# From smartphone data to route choice modeling

Michel Bierlaire

Transport and Mobility Laboratory

School of Architecture, Civil & Environmental Engineering

Ecole Polytechnique Fédérale de Lausanne, Switzerland

# EPFL

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

# Transport and Mobility Laboratory

- Transportation research
  - Airlines, ports, buses, car traffic, land-use, etc.
- Operations research
  - Nonlinear optimization, column-generation, simulation, Markov chains, etc.
- Discrete choice models
  - Multivariate Extreme Value models, mixtures models, latent variables, Biogeme, etc.

transp-or.epfl.ch

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

# Collaborators

- 5 research associates
- 10 PhD students

On this research:

- Jingmin Chen, PhD student.
- Gunnar Flötteröd, postdoc.

# Outline

- Smartphone data
- Route choice: the chosen route
- Route choice: the non chosen routes

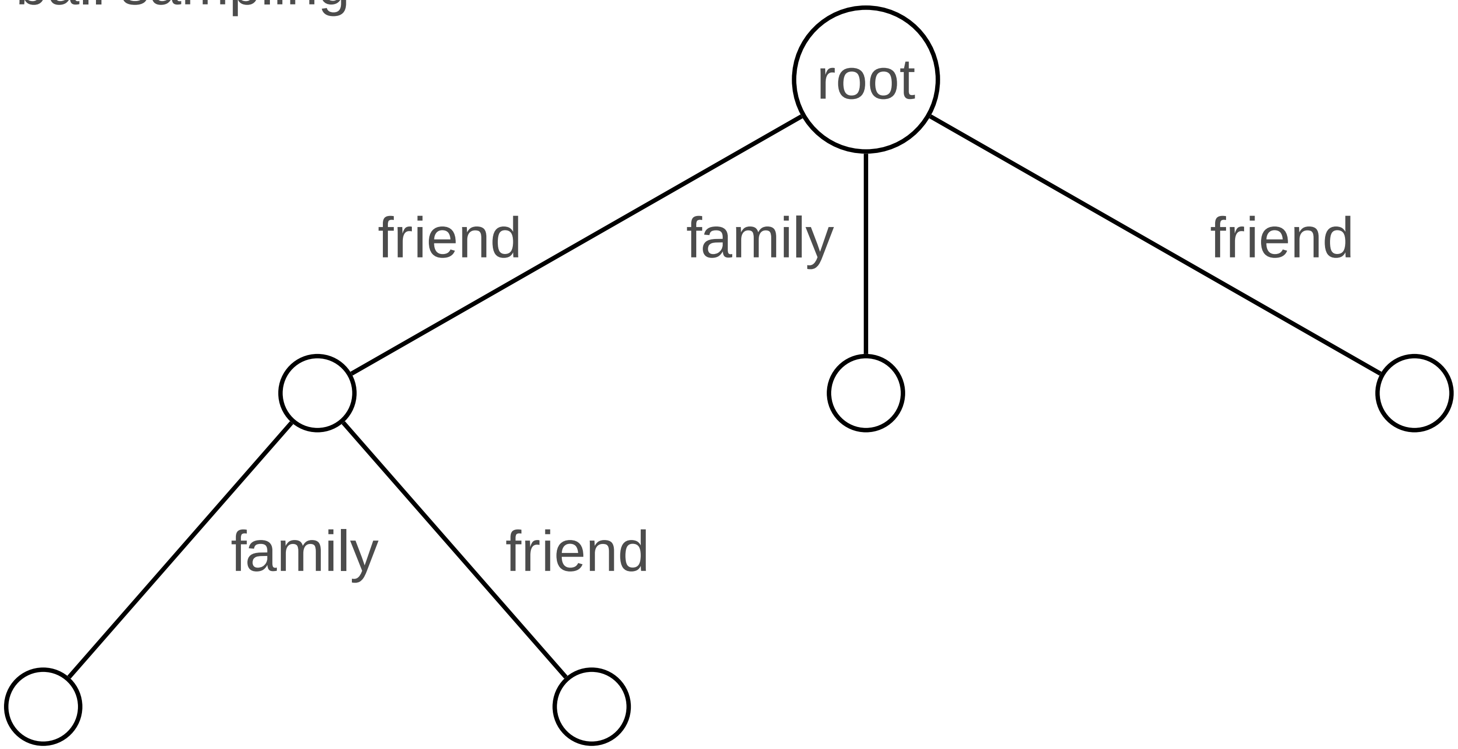
# Nokia data collection campaign

# Nokia data collection campaign

- Funding source: Nokia Research Center (NRC) at EPFL.
- Participants: About 185.
- Since: September 2009.
- Phone: Nokia N95.
- Collaborators: NRC Lausanne, IDIAP (Switzerland).

# Recruitment

snow ball sampling



# Participants

- About $185$ participants.
- Mostly from Lausanne area.
- $\sim 1/3$ females.
- $< 1/4$ students.

# Software design

Phone software (EPFLSCOPE)

- written in python Symbian S60;
- starts with the operating system, runs in backend;
- cannot be turned off by users;
- records data constantly;
- uploads data automatically to DB A via wireless network (WIFI, 3G), every 2 hours.

Databases

- are administrated by Nokia;
- a remote database (DB A) with data access API (httprequest, JSON format);
- another geographical database (DB B) copies data from DB A with $\sim$ 12 hours lag (SQL access).

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

# Energy performance

The original software was developed by Nokia.

- With GPS on, one fully charged battery lasts less than 4 hours.

The energy performance was improved by TRANSP-OR, IDIAP and NRC Lausanne.

- Turn off GPS if stationary.
- Determines stationary/moving: GPS, known WLAN, cell ID, accelerometer.
- One fully charged battery can last $\sim 10$ hours.

# Privacy and security

- Data is owned by participants. They can delete their data from DB A.
- The campaign is permitted and controlled by an ethical committee.
- Nokia and authorized research partners (in CH) get access to the data.

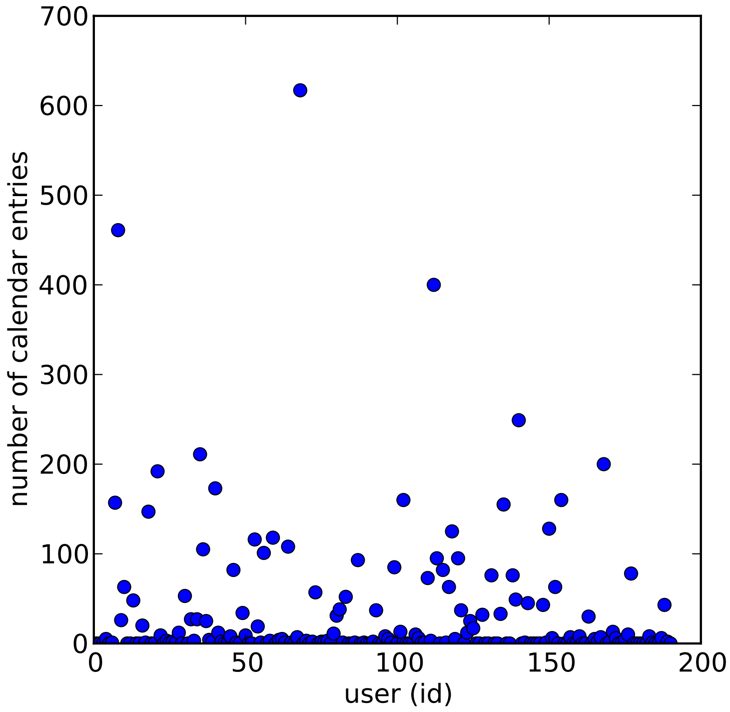
It took ONE YEAR for EPFL to get data access (although data had already been in Nokia's databases).

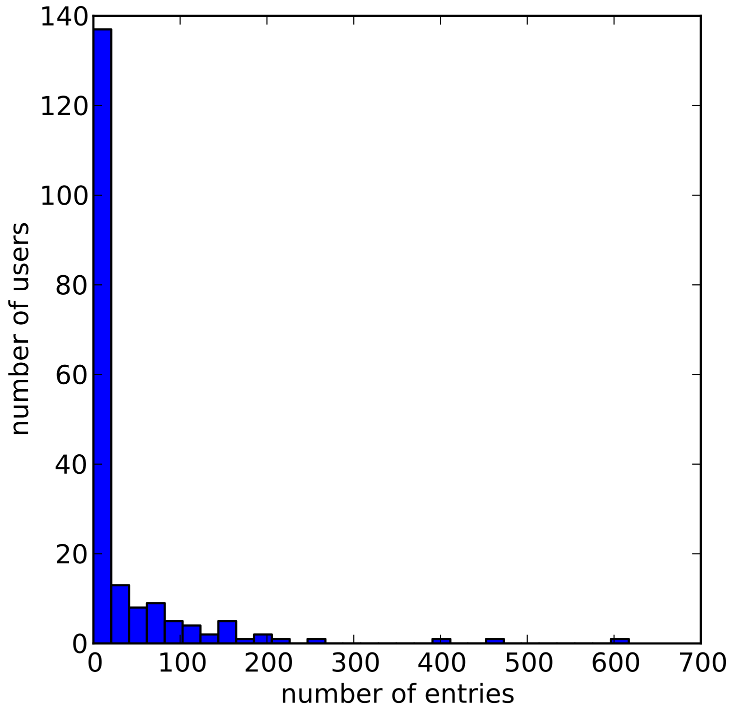
# Data volume

$\sim$ 150k-entries/100MB of data per user per month

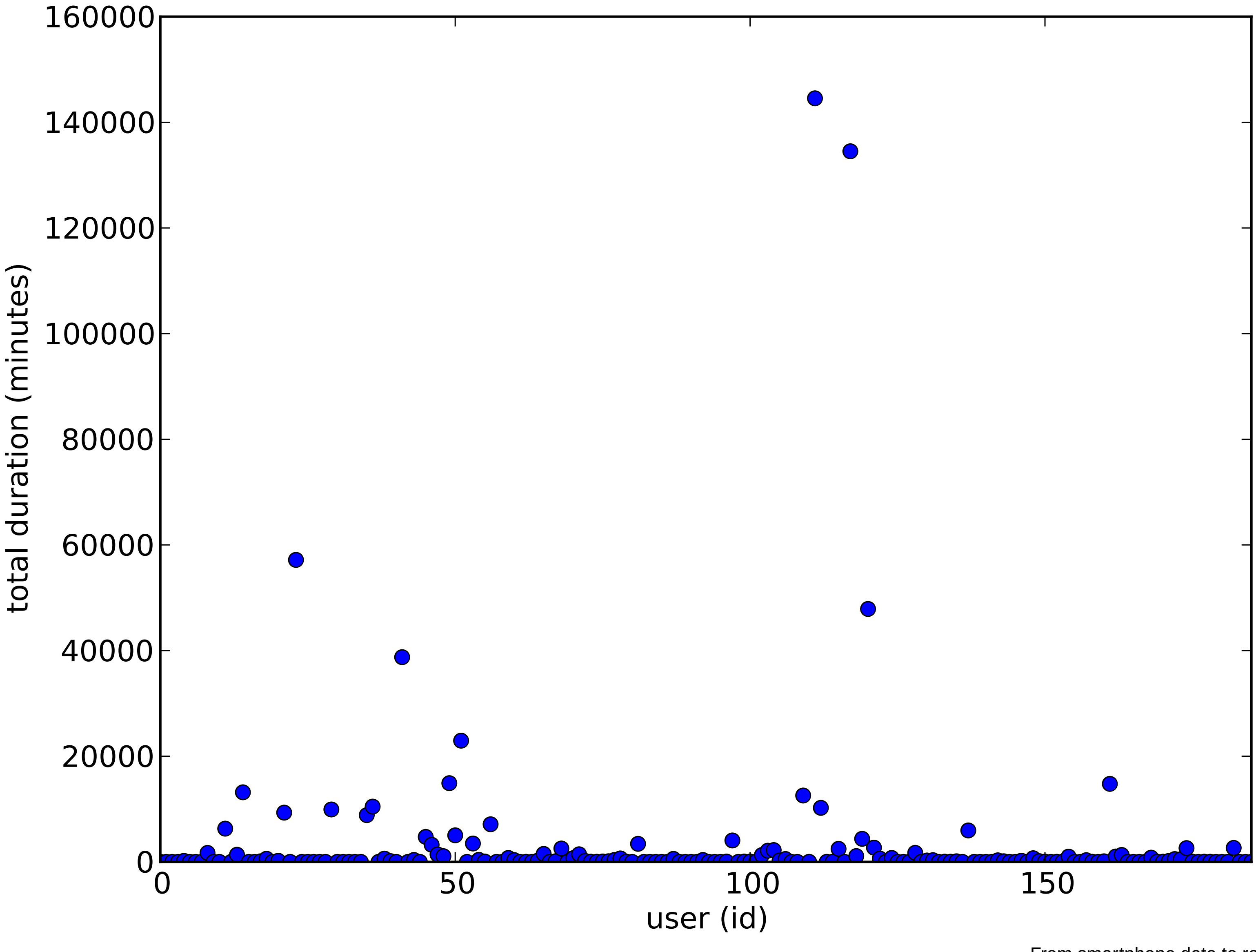
| | |
|---|---|
| Number of GPS points | 11,531,652 |
| Number of calls | 247,448 |
| Duration of calls | 6,903h |
| Number of sms | 179,358 |
| Number of video made | 3,890 |
| Number of pictures taken | 54,537 |
| Number of unique BT | 543,517 |
| Number of unique WIFI | 572,910 |
| Number of unique cell towers (63 countries) | 100,505 |
| Number of unique cell towers (CH) | 28,945 |
| Number of acceleration samples | 1,344,198 |
| Number of application events captures | 8,280,554 |
| Number of phone book entries | 115,134 |

# Calendar: number of entries

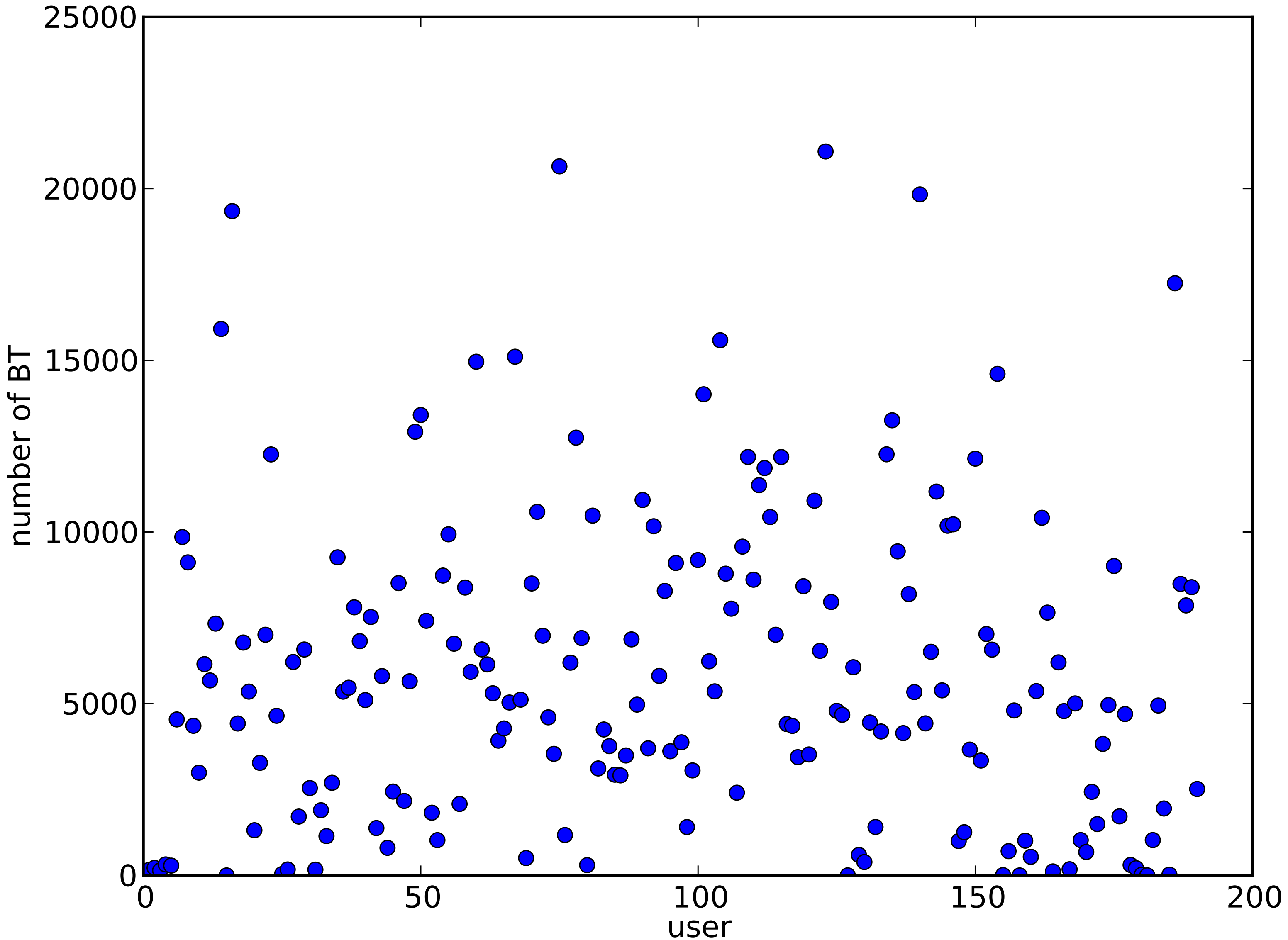


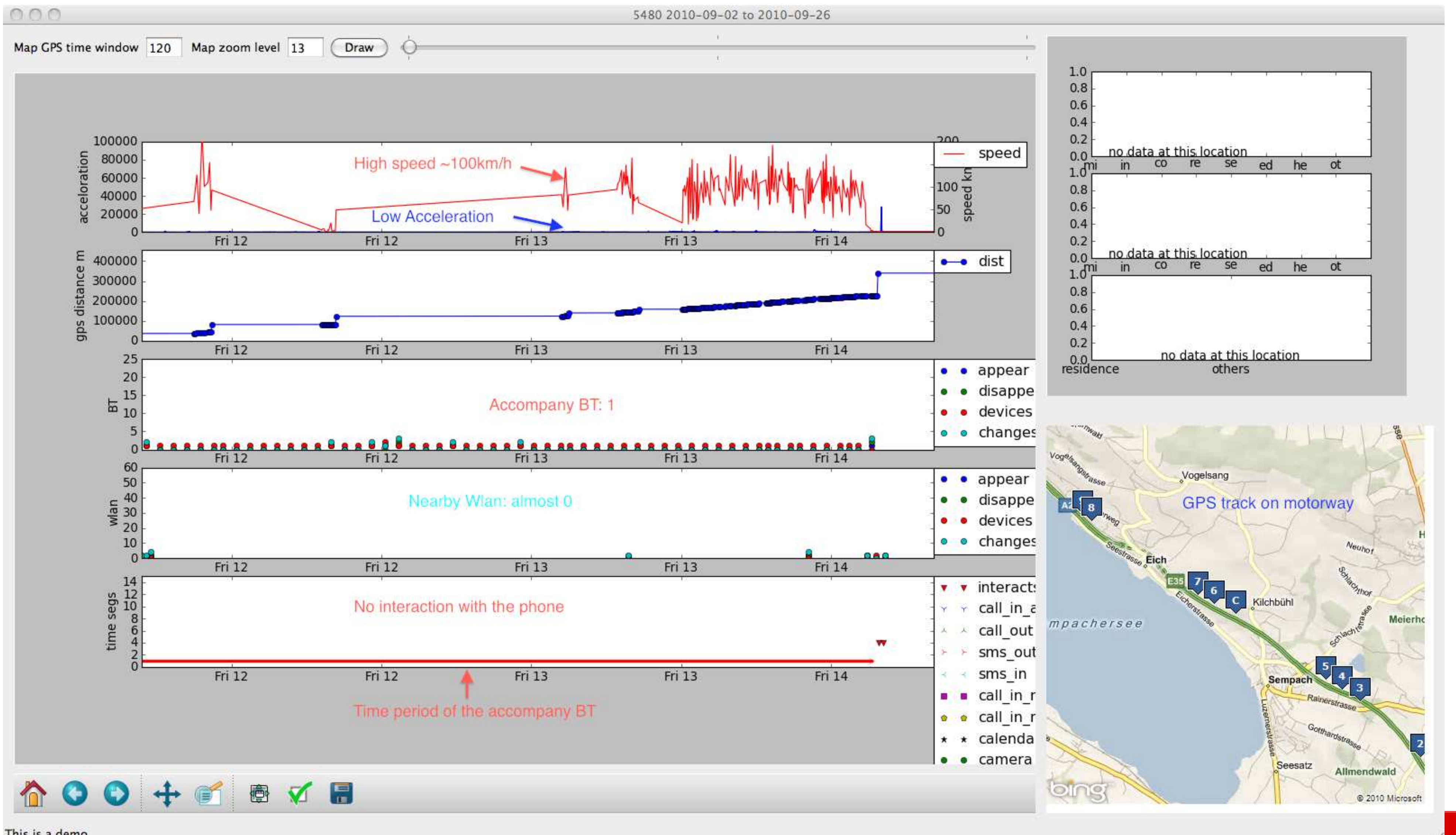
# Media play


160000
140000
120000
100000
80000
60000
40000
20000
0
total duration (minutes)
0
50
100
150
user (id)

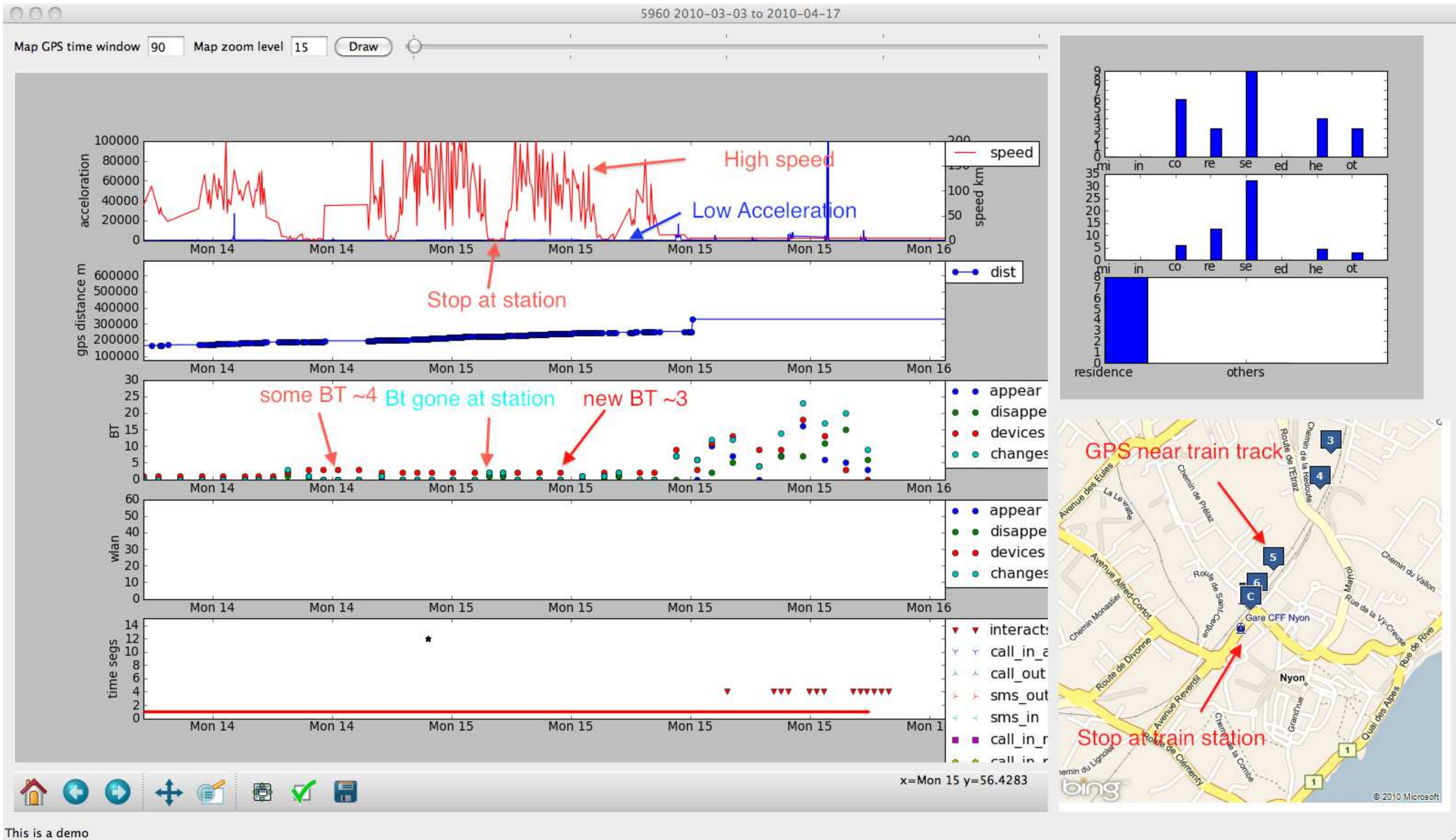
# Number of Bluetooth devices



# Mobility patterns: car



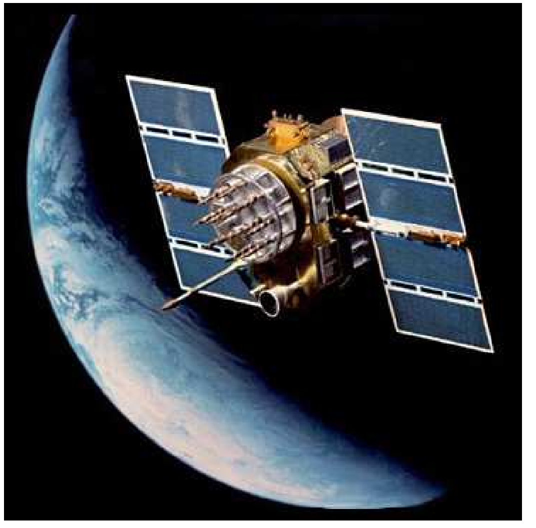
# Mobility patterns: train




TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

# Route choice: the chosen route

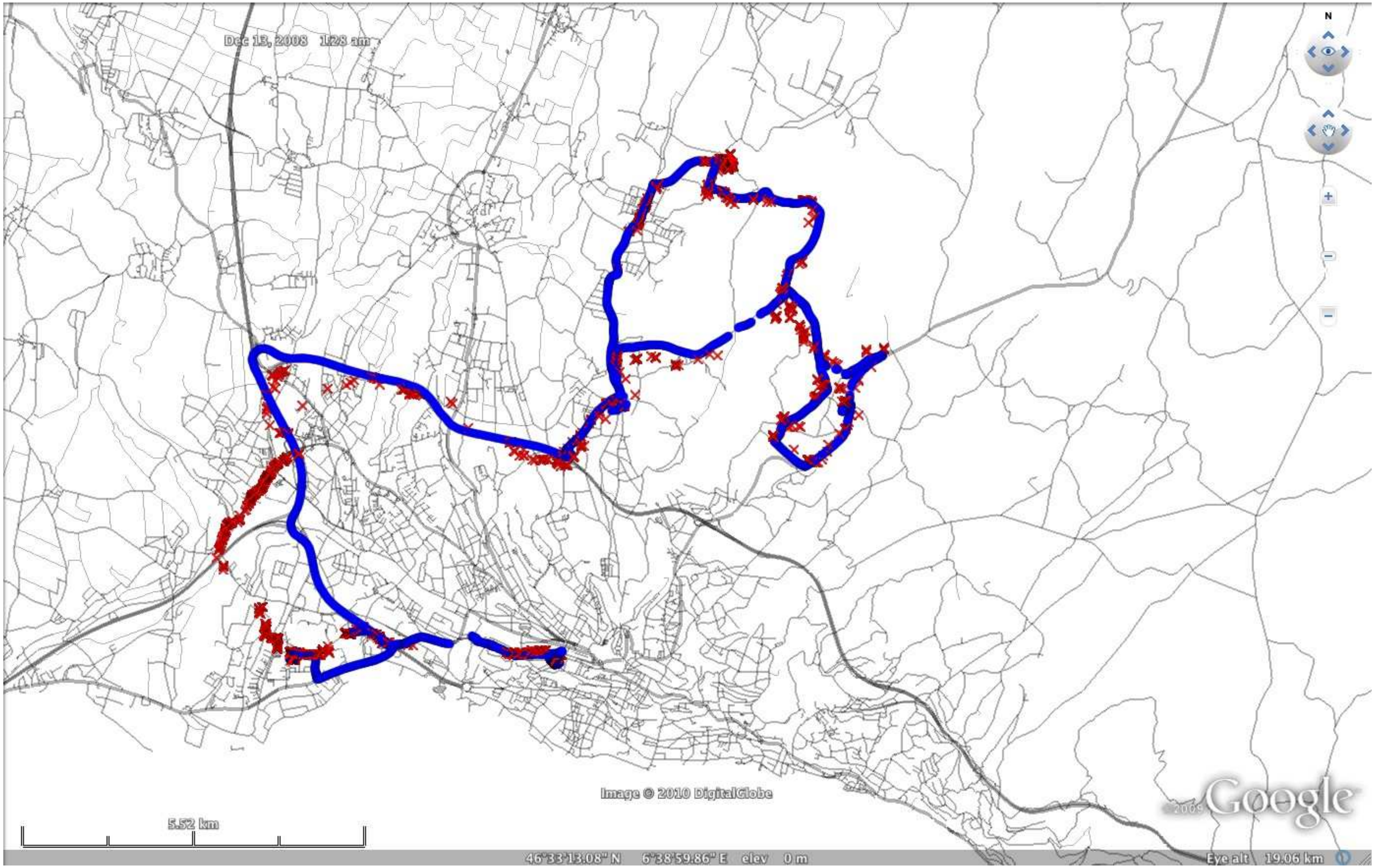


- Focus on GPS data from smartphone
- Objective: reconstruct actual paths

TRANSP-OR

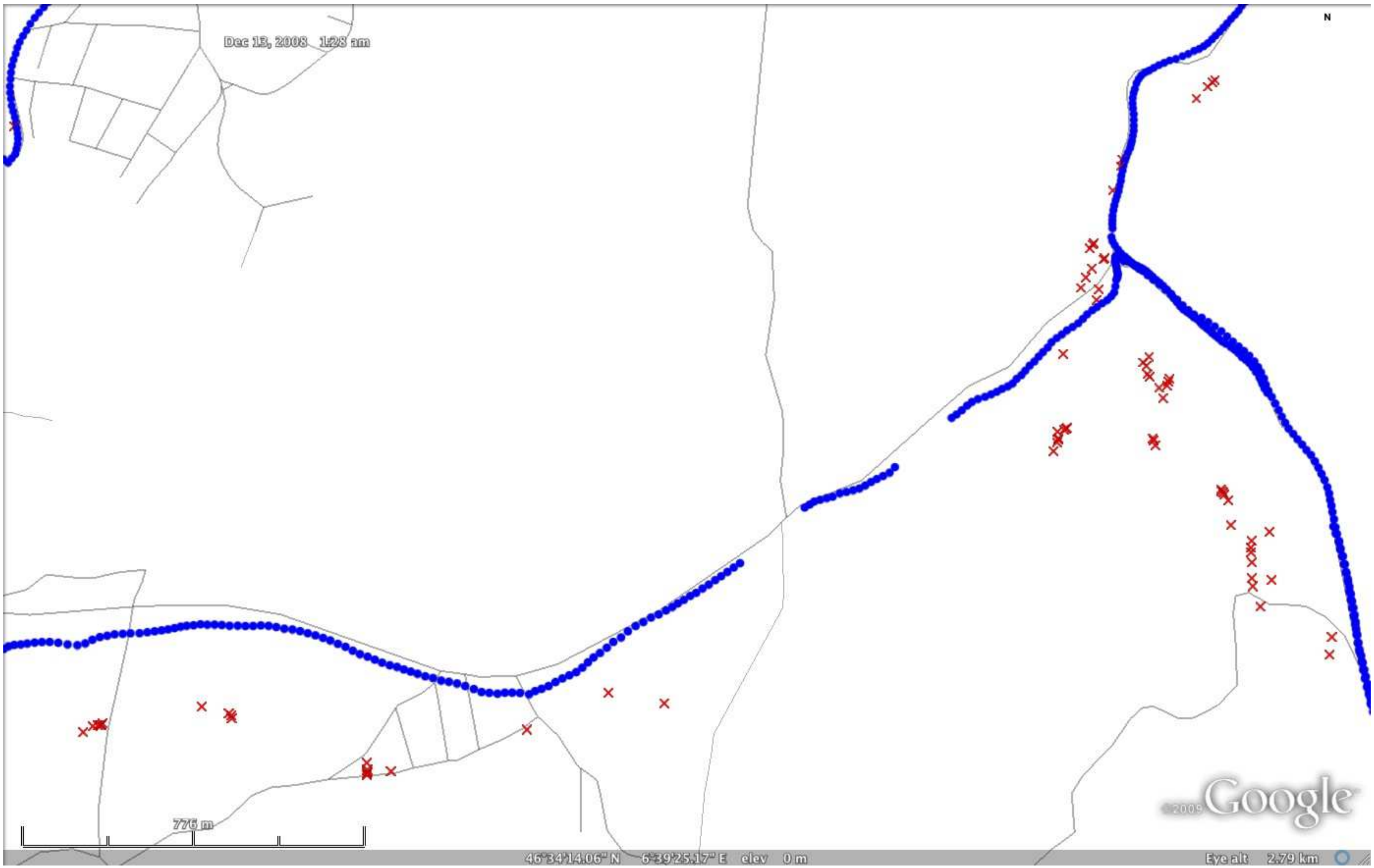
ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

# Issues



Dec 13, 2008 1:28 am
Image © 2010 DigitalGlobe
5.52 km
46°33'13.08" N 6°38'59.86" E elev 0 m
Eye alt 19.06 km
Google

# Issues



Dec 13, 2008 1:28 am
N
776 m
46°34'14.06" N 6°39'25.17" E elev 0 m
Eye alt 2.79 km
©2009 Google

# Issues

- Low data collection rate to save battery (every 10 seconds)
- Inaccuracy due to technological constraints
- Smartphone carried in bags, pockets: weaker signal
- Map matching algorithms do not work with this data

# Context

- Network: $G = (N, A)$
- Node coordinates: $x_n = \{\text{lat}, \text{lon}\}$
- Arc geometry:

$$\mathcal{L}_a : [0, 1] \to \mathbb{R}^2.$$

Example: straight line

$$\mathcal{L}_a(\ell) = (1 - \ell)\, x_u + \ell x_d.$$

- Model for the movement of the mobile phone:

$$x = S(x^-, t^-, t, p)$$

  - Ideally a traffic simulator
  - Simpler models are used in practice
  - Random variable with density $f_x(x|x^-, t^-, t, p)$

# Data

One measurement: $\widehat{g} = \left(\widehat{t}, \widehat{x}, \widehat{\sigma}^x, \widehat{v}, \widehat{\sigma}^v, \widehat{h}\right),$

- $\widehat{t}$, a time stamp ;
- $\widehat{x} = (\widehat{x}_{\text{lat}}, \widehat{x}_{\text{lon}})$, a pair of coordinates;
- $\widehat{\sigma}^x$, the standard deviation of the horizontal error in the location measurement;
- $\widehat{v}$, a speed measurement (km/h) and,
- $\widehat{\sigma}^v$, the standard deviation of the error in that measurement;
- $\widehat{h}$, a heading measurement, that is the angle to the north direction, from 0 to 359, clockwise.

Sequence: $(\widehat{g}_1, \ldots, \widehat{g}_T)$

# Measurement equations

Objective (derivation in the appendix):

- Given a path $p$
- Given a sequence $(\widehat{g}_1, \ldots, \widehat{g}_T)$
- What is the likelihood that the sequence has been generated by a smartphone moving along path $p$?
- Note: different approach from map matching, which is essentially a projection procedure.
- We focus on the position only
- We derive

$$\Pr(\widehat{x}_1, \ldots, \widehat{x}_T | p),$$

- ... recursively

$$\Pr(\widehat{x}_1, \ldots, \widehat{x}_T | p) = \Pr(\widehat{x}_T | \widehat{x}_1, \ldots, \widehat{x}_{T-1}, p) \Pr(\widehat{x}_1, \ldots, \widehat{x}_{T-1} | p).$$

# Case study: true path

# Case study: path with a deviation (1)


TRANSP-OR

EPFL
ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

# Case study: path with a deviation (2)


TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

# Case study: log likelihood from measurement equation

| | |
|---|---|
| True path | -11.3 |
| Deviation 1 | -12.9 |
| Deviation 2 | -13.2 |

- Results are consistent with intuition

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

# Route choice: the non chosen routes

- Choice model: $P_n(i|\mathcal{C}_n)$
- Route choice: what is $\mathcal{C}_n$?
- Many “behaviorally motivated” heuristics proposed in the literature.
- Most of the time, the chosen route is not included.
- `Frejinger, Bierlaire and Ben-Akiva (2009)` propose an econometric approach.
- Idea:
  - Assumption: **all** paths connecting the OD pair are relevant.
  - Issue: enumeration is prohibitive.
  - Solution: sampling of alternatives.

# Sampling of alternatives

- Sample $\mathcal{C}_n$ with replacement from $\mathcal{C}$ according to $\{q(i)\}_{i\in\mathcal{C}}$
- Add the chosen alternative
- $k_{in}$ is the number of times alternative $i$ is contained in $\mathcal{C}_n$
- Correct for sampling when estimating logit model

$$P(i|\mathcal{C}_n) = \frac{e^{\mu V_{in} + \ln\left(\frac{k_{in}}{b(i)}\right)}}{\sum_{j\in\mathcal{C}_n} e^{\mu V_{jn} + \ln\left(\frac{k_{jn}}{b(j)}\right)}}$$

where $\{b(i)\}_{i\in\mathcal{C}}$ is such that $q(i) = b(i)/\sum_{j\in\mathcal{C}} b(j)$

Objective: sample paths according to pre-specified $\{b(i)\}_{i\in\mathcal{C}}$

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

# Using Markov Chains

- Finite state space
- Discrete time $k = 0, 1, \ldots$
- At time $k$, process is in state $i^k$
- $p(i, j)$ is one-step probability to go from state $i$ to state $j$
- Process has a unique stationary distribution if
  - every state eventually reaches every other state
  - there is at least one state $i$ with $p(i, i) > 0$

Objective: build MC of routes with stationary distribution $\{q(i)\}_{i \in \mathcal{C}}$

# Metropolis-Hastings algorithm

- Given
  - a finite state space,
  - positive weights $\{b(i)\}_i$,
  - and irreducible proposal transition distribution $q(i,j)$,
- the Metropolis-Hastings algorithm generates a Markov chain that converges to

$$q(i) = b(i) / \sum_j b(j).$$

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

# Metropolis-Hastings algorithm

1. Set iteration counter $k = 0$
2. Select arbitrary initial state $i^k$
3. Repeat beyond stationarity
   (a) Draw candidate state $j$ from $\{q(i^k, j)\}_j$
   (b) Compute acceptance probability

$$\alpha(i^k, j) = \min\left(\frac{b(j)q(j, i^k)}{b(i^k)q(i^k, j)}, 1\right)$$

   (c) With probability $\alpha(i^k, j)$, let $i^{k+1} = j$; else, let $i^{k+1} = i^k$
   (d) Increase $k$ by one

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

# Using MH for path sampling

- State space comprises $\mathcal{C}$
- Weights $b(i)$ favor plausible paths (importance sampling)
- Typically, paths with length close to the shortest path have high probability to be sampled
- Transition distribution $q(i, j)$ creates local path modifications
  - too little variability: slow convergence
  - too much variability: random search

# State space

- a state $i = (\Gamma, a, b, c)$ consists of
  - a path $\Gamma$
  - three node indices $a < b < c$ within that path
- node indices simplify computation of transition probabilities

# Proposal transition distribution

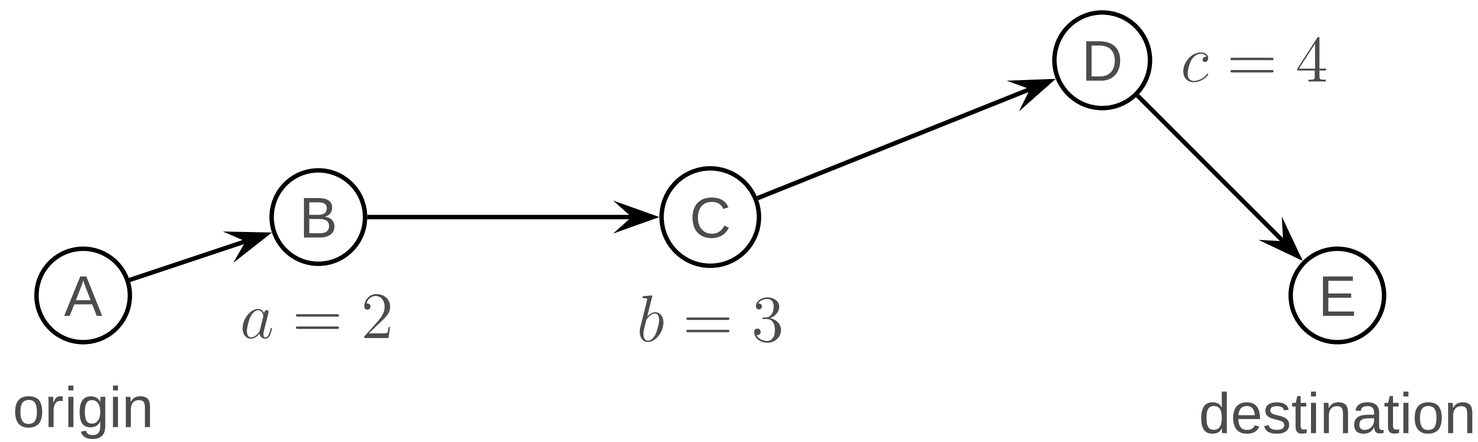
- SHUFFLE operation
  - Re-sample (uniformly) $a < b < c$ within path $\Gamma$
- SPLICE operation
  - Sample a node $v$ “near” the path segment $\Gamma(a, c)$
  - Connect $\Gamma(a)$ to $v$
  - Connect $v$ to $\Gamma(c)$
  - Let new $b$ point at $v$, update $c$
- Overall transition: randomly select one procedure

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

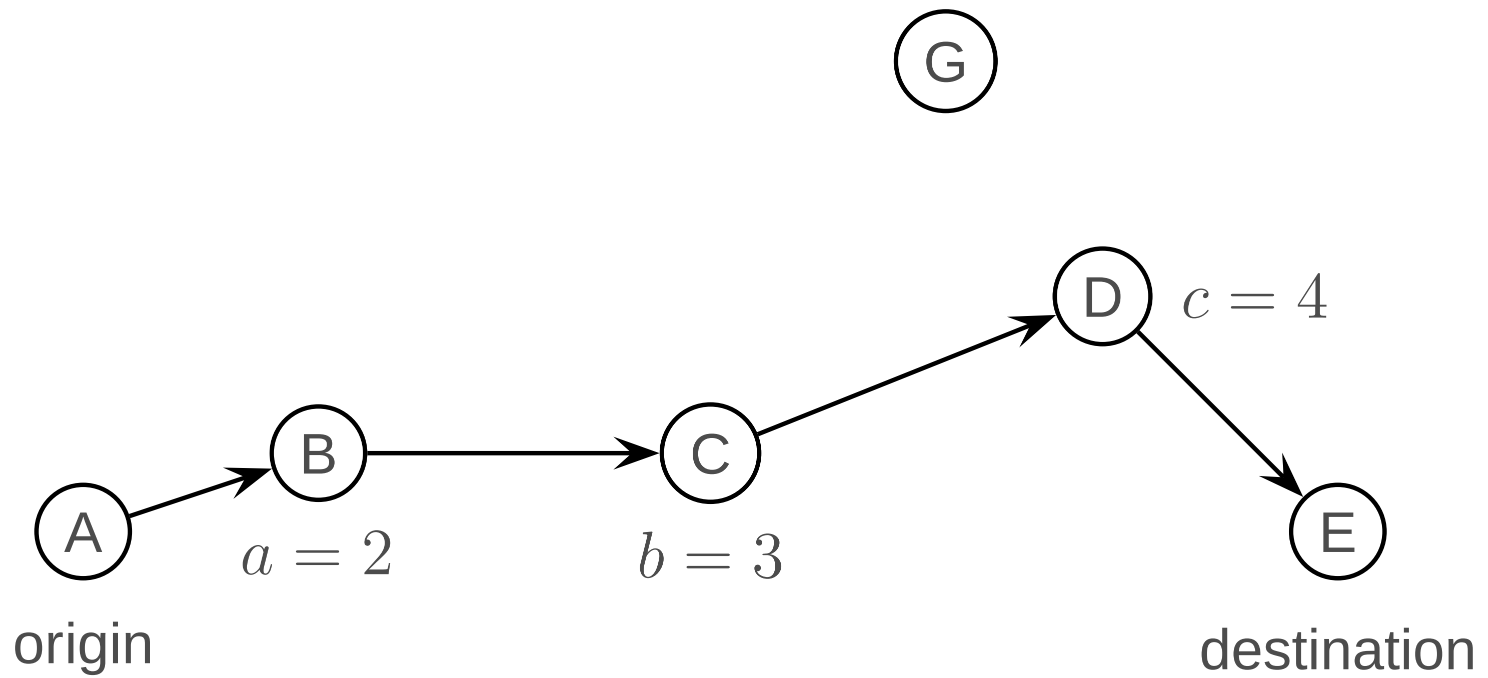
# Proposal transition distribution

A state $(\Gamma, a, b, c)$



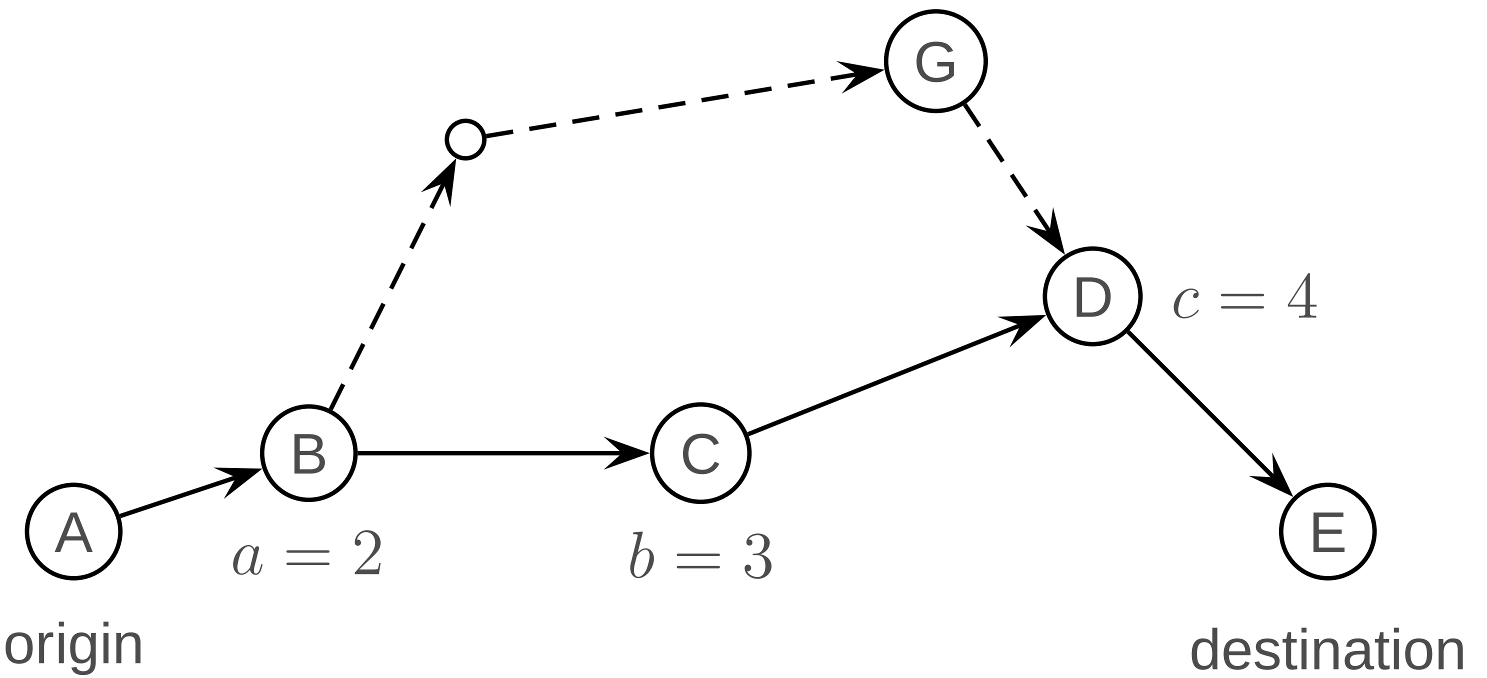
# Proposal transition distribution

SPLICE: a new node $G$ is sampled



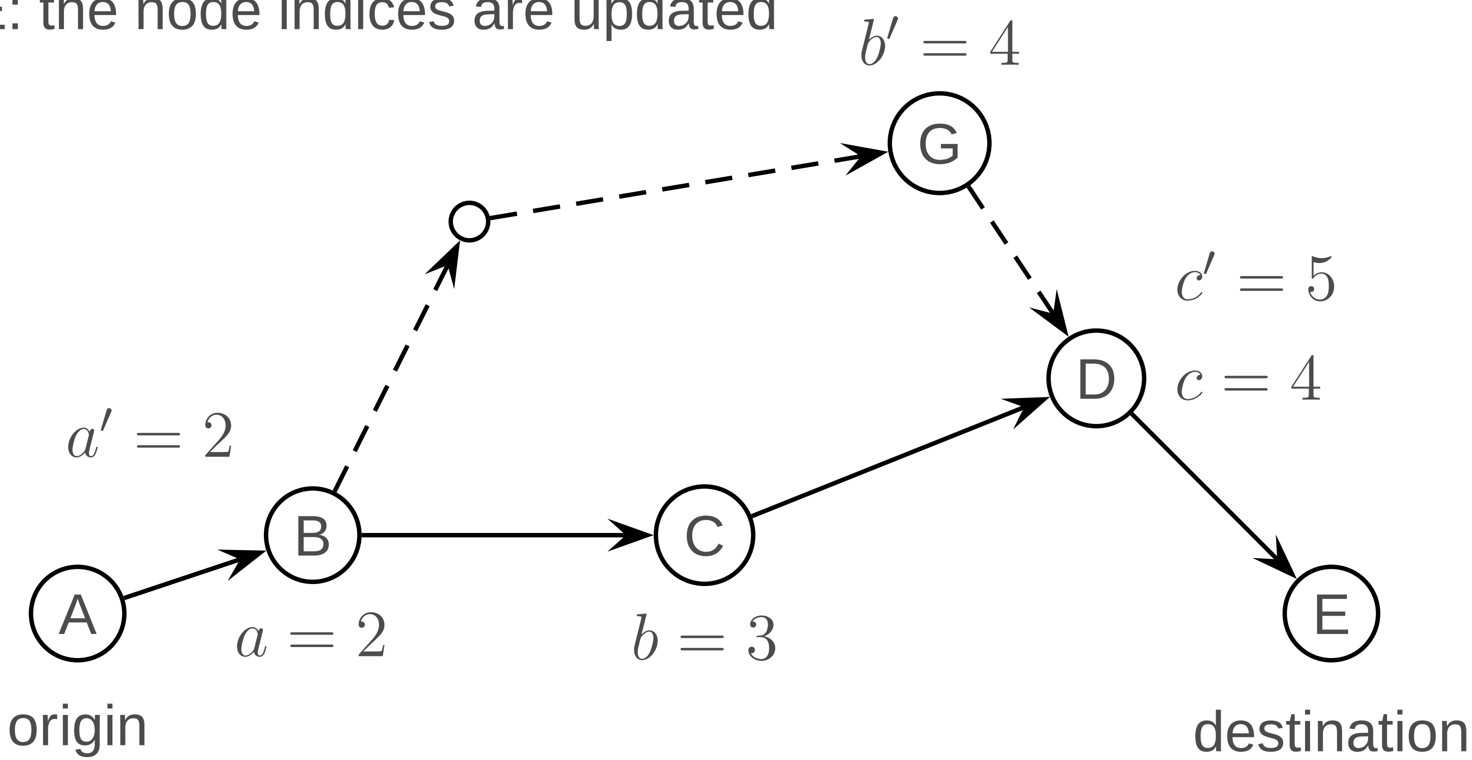
# Proposal transition distribution

SPLICE: $G$ is connected to the path



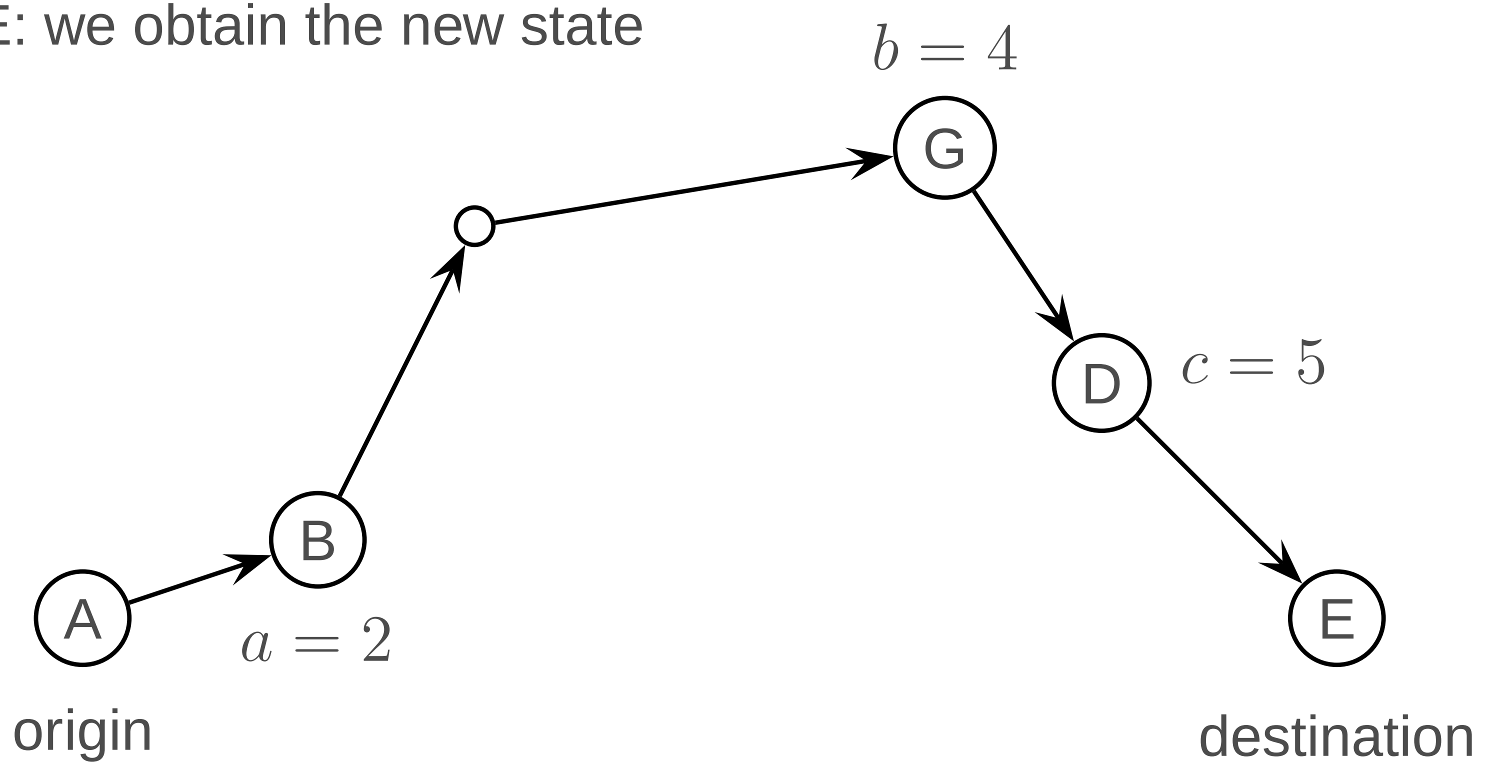
# Proposal transition distribution

SPLICE: the node indices are updated



# Proposal transition distribution

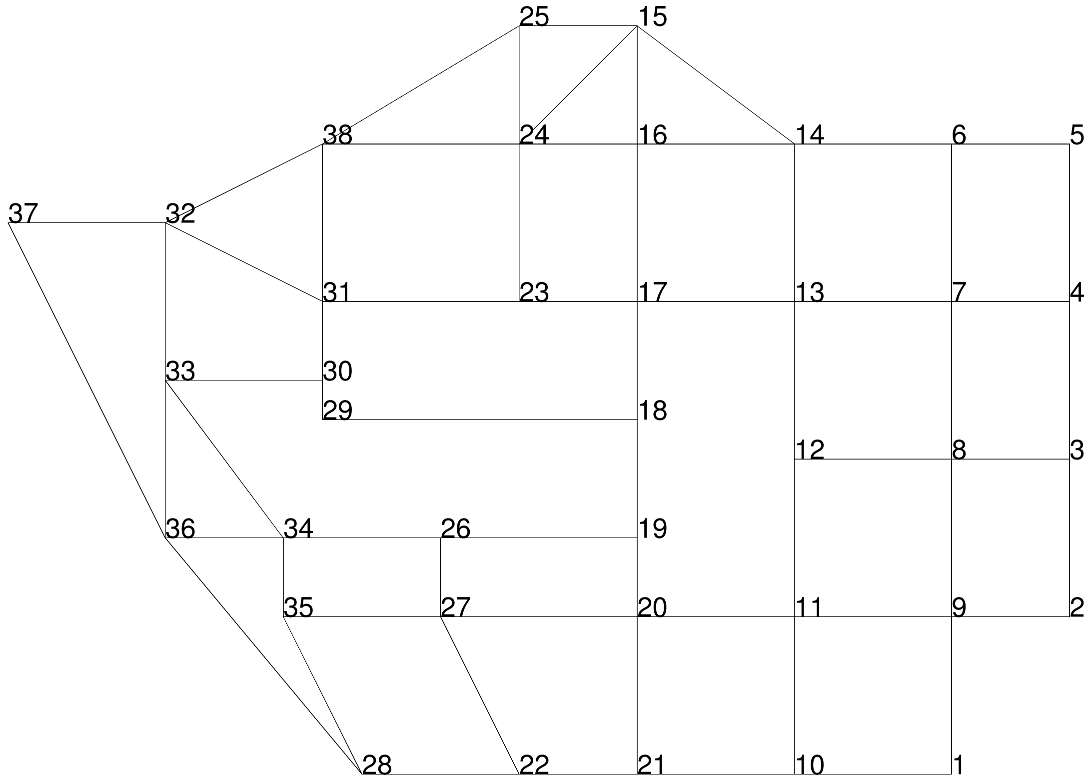
SPLICE: we obtain the new state



# Path generation algorithm

- A great deal of technical difficulties have to be addressed
- Implementation in Java
- Runs fast on real networks

# Simple example



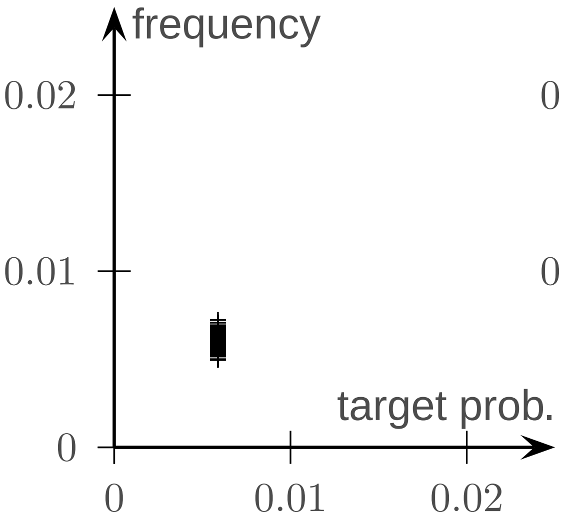
# Simple example

- Target weights:

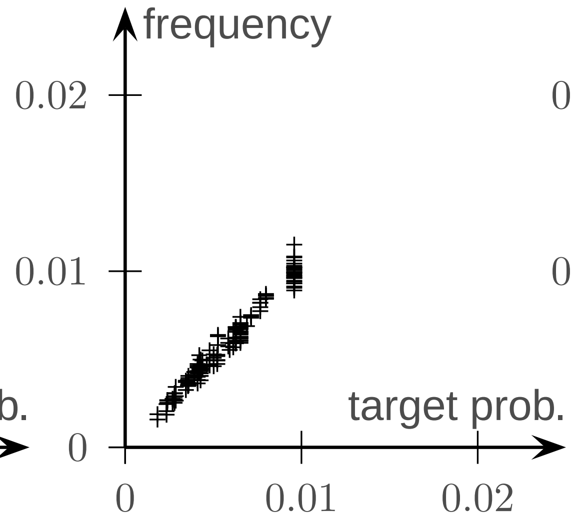
$$b(i) = \exp[-\mu\delta(\Gamma)]$$

  where $\delta(\Gamma)$ is the length of path $\Gamma$.
- Note: $\mu = 0$ means equal probability.

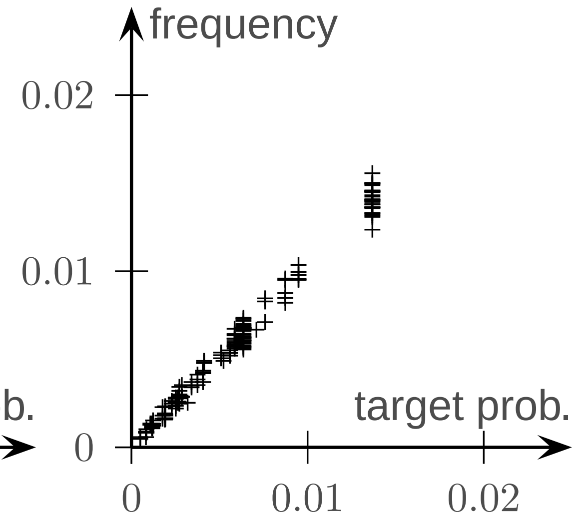
# Scatter plots




(a) $\mu = 0.0$




(b) $\mu = 2.0$



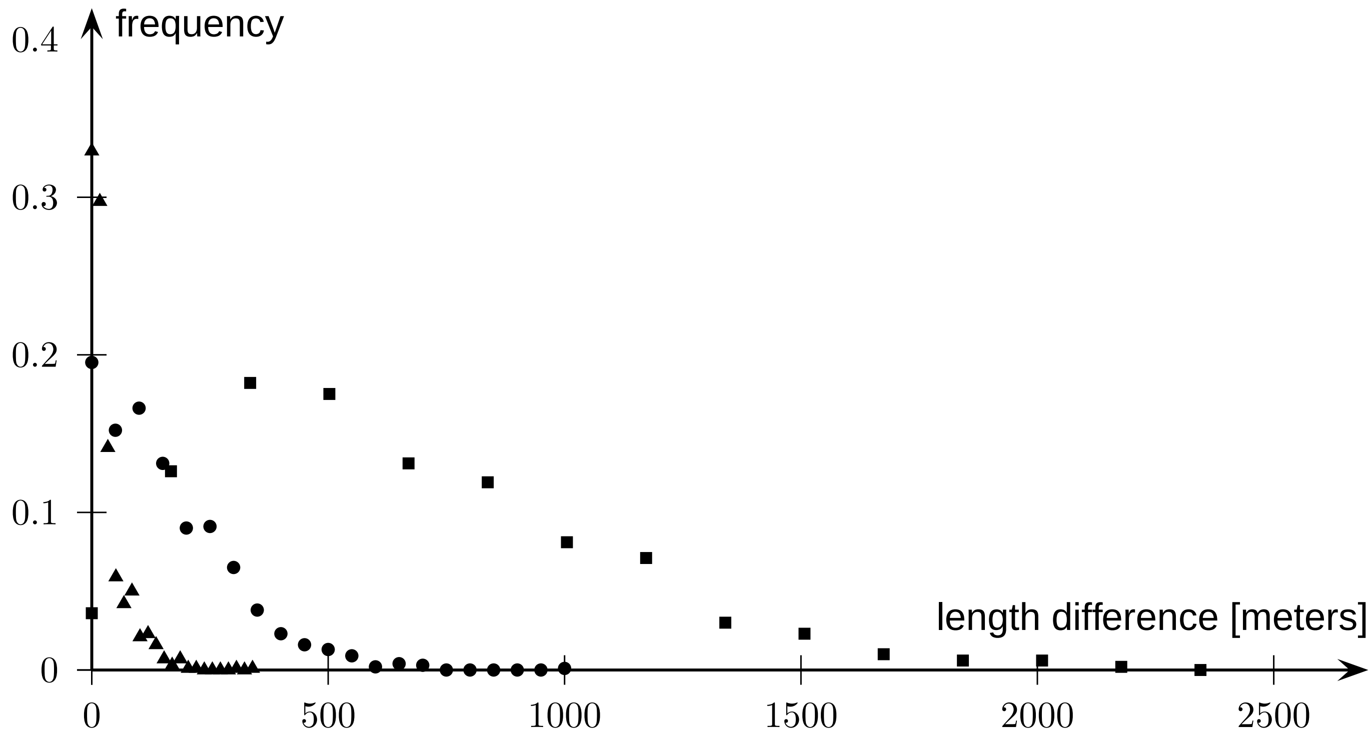

(c) $\mu = 4.0$

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

# Tel-Aviv example

# Tel-Aviv: length distribution




Squares: $\mu = 0.01$, circles: $\mu = 0.02$, triangles: $\mu = 0.04$.

# Conclusion

- Route choice modeling is difficult.
- Data: smartphones
- Identify the chosen route
  - Deal with inaccuracy and low rate
  - Probabilistic map matching
- Identify the non chosen routes
  - Sampling of paths
  - Markov Chain Monte-Carlo method
  - The devil is in the details...
  - but it works!

# References


- Bierlaire, M., and Frejinger, E. (2008). Route choice modeling with network-free data, Transportation Research Part C: Emerging Technologies 16(2):187-198.

  http://dx.doi.org/10.1016/j.trc.2007.07.007

- Bierlaire, M., Chen, J., and Newman, J. P (2010). Modeling Route Choice Behavior From Smartphone GPS data. Technical report TRANSP-OR 101016. Transport and Mobility Laboratory, ENAC, EPFL.

  http://transp-or.epfl.ch/php/abstract.php?type=1&id=BierChenNewm10

# References


- Bierlaire, M., and Flötteröd, G. (2010). Metropolis-Hastings sampling of alternatives for route choice models. Proceedings of the Swiss Transport Research Conference (STRC) September 1-3, 2010.

  http://transp-or.epfl.ch/documents/proceedings/BierFloeSTRC2010.pdf


TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

# Appendix

- Derivation of the measurement equation for the probabilistic map matching.

# Recursion: first step

$$\Pr(\widehat{x}_1|p) = \int_{x_1 \in p} \Pr(\widehat{x}_1|x_1, p) \Pr(x_1|p) dx_1,$$

- integral spans all locations $x_1$ on path $p$
- no prior information on $x_1$

$$\Pr(x_1|p) = 1/L_p$$

- a smarter way would be to assign more probability in the beginning of the path
- measurement error of the device:

$$\Pr(\widehat{x}_1|x_1, p) = \Pr(\widehat{x}_1|x_1)$$

# Measurement error of the device

- Assume that latitudinal and longitudinal errors are i.i.d. normal with variance $\sigma^2$
- Measurement error is Rayleigh
- $\sigma^2$ unknown, estimate:

$$\widehat{\sigma}^2 = \sigma^2_{\text{network}} + (\widehat{\sigma}^x_1)^2$$

where

- $\sigma^2_{\text{network}}$: network coding errors
- $(\widehat{\sigma}^x_1)^2$: GPS errors.

$$\Pr(\widehat{x}_1 | x_1) = \exp\left(-\frac{\|\widehat{x}_1 - x_1\|_2^2}{2\widehat{\sigma}^2}\right).$$

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

# Recursion: first step

$$\Pr(\widehat{x}_1|p) = \frac{1}{L_p}\int_{x_1} \exp\left(-\frac{\|\widehat{x}_1 - x_1\|_2^2}{2\widehat{\sigma}^2}\right) dx_1.$$

- Integral may be cumbersome for long paths
- Can be simplified using the concept of Domain of Data Relevance
- See Bierlaire & Frejinger (2008) and Bierlaire, Chen and Newman (2010)

# Recursion: second step

$$\Pr(\widehat{x}_1, \widehat{x}_2 | p) = \Pr(\widehat{x}_2 | \widehat{x}_1, p) \Pr(\widehat{x}_1 | p),$$

Focus now on

$$\Pr(\widehat{x}_2 | \widehat{x}_1, p) = \int_{x_2 \in p} \Pr(\widehat{x}_2 | x_2, \widehat{x}_1, p) \Pr(x_2 | \widehat{x}_1, p) dx_2.$$

- first term = $\Pr(\widehat{x}_2 | x_2)$ measurement error, same as before
- second term: predicts the position at time $\widehat{t}_2$ of the traveler

$$\Pr(x_2 | \widehat{x}_1, p) = \int_{x_1 \in p} \Pr(x_2 | x_1, \widehat{x}_1, p) \Pr(x_1 | \widehat{x}_1, p) dx_1.$$

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

# Position predictor

$$\Pr(x_2|\widehat{x}_1, p) = \int_{x_1 \in p} \Pr(x_2|x_1, \widehat{x}_1, p) \Pr(x_1|\widehat{x}_1, p) dx_1.$$

- First term: movement model

$$\Pr(x_2|x_1, \widehat{x}_1, p) = f_x(x_2|x_1, \widehat{t}_1, \widehat{t}_2, p),$$

- Second term: Bayes rule

$$\Pr(x_1|\widehat{x}_1, p) = \frac{\Pr(\widehat{x}_1|x_1, p) \Pr(x_1|p)}{\int_{x_1} \Pr(\widehat{x}_1|x_1, p) \Pr(x_1|p) dx_1}.$$

simplifies to

$$\Pr(x_1|\widehat{x}_1, p) = \frac{\Pr(\widehat{x}_1|x_1, p)}{\int_{x_1} \Pr(\widehat{x}_1|x_1, p) dx_1}$$

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

# Measurement equations

- Step $k$ of the recursion based on same principles
- but requires some technical simplifications

$$\Pr(x_{k-1}|\widehat{x}_{k-1},p) = \frac{\Pr(\widehat{x}_{k-1}|x_{k-1},p)}{\int_x \Pr(\widehat{x}_{k-1}|x,p)dx}.$$

- Integrals can be simplified using the DDR