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

Yang Yang[1], **Zongtao Liu[1]**, Chenhao Tan[2], Fei Wu[1], Yueting Zhuang[1], Yafeng Li[3]

[1]**Zhejiang University**

[2]University of Colorado Boulder

[3]China Telecom

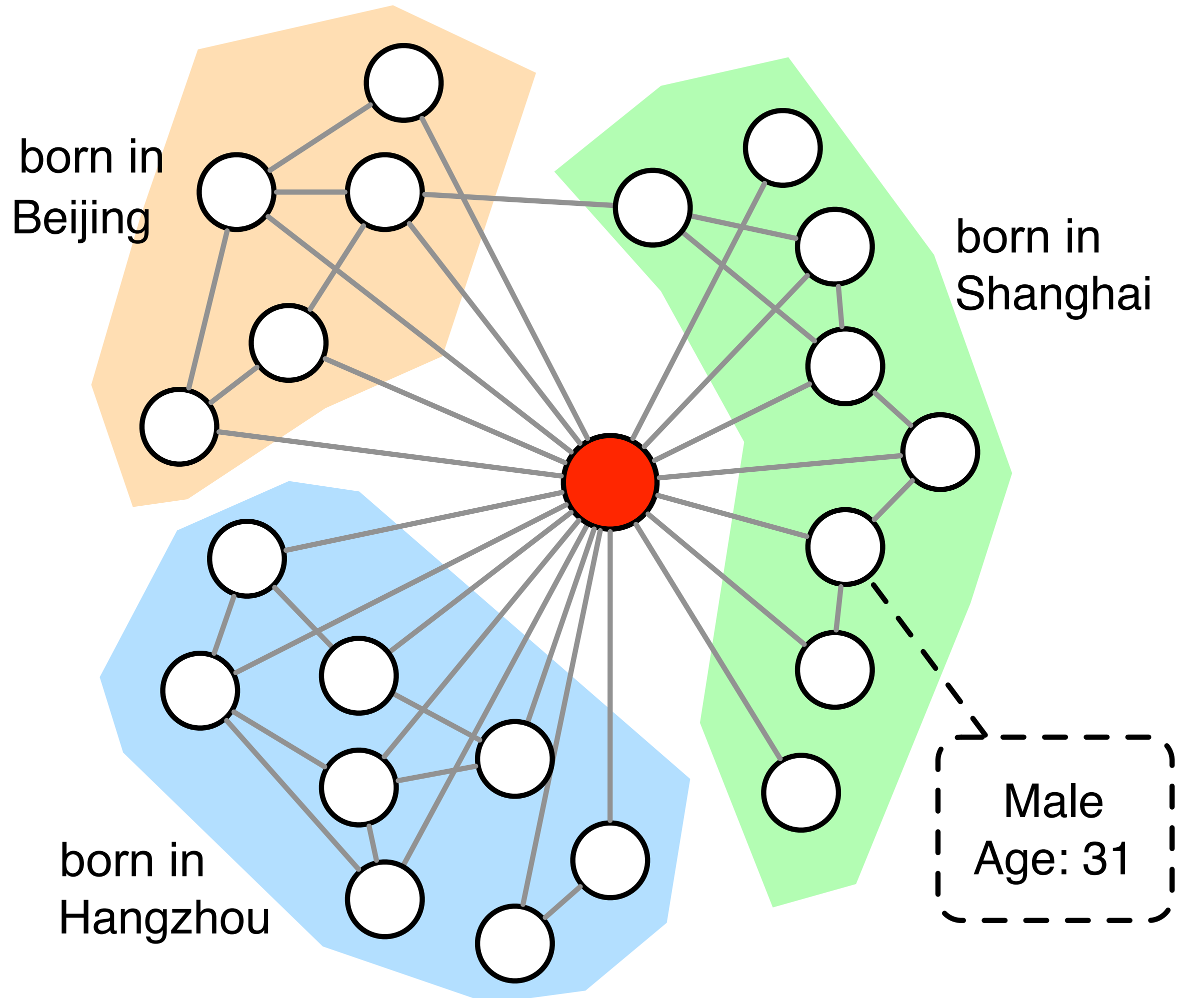
Urban Migrants

- In China, **260 million** people migrate to cities to realize their urban dreams.
- Urban migrants also pose great challenges including **segregation** and **social inequality.**
- Understanding migrant integration helps policymakers with urban planning.
- **We conduct quantitative explorations of migrant integration based on mobile communication networks.**

Telecommunication Metadata

One-month complete call data in Shanghai

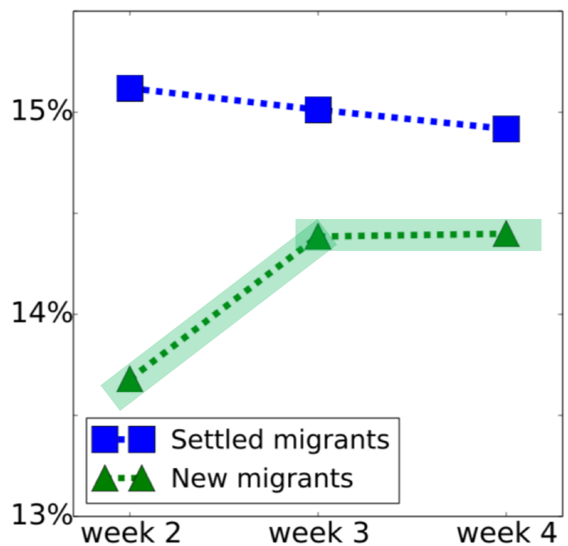
698M+ call logs and **54M+** users provided by China Telecom[1]



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

Integration and Disintegration

- Migrant Integration
  - We observe an increasing trend for new migrants misclassified as locals over the three weeks .[1]



Fraction of migrants classified as locals.

1. Yang Yang, Chenhao Tan, Zongtao Liu, Fei Wu, and Yueting Zhuang. Urban Dreams of Migrant: A Case Study of Migrant Integration in Shanghai. **AAAI'18.**

Integration and Disintegration

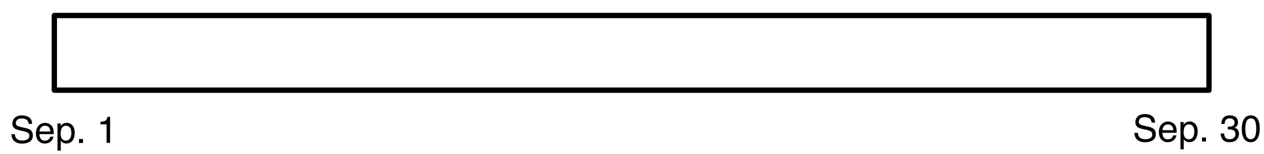
- Migrant Integration
  - We observe an increasing trend for new migrants misclassified as locals over the three weeks .[1-]
- Departure of New Migrants
  - Around 4% of new migrants ended up leaving early.
- To Stay or to leave?
  - Initial period of a migrant's integration process in Shanghai

A migrant's first step -> Eventual integration

1. Yang Yang, Chenhao Tan, Zongtao Liu, Fei Wu, and Yueting Zhuang. Urban Dreams of Migrant: A Case Study of Migrant Integration in Shanghai. **AAAI'18.**

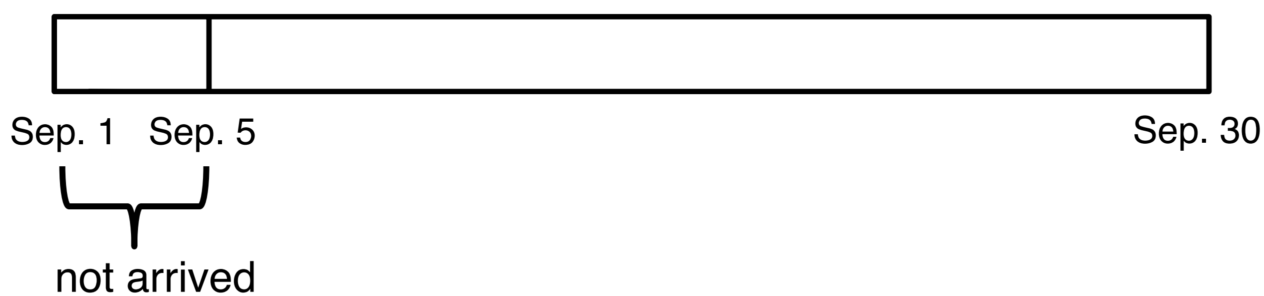
How Many Migrants are Leaving in the First Weeks?

- Based on people's birthplaces and call history, we define locals and new migrants:
  - Locals: who were born in Shanghai
  - New migrants: who were not born in Shanghai and had no call logs in the first 4 days in our dataset



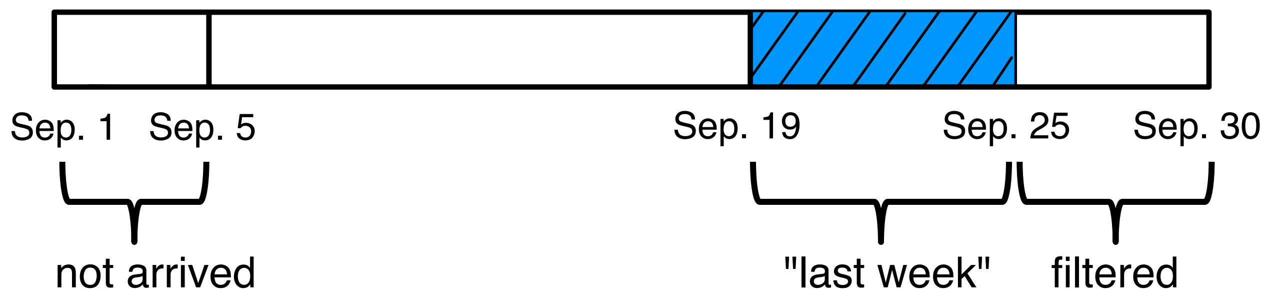
How Many Migrants are Leaving in the First Weeks?

- Based on people's birthplaces and call history, we define locals and new migrants:
  - Locals: who were born in Shanghai
  - New migrants: who were not born in Shanghai and had no call logs in the first 4 days in our dataset



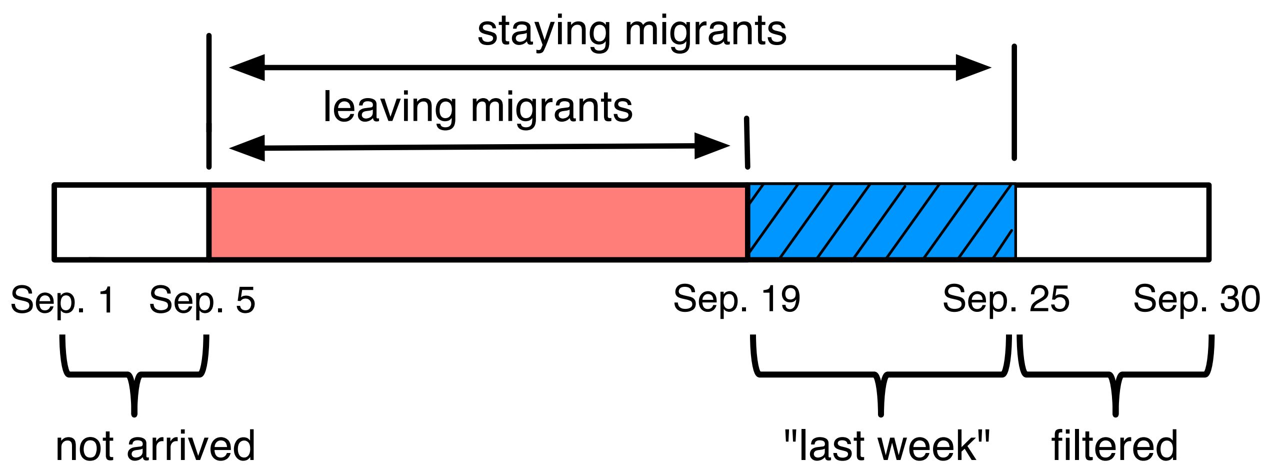
How Many Migrants are Leaving in the First Weeks?

- Based on people's birthplaces and call history, we define locals and new migrants:
  - Locals: who were born in Shanghai
  - New migrants: who were not born in Shanghai and had no call logs in the first 4 days in our dataset



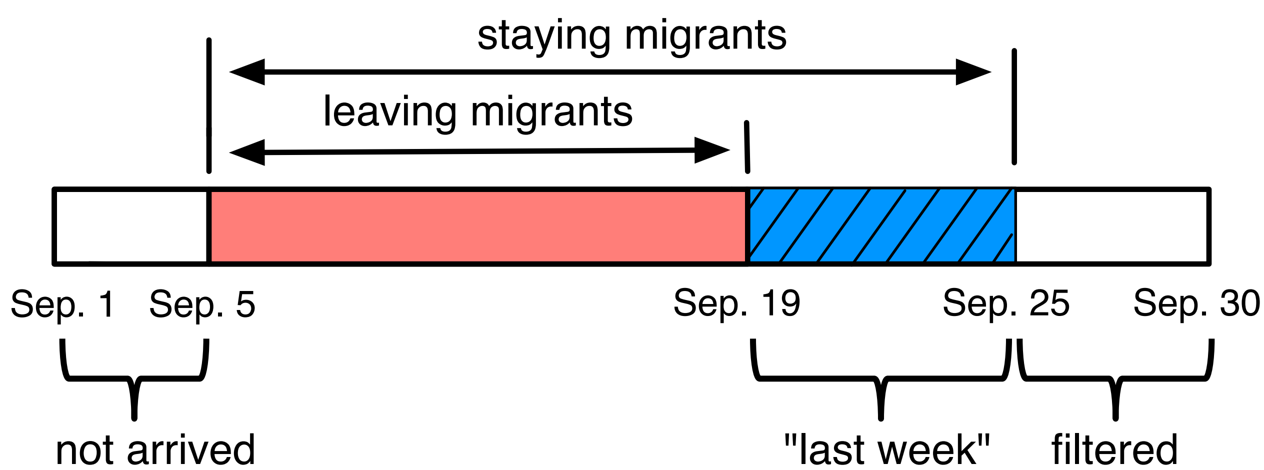
How Many Migrants are Leaving in the First Weeks?

- Based on people's birthplaces and call history, we define locals and new migrants:
  - Locals: who were born in Shanghai
  - New migrants: who were not born in Shanghai and had no call logs in the first 4 days in our dataset



How Many Migrants are Leaving in the First Weeks?

- Based on people's birthplaces and call history, we define locals and new migrants:
  - Locals: who were born in Shanghai
  - New migrants: who were not born in Shanghai and had no call logs in the first 4 days in our dataset



1.8M locals, 34K staying migrants and 1.5K leaving migrants.

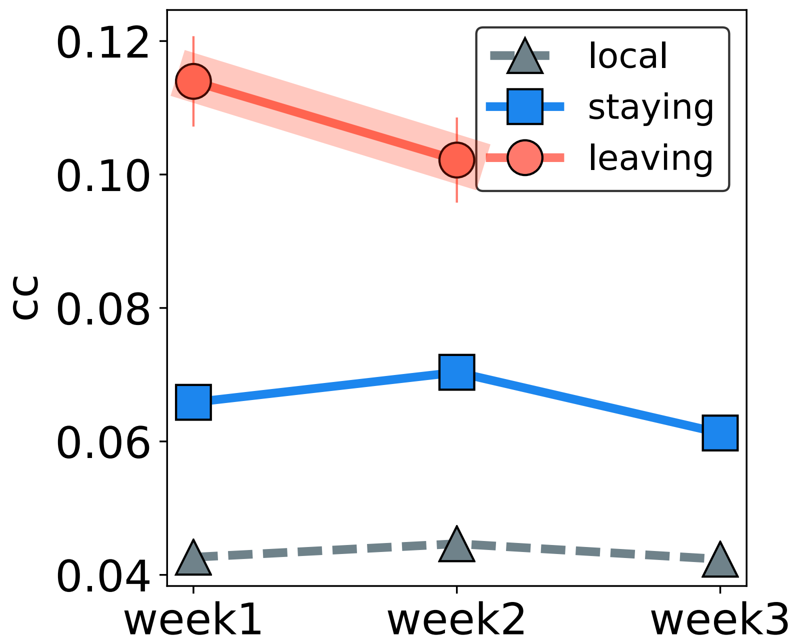
The (Dis)integretion of Migrants

- Q1: What kind of people tend to start with less dense ego networks? Leaving migrants or staying migrants?

Leaving migrants start with denser ego networks

- Q1: What kind of people tend to start with less dense ego networks? Leaving migrants or staying migrants?

clustering coefficient: the fraction of triangles in the ego-network and indicates how likely a person's contacts know each other



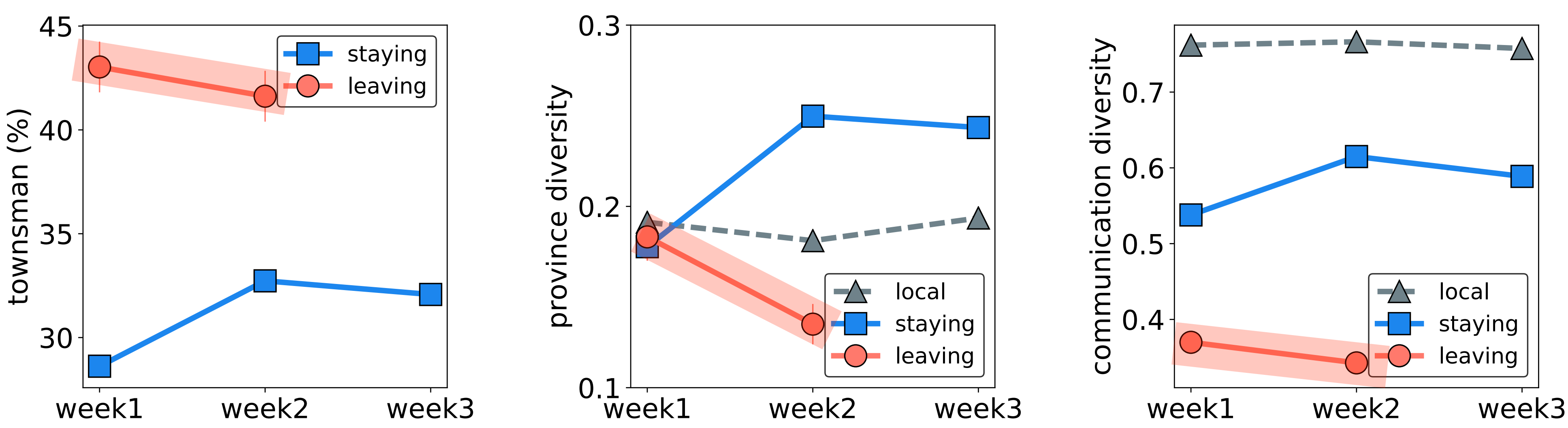
The (Dis)integretion of Migrants

- Q2: What kind of people tend to have less diverse connections? Leaving migrants or staying migrants?

Leaving migrants tend to have less diverse connections

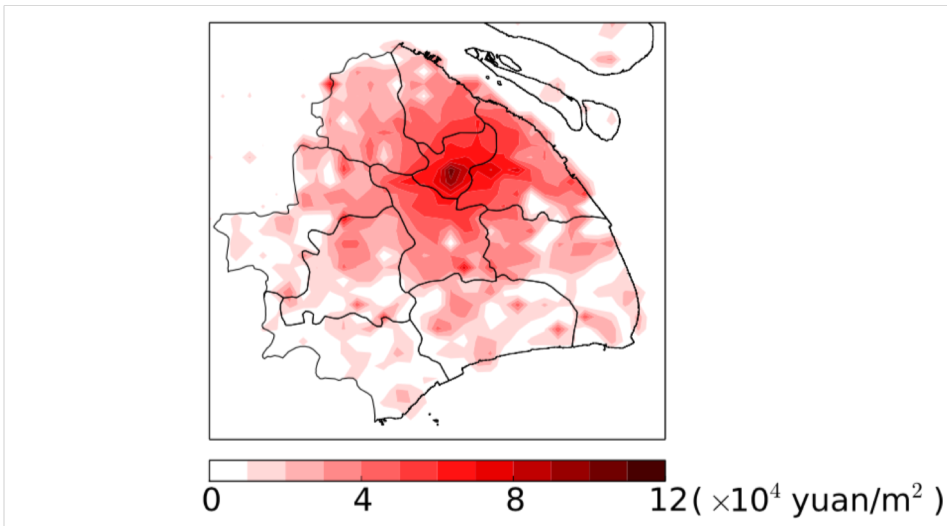
- Q2: What kind of people tend to have less diverse connections? Leaving migrants or staying migrants?

townsman: the fraction of v 's contacts born in the same province
province diversity: entropy of the distribution of birth provinces among v 's contacts
communication diversity: Shannon entropy of the distribution of the number of calls to their contacts



The (Dis)integretion of Migrants

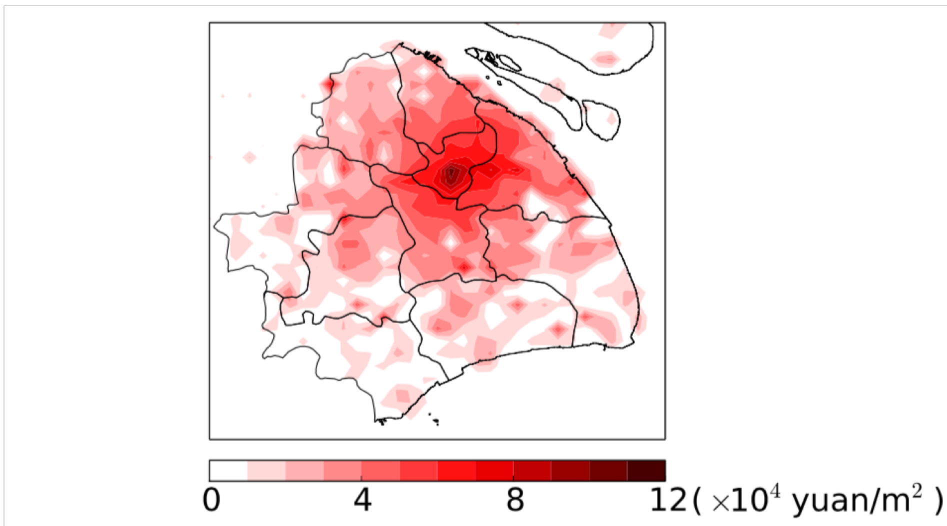
- Q3: What kinds of people tend to be active at more expensive area? Leaving migrants or staying migrants?



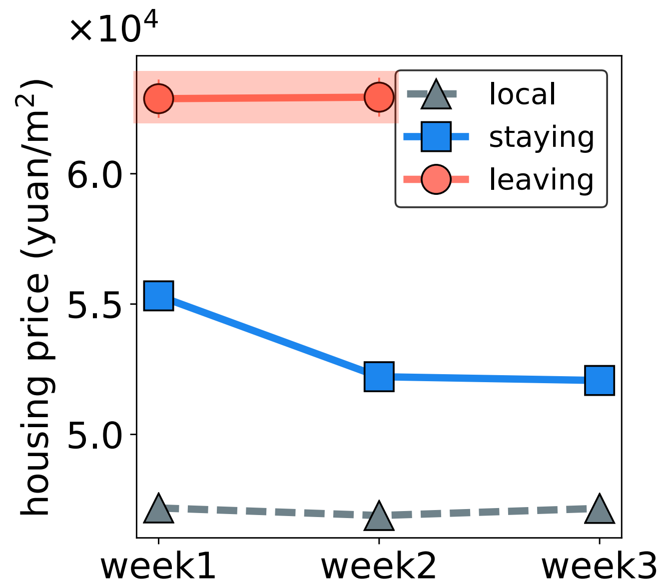
(a) Housing price distribution in Shanghai

Leaving migrants tend to stay in most expensive area

- Q3: What kinds of people tend to be active at more expensive area? Leaving migrants or staying migrants?



(a) Housing price distribution in Shanghai



(b) Avg. housing price of users' active areas.

The (Dis)integration of Migrants

- Feature sets:
  - Ego network properties
  - Call behavior
  - Geographical patterns
  - Housing price information

Classification Tasks

- New Migrants (*35K*) vs. Locals (*1.7M*)
- Leaving Migrants (1.4K) vs. Staying Migrants(*34K*)

Mobile networks，user v → **new migrant?** → **leaving migrant?**

new migrant? → **staying migrant?**

Mobile networks，user v → **local?**

Task 1

Task 2

New Migrants from Locals

- New Migrants(*35K*) vs. Locals(*1.7M*)
- Classifier: random forest
- 5-fold cross-validation

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.

New Migrants from Locals

- New Migrants(*35K*) vs. Locals(*1.7M*)
- Classifier: random forest
- 5-fold cross-validation

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.

Churn prediction problem

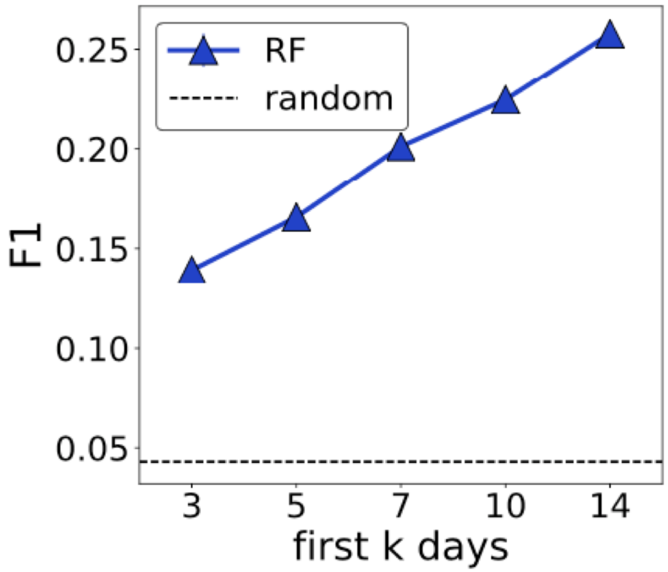
- Leaving Migrants(1.4K) vs. Staying Migrants(*34K*)
- Classifier: random forest
- 5-fold cross-validation

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.

Churn prediction problem

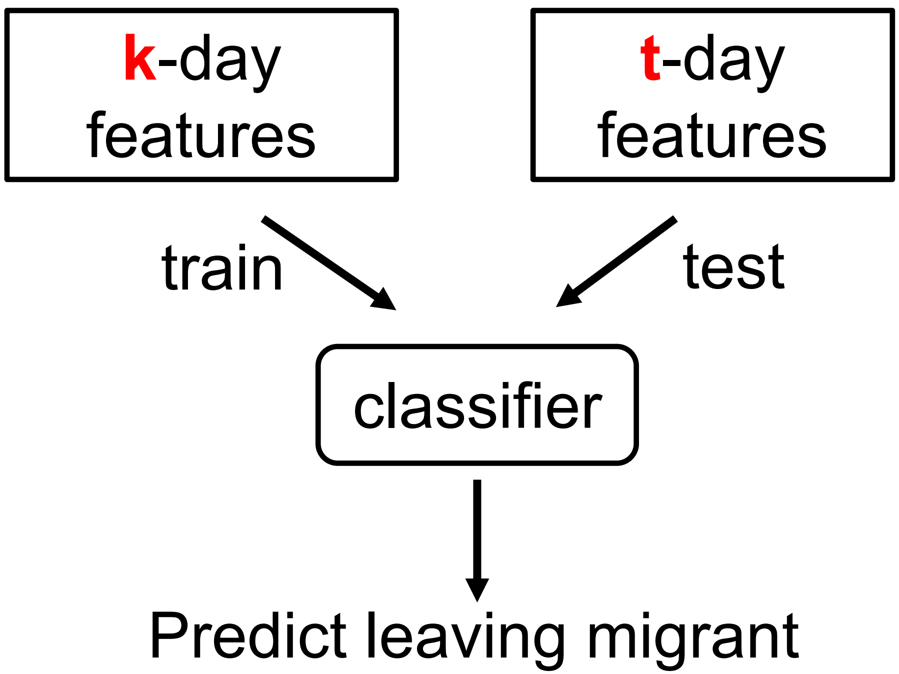
- Early detection of leaving migrant
  - Is it possible to detect leaving migrants sooner than two weeks?
    - If so, we may be able to provide integration service.
  - We extract features based on one's information from the first k days.



(c) F1.

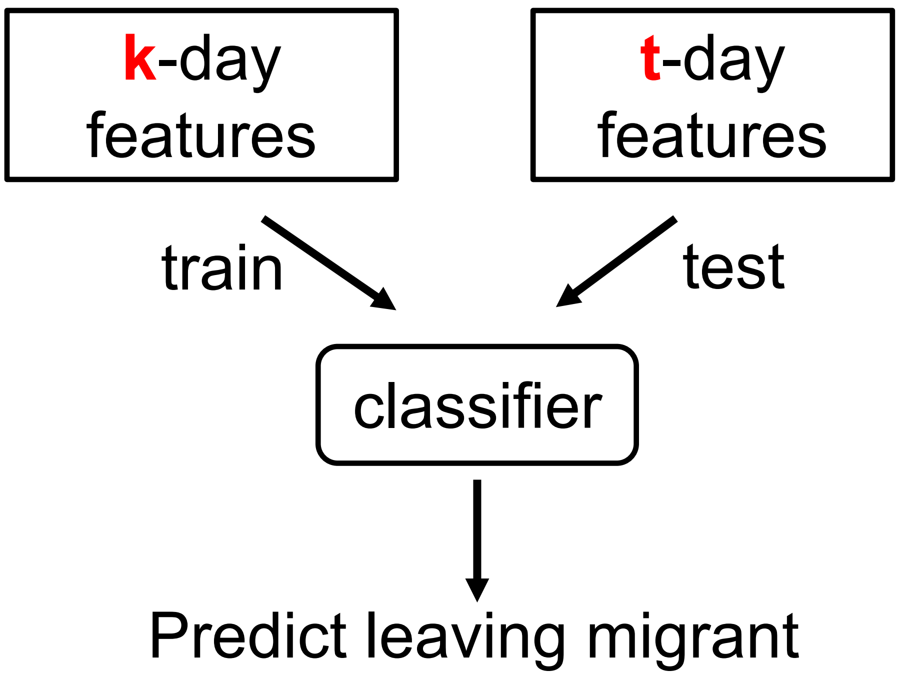
Churn prediction problem

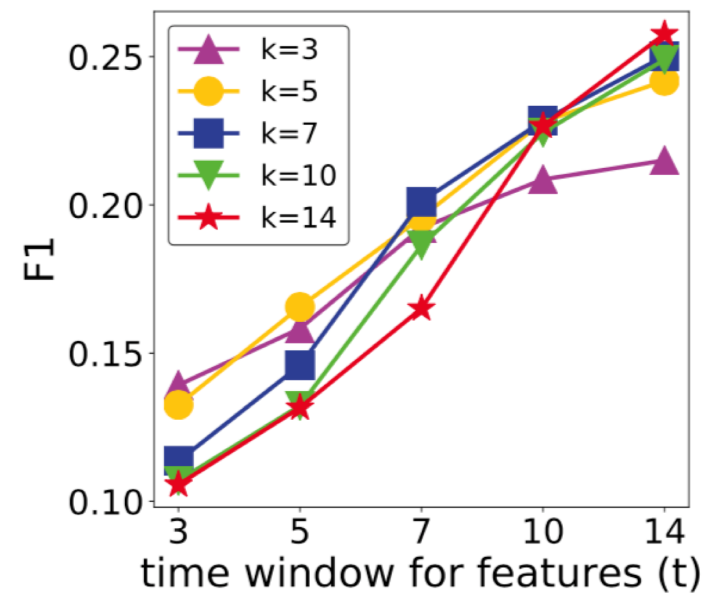
- Why does the performance improve?
  - We disentangle the improvement due to feature quality or classifier quality



With the first 5 days' data, the classifier performs as well as those trained using 14 days

- Why does the performance improve?
  - We disentangle the improvement due to feature quality or classifier quality





(d) Disentangling performance improvement.

Summary

- We study the problem of **early departure of new migrants**.
- Leaving migrants develop **less diverse connections** and their active areas also **have higher housing prices** than that of staying migrants.
- Classification performance improves over time, mainly because the features become more robust.

Summary

- We study the problem of **early departure of new migrants**.
- Leaving migrants develop **less diverse connections** and their active areas also **have higher housing prices** than that of staying migrants.
- Classification performance improves over time, mainly because the features become more robust.

Thank you!

Q&A

QR code for housing price data:

Appendix: Telecommunication in China

- Obtaining a local number is the first integration step for a new migrant
  - Long-distance call cost
- It is uncommon for a temporary visitor to obtain a local number
  - obtaining a phone number is nontrivial and requires personal identification
- We can identify people who just obtained a local number but were not from Shanghai originally.
  - Personal identification allows us to extract the birthplace of a person.

Appendix: Data Privacy

- All data we used was anonymized by China Telecom
- We only have meta data, without contents.

Appendix: Feature Sets

	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, $\frac{\lvert\{(s,t) \mid (s,t) \in E_t\}\rvert}{d_v(d_v - 1)}$, where s and t are v's contacts, and d_v is v's degree.

Appendix: Feature Sets

	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.

Appendix: Feature Sets

Geographical features of v at time t	
center	The latitude and longitude of a user v's center of mass l_{CM}, $l_{\mathrm{CM}} = \frac{1}{\lvert L_v^t \rvert} \sum_{l \in L_v^t} l$.
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., $\frac{1}{\lvert L_v^t \rvert} \sum_{l \in L_v^t} \lvert l - l_{\mathrm{CM}} \rvert$.
max radius	The maximal distance of v from her center of mass, i.e., $\max_{l \in L_v^t} \lvert l - l_{\mathrm{CM}} \rvert$.
moving distance	The total distance that v moves, $\sum_i \lvert l_i - l_{i-1} \rvert$.
average distance	The average distance that v moves, $\frac{1}{\lvert L_v^t \rvert} \sum_i \lvert l_i - l_{i-1} \rvert$.
home distance	The distance between v's workplace and home.

Appendix: Feature Sets

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.