Regularity and Conformity: Location Prediction Using Heterogeneous Mobility Data

KDD'15

Yingzi Wang[12], Nicholas Jing Yuan[2], Defu Lian[3], Linli Xu[1]

Xing Xie[2], Enhong Chen[1], Yong Rui[2]

[1]University of Science and Technology of China

[2]Microsoft Research

[3]University of Electronic Science and Technology of China

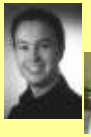
The State of Machine Learning

Classification models

Decision tree 
Bayesian classifier
Perceptron
Neural networks

Maximal Margin

SVM (Vapnik)

Maximal margin sequential learning

M3N network (Ben Taskar)


Deep learning

Deep belief networks


Topic models

PLSI, LDA, etc.



Sequential learning

HMM

Sequential learning

MEMM, CRF, voted perceptron


Graphical models

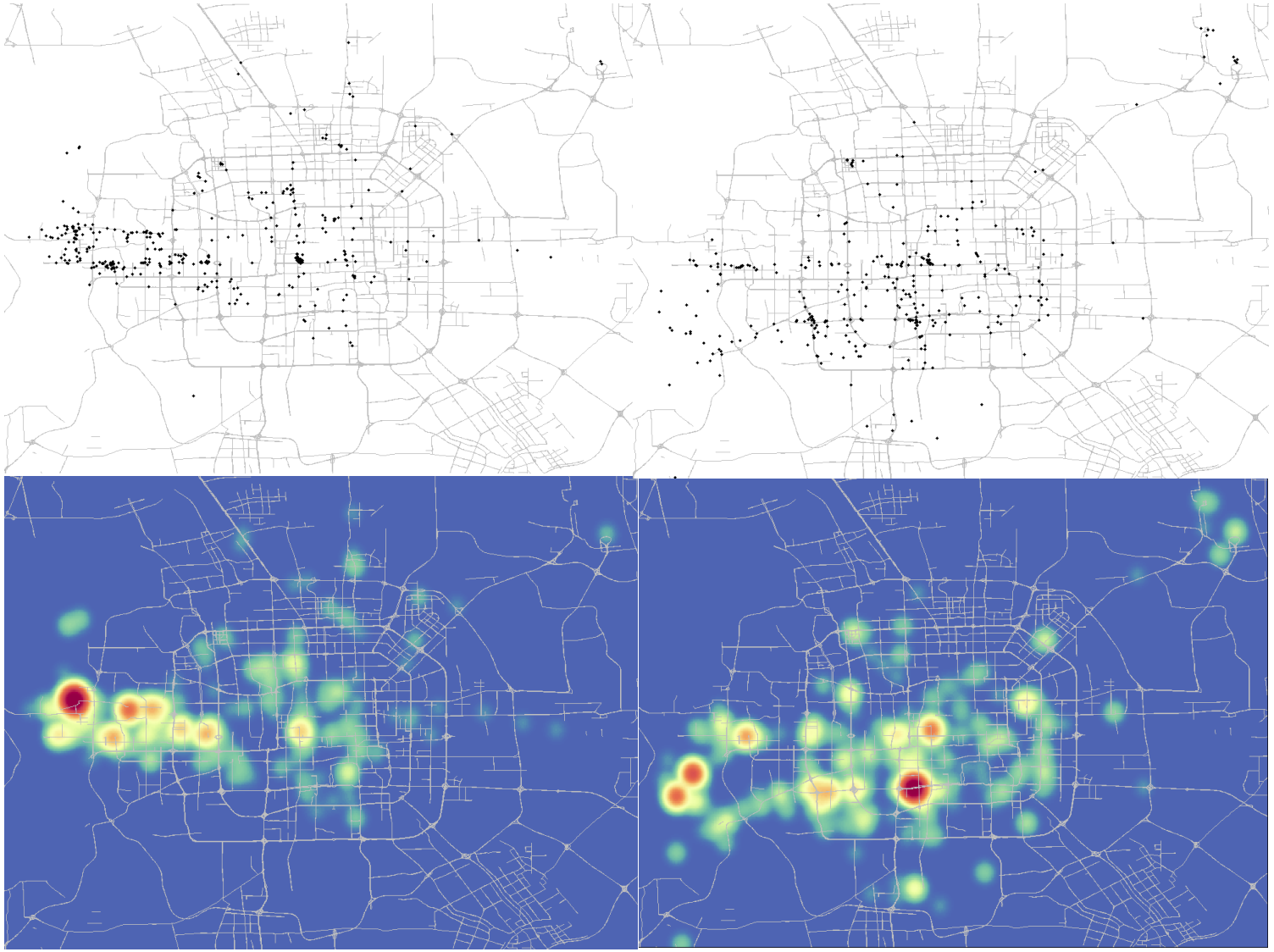
Factor graph, Exponential model




For more: http://arnetminer.org/expertisesearch.do?keyword=Machine%20Learning


Regularity
Conformity

Regularity



93%: the limit of predictability of human mobility(Song et al. [1])

Major hubs: homes, workplaces
Minor hubs: shopping malls, gyms, and restaurants

[1]C. Song, Z. Qu, N. Blumm, and A.-L. Barabási. Limits of predictability in human mobility. Science, 327(5968):1018–1021, 2010.

Conformity

Related work

Individual models

methods	target			feature				
	CI	GPS	Wifi	SMP	TC	IT	SR	CF
NextPlace(Scellato et al.)		√	√			√		
WhereNext(Monreale et al.)		√				√		
W^4(Yuan et al.)	√			√	√	√		

Collaborative models

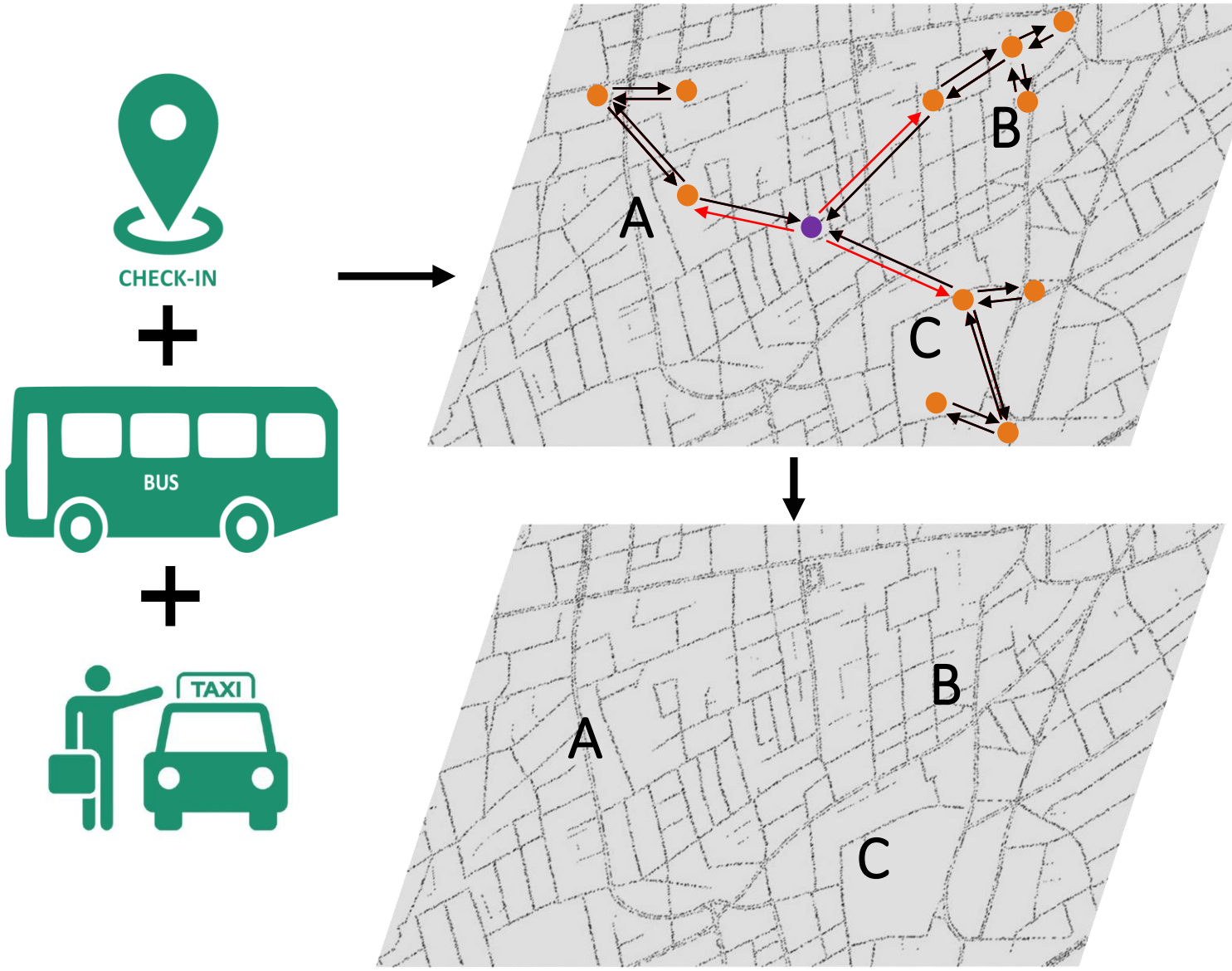
methods	target			feature				
	CI	GPS	Wifi	SMP	TC	IT	SR	CF
PSMM(Cho et al.)	√			√		√	√	
SHM(Gao et al.)	√					√	√	
gSCorr(Gao et al.)	√					√	√	
CEPR(Lian et al.)	√			√		√		√

CI: check-in, SMP: spatial mobility pattern, TC: text content, IT: individual temporal patterns, SR: social relationship, CF: collaborative filtering, HT: heterogeneous mobility datasets

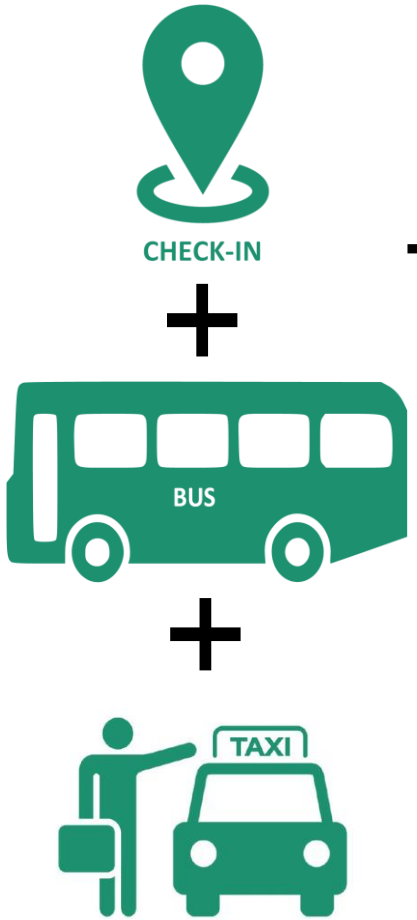
Limits:

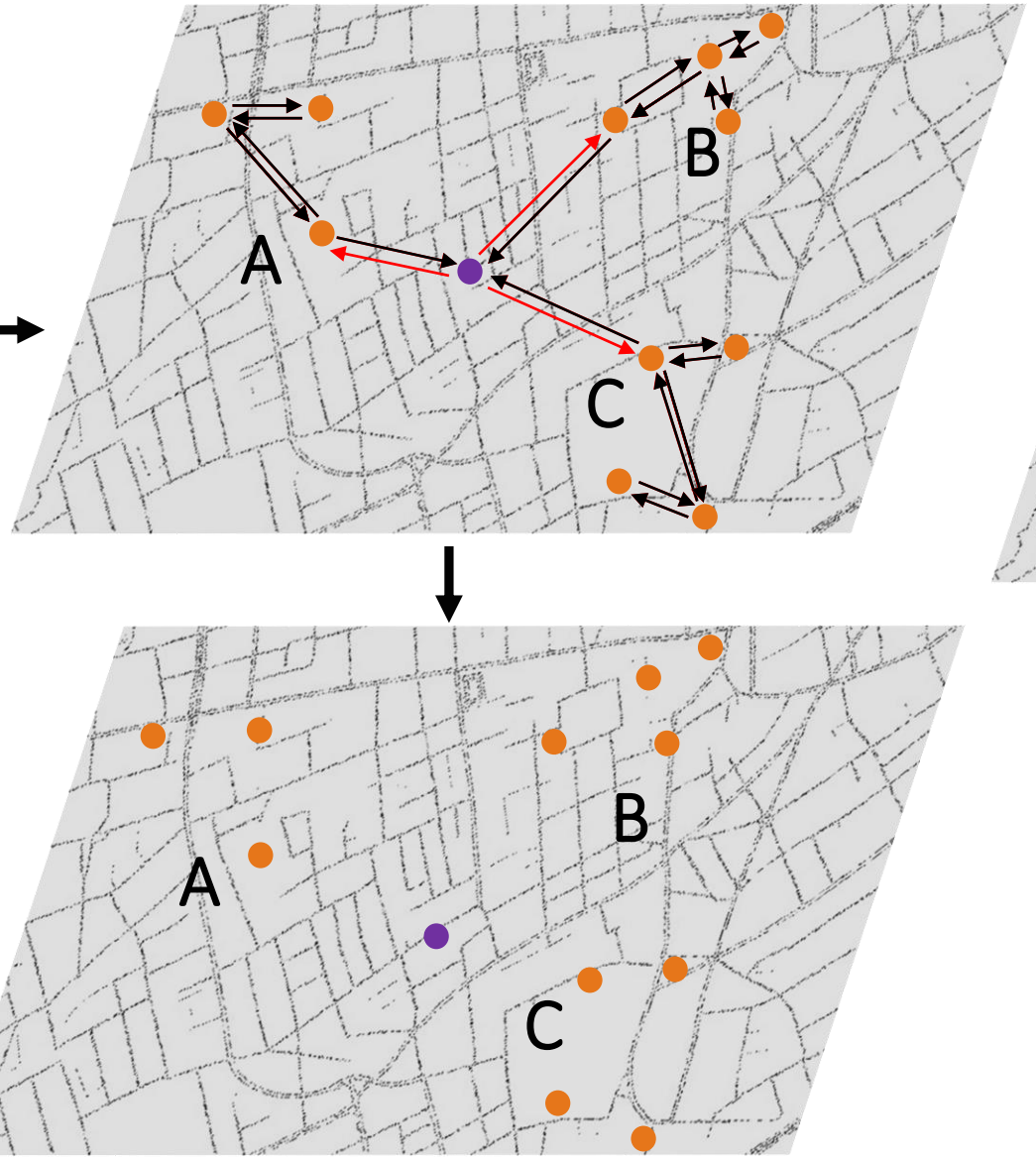
1. Failed to Incorporate both regularity and conformity of human mobility
2. Static, not time-aware
3. Homogeneous data

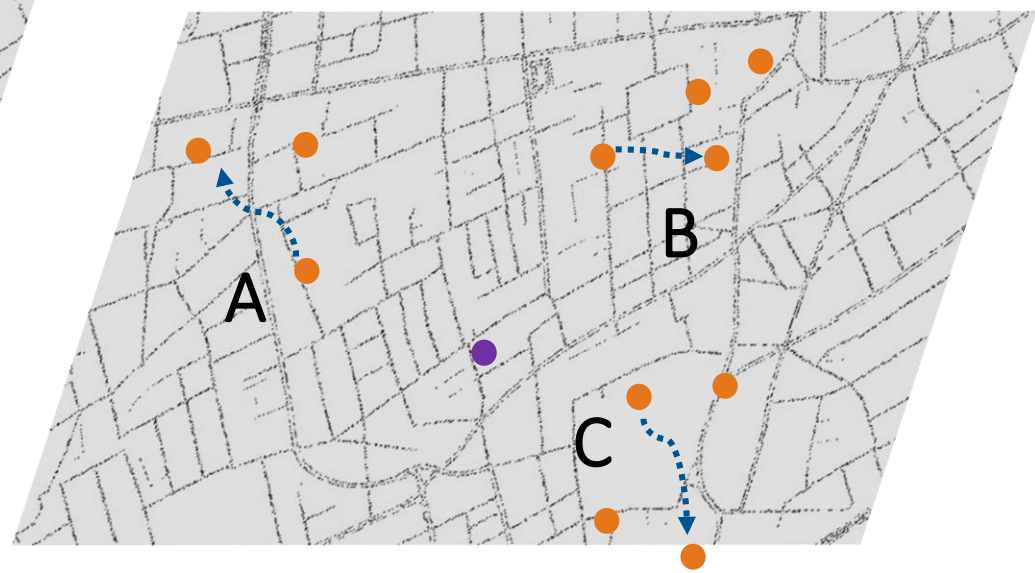
Problem definition



Problem definition







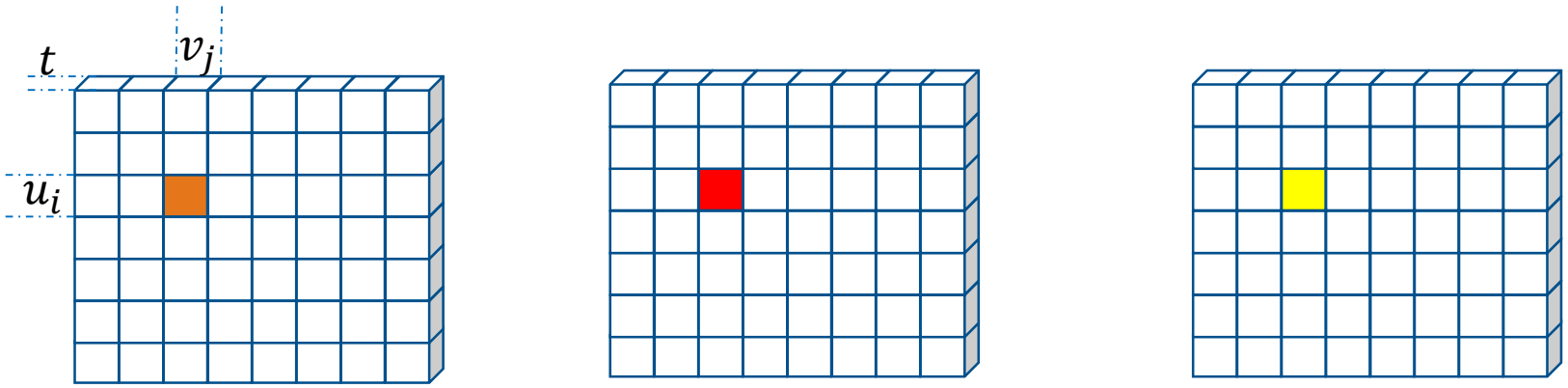
Regularity

Conformity

Main idea overview

- Split days into T time slots $\mathcal{T} = \{t_1, t_2, \dots, t_T\}$
- ***M*** users and ***N*** venues
  - $\mathcal{U} = \{u_1, u_2, \dots, u_M\}$
  - $\mathcal{V} = \{v_1, v_2, \dots, v_N\}$
- Preference matrix of $\mathcal{U}$ to $\mathcal{V}$ at time $\boldsymbol{t}$: $\mathbf{R}(t) \in \mathbb{R}^{M \times N}$

$R_{ij}(t) = R_{ij}^{(r)}(t) + R_{ij}^{(c)}(t)$

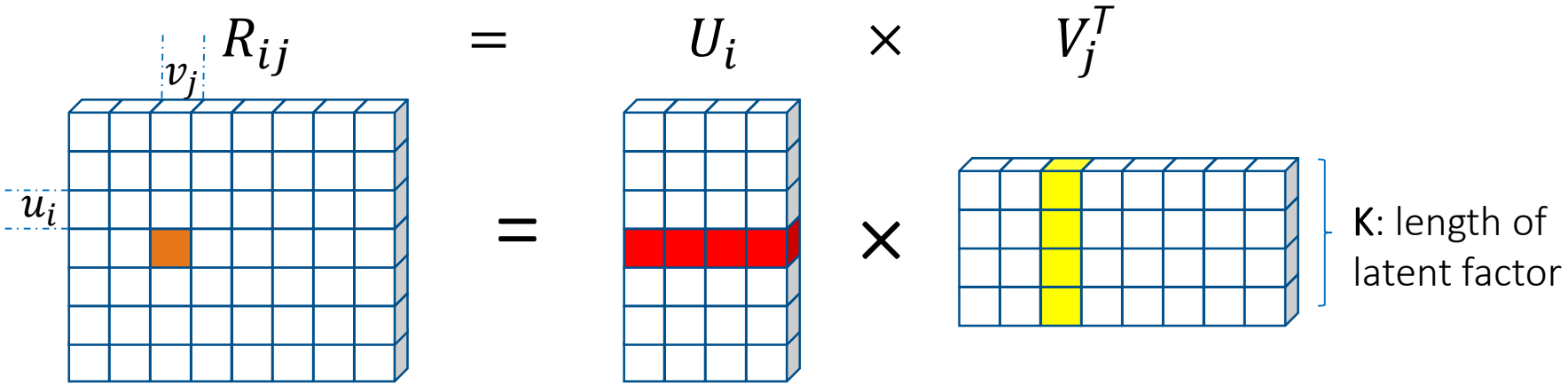


Regularity term

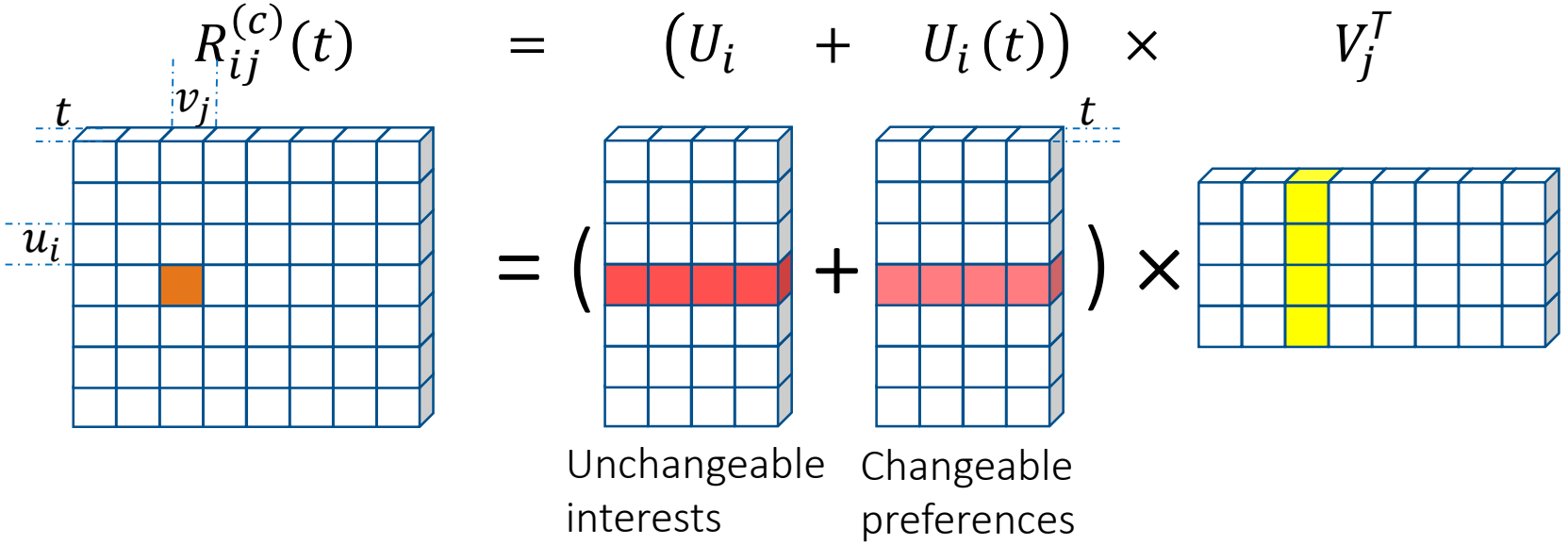
conformity term

Conformity term (check-in data)

- Traditional collaborative model: Matrix Factorization



- Time-aware Matrix Factorization



Regularity term(heterogeneous data)

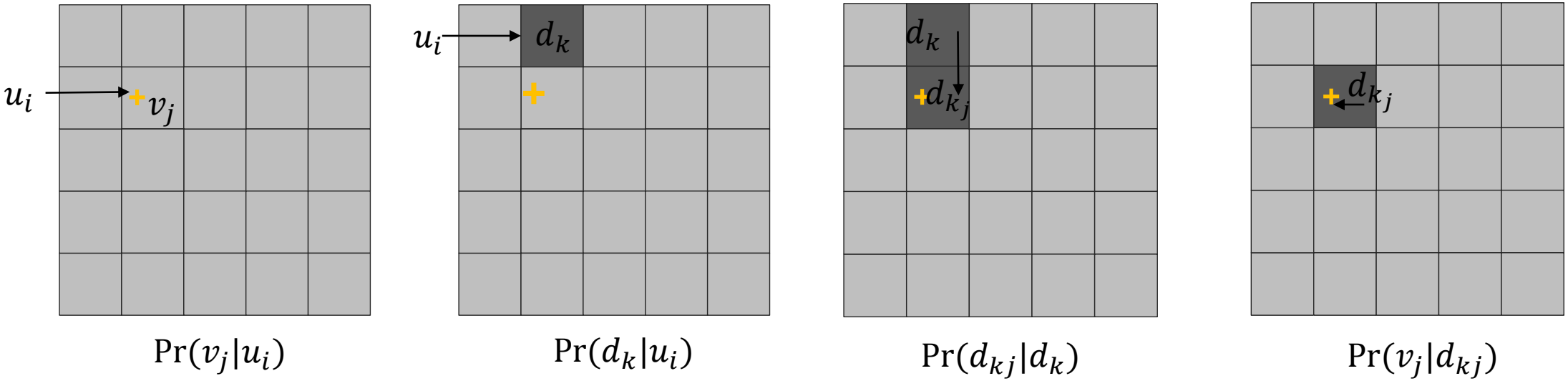
- Split the city into I grid cells: $C = \{d_1, d_2, \dots, d_I\}$
- v_j belongs to a grid d_{kj}
- u_i travels from a grid d_k to v_j

$$\Pr(v_j|u_i) \propto \sum_{k=1}^{I} \Pr(d_k|u_i) \cdot \Pr(v_j|d_k)$$

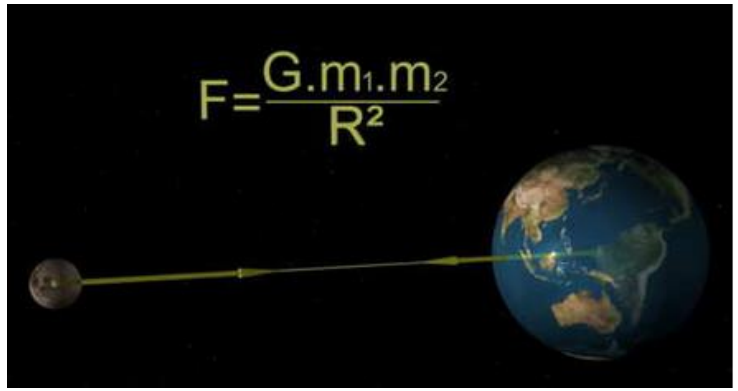
$$= \sum_{k=1}^{I} \Pr(d_k|u_i) \cdot \Pr(d_{kj}|d_k) \cdot \Pr(v_j|d_{kj})$$

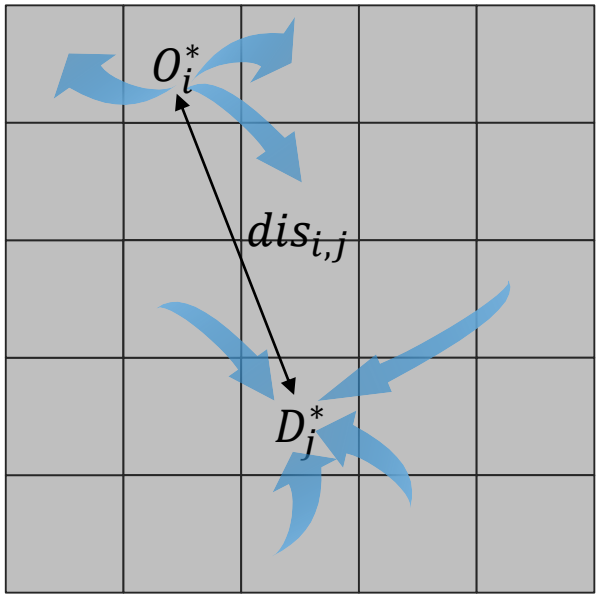
H_{ik} Q_{jk}

$$\longrightarrow R_{ij}^{(r)} = H_i \cdot Q_j^T$$



Gravity model





$$T_{i,j}^* = c\frac{(O_i^*)^a \cdot (D_j^*)^b}{\exp(r \cdot dis_{i,j})},$$

$$* \in \{B, A, C\}$$

$m_1 \rightarrow (O_i^*)^a$, O_i^*: number of individuals leaving grid d_i in data*

$m_2 \rightarrow (D_j^*)^b$, D_j^*: number of people going toward d_j in data*

$R^2 \rightarrow \exp(r \cdot dis_{i,j})$, $dis_{i,j}$: distance between d_i and d_j

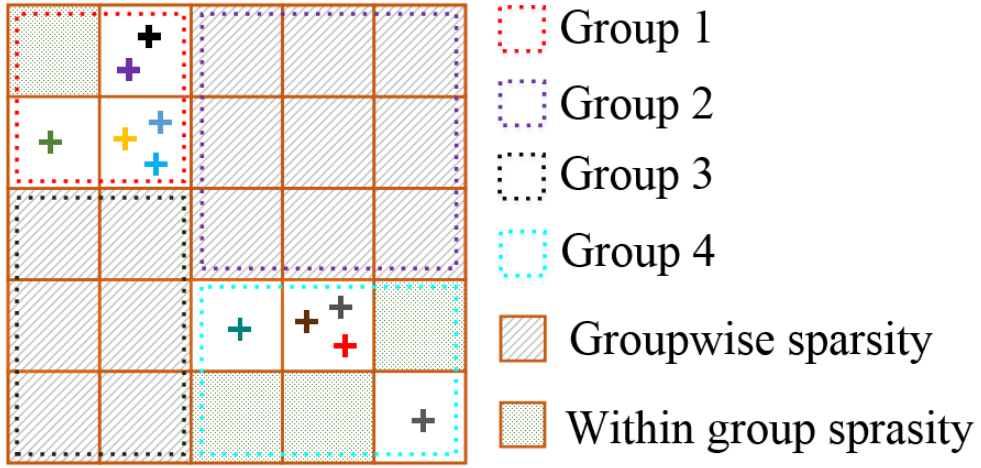
B: bus data

A: taxi data

C: check-in data

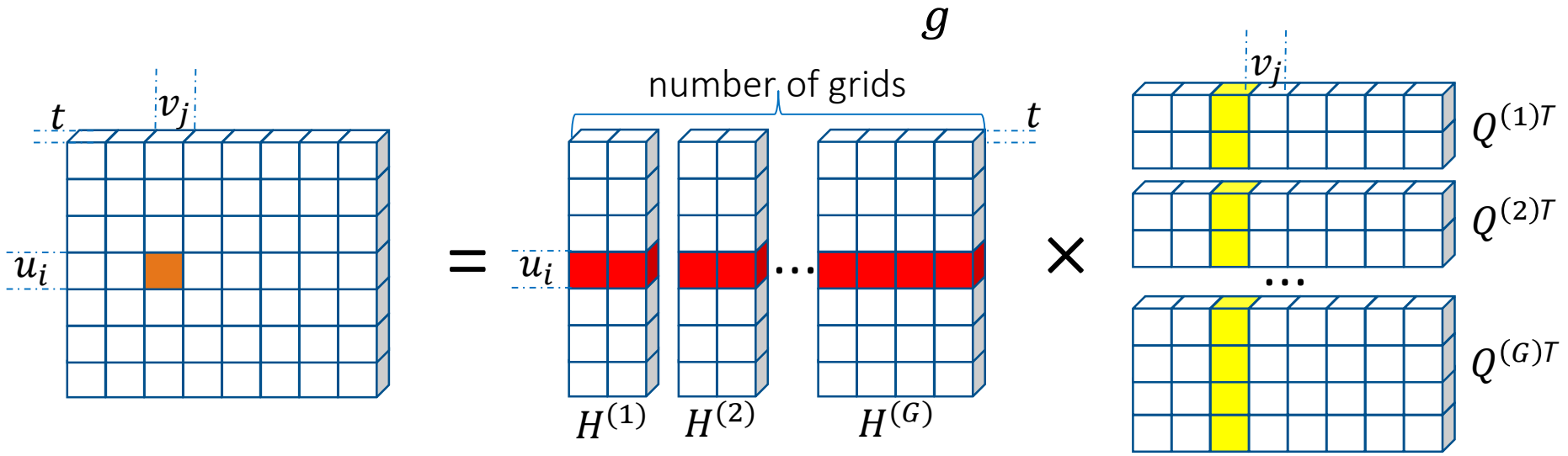
c,a,b,r: constants

Achieve the two-level sparsity



Cluster grids into G group

$$R_{ij}^{(r)} = H_i \cdot Q_j^T \longrightarrow R_{ij}^{(r)} = H_i^{(g)} \cdot Q_j^{(g)T}$$



RCH model

- **Sparse group lasso**
- Objective function:

$$\boldsymbol{\Theta}(\boldsymbol{U}, \boldsymbol{U}(t), \boldsymbol{V}, \boldsymbol{H}(t), \theta^{B}, \theta^{A})$$

$$= \sum_{t \in \mathcal{T}} \left|\left| \boldsymbol{R}(t) - \sum_{g \in \mathcal{G}} \boldsymbol{H}^{(g)}(t) \sum_{* \in \mathcal{P}} \theta^{*} \left(\boldsymbol{Q}^{*(g)}\right)^{T} - (\boldsymbol{U} + \boldsymbol{U}(T))\boldsymbol{V}^{T} \right|\right|_F^2$$

$$+ \sum_{t \in \mathcal{T}} \left((1-\alpha)\sigma \sum_{j=1}^{M} \sum_{g \in \mathcal{G}} \left\| \boldsymbol{H}_j^{(g)}(t) \right\|_2 + \alpha\sigma \sum_{j=1}^{M} \left\| \boldsymbol{H}_j(t) \right\|_1 \right)$$

$$+ \gamma(\|\boldsymbol{U}\|_F^2 + \|\boldsymbol{V}\|_F^2) + \beta \sum_{t \in T} \|\boldsymbol{U}(t)\|_F^2,$$

where $* \in \mathcal{P} = \{A, B, C\}$ and $\theta^{C} = 1$.

Offers group-wise sparsity

Offers within-group sparsity

- Optimization: alternative minimization

Heterogeneous mobility data

Data Set	Check-in(Sina Weibo)	Bus Data	Taxi Data
City	Beijing	Beijing	Beijing
Scale of Data	12,133,504 check-ins	3,000,000 bus-trips	19,400,000 taxi transitions
Period	Mar. 2011 to Sep. 2013	Aug. 2012 to May 2013	Mar. 2011 to Aug. 2011
Content	user ID, check-in time, venue Id, venue's geo-coordinates	card Id, alighting time, boarding and alighting stops	times, geo-coordinates of boarding and alighting

Experiments

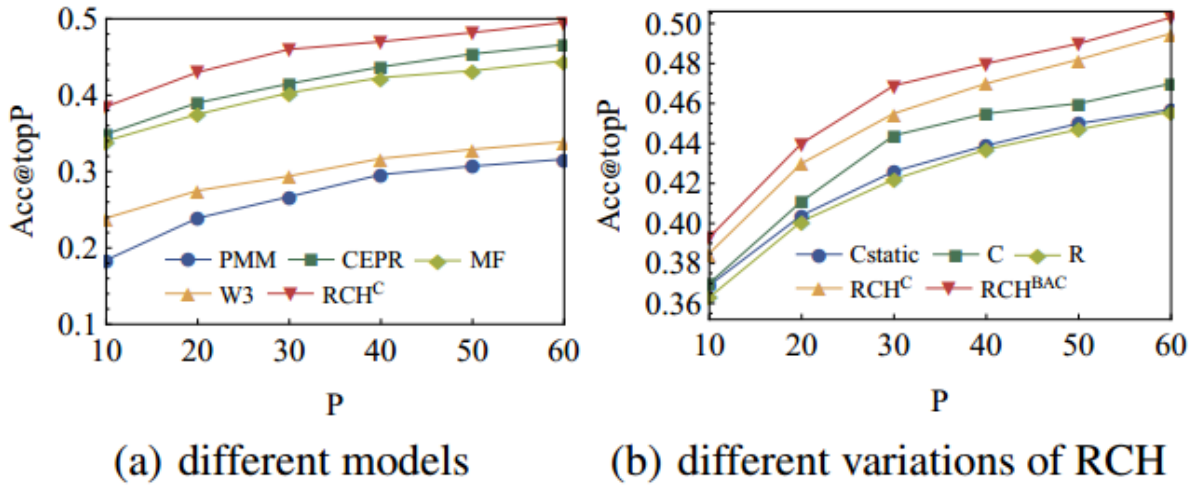
- **Baselines**
  - **MF**(Most Frequent Model)
    - Calculate the frequencies of users' check-ins
  - **PMM** (Periodic Mobility Model)
    - 2-dimentional (home, work)
    - Time-independent spatial Gaussian Mixture
  - $\mathbf{W^3}$(Who, When, Where)
    - Probabilistic model
  - **CEPR**
    - Human mobility: regular and novel ones

Divide the check-in data into two parts by time order: training part and testing part (7:3)

Metrics:
Acc@topP: prediction accuracy for prediction list with length P
APR(average percentile rank): for the actually visited venues

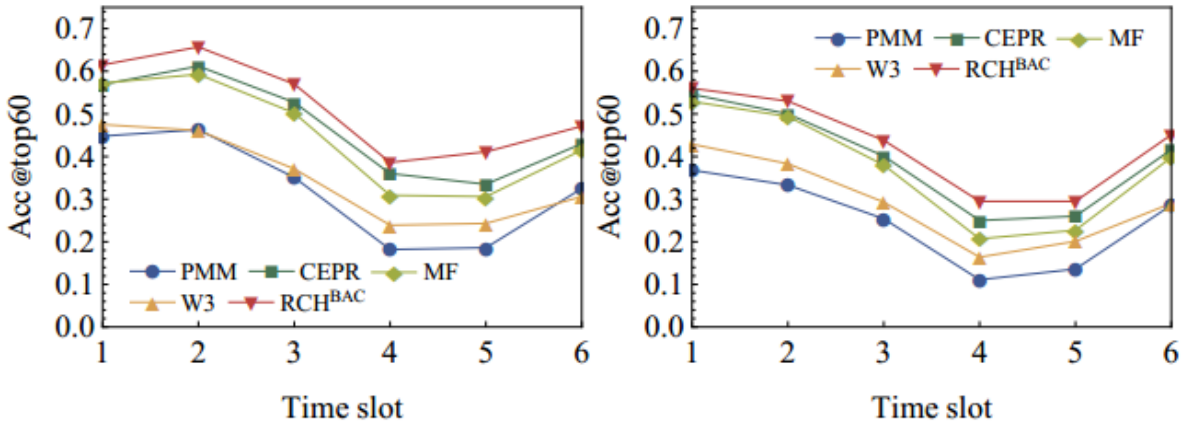
Parameters:
Default values of $\{\theta^A, \theta^B, \gamma, \beta, \alpha, \sigma\}$ are $\{1, 1, 0.005, 0.005, 0.95, 10^{-5}\}$

Results



(a) different models (b) different variations of RCH

Acc@top*P*



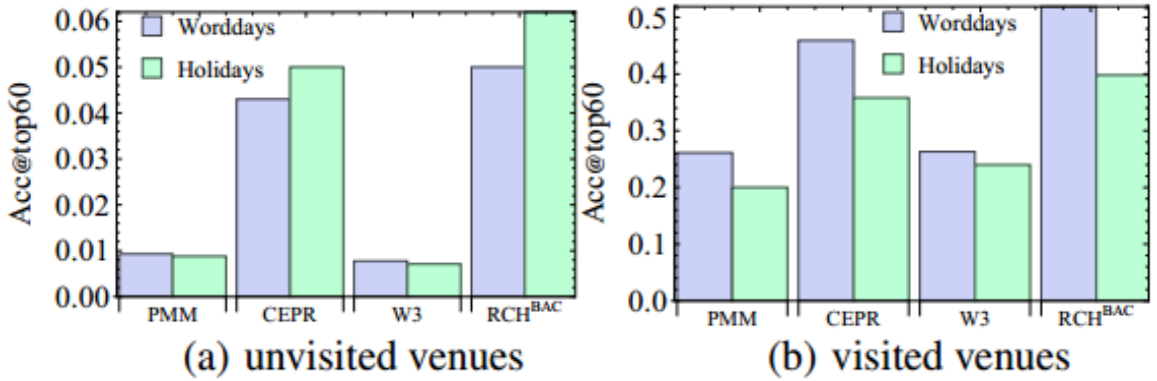
(a) workdays (b) holidays

Acc@top*P* for different type of days and time slots

$\mathrm{RCH}^{\mathrm{C}}$ shows improvements over all the 4 baselines;
$\mathrm{RCH}^{\mathrm{BAC}}$ has better performance than $\mathrm{RCH}^{\mathrm{C}}$;

Workdays have better prediction accuracies than holidays;
The time slot 2 of workdays has the highest accuracy;

Results



Acc@top*60* of visited and unvisited venues

Models	Workdays						Holidays					
t	12-4am	4-8am	8am-12pm	12-4pm	4-8pm	8pm-12am	12-4am	4-8am	8am-12pm	12-4pm	4-8pm	8pm-12am
C_{static}	0.884	0.899	0.865	0.799	0.801	0.832	0.872	0.848	0.807	0.753	0.761	0.840
C	0.885	0.908	0.869	0.820	0.825	0.848	0.868	0.854	0.818	0.781	0.787	0.844
R	0.887	0.884	0.873	0.826	0.831	0.860	0.859	0.843	0.814	0.768	0.790	0.837
RCH^{C}	0.896	0.911	0.880	**0.835**	0.838	0.870	0.881	0.859	0.823	0.781	0.793	0.849
RCH^{BAC}	**0.899**	**0.912**	**0.883**	**0.835**	**0.840**	**0.871**	**0.890**	**0.863**	**0.829**	**0.786**	**0.795**	**0.850**

APR of our models in different time slots

CEPR and $\mathbf{RCH^{BAC}}$: outperform PMM and $\mathbf{W^3}$ apparently for unvisited venues benefit from collaborative filtering;
CEPR and $\mathbf{RCH^{BAC}}$: accuracy of unvisited venues on holidays is higher than workdays;
$\mathbf{RCH^{ABC}}$ has highest APR in our models in different time slots;

Summary

- Integrate both the regularity and conformity of human mobility
- Provide a time-aware collaborative model
- Incorporate heterogeneous mobility data into prediction model
- Learn spatial influence and group structure based on gravity model and sparse group Lasso
- RCH model: significantly outperforms existing approaches

Yingzi Wang

Microsoft Research


可以啊，我负责出钱就行了
女
生
祝
MobiSense
乐
节
快