JOURNAL OF SOCIAL COMPUTING
ISSN 2688-5255 06/06 pp71–88
Volume 2, Number 1, March 2021
DOI: 10.23919/JSC.2021.0003

Estimating Multiple Socioeconomic Attributes via Home Location—A Case Study in China

Shichang Ding, Xin Gao, Yufan Dong, Yiwei Tong, and Xiaoming Fu*

Abstract: Inferring people's Socioeconomic Attributes (SEAs), including income, occupation, and education level, is an important problem for both social sciences and many networked applications like targeted advertising and personalized recommendation. Previous works mainly focus on estimating SEAs from peoples' cyberspace behaviors and relationships, such as the content of tweets or the social networks between online users. Besides cyberspace data, alternative data sources about users' physical behavior, like their home location, may offer new insights. More specifically, in this paper, we study how to predict a person's income level, family income level, occupation type, and education level from his/her home location. As a case study, we collect people's home locations and socioeconomic attributes through a survey involving 9 provinces and 85 cities in China. We further enrich home location with the knowledge from real estate websites, government statistics websites, online map services, etc. To learn a shared representation from input features as well as attribute-specific representations for different SEAs, we propose H2SEA, a factorization machine-based multi-task learning method with attention mechanism. Extensive experiment results show that: (1) Home location can clearly improve the estimation accuracy for all SEA prediction tasks (e.g., 80.2% improvement in terms of F1-score in estimating personal income level); (2) The proposed H2SEA model outperforms alternative models for SEA inference in terms of various evaluation metrics, such as Area Under Curve (AUC), F-measure, and specificity; (3) The performance of specific SEA prediction tasks (e.g., personal income) can be further improved if H2SEA only focuses on cities or villages due to urban-rural gap in China; (4) Compared with online crawled housing price data, the area-level average income and Points Of Interest (POI) are more important features for SEA inferences in China.

Key words: personal income; family income; occupation; education; multi-task learning

1 Introduction

Inferring people's Socioeconomic Attributes (SEAs), such as income level, education level, and occupation types, is an important problem for social computing[1]. These attributes play an important role in studies like social stratification and social welfare. They are also basic factors to calculate people's Socioeconomic Status (SES), which is a key concept in sociology[2, 3]. They can help governments to design and evaluate social policies, especially for welfare policy. SEAs also offer hints for online service providers to design personalized services in recommendation and advertisement[4–7]. However, these attributes are hard to collect for both researchers and companies, since people are reluctant to expose their income or job information or the legal privacy framework

• Shichang Ding is with State Key Laboratory of Mathematical Engineering and Advanced Computing, Zhengzhou 276800, China. E-mail:sding@cs.uni-goettingen.
• Xin Gao is with the Department of Sociology, Tsinghua University, Beijing 100085, China. E-mail: gaoxinxg@126.com.
• Yufan Dong and Xiaoming Fu are with the Institute of Computer Science, University of Göttingen, Göttingen 37077, Germany. E-mail: yufan.dong@stud.uni-goettingen.de; fu@cs.uni-goettingen.de.
• Yiwei Tong is with Shanghai Hejin Information Technology Company, Shanghai 200100, China. E-mail: yiwei@heywhale.com.
• This work has been conducted while the first author was a PhD student at the University of Göttingen, Germany.
* To whom correspondence should be addressed.
Manuscript received: 2020-12-18; accepted: 2021-01-18

does not allow. As previous studies[8] point out, in real-world datasets, only partial demographic attributes are known for a great number of users, while some users do not provide any attributes at all.

Given its importance, various machine learning methods have been proposed to automatically estimate people's SEA from their cyberspace behavior[9–15]. For example, the language patterns, topics, or even emotions in tweet content have been used to estimate people's income or occupation[13–15]. People's mobile phone usage habits are also leveraged to predict people's SES or family income[9–12]. More recently, researchers begin to get interested in inferring SEAs from people's physical behaviors. For instance, many researchers estimate people's income and education level based on how people purchase items from offline retailers[8, 16], and infer people's income level from people's mobility pattern in subway systems[17].

However, home location, as a rather informative and fundamental user behavior indicator, has been overlooked by most previous works for SEA inference. Previous works mainly focus on utilizing people's home location for targeted ads of local business[18], urban planning[19], location-aware recommendations[20, 21], etc. Home location has been observed to be related to people's attributes. For example, racial segregation in housing is quite clear in the USA[22, 23]. In China, researchers have also already observed that SEAs, such as income and occupation, are related to the home location[12, 24, 25]. In the USA, racial segregation in housing is most visible and prominent. As far as this paper is concerned, we need to consider whether Chinese people are segregated by income, education, and occupation. Since China's housing system reformed in 1998, housing prices have risen rapidly, with an average annual growth rate of 7%[26]. During this period, the inter-city and intra-city differences of housing prices in China become gradually significant[27, 28]. Wang et al.[29] find that the level of residents' income and wealth is important for the spatial differentiation of housing prices. Besides the differences among cities, the spatial differentiation of housing prices within the city is also relatively large. For example, Wang et al.[30] found out that in Yangzhou city, the housing prices are high in the center and low in the periphery, high in the west and low in the east. And Wang et al.[30] also found out that the spatial aggregation of specific income class is one of the main influencing factors for this distribution pattern of housing prices in Yangzhou city: in the high-income areas, the grade of houses are often higher, where the residential types are mainly villas and high-grade ordinary commercial housings; in the middle- and low-income areas, the residential types are mainly affordable housing, middle- and low-grade housing. Therefore, the location orientation of specific residential type and grade and the spatial agglomeration of specific income class often interact and influence each other[30]. The related studies in other Chinese cities, such as Guangzhou, Nanjing, Beijing, Zhengzhou, and Dongguan[24, 25, 31, 32], also find there are correlations between the spatial distribution of housing prices and income class and occupation.

Though important, investigating the relationship between people's SEAs and the home location is quite challenging for the following reasons. First, though datasets containing both personal SEAs and home location are critical for meaningful experiments, there are almost no open datasets including such information as far as we checked. Second, the home location itself only contains limited information (features) for prediction, which makes it difficult to predict attributes furthermore, and personal income, occupation, and education levels are complex attributes that are hard to predict even with rich human behavior data[10, 33].

To tackle these problems, we propose a Home to SEA (H2SEA) method to infer people's attributes, including *personal's monthly income level*, *family yearly income level*, *family yearly consumption level*, *occupation type*, and *education level* from their home location. To the best of our knowledge, this is the first work focusing on SEA inference through the home location. The main contributions are summarized as follows:

- We enrich the home location with knowledge from various aspects，such as area-level economic statistics, housing price, Point Of Interest (POI), and administrative division. We design multiple SEA-related features according to this knowledge. The source data of these features are mined from multiple commercial real-estate websites, official statistic bureau websites, online maps, etc.
- We propose a factorization machine-based multi-task learning method with attention mechanism, to learn a shared representation from input features as well as attribute-specific representations for different SEA prediction tasks. The multi-task method can additionally leverage the potential relationship between income, education, and occupation. Comparing with existing multi-task learning methods for attribute inference, the

proposed model further improves the performance with limited features by modeling the second-order feature interactions with Factorization Machine (FM).

• As a case study, we carry out a survey to collect people's personal income level, family income level, occupation types, and education level in China. In the end, we collect a dataset that includes 9 provinces and 85 cities in China. The experiments on this dataset demonstrate that (1) home location can clearly improve the performance of predicting people's SEAs; (2) the proposed method outperforms compared methods on all SEA prediction tasks in terms of multiple metrics, such as F1-measure.

• By further analyzing the relationship between SEAs and home location, we find that the most important features in most SEA predictions are county-level average income and POI distribution, instead of housing price crawled from real estate agency websites. We conjecture that these are caused by various kinds of non-commercial accommodations in China. People living in these kinds of accommodations usually paid much lower prices compared with the commercial ones nearby.

The rest of this paper is structured as follows. Section 2 introduces the ground truth dataset. Section 3 discusses how to design and collect data for SEA-related features. The H2SEA model is proposed in Section 4. Experimental results are presented in Section 5. Section 6 further analyzes the relationship between housing prices and income in China. Related work is then reviewed in Section 7. The paper is concluded in Section 8 with a brief discussion of limitations and directions to future research.

2 Ground Truth Dataset

We collected a dataset covering a sampled population's personal SEAs from 2015 to 2017 in China. In terms of sampling method, we regard the province as an independent population and adopt multi-stage Probability Proportionate to Size (PPS) sampling stratified sampling[34]. This is one of the methods often used in social, economic, and demographic surveys. It is a kind of multi-stage sampling method combining stratified sampling with simple random sampling or unequal sampling. It contains different sample stages. In this survey, we have four stages according to China' administrative unit, including district, town, village, and household. China is the second-largest economy in the world and develops fast in the last several decades. However, the differences of development levels among different areas are quite big. During the survey, we choose the provinces and cities which are at different economic levels to get a better understanding of the whole China. In the end, the dataset covers 85 cities in 9 provinces of China, such as Zhejiang, Guangdong, Jiangsu, Sichuan, Shannxi, Hebei, etc. The sampling units at each stage are shown in Table 1.

The PPS extraction formula is

$$s_p = \frac{10G_\alpha}{S_p} \times \frac{2G_\beta}{G_\alpha} \times \frac{2G_\gamma}{G_\beta} \times \frac{30}{G_\gamma} = \frac{1200}{S_p} \quad (1)$$

where s_p is the sampling proportion, G_α is the sample size of districts, G_β is the sample size of towns, and G_γ is the sample size of villages. 10, 2, 2, and 30 are the sample volume to be selected from each level, respectively, and S_p is the number of samples to be selected from the overall population. The total sample size of each province is estimated to be 1200, and the total sample size of 9 provinces is about 10 800. In three years, we investigated 32 443 people. The household survey was conducted by knocking on the doors of households in the morning, noon, and evening. If no one was seen, this household would be replaced by nearby households until 30 households responded. However, the recorded home locations of these people are community-level, which often covers more than thousands of people in China. During early experiments, we find this is too coarse-grained for personal-level SEA-inference tasks. To get more accurate home locations, we re-collected the volunteers' check-in data on a famous online social network platform called QQ. Following previous methods[17], we combine the most visited check-in location during the night and the collected home location information to calculate the latitude and longitude of a person's home. Among 32 443 volunteers, 4509 of them (who live in their own house instead of rented one) reported at least one socioeconomic attribute and agreed to share their home location for research purposes. Each record in the dataset consists of an anonymous volunteer's ID, age, gender, home location,

Table 1 Sampling stage and corresponding sampling unit and quantity.

Sample stage	Unit	Quantity
The First stage Sampling Unit (FSU)	District, county	10
The Second stage Sampling Unit (SSU)	Township, town	2
The Third stage Sampling Unit (TSU)	Village, community	2
The Ultimate stage Sampling Unit (USU)	Household	30

SEAs, etc. SEAs include *personal monthly income level*, *family yearly income level*, *personal education level*, and *personal occupation type*.

The demographics description is shown in Table 2. To protect personal privacy, we ask most volunteers to choose general socioeconomic levels. Besides, about 2800 people also agreed to submit their exact monthly income number. In this paper, we only use the accurate income ranges to calculate the correlation coefficients between income and housing price in Section 6. The recorded ID is a random number, which has no relationship with the volunteers' identification information. All data were collected under a confidentiality agreement and only allowed for research purposes.

3 Feature Engineering

Assume a person's home location is H_i, and $[x_1, x_2, \ldots, x_i, \ldots, x_n]$ are the features generated only based on H_i. Given a person' features F, we aim to estimate his/her *personal monthly income level*, *family yearly income level*, *family yearly consumption level*, *occupation type*, and *education level*. In this paper, all the sub-problems are defined as three-level classification tasks. The latitude and longitude of the home location are too limited for multiple SEA predication. Hence, we need to design SEA-related features to enrich the home location. In this section, we introduce how to design SEA-related features as well as collecting corresponding data for these features.

Table 2 Demographics description.

Demographic		Fraction (%)
Personal monthly income	Under 2000 yuan	23.04
	2000–4000 yuan	30.74
	Over 4000 yuan	20.16
	Not answer	26.06
Family yearly income	Under 40 000 yuan	37.19
	40 000–75 000 yuan	35.82
	Over 75 000 yuan	25.86
	Not answer	1.13
Personal occupation type	Farmers, temporary worker, unemployed, etc	65.95
	Ordinary employers, freelancer, etc	24.91
	Middle and senior managers, etc	8.32
	Not answer	0.82
Personal education level	Lower secondary education	54.29
	Upper secondary education (high)	23.20
	University	22.51
Age	Under 30	6.18
	31–40	11.40
	41–50	24.36
	51–60	27.74
	Over 60	30.27
	Not answer	0.05

3.1 Features based on housing price

A common observation is that personal income or occupation may be related to people's housing price[12, 17]. The government usually only publishes area-level average housing prices (e.g., city-level or county-level in China), which may be too coarse-grained for personal SEA predication. Thus, it is hard to get the exact housing price of the house in which the targeted user lives in. Fortunately, some commercial real estate websites may publish the housing price of a house in or near a specific Global Positioning System (GPS) location which is now for selling. In this paper, we collect the housing prices, which are near one home location, from some real estate commercial websites.

HP denotes the collected housing price list of a home location: $\mathrm{HP} = \mathrm{hp}_1, \mathrm{hp}_2, \ldots, \mathrm{hp}_{\mathrm{numhp}}$. F_{hp} includes the number of houses for selling F_{numhp}, the average value F_{avhp}, median value F_{mvhp}, the max value F_{maxhp}, the min value F_{minhp}, and the standard deviation F_{stdhp} of housing prices.

The housing price dataset is mainly crawled from the Lianjia website (Lianjia.com, one of the biggest real estate agency service providers in China), which records the house prices and location information of apartments selling in China. From the Lianjia website, we can crawl the prices of the houses which are less than 2 kilometers away from one person's home location. We can find housing price information for 43% of volunteers. Lianjia only records the prices of houses which are sold in recent times. So there may be no housing price data for one person's home location if no nearby houses are for selling in recent times. For missing F_{hp}, we use the nearest known housing price as a substitute if the distance is less than 10 kilometers. If there is no housing price data nearby, we use the city-level average housing price published by local governments as a substitute.

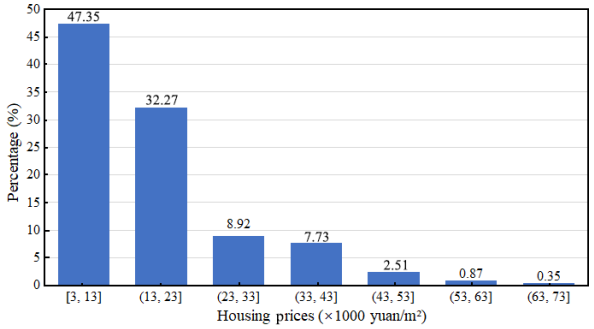
We also tried to find more housing prices on the other important real estate commercial websites. However, the data on other websites are not so reliable. First, one current owner may tend to show their housing prices on all of the important websites. So if we cannot find housing prices on Lianjia.com, it is highly possible that we cannot find the prices on other websites either. Besides, it is worthwhile to mention that, several other

websites tend to show a much lower housing price to attract customers. They only tell the real asking price during the telephone contact with a customer. This phenomenon actually confuses our early-stage analysis.

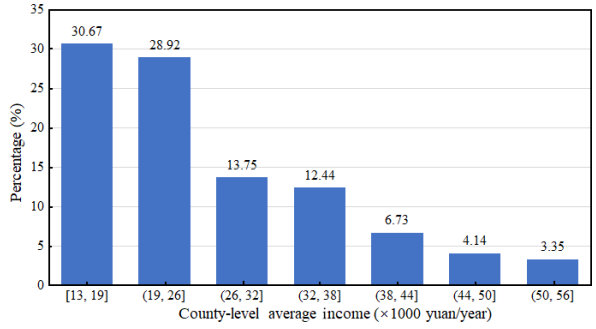
The distribution of housing prices is shown in Fig. 1a. The highest average housing price is 64 354 yuan/m^2 in Shenzhen city of Guangdong province and the lowest housing price is 3117 yuan/m^2 in Xuzhou city of Jiangsu Province.

3.2 Features based on renting price

Features based on renting price F_{rp} include a set of features related to the renting prices around a person's home location. Though in this paper, we mainly focus on the people who have their own house/apartment, the renting prices of an area could be also helpful in predicting people's SEAs for the following reasons. First, the number of crawled housing prices in many communities is not enough. It can not completely cover all home locations. And we observe that renting prices are usually high in high housing price areas. So renting price is an important supplement to housing prices. Second, renting prices may be related to the income or consumption level of people who have their own houses. For example, some people could gain more income by renting their house to others. It would be better to introduce more related features to alleviate the limitation of input features.



(a) Distribution of housing prices



(b) Distribution of county-level average income

Fig. 1 Distributions of housing prices and county-level income in China.

RP denotes the collected renting price list of a home location: RP $= \text{rp}_1, \text{rp}_2, \ldots, \text{rp}_{\text{numrp}}$. F_{rp} is just like F_{hp}. It includes the number F_{numrp}, the average value F_{avrp}, the median value F_{mvrp}, the max value F_{maxrp}, the min value F_{minrp}, and the standard deviation F_{stdrp} of RP.

We use similar methods, like housing prices, to collect renting prices for each home location. We also collect the renting prices from commercial websites like Lianjia.com. We can find renting price data inside the 2 km radius of 32% home location. The others are using the nearest known renting prices as an approximation.

3.3 Features based on official area-level economic statistics

Features based on area-level economic statistics include several kinds of features, such as average income, Gross Domestic Product (GDP), and government budget and tax. These area-level economic statistics mainly reflect the economic development level of one administrative division. Some statistics are directly related to the SEAs of people living in the area. They are usually published by governments and could be found on government websites (https://www.census.gov/quickfacts/bostoncitymassachusetts). In some developed countries, there may be fine-grained statistics. For example, French governments publish a composite index called SEL[35]. The SEL of a district is calculated based on the income, assets, and education of people who live in this district. The area of one district is only 1–4 km^2. However, in developing countries like China, most local governments only publish coarse-grained statistics. We find that Chinese governments only publish county-level average income for most areas. A county in China can cover hundreds of thousands of people and hundreds of square kilometers (https://en.wikipedia.org/wiki/List_of_counties_in_China). Though quite coarse-grained, these statistics could be helpful in prediction, because they are all related to the economic levels of an area in which the home location belongs to. In this paper, we mainly collect three types of Chinese area-level economic statistics features.

County-level average income F_{clai}. If we could collect very fine-grained average income inside the area of a home location, the feature F_{clai} is calculated just like F_{hp} and F_{rp}. However, in actual scenarios, the published county-level average income in China covers a very large area. So for home locations in one county,

their values of F_{clai} share one same value.

Town-level budget and tax, F_{tlb} and F_{tlt}. The town is an administrative division smaller than the county and larger than the community or village. The town is the smallest administrative division, of which the economic statistics can be found on Chinese government websites. We cannot find town-level statistics that are directly related to personal income or occupation, like average income. In this paper, we use the town-level budget and tax, which may be indirectly related to people's SEAs. And they are easier to be found on the websites of bureaus of statistics (http://gc.public.zhengzhou.gov.cn/02SCB/1419273.jhtml).

Unlike commercial websites which publish updated housing price, governments tend to publish the economic statistics for several years. Economic statistics in a long time could be also helpful in prediction. A long time of economic statistics can avoid anomalies and also reflect the potential in development. In this paper, we collect the history of the county-level average income, town-level budget, and town-level tax (from 2012 to 2017). Most county-level average income, town-level budget, and town-level tax can be found on government websites. For example, a county-level average income can be found in some government websits (http://www.cdstats.chengdu.gov.cn/htm/detail_63138.html). In some cities, the average income is also written as average disposable income. We can find the average income for 97% of counties. For the other counties, we use city-level average income as an approximation. The distribution of county-level average income is shown in Fig. 1b. The largest county-level average income is 56 442 yuan/year in Yuexiu county, Guangzhou city of Guangdong province. The smallest is 13 314 yuan/year in Xuanhan county, Dazhou city of Sichuan province. We could also find 27% of towns' budget data and 16% of towns' tax data.

3.4 Features based on point of interests

The urbanization process leads to different functional regions in a city, e.g., entertainment areas, business districts, and residential areas[36]. The function of living areas may be related to people's occupation and education level. POIs can be used to give a description of the function of one area. If there are many restaurants and few schools/universities in a living area, we may think this is an entertainment area. As the number of restaurants in most areas is typically much larger than that in schools/universities, we need to carefully check the overall distribution. If compared to the overall distribution, the percentage of schools/universities is higher while that of the restaurants is lower, then this should be an education area rather than an entertainment area.

First, the POI information of all home locations should be collected. Then for the j-th POI category, its overall frequency of all home locations is the ratio of the number of the j-th category to the number of all POIs. The frequency of the j-th POI category in one home location H_i is the ratio of the number of the j-th category in H_i to the number of all POIs in H_i.

Then features based on point of interests F_{poi} are

$$F_{\text{poi\_dis}} = \{\text{of}_1/\text{OF}_1, \text{of}_2/\text{OF}_2, \ldots, \text{of}_l/\text{OF}_l\} \quad (2)$$

where l is the number of collected POI categories. OF_j means the proportion of the j-th POI category in the whole collected POI dataset. of_j means the proportion of the j-th POI category collected for one home location. For a home location, if of_j/OF_j is larger than 1, it means this area has more POIs in the j-th category compared with the overall distribution. Then the j-th category is more important to determine the function of this area.

POI dataset is also crawled based on Baidu Map API Service (lbsyun.baidu.com). We collect all POI records which are less than 2 km away from the home location. We can find POI information in all communities. There are 2 levels of POI in Baidu Map services. The first level includes 21 categories including public facility, domestic services, education, business residence, hospital, hotel, car services, sport, leisure, scenery, restaurant, public transportation, financial services, etc. Each first-level POI category consists of various kinds of second-levels POI categories. For example, the food category includes 9 second-level categories, like Chinese restaurants, foreign restaurants, coffee bars, etc. We use the second-level POI category opened by Baidu Map. There are 114 different kinds of second-level POI categories involved in our dataset.

3.5 Categorical features

Here we introduce several categorical features. The categorical features are different from the above continuous features. It usually contains a number of categories or distinct groups. And there might not be a logical order between different categories or groups.

Zip code, F_{zp}, can be used as a home-based feature. as a home-based feature. Zip code or postal code is used by postal service. Its basic format consists of several digits (e.g., 6 digits in China). There is a zip code for

each town in China. The zip code of a town to which a home location belongs can be found on official websites.

Location names can also be used in prediction. In China, we can use the province name (F_{pn}), city name (F_{cn}), county name (F_{con}), town name (F_{tn}), and street name (F_{sn}). They are corresponding to people's living areas. Location names are useful, because the gap between different places in China is quite big. For instance, people living in eastern coastal provinces or province capitals usually have much higher incomes than others.

During our data collection stage, we also collect the urban type of a home location. Urban types include 3 categories: city center, city border, and rural area. There are very serious urban-rural income gap and inequality in China[37]. People in the city-center may have higher income, better education, and more working career opportunities than rural areas.

The province name, county name, town name, and zip code can be found easily on official websites (https://worldpostalcode.com/china/). There are 3 different types among all home locations: city center (30%), city border (24%), and villages (46%). 54% of home locations are or near urban areas. This is very close to China's urbanization rate 58.7%, investigated in 2017 (Urbanization rate in China, https://en.wikipedia.org/wiki/Urbanization_in_China).

4 H2SEA

The overall architecture of the proposed method H2SEA is presented in Fig. 2. In this section, we present the details of H2SEA. H2SEA model predicts a person's N kinds of socioeconomic attributes (denoted as $\boldsymbol{Y} = \{Y^1, Y^2, \ldots, Y^N\}$) based on his/her home location (denoted as H).

4.1 FM-based shared embedding layers

FM-based shared embedding layers consists of an embedding layer and an FM layer. The features are fed into the embedding layer to get an initial representation that is shared for all tasks. Previous works[8] usually use one feedforward neural network layer to get the initial embeddings for the input basic features. However, there is one problem: the features based on a single home location maybe not enough for predicting people's SEA. To tackle these problems, we leverage FM[38] to generate the embeddings. Compared with the feedforward neural network layer, FM additionally considers the value of feature interactions. Feature interactions can improve SEA prediction by modeling the underlying relationship between different features. Simply put, it generates new second-order features based on basic input features. FM can automatically learn feature interactions. It embeds features into a latent space and models the second-order interactions between features via inner product of their

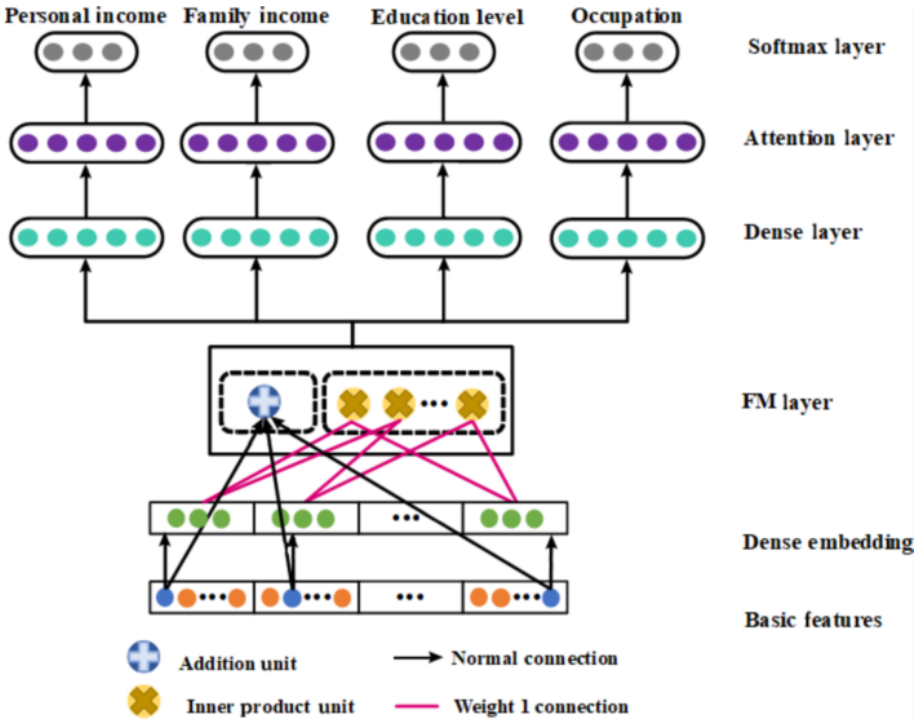


Fig. 2 Architecture of H2SEA model.

embedding vectors.

In this paper, we denote one socioeconomic attribute of a person u as $\boldsymbol{Y}$ and the basic features as $\boldsymbol{X}$. $\boldsymbol{X}$ includes both continuous fields (e.g., features based on housing price) and categorical fields (e.g., features based on zip code). We represent every categorical feature as a vector of one-hot encoding and every continuous feature as the value itself. $\boldsymbol{X} = [x_1, x_2, \ldots, x_i, \ldots, x_n]$. x_i is the vector representation of the i-th feature, like F_{hp}, F_{poi}, or F_{zp}, n is the number of all basic features.

FM could be seen as a combination of the embedding layer and inner product layer. Here we actually use the latent feature vectors in FM as embedding network weights. The output of FM layer is the summation of an additional unit and a number of inner product units. The FM-based shared embedding e_{fm} is defined as

$$e_{\text{fm}} = \langle w_{\text{fm}}, x\rangle + \sum_{i=1}\sum_{j=i+1} \langle \boldsymbol{V}_i, \boldsymbol{V}_j\rangle x_i \cdot x_j \quad (3)$$

where $w_{\text{fm}} \in \mathbf{R}^k$ and $\boldsymbol{V}_i \in \mathbf{R}^d$. k is the dimension of the one-hot vector and d is the dimension size of embedding layers. For a person, w is used to weigh its basic features' order-1 importance ((w_{fm}, x)). The latent vector $\boldsymbol{V}_i$ can measure the impact of interactions between the feature x_i and all the other features by the inner product units. FM can train latent vector $\boldsymbol{V}_i$ (or $\boldsymbol{V}_j$) whenever i (or j) appears in a data record.

4.2 Attention-based attribute specific layers

Attention-based attribute specific layers consist of a dense layer and an attention layer. The shared representations have captured the global signal shared by all attribute predication tasks. Next, we need to refine the shared representations to adapt to the different tasks. For each task, we use a dense layer and an attention layer to generate attribute-specific representations. First, we use the dense layer to learn an primary attribute-specific representation for the n-th SEA,

$$d^n = \text{relu}(e_{\text{fm}} \times w_d^n + b_d^n) \quad (4)$$

where relu is a non-linear activation. w_d^n and b_d^n are weight and bias parameters for the n-th task. It is reasonable to assume that some input feature maybe more related with certain SEAs than others. For example, area-level average income may be more related with income level, while the POI distribution may be more related with occupation types. To model the varying importance of features for different attribute, here we use an attention layer,

$$t^n = \text{relu}(d^n \times w_t^n + b_t^n) \quad (5)$$

$$a^n = \text{softmax}(t^n) \quad (6)$$

where a^n denotes the attention weights for the n-th tasks. The sum of a^n equals to 1. The distribution of a^n can be seen as the importance of each feature embedding for the n-th transaction. The final representation for the n-th SEA predication task u^n is the weighted sum of all shared embeddings,

$$u^n = \sum_{i=1}^{k} a_i^m \times d_i^m \quad (7)$$

4.3 Prediction layers

Prediction layers consist of all prediction layers for 4 SEA inference tasks. In the end, the output of attention-based attribute specific layer u^n is fed into the softmax (or sigmoid layer) to estimate the SEAs of a person. Take the n-th SEA as an example, the predication probability $\hat{y}_n$ is defined as

$$\hat{y}^n = \text{softmax}(u^n) \quad (8)$$

If the distribution of attributes is even, the loss function $\mathcal{L}_n$ is computed as follows:

$$\mathcal{L}_n = -\frac{1}{M_n}\sum_{j=1}^{M_n}\sum_{k=1}^{C_n} y_{j,k}^n \log(\hat{y}_{j,k}^n) \quad (9)$$

where M_n is the number of users whose n-th SEA is not missing. C_n is the number of n-th attribute category. $y_{j,k}^n$ and $\hat{y}_{j,k}^n$ are the ground truth and estimated SEA labels, respectively.

If the distribution of an SEA is quite imbalanced, we leverage a weighted cross-entropy function to calculate the prediction loss $\mathcal{L}_n$ as follows:

$$\mathcal{L}_n = -\frac{1}{M_n}\sum_{j=1}^{M_n}\sum_{k=1}^{C_n} w_{y_{j,k}^n} y_{j,k}^n \log(\hat{y}_{j,k}^n) \quad (10)$$

where $w_{y_{j,k}^n} = \dfrac{\sum_{k=1}^{C_n}\sqrt{M_n^k}}{\sqrt{M_n}}$ is a parameter to control the cost weight of each attribute category, M_n^k is the number of people with the n-th attribute label. The total loss of all SEA tasks can be defined as

$$\mathcal{L}_{\text{total}} = \sum_{n=1}^{N} \lambda_n \mathcal{L}_n + \alpha||\Theta|| \quad (11)$$

where λ_n are hyper-parameters controlling the relative importance of the n-th SEA predication task. We enforce that $\sum_{n=1}^{N}\lambda_n = 1$ to facilitate the tuning of the hyper-parameters. Θ denotes all trainable parameters of H2SEA model. We adopt L2-normalization[39] and dropout[40] to prevent overfitting. α controls the L_2 regularization strength. By optimizing the entire loss $\mathcal{L}_{\text{total}}$, our model can get the best results for

recommending task. H2SEA model is trained via back-propagation and Adam[41].

5 Experiment

In this section, through experiments based on the actual dataset, we want to answer the following questions: (1) Whether home location has predictive power for socioeconomic attributes? (2) Whether H2SEA model outperforms widely-used baselines? (3) What are the most important home-based features for income, occupation, or education prediction? (4) How different settings (e.g., dropout and λ_n) affect the performance?

5.1 Experiment setup

We use the following SEA prediction tasks to test the predictive power of home location:

• **Personal Income Level (PIL)**. This is a three-level personal income prediction task. The boundary lines are 2000 yuan and 4000 yuan every month. The percentage of people in low-income-level is 31.2%, middle-income-level is 41.5%, while high-income-level is 27.3%.

• **Family Income Level (FIL)**. This is a three-level family income prediction task. The boundary line are 40 000 yuan and 75 000 yuan every year. The percentage of low-level is 37.6%, middle-level is 36.2%, while high-level is 26.2 %.

• **Education Level (EL)**. This is a three-level education level prediction task. This task aims to predict whether a person has a university degree, high school degree, or junior high school degree. The percentage of the junior high school is 54.29%, the high school degree is 23.20%, while the university degree is 22.51%.

• **Occupation Type (OT)**. This is a three-level occupation prediction task. This task aims to predict people's occupation type. The people in low-level (farmers, temporary worker, unemployed) is 66.49%, middle-level (ordinary employers, freelancer) is 25.12%, while high-level(manager) is 8.39%.

Evaluation metrics. We use the following evaluation metrics: macro-F1, Area Under Curve (AUC), G-Mean, and accuracy for all tasks. In unbalanced tasks like OT, macro-F1 is the most important metric.

Baselines. To the best of our knowledge, there exists no model focusing on estimating personal SEAs from home location. Here we use the following widely-used standard classification methods as baselines:

• **POP:** POP simply estimate an individual' SEA as the majority classes[8]. This model ignores all input features.

• **Logistic Regression (LR):** We use 2-degree LR to model the linear combination of basic features and all order-2 feature interactions.

• **Gradient Boosting Decision Tree (GBDT).** The gradient boosting model is famous for its outstanding performance and efficiency for general classification tasks. Xgboost is an open-source gradient boosting library[42]. We use all features to train Xgboost model.

• **Embedding Transformation Network with Attention (ETNA).** This is one of the state-of-the-art multi-task demographic inference models. It also uses an attention mechanism to refine the shared embeddings for different demographics. Its original version is designed for sequential input data. Here we use a simple embedding layer instead. Compared with H2SEA, ETNA neglects the effect of feature interactions. So we can compare the results of H2SEA and ETNA to see the effect of FM-based feature interaction.

• **H2SEA-No-Attention (H2SEA-NA).** For the ablation study, here we use H2SEA-NA to check the effect of the attention mechanism of H2SEA. H2SEA-NA is H2SEA without the attention mechanism.

70% of people are chosen as the training dataset, 20% as validation dataset, and 10% as test dataset. Our model is implemented based on Keras. Hyper-parameters of H2SEA are tuned by grid-searching on the validation set. Due to limited space, here we only show the best settings of PIL as an example. The latent dimension of the FM component (or field embedding size) is 6, the dropout is 0.3, the number of neurons per layer (deep component) is 32, the number of hidden layers (deep component) is 3, the learning rate of Adam is 0.001, the activation function is relu, and the L2-norm ratio is 0.000 01.

5.2 Results analysis

This section mainly answers whether personal SEAs can be predicted based on home location and how H2SEA model performs compared with baselines.

The results of all tasks are showed in Table 3. The numbers in Table 3 are averaged by 10 times of train-testing. To achieve the best performance, we carefully conducted parameter tuning of all methods. From Table 3, we have the following observations.

• Home location clearly improves the performance in estimating personal income, family income, occupation, and education level. Especially, compared with random guess, H2SEA model can increase 80.22% in F1-score and 42.57% in G-Mean in personal income prediction; 64.57% in F1-score and 41.08% in G-Mean in family

Table 3 Performance comparison.

Task	Method	F1-score	AUC	G-Mean	Accuracy
Personal income	POP	0.3329	0.4894	0.3521	0.3561
	LR	0.5314	0.7231	0.4838	0.5233
	Xgboost	0.5734	0.7482	0.4652	0.5296
	ETNA	0.5863	0.7630	0.4719	0.5394
	H2SEA-NA	0.5936	0.7702	0.4987	0.5289
	H2SEA	**0.5999**	**0.7786**	**0.5020**	**0.5501**
Family income	POP	0.3247	0.4978	0.3729	0.3648
	LR	0.4815	0.7035	0.4354	0.5233
	Xgboost	0.5050	0.7181	0.4676	0.5296
	ETNA	0.5183	0.7351	0.4741	0.5434
	H2SEA-NA	0.5315	0.7462	0.4656	0.5588
	H2SEA	**0.5345**	**0.7546**	**0.5261**	**0.5576**
Education level	POP	0.3259	0.4997	0.5529	0.5463
	LR	0.4825	0.7449	0.5676	0.6006
	Xgboost	0.4927	0.7697	0.6568	0.6585
	ETNA	0.5083	0.7975	0.6645	0.6595
	H2SEA-NA	0.5227	0.8171	0.6739	0.6578
	H2SEA	**0.5272**	**0.8289**	**0.7039**	**0.6640**
Occupation type	POP	0.3391	0.5075	0.5562	0.5881
	LR	0.4681	0.7088	0.5618	0.5858
	Xgboost	0.4717	0.6997	0.5835	0.5952
	ETNA	0.4848	0.7201	0.5877	0.5833
	H2SEA-NA	0.4981	0.7325	0.6244	0.5804
	H2SEA	**0.5003**	**0.7434**	**0.6633**	**0.5869**

income prediction; 61.76% in F1-score and 27.31% in G-Mean in education level prediction; 47.55% in F1-score and 19.26% in G-Mean in occupation prediction.

• Considering the relative improvements compared with random guess, personal income level achieves better results than family income, occupation type, and education. It is quite surprising that home location achieves weaker results in family income than personal income. Because a house/apartment is often bought by a family rather than one individual, the housing price may be more related to the family income level than the personal income level. We conjecture the weak predictability may be caused by the weak relationship between housing price and family/personal income level. The most important feature for income is county-level average personal income, which is clearly more related to personal income level. We will further analyze why the relationship between housing prices and income is weak in Section 6. However, we should note that the H2SEA model still performs much better than random guess. The performances of occupation type and education level are weaker than income prediction, indicating that home location alone is not enough to predict these two attributes. Besides, the imbalance of these two attributes also increases the difficulty in estimation.

• H2SEA model outperforms all baselines in terms of F1-score and G-Mean. The second best classifier is ETNA. Here we ignore the results of H2SEA-NA. Because it is mainly for ablation test and not a baseline method proposed in previous works. H2SEA outperforms ETNA in all tasks by 2.43%–3.71%, 2.70%–4.24%, 6.06%–12.94%, and 0.62%–2.84% in terms of F1-score, AUC, G-Mean, and accuracy, respectively. It indicates that second-order feature interactions can clearly improve performance. ETNA is better than all the other single-tasks models, like Xgboost and LR. It demonstrates that the multi-task learning method can model the underlying relationships between various attributes. It is worth to point out that the accuracy of H2SEA is worse than Xgboost in occupation level by 1.39%. This is caused by imbalance. Only 9.63% of people are in higher-level (middle and senior managers, etc). We mainly consider more about AUC, macro-F1, and G-Mean in an unbalanced task. For example, the G-Mean of H2SEA are 13.67% better than that of Xgboost. Besides, the G-Mean of H2SEA is 12.94% better than ETNA in occupation estimation compared to only 6.06% in personal income level estimation. This indicates that H2SEA may better handle imbalanced datasets through the weighted softmax loss function.

• For ablation study, we can compare H2SEA with two models: ETNA and H2SEA-NA. ETNA is one of the state-of-the-art multi-task attribute inference models. It can be seen as H2SEA without a feature interaction mechanism. We already compare ETNA with H2SEA in the third observation of Section 5.2. H2SEA outperforms ETNA in all tasks by 2.31%, 3.13%, 3.72%, and 3.20% in terms of F1-score. We also compare the performance of H2SEA-NA and H2SEA. H2SEA outperforms H2SEA-NA in all tasks by 1.06%, 0.56%, 0.86%, and 0.44% in terms of F1-score. The results show that both the FM-based feature interaction and attention mechanism are helpful in improving the performance of SEAs inference. And the FM-based feature interaction mechanism is also more useful. This shows the importance of creating new second-order features by the feature interaction mechanism when facing the limited input data problem in the SEA inference task.

5.3 Feature importance analysis

This section discusses the most important features in each task. We mainly show the metrics of the top 5 important features in each task. The metrics are calculated when only using one feature for prediction. The importance of home-based features can help to understand the relationship between home location and different SEAs. In Table 4, for each task, the feature importances are decreasing from top to bottom. The importance is mainly ordered by the combined improvement of F1-score, AUC, and G-Mean. From Table 4, we can observe that:

• **County-income is the most important feature for income prediction**. It shows that even coarse-grained area-level income statistics may be of great help for income prediction. Besides, county-income is also the second important feature for occupation prediction and the third important feature for education prediction. This result indicates that county-income is highly related to people's occupation and education level. This is reasonable because some occupation types earn much more money than others, and the education resources in high-income-level-areas are usually richer than low-income-level-areas.

• **POI is the most important feature for education and occupation predication**. It is also the second important feature of income prediction. POI reflects the function of the living areas. The results demonstrate that the function of one living area is highly related to people's occupation and educational background. For example, we find that people with university degrees are more likely to lives in the areas, where the most important POI categories are related to universities, governments, or high-tech companies.

Table 4 Metrics of top 5 features in each task.

Task	Feature	F1-score	AUC	G-Mean	Accuracy
Personal income	County-income	0.4767	0.7468	0.4149	0.4566
	POI	0.4752	0.7434	0.4131	0.4576
	City-name	0.4787	0.7361	0.4161	0.4171
	Province-name	0.4338	0.7295	0.3794	0.3984
	Average housing price	0.3623	0.6915	0.3653	0.3331
Family income	County-income	0.4609	0.6833	0.4750	0.5031
	POI	0.4445	0.6572	0.4713	0.4930
	City-name	0.4214	0.6393	0.4706	0.4774
	Province-name	0.4237	0.6727	0.4557	0.4376
	County-name	0.3847	0.6082	0.4582	0.3961
Education level	POI	0.5061	0.6647	0.6686	0.5255
	Urban type	0.5224	0.5964	0.7018	0.5058
	County-income	0.4839	0.6441	0.6280	0.4910
	County-name	0.5425	0.4952	0.6863	0.4417
	Average housing price	0.5296	0.4736	0.6924	0.4271
Occupation type	POI	0.4641	0.6839	0.6073	0.5097
	County-income	0.4751	0.6459	0.6324	0.5411
	City-name	0.4803	0.5877	0.6511	0.4796
	Urban type	0.4346	0.6454	0.5771	0.4781
	County-name	0.5057	0.5409	0.6085	0.4344

• **Housing prices are not so effective in SEA prediction tasks**. Housing prices may be one of the most widely used home-based features and are often used as a proxy of people's income in previous works[43]. This is mainly because people usually believe that housing price is highly related to income. However, our study shows the average housing price is only the fifth important feature for personal income and education level prediction. This may be caused by data missing. Besides data missing problems, we also analyze other possible reasons in Section 6.

• **Town-level budget and tax are also not effective**. None of them shows up in the top 5 features. In China, it seems that a higher budget and tax of one area does not necessarily mean a higher income, or higher education level in that area. For example, the town-level budget may change significantly in a short time. For example, in 2016, the town-level budget in Hangzhou city is more than 2 times that in 2015. Then the budget in 2017 is only about 120% of that in 2015. That is because this city held an important G20 Summit in 2016. This indicates that budget data may not be so related to personal SEAs. These statistics are more useful in developed countries[35]. Though they are fine-grained, they are not so related to personal SEAs. This may reflect that the Chinese economy still relies more on investment: a large part of profits go to investors, company owners and governments instead of ordinary workers[44].

• Besides county-income and POI, province/city/county-name and urban types are also important to SEA prediction. This implies there might be a big gap in income, education, and occupation between different areas of China. For example, we can conclude that people tend to have better opportunities to get into university if they are living in cities instead of rural areas (urban-type).

5.4 Difference between city and village

In previous sections, we try to give a unified model for all Chinese people's SEAs based on their home locations. However, previous studies[45,46] point out that

the income level, education level, and occupation types are quite different in China's cities and villages. We think the relationships between home location and the SEAs may be also different between cities and villages. So one unified model may lead to sub-optimal results for people living in cities or villages. In this section, we re-tune the hyper-parameters of our H2SEA model to predict the SEAs of people who live in cities and rural areas, respectively. The city-only or village-only dataset's change ratios (F1-measure) compared with the full dataset (city + village) are shown in Table 5. City-only dataset only consists of urban subsamples while village-only dataset only consists of rural subsamples.

From Table 5, we can see:

• For personal income, the performance of city-only dataset is 2.69% better than that of city+village dataset while the village-only dataset decreases by 7.33%. We re-calculate the feature importance and completeness of features, we find that: (1) housing price becomes the third important feature in the city-only dataset, more important than that in city+village dataset while housing price is only the eighth important feature in the village-only dataset; (2) housing price is missing for most people in villages: 17% of people in the village-only dataset and 57% of that in the city-only dataset have housing price data. So the performance of village-only dataset is worse, because one kind of useful feature (housing price) is much less than that in the city-only dataset.

• For family income, the performance of the city-only dataset is almost the same as that in the city+village dataset, while the village-only dataset increases by 2.23%. The housing price increases to the seventh important feature in the city-only dataset while decreasing the last important feature in the village-only dataset. Compared with personal income, housing price is not so important for family income. So the incompleteness of housing price did not clearly affect the performance of the village-only dataset. Compared with city+village and city datasets, village-name has become more important (third important) for the village-only dataset. The relationships between features and family income are quite different for village-only and city-only datasets.

• For education level, the change ratio is quite similar with Personal Income: the performance of the city-only dataset increases by 7.98% and the village-only dataset decreases by 14.43%. The reason is also similar to Personal Income: housing price is the second important feature while it is missing for most villagers.

• For occupation type, the change ratio is just the opposite with family income: the performance of village-only dataset is almost the same as that of city+village dataset, while the city-only dataset increases by 1.15%. Housing price or the substitute housing price is also not important for occupation type in all datasets. So the performance of village is not affected.

Table 5 Change ratio of F1-measure of people living in city or village.

(%)

Dataset	Personal income	Family income	Education level	Occupation type
City+village	100.00	100.00	100.00	100.00
City-only	102.69	100.24	107.98	101.15
Village-only	92.67	102.23	85.57	100.19

From the above results, we can see the performance of H2SEA is different between cities and villages, especially for personal income and education level. The main reason is the different missing ratio of housing price in cities and villages. Housing price is easy to be collected for people living in cities, while very hard for villagers. Housing price is also an important indicator for urban residents' personal income and education level. So the performance of H2SEA model is much better if we only focus on city.

The importance of housing price is quite low for villagers. However, we do not know it is due to the serious incompleteness of housing prices or the weak housing-price-income relationship in the villages. This will leave to be a future work involving more housing price data in villages. The housing price data collection method via Lianjia.com in this paper is not suitable in villages. The housing prices of most villages in our dataset are not shown on websites like Lianjia.com. We may need to manually collect the housing prices of villages in the future work.

Beside housing price, we also use the substitute housing price as a proxy for people whose housing price cannot be found online. However, this is not working so well for villagers. The first kind of substitute housing price is the nearest collected housing price and the distance should be less than 10 km. For most villagers, they also do not have the first kind of substitute housing price, except for those who live near the border of cities. So most villagers are using the second kind of substitute housing price: city-level average housing price. This is one of the least important features because: (1) all people in one city are sharing the same value; (2) the published average housing price is calculated mainly based on the

housing price data in cities, and quite different from the real housing price in villages; (3) many villagers are living in self-built houses with very low housing prices compared with commercial houses in the cities. So the substitute housing price cannot help in predicting the SEAs of villagers (e.g., farmers and temporary workers who move to cities to work) who do not have real housing prices.

6 Relationship Between Housing Price and Income

Usually, people think that housing price is a very important feature when studying the relationship between home and socioeconomic attributes. Richer people live in high price-level areas and poorer people live in low-price-level areas. Previous studies[12, 17] also show that the housing price has a strong correlation with personal income in Singapore and Shanghai. However, in our case, housing price is not so effective in income prediction. Here we try to give an analysis of possible reasons.

The first reason may be caused by size. Singapore and Shanghai are just two cities, while China is a country with nearly 700 cities (https://en.wikipedia.org/wiki/List_of_cities_in_China, accessed on January 10, 2021) and our dataset covers only 85 cities. The relationship between housing prices and personal income are different in different cities. For example, the correlation coefficient between housing prices and personal income level is 0.37 in Harbin city and 0.05 in Nanjing city. As a whole, the correlation coefficient between housing prices and personal income level over China is only 0.185, much weaker than that over Singapore (0.8[12]) and Shanghai (0.68[17]). Figure 3 shows the relationship between 2800 people's housing prices and income values over China. Though the coefficient is a little higher (0.34), it is still much weaker than that of Singapore and Shanghai. The housing-price-income relationship across many different cities is more complicated than one city, which affects the income prediction accuracy. Please notice that here we only consider the home location of which house prices can be found from websites, and most of these house prices are meant for cities.

The second reason is that the housing-price-income-level relationships in many cities are much weaker than in Shanghai and Singapore. In our dataset, the strongest city-level housing-price-income correlation coefficient

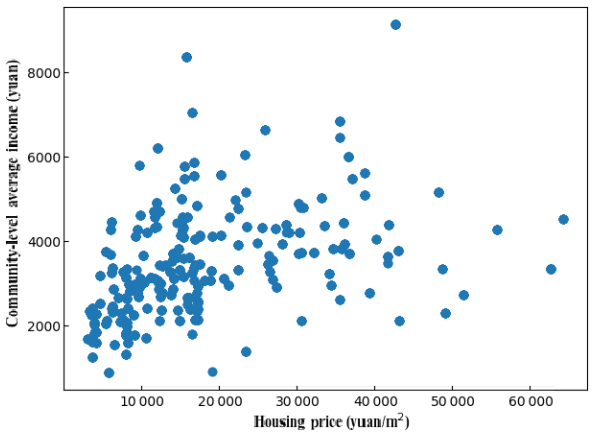


Fig. 3 Relationship between housing price and community-level average personal income in China.

is 0.37 in Harbin City. The others are all below 0.37. Besides the dataset difference from previous works[12, 43], we notice that an interesting phenomenon: *many low-and and middle-income people live in high-price areas*. Previous studies[47–50] show that there are many kinds of houses in China, such as publicly funded housing, housing-reform housing (housing obtained from housing reform), commercial housing, affordable/economic housing, resettlement housings, etc. Detailed analysis of the difference between various kinds of housings is beyond the scope of this work. In this paper, we simply divide them into two groups: commercial housing and the others. For commercial housing, most people need to pay the full prices by themselves. So for people who live in commercial housing, their incomes are highly related to the housing prices. For the other kinds of houses, the governments or the state-owned companies would help to pay at least part of the cost of the houses as welfare or compensation. For example, for some people who worked in many state-owned companies or governments, they can buy the housing-reform houses at a much lower price compared with commercial housing as a welfare[48]; for farmers whose houses are bought and demolished by governments or real estate companies, they may get resettlement housings as compensatio[50, 51]. In short, for people who live in the second kind of houses, their incomes are not so related to the housing prices.

7 Related Work

In this paper, we mainly investigate whether people's home location can be used to infer personal SEAs. Our topic is related to two domains: socioeconomic attributes prediction and multi-task learning.

7.1 Personal socioeconomic attributes prediction

Investigating the possibility of predicting personal SEAs is not only important for researchers working on location privacy protection. On one hand, personal SEA prediction itself can be used to improve personal recommendation, user profiling, and precise marketing. On the other hand, it may serve as a proxy method to collect economic or social statistics in some developing countries[52]. Given its importance, a great number of approaches have been proposed to estimate income level, occupation, and education. Most of them try to predict SEAs from people's cyberspace behavior data, like mobile phone calls[10] and Twitter contents[14].

Taking personal income prediction as an example, the two most widely studied data source types are from Online Social Networks (OSN) and mobile phones (which mainly include call detail records and usage data). As shown in Table 6, quite a few studies are focusing on OSN-based personal income prediction. Note that we also include part of studies which predict personal SES. SES is a combined index calculated based on personal income, work type, and education level[3].

Famous OSN platforms, like Twitter and Facebook, develop fast in recent years. Previous works have established that people's SEAs can be predicted by analyzing their tweets, social links, or profiles recorded by OSN[1, 13–15, 35, 54, 56, 58]. For example, Preoţiuc-Pietro et al.[14] found that higher income users express more fear and anger, whereas lower income users express more of the time emotion and opinions. Volkova et al.[59] extracted lexical features from tweets to predict users' income and education level. Recently, Matz et al.[61] found that Facebook likes and status updates are important to personal income prediction. Abitbol et al.[35] investigated the potential of census, occupation, and housing price in predicting Twitter users' SES.

Another important source data type is mobile phone-related data. Some researchers try to predict people's income based on multiple factors, like communication, the structure of contact network, and mobility pattern. For example, Soto et al.[9] showed that cell phone behavior, social network, and mobility data can be used to identify the socioeconomic levels of a population living in a community. The SES/socioeconomic level information was provided by a national statistical institute, which considers 134 indicators including the level of studies of the number of cell phones, computers, combined income, occupation of the members of the household, etc. Blumenstock et al.[10] and Blumenstock[52] estimated Rwandans and Afghans' personal income by extracting features from mobile phone communication, contact network, and mobility patterns. Sundsøy et al.[55] found that location dynamics, handset brands, and even top-up patterns of mobile phones can also be used in income prediction. Recently, researchers begin to pay more attention to predict SEA based on people's physical behaviors, like retail transaction record[8, 16] and transportation records[17, 63]. Different from these works, we predict multiple sensitive SEAs based on people's home location.

7.2 Multi-Task Learning (MTL)

MTL is a learning paradigm in machine learning. The main purpose of MTL is to take the advantage of useful information shared in multiple tasks to improve the generalization performance of all the tasks[64]. All of these learning tasks are assumed to be related to each other. Considering the cost of data collection, researchers may need to predict multiple users' attributes from one dataset. Therefore some efforts have been put in studying how to apply multi-task learning in user attribute inference[8, 65].

One of the first multi-task models proposed for socioeconomic attribute inference is Structured Neural Embedding (SNE)[8]. SNE uses a simple dense layer to

Table 6 Related works of personal income prediction.

Work	Predicted attribute	Source data	Work	Predicted attribute	Source data
Ref. [35]	SES	Tweets	Ref. [52]	Personal income	Mobile phone metadata
Ref. [14]	Income	Tweets	Ref. [9]	SES	Mobile phone records
Ref. [15]	Income	Tweets	Ref. [53]	Income	Mobile phone call detail records
Ref. [54]	Education and income	Tweets	Ref. [10]	Income	Mobile phone metadata
Ref. [1]	Occupation and income	Tweets	Ref. [55]	Income	Mobile phone metadata
Ref. [56]	Income	Tweets	Ref. [57]	Income and education level	Cookie
Ref. [58]	Education and income	Tweets	Ref. [16]	Income and education level	Retail transaction records
Ref. [59]	Education and income	Tweets	Ref. [8]	Income and education level	Retail transaction records
Ref. [60]	Family income	Tweets	Ref. [17]	SES	Smart card transportation records
Ref. [61]	Income	Facebook likes	Ref. [62]	Education and income	WiFi log

generate initial embedding vectors for all input features. Average pooling is conducted on these vectors and then fed into a linear prediction layer for each SEA estimation task. Different from the conventional multi-task method, SNE ignores the correlation between attributes, since it assumes the correlation is hard to model without explicit knowledge of relationships among tasks. Instead of summation of each task, SNE uses a single structured prediction task to combine all tasks, attempting to reveal the patterns of the correlation among attributes. However, the output space of the SNE is much larger than the conventional multiple task learning method. As a result, SNE is not suitable if the scale of input data sources is limited, or it will lead to overfitting.

Recently, Kim et al.[65] proposed a new multi-task method to predict age, gender, and marital status from people's transaction records. Researchers collect the purchasing histories and user attributes of 56 thousand users. The input data is quite similar to SNE[8]. Compared with SNE, ETNA[65] transforms shared embeddings into task-specific embedding and detects more important signals with an attention mechanism. The results show that the attention mechanism not only increases the performance, but also help to interpret how the customers' attributes related to different items. ETNA simply uses the initial embeddings of items as input, which are also not sufficient for limited input data sources. Different from these works, we propose to utilize second-order feature interactions to improve the performance for limited basic features.

8 Conclusion

This paper tries to examine whether people's multiple socioeconomic attributes (e.g., income and occupation) can be estimated only based on their home location. This study first designs and collects multiple types of SEA-related features, such as housing price, county-level income, and urban types. Then an FM-based multi-task learning method named H2SEA is proposed to model both second-level feature interactions to achieve good prediction accuracy. Based on a dataset collected in 9 provinces of China, our experiment shows that home location and home-based features can clearly improve the performance in predicting people's income, education, and occupation. The H2SEA model also outperforms baseline methods in terms of various metrics like AUC and F1-score.

This paper is the first effort to test the predictive power of home location data on personal SEAs. There are still some open issues that may affect our study. Collecting ground truth and building basic feature datasets cost us a lot of time. We are not able to collect enough data in other countries or less developed areas in China. As a result, some of the conclusions may not hold in other regions. For example, housing prices are not so effective in our experiments, but this may be different in other countries. In the future, we plan to develop a more general model between SEAs and home locations.

Acknowledgment

The research work was partly funded by the European Union's Horizon 2020 Research and Innovation Program under the Marie Sklodowska-Curie (No. 824019), and the Tsinghua-Göttingen Student Exchange Project (No. IDS-SSP-2017001).

References

[1] N. Aletras and B. Chamberlain, Predicting twitter user socioeconomic attributes with network and language information, in *Proc. of the 29th on Hypertext and Social Media*, Baltimore, MD, USA, 2018, pp. 20–24.

[2] R. Bradley and R. Corwyn, Socioeconomic status and child development, *Annual Review of Psychology*, vol. 53, no. 1, pp. 371–399, 2002.

[3] S. Sirin, Socioeconomic status and academic achievement: A meta-analytic review of research, *Review of Educational Research*, vol. 75, no. 3, pp. 417–453, 2005.

[4] T. Szopiński, Factors affecting the adoption of online banking in poland, *Journal of Business Research*, vol. 69, no. 11, pp. 4763–4768, 2016.

[5] D. Chen, D. Jin, T. Goh, N. Li, and L. Wei, Context-awareness-based personalized recommendation of anti-hypertension drugs, *Journal of Medical Systems*, vol. 40, no. 9, p. 202, 2016.

[6] L. Hung, A personalized recommendation system based on product taxonomy for one-to-one marketing online, *Expert Systems with Applications*, vol. 29, no. 2, pp. 383–392, 2005.

[7] Y. Wu, N. Carnt, and F. Stapleton, Contact lens user profile, attitudes and level of compliance to lens care, *Contact Lens and Anterior Eye*, vol. 33, no. 4, pp. 183–188, 2010.

[8] P. Wang, J. Guo, Y. Lan, J. Xu, and X. Cheng, Your cart tells you: Inferring demographic attributes from purchase data, in *Proc. of the 9th ACM Int. Conf. on Web Search and Data Mining*, San Francisco, CA, USA, 2016, pp. 173–182.

[9] V. Soto, V. Frias-Martinez, J. Virseda, and E. Frias-Martinez, Prediction of socioeconomic levels using cell phone records, in *Proc. of Int. Conf. on User Modeling, Adaptation, and Personalization*, Girona, Spain, 2011, pp. 377–388.

[10] J. Blumenstock, G. Cadamuro, and R. On, Predicting poverty and wealth from mobile phone metadata, *Science*, vol. 350, no. 6264, pp. 1073–1076, 2015.

[11] A. Almaatouq, F. Prieto-Castrillo, and A. Pentland, Mobile communication signatures of unemployment, in *Proc. of Int. Conf. on Social Informatics*, Bellevue, WA, USA, 2016, pp. 407–418.

[12] Y. Xu, A. Belyi, I. Bojic, and C. Ratti, Human mobility and socioeconomic status: Analysis of Singapore and Boston, *Computers, Environment and Urban Systems*, vol. 72, pp. 51–67, 2018.

[13] D. Preoţiuc-Pietro, V. Lampos, and N. Aletras, An analysis of the user occupational class through twitter content, in *Proc. of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th Int. Joint Conf. on Natural Language Processing*, Beijing, China, 2015, pp. 1754–1764.

[14] D. Preoţiuc-Pietro, S. Volkova, V. Lampos, Y. Bachrach, and N. Aletras, Studying user income through language, behavior and affect in social media, *PloS One*, vol. 10, no. 9, p. e0138717, 2015.

[15] V. Lampos, N. Aletras, J. Geyti, B. Zou, and I. Cox, Inferring the socioeconomic status of social media users based on behavior and language, in *Proc. of European Conf. on Information Retrieval*, Padua, Italy, 2016, pp. 689–695.

[16] M. Oyamada and S. Nakadai, Relational mixture of experts: Explainable demographics prediction with behavioral data, in *Proc. of 2017 IEEE Int. Conf. on Data Mining* (*ICDM*), New Orleans, LA, USA, 2017, pp. 357–366.

[17] S. C. Ding, H. Huang, T. Zhao, and X. M. Fu, Estimating socioeconomic status via temporal-spatial mobility analysis—A case study of smart card data, in *Proc. of 28th Int. Conf. on Computer Communication and Networks, ICCCN 2019*, Valencia, Spain, 2019, pp. 1–9.

[18] A. Ahmed, Y. Low, M. Aly, V. Josifovski, and A. Smola, Scalable distributed inference of dynamic user interests for behavioral targeting, in *Proc. of the 17th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, San Diego, CA, USA, 2011, pp. 114–122.

[19] M. C. Gonzalez, C. A. Hidalgo, and A. L. Barabasi, Understanding individual human mobility patterns, *Nature*, vol. 453, no. 7196, pp. 779–782, 2008.

[20] C. Huang and D. Wang, Unsupervised interesting places discovery in location-based social sensing, in *Proc. of 2016 Int. Conf. on Distributed Computing in Sensor Systems* (*DCOSS*), Washington, DC, USA, 2016, pp. 67–74.

[21] B. Srilakshmi and K. S. Kumar, An efficient and scalable location-aware recommender system, *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 6, pp. 1384–1399, 2014.

[22] J. Goering, A. Kamely, and T. Richardson, Recent research on racial segregation and poverty concentration in public housing in the united states, *Urban Affairs Review*, vol. 32, no. 5, pp. 723–745, 1997.

[23] P. Bqjari and M. E. Kahn, Estimating housing demand with an application to explaining racial segregation in cities, *Operations Research*, vol. 45, no.4, pp. 419–422, 2005.

[24] B. Qin and Y. L. Jiao, Housing price distribution and urban spatial restructuring in Beijing, (in Chinese), *Economic Geography*, vol. 30, no. 11, pp. 1815–1820, 2010.

[25] J. Chang, Research of spatial distribution and driving mechanism of housing price in Dalian city, (in Chinese), *Journal of Liaoning Normal University*: *Natural Science Edition*, vol. 33, no. 4, pp. 503–506, 2010.

[26] Y. N. Shih, H. C. Li, and B. Qin, Housing price bubbles and inter-provincial spillover: Evidence from China, *Habitat Int.*, vol. 43, no. 4, pp. 142–151, 2014.

[27] X. J. Song, H. Wei, and L. Wang, Research of spatial structure and differentiation pattern of housing price in Xi'an based on esda and geostatistical analysis, (in Chinese), *Science of Surveying and Mapping*, vol. 36, no. 2, pp. 171–174, 2011.

[28] Z. X. Zhao, Q. Xu, S. Peng, and L. Hong, Analyzing spatial-temporal patterns of house price based on network big data in the main city zone of Kunming, in *Proc. of the 2020 Artificial Intelligence and Complex Systems Conf.*, Wuhan, China, 2020, pp. 5–10.

[29] Y. Wang, D. L. Wang, and S. J. Wang, Spatial differentiation patterns and impact factors of housing prices of China's cities, (in Chinese), *Scientia Geographica Sinica*, vol. 10, pp. 1157–1165, 2013.

[30] Y. Wang, Q. Li, S. J. Wang, and J. Qin, Determinants and dynamics of spatial differentiation of housing price in Yangzhou, (in Chinese), *Progress in Geography*, vol. 68, no. 8, pp. 1082–1096 2013.

[31] J. Gao, C. Zhou, and C. Ye, The equitable distribution of public services in Guangzhou, *Planners*, vol. 26, no. 4, pp. 12–18, 2010.

[32] Z. Feng and M. Zhen, The spatial distribution of commodity housing and price in Nanjing based on the spatial analysis, (in Chinese), *Modern Urban Research*, vol. 7, pp. 47–53, 2008.

[33] F. L. Xu, T. Xia, H. C. Cao, Y. Li, F. N. Sun, and F. C. Meng, Detecting popular temporal modes in population-scale unlabelled trajectory data, in *Proc. of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 2, no. 1, pp. 46:1–46:25, 2018.

[34] R. G. Guo, *Practical Sampling Audit*, (in Chinese). Beijing, China: China Audit Press, 1990.

[35] J. L. Abitbol, M. Karsai, and E. Fleury, Location, occupation and semantics based socioeconomic status inference on twitter, in *Proc. of 2018 IEEE Int. Conf. on Data Mining Workshops* (*ICDMW*), Singapore, 2018, pp. 1192–1199.

[36] N. J. Yuan, Y. Zheng, X. Xie, Y. Z. Wang, K. Zheng, and H. Xiong, Discovering urban functional zones using latent activity trajectories, *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 3, pp. 712–725, 2015.

[37] T. Sicular, X. Yue, B. Gustafsson, and S. Li, The urban-rural income gap and inequality in China, *Review of Income and Wealth*, vol. 53, no. 1, pp. 93–126, 2007.

[38] S. Rendle, Factorization machines with libFM, *ACM*

Transactions on Intelligent Systems and Technology (*TIST*), vol. 3, no. 3, pp. 1–22, 2012.

[39] F. Wu, X. H. Yang, A. Packard, and G. Becker, Induced l2-norm control for LPV systems with bounded parameter variation rates, *Int. Journal of Robust and Nonlinear Control*, vol. 6, nos. 9&10, pp. 983–998, 1996.

[40] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, Dropout: A simple way to prevent neural networks from overfitting, *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[41] D. P. Kingma and J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980, 2014.

[42] T. Q. Chen and C. Guestrin, Xgboost: A scalable tree boosting system, in *Proc. of the 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, San Francisco, CA, USA, pp. 785–794.

[43] H. Huang, B. Zhao, H. Zhao, Z. Zhuang, Z. W. Wang, X. M. Yao, X. G. Wang, H. Jin, and X. M. Fu, A cross-platform consumer behavior analysis of large-scale mobile shopping data, in *Proc. of the 2018 World Wide Web Conf. on World Wide Web*, Lyon, France, 2018, pp. 1785–1794.

[44] G. Q. Chen and C. L. Luo, Analysis on factors to impact proportion of China's urban and rural residents' per capita income in GDP—Based on perspective of time and region, (in Chinese), *Journal of Beijing Technology and Business University* (*Social Sciences*), vol. 30, no. 5, pp. 116–126, 2015.

[45] T. Sicular, X. M. Yue, B. Gustafsson, and S. Li, The urban-rural income gap and inequality in China, *Review of Income and Wealth*, vol. 53, no. 1, pp. 93–126, 2010.

[46] X. L. Qian and R. Smyth, Measuring regional inequality of education in China: Widening coast-inland gap or widening rural-urban gap? *Journal of Int. Development*, vol. 20, p. 2, 2010.

[47] A. G. Walder and X. B. He, Public housing into private assets: Wealth creation in urban China, *Social Science Research*, vol. 46, pp. 85–99, 2014.

[48] S. M. Li, China's housing reform and outcomes, *Housing Studies*, vol. 27, no. 8, pp. 1–2, 2012.

[49] Y. P. Wang and A. Murie, Commercial housing development in urban China, *Urban Studies*, vol. 36, no. 9, p. 1475, 1999.

[50] X. Zhang, J. Wang, M. P. Kwan, and Y. W. Chai, Reside nearby, behave apart? Activity-space-based segregation among residents of various types of housing in Beijing, China, *Cities*, vol. 88, pp. 166–180, 2019.

[51] X. T. Cao and P. H. Liao, Discussion on the issues of the resettlement compensation policy for landless peasant under the background of coordinated urban and rural development, (in Chinese), *Journal of Anhui Agricultural Science*, vol. 40, no. 14, pp. 8360, 8361&8363, 2012.

[52] J. Blumenstock, Estimating economic characteristics with phone data, *AEA Papers and Proc.*, vol. 108, pp. 72–76, 2018.

[53] M. Fixman, A. Berenstein, J. Brea, M. Minnoni, M. Travizano, and Carlos Sarraute, A Bayesian approach to income inference in a communication network, in *Proc. of 2016 IEEE/ACM Int. Conf. on Advances in Social Networks Analysis and Mining* (*ASONAM*), San Francisco, CA, USA, 2016, pp. 579–582.

[54] S. Volkova and Y. Bachrach, Inferring perceived demographics from user emotional tone and user-environment emotional contrast, in *Proc. of the 54th Annual Meeting of the Association for Computational Linguistics* (*Volume 1*: *Long Papers*), Berlin, Germany, 2016, pp. 1567–1578.

[55] P. Sundsøy, J. Bjelland, B. A. Reme, A. M. Iqbal, and E. Jahani, Deep learning applied to mobile phone data for individual income classification, in *Proc. of 2016 Int. Conf. on Artificial Intelligence: Technologies and Applications*, Bangkok, Thailand, 2016, pp. 96–100.

[56] M. Hasanuzzaman, S. Kamila, M. Kaur, S. Saha, and A. Ekbal, Temporal orientation of tweets for predicting income of users, in *Proc. of the 55th Annual Meeting of the Association for Computational Linguistics* (*Volume 2*: *Short Papers*), Vancouver, Canada, 2017, pp. 659–665.

[57] P. Atahan, Learning profiles from user interactions, master thesis, The University of Texas at Dallas, Dallas, TX, USA, 2009.

[58] S. Volkova and Y. Bachrach, On predicting sociode mographic traits and emotions from communications in social networks and their implications to online self-disclosure, *Cyberpsychology, Behavior and Social Networking*, vol. 18, no. 12, pp. 726–736, 2015.

[59] S. Volkova, Y. Bachrach, M. Armstrong, and V. Sharma, Inferring latent user properties from texts published in social media, in *Proc. of 29th AAAI Conf. on Artificial Intelligence*, Austin, TX, USA, 2015, pp. 4296–4297.

[60] G. R. Borges, J. M. Almeida, and G. L. Pappa, Inferring user social class in online social networks, in *Proc. of the 8th Workshop on Social Network Mining and Analysis*, New York, NY, USA, 2014, p. 10.

[61] S. C. Matz, J. I. Menges, D. J. Stillwell, and H. A. Schwartz, Predicting individual-level income from facebook profiles, *PloS One*, vol. 14, no. 3, p. e0214369, 2019.

[62] Y. L. Ren, M. Tomko, F. D. Salim, J. Chan, and M. Sanderson, Understanding the predictability of user demographics from cyber-physical-social behaviours in indoor retail spaces, *EPJ Data Science*, vol. 7, no. 1, p. 1, 2018.

[63] Y. D. Zhu, F. Chen, M. Li, and Z. J. Wang, Inferring the economic attributes of urban rail transit passengers based on individual mobility using multisource data, *Sustainability*, vol. 10, no. 11, p. 4178, 2018.

[64] Y. Zhang and Q. Yang, A survey on multi-task learning, arXiv preprint arXiv:1707.08114, 2017.

[65] R. Kim, H. Kim, J. Lee, and J. Kang, Predicting multiple demographic attributes with task specific embedding transformation and attention network, in *Proc. of the 2019 SIAM Int. Conf. on Data Mining*, Calgary, Canada, 2019, pp. 765–773.

Shichang Ding received the PhD degree in computer science from University of Göttingen, Germany in 2020. He is now a lecturer at State Key Laboratory of Mathematical Engineering and Advanced Computing. His research interests include data-driven human behavior analysis, personalized recommendation, and IP geolocation.

Xin Gao received the BS degree from Northeastern University in 2018. She is a PhD candidate at Department of Sociology, Tsinghua University. Her research interests include social network studies and organization behavior.

Yufan Dong received the BS degree from University of Göttingen, Germany in 2020. He is a master student at Department of Computer Engineering, Technical University of Berlin. His research interests include cognitive systems and artificial intelligence.

Yiwei Tong received the MS degree from Department of Sociology at Tsinghua University in 2019. He is working at Shanghai Hejin Information Technology Co., Ltd, which provides service about data science. His research interests include social stratification, educational sociology, and how to improve the efficiency of collaboration in the field of computational social sciences.

Xiaoming Fu received the PhD degree in computer science from Tsinghua University, Beijing, China in 2000. He was then a research staff at Technical University of Berlin until joining the University of Göttingen, Germany in 2002, where he has been a professor in computer science and heading the computer networks group since 2007. He has spent research visits at Cambridge, Columbia, UCLA, Tsinghua University, Uppsala, and UPMC, and is an IEEE senior member and a distinguished lecturer. His research interests include Internet-based systems, applications, and social networks. He is currently an editorial board member of *IEEE Communications Magazine*, *IEEE Transactions on Network and Service Management*, *Elsevier Computer Networks*, and *Computer Communications*, and has published over 150 peer-reviewed papers in renowned journals and international conference proceedings.