# Route choice modeling based on network-free data

**Michel Bierlaire and Emma Frejinger**

**Transport and Mobility Laboratory, EPFL,** `transp-or.epfl.ch`

# Outline

- Introduction to route choice modeling
  - Modeling framework
  - Estimation
  - Issues
- Choice data
- Modeling based on network-free data
- Case study

# Route choice modeling

*Given a transportation network composed of nodes, links, origin and destinations.*
*For a given transportation mode and origin-destination pair, which is the chosen route?*

# Route choice modeling

- Deterministic approach: Travelers use the shortest (with regard to any arbitrary generalized cost) route among all
  - Behaviorally unrealistic
- Random utility models (discrete choice models)

# Framework

- Utility maximization
- An individual $n$ associates a utility $U_{pn}$ with each path $p$ in his/her choice set $\mathcal{C}_n$ and chooses the alternative with the highest utility

# Random Utility Models

$$U_{pn} = V_{pn} + \varepsilon_{pn}$$

$V_{pn}$: Deterministic part $V_{pn} = \beta^T X_{pn}$

$\varepsilon_{pn}$: Random term

- Uncertainty is introduced with the motivation that the analyst does not have complete information
  - Unobserved attributes
  - Unobserved taste variations
  - Measurement errors

# Estimation

- Likelihood function: the probability that the model reproduces all $N$ observations

$$L(\beta) = \prod_{n=1}^{N} P_n(\beta)$$

$P_n(\beta)$: probability of the observed alternative for individual $n$

$\beta$: vector of $K$ parameters

# Estimation

- Maximum likelihood estimation uses the logarithm of the likelihood function, $\mathcal{L}$

$$\mathcal{L}^*(\hat{\beta}_1, ..., \hat{\beta}_K) = \max_{\beta \in \mathbb{R}} \mathcal{L}(\beta) = \sum_{n=1}^{N} \ln P_n(\beta)$$

- BIOGEME: estimation software
  **Bi**erlaire's **O**ptimization Toolbox for **GE**V **M**odel **E**stimation

# Issues

- Problem characteristics
  - Universal choice set very large
  - Individual specific choice set unknown
  - Correlated alternatives due to overlapping paths
  - Data issues

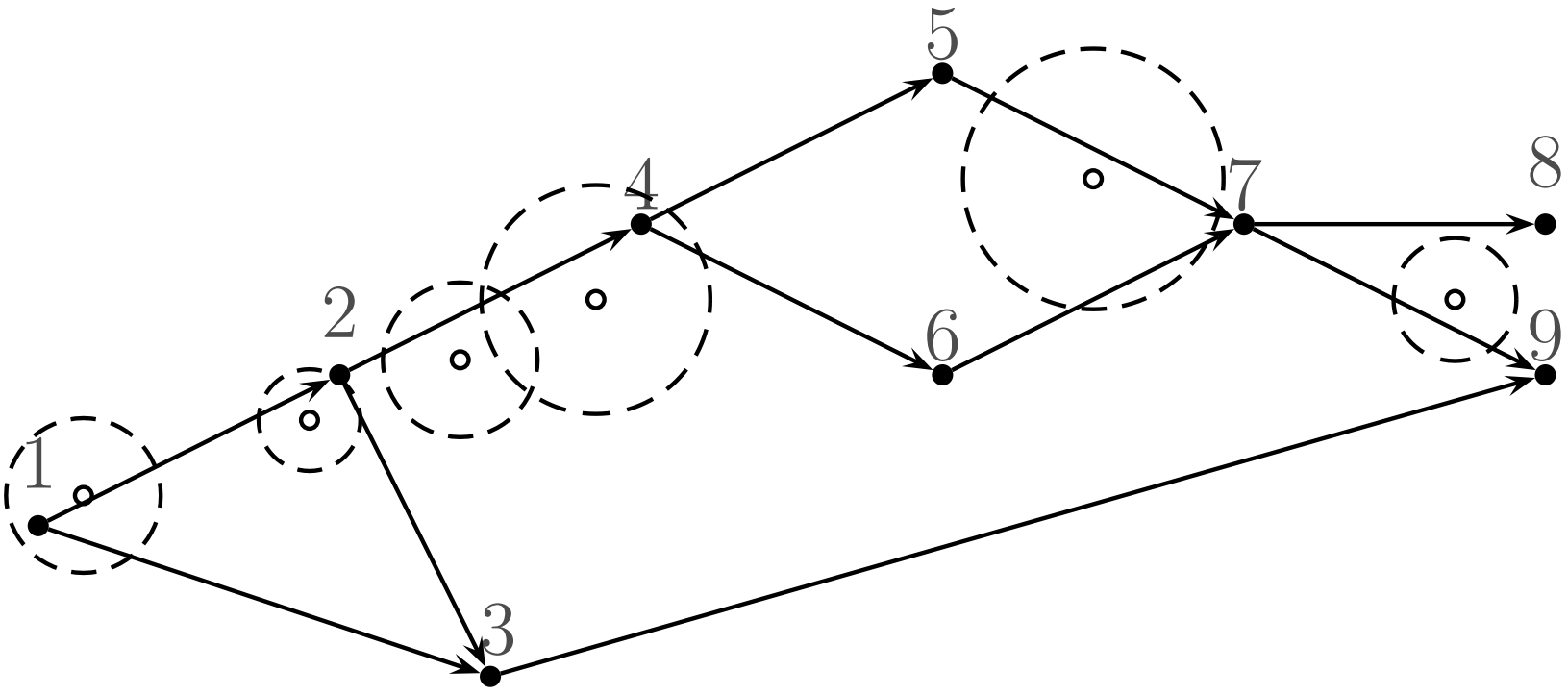
# Choice data

- Link-by-link descriptions of chosen routes are never directly available
- Data processing in order to obtain network compliant paths
  - Assumptions about missing data
  - Link matching of GPS points
- Data manipulation difficult to verify and may introduce bias and errors

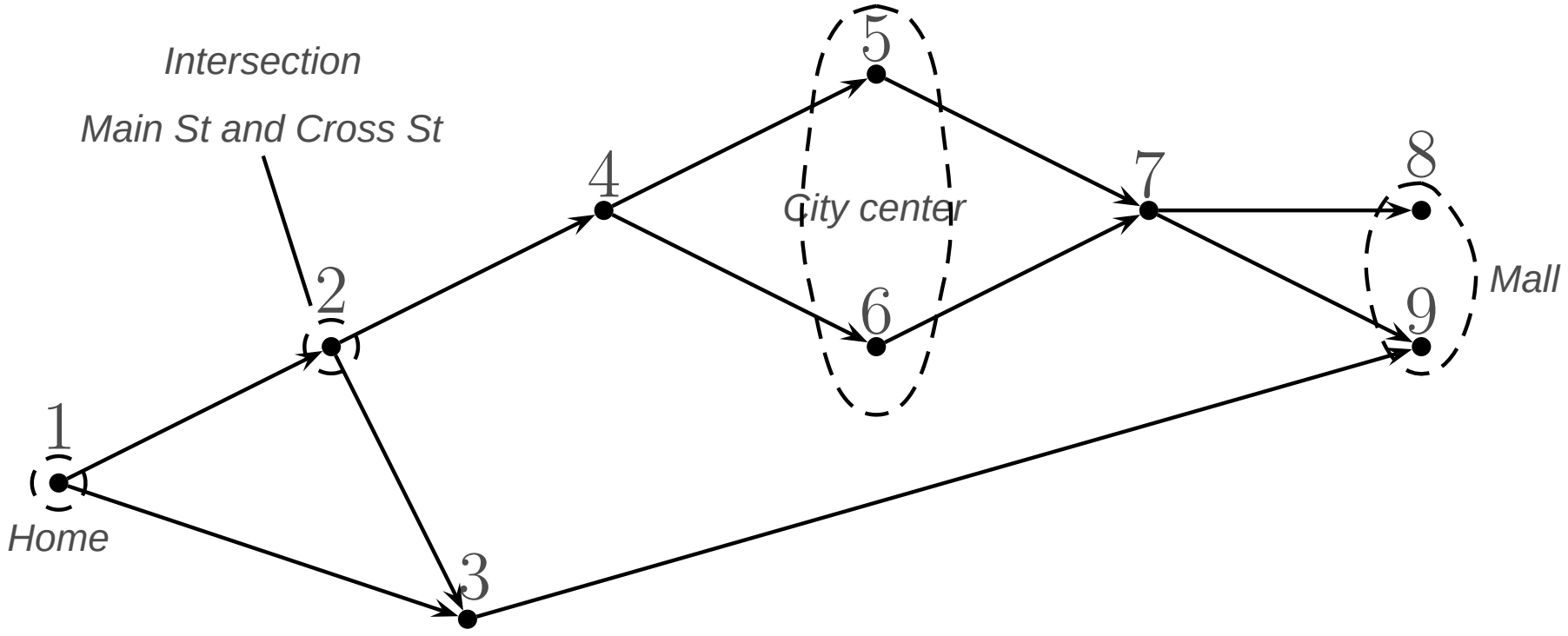
# Modeling with network-free data

- An observation $i$ is a sequence of individual pieces of data related to an itinerary. Examples: sequence of GPS points or reported locations
- For each piece of data we define a Domain of Data Relevance (DDR) that is the physical area where it is relevant
- The DDRs bridge the gap between the network-free data and the network model

# Example - GPS data



# Example - Reported trip



# Model estimation

- We aim at estimating the parameters $\beta$ of route choice model $P(p|\mathcal{C}_n(s);\beta)$
- We have a set $\mathcal{S}_i$ of relevant od pairs
- The probability of reproducing observation $i$ of traveler $n$, given $\mathcal{S}_i$ is defined as

$$P_n(i|\mathcal{S}_i) = \sum_{s \in \mathcal{S}_i} P_n(s|\mathcal{S}_i) \sum_{p \in \mathcal{C}_n(s)} P_n(i|p) P_n(p|\mathcal{C}_n(s);\beta)$$

# Model estimation

- Measurement equation $P_n(i|p)$
  - Reported trips

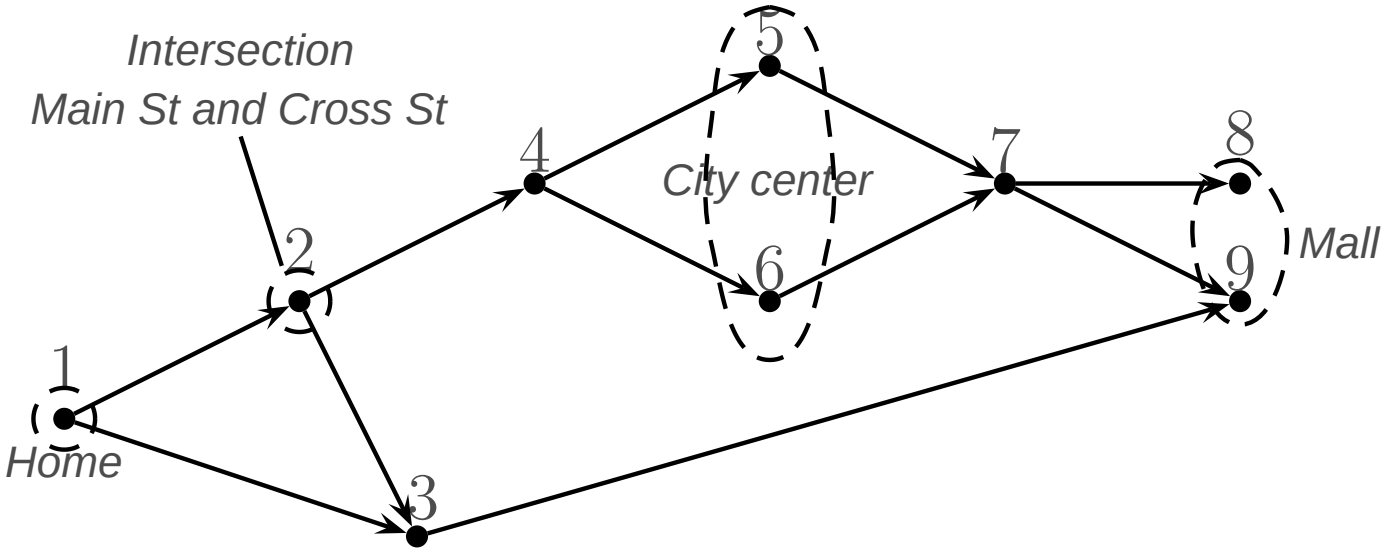
$$P_n(i|p) = \begin{cases} 1 & \text{if } i \text{ corresponds to } p \\ 0 & \text{otherwise} \end{cases}$$

  - GPS data

    $P_n(i|p) = 0$ if $i$ does not correspond to $p$

    If $i$ corresponds to $p$ then $P_n(i|p)$ is a function of the distance between $i$ and $p$

# Model estimation - Example




$$P_n(i|\mathcal{S}_i) = \sum_{s \in \mathcal{S}_i} P_n(s|\mathcal{S}_i) \sum_{p \in \mathcal{C}_n(s)} P_n(i|p) P_n(p|\mathcal{C}_n(s); \beta)$$

$$P(i|\mathcal{S}_i) = \frac{1}{2}\Big[P(p_1|\mathcal{C}(\{1,8\}); \beta) + P(p_2|\mathcal{C}(\{1,8\}); \beta)\Big] + \frac{1}{2}\Big[P(p_3|\mathcal{C}(\{1,9\}); \beta) + P(p_4|\mathcal{C}(\{1,9\}); \beta)\Big]$$

# Case study

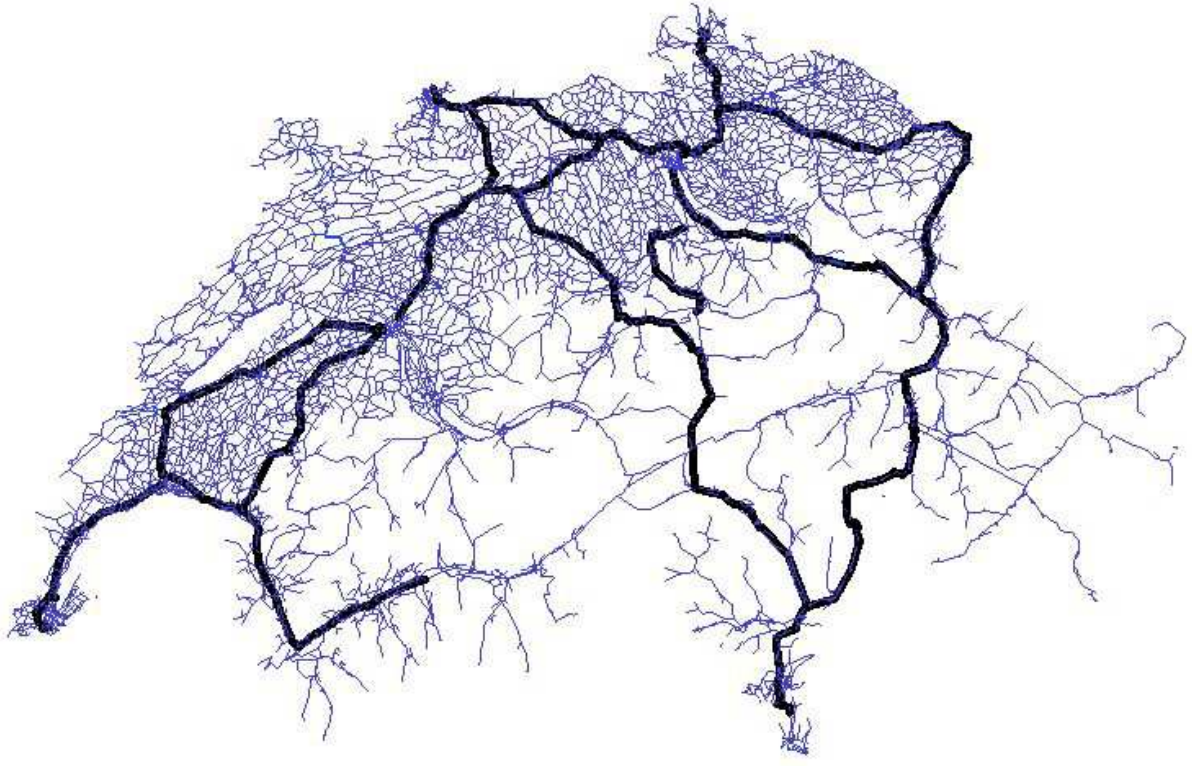
- Data collected in Switzerland for the evaluation of mobility pricing policies
- Long distance route choice
- Observations described by a sequence of locations (sample size: $940$)
- Simplified Swiss network ($39411$ links and $14841$ nodes)
- Estimation of two models: Path Size Logit (Ben-Akiva and Ramming, 1998) and Subnetwork model (Frejinger and Bierlaire, 2007)

# Case study


TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

# Case study



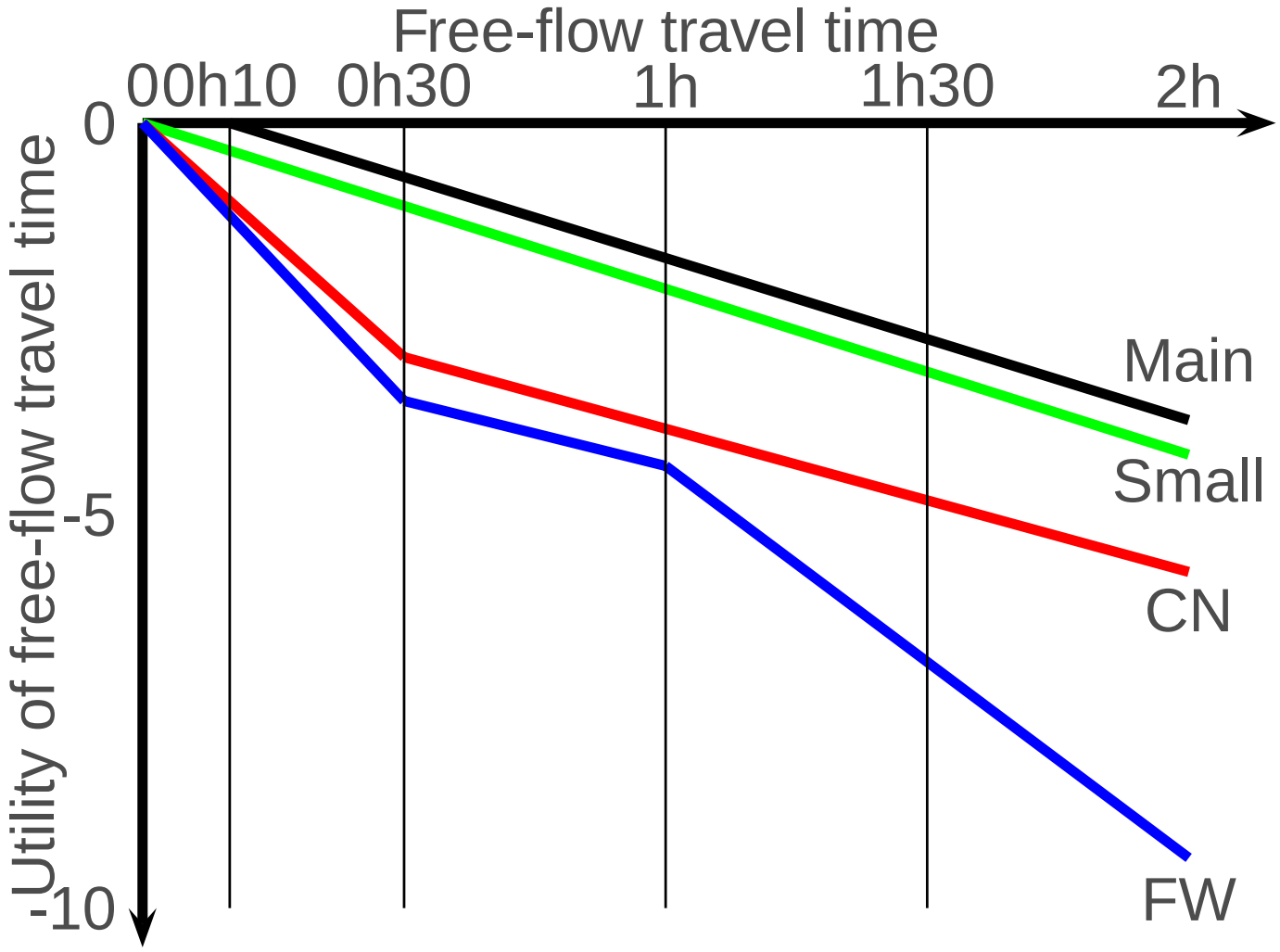
# Case study - Results

| Coefficient | PSL | Subnetwork |
|---|---|---|
| **ln(Path Size) based on free-flow time** | **1.04** | **1.10** |
| *Scaled Estimate* | *1.04* | *1.05* |
| (Rob. Std. Error) Rob. T-test | (0.134) 7.81 | (0.141) 7.78 |
| **Covariance** | | **0.217** |
| *Scaled Estimate* | | *0.205* |
| (Rob. Std. Error) Rob. T-test | | (0.0543) 4.00 |
| Number of simulation draws | - | 1000 |
| Number of parameters | 11 | 12 |
| Final log-likelihood | -1164.850 | -1161.472 |
| Adjusted rho square | 0.145 | 0.147 |
| Null log-likelihood: -1375.851 | | |

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

# Case study - Results



# Conclusion

- Link-by-link descriptions of chosen routes are never directly available
- Modeling framework that reconcile network-free data with a network based model without data manipulation
- Illustration on a case study using revealed preferences data
- Case study with GPS data left for future research