Reverse-Nearest Neighbor Queries on Uncertain Moving Object Trajectories

Tobias Emrich[1], Hans-Peter Kriegel[1], Nikos Mamoulis[2], Johannes Niedermayer[1], Matthias Renz[1], Andreas Züfle[1]

[1] LMU Munich
[2] HKU Hong Kong

Roadmap

› Motivation
  – Trajectory Data
  – Uncertainty in Trajectory Data

› Preliminaries
  – Markov Model
  – Bayesian Model Adaption
  – Indexing

› Reverse-Nearest Neighbor Search on Uncertain Trajectories
  – Problem Definition
  – Pruning Techniques
  – Evaluation

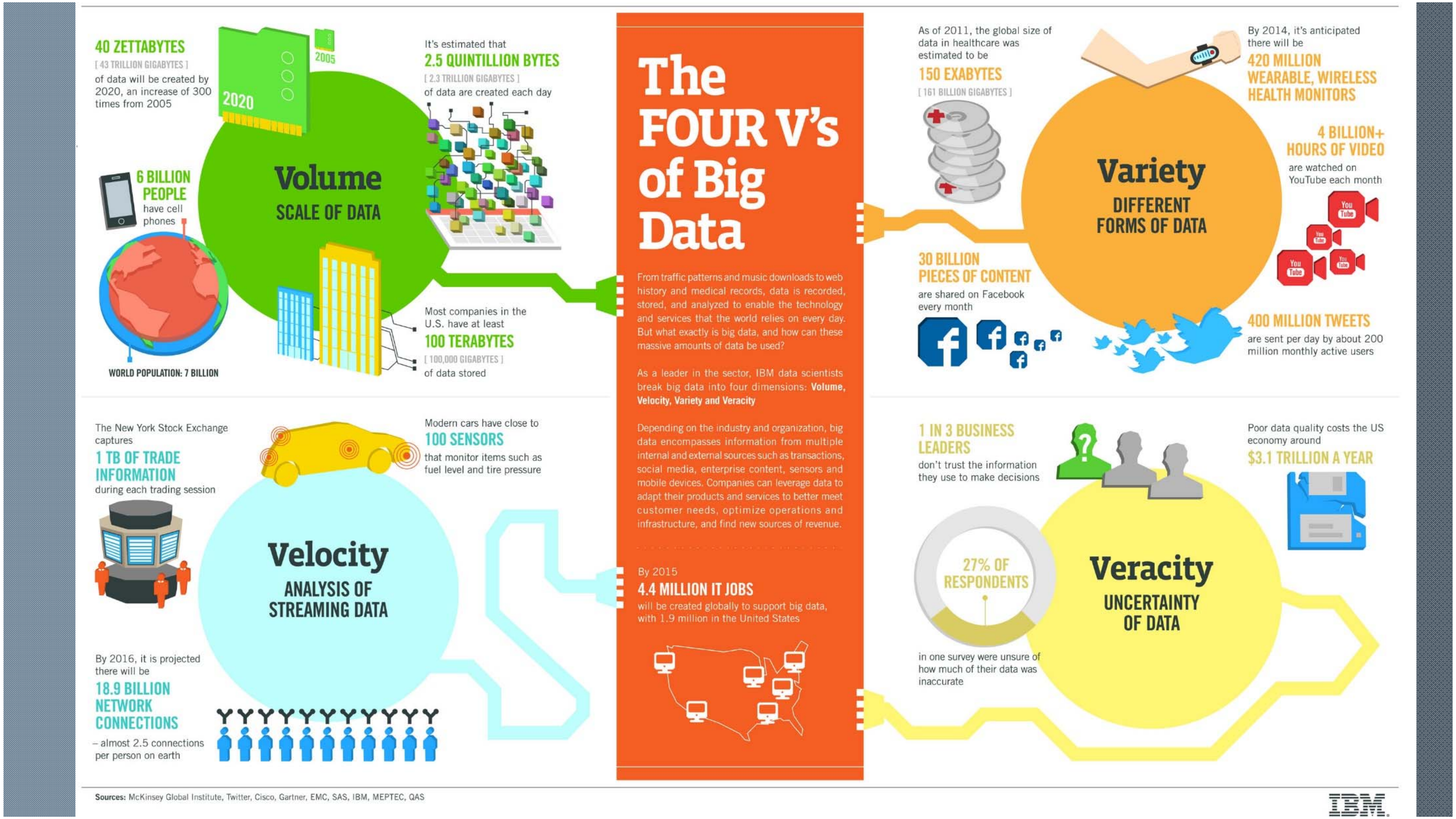
- Motivation
  - Trajectory Data
  - Uncertainty in Trajectory Data
- Preliminaries
- Reverse-Nearest Neighbor Search on Uncertain Trajectories

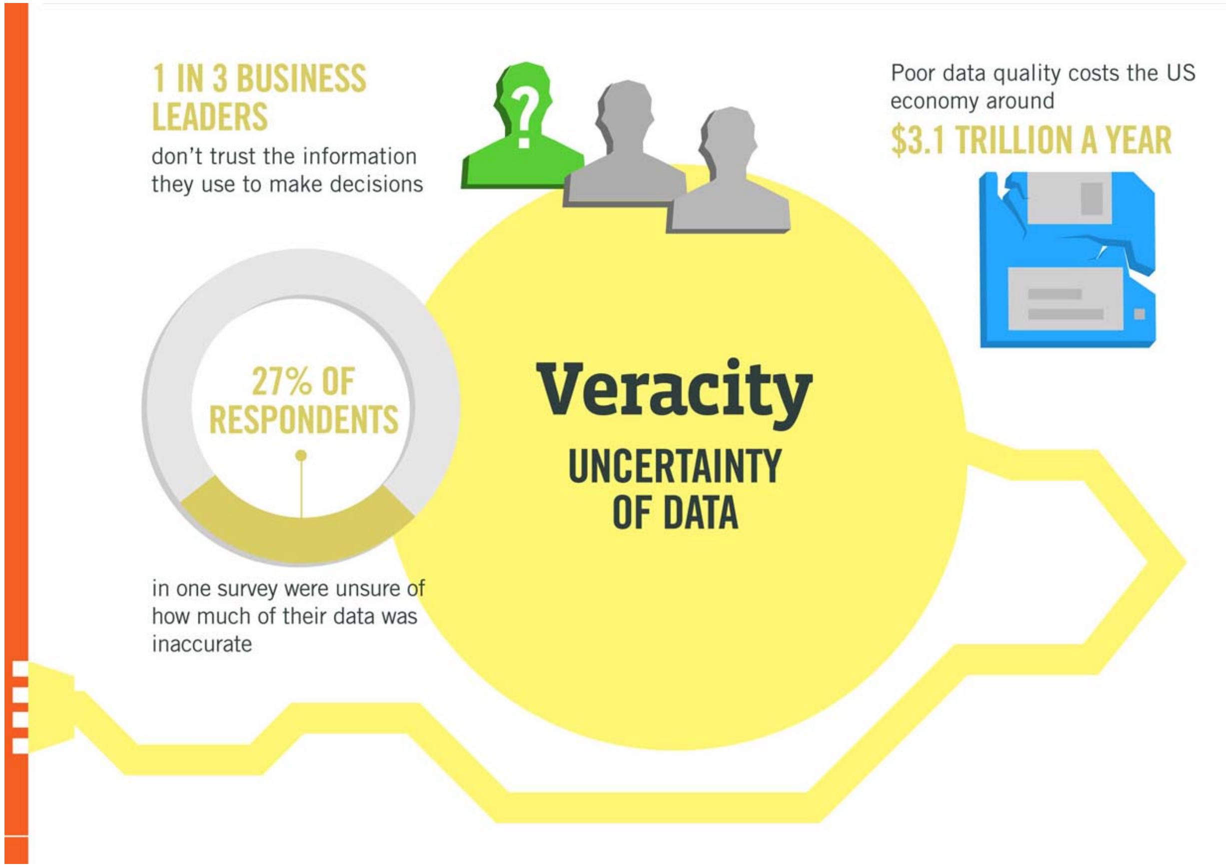
Moving Object Trajectories: Motivation

- Huge flood of geo-spatial data
  - Modern technology
  - New user mentality

- Great research potential
  - New applications
  - Innovative research
  - Economic Boost
    - “$600 billion potential annual consumer surplus from using personal location data” [1]

[1] McKinsey Global Institute. Big data: The next frontier for innovation, competition, and productivity. June 2011.


The FOUR V's of Big Data
From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?
As a leader in the sector, IBM data scientists break big data into four dimensions: Volume, Velocity, Variety and Veracity
Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.
By 2015
4.4 MILLION IT JOBS
will be created globally to support big data, with 1.9 million in the United States
Volume
SCALE OF DATA
40 ZETTABYTES
[43 TRILLION GIGABYTES]
of data will be created by 2020, an increase of 300 times from 2005
2020
2005
It's estimated that
2.5 QUINTILLION BYTES
[2.3 TRILLION GIGABYTES]
of data are created each day
6 BILLION PEOPLE
have cell phones
WORLD POPULATION: 7 BILLION
Most companies in the U.S. have at least
100 TERABYTES
[100,000 GIGABYTES]
of data stored
Velocity
ANALYSIS OF STREAMING DATA
The New York Stock Exchange captures
1 TB OF TRADE INFORMATION
during each trading session
Modern cars have close to
100 SENSORS
that monitor items such as fuel level and tire pressure
By 2016, it is projected there will be
18.9 BILLION NETWORK CONNECTIONS
– almost 2.5 connections per person on earth
Variety
DIFFERENT FORMS OF DATA
As of 2011, the global size of data in healthcare was estimated to be
150 EXABYTES
[161 BILLION GIGABYTES]
By 2014, it's anticipated there will be
420 MILLION WEARABLE, WIRELESS HEALTH MONITORS
4 BILLION+ HOURS OF VIDEO
are watched on YouTube each month
30 BILLION PIECES OF CONTENT
are shared on Facebook every month
400 MILLION TWEETS
are sent per day by about 200 million monthly active users
Veracity
UNCERTAINTY OF DATA
1 IN 3 BUSINESS LEADERS
don't trust the information they use to make decisions
Poor data quality costs the US economy around
$3.1 TRILLION A YEAR
27% OF RESPONDENTS
in one survey were unsure of how much of their data was inaccurate
Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS
IBM


1 IN 3 BUSINESS LEADERS
don't trust the information they use to make decisions
27% OF RESPONDENTS
in one survey were unsure of how much of their data was inaccurate
Veracity
UNCERTAINTY OF DATA
Poor data quality costs the US economy around
$3.1 TRILLION A YEAR

Research Challenge

Include the uncertainty, which is inherent in trajectory data, directly in the querying process.

Assess the reliability of query results.

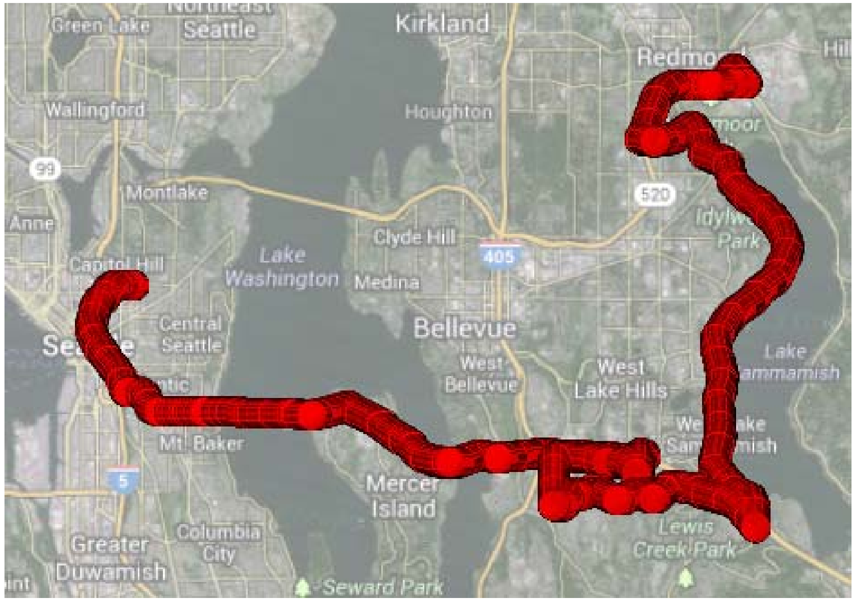
Enhance the underlying decision-making process.

Trajectory Data

- (object, location, time) triples

- Queries:
  - "Find friends that attended the same concert last saturday"

- Best case: Continuous function $time \rightarrow space$

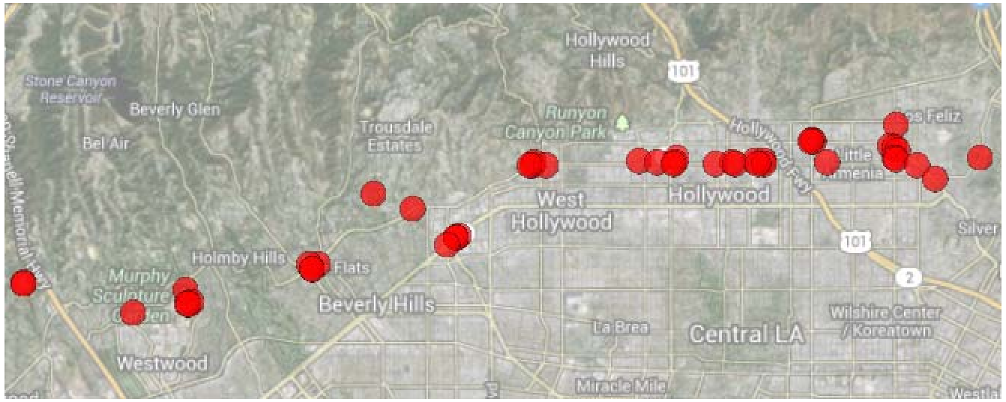


GPS log taken from a thirty minute drive through Seattle

Dataset provided by: P. Newson and J. Krumm. Hidden Markov Map Matching Through Noise and Sparseness. ACMGIS 2009.

Sources of Uncertainty

- Missing Observations
  - Missing GPS signal
  - RFID sensors available in discrete locations only
  - Wireless sensor nodes sending infrequently to preserve energy
  - Infrequent check-ins of users of geo-social networks

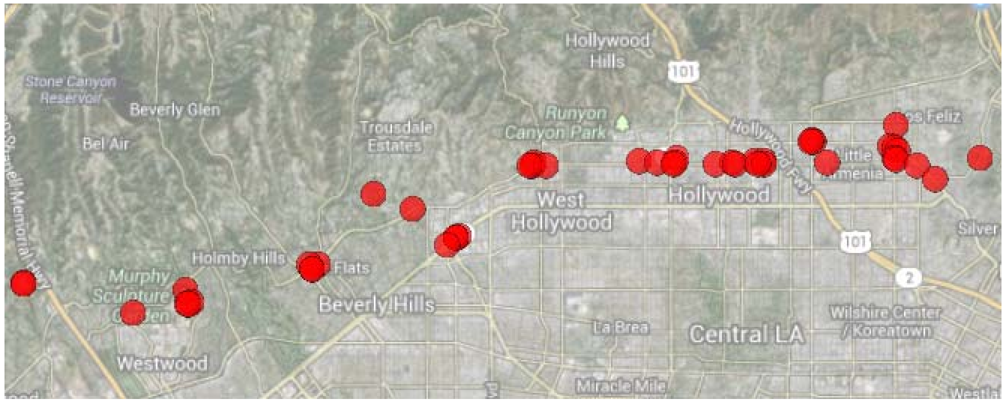


Dataset provided by: E. Cho, S. A. Myers and J. Leskovek. Friendship and Mobility: User Movement in Location-Based Social Networks. SIGKDD 2011.

Sources of Uncertainty

- Uncertain Observations
  - Imprecise sensor measurements (e.g. radio triangulation, Wi-Fi positioning)
  - Inconsistent information (e.g. contradictive sensor data)
  - Human errors (e.g. in crowd-sourcing applications)

➢ From database perspective, the position of a mobile object is uncertain



Dataset provided by: E. Cho, S. A. Myers and J. Leskovek. Friendship and Mobility: User Movement in Location-Based Social Networks. SIGKDD 2011.

- Motivation
- Preliminaries
  - Markov Model
  - Bayesian Model Adaption
  - Indexing
- Reverse-Nearest Neighbor Search on Uncertain Trajectories

Markov Model

› Discretization of time and space

  – treat intersections as states and add additional states on long streets

  – The time interval corresponding to a tick is e.g. 20 sec

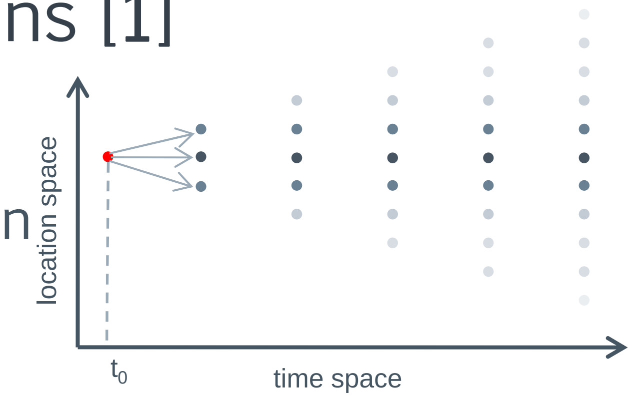
› Estimation of model parameters

  – Transition probabilities from one state to another are learned from historical data (very sparse matrix!!)

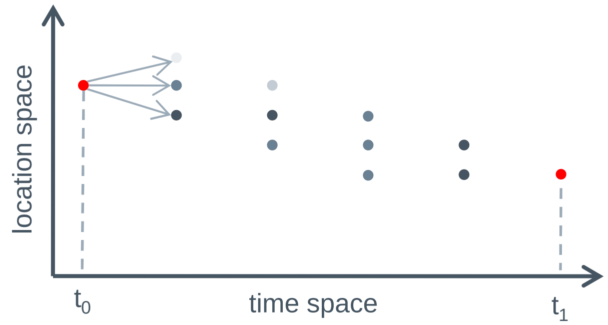
  – Transition matrix can change over time and for different object groups

Model Adaption: Observations [1]

› So far we had only one observation from which we could extrapolate



› This is not really of interest since cars do not move randomly

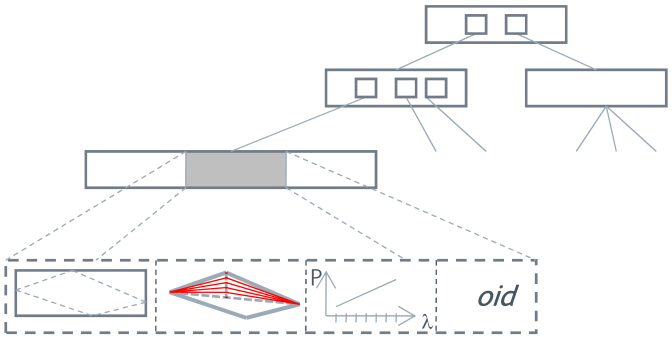


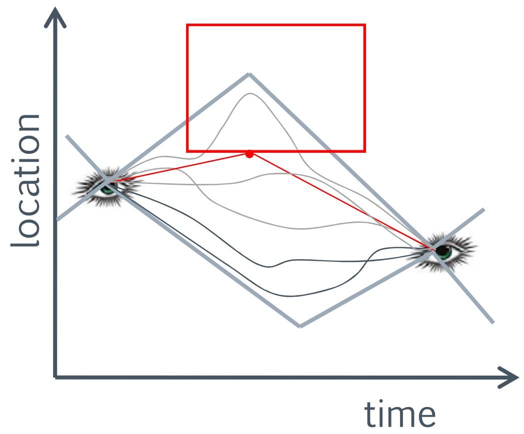
› With two observations we have to introduce more artificial states and adapt the techniques

[1]J. Niedermayer, A. Züfle, T. Emrich, M. Renz, N. Mamoulis, L. Chen, H.-P. Kriegel: *Probabilistic Nearest Neighbor Queries on Uncertain Moving Object Trajectories.* PVLDB 2013

Indexing Uncertain Trajectory Data [2]

- With the above techniques each object in the database has to be processed
- Index Structure based on R-Tree indexing the ST-Space





[2] T. Emrich, H.-P. Kriegel, N. Mamoulis, M. Renz, and A. Züfle. *Indexing uncertain spatio-temporal data.* CIKM 2012.

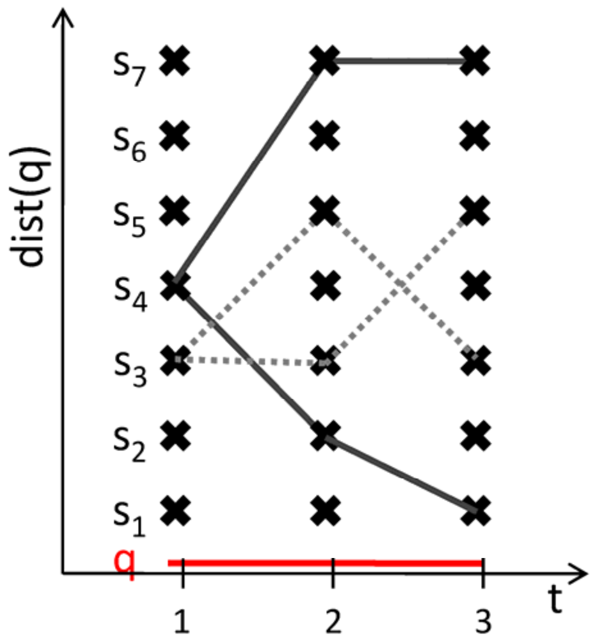
› Motivation

› Preliminaries

› Reverse-Nearest Neighbor Search on Uncertain Trajectories
  - Problem Definition
  - Pruning Techniques
  - Evaluation

Reverse Nearest Neighbor Queries: Example

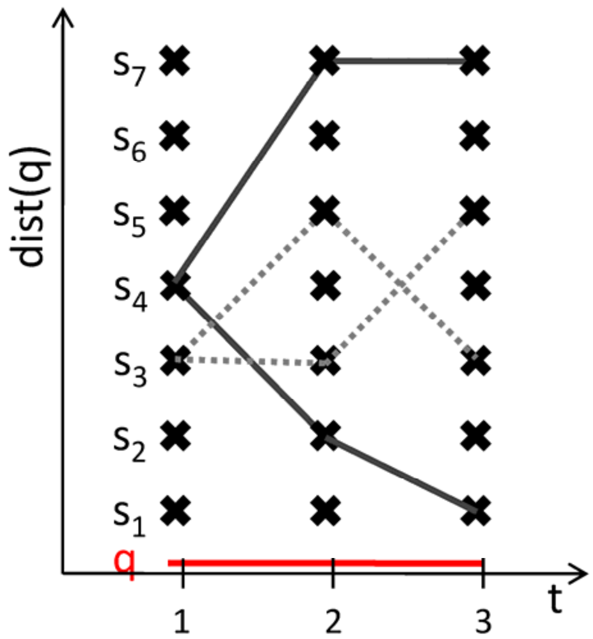
A probabilistic ∃(∀) reverse nearest neighbor query retrieves all objects having a sufficiently high probability to be the reverse nearest neighbor of a query trajectory q for at least one point of time (each point of time) in a query time window T.



object	trajectory	P(tr)
o_1 ——	$tr_{1,1} = s_4, s_2, s_1$	0.4
o_1 ——	$tr_{1,2} = s_4, s_7, s_7$	0.6
o_2	$tr_{2,1} = s_3, s_3, s_5$	0.2
o_2	$tr_{2,2} = s_3, s_5, s_3$	0.8

Reverse Nearest Neighbor Queries: Example

A probabilistic ∃(∀) reverse nearest neighbor query retrieves all objects having a sufficiently high probability to be the reverse nearest neighbor of a query trajectory q for at least one point of time (each point of time) in a query time window T.

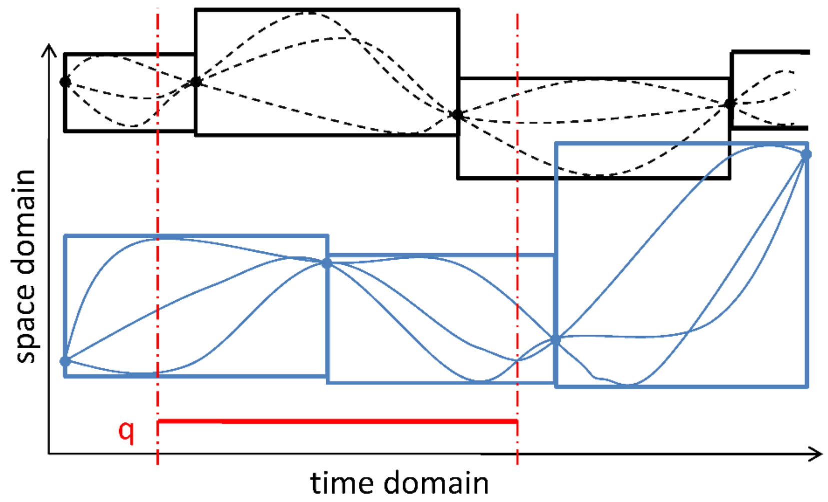


object	trajectory	P(tr)
o_1 —	$tr_{1,1} = s_4, s_2, s_1$	0.4
o_1 —	$tr_{1,2} = s_4, s_7, s_7$	0.6
o_2	$tr_{2,1} = s_3, s_3, s_5$	0.2
o_2	$tr_{2,2} = s_3, s_5, s_3$	0.8

Example Query: Return objects having a non-zero probability to be the RNN of q at time t=2 and t=3.

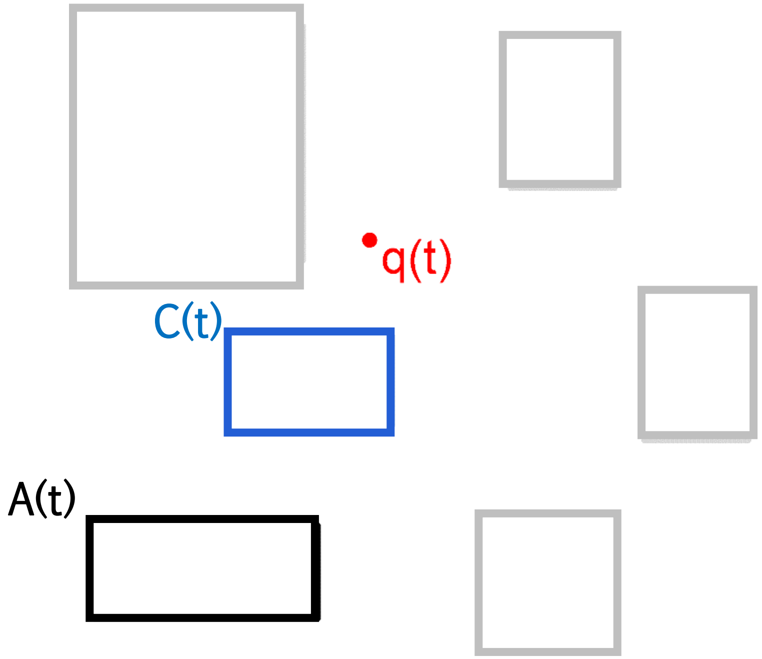
Temporal Pruning

Prune page regions of the index not overlapping the query in time.



Spatial Pruning

- Rectangle based pruning at single points of time of the query.[3]
  - Can q(t) be closer to C(t) than A(t)?
    => C is a candidate at time t
  - Must q(t) be closer to C(t) than A(t)?
    => C is a true hit at time t
- Union(Intersect) to obtain probabilistic $\exists(\forall)$ reverse nearest neighbor candidates



[3] T. Emrich, H.-P. Kriegel, P. Kröger, M. Renz, and A. Züfle. *Boosting spatial pruning: on optimal pruning of MBRs.* SIGMOD 2010.

Verification

› Monte-Carlo-Sampling

– Using a-posteriori transition matrices conditioned to observations

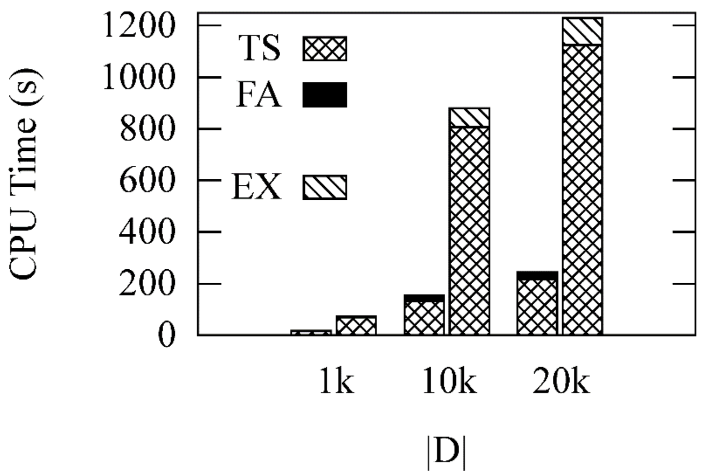
– For each sampled world

  › Check for each object o if it is a ∃(∀) reverse nearest neighbor

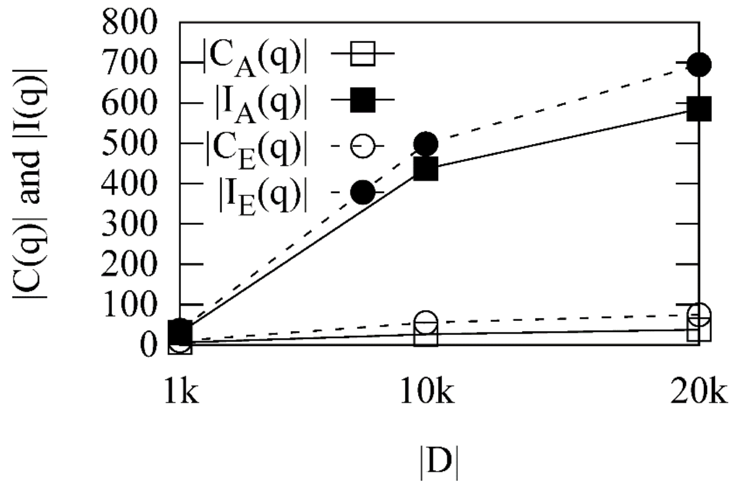
– Use the relative number of sampled worlds where o is a ∃(∀) reverse nearest neighbor as an unbiased estimator of the probability p that o is a ∃(∀) reverse nearest neighbor.

– Use standard techniques to obtain a confidence interval of the probability of a binomial random variable.

Evaluation



CPU Time (s)
1200
1000
800
600
400
200
0
TS
FA
EX
1k
10k
20k
|D|



|C(q)| and |I(q)|
800
700
600
500
400
300
200
100
0
$|C_A(q)|$
$|I_A(q)|$
$|C_E(q)|$
$|I_E(q)|$
1k
10k
20k
|D|

Thank you for your attention!

Does the Markov assumption hold in reality ?

- Of course single cars do not follow the Markov Chain (weighted random walk)
- However the Markov Model is just the apriori Model in which we infer the observations