Data management project in Identifying Insufficient Data Coverage

by: Gili Weissberg, Noga Mendler

The paper "Identifying Insufficient Data Coverage in Databases with Multiple Relations" addresses a critical issue in data quality and fairness: insufficient data coverage in large relational databases. This issue is especially significant in the era of big data, where decision-making models and machine learning algorithms heavily rely on diverse datasets. Poor data coverage can introduce bias and inaccuracies in downstream decision-making systems.

We chose to use the **SF Bay Area Bike Share** dataset for our analysis due to its comprehensive and detailed information about bike trips, including start and end stations, trip durations, and start times. This dataset provides a rich foundation for understanding bike usage patterns in the San Francisco Bay Area.
For our analysis, we focused on the **first 100,000 records** of the dataset to ensure manageable data size while still capturing a representative sample of bike trip behaviors.

To better analyze and visualize the data, we structured it into a three-layer graph, organized hierarchically as follows:

1. **Start Stations Layer (Bottom Layer):**

This layer represents the starting stations from which bike trips originate. Each node in this layer corresponds to a unique start station, and the data includes metrics such as the number of trips originating from each station.

2. **Duration Ranges Layer (Middle Layer):**

Above the start stations layer, we have the duration ranges layer. This layer categorizes bike trips based on the duration of the rental period. Each node in this layer represents a specific range of rental durations (e.g., 0-60 minutes, 61-120 minutes, etc.).

3. **Hour Ranges Layer (Top Layer):**

The top layer represents the hour ranges during which bike trips begin. Each node in this layer corresponds to a specific time range (e.g., 6-12 AM, 12-6 PM, etc.)..

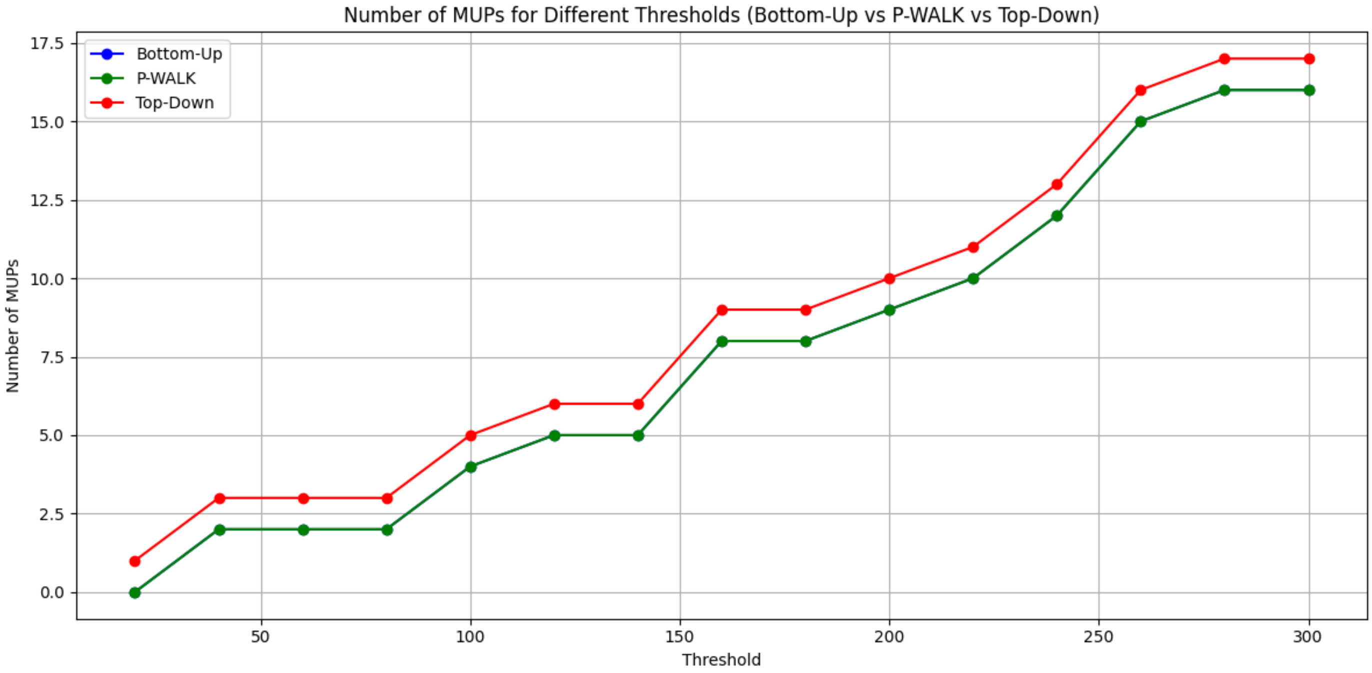
Experimental Results and Analysis

In our study, we developed and implemented five algorithms: P-WALK, Bottom-Up, Top-Down, Distinct Sampler, and Correlated Sampler. These algorithms were designed to analyze and extract meaningful patterns from the dataset, with P-WALK serving as the core method for identifying Maximally Uncovered Patterns (MUPs), while the Bottom-Up and Top-Down approaches provided complementary hierarchical traversal strategies.

The Distinct Sampler and Correlated Sampler were applied on P-WALK to ensure diverse and attribute-focused sampling, respectively. Together, these algorithms enabled a comprehensive evaluation of the dataset, allowing us to assess their performance and effectiveness in uncovering key insights.

Sanity Check:

It can be observed that all three algorithms discover nearly the same MUPs, as required, and the number of MUPs increases with the thresholds, as expected.

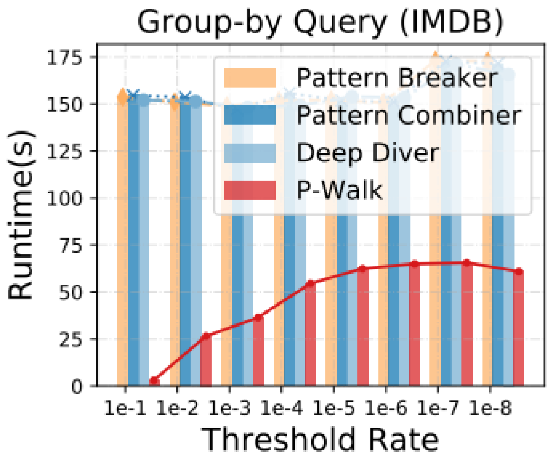


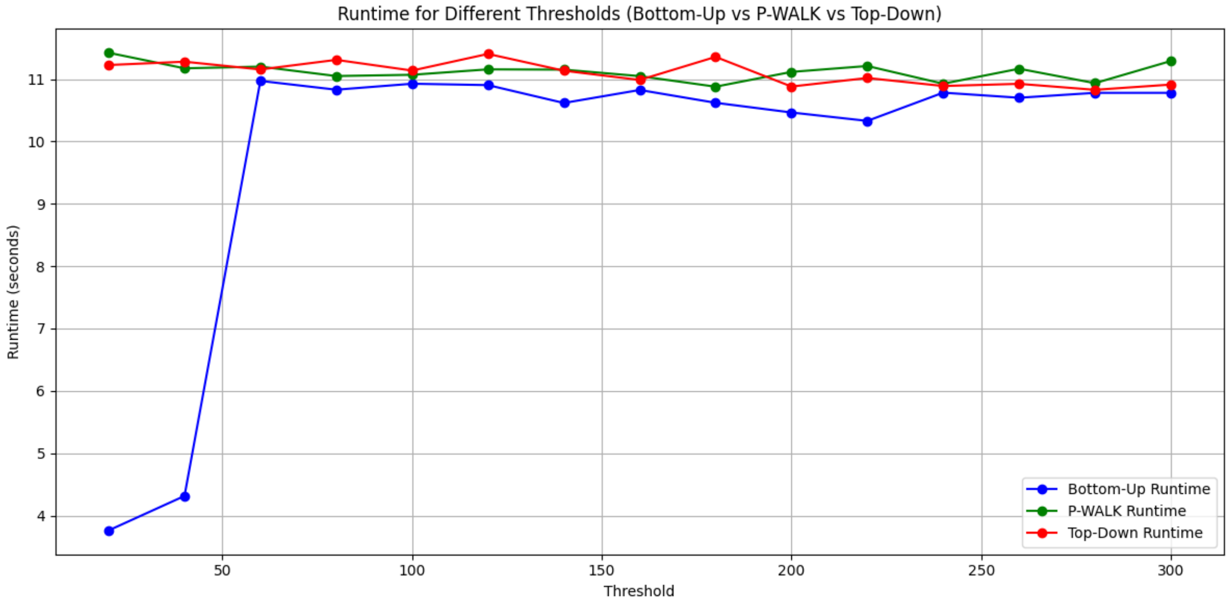
Experiments and Graphs Conducted Based on the Paper:

Evaluation Aspect	**Parameters Checked**	**Algorithms Evaluated**	**Datasets**
Runtime for MUP Identification	Use group by query ,Runtime under varying thresholds	Top-down (Pattern Breaker), Bottom-up, (Pattern combiner) , P-WALK	SF Bay Bike Share
Approximate MUP Identification and Accuracy	Classification accuracy for covered/uncovered patterns under varying thresholds	Distinct sampler, Correlated sampler	SF Bay Bike Share,
Approximate MUP Identification Efficiency	Runtime under varying sampling rates	Distinct sampler, Correlated sampler	SF Bay Bike Share,
MUP Distribution Analysis	MUP distribution across layers, Variation in MUP count under different thresholds	Top-down (Pattern Breaker), Bottom-up (Pattern Combiner), P-WALK	SF Bay Bike Share

Runtime for MUP Identification

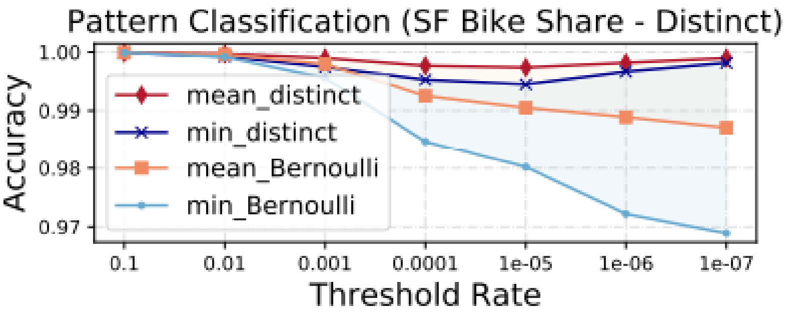
It can be observed that all the algorithms in both graphs are represented by horizontal lines parallel to the x-axis and remain constant. The reason why P-WALK in our graph also exhibits this behavior is due to the structure of our data graph, where most of the MUPs are concentrated in the leaves. As a result, it sometimes coincides with the Bottom-Up algorithm.

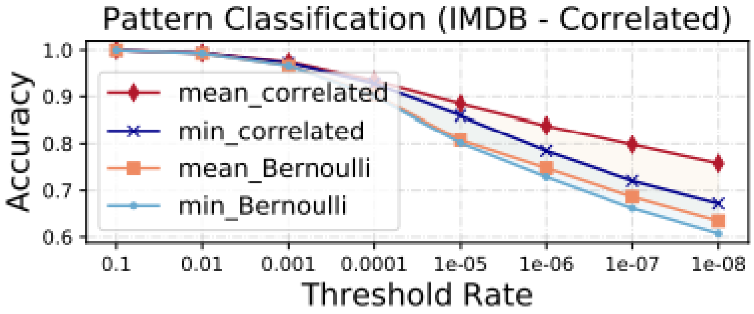


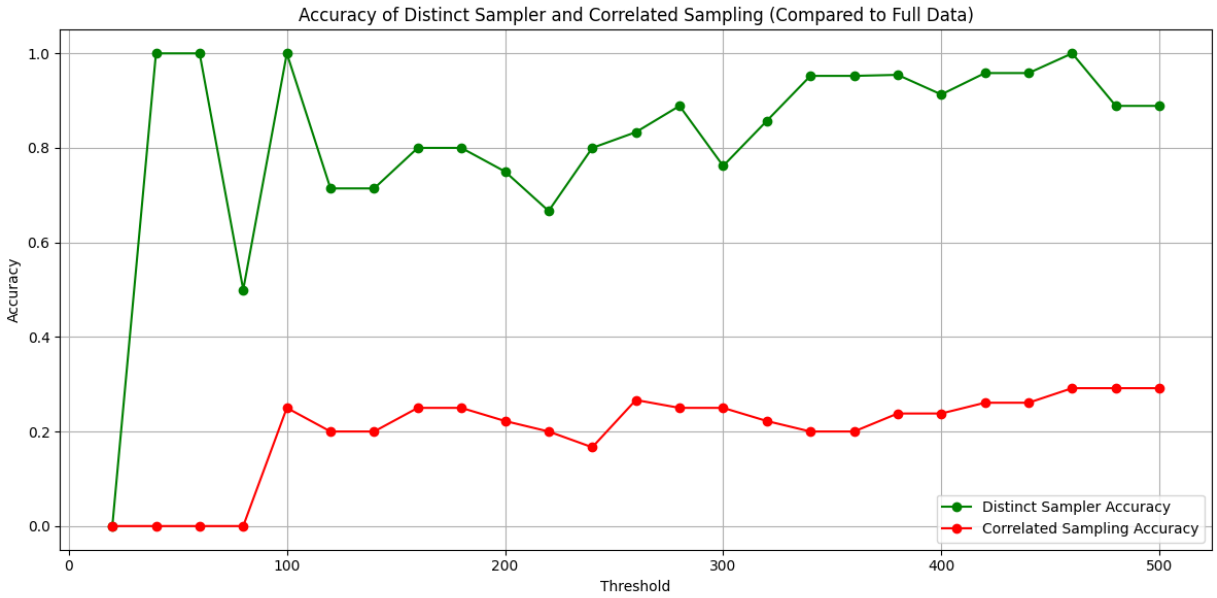


Approximate MUP Identification and Accuracy

It might be confusing and mistakenly thought that there is no correlation between the graph from the paper and the graph from our experiment. However, in both, the accuracy order increases as the threshold rises. The difference is simply that in the graph from the paper, the threshold order is arranged from largest to smallest.

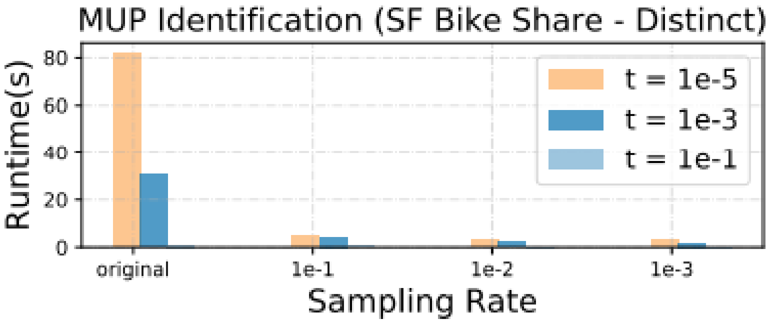


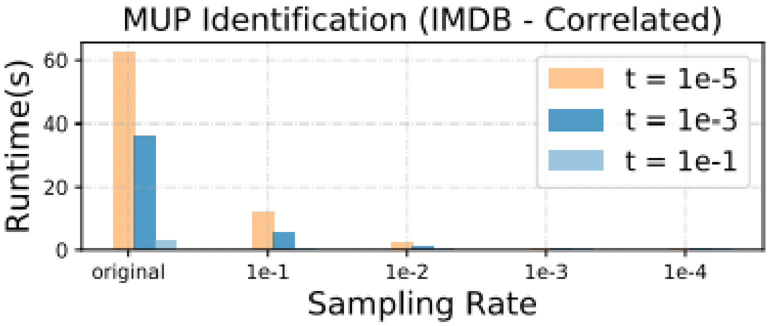


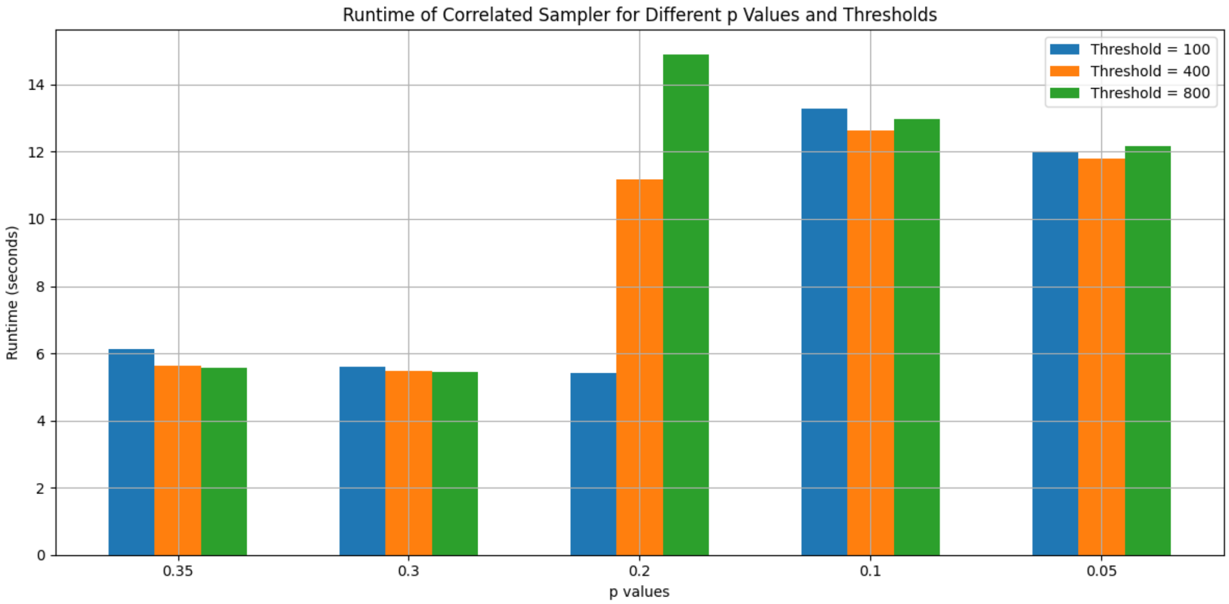


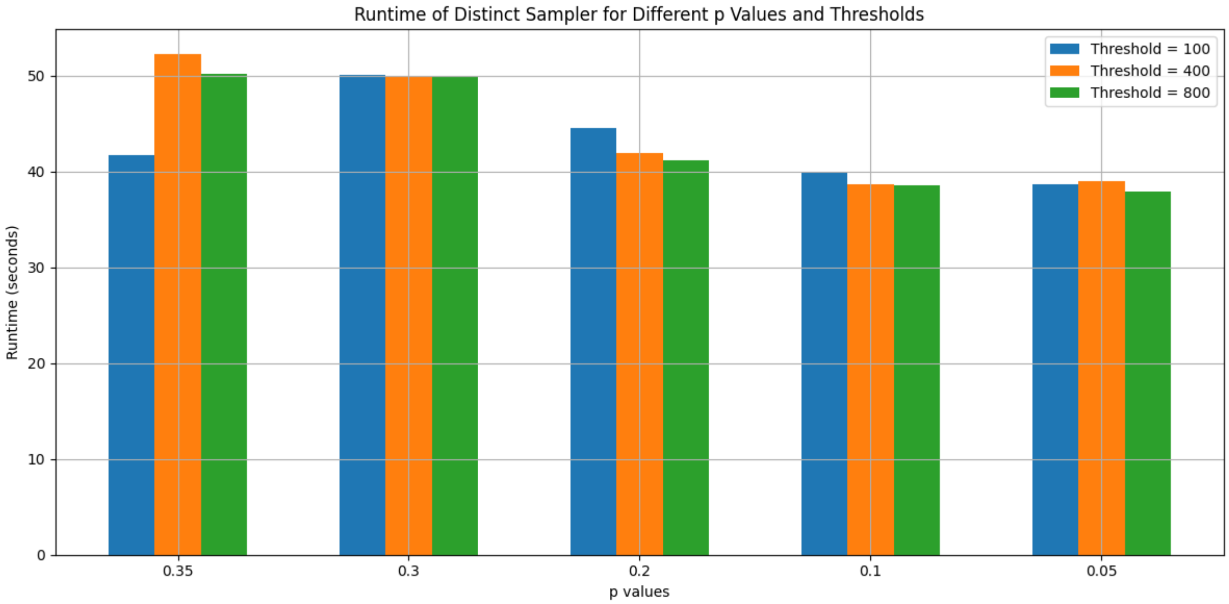
Approximate MUP Identification Efficiency

For the **Correlated Sampler** algorithm, a clear correlation can be observed between the experimental results and the graph from the paper, with a decrease in runtime as the value of **p** decreases. However, for the **Distinct Sampler**, no such correlation is evident. This can be explained by the fact that this graph is influenced by multiple parameters, where different weights are assigned during the decision-making process at each step. In addition to **p**, there are other dynamic parameters, such as **f** and **s**, that further complicate the relationship and contribute to the lack of a clear correlation.



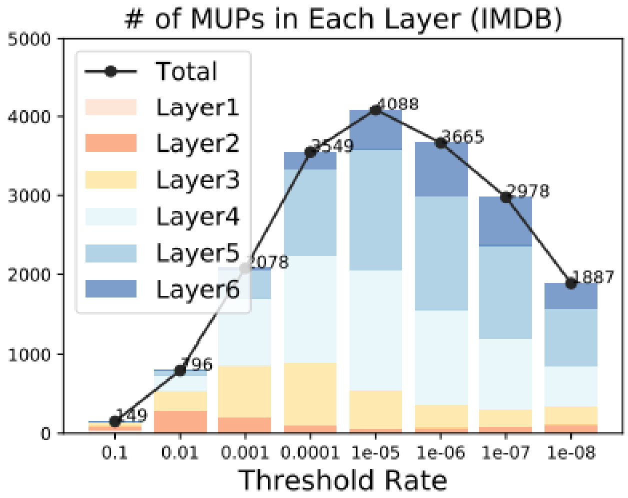


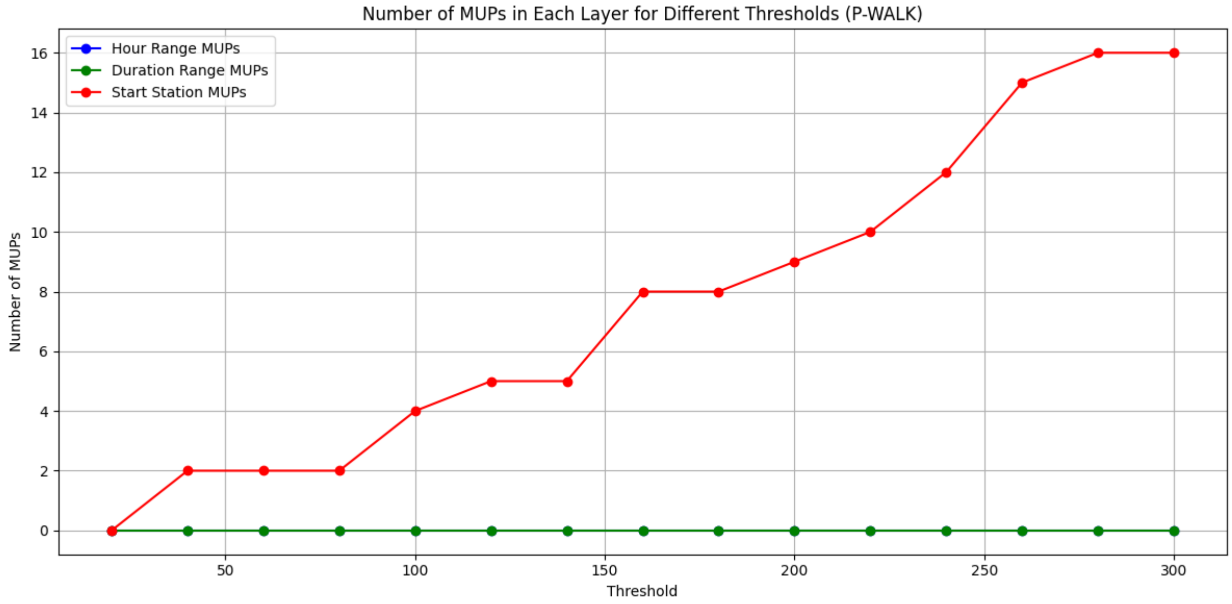

Runtime of Correlated Sampler for Different p Values and Thresholds
Threshold = 100
Threshold = 400
Threshold = 800
Runtime (seconds)
0
2
4
6
8
10
12
14
0.35
0.3
0.2
0.1
0.05
p values


Runtime of Distinct Sampler for Different p Values and Thresholds
Threshold = 100
Threshold = 400
Threshold = 800
Runtime (seconds)
0
10
20
30
40
50
0.35
0.3
0.2
0.1
0.05
p values

MUP Distribution Analysis

Due to the layered structure of the data graph, the vast majority of the MUPs (Maximally Uncovered Patterns) are concentrated in the leaf layer, which represents the start stations. This hierarchical arrangement causes the MUPs to predominantly appear at the lowest level of the graph, where the most granular and specific patterns are found.


# of MUPs in Each Layer (IMDB)
Total
Layer1
Layer2
Layer3
Layer4
Layer5
Layer6
1e-05
1e-06
1e-07
1e-08
Threshold Rate


Number of MUPs in Each Layer for Different Thresholds (P-WALK)
Hour Range MUPs
Duration Range MUPs
Start Station MUPs
Number of MUPs
Threshold