Urban Knowledge Graph Aided Mobile User Profiling

YU LIU, ZHILUN ZHOU, YONG LI, and DEPENG JIN, Beijing National Research Center for Information Science and Technology (BNRist), Department of Electronic Engineering, Tsinghua University, China

Nowadays, the explosive growth of personalized web applications and the rapid development of artificial intelligence technology have flourished the recent research on mobile user profiling, i.e., inferring the user profile from mobile behavioral data. Particularly, existing studies mainly follow the data-driven paradigm to develop feature engineering and representation learning on such data, which however suffer from the robustness issue, i.e., generalizing poorly across datasets and profiles without considering semantic knowledge therein. In comparison, the rising knowledge-driven paradigm built upon the knowledge graph (KG) offers a potential solution to mitigate such weakness. Therefore, in this article, we propose a Knowledge Graph aided framework for Mobile User Profiling (KG-MUP). Specifically, to distil semantic knowledge among data, we firstly construct an urban knowledge graph (UrbanKG) with domain entities like users, regions, point of interests (POIs), and so on. identified, as well as semantic relations for home, workplace, spatiality, and so on. extracted. Moreover, we leverage tensor decomposition and graph neural network to obtain knowledgeable user representations from UrbanKG. In addition, we introduce several customized features to quantify individual mobility characteristics for mobile user profiling. Extensive experiments on three real-world mobility datasets demonstrate that KG-MUP achieves state-of-the-art performance on user profile inference tasks. Moreover, further results also reveal the importance of various semantic knowledge to user profile inference, which provides meaningful insights on user modeling with mobile behavioral data.

CCS Concepts: • **Information systems** → **Personalization**; • **Computing methodologies** → **Knowledge representation and reasoning;**

Additional Key Words and Phrases: Mobile user profiling, knowledge graph, representation learning, human mobility

ACM Reference format:
Yu Liu, Zhilun Zhou, Yong Li, and Depeng Jin. 2023. Urban Knowledge Graph Aided Mobile User Profiling. *ACM Trans. Knowl. Discov. Data.* 18, 1, Article 28 (October 2023), 30 pages.
https://doi.org/10.1145/3596604

1 INTRODUCTION

Owing to the booming personalized applications in recent years, the user profile, a summary of the user's demographics, interests and characteristics, has become essential for user experience enhancement and company profit growth [9, 16, 37, 48, 70]. For example, online platforms such as

Y. Liu and Z. Zhou contributed equally to this research.

Authors' address: Y. Liu, Z. Zhou, Y. Li (corresponding author), and D. Jin, Beijing National Research Center for Information Science and Technology (BNRist), Department of Electronic Engineering, Tsinghua University, Beijing, China, 100084; e-mails: liuyu2419@126.com, sz_alan@126.com, liyong07@tsinghua.edu.cn, jindp@tsinghua.edu.cn.

1556-4681/2023/10-ART28 $15.00
https://doi.org/10.1145/3596604

Google and Alibaba benefit from the user's demographics, e.g., gender and age, for personalized recommendations as well as targeted advertisements [7]. To obtain user profiles, early studies are performed in the form of interviews and questionnaires [12], in which case users are reluctant to provide ground truth for privacy concern. In comparison, owing to the diverse behavior data collected from user-centric web services, several studies explore the *user profiling* task [9, 70], to infer the user profile from user-generated data. Especially, the prevalence of personalized mobile applications as well as the accessibility of massive mobile behavioral data have attracted intense research focus on *mobile user profiling* recently [3, 36, 46, 57–59, 63, 71], i.e., inferring the user profile from mobile behavioral data.

Specifically, most existing mobile user profiling studies follow the *data-driven paradigm*, which directly leverage massive data to extract the user profile, including the feature engineering based and representation learning based studies. The feature engineering based studies [3, 29, 62, 67] explicitly define mobility features like the radius of gyration, the mobility entropy, and so on, and train a classifier like XGBoost [6] to classify the user profile, while the representation learning based ones [46, 54, 57–59, 63] utilize various neural networks to implicitly learn low-dimensional vectors for mobile user representations. However, such data-driven studies heavily depend on training data and targeted profiles, leading to the *robustness issue*. For instance, the commonly used radius of gyration characterizes the user's mobility level, which fails in user gender inference [67]. The neural networks with large parameter space are prone to overfit and lack interpretability, which implies poor generalization across datasets [57]. Moreover, such data-driven studies largely ignore the semantic knowledge in user mobile behaviors, e.g., the complex relationships between different users and environment [55].

Meanwhile, several recent studies have introduced the *knowledge-driven paradigm* for other user modeling applications [15, 30, 31, 55, 56, 76], built upon the **knowledge graph** (**KG**) where domain entities and semantic relations form nodes and edges therein, respectively. Specifically, these studies firstly construct specific KGs to extract knowledge from massive data, then employ customized KG embedding techniques [60] for task-specific knowledge representations, such as spatio-temporal KG for mobility prediction [55], urban movement KG for flow prediction [31, 76] as well as KG for user recommendation [15, 30, 56]. Such knowledge-driven studies successfully extract the semantic knowledge like complex relationships in various user behaviors, and thus motivate us to the knowledge-driven paradigm for mobile user profiling, while directly copying such paradigm faces the *incompleteness issue* [38]. For example, the KG mainly focuses on modeling the semantic interconnections between entities, but misses quantitative/numerical information like the number of stay points in user mobility trajectories, which is definitely important and necessary [62]. Moreover, how to construct the specific KG for mobile user profiling is still unclear and underexplored. Consequently, leveraging the knowledge-driven paradigm for mobile user profiling suffers from limitations, and it remains an open problem to be addressed.

Therefore, to fully exploit the advantages of knowledge-driven paradigm with limitations overcome, in this article, we design a Knowledge Graph aided framework for Mobile User Profiling, termed as KG-MUP. The designed framework consists of two main modules, i.e., KG-based module and feature-based module. The KG-based module introduces the KG structure to capture the semantic information in user mobile behaviors and urban environment with the robustness issue solved, while the feature-based module builds upon the feature engineering explored by existing data-driven studies with the incompleteness issue overcome. To be specific, for the KG-based module, we firstly construct the **Urban Knowledge Graph** (**UrbanKG**), and then leverage representative KG embedding techniques for knowledgeable user representations. Specifically, in UrbanKG, both users and environment elements like regions, **points of interest** (**POIs**), commercial areas, and so on. form entities, while their complex relationships like home, workplace, spatiality, attribute, and so on. form relations, which together provide a knowledgeable description of users and environment. Moreover, both tensor decomposition and graph neural network are adopted in UrbanKG embedding for robustness. As for the feature-based module, to quantify the mobility characteristics of users, we extract various features from user mobility trajectories with individual property and environment status considered. Finally, the knowledgeable representations from KG-based module and the mobility features from feature-based module are integrated together to infer the user profile. On one hand, the KG-based module constructs the UrbanKG to extract the semantic knowledge in user mobile behaviors, which potentially mitigates the robustness issue of feature-based one. The identified semantic knowledge stands out as the novel point compared with existing data-driven studies. On the other hand, the feature-based module quantifies several mobility features to capture numerical information in massive data, which also mitigates the incompleteness issue of KG-based one. Overall, both modules are integrated into a general KG-aided framework with better mobile user profiling achieved, as validated by extensive experiments.

To conclude, the main contributions of this article lie in the following three aspects:

- To the best of our knowledge, we are the first to propose a KG-aided framework for mobile user profiling, where KG-based module and feature-based module are designed to capture semantic knowledge and mobility features, respectively. Especially, such general framework also provides novel insights into integrating knowledge-driven paradigm with traditional studies together for better user modeling research.
- We present a systematic investigation of extracting the semantic knowledge from mobile behavioral data for mobile user profiling, which identifies the essential entities of users, regions, POIs, commercial areas, and so on. as well as their complex relationships of home, workplace, spatiality, attribute, and so on, and further constructs the urban knowledge graph.
- We conduct extensive experiments on three datasets for mobile user profiling task of profile inference, and our proposed KG-MUP framework achieves considerable performance across three datasets and five user profiles, which outperforms the best baseline with a precision margin of 2%–13%. Several in-depth studies further reveal the effectiveness and interpretability of our framework design on various user profiles.

The rest of this article is organized as follows. We present preliminaries and problem statement in Section 2, and then introduce the details of framework design in Section 3. We conduct experiments and discuss empirical results in Section 4. After reviewing the related works in Section 5, we summarize our article in Section 6.

2 PRELIMINARIES AND PROBLEM STATEMENT

In this section, we firstly introduce key concepts of mobile behavioral data as well as KG, and then formally define our research problem. Specifically, the mobile behavioral data used in this study mainly refers to the mobility trajectories of users, which are defined as follows.

Definition 2.1 (Mobility Record). A mobility record is defined as a triplet $p = (u, l, \tau)$, which means that user u visits location l at time τ, and $l = (lng, lat)$ is the longitude-latitude pair of the location.

Definition 2.2 (Mobility Trajectory). Given the set of users $\mathcal{U}$, the mobility trajectory of a user $u \in \mathcal{U}$ is defined as a sequence of mobility records $tr^u = \{p_1^u, p_2^u, \ldots, p_n^u\}$, where $p_i^u = (u, l_i^u, \tau_i^u)$ is a mobility record. Besides, the mobility records are in time order, i.e., for $\forall i, j$ with $1 \leq i \leq j \leq n$, we have $\tau_i^u \leq \tau_j^u$. The mobile behavioral data is denoted as $\mathcal{D}_{\text{mobi}} = \{tr^u | u \in \mathcal{U}\}$.

In addition, to capture the interaction between users and the urban environment, we also explore multi-source urban environmental data for mobile user profiling, which is defined as follows.

Definition 2.3 (Urban Environmental Data). The urban environmental data $\mathcal{D}_{\text{urban}}$ mainly covers spatial data $\mathcal{D}_{\text{spatial}}$, attribute data $\mathcal{D}_{\text{attr}}$ and aggregated mobility data $\mathcal{D}_{\text{agg-m}}$, denoted as $\mathcal{D}_{\text{urban}} = \{\mathcal{D}_{\text{spatial}}, \mathcal{D}_{\text{attr}}, \mathcal{D}_{\text{agg-m}}\}$. $\mathcal{D}_{\text{spatial}}$ includes spatial related concepts such as POI, commercial area and road network, while $\mathcal{D}_{\text{attr}}$ records category attributes of POIs and other available attribute information. $\mathcal{D}_{\text{agg-m}}$ is aggregated from individual mobility data, e.g., crowd flow data.

Since the knowledge-driven studies are built upon KG, here we present the definition of KG as follows [17, 18, 60].

Definition 2.4 (Knowledge Graph). A KG is defined as a multi-relational graph $\mathcal{G} = (\mathcal{E}, \mathcal{R}, \mathcal{F})$, where $\mathcal{E}$ is the set of entities, $\mathcal{R}$ is the set of relations and $\mathcal{F}$ is the fact set with $\mathcal{F} = \{(h, r, t) | h, t \in \mathcal{E}, r \in \mathcal{R}\}$. Each triplet $(h, r, t) \in \mathcal{F}$ denotes a directional edge in KG from entity h to entity t with relation type r.

Furthermore, the user profiling task is traditionally defined as follows [9, 12, 70].

Definition 2.5 (User Profiling). The user profiling is to infer user profiles on demographics, interests and characteristics by observable information about users, e.g., users' behavior data.

Based on the concepts above, we formally define our research problem as follows.

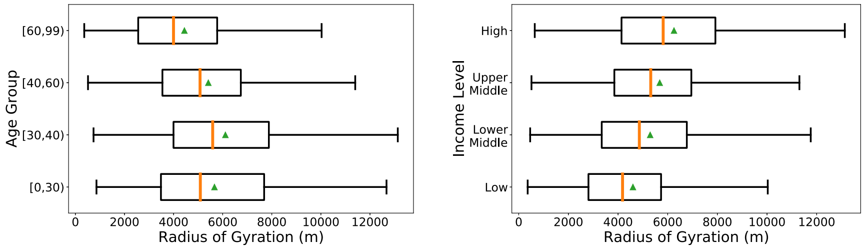
Problem 1 (Knowledge Graph Aided Mobile User Profiling Problem). *Given a set of users $\mathcal{U}$, their mobile behavioral data $\mathcal{D}_{\text{mobi}}$ and urban environmental data $\mathcal{D}_{\text{urban}}$, the knowledge-aided mobile user profiling problem aims at extracting semantic knowledge via KG construction and representation method f_{KG}, i.e., $(\mathcal{G}, \mathcal{X}_{\text{KG}}) = f_{\text{KG}}(\mathcal{U}, \mathcal{D}_{\text{mobi}}, \mathcal{D}_{\text{urban}})$, quantify mobility characteristics via feature extraction method f_{fea}, i.e., $\mathcal{X}_{\text{fea}} = f_{\text{fea}}(\mathcal{U}, \mathcal{D}_{\text{mobi}}, \mathcal{D}_{\text{urban}})$, and finally leverage the both to infer a type of user profile $y_u \in \{1, \ldots, K\}$ (K is the number of classes for the profile.) via integration method f, i.e., $y_u = f(\mathcal{G}, \mathcal{X}_{\text{KG}}, \mathcal{X}_{\text{fea}})$.*

3 METHODOLOGY

3.1 Design Motivation

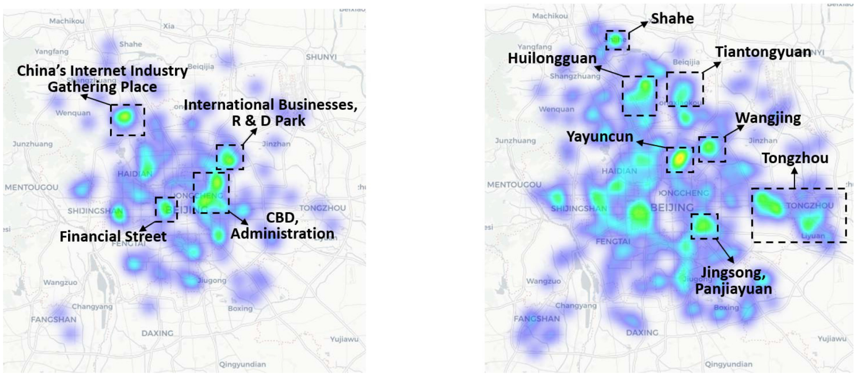
To motivate our proposed KG-MUP framework, here we have an investigation of the real-world dataset, which reveals the strong correlations between mobility data characteristics and user profiles, as well as the one between semantic knowledge and user profiles. The analysis is based on two datasets with user trajectory records and profiles in Beijing collected by a local mobile operator and a social network platform, called Mobile Operator dataset and Social Network dataset, respectively. The dataset details will be described in Section 4 later.

On one hand, the users with different profiles show significantly different mobility characteristics, as validated in previous studies [62, 63, 66] and our data analysis. Specifically, we present the radius of gyration distribution of users in different age groups and income levels in Figure 1(a) and (b), respectively. The radius of gyration characterizes the user's mobility level, and a large value indicates a large mobility range. For example, in Figure 1(a), we find that young people usually have a larger mobility range than older groups, which is in accord with our daily observations that younger groups especially in age 30–40 have large-range movement for commuting and business, while older groups might avoid overmuch activities for health reason. Note that students in school take up a large part of people in age 0–30, and thus drag down the overall radius of gyration for this group. As for the results of users with different income levels in Figure 1(b), we observe a strong correlation between income level and mobility range, i.e., users with higher income level have larger radius of gyration. A possible reason is that the income level has a strong positive correlation with car ownership, and such private transportation further increases the mobility range [8]. Thus, according to above data analysis, mobility data characteristics like the radius of gyration



(a) Age group (b) Income level

Fig. 1. The distribution of radius of gyration across users in (a) Age groups and (b) Income levels in Mobile Operator dataset.



(a) Workplace (b) Home

Fig. 2. The spatial distribution of (a) Workplace and (b) Home for users with the professional occupation in Social Network dataset. The identified regions in (b) are major residential areas in Beijing, China.

are quite useful for mobile user profiling problem, which thus motivates our feature-based module design in KG-MUP framework.

On the other hand, rich semantic knowledge also lies in mobile behavioral data, such as the complex relationships between users and environment, as well as the venue preferences of users with different profiles. In Figure 2, we visualize the spatial distribution of workplace and home for users with professional occupations in areas of Internet, finance, business, education, and so on. In particular, Figure 2(a) indicates that users' workplaces usually imply their profiles especially the occupation, e.g., users who work at Financial Street are probably financial professionals. Moreover, we observe that the spatial distribution of home in Figure 2(b) extends outward compared with that of workplace in Figure 2(a), which agrees with the location theory in urban planning, i.e., residential areas are further from city center than business areas for land price [23]. Besides, such results also reveal the commonsense knowledge that people usually choose residential areas near their workplaces. For example, most people working at China's Internet industry gathering place live in nearby residential areas of Huilongguan[1] and Tiantongyuan.[2] Similar connections can be observed between the working areas of CBD/Administration and the residential areas of Jingsong/Panjiayuan/Tongzhou. Hence, these observations indicate the importance of semantic

[1] https://en.wikipedia.org/wiki/Huilongguan.
[2] https://en.wikipedia.org/wiki/Tiantongyuan.

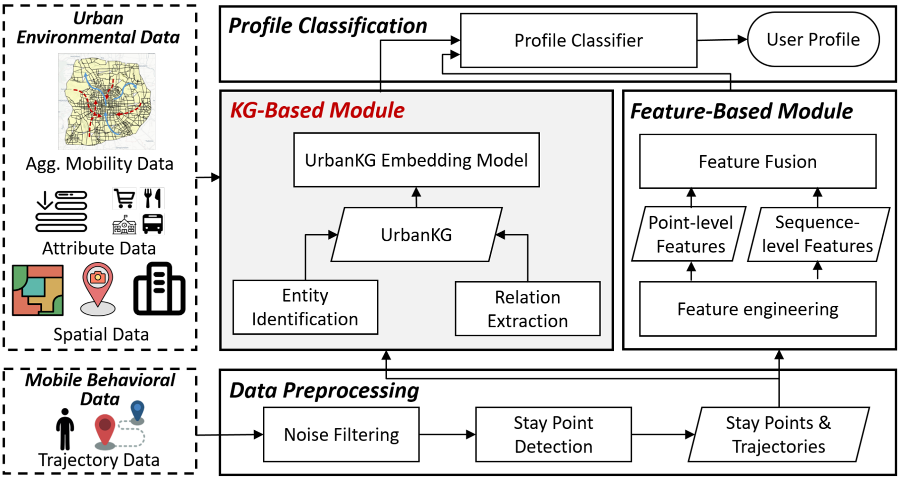


Fig. 3. The architecture of our proposed KG-aided mobile user profiling framework.

knowledge to mobile user profiling problem, which supports our KG-based module design in KG-MUP framework. Furthermore, based on above observations from both feature engineering and semantic knowledge, a promising solution to mobile user profiling problem is to integrate both aspects into our framework design.

3.2 Framework Overview

As illustrated in Figure 3, we propose a framework for mobile user profiling that consists of KG-based and feature-based modules. Specifically, we first take a series of data preprocessing procedures to denoise the mobile behavioral data, i.e., users' trajectories, and extract meaningful stay points therein. Secondly, for KG-based module, we combine urban environmental data with user trajectories and construct an UrbanKG, which incorporates users, key elements in the city as well as various relations between them, to capture urban knowledge relevant to user profiles. The UrbanKG is then fed into a KG embedding model to learn knowledgeable user representations that can preserve urban knowledge therein. Hence, the semantic information among user mobile behaviors and urban environment is fully explored. Thirdly, we adopt feature engineering to calculate point-level and sequence-level features in feature-based module, which aims at characterizing quantitative features of trajectories that are missed in KG-based module. Finally, users' mobility features and knowledgeable representations are integrated together and fed into a traditional classifier for user profile inference.

3.3 Data Preprocessing

3.3.1 Noise Filtering. Due to possible signal blockage, errors of mobile devices and other reasons, noise probably exists in mobility records, which may badly influence mobile user profiling. Thus, we take the following procedures to denoise the trajectory [67]. First, we calculate the user's speed of each mobility record according to contextual records and filter out the records with a speed higher than 120 kilometers per hour. Then, we filter out users with less than 100 mobility records since their trajectories are less representative.

3.3.2 Stay Point Detection. As investigated in previous works [19, 26, 41], the stay point, where the user stays for a period of time, implies mobility semantics. Here, we identify users' stay points from their trajectories. Specifically, in a trajectory, if a sequence of consecutive points lie within a

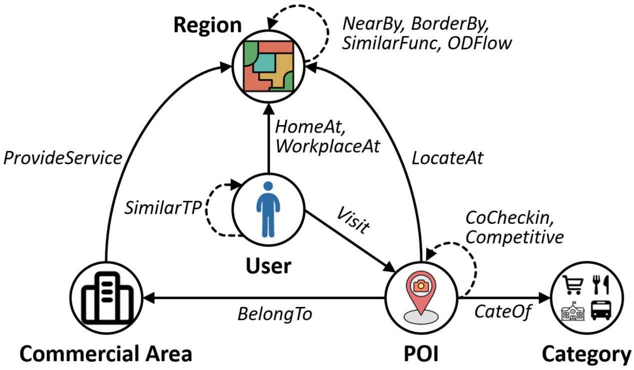


Fig. 4. The schema of urban knowledge graph, where solid edges and dash edges are relations between entities in different types and the same type, respectively.

spatial threshold d_0 and have a stay time longer than a time threshold τ_0, these points are merged into a stay point, whose location is set as the centroid of all these points. In our study, the spatial threshold and temporal threshold are set as $d = 500$ meters and $\tau = 10$ minutes, respectively. Once a stay point is identified, we match it to the nearest POI in the city and define the stay point as the user's visit to the POI.

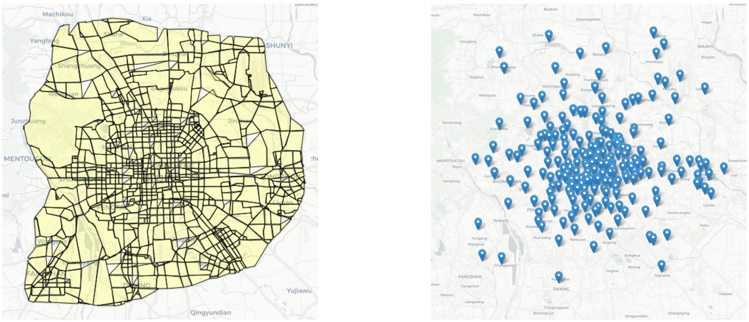
There are three reasons for using stay points instead of raw trajectories. Firstly, single noisy points can be filtered out in this process because a stay point is identified only when a series of trajectory points lie in the same location. Secondly, users' raw trajectories within different datasets may have different spatial and temporal granularities. By merging raw trajectories into stay points with the same spatial and temporal thresholds, we can get trajectories with a rather uniform granularity on different datasets, which helps data processing afterward. Thirdly, in our study, we focus on locations a user visits instead of those the user simply passes by, because only locations where a user stays for a period of time reflect the semantics of trajectory, i.e., the user's intention. For example, a student goes to school from home and passes by a supermarket. In this case, if he intends to visit the supermarket first and spends more than 10 minutes shopping there, then the supermarket will be identified as a stay point. On the contrary, if he does not intend to go shopping and simply passes by the supermarket, then it will not be identified as a stay point.

3.4 KG-based Module for Mobile User Profiling

In this section, we introduce the KG-based module of mobile user profiling, including how we construct the UrbanKG and how we learn knowledgeable user representations from UrbanKG.

3.4.1 UrbanKG Construction. Motivated by the KG application in previous urban computing studies [33, 35, 55, 58, 74], here we construct the UrbanKG from both mobile behavioral data and urban environmental data for urban knowledge. The high-level structure, i.e., schema of UrbanKG is presented in Figure 4. The UrbanKG contains five types of entities including users and key elements in the city, i.e., regions, commercial areas, POIs, and categories. Moreover, various semantic relations between entities are also considered in UrbanKG, including user-to-environment relations, user-to-user relations as well as environment-to-environment relations. Based on the defined schema, the UrbanKG construction process is divided into two parts of entity identification and relation extraction, which are described as follows:

Entity Identification. We first identify several types of entities from mobile behavioral data and urban environmental data, including users, regions, commercial areas, POIs, and categories.



(a) Regions (b) Commercial areas

Fig. 5. The visualization of region entities and commercial area entities within the Sixth Ring Road, Beijing, China.

- **Users.** Users lie in the center of mobile user profiling problem, so it is essential to add users into UrbanKG, and we identify each user as an entity in UrbanKG.
- **Regions.** Regions are identified by dividing the city into small areas according to the main road network, which can represent basic functional areas in cities. We visualize the region entities within the Sixth Ring Road, Beijing in Figure 5(a), from which we can observe that in downtown areas the regions are rather smaller while in suburban areas the regions are larger.
- **Commercial Areas.** Commercial areas are core areas of business and commercial activities such as Sanlitun[3] in Beijing, and we visualize the commercial area entities within the Sixth Ring Road, Beijing in Figure 5(b). Accordingly, the commercial areas distribute mostly in downtown areas, while much sparser in suburban areas.
- **POIs.** POIs represent the basic functional units in a city such as restaurants and parks, and users' activities can be seen as visits to different POIs. As a result, it is necessary to integrate POIs into UrbanKG. Note that the number of POIs in a city is quite large, many of which are not relevant to users, so in this study, we only consider POIs visited by users.
- **Categories.** POIs are classified into various categories, which indicate the function and property. Specifically, we identify three levels of coarse-level, mid-level, and fine-grained categories, e.g., Food Category, Western Food Category, and German Cuisine Category.

Relation Extraction. Now we introduce how to extract the complex relationships between entities. The relations can be classified into three categories of user-to-environment relations, user-to-user relations and environment-to-environment relations, which characterize the relationships between entities from different aspects. Table 1 summarizes the relations and corresponding knowledge in UrbanKG.

- **User-to-Environment Relations.** To link users with city elements that they interact with, we first identify two meaningful types of stay points, i.e., user's home and workplace. Following previous study [19] and daily life patterns in megacities like Beijig and Shanghai in China, the most frequently visited location during weekday nights (between 7pm of the first day and 9am of the next day) is identified as the home of a user. Similarly, the workplace is identified as the most frequently visited location during weekday daytime (between 8am and 9pm). We also restrict that a user's workplace should be at least 1km from the user's home. These two types of stay points represent long-term characteristics of users' mobility

[3] https://en.wikipedia.org/wiki/Sanlitun.

Table 1. The Triplet form and Corresponding Semantic Knowledge Captured in UrbanKG

Triplet Form	**Semantic Knowledge**
(User, *HomeAt*, Region)	The spatial knowledge about users' home.
(User, *WorkplaceAt*, Region)	The spatial knowledge about users' workplace.
(User, *Visit*, POI)	The visiting habits in users' daily life.
(User, *SimilarTP*, User)	The connectivity of users with similar temporal pattern in mobility.
(Region, *NearBy*, Region)	The spatial knowledge about regions within certain distance.
(Region, *BorderBy*, Region)	The spatial knowledge about regions sharing the boundary.
(Region, *SimilarFunc*, Region)	The socioeconomic knowledge about regions with similar functional types.
(Region, *ODFlow*, Region)	The significant population flow transition pattern between regions.
(CA, *ProvideService*, Region)	The service relationships between commercial areas and regions.
(POI, *CateOf*, Category)	The attribute knowledge of POIs.
(POI, *CoCheckin*, POI)	The geographical influence between POIs with concurrent check-ins.
(POI, *Competitive*, POI)	The competitive relationships between POIs.
(POI, *LocateAt*, Region)	The spatial relationships between POIs and regions.
(POI, *BelongTo*, CA)	The spatial relationships between POIs and commercial areas.

Here CA denotes commercial area.

behavior, and they also have explainable semantics, i.e., home and workplace. Therefore, identifying these stay points is of significant importance to mobile user profiling. We then identify the home and workplace of a user to physical regions they are located at, and use relations *HomeAt* and *WorkplaceAt* to link them with the user.

In addition, we draw *Visit* relational links between each user and corresponding visited POIs. By adding these relations, users are connected with city elements that they interact with, which helps distil urban knowledge for mobile user profiling.

- **User-to-User Relations.** To capture the semantic similarity between user mobility patterns, we identify similar users based on **time partition** (**TP**) proposed in [64], and link users with the relation *SimilarTP*. The core idea of this algorithm is to consider only the time allocation patterns of users instead of their physical locations. For example, two users lead the same lifestyle that they both go to work from 9am to 6pm and stay at home from 6pm to 9am the next day. In this case, even if they live in different parts of the city, they are identified as similar users because they have the same time allocation pattern. Hence, the temporal patterns are decoupled from users' locations to discover users with similar lifestyles but totally different physical locations.
- **Environment-to-Environment Relations.** Since the urban environment is strongly correlated with users' daily life, we further extract various kinds of environment-to-environment relations, i.e., relationships between city elements. For example, *BorderBy* and *NearBy* depict the spatial relationship while *SimilarFunc* links regions with similar functions. We also use *ODFlow* relation to link two regions with significant crowd flow transition. Besides, a commercial area is connected to its served regions by *ProvideService* relation. As for POIs, relations between POIs include *CoCheckin* and *Competitive*, which model the geographical influence and competitive relationship between POIs, respectively. Each POI is connected to category by relation *CateOf*. In addition, *LocateAt* and *BelongTo* link the POI with the region and commercial area that it locates at, respectively.

Since the categories are in three levels, we link categories in different levels with the relations of $SubCateOf_{ij}$, where i and j denote category levels. In addition, for each relation, we further introduce the corresponding reverse relation [21, 32–34] in UrbanKG construction.

3.4.2 UrbanKG Embedding Models. In order to obtain user representations that can describe the knowledge in UrbanKG, we leverage KG embedding models for solution. A KG embedding model learns low-dimensional vectors for entities and relations in KG [60]. Specifically, for each triplet (h, r, t), the KG embedding model designs a scoring function ϕ with corresponding vectors of $\boldsymbol{e}_h, \boldsymbol{e}_r, \boldsymbol{e}_t \in \mathbb{R}^d$ to measure the plausibility, so that the true fact obtains a high score $\phi(h, r, t)$. Here d is the embedding dimension. In this study, we leverage tensor decomposition and graph neural network for UrbanKG embedding, which are described as follows:

Tensor Decomposition. A state-of-the-art model TuckER[2] leverages the Tucker decomposition to measure the plausibility of triplets, whose scoring function is designed as

$$\phi(h, r, t) = \mathcal{W} \times_1 \boldsymbol{e}_h \times_2 \boldsymbol{e}_r \times_3 \boldsymbol{e}_t, \tag{1}$$

where $\mathcal{W} \in \mathbb{R}^{d \times d \times d}$ is the core tensor of Tucker decomposition, $\times_n$ represents the tensor product along the nth dimension, and $\boldsymbol{e}_h, \boldsymbol{e}_t \in \mathbb{R}^d$, $\boldsymbol{e}_r \in \mathbb{R}^d$ are the embeddings of head entity h, tail entity t and relation r, respectively.

Graph Neural Network. A recent study R-GCN [43] further considers the structural information of KG, and develops an encoder-decoder framework for KG embedding. For the encoder part, R-GCN applies relational-aware graph convolutional networks to aggregate structural information from entity neighborhood, and then calculate the plausibility score via a bilinear product in the decoder part, which is calculated as

$$\phi(h, r, t) = (\boldsymbol{e}_h^K \odot \boldsymbol{e}_r)^\top \boldsymbol{e}_t^K, \tag{2}$$

where $\boldsymbol{e}_h^K$ and $\boldsymbol{e}_t^K$ are the representations of head entity h and tail entity t after K layers of graph encoding. Specifically, the representation of entity e_i at the $(l+1)$-th layer can be obtained as

$$e_i^{(l+1)} = \sigma\left(\sum_{r \in \mathcal{R}} \sum_{j \in \mathcal{N}_i^r} W_r^{(l)} e_j^{(l)} + W_0^{(l)} e_i^{(l)}\right),$$

where $e_i^{(l)}$ is the representation of entity e_i in the lth graph encoding layer, $\mathcal{N}_i^r$ stands for the set of entities connected to entity e_i via relation r and $W_r^{(l)}$, $W_0^{(l)}$ are learnable weight matrices in the lth graph encoding layer. σ can be a nonlinear activation function.

According to our problem statement in Section 2, to extract semantic knowledge for mobile user profiling, here we design above KG-based module with UrbanKG construction process and embedding models developed, which obtains user representations $\mathcal{X}_{\text{KG}}$, corresponding to the knowledge-driven method f_{KG} .

3.5 Feature-based Module for Mobile User Profiling

To quantify individual mobility characteristics, we also propose a feature-based module for mobile user profiling. Following previous studies [5, 13, 29, 62, 67], we design some features to characterize users' mobility patterns. Specifically, given a user u's stay points $T^u = (l_1^u, \dots l_{N^u}^u)$ with $l_i^u = (lng_i^u, lat_i^u)$ as the longitude-latitude pair, the features can be classified into point-level features and sequence-level features.

Point level features measure the characteristics of stay points, which are defined as follows:

- **Number of stay points,** N^u**.** The total number of stay points of a user u.
- **Number of unique stay points,** n^u**.** The number of unique stay points among the total N^u stay points for user u.
- **Trajectory centroid,** l_C^u**.** The average coordinates of all stay points in the trajectory, which depicts the center of a user u's activity area, and is computed as

$$l_C^u = (lng_C^u, lat_C^u) = \left(\frac{1}{N^u}\sum_{i=1}^{N^u} lng_i^u, \frac{1}{N^u}\sum_{i=1}^{N^u} lat_i^u\right). \tag{3}$$

- **Radius of gyration,** ρ^u**.** This feature measures the deviation of stay points from their centroid l_C^u, depicting the range of trajectory, which is computed as

$$\rho^u = \sqrt{\frac{1}{N}\sum_{i=1}^{N^u} \text{dist}(l_i^u, l_C^u) \cdot \text{dist}(l_i^u, l_C^u)}. \tag{4}$$

  Note that here we use $\text{dist}(\cdot,\cdot)$ to denote the distance calculation between to longitude-latitude pairs.
- **Stay point entropy,** E_s^u**.** This feature measures the diversity of a user u's stay points, which is computed as

$$E_s^u = -\sum_{i=1}^{n} freq\,(l_i^{u\prime}) \log_2 freq\,(l_i^{u\prime})\,, \tag{5}$$

  where $l_1^{u\prime}, l_2^{u\prime}, \ldots, l_n^{u\prime}$ are unique stay points and $freq(l_i^{u\prime})$ is the visiting frequency to $l_i^{u\prime}$. Besides, users that mostly stay in the same location usually have a small stay point entropy.
- **Region entropy,** E_r^u**.** This feature measures the distribution entropy of a user's visited regions. For each stay point, we identify its located region. Let n_r^u be the number of regions user u visits, and n_i^u be the times the user u visits the ith region. Then the region entropy is defined as

$$E_r^u = 1 - \sum_{i=1}^{n_r^u} \left(\frac{n_i^u}{N^u}\right)^2. \tag{6}$$

  A small E_r^u implies that the user u's stay points mostly lie in a small number of regions. Especially, stay point entropy E_s^u and region entropy E_r^u reflect the heterogeneity of mobility trajectories.

Apart from the point-level features above, we further design the following sequence-level features to characterize the travel properties of user mobility.

- **Travel distance,** D^u**.** A travel is defined as a transition part between two consecutive stay points. Thus, the travel distance is defined as the total distance of all travels, i.e., trajectory length, for the user u.

$$D^u = \sum_{i=1}^{N^u-1} D_i^u = \sum_{i=1}^{N^u-1} \text{dist}(l_i^u, l_{i+1}^u). \tag{7}$$

- **Standard deviation of travel distances,** σ^u**.** This feature measures the regularity of a user u's movement, which is computed as

$$\sigma^u = \sqrt{\frac{\sum_{i=1}^{N^u-1}(D_i^u - \bar{D^u})^2}{N^u - 1}}, \tag{8}$$

  where $\bar{D^u} = \frac{\sum_{i=1}^{N^u-1} D_i^u}{N^u-1}$ is the average of all travel distances. A large σ^u means that the travel distance D_i^u varies greatly, which indicates an irregular mobility pattern.
- **Travel entropy,** E_{tr}^u**.** This feature indicates the diversity of travels, which is computed as

$$E_{tr}^u = -\sum_{i=1}^{n_{tr}^u} freq\,(l_i^u \rightarrow l_{i+1}^u) \log_2 freq\,(l_i^u \rightarrow l_{i+1}^u)\,, \tag{9}$$

  where n_{tr}^u is the number of unique origin-destination travels of a user, and $freq(l_i^u \rightarrow l_{i+1}^u)$ is the frequency of travel $l_i^u \rightarrow l_{i+1}^u$ in all N^u-1 travels. The travel entropy E_{tr}^u will be small if a few travels take up the majority of all travels. For example, a user goes to work from home and goes home from workplace everyday, and seldom visits other places, which results in a small travel entropy.

The point-level features and sequence-level features introduced above are able to capture the quantitative characteristics of users' trajectories from different aspects, which correspond to $\mathcal{X}_{\text{fea}}$ with the feature extraction method f_{fea}. So far we have obtained users' knowledgeable representations $\mathcal{X}_{\text{KG}}$ via knowledge-driven method f_{KG} and users' mobility features $\mathcal{X}_{\text{fea}}$ through feature extraction method f_{fea}. After that, we use simple yet effective concatenation to combine $\mathcal{X}_{\text{KG}}$ and $\mathcal{X}_{\text{fea}}$ together and leverage traditional classifier f to infer the user profile. Without loss of generality, we adopt the widely used random forests classifier in this study.

3.6 Framework Optimization

The proposed KG-MUP framework is optimized in a two-step training way. In the first step, the KG-based module f_{KG} is optimized with the objective in traditional KG embedding studies [2, 60], while the feature-based module f_{fea} is developed to extract features from mobile behavioral data, both of which are independent with the user profile data. Then, in the second step, with users' knowledgeable embeddings $\mathcal{X}_{\text{KG}}$ and mobility features $\mathcal{X}_{\text{fea}}$ available, a supervised classifier f like random forests [27] is trained for user profile inference.

Especially, for the KG-based module f_{KG}, with UrbanKG $\mathcal{G} = (\mathcal{E}, \mathcal{R}, \mathcal{F})$ constructed, the UrbanKG embedding model $\phi \in f_{\text{KG}}$ is optimized to correctly measure the plausibility of triplets in $\mathcal{G}$, i.e., the true/positive triplet (h, r, t) should be scored higher than the negative triplets (h, r, t') with $t' \in \mathcal{E} \setminus \{t\}$. Therefore, we adopt the cross-entropy loss for objective function [2, 20], which is formulated as,

$$\min_{\Theta} \sum_{(h,r,t)\in\mathcal{F}} -\log \frac{e^{\phi(h,r,t)}}{\sum_{t'\in\mathcal{E}} e^{\phi(h,r,t')}} + \lambda \cdot \|\Theta\|, \tag{10}$$

where Θ is the learnable parameters in ϕ with $\mathcal{X}_{\text{KG}} \in \Theta$, and λ is the regularization parameter.

Moreover, with the feature-based module f_{fea} designed above, for the user $u \in \mathcal{U}$, we leverage the classifier f to infer the user profile y_u, i.e.,

$$\begin{aligned} y_u = & \mathop{\arg\max}_{k\in\{1,\ldots,K\}} \; f(k, \mathcal{G}, \boldsymbol{x}_u^{\text{KG}}, \boldsymbol{x}_u^{\text{fea}}) \qquad (11) \\ \text{s.t.} \quad & \boldsymbol{x}_u^{\text{KG}} \in \mathcal{X}_{\text{KG}}, \; \boldsymbol{x}_u^{\text{fea}} \in \mathcal{X}_{\text{fea}} \\ & (\mathcal{G}, \mathcal{X}_{\text{KG}}) = f_{\text{KG}}(\mathcal{U}, \mathcal{D}_{\text{mobi}}, \mathcal{D}_{\text{urban}}) \\ & \mathcal{X}_{\text{fea}} = f_{\text{fea}}(\mathcal{U}, \mathcal{D}_{\text{mobi}}, \mathcal{D}_{\text{urban}}), \end{aligned}$$

where K is the number of profile classes, $\boldsymbol{x}_u^{\text{KG}}$ and $\boldsymbol{x}_u^{\text{fea}}$ are the outputs of KG-based module and feature-based module for u, respectively. Hence, Problem 1 is successfully solved by our proposed KG-MUP framework.

We summarize the learning procedure of KG-MUP framework in Algorithm 1. We first use representation learning and feature engineering methods to obtain user embeddings and features. Specifically, for KG-based module in lines 4–12, we construct UrbanKG, and randomly initialize the embeddings of all entities and relations therein, in lines 5–6. Then in lines 7–12, we calculate the loss according to Equation (10) for each triplet in UrbanKG, and update the embeddings of entities and relations based on the gradient of loss. As for feature-based module in lines 13–14, we calculate users' features from mobility trajectories. In lines 15–17, we train a classifier f with user embeddings and features as well as observed user profile data, which is further used for profile inference. According to the learning procedure, both time complexity and space complexity of the algorithm are constrained by the representation learning in Step 1. To be specific, let n_e and n_r

ALGORITHM 1: Learning procedure for KG-MUP framework.

```
Input: The user set $\mathcal{U}$, mobile behavioral data $\mathcal{D}_{\text{mobi}}$ and urban environmental $\mathcal{D}_{\text{urban}}$
Output: The user profile $y_u, \forall u \in \mathcal{U}$.
Step 1: REPRESENTATION LEARNING & FEATURE ENGINEERING
    ▹ KG-based module, $f_{\text{KG}}$
    Construct UrbanKG $\mathcal{G} = (\mathcal{E}, \mathcal{R}, \mathcal{F})$ from $\mathcal{D}_{\text{mobi}}$ and $\mathcal{D}_{\text{urban}}$;
    Initialize entity embeddings $\boldsymbol{E}$ and relation embeddings $\boldsymbol{R}$;
    for $i = 1, 2, \ldots, n_{\text{iter}}$ do
        for $(h, r, t) \in \mathcal{F}$ do
            Compute the score $\phi(h, r, t')$ for all entities $t' \in \mathcal{E}$ according to (1) or (2);
            Compute the loss according to (10);
            Update parameters of embeddings through backward propagation;
    Obtain user embeddding $\boldsymbol{x}_u^{\text{KG}}$ from $\boldsymbol{E}$, $\forall u \in \mathcal{U}$;
    ▹ feature-based module, $f_{\text{fea}}$
    Compute user features $\boldsymbol{x}_u^{\text{fea}}$ from $\mathcal{D}_{\text{mobi}}$ and $\mathcal{D}_{\text{urban}}$, $\forall u \in \mathcal{U}$, according to basic statistics
and (3)–(9) ;
Step 2: PROFILE INFERENCE
    Initialize the classifier $f$, and train $f$ with observed user profile data;
    Infer the user profile $y_u$ with $f, \boldsymbol{x}_u^{\text{KG}}$ and $\boldsymbol{x}_u^{\text{fea}}$, $\forall u \in \mathcal{U}$, according to (11).
```

Table 2. The Basic Information of Three Real-world Mobile Behavioral Datasets

Source	City	Time Duration	# User	#Record	#Stay Point	Spatial	Temporal
Mobile Operator	Beijing	1st July - 31st August, 2017	4,255	3,612,165	264,409	1 m	1 min
Social Network Platform	Beijing	1st October - 31st December, 2019	100,000	57,794,023	1,536,797	425 m	1 s
Life Service Platform	Shanghai	1st July - 31st August, 2021	27,673	1,038,648	765,023	1 m	1 s

The column of “User” denote the number of original users in raw data. The columns of “Spatial” and “Temporal” denote the resolution of raw data in space and time, respectively.

denote the number of entities and relations in UrbanKG, then the time complexity for tensor decomposition model is $O(d)$ with effective tensor product, while the one for graph neural network model is $O(d^2)$ with encoder-decoder framework. As for the space complexity, tensor decomposition model costs $O(n_e d + n_r d + d^3)$ parameters with entity/relation embeddings and core tensor considered, while graph neural network model costs $O(n_e d + n_r d^2)$ with entity embeddings and relation weight matrices considered. Owing to GPU acceleration, KG-MUP is in high efficiency in practical, as shown in experiments later.

4 EXPERIMENTS AND RESULTS

4.1 Datasets

4.1.1 Real-world Datasets. We evaluate our models on three real-world mobile behavioral datasets collected from three data sources across two cities of Beijing and Shanghai in China. Their basic information is listed in Table 2, and other details are as follow.

- **Mobile Operator Dataset.** The dataset is from [29], collected from a local mobile operator, China Mobile,[4] which records the time and location of connected base station whenever

[4] http://www.10086.cn/index/bj/index_100_100.html.

Table 3. User Profile Distribution in Mobile Operator Dataset

Profile	Category
Income	low(26.39%), lower medium(28.30%), upper medium(17.15%), high(28.16%)
Gender	male(34.16%), female(65.84%)
Education	junior high school(8.82%), senior high school(21.95%), undergraduate(58.15%), postgraduate(11.08%)
Age	0-30(13.34%), 30-40(24.84%), 40-60(43.97%), 60-99(17.85%)

Table 4. User Profile Distribution in Social Network Dataset

Profile	Category
Income	very low(2.72%), low(13.58%), medium(53.25%), high(28.75%), very high(1.70%)
Gender	male(62.23%), female(37.77%)
Education	senior high school(23.71%), undergraduate(59.70%), postgraduate(16.58%)
Age	20–25(26.56%), 25–30(24.92%), 30–40(38.60%), 40–50(9.92%)
Occupation	administration support(12.35%), healthcare and technicians(3.94%), managers(0.80%), professionals(61.93%), sales workers(11.59%), services(6.45%), transport and production(2.93%)

Table 5. User Profile Distribution in Life Service Dataset

Profile	Category
Income	low(6.17%), medium(36.05%), high(57.78%)
Gender	male(41.55%), female(58.45%)
Age	0–20(3.14%), 20–25(24.35%), 25–30(30.30%), 30–35(18.21%), 35–40(11.15%), 40–99(12.85%)
Occupation	white collars(76.68%), students(3.41%), others(19.90%)

users access cellular network. In addition, the user profiles are also collected through digital questionnaires from 11 hospitals in Beijing, whose distribution is shown in Table 3.

- **Social Network Dataset.** The dataset is from [63], collected from a social network platform, Tencent,[5] and mobile behaviors are recorded when users invoke location-based services. Besides, this dataset also collects user profile information through large-scale user survey, where the occupations are merged into 7 categories following previous work [42, 63]. The distribution of user profiles is shown in Table 4.
- **Life Service Dataset.** The dataset is collected from a life service platform, Meituan,[6] and mobile behaviors are also recorded when users invoke location-based services. The user profiles are collected through user survey as well as expert rules, whose distribution is shown in Table 5.

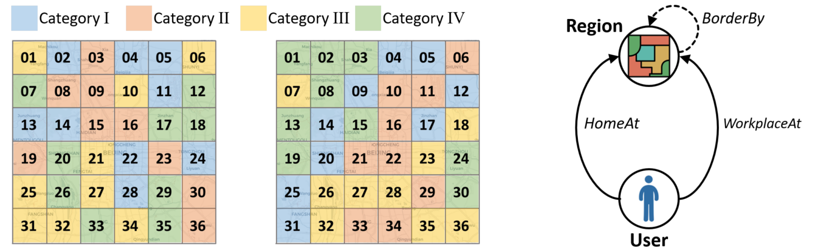
It can be observed that the evaluated profiles vary across three datasets from different sources, which guarantees the robust evaluation. Especially, for Mobile Operator dataset and Social Network dataset in Beijing, we filter mobility records within the Sixth Ring Road in Beijing, which are not far from downtown area. For Life Service dataset, we preserve mobility records in Shanghai. The age profile in Social Network dataset is divided into four groups for group balance, while other profiles in three datasets follow the divisions in their original articles or providers [29, 63].

[5] https://www.tencent.com.
[6] https://www.meituan.com.

Table 6. The Basic Information of the Constructed UrbanKGs on Three Datasets

Dataset	Overall Statistics				Entity Types			
	#Entity	#Relation	#Fact	#User	#Region	#CA	#POI	#Category
Mobile Operator	13,164	19*2	11,097	1,417	1,195	226	10,108	12/37/169
Social Network	15,971	19*2	160,246	8,599	1,195	226	5,744	12/37/158
Life Service	100,701	18*2	659,318	7,733	2,488	265	89,970	12/37/196

Here CA denotes commercial area. #Category denotes the number of coarse-level/mid-level/fine-grained categories for POIs. *2 in #Relation column considers reverse relations in statistics.



(a) Home regions of users (b) Workplace regions of users (c) Schema of simplified UrbanKG

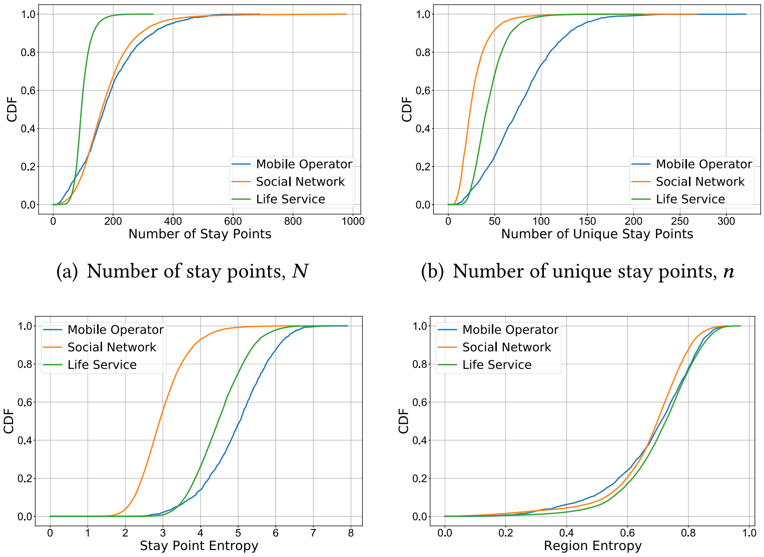
Fig. 6. The description of the synthetic dataset, including (a) Home regions and (b) Workplace regions of users in the dataset and (c) The schema of simplified UrbanKG. The region color corresponds to the profile category of users who live/work there.

Following the KG construction in Section 3.4.1 as well as the schema in Figure 4, we construct three UrbanKGs for corresponding datasets, whose statistics are summarized in Table 6. The POI data used for construction is collected in November, 2018 by Tencent [44], which can be accessed via map service.[7] The mobility records are rather sparse in Life Service dataset, thus we ignore *SimilarTP* for simplicity.

4.1.2 Synthetic Datasets. As described before, the semantic knowledge in mobile user behaviors like the distributions of home and workplace is important for user profile inference. To validate that the KG-based module in KG-MUP framework can capture such knowledge, we further develop two synthetic datasets, i.e., Balanced dataset and Unbalanced dataset. Both datasets contain 1,000 users and each user belongs to one of four profile categories (Category I–IV), where users in Balanced dataset are uniformly divided into each category, while users in Unbalanced dataset are divided into Category I–IV with the ratio of 5%, 15%, 30%, and 50%, respectively. Moreover, we partition a city into 6 × 6 regions, which are uniformly assigned to each category of users as their home regions and workplace regions, as shown in Figure 6(a) and (b). Following the generation process of synthetic datasets, we construct the simplified UrbanKG, whose schema is shown in Figure 6(c).

4.1.3 Privacy and Ethical Considerations. Given the importance and sensitivity of such mobile behavioral data, we enforce the following three protocols for privacy and ethical concerns in this research. First, all datasets are properly anonymized by data owners before sharing with researchers, and real user IDs are never made available to researchers. Second, all researchers that have been authorized to access the datasets are bounded by strict non-disclosure agreements, and the research protocols are approved by the local institutional board. Third, we store all the data in a secure off-line server, which only the authorized core researchers can access.

[7] https://lbs.qq.com/service/webService/webServiceGuide/webServiceSearch.



(a) Number of stay points, N

(b) Number of unique stay points, n

(c) Stay point entropy, E_s

(d) Region entropy, E_r

Fig. 7. The CDF of mobility characteristics on three real-world datasets in respective of (a) Number of stay points, N, (b) Number of unique stay points, n, (c) Stay point entropy, E_s, (d) Region entropy E_r.

On the other hand, inferring user profiles by mobile user profiling models might also face ethical and moral concerns. First, such inferring progress is often defended as being legal, which however is not in line with the society's or users' ethical and moral standings [39], e.g., the users resist such passive inference on personal information. Second, the inferred results might be abused for a number of purposes and further result in disclosure of personal information. Third, the overuse of inferred results by web platforms especially recommender systems can aggravate the echo chamber phenomenon [1], i.e., the users are exposed only to certain types of information. To mitigate such risks associated with using mobile user profiling results, we encourage research to investigate the impacts of using inferred results in particular real-world scenarios, and call for making legal requirement on personal information usage like the General Data Protection Regulation [40].

4.1.4 Basic Statistics. To provide a comprehensive understanding of the datasets, we present the cumulative distribution of some basic features on three datasets in Figure 7. It can be observed in Figure 7(a) that the users' number of stay points in two Beijing datasets, Mobile Operator dataset and Social Network dataset, follow a similar distribution, and most of users have less than 500 stay points. In comparison, users' stay points in Life Service dataset are much less. However, in Figure 7(b), the users in Social Network dataset generally have a smaller number of unique stay points than the other two datasets, which indicates that users in Social Network dataset are likely to visit the same places many times. This phenomenon can be also reflected by stay point entropy shown in Figure 7(c). Specifically, users' stay point entropies in Social Network dataset are generally smaller, which shows that they have a smaller diversity in the place they visited. Nevertheless, in Figure 7(d), the three datasets have a similar distribution in terms of region entropy. Note that regions are partitioned by the main road networks in a city, which represent the basic functional areas of cities. On one hand, the similar region entropy distribution shows that although users come from different datasets or even cities, their mobility behaviors are similar in the diversity of regions they visited. On the other hand, this finding also indicates that the partition of regions by road networks is reasonable and appropriate for user movement modeling.

4.2 Experiment Settings

4.2.1 Baseline. We choose several representative baselines for mobile user profiling to compare with our model.

- Random Guess. This is the simplest method for prediction task. For each user, we randomly choose a profile category as the prediction.
- DPLink [10]. A recurrent neural network based model designed for user identity linkage task. Here we combine DPLink and multilayer perceptron to build an end-to-end model for user profile inference.
- GCN [22]. A classic graph neural network based model that aggregates information from neighbors for each node. Here we apply GCN on the heterogeneous graph proposed in SUME [63], to obtain user embeddings.
- GAT [52]. A graph neural network based model that uses attention modules to aggregate information from different neighbor nodes. Here, GAT adopts the same heterogeneous graph with GCN as input.
- DGI [53] A graph neural network based model that summarizes subgraphs centered around each node in an unsupervised manner for representation learning. Here, DGI follows the same input graph with GCN.
- IMUP [58]. A reinforcement learning model for mobile user profiling, which trains an agent to imitate a mobile user based on the user profiles.
- SUME [63]. It constructs a heterogeneous graph based on similarity between users and locations, and further proposes a network embedding algorithm to obtain user embeddings.

For our proposed KG-MUP framework, we develop two KG embedding models of TuckER [2] and R-GCN [43] in the KG-based module, denoted as KG-MUP/T and KG-MUP/R, respectively. Especially, all the models above except for Random Guess and DPLink, are trained in an unsupervised way to obtain user embeddings/representations, which are further fed into the random forests classifier for user profile inference task. To ensure fair comparison, the embedding dimensions of all these models are set to 32. We tune all baselines following the reported settings in the original study.

4.2.2 Evaluation Metrics. In our experiment, we use Precision, Recall and F1-score as evaluation metrics, which are widely used in mobile user profiling studies [29, 46, 63]. Given a profile with M categories, the metrics are calculated as,

$$Pre = \frac{1}{M}\sum_{m=1}^{M}\frac{TP_m}{TP_m + FP_m}, \quad Rec = \frac{1}{M}\sum_{m=1}^{M}\frac{TP_m}{TP_m + FN_m}, \quad F1 = \frac{1}{M}\sum_{m=1}^{M}\frac{2\cdot TP_m}{2\cdot TP_m + FP_m + FN_m},$$

where TP_m, FP_m, and FN_m stand for true positive, false positive and false negative rates for the m-th category. Without loss of generality, we follow previous studies [29, 63] and adopt the top-k profiles that a user exhibits for metric calculation, setting $k = 2$ for profiles with more than three categories and $k = 1$ otherwise. In addition, we randomly split the datasets into five subsets and report the average performance of 5-folds cross-validation to guarantee the robustness of the results.

4.3 User Profile Inference

We present the overall performance of our model and baselines on three datasets in Tables 7–9, respectively, from which we have the following findings. Firstly, it can be observed from the results that our proposed model outperforms all the baselines on most profiles. For example, on

Table 7. Performance Comparison with Baselines on Mobile Operator Dataset

	Income			**Gender**			**Education**			**Age**		
Model	Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1
Random	0.509	0.514	0.508	0.501	0.500	0.486	0.439	0.497	0.435	0.469	0.492	0.465
DPLink	0.559	0.527	0.528	0.542	0.539	0.538	0.734	0.513	0.532	0.590	0.525	0.513
GCN	0.608	0.596	0.598	0.540	0.543	0.540	0.846	0.549	0.593	0.674	0.544	0.558
GAT	0.587	0.588	0.585	0.548	0.551	0.547	0.714	**0.570**	**0.606**	0.650	0.588	0.598
DGI	0.645	0.646	0.643	0.546	0.547	0.544	0.650	0.557	0.583	0.594	0.602	0.588
IMUP	0.551	0.534	0.535	0.525	0.525	0.524	0.836	0.522	0.571	0.736	0.515	0.533
SUME	0.587	0.582	0.582	0.551	0.543	0.541	0.820	0.547	0.593	0.701	0.620	0.635
KG-MUP/T	0.677	0.663	0.666	0.553	**0.555**	**0.553**	0.723	0.568	0.602	**0.780**	0.607	0.605
KG-MUP/R	**0.687**	**0.667**	**0.673**	**0.569**	0.545	0.534	**0.904**	0.541	0.586	0.709	**0.632**	**0.650**
Improv.	6.5%	3.3%	4.7%	3.3%	0.7%	1.1%	6.9%	-0.4%	-0.7%	6.0%	1.9%	2.4%

Best results are in bold and the second best results are underlined. The last row shows relative improvement in percentage.

Table 8. Performance Comparison with Baselines on Social Network Dataset

	Income			**Gender**			**Education**			**Age**			**Occupation**		
Model	Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1
Random	0.340	0.386	0.300	0.489	0.489	0.481	0.336	0.336	0.310	0.481	0.499	0.477	0.229	0.275	0.203
DPLink	0.509	0.407	0.420	0.518	0.517	0.516	0.360	0.344	0.313	0.631	0.523	0.501	0.413	0.285	0.300
GCN	0.501	0.431	0.446	0.550	0.551	0.550	0.376	0.374	0.374	0.579	0.564	0.569	0.327	**0.313**	0.318
DGI	0.535	0.416	0.438	0.542	0.544	0.542	0.356	0.358	0.356	0.582	0.575	0.576	0.386	0.292	0.319
GAT	0.622	0.430	0.447	0.549	0.551	0.549	0.381	0.388	0.379	0.586	0.570	0.575	0.370	0.305	0.323
IMUP	0.468	0.418	0.433	0.503	0.503	0.503	0.349	0.349	0.349	0.523	0.511	0.514	0.413	0.286	0.316
SUME	0.451	0.432	0.436	0.544	0.542	0.542	0.374	0.386	0.371	0.624	0.558	0.571	0.370	0.300	0.314
KG-MUP/T	0.506	**0.456**	**0.466**	**0.565**	**0.567**	**0.565**	**0.398**	**0.410**	**0.399**	0.630	**0.601**	**0.610**	**0.433**	0.311	**0.336**
KG-MUP/R	**0.629**	0.431	0.442	0.561	0.564	0.560	0.389	0.399	0.390	**0.669**	0.581	0.598	0.429	0.306	0.324
Improv.	1.1%	5.6%	4.3%	2.7%	2.9%	2.6%	4.5%	5.7%	5.3%	6.0%	4.5%	5.9%	4.8%	-0.6%	4.0%

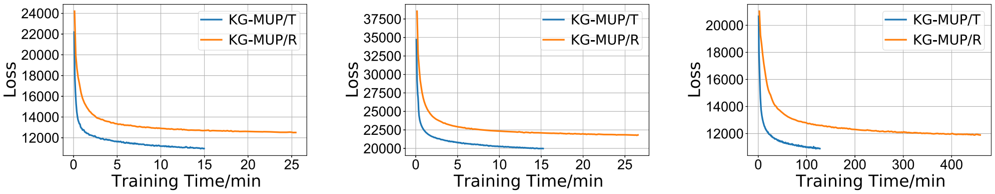
Best results are in bold and the second best results are underlined. The last row shows relative improvement in percentage.

Table 9. Performance Comparison with Baselines on Life Service Dataset

	Income			**Gender**			**Age**			**Occupation**		
Model	Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1
Random	0.328	0.331	0.287	0.494	0.494	0.491	0.318	0.334	0.306	0.330	0.309	0.257
DPLink	0.331	0.351	0.339	0.519	0.518	0.517	0.357	0.345	0.318	0.350	0.340	0.319
GCN	0.381	0.383	0.381	0.530	0.530	0.525	0.368	0.371	0.368	0.346	0.348	0.345
GAT	0.385	0.387	0.383	0.529	0.530	0.526	0.369	0.373	0.367	0.367	0.351	0.352
DGI	0.375	0.378	0.369	0.540	0.540	0.540	0.349	0.352	0.349	0.350	0.354	0.348
IMUP	0.343	0.345	0.343	0.518	0.519	0.518	0.368	0.350	0.350	0.352	0.356	0.352
SUME	0.406	0.414	0.408	0.514	0.514	0.513	0.382	0.388	0.381	0.360	0.357	0.357
KG-MUP/T	**0.426**	**0.438**	**0.429**	0.569	**0.567**	0.566	**0.410**	**0.415**	**0.406**	**0.385**	**0.400**	**0.389**
KG-MUP/R	0.417	0.420	0.416	**0.570**	**0.567**	**0.567**	0.406	0.406	0.400	0.383	0.385	0.382
Improv.	4.9%	5.8%	5.1%	5.6%	5.0%	5.0%	7.3%	7.0%	6.6%	4.9%	12.0%	9.0%

Best results are in bold and the second best results are underlined. The last row shows relative improvement in percentage.

Mobile Operator dataset, our proposed KG-MUP framework improves baselines by 3.3%–6.9% with precision. Such considerable performance gains demonstrate the superiority of knowledge-driven paradigm to capture the user characteristics from different aspects. In addition, the two variants of KG-MUP, i.e., KG-MUP/T and KG-MUP/R achieve close inference performance on most profiles across datasets. Considering the three datasets from different sources and cities, the consistent improvements on all datasets further show the robustness of our model and the effectiveness of UrbanKG to model complex relationships between users and the environment.



(a) Mobile Operator

(b) Social Network

(c) Life Service

Fig. 8. Comparison on clock time of model training v.s. training loss.

Secondly, among baselines, the performance of DPLink model is rather worse since it only utilizes user trajectories without complex environment considered. Meanwhile, we notice that although IMUP model constructs a graph for mobile user profiling, its performance is still worse than other graph based baselines, which is due to the lack of users considered in graph and thus fails to model the relationship between users and the environment. Compared with GCN, GAT and SUME, our proposed KG-MUP framework considers not only more semantic knowledge but also relation-aware representation, which contributes to the relative improvements.

Thirdly, the performance improvement varies across different profiles. Taking Social Network dataset as an example, the improvement on education inference achieves about 5%, while the improvement on gender inference is less than 3%. Note that we use the same user embeddings to infer different user profiles on each dataset. On one hand, this finding shows the robustness of learnt user embeddings in inferring different user profiles. On the other hand, it also suggests that the correlations between mobility patterns and user profiles are different, which affects the performance of inferring profiles. This will be further verified later in Section 4.4. Besides, the prediction performance highly depend on the human behavior characteristics as well as profile semantics captured in corresponding datasets.

Moreover, we also plot the learning curves of KG-MUP/T and KG-MUP/R on three datasets in Figure 8. The results show that both models reach convergence within half an hour on Mobile Operator dataset and Social Network dataset with over ten thousands of entities. Since the number of entities in Life Service dataset is over one hundred, the corresponding convergence time costs about five hours. Therefore, both models achieve nearly linear time complexity with the number of entities in training dataset, which is in accord with theoretical analysis aforementioned.

Overall, our proposed KG-MUP framework achieves both effective and robust performance on various user profile inference tasks across three datasets, and the significant performance gain validates the capability of UrbanKG to capture the urban knowledge in mobile behavioral data for mobile user profiling.

4.4 Effectiveness of KG-based Module

As the core part of our proposed KG-MUP framework design, KG-based module unleashes the power of UrbanKG for mobile user profiling, and thus here we present experiments on two synthetic datasets to verify its effectiveness. Specifically, a direct validation is to examine whether KG-based module can help mobile user profiling under the profile-region knowledge assumption, i.e., users' homes and workplaces are correlated to their profiles.

We first compare the performance of two traditional graph neural network models (GCN, GAT) and two KG embedding models (TuckER, R-GCN). Note that here we only use the KG-based module in KG-MUP. As shown in Table 10, in both balanced and unbalanced datasets, all these models achieve great performance gains over random guess, which shows that the learnt embeddings can help infer user profiles to some degree. What's more, KG embedding models significantly

Table 10. Performance Comparison on Synthetic Datasets

	Balanced			**Unbalanced**		
Model	Pre	Rec	F1	Pre	Rec	F1
Random	0.234	0.237	0.235	0.261	0.268	0.233
GCN	0.555	0.560	0.553	0.668	0.578	0.601
GAT	0.871	0.872	0.869	0.917	0.759	0.793
TuckER	1.000	1.000	1.000	0.994	0.967	0.978
R-GCN	0.994	0.994	0.994	0.986	0.963	0.972

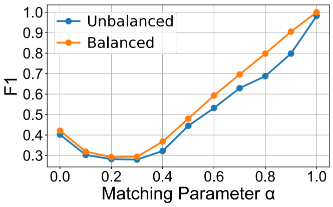


Fig. 9. Performance on synthetic datasets under different matching parameters.

outperform graph neural network models on both datasets, which indicates that taking relation types into consideration is quite important to learn better representations. Note that in real world, users with the same profile may not live or work in the same set of regions. To simulate this case, we define a matching parameter α to measure how much the targeted profiles match with profile-region assumption. Specifically, for each user, its home and workplace have a probability α to lie in one of the candidate regions shown in Figure 6, and a probability $1-\alpha$ to locate in other regions. The results in Table 10 introduced above are actually an ideal case where $\alpha = 1$.

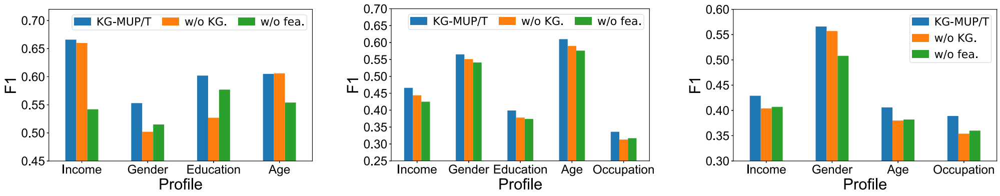
Therefore, we examine the performance of TuckER model under different matching parameters as shown in Figure 9. We observe that on both balanced and unbalanced datasets, as α decreases from 1.0 to about 0.3, the F1-score of profile inference drops greatly to around 0.3, which suggests that although KG can model the semantic knowledge between users and city elements, the performance is greatly affected by how strong the profile-region assumption holds. In addition, the F1-score slightly increases when α decreases from 0.3 to 0, which is because the correlation between user profile and home, workplace actually becomes stronger then. Taking the $\alpha = 0$ case as an example, it is actually equivalent to the case that the home and work regions of each user are selected from the other 27 candidate regions. In summary, KG can perfectly model the semantic knowledge between users and city elements in the ideal case ($\alpha = 1$). As long as there exists a correlation between users' profiles and their home and workplace, KG-based module can learn user representations that help for profile inference, while traditional graph neural network models cannot achieve that good performance without semantic relations considered.

In real-world datasets, some profiles are strongly correlated to the homes and workplaces of users such as income and occupation, which corresponds to a larger matching parameter α, while this correlation on some other profiles such as gender may be weaker with a smaller α. Consequently, the performance gain of our model on different profiles may be quite different, as depicted in Tables 7–9 and result analysis before.

4.5 Ablation Study

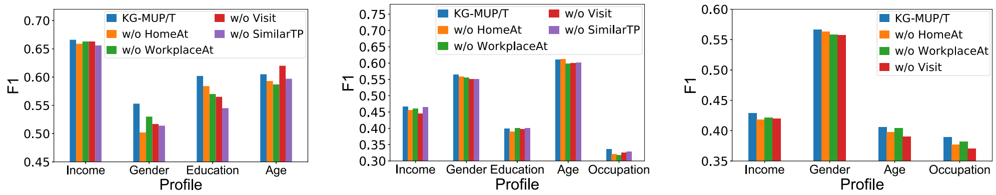
We conduct the ablation study to verify the effectiveness of KG-based and feature-based modules as well as various semantic knowledge in UrbanKG. In Figure 10, we visualize the performance of our framework and two variants that only contains one of the modules, i.e., w/o KG-based module and w/o feature-based module. To avoid redundancy, we only show the F1-score of each profile on three real-world datasets, and similar observations are made on other metrics.

As shown in Figure 10, the performance becomes worse on almost all profiles when omitting KG-based or feature-based modules, which demonstrates the effectiveness of integrating both KG and feature engineering. For example, in Mobile Operator dataset with Figure 10(a), the performance drops without KG-based module are in 0.17%–12.46% across profiles, while the ones without feature-based module are in 4.15%–18.62%. Similarly, the corresponding performance drops in Social Network dataset (Figure 10(b)) and Life Service dataset (Figure 10(c)) are 2.48%–6.85% (w/o



(a) Mobile Operator (b) Social Network (c) Life Service

Fig. 10. Performance comparison of models without KG-based or feature-based module. KG. and fea. represent KG-based module and feature-based module, respectively.



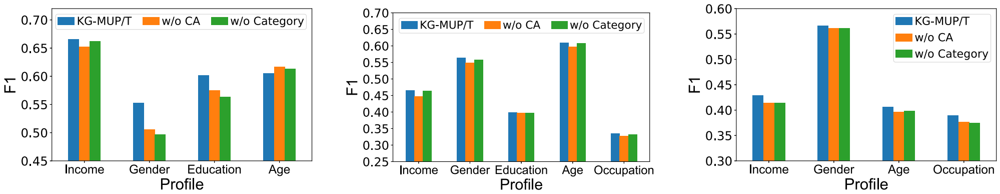
(a) Mobile Operator (b) Social Network (c) Life Service

Fig. 11. Performance comparison of models without a specific relation in UrbanKG.

KG.)/4.25%–8.80% (w/o fea.) and 1.59%–9.00% (w/o KG.)/5.13%–10.25% (w/o fea.), respectively. Besides, we observe that on some profiles, like income and age on Mobile Operator dataset, the performance of feature-based module is better than that of KG-based module, while on some other profiles like gender and education, KG-based module performs better. This further indicates that KG-based and feature-based modules are both essential because they actually depict the characteristics of users from different aspects. For example, in feature-based module, we can easily calculate some quantitative features of a user like the total distance of its trajectory, which is hard to integrate into KG-based module. On the other hand, KG-based module can model the complex relationship between users and city elements but feature-based module cannot. Hence these two modules are actually complementary to each other for better mobile user profiling performance.

Besides, we analyze the effectiveness of different semantic knowledge in UrbanKG by removing one of the relations directly relevant to users, i.e., *HomeAt*, *WorkplaceAt*, *Visit*, and *SimilarTP*. As shown in Figure 11, the performance drops on most profiles on all three datasets when removing each relation, demonstrating the importance of such semantic knowledge to infer user profiles, which furthers validate the effectiveness of UrbanKG construction. Especially, on both Social Network dataset and Life Service dataset, we observe significant performance drops on occupation inference when removing relation *WorkplacAt*, which implies that the workplace is highly correlated to a user's occupation. Moreover, relations show diverse effects on three datasets, e.g., the relation *HomeAt* on Mobile Operator dataset (Figure 11(a)) with performance drop of 1.05%–9.22% across profiles, and the relation *Visit* on Life Service dataset (Figure 11(c)) with performance drop of 1.59%–4.88% across profiles. Also, we notice that the gains by some relations are not significant on certain profile prediction, which might be attributed to the dataset characteristics and profile semantics. Hence, we build the KG-MUP framework as well as UrbanKG for an overall prediction performance on three datasets across multiple profiles.

Furthermore, we remove different types of entities in UrbanKG that are not directly linked to users, i.e., POI categories and commercial areas. As shown in Figure 12, the performance on most profiles reduces when removing these entities, which shows that although these entities are not directly linked to users, they are important in the UrbanKG and can help enrich semantics of users.



(a) Mobile Operator (b) Social Network (c) Life Service

Fig. 12. Performance comparison of models without a certain type of entities in UrbanKG.

Table 11. Gender Profile Inference Performance Comparison of KG-MUP/T Model Trained and Tested on Different Datasets

F1-score		**Testing Dataset**		
		Mobile Operator	Social Network	Life Service
Training Dataset	Mobile Operator	0.553	0.505	0.510
	Social Network	0.399	0.565	0.473
	Life Service	0.521	0.501	0.566

For example, commercial areas are places where people usually do shopping, thus closely related to users' income level. As a result, removing commercial areas affects the performance of income inference on all datasets. In addition, the performance of gender inference on Mobile Operator dataset greatly decreased when removing POI categories, which indicates that the preference for POIs is quite different among males and females.

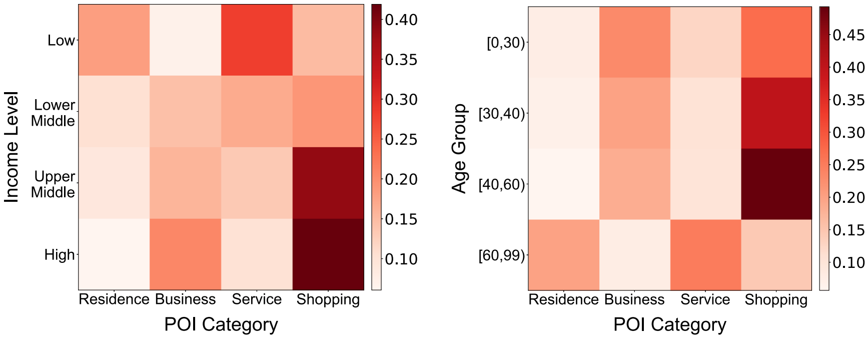
4.6 Generalization Study

The experiment results above validate the effectiveness of KG-MUP framework trained and tested on the same dataset. In this part, we consider a more practical case of out-of-sample scenario, where the proposed KG-MUP/T model is trained on one dataset, and tested on another dataset. Such case can validate the model generalization capability, and Table 11 summarizes the prediction results across three datasets on gender profile inference.

According to the results, the values in diagonal blocks are obtained from models trained and tested on the same dataset, which achieve reasonably best result. Moreover, the results in the first and third rows indicate that models trained on Mobile Operator dataset and Life Service dataset achieve comparable performance on other datasets, which further confirm the model generalization capability. This is because UrbanKG successfully capture the essential relationships between users, which help generalize to user profile inference across datasets. However, models trained on Social Network dataset perform poorly on other datasets, which might be explained by different mobility characteristics between Social Network dataset and others, as shown in Figure 7(b)–(c). Therefore, the proposed KG-MUP framework can generalize to datasets with similar characteristics/distribution, while further modifications are necessary for generalization across dissimilar datasets, which can be a further work.

4.7 Interpretability Study

4.7.1 User Profile & Visiting Pattern. To further investigate the capability of KG-based module to capture the semantics of user trajectories, we analyze the visiting preference of users with different profiles based on their embeddings in UrbanKG. Specifically, we calculate the average user embeddings in UrbanKG with the same category of profile, and then select the top-5 regions based on cosine similarity between region embeddings and user embeddings. After that, we calculate the



(a) Income Level (b) Age Group

Fig. 13. The distribution of POI categories in the top-5 regions related to user profiles. Service represents life services.

POI category distribution in the top-5 regions, and visualize the results on Mobile Operator dataset in Figure 13. Note that the cosine similarity of embeddings measures the correlation between entities in UrbanKG, and POI category distribution reflects the function of a region. Therefore, the values in heatmap actually depict users' visiting preferences.

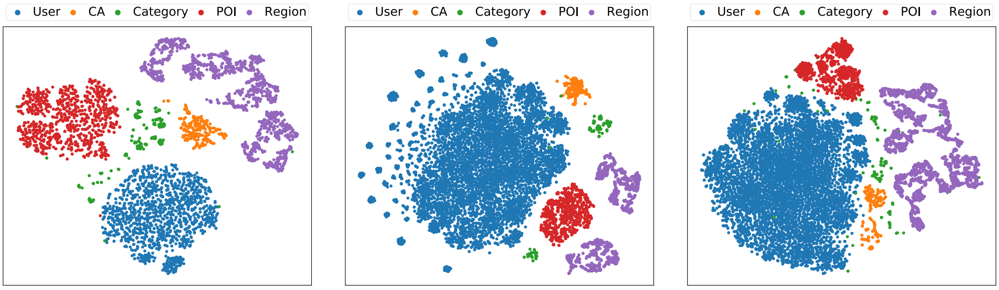
For users with different income levels, it can be observed from Figure 13(a) that users with higher income levels are more likely to visit places related to business and shopping. Besides, users with lower income levels are more closely related to POIs in residence and life service categories, suggesting that they tend to stay at home and travel less, which is also consistent with our findings in Figure 1(b). Furthermore, we can observe from Figure 13(b) that old people are likely to visit POIs in residence and life service categories, and their relationships with business are much weaker than younger ones. This is because many old people have retired and mostly stay at home. In addition, there is a strong correlation between users aging from 40 to 60 and POIs in shopping category, which results from the shopping inclination of most females in the middle age [72]. The findings above are all reasonable with commonsense knowledge, which further demonstrates that the learnt KG embeddings can successfully capture semantics in user mobility trajectories, and provide an understanding of the influence of user profiles to their daily activities in urban environment.

4.7.2 UrbanKG Embedding Visualization. To further examine the capability of our learnt embeddings for mobile user profiling, we use t-SNE [50] to visualize the embeddings of different types of entities on three real-world datasets, as shown in Figure 14. Specifically, we randomly sample 1000 POIs and preserve all entities of other types. It can be obviously observed that there are several clusters in space and entities with the same type mostly cluster together, which demonstrates the learning capability of our proposed KG-based module.

Moreover, we visualize the learnt user embeddings from our KG-MUP framework on Mobile Operator dataset. As shown in Figure 15, users with different profiles are plotted in different colors. Specifically, in Figure 15(a), users aging from 30 to 40 mostly lie in the upper right part of the figure while users over 60 distribute more in the lower left part. Similarly, as shown in Figure 15(b), embeddings of users with low income level are spatially separated from those in high income levels, which indicates the feasibility of KG-aided mobile user profiling, i.e., constructing UrbanKG and leveraging knowledgeable representations for user profile inference.

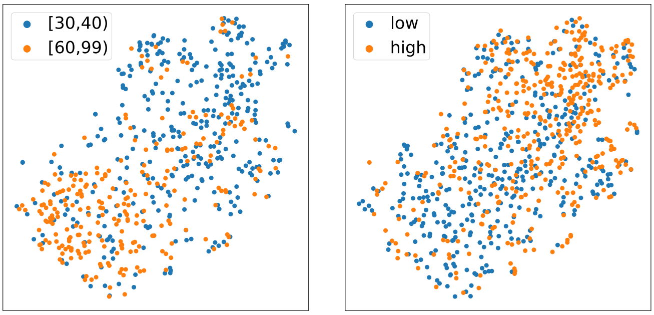
4.8 Parameter Study

As described in Section 3.3, we first adopt the stay point detection algorithm to identify the stay points in user trajectories, which includes spatial threshold and temporal threshold of d and τ,



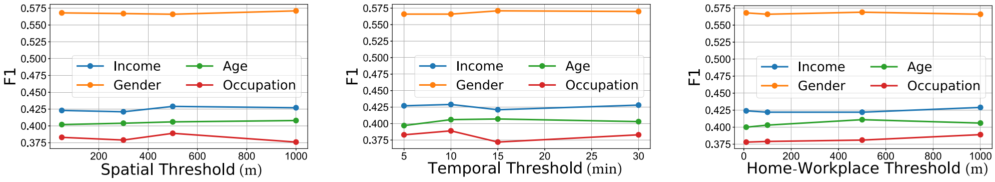
(a) Mobile Operator (b) Social Network (c) Life Service

Fig. 14. Visualization of different types of entities in UrbanKG.



(a) Age (b) Income

Fig. 15. Visualization of learnt user embeddings with different profiles on Mobile Operator dataset.



(a) Spatial Threshold (b) Temporal Threshold (c) Home-Workplace Threshold

Fig. 16. Effects of (a) spatial threshold, (b) temporal threshold, and (c) home-workplace threshold on Life Service dataset across profiles.

respectively. Moreover, in user-to-environment relations extraction, we restrict the distance between user home and workplace to be larger than a threshold of 1 km. Since the whole KG-MUP framework and experiment validation are built upon such data preprocessing, here we investigate the influence of such threshold parameters to KG-MUP/T model on Life Service dataset across profiles, as shown in Figure 16.

According to the curves in Figure 16(a) and (b), the profile inference performances under different spatial thresholds and temporal thresholds are consistent, i.e., the proposed KG-MUP framework is robust to the stay point detection processing. The choice of $d = 500$ meters and $\tau = 10$ min can achieve the best performance across different profiles. As for the home-workplace threshold in Figure 16(c), the results show similar robustness with aforementioned results. Specifically, on occupation inference task, the F1-score drops when we gradually relax the restriction, which indicates that such restriction can help identify workplace more accurately and thus improve occupation inference performance.

5 RELATED WORKS AND DISCUSSION

In this section, we present a detailed review of related works to our study, especially from the two aspects of mobile user profiling as well as KG embedding and application.

5.1 Mobile User Profiling

As described before, user profiling aims to infer the user profile from user-generated data, and our study focuses on mobile user profiling, by applying mobile behavioral data for the task. Especially, previous studies on mobile user profiling can be categorized into explicit modeling with feature engineering [3, 14, 24, 25, 28, 29, 62, 67, 68] and implicit modeling with representation learning [46, 46, 54, 57, 58, 63].

The explicit modeling studies manually define features or rules to characterize mobility characteristics, which are then fed into classifiers for profile inference. For example, a representative work [62] defines spatiotemporal features from user trajectories and semantic features from geographical context to infer user profiles like gender and age, while another work [3] extracts the significant stay points in user trajectories with their semantic descriptions to infer personal information such as religious beliefs and political opinions [3]. Moreover, some studies [14, 28] extract features with geographical and social influence considered to infer user's home location. People's health conditions can also be inferred with several mobility features extracted from trajectories, as investigated in [29, 67, 68]. In addition, other types of mobile behavioral data like vehicle mobility data [25] and travel data [24] are also explored to infer social relationships as well as basic attributes.

In comparison, the implicit modeling studies leverage various neural networks and tensor decomposition techniques to learn latent representations for users, which can be applied for profile inference. In early studies [59, 73], user trajectories are represented in a three-order tensor to indicate users, spatial points and temporal points, where various tensor decomposition techniques are utilized to obtain user latent representations. Furthermore, the Word2vec algorithm as well as recurrent neural network are applied for trajectory encoding and user representation learning [45, 46]. Recently, several studies leverage advanced learning frameworks for mobile user profiling, such as adversarial training [57], reinforcement learning [58], imitation learning [54], and graph embedding [63]. Specifically, StructRL [57] constructs activity graphs to represent user behaviors, on which a deep adversarial substructured learning framework is developed for user representations. IMUP [58] formulates the mobile user profiling problem into a reinforcement learning task where users are modeled as agents and spatial entities like POIs and commercial areas are integrated as a spatial KG, then updates user representations in an incremental learning way. RIRL [54] proposes an imitation-based framework with reinforcement learning, where the agent is trained to imitate users' mobile behaviors in order to achieve optimal mobile user profiling. Moreover, the state-of-the-art model SUME [63] constructs three types of graphs from user trajectories to capture user-user, user-location and location-location interactions, and further proposes a semantic-enhance embedding algorithm to learn user representations. In addition, there are some recent studies of applying deep neural networks for trajectory representations [10, 11], which can be further extended to mobile user profiling.

While many efforts have been made for mobile user profiling, existing studies above still follow the data-driven paradigm, which suffers from the robustness issue. For the explicit modeling studies, the effectiveness of the feature engineering highly depends on the researchers' manually defined features, which is labor-intensive and time-consuming [33]. Moreover, the importance of features significantly varies on different user profiles, and important information like semantic connectivity in mobile behavioral data is missed, leading to inferior performance [58]. As for the implicit modeling studies, neural network based models with tremendous parameters usually overfit the training data and thus generalize poorly across datasets and user profiles [57]. Besides, deep neural networks are often limited by the black box problem, and thus seem unreliable in practical mobile user profiling applications. In contrast, our study introduces KG aided framework, which mitigates the robustness issue, providing both effective and interpretable performance.

5.2 Knowledge Graph Embedding and Application

The KG is usually stored in symbolic triplets, which are hard to manipulate for specific applications. Thus, KG embedding is proposed to represent entities and relations therein into low-dimensional vector spaces, i.e., embeddings, with inherent semantics preserved [17, 18, 60]. Existing studies on KG embedding are categorized into three types of translation distance models [4, 61], tensor decomposition models [2, 21, 49], and neural network models [43, 51, 65, 69, 75]. The translation distance models such as TransE [4] and TransH [61] translate entity embeddings into relation-specific space via various operations like addition and hyperplane, and measure the corresponding triplet plausibility for embedding learning. The tensor decomposition models formulate the KG embedding problem into a tensor decomposition task, with typical Canonical Polyadic (CP) decomposition and Tucker decomposition utilized. For example, ComplEx [49] and SimplE [21] extend CP decomposition for KG embedding with complex-valued embeddings and inverse relation embeddings, respectively, while TuckER [2] extends Tucker decomposition for KG embedding with entity and relation embeddings as factor matrices therein, which achieves competitive performance across benchmarks. Moreover, several studies leverage neural networks especially graph neural networks to capture the structural information for KG embedding. Especially, R-GCN [43] proposes relational graph convolutional networks to the multi-relational structure of KG, while both CompGCN [51] and KE-GCN [65] design relation-specific message passing mechanisms for both semantic and structural information encoding. Besides, a recent study NBFNet [75] generalizes Bellman-Ford algorithm to graph neural networks for KG embedding. Since our proposed KG-MUP framework is generally based on UrbanKG, various KG embedding models above can be flexibly applied for knowledgeable representations, and we choose two representative models of TuckER [2] and R-GCN [43] for effective and efficient performance.

Owing to powerful KG embedding techniques as well as semantic knowledge therein, KG has been widely applied in various downstream applications, including recommender system [15, 30, 56], mobility prediction [55], flow prediction [31, 76], traffic prediction [47] and site selection [33]. Specifically, KGNN-LS [56] investigates the utility of item KG for graph neural network based recommendation, while UKGC [30] and KnowSite [33] construct an urban KG for location recommendation and site selection, respectively. Furthermore, several studies extract the movement knowledge from mobile behavioral data for spatio-temporal applications. For example, STKG [55] constructs a spatio-temporal KG and extends ComplEx [49] to predict user trajectory with the KG. Both KGPMF [76] and RFP-KMN [31] transform flow transitions into relations of KG, which are further utilized to predict unobserved flow patterns. Besides, a recent study [47] benefits from the interpretability of KG to understanding urban traffic like road congestion and traffic relationship discovery. Hence, these studies motivate our study to firstly introduce the UrbanKG as well as the knowledge-driven paradigm for mobile user profiling.

6 CONCLUSION AND FUTURE WORK

In this article, we propose KG-MUP, a KG aided framework for mobile user profiling problem. To capture the semantic knowledge in mobile user behaviors, we identify the essential entities and complex relationships with UrbanKG constructed, and further develop KG embedding models for knowledgeable user representations in a knowledge-driven paradigm. Moreover, we introduce representative features to quantify mobility characteristics for individual mobility behaviors. Furthermore, both knowledgeable representations and quantified features are cohesively fused for user profile inference. Extensive experiments demonstrate that KG-MUP outperforms the state-of-the-art baselines with both effectiveness and interpretability achieved.

As for future work, one promising direction is to embed quantified features into UrbanKG construction so as to solve the mobile user profiling problem directly built upon the KG. We further consider to distil more semantic knowledge from multi-source user-generated data for more generalized UrbanKG construction and corresponding downstream applications.

REFERENCES

[1] Eytan Bakshy, Solomon Messing, and Lada A. Adamic. 2015. Exposure to ideologically diverse news and opinion on Facebook. *Science* 348, 6239 (2015), 1130–1132.

[2] Ivana Balazevic, Carl Allen, and Timothy Hospedales. 2019. TuckER: Tensor factorization for knowledge graph completion. In *EMNLP*. 5188–5197.

[3] Benjamin Baron and Mirco Musolesi. 2020. Where you go matters: A study on the privacy implications of continuous location tracking. *UbiComp* 4, 4 (2020), 1–32.

[4] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems.* 2 (2013), 2787–2795.

[5] Luca Canzian and Mirco Musolesi. 2015. Trajectories of depression: Unobtrusive monitoring of depressive states by means of smartphone mobility traces analysis. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 1293–1304.

[6] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 785–794.

[7] Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, Rohan Anil, Zakaria Haque, Lichan Hong, Vihan Jain, Xiaobing Liu, and Hemal Shah. 2016. Wide & deep learning for recommender systems. In *Proceedings of the 1st Workshop on Deep Learning for Recommender Systems.* 7–10.

[8] Yi-Shih Chung and Chi-Hung Wu. 2020. Cross-state comparison of income and vehicle-ownership effects on household mobility. *Transportation Research Record* 2674, 11 (2020), 813–826.

[9] Christopher Ifeanyi Eke, Azah Anir Norman, Liyana Shuib, and Henry Friday Nweke. 2019. A survey of user profiling: State-of-the-art, challenges, and solutions. *IEEE Access* 7 (2019), 144907–144924.

[10] Jie Feng, Mingyang Zhang, Huandong Wang, Zeyu Yang, Chao Zhang, Yong Li, and Depeng Jin. 2019. Dplink: User identity linkage via deep neural network from heterogeneous mobility data. In *The World Wide Web Conference.* 459–469.

[11] Qiang Gao, Fan Zhou, Kunpeng Zhang, Goce Trajcevski, Xucheng Luo, and Fengli Zhang. 2017. Identifying human mobility via trajectory embeddings. In *IJCAI* 17, (2017), 1689–1695.

[12] Daniela Godoy and Analia Amandi. 2005. User profiling in personal information agents: A survey. *The Knowledge Engineering Review* 20, 4 (2005), 329–361.

[13] Marta C. Gonzalez, Cesar A. Hidalgo, and Albert-Laszlo Barabasi. 2008. Understanding individual human mobility patterns. *Nature* 453, 7196 (2008), 779–782.

[14] Yulong Gu, Yuan Yao, Weidong Liu, and Jiaxing Song. 2016. We know where you are: Home location identification in location-based social networks. In *2016 25th International Conference on Computer Communication and Networks.* 1–9.

[15] Qingyu Guo, Fuzhen Zhuang, Chuan Qin, Hengshu Zhu, Xing Xie, Hui Xiong, and Qing He. 2020. A survey on knowledge graph-based recommender systems. *IEEE Transactions on Knowledge and Data Engineering* 34, 8 (2020), 3549–3568.

[16] Al-Shamri Mohammad Yahya H. 2016. User profiling approaches for demographic recommender systems. *Knowledge-Based Systems* 100 (2016), 175–187.

[17] Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia D'Amato, Gerard de Melo, Claudio Gutierrez, José Emilio Labra Gayo, Sabrina Kirrane, Sebastian Neumaier, Axel Polleres, Roberto Navigli, Axel-Cyrille Ngonga Ngomo, Sabbir M. Rashid, Anisa Rula, Lukas Schmelzeisen, Juan Sequeda, Steffen Staab, and Antoine Zimmermann. 2021. Knowledge graphs. *Computing Surveys* 54, 4 (2021), 1–37.

[18] Shaoxiong Ji, Shirui Pan, Erik Cambria, Pekka Marttinen, and S. Yu Philip. 2021. A survey on knowledge graphs: Representation, acquisition and applications. *IEEE Transactions on Neural Networks and Learning Systems* 33, 2 (2021), 494–514.

[19] Shan Jiang, Yingxiang Yang, Siddharth Gupta, Daniele Veneziano, Shounak Athavale, and Marta C. González. 2016. The TimeGeo modeling framework for urban mobility without travel surveys. *Proceedings of the National Academy of Sciences* 113, 37 (2016), E5370–E5378.

[20] Rudolf Kadlec, Ondrej Bajgar, and Jan Kleindienst. 2017. Knowledge base completion: Baselines strike back. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*. 69–74. arXiv:1705.10744. Retrieved from https://arxiv.org/abs/1705.10744.

[21] Seyed Mehran Kazemi and David Poole. 2018. SimplE embedding for link prediction in knowledge graphs. In *Advances in Neural Information Processing Systems* 31 (2018), 4284–4295.

[22] Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*.

[23] Mariano Kulish, Anthony Richards, and Christian Gillitzer. 2012. Urban structure and housing prices: Some evidence from australian cities. *Economic Record* 88, 282 (2012), 303–322.

[24] Can Li, Lei Bai, Wei Liu, Lina Yao, and S. Travis Waller. 2021. Urban mobility analytics: A deep spatial-temporal product neural network for traveler attributes inference. *Transportation Research Part C: Emerging Technologies* 124 (2021), 102921.

[25] Jie Li, Fanzi Zeng, Zhu Xiao, Hongbo Jiang, Zhirun Zheng, Wenping Liu, and Ju Ren. 2020. Drive2friends: Inferring social relationships from individual vehicle mobility data. *IEEE Internet of Things Journal* 7, 6 (2020), 5116–5127.

[26] Quannan Li, Yu Zheng, Xing Xie, Yukun Chen, Wenyu Liu, and Wei-Ying Ma. 2008. Mining user similarity based on location history. In *Proceedings of the 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. 1–10.

[27] Andy Liaw, Matthew Wiener, et al. 2002. Classification and regression by RandomForest. *R news* 2, 3 (2002), 18–22.

[28] Ilaria Liccardi, Alfie Abdul-Rahman, and Min Chen. 2016. I know where you live: Inferring details of people's lives by visualizing publicly shared location data. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 1–12.

[29] Zongyu Lin, Shiqing Lyu, Hancheng Cao, Fengli Xu, Yuqiong Wei, Hanan Samet, and Yong Li. 2020. HealthWalks: Sensing fine-grained individual health condition via mobility data. *UbiComp* 4, 4 (2020), 1–26.

[30] Chang Liu, Chen Gao, Depeng Jin, and Yong Li. 2021. Improving location recommendation with urban knowledge graph. arXiv:2111.01013. Retrieved from https://arxiv.org/abs/2111.01013.

[31] Jia Liu, Tianrui Li, Shenggong Ji, Peng Xie, Shengdong Du, Fei Teng, and Junbo Zhang. 2021. Urban flow pattern mining based on multi-source heterogeneous data fusion and knowledge graph embedding. *IEEE Transactions on Knowledge and Data Engineering* 35, 2 (2021), 2133–2146.

[32] Yu Liu, Jingtao Ding, Yanjie Fu, and Yong Li. 2023. UrbanKG: An urban knowledge graph system. *ACM Transactions on Intelligent Systems and Technology* 14, 4 (2023), 1–25.

[33] Yu Liu, Jingtao Ding, and Yong Li. 2021. Knowledge-driven site selection via urban knowledge graph. arXiv:2111.00787. Retrieved from https://arxiv.org/abs/2111.00787.

[34] Yu Liu, Jingtao Ding, and Yong Li. 2022. Developing knowledge graph based system for urban computing. In *Proceedings of the SIGSPATIAL International Workshop on Geospatial Knowledge Graphs*. 3–7.

[35] Yu Liu, Xin Zhang, Jingtao Ding, Yanxin Xi, and Yong Li. 2023. Knowledge-infused contrastive learning for urban imagery-based socioeconomic prediction. In *Proceedings of the ACM Web Conference 2023*. 4150–4160.

[36] Eric Malmi and Ingmar Weber. 2016. You are what apps you use: Demographic prediction based on user's apps. In *Proceedings of the International AAAI Conference on Web and Social Media*. 635–638.

[37] Stuart E. Middleton, Nigel R. Shadbolt, and David C. De Roure. 2004. Ontological user profiling in recommender systems. *ACM Transactions on Informaiton Systems* 22, 1 (2004), 54–88.

[38] John Mylopoulos. 1980. An overview of knowledge representation. *ACM SIGART Bulletin* 74 (1980), 5–12.

[39] Harshvardhan J. Pandit and Dave Lewis. 2018. Ease and ethics of user profiling in black mirror. In *Companion Proceedings of the The Web Conference*. 1577–1583.

[40] European Parliament and Council of the European Union. 2016. Regulation on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/EC (data protection directive). *Official Journal of the European Union* 119 (2016), 1–88.

[41] Sijie Ruan, Zi Xiong, Cheng Long, Yiheng Chen, Jie Bao, Tianfu He, Ruiyuan Li, Shengnan Wu, Zhongyuan Jiang, and Yu Zheng. 2020. Doing in one go: Delivery time inference based on couriers' trajectories. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2813–2821.

[42] Sunhee Sang, Morton O'Kelly, and Mei-Po Kwan. 2011. Examining commuting patterns: Results from a journey-to-work model disaggregated by gender and occupation. *Urban Studies* 48, 5 (2011), 891–909.

[43] Michael Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. 2018. Modeling relational data with graph convolutional networks. In *ESWC* (2018), 593–607.

[44] Erzhuo Shao, Huandong Wang, Jie Feng, Tong Xia, Hedong Yang, Lu Geng, Depeng Jin, and Yong Li. 2021. DeepFlowGen: Intention-aware fine grained crowd flow generation via deep neural networks. *IEEE Transactions on Knowledge and Data Engineering* 34, 12 (2021), 5693–5707.

[45] Adir Solomon, Ariel Bar, Chen Yanai, Bracha Shapira, and Lior Rokach. 2018. Predict demographic information using Word2vec on spatial trajectories. In *Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization*. 331–339.

[46] Yiwei Song, Dongzhe Jiang, Yunhuai Liu, Zhou Qin, Chang Tan, and Desheng Zhang. 2021. HERMAS: A human mobility embedding framework with large-scale cellular signaling data. *UbiComp* 5, 3 (2021), 1–21.

[47] Jiyuan Tan, Qianqian Qiu, Weiwei Guo, and Tingshuai Li. 2021. Research on the construction of a knowledge graph and knowledge reasoning model in the field of urban traffic. *Sustainability* 13, 6 (2021), 3191.

[48] Jie Tang, Limin Yao, Duo Zhang, and Jing Zhang. 2010. A combination approach to web user profiling. *ACM Transactions on Knowledge Discovery from Data* 5, 1 (2010), 1–44.

[49] Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. 2016. Complex embeddings for simple link prediction. In *International Conference on Machine Learning* 48 (2016), 2071–2080.

[50] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research* 9, 11 (2008), 2579–2605.

[51] Shikhar Vashishth, Soumya Sanyal, Vikram Nitin, and Partha Talukdar. 2019. Composition-based multi-relational graph convolutional networks. In *International Conference on Learning Representations*.

[52] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph attention networks. In *International Conference on Learning Representations*.

[53] Petar Velickovic, William Fedus, William L. Hamilton, Pietro Liò, Yoshua Bengio, and R. Devon Hjelm. 2019. Deep graph infomax. *ICLR* 2, 3 (2019), 4.

[54] Dongjie Wang, Pengyang Wang, Kunpeng Liu, Yuanchun Zhou, Charles E. Hughes, and Yanjie Fu. 2021. Reinforced imitative graph representation learning for mobile user profiling: An adversarial training perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 4410–4417.

[55] Huandong Wang, Qiaohong Yu, Yu Liu, Depeng Jin, and Yong Li. 2021. Spatio-temporal urban knowledge graph enabled mobility prediction. *UbiComp* 5, 4 (2021), 1–24.

[56] Hongwei Wang, Fuzheng Zhang, Mengdi Zhang, Jure Leskovec, Miao Zhao, Wenjie Li, and Zhongyuan Wang. 2019. Knowledge-aware graph neural networks with label smoothness regularization for recommender systems. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 968–977.

[57] Pengyang Wang, Yanjie Fu, Hui Xiong, and Xiaolin Li. 2019. Adversarial substructured representation learning for mobile user profiling. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 130–138.

[58] Pengyang Wang, Kunpeng Liu, Lu Jiang, Xiaolin Li, and Yanjie Fu. 2020. Incremental mobile user profiling: Reinforcement learning with spatial knowledge graph for modeling event streams. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 853–861.

[59] Pinghui Wang, Feiyang Sun, Di Wang, Jing Tao, Xiaohong Guan, and Albert Bifet. 2017. Inferring demographics and social networks of mobile device users on campus from AP-trajectories. In *Proceedings of the 26th International Conference on World Wide Web Companion*. 139–147.

[60] Quan Wang, Zhendong Mao, Bin Wang, and Li Guo. 2017. Knowledge graph embedding: A survey of approaches and applications. *IEEE Transactions on Knowledge and Data Engineering* 29, 12 (2017), 2724–2743.

[61] Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. 2014. Knowledge graph embedding by translating on hyperplanes. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

[62] Lun Wu, Liu Yang, Zhou Huang, Yaoli Wang, Yanwei Chai, Xia Peng, and Yu Liu. 2019. Inferring demographics from human trajectories and geographical context. *Computers, Environment and Urban Systems* 77 (2019), 101368.

[63] Fengli Xu, Zongyu Lin, Tong Xia, Diansheng Guo, and Yong Li. 2020. SUME: Semantic-enhanced urban mobility network embedding for user demographic inference. *UbiComp* 4, 3 (2020), 1–25.

[64] Fengli Xu, Tong Xia, Hancheng Cao, Yong Li, Funing Sun, and Fanchao Meng. 2018. Detecting popular temporal modes in population-scale unlabelled trajectory data. *UbiComp* 2, 1 (2018), 1–25.

[65] Donghan Yu, Yiming Yang, Ruohong Zhang, and Yuexin Wu. 2021. Knowledge embedding based graph convolutional network. In *Proceedings of the Web Conference 2021*. 1619–1628.

[66] Yanwei Yu, Hongjian Wang, and Zhenhui Li. 2018. Inferring mobility relationship via graph embedding. *UbiComp* 2, 3 (2018), 1–21.

[67] Yunke Zhang, Fengli Xu, Tong Li, Vassilis Kostakos, Pan Hui, and Yong Li. 2021. Passive health monitoring using large scale mobility data. *UbiComp* 5, 1 (2021), 1–23.

[68] Yunke Zhang, Fengli Xu, Tong Xia, and Yong Li. 2021. Quantifying the causal effect of individual mobility on health status in urban space. *UbiComp* 5, 4 (2021), 1–30.

[69] Zhanqiu Zhang, Jie Wang, Jieping Ye, and Feng Wu. 2022. Rethinking graph convolutional networks in knowledge graph completion. In *Proceedings of the ACM Web Conference 2022*. 798–807.

[70] Sha Zhao, Shijian Li, Julian Ramos, Zhiling Luo, Ziwen Jiang, Anind K. Dey, and Gang Pan. 2019. User profiling from their use of smartphone applications: A survey. *Pervasive and Mobile Computing* 59 (2019), 101052.

[71] Sha Zhao, Feng Xu, Zhiling Luo, Shijian Li, and Gang Pan. 2018. Demographic attributes prediction through app usage behaviors on smartphones. In *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*. 870–877.

[72] Sha Zhao, Yizhi Xu, Xiaojuan Ma, Ziwen Jiang, Zhiling Luo, Shijian Li, Laurence Tianruo Yang, Anind Dey, and Gang Pan. 2019. Gender profiling from a single snapshot of apps installed on a smartphone: An empirical study. *IEEE Transactions on Industrial Informatics* 16, 2 (2019), 1330–1342.

[73] Yuan Zhong, Nicholas Jing Yuan, Wen Zhong, Fuzheng Zhang, and Xing Xie. 2015. You are where you go: Inferring demographic attributes from location check-ins. In *Proceedings of the 8th ACM International Conference on Web Search and Data Mining*. 295–304.

[74] Zhilun Zhou, Yu Liu, Jingtao Ding, Depeng Jin, and Yong Li. 2023. Hierarchical knowledge graph learning enabled socioeconomic indicator prediction in location-based social network. In *Proceedings of the ACM Web Conference 2023*. 122–132.

[75] Zhaocheng Zhu, Zuobai Zhang, Louis-Pascal Xhonneux, and Jian Tang. 2021. Neural bellman-ford networks: A general graph neural network framework for link prediction. *NeurIPS* 34 (2021), 29476–29490.

[76] Chenyi Zhuang, Nicholas Jing Yuan, Ruihua Song, Xing Xie, and Qiang Ma. 2017. Understanding people lifestyles: Construction of urban movement knowledge graph from GPS trajectory. In *Proceedings of the IJCAI*. 3616–3623.

Received 30 May 2022; revised 1 February 2023; accepted 3 May 2023