Data-Intensive Routing in Spatial Networks

Christian S. Jensen
www.cs.aau.dk/~csj

Roadmap

- Setting: big data
- Road network travel cost modeling and computation
  - Time-varying, uncertain weights
    - Histograms
    - GMMs
- Routing
  - Stochastic skyline routing
  - Personalized routing
  - Routing based on local-driver behavior
- Closing
  - Demos, the future, challenges, acknowledgments, readings

Setting: Big Data

Hype or Substance?

- We have been pushing the boundaries for decades
  - How much data we can handle
  - How fast
  - Data integration
- Examples
  - VLDB: International Conference on Very Large Database
  - TODS: ACM Transactions on Database Systems
- So is it all hype?
  - No

Instrumentation and Digitization

- Instrumentation of reality
  - Notably, smartphones
- Digitization of processes
  - E.g., e-commerce, public services, communications, social interactions

2005 vs. 2013
Big Data

Every day, we create 2.5 quintillion bytes of data — so much that 90% of the data in the world today has been created in the last two years alone. This data comes from everywhere: sensors used to gather climate information, posts to social media sites, digital pictures and videos, purchase transaction records, and cell phone GPS signals to name a few. This data is big data.

http://www-01.ibm.com/software/data/bigdata/

Big Data – Synthesis

• The result is new opportunity.
• Lots of data and unprecedented computing infrastructure combine to offer potentials for value creation from data.
• To be competitive, society and businesses must be able to create value from data.
• Data-based decisions and data-driven processes
  ■ Decisions based on good data beat decisions based on feelings or opinions.
• A finer granularity of services
• Entirely new services

Motivation – ITS

• A safer, greener, and more efficient and cost-effective transportation infrastructure
• Congestion, greater Copenhagen region
  ■ ~10 billion DKK/year (2004)
• Bad setting of signalized intersections in Denmark
  ■ ~9.3 billion DKK/year (2012)

Big Data in Routing

Motivation – Eco-Routing

• The transportation sector is the second largest greenhouse gas (GHG) emitting sector and also causes substantial pollution.
  ■ By 2020, it’s projected that $6.9 trillion worth will be moved every day, worldwide.
• The reduction of greenhouse gas (GHG) emissions from transportation is essential to combat global climate change.
  ■ EU: reduce GHG emissions by 30% by 2020.
  ■ GB: a 50% GHG reduction by 2050.
  ■ China: a 17% GHG reduction by 2015.
• Eco-routing can reduce vehicular impact by up to 20%.
• General context: Smart City

Motivation – Eco-Weights

• The capture of the environmental costs of traversing road network edges is key to eco-routing.
  ■ Eco-weights are uncertain.
  ■ Eco-weights are time-dependent.
Time-Varying Uncertain Eco-Weights

Outline
- Approach I – histograms
  - Setting
  - Framework
  - ERN building
  - GHG emissions estimation
- Approach II – GMMs
  - Setting
  - MTUG building
  - Cost estimation

Setting
- Eco Road Network $G = (V, E, F)$
  - $V$: Vertex set. Each vertex indicates a road intersection.
  - $E$: Edge set. Each edge indicates a road segment.
  - Function $F$ assigns a time-dependent, uncertain eco-weight to each edge in $E$.
- Input
  - A set of map-matched trajectories $TR$.
  - An accompanying road network $G' = (V, E, \text{Null})$.
- Output
  - The Eco Road Network $G = (V, E, F)$.

Framework
- Time-dependent uncertain histograms.
  - A vector of $(\text{period, histogram})$ tuples $<T_i, H_i>$.
  - $H_i$ is the histogram describing the distribution of cost values observed in period $T_i$.
  - Used to represent the eco-weights of road network edges
  - Two types of compression are applied to reduce the storage space while retaining acceptable accuracy.

Framework
- GPS records are map-matched to the corresponding edges.
- Map-matched records are transformed into traversal records.
  - A traversal record $r = (e, t, t_t, g_e)$ indicates that edge $e$ is traversed by a trajectory $tr$ starting at time $t$ and has travel time $t_t$ and GHG emissions $g_e$.
  - The VT-micro environmental impact model is used to estimate the GHG emissions of each traversal record.

Traversal Record Analysis

<table>
<thead>
<tr>
<th>Trajectories TR</th>
<th>Road Network $G = (V, E, \text{NULL})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Traversal Record Analysis</td>
<td></td>
</tr>
<tr>
<td>Initial Histogram Construction</td>
<td></td>
</tr>
<tr>
<td>Histogram Merging</td>
<td></td>
</tr>
<tr>
<td>Bucket Reduction</td>
<td></td>
</tr>
<tr>
<td>Eco Road Network $G = (V, E, F)$</td>
<td></td>
</tr>
</tbody>
</table>
**Initial Histogram Building**
- Each edge is associated with a set of traversal records.
- Divide the time space into intervals with equal width.
  - The default value is 1 hour, (24 intervals in total).
- For each edge e:
  - Build equi-width histograms for each time interval.
  - The number of buckets per time interval is configurable.
  - The histograms are isomorphic.

**Histogram Bucket Reduction**
- Further reduce the storage size of an individual histogram by merging adjacent buckets.
  - Use SSE to measure the merge cost (accuracy loss).
  - Merge buckets when the cost does not exceed threshold $T_{\text{nop}}$.
  - Iteratively merge adjacent buckets in all the histograms of a road segment.

**Histogram Merging**
- For each edge, merge two temporally adjacent histograms if they are sufficiently similar.
- Use cosine similarity to quantify similarity:
  $$\text{sim}(H_i, H_j) = \frac{V(H_i) \cdot V(H_j)}{\|V(H_i)\| \cdot \|V(H_j)\|}$$
- We use a merge threshold $T_{\text{merge}}$ to decide when to stop merging.

**Route Cost Estimation**
- For a route:
  - Estimate the distribution of GHG emissions as a histogram.
  - Aggregate the histograms of the edges in the route.
- Given two histograms $H_i$ and $H_j$ for adjacent edges:
  - A histogram $H'$ is computed that represents the aggregated GHG emissions distribution for traversing both edges.
  $$H' = H_i + H_j$$

**Outline**
- **Approach I** – histograms
  - Setting
  - Framework
  - ERN building
  - GHG emissions estimation
- **Approach II** – GMMs
  - Setting
  - MTUG building
  - Cost estimation

**Road Network Model**
- **MTUG**: Multi-cost, Time-dependent, Uncertain Graph
- Assume $N$ different costs of interest:
  - Distance (DI), travel time (TT), GHG emissions (GE)
- $G = (V, E, MM, W)$
  - $V$ is the vertex set, and $E$ is the edge set.
  - $MM = \langle MM^{\langle 1 \rangle}, \ldots, MM^{\langle N \rangle} \rangle$
  - Function $MM^{\langle i \rangle}$ maps an edge to the minimum and maximum $i$-th cost of using the edge.
    - $MM^{\langle DI \rangle}(e_i) = (150 \text{ seconds}, 500 \text{ seconds})$
    - $MM^{\langle GE \rangle}(e_i) = (10 \text{ ml}, 85 \text{ ml})$
  - $W = \langle W^{\langle 1 \rangle}, \ldots, W^{\langle N \rangle} \rangle$
  - Function $W^{\langle i \rangle}$ maps an edge to a set of (interval, random variable) pairs of the $i$-th cost type.
    - $W^{\langle DI \rangle}(e_i) = \{[[0:00, 7:15], N(300, 120)), ([7:15, 8:45], N(450, 100)), \ldots\}$
    - $W^{\langle GE \rangle}(e_i) = \{[[0:00, 7:00), N(30, 100)), ([7:00, 9:00), N(50, 85)), \ldots\}$
Instantiation of **MM** in an MTUG

- **MM** and **W** are instantiated using GPS records.
- GPS records are map matched to edges.
- Each edge is associated with a set of traversal records of the form \((e, t, C)\).
  - An edge record indicates that a traversal on edge \(e\) at time \(t\) takes costs \(C\), where \(C\) is a vector of all costs of interest.
  - \((e_1, 8:08, <55 \text{ seconds}, 80 \text{ ml})\)
  - \((e_2, 9:18, <45 \text{ seconds}, 63 \text{ ml})\)
  - \((e_3, 10:10, <43 \text{ seconds}, 60 \text{ ml})\)
  - \((e_4, 21:03, <45 \text{ seconds}, 62 \text{ ml})\)
- Based on the edge records on an edge, functions **MM** on the edge can be instantiated.
  - **MM**(TT)\((e_i) = (43 \text{ seconds}, 55 \text{ seconds})\)
  - **MM**(GE)\((e_i) = (60 \text{ ml}, 80 \text{ ml})\)

Instantiation of **W** in an MTUG (cont.)

- If two RVs in two adjacent intervals are similar, we combine the two intervals into a long interval.
  - Use KL-divergence to measure the similarity between two RVs.
  - Re-estimate a new RV for the long interval using the costs in the long interval.
  - The whole procedure works iteratively until no RVs from consecutive intervals are similar enough to be combined.
  - The long intervals along with their RVs instantiate **W**.

**Route Costs in MTUG**

- Given a route \(R = <r_1, r_2, \ldots, r_X>\), where \(r_i \in E\) is an edge.
- **RC**\((R_i, t)\) indicates the costs of using route \(R_i\) at time \(t\)
  - **RC**\((R_i, t) = <\text{RV}_{DI}, \text{RV}_{TT}, \text{RV}_{GE}>\) is a vector of RVs, and each RV corresponds to a travel cost.
  - **RV**\(_{DI}\) is a deterministic value, which equals to the sum of the length of each edge in route \(R_i\).
  - **RV**\(_{TT}\) is the *convolution* of the corresponding travel time RV of each edge in route \(R_i\).
    - Deciding the travel time RV of the first edge \(r_i\) is dependent on the trip start time \(t\).
    - Deciding the travel time RV of the \(k\)-th edge \(r_i\) is dependent on the travel time of the previous \(k-1\) edges, which may be uncertain.
  - **RV**\(_{GE}\) is the *convolution* of the corresponding GHG emission RV of each edge in route \(R_i\).

**Deterministic Skyline Routes**

- Route cost: \(\text{cost}(R_i) = <\text{DI}, \text{TT}, \text{GE}>\)
  - A vector of deterministic values.
  - Each value corresponds to a travel cost.
- Dominance relationship
  - \(R_i\) dominates \(R_j\) if all the costs of \(R_i\) are no greater than those of \(R_j\), and there is at least one cost of \(R_i\) is smaller than that of \(R_j\).
- Consider multiple routes for the same source-destination.
  - \(R_1\): 3.5 km, 230 mg, 10 min;
  - \(R_2\): 5.1 km, 250 mg, 11 min;
  - \(R_3\): 5.1 km, 200 mg, 12 min;
- The skyline routes are the non-dominated routes.
  - Since \(R_2\) is dominated by \(R_1\), \(R_1\) and \(R_3\) are the skyline routes.

**Stochastic Skyline Route Planning**

**Under Time-Varying Uncertainty**
Stochastic Dominance

- Route cost: \( RC(R_i) = \langle RV_{DI}, RV_{TT}, RV_{GE} \rangle \)
  - A vector of random variables (RVs), where each RV represents the distribution of a travel cost.

- Stochastic Dominance between two RVs
  - Given two RVs \( X \) and \( Y \), if \( \text{cdf}_X(a) \geq \text{cdf}_Y(a) \), for all possible value \( a \) in \( R \), we say \( X \) stochastically dominates \( Y \).

\[
\text{Cost}(R_1) \text{RV}_{TT} \text{ stochastically dominates Cost}(R_3). \text{ RV}_{TT}.
\]
\[
\text{Cost}(R_2) \text{RV}_{TT} \text{ stochastically dominates Cost}(R_4). \text{ RV}_{TT}.
\]
\[
\text{No stochastic dominance between Cost}(R_3). \text{ RV}_{TT} \text{ and Cost}(R_4). \text{ RV}_{TT}.
\]

Example Result

- Skyline routes R1, R2, and R3, identified by our algorithm
  - R1: 94,849 m; R2: 106,216 m; R3: 91,382 m;
    - DI: R3 dominates R1 and R2.

\[
\begin{align*}
\text{(a) cdf of } TT \\
\text{(b) cdf of } GE
\end{align*}
\]

- TT: R1 dominates R2 and R3.
- GE: R2 dominates R1 and R3.

Stochastic Skyline Routes

- Dominance between two routes \( R_i \) and \( R_j \)
  - If each RV of cost(\( R_i \)) stochastically dominates the corresponding RV of cost(\( R_j \)), then \( R_i \) dominates \( R_j \).

- Stochastic skyline routes
  - Given a source-destination pair and a trip starting time
  - The stochastic skyline routes are the routes that are not dominated by any other routes.

Early Pruning Strategy

- Do the following for all travel cost types of interest.
  - We use travel time as an example.
  - We maintain a graph where each edge is associated with the minimum travel time, which is recorded in MM.
  - From the destination, run \( D^*/L^* \) algorithm on the graph.
    - As each vertex is associated with the minimum travel time, we get the "fastest" possible route from the source to the destination.
    - Change the route cost of the "fastest" route to \( \infty \) and add the route as a candidate Skyline route.

\[
\begin{align*}
\text{Candidate skyline route 1} & \quad \text{Candidate skyline route 2} \\
V_s & \quad V_1 \\
V_2 & \quad V_d \\
V_3 & \quad \text{Candidate skyline route 3}
\end{align*}
\]

Each vertex \( v \) has the shortest distance least travel time least GHG emissions to the destination \( v_d \).
Early Pruning Strategy (cont.)
- Explore routes from source, until no more routes can be explored.
  - Estimate the least possible travel costs for a partially explored route.
  - If the partially explored route with its estimated least possible costs is dominated by an existing candidate skyline route, there is no need to explore the route any further.
  - Otherwise, continue exploring.
  - Update candidate skyline route if necessary.

Stochastic Dominance Checking
- Naïve approach: check according to the definition of stochastic dominance.
  - For each value of $a$, check whether $\text{cdf}_X(a) \geq \text{cdf}_Y(a)$
- An efficient approach
  - Consider one cost type at a time
  - Compute the minimum and maximum possible travel costs of a route
- Distinguish among three cases based on the min and max travel costs of two routes
  - Disjoint case: dominance

Stochastic Dominance Checking (cont.)
- Covered case: non-dominance

Summary
- Described a framework that enables stochastic skyline route planning in road networks with multiple, time-dependent, and uncertain travel costs.
- Enables eco-routing in a realistic setting.

Personalized Routing
- Different drivers may take different routes because they may have quite different preferences.
- The same drivers may take different routes in different contexts.
  - Morning: try to save time to avoid being late.
  - Weekend afternoon: try to save fuel consumption.
- Challenges
  - Identify contexts for drivers and identify driving preference in each context.
  - Deal with time-dependent uncertain travel costs, e.g., travel time and fuel consumption, while considering individual drivers' driving behaviors, e.g., aggressive vs. moderate driving.
**Vehicle Routing with User-Generated Trajectory Data**

**Introduction**
- Local travel
  - Knowledge of the surroundings
  - Follow familiar routes
- Travel in unfamiliar surroundings to unknown destinations
  - Depend on available routing services
  - Expect that the provided route is the best
- Idea: Use GPS data to let those who travel in unfamiliar surroundings benefit from the insights of local travelers

**Goal of the Study**
- Propose a routing framework that
  - Utilizes GPS data volunteered by local drivers
  - Exploits possibly hard-to-formalize insight into local conditions
  - Takes into account temporal variation in driver behavior
  - Recommends routes based on popularity and temporal aspects
- Evaluate the quality of proposed routes
  - Study based on trip length and pre-selected drivers
  - Quality comparison with existing routing service and route recommendation approaches

**Example Results**
- Dark, bold routes: actual routes used by drivers.
- Red routes: shortest routes.
- Green routes: fastest routes.
- Blue routes: predicted routes using the identified contexts and driving preferences.

**Framework**

**Offline Phase**
- Drivers’ Trajectories
- Context and Preference Identification

**Online Phase**
- Driver, Source, Destination, Time
- Contexts & Preferences
- Routing Module
- Optimal Routes

**Trips used:**
- Start or end at, or go through, source and destination locations
- Trips that start during the issued time period are preferred

**When the source and destination are not covered by available GPS data, an existing routing service is used.**
Data Preparation Methodology

- Trips that follow the same sequence of road segments are grouped into route usage objects.

<table>
<thead>
<tr>
<th>User</th>
<th>Route</th>
<th># Traversals</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>B</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>C</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>D</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>E</td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>F</td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>G</td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>H</td>
<td></td>
</tr>
<tr>
<td>H</td>
<td>I</td>
<td></td>
</tr>
</tbody>
</table>

- Route $r_2$ is taken by users $u_2$ and $u_4$ 3 and 5 times during peak hours and 5 and 6 times during off-peak hours.

$(r_2,p_i,\{(\text{peak},\{(u_2,3),(u_4,5)\}),\{(\text{off},\{(u_2,5),(u_4,6)\})\}))$

Scoring of Routes

Preferred routes
- Popular among drivers
- Taken by many distinct drivers
- Popular on the time of the day and day of the week of the query

Scoring of Routes

Route preference value:

$\text{pref}(r) = \alpha \cdot \text{users}(r) + (1 - \alpha) \cdot \text{traversals}(r)$

- $\alpha$: number of distinct drivers taking the route
- $\beta$: number of traversals of the route

Final route score:

$\text{score}(r) = \beta \cdot \text{pref}^M(r) + (1 - \beta) \cdot \text{pref}^N(r)$

- $\beta$: considers trips taken during the query temporal pattern
- $(1 - \beta)$: considers trips taken during other temporal patterns

Empirical Study: Data

- Monitoring period: 2 years
- Number of drivers: 285
- Number of GPS points (raw data): ~182,700,000
- Number of trips: ~275,000
- "Pay as You Speed" project (http://www.sparenfahrten.de)

Routing Quality Evaluation: Data

For this study, we randomly selected equal amounts of trips for different trip length intervals

Routing Quality: Match

A match is identified using LCSS

Trajectory DB

Routing Service

Compare route with trajectory

Source and destination

Preferred route

Total number of matches
Routing Quality Evaluation: Results

Our Proposal

Google Directions API (top-1)

Routing Quality Evaluation: Data

For this study, we considered the five drivers with the most trips.

Related Work

Four existing routing techniques use drivers’ trajectories

1. The road network is formed from the road segments that are covered by the trajectory data set
2. A route is formed by
   - Prioritizing parts of roads that are followed the most by a specific driver (personalized routes) [1]
   - Prioritizing parts of roads that are taken by other drivers [2],[4]
   - Possibly using sub-routes from multiple routes [3]
3. Trajectories used for scoring must contain the destination and must start and end during the provided time interval. [3]
4. Suggested routes are formed from the most popular routes or route parts in the available data set.

Conclusion and Research Directions

Conclusions

- The proposed framework utilizes trajectory data collected from local drivers for routing.
- A preferred route is selected using a flexible scoring function that considers
  - The number of traversals of the route
  - The number of distinct drivers taking the route
  - The time periods when the traversals occurred
  - Use of travel histories of local drivers can increase routing quality
- More details in the paper!

Research directions

- Additional aspects of the framework can be considered
  - Efficiency of route identification process (LCSS technique)
  - Inclusion of personalized routes
  - Better support for routes that are constructed from sub-routes

Closing
System of Sensors Model

- The setting may be modeled as a system of (logical) streams, one per edge.
  - Data is emitted from the stream of an edge when a vehicle traverses the edge
  - Spatial
  - Spatio-temporally correlated
  - Sparse
- Real, unlike early, envisioned smart dust applications!

Demos and Prototype Systems

- **EcoTour**: [http://daisy.aau.dk/its/]
  - Computes and compares the shortest, the fastest, and the most eco-friendly routes for arbitrary source-destination pairs in DK.
  - Best demo award at IEEE MDM 2013.
- **EcoSky**: [http://daisy.aau.dk/its/eco/]
  - Supports skyline eco-routing and personalized eco-routing
- **Sheafs**: [http://daisy.aau.dk/its/sheaf]
  - Trajectory based traffic sheafs
- **Strict-Path Queries**: [http://daisy.aau.dk/its/spgdemo]
  - Trajectory based, Strict-Path Queries
  - Trips (historical travel-time), route choice, Napoleon (road usage)
- **iPark**: identifying parking spaces from GPS trajectories
  - On-street parking lanes vs. parking zones

The Future

- Much more travel data
  - GPS data from vehicles
  - Inductive loop detectors, Wi-Fi/Bluetooth
  - Collective transport data, e.g., bus data,
  - Multimodal collective transport data, e.g., "Rejsekortet"
- Much more connected vehicles
- New services
  - Routing
  - Safety and warnings
  - Parking, fees, insurance, road pricing
  - Car sharing, multi-modality
- Self-driving vehicles

Challenges, Examples

- Detect: "black spots" before they occur.
- Modeling spatio-temporal congestion from data
- Characterize the effects of events
  - Accidents, malfunctioning of traffic signals, rain, a concert
- Real-time traffic management
  - In response to current or predicted situation, actuate traffic signals and drivers (via their smartphones or navigation devices) to optimize the use of the infrastructure and driver experience
- Automated trade-off between weight level of detail and available data.
- Stochastic routing at 20 milliseconds.
- Integrate with "point" data

Acknowledgments

- Colleagues at Aalborg University, Aarhus University, and beyond.
- The EU FP7 project, Reduction: [http://www.reduction-project.eu]
- The Obel Family Foundation: [http://www.obel.com/en]

Readings

- V. Ceikute, C. S. Jensen: Vehicle Routing with User-Generated Trajectory Data. MDM (1) 2015
Readings

- C. Guo, Y. Ma, B. Yang, C. S. Jensen, Manohar Kaul: EcoMark: evaluating models of vehicular environmental impact. SIGSPATIAL/GIS 2012: 269-278