A spatiotemporal knowledge graph-based method for identifying individual activity locations from mobile phone data

Keywords: Human mobility, Mobile phone data, Activity location, Spatiotemporal knowledge graph, Community detection

1 INTRODUCTION

Human mobility, which refers to the movement of individuals between geographical locations, is a topic of interest across multiple disciplines (Barbosa et al. 2018). In the past decade, high spatiotemporal coverage from mobile phone data has created new opportunities for human mobility analysis. However, this data typically has sparse temporal and lower spatial resolution compared to household travel surveys, along with locational uncertainty due to cell tower distribution (Wang and Chen 2018). As a result, mobile phone data must first be processed to extract mobility information, particularly the foundational activity locations.

Current studies typically identify activity locations through spatial correlation in mobile phone records. One basic approach defines a fixed spatial range, such as a 500m x 500m grid, grouping records within each grid as a single location (Jiang et al. 2013). However, varying activity boundaries can cause accuracy issues with this fixed threshold. Spatial clustering methods like DBSCAN adaptively group nearby records without a preset range (Huang et al. 2023), but still require careful tuning of parameters like search radius and minimum points. Additionally, these methods rely solely on spatial similarity, making it difficult to distinguish spatially close but temporally distinct activities. Identifying activity locations by considering both spatial and temporal proximity is promising, but spatiotemporal clustering methods often depend on user-defined parameters. A viable alternative is to transform mobile phone records into a graph-based structure that intuitively captures complex spatiotemporal relationships. Parameter-free graph partitioning algorithms can then identify activity locations based on these strong relationships within the graph.

Therefore, this study proposes a spatiotemporal knowledge graph (STKG)-based method for identifying individual activity locations from mobile phone data. The STKG is constructed to capture individual mobility characteristics by transforming spatial and temporal relationships of stays into a spatiotemporal graph. A modularity-optimization community detection algorithm is then applied to identify stays with dense spatiotemporal connections as activity locations. A case study in Shanghai verifies the advantages of the proposed method.

2 METHODOLOGY

As shown in Figure 1, the spatiotemporal knowledge graph (STKG) based activity location identification method includes two steps: (1) Conceptualizing and constructing an STKG to store mobility-related spatiotemporal information from large-scale mobile phone data. Then, inferring spatial adjacency and temporal co-occurrence relationships between individual stays to expand the STKG. (2) Representing the spatiotemporal relationship as a spatiotemporal graph. Community detection algorithm is used to divide the graph into densely connected subgraphs, which are then mapped onto geographical locations to form activity locations.

Firstly, after data preprocessing, the individual traces based on mobile phone data are represented using time slots and grids as basic analysis units in the temporal and spatial dimensions. Thus, an individual's sequence of stays and movements can be represented as a spatiotemporal path that shows their presence in different grid locations at various time slots. In the STKG, the knowledge of what an individual is doing (stay), at what time (time slot), and in what location (grid) can be described using triples. These triples are then stored in a graph format.

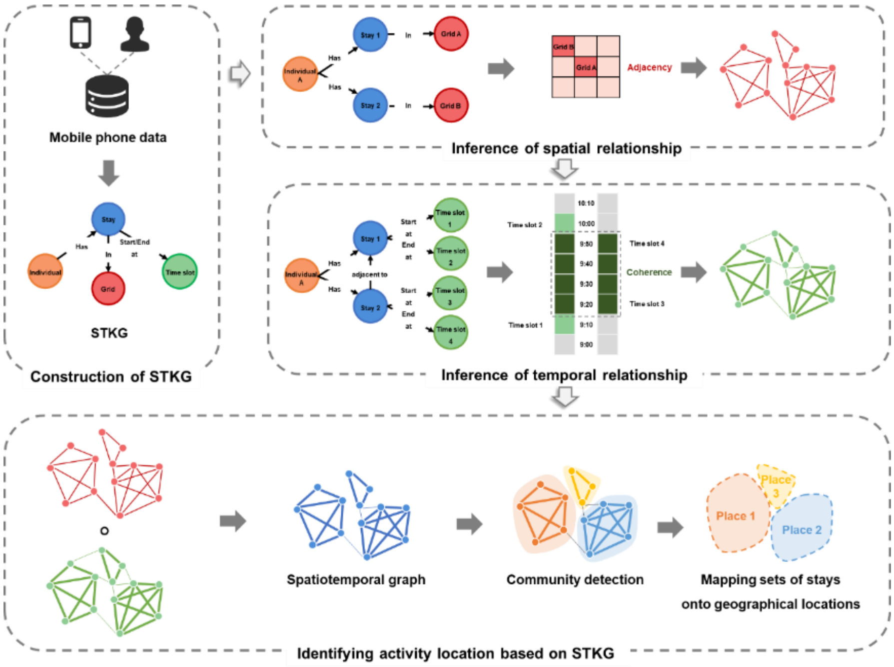


Figure 1 – *Research framework*

The spatial relationships between stays are further inferred based on the STKG. Queen contiguity is defined as two spatial objects being contiguous if they share one or more vertices. After measuring the spatial relationships between grids, spatial relationships between stays can be inferred through the associations between stays and grids. Let the 2D grid array be G, where each cell is uniquely determined by its row and column indices (r, c). For a grid cell $G_{(r,c)}$, its queen contiguity neighbors can be represented as Equation 1:

$$N_{(r,c)} = \{(r + a, c + b) \mid a, b \in \{-1,0,1\}, (a, b) \neq (0,0)\} \tag{1}$$

The co-occurrence of two stays is defined as both stays occurring within the same period. Let the time slots of stay1 and stay2 be represented by binary vectors $\boldsymbol{A}$ and $\boldsymbol{B}$, where each vector has a length of 144, corresponding to the 144 time slots (10 minutes each) in a day. To measure the temporal co-occurrence relationship between stay1 and stay2, the cosine similarity between the two vectors is used. This is calculated as Equation 2, where $\boldsymbol{A} \cdot \boldsymbol{B}$ represents the dot product of vectors $\boldsymbol{A}$ and $\boldsymbol{B}$, and $\parallel \boldsymbol{A} \parallel$ and $\parallel \boldsymbol{B} \parallel$ represent the Euclidean norms of the vectors.

$$CS(\boldsymbol{A}, \boldsymbol{B}) = \parallel \boldsymbol{A} \parallel \parallel \boldsymbol{B} \parallel \boldsymbol{A} \cdot \boldsymbol{B} \tag{2}$$

For an individual, the graph depict temporal and spatial relationships is represented as a single graph using the Hadamard product, denoted as $STG = SG \circ TG$, where SG represents the spatial relationship graph, and TG represents the temporal relationship graph. Fast Unfolding algorithm is used to divide STG into subgraphs where nodes with dense spatiotemporal relationship. Modularity Q is used to evaluate the quality of community partitioning, which is defined as Equation 3, where A_{ij} is the weight of the edge between nodes i and j, k_i is the strength of node i. $\delta(C_i, C_j)$ is the Kronecker delta, equal to 1 if nodes i and j belong to the same community ($C_i = C_j$) and 0 otherwise. C_i is the community assignment of node i.

$$Q = \frac{1}{2m} \sum_{i,j} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(C_i, C_j)$$

$$m = \frac{\sum_{ij} A_{ij}}{2} \tag{3}$$

$$k_i = \sum_j A_{ij}$$

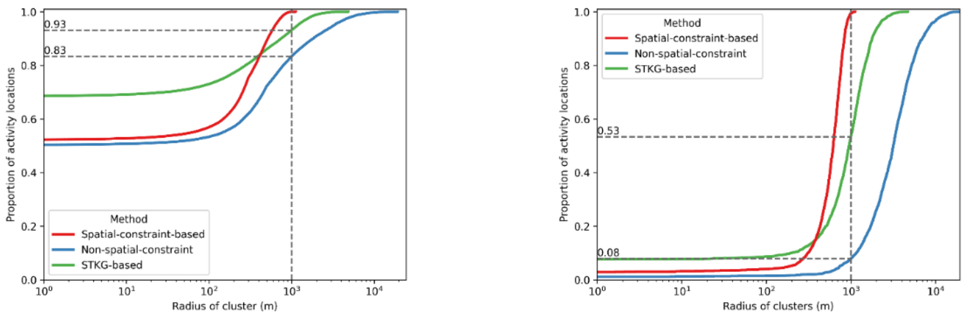
The modularity gain ΔQ is calculated as Equation 4, where W_{in} is the total weight of the edges inside community C_j, W_{tot} is the total weight of the edges connected to nodes in community C_j, k_i is the degree (total weight of edges) of node i, k_i^{in} is the sum of the weights of edges from node i to nodes in community C_j.

$$\Delta Q = \left[\frac{W_{in} + 2k_i^{in}}{2m} - \left(\frac{W_{tot} + 2k_i}{2m}\right)^2\right] - \left[\frac{W_{in}}{2m} - \left(\frac{W_{tot}}{2m}\right)^2 - \left(\frac{k_i}{2m}\right)^2\right] \tag{4}$$

The community detection helps identify meaningful communities within the graph, grouping stays that exhibit strong spatiotemporal correlations into clusters. By mapping stays that belong to the same community onto geographical locations, one can derive individual activity locations.

3 RESULTS

To evaluate activity location identification methods spatially, we calculated the cluster radius for all activity locations, as shown in Figure 2(a). For locations identified by the spatial-constraint-based method, cluster radii were generally below 1500 meters, with most under 1000 meters. Using 1000 meters as a threshold, we found that 83% of activity locations from the STKG-based method and 93% from the non-spatial-constraint method fell below this threshold, indicating that the STKG-based method, by considering temporal correlation, achieves 10% greater accuracy. We further analyzed the cluster radius for each user's primary daytime activity location, as shown in Figure 2(b). Results indicate a more significant advantage for the STKG-based method over the non-spatial-constraint method. While the non-spatial-constraint method may require parameter adjustments to increase the proportion of locations within a 1000-meter radius, the STKG-based method can contain an additional 45% of daytime hotspots within this range without parameter changes.



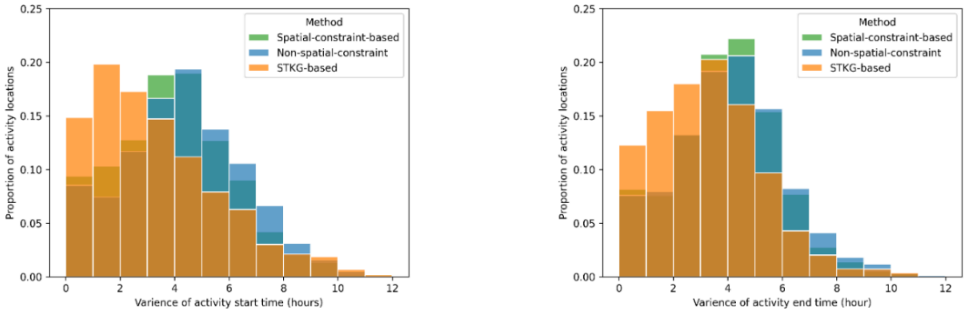
(a) All activity locations

(b) Activity locations with the longest daytime stay for each user

Figure 2 – *Radius of activity locations*

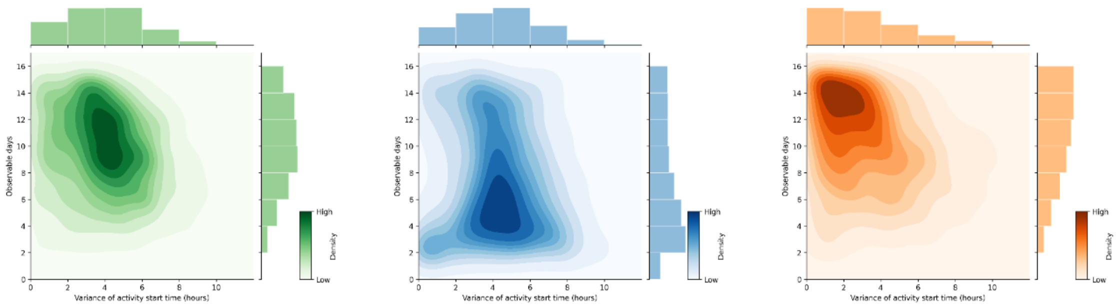
To evaluate activity location identification methods temporally, two indicators are used: the variance of start and end times for activities at the same location on different days and the frequency of days work activities are observed. As shown in Figure 3, we calculated the start and end time variance at the location with the longest daytime stay (potentially the workplace). The STKG-based method outperforms baseline methods, maintaining 10–20% lower variance.

Figure 4 shows the joint distribution of start time variance and workday frequency at these locations. The spatial-constraint method identifies more workdays (8–12 days) than the non-spatial-constraint method (4–8 days), though their start time variance is similar (around 4 hours). In contrast, the STKG-based method has high-density values in regions with more observed workdays (13–15 days) and lower start time variance (around 2 hours), indicating strong regularity and stability in identified activity locations.



(a) Variance of activity start time (b) Variance of activity end time

Figure 3 – *Variance in start and end times of activities with the longest daytime stay*



(a) Spatial-constraint-based method (b) Non-spatial-constraint method (c) STKG-based method

Figure 4 – *Joint distribution of variance in start times of activities with the longest daytime stay and the number of days these activities are observable*

4 CONCLUSION

This study introduces an STKG-based method for identifying individual activity locations from mobile phone data. The STKG models mobility patterns by integrating spatial and temporal connections between stays, forming a spatiotemporal graph. A modularity-optimization algorithm then detects dense spatiotemporal clusters as activity locations. A Shanghai case study validates the effectiveness of proposed method. Future work will involve recruiting volunteers to compare passive mobile phone-based location identification with actively collected activity data for further validation.

References

Barbosa, Hugo, Marc Barthelemy, Gourab Ghoshal, Charlotte R James, Maxime Lenormand, Thomas Louail, Ronaldo Menezes, José J Ramasco, Filippo Simini, and Marcello Tomasini. 2018. “Human Mobility: Models and Applications.” *Physics Reports* 734: 1–74.

Huang, Le, Fan Xia, Hui Chen, Bowen Hu, Xiao Zhou, Chunxiao Li, Yaohui Jin, and Yanyan Xu. 2023. “Reconstructing Human Activities via Coupling Mobile Phone Data with Location-Based Social Networks.” *Travel Behaviour and Society* 33. Elsevier: 100606.

Jiang, Shan, Gaston A Fiore, Yingxiang Yang, Joseph Ferreira Jr, Emilio Frazzoli, and Marta C González. 2013. “A Review of Urban Computing for Mobile Phone Traces: Current Methods, Challenges and Opportunities.” In *Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing*, 1–9.

Wang, Feilong, and Cynthia Chen. 2018. “On Data Processing Required to Derive Mobility Patterns from Passively-Generated Mobile Phone Data.” *Transportation Research Part C: Emerging Technologies* 87: 58–74.