Distributed Trajectory Similarity Search

Dong Xie, Feifei Li, Jeff M. Phillips
{dongx, lifeifei, jeffp}@cs.utah.edu
University of Utah

Motivation

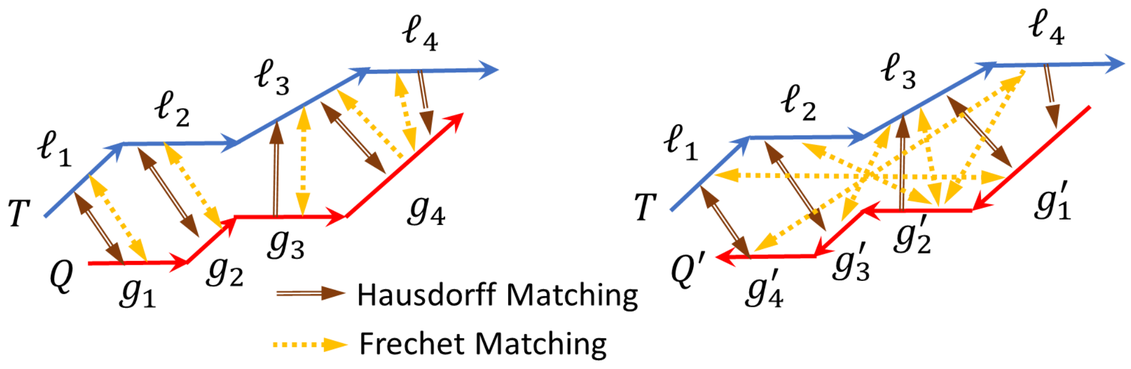
Huge amount of trajectory data are being generated everyday.
Widely used in traffic analysis, transportation planning.
Classic problem: find **'similar' trajectories.**
Require distributed solutions to **scale out**.

Trajectory Similarity Search

***k* nearest neighbor query** over trajectories under a specific distance metric D.

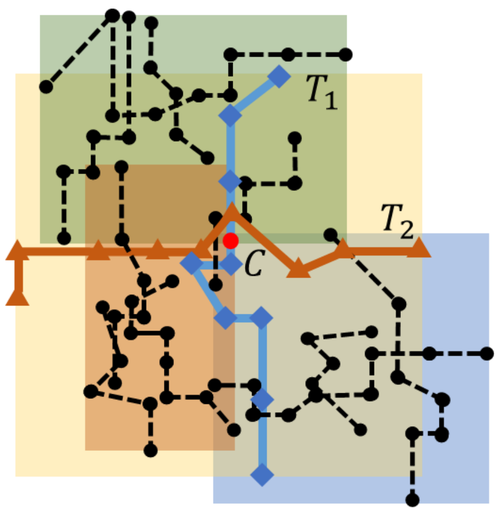
Not yet studied under a distributed environment.

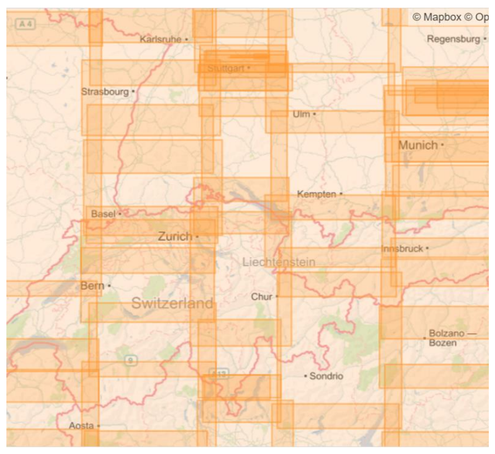
Different metrics: *Hausdorff distance* **vs.** *Frechet distance*



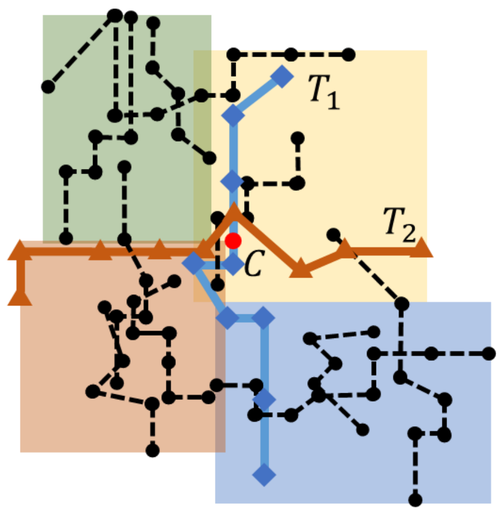
Segment-based vs. Trajectory-based Indexing

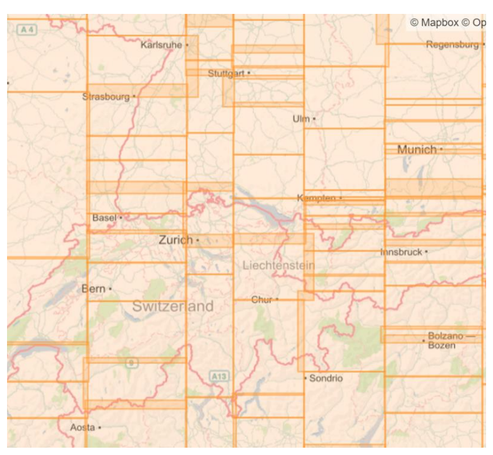
Partition **trajectories** as individual objects



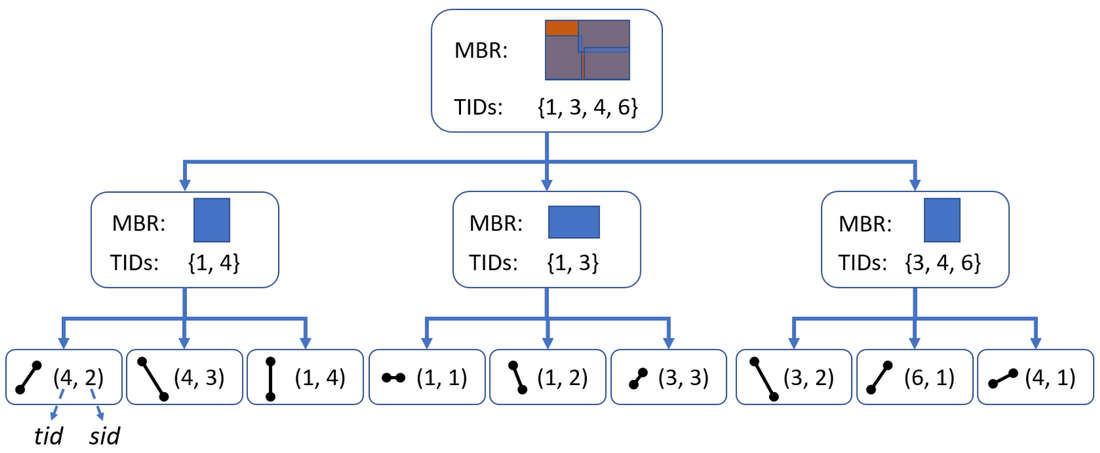


Partition all **segments** in trajectories



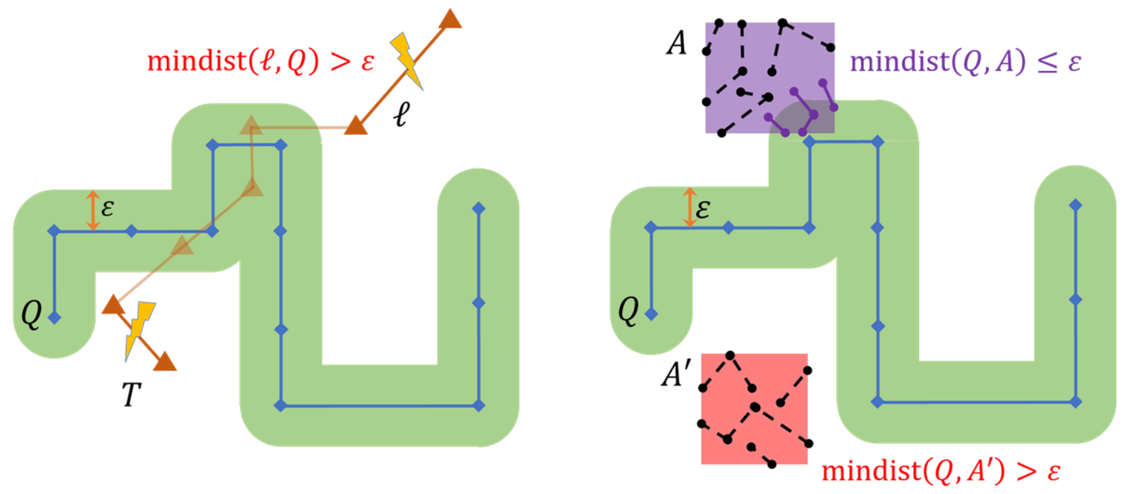


Customized R-Tree Local Index



Pruning Theorem

Given a distance threshold $\varepsilon > 0$, and two trajectories Q and T. If there exists a segment $\ell_i \in T$ such that $\text{mindist}(\ell_i, Q) > \varepsilon$, *then we have $D_H(Q,T) > \varepsilon$ and $D_F(Q,T) > \varepsilon$*



Search Procedure

Step 1: Pruning Bound Selection

Find a safe pruning bound ε covering at least k data trajectories.
Sample $c \cdot k$ trajectories passing similar regions as the query trajectory.
Find the k-th closest distance as the pruning bound ε.
Theory beneath: quantile estimation based on samples.

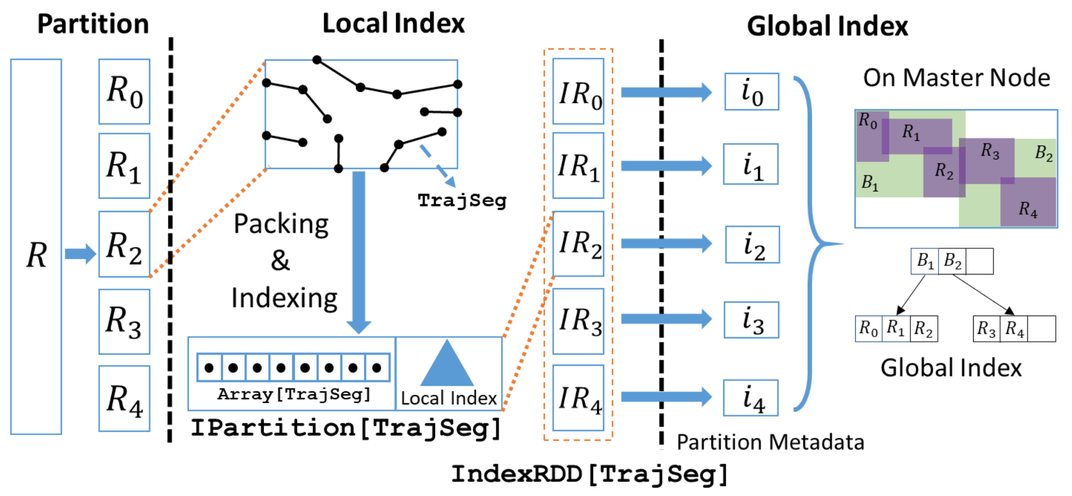
Step 2: Index-based Pruning

Utilize the distributed index to find the set of trajectory IDs can be safely pruned by ε.

Step 3: Finalizing Results

Rebuild all candidate trajectories, then launch a distributed top-K.

Two Level Indexing in Apache Spark



Design Choice and Optimization

Concise Data Structure for TID sets in customized R-Trees.
Roaring Bitmap: a ***concise and flexible*** compressed bitmap

Dual Indexing Strategy.

Keep **another data copy** organized in trajectory objects.
Eliminate the procedure of regrouping candidate trajectories.

Experiment Results