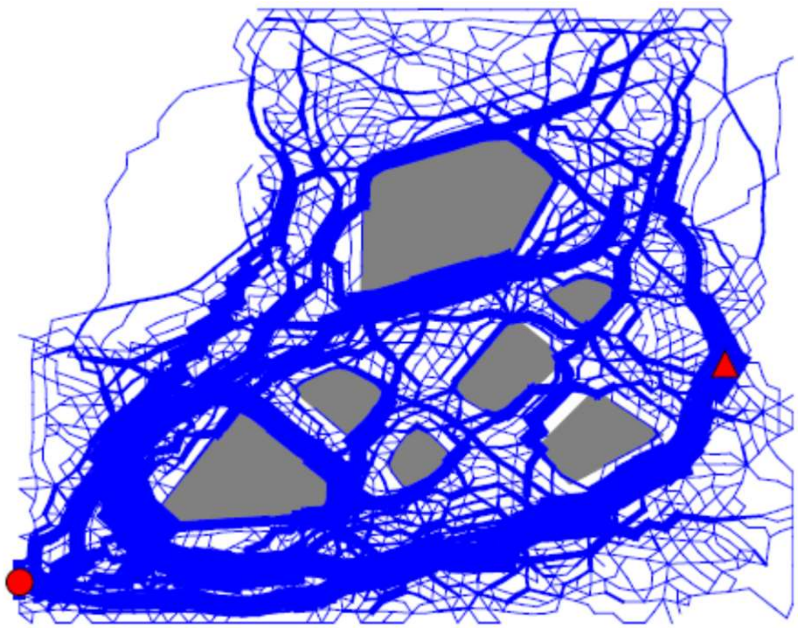
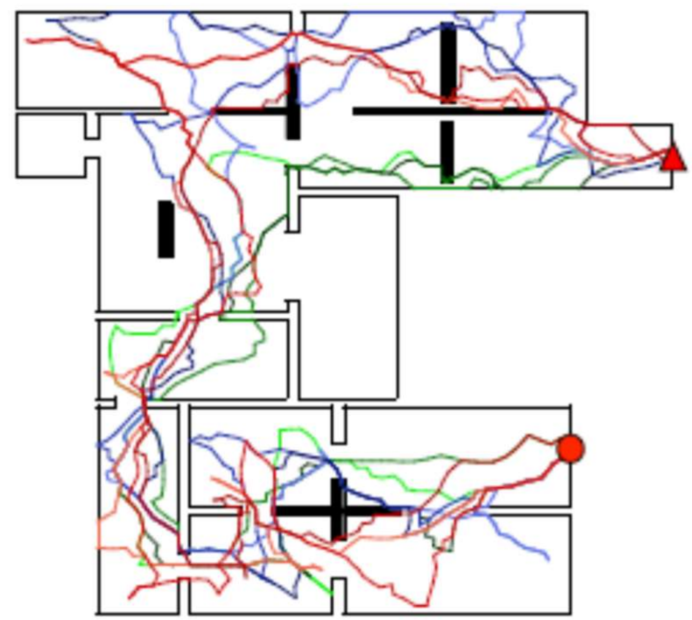
Don’t Collect Too Much -- Geometric Approaches for Protecting Trajectory Privacy

Jie Gao
Stony Brook University
April 26th, 2017
Dagstuhl Workshop on CG

Location and Trajectory Privacy

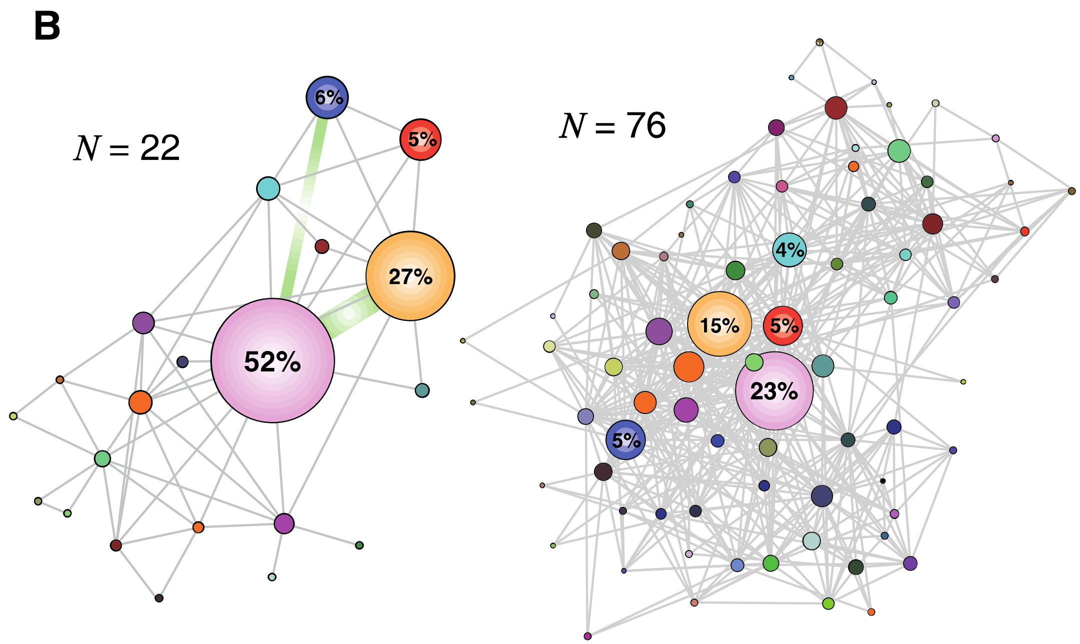
- Locations/trajectories are collected.





Trajectories are sensitive & identifying

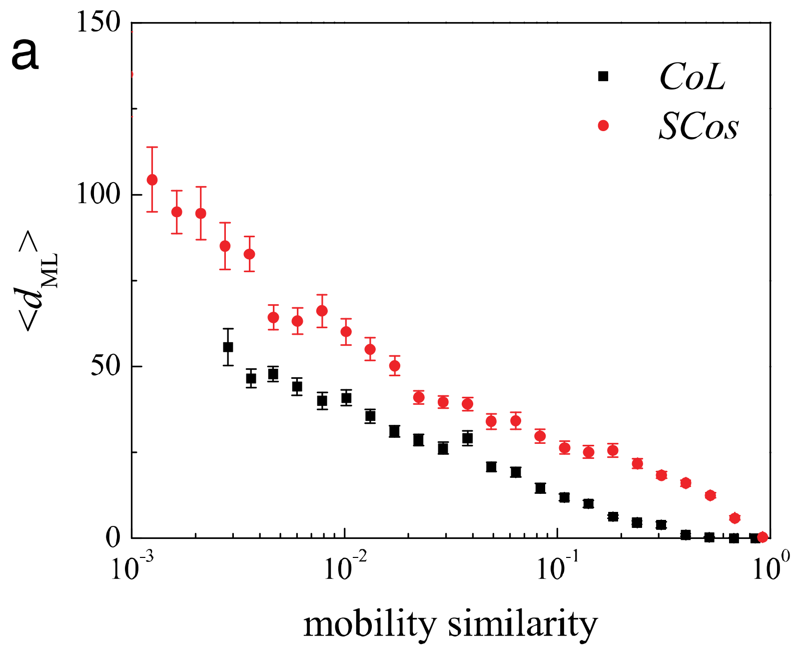
- Frequently visited locations → home/work address; predictability of location > 93%



Limits of Predictability in Human Mobility, Science, 2010.

Trajectories are sensitive & identifying

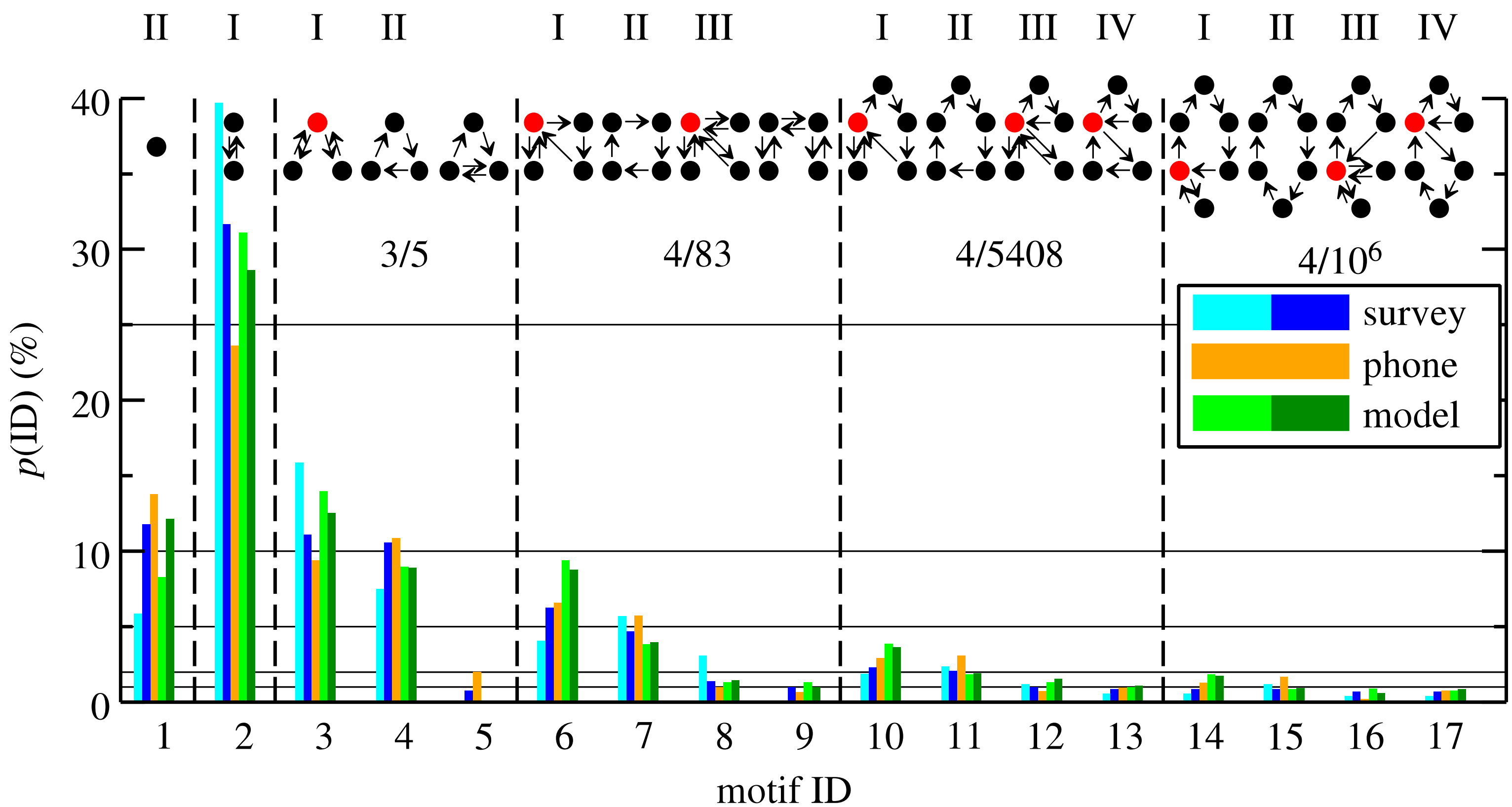
- Frequent co-location patterns → social ties



Human Mobility, Social Ties, and Link Prediction, KDD'11.

Trajectories are sensitive & identifying

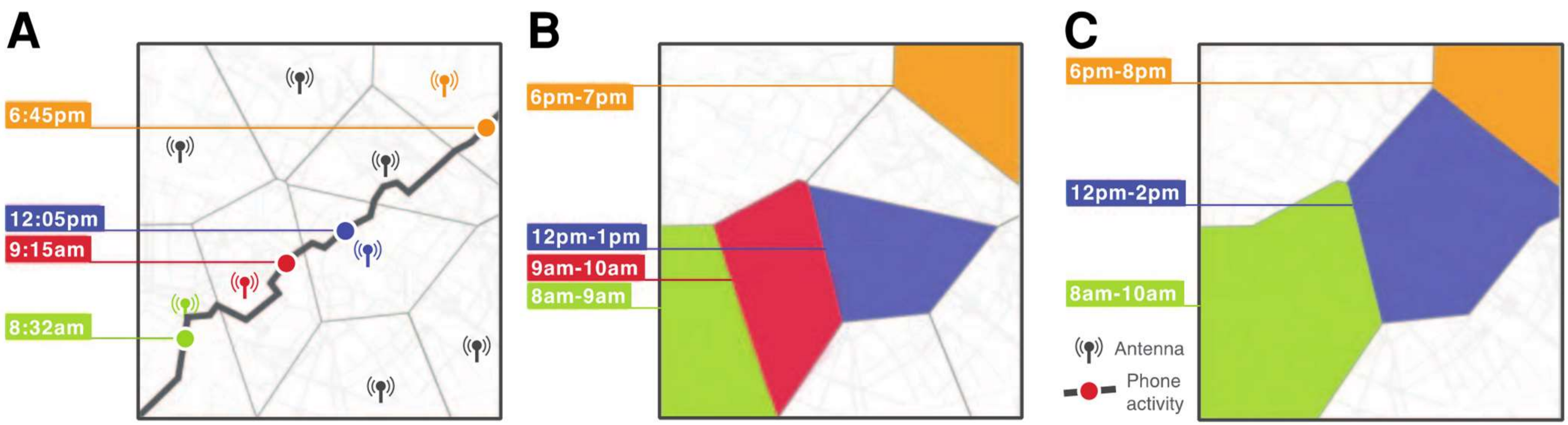
- Motifs – revealing activities



Unravelling daily human mobility motifs, 2013.

Trajectories are sensitive & identifying

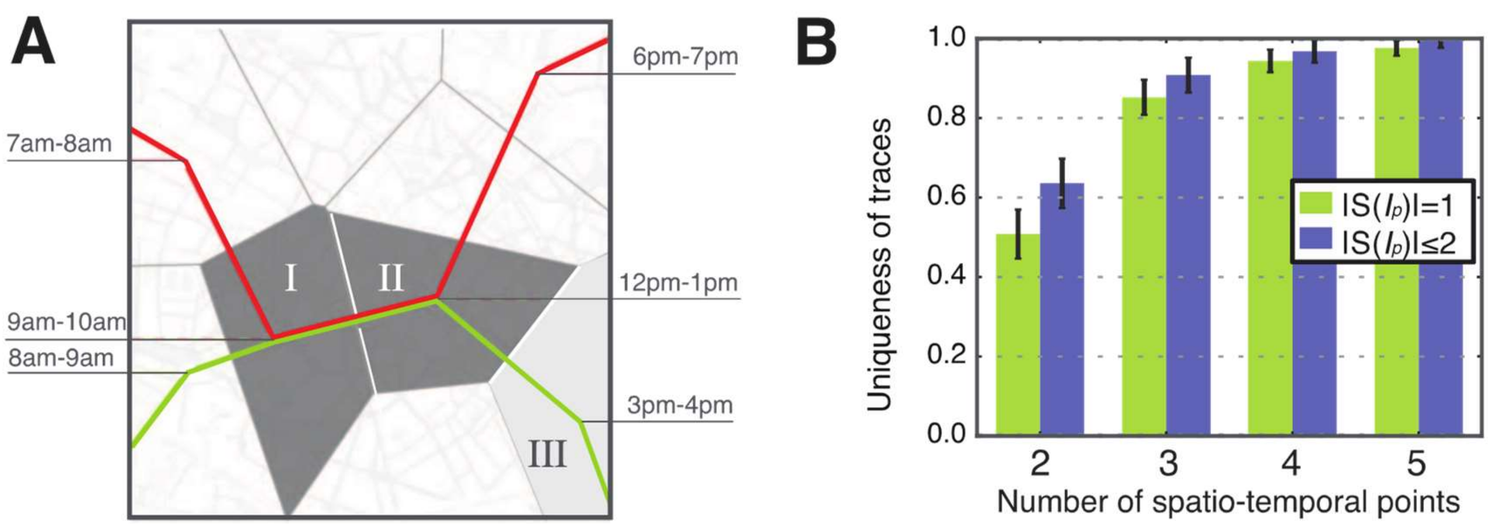
- Unique signature.



Unique in the Crowd: The privacy bounds of human mobility, Nature, 2013.

Trajectories are sensitive & identifying

- Unique signature – 4 spatial-temporal points are enough to identify 95% trajectories in 1.5 million users.



Trajectory Privacy Protection

- Utility vs Privacy:
  - Statistical patterns of group mobility: clustered motion; popular paths
  - Anomaly detection
- Privacy models
  - K-anonymity; “r-gather clustering”
  - Differential privacy.

Location/Trajectory Collection

- Location/trajectory collected by GPS and stored on the device.
  - Users voluntarily contribute such data.

- Wireless devices leave traces behind.
  - Cell towers.
  - WiFi AP.

Privacy Preserving with Sensing

1. Collect data;
2. Run anonymization;

Or, answer statistical queries with privacy added.

1. Collect **little** data
2. Derive group behaviors or statistical patterns.

One Network Setting; Two Case Studies

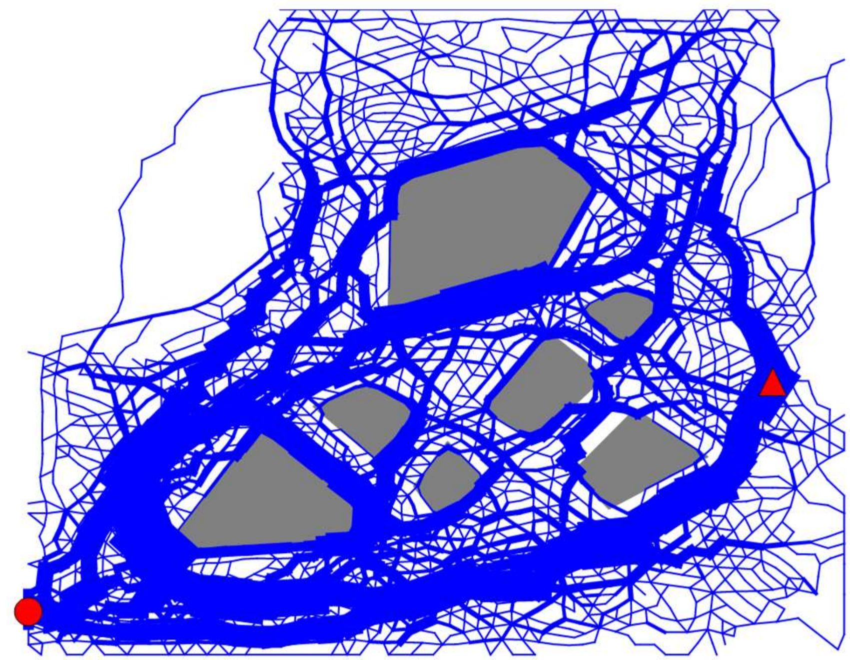
- Smart city environment: many checkpoints that record user mobility.
- What shall be collected at these checkpoints?
  - Low cost, w/ privacy protection.

1. Distributed trajectory clustering by homology.
2. Popular path mining and query.

Part I: Trajectory Clustering

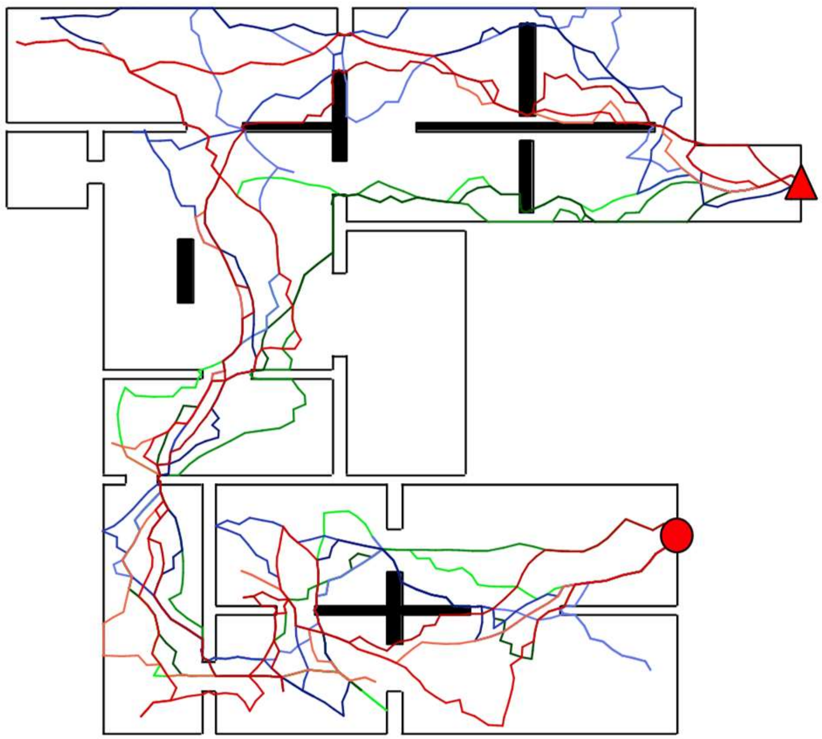
Clustering by Homology

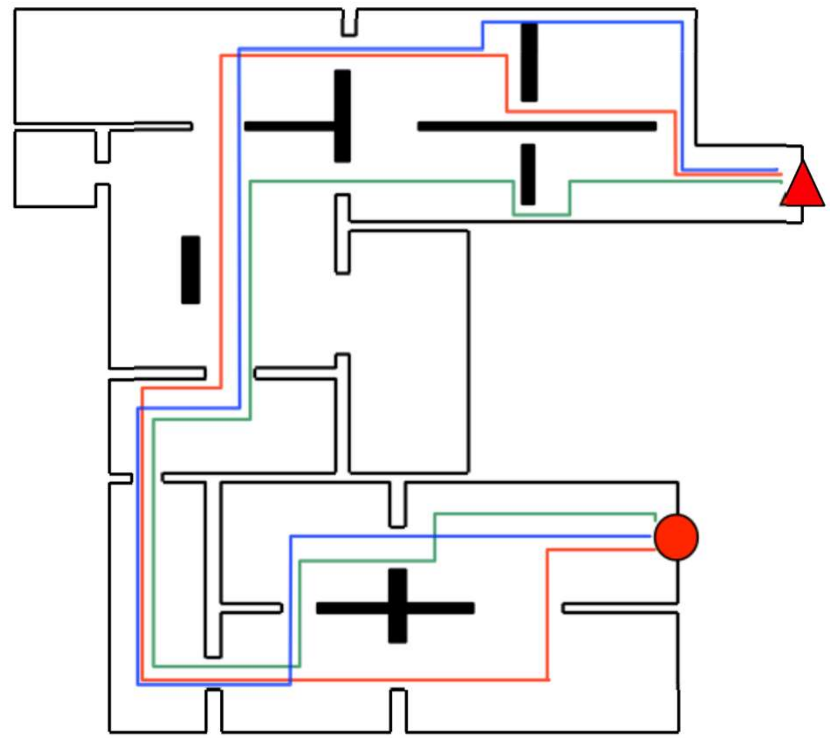
(a) 4 trajectories with different homology types



(b) Trajectory flow

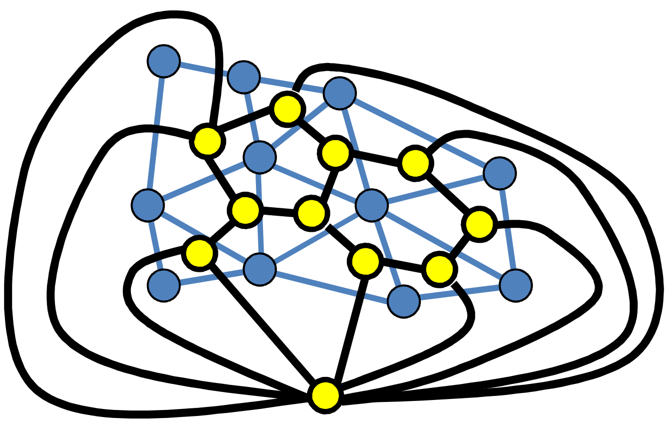
Clustering by Homology

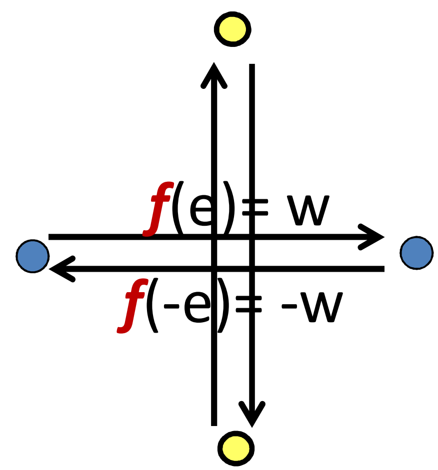




Differential 1-Form

- Planar graph G with **faces**.
- One-form: "**directed**" weights ***f*** on edges.
- Dual graph G': face➔vertex; vertex➔face; edges rotated by 90°.

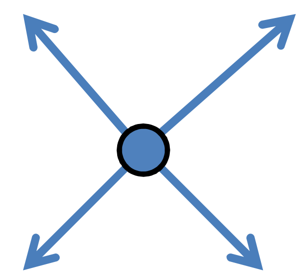




Harmonic 1-Form

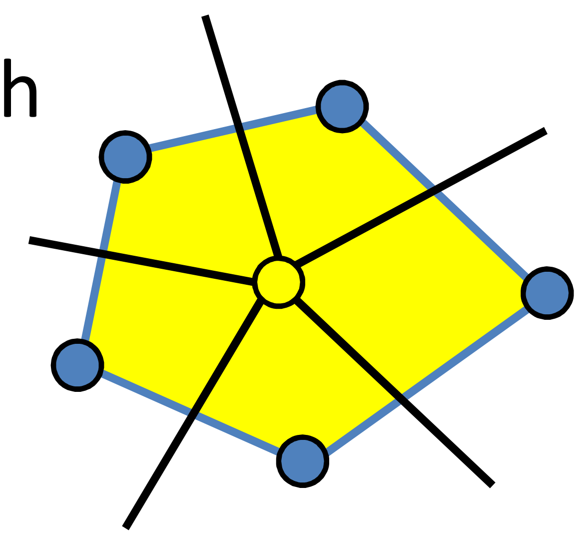
1. **Divergence-free**: $\sum_{\text{neighbor } v} f(u, v)=0$

i.e., no sources, no sinks



2. **Curl free**: $\sum_{\text{edge } e \text{ on a face}} f(e)= 0$

i.e., divergence-free in dual graph



Use Harmonic 1-form

- For a cycle **not enclosing any hole**, the integration of the harmonic 1-form is **zero**.
- **Preprocessing**: Compute a harmonic 1-form on the graph s.t. only cycles enclosing holes integrate to non-zero values
- **Homology check**: Simple integration along the trajectories.
- Distributed storage & computation.
- How to compute a harmonic 1-form?

Hodge Decomposition

- Start w/ an arbitrary 1-form ω.
- Hodge decomposition

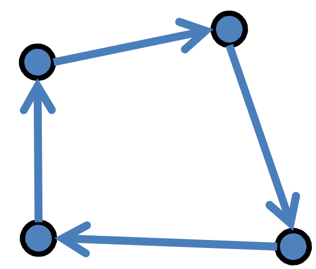
$$\omega = \alpha + \beta + \gamma$$

- **α: gradient flow**, $\alpha(u, v)= \tau(u)- \tau(v)$, τ is a potential function on **vertices, 0-form**.
- **Operation δ**: Integration along a face

$= \tau(u_1) - \tau(u_2) + \tau(u_2) - \tau(u_3) + \ldots$

$+ \tau(u_k) - \tau(u_1)$.

$= 0$



Hodge Decomposition

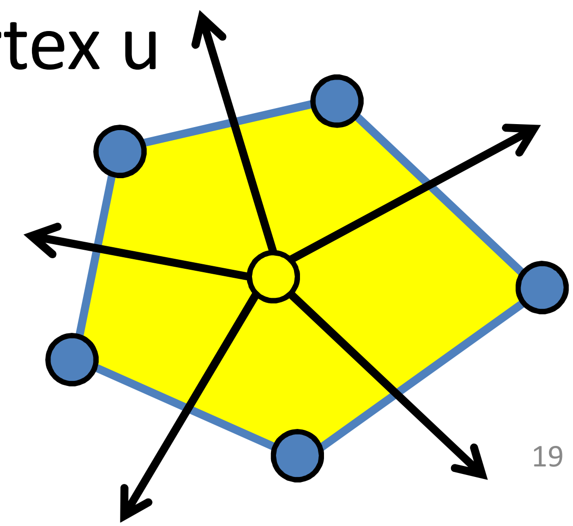
- Hodge decomposition

$$\omega = \alpha + \beta + \gamma$$

- **β: curl flow**, i.e., gradient flow in the dual graph, $\beta(u, v) = \eta(x) - \eta(y)$, x is the face to the right, y is the face to the left. η is a function on **faces, 2-form**.
- **Operation d**: ∑ β on edges of vertex u

= ∑ β dual edges on face u*

= 0



Hodge Decomposition

- Hodge decomposition

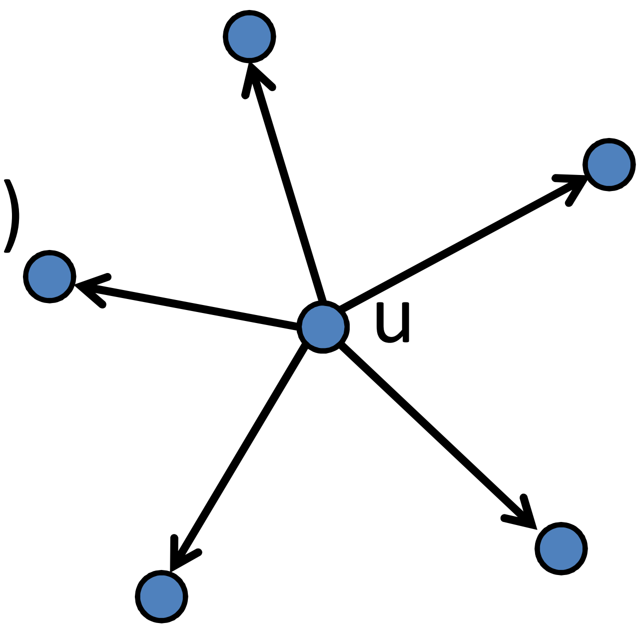
$$\omega = \alpha + \beta + \gamma$$

- γ: harmonic 1-form.
- Integration along a face = 0 (curl-free)
- Integration on edges of a vertex = 0 (divergence-free)

Gossip-style Implementation

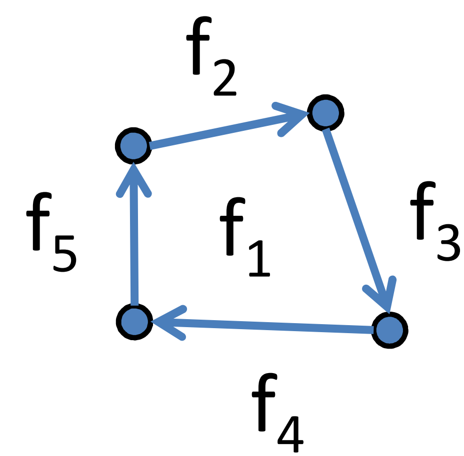
- Goal: find **0-form τ** and 2-form η.
- d: Integration of the edges of a vertex
- $d\omega = d\alpha + \cancel{d\beta} + \cancel{d\gamma}$
- $\sum w(e) = \sum_{(u, v)} \tau(u) - \tau(v)$
- $\tau(u) = [\sum w(e) + \sum_{(u, v)} \tau(v)]/d(u)$

- Initialize all $\tau(u) = 0$
- Run gossip with neighbors.



Gossip-style Implementation

- Goal: find 0-form τ and **2-form η**.
- δ: Integration along a face f
- $\delta\omega = \cancel{\delta\alpha} + \delta\beta + \cancel{\delta\gamma} = \delta d\tau$
- $\sum_{e \text{ on face } f} w(e) = \sum_i \eta(f) - \eta(f_i)$
- $\eta(f) = [\sum_{e \text{ on } f} w(e) + \sum_i \eta(f_i)]/d(f)$



- Initialize all $\eta(f) = 0$
- Run gossip with neighbors.

Homology Basis

- Harmonic 1-forms form a linear space of dim k, for **k holes**, or **2g** for a closed surface with **genus g**.
- Linear dependency can be checked locally.

- **Homology signature** of a trajectory: k-vector integration along k harmonic 1-forms.

Part II: Traffic Pattern Query

Differential Privacy

- **D** is the input dataset and **D'** differs from **D** by only one element **x**.
- A randomized mechanism M is **ε-differentially private** if for any S (set of output)

$$\Pr[M(D) \text{ in } S] \leq e^{\varepsilon} \Pr[M(D') \text{ in } S]$$

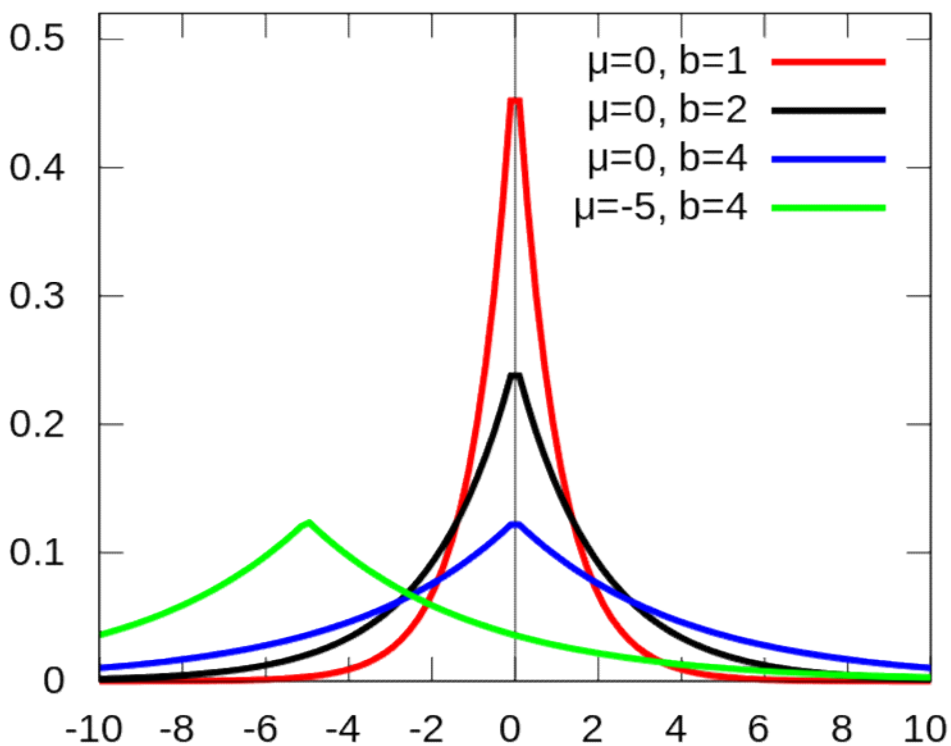
Counting Query

- Given a database of medical records.
- Q1: How many patients have disease y?
- Q2: How many patients, whose name are not x, have disease y?
- Add Laplace noise Lap(1/ε) ➔

$$\Pr[Q1=z]/\Pr[Q2=z] \leq e^{\varepsilon}.$$

Laplace Noise

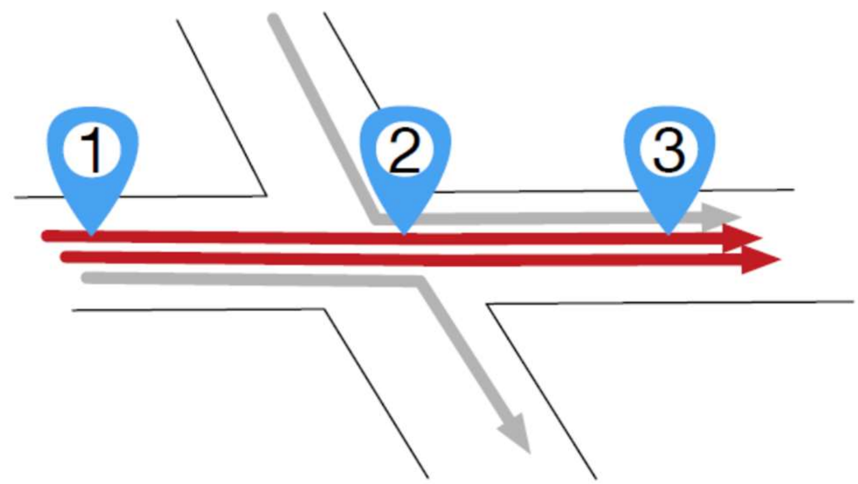
$$\mathrm{Lap}(x|b) = \frac{1}{2b}\exp\left(-\frac{|x|}{b}\right)$$



Popular Paths

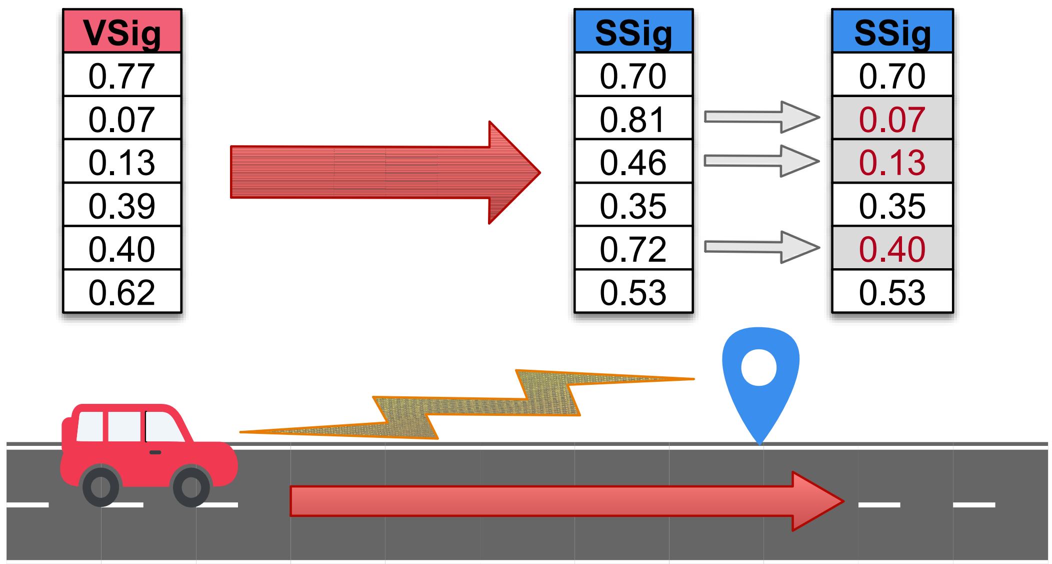
- A path travelled by ϕ-fraction of all vehicles that.

- A subpath of a popular path is still popular;
- A node stays on at most **1/ϕ** maximally popular paths.



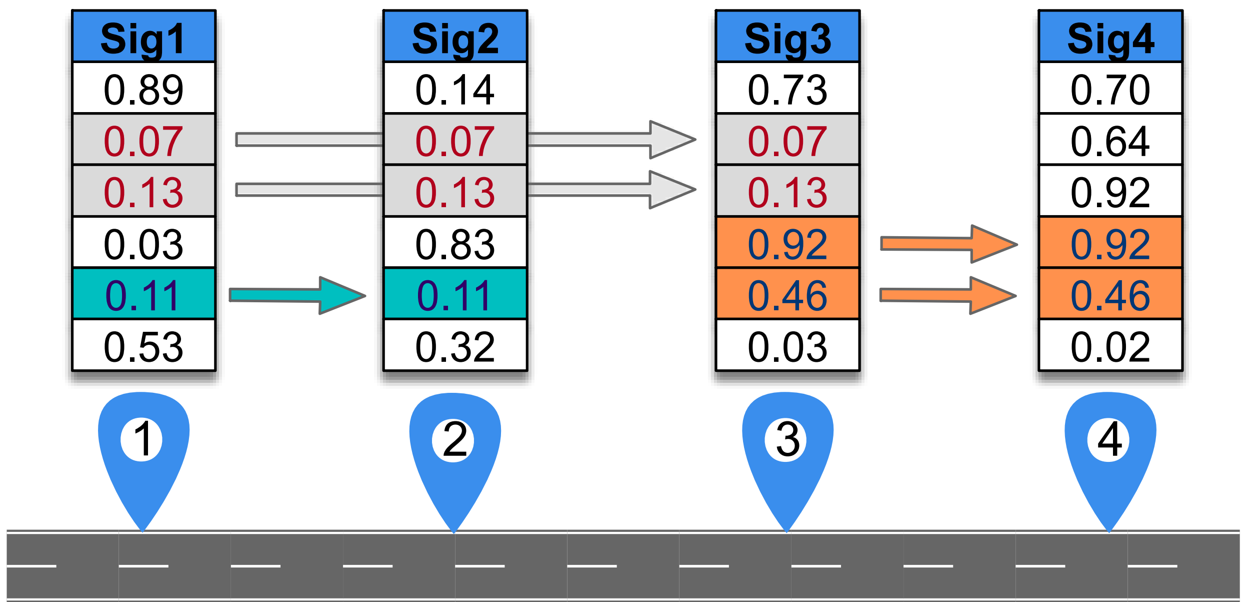
MinHash Signature

- Each node stores the MinHash of vehicles it has met so far.



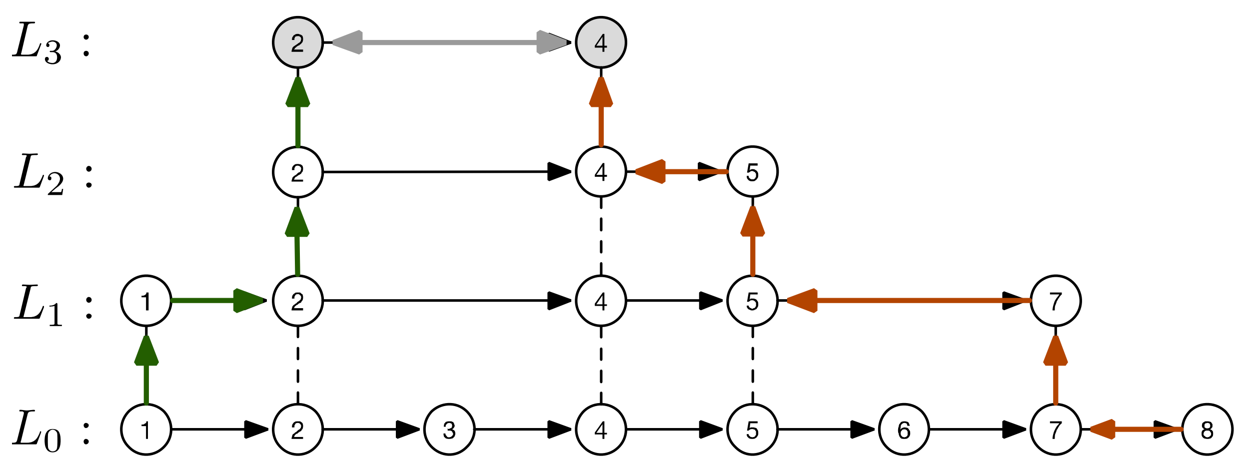
MinHash Signature

- Differentially private for dense traffic.
- # common MinHash entries along a path estimates path popularity.

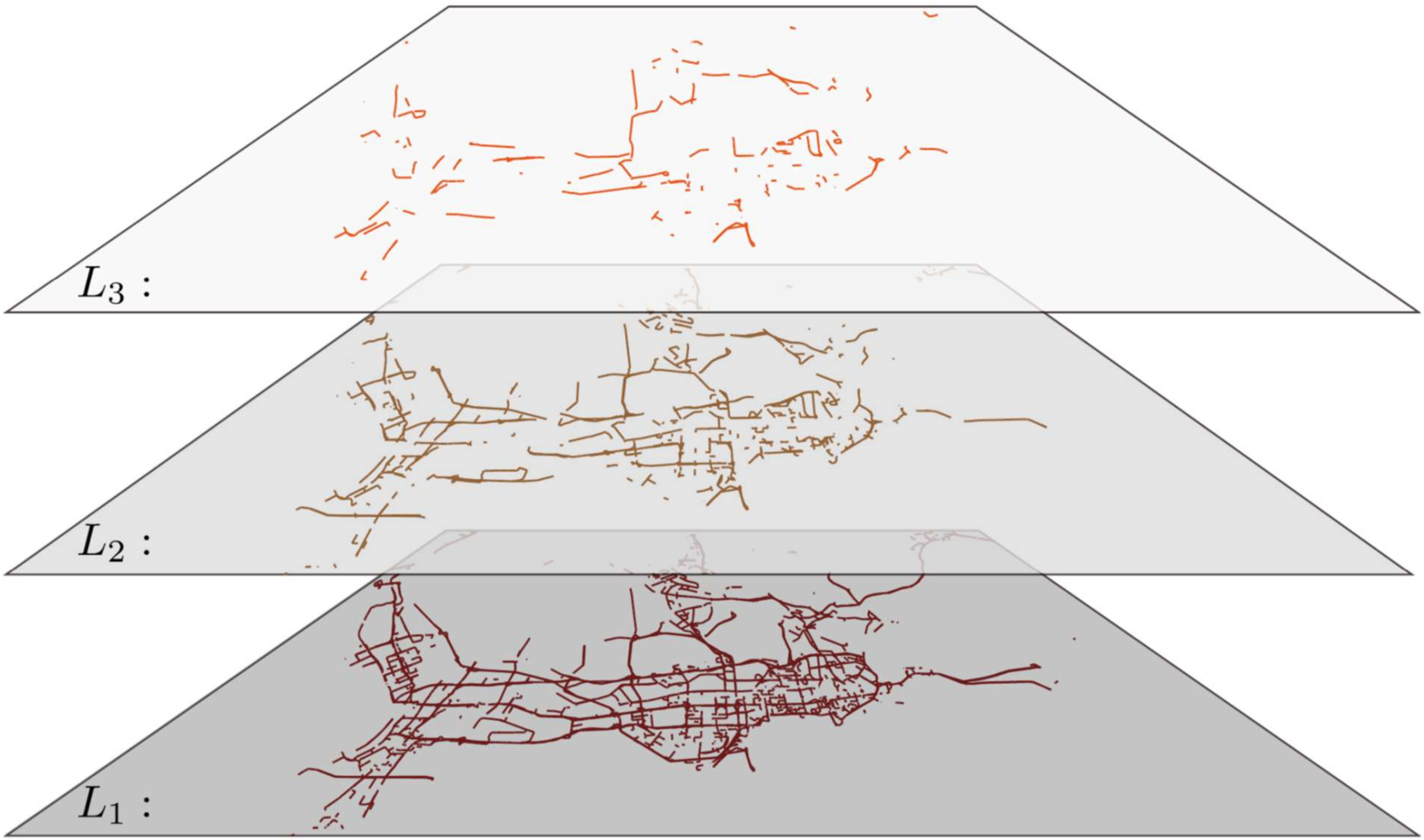


MinHash Hierarchy

- Recursively subsample checkpoint.
- Edge (u, v): if there is at least one consistent path from u to v



Traffic Pattern in a City



Traffic Pattern Queries

- By careful search in the hierarchy of m nodes.
  - Popular paths for (s, t) – $O(\log m)$
  - Popularity for a path P. – $O(\log m)$
  - All popular paths from s. - $O(\log^2 m)$

Summary

- Sensing with privacy consideration.
- Reduced communication cost.

Questions & Comments?

- http://www.cs.stonybrook.edu/~jgao
- Xiaotian Yin, Chien Chun Ni, Jiaxin Ding, Jie Gao, Xianfeng David Gu, **Decentralized Path Homotopy Detection Using Hodge Decomposition in Sensor Networks**, SigSpatial'15.

- Jiaxin Ding, Chien Chun Ni, Mengyu Zhou, Jie Gao, **MinHash Hierarchy for Privacy Preserving Trajectory Sensing and Query**, IPSN'17.