Empowered by Innovation NEC

Mobility Demand Prediction in Urban Scenarios through Multi-source, User-generated Data

Konstantinos Gkiotsalitis, Alexandros Alexandrou
NEC Laboratories Europe
Intelligent Transport Systems Group
Heidelberg, Germany

Research Scope

MDP with surveys

- Infrequent updates (~5-10 years)
- High labor cost
- Static Information

Develop methods and techniques that:

- enable the utilization of user-generated data (Smart Card logs, Social Media) for MDP
- identify automatically individuals' mobility patterns from users' data logs
- protect users' privacy by obfuscating users' IDs and geo-tagged locations

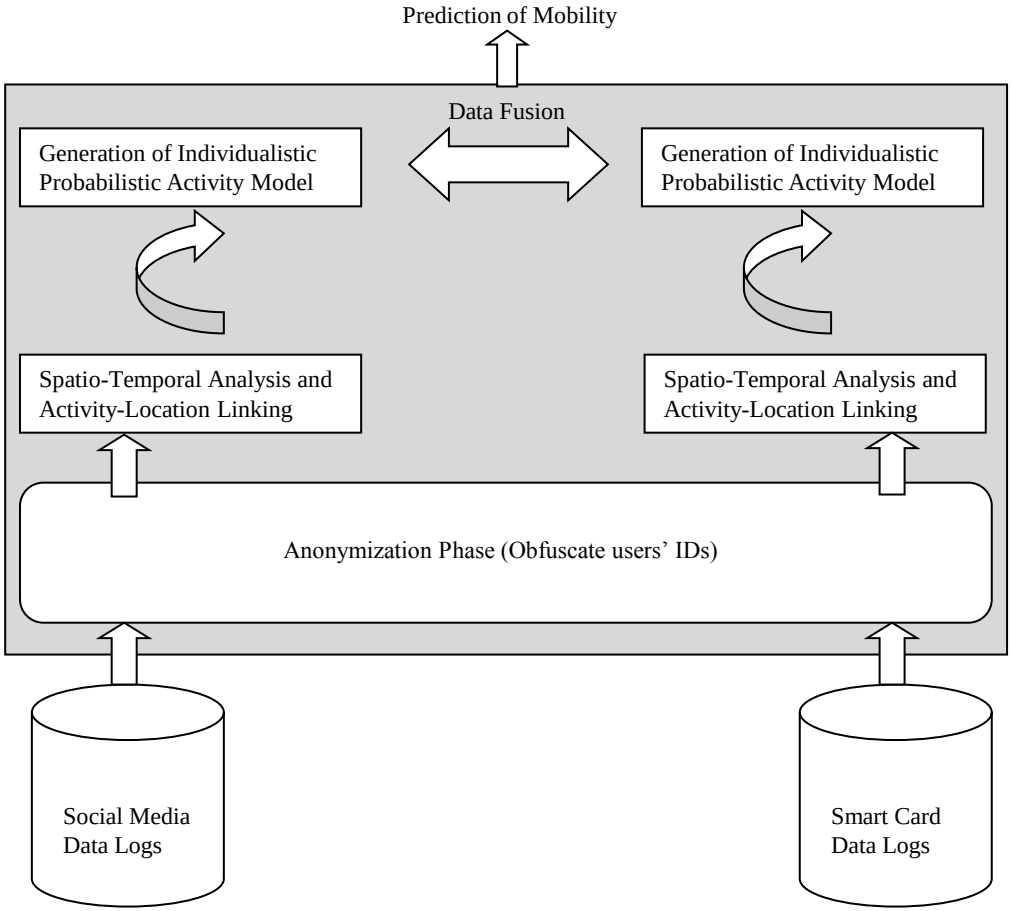
Motivation for utilizing user-generated data for MDP

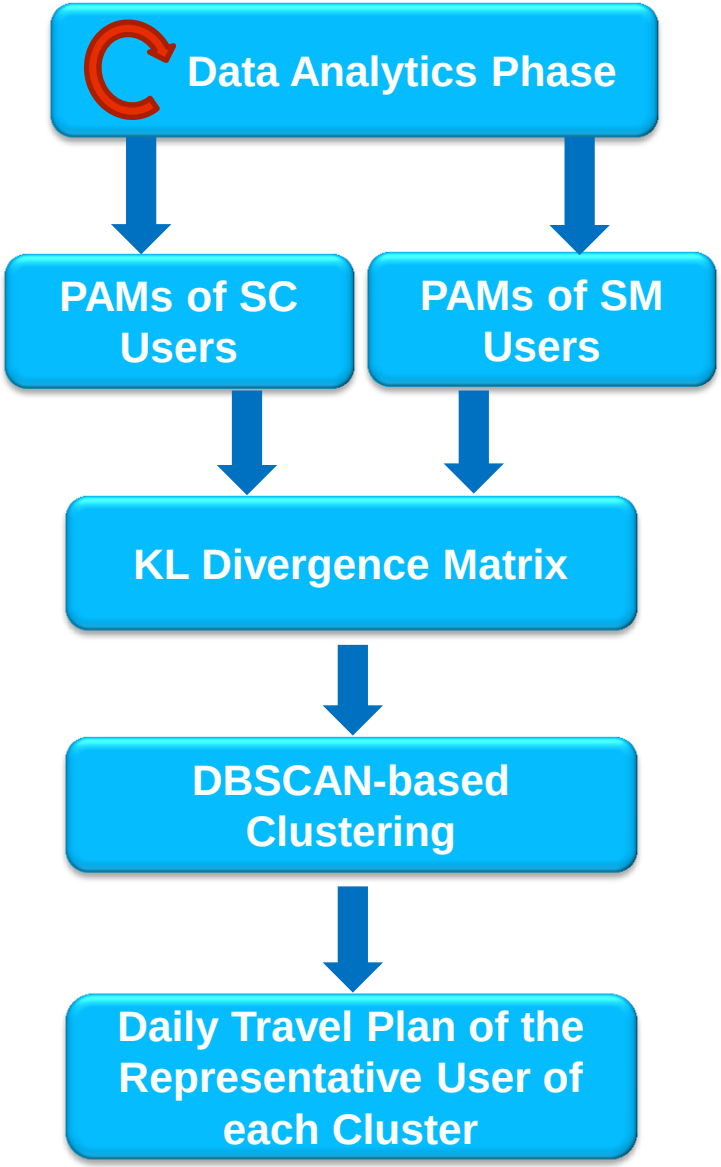
- Continuous generated information
- Vast amount of inexpensive data
- Capture seasonality effects on travel behavior
- information-rich data (i.e., embedded text)

But.. there are some drawbacks:

- Privacy issues
- Market penetration way below 100%
- Retrieved data cannot feed traditional travel demand prediction models

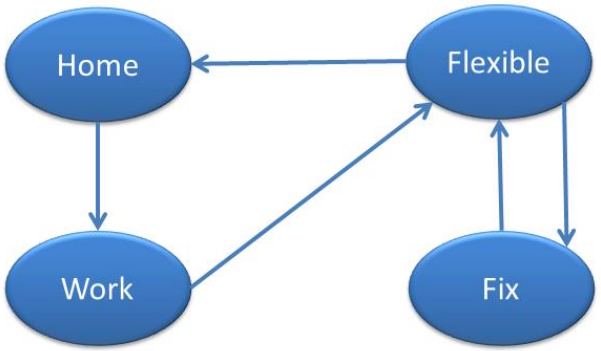
Overview of the Approach

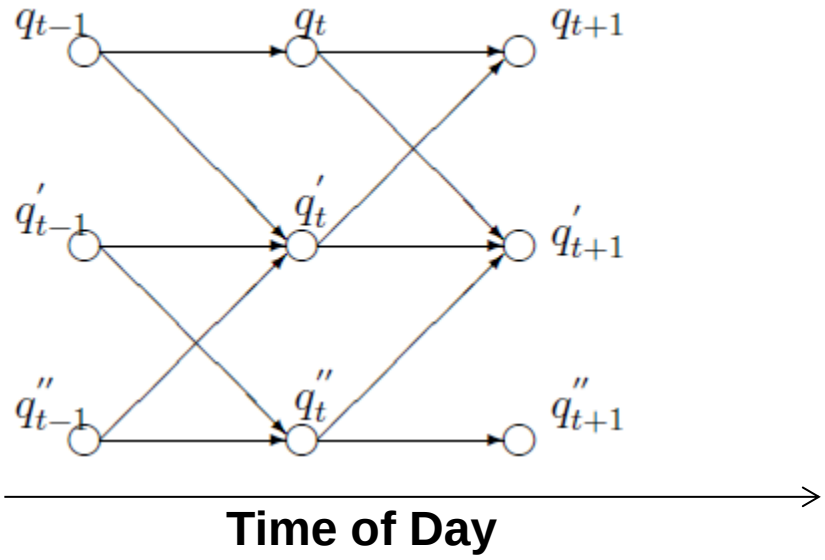




PAM: Model for forecasting users' daily schedules

Basic assumption: Individuals are characterized by their current state which evolves during the day



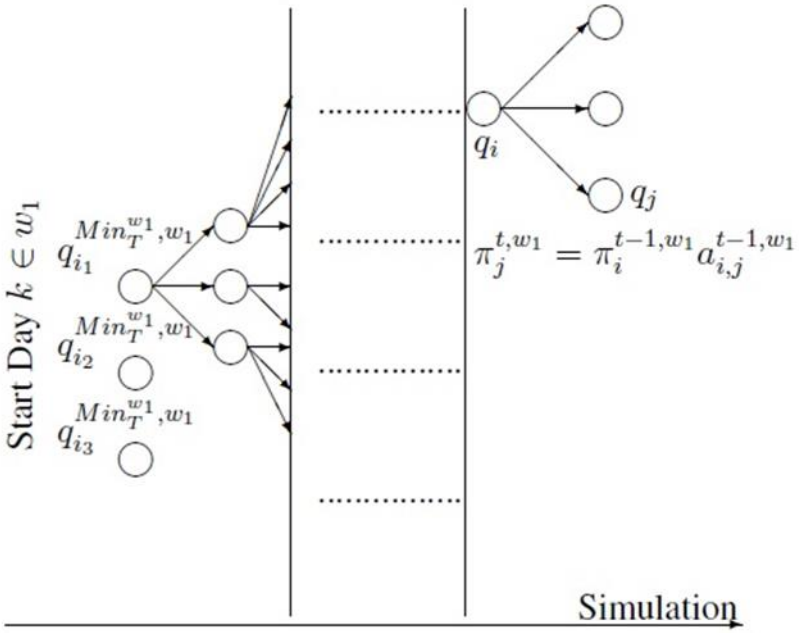


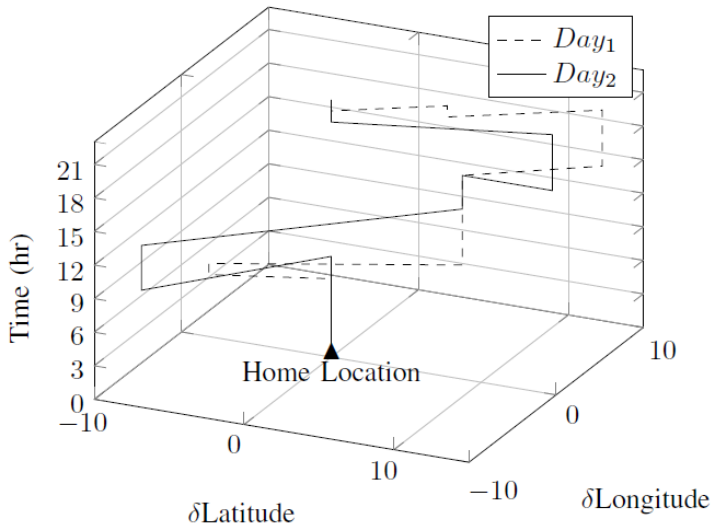
Probabilistic (Activity) Model – PAM

Learning Phase

Probabilistic Activity Model (PAM) of the User

Individualistic PAM: $P(I,k,t,L_m,A_n) = \dfrac{N(I,k,t,L_m,A_n)}{\sum_{A_n \in A}\sum_{L_m \in \Lambda} N(I,k,t,L_m,A_n)}$



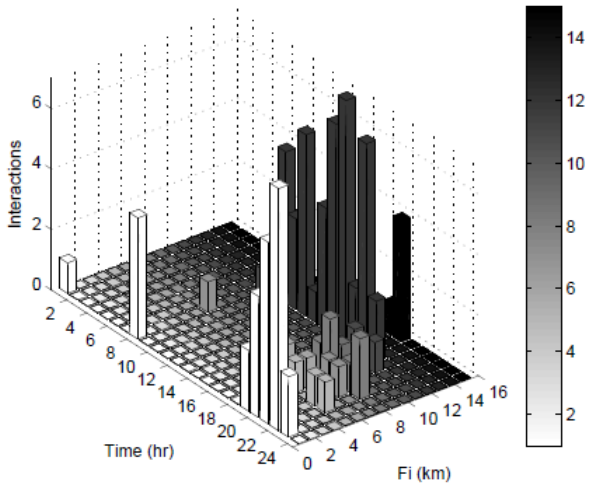


Generating user's PAM from Smart Card Data

- For each Smart Card User:
  - Revisited SC terminals are assigned to one of the following activity types

$$A_m = \begin{cases} A_1 : \text{Home or} \\ A_2 : \text{Fixed Activity, in close proximity to home location (<5km) or} \\ A_3 : \text{Fixed Activity, more than 5km away from home location or} \\ A_4 : \text{Flexible Activity, in close proximity to home location (<5km) or} \\ A_5 : \text{Flexible Activity, more than 5km away from home location} \end{cases}$$

  - For the assignment spatio-temporal analysis of users' tap-ins/outs is performed†

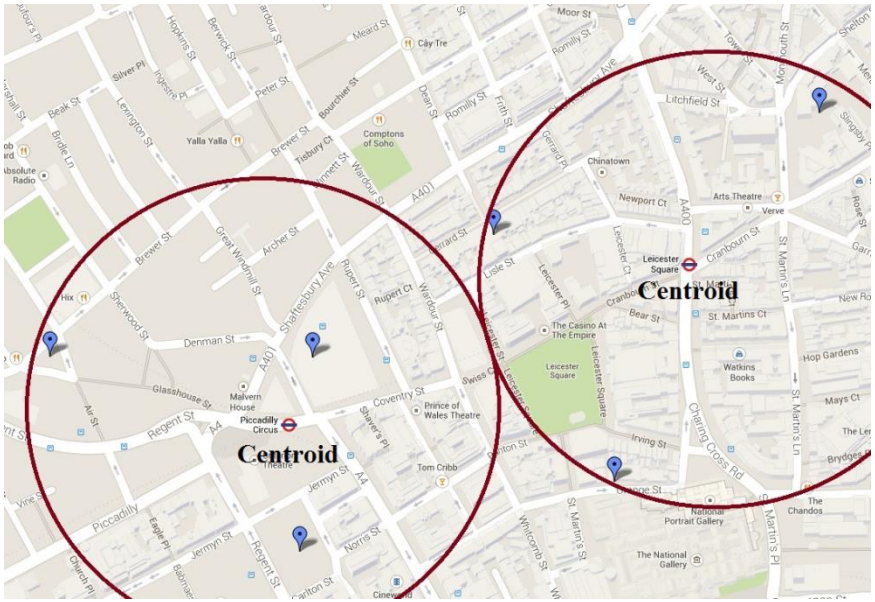


Spatio-temporal analysis of the observed individual's mobility patterns over time, where F_i is the distance from home

†GKIOTSALITIS, K., ALESIANI, F., BALDESSARI, R. (2014). Educated rules for the prediction of human mobility patterns based on sparse social media and mobile phone data. In Transportation Research Board: 93rd Annual Meeting Compedium of Papers, Paper: #14–0745.

Generating user's PAM from Social Media Data

- For each SM user:
  - Geo-tagged locations are represented by the nearest SC terminal station (centroid)
  - Users' IDs are obfuscated
  - Activities are assigned to centroids based on:



$$A_m = \begin{cases} A_1 : \text{Home or} \\ A_2 : \text{Fixed Activity, in close proximity to home location (<5km) or} \\ A_3 : \text{Fixed Activity, more than 5km away from home location or} \\ A_4 : \text{Flexible Activity, in close proximity to home location (<5km) or} \\ A_5 : \text{Flexible Activity, more than 5km away from home location} \end{cases}$$

  - For the assignment, spatio-temporal analysis of users' SM data is performed

KL-Divergence Matrix

- PAM probabilistic distance between users based on Kullback-Leibler Divergence:

$$KL_{I_1,I_2} = \frac{1}{|\lambda| t_{max}} \sum_{t=0}^{t=t_max} \sum_{L_m=0}^{L_m=|\lambda|} \sum_{k=0}^{k=1} \sum_{A_n=0}^{A_n=|A|} |P(I_1,k,t,L_m,A_n) - P(I_2,k,t,L_m,A_n)|^2$$

- K-L matrix (all pairs of users):

$$[KL] = \begin{pmatrix} KL_{0,0} & KL_{0,1} & \cdots & KL_{0,N-1} \\ \vdots & \vdots & \ddots & \vdots \\ KL_{j,0} & KL_{j,1} & \cdots & KL_{j,N-1} \\ \vdots & \vdots & \ddots & \vdots \\ KL_{N-1,0} & KL_{N-1,1} & \cdots & KL_{N-1,N-1} \end{pmatrix}$$

Description of the Data Sample

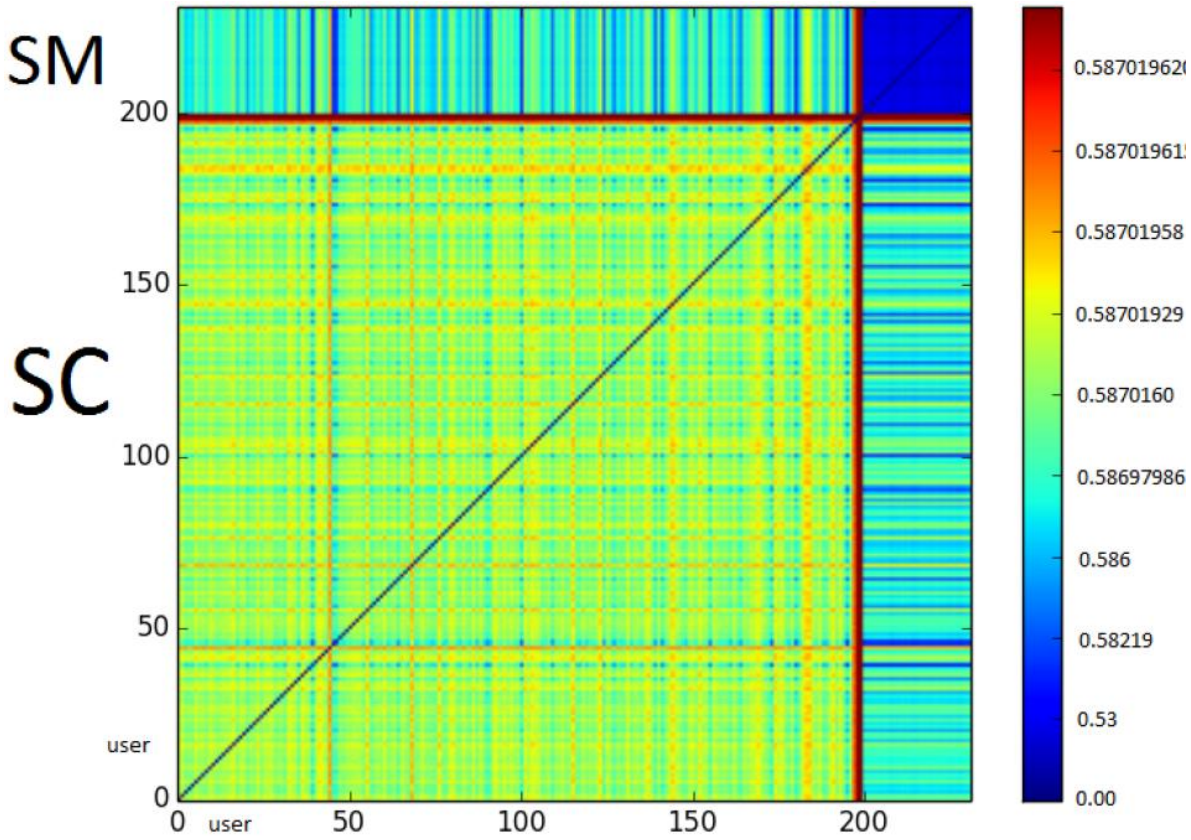
SC_User_ID	Day (0-6)	Enter_Minute	Exit_Minute	Tap_In_Station	Tap_Out_Station
0	0	1315	1423	421	217
0	2	101	164	170	398
0	4	904	950	318	463
0	1	1413	1440	510	282
0	3	656	745	129	549
0	2	1432	1440	277	346
0	1	992	1102	217	282
0	2	748	864	423	461
0	1	535	578	56	126
0	3	1082	1178	388	157
0	0	637	726	258	146
0	1	721	812	5	218
0	5	823	845	333	322
0	5	133	185	440	199
0	4	704	756	181	215
0	5	1241	1302	277	484
0	3	1380	1440	433	463
0	3	716	813	428	259
0	4	921	991	295	491
0	0	681	705	273	509
0	3	319	389	183	13
0	4	988	1020	558	587
0	4	1179	1182	103	420
0	1	425	470	194	420
0	0	1202	1282	315	161
0	0	1248	1295	400	562

SC: Data entries from 200 users in London (Oyster Card; 608 stations)

Day	Month	Date	Time	Latitude	Longitude	Recipients	Text Message
Tue	May	7	07:47:17	51.8374	-1.350075	None	""Magic #topiary #alphabet #unnatural @
Sun	May	5	17:41:04	51.84196	-1.361575	None	""Perfection #blue #green #summertime
Sat	May	4	17:02:31	51.84196	-1.361575	None	""Park Life #blenheimpalace #moat #brid
Sat	May	3	17:02:04	51.84762	-1.355256	None	""Metal Work #woodstock #pub #chilled
Fri	May	2	21:44:39	51.50704	-0.116489	None	""Bank Robber #savetheskatepark #south
Thu	May	30	22:37:05	51.5137	-0.130119	None	""One for the road with @bill_studio son
Tue	April	28	08:26:38	51.51734	-0.139902	None	""Sign of the Times #xx #exposure20 #its
Sun	April	27	20:32:14	51.51723	-0.140849	None	""The Don. Amongst Dons. Thanks @ther
Sat	April	27	20:28:13	51.52263	-0.103582	None	""Pre birthday warm up #cheese #redwin
Sat	April	27	15:16:03	51.54129	-0.145874	None	""Familiar Face #lookalike #joestrummer
Sat	April	27	15:10:54	51.54212	-0.147173	None	""Properly mental place... #cyberdog #ne
Thu	April	25	17:31:51	48.8738	2.294952	None	""L'Arc de Soleil #paris #sunny #blue #arc
Wed	April	24	17:58:12	51.5202	-0.135056	None	""Better than Easter #olliedabbous #delic
Sat	April	20	22:30:12	48.85414	2.333533	None	""Sit Still #cafeflore #reflections #shoot #
Sat	April	20	22:15:23	48.86099	2.34188	None	""Possibly the best meal I've ever had in
Sat	April	20	18:57:48	48.8573	2.33189	None	""Anyone home #legend #sergegainsbour
Fri	April	19	17:49:12	50.63921	3.07555	None	""Romantic escape #knitting sonstein #pa
Fri	April	19	08:42:08	51.51734	-0.139902	None	""What's the magic number? #xx #exposu
Thu	April	18	22:08:29	51.51604	-0.135258	None	""Anyone seen #ghostfacekillah #100club
Thu	April	18	22:04:51	51.51604	-0.135258	None	""The brilliant #ghostfacekillah #wutang
Thu	April	18	21:16:44	51.51604	-0.135258	None	""No two cups are the same #fragment #
Thu	April	18	21:12:28	51.51604	-0.135258	None	""It's gettin' hot in here #getdirty with kei
Thu	April	18	20:37:15	51.51604	-0.135258	None	""Boom! #doom #100club #converse #ho
Thu	April	18	18:59:28	51.51604	-0.135258	None	""Snap @therealnihal I blame keithtaper
Wed	April	17	07:47:30	51.51292	-0.122225	None	""Clean up the streets #graffitti #streetart
Tue	April	16	13:20:50	51.507	-0.129174	None	""Learn and pass it on #englisheffect #ori
Tue	April	16	08:25:02	51.507	-0.129174	None	""Pure beauty #markwallinger #whitehor
Tue	April	16	08:12:11	51.50799	-0.128049	None	""View from the top #nelsonscolumn #sp
Tue	April	16	08:10:26	51.52149	-0.138858	None	""Only way is up #bluesky #bttower #spri
Sun	April	14	16:04:26	51.51931	-0.121737	None	""Supercool innlondon #superpool #archi
Sun	April	14	16:00:40	51.51931	-0.121737	None	""Film focus innlondon #projections #sta
Sun	April	14	15:54:40	51.51931	-0.121737	None	""Architecture beyond boundaries innlon
Sun	April	14	15:39:17	51.51931	-0.121737	None	""Art hits the bullseye innlondon #istanb
Sun	April	14	11:39:31	51.54953	-0.168008	None	""Spring? Finally #sunshine @ Belsize Vill
Wed	April	10	18:30:16	51.51734	-0.139902	None	""Spell check malfunction? #xx #exposure

SM: Data entries of 32 users; time period varies from 2 to 12 months from November 2012-February 2014

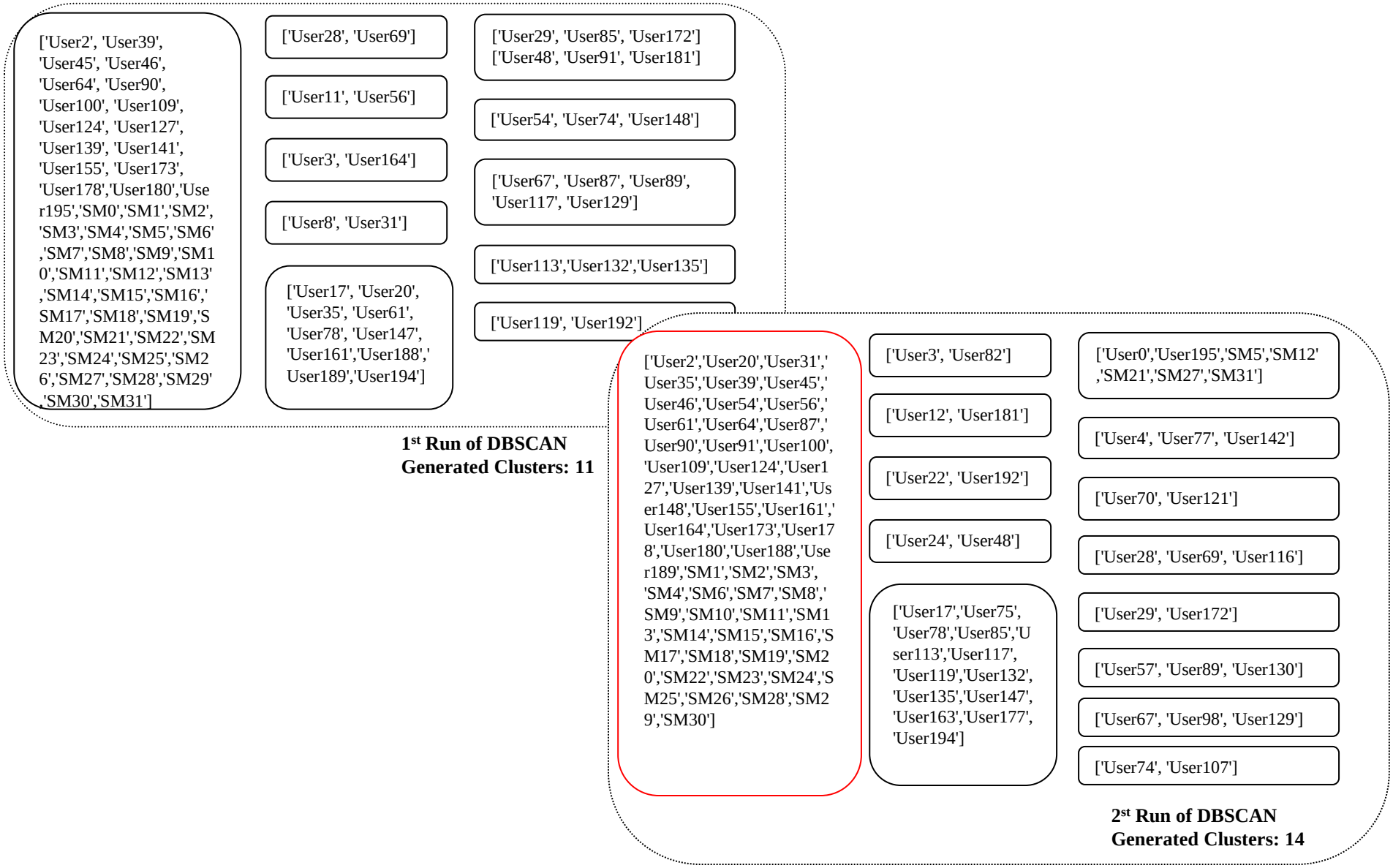
KL-Divergence Results



KL Distance Matrix. 0-199 are SC users and 200-231 are SM users

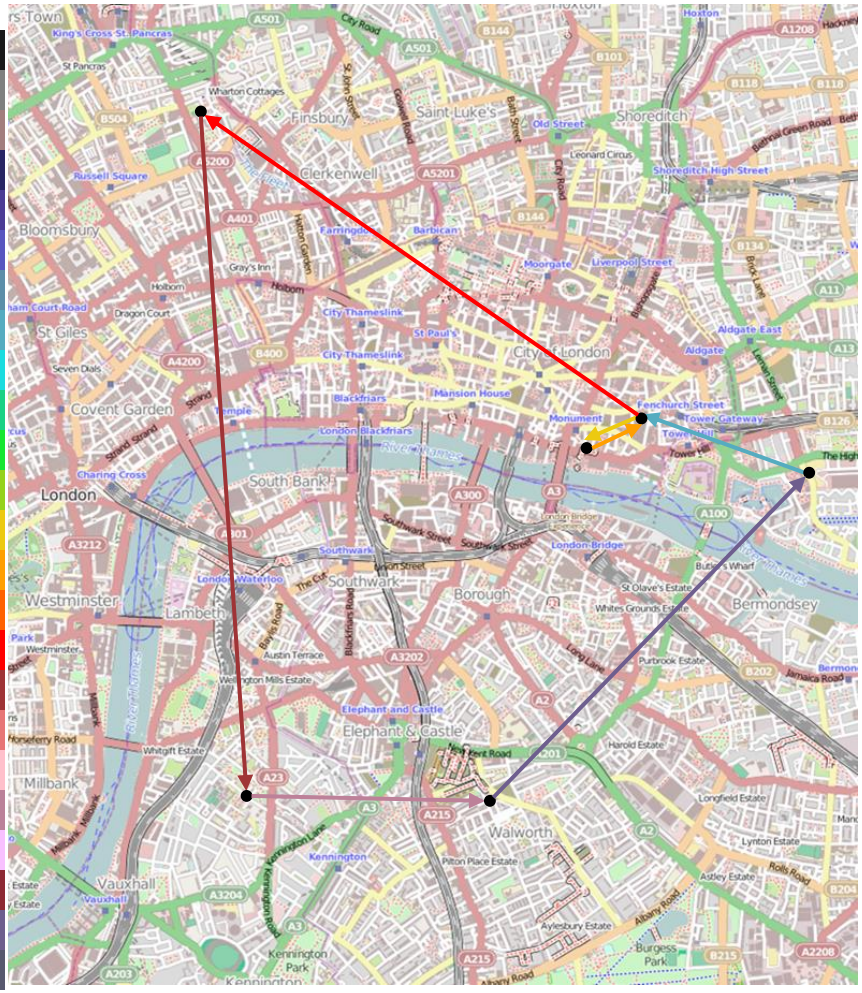
- The probability distance is bigger among SC users (range 0.58-0.587)
- SM users have more common mobility-activity patterns (range 0-0.53)
- The probability distance between SC and SM users is in the range of 0.53-0.587

Results of Users' Clustering based on a modified DBSCAN method



Daily Travel Plan

Time Period	Colour
0 - 1	
1 - 2	
2 - 3	
3 - 4	
4 - 5	
5 - 6	
6 - 7	
7 - 8	
8 - 9	
9 - 10	
10 - 11	
11 - 12	
12 - 13	
13 - 14	
14 - 15	
15 - 16	
16 - 17	
17 - 18	
18 - 19	
19 - 20	
20 - 21	
21 - 22	
22 - 23	
23 - 0	



Daily Mobility Patterns of the Representative User in Cluster 1 (OpenStreetMap View)

Discussion

- The developed techniques were tested with the use of SC and SM data demonstrating that:
  - User-generated data from different sources can be fused based on users' mobility-activity patterns without taking into account privacy-sensitive information
  - Linking users' from different sources can be used for:
    - Improving data completeness; supplement traditional surveys

- For future research
  - Link clustered users with census data to perform mobility demand prediction in a city-scenario
  - Test the performance of the techniques against O-D data from a city

ACKNOWLEDGMENT

This work was partially supported by the European Commission under TEAM, a large scale integrated project part of the FP7-ICT for Cooperative Systems for energy-efficient and sustainable mobility

Empowered by Innovation

Thank you for your attention