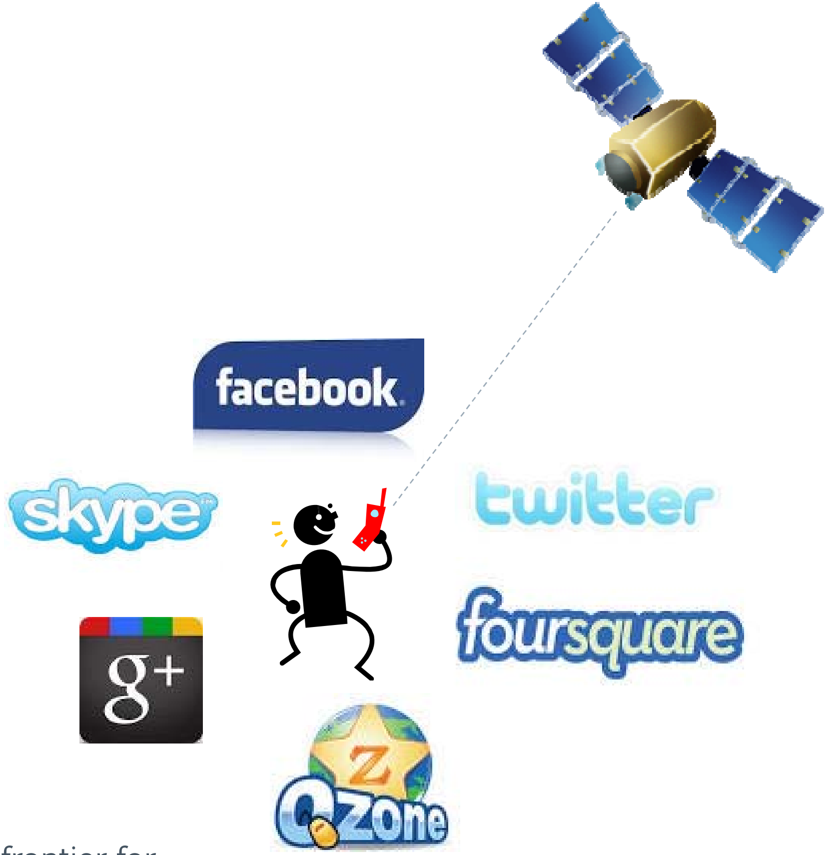
Similarity Search and Mining in Uncertain Spatial and Spatio-Temporal Databases

Andreas Züfle

Geo-Spatial Data

- 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]



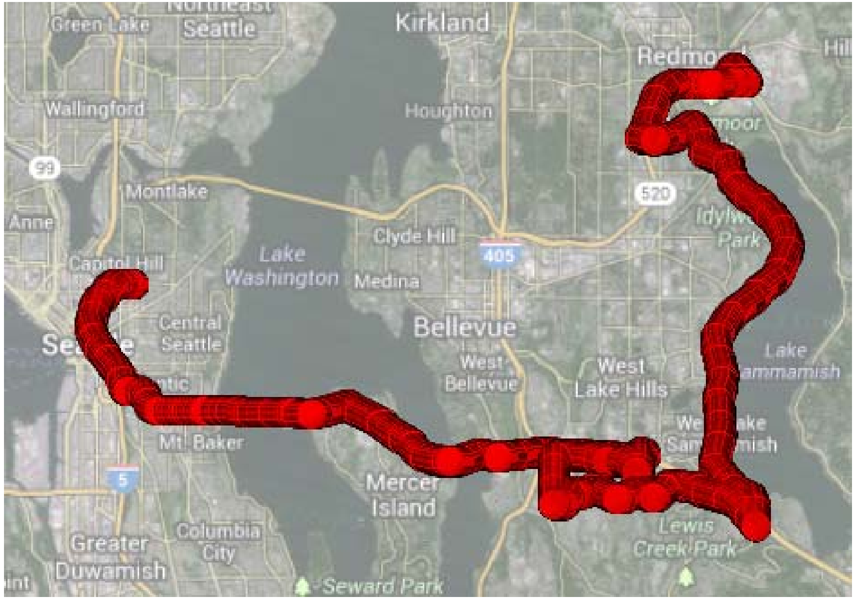
[1] McKinsey Global Institute. Big data: The next frontier for innovation, competition, and productivity. June 2011.

Spatio-Temporal Data

- (object, location, time) triples

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

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

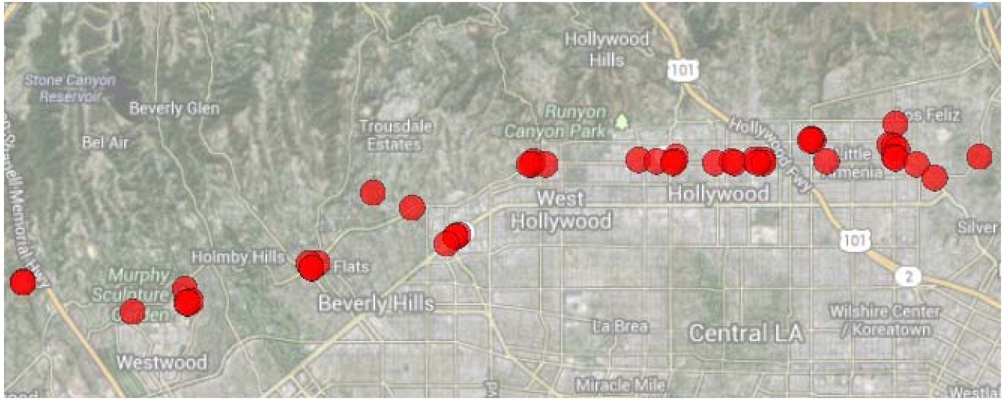


GPS log taken from a thirty minute drive through Seattle

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

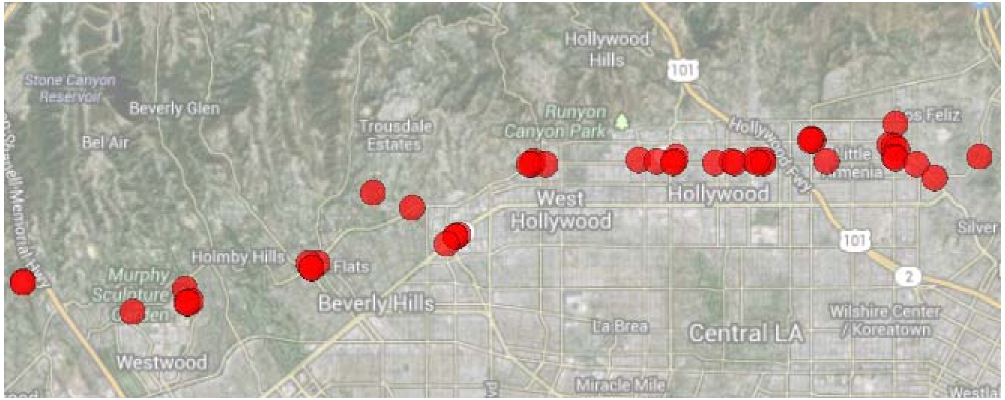


Dataset provided by: E. Cho, S. A. Myers and J. Leskovek. Friendship and Mobility: User Movement in Location-Based Social Networks. SIGKDD 2011.

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)

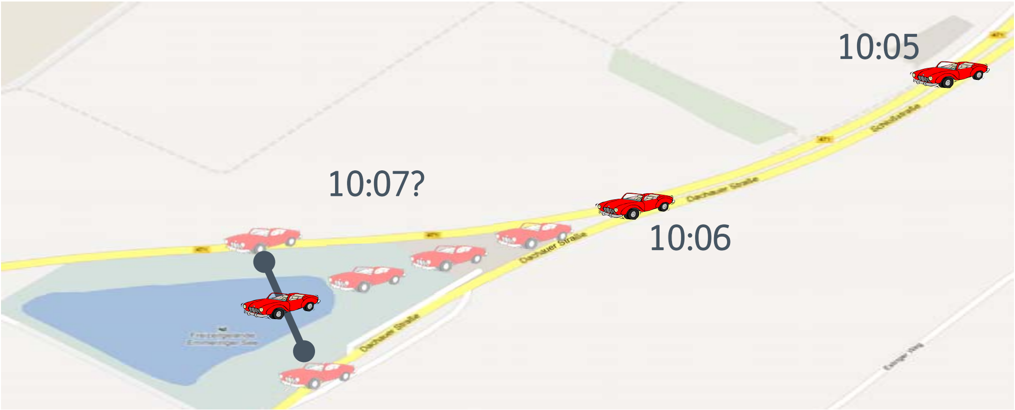
➢ From database perspective, the position of a mobile object is uncertain



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

Traditional Solutions

- Avoid uncertainty
  - Store aggregated positions in the database
    - Extrapolated positions
    - Expected positions
    - Most-likely positions
- Impossible to assess the confidence of results



Research Challenge

Include the uncertainty, which is inherent in spatial and spatio-temporal data, directly in the querying and mining process.

Research Challenge

Include the uncertainty, which is inherent in spatial and spatio-temporal data, directly in the querying and mining process.

↓

Assess the reliability of similarity search and data mining results, enhancing the underlying decision-making process.

Research Challenge

Include the uncertainty, which is inherent in spatial and spatio-temporal data, directly in the querying and mining process.

↓

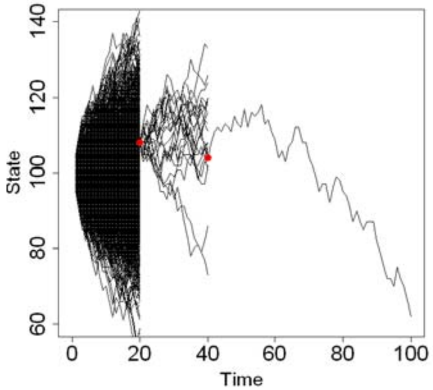
Assess the reliability of similarity search and data mining results, enhancing the underlying decision-making process.

↓

Improve the quality of modern location based applications and of research results in the field.

Uncertain Spatio-Temporal Data Model [1]

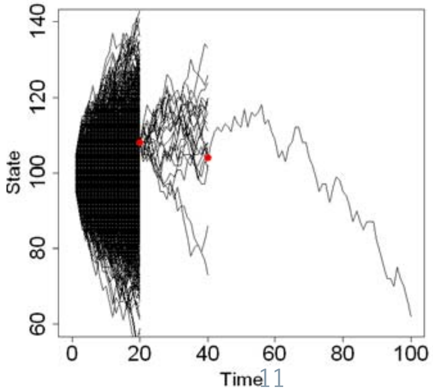
- Discretize Time and Space
- Model object movement as a Markov chain
  - Weighted Random Walk
- Learn transition probabilities empirically
- Rejected possible worlds that do not match all observations
- Exact Probabilities can be computed for special queries [1]
- General Approach: Monte-Carlo-Sampling
  - Draw a sufficiently high number of samples
  - Approximate result probability = ratio of samples that satisfy the query and total number of drawn samples



[1] T. Emrich, H.-P. Kriegel, N. Mamoulis, M. Renz, and A. Züfle. Querying uncertain spatio-temporal data. In Proceedings of the 28th International Conference on Data Engineering (ICDE), Washington, DC, 2012.

Uncertain Spatio-Temporal Data Model [1]

- Discretize Time and Space
- Model object movement as a Markov chain
  - Weighted Random Walk
- Learn transition probabilities empirically
- Rejected possible worlds that do not match all observations
- Exact Probabilities can be computed for special queries [1]
- General Approach: Monte-Carlo-Sampling
  - Draw a sufficiently high number of samples
  - Approximate result probability = ratio of samples that satisfy the query and total number of drawn samples
- But how to draw samples efficiently such that they are conform with the observations?

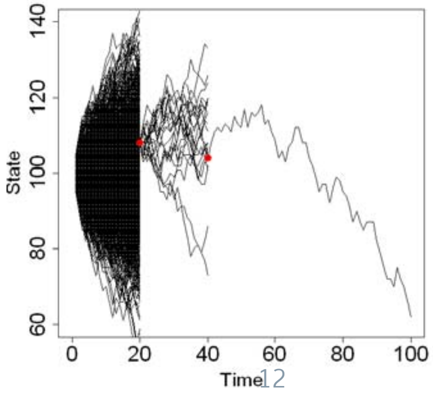


11

[1] T. Emrich, H.-P. Kriegel, N. Mamoulis, M. Renz, and A. Züfle. Querying uncertain spatio-temporal data. In Proceedings of the 28th International Conference on Data Engineering (ICDE), Washington, DC, 2012.

Uncertain Spatio-Temporal Data Model [1]

- Discretize Time and Space
- Model object movement as a Markov chain
  - Weighted Random Walk
- Learn transition probabilities empirically
- Rejected possible worlds that do not match all observations
- Analytic solutions for special query types [1]
- General Approach: Monte-Carlo-Sampling
  - Draw a sufficiently high number of samples
  - Approximate result probability = ratio of samples that satisfy the query and total number of drawn samples
- But how to draw samples efficiently such that they are conform with the observations?
- Solution: Adaption of transition matrices



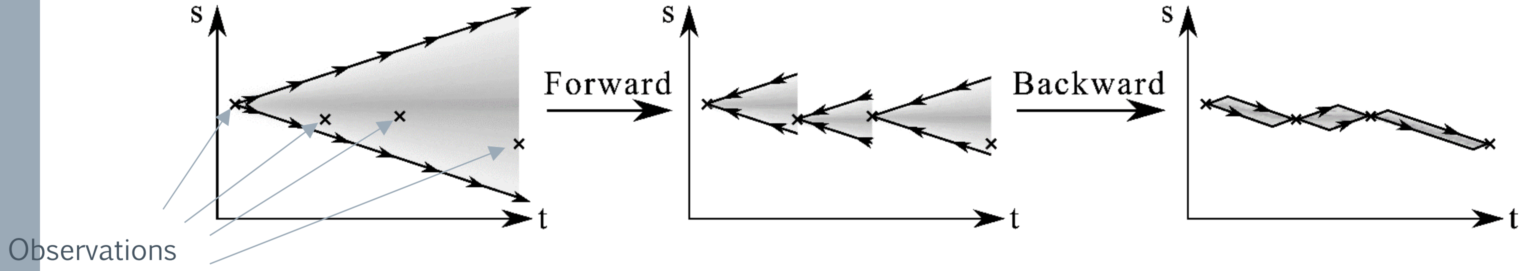
[1] T. Emrich, H.-P. Kriegel, N. Mamoulis, M. Renz, and A. Züfle. Querying uncertain spatio-temporal data. In Proceedings of the 28th International Conference on Data Engineering (ICDE), Washington, DC, 2012.

Probabilistic NN-Queries

- Extension of Nearest-Neighbor-Querys on (certain) trajectories to the uncertain case
- Certain Case:
  - For a query trajectory q, and a time interval T, a ∀-Nearest-Neighbor Query returns all objects having the smallest distance to q during the whole interval T.
  - For a query trajectory q, and a time interval T, a ∃-Nearest-Neighbor Query returns all objects having the smallest distance to q during any time in T.
- Uncertain Case:
  - For an uncertain trajectory q, a probabilistic ∀(∃)-Nearest-Neighbor Query returns, for each object in the database, the probability to be a ∀(∃)-Nearest-Neighbor of q.
  - Both variants are NP-hard to solve analytically. (Proofs given in the paper)

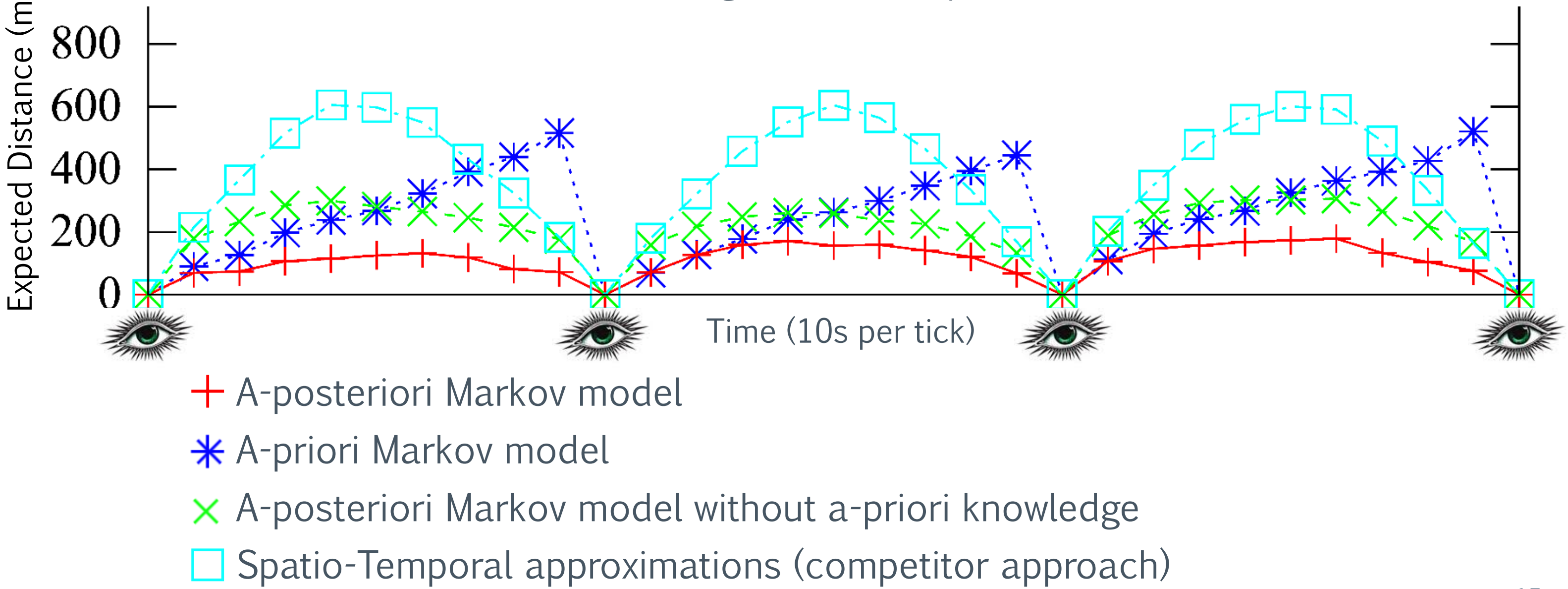
Adding knowledge to the model: Bayesian Inference

- Using Bayesian inference, additional knowledge can be added to the model, such as discrete observations of an object.
- Use empirically learned transition probabilities as a-priori model
- Adapt this model to an a-posteriori given information about discrete observations of an object.
- Model adaption using a Forward-Backward approach



Adding knowledge to the model: Bayesian Inference

- The adapted a-posteriori model allows to effectively interpolate positions between discrete observations.
- Significant improvement to existing approaches.
- Good results even without having a trained a-priori model



Summary & Other Contributions

- Theoretical Analysis: NP-hardness of NN-Queries using this model.

- Efficient Markov model adaption, given observations by Bayesian Inference
  - Using the empirically learned Markov chain as prior
  - Using a forward-backward approach to derive the posterior

- Efficient sampling approach using the posterior model
  - Applicable for any query having a solution for the certain certain case!

- Index support for ∀(∃)-Nearest-Neighbor Queries
  - Based on an existing index structure
  - Algorithms for efficient query processing provided

- Strong Experimental Results
  - Probabilistic Models can vastly reduce the expected prediction error
  - Compared to
    - Traditional approaches predicting a single location
    - Existing approaches for uncertain spatio-temporal data

Thank you for your attention
Questions?!