Name:	Suhas Kamath	**Assignment No:**	Assignment 1
SR No:	06-18-01-10-51-25-1-25945	**Course Code:**	DS221
Email ID:	suhaskamath@iisc.ac.in	**Course Name:**	Intro. to Scalable Systems
Date:	September 8, 2025	**Term:**	AUG 2025

Question 1

Solution Approach

In this question, we had to find the minimum weight of the duplicate parcels. The list of parcels was provided in the input file. An initial approach could be to run a nested for loop, which should read the list and identify the minimum weight of the duplicate parcels. This would be an inefficient algorithm and would result in a time complexity of $O(n^2)$. A much better algorithm is as follows:

1. Create a hash map to store the id and the minimum weight of each parcel.
2. Create a set to store the list of package ids that have duplicates.
3. Loop through the list of parcels and update the minimum weight with respect to each id in the above hash map.
4. During this loop, if a package is found to be duplicated, it is added to the set.
5. By the end of the for loop, we have two pieces of data: the id numbers of the duplicated parcels, and the minimum weight of each parcel.
6. Using the above two pieces of data, we create a separate two-dimensional vector array to store the list of duplicated packages with their minimum weight.
7. We then sort and return this vector array.

Time and Space Complexity Analysis

Time Complexity

First, let us analyze the program to determine its time complexity. The code can be divided into the following major sections, namely:

1. Creation and initialization of the required data structures.
2. The for loop that calculates the list of duplicate package ids and the minimum weight of each parcel.
3. The for loop that compiles the final result, i.e duplication parcel ids and their respective minimum weights.
4. The sort() function that sorts the package ids based on their id number.

The detailed time complexity analysis is as follows:

Let m = number of duplicate parcels and n = total number of parcels in the list.

Section	**Best Case**	**Average Case**	**Worst Case**	**Reasoning**
Section 1	$O(n)$	$O(n)$	$O(n)$	A single pass is made through all n parcels. Hash map and set operations are $O(1)$ on average.
Section 2	$O(1)$	$O(m)$	$O(m)$	In the best case, no duplicates exist ($m = 0$). In the worst case, nearly every unique parcel is a duplicate.
Section 3	$O(1)$	$O(m \log m)$	$O(n \log n)$	The cost depends entirely on the number of duplicates, m. In the worst case, m is proportional to n.
Total	$O(n)$	$O(n + m \log m)$	$O(n \log n)$	The best case is linear if no sorting is needed. The worst case is dominated by sorting all duplicate IDs.

Space Complexity

There are three main data structures. They are as follows:

- Hash map
- Set
- Result vector

The detailed space complexity analysis is as follows:

Data Structure	**Best Case**	**Average Case**	**Worst Case**	**Reasoning**
Map	$O(n)$	$O(n)$	$O(n)$	In all cases, this map may store up to n unique parcel IDs.
Set	$O(1)$	$O(m)$	$O(n)$	Stores duplicate IDs. The best case is no duplicates ($m = 0$); the worst is that nearly all n items correspond to unique duplicate IDs.
Vector	$O(1)$	$O(m)$	$O(n)$	The output size is determined by m, the number of unique duplicates.
Total	$O(n)$	$O(n)$	$O(n)$	The space is dominated by the map, which depends on the number of unique parcels, bounded by n.

Name:	Suhas Kamath	**Assignment No:**	Assignment 1
SR No:	06-18-01-10-51-25-1-25945	**Course Code:**	DS221
Email ID:	suhaskamath@iisc.ac.in	**Course Name:**	Intro. to Scalable Systems
Date:	September 8, 2025	**Term:**	AUG 2025

Experimental Setup

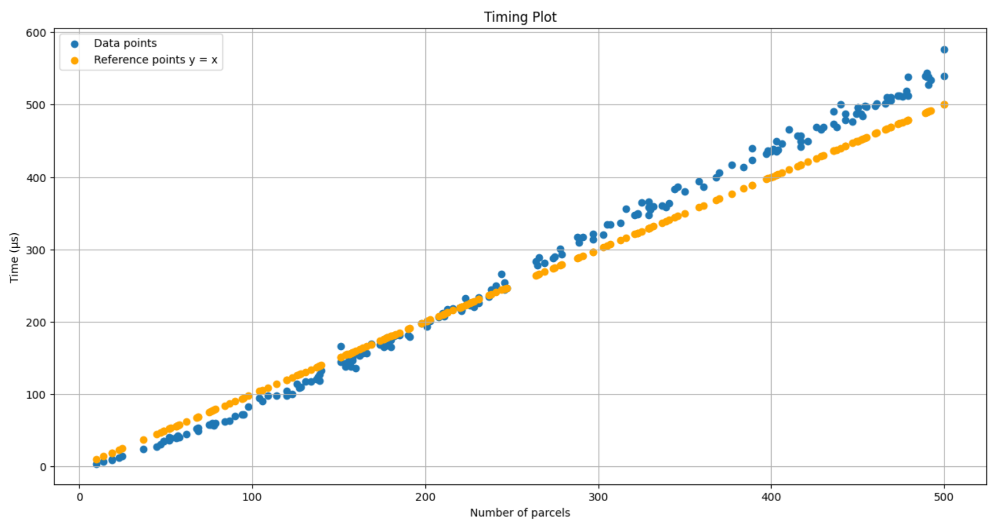
I first wrote the code on my personal computer. I had downloaded version 14 of the C++ compiler, and was using several new features provided in this version. This includes declaring variables with the auto keyword, and for-each loops. When I got access to the cluster, I realized that the cluster was still using C++ version 8.5. I then had to manually degrade the code, such that it can run on the cluster.
The code was then executed on the cluster provided. I created two files, namely q1_input.csv and q1_output.csv. I then wrote a program that does the following:

1. Reads one input from the q1_input.csv file.
2. Processes it using the function present in user_code.h.
3. Compares the output generated by the user_code.h file with the real output present in q1_output.csv.
4. Notes down the time taken with respect to the number of nodes in the q1_timings.csv file.

I then moved the q1_timings.csv file to my personal computer, where I drew a graph based on the data in the csv file. For this, I decided to use Python and its matplotlib library, which I am familiar with. The plotted graph is attached the next section.

Empirical Observations

The following time graph was obtained:



Most of these experiments are in the average case, which is why the reference points are in the order of $O(n)$.

This algorithm is scalable because $O(n \log n)$ is a decent time complexity. It is most definitely better in scalability terms than the obvious approach, which has a time complexity of $O(n^2)$.

Additional Insights

Some further optimizations can be made if the input data was sorted on basis of ID. In such a case, no sorting shall be required. This would essentially enhance the time complexity to $O(n)$. The space complexity, however, shall remain at $O(n)$.

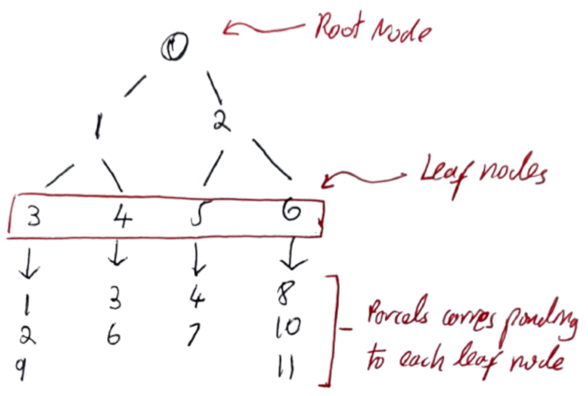
Name: Suhas Kamath
SR No: 06-18-01-10-51-25-1-25945
Email ID: suhaskamath@iisc.ac.in
Date: September 8, 2025
Assignment No: Assignment 1
Course Code: DS221
Course Name: Intro. to Scalable Systems
Term: AUG 2025

Question 2

Solution Approach

In this problem, we are given the pre-order and in-order traversals of a binary tree, along with a list of parcels stored at each leaf. We are also given several queries, each asking for the highest junction in the tree where all the parcels in that query are present in the sub-tree of that junction.

An example of such a tree is as follows. This tree has been constructed using the input in the sample_tests folder:



The most obvious first thing to do is to reconstruct the tree using the in-order and pre-order traversals. We can assign parcels to their respective leaves. Then, to actually solve the problem, we can propagate the parcels upward to each ancestor. Then, we can check the queries against each node to get the solution. This solution, however is very wasteful because we are checking every single node.

We can optimize the code further by using the fact that each parcel can be present only on one leaf node. Just like before, we reconstruct the tree from the pre-order and in-order traversals. Next, we pre-process the tree using a depth-first search to record the depth of each node and to fill a binary lifting table. This binary lifting table is used to quickly access the ancestor of a particular node. After this, we traverse the leaves and map every parcel to the ID of the leaf node where it resides.

Now that the tree is constructed and the parcels are assigned to the respect leaf nodes, we can start processing queries. For each query, we look up the leaf nodes corresponding to the parcels in the query and compute their lowest common ancestor. The node we obtain at the end is the junction we are looking for.

This method avoids the redundancy of storing parcel sets at many nodes and essentially reduces the problem to a series of LCA computations.

Name:	Suhas Kamath	**Assignment No:**	Assignment 1
SR No:	06-18-01-10-51-25-1-25945	**Course Code:**	DS221
Email ID:	suhaskamath@iisc.ac.in	**Course Name:**	Intro. to Scalable Systems
Date:	September 8, 2025	**Term:**	AUG 2025

Time and Space Complexity Analysis

Time Complexity

First, let us analyze the program to determine its time complexity. The code can be divided into the following major sections, namely:

1. Building the tree from the in-order and pre-order traversals.
2. Traversing the tree in DFS to assign a ID to each node.
3. Pre-processing the required parameters for Lowest Common Ancestor (LCA).
4. Mapping each parcel to a leaf node.
5. Solving each query one by one.

The detailed time complexity analysis is as follows:

Let n = total number of nodes (including leaf and internal nodes), p = total number of parcels across all leaf nodes and q = number of queries.

Section	Best Case	Average Case	Worst Case	Reasoning
Section 1	$O(n)$	$O(n)$	$O(n)$	Tree construction from traversals is always linear.
Section 2	$O(n)$	$O(n)$	$O(n)$	A standard DFS to compute depths visits each node once.
Section 3	$O(n \log n)$	$O(n \log n)$	$O(n \log n)$	Building the binary lifting table is a mandatory, fixed-cost pre-computation.
Section 4	$O(n)$	$O(n)$	$O(n)$	A full tree traversal is required to map parcels to leaves.
Section 5	$O(\log n)$	$O(q \cdot p \cdot \log n)$	$O(q \cdot p \cdot \log n)$	The best case is a single, two-parcel query. The average and worst cases scale with query load.
Total	$O(n \log n)$	$O(n \log n + q \cdot p \cdot \log n)$	$O(n \log n + q \cdot p \cdot \log n)$	The total time is dictated by the mandatory pre-processing and the total number of LCA lookups.

Space Complexity

There are four main data segments. They are as follows:

- Tree
- Binary lifting table (LCA Table)
- Parcel mapping
- Query storage

The detailed space complexity analysis is as follows:

Data Structure	Best Case	Average Case	Worst Case	Reasoning
Tree	$O(n)$	$O(n)$	$O(n)$	Storing the tree nodes, their depths, and index mappings requires space linear to the number of nodes, n.
Binary lifting table (LCA Table)	$O(n \log n)$	$O(n \log n)$	$O(n \log n)$	The binary lifting table is the largest data structure, and its size is fixed based on n, regardless of the case.
Parcel mapping	$O(p)$	$O(p)$	$O(p)$	Space is required to map every unique parcel to its leaf node.
Query storage	$O(q)$	$O(q)$	$O(q)$	Space is needed for the final output vector, where q is the number of queries.
Total	$O(n \log n + p)$	$O(n \log n + p)$	$O(n \log n + p)$	The space complexity is consistently dominated by the pre-computation data structures, primarily the LCA table.

Experimental Setup

I followed a similar setup as Question 1 for the experimental setup.

I first wrote the code on my personal computer. I had downloaded version 14 of the C++ compiler, and was using several new features provided in this version. This includes declaring variables with the auto keyword, and for-each loops. When I got access to the cluster, I realized that the cluster was still using C++ version 8.5. I then had to manually degrade the code, such that it can run on the cluster.

The code was then executed on the cluster provided. I created two files, namely q2_input.csv and q2_output.csv. I then wrote a program that does the following:

1. Reads one input from the q2_input.csv file.
2. Processes it using the function present in user_code.h.
3. Compares the output generated by the user_code.h file with the real output present in q2_output.csv.
4. Notes down the time taken with respect to the number of nodes in the q2_timings.csv file.

I then moved the q2_timings.csv file to my personal computer, where I drew a graph based on the data in the csv file. For this, I decided to use Python and its matplotlib library, which I am familiar with.

Later, I realised that matplotlib is also capable of drawing three-dimensional graphs. I hence wrote some more Python code, in order to generate this 3-D graph.

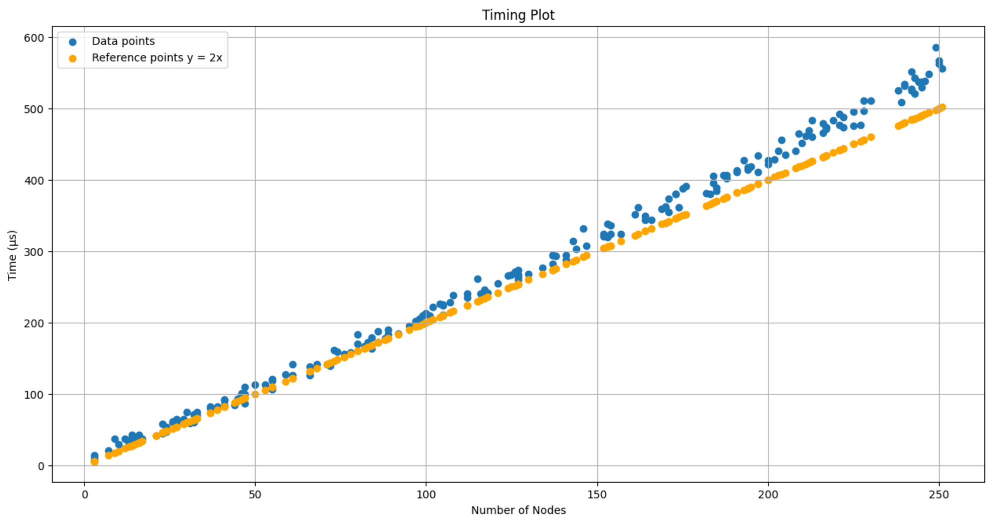
Both graphs are attached in the next section.

Empirical Observations

As mentioned in the above section, the time varies with respect to three variables:

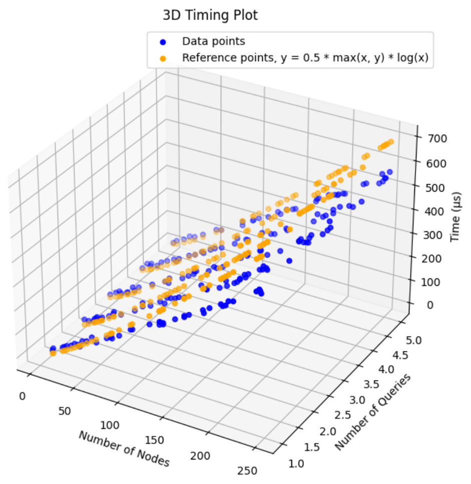
1. Number of nodes (junctions) in the tree
2. Number of queries
3. Total number of parcels

The following time graph was obtained when the time was plotted against the number of nodes:



Name:	Suhas Kamath	**Assignment No:**	Assignment 1
SR No:	06-18-01-10-51-25-1-25945	**Course Code:**	DS221
Email ID:	suhaskamath@iisc.ac.in	**Course Name:**	Intro. to Scalable Systems
Date:	September 8, 2025	**Term:**	AUG 2025

The following three-dimensional graph was obtained when the time was plotted against the number of nodes and number of queries:



Because we are unable to draw a four-dimensional graph, we are assuming that the number of parcels is constant. Furthermore, it was noticed that the number of parcels has a lesser effect on the overall time relative to the number of nodes and queries.

Additional Insights

Currently, we are using binary lifting, such that each lowest common ancestor problem is answered in $O(n \log n)$. If there are much more queries than nodes, i.e. $q >> n$, we can use Euler Tour and Range Minimum Queries such that each lowest common ancestor problem can be solved in constant time.

Another improvement that can be made is that we can modify the code such that k-ary trees are supported, instead of just binary trees.

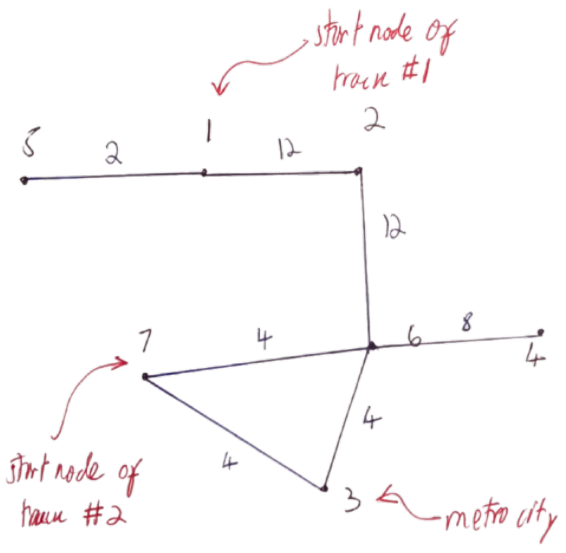
Name:	Suhas Kamath	**Assignment No:**	Assignment 1
SR No:	06-18-01-10-51-25-1-25945	**Course Code:**	DS221
Email ID:	suhaskamath@iisc.ac.in	**Course Name:**	Intro. to Scalable Systems
Date:	September 8, 2025	**Term:**	AUG 2025

Question 3

Solution Approach

The first step in the solution is to represent the network efficiently. The input is given as an edge list, and we convert it into an graph adjacency list so that we can quickly traverse all neighbors of a given node.

An example of such a graph is as follows. This graph has been constructed using the input in the sample_tests folder:



The next step is to determine travel times for both trucks. We will use Dijkstra's algorithm for this. However, an issue arises with this. The original (vanilla) Dijkstra's algorithm does not have support for the booster concept of the question. Hence, we need to modify the algorithm. In this modified algorithm, each state includes both the node and whether the booster has already been used. Thus, the distance array becomes dist[node][booster_state].

We will have to execute this modified Dijkstra algorithm twice:

1. Once from node 1 (the starting city of the first truck)
2. Once from node n (the starting city of the second truck)

When moving from a node to the next node, there are two cases:

1. The truck has not used its booster yet:
   (a) If the current node is a metro city, we allow the truck to apply the booster. After this, whenever we traverse an edge, we will cut down the travel time to half.
   (b) If the current city is not a metro city, we have no choice but to use the complete travel time.
2. The truck has used its booster: In this case, there is no point in checking if the current city is a metro city or not. We just continue to the next node, with half of the travel time.

At last, we can determine the meeting point of the trucks. Basically, we calculate the meeting time of the trucks at every single node. Because both trucks have to be there at the node to meet, the meeting time is basically max of the time each truck takes to get there. From this list, we basically identify the minimum time taken. This will be the fastest that the two trucks can meet at a node.

Name:	Suhas Kamath	**Assignment No:**	Assignment 1
SR No:	06-18-01-10-51-25-1-25945	**Course Code:**	DS221
Email ID:	suhaskamath@iisc.ac.in	**Course Name:**	Intro. to Scalable Systems
Date:	September 8, 2025	**Term:**	AUG 2025

Time and Space Complexity Analysis

Time Complexity

First, let us analyze the program to determine its time complexity. The code can be divided into the following major sections, namely:

1. Building the graph by converting from list of edges to adjacency list.
2. Running the modified Dijkstra algorithm.
3. Finding the meeting time at each node, and taking the minimum of these times.

The detailed time complexity analysis is as follows:

Let n = total number of nodes (cities) and m = total number of edges (roadways) between two nodes (cities).

Section	Best Case	Average Case	Worst Case	Reasoning
Section 1	$O(n+m)$	$O(n+m)$	$O(n+m)$	Building the adjacency list is always linear in the size of the graph.
Section 2	$O((n+m)\log n)$	$O((n+m)\log n)$	$O((n+m)\log n)$	The complexity of Dijkstra's with a priority queue is consistent across best, average, and worst cases. This is the dominant operation.
Section 3	$O(n)$	$O(n)$	$O(n)$	A linear scan through all cities is always required to find the optimal meeting point.
Total	$O((n+m)\log n)$	$O((n+m)\log n)$	$O((n+m)\log n)$	The overall complexity is consistently determined by the two runs of Dijkstra's algorithm.

Space Complexity

There are three main data segments. They are as follows:

- Adjacency list
- Distance table
- Priority Queue

Name:	Suhas Kamath	**Assignment No:**	Assignment 1
SR No:	06-18-01-10-51-25-1-25945	**Course Code:**	DS221
Email ID:	suhaskamath@iisc.ac.in	**Course Name:**	Intro. to Scalable Systems
Date:	September 8, 2025	**Term:**	AUG 2025

Data Structure	**Best Case**	**Average Case**	**Worst Case**	**Reasoning**
Adjacency List	$O(n+m)$	$O(n+m)$	$O(n+m)$	The space to store the graph is always proportional to the number of its nodes and edges.
Distance Table	$O(n)$	$O(n)$	$O(n)$	Two tables of size $n \times 2$ are needed to store distances from the start and end nodes.
Priority Queue	$O(n)$	$O(n+m)$	$O(n+m)$	In the worst case, the priority queue can hold an entry for each edge. In sparse graphs, this is $O(n)$.
Total Asymptotic	$O(n+m)$	$O(n+m)$	$O(n+m)$	The space complexity for all cases is determined by the size of the input graph, which is needed to store the adjacency list.

Experimental Setup

I followed a similar setup as Question 1 and 2 for the experimental setup.

I first wrote the code on my personal computer. I had downloaded version 14 of the C++ compiler, and was using several new features provided in this version. This includes declaring variables with the auto keyword, and for-each loops. When I got access to the cluster, I realized that the cluster was still using C++ version 8.5. I then had to manually degrade the code, such that it can run on the cluster.
The code was then executed on the cluster provided. I created two files, namely q3_input.csv and q3_output.csv. I then wrote a program that does the following:

1. Reads one input from the q3_input.csv file.
2. Processes it using the function present in user_code.h.
3. Compares the output generated by the user_code.h file with the real output present in q2_output.csv.
4. Notes down the time taken with respect to the number of nodes in the q3_timings.csv file.

I then moved the q3_timings.csv file to my personal computer, where I drew two graphs based on the data in the csv file. One is time taken vs. number of nodes. The other is time taken vs. number of edges. For this, I decided to use Python and its matplotlib library, which I am familiar with. I wrote some additional Python code, in order to generate a 3-D graph.

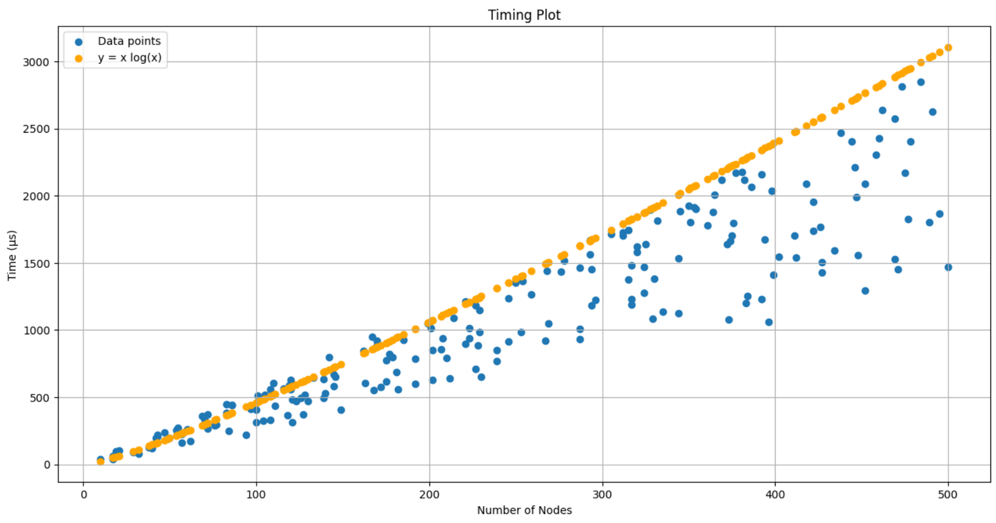
All three graphs are attached in the next section.

Empirical Observations

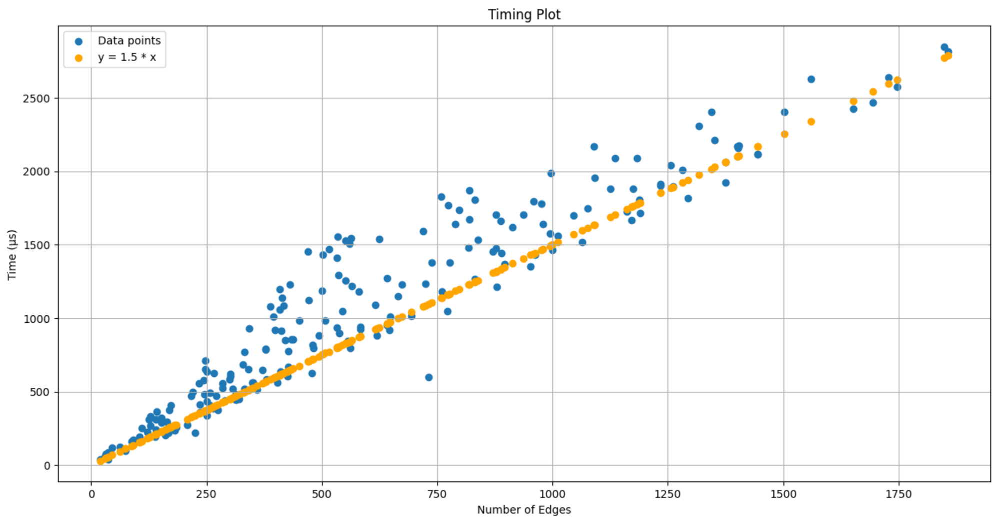
As mentioned in the above section, the time varies with respect to two variables:

1. Number of nodes (cities) in the graph
2. Number of edges

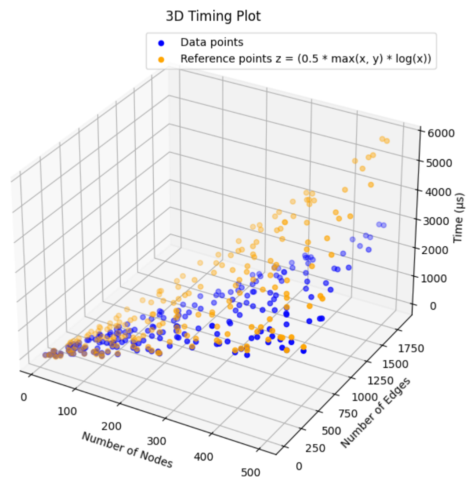
The following time graph was obtained when the time was plotted against the number of nodes (cities):



The following time graph was obtained when the time was plotted against the number of edges (roadway between cities):



The following three-dimensional graph was obtained when the time was plotted against the number of nodes and number of edges:



Additional Insights

If the algorithm is on an edge case, it can be further optimized to reduce the time. Some of the possible optimizations are as follows:

1. If the graph is disconnected, and the two trucks are on disconnected segments of the graph, there is no point in doing any calculation. The two trucks simply cannot meet, and we can return -1 instantly.
2. If the starting city of both trucks are metro cities, we can apply the boosters in the first step itself. After this, we do not need to accommodate for two different scenarios. Similarly, if none of the cities are metro cities, we can never apply the booster. Again, there shall be no need to accommodate two scenarios.