UNIVERSITÉ Concordia UNIVERSITY

Cloud-based Framework for Spatio-Temporal Trajectory Data Segmentation and Query

Master Thesis Defense

Huaqiang Kang

Supervisor: Yan Liu

Committee Members:

Overview

- This thesis presents a parallel trajectory segmentation framework that *scales horizontally* as the size of the trajectory data. This framework supports trajectory *segmentation*, *indexing*, *storage* and *query*.

- The study applies the framework to evaluate two metrics: "*trajectory similarity*" and "*chance of collision*" and one pattern: "*objects moving in groups*".

Outline

- Background
- Problem Statement, Objective, Contributions
- Research method
- Evaluation method
- Case Study
- Reliability and validity
- Future work and conclusion

Why Spatio-temporal Analysis Important

- **What is spatio-temporal data**
  data collected across time as well as space have at least one spatial and one temporal property.

- **Trajectory generation**
  Vehicle GPS, wearable device accelerometer, Video stream, Wifi network, RFID,GMS

- **Web2.0 Data Mining**

Geological dating app
Find the people you've crossed paths with.

Share rides and fare
uberPool matches riders heading in the same direction.

How Trajectory Data Are Represented

- **A trajectory can be treated as $T=<p_1, \ldots p_n>$ where $p_k=(id_k, loc_k, t_k, A_k)$ is the k^{th} position**

  id_K is the position identifier

  loc_k is the spatial location of the position

  t_k is the time at which the position was recorded

  A_k is the additional data

$Loc_k = (x, y)$

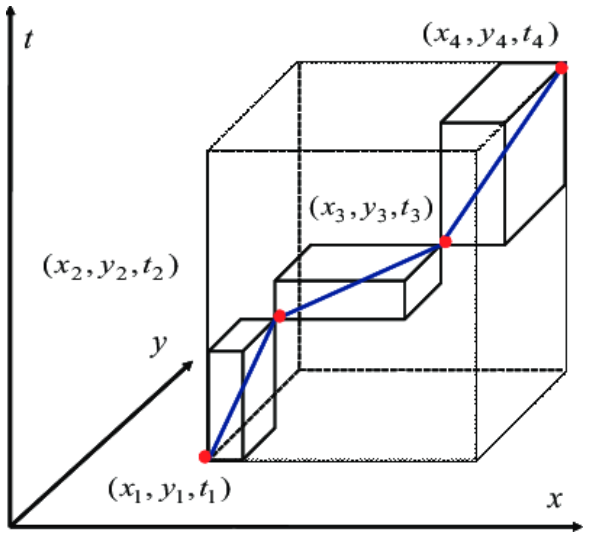
For GPS-based data

Loc_k = cell ID

For GMS-based data

How Trajectory Data Are Represented(Cont'd)

- **minimum bounding rectangle (MBR)**
- **is the bounding geometry formed by the minimum and maximum (X,Y) coordinates**
- **We use MBR for approximation spatial query, spatial indexing.**



Outline

- Background
- Problem Statement, Objective, Contributions
- Research method
- Evaluation method
- Case Study
- Reliability and validity
- Future work and conclusion

Problem Statement

- **How to partition and segment trajectories in a distributed framework ?**
- **What are the key factors of parallelism that affect the scalability of trajectory segmentation and queries ?**
- **How to evaluate the performance gain in parallel ?**

Objective Challenges

- A method to partition trajectories into multiple nodes.
- Requires an algorithm to segment trajectories in parallel.
- A parallel query method to ensure the results are identical to sequential query.
- A test case to evaluate the framework performance.

Existing Spatio-temporal Processing Framework

Features	GeoSpark	Simba	SpatialHadoop	Our Framework
Trajectory Processing	No	No	No	Yes
Spatial Indexing	Yes	Yes	Yes	Yes
Range Query, KNN, Distance Join	Yes	Yes	Yes	Yes
Trajectory Similarity, Collision Detection	No	No	No	Yes
Data Persistent	Index Only	Index Only	No	Yes

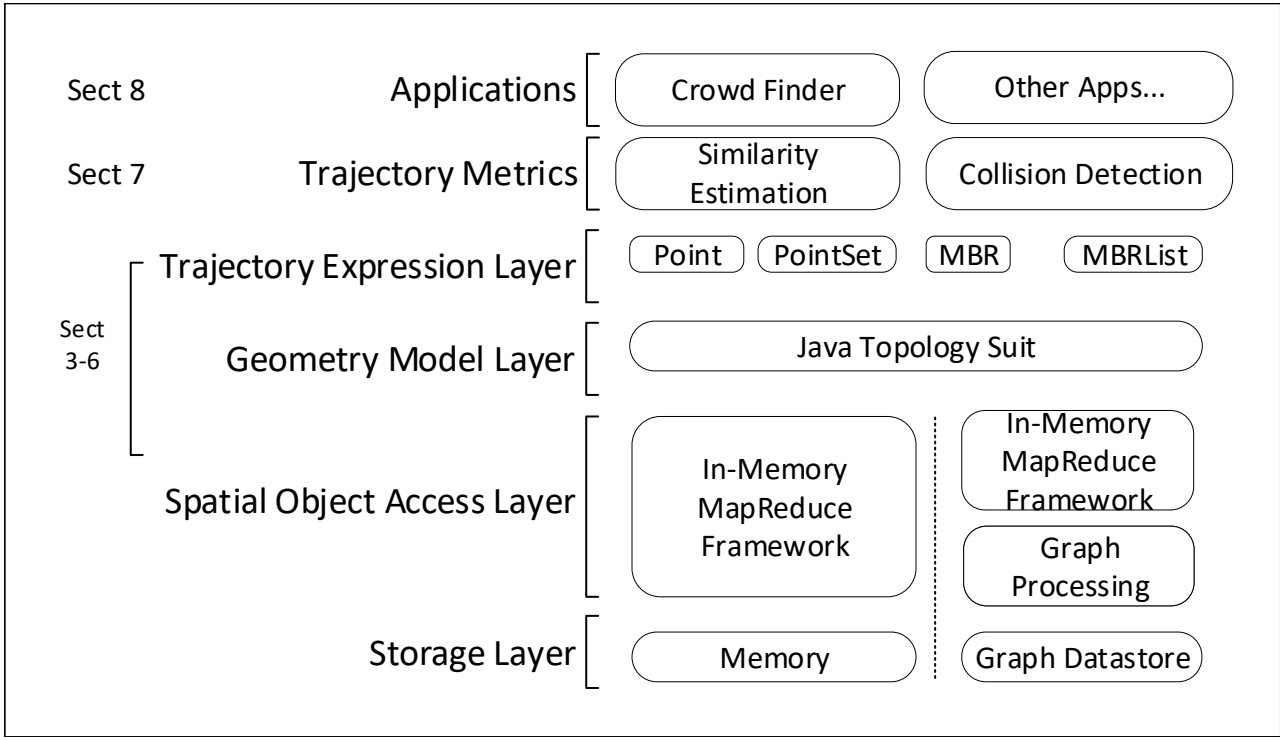
Our Contributions

- **Design an algorithm for parallel segmentation of trajectories in a distributed system.**
- **Develop a system workflow that queries trajectories segments in-memory and in a graph database.**
- **Define metrics that are further utilized by applications such crowd analysis.**
- **Develop data skew migrations solution to balance the workload as both the size of data and the computing cluster grow.**

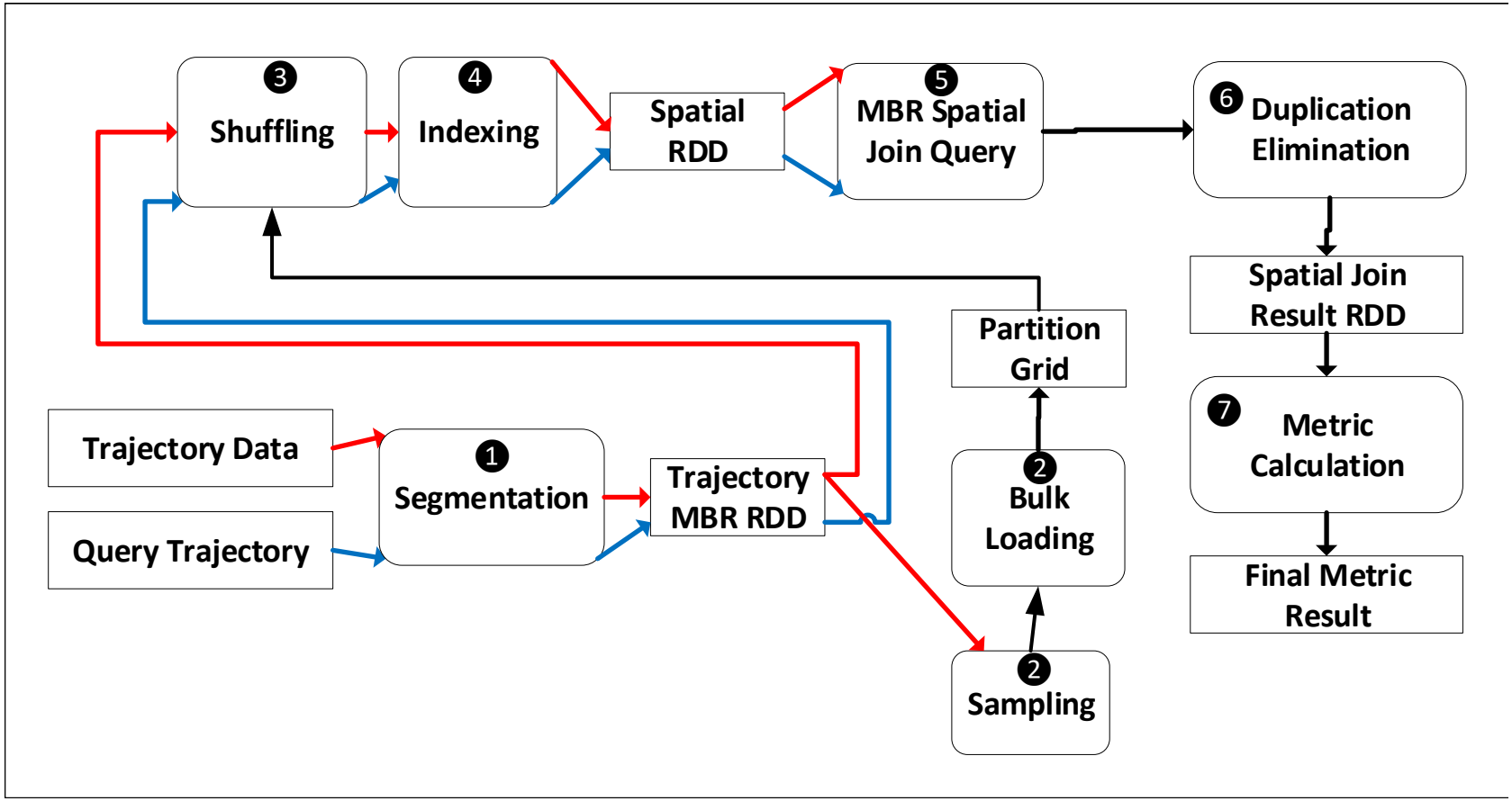
Outline

- Background
- Problem Statement, Objective, Contributions
- Research method
- Evaluation method
- Case Study
- Reliability and validity
- Future work and conclusion

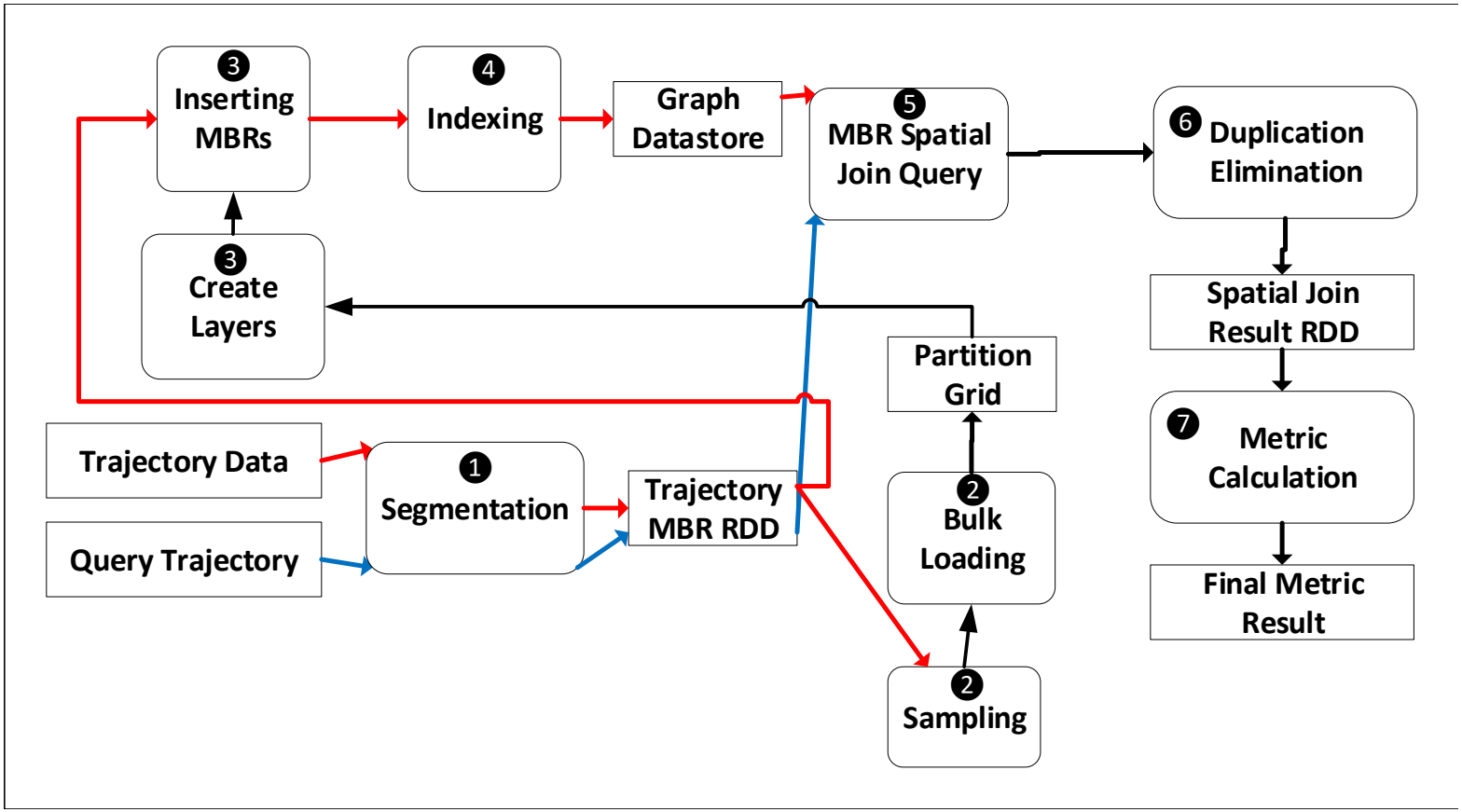
System Architecture



In-memory Processing Framework

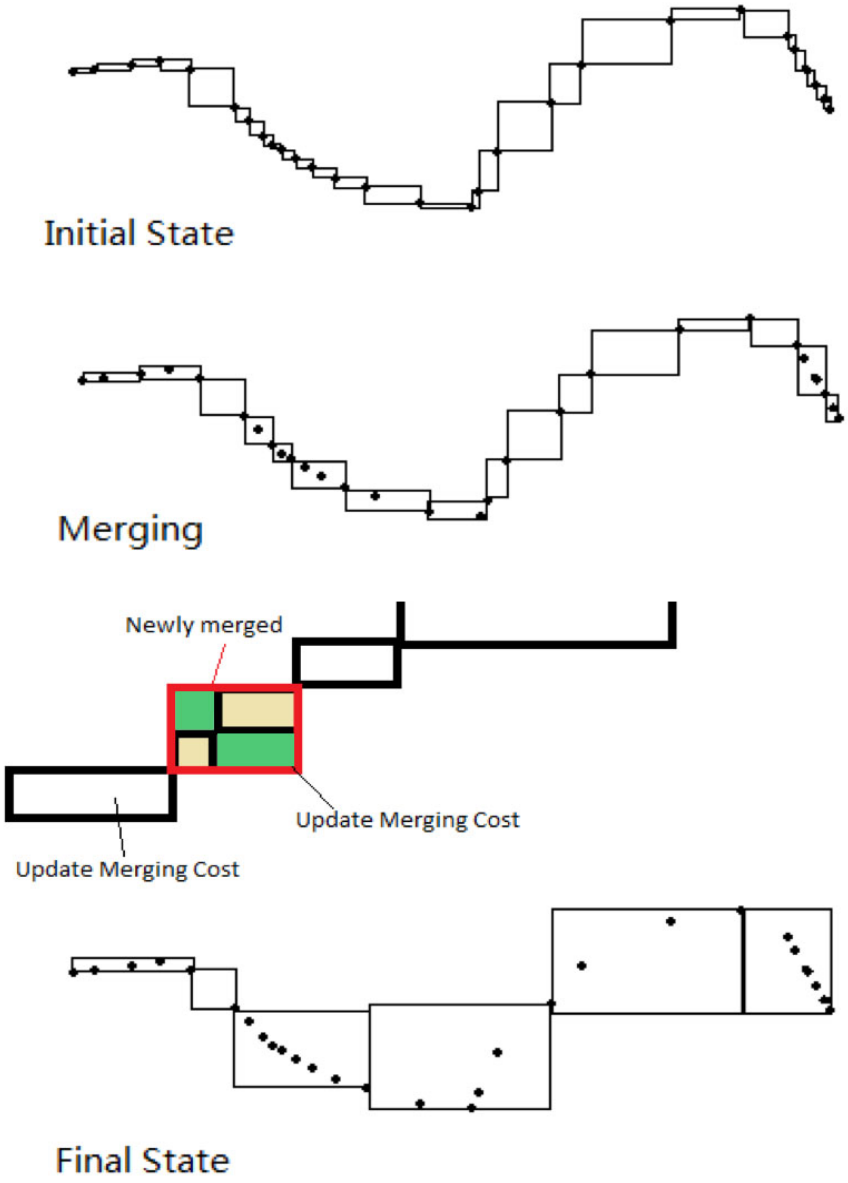


Graph Storage Processing Framework



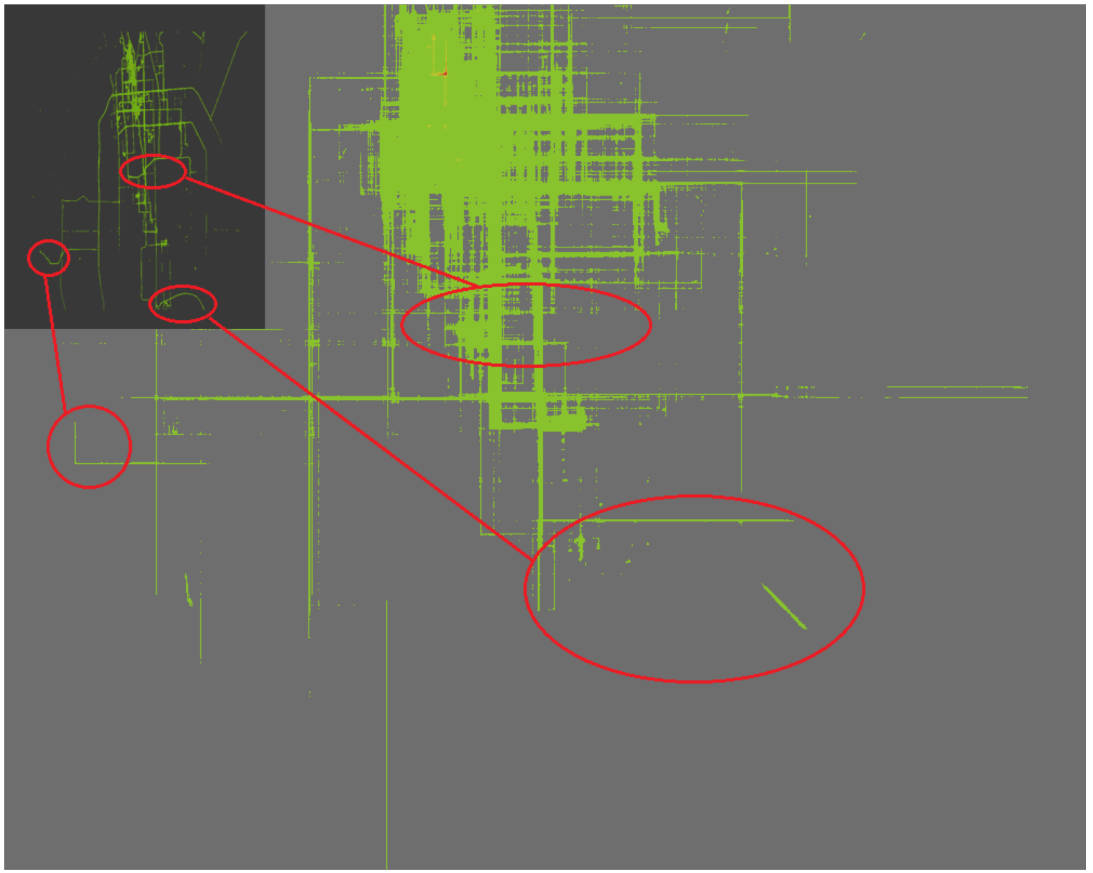
Greedy Split

The less area covered,
The more precise the segments are.

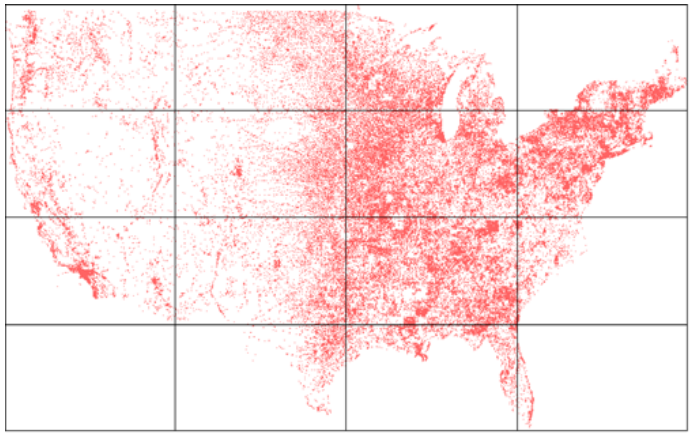


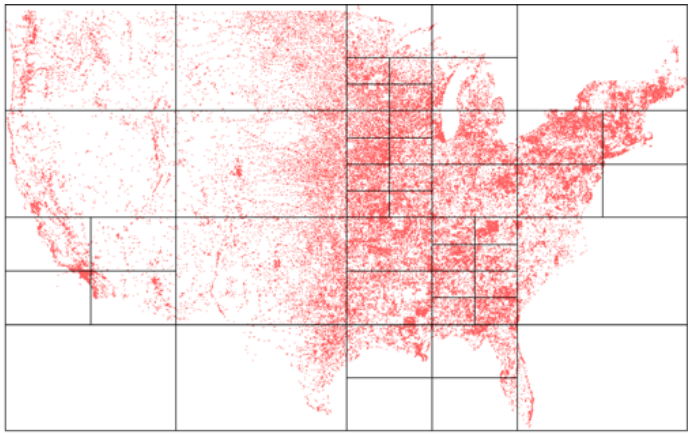
Trajectory Transforming Virtualization

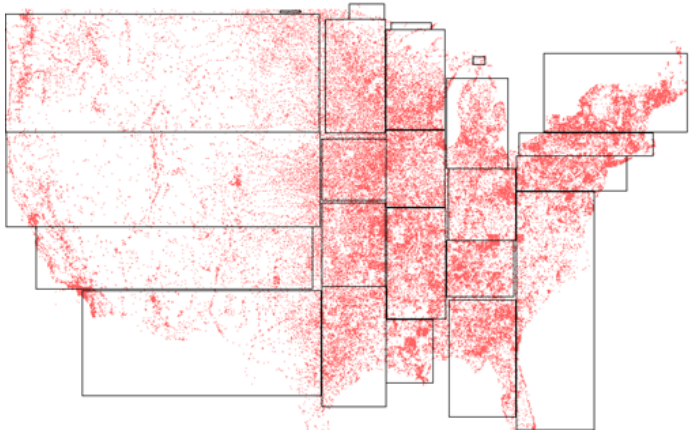
We draw the heatmap of trajectories and the heatmap of MBRs to compare illustrate the transformation from sub-trajectories to MBRs.



Spatial Partitioning







Uniform Grid

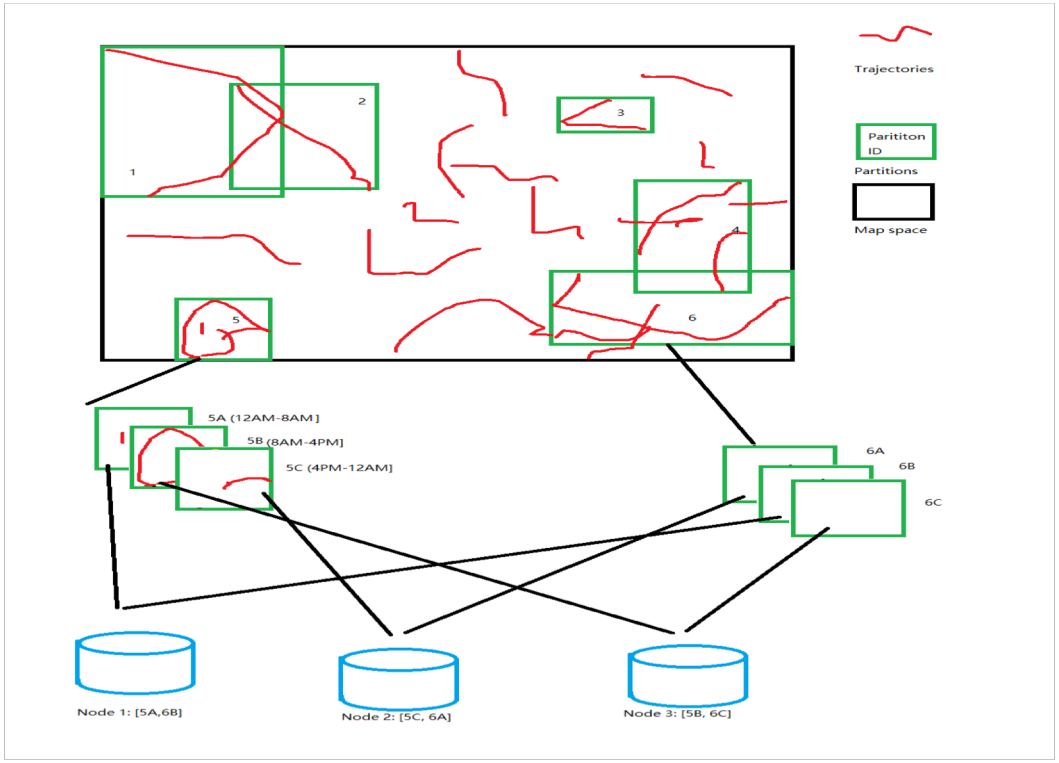
Quad-Tree Grid

R-Tree Grid

Spatial data management in apache spark: the GeoSpark perspective and beyond, 2018, Jia Yu et al.

Data Skew Solution

- A better 2-D partitioning solution is always desired.
- Introducing another attribute like time dimension.



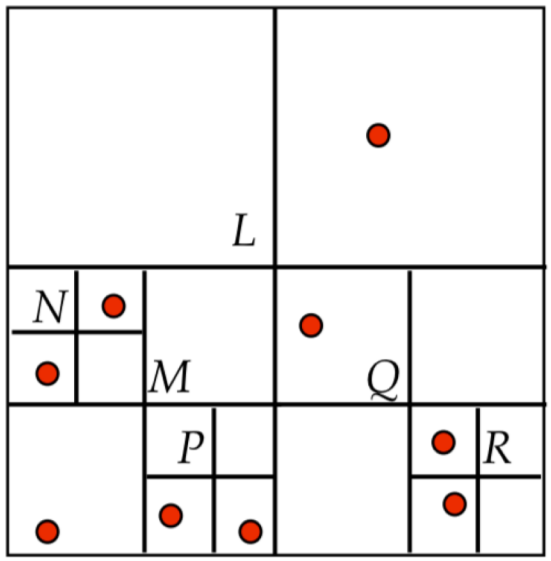
QuadTree

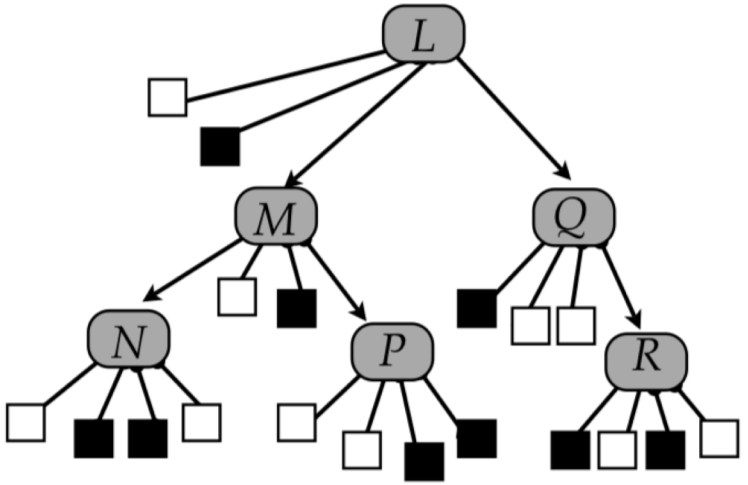
Recursively subdivide cells into 4 equal-sized subcells until a cell has only one point in it.

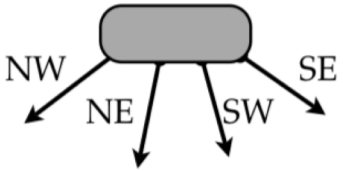
Each division results in a single node with 4 child pointers.

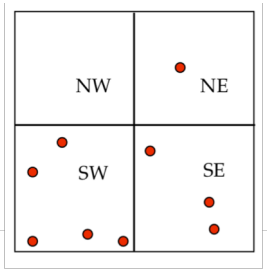
When cell contains no pints, add special "no-point" node.

When cell contains 1 point, add node containing point + data associated with that point.



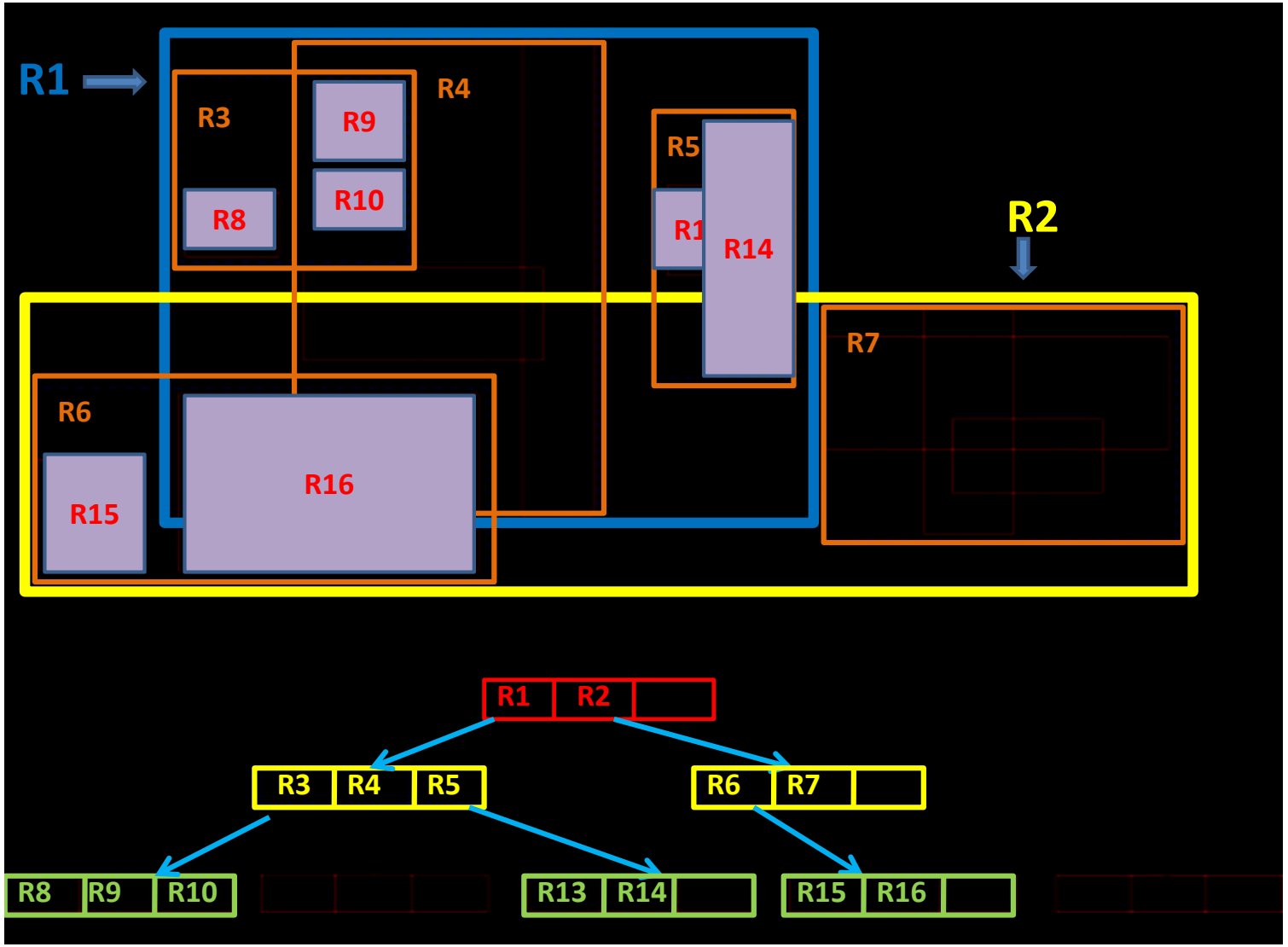






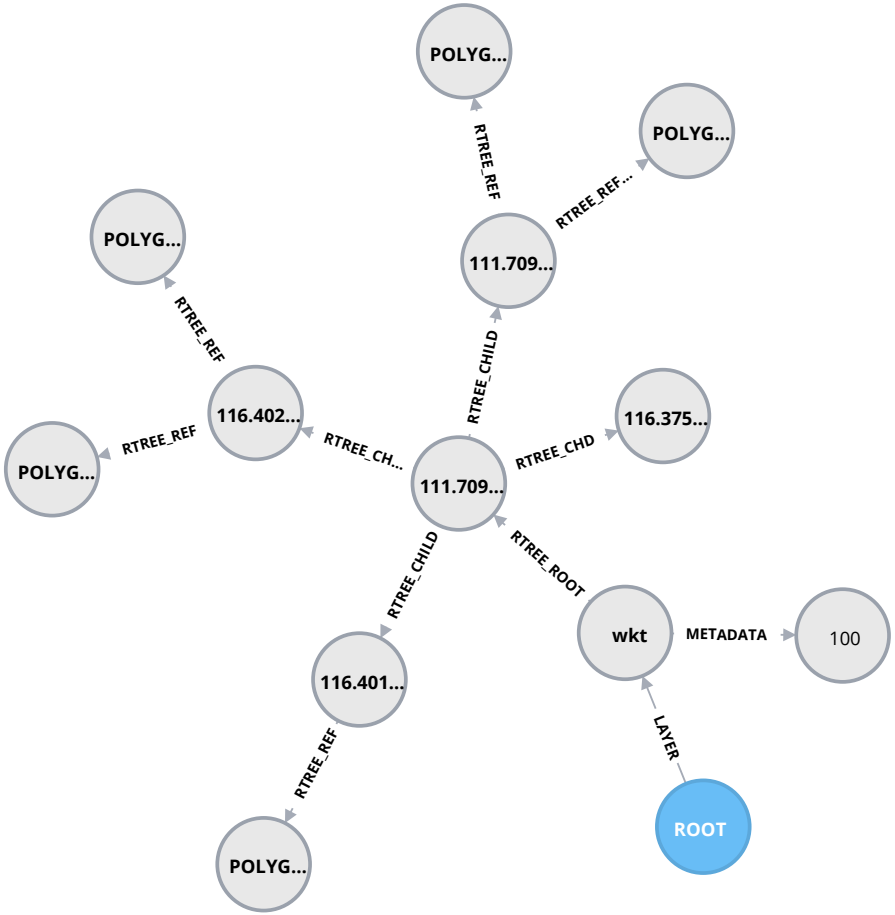
https://www.cs.cmu.edu/~ckingsf/bioinfo-lectures/quadtrees.pdf

R-Tree Indexing



R-trees (data structure), slideshare, mahesamrin, 2014

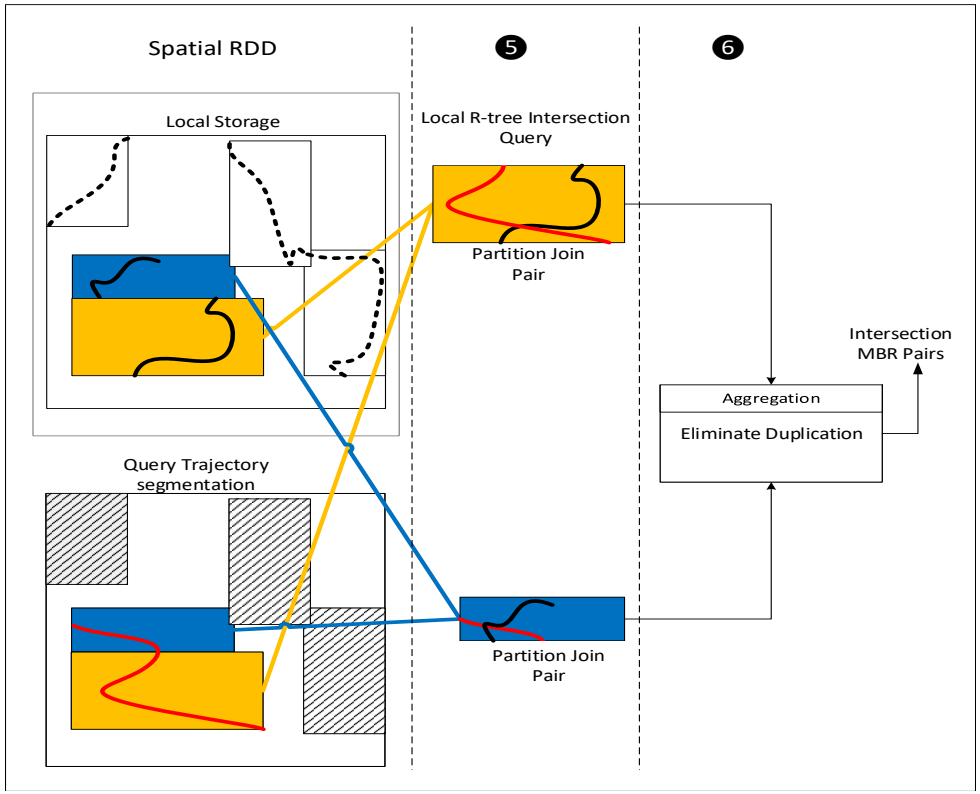
R-Tree in Graph Storage



- Root node in blue
- wkt layer connected to root
- More layers can be created
- Bounding Boxes linked to layers
- Objects linked to BBoxes

R-trees (data structure), slideshare, mahesamrin, 2014

Parallel Query

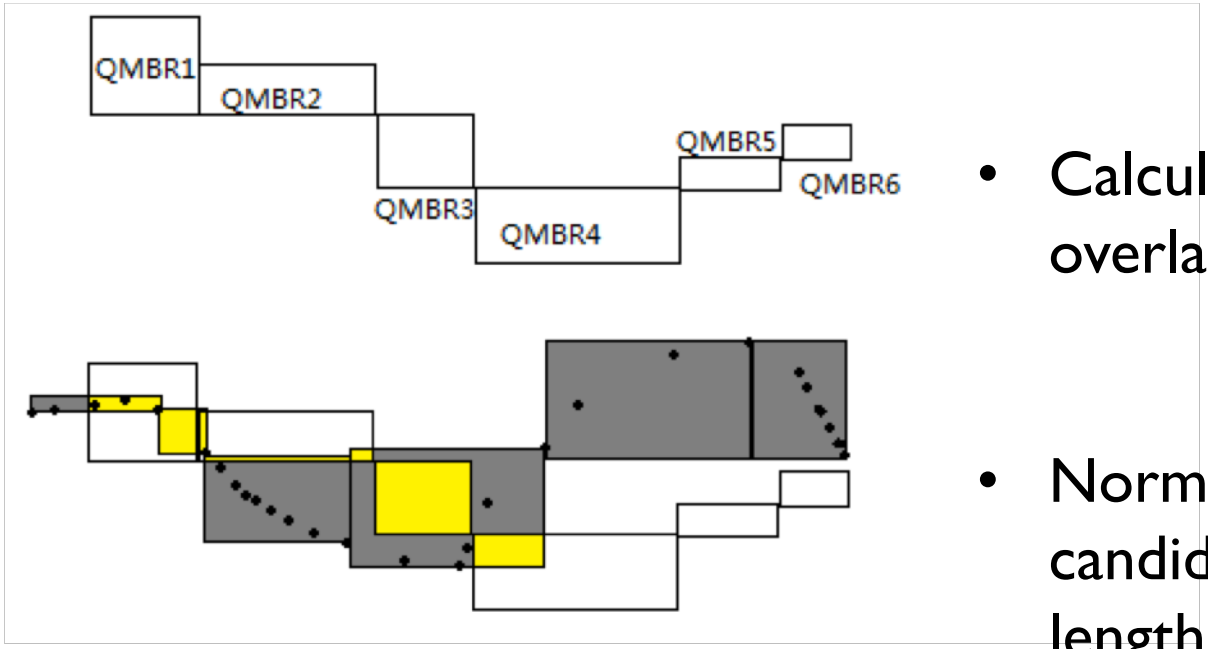


- Query Trajectory in Red
- Trajectory distributed in two partitions (Yellow and blue)
- Query proceeding in two partitions
- Aggregate to eliminate duplications

Outline

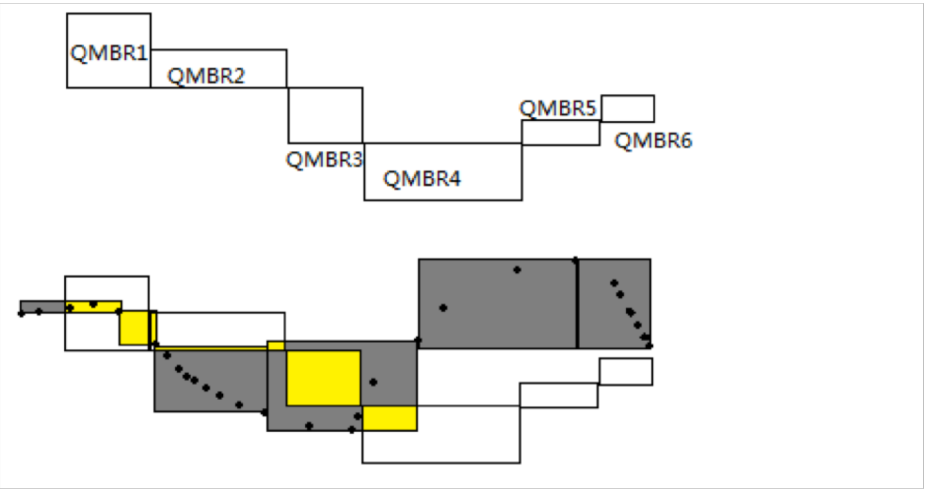
- Background
- Problem Statement, Objective, Contributions
- Research method
- Evaluation method
- Case Study
- Reliability and validity
- Future work and conclusion

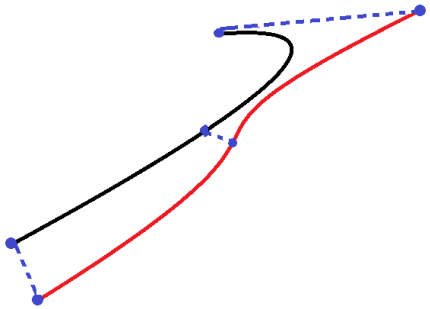
Similarity Estimation



- Calculate the yellow overlapping area.
- Normalized by candidate trajectory's length.

Collision Detection



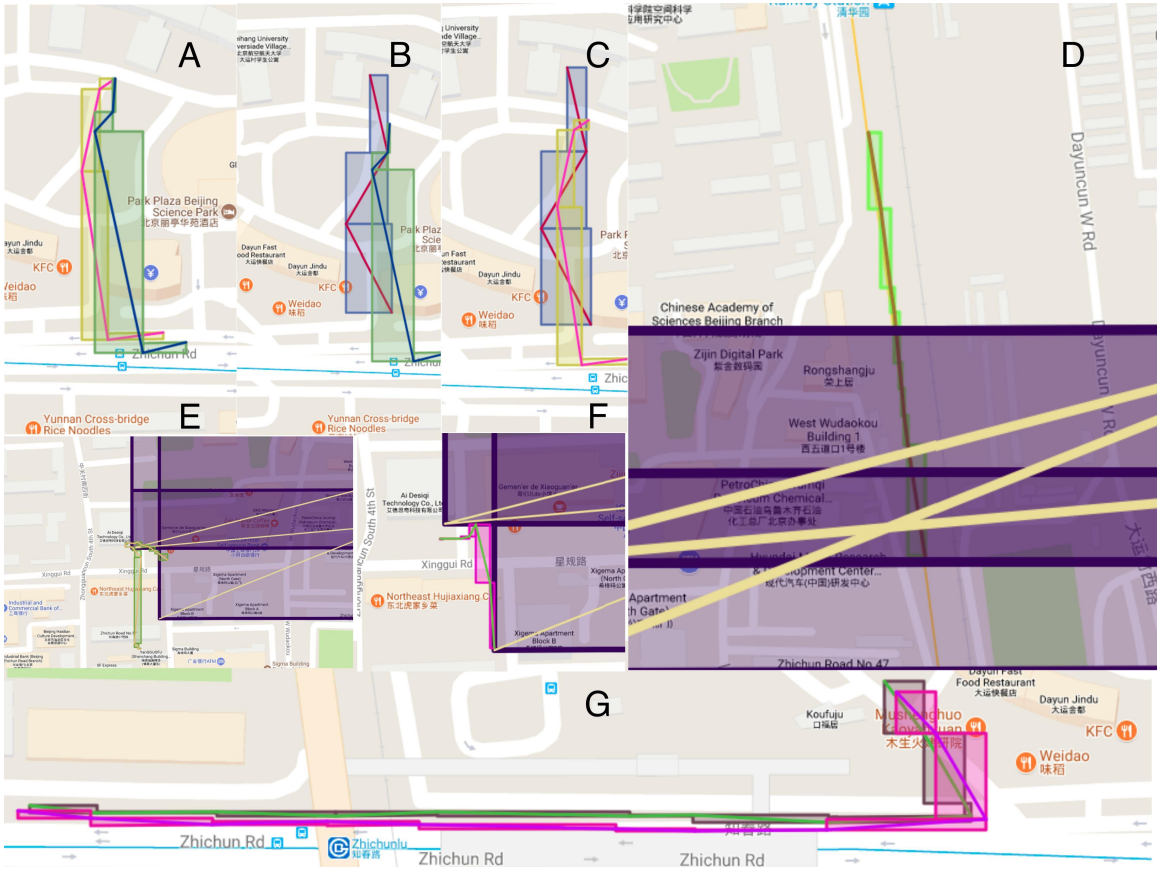


Similarity Estimation > 0.

Examine three checkpoints' distance.

- A threshold is set to detect the collision.

Collision Detection Result

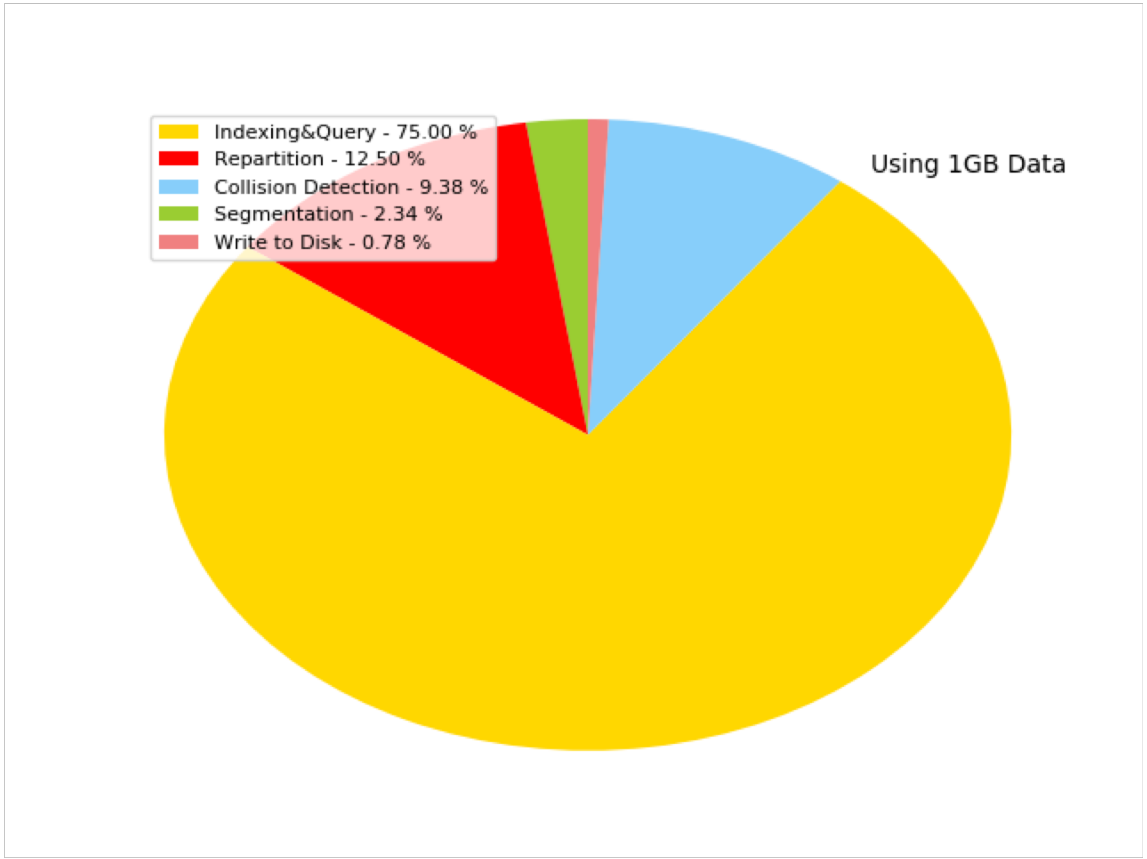


We select 13 trajectories to see the collision detection results.

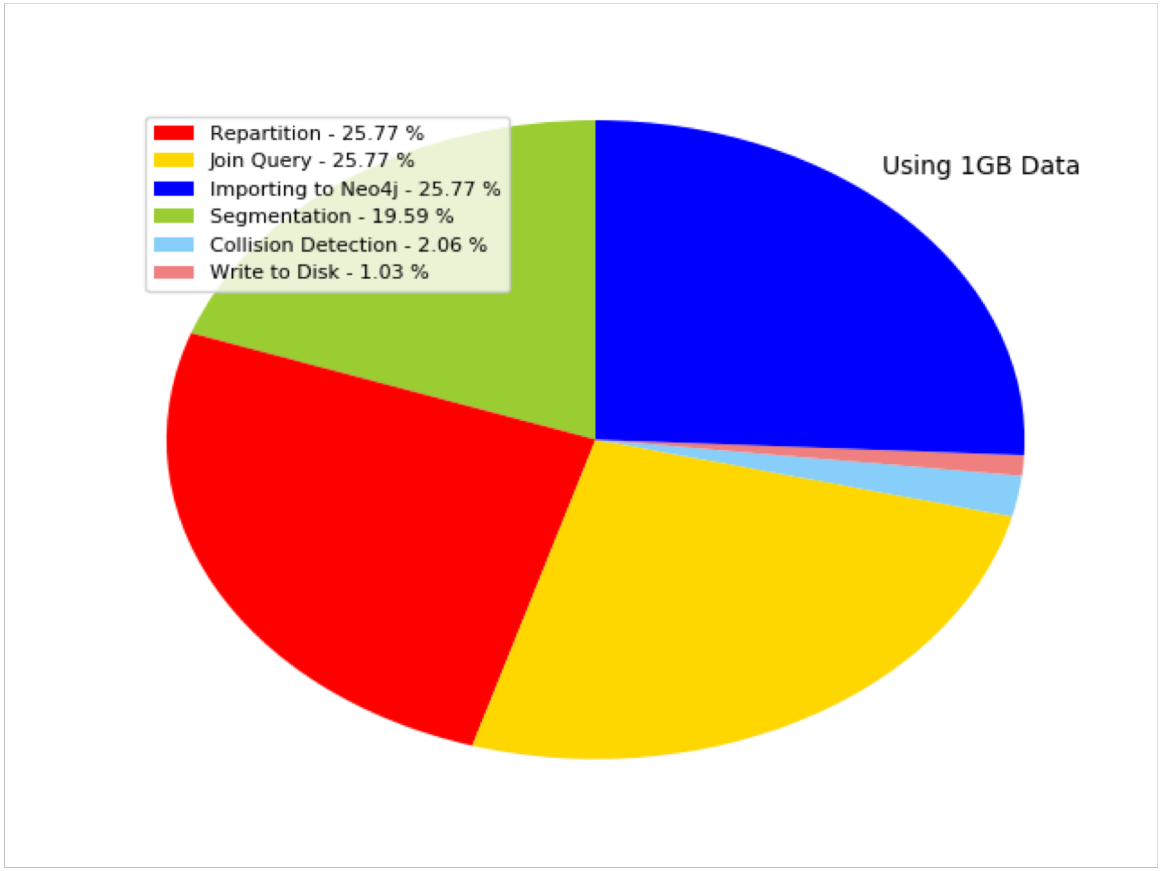
Benchmark

- **Platform: Amazon AWS EMR platform.**
- **Amazon EMR provides a managed platform to run populated framework such as Apache Spark, Hbase, Presto and Flink.**
- **No manual setup, elastic, flexible, low cost.**
- **R4.2 Xlare instance, 8 core 61GB memory each node.**
- **Dataset: Microsoft GeoLife trajectories.**
- **1.6GB, 17621 trajectories, 1,300,000 KM, 50, 000 hours.**

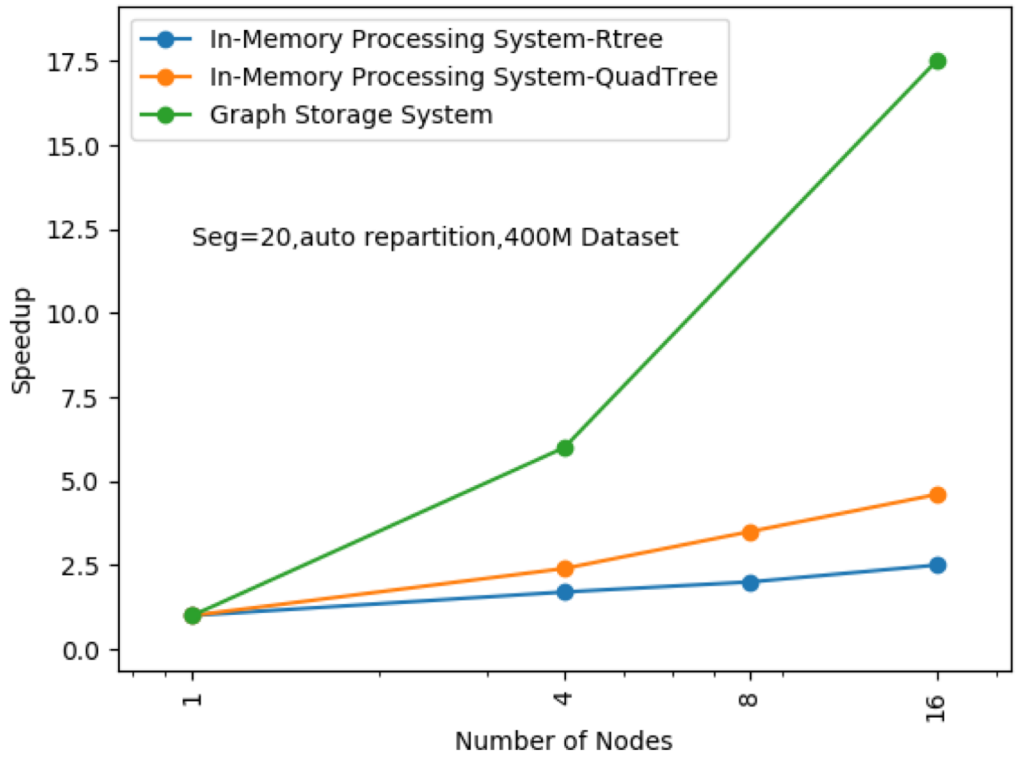
Framework with GeoSpark Stage Time Decomposition



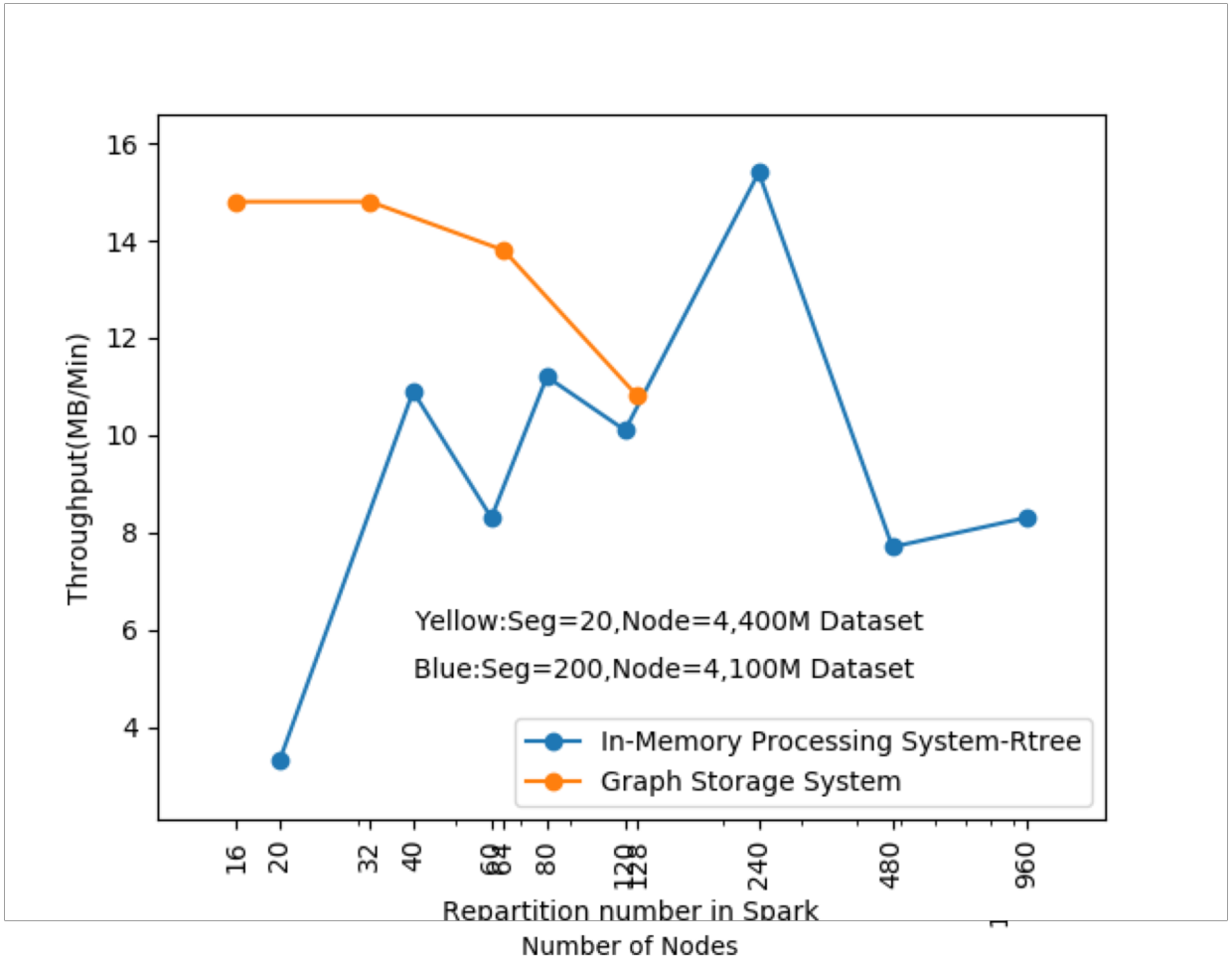
Framework with Neo4j Stage Time Decomposition



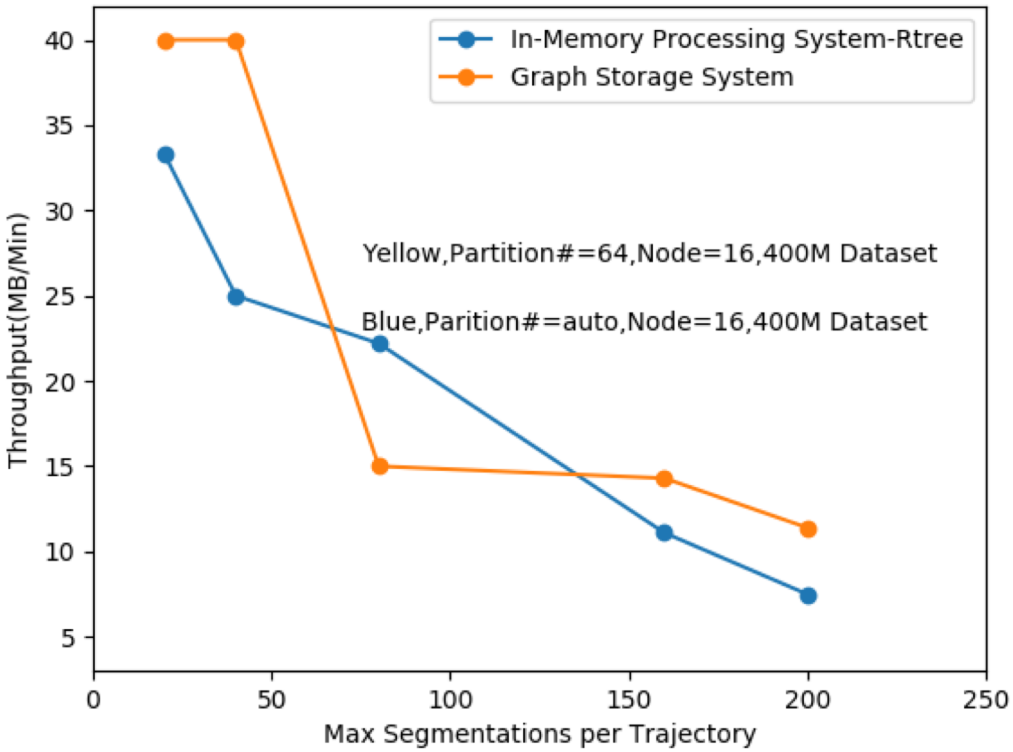
Framework Speedup



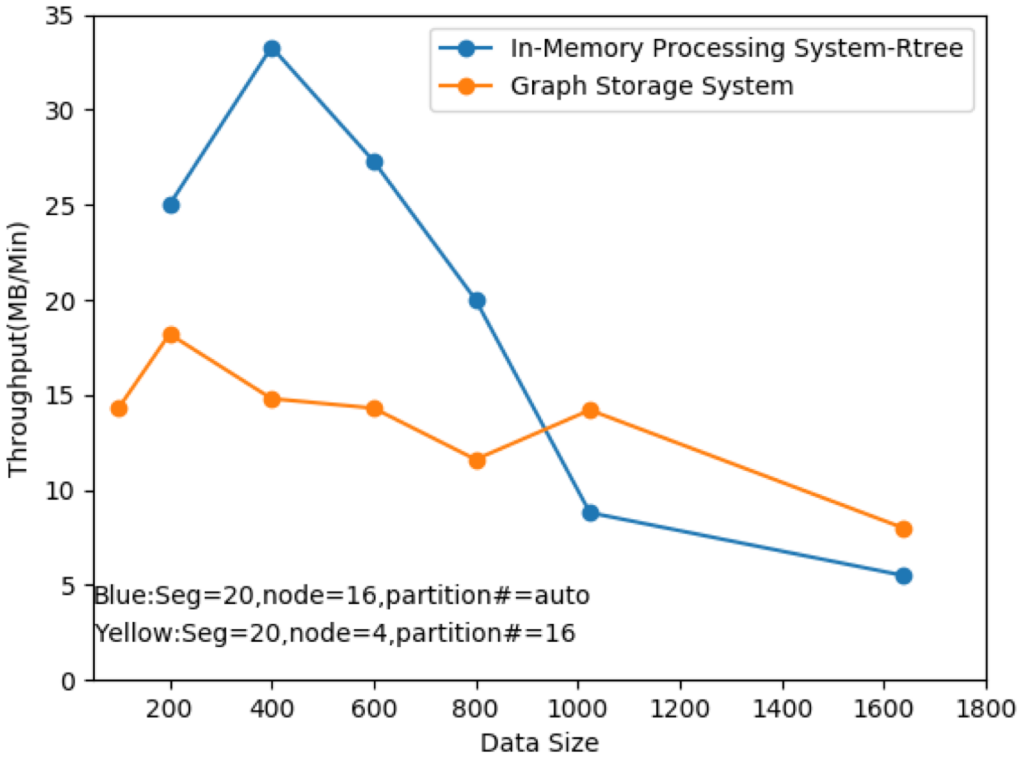
Repartitioning Effects to Throughput



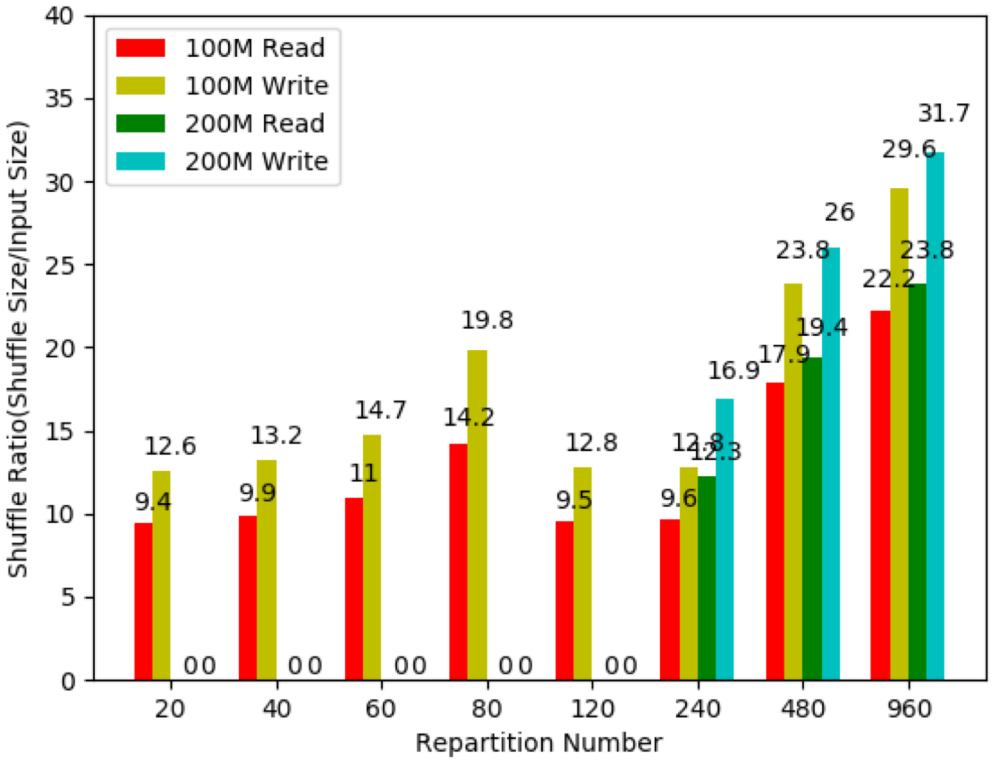
Trajectory Segmentation Number Effects to Throughput



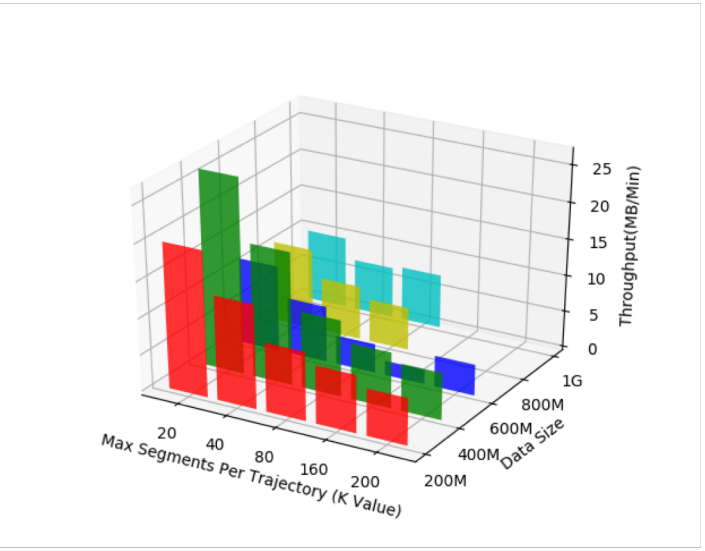
Datasize Effects to Throughput



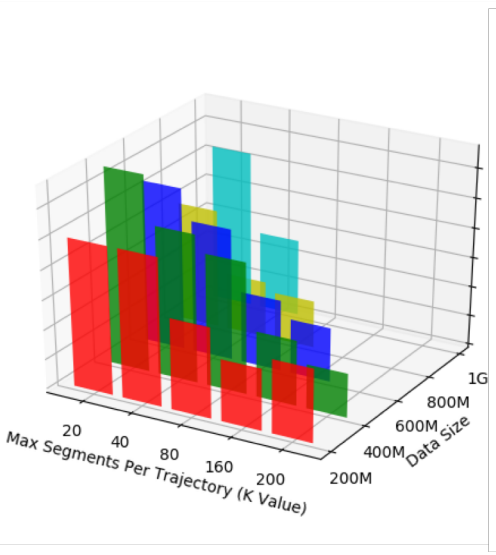
Shuffle Ratio to Repartitioning number



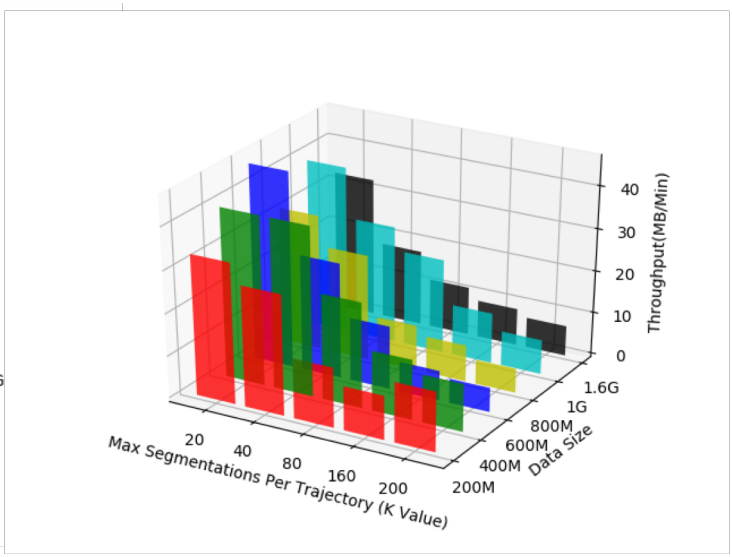
Framework Throughput



8 node in-memory framework throughput



16 node in-memory framework throughput



16 node graph storage framework

Outline

- Background
- Problem Statement, Objective, Contributions
- Research method
- Evaluation method
- Case Study
- Reliability and validity
- Future work and conclusion

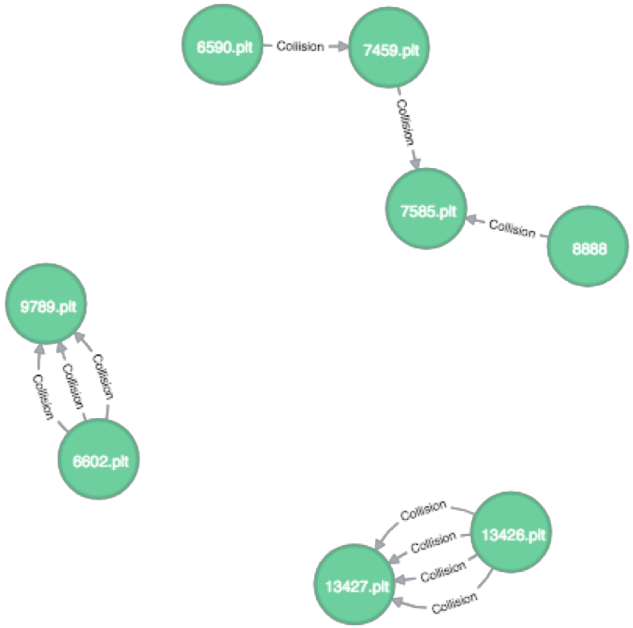
What Is Moving Crowd

A Graph G consisting of edges E(G) and vertices V(G)

An MBR expressed as u ∈ V (G)

A collision event expressed (u,v) ∈ E(G)

Searching for the connected components



a *connected component* of an undirected graph is a subgraph in which any two vertices are connected **to each other by paths, and which is** connected **to no additional vertices in the supergraph.**

Crowd Result Cross Validation

- **We compare the Crowd results with GPFinder[1] for validation.**
- **We set a trajectory positive if it forms a crowd with any other trajectories.**
- **We set a trajectory negative if it is isolated and has no collision with other trajectories.**
- **We randomly selected 3330 trajectories.**

		GPFinder	
		Positive	Negative
Our Prediction	Positive	2484	256
	Negative	421	169

[1] Y. Xian, Y. Liu, and C. Xu, "Parallel gathering discovery over big trajectory data," in *2016 IEEE International Conference on Big Data (Big Data)*, Dec 2016, pp. 783–792.

Crowd Result Cross Validation(Ctn'd)

- **The second round we extract the 2640 Positive trajectories from our application as input.**

- **Our results are stable that no negative found but since GPFinder is based on DBSCAN, the results changed.**

		GPFinder	
		Positive	Negative
Our Prediction	Positive	2671	69
	Negative	No Input	No Input

Outline

- Background
- Problem Statement, Objective, Contributions
- Research method
- Evaluation method
- Case Study
- Reliability and validity
- Future work and conclusion

Threads of Reliability and Validity

- **Test-retest reliability: Random sampling cannot ensure the constant partitioning and constant throughput.**
- **Parallel-forms reliability: We lack of throughput test on other cloud computing platform like Google Crowd or Microsoft Azure.**
- **Selection biases: The selection of test trajectories is constrained in one city and no mechanism to ensure the even distribution of trajectory length.**

Outline

- Background
- Problem Statement, Objective, Contributions
- Research method
- Evaluation method
- Case Study
- Reliability and validity
- Future work and conclusion

Conclusion and Future Work

- **We develop a distributed trajectory segmentation framework that transforms sequences of trajectories into queryable data blocks to build trajectory analysis applications.**
  - **System architecture**
  - **Workflows (in memory framework and graph storage)**
  - **Evaluation of performance**
  - **Observe the bottleneck and give data balancing method**
- **Next**
  - **Other indexing methods**
  - **Increase Robustness and redundancy**

Questions?

CONCORDIA.CA