Prediction and Privacy for Human Mobility Data

Matthias Grossglauser
EPFL

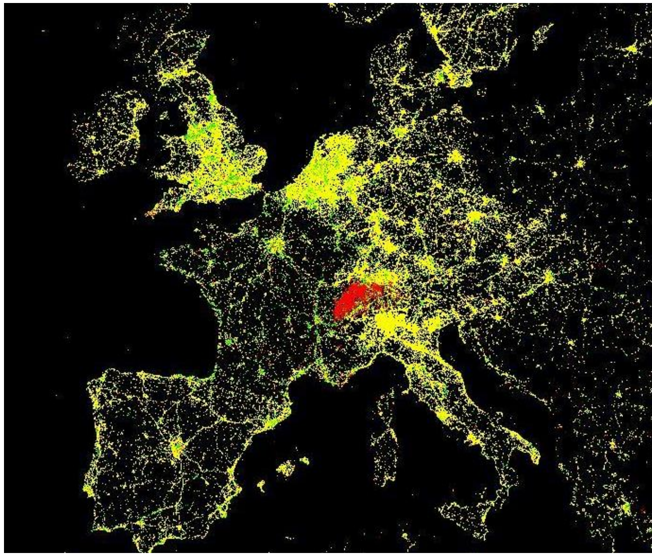
With Mohamed Kafsi and Patrick Thiran

Allerton, Oct 2014

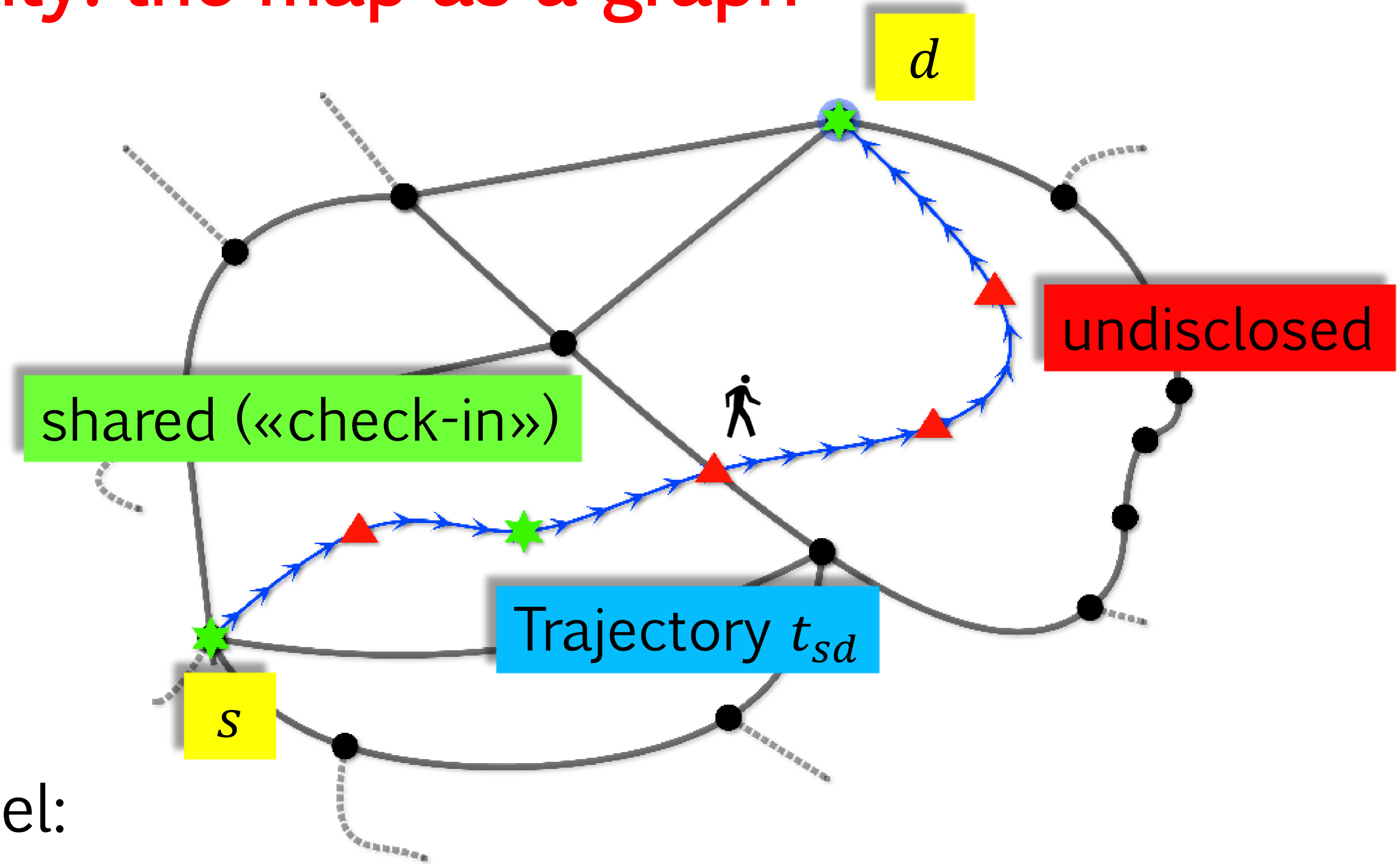
ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Mobility mining

- Mobility patterns say a lot about us:
  - Activities, social contacts & communities, work, travel,…
  - People share location info: "check-ins" (foursquare etc.)
- Opportunities:
  - Optimizing services, anticipating needs (aka targeted advertisement)
  - Infrastructure optimization, store placement,…
- Threats:
  - Personal privacy: profiling, revealing locations,…



Mobility: the map as a graph



- Model:
  - World = a graph
  - User mobility = sequence of vertices (trajectory)
- Question:
  - How undisclosed are undisclosed locations?

Model

- Assumptions:
  - Markov chain capturing mobility patterns
  - Check-in = conditioning on an intermediate state
  - Privacy = uncertainty about trajectory T_{sd}: conditional entropy
- Result:
  - Formulate as conditional entropy of Markov trajectories given intermediate states
  - Exact results on "number of bits" revealed about trajectory [KGT13]
  - Extension of classical result by [Ekroot & Cover 1993]

Entropy of Markov trajectories

- Measuring uncertainty about the trajectory: Shannon entropy of the trajectory from s to d:

$$H_{sd} \stackrel{\text{def}}{=} H(T_{sd}) = -\sum_{t_{sd} \in \mathcal{T}_{sd}} p(t_{sd}) \log p(t_{sd})$$

- $\mathcal{T}_{sd}$= set of trajectories starting at s, ending at d, with no intermediate state d
  - Cardinality is typically infinite
- H: matrix of trajectory entropies
  - General closed-form expression [Ekroot & Cover, 1993] for irreducible MC

Conditional entropy of Markov trajectories

- How does the predictability of a trajectory evolve when we condition on a sequence of intermediate states $\boldsymbol{u} = (u_1, u_2, \dots, u_l)$?
- Conditional entropy of the trajectory from s to d visiting all intermediate states $\boldsymbol{u}$:

$$H_{sd|\boldsymbol{u}} = -\sum_{t_{sd} \in \mathcal{T}_{sd}^{\boldsymbol{u}}} p(t_{sd} | t_{sd} \in \mathcal{T}_{sd}^{\boldsymbol{u}}) \log p(t_{sd} | t_{sd} \in \mathcal{T}_{sd}^{\boldsymbol{u}})$$

- $\mathcal{T}_{sd}^{\boldsymbol{u}}$: set of trajectories starting at s, ending at d, with no intermediate state d, and $\boldsymbol{u}$ as a subsequence
- Again, enumerating all trajectories costly or impossible (infinite)

Computing conditional entropy: step 1

- Show that conditional entropy given subsequence $u = (u_1, u_2, \ldots, u_l)$ can be decomposed into segments:

$$H(T_{sd} | T_{sd} \supset s u_1 \ldots u_l d) = \sum_{k=0}^{l-1} H_{u_k u_{k+1} | \bar{d}} + H_{u_l d}$$

- Problem: trajectory entropy $H_{s' d' | \bar{d}}$ conditioned on not going through state d
- Computing $H_{s' d' | \bar{d}}$:
  - Derive new matrix P', such that unconditional entropy in P'=conditional entropy in P

Step 2: transforming $\boldsymbol{P}$ into $\boldsymbol{P'}$

$$P$$

$$H(T_{s'd'} | T_{s'd'} \notin \mathcal{T}^{d}_{s'd'})$$

d' and d are made absorbing

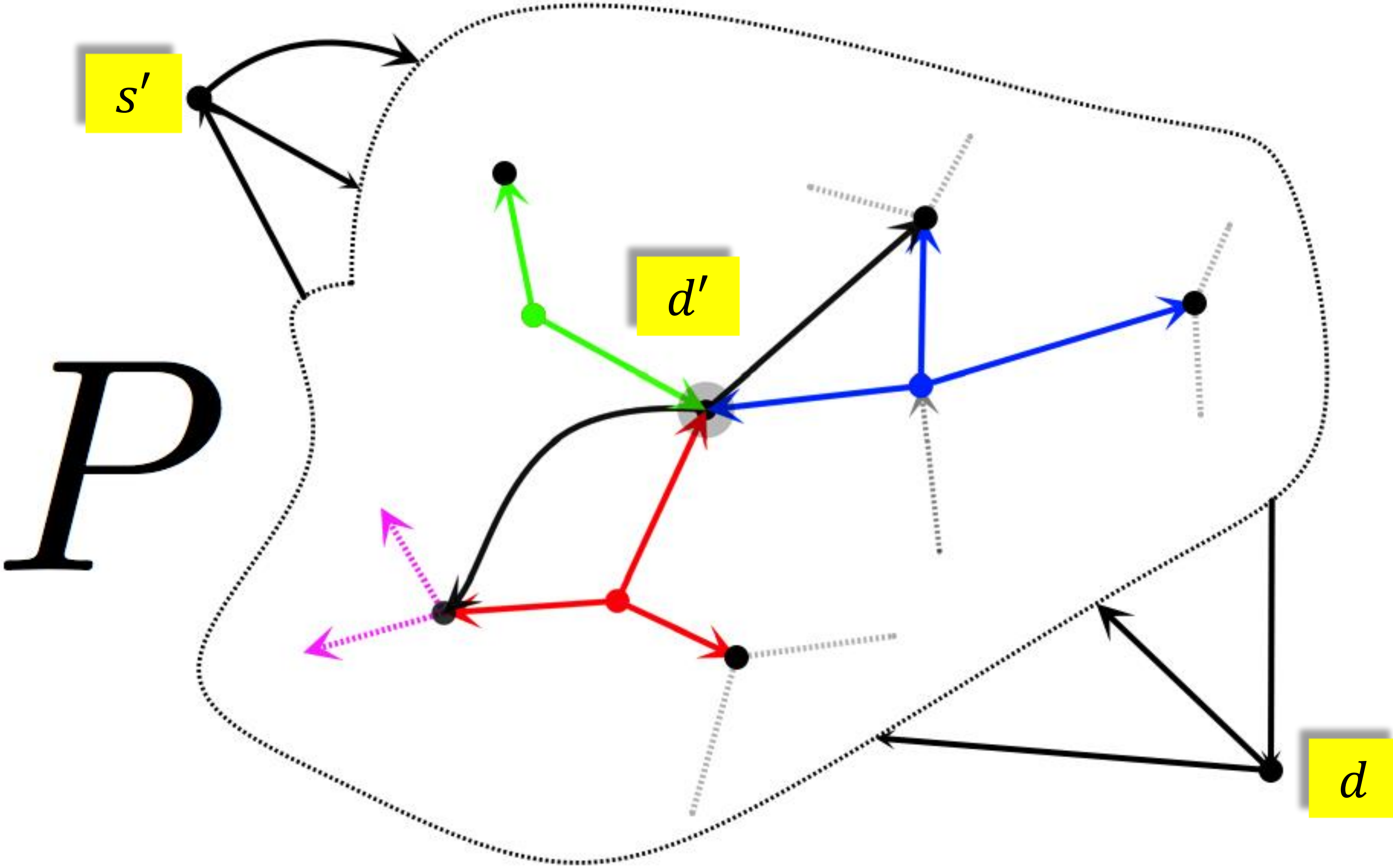
$$\bar{P}$$

$$P'_{ij} = \begin{cases} \dfrac{\alpha_{jd'd}}{\alpha_{id'd}} \bar{P}_{ij} & if\ \alpha_{id'd} > 0 \\ \bar{P}_{ij} & \text{otherwise} \end{cases}$$

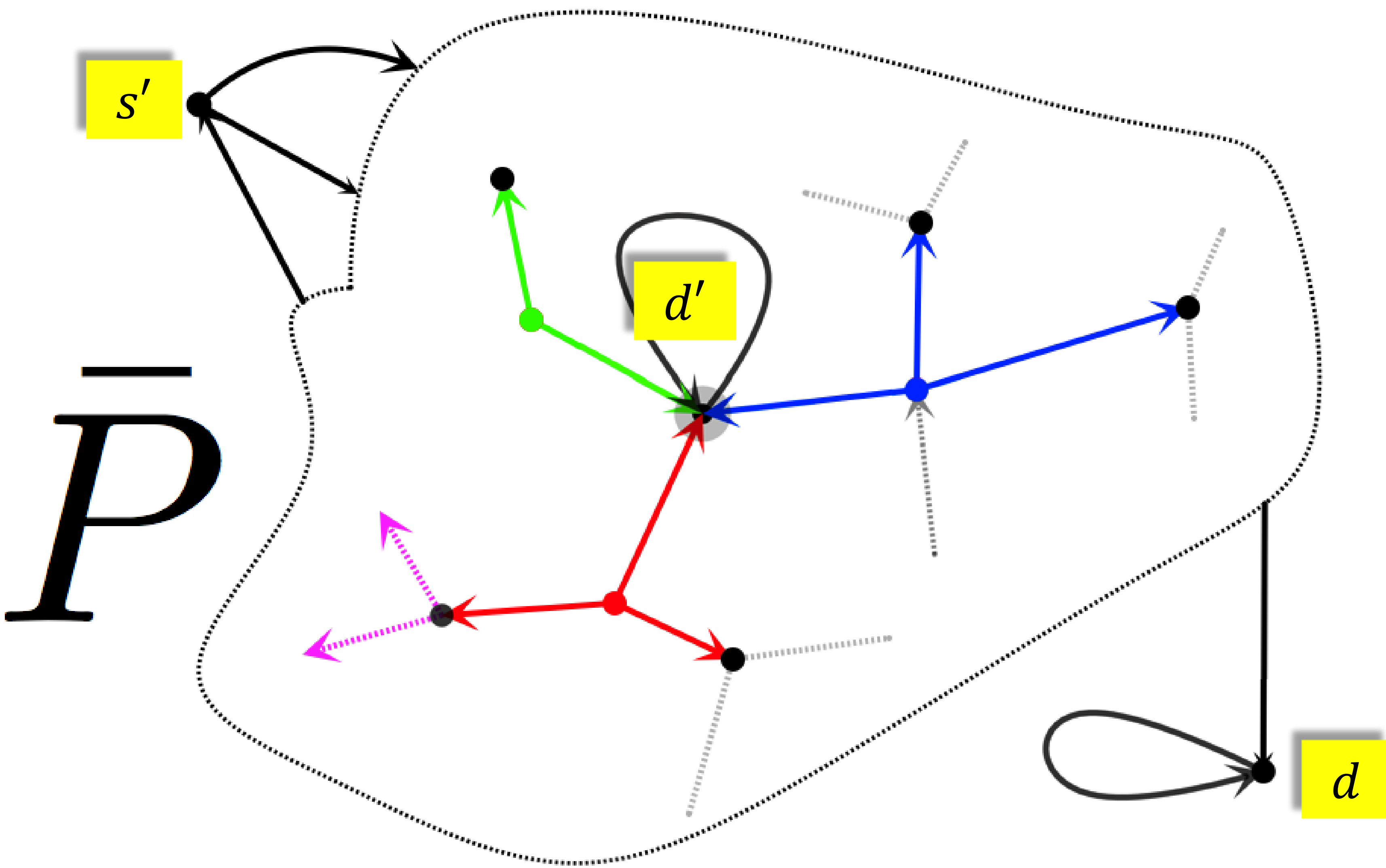
$$H(T'_{s'd'})$$

$$P'$$

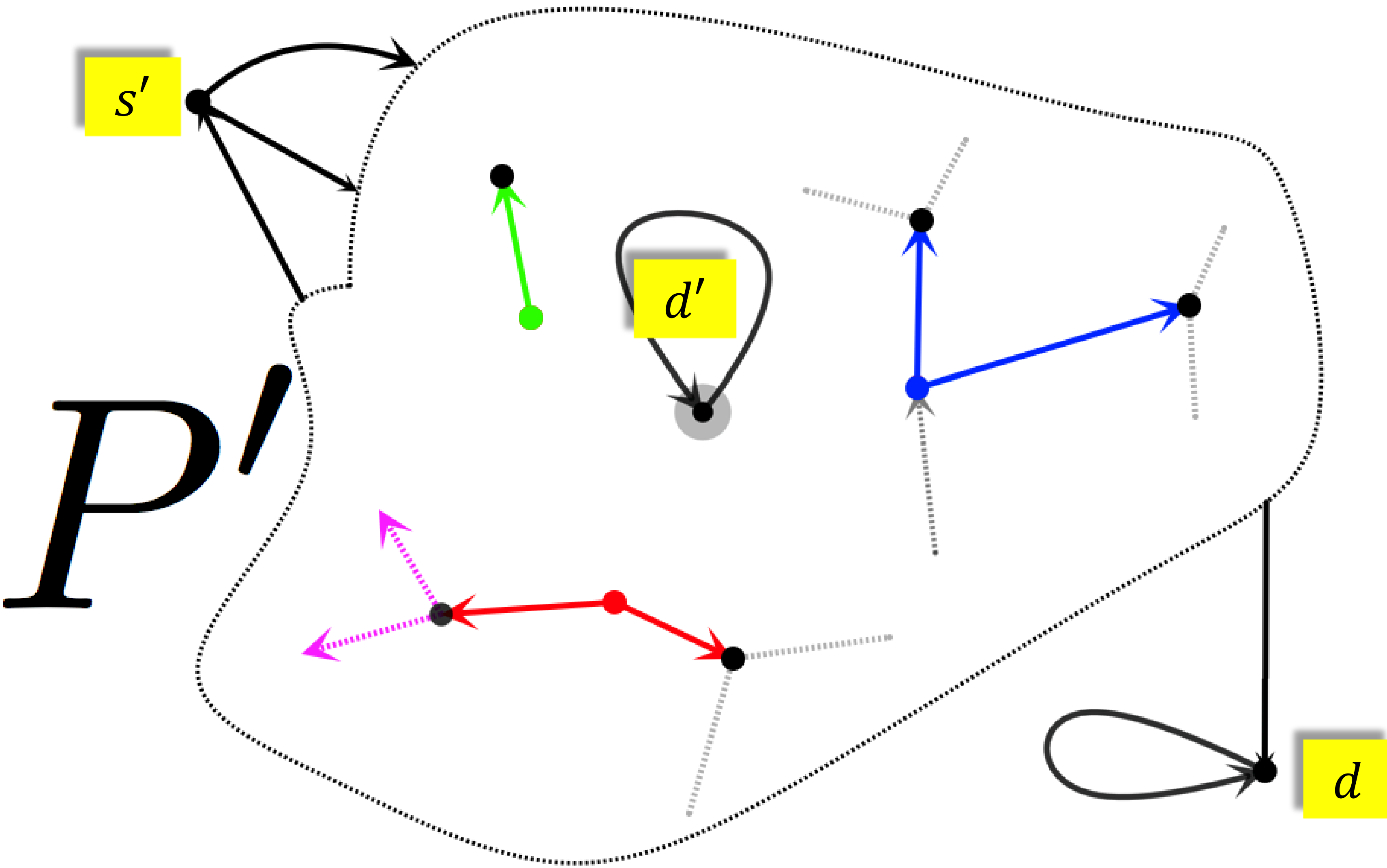
Step 2: P



Step 2: $\bar{P}$ has d, d' absorbing



Step 2: P': normalized transition probabilities



Step 2: computing $H_{s'd'|\bar{d}}$

- Basic idea: reduce computing conditional entropy → unconditional entropy over a modified MC
- Relationship between original chain and P':
  - $t_{s'd'} \in \mathcal{T}^{d}_{s'd'}$ → $p'(t_{s'd'}) = 0$
  - $t_{s'd'} \notin \mathcal{T}^{d}_{s'd'}$ →

$$p'(t_{s'd'}) = P'(s',x_2)P'(x_2,x_3)\ldots P'(x_k d')$$
$$= \frac{\alpha_{x_2 d' d}}{\alpha_{s'd'd}} P(s',x_2) \frac{\alpha_{x_3 d' d}}{\alpha_{x_2 d' d}} P(x_2,x_3) \ldots \frac{\alpha_{d'd'd}}{\alpha_{x_k d' d}} P(x_k, d')$$
$$= \frac{\alpha_{d'd'd}}{\alpha_{s'd'd}} P(s,x_2)P(x_2,x_3)\ldots P(x_k,d')$$
$$= \frac{p(t_{s'd'})}{p\left(T_{s'd'} \notin \mathcal{T}^{d}_{s'd'}\right)} = p(t_{s'd'} | T_{s'd'} \notin \mathcal{T}^{d}_{s'd'})$$

Filtering trajectories hitting d first

Step 3: unconditional entropy for general MC

- Relaxing the irreducibility condition of [Ekroot&Cover93]
- Express the entropy as a linear combination of local entropies

$$H_{s'd'} = \sum_{i \neq d'} \left((I - Q_{d'})^{-1} \right)_{s'i} H(P_{i\cdot})$$

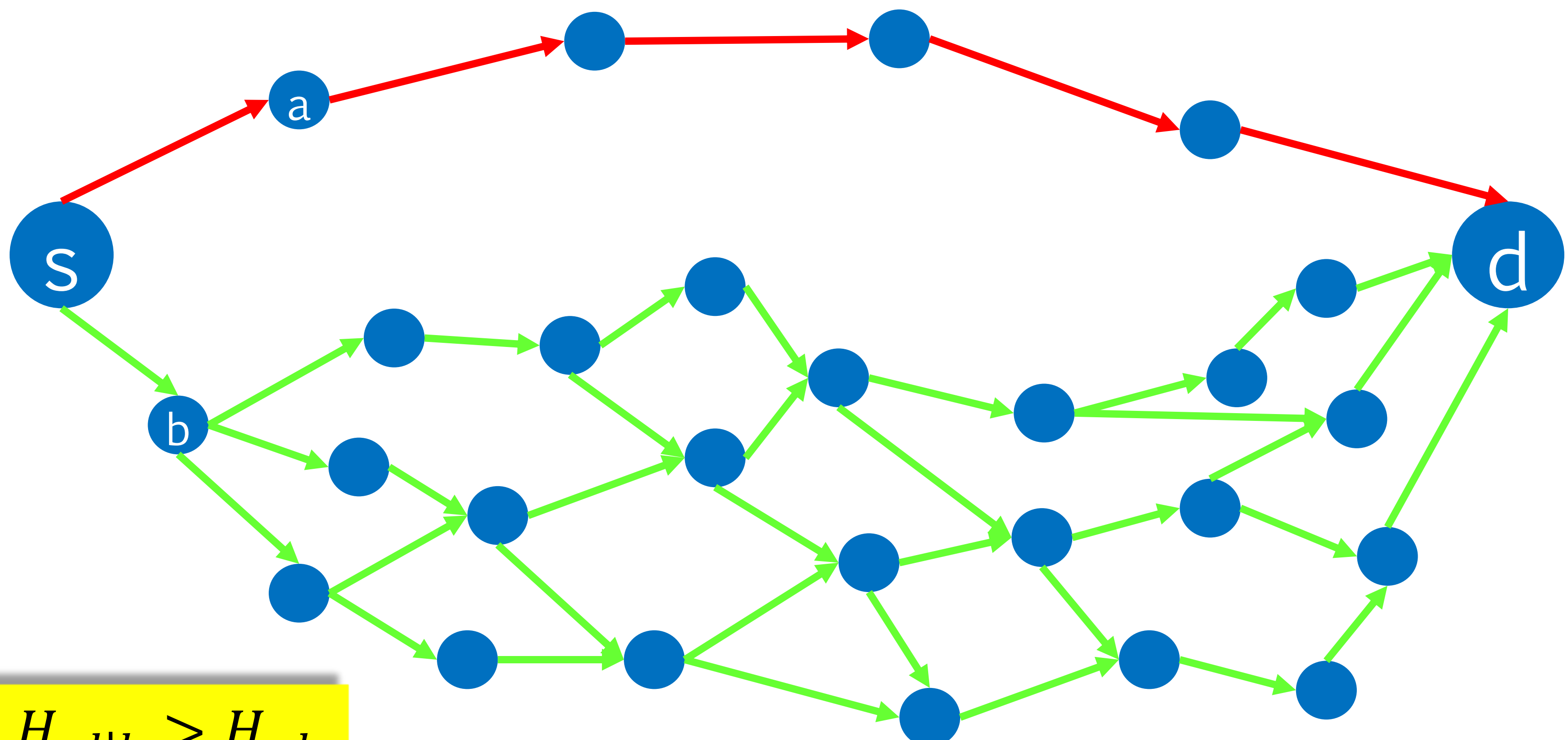
Expected number of visits to state i

Entropy of next transition out of state i

Conditional trajectory entropy: not monotonic!

- Counter-example:

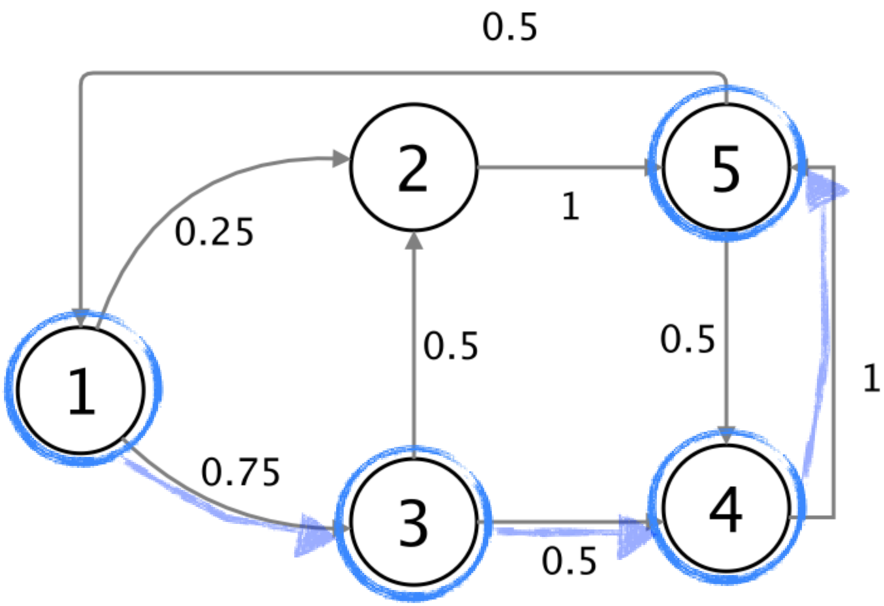
$H_{sd|a} = 0 < H_{sd}$



$H_{sd|b} > H_{sd}$

Conditional trajectory entropy: not additive!

- Counter-example:



$$\begin{pmatrix} 3.56 & 3.69 & 1.74 & \boxed{3.18} & 1.56 \\ 2 & 5.69 & 3.74 & 2.59 & 0 \\ 3 & 3.84 & 4.74 & 2.29 & 1 \\ 2 & 5.69 & 3.74 & 2.59 & \boxed{0} \\ 2 & 5.69 & 3.74 & 2.59 & 1.78 \end{pmatrix}$$

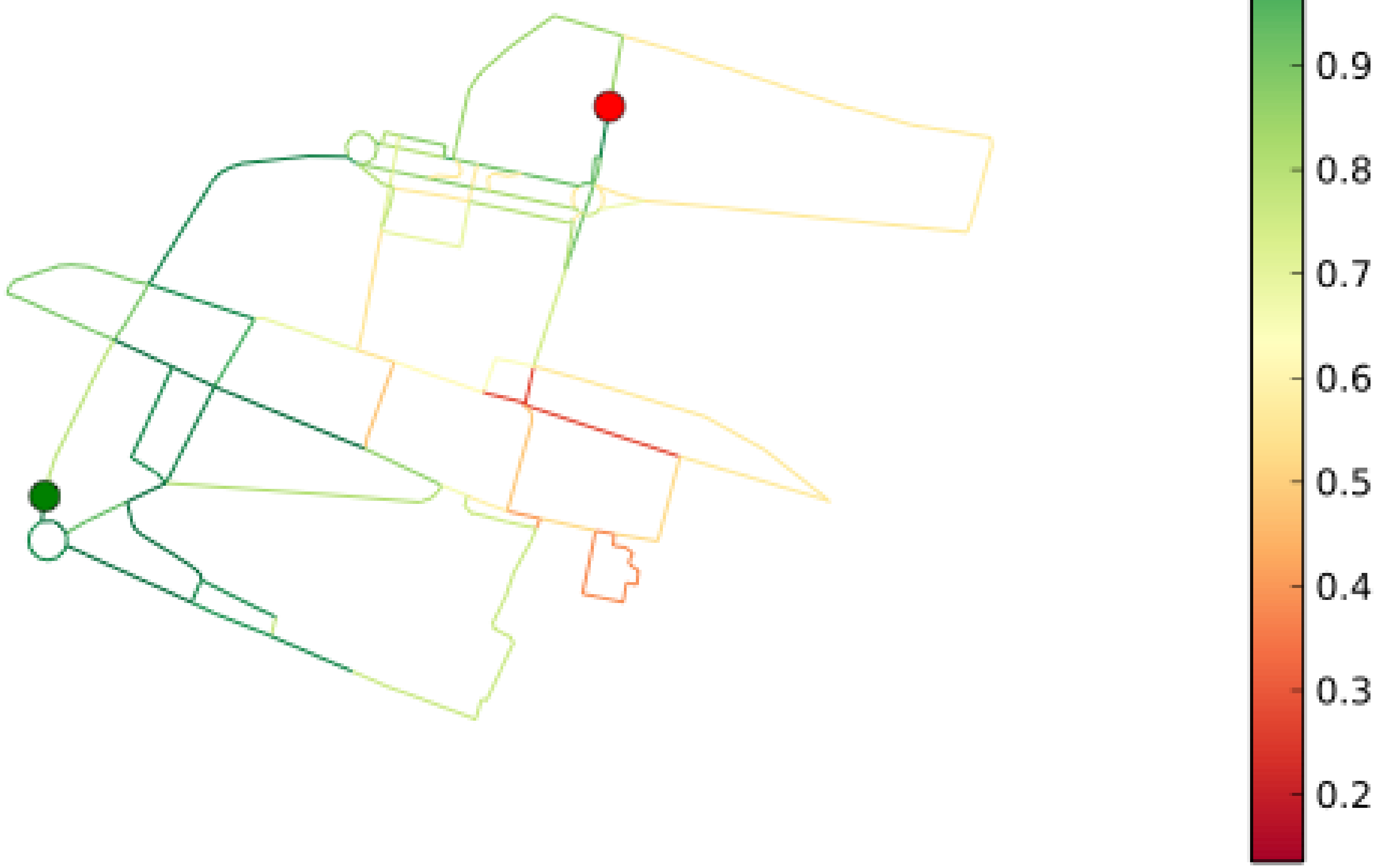
$$H_{15|4} \neq H_{14} + H_{45}$$

Computational cost

- Worst-case complexity: $O(ln^3)$
  - l: length of conditioning vector
  - n: number of states
  - Dominated by computation of $(I - Q_d)^{-1}$
  - Linear in length l of conditioning vector → efficient to process long trajectories
- Processing individual trajectory:
  - Only row s of $(I - Q_d)^{-1}$ needed → rely on efficient methods for sparse matrix inversion
- Processing large batch of trajectories:
  - Computation of $(I - Q_d)^{-1}$ amortized → linear in total # of conditioning states (over all trajectories)

Application: trajectory privacy with check-ins

Normalized conditional entropy: $\dfrac{H_{sd|u}}{H_{sd}}$

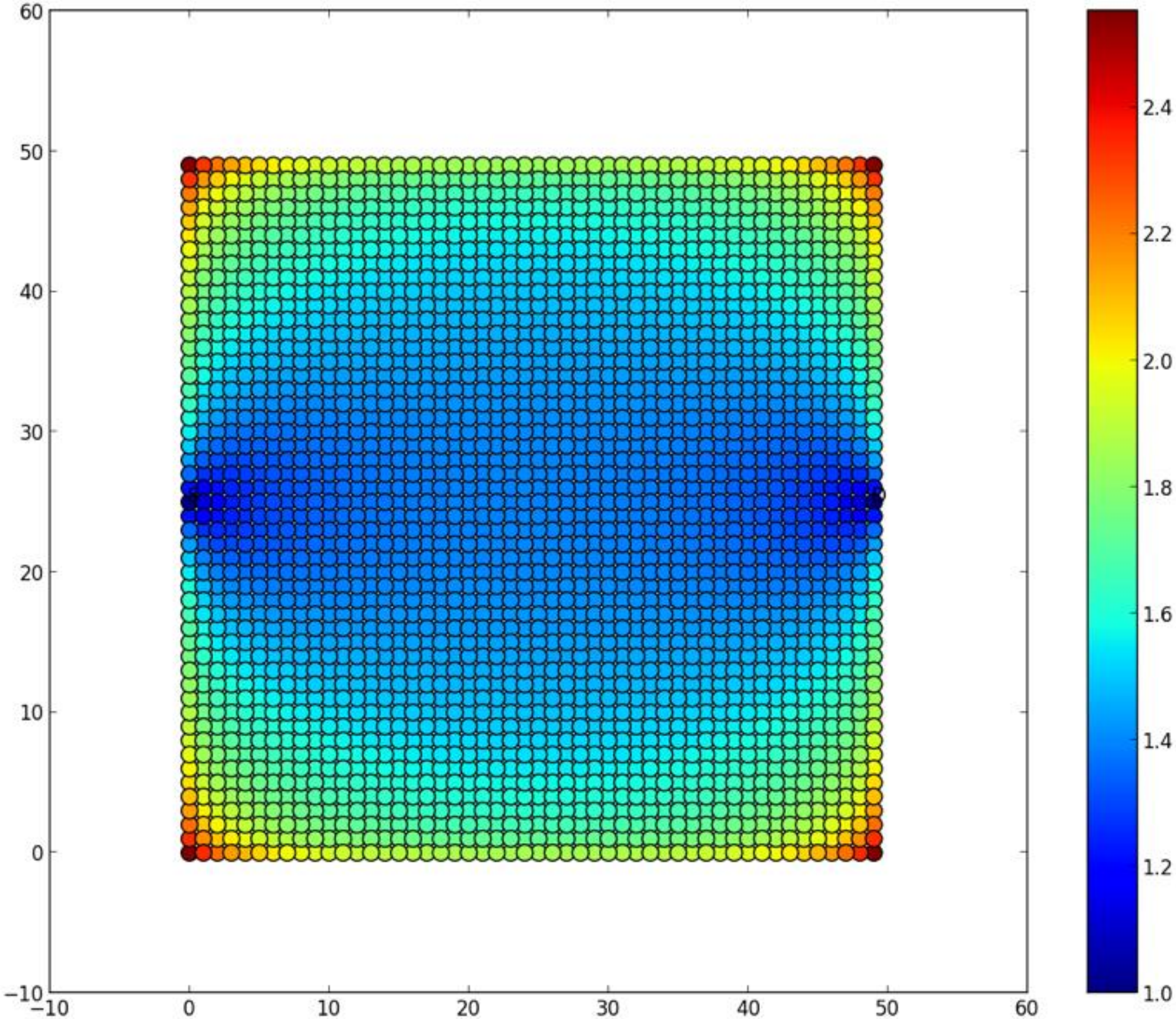


Application: trajectory segmentation

- Human mobility:
  - Serves to reach a set of "waypoints" = intermediate destinations
- Waypoints: personal choices
  - Work; school; shopping; doctor's appointment; ...
- Between waypoints: generic behavior
  - Optimization of travel time & cost; reacting to conditions; incomplete information
- Question:
  - Given only a low-order mobility model trained from a whole population, can we infer waypoints for individual users?
- Intuition:
  - Adding "out of the way" waypoints enriches the set of plausible trajectories → $H_{sd|\boldsymbol{u}} > H_{sd}$

Example:

- $H_{sd|u}/H_{sd}$ as a function of u, for unbiased random walk

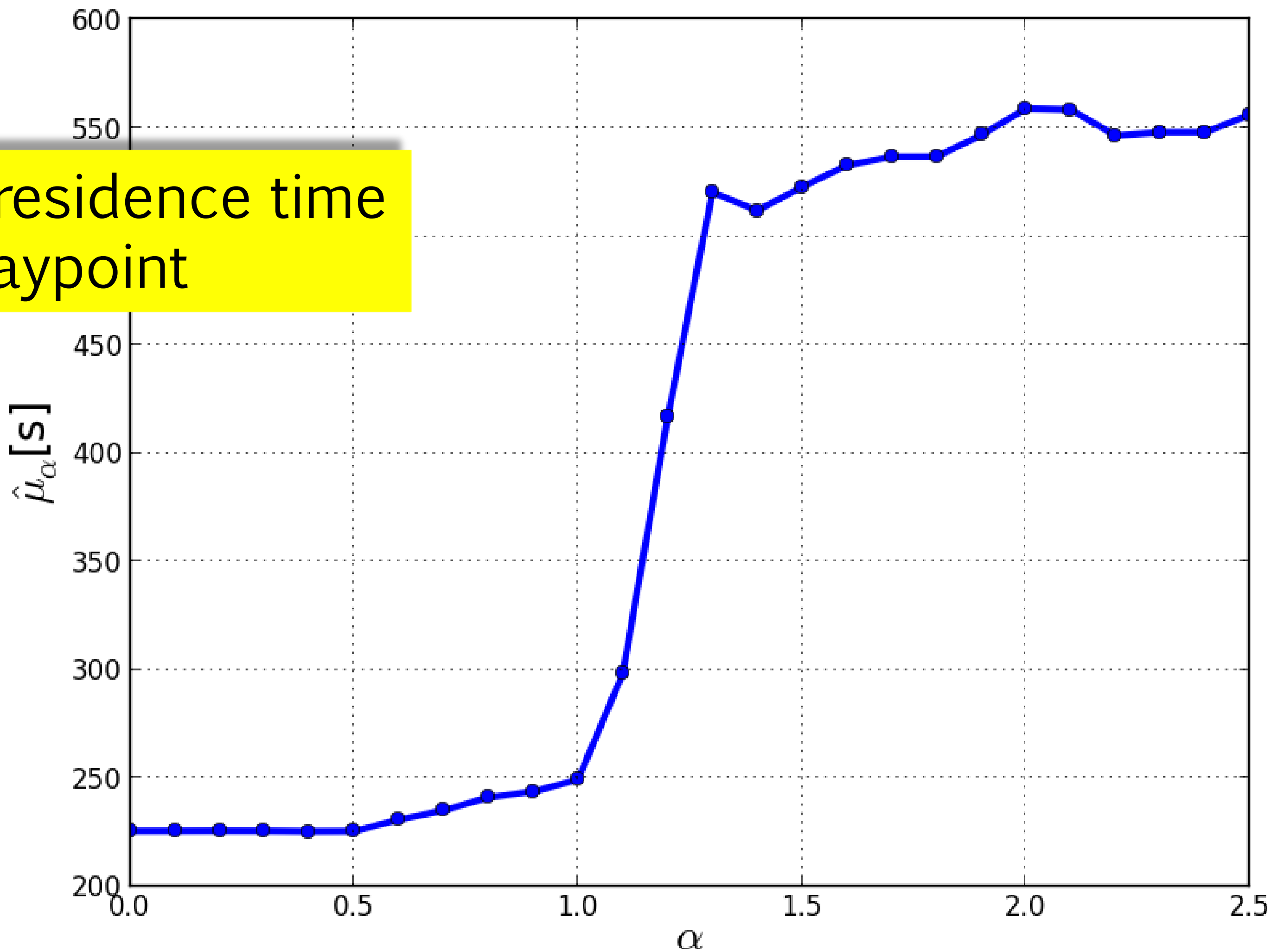


Segmentation of mobility traces

- Geolife project: ~ 200 users, 20k trajectories

Residence time vs relative conditional entropy

high residence time
→ waypoint



- Expected residence time vs $\dfrac{H_{sd|u}}{H_{sd}} > \alpha$

Conclusion

- Principled way to quantify mobility uncertainty
  - Conditional entropy given start, end, intermediate states
  - With respect to a Markov mobility model
  - Low-order: easy to learn (dense) & compute; representative for population; overfitting control
  - Efficient to process large batches of trajectories
- Privacy:
  - Information loss (or gain!) by revealing set of locations
  - Not monotonic, not additive
  - Inverse problem: trajectory compression
- Segmentation:
  - Idea: trajectory = reaching a sequence of waypoints
  - Expect high $H_{sd|\boldsymbol{u}}$ for waypoints $\boldsymbol{u}$
  - Can segment without time stamps & spatial coordinates, and relative to generic model

Prediction and Privacy for Human Mobility Data

Thanks!

Questions?

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE