To Stay or to Leave: Churn Prediction for Urban Migrants in the Initial Period

Yang Yang
Zhejiang University
yangya@zju.edu.cn

Zongtao Liu
Zhejiang University
tomstream@zju.edu.cn

Chenhao Tan
University of Colorado Boulder
chenhao@chenhaot.com

Fei Wu
Zhejiang University
wufei@zju.edu.cn

Yueting Zhuang
Zhejiang University
yzhuang@zju.edu.cn

Yafeng Li
China Telecom.
liyafeng@chinatelecom.cn

ABSTRACT

In China, 260 million people migrate to cities to realize their urban dreams. Despite that these migrants play an important role in the rapid urbanization process, many of them fail to settle down and eventually leave the city. The integration process of migrants thus raises an important issue for scholars and policymakers.

In this paper, we use Shanghai as an example to investigate migrants' behavior in their first weeks and in particular, how their behavior relates to early departure. Our dataset consists of a one-month complete dataset of 698 telecommunication logs between 54 million users, plus a novel and publicly available housing price data for 18K real estates in Shanghai. We find that migrants who end up leaving early tend to neither develop diverse connections in their first weeks nor move around the city. Their active areas also have higher housing prices than that of staying migrants. We formulate a churn prediction problem to determine whether a migrant is going to leave based on her behavior in the first few days. The prediction performance improves as we include data from more days. Interestingly, when using the same features, the classifier trained from only the first few days is already as good as the classifier trained using full data, suggesting that the performance difference mainly lies in the difference between features.

CCS CONCEPTS

• **Applied computing** → **Law, social and behavioral sciences**; • **Information systems** → **Social networks**; *Data mining*; • **Computing methodologies** → **Artificial intelligence**;

KEYWORDS

urban migrants, migrant integration, churn prediction

ACM Reference Format:
Yang Yang, Zongtao Liu, Chenhao Tan, Fei Wu, Yueting Zhuang, and Yafeng Li. 2018. To Stay or to Leave: Churn Prediction for Urban Migrants in the Initial Period. In *WWW 2018: The 2018 Web Conference, April 23–27, 2018, Lyon, France.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3178876.3186144

WWW 2018, April 23–27, 2018, Lyon, France

ACM ISBN 978-1-4503-5639-8/18/04.
https://doi.org/10.1145/3178876.3186144

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 [29]. 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, 29]. 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 [40]; their children may be excluded from schools [21]. 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 the integration of migrants into urban society.

In this paper, we focus on the initial period of a migrant's integration process because a migrant's first steps are important for her eventual integration. Despite the importance of the initial period [6, 10, 37], existing studies, which mostly rely on survey data, rarely have fine-grained data to examine this period. We take Shanghai as an example to investigate two aspects based on telecommunication data: how they develop their initial personal networks and how they move around the city. In particular, our dataset allows us to explore why some migrants decided to leave early. This problem of whether to stay in a new city resembles studies on whether users will stay in online communities, also known as churn prediction [2, 13, 14, 28, 34, 47], but presents more complex dynamics because moving offline requires considerable amount of efforts.

Organization and highlights of this work. We present a large-scale quantitative exploration on the first weeks after migrants arrive in Shanghai, one of the biggest cities in China. We employ a Shanghai *one-month complete* telecommunication metadata provided by China Telecom, the third largest mobile service providers in China. Our dataset consists of around 698 million call logs between 54 million users. In addition, we collect housing prices of over 20K real estates in Shanghai to study the role of housing price in migrant integration. The details of our datasets are introduced in Section 2. We are able to identity whether a user of China Telecom

is a migrant because 1) it is necessary for a migrant to apply for a local phone number in Shanghai due to long-distance costs; 2) it is uncommon for a temporary visitor to apply for a local phone number due to the burdensome application process; 3) applying for phone numbers require personal identification, which contains birthplace information. We use *locals*, who were born in Shanghai, as a comparison point to understand the integration process. We also differentiate *leaving migrants*, new migrants that left the city in three weeks after they moved to Shanghai, from *staying migrants*, new migrants that managed to stay for the first month. Our results indicate that around 4% of new migrants ended up leaving early. This work builds on our previous work [55], which employs the same telecommunication dataset to explore different characteristics of new migrants, settled migrants, and locals. The key differences lie in the churn prediction task to predict the early departure of new migrants and the novel combination of housing price information and telecommunication metadata.

We first explore how locals, leaving migrants, and staying migrants differ in their mobile communication networks and geographical locations in Section 3. The dynamic patterns over time allow us to study the integration process in the first month. Overall, locals are stable in all our proposed features during this period, whereas both staying migrants and leaving migrants go through significant changes despite the short time span. However, the changes of staying migrants and leaving migrants may happen to different extents, sometimes even in different directions. Specifically, we find that it is important for a new migrant to develop *diverse* social ties in the first few weeks. For instance, a migrant with contacts from more provinces tends to stay. As for geographical locations, leaving migrants are less active in terms of geographical movements, and move around more expensive housing areas than both locals and staying migrants, whereas staying migrants move around cheaper areas over time. This observation suggests that it is important for new migrants to find their own active area in a big city. Staying migrants are still different from locals in the last week of our dataset, which suggests that one month is far from enough for a migrant to integrate into a new city.

We then study to what extent we can separate the three groups by formulating prediction tasks in Section 4. Because of the class imbalance, we investigate two prediction tasks: distinguishing new migrants from locals and distinguishing leaving migrants from staying migrants. Our proposed features are effective in both tasks and clearly outperform random guessing. Random forest is the best performing classifier, indicating the importance of non-linearity. We focus our analysis on the second task as it holds promise for delivering personalized service to new migrants who find the integration process difficult. The prediction performance improves as we include features that span more days. Such a performance improvement mainly comes from better feature quality as we use data from more days, because the classifier trained from only the first few days can perform as well as the classifier trained using full data by using the same set of features when testing.

Our work aims to understand the (dis)integration of migrants. This challenging problem necessarily involves efforts from a wide range of disciplines, including anthropology, economics, and sociology. We thus provide an overview of related work in Section 5 and offer some concluding discussions in Section 6.

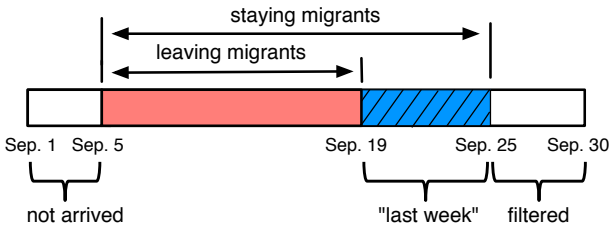


Figure 1: An illustration of how we define leaving migrants and staying migrants. The first several days are used to filter out new migrants. Since the last few days overlap with national holidays, we use Sep. 19 — Sep. 25 as the last week to make sure that leaving migrants left early instead of traveling temporarily.

2 EXPERIMENTAL SETUP

In this paper, our main dataset is the *complete* telecommunication records between mobile users using China Telecom in Shanghai over one month in 2016. Before introducing our dataset and experimental setup, we highlight several facts about telecommunication in China. First, obtaining a local number is the first integration step for a new migrant because of long-distance call costs, and we are able to differentiate whether a telephone number is a local number in Shanghai or from other regions. Second, since obtaining a phone number is nontrivial and requires personal identification, it is uncommon for a temporary visitor to obtain a local number. Personal identification allows us to extract the birthplace of a person. Therefore, we can identify people who just obtained a local number but were not from Shanghai originally. It is worth noting that long-distance call costs have been removed by China Telecom in September 2017, which makes our telecommunication metadata unique and valuable to understanding migrant integration. Insights from our work can nevertheless be used to analyze telecommunication patterns without personal identification information.

2.1 How Many Migrants are Leaving in the First Weeks?

Telecommunication dataset. Our telecommunication dataset is provided by China Telecom, the third largest mobile service provider in China. The dataset spans a month from September 1st, 2016 to September 30th, 2016. It includes over 698 million call logs between around 54 million users. For each user, we obtain her demographical information including age, sex, and birthplace by the personal identification binded with the phone number. Each entry in the call logs contains the caller's number, the callee's number, the starting time, and the ending time. In addition, for each call, we have the GPS location of the corresponding telecommunication tower, which is widely used to approximate user locations during the call. Our dataset was anonymized by China Telecom to protect user privacy. Throughout the paper, we report only average statistics without revealing any identifiable information of individuals.

Locals, staying migrants, and leaving migrants. We only consider users with local phone numbers in this work.[1] We categorize

[1]We merge numbers corresponding to the same user ID into one to account for users with multiple numbers. We also filtered around 15, 000 users that have abnormally

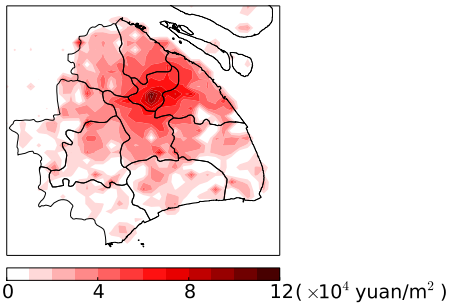


Figure 2: Housing price distribution over Shanghai.

users in our dataset into three groups based on their birthplaces and call history. We refer to people that were born in Shanghai as *locals*. We consider people that were not born in Shanghai and had no call logs in the first 4 days in our dataset as *new migrants*. Our focus in this paper is to understand the behavioral pattern of new migrants, which shed light on the integration process of migrants.

Our first question is how many new migrants are leaving in the first weeks, despite that they made efforts to obtain a local number. We identify new migrants that ended up leaving Shanghai early, i.e., before *the last week* in our dataset. To make sure that people did not leave temporarily, we omit the last 5 days' data for all users as the National Day holidays were close to that time, which may lead to temporal travel. That is, the last week in our dataset is defined as Sep. 19 - Sep. 25. We consider new migrants as *leaving migrants* if they were active in the first two weeks (Sep. 5 - Sep. 18) and have no record since Sep. 19, and as *staying migrants* if they were active in all the three weeks. Figure 1 gives an illustration.

Based on our definition, we identify *1.8M locals, 34K staying migrants, and 1.5K leaving migrants*. It follows that around ***4%*** of the new migrants left Shanghai in the first few weeks, which is a useful statistic for urban policymakers and complements existing survey based approaches. To the best of our knowledge, there is no public official report about this statistic. This categorization of users into locals, staying migrants, and leaving migrants constitutes the basis for our computational framework.

2.2 Housing Price Data

Economic theory suggests that individual migration depends on housing prices in different areas [45]. To validate and further study this in our dataset, we employ a housing price data from AnJuKe[2], an online platform for real estate sales and renting. Our data covers around 18K real estates at Shanghai in 2017[3]. Together with the GPS locations, we calculate the average housing price for a particular user's home, work place, and other active areas. Overall, the housing price in Shanghai spreads a wide range (Figure 2). For instance, in HuangPu area, the center of Shanghai, the average housing price has exceeded 100K yuan (∼15K US dollars) per square meter. Meanwhile, other areas like MinHang has the housing price below 30K yuan per square meter. The average housing price in our dataset is 54.3K, with a standard deviation of 29.4K.

2.3 Computational Framework

From a person's call logs, we can extract a mobile communication network, which can reasonably approximate a user's social network and how she develops her connections to others after moving to a new city. We can also obtain the geographical locations of users from our data, which is also valuable to understanding migrants' active areas in a new city. We then formulate the following notations, which are consistent with our previous work [55].

Mobile communication networks. Based on call logs, we establish a mobile communication network grouped by time periods. Formally, we build a directed graph $G_t = (V_t, E_t)$ for a time period t, where V_t is the set of users, and each directed edge $e_{ij} \in E$ indicates that v_i calls v_j during that time period ($v_i, v_j \in V_t$). Here t can refer to a week or several days. Note that only a subset of users in V_t are labeled as locals, staying migrants, or leaving migrants.

Geographical locations. For each call a person makes, we have access to the GPS location from the corresponding telecommunication tower. We then group a person's locations by time periods. We collect all the locations that a person makes calls in a time period t, and refer to this ordered list of locations for user v as $L_v^t = [l_1, \ldots, l_n]$, 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.

3 THE (DIS)INTEGRATION OF MIGRANTS

In this section, we study the integration process of migrants in the first weeks after moving to a new city, and the disintegration process for some migrants who left early. To do that, we examine a wide range of factors from people's mobile communication networks and geographical locations. We propose four types of features: ego network properties, call behaviors, geographical patterns, and housing price information. In order to understand the integration process, we use locals as a comparison point. Therefore, we examine the differences between locals, staying migrants, and leaving migrants in each week and how the features evolve over time. Here we focus on explaining the motivation and evolving patterns of each feature by itself, and will examine their prediction performance in Section 4. Please refer to Table 4 in the appendix for computational details of each feature. Note that we do not have feature values for leaving migrants at week 3 because they left in the third week.

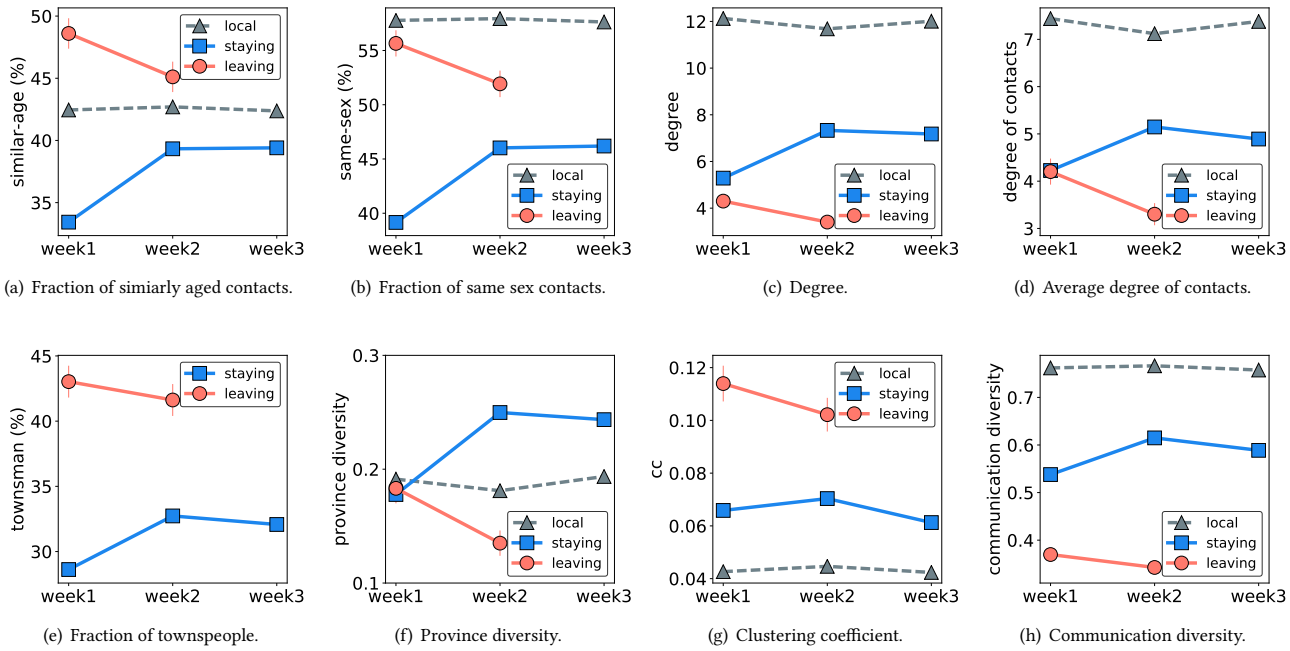
3.1 Ego network properties.

We first study how individuals build new connections and maintaining existing relationships in the first weeks after moving to a new city. We extract features based on a person's ego network, the subgraph consisting of a person and all her neighbors [19]. We treat the telecommunication network as an undirected graph when building ego networks, so each neighbor of a person v either called v or received v's calls. As Figure 4 illustrates, we study characteristics of v and her friends such as demographics, birthplaces, and connections to other people. Figure 3 presents the results. Overall,

high degrees, who likely correspond to fraudsters, delivery persons, or customer services according to a user type list provided by China Telecom.

[2]https://shanghai.anjuke.com/

[3]This data is publicly available: http://yangy.org/data.html



(a) Fraction of simiarly aged contacts. (b) Fraction of same sex contacts. (c) Degree. (d) Average degree of contacts.

(e) Fraction of townspeople. (f) Province diversity. (g) Clustering coefficient. (h) Communication diversity.

Figure 3: How locals, staying migrants, and leaving migrants build social connections in the first three weeks. y-axis represents feature values based of an individual's ego-network and x-axis represents time.

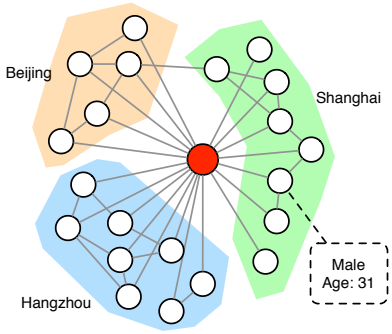


Figure 4: Example of an ego-network centered by a particular user v. We study how the structure of its ego-network evolve over time, and how the evolution pattern of a new migrant is correlated with her decision to settle down or leave the city.

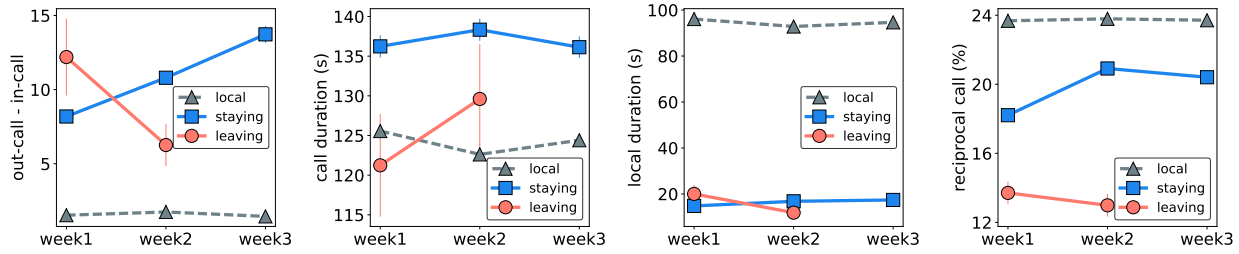
we find that the ego network features of locals are more stable than those of new migrants over time.

Demographics. Social homophily suggests that people tend to develop connections to those who are similar to themselves [35]. As Figure 3(a) shows, locals have a stable fraction of similarly-aged friends (around 0.41), while leaving migrants have a larger fraction and staying migrants have a smaller fraction. We see the results of leaving migrants and staying migrants are both getting closer to locals, which indicates the integration process of migrants. This process seems to happen rather quickly in terms of talking to similar-aged friends. As for sex (Figure 3(b)), locals show the strongest homophily in sex, i.e., locals have the largest fraction of contacts with the same sex. In comparison, new migrants who have a strong homophily in sex during the first week tend to be leaving migrants, while new migrants with more friends of the different sex tend to stay in Shanghai.

Degree. A people's degree reflects the number of contacts she has (Figure 3(c)(d)). As expected, locals and their contacts have the largest degree. In the first week, staying migrants and leaving migrants, and even their friends have very similar numbers of contacts. However, staying migrants develop significantly more connections than leaving migrants in the second week. Such better connectedness also applies to the contacts that staying migrants make in the second week.

Diversity of connections. We finally examine the diversity of one's connections from three aspects: the birthplace of contacts, clustering coefficient, and communication diversity over contacts.

We analyze the birthplace of contacts both in the fraction of people who came from the same hometown and the diversity across different provinces. Leaving migrants rely on people from the same hometown much more, while staying migrants start with a lower fraction of townspeople and grow a little bit over the integration (Figure 3(e)). Remember that locals were born in Shanghai, so they are not shown in this figure. To further study this, we define a person v's province diversity as the entropy of the distribution of birth provinces among v's contacts, i.e., $-\sum_x p_x \log_2 p_x$, where p_x is the probability that a contact of v was born in province x. Figure 3(f) again shows that locals are pretty stable over time, while staying migrants have the most diverse group of contacts and



(a) Difference between out-calls and in-calls.

(b) Average duration of calls.

(c) Average duration of calls made to locals.

(d) Fraction of reciprocal calls.

Figure 5: Call behavior of locals, staying migrants, and leaving migrants.

leaving migrants have the lowest. This suggests that contacting people from different regions may help integrate in a new city.

Clustering coefficient measures the fraction of triangles in the ego-network and indicates how likely a person's contacts know each other. From Figure 3(g), we see that leaving migrants present the largest clustering coefficient, while locals have the lowest. This suggests that new migrants that start with a close-knit group after moving to a big city may hinder themselves from integration.

Finally, inspired by the social diversity proposed in Eagle et al. [17], we define the communication diversity as a function of Shannon entropy to quantify how a person split the number of calls to her friends, i.e., $\frac{-\sum_j p_{ij} \log(p_{ij})}{log(k_i)}$. Here k_i is the out-degree and p_{ij} is the probability defined as $p_{ij} = \frac{n_{ij}}{\sum_l n_{il}}$, where n_{ij} is the number of calls user v_i makes to user v_j. The result in Figure 3(h), again, suggests that developing more diverse connections may help new migrants integrate into a big city like Shanghai.

3.2 Call Behavior

For users' call behaviors, we first examine the difference between a person's outgoing calls and incoming calls in Figure 5(a). The positive values suggest that the three groups of people in our dataset are more likely to make phone calls than to receive calls. We also see that the difference is larger for new migrants than locals from the figure. Making more out-calls may be a sign that new migrants are developing initial connections. Staying migrants make more out-calls as time grows, whereas leaving migrants make fewer. Note that this feature of staying migrants grows more dissimilar to locals, suggesting that two weeks is too short for staying migrants to integrate into locals.

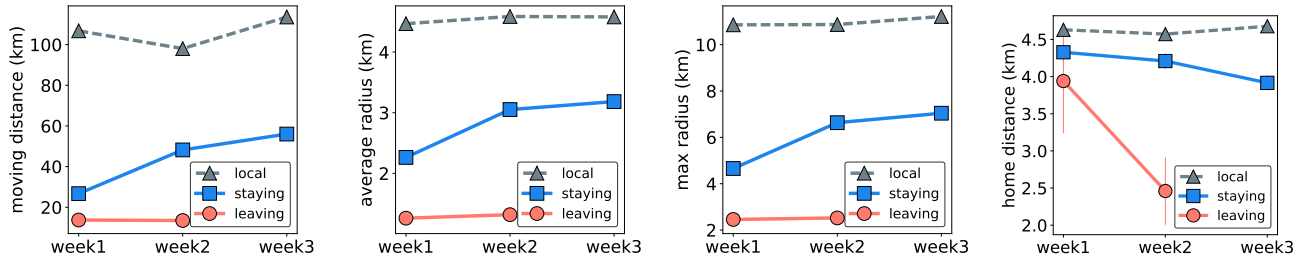
The duration of calls may reflect the strength of a relation between two person. Naturally, closed friends tend to have longer phone calls, while strangers are more likely to have a quick check-in. Figure 5(b) shows that staying migrants make much longer calls than locals and leaving migrants. One possible explanation is that staying migrants need more time to figure out their initial life in a new city. Lack of such relations, leaving migrants fail to integrate into Shanghai. Meanwhile, locals, who have stable relations in Shanghai, are less likely to make long calls. However, as Figure 5(c) shows, locals make significant longer calls to other locals, while new migrants do not develop strong relations with locals in the first three weeks after they arrive.

Finally, we investigate the fraction of reciprocal calls (i.e., two-way relationships between users). As Figure 5(d) shows, locals are more likely to have reciprocal relationships with their contacts, while the fraction of reciprocal calls is lower for staying migrants, and is lowest for leaving migrants. This again shows that the personal networks of new migrants are still nascent. At week 2, staying migrants have a larger likelihood to have reciprocal relationships, whereas leaving migrants' likelihood decreases.

3.3 Geographical Patterns

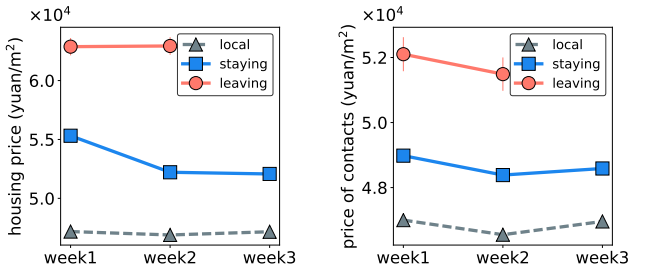
We use locations to measure the mobility of migrants. Given a user v's geographical locations $L_v^t = \{l_1, \cdots, l_n\}$, which is ordered by time and contains the latitude and the longitude of locations for user v during a time period t, we can measure that user's active area from three different aspects. First, we measure the *total distance* that the user moves as $\sum_i |l_i - l_{i-1}|$ (Figure 6(a)). Second, we compute a user v's radius based on her center of mass l_{CM} as $\frac{\sum_{l \in L_v^t} l}{|L_v^t|}$. We define *average radius* as the average distance of v from her center of mass, $\frac{\sum_{l \in L_v^t} |l - l_{CM}|}{|L_v^t|}$ (Figure 6(b)). Similarly, we define *max radius* as the max distance of v from her center of mass, i.e., $\max_{l \in L_v^t} |l - l_{CM}|$ (Figure 6(c)). The results using these three statistics are consistent: locals tend to move the most distance between calls on average and have much larger active area than new migrants. In comparison, staying migrants tend to expand their active areas, whereas leaving migrants' moving distance, average radius, and max radius only change slightly over time.

By assuming that most people work in day time and go back home at night, we can define a person v's workplace as her center of mass during 9:00am to 16:00pm and her home as the center of mass during 20:00pm to 7:00am. We study the distances between a person's home and her workplace in Figure 6(d). Locals have a stable distance over time and the small fluctuation can be explained by people's activity during the day and the night. New migrants live slightly closer to workplace in the first week, while the distance becomes smaller over time, suggesting that new migrants may find new places to live after they obtain a job. This decrease in distance between the day and the night can further alleviate the concern that these new migrants are temporary visitors.



(a) Moving distance. (b) Average radius. (c) Max radius. (d) Distance between home and work palce.

Figure 6: Geographical features of locals, staying migrants, and leaving migrants (in kilometers).



(a) Average housing price of users' active areas. (b) Average housing price of friends' active areas.

Figure 7: Housing price features of locals, staying migrants, and leaving migrants.

Feature sets	Precision	Recall	F1
all features	0.2355	0.8397	**0.3678**
ego network properties	0.2097	0.8499	0.3363
call behavior	0.1021	0.8358	0.1820
geographical patterns	0.0813	0.5971	0.1433
housing price information	0.0641	0.5347	0.1144
random guess	0.0198	0.0198	0.0198

Table 1: Distinguishing new migrants from locals using random forest with different set of features.

3.4 Housing Price Information

The soaring housing price has been a central issue in the urbanization process of China [12, 51]. Housing price can play an important role in a migrant's integration process. Figure 2 presents the overall distribution of housing price. We compute the average housing price of a person's active geographical locations as well as of a person's friends' active locations.

Surprisingly, we find that locals tend to be active in the least expensive areas, while leaving migrants stay in the most expensive places (Figure 7(a)). The average housing price of staying migrants drops significantly from the first week to the second week, but it is not the case for leaving migrants. Similar results can be observed in the average housing price of a person's home. We omit the figure for space reasons. A new migrant may leave Shanghai early because she fails to find a place with a reasonable renting price. Due to social homophily, the average housing prices of one's active areas is similar to that of their friends (Figure 7(b)).

Summary. Through the evolving patterns of our proposed features, we find that staying migrants are able to move towards locals in many dimensions, but three weeks is of course too short to finish the integration process. However, comparing staying migrants to leaving migrants, staying migrants have more active and diverse mobile communication contacts and geographical movements. This suggests that actively expanding one's social network after moving to a new city is an important step for migrant integration.

4 PREDICTING (LEAVING) MIGRANTS

Having established the dynamic patterns of single features, we explore to what extent locals, staying migrants, and leaving migrants are separable based on our proposed features. As these three groups of people have very different population sizes, we set up two prediction tasks. We first propose a binary classification task to predict if an individual is a local or a new migrant, and then work on distinguishing leaving migrants from staying migrants. Both tasks are challenging due to the sparsity of data: less than 2% people are new migrants and 4% (1.5K) migrants left early in our dataset. The second task is more difficult as the behavior patterns of leaving migrants and staying migrants are more similar than those of new migrants and locals. However, accurate prediction of leaving migrants may allow for personalized service to help integration and insights from the second task can potentially inform urban policymakers, so we focus on the second task. For both tasks, we use the same features listed in Table 4 in the appendix.

4.1 Distinguishing New Migrants from Locals

Our first binary classification task is to distinguish new migrants from locals. Formally, given a user v, a set of v's mobile communication networks $\{G_t\}$ over the first 14 days since she moves to Shanghai ($1 \le t \le 14$), and the geographical locations L_v^t of v at time t (the t-th day), our goal is to predict if v is a new migrant

Classification method	Precision	Recall	F1
Random forest	0.1597	0.6659	0.2576
Multilayer perceptron	0.1329	0.5533	0.2140
Support vector machine	0.1238	0.6815	0.2095
Logistic regression	0.1006	0.7082	0.1762
random guess	0.0437	0.0426	0.0431

Table 2: Performance in distinguishing leaving migrants from staying migrants with different classifiers. Each classifier uses all features extracted from the first $k = 14$ days.

Feature sets	Precision	Recall	F1
all features	0.1597	0.6659	0.2576
ego network properties	0.1347	0.6580	0.2234
housing price information	0.1067	0.5978	0.1809
call behavior	0.0984	0.5853	0.1683
geographical information	0.0863	0.5691	0.1498

Table 3: Distinguishing leaving migrants from staying migrants using random forest with different feature sets extracted from the first $k = 14$ days.

or a local. We conduct 5-fold cross-validation and use precision, recall, and F1-score for evaluation, with the minority class, i.e., new migrants, as the target class.

Table 1 demonstrate the results of random forest in this task. Recall that in our dataset, there are 1.8 million locals and 35.5 thousand new migrants (52:1). Thus random guessing would obtain an F1-score around 0.02. Our method is able to outperform this random baseline significantly with an F1 of 0.36. We further compare the effectiveness of different feature sets by training a classifier that includes a single feature set and excludes the other feature sets. Table 1 shows that using every singe feature set outperforms random guessing. Ego network properties performs the best, followed by geographical patterns.

4.2 Predicting Leaving Migrants

In the second task, we aim to separate staying migrants from leaving migrants, i.e., predict if a new migrant will leave Shanghai within the 3rd week. We again conduct 5-fold cross-validation and use precision, recall, and F1-score for evaluation, with leaving migrants as the target class.

Overall performance. We experiment with different classifiers including logistic regression, support vector machine, multilayer perceptron, and random forest. Recall that we have 34K staying migrants and 1.5K leaving migrants in our dataset. Random guessing would thus obtain an F1-score around 0.04. Table 2 shows that all the machine learning classifiers clearly outperform random guessing and random forest provides the best performance. Random forest and MLP outperform others in terms of F1, suggesting that non-linearity is important for this classification task.

Overall, the prediction performance suggests that the proposed features are not only effective in distinguishing new migrants from locals, but also useful in predicting a migrant's decision to stay or leave. However, the F1-score in this task is not as good as in the first task. The performance drop suggests that as expected, predicting a migrant's leaving decision is much harder than distinguishing new migrants from locals.

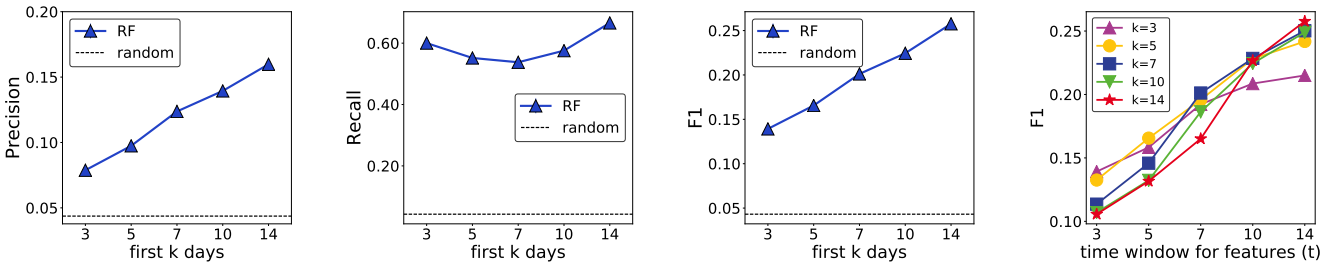
Table 3 also lists the performance of the classifier using a single feature set with random forest. Again, ego network properties perform the best. Geographical patterns, however, perform worse in this task than in the task of distinguishing new migrants from locals. In comparison, housing price features achieve a better F1-score than geographical patterns, which suggests that knowing meta information such as housing price of a person's active areas is more useful than simply knowing the active areas for predicting migrants' early departure.

Early detection of leaving migrants. We next explore whether it is possible to detect leaving migrants sooner than two weeks. If we can detect departure early on, we may be able to provide integration service. To do that, we extract features based on a person's information from the first k days. Figure 8 shows that precision and F1 score follow a very similar trend: knowing a person's behavior for longer after she moves to a new city (as k increases) allows us to better predict her decision to leave or stay. Since leaving migrants only take a very small fraction of our data, recall is relatively stable over different ks, which is already around 0.6 when $k = 3$, whereas improvement in precision is the main reason for overall performance improvement. Even when observing only 3 days, the classifier can outperform random guessing.

Why does the performance improve? To understand why the performance improves as we observe new migrants for longer, we propose a novel set of experiments. We attempt to disentangle the improvement due to feature quality or classifier quality by replacing the features with future information when applying the classifier trained from only a small number of days. Specifically, we first train a classifier with data in the first k days, and then use features extracted from the first t days to predict if the user will leave the city within the 3rd week after she arrives in Shanghai. We vary different k and t to see how they influence the performance.

Surprisingly, we see that classifiers trained with only the first 5 days' data perform as well as those trained using 14 days when testing with features extracted from the first 10 or 14 days in Figure 8(d). This result indicates that the classifier can be well trained using data from a small number of days and the performance improvement is mainly due to improved feature quality. In other words, as new migrants stay longer, we have more reliable information regarding how she behaves, but even with less reliable information from the first 5 days, we can already know how different features relate to leaving migrants.

Feature importance We finally discuss the important features in the learned random forests. For each feature, we calculate its Gini importance (also known as mean decrease impurity) in the learned random forest model with the first 14 days' data. Figure 9 lists the most important 10 features. Degree (#contacts) ranks the first, while in-calls (#incoming calls), out-degree (#contacts making outgoing calls), and CC (clustering coefficient) are both in top 10. It, again, suggests that expanding more (diverse) connections is critical for new migrants and ego network properties are useful features. Two of the top 3 features are relevant to housing price,



(a) Precision. (b) Recall. (c) F1. (d) Disentangling performance improvement.

Figure 8: The left three figures present prediction performance in distinguishing leaving migrants from staying migrants by employing features extracted from the first k days since new migrants moved to Shanghai. x-axis represents the number of days that we extract features from and train classifiers on, y-axis represents the evaluation metric. Figure 8(d) shows that the performance improvement mainly comes from improved feature quality. x-axis represents the number of days (t) that we extract features from when testing, while different lines track the performance of the classifier trained using k days. The classifier with small k (i.e., ≥ 5) shows similar performance with $k = 15$ as long as we use features from 14 days when testing.

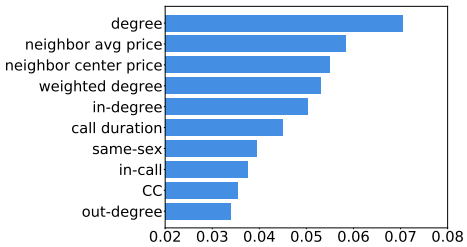


Figure 9: The most important 10 features. The x-axis denotes the relative Gini importance of features.

which is consistent with our finding in Section 3 that living in an area with reasonable pricing is important.

5 RELATED WORK

The urbanization process poses significant challenges for the society that require efforts from various disciplines. We summarize 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, 18, 20, 21, 42–44, 55]. For example, our previous work explores different characteristics of locals, settled migrants, and new migrants [55]. In addition to the effect of demographics (ethnic groups, rural vs. urban) on urban migrant integration, Schiller and Çauglar [42] 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 [44]. Beyond our scope, immigrants (migrants to a new country) and refugees (a subgroup of immigrants) have also received significant interests [4, 5, 24, 46, 50].

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, 11, 26, 32, 33, 49, 52–54, 56, 57]. There are at least three perspectives as suggested in [26]: 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, 53, 54]. 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, 16, 22, 25, 39, 41, 58, 59]. 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. [59] employ GPS data from taxicabs to evaluate transportation system in Beijing; De Nadai et al. [15] 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 studies in understanding the global mobility patterns [16].

Temporal social networks and online communities. Our work is also relevant to studies on the evolution of networks [23, 27, 30, 31, 36, 38, 48]. Using data from online social media, these studies explore the connection between individual behavior and global network properties. For instance, Viswanath et al. [48] 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. [30] develop a triangle-closing model to explain network evolution. In addition, studies have investigated the process of new user integration in online communities [2, 13, 14, 28, 34, 47]. In particular, McAuley and Leskovec [34] examine the process of how a new user becomes an expert on review websites.

6 CONCLUSION

In this paper, we examine the integration process of migrants in the first weeks after moving to a new city, and the disintegration process for some migrants who left early. We use Shanghai as a case study and employ a one-month complete dataset of telecommunication metadata in Shanghai with 54 million users and 698 million call logs, plus a housing price dataset for 20K real estates in Shanghai. We examine four types of features extracted from people's mobile communication networks and geographical locations: ego network properties, call behaviors, geographical patterns, and housing price information. Through our study, we find that some behavior patterns of staying migrants evolve towards locals over time (e.g., contacting similar-aged people) as new migrants integrate into a new city. The communication networks that a new migrant develops is associated with her decision to stay or leave. New migrants who manage to stay tend to develop diverse connections, move around the city in less expensive housing areas in the first weeks, compared to leaving migrants. Our proposed features are effective in both distinguishing new migrants from locals and distinguishing leaving migrants from staying migrants. It is more challenging to predict leaving migrants than to predict new migrants in general. The performance improves as we are able to observe new migrants' behavior for longer. Intriguingly, when using the same features, the classifier trained from only the first few days is already as good as the classifier trained using full data. This observation indicates that the performance improvement mainly lies in having better features.

Our work is limited by the data that we have access to. We provide the large-scale characterization of the first weeks after migrants move to a new city, but the integration process takes much longer, maybe even a life. Moreover, 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 a more in-depth understanding of the integration process. Last but not least, although China Telecom is a major service provider and Shanghai is an important global city, the selection bias 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 what difficulty a particular migrant have in fitting into the city and provide timely and useful support.

Acknowledgements. The work is supported by the Fundamental Research Funds for the Central Universities, NSFC (61702447, U1611461, 61625107), and a research funding from HIKVISION Inc.

A APPENDIX

In Table 4, we list features that we explore in Section 3 and use for the classification tasks in Section 4. We omit simple demographics features based on the user's personal attribute such as age and gender in the table for space reasons. 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.

Feature	Description
	Ego networks of user v in G_t
similar-age	The fraction of v's contacts that are at similar ages with v (±5 years).
same-sex	The fraction of v's contacts with the same sex with v.
local	The fraction of v's contacts born in Shanghai.
townsman	The fraction of v's contacts 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 4.1.
degree	The number of v's unique contacts.
in(out)-degree	The number of v's unique contacts having been called by v (called v)
neighbor degree	The average degree of v's contacts.
CC	Clustering coefficient of v's ego-network, $\frac{\|\{(s,t)\|(s,t)\in E_t\|\}}{d_v(d_v-1)}$, where s and t are v's contacts, and d_v is v's degree.
	Call behavior of user v in G_t
in(out)-call	The number of incoming (outgoing) calls.
out-call - in-call	The difference between the number of outgoing calls and incomming calls.
(local) call duration	v's average call duration (with locals).
(local) duration variance	The variance of v's call duration (with locals).
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.
communication diversity	Shannon entropy of the distribution of the number of calls to their contacts, defined as $\frac{-\sum_j p_{ij}\log(p_{ij})}{log(k_i)}$, where k_i is the out-degree, $p_{ij} = \frac{n_{ij}}{\sum_l n_{il}}$, n_{ij} is the number of calls user v_i makes to user v_j.
	Geographical features of v at time t
center	The latitude and longitude of a user v's center of mass l_{CM}, $l_{\text{CM}} = \frac{1}{\|L_v^t\|}\sum_{l\in L_v^t} l$.
workplace center	The center of user v during 9:00am to 16:00pm
home center	The center of user v during 20:00pm to 7:00am
average radius	The average distance of v from her center of mass, i.e., $\frac{1}{\|L_v^t\|}\sum_{l\in L_v^t} \|l - l_{\text{CM}}\|$.
max radius	The maximal distance of v from her center of mass, i.e., $\max_{l\in L_v^t} \|l - l_{\text{CM}}\|$.
moving distance	The total distance that v moves, $\sum_i \|l_i - l_{i-1}\|$.
average distance	The average distance that v moves, $\frac{1}{\|L_v^t\|}\sum_i \|l_i - l_{i-1}\|$.
home distance	The distance between v's workplace and home.
	Housing price features of user v
average price	The average housing price of v's active areas.
center price	The housing price of v's center of mass.
neighbor avg(center) price	The average value of the average(center) price of v's contacts.
workplace avg(center) price	The average(center) price of user v during 9:00am to 16:00pm.
home avg(center) price	The average(center) price of user v during 20:00pm to 7:00am.

Table 4: List of features.

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 Brockerhoff. 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] S Chandrasekhar, Mousumi Das, and Ajay Sharma. 2015. Short-term migration and consumption expenditure of households in rural India. *Oxford Development Studies* 43, 1 (2015), 105–122.

[11] 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.

[12] Junhua Chen, Fei Guo, and Ying Wu. 2011. One decade of urban housing reform in China: Urban housing price dynamics and the role of migration and urbanization, 1995-2005. *Habitat International* 35, 1 (2011), 1–8.

[13] 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*.

[14] 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*.

[15] 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*.

[16] 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).

[17] Nathan Eagle, Michael W Macy, and Rob Claxton. 2010. Network diversity and economic development. *Science* 328, 5981 (2010), 1029–1031.

[18] Claude S Fischer. 1982. *To dwell among friends : personal networks in town and city*. University of Chicago Press, Chicago.

[19] Linton C. Freeman. 1982. Centered graphs and the structure of ego networks. *Mathematical Social Sciences* 3, 3 (1982), 291–304.

[20] Edward L Glaeser and David C Mare. 2001. Cities and skills. *Journal of labor economics* 19, 2 (2001), 316–342.

[21] 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.

[22] 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*.

[23] 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*.

[24] 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.

[25] Shan Jiang, Gaston A Fiore, Yingxiang Yang, Joseph Ferreira Jr, Emilio Frazzoli, and Marta C GonzÃąlez. 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*.

[26] 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.

[27] Gueorgi Kossinets and Duncan J Watts. 2006. Empirical analysis of an evolving social network. *Science* 311, 5757 (2006), 88–90.

[28] Cliff Lampe and Erik Johnston. 2005. Follow the (slash) dot: effects of feedback on new members in an online community. In *Proceedings of GROUP*.

[29] June J.H. Lee. 2015. *World migration report 2015: Migrants and cities: New partnerships to manage mobility*.

[30] Jure Leskovec, Lars Backstrom, Ravi Kumar, and Andrew Tomkins. 2008. Microscopic Evolution of Social Networks. In *Proceedings of KDD*.

[31] 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.

[32] Zai Liang. 2001. The age of migration in China. *Population and development review* 27, 3 (2001), 499–524.

[33] 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.

[34] Julian John McAuley and Jure Leskovec. 2013. From amateurs to connoisseurs: Modeling the evolution of user expertise through online reviews. In *Proceedings of WWW*.

[35] 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.

[36] 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.

[37] K Bruce Newbold and Pat Deluca. 2007. Short-term residential changes to Toronto's immigrant communities: Evidence from Lsic Wave 1. *Urban Geography* 28, 7 (2007), 635–656.

[38] Ashwin Paranjape, Austin R Benson, and Jure Leskovec. 2017. Motifs in temporal networks. In *Proceedings of WSDM*.

[39] Daniele Quercia, Luca Maria Aiello, Rossano Schifanella, and Adam Davies. 2015. The digital life of walkable streets. In *Proceedings of WWW*.

[40] Shahra Razavi and Silke Staab. 2010. Underpaid and overworked: A cross-national perspective on care workers. *International Labour Review* 149, 4 (2010), 407–422.

[41] Jonathan Reades, Francesco Calabrese, Andres Sevtsuk, and Carlo Ratti. 2007. Cellular census: Explorations in urban data collection. *IEEE Pervasive Computing* 6, 3 (2007).

[42] Nina Glick Schiller and Ayse Çauglar. 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.

[43] Nina Glick Schiller and Ayse Simsek-Caglar. 2011. *Locating migration: Rescaling cities and migrants*. Cornell University Press.

[44] 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.

[45] Kim Sui So, Peter F Orazem, and Daniel Otto. 2001. The effects of housing prices, wages, and commuting time on joint residential and job location choices. *American Journal of Agricultural Economics* 83, 4 (2001), 1036–1048.

[46] Alison Strang and Alastair Ager. 2010. Refugee integration: Emerging trends and remaining agendas. *Journal of Refugee Studies* 23, 4 (2010), 589–607.

[47] Chenhao Tan and Lillian Lee. 2015. All who wander: On the prevalence and characteristics of multi-community engagement. In *Proceedings of WWW*.

[48] 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*.

[49] 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.

[50] Mary C Waters and Tomás R Jiménez. 2005. Assessing immigrant assimilation: New empirical and theoretical challenges. *Annu. Rev. Sociol.* 31 (2005), 105–125.

[51] Haizhen Wen and Yulong Tao. 2015. Polycentric urban structure and housing price in the transitional Chine: Evidence from Hangzhou. *Habitat International* 46 (2015), 138–146.

[52] Weiping Wu. 2004. Sources of migrant housing disadvantage in urban China. *Environment and Planning A* 36, 7 (2004), 1285–1304.

[53] Xiaogang Wu and Donald J Treiman. 2004. The household registration system and social stratification in China: 1955–1996. *Demography* 41, 2 (2004), 363–384.

[54] 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.

[55] Yang Yang, Chenhao Tan, Zongtao Liu, Fei Wu, and Yueting Zhuang. 2018. Urban dreams of migrants: A case study of migrant integration in Shanghai. In *Proceedings of AAAI*.

[56] 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.

[57] 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.

[58] 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.

[59] Yu Zheng, Yanchi Liu, Jing Yuan, and Xing Xie. 2011. Urban computing with taxicabs. In *Proceedings of UbiComp*.