社交媒体中的时空轨迹模式挖掘

Trajectory Pattern Mining in Social Media

Wen-Chih Peng (彭文志)
Dept. of Computer Science
National Chiao Tung University
wcpeng@cs.nctu.edu.tw

Social media is everywhere

* Essential entities
  * Users
  * Locations
  * User check in location

photo

title

description

comment

address, phone and link

times of check-in

map

check-in of friend and comment

nearby landmark

高雄國際機場

讚

機場 · Kaohsiung, Taiwan

編輯

資訊

朋友的動態 (1)

維基百科

1,906
個讚

36,731
個打卡次

另外在：

Gowalla

你是此地負責人嗎？

你認識此地的負責人嗎？

建立粉絲專頁

檢舉粉絲專頁

分享

資料

編輯

簡介

高雄國際機場（；IATA代碼：**KHH**；ICAO代碼：**RCKH**），是臺灣一座位於高雄市小港區的中型商用機場。高雄國際機場為台灣南部唯一的聯外國際機場與主要的國際貨運、旅客出入吞吐地之一，是臺灣的第二大機場，並做為桃園國際機場的備用降落場。其管理及營運單位為中華民國交通部民用航空局**高雄國際航空站**。因位於高雄市小港區，故高雄國際機場也通稱為**小港機場**或**高雄小港機場**。

地址 Kaohsiung, Taiwan · 路線

聯絡電話 新增電話號碼

網址 http://www.kia.gov.tw/

朋友的動態

顯示全部

Mick Yu-Min Lin 和 Kao Alfred 與 Karban Feynman 一起來過這裡。

讚 · 留言 · 6月4日 16:50 來自手機 ·

Sj Yen 、 Karban Feynman 和其他 3 人都說讚。

查看全部 4 則留言

Adrian Huang 做高鐵應該比坐飛機快吧XD~
6月4日 17:09 · 讚

Sj Yen 剛才電話中叫我快點從高雄搭高鐵去桃園集合是怎麼一回事XD
6月4日 17:28 傳自手機 · 讚

留言……

附近地標

建立粉絲專頁 · 顯示全部

美麗島車站
大眾運輸
10,226 個打卡次 · 851 個讚

星巴克starbucks
咖啡廳
4,854 個打卡次 · 511 個讚

編輯

推薦人數 (1)

查看全部

楊豐光 在我們高雄的機場!!!!!!!!!
1

寫一段推薦評論……

動態贊助

顯示全部

簡俊仁覺得 Hennessy Artistry Taiwan 很讚。

Hennessy Artistry Taiwan
讚

贊助

顯示全部

打我阿笨蛋！

憤怒豬集體嗆聲
好玩才敢大聲說
過關簡單、升級容易
5歲小孩都會玩
【點擊創角挑戰】

《明朝時代 2》征戰明朝

讓你親身感受叛亂、侵略、守城、衛國、無恥、不屈、狡詐、氣節！

巨星小S陪你打麻將

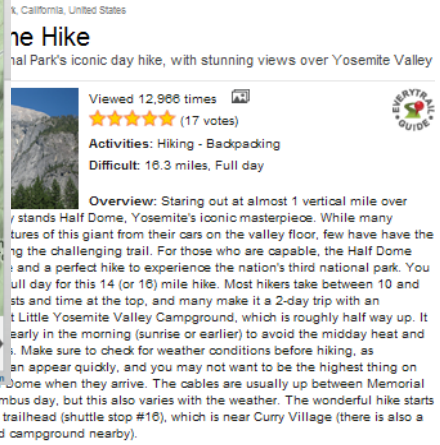
下課(班)十分鐘打麻將
綜藝天后小S陪你打麻將邀你無聊Out~！！
【免下載、立即玩】
咖緊來¢新明星3缺1

陳貽婷曾玩過麻將—新明星3缺1網頁遊戲 免費的麻將遊戲。

Social Media with Explicit Trajectory Data

- Many Geo-Web sites (i.e., EveryTrail, Bikely) allow users to share their trajectories

Every trail



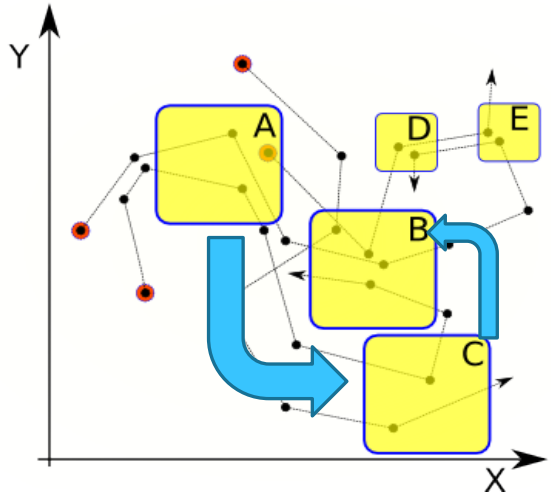
My tracks

Trajectory Types

* Trajectory data with the detailed routes is generated in a high sampling rate.
* Trajectory data from check-in data is viewed as low-sampling trajectories.

Mining Knowledge in Trajectories: Trajectory Pattern

* A trajectory pattern represents the frequent movement behavior, which consists of:
  * **Hot region:** areas with trajectories densely passed by
  * **Relation:** sequential relationships among hot regions



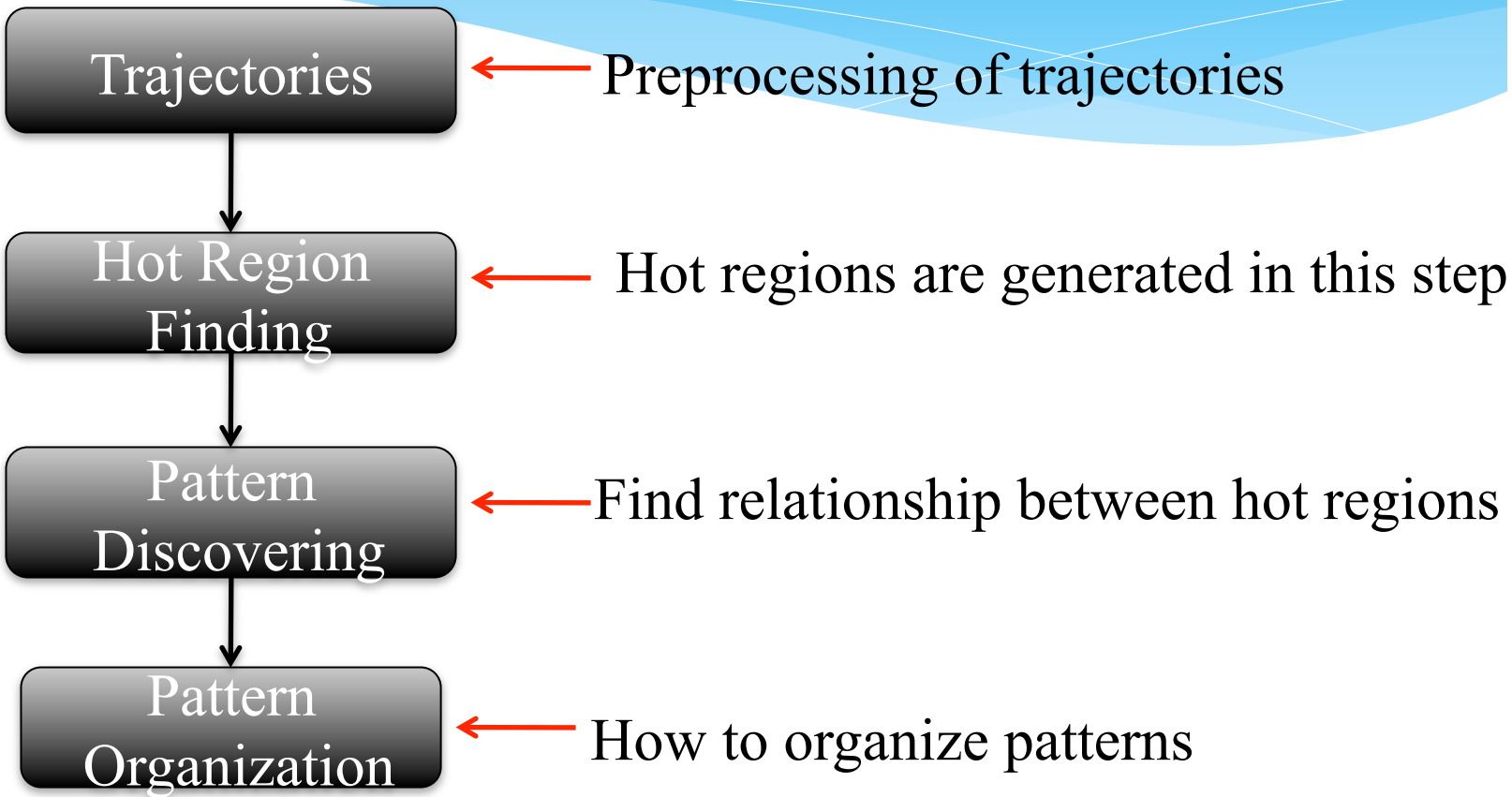
A Trajectory Pattern: A → C → B

6

Trajectory Pattern Mining

* Input: a set of trajectories
* Output: trajectory patterns
* Trajectory patterns:
  * **Hot regions**: areas an object usually stays
  * **Relations:** sequential or temporal relationships among hot regions

A General Framework



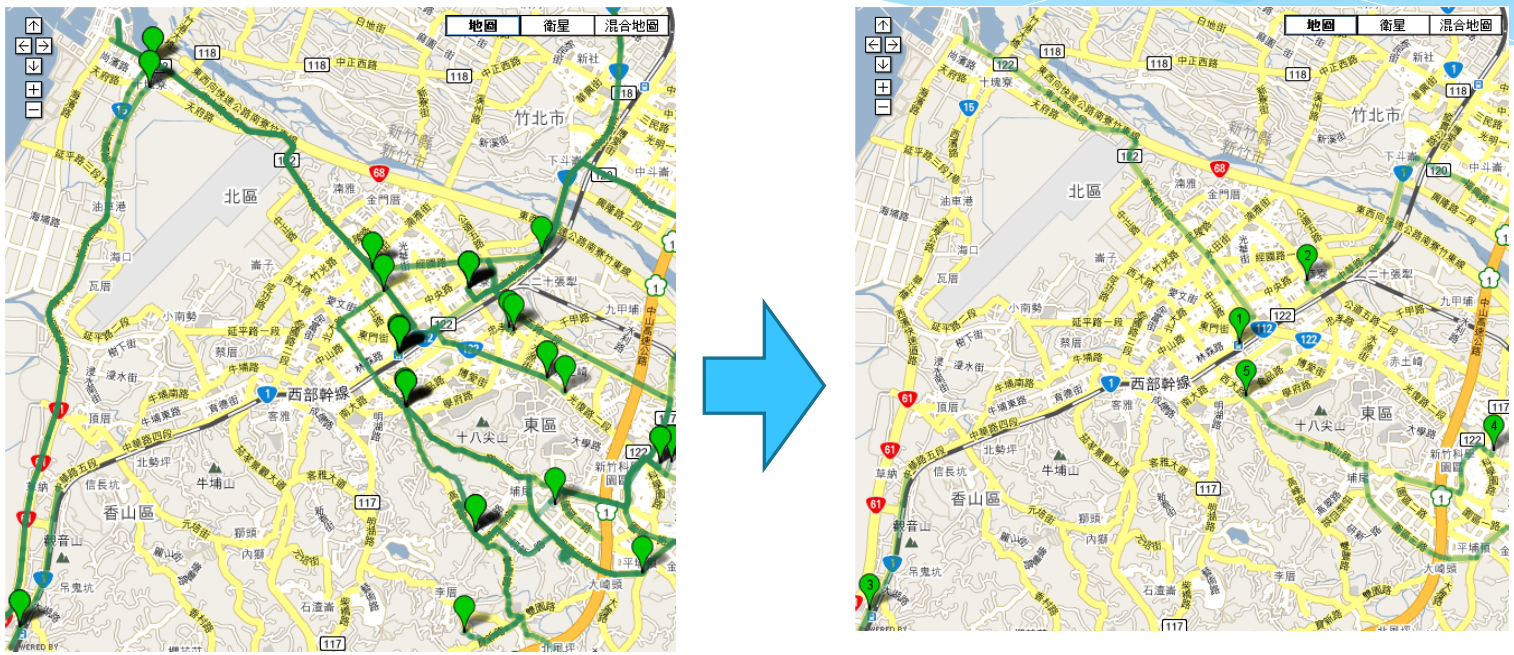
8

Applications of Trajectory Patterns

* Pattern-aware trip planning
* Smart navigation
* Location-based Communities

PATS:基於使用者移動模式之旅遊路徑推薦系統

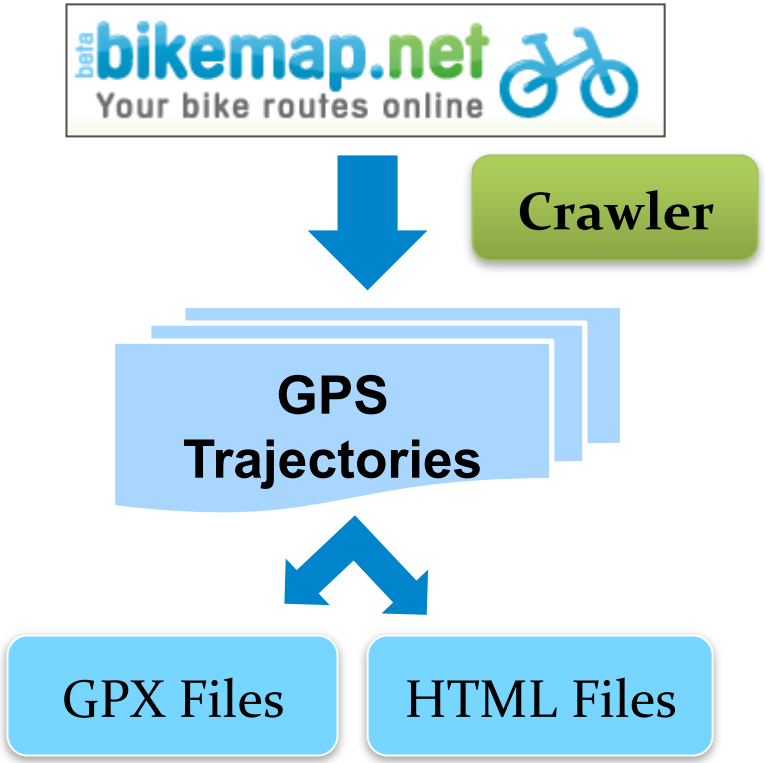
- Input: a spatial range
- Output: top-K trajectories passing the query range



Trajectory Dataset

L.-Y. Wei, W.-C. Peng, B.-C. Chen, and T.-W. Lin, ``PATS: A Framework of Pattern-Aware Trajectory Search," *Proceedings of the 1st Workshop on Uncertain Mobile Data Management and Mining (In conjunction with MDM)*, 2010.
教育部 99學年度大學校院網路通訊軟體與創意應用競賽 手機應用組第三名

Real Dataset



ADSLab05.cs.nctu.edu.tw - PuTTY

```
#! /bin/bash
#from 170-150000
for ((id=150001; id <= 250000 ; id++))
do
  echo -n "$id ";
  filename=$id.gpx
  /usr/bin/wget URL http://www.bikemap.net/route/$id/export.gpx
  cp export.gpx ${filename}
  #echo ${filename};
  rm export.gpx
done
```

```
<?xml version="1.0" encoding="UTF-8"?>
<gpx
 version="1.0"
creator="GPSBabel - http://www.gpsbabel.org"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xmlns="http://www.topografix.com/GPX/1/0"
xsi:schemaLocation="http://www.topografix.com/GPX/1/0 http://www.topografix.com/GPX/1/0/gpx.xsd">
<time>2009-10-14T02:36:06Z</time>
<bounds minlat="24.505800000" minlon="121.086650000" maxlat="24.661400000" maxlon="121.126700000"/>
<trk>
<trkseg>
<trkpt lat="24.661400000" lon="121.121770000">
  <ele>0.000000</ele>
</trkpt>
<trkpt lat="24.660560000" lon="121.122320000">
  <ele>0.000000</ele>
</trkpt>
<trkpt lat="24.660330000" lon="121.122420000">

  <div id="routeninfo">
    <p><strong>建立者:</strong> <a href="/user/Carson">Carson</a></p>
  <div>
    <p><br /><a href="http://www.bikemap.net/events/create?event__route_id=92082">Create event at this route</a></p>
  </div>
  <ul>
    <li><strong>距離:</strong> 35.5 km</li>
    <li><strong>垂直爬升:</strong>
        ca. 240 m</li>
    <li><strong>坡度:</strong> 部分起伏-部分平坦</li>
    <li><strong>路面:</strong> 有鋪面</li>
```

Real Dataset

Crawler

ADSLab05.cs.nctu.edu.tw - PuTTY

```
#! /bin/bash
#from 170-150000
for ((id=150001; id <= 250000 ; id++))
do
 echo -n "$id ";
 filename=$id.gpx
 /usr/bin/wget URL http://www.bikemap.net/route/$id/export.gpx
 cp export.gpx ${filename}
 #echo ${filename};
 rm export.gpx
done
```

GPS Trajectories

GPX Files

HTML Files

Parser

Landmarks

Postgres Database

ADSLab05.cs.nctu.edu.tw - PuTTY

```
<?php
    error_reporting(E_ERROR & ~E_NOTICE & ~E_WARNING );
    //error_reporting(E_ALL);

    set_time_limit(0);

    ini_set("memory_limit",-1);

    include("PostgreDB.class.php");


    function KPLAT_between_Longitude($lat)
    {
            $circle=40075.02;       //赤道週長
            $radius=6378.137;       //地球半徑
            $radian=57.2957795131;  //一弧度幾度

            return $circle*cos(deg2rad($lat))/360;
    }

    function two_point_degree($x1,$y1,$x2,$y2)
    {

            $y_unit=111.222;

            $north_vector_x=0;
            $north_vector_y=1;
            $north_length=sqrt($north_vector_x*$north_vector_x+$north_vector_y*$north_vector_y);

            $in_vector_x=$x2*KPLAT_between_Longitude($y2)-$x1*KPLAT_between_Longitude($y1);
            $in_vextor_y=($y2-$y1)*$y_unit;
            $in_length=sqrt($in_vector_x*$in_vector_x+$in_vextor_y*$in_vextor_y);
            if($in_length==0) $in_length=0.00000001;
            $tmp=($north_vector_x*$in_vector_x+$north_vector_y*$in_vextor_y)/($north_length*$in_l
ength);

            if( ($x2-$x1)>=0 )
                    return rad2deg(acos($tmp));
            else
                    return 360-rad2deg(acos($tmp));
    }
    //x:lon y:lat
```

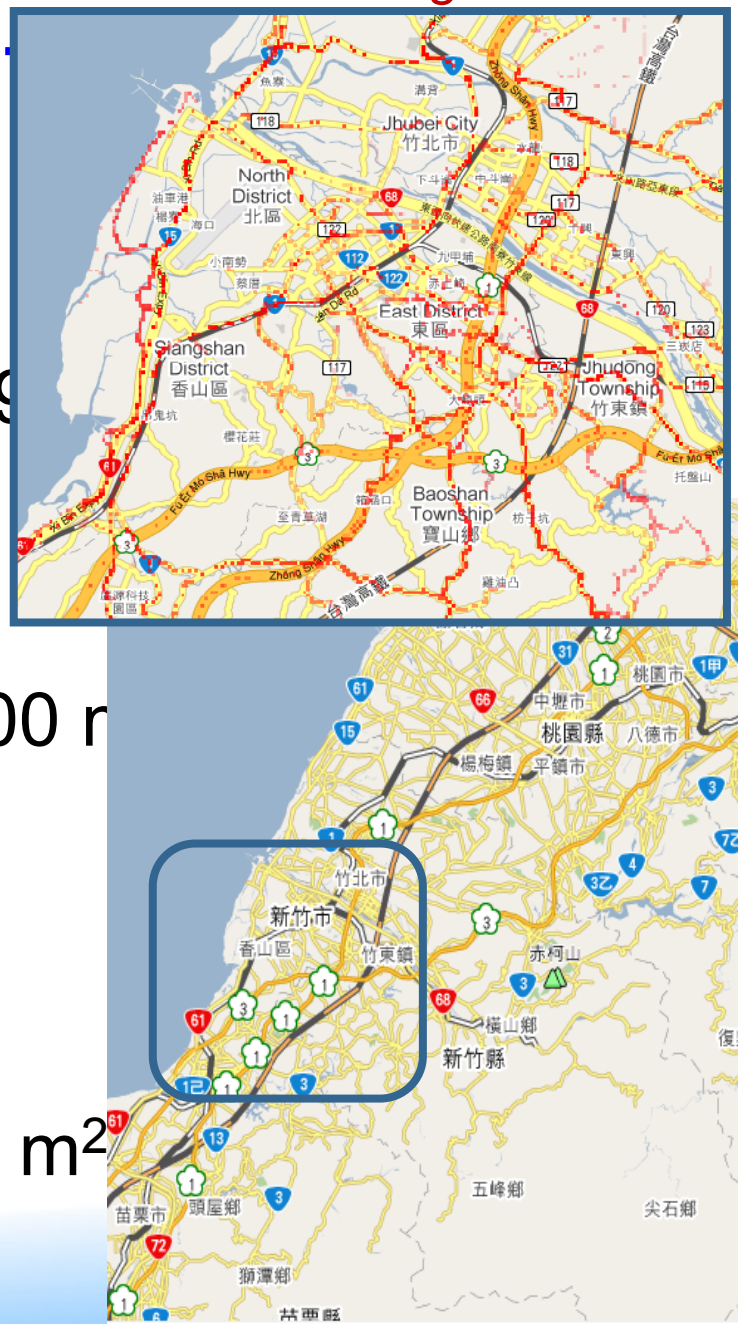
41,2-5 Top

CarWeb 暢行天下!

Trajectory Da

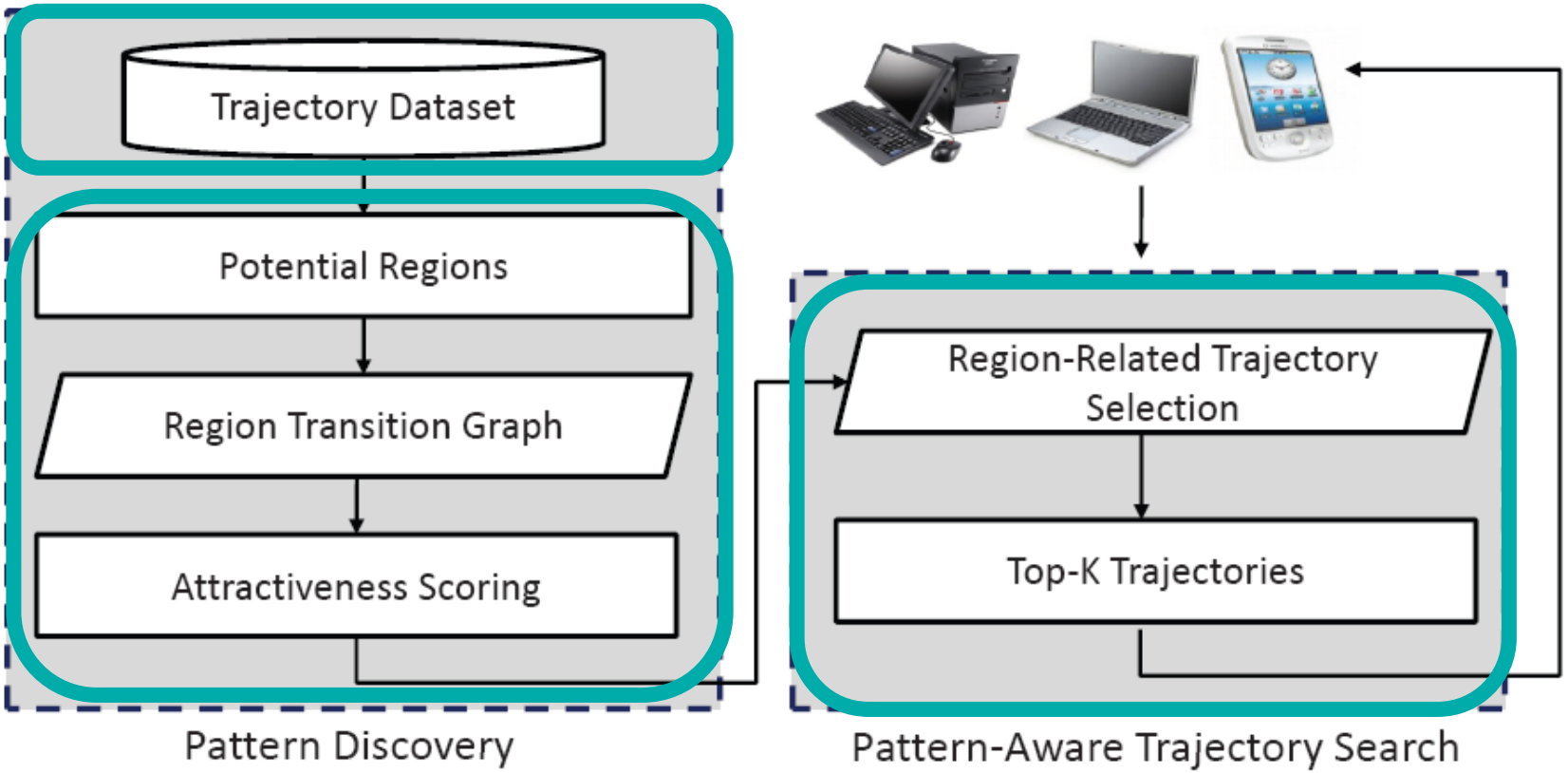
- Taiwan
  - Time: Jun. 2008 ~ Oct. 200
  - 6,548 trajectories
  - 1,301,192 GPS points
  - 7,311,763 regions (100 x 100 r
    - Density >0: 135,829 regions
- Around Hsinchu
  - 1300 trajectories
  - 456,000 regions (100 x 100 m²
    - Density >0: 23,600 regions

Hot Regions



CarWeb 暢行天下!

Framework of PATS

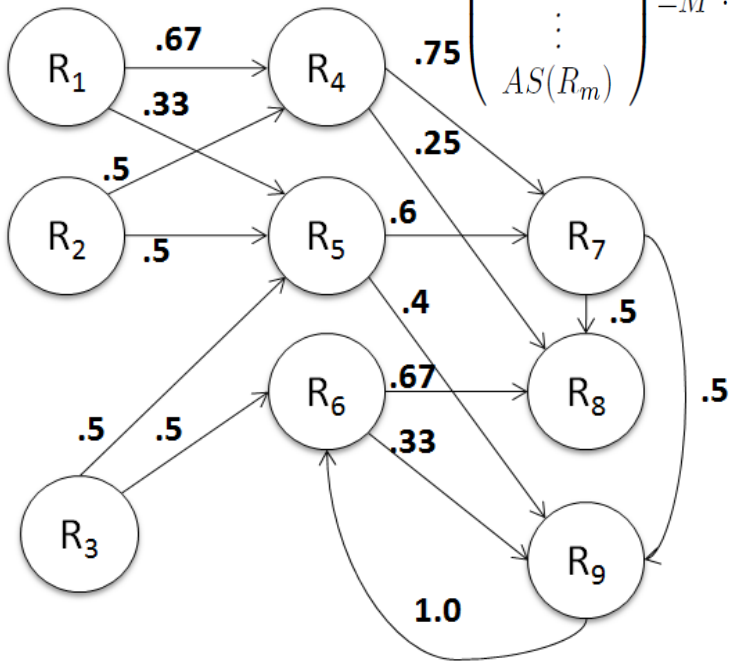


Pattern Discovery

* Propose a random walk based algorithm to score spatial regions

$$M = \begin{pmatrix} 1-\alpha & \alpha \cdot w_{<R_1,R_1>} & \cdots & \alpha \cdot w_{<R_m,R_1>} \\ 1-\alpha & \alpha \cdot w_{<R_1,R_2>} & \cdots & \alpha \cdot w_{<R_m,R_2>} \\ \vdots & \ddots & & \\ 1-\alpha & \alpha \cdot w_{<R_1,R_m>} & \cdots & \alpha \cdot w_{<R_m,R_m>} \end{pmatrix}$$

Tid	A sequence of potential regions
Tra_1	$R_1 \dashrightarrow R_4 \dashrightarrow R_7$
Tra_2	$R_1 \dashrightarrow R_4 \dashrightarrow R_7$
Tra_3	$R_1 \dashrightarrow R_5 \dashrightarrow R_7$
Tra_4	$R_2 \dashrightarrow R_4 \dashrightarrow R_8$
Tra_5	$R_2 \dashrightarrow R_5 \dashrightarrow R_7$
Tra_6	$R_5 \dashrightarrow R_9$
Tra_7	$R_3 \dashrightarrow R_5$
Tra_8	$R_3 \dashrightarrow R_6 \dashrightarrow R_8$
Tra_9	$R_6 \dashrightarrow R_8$
Tra_{10}	$R_6 \dashrightarrow R_9$
Tra_{11}	$R_5 \dashrightarrow R_9 \dashrightarrow R_6$
Tra_{12}	$R_4 \dashrightarrow R_7 \dashrightarrow R_9$
Tra_{13}	$R_5 \dashrightarrow R_7 \dashrightarrow R_8$



$$\begin{pmatrix} AS(R_1) \\ AS(R_2) \\ \vdots \\ AS(R_m) \end{pmatrix} = M \cdot \begin{pmatrix} 1 \\ AS(R_1) \\ \vdots \\ AS(R_m) \end{pmatrix}$$

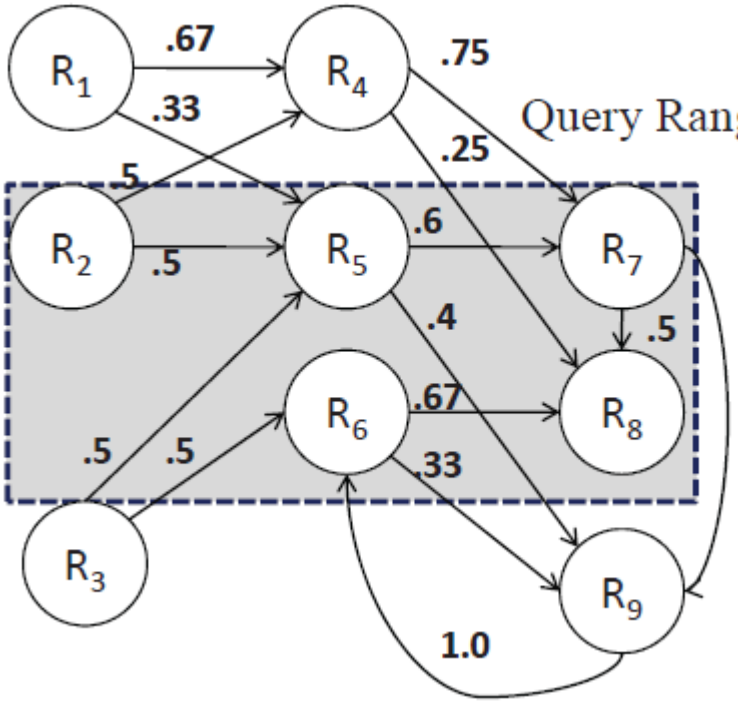
R_1: 0.15
R_2: 0.15
R_3: 0.15
R_4: 0.299
R_5: 0.32
R_6: 0.214
R_7: 0.501
R_8: 0.356
R_9: 0.301

Pattern-Aware Trajectory Search

Input: a query range Q

Output: top-k trajectories passing Q

* Trajectory score S(Tra)= $\sum_{R_i \in Tra \cap Q} AS(R_i)$



E.g.,

$$S(Tra_4) = \sum_{R_i \in Tra_4 \cap Q} AS(R_i)$$

$$= AS(R_2) + AS(R_8) = 0.15 + 0.36 = 0.51$$

Tid	A sequence of potential regions
Tra_1	$R_1 \dashrightarrow R_4 \dashrightarrow R_7$
Tra_2	$R_1 \dashrightarrow R_4 \dashrightarrow R_7$
Tra_3	$R_1 \dashrightarrow R_5 \dashrightarrow R_7$
Tra_4	$R_2 \dashrightarrow R_4 \dashrightarrow R_8$
Tra_5	$R_2 \dashrightarrow R_5 \dashrightarrow R_7$
Tra_6	$R_5 \dashrightarrow R_9$
Tra_7	$R_3 \dashrightarrow R_5$
Tra_8	$R_3 \dashrightarrow R_6 \dashrightarrow R_8$
Tra_9	$R_6 \dashrightarrow R_8$
Tra_{10}	$R_6 \dashrightarrow R_9$
Tra_{11}	$R_5 \dashrightarrow R_9 \dashrightarrow R_6$
Tra_{12}	$R_4 \dashrightarrow R_7 \dashrightarrow R_9$
Tra_{13}	$R_5 \dashrightarrow R_7 \dashrightarrow R_8$

Top-1 trajectory: Tra_{13}

Smart Navigation

Community-based traffic sharing platforms

* Predicting your destination
* Deriving personalized routes
* Estimating traffic status

Predicting Destinations

Historical GPS Logs

Data Cleaning

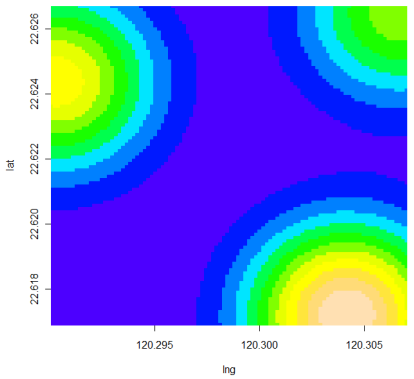
Representative Region Extraction

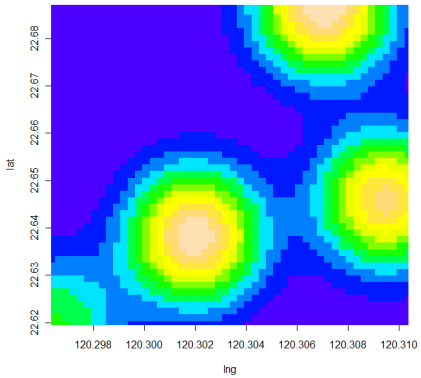
Rule Learning

Query: Source Location

Rule-Based Prediction Model

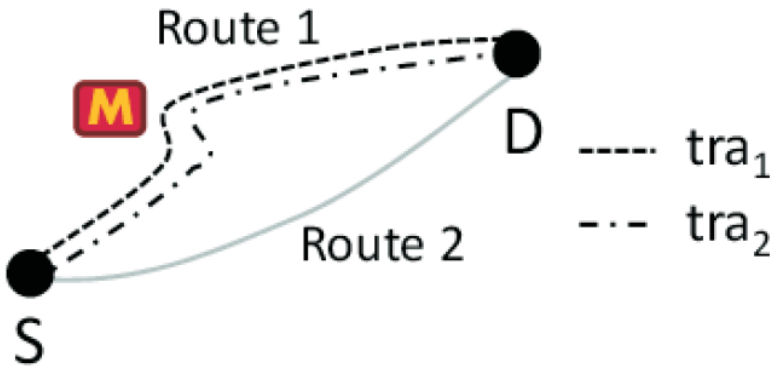
Top-k Destinations





Personalized Route Planning

* General route planning
  * Query: a start S, a destination D
  * Output: a shortest/fastest route
* Issue
  * People follow some regular routes for specific purposes and these regular routes may be not the shortest/fastest routes
  * Example
    * A user usually drives Route 1 to work
    * Route 2 is shorter/faster than Route 1



19

Personalized Route Planning

* Benefits of personalized route planning
  * Provide a user with useful real-time traffic information
    * A user follows some regular routes for specific purposes and these regular routes are usually not the shortest/ fastest routes
  * Facilitate a user's life

20

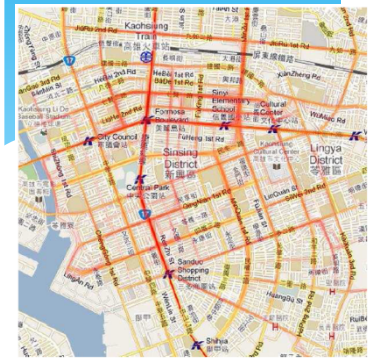
Personalized Route Planning

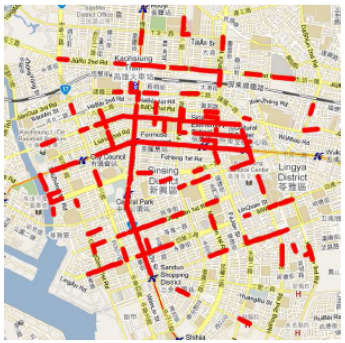
[Problem Definition]

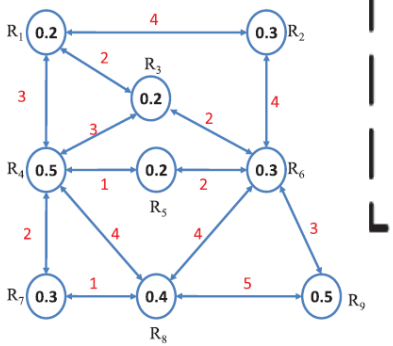
* Input
  * Road network
  * One user's historical trajectories
  * Query: a start S, a destination D, a rank-threshold k
* Output
  * The top-k personalized routes from S to D
    * Length
    * Familiarity

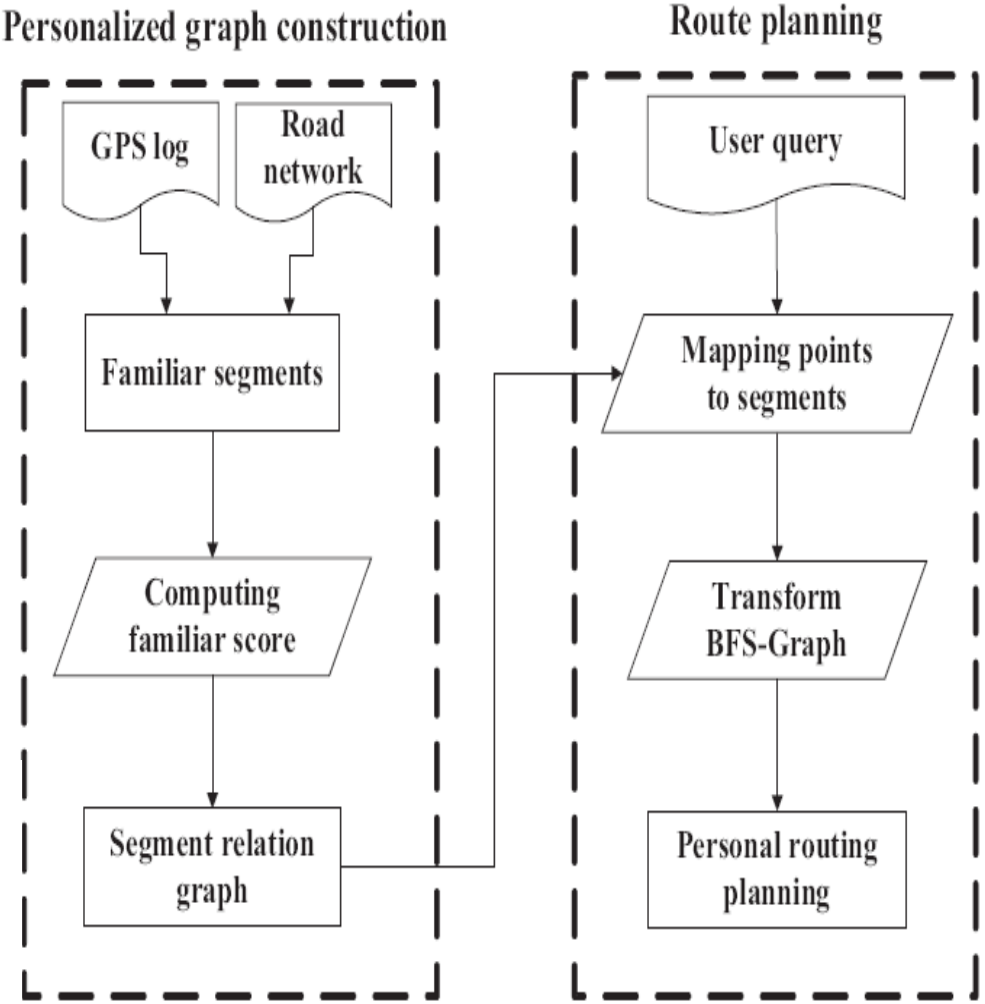
21

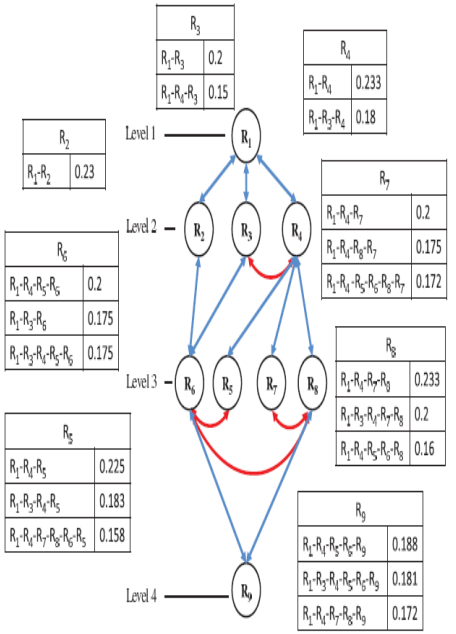
System Framework











Traffic Estimation on Personalized Routes

* Goal
  * Estimate speeds on road segments
* Data
  * Traffic database: Road network, GPS data (from vehicles)
* Input
  * Road segments & time
* Output
  * Estimated speeds on the road segments at the specified time

Spatio-Temporal Weighted Approach

Query
(Time, Road Segment)

Traffic Database

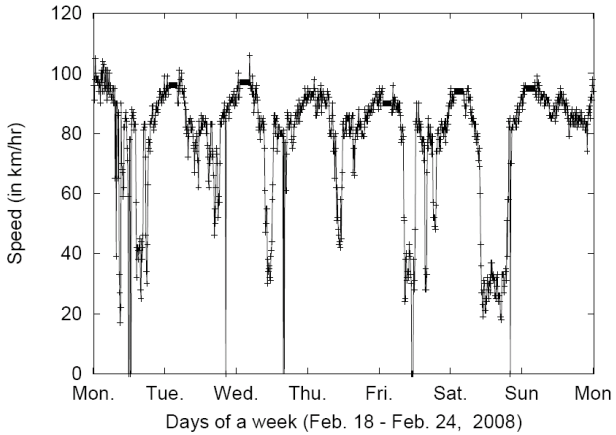
Data Selection
- Temporal Factor
- Spatial Factor

Historical Data

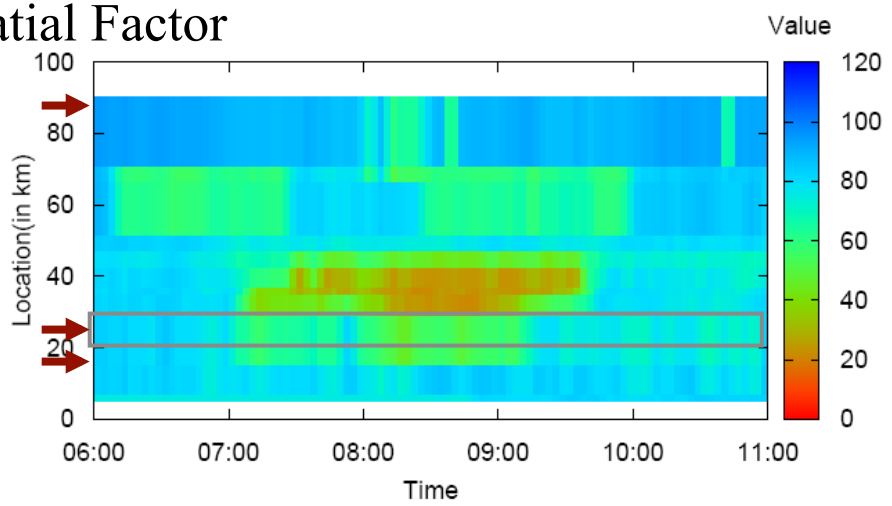
WMA-Like

Status of the Query Road

- Temporal Factor



- Spatial Factor



Mining trajectory profiles for identifying user communities

* Mine user communities
  * Users who have similar moving behaviors
* Utilize user communities
  * Geo-web sites
    * Look for some interesting traveling/training paths
    * Find traveling/training partners
  * Carpooling service
  * Group-buying

Group-Buying in Taiwan

100% 土鳳梨酥 喜餅專用禮盒，

可裝　　１２入１５０元

一盒一袋，訂婚送禮大方體面

適用滿１６００元免運　滿５０００元９折

囍字可貼或不貼，要貼的要註明

Where to meet for picking up your orders

合購: 徵1600免運 過年團,到貨日期：1/18(三)or
統計: 3 人　106 份　1750 元
面交: [台北市]大同區 延平北路民權西路(主購家)近
主購: annie77649 (315) | 3 天前 開團

合購: 很想試試10元的土鳳梨酥 徵1,600元免運
統計: 5 人　115 份　1150 元
面交: [新北市]三重區 三重徐匯中學門口 [新北市]
主購: agnes (238) | 3 天前 開團

1 - 10 of 1149　　1 / 115　1

商店粉絲

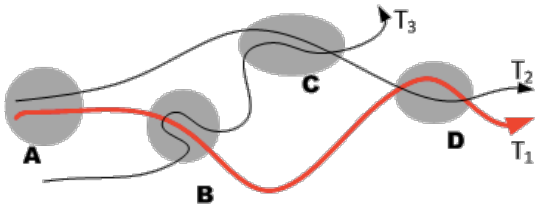
加入商店粉絲

請勾選「加入商店粉絲」，當有新合購會立刻通知您喔！

有 663 人想要一起合購喔！

Discovering Communities

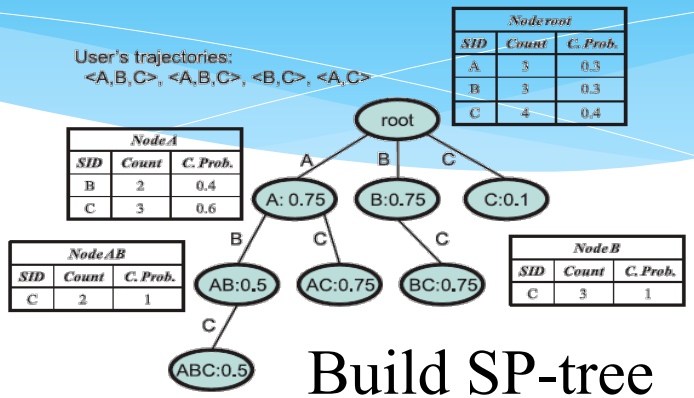
1. Preprocess



Find hot regions

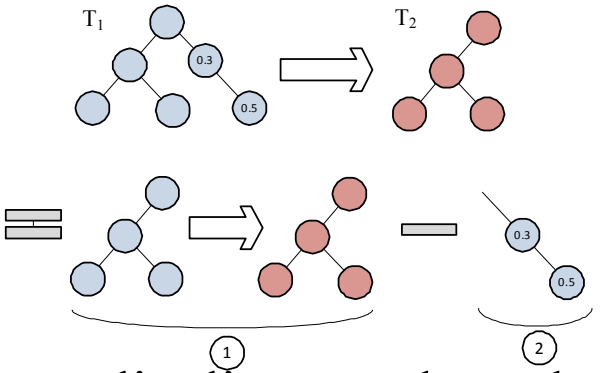
2. Construct User's Profile

User's trajectories:
<A,B,C>, <A,B,C>, <B,C>, <A,C>



Build SP-tree

3. Formulate Distance function



Use an edit-distance-based approach

4. Identify Community

Community 1 Community 2

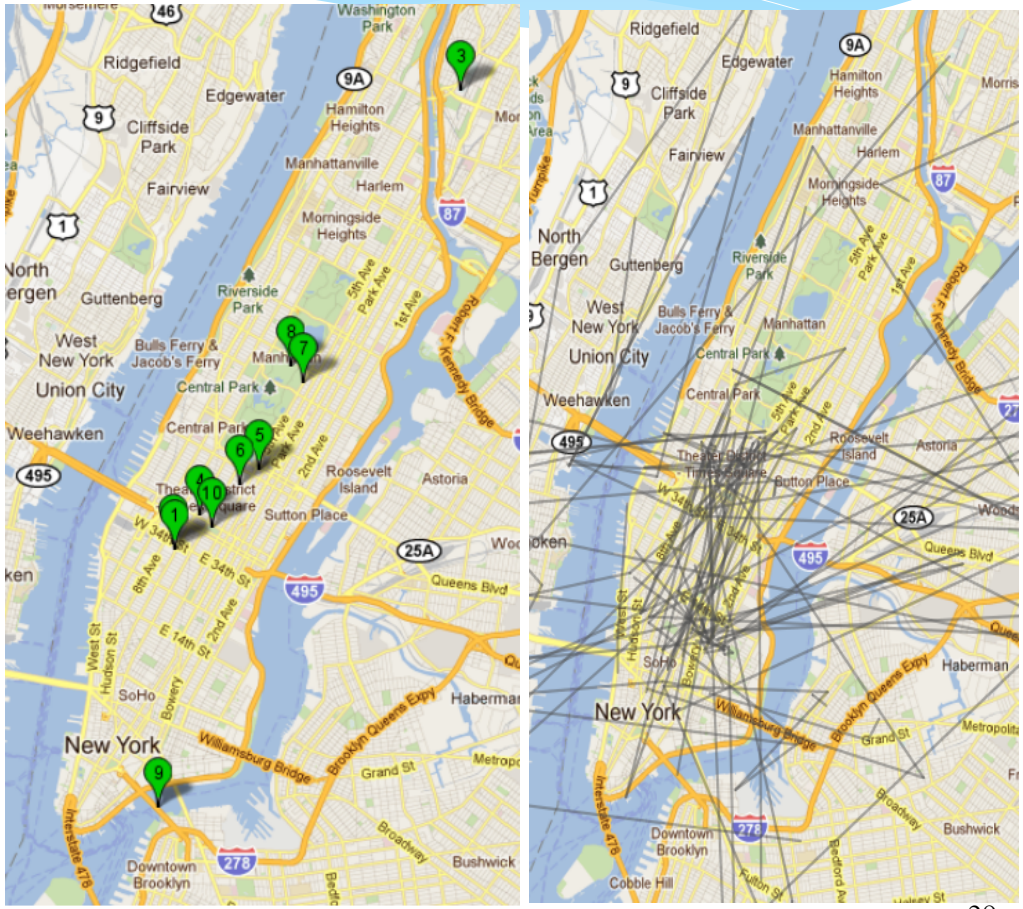
Cluster users by their SP-trees

Trajectories are not “Perfect”

* Check-in data are sample data points from trajectories
* Given a set of check-in data, trajectory data mining work is a challenging task

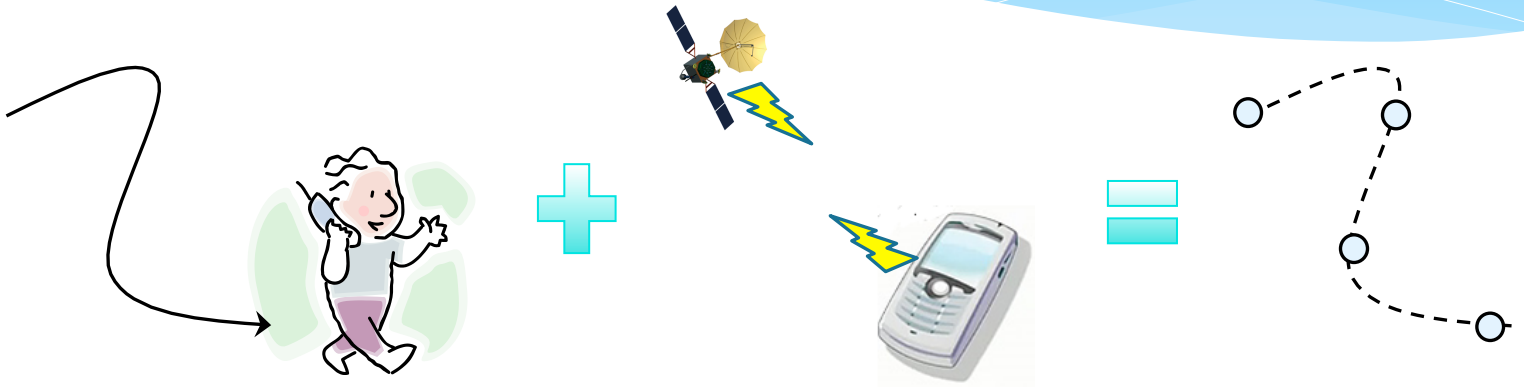
Foursqaure: Check-In records

* Top-10 check-in locations in New York
* In this area
  * 10306 users
  * 21704 trips
  * 6 check-ins per trip



Trajectory Characteristics

* The generation of trajectories:



Movement + **Positioning** = **Trajectory**

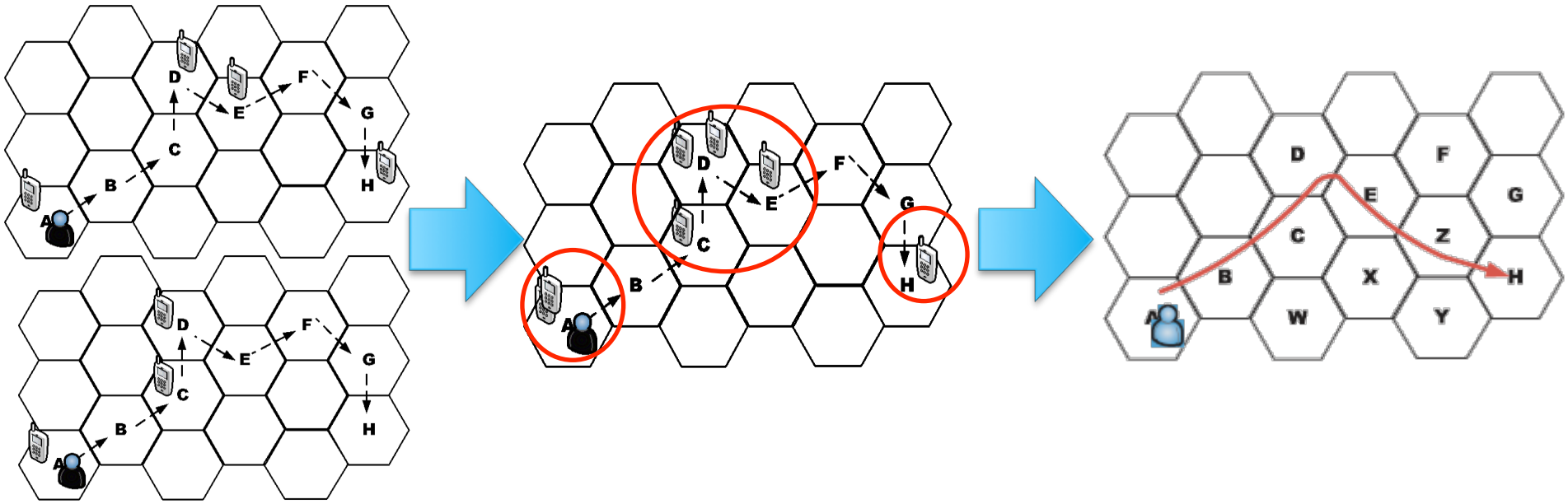
Spatial and Temporal Shifting

Silent Duration

30

A Regression Approach

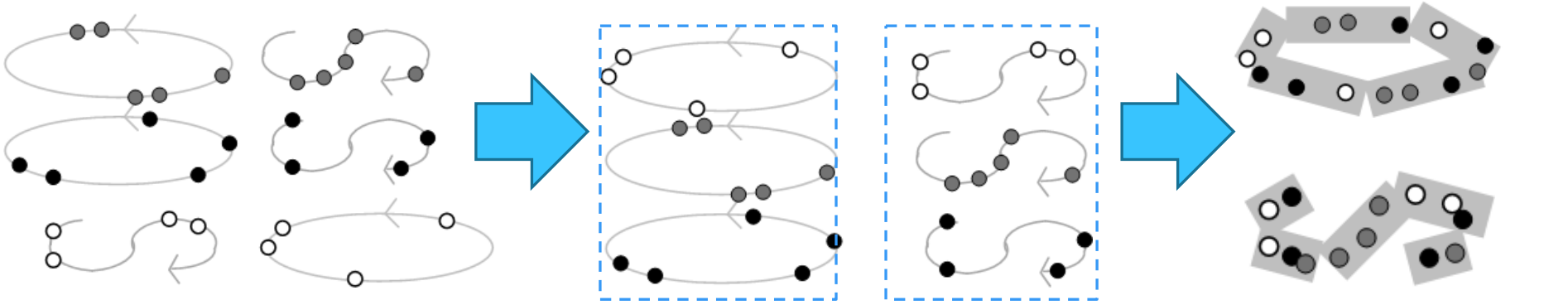
* Exploring frequent and regularity of users to approximate user movement paths



C.-C. Hung and W.-C. Peng, ``A Regression-based Approach for Mining User Moving Patterns from Random Sample Data", to appear in *Data and Knowledge Engineering*.

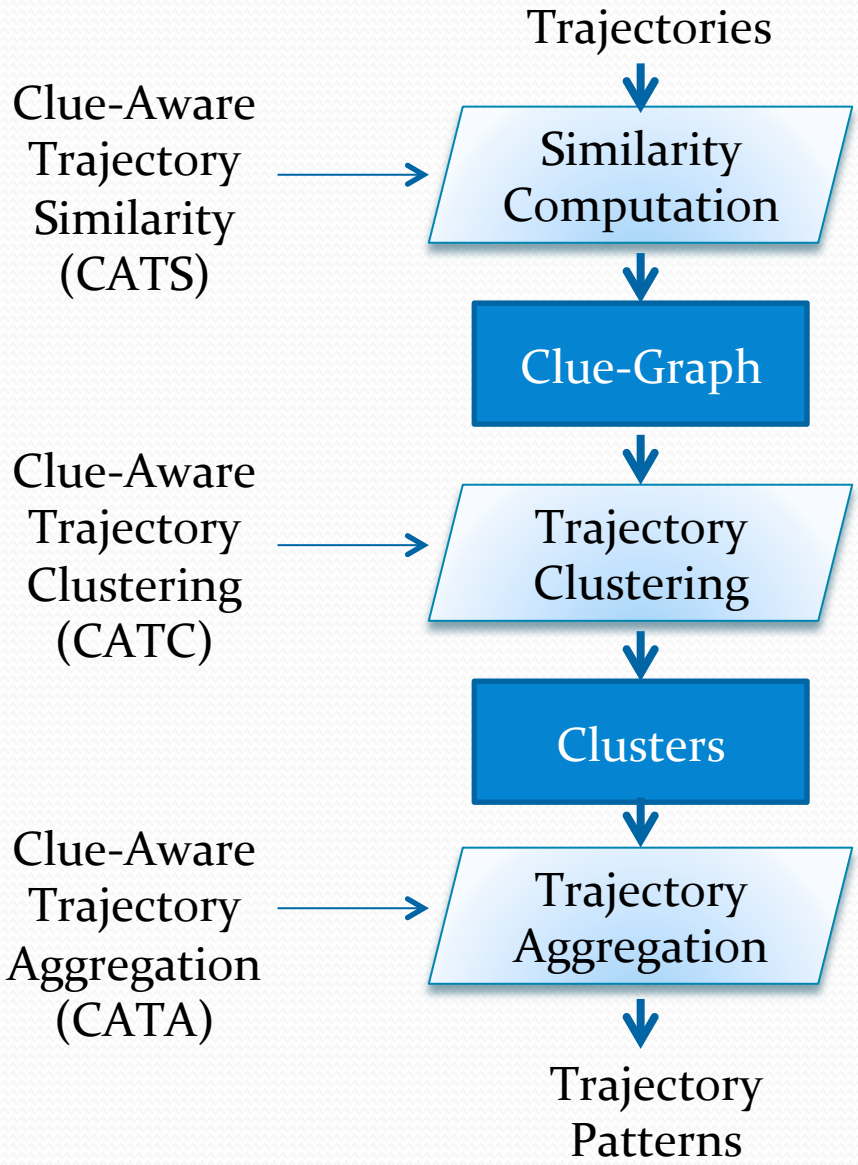
CACT: Clustering and Aggregating Clues of Trajectories for Trajectory Patterns and Routes

* Formulate a new similarity measure based on clues in trajectories
* Propose a clue-aware trajectory clustering algorithm



C.-C. Hung and W.-C. Peng and W.-C. Lee ``CACT: Clustering and Aggregating Clues of Trajectories for Trajectory Patterns and Routes", **Very Large Data Base Journal (top 1 journal in databases)**

Solution Framework

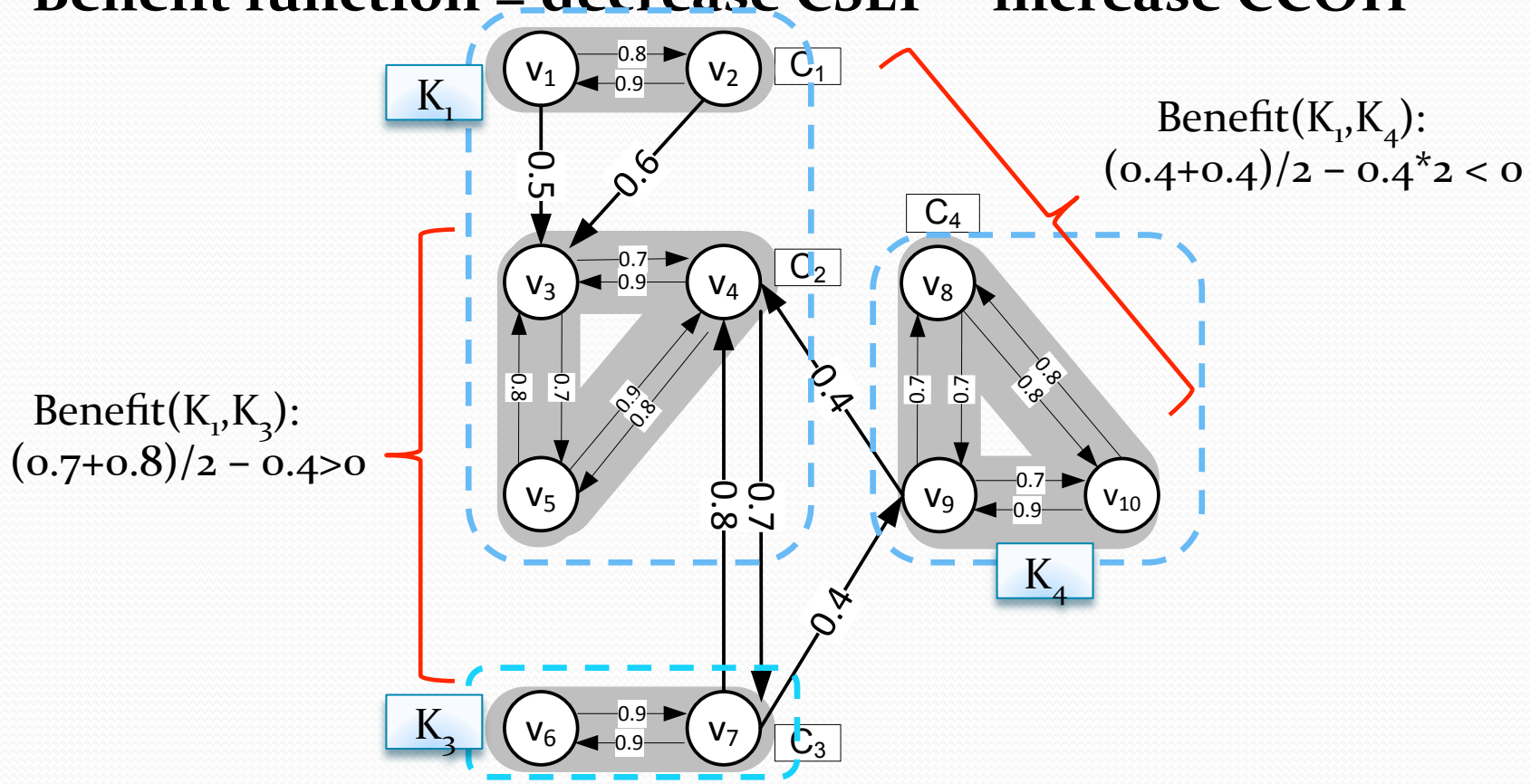


Clue-Aware Trajectory Clustering

* **Step 1: Find core-sets**
* **Step 2: Exploit a benefit function to merge candidate-clusters**
  * In the beginning, a core set is a candidate-cluster.
  * More common edges two core-set are, more confident they follow the same behavior.
* **Step 3: Repeat Step 2 until merging cannot bring any benefit**
  * Merging any two candidate-clusters leads to negative benefit function
* **Step 4: Eliminate infrequent clusters**
  * Infrequent clusters contain less than min_sup trajectories.

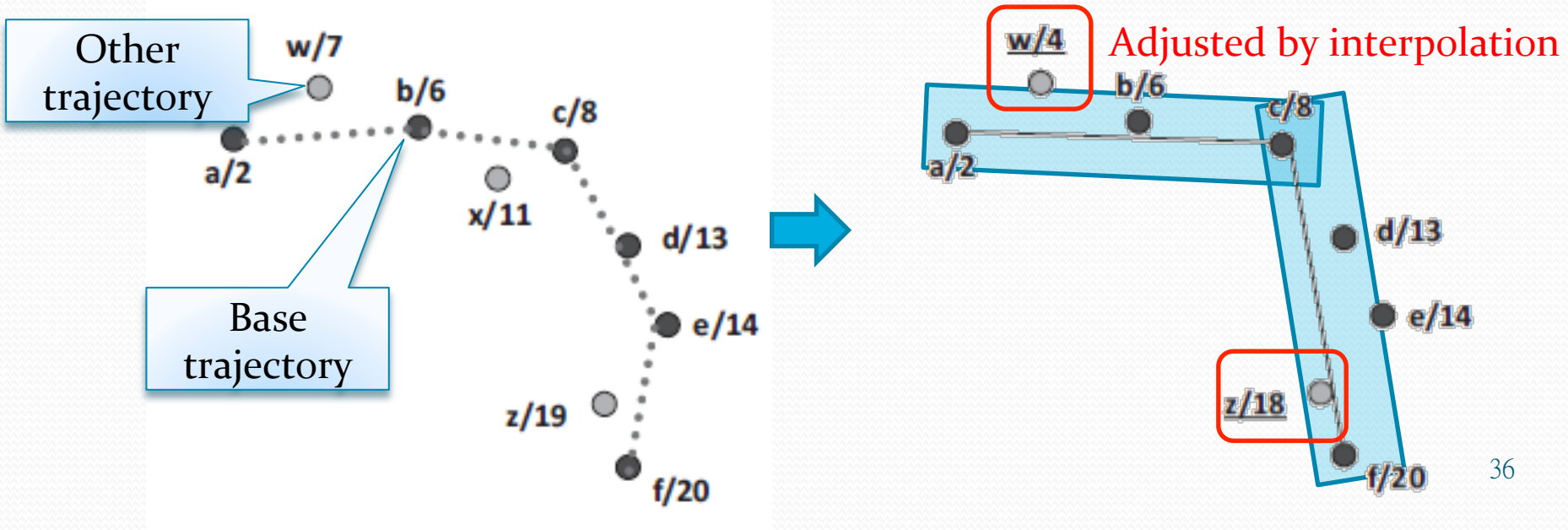
Benefit Function

- **Merge two candidate-clusters into one:**
  - Total CCOH increases
  - Total CSEP decreases
- **Benefit function = decrease CSEP – increase CCOH**



Aggregation Phase

- **Goal:** derive trajectory patterns with representative lines.
  - **Input:** Trajectory clusters $\{K_1, K_2, ..., K_n\}$
  - **Output:** Trajectory patterns $\{P_1, P_2, ..., P_n\}$
- **Concept:**
  - Select a base trajectory
  - Adjust spatial/temporal information of other trajectories
  - Use Douglas-Peucker simplifier to obtain the representative lines



Social media is growing

- "Location, Location, Location", **in MDM 2009,** Christian S. Jensen already pointed out.
- Next step: understanding trajectory patterns behind users
- On-going projects:
  - Sensor-enabled cloud services (with NSC, Mio and ITRI)
  - BuddySquare (with ITRI)
  - Mining App. Usage behaviors (with HTC)

Selected publication

•C.-C. Hung, W.-C. Peng, and W.-C. Lee, ``Clustering and Aggregating Clues of Trajectories for Mining Trajectory Patterns and Routes," *Very Large Data Base Journal (VLDB J)* (to appear).

•C.-C. Hung and W.-C. Peng, ``A Regression-based Approach for Mining User Movement Patterns from Random Sample Data," *Data and Knowledge Engineering*, Vol. 70, No. 1, pp. 1-20, 2011.

•T.-W. Lin, B. Zheng, L.-Y. Wei,and W.-C. Peng, ``Exploring Dynamic Query Shapes for Nearby Traffic Monitoring Queries in Road Networks," *Proceedings of ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, Chicago, Illinois, USA, Nov. 1-4, 2011.

•K.-P. Chang, L.-Y. Wei, Mi-Yeh Yeh and W.-C. Peng, ``Discovering Personalized Routes from Trajectories," *Proceedings of the third International Workshop on Location-based Social Networks (in conjection with ACM SIGSPATIAL GIS 2011)*, Chicago, Illinois, USA, Nov. 1-4, 2011.

•P.-R. Lei, Acrt Shen and W.-C. Peng, ``Exploring Spatial-Temporal Trajectory Model for Location Prediction,"Proceedings of the 12th International Conference on Mobile Data Management, Luleå, Sweden, June 6-9, 2011. (Best Student Paper Award)

•C.-C. Hung and W.-C. Peng, ``Clustering Fragmented Trajectories for Mining Movement Behaviors," *Proceedings of the 2011 International Workshop on Behavior Informatics (In conjunction with PAKDD),* Shenzhen, China, May 24, 2011.

•C.-C. Hung and W.-C. Peng, ``Model-driven Traffic Data Acquisition in Vehicular Sensor Networks," *Proceedings of the 39th International Conference on Parallel Processing (ICPP 2010)*, San Diego, CA, USA, Sept. 13-16, 2010.

•L.-Y. Wei, W.-C. Peng, B.-C. Chen, and T.-W. Lin, ``PATS: A Framework of Pattern-Aware Trajectory Search," *Proceedings of the 1st Workshop on Uncertain Mobile Data Management and Mining (In conjunction with MDM)*, Kansas, Missouri, USA, May 23, 2010.

•L.-Y. Wei, W.-C. Peng, C.-S. Lin and C.-H. Jung, ``Exploring Spatio-Temporal Features for Traffic Estimation on Road Networks," *Proceedings of the 11th International Symposium on Spatial and Temporal Databases (SSTD 2009)*, Aalborg, Denmark, July 8-10, 2009.