电子科技大学
University of Electronic Science and Technology of China

Similarity and Outlier on Spatial Temporal Data

Spatial temporal data group

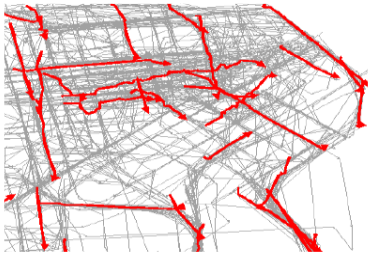
Data Mining Lab,
Big Data Research Center, UESTC
Email： ruizhiwuuestc@gmail.com

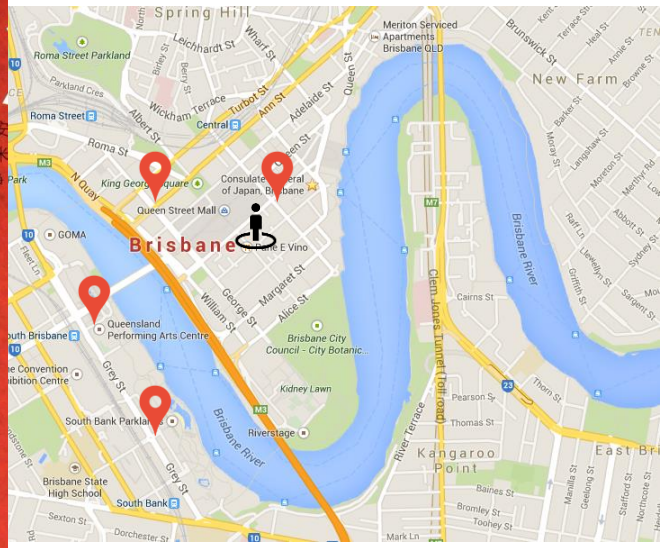
Similarity and Outlier

数据挖掘实验室
Data Mining Lab

- Similarity
- Outlier

Why choose this topic?

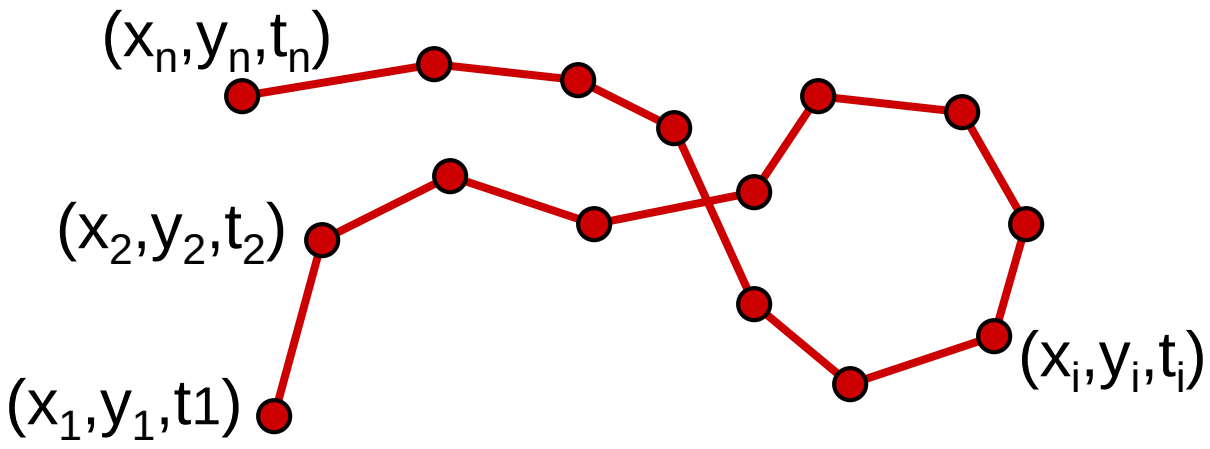




Similarity and Outlier

ID	Timestamp	GPS
“Peter”	2010-04-02 13:12	37.5, -122.5
“Peter”	2010-04-02 15:22	37.2, -123.5
…	…	…

ID	Timestamp	Trajectory
“Peter”	2010-04-02 13:12	<37.5, -122.5>,<37.5, -123.5>…



Similarity and Outlier

- Same time stamp

$$|D| = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

- Different time stamp

$$|D| = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (t_1 - t_2)^2}$$

Similarity and Outlier

Example1

You have entered your facebook in your dorm, half hour ago.

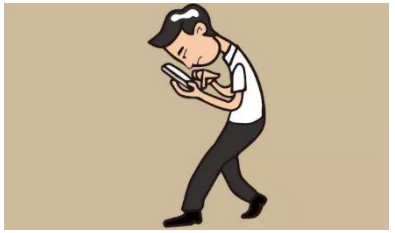
You have entered your facebook in your dorm, now.

Example2

You have entered your facebook in your dorm, half hour ago.

You have entered your facebook in your lab, now.

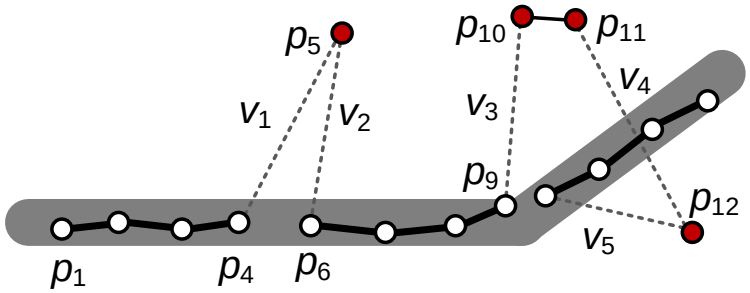
$$|D| = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (t_1 - t_2)^2}$$



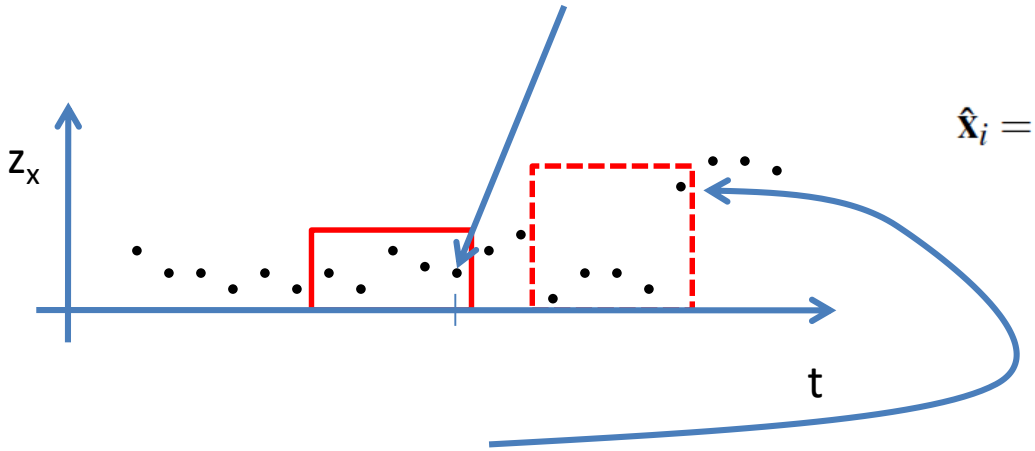
$$|D_2| > |D_1|$$

Anomaly Point ?

Two locations were far away or exceeded a threshold value in a very short period of time .



Filtered version of this point is mean of points in solid box



$$\hat{\mathbf{x}}_i = \frac{1}{n} \sum_{j=i-n+1}^{i} \mathbf{z}_j$$

The Kalman filters

1) Generate P particles

$x_i^{(j)}, j = 1,2,..P$

3) Importance weights

$\omega_i^{(j)} = P(z_i|\hat{x}_i^{(j)})$

5) Compute a weight sum

$\hat{x}_i = \sum_{i=1}^{P} \omega_i^{(j)} \hat{x}_i^{(j)}$

2) Importance sampling

$P(x_i|x_{i-1})$

4) Selection step

$\omega_i^{(j)}$ (normalized)

Heuristics-Based Outlier Detection

Insight: the number of noise points is much smaller than common points

Calculates the travel speed of each point

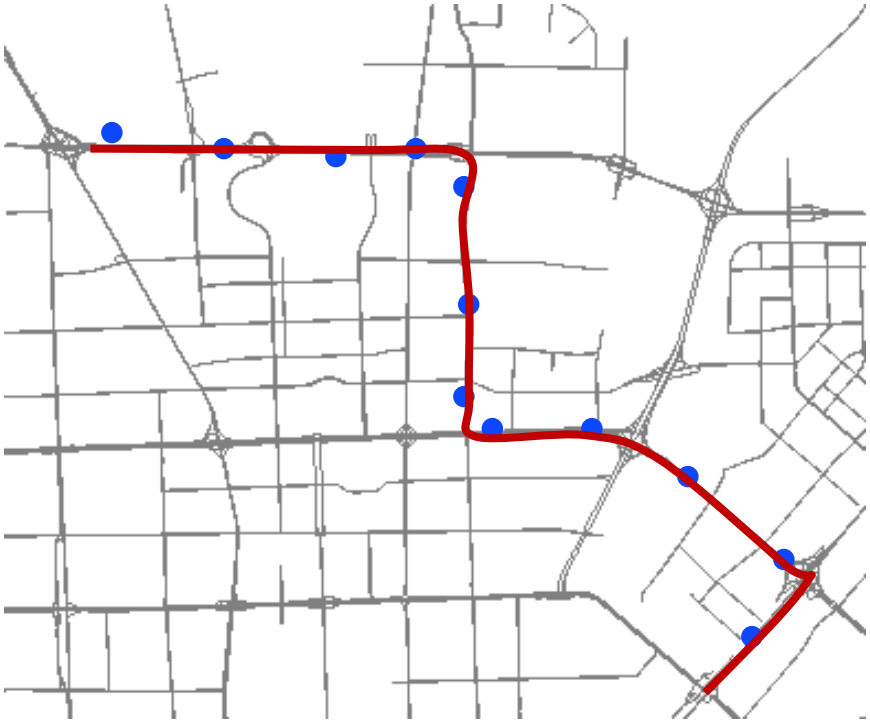
The segments with a speed larger than a threshold are cut off

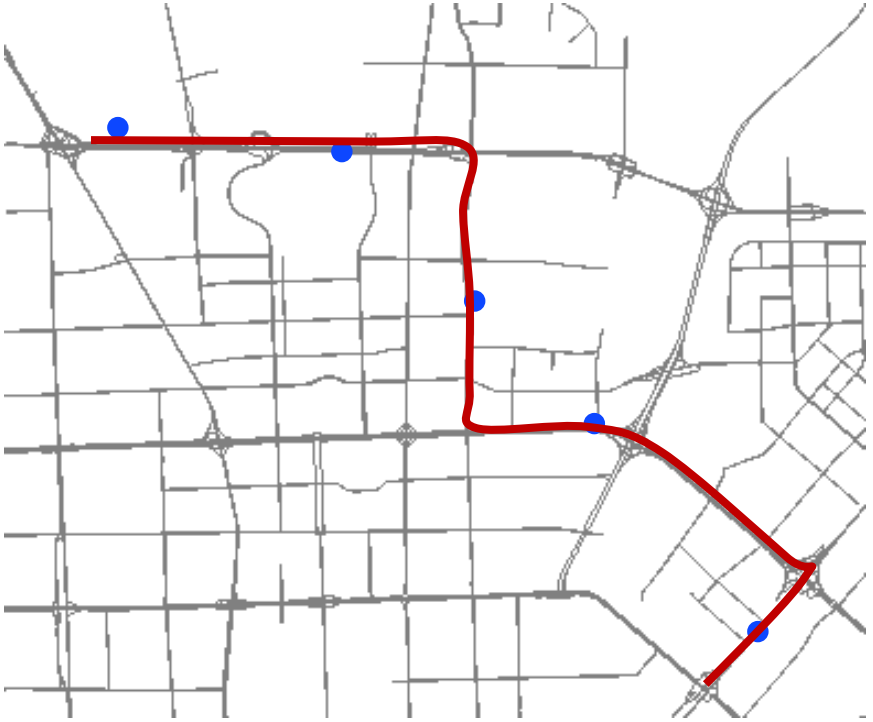
Distance-based outlier detection

Similarity and Outlier

Trajectory

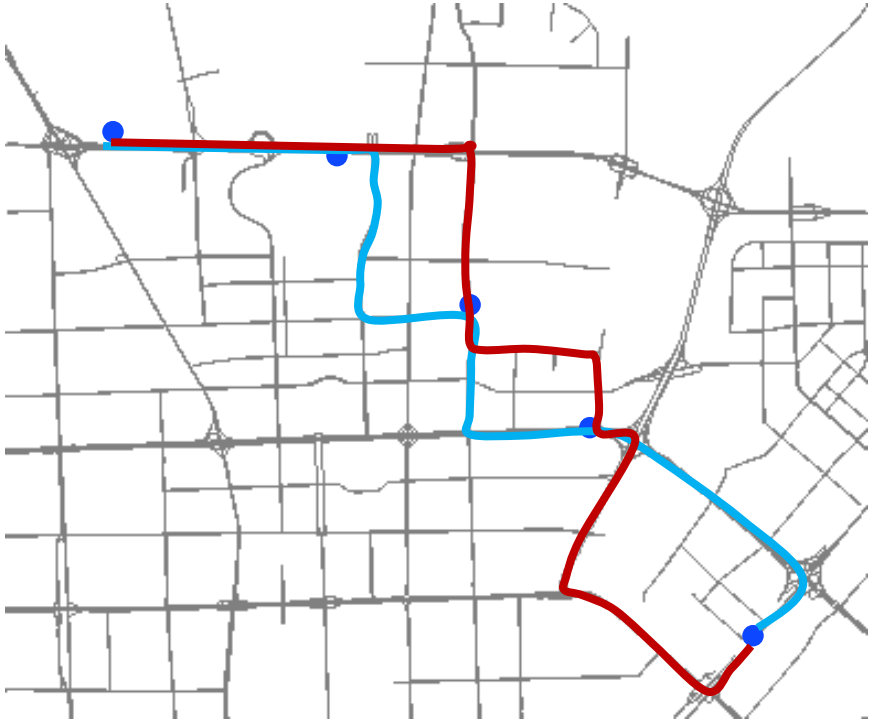
Same time period





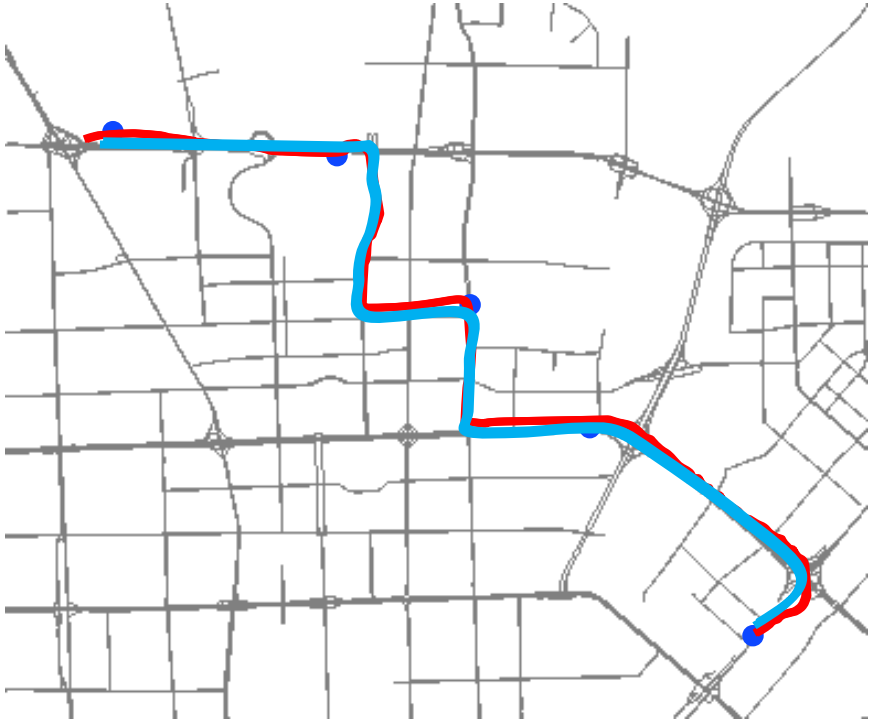
Similarity and Outlier

Same time period



Similarity and Outlier

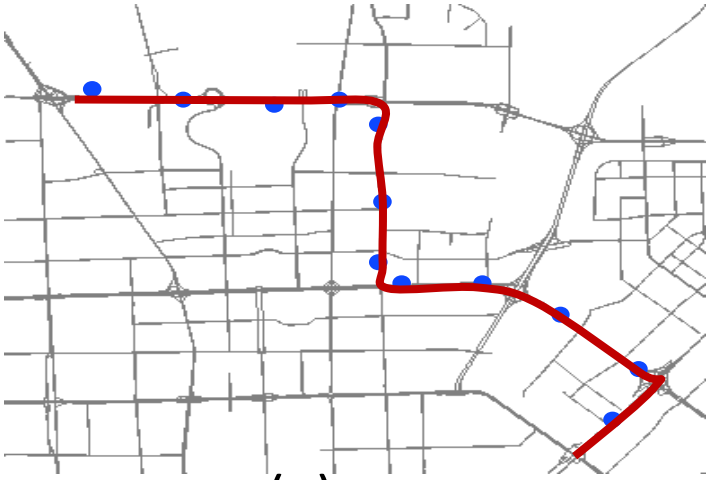
Same time period



Same?

Similarity and Outlier

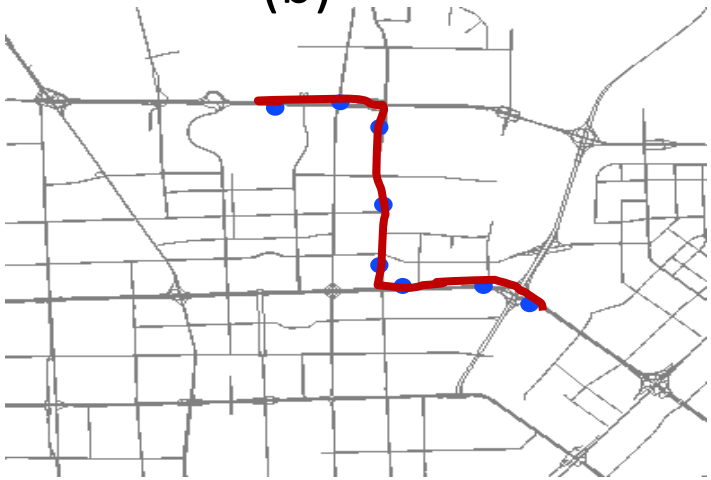
Same time period



(a)

(b)

(c)



(d)

Similarity and Outlier

Why difficulty?

Problem 1. Different sampling rate because of different device

5 minutes sample rate of a GPS device on taxi; 1minutes sample rate Of a GPS device on a privacy car.

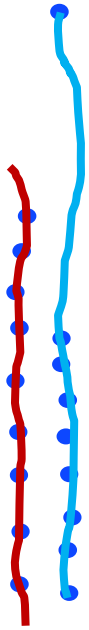
Problem 2. Different path on low sample rate

Problem 3. Direction

Problem 4. Different length

Similarity and Outlier

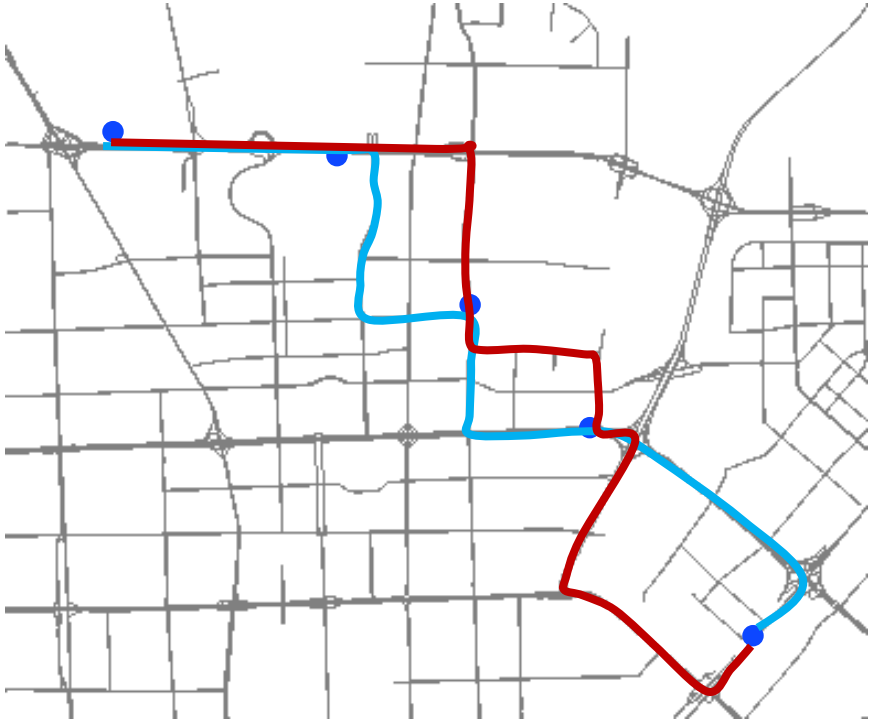
Euclidean distance?



Closest-Pair Distance: $CPD(A,B) = min_{p \in A, p' \in B} D(p,p')$

Similarity and Outlier

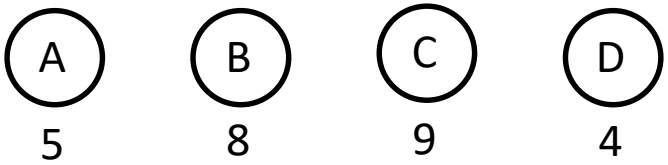
Same time period



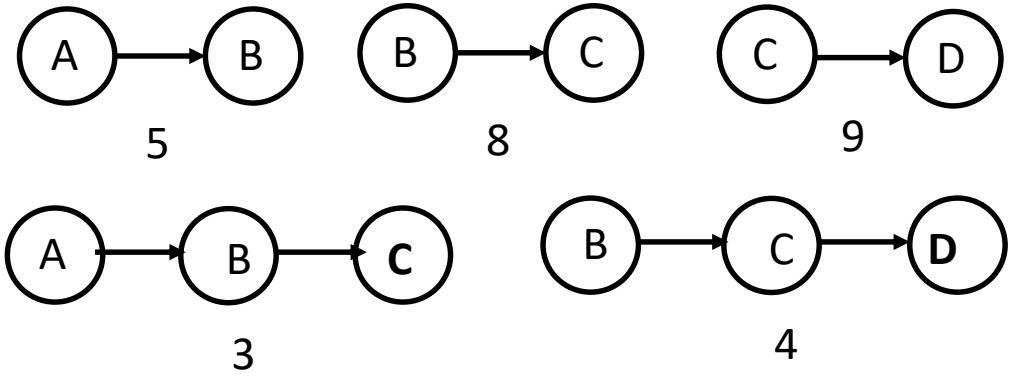
Similarity and Outlier

Trajectory convert to location set {A,B,C,D…..G}.
Trajectory similarity convert to measure set similarity.

- We measure the distribution of two set.

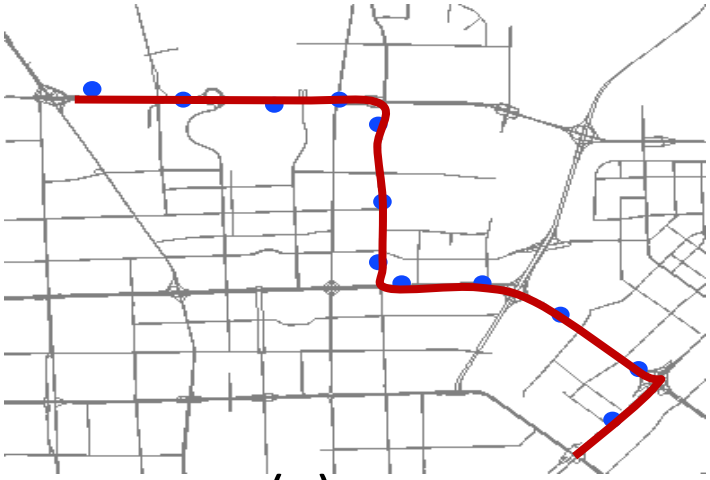


- Building two items or three items in set measures the distribution or common items.



Similarity and Outlier

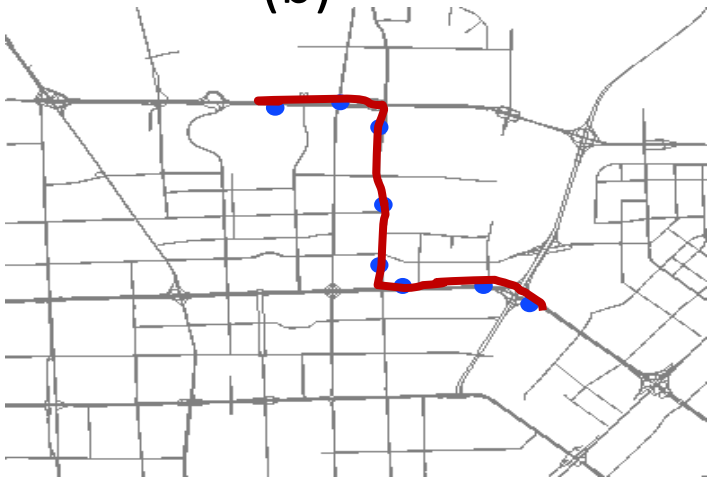
Same time period



(a)

(b)

(c)



(d)

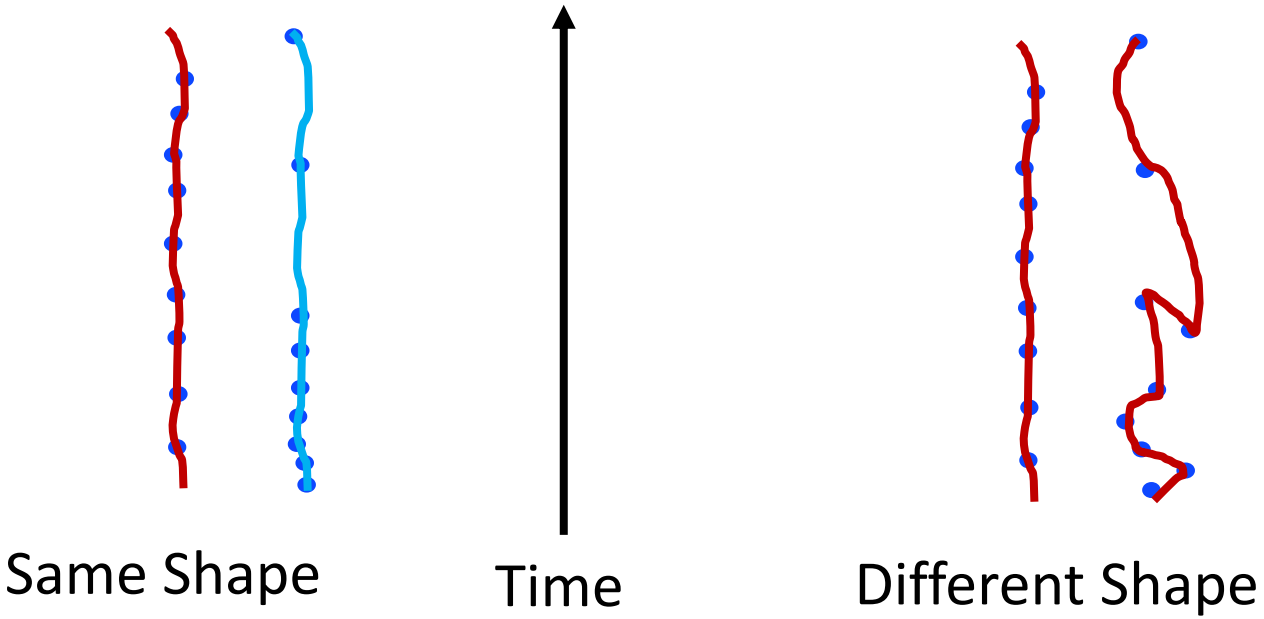
Longest common subsequence (LCSS)

skip some noise points when calculating the distance
A threshold ε is used to determine whether two points are matched

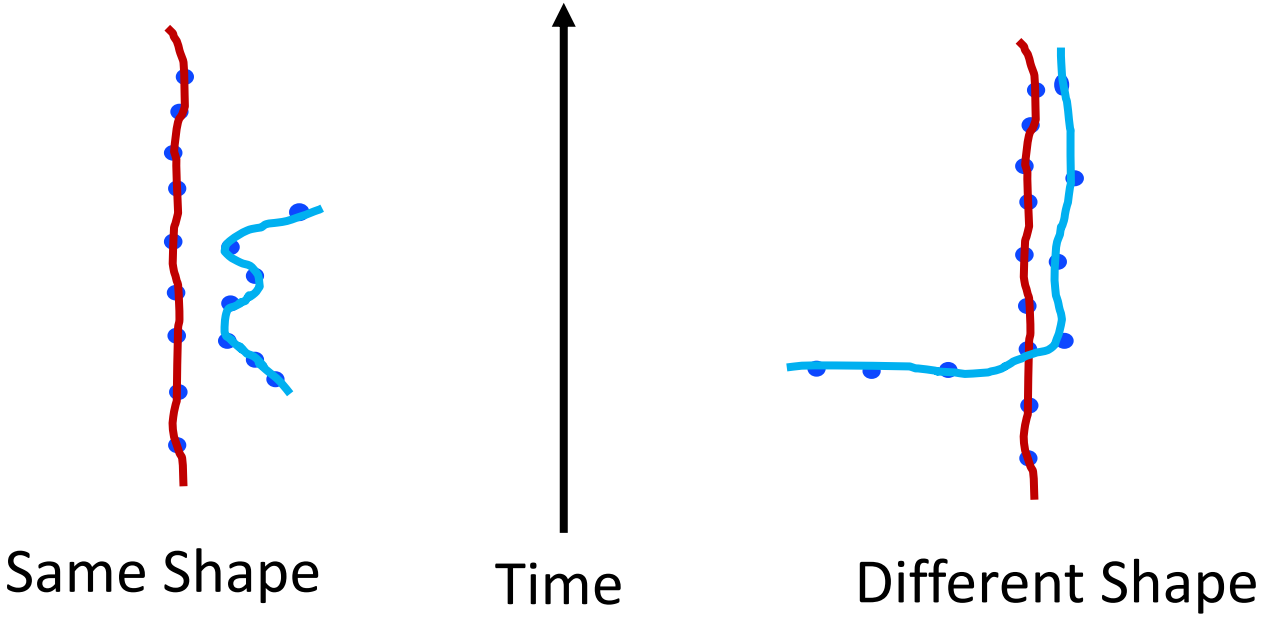
$$LCSS_{\delta,\varepsilon}(T_A, T_B) = \begin{cases} 0, & \text{if } T_A \text{ or } T_B \text{ is empty} \\ 1 + LCSS_{\delta,\varepsilon}(Head(T_A), Head(T_B)), & \text{if } |m-k| \le \delta \text{ and } |a_{m,1} - b_{k,1}| \le \varepsilon \\ & \text{and } \ldots \text{ and } |a_{m,n} - b_{k,n}| \le \varepsilon \\ max(LCSS_{\delta,\varepsilon}(Head(T_A), T_B), & \\ \quad LCSS_{\delta,\varepsilon}(T_A, Head(T_B)), & otherwise, \end{cases}$$

Count common subsequence

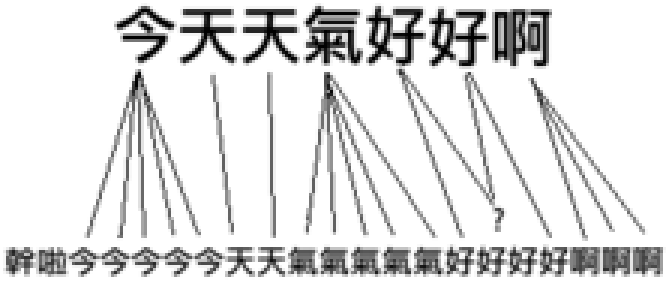
Similarity and Outlier



Similarity and Outlier



Dynamic time warping (DTW)

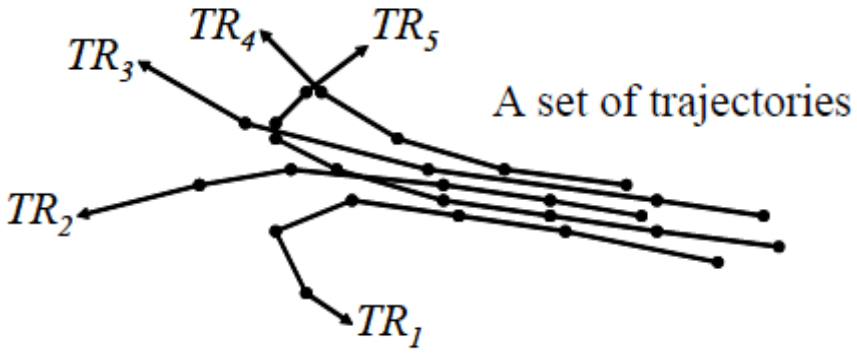


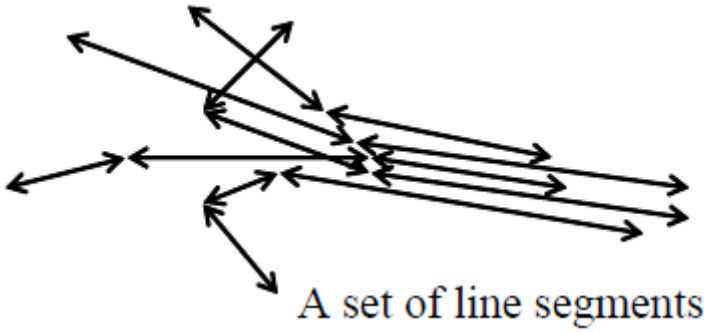
Edit distance

$$EDR_{\varepsilon}(T_A, T_B) = \begin{cases} k, & \text{if } m = 0 \\ m, & \text{if } k = 0 \\ min(EDR_{\varepsilon}(Rest(T_A), Rest(T_B)) + subcost, & \\ \quad EDR_{\varepsilon}(Rest(T_A), T_B) + 1, & otherwise \\ \quad EDR_{\varepsilon}(T_A, Rest(T_B)) + 1), & \end{cases}$$

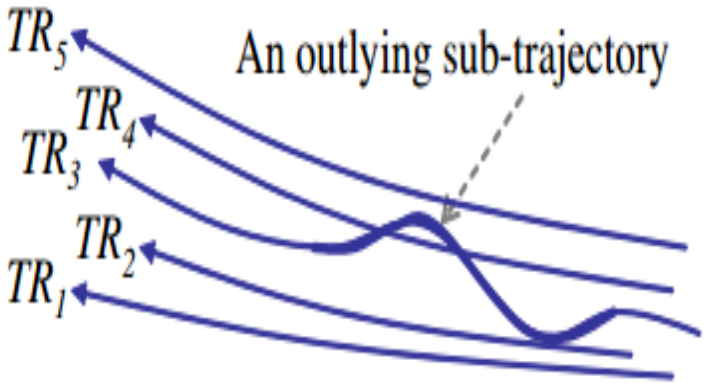
Similarity and Outlier

Same?





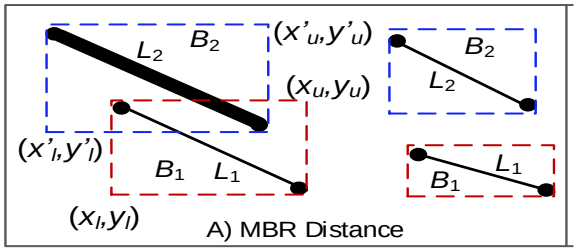
Partition?



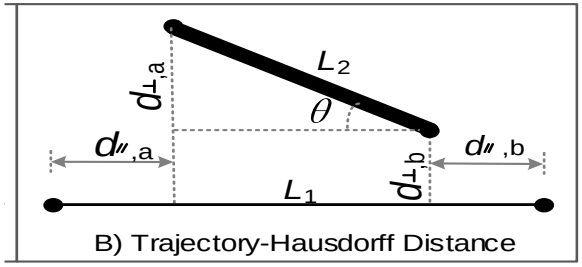
Outlier

Similarity and Outlier

- The distance between two trajectory segments
  - the Minimum Bounding Rectangle (MBR)-based
    - $\sqrt{(\Delta([x_l, x_u], [x'_l, x'_u]))^2 + (\Delta([y_l, y_u], [y'_l, y'_u]))^2}$
    - $\Delta([x_l, x_u], [x'_l, x'_u]) = \begin{cases} 0 & [x_l, x_u] \cap [x'_l, x'_u] \neq \emptyset \\ x'_l - x_u & x'_l > x_u \\ x_l - x'_u & x_l > x'_u \end{cases}$



  - Trajectory-Hausdorff Distance
    - The aggregate perpendicular distance ($d_\perp$)
    - The aggregate parallel distance ($d_{//}$)
    - The angular distance (d_θ)
    - $D_{Haus} = w_1 d_\perp + w_2 d_{//} + w_3 d_\theta$



Similarity and Outlier

➢ Trajectory feature representation

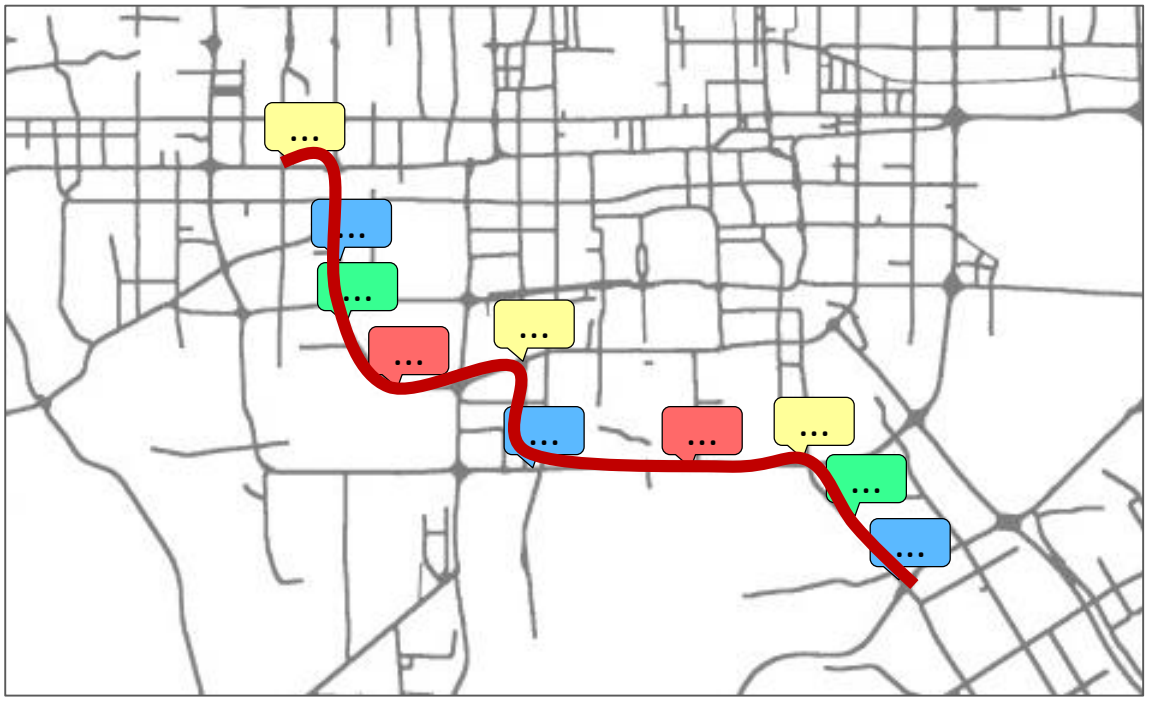
Trajectories: Overall Characteristics

1. Geometric shape
2. Length (traveled distance)
3. Duration (in time)
4. Speed
   - Mean, median, and maximal Speed
   - Periods of constant speed, acceleration, deceleration
5. Direction:
   - Periods of straight, curvilinear, circular movement;
   - Major turns ('turning points') in: time, position, angle, initial and final directions, and speed in the moment of the turn;

Anything else?

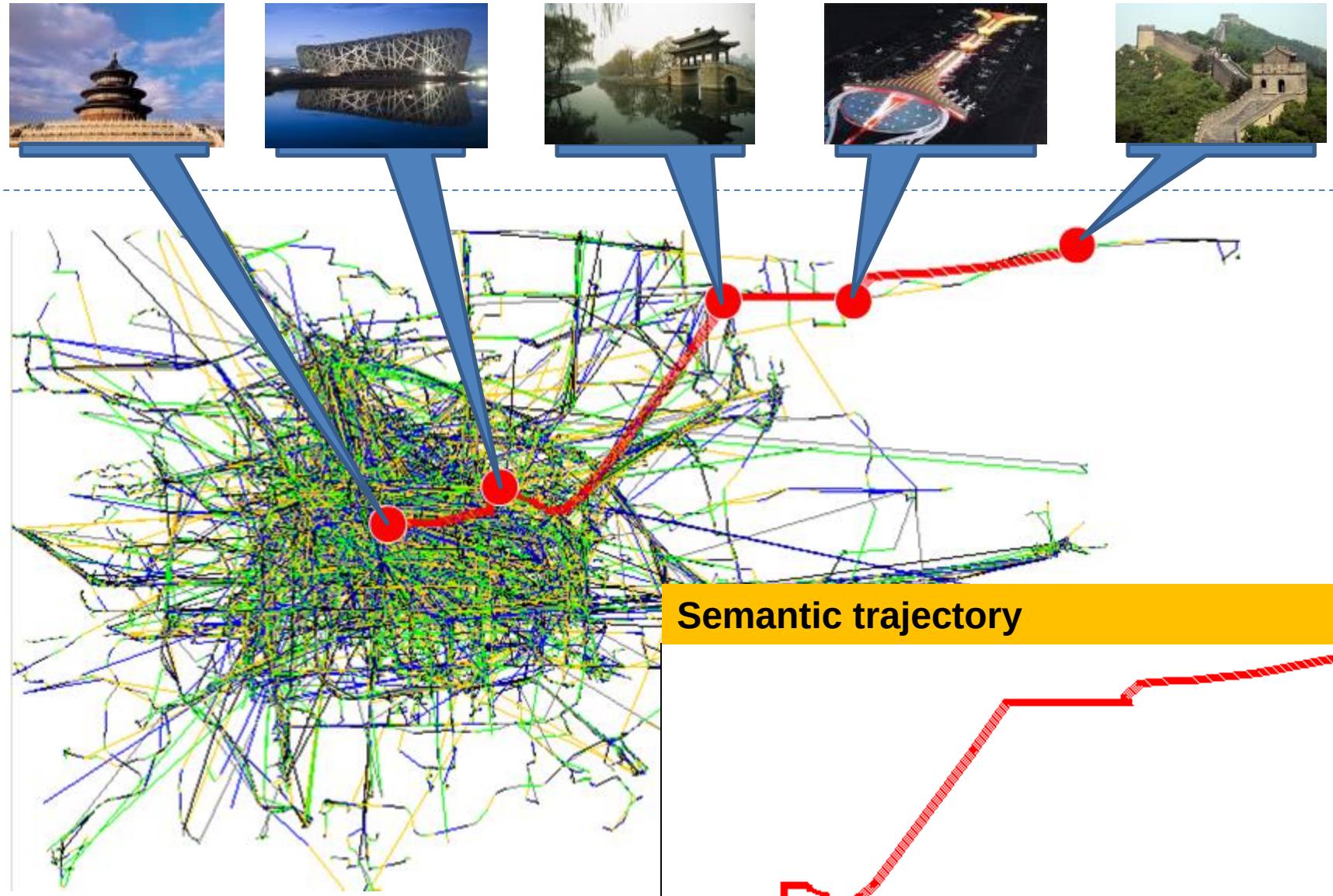
Similarity and Outlier

POIs based method



Similarity and Outlier

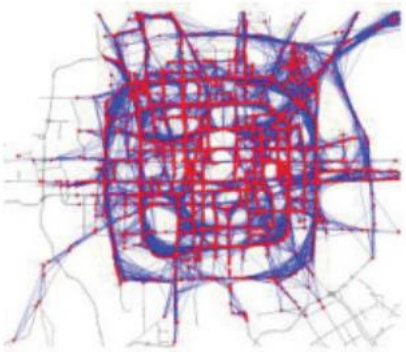
Previous information based



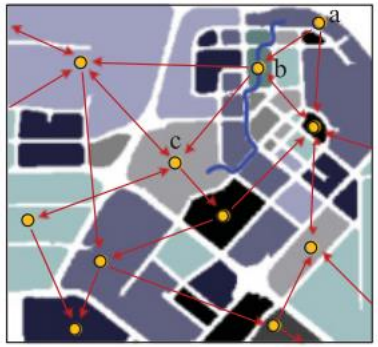
Similarity and Outlier

- Others methods representation

Graph



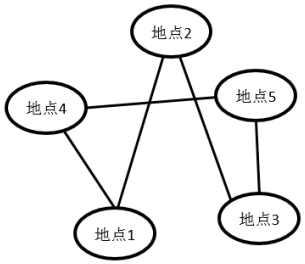
(a) A landmark graph



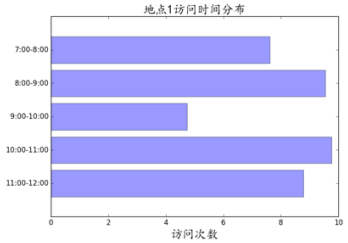
(b) A region graph

Network

用户轨迹:<地点3，6:30>，<地点3，7:00>，……<地点2，7:30>，<地点3，8:00>，<地点1，8:45>，<地点1，9:00>，……<地点1，10:30>，<地点5，11:30>，<地点4，12:30>，<地点3，14:30>，

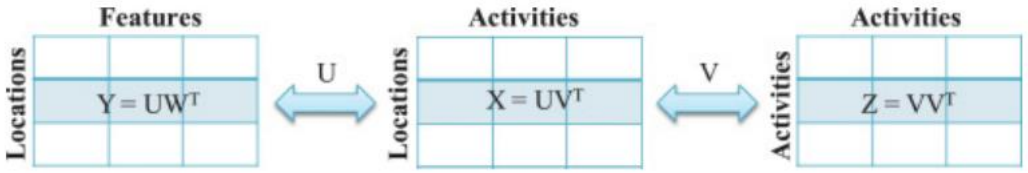


用户轨迹网络

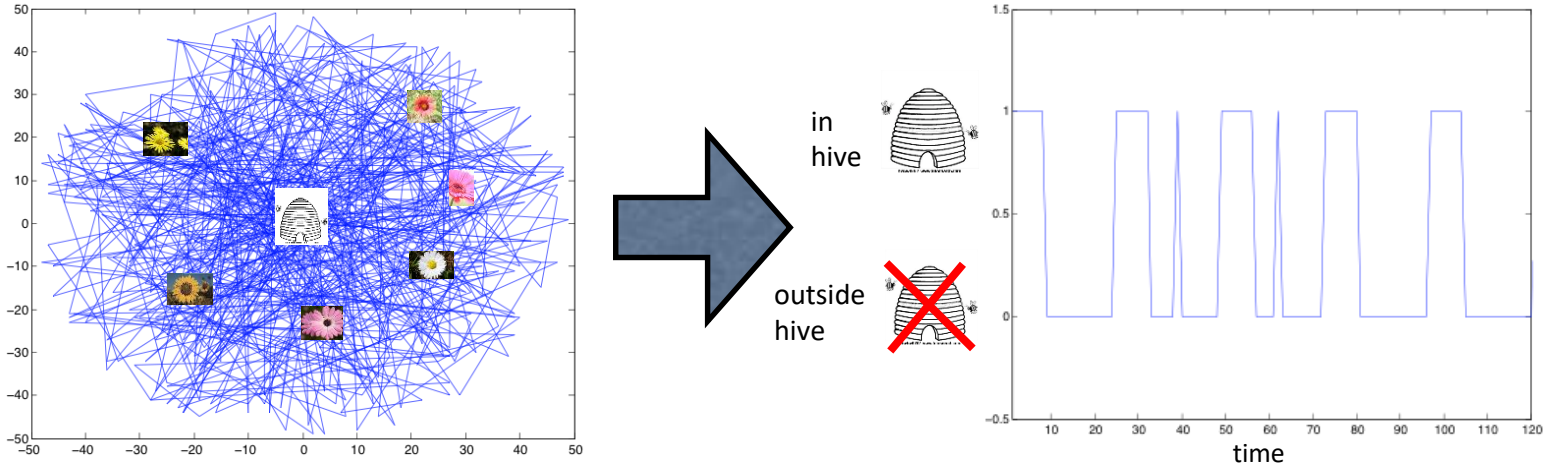


节点：地点1所存储时间分布信息

Matrix



Similarity and Outlier



- We can observe its movement from the hive (in or out).

Outlier is different from pattern.

Thanks

By R. Wu