j \in V_t$). Note that only a subset of users in V_t are labeled as locals, settled migrants or new migrants (in total around 3 million users). This subset is the focus of our study.

Geographical locations. Another component is the geographical locations that a person is active at. Specifically, for each call a person makes, we have access to the GPS location from the corresponding telecommunications tower. We use each week as a window and collect all the locations that a person makes calls in that week, and refer to this ordered list of locations for user v at week t as $L_v^t = [l_1, \dots, l_n]$, where l_i contains the latitude and the longitude. We have geographical locations for the subset of users with labels since they are all users of China Telecom by definition.

We will define computational features based on these two components in Section 3.

Table 1: List of features in this paper. We view all directed edges as undirected except in measuring reciprocal calls. For demographics related features, we only include users for whom we have the corresponding information. In this work, we use contacts and friends interchangeably.

Feature	Description
Demographics of user v’s friends in G_t	
similar-age	The fraction of v ’s friends that are at similar ages with v (± 10 years).
same-sex	The fraction of v ’s friends having the same sex with v .
local	The fraction of v ’s friends who were born in Shanghai.
townsmen	The fraction of v ’s friends that were born in the same province with v but not in Shanghai. This feature is always 0 for locals, so it is not included in prediction experiments in Section 5.
Ego-network characteristics of user v in G_t	
degree	The number of v ’s unique contacts.
weighted degree	The number of calls v makes.
neighbor degree	The average degree of v ’s contacts.
CC	Clustering coefficient of v ’s ego-network, i.e., $\frac{ \{(s,t) (s,t) \in E_t \} }{d_v(d_v-1)}$, where s and t are v ’s friends, and d_v is v ’s degree.
Call behavior in G_t	
call duration	v ’s average call duration.
duration variance	variance of v ’s call duration.
province diversity	Entropy of the distribution of birth provinces among v ’s contacts, defined as $-\sum_i p_i \log_2 p_i$, where p_i is the probability that a contact of v was born in province i .
reciprocal call	The probability that v ’s contacts also call v in week t .
Geographical features of v at week t	
center	The latitude and longitude of a user v ’s center of mass l_{CM} , $l_{CM} = \frac{1}{ L_v^t } \sum_{l \in L_v^t} l$.
max radius	The maximal distance of v from her center of mass, i.e., $\max_{l \in L_v^t} l - l_{CM} $.
average radius	The average distance of v from her center of mass, i.e., $\frac{1}{ L_v^t } \sum_{l \in L_v^t} l - l_{CM} $.
moving distance	The total distance that v moves, i.e., $\sum_i l_i - l_{i-1} $.
average distance	The average distance that v moves, i.e., $\frac{1}{ L_v^t } \sum_i l_i - l_{i-1} $.

3 LOCALS, SETTLED MIGRANTS, NEW MIGRANTS

To understand how locals, settled migrants and new migrants differ from each other, we examine a wide range of features from people’s mobile communication networks and geographical locations. To observe the fresh state of urban migrants without much integration, we use the data from the first week that new migrants

²We obtain a person’s birthplace using the personal identity card number.

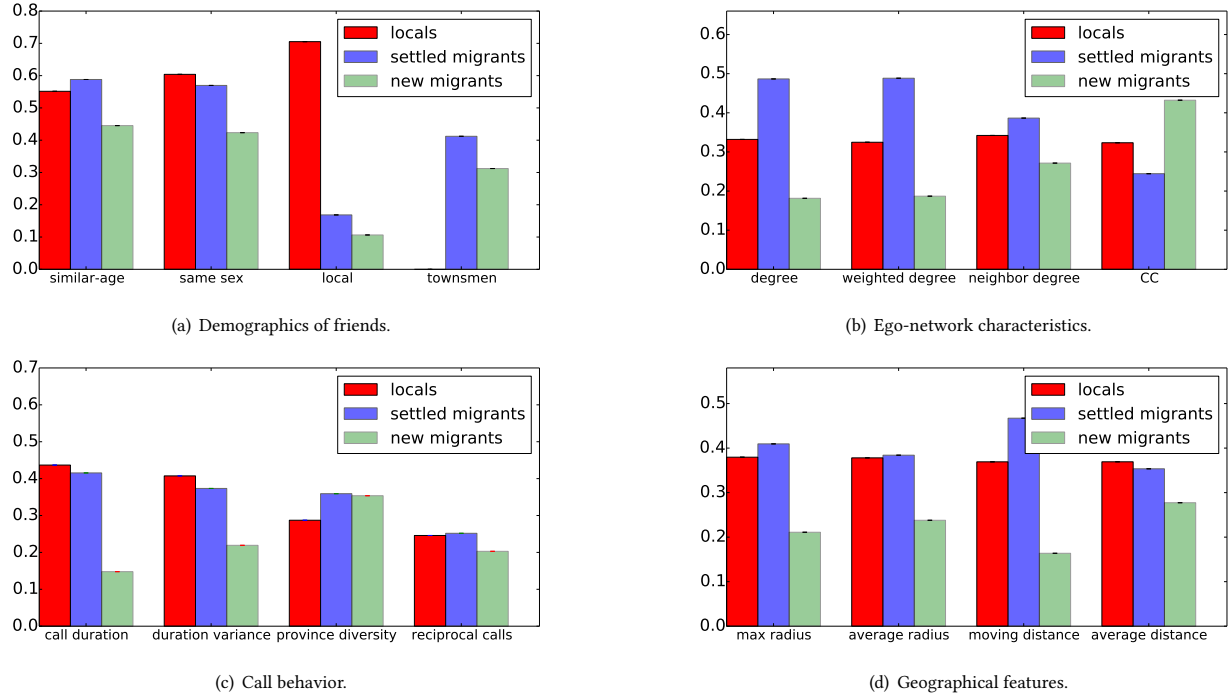


Figure 2: Feature comparison between locals, settled migrants and new migrants. Different colors represent different groups of people. Since different features may end up in very different scales, we normalize each feature group in this figure so that their means sum up to 1, except for demographics of contacts and fraction of reciprocal calls (they all naturally fall between 0 and 1). Error bars represent standard errors, and they are tiny.

joined China Telecom (week 2) in this section. We will focus on the integration process in Section 4, in which we also show that most features do not change for locals and settled migrants in future weeks. Table 1 provides an overview of the features that we consider. Demographics of users’ friends,³ ego-networks features, and call behavior derive from the mobile communication networks, while geographical features come from location information. In the following, we will explain the motivation and related theories of each feature.

Demographics of contacts (Figure 2(a)). A person’s mobile communication network can reasonably approximate her social network. Locals likely maintain very different social networks from migrants since they have grown up in this city. Also, as a person settle down in a new city, her social network may change dramatically. Existing studies suggest that kin relationships play an important role in determining the destination of migration [20] and relationship with locals are crucial for migrant integration [32]. We look at the demographics of contacts in age, sex, and birthplaces.

Homophily in sex and age. It is well recognized that people tend to make friends with those who are similar to themselves, also

known as homophily [34]. We observe interesting contrasts regarding homophily of age and sex. Locals show the strongest homophily in sex, i.e., locals have the largest fraction of contacts with the same sex. Surprisingly, in terms of the absolute fraction, new migrants have more contacts with a different sex than with the same sex (only around 40%). In contrast, locals are less likely to have contacts at similar ages than settled migrants, but more than new migrants.

Birthplaces. The most striking difference lies in that around 70% of a local person’s contacts are also locals. This number is much lower for settled migrants, and the lowest for new migrants. Townsmen, people who share the same hometown, are an important component of a new migrant’s initial network (30% of new migrants’ contacts are townsmen). This observation echoes existing findings regarding kin relationship. In comparison, settled migrants have an even higher fraction of townsmen in their contacts, which suggests that new migrants get to know more people from the same hometown as they integrate into a city. These observations are consistent with homophily, but they also indicate that urban migrants maintain a relatively separate personal network from locals.

Ego-network characteristics (Figure 2(b)). As expected, new migrants have the smallest degree and weighted degree. However, settled migrants tend to have the largest degree, larger than locals. This indicates an interesting transition that migrants may undergo. Maybe because of homophily, neighbors of settled migrants also

³Note that although we do not have demographics information for users using other service providers, we have demographics information for a much larger set than the ones labeled locals, settled migrants or new migrants because we require these users to be active in each week.

have the largest average degree, and neighbors of new migrants have the smallest average degree.

Clustering coefficient measures the fraction of triangles in the ego-networks. It roughly reflects how connected a person’s contacts are to each other. Interestingly, new migrants present the largest clustering coefficient, while settled migrants have the lowest. It may suggest that new migrants start with a close-knit group when they move to a big city like Shanghai. Connecting with our previous observations, this close-knit group tend to come from the same province as the new migrants. It is worth noting that this could also relate to that new migrants have the smallest ego-networks.

Calling behavior (Figure 2(c)). The duration of calls may reflect the nature of the relations between a person and her contacts. Calls of long duration likely involve intimate relations or are driven by substantial businesses, while calls of short duration tend to be quick check-ins or relate to small incidences. Our data show that locals and settled migrants have similar levels of average call duration, much larger than new migrants. Similar trends show up in the variance of call duration.

Regarding the diversity of provinces in a person’s contacts, settled migrants have the most diverse group of contacts, while locals have the lowest. This pattern resonates with previous observations that locals have about 70% of contacts that are also locals.

Finally, we find that locals and settled migrants are more likely to have reciprocal relationships with their contacts, while the fraction of reciprocal calls is the lowest for new migrants. This again shows that the personal networks of new migrants are still nascent. Note that the difference is much less dramatic than that in call duration.

Geographical patterns (Figure 2(d)). The mobility of people in different groups can be reflected by their locations over time. As we have discussed in the introduction, both settled migrants and new migrants tend to move around the central part of Shanghai, while locals are more disproportionately frequent in the periphery. Regarding the radius of a person’s movement around her own center, we observe that settled migrants have the largest radius both in terms of max radius and average radius. This suggests that although new migrants start with a smaller active area than locals, settled migrants move in an even larger area than locals.

Total moving distance is correlated with the total number of calls that a person makes. We thus discover the same ordering as in weighted degree. However, locals tend to move the most distance between calls on average, while new migrants move the shortest distance. This further suggests that new migrants have a smaller active area than locals.

Summary. Comparing settled migrants to locals, we observe that settled migrants have more active and diverse behavior patterns both in mobile communication networks and in geographical movements. Meanwhile, new migrants present different characteristics from both settled migrants and locals. This suggests that new migrants go through significant changes in their communication networks and geographical locations as they slowly settle down.

4 INTEGRATION OF NEW MIGRANTS

Given the differences between locals, settled migrants and new migrants that we have observed, we now investigate the integration process of new migrants. Since a subset of new migrants eventually

locals > settled > new, locals < settled < new	New migrants are moving towards locals and settled migrants are in the middle of this process.
new > locals > settled, new < locals < settled	New migrants move towards locals initially, but eventually move away from them and remain different from locals after they settle down.
locals > new > settled, locals < new < settled	Settled migrants and locals are different, and new migrants never move towards locals.

Table 2: All possible orderings of feature values between locals, settled migrants, and new migrants and their corresponding implications.

become settled migrants, we hypothesize that the features of new migrants will grow more similar to those of settled migrants in week 3 and week 4. Indeed, we find that new migrants are slowly “becoming” settled migrants in most features. Figure 3 presents how some features of locals, settled migrants and new migrants change over the four weeks (new migrants only moved to Shanghai in week 2). Although existing studies have argued that different generations of migrants can exhibit different characteristics [11, 37], our observation shows that the features that we propose are robust to generation differences, or Shanghai is too young a city to observe generation gaps from telecommunication records.

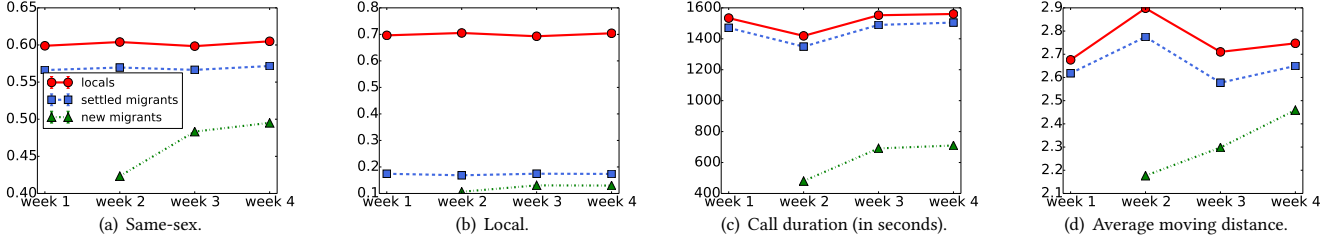
The more interesting comparison is with locals. One possible way to evaluate migrant integration is whether they become more similar to locals over time. Depending on how the features of locals compare to new migrants and settled migrants, we can observe several possible trajectories as shown in Table 2.

An ideal integration process suggests that new migrants become more similar to locals, and settled migrants are a middle state in this process, i.e., the orderings should follow *locals > settled > new* or *locals < settled < new*, and we should observe that the features of new migrants move towards settled migrants in week 3 and week 4. Some features indeed show consistent trajectories with this ideal integration process, including fraction of same-sex contacts, fraction of local contacts, call duration, duration variance, and average moving distance. It makes sense that migrants are probably never going to match locals in fraction of local contacts, but such matching may happen in average moving distance and call duration.

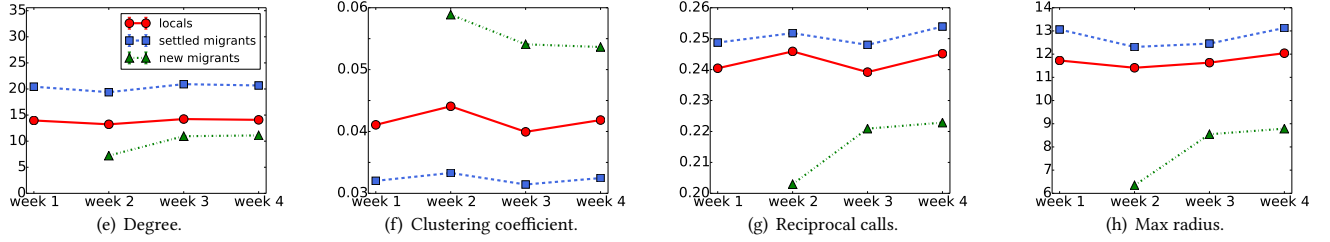
However, for the majority of features, we observe that although new migrants initially move towards locals, they may eventually become further away from locals after settling down. These features include degree, weighted degree, average degree of neighbors, clustering coefficient, fraction of friends with similar age, fraction of reciprocal calls, max radius, average radius, total moving distance. In particular, all features in ego-network characteristics follow this trajectory, suggesting that new migrants eventually build quite different communication networks from locals.

It is rare that new migrants do no move towards locals at all but become more different from locals in the integration process. This

new migrants move towards settled migrants and both move towards locals



new migrants move towards locals initially, but will likely eventually move away from locals



new migrants move towards settled migrants but away from locals

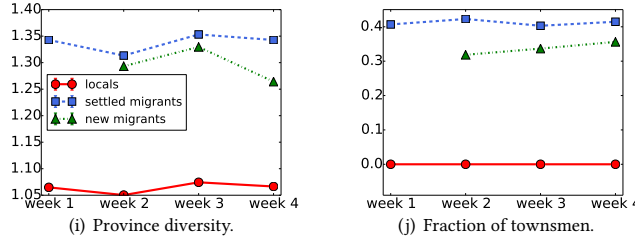


Figure 3: Integration process of new migrants. Each figure presents how the values of a feature evolves over the four weeks for locals, settled migrants, and new migrants. We choose four samples in the first two rows because there are more than four features that belong to those categories. Error bars represent standard errors.

only happens in province diversity and townsmen.⁴ Both point to the fact that new migrants start with a more diverse communication network in terms of birthplaces than locals, and their networks get even more diverse over their stay in Shanghai. Note that there is a decline in province diversity for new migrants in week 4 but they are still closer to settled migrants than to locals.

Interestingly, in some features, we observe that the integration slows down or converges in week 4 for new migrants. This is likely due to the fact that not all new migrants are going to become settled migrants. As a result, we can already see that the integration process stops or slows down in week 4 for a subset of these people. In addition to these summary statistics, geographically, we also observe that the centers of new migrants are expanding over time, as shown in Figure 4.

Discussion. Overall, we find that new migrants are settling down and gradually becoming settled migrants, and this observation is robust with potential generation gaps. However, in a substantial fraction of the features, although new migrants are temporarily

moving towards locals, they are probably going to become different from locals as settled migrants do. In other words, despite settling down, settled migrants remain fairly different from locals.

5 DIFFICULTY OF DISTINGUISHING MIGRANTS

We set up two prediction tasks to assess the difficulty of distinguishing migrants from locals with the features that we propose. Since the number of new migrants is much smaller compared to settled migrants and locals (22K vs. 1.0M and 1.7M), we employ two formulations in this section. First, we propose a binary classification task to distinguish settled migrants from locals. We then apply this binary classifier to new migrants to evaluate how often a new migrant would be mistakenly thought of as a local by our classifier. This misclassification rate can reflect how well new migrants have integrated, at least in terms of fooling our classifier. Second, we work on the more challenging three-way classification problem to identify new migrants, settled migrants, and locals.

Experiment setup. In both prediction tasks, each instance consists of features based on a user’s calling logs within a particular

⁴It is tricky for townsmen, since locals do not have townsmen that are not from Shanghai and always have 0 in this feature.

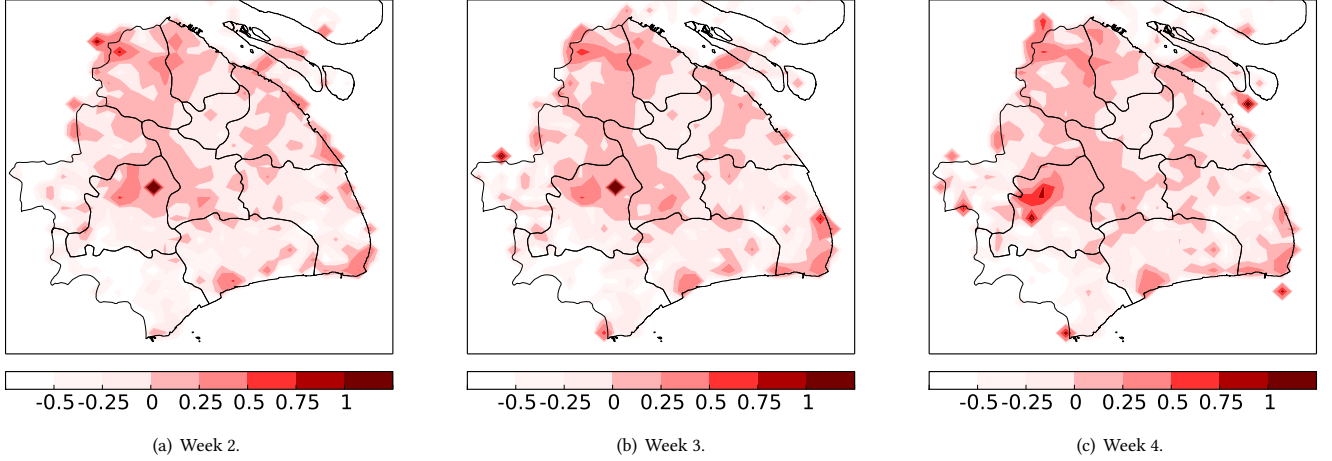


Figure 4: Evolution of geographical distributions from week 2 to week 4 for new migrants. The computation procedure is the same as Figure 1 and each figure shows the log odds ratio compared to the overall average in each week. New migrants are slowly expanding into larger parts of the city.

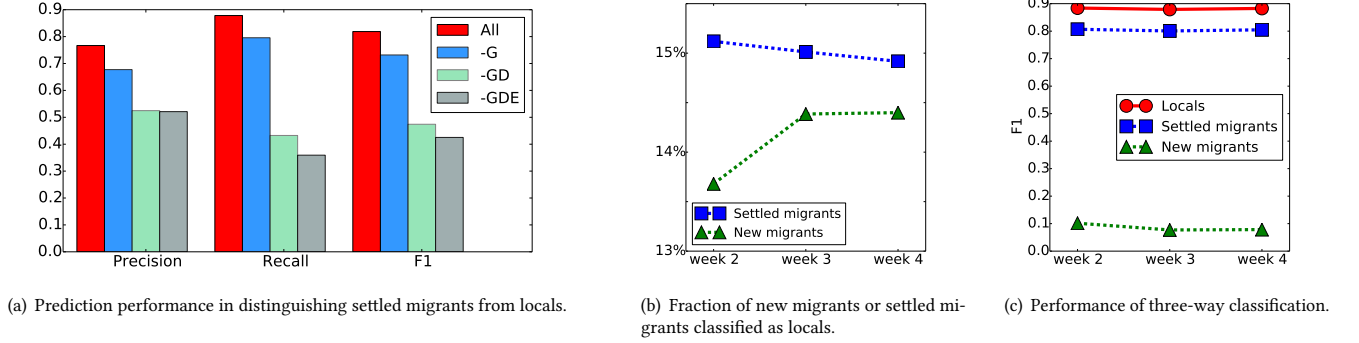


Figure 5: Performance of distinguishing migrants. Figure 5(a) and Figure 5(b) are results from the binary classification between settled migrants and locals. Figure 5(a) shows the performance of feature ablation (we show the feature class that is the most influential in each ablation step), while Figure 5(b) shows the fraction of settled migrants and new migrants that is mistakenly classified as locals over time. Figure 5(c) shows the F1 scores in the three-way classification problem over time.

week. We randomly draw 50% of users and use their calling logs in week 2 to train the classifier. We use the remaining data to test the classifier (50% of data in week 2, and 100% of data in week 3 and week 4). Based on the calling logs of a given user, we extract all features listed in Table 1 except “townsmen”, as measuring the fraction of townsmen relies on the user’s label (the user’s birthplace). We use precision, recall, and F1-score for evaluation, with the minority class (i.e., migrants) as the target class. For the classifier, we use ℓ_2 -regularized logistic regression. We choose the best ℓ_2 penalty coefficient using 5-fold cross-validation in training data.

5.1 Settled Migrants vs. Locals (Binary Classification)

Prediction performance (Figure 5(a)). It turns out to be relatively easy to distinguish settled migrants from locals. We can achieve an F1-score of 0.82 with all the features that we propose.

We further analyze the contribution of each type of features by removing them one by one. In the first removal step, we find that geographical features were the most influential feature set, i.e., F1 drops the most (0.11) if we remove geographical features (-G in Figure 5(a)). Demographics is the most important in the second step (-GD in Figure 5(a)). In the third step, removing ego-network features is the choice, leaving us with a classifier that only uses call behavior (-GDE in Figure 5(a)). F1 drops almost 50% to 0.43 after removing these three types of features. In addition, the prediction performance of the binary classifier is robust over time: F1-scores on each week vary little (<0.0007).⁵

Integration of new migrants (Figure 5(b)). One way to evaluate the integration of new migrants is to measure how often this binary classifier would mistakenly classify a new migrant as a local. We present the fraction of misclassified locals among settled

⁵We do not show the plot in the interest of space.

migrants as a comparison point. Overall, settled migrants are more likely to be misclassified as locals than new migrants (e.g., 15.2% vs 13.6% in week 2). However, we observe an increasing trend for new migrants over the three weeks. In week 3, the misclassified fraction of new migrants increases to 14.4%, suggesting that they become similar to locals over time. The growth slows down in week 4, which is consistent with our findings in Section 4. To our surprise, the fraction of settled migrants misclassified as locals slightly decreases over time. This suggests that some settled migrants could have stopped integrating with locals after settling down, but build their own communities and keep their own lifestyles instead.

5.2 Identifying New Migrants (Three-way Classification)

The three-way classification problem is challenging due to the relatively small number of new migrants (about 0.8% of all instances). The classifier only achieves an F1-score of 0.1 on identifying new migrants and this performance drops over time, while the performance on settled migrants and locals remains similar to the binary classification task (Figure 5(c)). We find that more new migrants are classified as settled migrants or locals incorrectly by the classifier over time. This is consistent with the observation that new migrants are becoming similar to settled migrants or locals in most characteristics despite the short time span, while settled migrants and locals tend to stay constant.

6 RELATED WORK

The urbanization process poses significant challenges for the society that require efforts from various disciplines. We review relevant studies in the following four aspects.

Migrant integration. Migrant integration is a well-recognized research question in many disciplines, including anthropology, economics, sociology and urban planning [9]. Most relevant to our work is the study of urban migration [7, 8, 16, 18, 19, 41–43, 49]. In addition to the effect of nation-states and demographics (ethnic groups, rural vs. urban) on urban migrant integration, Schiller and Çağlar [41] argue that the role of migrants in the cities depends on the rescaling of the cities themselves. Government policy and agenda-setting also play an important role in the integration process [43]. Beyond our scope, immigrants (migrants to a new country) and refugees (a subgroup of immigrants) have also received significant interests [4, 5, 23, 44, 48].

Urban migrants in China. The unprecedented speed of development and the huge population in China have sparked a battery of studies on urban migrants in China [1, 10, 25, 31, 32, 47, 50–54]. There are at least three perspectives as suggested in [25]: those of the migrants themselves, of the urban employers, and of the government. Our work presents the perspective of migrants based on their telecommunication patterns. It is worth noting that a central topic in public policy regarding migrants in China is the impact of the “hukou” system, a household registration system that limits the benefits and social welfare of migrants [1, 51, 52]. Finally, although satisfying migrants’ needs is among the challenges that Bai et al. [3] highlight in the Chinese government’s urbanization strategy, little attention is paid to the social integration of migrants.

Urban computing. Data-driven studies related to cities have gained importance recently and led to a new term, urban computing [1, 15, 21, 24, 38, 40, 55, 56]. These studies combine heterogeneous data sources, including location data, social media activity data, mobile phone data and survey data, to propose metrics for city development and potentially guide urban policies. For instance, Zheng et al. [56] employ GPS data from taxicabs to evaluate transportation system in Beijing, while De Nadai et al. [14] use mobile phone data to extract human activity and propose metrics to measure urban diversity. Recently, Twitter has also been used as a tool for demographical studies in understanding the global mobility patterns [15].

Temporal social networks and online communities. Our work is also relevant to studies on the evolution of networks [22, 26, 29, 30, 35, 36, 46]. Using data from online social media, these studies explore the connection between individual behavior and global network properties. For instance, Viswanath et al. [46] find that links in the activity network tend to come and go rapidly over time, and the strength of ties exhibits a general decreasing trend as the social network link ages on Facebook. Leskovec et al. [29] develop a triangle-closing model to explain network evolution. In addition, studies have investigated the process of new user integration in online communities [2, 12, 13, 27, 33, 45]. In particular, McAuley and Leskovec [33] examine the process of how a new user becomes an expert on review websites.

7 CONCLUDING DISCUSSIONS

We present the first large-scale study on migrant integration based on telecommunication metadata. We demonstrate the differences between locals, settled migrants, and new migrants, and show how characteristics of a migrant’s communication network and geographical locations evolve in the integration process. For instance, although migrants build connections with locals in the integration process, townsmen become increasingly important in their networks. We also find that migrants increase the size of their active area and their moving distance in the city over time. Overall, in most features, we are able to observe how new migrants gradually become settled migrants, but that does not necessarily indicate that they become more similar to locals.

We further formulate prediction problems to distinguish migrants from locals. A classifier based on the features that we propose can achieve a high F1-score on settled migrants, around 0.82. This confirms that migrants are still fairly different from locals in their behavior patterns, supporting studies on the segregation of migrants. Meanwhile, we also observe that a larger fraction of new migrants is classified as locals over time, partly documenting the integration process.

Our work is limited by the data that we have access to. One month is too short to capture the full integration process of migrants. Also, people’s lives are much richer and more dynamic than what we are able to capture using telecommunication metadata. Job situations, health records, and daily interactions can all potentially offer us a more in-depth understanding of migrant integration. In addition, it is challenging to address the imbalance problem between the three groups for realistic applications because new migrants are the minority by definition. Last but not least, although China

Telecom is a major service provider and Shanghai is an important global city, the selection bias presented in our data may limit the generalizability of our findings.

As urbanization is happening at an unprecedented rate and data collection becomes ubiquitous in smart cities, there are tremendous opportunities for data-driven approaches to understanding and improving migrant integration. For instance, it would be useful to identify migrants that have trouble fitting into the city and provide timely and useful support. We hope that our study can encourage more researchers in our community to examine this problem from different perspectives and eventually lead to methodologies and applications that benefit policymaking and millions of migrants.

REFERENCES

- [1] Farzana Afridi, Sherry Xin Li, and Yufei Ren. 2015. Social identity and inequality: The impact of China's hukou system. *Journal of Public Economics* 123 (2015), 17–29.
- [2] Lars Backstrom, Dan Huttenlocher, Jon Kleinberg, and Xiangyang Lan. 2006. Group Formation in Large Social Networks: Membership, Growth, and Evolution. In *Proceedings of KDD*.
- [3] Xuemei Bai, Peijun Shi, and Yansui and Liu. 2014. Society: Realizing China's urban dream. *Nature News* 509, 7499 (2014), 158.
- [4] Frank D Bean and Gillian Stevens. 2003. *America's Newcomers and the Dynamics of Diversity*. Russell Sage Foundation.
- [5] Gary S Becker and Diane Coyle. 2011. The challenge of immigration: a radical solution. *Institute of Economic Affairs Monographs Occasional Paper* 145 (2011).
- [6] Donatien Beguy, Philippe Bocquier, and Eliya Msiyaphazi Zulu. 2010. Circular migration patterns and determinants in Nairobi slum settlements. *Demographic Research* 23 (2010), 549.
- [7] Martin Brouckerhoff. 1995. Child survival in big cities: the disadvantages of migrants. *Social Science & Medicine* 40, 10 (1995), 1371–1383.
- [8] Lawrence A Brown and Eric G Moore. 1970. The intra-urban migration process: a perspective. *Geografiska Annaler. Series B, Human Geography* 52, 1 (1970), 1–13.
- [9] Stephen Castles, Hein De Haas, and Mark J Miller. 2013. *The age of migration: International population movements in the modern world*. Palgrave Macmillan.
- [10] Juan Chen, Deborah S. Davis, Kaming Wu, and Haijing Dai. 2015. Life satisfaction in urbanizing China: The effect of city size and pathways to urban residency. *Cities* 49 (2015), 88–97.
- [11] Barry R Chiswick and Noyna DebBurman. 2004. Educational attainment: analysis by immigrant generation. *Economics of Education Review* 23, 4 (2004), 361–379.
- [12] David Crandall, Dan Cosley, Daniel Huttenlocher, Jon Kleinberg, and Siddharth Suri. 2008. Feedback Effects Between Similarity and Social Influence in Online Communities. In *Proceedings of KDD*.
- [13] Cristian Danescu-Niculescu-Mizil, Robert West, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. No Country for Old Members: User Lifecycle and Linguistic Change in Online Communities. In *Proceedings of WWW*.
- [14] Marco De Nadai, Jacopo Staiano, Roberto Larcher, Nicu Sebe, Daniele Quercia, and Bruno Lepri. 2016. The death and life of great Italian cities: a mobile phone data perspective. In *Proceedings of WWW*.
- [15] Mark Dredze, Manuel García-Herranz, Alex Rutherford, and Gideon Mann. 2016. Twitter as a Source of Global Mobility Patterns for Social Good. *CoRR abs/1606.06343* (2016).
- [16] Claude S Fischer. 1982. *To dwell among friends : personal networks in town and city*. University of Chicago Press, Chicago.
- [17] William H Frey. 1979. Central city white flight: Racial and nonracial causes. *American Sociological Review* (1979), 425–448.
- [18] Edward L Glaeser and David C Mare. 2001. Cities and skills. *Journal of labor economics* 19, 2 (2001), 316–342.
- [19] Charlotte Goodburn. 2009. Learning from migrant education: A case study of the schooling of rural migrant children in Beijing. *International Journal of Educational Development* 29, 5 (2009), 495–504.
- [20] Douglas T Gurak and Fee Caces. 1992. Migration networks and the shaping of migration systems. *International migration systems: A global approach* (1992), 150–176.
- [21] Desislava Hristova, Matthew J Williams, Mirco Musolesi, Pietro Panzarasa, and Cecilia Mascolo. 2016. Measuring Urban Social Diversity Using Interconnected Geo-Social Networks. In *Proceedings of WWW*.
- [22] Abigail Z Jacobs, Samuel F Way, Johan Ugander, and Aaron Clauset. 2015. Assembling Thefacebook: Using Heterogeneity to Understand Online Social Network Assembly. In *Proceedings of Web Science*.
- [23] Karen Jacobsen and Loren B Landau. 2003. The dual imperative in refugee research: some methodological and ethical considerations in social science research on forced migration. *Disasters* 27, 3 (2003), 185–206.
- [24] Shan Jiang, Gaston A Fiore, Yingxiang Yang, Joseph Ferreira Jr, Emilio Frazzoli, and Marta C Gonzalez. 2013. A review of urban computing for mobile phone traces: current methods, challenges and opportunities. In *Proceedings of the 2nd ACM SIGKDD international workshop on Urban Computing*.
- [25] John Knight, Lina Song, and Jia Huaibin. 1999. Chinese rural migrants in urban enterprises: three perspectives. *The Journal of Development Studies* 35, 3 (1999), 73–104.
- [26] Gueorgi Kossinets and Duncan J Watts. 2006. Empirical analysis of an evolving social network. *Science* 311, 5757 (2006), 88–90.
- [27] Cliff Lampe and Erik Johnston. 2005. Follow the (slash) dot: effects of feedback on new members in an online community. In *Proceedings of GROUP*.
- [28] June J.H. Lee. 2015. *World migration report 2015: Migrants and Cities: New Partnerships to manage mobility*.
- [29] Jure Leskovec, Lars Backstrom, Ravi Kumar, and Andrew Tomkins. 2008. Microscopic Evolution of Social Networks. In *Proceedings of KDD*.
- [30] Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. 2007. Graph evolution: Densification and shrinking diameters. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 1, 1 (2007), 2.
- [31] Zai Liang. 2001. The age of migration in China. *Population and development review* 27, 3 (2001), 499–524.
- [32] Ye Liu, Zhigang Li, and Werner Breitung. 2012. The social networks of new-generation migrants in China's urbanized villages: A case study of Guangzhou. *Habitat International* 36, 1 (2012), 192–200.
- [33] Julian John McAuley and Jure Leskovec. 2013. From Amateurs to Connoisseurs: Modeling the Evolution of User Expertise Through Online Reviews. In *Proceedings of WWW*.
- [34] Miller McPherson, Lynn Smith-Lovin, and James M Cook. 2001. Birds of a feather: Homophily in social networks. *Annual review of sociology* 27, 1 (2001), 415–444.
- [35] Ron Milo, Shalev Itzkovitz, Nadav Kashtan, Reuven Levitt, Shai Shen-Orr, Inbal Ayzenshtat, Michal Sheffer, and Uri Alon. 2004. Superfamilies of evolved and designed networks. *Science* 303, 5663 (2004), 1538–1542.
- [36] Ashwin Paranjape, Austin R Benson, and Jure Leskovec. 2017. Motifs in Temporal Networks. In *Proceedings of WSDM*.
- [37] Alejandro Portes and Lingxin Hao. 2002. The price of uniformity: Language, family and personality adjustment in the immigrant second generation. *Ethnic and Racial Studies* 25, 6 (2002), 889–912.
- [38] Daniele Quercia, Luca Maria Aiello, Rossano Schifanella, and Adam Davies. 2015. The Digital Life of Walkable Streets. In *Proceedings of WWW*.
- [39] Shahra Razavi and Silke Staab. 2010. Underpaid and overworked: A cross-national perspective on care workers. *International Labour Review* 149, 4 (2010), 407–422.
- [40] Jonathan Reades, Francesco Calabrese, Andres Sevtsuk, and Carlo Ratti. 2007. Cellular census: Explorations in urban data collection. *IEEE Pervasive Computing* 6, 3 (2007).
- [41] Nina Glick Schiller and Ayse Çağlar. 2009. Towards a comparative theory of locality in migration studies: Migrant incorporation and city scale. *Journal of ethnic and migration studies* 35, 2 (2009), 177–202.
- [42] Nina Glick Schiller and Ayse Simsek-Caglar. 2011. *Locating migration: Rescaling cities and migrants*. Cornell University Press.
- [43] Peter WA Scholten. 2013. Agenda dynamics and the multi-level governance of intractable policy controversies: the case of migrant integration policies in the Netherlands. *Policy Sciences* 46, 3 (2013), 217–236.
- [44] Alison Strang and Alastair Ager. 2010. Refugee integration: Emerging trends and remaining agendas. *Journal of Refugee Studies* 23, 4 (2010), 589–607.
- [45] Chenhao Tan and Lillian Lee. 2015. All who wander: On the prevalence and characteristics of multi-community engagement. In *Proceedings of WWW*.
- [46] Bimal Viswanath, Alan Mislove, Meeyoung Cha, and Krishna P Gummadi. 2009. On the Evolution of User Interaction in Facebook. In *Proceedings of the 2nd ACM Workshop on Online Social Networks*.
- [47] Feng Wang and Xuejin Zuo. 1999. Inside China's cities: Institutional barriers and opportunities for urban migrants. *The American Economic Review* 89, 2 (1999), 276–280.
- [48] Mary C Waters and Toms R Jimnez. 2005. Assessing immigrant assimilation: New empirical and theoretical challenges. *Annu. Rev. Sociol.* 31 (2005), 105–125.
- [49] Carolyn Whitzman. 2006. At the intersection of invisibilities: Canadian women, homelessness and health outside the 'big city'. *Gender, Place & Culture* 13, 4 (2006), 383–399.
- [50] Weiping Wu. 2004. Sources of migrant housing disadvantage in urban China. *Environment and Planning A* 36, 7 (2004), 1285–1304.
- [51] Xiaogang Wu and Donald J Treiman. 2004. The household registration system and social stratification in China: 1955–1996. *Demography* 41, 2 (2004), 363–384.
- [52] Xiaogang Wu and Donald J Treiman. 2007. Inequality and equality under Chinese socialism: The hukou system and intergenerational occupational mobility. *American Journal of Sociology* 113, 2 (2007), 415–445.
- [53] Zhongshan Yue, Shuzhuo Li, Xiaoyi Jin, and Marcus W Feldman. 2013. The role of social networks in the integration of Chinese rural–urban migrants: A migrant–resident tie perspective. *Urban Studies* 50, 9 (2013), 1704–1723.

- [54] Kevin Honglin Zhang and Shunfeng Song. 2003. Rural-urban migration and urbanization in China: Evidence from time-series and cross-section analyses. *China Economic Review* 14, 4 (2003), 386–400.
- [55] Yu Zheng, Licia Capra, Ouri Wolfson, and Hai Yang. 2014. Urban Computing: Concepts, Methodologies, and Applications. *ACM Transactions of Intelligent System and Technology* 5, 3 (2014), 38:1–38:55.
- [56] Yu Zheng, Yanchi Liu, Jing Yuan, and Xing Xie. 2011. Urban computing with taxicabs. In *Proceedings of UbiComp*.