



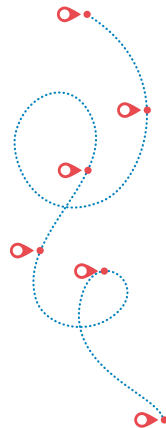
Higher-order Analysis of Human Mobility Data

In collaboration with Chen Zhang and Jürgen Hackl

HONS Satellite @ NetSci '26
August 3, 2026
Boston, MA

Talk Roadmap

1. Background: Human mobility, infrastructure, and higher-order networks
2. Case study: observation and simulation in Île-de-France
3. Discussion: Takeaways, limitations, and future work



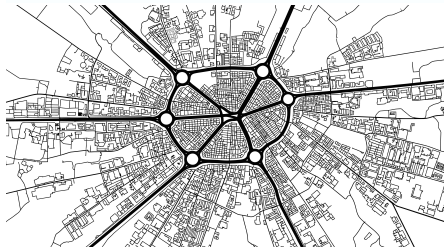
Human Mobility

Fundamental Questions

- ▶ How do people **move through the world**?
- ▶ What factors **explain or influence** mobility behavior?
- ▶ How does human mobility **evolve over time**?
- ▶ Can we make **accurate predictions** to help **guide decision making**?

Application areas

- ▶ planning and design
- ▶ transportation engineering
- ▶ public health and epidemiology
- ▶ point-of-interest recommendation
- ▶ ...and many more



Connecting Mobility and Infrastructure Networks

Why infrastructure?

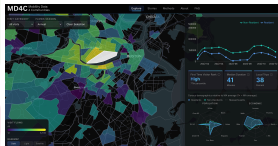
- ▶ Transportation infrastructure influences (and is influenced by) mobility patterns
- ▶ Policymaking requires understanding how people actually use physical infrastructure

Geospatial aggregation breaks the link

- ▶ Mobility studies often aggregate based on geospatial tessellation (grids, census blocks, H3)
- ▶ Simplifies network construction, but **loses connection to infrastructure**

Our work: mobility dynamics + infrastructure networks

Integrate **individual mobility** dynamics and **infrastructure topology** using **higher-order network analysis**



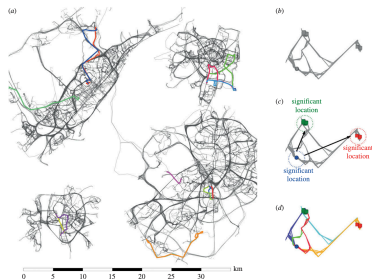
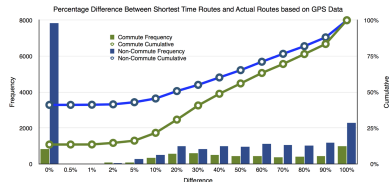
Higher-order Human Mobility

Humans are not (memoryless) random walkers

- ▶ People move with **intentionality** and **memory**
- ▶ Transportation networks impose **constraints**
- ▶ During a trip, individuals **take action** in response to both **internal goals** and the **external environment**

Humans are not shortest-path machines

- ▶ Preferences for efficiency are **contextual**, not based purely on shortest distance or travel time
- ▶ Factors like infrastructure quality and ease of use, local knowledge, urgency, and habit matter



Challenges in Individual Mobility Research

Data Collection

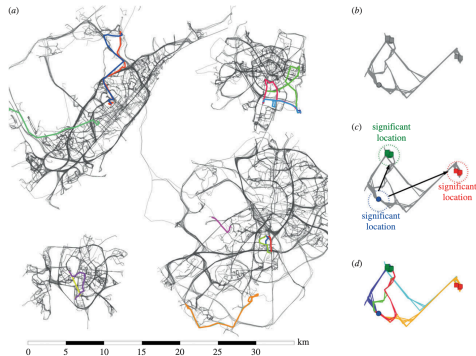
- ▶ **Difficult to collect** good data due to costs, logistics, and privacy considerations
- ▶ Even good data is inherently **large, noisy, and heterogeneous**

Data Integration

- ▶ Requires integrating **geospatial, network, and administrative/economic/environmental** datasets
- ▶ Messy and diverse data needs pre-processing, adding uncertainty

Our Question

Can simulations help overcome these challenges?



Simulating Individual Mobility

Dream: Individual mobility, without individual data

What if we could **simulate high-quality, realistic individual mobility** data by combining collective mobility patterns with agent-based simulations?

Advantages of Simulated Data

- ▶ **Obviates need to collect detailed individual data**, avoiding privacy, logistics, and cost issues
- ▶ **Enables transferability** to regions without GPS data
- ▶ **Uses existing mobility data** collected by governments, transportation firms, and NGOs
- ▶ **Facilitates simulation of artificial interventions** to help guide policymaking

Making the dream a reality

We need methods to **evaluate whether simulated datasets are reliable surrogates** for real human mobility.



Key Features

- Agent-Based Traffic Simulation**
Simulate a large number of individual travelers in your transport model.
- Fast and Dynamic**
MATSim simulates whole days second-by-second for fully dynamic results.
- Active Development**
MATSim continues to evolve and regularly gains new functionality.
- Open Source**
Adapt the code to your needs.
- Private and Public Traffic**
Simulate various transport modes, including car, transit, walk, bike, dirt, ...
- Supports large scenarios**
MATSim can simulate large metropolitan areas in high level of details.
- Versatile Analyses and Simulation Output**
Use preprocessed output or extract your own KPIs from the detailed simulation output.
- Modular Approach**
Extend or replace functionality and algorithms based on your needs.

Higher-order Perspective on Mobility

Our Approach

Compare real-world and simulated mobility paths from a **path-based higher-order perspective**.

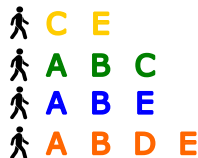
Research Questions

- ▶ Do simulated mobility paths **capture higher-order patterns** observed in real mobility paths?
- ▶ How well does a simulated dataset **align with observed mobility** in the same region?

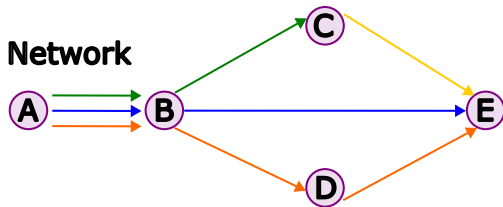
Methods Preview

- ▶ Path length distributions
- ▶ Network coverage
- ▶ Higher-order network properties
- ▶ Next-step prediction

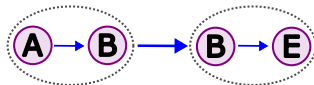
Path Data



Network

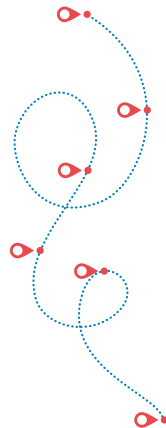


Higher-order Network



Talk Roadmap

1. Background: Human mobility, infrastructure, and higher-order networks
- 2. Case study: observation and simulation in Île-de-France**
3. Discussion: Takeaways, limitations, and future work



Case Study: Île-de-France

The NetMob 2025 Data Challenge is organized with the support of the following institutions and research projects:



NetMob 2025 Data Challenge

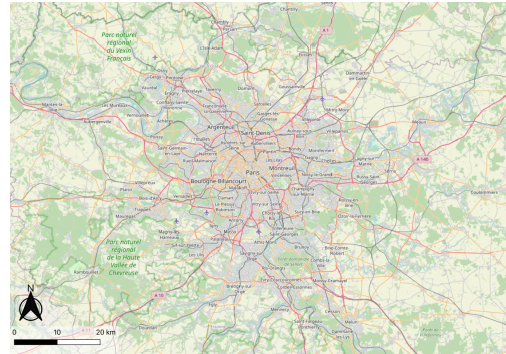
- ▶ Mobility trajectories for 3,320 people (representative sample) living in Île-de-France (Chasse *et al.* 2025)
- ▶ Participants carried dedicated GPS devices for 1 week and logged trip information
- ▶ Each trip is a path through the transportation network

Open Île-de-France MATSim Simulation

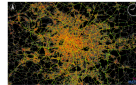
- ▶ Synthetic population of Île-de-France built from open census and regulatory data (Hörl & Balać, 2021)
- ▶ MATSim simulation creates synthetic mobility paths through the transportation network

OpenStreetMaps Network

We use the OpenStreetMaps (OSM) network as the underlying transportation network in both cases.



An open synthetic population of Île-de-France



This repository contains the code to create an open data synthetic population of the Île-de-France region around in Paris and other regions in France.



Preprocessing (in brief)

Network simplification

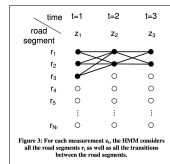
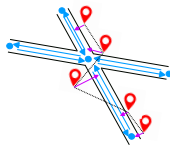
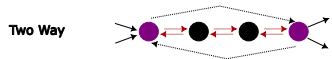
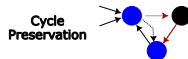
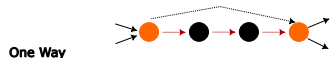
- ▶ Remove **redundant nodes** that do not represent meaningful choice points for travelers

NetMob Map matching

- ▶ Match GPS coordinates to sequences of OpenStreetMaps network edges
- ▶ Implementation: Hidden Markov Model in Valhalla open source routing engine

MATSim Path extraction

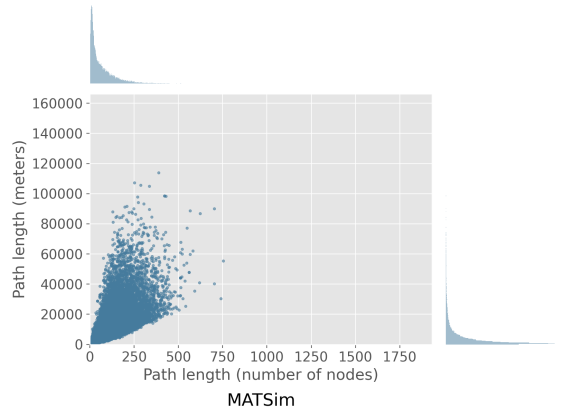
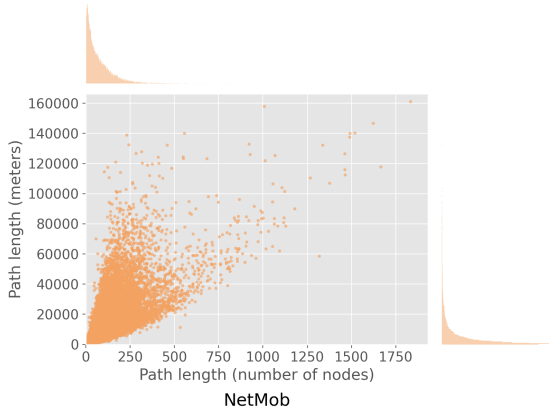
- ▶ Extract sequences of OSM nodes from MATSim output
- ▶ Interpolate walking and biking modes with shortest paths
- ▶ Map transit (bus) paths from stops back to OSM edges



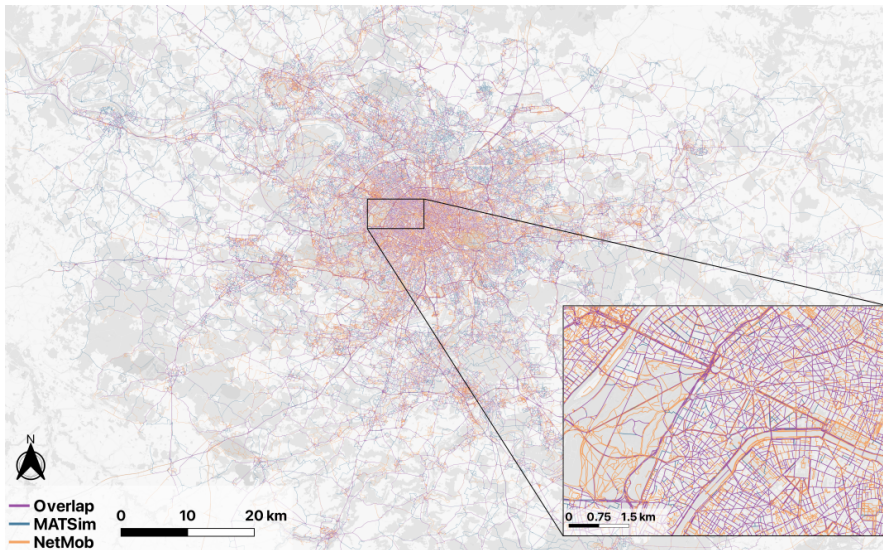
Path Data

Mode	NetMob Mode Name	NetMob	MatSim	Combined Total
Vehicle	PRIV_CAR_DRIVER	20,915		
	PRIV_CAR_PASSENGER	2,952		
	LIGHT_COMM_VEHICLE	44		
	TAXI	195		
	TWO_WHEELER	530		
	Total Vehicle	24,636	22,861	47,253
Bicycle	BIKE	3,693		
	ELECT_BIKE	1,685		
	ELECT_SCOOTER	348		
	Total Bicycle	5,726	1,223	6,902
Walking	WALKING	17,571	18,685	35,260
Public Transportation	BUS	2,728	10,947	13,648
Total	–	50,661	53,716	103,063

Path Length Distributions



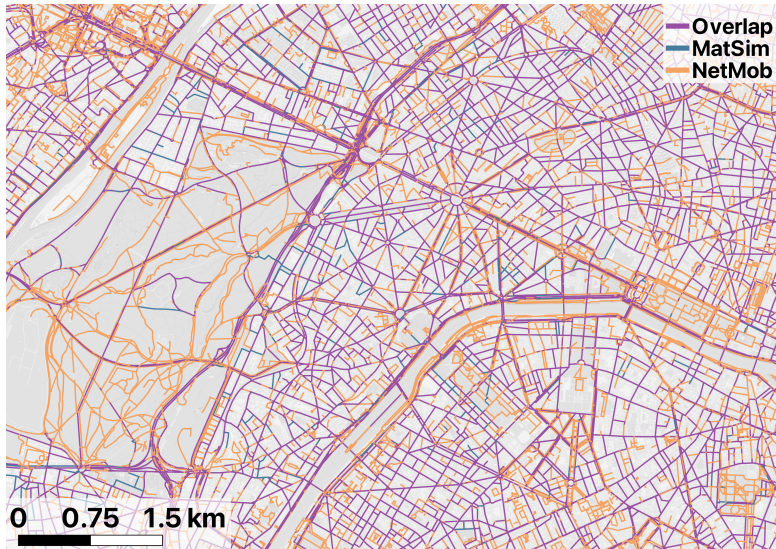
Mapped Data



Mapped Data

Observed paths are more diverse

- ▶ Region between Bois de Boulogne in the west to the Louvre in the east
- ▶ NetMob coverage **much more diverse in pedestrian areas**
- ▶ MATSim provides only origin and destination for walking and cycling, we impute shortest paths



Random Walk Dataset

Purpose

- ▶ Memoryless random walk is the simplest null model for mobility paths
- ▶ Point of comparison that “mirrors” combined statistics of the two path datasets

Random walk design

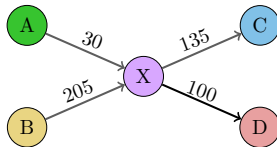
- ▶ **Start node** and **path length** distributions mirror combination of both datasets
- ▶ Compute random walks over the full OSM network (i.e., without simplification)
- ▶ Filter paths in the same way as the original datasets

Discrete time random walk

- ▶ Random walker placed on node u at time $t = t_0$
- ▶ Let η_u denote the out-neighbors of u
- ▶ At $t + 1$, the walker moves to a node $v \in \eta_u$ with the following probability:

$$\Pr(v|u) = \frac{w_{uv}}{\sum_{x \in \eta_u} w_{ux}}$$

where w_{uv} , $v \in \eta_u$ is either 1 (*unweighted* or *uniform*) or the weight of the edge between u and v (*edge weighted*).



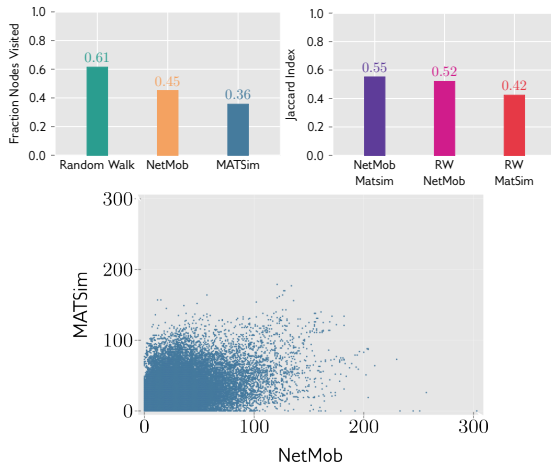
Network Coverage

Node visits

- ▶ What **proportion of nodes** are visited by paths from each dataset?
- ▶ How much do NetMob participant and MATSim agent **visited nodesets overlap**?
- ▶ Are **visit frequencies** aligned?

Takeaways

- ▶ Fewer than **half of nodes in the network visited**
- ▶ Around **half of nodes visited in common**
- ▶ Visitation **frequencies correlate, but with some outliers**



DeBruijn Graphs for HON Analysis

*k*th-order de Bruijn graph

- ▶ Nodes represent paths of length $k - 1$ edges
- ▶ Higher-order nodes connect if they overlap in $k - 1$ first-order nodes
- ▶ Each edge represents a path of length k edges

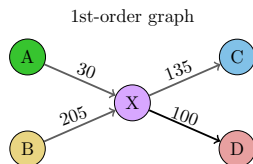
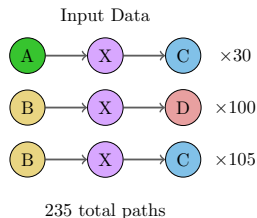
Useful properties

- ▶ Natural representation for incorporation of dynamical memory into network topology
- ▶ Directed, weighted graph that can be analyzed using network science tools

Higher-order transition probabilities

- ▶ Normalized edge weights $\rightarrow$ k th-order random walk

$$\Pr(v_t | v_{t-1}, v_{t-2}, \dots, v_0) = \Pr(v_t | v_{t-1}, \dots, v_{t-k})$$



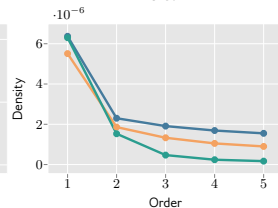
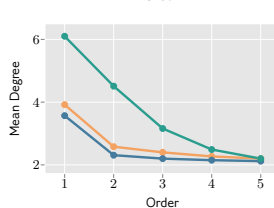
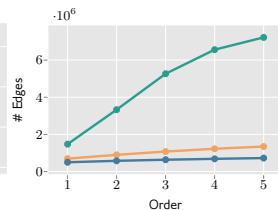
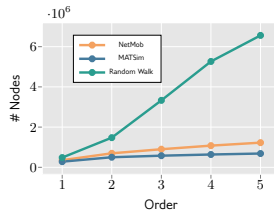
Higher-order Network Analysis

Mobility Network Alignment

- ▶ NetMob and MATSim: nodes and edges increases with order, while average degree and density decrease
- ▶ Random walk visits many more nodes and edges

Optimal Order estimation

- ▶ Can compute likelihood of path dataset based on multi-order transition probabilities
- ▶ Apply likelihood ratio test between adjacent orders to find **optimal order** (high likelihood, balancing parameter blowup)
- ▶ NetMob and MATSim have estimated optimal order $k = 2$, while random walks has $k = 1$ (memoryless)



Next-step prediction

Next-step prediction with backoff

- **Prediction:** Given partial path of length $k - 1$, use softmax of transition probabilities to predict next node:

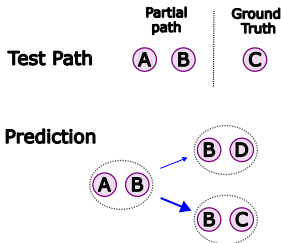
$$P(v_k = u) \propto P(u|v_{k-1}, \dots, v_0)$$

- **Backoff:** If subpath has not been observed in training, try decreasing orders:

$$P(v_k = u) \propto P(u|v_{k-1}, \dots, v_i) \text{ for } i \in 1, \dots, k - 1$$

where $i = k = 1$ is a memoryless random walk step:

$$P(v_k = u) \propto P(u|v_{k-1})$$



Experimental Setup

- Split paths into training and testing sets (stratified by log of path length, 10 splits)
- Use training split to construct k th-order network
- Split test data into sub-paths and use $k - 1$ symbols as input to predict k th step
- Measure average prediction accuracy

Next-step prediction results

Accuracy jump at optimal order

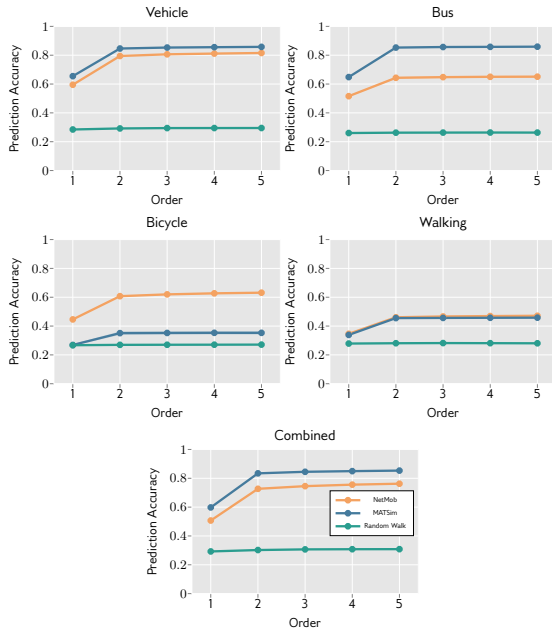
- ▶ Both datasets show jump in accuracy at optimal order $k = 2$, then only marginal benefit
- ▶ Prediction accuracy essentially the same across orders for random walk dataset

MATSim vehicle paths more predictable

- ▶ Next-step prediction accuracy higher for both vehicle and bus modes
- ▶ Especially strong for bus, where MATSim is based on true bus routes without noise

Pedestrian paths poorly predicted

- ▶ Prediction accuracy lower for both bicycle and walking
- ▶ Especially low for MATSim bicycle (interpolated shortest paths)



Next-step entropy

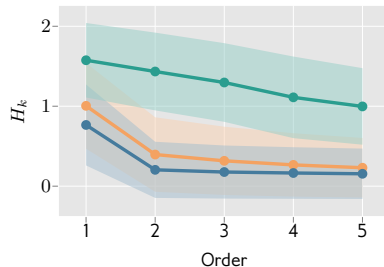
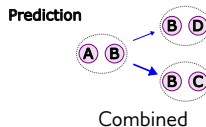
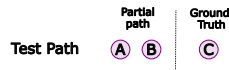
Next-step entropy

- ▶ Average entropy of transition probability distributions from higher-order nodes:

$$\langle H_k \rangle = \frac{1}{N_k} \sum_{s \in \mathcal{M}_k} n_s H(\hat{p}_s).$$

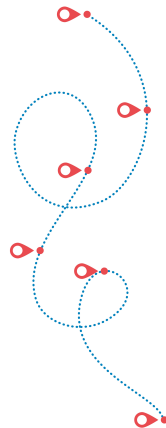
Higher entropy in real paths

- ▶ NetMob entropy higher for all modes
- ▶ Reinforces that observed trajectories are less predictable than simulated



Talk Roadmap

1. Background: Human mobility, infrastructure, and higher-order networks
2. Case study: observation and simulation in Île-de-France
3. Discussion: Takeaways, limitations, and future work



Takeaways

Mobility analysis with higher-order networks

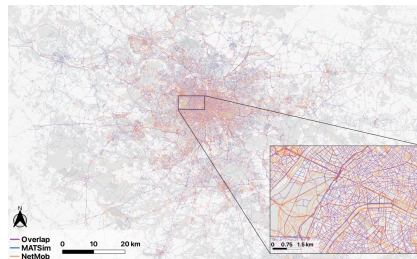
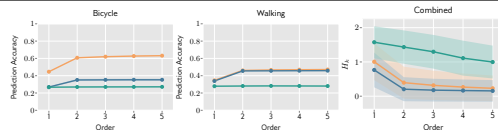
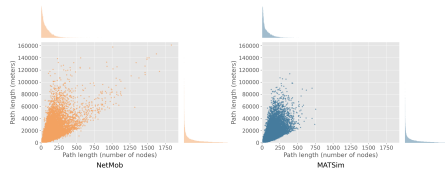
Higher-order network analysis gives us insight into the relationship between observed and simulated mobility.

Good News

- ▶ Open Île-de-France MATSim simulations and NetMob 2025 Data Challenge datasets align well overall
- ▶ Similar higher-order network structural features and overall node visitation rates

Less Good News

- ▶ MATSim does not generate paths at the extreme end of the NetMob distribution
- ▶ Node visitation patterns do not align directly
- ▶ MATSim paths are more predictable and have lower expected conditional entropy than NetMob paths
- ▶ Uncertainty with pedestrian and cycling paths



Limitations & Future Work

Map matching quality

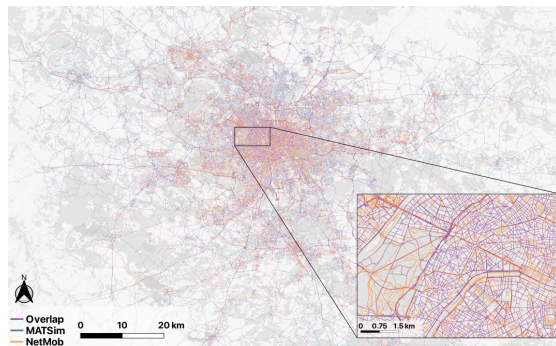
- ▶ Without ground truth, we need methods to validate map matching on new datasets
- ▶ Multi-modal trip segmentation (pre-processing step by data provider)

Further Data Sources

- ▶ NetMob Data Challenge data contains more information (trip purpose, sociodemographics)
- ▶ Simulations beyond MATSim (SUMO, existing Digital Twin models) and augmented pedestrian simulations
- ▶ Comparison with geospatial aggregation and point-of-interest analysis methods

Additional Higher-order Methods

- ▶ Statistical anomaly detection (HYPA)
- ▶ Sequential motifs
- ▶ Direct comparison of higher-order models



Thank You!

Contact

Dr. Timothy LaRock
tlarock@princeton.edu
<https://tlarock.github.io>

Preprint coming very soon...

Preprint in collaboration with Chen Zhang and
Jürgen Hackl will post to arxiv tonight!



OpenStreetMap

The NetMob 2025 Data Challenge is organized with the support of the following institutions and research projects:



Random Walks on Graphs

Discrete time random walk

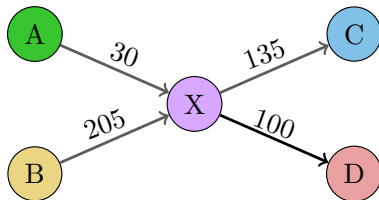
- ▶ Random walker placed on node u at time $t = t_0$
- ▶ Let η_u denote the out-neighbors of u
- ▶ At $t + 1$, the walker moves to a node $v \in \eta_u$ with the following probability:

$$\Pr(v|u) = \frac{w_{uv}}{\sum_{x \in \eta_u} w_{ux}}$$

where w_{uv} , $v \in \eta_u$ is either 1 (*unweighted* or *uniform*) or the weight of the edge between u and v (*edge weighted*).

Note

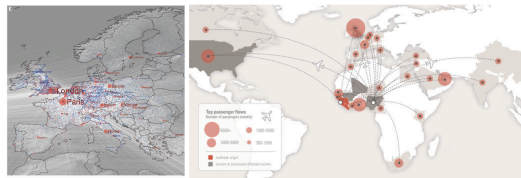
- ▶ This process is called **memoryless** (also **Markovian**) because the next node to be visited depends only on the current node



From Individual to Collective Mobility

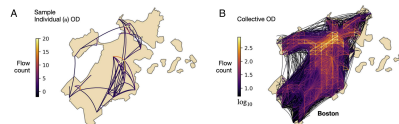
Individual Mobility

- ▶ **Individual trajectories** through the built environment
- ▶ Derived from things like: trip logs, historical records, and GPS trajectory data
- ▶ Examples: GPS location records (mobile phones, exercise apps, dedicated devices), vehicle GPS (personal, taxi, rideshare), etc.



Collective Mobility

- ▶ **Aggregate patterns** across spatial and temporal scales
- ▶ Origin-destination flows derived from things like: large-scale surveys and **aggregations of individual mobility datasets**
- ▶ Examples: local and global migration, commuting flow patterns, census transportation data, etc.



Dataset Processing: Network Simplification

Network simplification

- ▶ Remove “redundant” nodes that do not represent meaningful choice points for travelers and artificially reduce navigational complexity
- ▶ Removal also reduces downstream computation

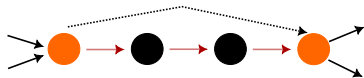
Redundant nodes

- ▶ One-way roads: Nodes with degree exactly 2 and different in/out neighborhoods
- ▶ Two-way roads: Nodes with degree exactly 4 and equivalent in/out neighborhoods

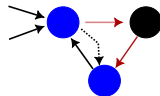
Consequences

- ▶ Nodes: Reduced from 3,016,582 to 788,018 (26%)
- ▶ Edges: Reduced from 6,655,074 to 2,197,946 (33%)
- ▶ Cycles and overall navigability preserved

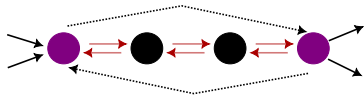
One Way



Cycle Preservation



Two Way



Dataset Processing: NetMob

Map matching

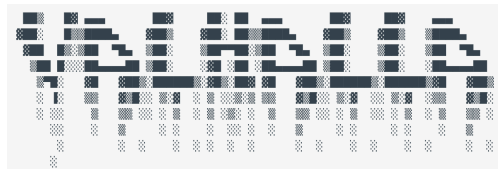
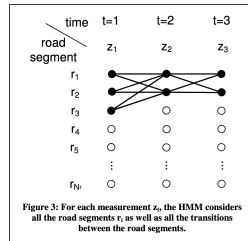
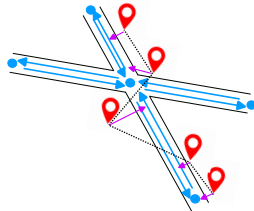
- ▶ Match sequences of GPS coordinates to OpenStreetMaps network edges
- ▶ Matched sequences of unique edges are paths

Implementation

- ▶ Hidden Markov Model approach
- ▶ Valhalla Open Source Routing Engine implementation

Post processing

- ▶ Discontinuities interpolated with shortest paths
- ▶ Redundant nodes removed from paths



Dataset Processing: MATSim

Path extraction

- ▶ Extract sequences of OSM nodes from MATSim XML output events
- ▶ Interpolate walking and biking modes (MATSim only provides origin/destination)
- ▶ Map transit (bus) paths from stops (GTFS data) back to OSM edges

Optimal Order Estimation

Multi-order model likelihood

Given a set of N observed paths $\mathcal{P} = \{p_j\}_{j=1}^N$, we calculate the likelihood of the data under a k th-order model as

$$\mathcal{L}(\bar{M}_k \mid \mathcal{P}) = \prod_{j=1}^N \bar{P}^{(k)}(p_j),$$

where $\bar{P}^{(k)}(p_j)$ is the probability of path p_j under the multi-order model $\bar{M}_k$ with maximum order k .