Veracity in Big Data
Reliability of Routes

Dr. Tobias Emrich
Post-Doctoral Scholar
Integrated Media Systems Center (IMSC)
Viterbi School of Engineering
University of Southern California
Los Angeles, CA 900890781
emrich@usc.edu
OUTLINE

• Big (Uncertain) Data
• Reliability in Traffic Networks
• From Uncertainty to Reliability
• Outlook
BIG DATA

“Big data is like ...
everyone talks about it,
nobody really knows how to do it,
everyone thinks everyone else is doing it,
so everyone claims they are doing it...”

Dan Ariely
Center for Advanced Hindsight at Duke University
"Big data is like ...: everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it...”

Dan Ariely
Center for Advanced Hindsight at Duke University

Median Hadoop-job is ~13 GB
90% of the jobs are < 100 GB

BIG DATA

Variety

Volume

Veracity

Velocity

Value
Uncertainty in (not so big) Data


Uncertainty Databases

Uncertainty is inherent in many datasets:
- Automated Extraction of Information from HTML
- **Sensor Readings**
- Human Observations
- **Predictions**
- ...
Reliability in Traffic Networks

Usually, I take route A when I have a meeting in the morning,

I take route B
Reliability in Traffic Networks

Travel time prediction, incorporating uncertainty

Route A: ~53 min
Route B: ~58 min

Travel Time

Probability

Route A
Route B
Reliability in Traffic Networks

- Predicted travel time of a route vary due to:
  - Imprecise Prediction of traffic flow
  - Unpredictable accidents
  - Changing weather conditions
  - ...
- Routes differ in variance of travel time
- Starting at 9am
  - Route A arrives before 10am with 89%
  - Route B arrives before 10am with 99.2%
  - To have 99.2% on route A I have to leave 10 mins earlier
From Uncertainty to Reliability

- Current approaches

9.00pm  \[ tt(9.00) = 5 \text{ min} \]

\[ tt(9.00) = 6 \text{ min} \]

\[ tt(9.00) = 3 \text{ min} \]

S  \quad \text{Traditional} \quad D

\[ TT = 14 \text{ min} \]
From Uncertainty to Reliability

• Current approaches

Traditional

ClearPath

9.00pm

S

tt(9.00) = 5 min

tt(9.05) = 5 min

D

TT = 14 min

TT = 12 min
From Uncertainty to Reliability

• Considering uncertainty of predictions

9.00pm

S

tt(9.00) = 5 min

D

tt(9.05) = ? min

How to predict uncertain travel time?

How to model uncertain travel time?
From Uncertainty to Reliability

• Considering uncertainty of predictions

9.00pm

S

D

tt(9.00) = 5 min

tt(9.05) = 
4 min [20%]
5 min [50%]
6 min [30%]

How to predict uncertain travel time?

How to model uncertain travel time?
From Uncertainty to Reliability

- Considering uncertainty of predictions

9.00pm

S

\[ \text{tt}(9.00) = 5 \text{ min} \]

\[ \text{tt}(9.05) = \begin{cases} 4 \text{ min [20%]} \\ 5 \text{ min [50%]} \\ 6 \text{ min [30%]} \end{cases} \]

D

\[ \text{tt}(?) = ? \text{ min} \]

How to predict uncertain travel time?

What time to use?

How to model uncertain travel time?
From Uncertainty to Reliability

- Considering uncertainty of predictions

How to predict uncertain travel time?

What time to use?

How to model uncertain travel time?
From Uncertainty to Reliability

• Considering uncertainty of predictions

\[ \text{tt}(9.00) = 5 \text{ min} \]

9.00pm

\[ \text{tt}(9.05) = \begin{cases} 4 \text{ min [20\%]} \\ 5 \text{ min [50\%]} \\ 6 \text{ min [30\%]} \end{cases} \]

\[ \text{tt}(9.04) = \begin{cases} 2 \text{ min [60\%]} \\ 3 \text{ min [40\%]} \\ 4 \text{ min [10\%]} \end{cases} \]

\[ \text{tt}(9.05) = \ldots \]

\[ \text{tt}(9.06) = \ldots \]

\[ \text{TT} = ? \text{ min} \]

How to predict uncertain travel time?

What time to use?

How to model uncertain travel time?

How to add up uncertain travel times?
From Uncertainty to Reliability

- Considering uncertainty of predictions

- How to predict uncertain travel time?
- What time to use?
- How to model uncertain travel time?
- How to add up uncertain travel times?
- How to deal with correlations?

tt(9.00) = 5 min

S

9.00pm

D

tt(9.05) = 4 min [20%], 5 min [50%], 6 min [30%]

tt(9.04) = 2 min [60%], 3 min [40%], 4 min [10%]

tt(9.05) = ...

tt(9.06) = ...

TT = ? min

USC Viterbi
School of Engineering
Integrated Media Systems Center
From Uncertainty to Reliability

- Considering uncertainty of predictions

How to deal with correlations?

How to predict uncertain travel time?

What time to use?

How to model uncertain travel time?

How to add up uncertain travel times?

9.00pm

\[ TT = ? \text{ min} \]

\[ \text{tt}(9.00) = 5 \text{ min} \]

\[ \text{tt}(9.04) = 2 \text{ min [60%]} \]

\[ \text{tt}(9.05) = 3 \text{ min [40%]} \]

\[ \text{tt}(9.06) = 4 \text{ min [10%]} \]

\[ \text{tt}(9.05) = \ldots \]

\[ \text{tt}(9.06) = \ldots \]
Outlook

- Evaluation of the quality of the result
- Efficient online prediction

- Extension to new query mechanisms:
  “When do I have to start (and which route do I have to take) when I want to be at USC at 8.00am (or before) with a probability of 99%?”
Questions?

Dr. Tobias Emrich
emrich@usc.edu
Uncertainty in Databases

• Uncertainty is inherent in many datasets:
  – Automated Extraction of Information from HTML
    (i.e. John works at Google vs. John works at Microsoft)
  – Sensor Readings
    (i.e. RFID sensors tracking the position of customers)
  – Human Observations
    (i.e. the seen Bird was either a Raven (75%) or a Crow (25%))
  – Predictions
    (i.e. tomorrow its going to rain (10%) or not(90%))
  – ...

• Two approaches to solve this
  – Cleaning (e.g. get rid of uncertainty)
  – Management (e.g. handle the uncertainty)