IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 17, 2020, accepted August 7, 2020, date of publication August 24, 2020, date of current version September 3, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3018997

Inferring Location Types With Geo-Social-Temporal Pattern Mining

TARIQUE ANWAR[1], (Member, IEEE), KEWEN LIAO[2,4], ANGELIC GOYAL[3], TIMOS SELLIS[4], A. S. M. KAYES[5], AND HAIFENG SHEN[2]

[1]Department of Computing, Macquarie University, Sydney, NSW 2109, Australia
[2]Peter Faber Business School, Australian Catholic University, Sydney, NSW 2060, Australia
[3]Department of Computer Science and Engineering, Indian Institute of Technology Ropar, Punjab 140001, India
[4]Swinburne's Data Science Research Institute, Swinburne University of Technology, Melbourne, VIC 3122, Australia
[5]Department of Computer Science and Information Technology, La Trobe University, Melbourne, VIC 3086, Australia

Corresponding author: Tarique Anwar (tarique.anwar@mq.edu.au)

This work was supported in part by ISIRD Research Grant from IIT Ropar.

ABSTRACT With a rapid growth in the global population, the modern world is undergoing a rapid expansion of residential areas, especially in urban centres. This continuously demands for increased general services and basic amenities, which are required according to the kind of population associated with the places. The advent of location-based online social networks (LBSNs) has made it much easier to collect voluminous data about users in different locations or spatial regions. The problem of mining location types from the LBSN data is largely unexplored. In this paper, we propose a pattern mining approach, using the geo-social-temporal data collected from LBSNs, to infer types of different locations. The proposed method first mines frequent co-located users and user components from an LBSN and then performs a temporal pattern analysis to finally categorize the locations. Extensive experiments are conducted on two real datasets that demonstrate the efficacy of the proposed method in terms of mean reciprocal rank (MRR), visualisations, and insights. The resulting inference mechanism would be very useful in several application domains including urban planning, billboard placement, tour planning, and geo-social event planning.

INDEX TERMS Location based social networks, spatial data mining, co-located friendships, geo-social-temporal patterns.

I. INTRODUCTION

The modern world is going through an expansion in both urban and rural areas. While the rural areas are growing at a relatively slower pace, there is a rapid growth of our cities, horizontally as well as vertically. This is continuously and consistently raising the demands for general services and basic amenities. With the development of wireless communication technologies and ubiquitous GPS-equipped mobile devices, the online social networking (OSN) sites rapidly took a new form, called location-based social networks (LBSNs). These social networks allow the registered users to share their location along with the performed activity, referred as ''check-in'' (e.g., visiting Taj Mahal, eating at a local restaurant), and discuss on them as part of their online social interactions. Some popular LBSNs are Foursquare, Facebook, Twitter, Weibo, BrightKite, and Gowalla. In recent years, LBSNs have been quite successful in attracting a large portion of online users. Meanwhile, an enormous amount of the combined geo-social-temporal data is being generated everyday from user activities. This brings us a huge potential of solving a range of crucial problems of the growing society. These data provide opportunities of research in three main aspects of human mobility: *geographic movement* - the places we visit; *temporal dynamics* - periodicity constraints in our movement; and *social network* - evolution of offline and online relationships. An analysis of all these aspects together leads to the discovery of various interesting structural patterns, subject to geographical and social constraints, therefore enhancing the knowledge discovery process from the view of data miners. Location type inference based on LBSN data is one of such research problems with a significant impact. Important applications of automatic location inference were seen in urban planning [1], sophisticated tourism [2], real estate management [3], and geo-social event planning [4]. With ever-expanding urban areas, it becomes difficult to manually identify and organize many different types of regions of interest (ROIs) and points of interest (POIs). Consider

The associate editor coordinating the review of this manuscript and approving it for publication was Hocine Cherifi.

the example of a plan to set up a new entity, such as a hospital, a train station, or a small government office, in a city. Automatic inference about the type of a potential venue and surrounding locations would help in making a careful decision.

In the past decade, there has been significant research interests on mining and analyzing socio-spatio-temporal patterns from LBSN data [5]. A great deal of research work has been recently devoted on profiling users or mining user social behaviours based on their mobility patterns [6] (for example, friend recommendation based on user check-in patterns). However, another important direction of profiling locations based on social relationships has not received much attention. In particular, the problem of mining location types from visitor social relationships on an LBSN is largely unexplored, despite its necessity. Some works also exist along the line of mining different functional zones in an urban area based on user movement trajectories. This is a very different problem that is limited to an urban context. In contrast, the problem of inferring the type of a location (e.g., workplace or residential area) based on geo-social data is important in a global context in order to aid intelligent decision making where location type plays a crucial role. Decision contexts include smart urban planning, strategic development of tourism, real estate management, geo-social event planning, and business development. There are three major challenges in this problem. The first challenge is to model the relationship between the social network connections and spatial check-ins. The second challenge is to characterise the spatial, social, and temporal patterns individually as well as combined altogether. The last challenge is to process the large voluminous data of millions of records and identify the patterns. It needs to be done in an intelligent and efficient manner.

In this paper, we study the problem of inferring location types from LBSNs and invent a step-by-step geo-social-temporal pattern mining approach as the inference mechanism. The method starts with mining the spatial patterns in the form of frequent co-located users, and then this is followed by mining the geo-social patterns in the form of frequent co-located friendship components. The resulting component patterns help in determining whether a location is public or private. The patterns are then expanded based on a solid temporal analysis to form geo-social-temporal patterns. These patterns in the end decide specific types of locations of interest. In summary, our work makes the following main contributions:

- A two-step geo-social pattern mining method is developed to compute frequent co-located friendship components.
- A sophisticated temporal analysis is then performed as the final step of an overall geo-social-temporal pattern mining method. The complete method ultimately infers specific location types of interest from LBSN.
- Extensive experiments are performed on two real datasets. The obtained results are convincing, which validates the efficacy of the proposed method.

The rest of the paper is organized as follows. Section II presents a basic background and the problem definition, which is followed by a geo-social pattern mining method in Section III. Temporal analysis of the mined geo-social patterns is provided in Section IV. In Section V, extensive experimental results are presented, before the surveyed related works in Section VI. Finally, Section VII concludes the whole paper with a concise summary and future direction.

II. PRELIMINARIES

This section gives a brief background of LBSN and its formal definition. This is followed by the problem formulation of location type inference.

A. LOCATION-BASED SOCIAL NETWORKS

The existing social networks like Instagram, Flickr, Twitter, all have a common feature of geo-tagging locations by the registered users. In these location based social networks (LBSNs), the social interactions are depicted by online network structures, and the location-based geographical activities are represented as check-in records, which consist of sequences of data points with latitude-longitude records, time stamps, and venue information. Due to the pervasive mobility of users that leads to their ubiquitous social interactions, a huge amount of user-generated geo-social data is rapidly generated and accumulated. Such big geo-social data not only collectively represent the diverse kinds of real-world human activities, but also serve as a handy resource for various geo-social applications.

For simulation of the proposed solution - the data from `Brightkite` and `Gowalla` are used, which have been active and popular LBSN sites in the past. In these sites, registered users could share their location through check-in, and could also see the other nearby users and those who have checked-in at that place in the past. Along with check-ins, online friendship data among users is also available. This data allows studying the three main aspects of human mobility: geographic movement - the places we visit; temporal dynamics - periodicity constraints in our movement; and the social network - evolution of offline and online friendships. All these aspects when analyzed together exhibit various interesting structural patterns subject to geographical and social constraints, therefore enhancing the knowledge discovery process from the view of data miners. In the following, we formally define LBSNs.

Definition 1: *(Social Network):* A social network is defined as a graph $N = (U, R)$, where U is the set of users (represented as nodes), and R is the set of relationships or connections between the users (represented as edges between the nodes). If two users $u, v \in U$ are related or connected in the social network, then there exists an edge $r_{uv} \in R$ in N. ■

Definition 2: *(Location):* A location l is defined as a geographic place on earth marked by its geographic coordinates, (latitude, longitude) $= (l.lat, l.lon)$. ■

TABLE 1. A sample of data from `Brightkite` dataset.

Location ID (LID)	Time stamp	Frequent co-located users (User IDs)
4	12	1168, 1697, 0, 167, 969, 875, 12, 43
103	7	1168, 1009, 0, 167, 969, 875, 12, 43
56	3	0, 1009, 1697, 167, 43, 12
74	11	0, 1697, 167, 969, 875, 12, 43, 1168, 1009
23	1	0, 167, 969, 875, 12, 43, 1168, 1009, 86
256	8	0, 1697, 167, 969, 875, 12, 43, 1009, 86

Definition 3: *(Check-In)* A check-in c is defined as an explicit record of a location $c.l$ visited by a user $c.u$ at time $c.t$. ■

Definition 4: *(Location-Based Social Network):* A location based social network (aka LBSN) is defined as a graph $\mathcal{L} = (U, R, C)$, where U is the set of users (represented as nodes), R is the set of relationships or connections between the users (represented as edges between the nodes), and C is the set of all check-ins logged by the users in U. ■

B. PROBLEM STATEMENT

We consider few selected major types of locations, defined in Definition 5, in our problem. These locations types are broadly of either public space or private space in nature.

Definition 5: *(Location Type):* The type of a location l is defined to be one of the following: *i*) Public - Education/Workspace, *ii*) Public - Marketplace, *iii*) Public - Recreation spot, *iv*) Public - All-time operational, *v*) Private - Workspace, and *vi*) Private - Residence. This complete set of types is denoted by $\mathcal{T}$. ■

Definition 6: *(Location Type Inference):* Given an LBSN $\mathcal{L}$, a location l, and a set of location types $\mathcal{T}$, the problem of location type inference is to identify the type of l as one of the types in $\mathcal{T}$ on the basis of $\mathcal{L}$. ■

Our aim is to identify the different types of regions based on the analysis of the patterns found in the location based data which consists of (latitude, longitude) coordinates and the time stamps at which the users checked-in. The users check-in at different places and so do their friends. If multiple different and non-related group of friends are present at a location in the same time-period, it can be intuitively concluded that the location and its surrounding region is a public area; further public areas can be classified into different types based on active time periods like the location which is 24 hours active can be a hospital complex or a multipurpose building.

III. GEO-SOCIAL PATTERN MINING

The section presents a two-step geo-social pattern mining method to compute frequent co-located friendship components. With the mined user components information, location types can be initially classified as either public or private.

A. DATA PREPROCESSING

Location based dataset provides two types of information - i) details about the location where users checked in, and ii) user social friendship/graph data. The check-in data actually contains information not being used. Therefore data cleaning is applied first. Initially, each tuple contains: *node id* (users who checked in), *time stamp* (time and date of check-in), *check-in latitude and longitude*, and *check-in location id*. The processed data instead consists two dictionaries D_u and D_t: D_u contains the location coordinates as key, and the array of user ids who checked in at the corresponding location as value; D_t contains the location coordinates as key, and the array check-in times as value. Further, time and location coordinates are indexed according to rounded values to consider them as ranges. Specifically, minutes and seconds are truncated so time slots per day is reduced to 24 hours. Similarly the location coordinates (longitude, latitude) are rounded to deliberately group close-by users.

Example 1: Table 1 below shows a sample representation of the processed data. The first column contains the *key* shared in both D_t and D_u, and the remaining columns contain their respective values. ■

B. IDENTIFYING FREQUENT CO-LOCATED USERS

To conduct frequent pattern mining, the following definitions are introduced:

Definition 7: *(Co-Located Users):* Two users u_i and u_j are said to be co-located, if both u_i and u_j have checked-in at the same location in the same time range at least once. ■

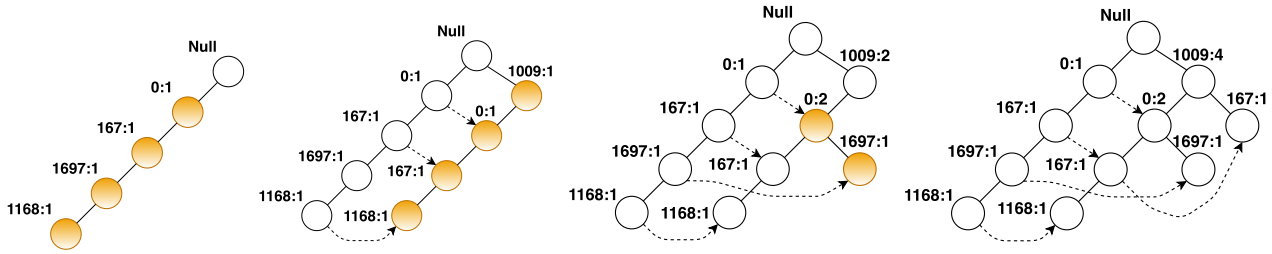
Definition 8: *(Co-Location Support):* Co-location support $colsup(u_i, u_j)$ between two users u_i and u_j is defined as the count of locations checked-in by both u_i and u_j in the same time range. ■

Definition 9: *(Frequent Co-Located Users):* Two users u_i and u_j are said to be frequent co-located, if the co-location support $colsup(u_i, u_j)$ is greater than or equal to a predefined minimum support threshold *minsup*. ■

The co-located users refer to the users who are checking in at the same location and in the same time range, and the basic parameter of minimum support is used to define the degree by how frequently the users are co-located.

Example 2: Consider Table 1. The support count of the set of user-ids (1697, 969, 875) is 3, as they appear together 3 times. ■

In this step, we mine the set of all the frequent co-located users F from the constructed dictionary D_u. Apriori algorithm can be applied for this task, but it requires $n + 1$ scans of the set of locations L, where n is the length of the longest pattern. Instead, our mining approach is developed based on the ideas of FP-growth algorithm [7]. It requires only two passes of



(a) After reading LID = 1 (b) After reading LID = 2 (c) After reading LID = 3 (d) After reading LID = 5

FIGURE 1. FP-Tree construction from the sample data given in Table 2: (a) shows the FP-Tree constructed after traversal of LID = 1, similarly (b) and (c) show the FP-Tree after LID=2 and LID=3, and (d) shows the final FP Tree constructed.

TABLE 2. A sample data of Location IDs (LID) and their associated User IDs.

LID	User IDs
1	1168, 1697, 0, 167
2	1168, 1009, 0, 167
3	1009, 0, 1697
4	1009, 969, 875
5	1009, 167, 12

the location database which is much faster. Following the divide and conquer approach, it first compresses the data in the form of an index structure called FP-Tree and then divides the indexed data into a set of conditional patterns. Each of the conditional patterns are mined for the frequent co-located users recursively.

FP-Tree Construction: Check-in data are indexed by an FP-Tree first. The tree represents the co-located users in a compressed manner. Its construction starts with a scan of all the locations in the dictionary D_u. All the unique users are identified, and if their support is greater than *minsup* they are retained as frequent users F. All the users in F are sorted in the descending order of their support count. Denoted by F_1, this is the list of frequent co-located users of length 1. The root of the FP-tree is created and labelled as "null". For each check-in transaction *Trans* corresponding to a location in L, the frequent users in *Trans* are sorted and selected according to the order of F_1. This sorted list is denoted by $[p|P]$, where p is the first element and P is the remaining list. $[p|P]$ is inserted into the tree *Tree* as follows. If *Tree* has a child N such that $N.userId = p.userId$, then $N's$ count is incremented by 1; else a new node N is created with a count initialized to 1, linked to its parent *Tree*, and linked to all other nodes with the same *userId* via the node-link structure. If P is nonempty, P is recursively inserted to N in the same manner. Table 2 and Figure 1 illustrate the construction of the FP-Tree and Table 3 shows the mined frequent co-located users from the tree. The details of the approach are given in Algorithm 1. The input to the algorithm are the FP-Tree, an empty set α for the of frequent users obtained so far, and the minimum support threshold *minsup*. If the tree contains a single path, then all possible combinations of the nodes (representing users) β in the path are formed, and $\beta \cup \alpha$ are accepted as frequent co-located users with $support = minsup$. Otherwise, the

Algorithm 1 Mining Frequent Co-Located Users

Input: FP-Tree *Tree*, minimum support count threshold *minsup*

Output: Complete set of frequent co-located users F

```
procedure FP-growth(Tree, α, minsup)
    if Tree contains a single path P then
        F ← initialize an empty set
        for all combination (denoted as β) of the nodes in the path P do
            F ← F∪ pattern β ∪ α with support count = minsup of nodes in β
        Return F
    else
        for all a_i in the header of Tree do
            generate pattern β = a_i∪α with support count = a_i.support count
            construct β's conditional pattern base and FP_tree Tree_β
            if Tree_β ∉ φ then
                call FP-growth(Tree_β, β)
```

co-location patterns are generated as $\beta = a_i \cup \alpha$ corresponding to each header $a_i \in$ FP-Tree, their conditional FP-Trees are constructed from their conditional pattern bases. If those trees are non-empty, the FP-growth algorithm is recursively applied on them to obtain the final frequent co-located users.

C. COMPUTING CO-LOCATED FRIENDSHIP COMPONENTS

The frequent co-located users F mined in the previous section are stored in the form of another dictionary D_f, where location is the key, and the array of maximal frequent co-located users as value. It captures the spatial patterns of users. In this step, we further mine geo-social patterns from spatial patterns by exploring the social relationships among the frequent co-located users F. The relationships between each pair of users in each record of D_f are checked against the social network in L to extract the connected components of users corresponding to each location. Algorithm 2 based on depth first search (DFS) shows the method for connected components discovery. It takes the dictionaries of frequent co-located users D_f and the user friendships W as input, and produces

TABLE 3. Mining the FP-Tree constructed in Figure 1.

User	Conditional Pattern Base	Conditional FP-tree	Frequent Patterns Generated
1168	{{0, 167, 1697:1}, {1009, 0, 167:1}}	{0:2, 167:2}	{1168, 0:2}, {1168, 167:2}
1697	{{0, 167:1}, {1009, 0:1}}	{0:2}	{0, 1697:2}
0	{1009:2}	{1009:2}	{0, 1009}
167	{{0:1}, {1009, 0:1}, {1009:1}}	{1009:2, 0:2}	{167, 1009:2}, {167, 0:2}

Algorithm 2 Computing Co-Located Friendship Components

Input: Dictionary of frequent co-located users D_f, Dictionary of user friendships D_w
Output: Array of number of friendship components $\mathcal{C}$

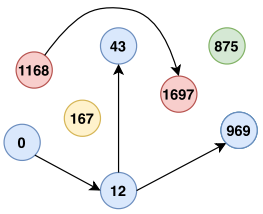
1: **procedure** ColocatedFriendshipComponents
2: $\mathcal{C}, s \leftarrow$ initialize an empty array of size $size(F)$ and an empty stack
3: **for all** $f \in D_f$ **do**
4: mark all users $u \in D_f.value(f)$ as not-visited
5: **for all** $u \in D_f.value(f)$ **do**
6: **if** u is not visited **then**
7: mark u as visited
8: push u into s
9: $\mathcal{C}(f) \leftarrow \mathcal{C}(f) + 1$
10: **while** s is not empty **do**
11: $v \leftarrow$ pop an element out of s
12: **for all** $w \in W.value(v)$ **do**
13: **if** w is not visited **then**
14: Mark w as visited
15: Push w into s
16: **Return** C

the output $\mathcal{C}$ as an array containing the number of components at each location. It starts with initializing an empty array with values set to zero for the number of components, and an empty stack to be used for the DFS-based exploration of the friendship network (line 2). For each record (set of frequently co-located users corresponding to one location-time entry), all the users are initially marked as *not-visited* (line 4). Then each *not-visited* user is accessed (lines 5-6), marked *visited* (line 7), and pushed into the stack (line 8). Each such *non-visited* user increments the count of the number of friendship components obtained so far (line 9). Upon accessing each user, all the elements in the stack are explored until the stack becomes empty (lines 10-15). While processing, each element v is popped out from the stack (line 11), all other *not-visited* users w related to v in the friendship dictionary D_w are accessed (lines 12-13), marked *visited* (line 14), and pushed into the stack (15). Upon its completion of execution, the array $\mathcal{C}$ would have the total number of friendship components for each location-time record, and therefore, returned (line 16).

PUBLIC AND PRIVATE PLACES

With the mined user components information at each location, the public and private location types can be identified from the rationale that *generally the public places have visitors of different backgrounds or socially disconnected with each other, whereas the private places are visited by people who are similar or socially connected with each other*. For instance, a place is marked as *public*, if its number of friendship components is greater than or equal to a predefined threshold ϵ (determined experimentally). Otherwise, it is marked as *private*.

Example 3: Figure 2 illustrates the idea used to mark the places as public and private. The data is a sample from our `BrightKite` dataset, continuing from Table 1. The nodes in the figure represent the obtained frequent co-located users at a particular location. The nodes in the same color are connected together via the friendship relation in the social network. The first figure shows the frequently co-located users at a particular location. As nodes 0, 12, 43, and 969 are connected together, they form one co-located friendship component. Similarly, three other components are obtained from this sample, and therefore marked as public. ■



Sl. No.	Connected Components
1	0, 12, 43, 969
2	1168, 1697
3	875
4	167

FIGURE 2. Illustration of computing co-located friendship components.

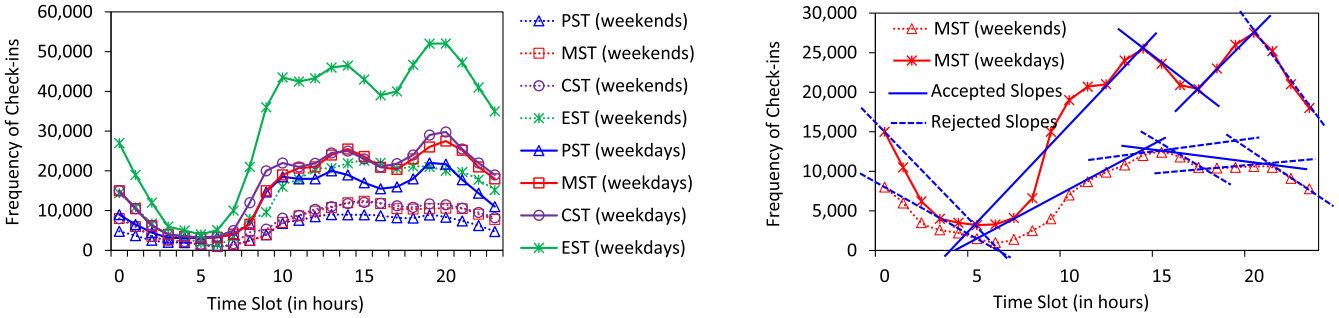
IV. TEMPORAL PATTERN ANALYSIS

After classifying locations into private and public by geo-social pattern mining, we further analyze the temporal patterns to determine the final types of private locations (e.g. residence and work studios) and public locations (e.g. marketplace, corporate area etc.). As the temporal analysis is data-centric, we start with introducing the datasets (also used in experiments).

A. DATASET

We use the publicly available datasets of `Brightkite`[1] and `Gowalla`.[2] `BrightKite` was a popular LBSN during 2007-2012. This dataset consists of two files, each for the check-in data and the friendship network. The check-ins data consists 4,491,143 check-ins over the period

[1]https://snap.stanford.edu/data/loc-Brightkite.html
[2]https://snap.stanford.edu/data/loc-gowalla.html



(a) Check-in trend in different time zones (b) Mountain Standard Time

FIGURE 3. Trends showing frequencies of hourly check-ins for main time zones of USA.

TABLE 4. Time-intervals for for weekdays.

Time-interval	6-14 (6am-2pm)	14-17 (2pm-5pm)	17-20 (5pm-8pm)	20-6 (8pm-6am)
Notation	A	B	C	D

TABLE 5. Time-intervals for weekends.

Time-interval	6-14 (6am-2pm)	14-20 (2pm-8pm)	20-6 (8pm-6am)
Notation	A	B∪C	D

of April 2008 - October 2010, and the friendship network consists of 58,228 nodes and 214,078 undirected edges, where the nodes represent the users and the edges represent bi-directional friendships between users. `Gowalla` was another LBSN similar to `Brightkite`. This dataset consists of 196,591 nodes (users) and 950,327 edges (friendships) in the friendship network, and 6,442,890 check-ins over the period of Feb. 2009 - Oct. 2010.

B. TEMPORAL SEGMENTATION

All check-ins are normally recorded in LBSN in a standard time zone. Therefore the recorded time is generally different than its local time. Both the `BrightKite` and `Gowalla` datasets provide time in the UTC format. The check-ins in most of regions are sparse. Therefore, we create a smaller datasets by extracting region between $75°W - 135°W$, which covers most of North America. The extracted region is divided into four main time zones EST, CST, MST, and PST. So, before conducting temporal analysis, we convert check-in times of all extracted locations from UTC to local time corresponding the check-ins to local events.

The temporal analysis starts with grouping the hourly time-slots that follow the same check-in pattern. For example, the evening hours may have a large number of check-ins, conveying that these hours having similar number of check-ins should be considered in the same time-interval. To properly group hourly time-slots into time-intervals, we use the elbow method, as illustrated in Figure 3. Figure 3(a) shows the frequency of check-ins in each hour of the day, across different region time-zones and for both weekdays and weekends in the `BrightKite` dataset. In each of the line curves, we consider all the *peaks* and *troughs* as candidates for potential boundaries of time-intervals. Between each pair of consecutive peak and trough, the slope of the line connecting them is calculated from $Slope(M_i, M_{i+1}) = \frac{|freq_i - freq_{i+1}|}{|hour_{i+1} - hour_i|}$, where M is the list of consecutive peaks and troughs, $freq_i$ is the frequency of check-ins in the hour of the i-th candidate, and $hour_i$ is the hour of the i-th candidate. Figure 3(b) shows the lines connecting the consecutive candidates and their slopes. The lines with steep slopes indicate a significant deviation in the check-in patterns, whereas lines with gentle slopes indicate an insignificant deviation. We experimentally set the slope thresholds, separately for the weekdays as $\theta_{wd} = 2000$, and weekends as $\theta_{wn} = 1000$, over all time zones. Weekdays and weekends are separated due to the large difference in their usual check-in patterns. Moreover, to form the time-interval segments, consecutive time ranges for which the slopes do not reach the threshold are merged. While the solid lines in the figure, showing slopes above the threshold, are accepted as segmentation points; the dotted lines, having slopes below the threshold, are rejected. Table 4 and 5 show the final formed time-intervals with their corresponding notations.

C. LOCATION TYPE IDENTIFICATION

The final step is to further categorize public/private locations from temporal patterns. For example, if a public location $l \in L$ has active night hours, intuitively we may say that l belongs to some stadium grounds organizing evening concert events. With such observations, we manually establish a relationship between the check-ins patterns of a location at different time-intervals and the possible types, shown

TABLE 6. Possible public and private types based on weekdays.

Time-interval	Ranked public types	Ranked private types
{A}	Education/Work-space(1)	Work-space(1), Residence(2)
{B}	Market(1), Recreation(2), Education/Work-space(3)	Residence(1), Work-space(2)
{C}	Market(1), Recreation(2), Education/Work-space(3),	Residence(1), Work-space(2)
{D}	Recreation like Stadium/club(1), Education/Work-space (2)	Residence(1), Work-space(2)
{A ∩ B}	Education/Work-space(1), Market(2), Recreation(3)	Work-space(1), Residence(2)
{A ∩ C}	–	–
{A ∩ D}	–	–
{B ∩ C}	Market(1), Recreation(2), Education/Work-space(3)	Residence(1), Work-space(2)
{B ∩ D}	–	–
{C ∩ D}	Recreation(1), Education/Work-space(2)	Residence(1), Work-space(2)
{ A ∩ B ∩ C}	Education/Work-space(1), Market(2), Recreation(3)	Work-space(1), Residence(2)
{ A ∩ C ∩ D}	–	–
{B ∩ C ∩ D}	Recreation(1), Education/Work-space(2)	Residence(1), Work-space(2)
{A∩B∩C∩D}	Public Transport like Airport(1), University area(2), Hospital(3), Recreation(4), Work-space like Customer Care Centre(5)	Residence(1), Work-space(2)

TABLE 7. Final public types based on weekdays.

Clubbed time-intervals	Ranked public types	Final public type
{A∪(A∩B)∪(A∩B∩C)}	Education/Work-space(1), Market(2), Recreation(3), All-time operational(4)	Education/work-space
{B∪C∪(B∩C)}	Market(1), Recreation(2), Education/Work-space(3), All-time operational(4)	Market
{D∪(C∩D)∪(B∩C∩D)}	Recreation(1), Education/Work-space(2), All-time operational(3), Market(4)	Recreation
{A∩B∩C∩D}	Public Transport like Airport(1), University area(2), Hospital(3), Recreation(4), Work-space like Customer Care Centre(5)	All-time operational

TABLE 8. Final private type based on weekdays.

Clubbed time-intervals	Ranked private types	Final private type
{A ∪ (A∩B) ∪ (A∩B∩C)}	Work-space(1), Residence(2)	Work-space
{B ∪ C ∪ D ∪ (B∩C) ∪ (C∩D) ∪ (B∩C∩D) ∪ (A∩B∩C∩D)}	Residence(1), Work-space(2)	Residence

TABLE 9. Possible public and private types based on weekends.

Time-interval	Ranked public type	Ranked private type
All	Recreation(1), Market(2), Work-space(3)	Residence(1), Work-space(2)

in Table 6. These relationships are based on real scenarios, realistic assumptions, and the existing related works on functional zones [8]–[10]. The table differentiates both public and private places. The second column shows different types of places based on the time of check-in along with their possible ranking (lower the order in brackets, higher is the possibility of the type of place). If a location has been identified as a public place and been active from 6am to 2pm (interval A), it is marked as a place of *education* like school or *workplace* (1). If the same location has been identified as a private place, it is marked as *workplace* with highest likelihood (1) and *residence* with the second highest likelihood (2). The table also considers if a location has been active in multiple time intervals. If a location has been active in intervals A as well as B and is public, then it is marked as a place of *education* or *workplace* with highest likelihood (1), *market* with second highest likelihood (2), and *recreation* with third highest likelihood (3). If the same location has been identified as private, then it is marked as *workplace* with highest likelihood (1), and *residence* with second highest likelihood (2). In the same way, we consider all possible combinations and present the possible types in the table along with their rank likelihood. This table is further simplified in Table 7 and 8, by clubbing together the time-interval combinations that show the same top-ranked types, resulting in the union of time-intervals. For example, if there are check-ins at a location in time-intervals A, B, and C (represented as {A∩B∩C}), or in time-intervals A and B (represented as {A∩B}), or in A, above the threshold, we can intuitively conclude that there is a high possibility of the region being an educational or work space like corporate offices area (therefore, marked by 1). Next possibility is a marketplace or mall area or some other area that provides various services/amenities (marked by 2), and the least possible is the area of recreation like restaurant, resort, club, etc. (thus, marked by 3). The rankings of super-set categories are given higher priority when clubbing and searching for final types. Table 9 simply shows the possible location types during weekends.

V. EXPERIMENTS

In this section, we present the details of our experimental evaluation. Section V-A presents our evaluation strategy, and Section V-B presents our experimental results.

A. EVALUATION STRATEGY

We compare the results obtained by geo-social-temporal mining with that of manually created Gold standard benchmark, using the metric *mean reciprocal rank* (MRR) that focuses mainly on the rank of inferred location type. Its value is calculated as shown in Equation 1, where G is the set of gold standard and $rank_i$ is the rank of i^{th} location type of G in the ranked list of inferred location types. Further, MRR ratios are measured by comparing the expected (Gold standard, created manually) and observed (inferred by our proposed method) type of locations for several sets of locations. The higher the measure, the better the result quality.

$$MRR = \frac{1}{|G|} \times \sum_{i=1}^{|G|} \frac{1}{rank_i} \tag{1}$$

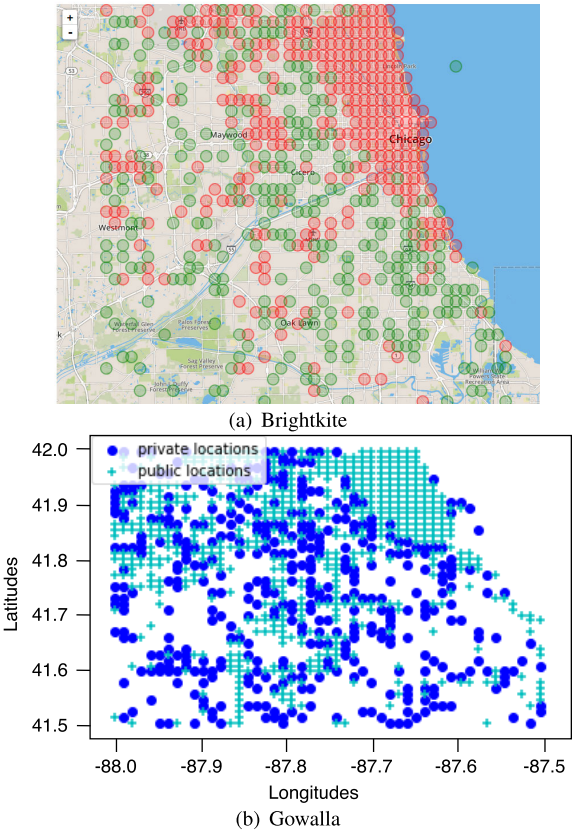
Example 4: Let us consider a gold standard $g_i = market$, which is the actual type of a location l_i identified manually. Let $observed = \{recreation, market, education/workspace, residence\}$ be the ranked set of types of l_i inferred by the proposed method, where *recreation* has the highest likelihood (1) and *residence* has the lowest (4). MRR for this single location is calculated as $\frac{1}{\text{rank } (market) \text{ in observed}} = \frac{1}{2} = 0.5$. MRR of multiple locations is computed as the average of their individual MRR values. ■

To perform the evaluation, we first manually create a Gold standard, for nine sets of locations for Brightkite and another nine sets for Gowalla. This is done as follows. In the first set, we identify the 10 most visited locations in the overall considered data (10 for Brightkite and 10 for Gowalla, separately), manually identify their types, and consider them as a Gold standard. The second set is created by selecting the 10 most visited locations that are inferred as public by our method (10 for Brightkite and 10 for Gowalla, separately), their types are manually identified and considered as a Gold standard. The third set is created in a similar way for private locations. Similarly, there are six more sets, each set for one particular category {Education/Workspace (public), Market (public), Recreation (public), All-time operational (public), Workspace (private), Residence (private)}. With these Gold standard types, we compare the types inferred by our method, using MRR.

B. EXPERIMENTAL RESULTS

In this section, we present our experimental results in two levels. First, the initial results are presented, where the check-in locations are identified as public or private locations, and then the results of the exact inferred type of locations are evaluated.

Figure 4 shows the public and private locations for Chicago region, here we can see that for both Brightkite (Figure 4(a)) and Gowalla (Figure 4(b)) dataset, the type of locations identified as private and public are similar. Also in general, public locations are surrounded by private locations and as we move towards the centre of the city, the frequency of public locations increases implying the general region distribution of a city, thus validating our results.



(a) Brightkite

(b) Gowalla

FIGURE 4. Locations identified as public and private in Brightkite and Gowalla Datasets.

Further, we show the trend of check-ins when only public vs private division in the dataset has been done for both Brightkite and Gowalla in Figure 5. The number of check-ins for each time zone in both the datasets show similar kind of trend. The dashed lines show that the number of weekday check-ins is larger than that of the weekend check-ins (represented by dotted lines), which is also evident as the number of weekdays are more than weekend days. Observe that the trend for weekdays vs weekends check-in is similar for all the four regions of different time zones. The private and public locations show the same pattern as well.

Table 10 presents the MRR measures obtained for the different types of locations in both Brightkite and Gowalla datasets. The measures are obtained for the 10 most visited locations of each type shown with IDs *a-h*, by comparing the location type inferred by the proposed method to the Gold standard created manually. IDs *a-d* are the types for the locations labelled as public, *e-f* are different types of private locations, and *g-h* consider the 10 most visited location in

TABLE 10. MRR measure of 10 most checked-in locations of different types.

ID	Type of category	Brightkite	Gowalla
a	Education/ Workspace	0.95	0.7
b	Market	0.733	0.783
c	Recreation	0.8	0.8
d	All-time operational	0.9	0.925
e	Workspace	0.6	0.75
f	Residence	0.75	1
g	Public Locations	0.9	0.9
h	Private Locations	0.9	0.8

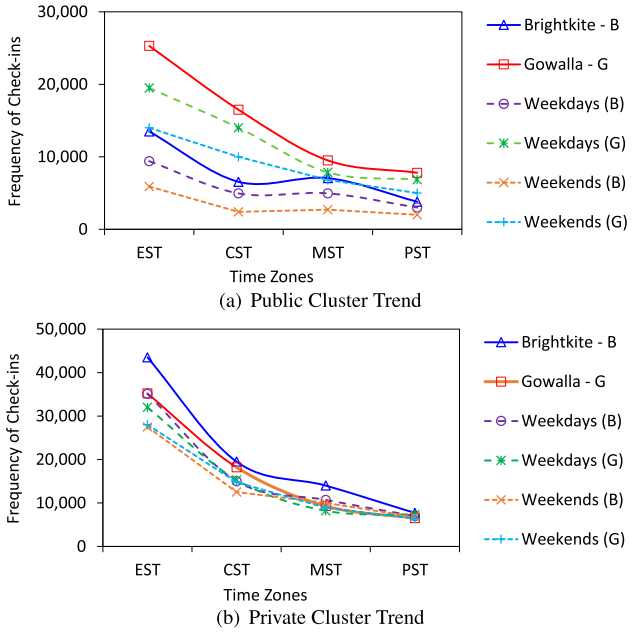


FIGURE 5. Frequency of checkins for the four main time zones of USA after public private division for Brightkite and Gowalla dataset.

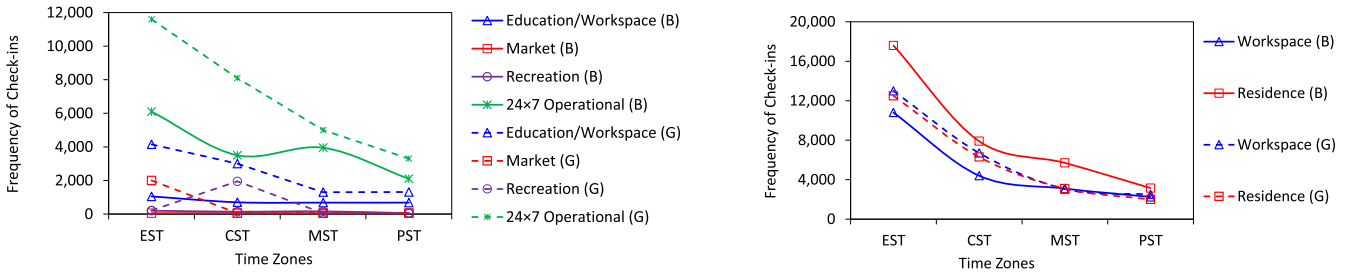
the public and private supersets without delving into further grained categories. Observe that the MRR measures are quite satisfactory with values ranging from 60% to 100%.

Figure 6 which shows the trend for final location type during weekdays of the four regions in different timezones, for both Brightkite and Gowalla datasets (Solid line represents Brightkite and dashed line represents Gowalla). In case of public locations, shown in Figure 6(a), the minimum number of check-ins are for market/services locations for both Brightkite and Gowalla, whereas for other types the trend differs. In Brightkite, a major proportion of check-ins is of all-time operational type, followed by education/workspace region and then recreation. And similar pattern follows for Gowalla dataset as well, with maximum check-ins in all-time operational area and further less education/workspace check-ins, followed by market and recreation. In Figure 6(b) for private locations, the trend of frequency of check-in is opposite for Brightkite and Gowalla, with workspace being more active for gowalla and residential area for Brightkite users.

In many application domains of planning, knowing the type of locations, such as workspace, residential area, recreation area, marketplace, or all-time operational service areas, plays an important role. Traditionally, this kind of tasks were done manually. But as our cities and frequently travelled places are expanding rapidly, these tasks demand smarter ways for automatic profiling of locations. The proposed method solves this problem by considering the LBSN data. The place identified as a private location is not fit for some particular class of activities such as shopping or new entities such as a shopping centre. On the other hand, the place identified as a public location is not fit for residence. A further detailed analysis can be done by considering the specific type, and assessing its suitability with the proposed activity or entity. The real estate industry can utilize the location profile obtained by the proposed method is assessing the value of a property or its future prospects. Tour organizers or tourists themselves can make use of the resulting location types of an untravelled city or country, to plan their spots and stay locations in such a way that match their interests. One major advantage of such method in the planning is that one does not require any local and detailed knowledge of the location, which makes it easy even for those completely unaware about the city or country.

VI. RELATED WORK

In the past decade, a significant research has been carried out in mining interesting patterns from LBSN data, aiming to assist in different application domains, but the problem of location type inference has remained unexplored. The most closely related works are hotspot identification [11], POI inference [12], and functional zones identification [8], all of which are in an urban context. It makes them inherently different than the problem considered in this paper. [11] uses probabilistic topic modelling based approach to extract hotspots by finding interesting patterns from twitter user tags. It can help in applications like traffic control management by detecting crowded regions. [13] gives a detailed analysis of the different spatial patterns found in city, and provides a case study on the cities of the U.S. A topic modeling based approach is used in [8], [12]to find the POIs (places of interests) and identify various functional zones or different type of regions (for example educational areas, recreational areas etc.) in a city. Note that functional zone is a zone of the city (a city can be divided into different zones), whereas location type, as studied in the paper, is inherently different as property of the location. They extend the same problem in [8] by analyzing human movement trajectories obtained from



(a) Public Cluster Trend (b) Private Cluster Trend

FIGURE 6. Trend showing the frequency of checkins of the four main time zones of USA after temporal segmentation in Brightkite and Gowalla datasets: Brightkite (B) shown by solid line, Gowalla (G) shown by dashed line.

the source-destination data of public transport commuters from subways and bus stops. The problem of identifying functional zones can be further extended to specific domains. A cluster ranking based framework is proposed in [14] to form hierarchical generative structures and rank real estate locations according to size, price etc. In [15], the authors take into account all the possible factors that affect the tourist interests in visiting in-home and out-of-town places, to infer POIs. Further, a framework based on spatio-temporal LDA is proposed in [16]. [17], [18] find the daily activity patterns based on spatio-temporal data. [17] uses a statistical method to find human activity patterns according to different types of human groups like students, worker and non-worker, the data for which is obtained by offline survey; and [18] performs the same task by using a kernel density estimation to identify the groups and further cluster the activities using k-means via Principal Component Analysis. A method for mining the spatio-temporal patterns in scientific data is presented in [19]. [20] presents a general framework using Apriori algorithm to identify the spatio-temporal co-occurrence patterns for continuously evolving spatio-temporal events that have polygon-like representations focused on solar events and astronomy to forecast weather more accurately. [21] further improves the efficiency of the approach by introducing a spatio-temporal index. Similarly [10] employs Apriori algorithm to mine spatio-temporal patterns among different regions for location based data based on user trajectories from incoming and outgoing trends of a region. Other works include predicting new check-ins based on users' previous check-in trajectory [22], prediction of location from human activity footprints [23], inferring friend recommendations and analyze social circle [24], inferring demographics of users by finding patterns of call logs [25], predicting social relations [26], and inferring different motifs from mobile users trajectories [27]. [28] and [29] are recent works on human mobility modeling from LBSN data.

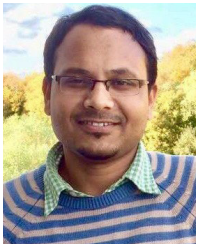
VII. CONCLUSION

In this paper, we proposed a geo-social-temporal mining approach to infer location types from location based social networks data. In particular, user check-in data is first mined to compute frequent co-located users, upon which the frequent co-located connected components are then mined from the social graph for each location. The components give an initial idea about the type of location, as being either public or private. Finally, the temporal patterns are analysed for a finer grained classification to narrow down public locations further into the categories of workspace/education, marketplace, all-time operational and recreation; and private locations into workspace and residence. Experiments conducted on real datasets show convincing results on level-by-level identifications from the generic public/private to specific location types. A promising future research direction is to infer location types across multiple heterogeneous LBSNs to achieve categorization with higher accuracy.

REFERENCES

[1] F. Valls, E. Redondo, D. Fonseca, R. Torres-Kompen, S. Villagrasa, and N. Martí, "Urban data and urban design: A data mining approach to architecture education," *Telematics Informat.*, vol. 35, no. 4, pp. 1039–1052, Jul. 2018.
[2] V. Shapoval, M. C. Wang, T. Hara, and H. Shioya, "Data mining in tourism data analysis: Inbound visitors to japan," *J. Travel Res.*, vol. 57, no. 3, pp. 310–323, 2018.
[3] E. Hromada, "Mapping of real estate prices using data mining techniques," *Procedia Eng.*, vol. 123, pp. 233–240, 2015.
[4] J. Goldblatt, "A future for event management: The analysis of major trends impacting the emerging profession in settings the agenda," in *Proc. Conf. Event Eval., Res. Educ., Events Beyond, Setting Agenda*, J. Allen, R. Harris, L. K. Jago, and A. J. Veal, Eds. Sydney, NSW, Australia: Univ. of Technology, 2000, pp. 1–8.
[5] P. Martí, L. Serrano-Estrada, and A. Nolasco-Cirugeda, "Social media data: Challenges, opportunities and limitations in urban studies," *Comput., Environ. Urban Syst.*, vol. 74, pp. 161–174, Mar. 2019.
[6] J. C. Valverde-Rebaza, M. Roche, P. Poncelet, and A. D. A. Lopes, "The role of location and social strength for friendship prediction in location-based social networks," *Inf. Process. Manage.*, vol. 54, no. 4, pp. 475–489, Jul. 2018.
[7] J. Han, J. Pei, Y. Yin, and R. Mao, "Mining frequent patterns without candidate generation: A frequent-pattern tree approach," *Data Mining Knowl. Discovery*, vol. 8, no. 1, pp. 53–87, Jan. 2004.
[8] N. J. Yuan, Y. Zheng, X. Xie, Y. Wang, K. Zheng, and H. Xiong, "Discovering urban functional zones using latent activity trajectories," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 3, pp. 712–725, Mar. 2015.
[9] Y. Hu and Y. Han, "Identification of urban functional areas based on POI data: A case study of the guangzhou economic and technological development zone," *Sustainability*, vol. 11, no. 5, p. 1385, Mar. 2019.
[10] X. Kong, M. Li, J. Li, K. Tian, X. Hu, and F. Xia, "CoPFun: An urban co-occurrence pattern mining scheme based on regional function discovery," *World Wide Web*, vol. 22, no. 3, pp. 1029–1054, May 2019.
[11] L. Ferrari, A. Rosi, M. Mamei, and F. Zambonelli, "Extracting urban patterns from location-based social networks," in *Proc. 3rd ACM SIGSPATIAL Int. Workshop Location-Based Social Netw.*, 2011, pp. 9–16.

[12] J. Yuan, Y. Zheng, and X. Xie, "Discovering regions of different functions in a city using human mobility and POIs," in *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2012, pp. 186–194.
[13] H. N. Huynh, E. Makarov, E. F. Legara, C. Monterola, and L. Y. Chew, "Characterisation and comparison of spatial patterns in urban systems: A case study of U.S. cities," *J. Comput. Sci.*, vol. 24, pp. 34–43, Jan. 2018.
[14] Y. Fu, H. Xiong, Y. Ge, Z. Yao, Y. Zheng, and Z.-H. Zhou, "Exploiting geographic dependencies for real estate appraisal: A mutual perspective of ranking and clustering," in *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2014, pp. 1047–1056.
[15] H. Yin, B. Cui, X. Zhou, W. Wang, Z. Huang, and S. Sadiq, "Joint modeling of user check-in behaviors for real-time Point-of-Interest recommendation," *ACM Trans. Inf. Syst.*, vol. 35, no. 2, pp. 1–44, Dec. 2016.
[16] H. Yin, X. Zhou, B. Cui, H. Wang, K. Zheng, and Q. V. H. Nguyen, "Adapting to user interest drift for POI recommendation," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 10, pp. 2566–2581, Oct. 2016.
[17] S. Jiang, J. Ferreira, and M. C. González, "Clustering daily patterns of human activities in the city," *Data Mining Knowl. Discovery*, vol. 25, no. 3, pp. 478–510, Nov. 2012.
[18] S. Jiang, J. Ferreira, and M. C. Gonzalez, "Discovering urban spatial-temporal structure from human activity patterns," in *Proc. ACM SIGKDD Int. Workshop Urban Comput.*, 2012, pp. 95–102.
[19] H. Yang, S. Parthasarathy, and S. Mehta, "A generalized framework for mining spatio-temporal patterns in scientific data," in *Proc. 11th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2005, pp. 716–721.
[20] K. G. Pillai, R. A. Angryk, J. M. Banda, M. A. Schuh, and T. Wylie, "Spatio-temporal co-occurrence pattern mining in data sets with evolving regions," in *Proc. IEEE 12th Int. Conf. Data Mining Workshops*, Dec. 2012, pp. 805–812.
[21] B. Aydin, D. Kempton, V. Akkineni, S. R. Gopavaram, K. G. Pillai, and R. Angryk, "Spatiotemporal indexing techniques for efficiently mining spatiotemporal co-occurrence patterns," in *Proc. IEEE Int. Conf. Big Data*, Oct. 2014, pp. 1–10.
[22] H. Gao, J. Tang, and H. Liu, "GSCorr: Modeling geo-social correlations for new check-ins on location-based social networks," in *Proc. 21st ACM Int. Conf. Inf. Knowl. Manage.*, 2012, pp. 1582–1586.
[23] H. Assem and D. O'Sullivan, "Discovering new socio-demographic regional patterns in cities," in *Proc. 9th ACM SIGSPATIAL Workshop Location-Based Social Netw.*, 2016, pp. 1–9.
[24] Q. Gao, G. Trajcevski, F. Zhou, K. Zhang, T. Zhong, and F. Zhang, "Trajectory-based social circle inference," in *Proc. 26th ACM SIGSPATIAL Int. Conf. Adv. Geographic Inf. Syst.*, Nov. 2018, pp. 369–378.
[25] Y. Dong, Y. Yang, J. Tang, Y. Yang, and N. V. Chawla, "Inferring user demographics and social strategies in mobile social networks," in *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2014, pp. 15–24.
[26] T. M. T. Do, K. Kalimeri, B. Lepri, F. Pianesi, and D. Gatica-Perez, "Inferring social activities with mobile sensor networks," in *Proc. 15th ACM Int. Conf. multimodal Interact.*, 2013, pp. 405–412.
[27] C. M. Schneider, V. Belik, T. Couronné, Z. Smoreda, and M. C. González, "Unravelling daily human mobility motifs," *J. Roy. Soc. Interface*, vol. 10, no. 84, Jul. 2013, Art. no. 20130246.
[28] P. Wang, J. Zhang, G. Liu, Y. Fu, and C. Aggarwal, "Ensemble-spotting: Ranking urban vibrancy via poi embedding with multi-view spatial graphs," in *Proc. SIAM Int. Conf. Data Mining*, 2018, pp. 351–359.
[29] P. Wang, Y. Fu, H. Xiong, and X. Li, "Adversarial substructured representation learning for mobile user profiling," in *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2019, p. 130.



TARIQUE ANWAR (Member, IEEE) received the Ph.D. degree in computer science from the Swinburne University of Technology. He is currently working as a Postdoctoral Research Fellow with Macquarie University and CSIRO Data61, in Australia. His research interests include data science, road traffic networks, social networks, and big data analytics.

KEWEN LIAO is currently a Senior Lecturer in information technology with the Peter Faber Business School, Australian Catholic University (ACU). His research interests include data science and algorithms.

ANGELIC GOYAL is currently pursuing the master's degree with the Indian Institute of Technology Ropar, India. Her research interests include data science and algorithms.

TIMOS SELLIS received the Ph.D. degree in computer science from the University of California, Berkeley, in 1986. He is currently a Professor and the Director of the Data Science Research Institute, Swinburne University of Technology, Australia. From 2013 to 2015, he was a Professor with RMIT University, Australia, and before 2013, he was the Director of the Institute for the Management of Information Systems (IMIS) and a Professor with the National Technical University of Athens, Greece. His research interests include big data, data streams, personalization, data integration, and spatio-temporal database systems. He is a Fellow of the ACM.

A. S. M. KAYES received the Ph.D. degree from the Swinburne University of Technology, Australia, in 2014. He is currently a Lecturer with the Cyber Security, La Trobe University, Australia. His research interests include data privacy and security, access control, cyber security, and advanced data analytics. He has served on the research tracks and review panels of many prestigious journals and conferences. He is a member of the Australian Computer Society and the IEEE Computer Society. He has published more than 40 research papers for peer-reviewed journals and conferences. He is an Assessor for the Australian Research Council (ARC).

HAIFENG SHEN is currently an Associate Professor and the Discipline Leader of information technology with the Peter Faber Business School, Australian Catholic University. His primary research expertise is in human-centered artificial intelligence and software technology that blends human expertise and artificial intelligence for better decision making through advanced analytics and interactive visualizations. His research interests include computer supported cooperative work, human computer interaction, software engineering, DevOps, and social and collaborative computing.