Dr. Tim LaRock In collaboration with Chen Zhang and Prof. Jürgen Hackl (arXiv preprint 2606.00733)

PRINCETON UNIVERSITY



Do simulated agents move like real people?

A path-based approach to comparing observed and simulated mobility

VISIM Seminar
November 6th, 2026
Online

Connecting Human Mobility and Infrastructure Networks

Human Mobility Research

- Mobility looks for patterns in how people move, using surveys, cellular records, and **GPS trajectories**
- Analyses often aggregate GPS to **geospatial tesselations** (grids, census blocks, H3) or **points of interest** (visits to shops, restaurants, etc.)
- Simplifies network construction, but **loses connection to underlying infrastructure networks**

Infrastructure networks facilitate mobility

- Transportation infrastructure (roads, sidewalks, transit, etc.) influences mobility patterns, and *vice versa*
- Policymaking requires understanding how people actually use physical infrastructure

Overarching vision: mobility + infrastructure networks

Paths of individual mobility through infrastructure topology



Map data © OpenStreetMap contributors (ODbL), https://www.openstreetmap.org/copyright. Hexagons: H3, https://h3geo.org/. GPS trace is synthetic.

Individual Mobility Data

Perennial Challenge

- Collecting individual mobility data (*i.e.*, GPS traces) is often **difficult**, **expensive**, and even **risky**
- Dream: What if we could **simulate high-quality, realistic individual mobility**?

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.

https://matsim.org/

Path-based Perspective

Flows vs. paths

- **Flows** (*e.g.*, link counts) give limited behavioral insight, since we must infer how segments combine into trips
- **Detailed paths** through the underlying network show how individuals navigate from origin to destination

Today's Question

How well does a simulated dataset **align with observed mobility** in the same region from a **path-based** perspective?

Case Study: Île-de-France

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

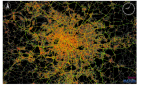
OpenStreetMap Network

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

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



An open synthetic population of Île-de-France

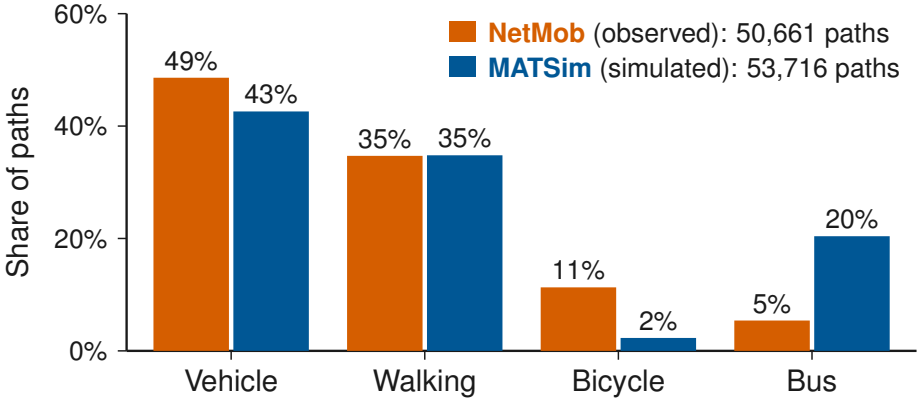


This repository contains the code to create an open data synthetic population of the Île-de-France region around in Paris and other regions in France.

MATSim
Multi-Agent Transport Simulation

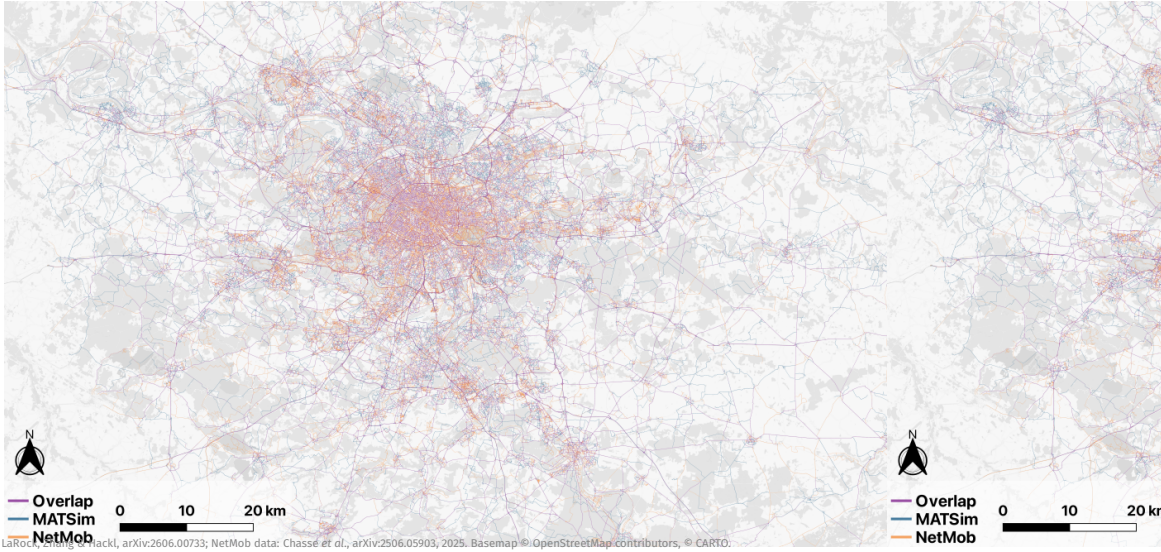
Map: OSM Contributors; Chasse *et al.*, arXiv:2506.05903, 2025; Hörl & Balać, Transportation Research Part C, 2021; https://github.com/eqasim-org/eqasim-france; https://netmob.org/www25/datachallenge

Path Data



Path counts: LaRock, Zhang & Hackl, arXiv:2606.00733 (Table 1). Vehicle: car (driver or passenger), taxi, light commercial vehicle, two-wheeler. Bicycle: bike, e-bike, e-scooter.

Mapped Data



Map: LaRock, Shang & Hackl, arXiv:2606.00733; NetMob data: Chasse *et al.*, arXiv:2506.05903, 2025. Basemap © OpenStreetMap contributors, © CARTO.

Mapped Data

Observed paths are more diverse

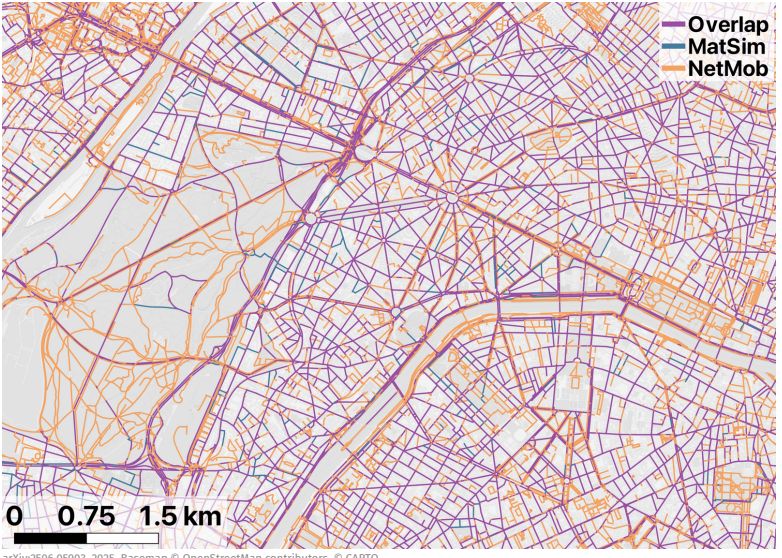
NetMob coverage **much more diverse in pedestrian areas**

Why?

MATSim provides only origin and destination for walking and cycling, we impute shortest paths

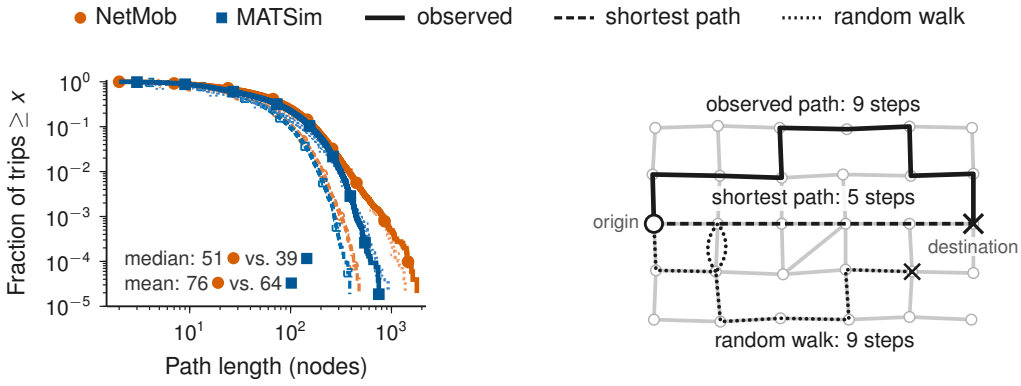
Takeaway

Need to integrate **better models of pedestrian mobility**



Map: LaRock, Zhang & Hackl, arXiv:2606.00733; NetMob data: Chasse *et al.*, arXiv:2506.05903, 2025. Basemap © OpenStreetMap contributors, © CARTO.

Path Length Distributions



Simulated paths are shorter

Use fewer nodes (93 NetMob paths are longer than any MATSim path) and travel shorter distances (not shown)

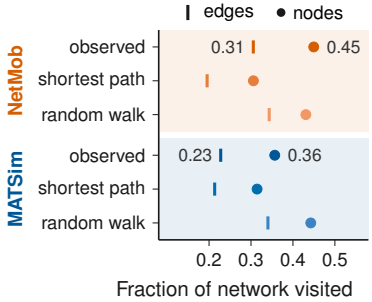
Baseline Comparison

For each observed/simulated path, replace with:

- shortest path between same origin/destination
- random walk with same origin and number of steps

Left: LaRock, Zhang & Hackl, arXiv:2606.00733.

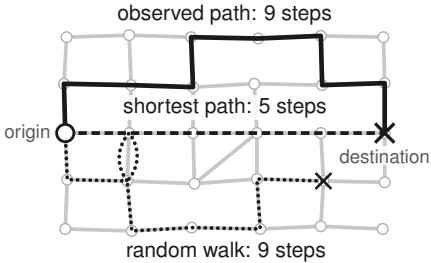
Network Coverage



Node visits

- Fewer than **half of nodes in the network visited**
- NetMob covers **more of the network** than MATSim: 45% of nodes and 31% of edges, against 36% and 23%
- Random walks cover fewer nodes than NetMob, but more than MATSim

How well do paths cover nodes?

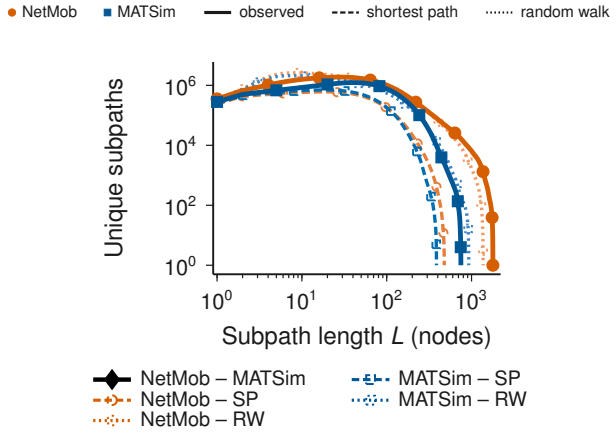


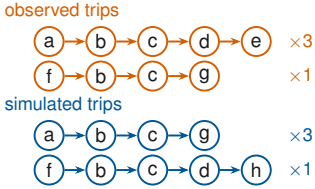
Shared and exclusive use

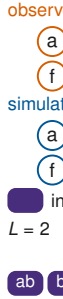
- **Only 55% of visited nodes (41% of edges) are used by both datasets**
- **31% of nodes are used only by NetMob, 13% only by MATSim**

Figures: LaRock, Zhang & Hackl, arXiv:2606.00733.

Beyond Nodes: Subpath Alignment











Left: LaRock, Zhang & Hanna, arXiv:2606.00733. SP: shortest path, RW: random walk. Right: toy example.

Next-step prediction

Gote *et al.*, *Applied Network Science*, 2023.

Next-step prediction results

Accuracy jump at optimal order

Both datasets show jump in accuracy at optimal order $k = 2$, then only marginal benefit

Simulated paths are more predictable

Across orders, MATSim paths are roughly 10% more predictable than NetMob

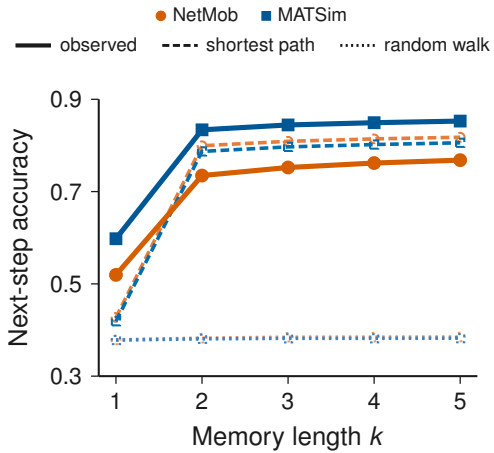


Figure: LaRock, Zhang & Hackl, arXiv:2606.00733.

Next-step prediction by mode

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/cyclist paths poorly predicted

- Both show low prediction accuracy lower for both walking
- Especially low for MATSim bicycle, while NetMob bicycle is actually more predictable
- MATSim paths for walking and cycling are shortest paths

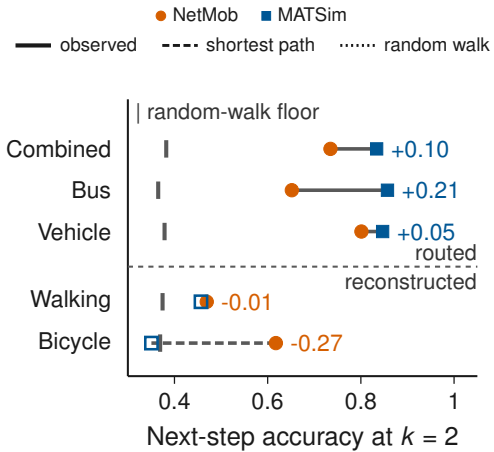


Figure: LaRock, Zhang & Hackl, arXiv:2606.00733.

Thank You!

Takeaways

Broad alignment between observed and simulated mobility, but some key differences:

- Observed mobility behavior is more diverse
- Simulations do not reproduce the tail of path lengths
- Simulated paths are more predictable
- Pedestrian mobility still open (simulation models, deeper understanding of observed data)

Future work

- Validating map matching without ground truth
- Further data: trip purpose and sociodemographics, simulations beyond MATSim
- Integration of better pedestrian mobility models
- Additional higher-order methods: anomaly detection (HYPA), sequential motifs

Contact

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

Do Simulated Agents Move Like People?

Timothy LaRock, Chen Zhang, Jürgen Hackl

https://arxiv.org/abs/2606.00733

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











Takeaways

Mobility analysis with higher-order networks

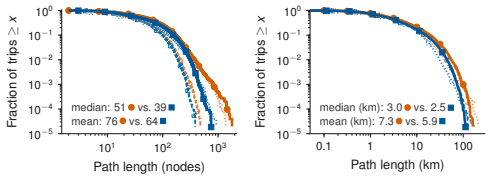
Higher-order network analysis gives us insight into the relationship between observed and simulated mobility.

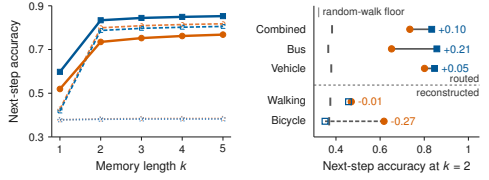
Good News

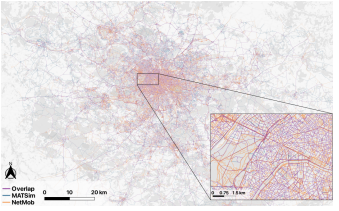
- Open Île-de-France MATSim simulations and NetMob 2025 Data Challenge datasets align well overall
- Similar higher-order network structural features and overall node visitation rates

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







Figures and map: LaRock, Zhang & Hackl, arXiv:2606.00733; NetMob data: Chasse *et al.*, arXiv:2506.05903, 2025. Basemap © OpenStreetMap contributors, © CARTO.

Limitations & Future Work

Map matching quality

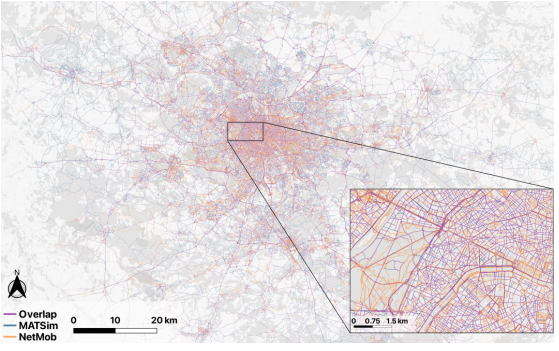
Without ground truth, we need methods to validate map matching on new datasets

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



Map: LaRock, Zhang & Hackl, arXiv:2606.00733; NetMob data: Chasse *et al.*, arXiv:2506.05903, 2025. Basemap © OpenStreetMap contributors, © CARTO.

Human Mobility

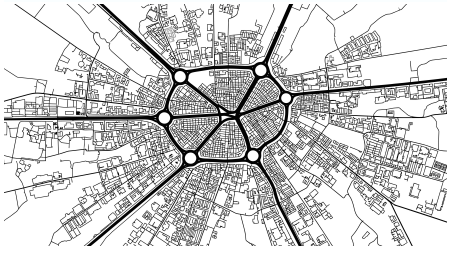
Fundamental Questions

- 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





Images: Adobe Stock

Challenges in Individual Mobility Research

Data Collection

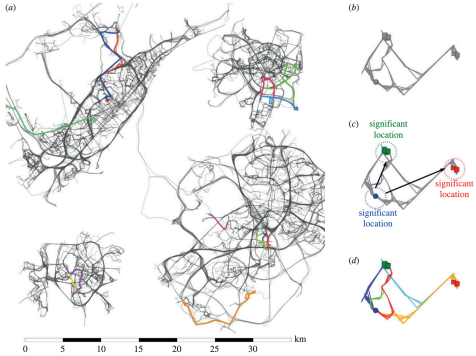
- **Difficult to collect** good data due to costs, logistics, and privacy considerations
- Even good data is inherently **large**, **noisy**, and **heterogeneous**

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?



Lima *et al.*, Royal Society Interface, 2016.

Random Walks on Graphs

Discrete time random walk

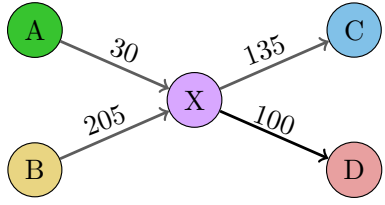
- Random walker placed on node u at time $t = t_0$
- Let η_u denote the out-neighbors of u
- At $t + 1$, the walker moves to a node $v \in \eta_u$ with the following probability:

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

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



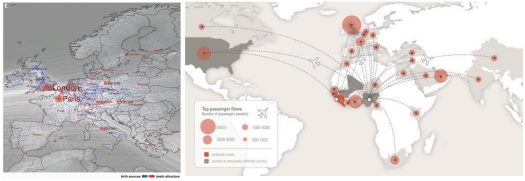
From Individual to Collective Mobility

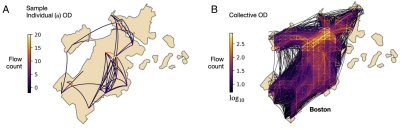
Individual Mobility

- **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.





Schich *et al.*, Science, 2014; Pastore-Piontti *et al.*, Mathematical and Statistical Modeling for Emerging and Re-emerging Infectious Diseases, 2016; Bontorin *et al.*, PNAS, 2025

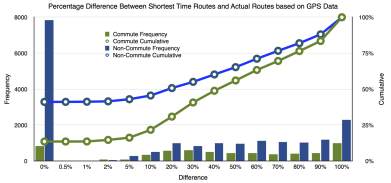
Higher-order Human Mobility

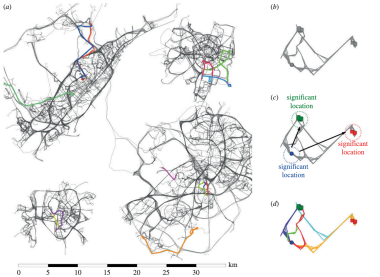
Humans are not (memoryless) random walkers

- 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





Zhu & Levinson, PLOS One, 2014; Lima *et al.*, Royal Society Interface, 2016.

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	47,497
Bicycle	BIKE	3,693		
	ELECT_BIKE	1,685		
	ELECT_SCOOTER	348		
	Total Bicycle	5,726	1,223	6,949
Walking	WALKING	17,571	18,685	36,256
Public Transportation	BUS	2,728	10,947	13,675
Total	–	50,661	53,716	104,377

Preprocessing (in brief)

Network simplification

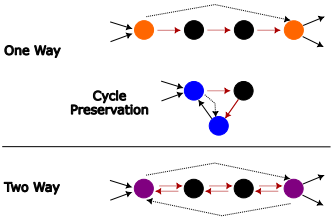
- Remove **redundant nodes** that do not represent meaningful choice points for travelers

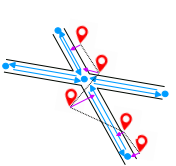
NetMob Map matching

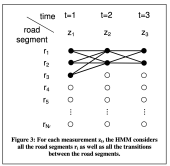
- Match GPS coordinates to sequences of OpenStreetMap network edges
- Implementation: Hidden Markov Model in Valhalla open source routing engine

MATSim Path extraction

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







Newsom & Krumm, ACM SIGSPATIAL, 2009; https://valhalla.github.io/valhalla/meili/implementation/details/

Dataset Processing: Network Simplification

Network simplification

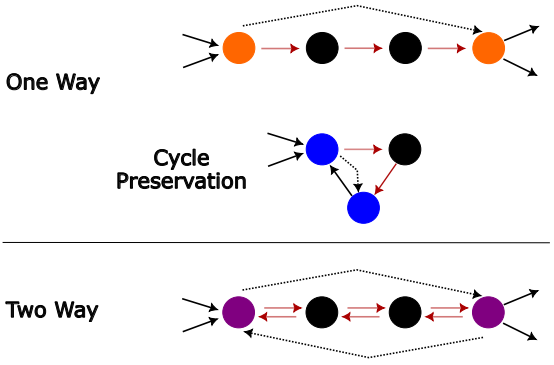
- Remove “redundant” nodes that do not represent meaningful choice points for travelers and artificially reduce navigational complexity
- Removal also reduces downstream computation

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



Dataset Processing: NetMob

Map matching

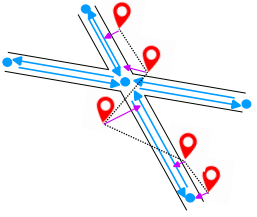
- Match sequences of GPS coordinates to OpenStreetMaps network edges
- Matched sequences of unique edges are paths

Implementation

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

Post processing

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



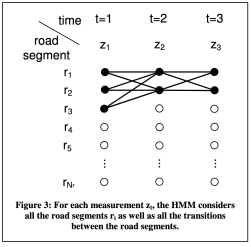
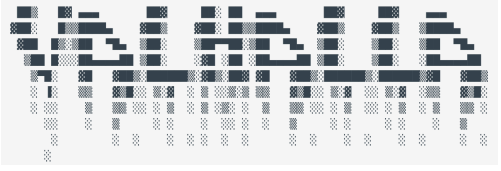


Figure 3: For each measurement z_t, the HMM considers all the road segments r_i as well as all the transitions between the road segments.



Newsom & Krumm, ACM SIGSPATIAL, 2009; https://valhalla.github.io/valhalla/

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

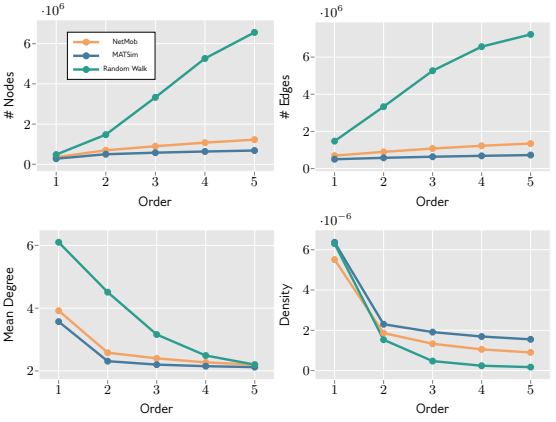
Higher-order Network Analysis

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)



Scholtes, ACM SIGKDD, 2017.

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

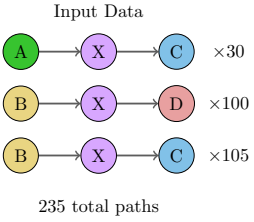
Useful properties

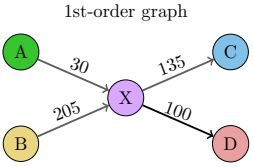
- Natural representation for incorporation of dynamical memory into network topology
- Directed, weighted graph that can be analyzed using network science tools

Higher-order transition probabilities

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

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





Scholtes, ACM SIGKDD, 2017; De Bruijn, Proceedings of the Koninklijke Nederlandse Akademie van Wetenschappen, 1946

Next-step entropy

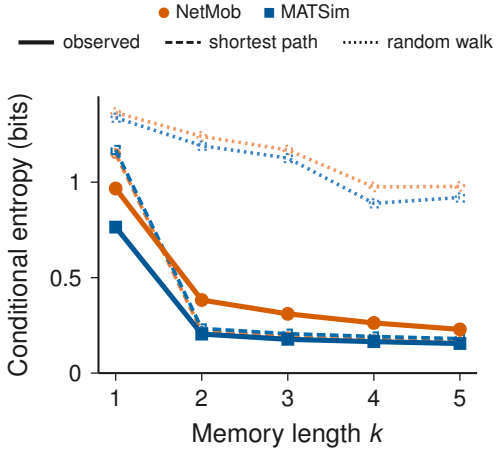
Next-step entropy

- 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 at every memory length
- Reinforces that observed trajectories are less predictable than simulated

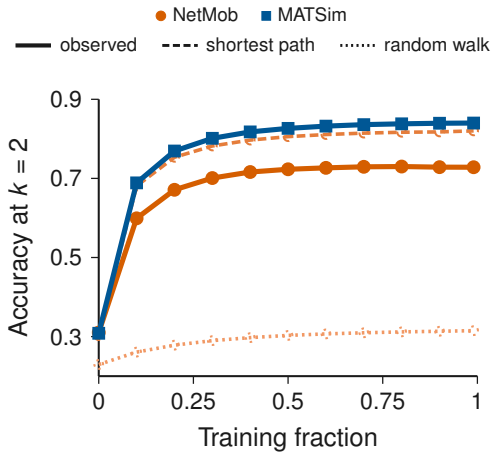


Figures: LaRock, Zhang & Hackl, arXiv:2606.00733.

Next-step prediction: Training Data

Varying the training fraction at $k = 2$

- Gap between MATSim and NetMob remains with 99% of paths used for training
- Limited training data does not explain the higher predictability of MATSim



Figures: LaRock, Zhang & Hackl, arXiv:2606.00733.

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 kth-order model as

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

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.

Network Coverage: Elements Unused by the Other Dataset

Stratified by how often an element is used

- Of the nodes and edges used at least t times by one dataset, what fraction is **never used by the other**?

Takeaways

- Agreement is better for **heavily used elements**: the unmatched fraction falls as t increases
- Across most of the range, **more of what NetMob uses is absent from MATSim** than the reverse

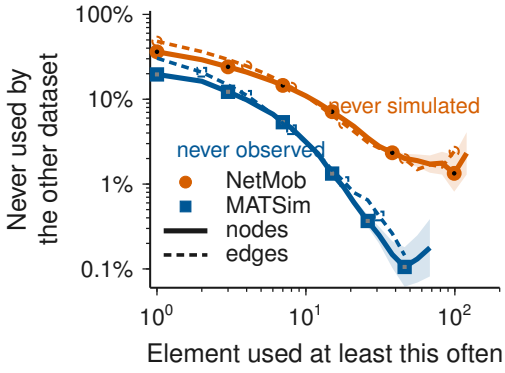


Figure: LaRock, Zhang & Hackl, arXiv:2606.00733.

Node-set Overlap Across All Pairs

Node-set Jaccard

- Jaccard similarity of the **sets of visited nodes** for every pair of datasets

Takeaways

- Observed and simulated mobility overlap substantially but incompletely: **Jaccard 0.55**
- The largest overlap is MATSim with its own shortest-path baseline (0.63), mostly because MATSim walking and cycling paths are shortest paths by construction (0.45 for vehicles only)

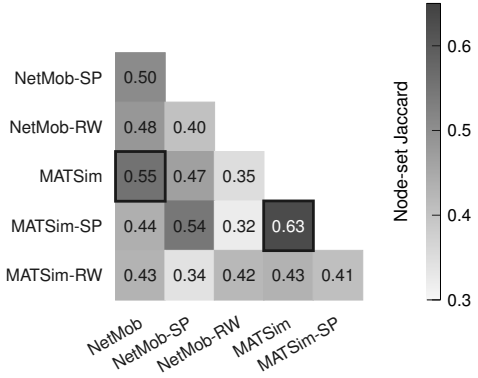
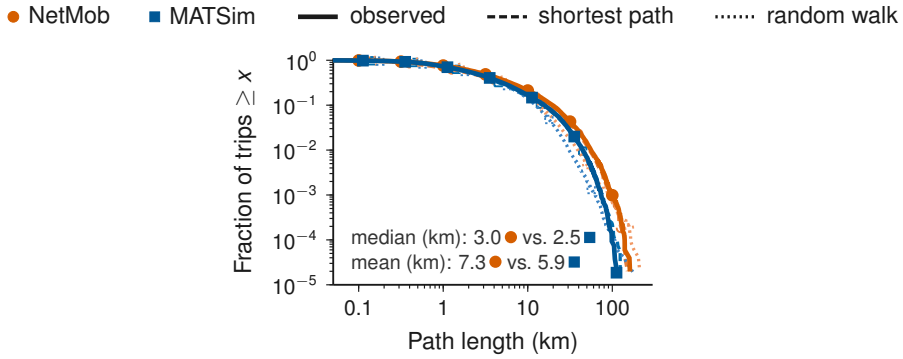


Figure: LaRock, Zhang & Hackl, arXiv:2606.00733. SP: shortest path, RW: random walk.

Path Length: Distance



Computed for this talk from the paths in LaRock, Zhang & Hackl, arXiv:2606.00733.

Path Length: Distance per Node

● NetMob ■ MATSim — observed ---- shortest path ····· random walk

Distance per node differs by length

- MATSim covers more distance per node for short paths
- Suggests routing prefers longer edges

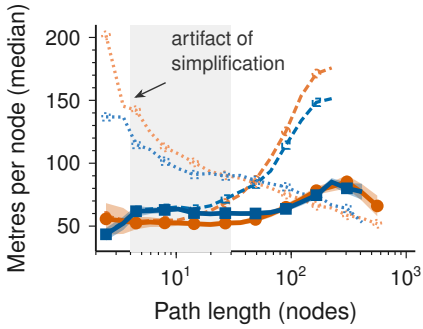
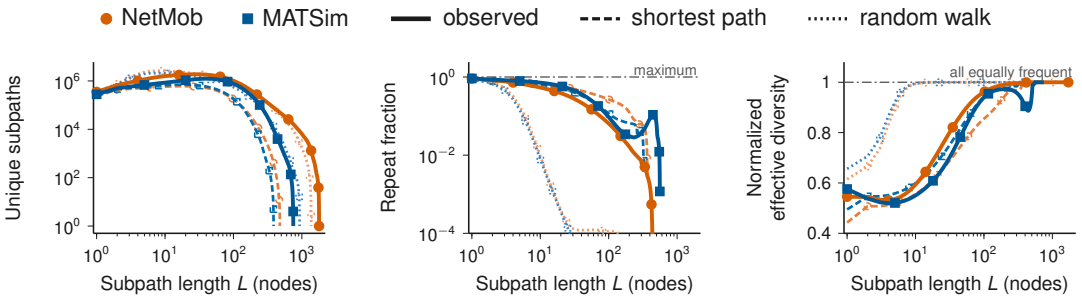


Figure: LaRock, Zhang & Hackl, arXiv:2606.00733 (baseline curves computed for this talk).

Subpath Diversity



Observed mobility has more distinct long subpaths, with less repetition

- Unique subpaths peak at L = 27 for NetMob and L = 39 for MATSim
- Repeat fraction similar at short lengths; for long subpaths it keeps falling for NetMob but rises again for MATSim
- Subpath frequencies similarly concentrated in both (normalized effective diversity)

Figures: LaRock, Zhang & Hackl, arXiv:2606.00733.