Project ID: P25
Project Title: TOURIST TRAJECTORIES Bangkok
Group ID: H5
Team Members with ID numbers: Ashutosh Wagh (2022H1030052H), S Shashank (2022H1030067H), Srinidhi P Katte (2022H1030075H), Kushal Chakraborty (2022H1030089H)

1. Introduction

In the dynamic realm of tourism, understanding and identifying the most frequented destinations by tourists is imperative for effective destination management, strategic marketing, and efficient resource allocation. With the advent of social media, a vast and invaluable source of data has emerged, capturing the diverse preferences and behaviors of tourists. Effectively harnessing this wealth of information presents a unique challenge: discerning patterns, uncovering the most sought-after destinations and finding the communities of tourists preferring certain types of destination trips. This problem statement centers on leveraging the power of social media analytics and community detection techniques to observe and identify tourist patterns. The ultimate goal is to pinpoint the most popular places visited, thereby enhancing tourist experiences and positively impacting the tourism industry of various regions, consequently contributing to the economic development of these areas.

2. Motivation

The motivation behind this endeavor lies in the transformative potential of harnessing social media data for tourism analysis. As tourists increasingly share their experiences, preferences, and travel itineraries on social platforms, a treasure trove of information becomes available. Extracting meaningful insights from this data can revolutionize destination management, enabling a deeper understanding of tourist behaviors. By employing advanced analytics and community detection algorithms, we aim to unlock patterns within this rich dataset, revealing clusters of users with similar travel patterns and preferences. This knowledge can empower stakeholders in the tourism industry to tailor their services, marketing strategies, and resource allocation to align with the identified patterns, ultimately enhancing the overall tourist experience.

3. Related Work

A. Paper Title and Authors: GIS Based Route Network Analysis for Tourist Places: A Case Study of Greater Imphal, T. Prameshwori, Wangshimenla. J, L. Surjit, L. Ramananda

Summary: This study uses ArcGIS network analysis to boost tourism development by establishing efficient travel routes, providing travel directions, locating adjacent facilities, and determining service regions by journey time directions, locating adjacent facilities, and determining service regions by journey time and distance. GIS technology aids tourist site visitation and decision-making. Modularity maximization technique is used to perform community detection in order to find the most sought after regions of Imphal. Through this, they have tried to infer the quality of service provided to the tourists in certain regions.

B. Paper Title and Authors: Xu, Dong, et al. "Tourism community detection: A space of flows perspective." Tourism Management 93 (2022): 104577.

Summary: The "space of flows" theory, focusing on information exchange and material movement at the time-space level, provides a novel scientific perspective for examining interactive tourism destination networks. This study, integrating tourism flow data with network scientific approaches, identifies seven actual and six potential tourism communities in the Yangtze River Delta urban agglomeration. Notably, distinct characteristics in tourism flow exist within and across these communities, emphasizing the need for considering spatiotemporal attributes in predicting regional tourism demand accurately. The study underscores the roles of informatization and government

behavior as key factors in shaping destination networks. Finally, the implications of tourism community detection for the integrated development of regional tourism are discussed.

C. **Paper Title and Authors:** Hu, Fei, et al. "A graph-based approach to detecting tourist movement patterns using social media data." *Cartography and Geographic Information Science* 46.4 (2019): 368-382.

   Summary: Understanding tourist movement is crucial for tourism studies, influencing strategies from attraction planning to product development. Traditional methods are limited by scale and cost. This paper proposes a graph-based method using Twitter data to detect tourist movement patterns. Geo-tagged tweets are cleaned, and community detection techniques like modularity maximization creates tourist graphs. Communities have also been created using clustering methods and their performance has been compared and evaluated in order to find the better performing community detection technique. Network analysis identifies patterns like popular attractions and tour routes. The approach is demonstrated using New York City, offering utility for businesses and government in planning tours, transportation, and developing centers.

D. **Paper Title and Authors:** Hu, Mingxing, et al. "The Communities Detection of the Tourist Flow Network using Mobile Signaling Data in Nanjing, China." Applied Spatial Analysis and Policy (2023): 1-24.

   Summary: This paper uses mobile signaling data and community detection techniques to analyze tourist flow communities. It identifies four spatial communities, evaluates core nodes using betweenness centrality and degree centrality, and analyzes factors influencing popularity response. Findings highlight spatial differentiation and varying roles of core nodes in different staying periods. The study emphasizes the influence of accommodation density, core nodes, key attractions, and accessibility on community formation. Analysis has been done about the formation of different community structures at different time durations. Community analysis has been done using modularity maximization and asynchronous label propagation techniques and by taking into account the modularity score. Incorporating a detailed time dimension contributes to understanding tourism communities, core nodes, and factors affecting popularity response, informing tourism community development and marketing.

4. Problem Statement

The challenge at hand revolves around effectively utilizing social media analytics and community detection techniques to discern patterns and identify the certain tourist communities preferring certain types of destination trips. This entails representing the social interactions among users as a graph, where nodes represent trip destinations of tourists, and edges represent connections such as interactions, distance or trip similarities.

The objective is to partition these nodes into communities in a way that maximizes the density of connections within communities while minimizing the density of connections between communities.

The mathematical formulation involves defining communities, introducing the modularity metric as an objective function, and employing community detection algorithms like the Louvain method to optimize this function.
The result is a partitioning of users into communities, each representing a cluster with similar travel patterns. This analysis aims to provide a data-driven foundation for destination management, marketing strategies, and resource allocation, ultimately contributing to the growth and sustainability of the tourism industry in various regions.

Let's define the problem mathematically:

A. Graph Representation

Let G=(V,E) be an undirected graph, where V is the set of nodes representing tourists' trip todestinations, and E is the set of edges representing connections between tourist destinations using various trip attributes like distance, categories etc.

B. Community Assignment

Let C={C_1, C_2, … , C_3} be the set of communities, where each C_i is a subset of V representing a community of tourists traveling to certain destination trips.

C. Community Detection Objective

The objective is to partition the set of nodes V into communities C in a way that maximizes the density of connections within communities and minimizes the density of connections between communities.

D. Modularity

The modularity of a partition is a measure of how well the partition captures the density of connections within communities compared to a random assignment. It is often denoted by Q and is defined as:

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

where A_{ij} is the adjacency matrix of the graph, k_i is the degree of node i, m is the total number of edges in the graph, and the value of $\delta(C_i, C_j)$ is 1 if $C_i = C_j$ and 0 otherwise.

E. Louvain Method

The Louvain method is an algorithm that optimizes the modularity of a partition. It works in a two-step iterative process:

First Step (Greedy Optimization): Each node is initially assigned to its own community. Nodes are then iteratively moved to the community that maximizes the increase in modularity.
Second Step (Agglomeration): Communities identified in the first step are treated as nodes in a new network, and the process is repeated.

F. Optimization Problem

The community detection problem can be framed as an optimization problem where the goal is to find the partition of nodes that maximizes the modularity:

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

subject to the constraint that each node belongs to exactly one community.

G. Label Propagation

Assign an initial label to each node. Initially, each node is assigned a unique label, i.e., label(v_i) = i for $v_i \in$ V.
The label update rule for node v_i can be formulated as follows:

$$label(v_i) \leftarrow arg\,max_c \sum_{v_j \in N(v_i)} \delta(label(v_j), c)$$

where $N(v_i)$ is the set of neighbors of v_i and $\delta(label(v_j), c)$ is equal to 1 if $label(v_j) = c$ and 0 otherwise.

Repeat the label update rule for a certain number of iterations or until convergence. At each iteration, nodes update their labels based on the labels of their neighbors. The final community assignment is determined by the labels after the label propagation process. Nodes with the same label belong to the same community.

5. Hypothesis

Challenges faced with the dataset:

Each of the data points in the dataset has source and destination details. Upon performing exploratory data analysis, we have discovered that the number of sources and destinations is equal to the number of rows, indicating that each trip is unique.

The source and destination details, as mentioned above, are not repeated in any other rows or data points. This leads to the formation of a disjoint bipartite graph, where node 1 represents the source and node 2 represents the destination, and the edge represents the correspondence between the source and destination. Due to the uniqueness of the source and destination details in each row, we consider that each data point represents the travel details of a unique customer, and there is no relation between other customers. Therefore, we propose different hypotheses to form edges or connections between various data points.

Hypothesis 1

We have taken the latitude and longitude of destinations and calculated the Haversine distance between every pair of destinations, creating the distance matrix. The dimension of the distance matrix formed is 33029x33029. Upon analyzing the values of the distance matrix, we found that the statistical measures of the distance values of the destinations are as follows:

Mean Distance: 11.870405936237502
Median Distance: 11.063895582669545
10th Percentile: 2.72303357925085
75th Percentile: 17.449444087284803
90th Percentile: 22.02091283349378

Distance values provide a measure of the closeness of the destinations to each other, offering insight into the similarity of the nodes. We formed a graph in which each node represents a destination. The pair of nodes with a distance greater than or equal to the threshold value forms an edge in the graph. Initially, we set the threshold value for similarity between nodes to be 11.87, but the number of edges was excessively high, leading to memory errors due to limited computational resources. After various analyses, we selected a subset of a thousand data points and computed the distance matrix as mentioned above. Using a threshold distance of 30, we obtained a manageable number of edges (26778) without causing computational issues. We utilized the formed graph for community detection to identify clusters of tourists with similar travel patterns.

Hypothesis 2

In this, we initially form a category similarity matrix, where

Now, we form a graph in which each node represents the destination, and there is an edge between two nodes if they belong to the same destination category. Then, we create a graphical representation where individual destinations are represented as nodes, and an edge is established between any two nodes if they share a common destination category. In this context, a destination category serves as a grouping criterion, and the graph provides a visual depiction of the relationships between destinations based on their shared category memberships. We use the above-formed graph to perform community detection and find out the cluster of tourists with similar travel patterns.

Hypothesis 3

In this we initially form a vehicle_id_trip_sim_matrix, where

Now, we form a graph in which each node represents the destination, and there is an edge between two nodes if the vehicle ID of the vehicle used for the trip to the two destination nodes is the same. Let's create a graphical representation where individual destinations are represented as nodes, and an edge is established between any two nodes if they use the same vehicle for the trip. In this context, a vehicle ID serves as a grouping criterion, and the graph provides a visual depiction of the relationships between destinations based on the intersection of the vehicles used for the destination trips. We use the above-formed graph to perform community detection and find out the cluster of tourists with similar travel patterns.

6. Solution Approach

We have performed the following steps to approach to the solution of the problem statement:

1) Input : Latitude and Longitude of the destinations
Graph G(V,E): V = {Node_id of the destination nodes}
E = {}
The graph has been created using different approaches for different hypotheses.

a) Hypothesis 1:

```
Step 1)   Adjacency_matrix[|V|][|V|]={-1}
Step 2)   Distance_matrix[|V|][|V|]={-1}
Step 3)   threshold_distance
Step 4)   For i in V :
Step 5)     For j in V:
Step 6)         Calculate the haversine distance between destinations indexed by  i
                and j using the corresponding latitude and longitude.
Step 7)    Haversine_distance =  EarthRadius * 2 * atan2(sqrt(
           sin(radians(lat_of_dest2 -
           lat_of_dest1)/2)**2 + cos(radians(lat_of_dest1)) *
            cos(radians(lat_of_dest2)) * sin(radians(lon_of_dest2 -
            lon_of_dest1)/2)**2 , 1 -sqrt( sin(radians(lat_of_dest2 -
            lat_of_dest1)/2)**2 + cos(radians(lat_of_dest1)) *
            cos(radians(lat_of_dest2)) * sin(radians(lon_of_dest2 -
            lon_of_dest1)/2)**2 )
Step 8)    Distance_matrix[i][j] = haversine distance
Step 9)   Find the threshold_distance from Distance_matrix using various
          statistical measures.
Step 10)     For i in V:
Step 11)       For j in V:
Step 12)           If Distance_matrix[i][j] >= threshold_distance:
Step 13)               Adjacency_matrix[i][j] = 1
Step 14)               E = E U {(i,j)}
```

```
Step 15)          Else:
Step 16)              Adjacency_matrix[i][j] = 0
```

Since we did not have any information about the distances between destinations, we have used haversine distance. Haversine distance is more realistic in nature than other distance metrics like Euclidean distance and Manhattan distance since it uses latitude and longitude to find out the actual earth-based distances.

b) Hypothesis 2:

```
Step 1)  Adjacency_matrix[|V|][|V|]={-1}
Step 2)  Category_Similary[|V|][|V|]={-1}
Step 3)    For i in V:
Step 4)      For j in V:
Step 5)          If  i.category_d == j.category_d:
Step 6)              Category_Similary[i][j] = 1
Step 7)               Adjacency_matrix[i][j] = 1
Step 8)               E = E U {(i,j)}
Step 9)          Else:
Step 10)             Category_Similary[i][j] = 0
Step 11)             Adjacency_matrix[i][j] = 0
```

c) Hypothesis 3:

```
Step 1)  Adjacency_matrix[|V|][|V|]={-1}
Step 2)  Vehicle_id_trip_sim_matrix[|V|][|V|]={-1}
Step 3)    For i in V:
Step 4)      For j in V:
Step 5)          If  i.vehicle_id == j.vehicle_id:
Step 6)              Vehicle_id_trip_sim_matrix[i][j] = 1
Step 7)               Adjacency_matrix[i][j] = 1
Step 8)               E = E U {(i,j)}
Step 9)          Else:
Step 10)             Vehicle_id_trip_sim_matrix[i][j] = 0
Step 11)             Adjacency_matrix[i][j] = 0
```

2) After forming graphs using the above approaches for different hypotheses, now we use various community detection techniques to identify the trip patterns of the tourists.

a) Disjoint Community Detection Technique

Step 1) Initialize the community_id of the node with the node_id. The number of communities are equal to the number of nodes ie |V|.

Step 2) Now for each node j in V - {i}, add it to the node i's community and calculate the modularity score.

Step 3) Keep only that node in i's community which will give the maximum modularity score. This is the merging operation.

Step 4) Perform step 3 for different pairs of nodes until there is an increase in modularity score.

Step 5) The next stage is node aggregation, in which all the nodes that belong to the the same community is aggregated into a single super node. The edges between different super nodes are formed and the weight between the super nodes denotes the number of edges from nodes of one super node to nodes of another super node.

Step 6) The above steps are carried out until convergence.
Step 7) The number of super nodes provides the number of communities formed using this technique.

Reason: We have used Louvain technique for disjoint community detection for two reasons:

i) We want to find out the destination categories of the tourist trips which form strong communities among themselves. For example, Tourists going to NightClubs may form communities with the tourists going to Restaurants.

ii) Modularity is a network-centric global metric, that when used for maximization, considers the entire network structure. To infer the travel patterns of tourists, analyzing the complete graph is crucial. Given that the Disjoint community detection technique referenced above operates globally, utilizing the entire graph, we deemed it suitable for addressing the problem at hand.

b) Overlapping Community Detection Technique

Step 1) Initialize unique community labels for all nodes in the network. Initialize the community_id of each node equal to its node_id.
Step 2) Inner Iteration: Update labels for all the nodes in the networks
  a. Update with the label having the highest frequency in its neighbors' current labels
  b. Break the tie in case of discrepancy at random
Step 3) Outer Iteration: Stop if the label is unchanged compared to earlier iteration. Else, continue

Reason: The reason for using the Label Propagation Asynchronous approach for overlapping community detection is because of the fact that a destination category may be a part of more than one community. Example: The tourists going to Spa may be a part of a community where tourists go to NightClubs and Restaurants as well as they may be a part of a community where tourists have trips to Leisure and Shopping Mall.

3) After performing community detection, we evaluate the performance of disjoint community Detection using the following approach.

a) The categories data of the destinations obtained from the 'category_d' attribute of the dataset forms the ground-truth category values or the community structure.
b) The communities for the destinations detected by the Louvain greedy method form the detected category values.
c) We use these two values to find out the purity score, true positive(tp),false positive(fp),true negative(tn), false negative(fn) and the other associated metrics like precision, recall, f1-score.

7. Results

Ground Truth Value of Number of communities = 6

Metric	Hyp1	Hyp2	Hyp3
No. of communities detected by Louvain method	**5**	**6**	**159**
Purity Score	**0.408**	**1.0**	**0.666**
Accuracy	**0.580**	**1.0**	**0.705**

F1 Score	**0.293**	**1.0**	**0.009**
Precision	**0.291**	**1.0**	**0.352**
Recall	**0.296**	**1.0**	**0.004**
True Positive	**21551**	**143379**	**93**
True Negative	**121617**	**356121**	**47026**
False Positive	**52438**	**0**	**171**
False Negative	**51147**	**0**	**19505**
True Negative Rate	**0.698**	**1.0**	**0.996**
False Positive Rate	**0.301**	**0.0**	**0.003**

8. **Insights**

Insights with respect to results from hypothesis 1

1. From the detected community versus destination category heatmap, we can infer that tourists with destination trips to nightclubs/bars, restaurants, and spas form dense communities. Hence, the trip pattern we can infer is that tourists with destinations to nightclubs/bars, restaurants, and spas have similar destination preferences and are likely to visit the same destination category in the future.
2. We have used the "category_d" of the dataset to form the ground truth community structure, in which each destination belongs to the community as mentioned in the "category_d" of the dataset. We have applied two community detection techniques.
    a. Using modularity maximization, we have received a modularity score of 0.411. The positive value of modularity indicates the presence of a strong community structure. However, as it is not very close to 1, the density of the community structure is not very high when compared to dense community structures with a modularity value of 1.
    b. The results of Modularity Maximization for disjoint community detection, which include purity score, true positives (tp), true negatives (tn), false positives (fp), false negatives (fn), confusion matrix, and F1-score, as well as accuracy, show that the Modularity Maximization technique performs fairly well in the detection of communities. However, as the values are not very close to the ideal values of the metrics, it is unable to detect communities exactly the same as the ground truth communities found from "category_d."
    c. The label propagation technique for overlapping community detection forms only two communities, indicating that it is unable to accurately capture and represent the communities as per the ground truth. It indicates a large amount of overlap of categories of tourist trip destinations which is not very realistic.

Insights with respect to results from hypothesis 2

1. From the detected community versus destination category heatmap, we can infer that tourist destinations, with the ground truth structure given by each "category_d" attribute, are perfectly placed in separate communities as formed by the Modularity Maximization technique. Based on this graph, constructed using hypothesis 2, we can infer that tourists going to destinations in the same categories as mentioned in the "category_d" attribute form highly clustered communities. This indicates that tourists in the particular community are more likely to plan a trip with a pattern featuring the same category of destination in the future.

2. We have used the "category_d" of the dataset to form the ground truth community structure, in which each destination belongs to the community as mentioned in the "category_d" of the dataset. We have applied two community detection techniques.
    a. Using modularity maximization, we have received a modularity score of 0.64. The high positive value of modularity indicates the presence of a strong community structure. From this, we infer that modularity maximization performs better for the graph constructed using hypothesis 2 when compared with the graph in hypothesis 1.
    b. The results of Modularity Maximization for disjoint community detection, including purity score, true positives (tp), true negatives (tn), false positives (fp), false negatives (fn), confusion matrix, and F1-score, as well as accuracy, show that the Modularity Maximization technique performs well in the detection of communities.
    c. The label propagation technique forms 6 communities, which is the same as the number of ground truth community structures constructed using "category_d."

Insights with respect to results from hypothesis 3

1. We observe that the number of communities formed using modularity maximization is greater than the number of communities in the ground truth structure.
2. We have used the "category_d" of the dataset to form the ground truth community structure, in which each destination belongs to the community as mentioned in the "category_d" of the dataset. We have applied two community detection techniques.
    a. Using modularity maximization, we have received a modularity score of 0.98. The high positive value of modularity indicates the presence of a strong community structure, and the density of the community structure formed through modularity maximization is high. From this, we infer that vehicles that tourists use to travel to a particular destination category are likely to be used for trips to the same destination categories by other tourists in the future.
    b. The results of Modularity Maximization for disjoint community detection, which include purity score, true positives (tp), true negatives (tn), false positives (fp), false negatives (fn), confusion matrix, F1-score, and accuracy, show that the Modularity Maximization technique performs extremely well in the detection of high-density communities.
    c. The label propagation technique forms a high number of well-clustered communities.

9. References

1. GIS Based Route Network Analysis for Tourist Places: A Case Study of Greater Imphal, T. Prameshwori, Wangshimenla. J, L. Surjit, L. Ramananda
2. Xu, Dong, et al. "Tourism community detection: A space of flows perspective." Tourism Management 93 (2022): 104577
3. Hu, Fei, et al. "A graph-based approach to detecting tourist movement patterns using social media data." *Cartography and Geographic Information Science* 46.4 (2019): 368-382.
4. Hu, Mingxing, et al. "The Communities Detection of the Tourist Flow Network using Mobile Signaling Data in Nanjing, China." Applied Spatial Analysis and Policy (2023): 1-24.
5. Chakraborty, Tanmoy, et al. "Metrics for community analysis: A survey." *ACM Computing Surveys (CSUR)* 50.4 (2017): 1-37.
6. Social Network Analysis, Tanmoy Chakraborty, Wiley, 2021
7. https://en.wikipedia.org/wiki/Haversine_formula