Inferring User Activity and Mobility in Location-based Social Networks

Presenter: Yu-Wen Wang
Advisor: Wen-Chih Peng

Outline

- Introduction
- Related Works
- Problem
- Solutions
- Experiment
- Conclusion

Motivations

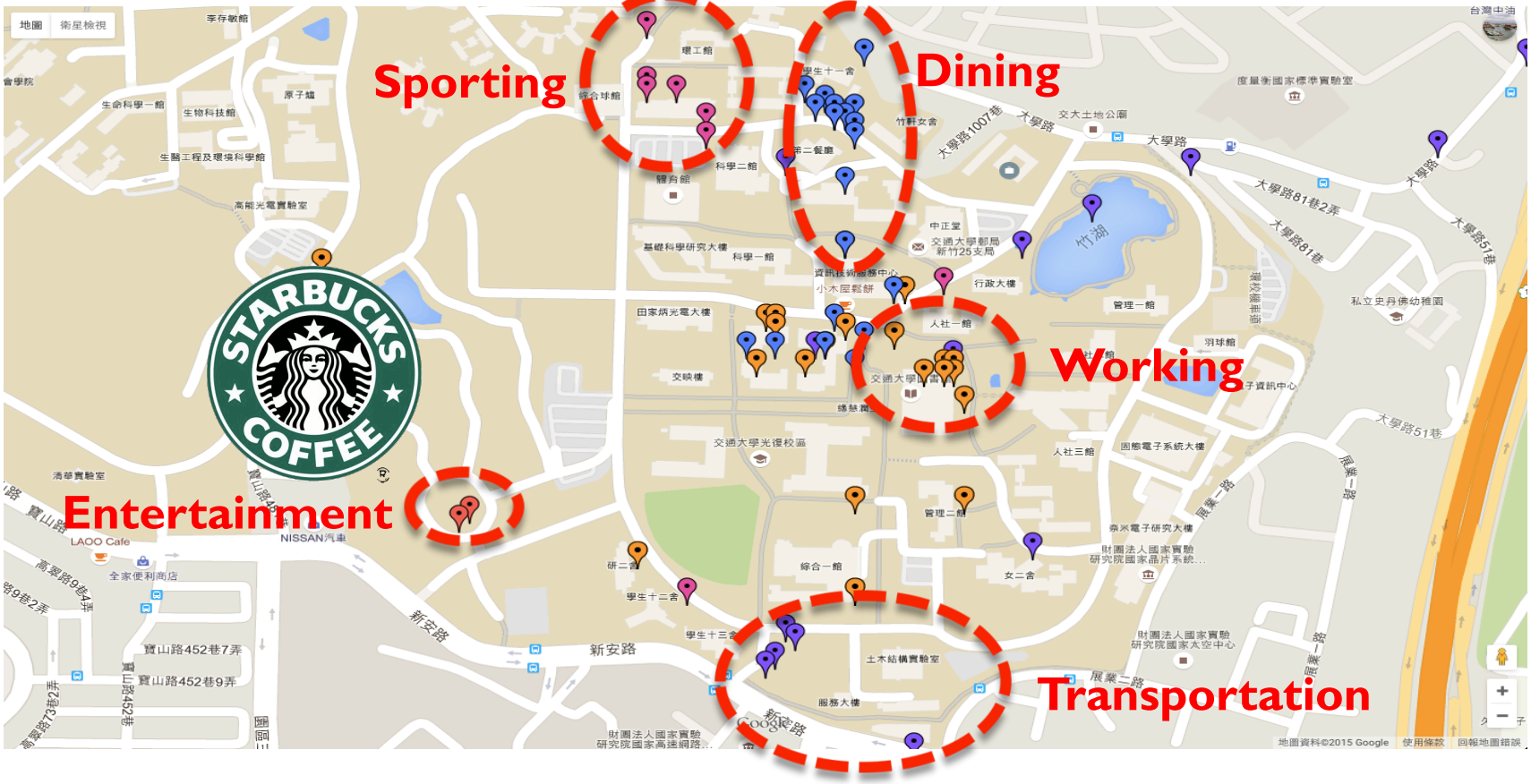
- People use smartphones almost everyday and everywhere
- With the check-in data from LBSN, the location data can be easily collected

Gowalla
foursquare
brightkite
myTown
gogshot
yelp

Applications

- Personalized location-based services
  - Coupons & Discount
  - Advertisements
  - Spreading business information

Observation

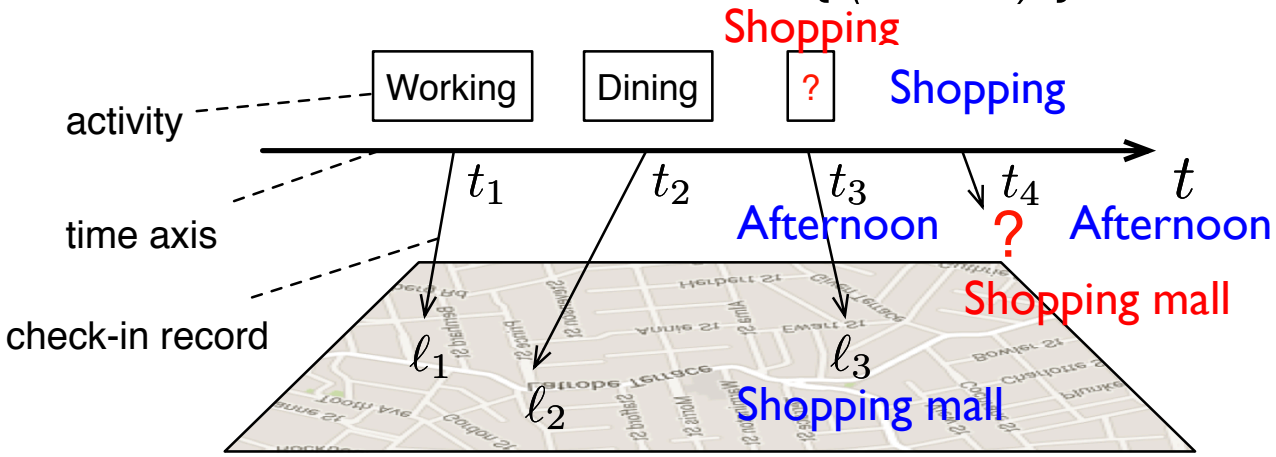


What is the user doing?

Where is the user?

Problem Definition

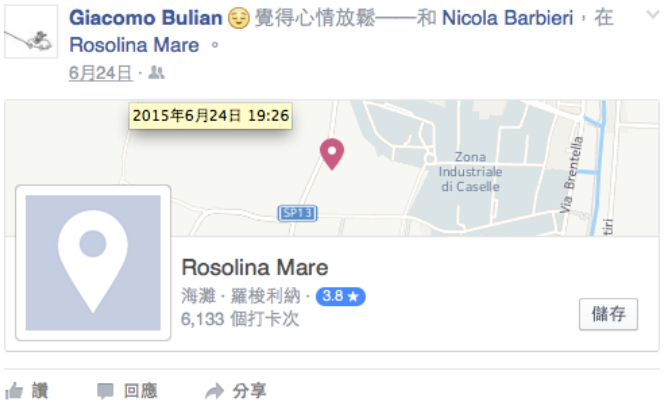
- Goal
  - Check-in Records in an LBSN : $C = \{ (u, a, l, t) \}$



  - **Infer user activity *P(a| u, l, t)***
    - Input: Location data and time stamp
    - Output: Activity
  - **Infer user mobility *P(l| u, a, t)***
    - Input: Activity Label and time stamp
    - Output: Location

Challenge Issue

- Data Insufficient
  - Lack of temporal information
    - Solution: Activity Transition Model
  - Lack of geographical information
    - Solution: Gaussian Mixture Model



- How to build a model that could infer both user activity and mobility?
  - Solution: Bayesian Network

Related Works

- Problem: Infer location P(l | t, u, s) Input t and act

$$\alpha \times$$ P(t | l, u) 

x P(l | u)

+

$$(1-\alpha) \times$$ P(t | l, s) 

x P(l | s)

Lack of handling new location

[1]Huiji Gao et al. Modeling temporal effects of human mobile behavior on location-based social networks (CIKM '13)

Related Works (cont.)

- Problem: Infer location

$$P(x(t)=x) = P(x_u(t)=x|c_u(t)=H) \cdot P(c_u(t)=H) + P(x_u(t)=x|c_u(t)=W) \cdot P(c_u(t)=W)$$

There are only two states

Input: Activity

[2]Eunjoon Cho et al. Friendship and mobility: user movement in location-based social networks (KDD '11)

Related Works (cont.)

- Problem: Predict location
  - Input: ~~A sequence of check-ins~~ Different: A check-in point

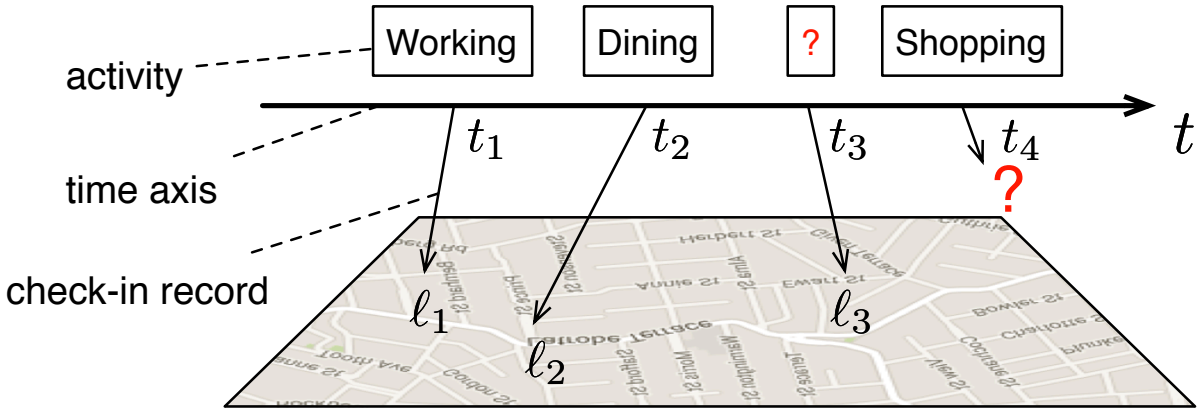
- Check-in count
- User count
- User count X Check-in count
- Max check-in count by user

Different: Predict **Next** step / Infer location

[3]Cheng Hong et al. What's Your Next Move: User Activity Prediction in Location-based Social Networks (SDM '13)

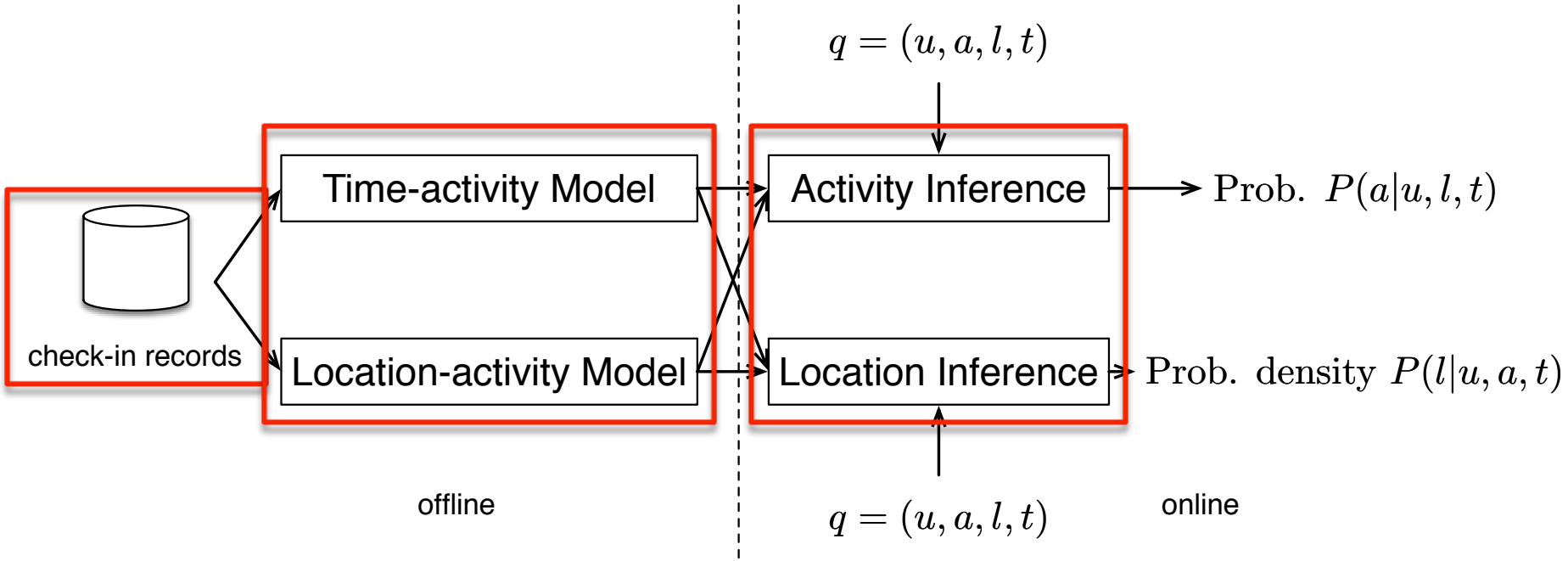
Problem Definition

- Goal
  - Check-in Records in an LBSN : $C = \{ (u, a, l, t) \}$



  - **Infer user activity *P(a| u, l, t)***
    - Input: Location data and time stamp
    - Output: Activity
  - **Infer user mobility *P(l| u, a, t)***
    - Input: Activity Label and time stamp
    - Output: Location

Framework


$q = (u, a, l, t)$
Time-activity Model
Activity Inference
Prob. $P(a|u, l, t)$
check-in records
Location-activity Model
Location Inference
Prob. density $P(l|u, a, t)$
offline
$q = (u, a, l, t)$
online

The Network Between Features

- Do regular activities
- Visit a place based the activity

18:00 Drive home

But! 18:00 Working

Network Based Activity Inference Model

- Activity is modeled by Bayesian Network

$$P(a|u,\ell,t) = \frac{P(a,\ell,t|u)}{P(\ell,t|u)}$$

$$\approx \frac{P(\ell|a,u)P(a|t,u)P(t|u)}{\sum_{a'} P(\ell|a',u)P(a'|t,u)P(t|u)}$$

$$= \frac{P(\ell|a,u)P(a|t,u)}{\sum_{a'} P(\ell|a',u)P(a'|t,u)}$$

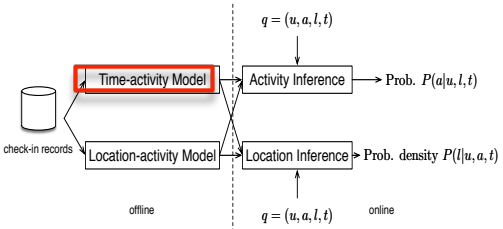
Location-Activity Correlation

Gaussian Mixture Model

Time-Activity Correlation

Activity Transition Model

Time-Activity Correlation



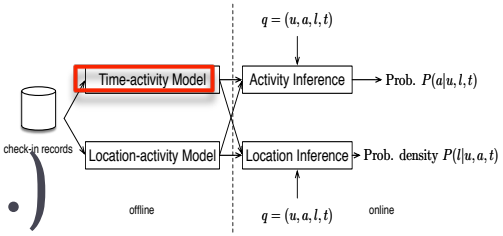
- Issue: Data Sparse Problem
  - About 50% / 80% users have 30 / 50 check-in records or less
  - Check-in at a time point
- Example

One day

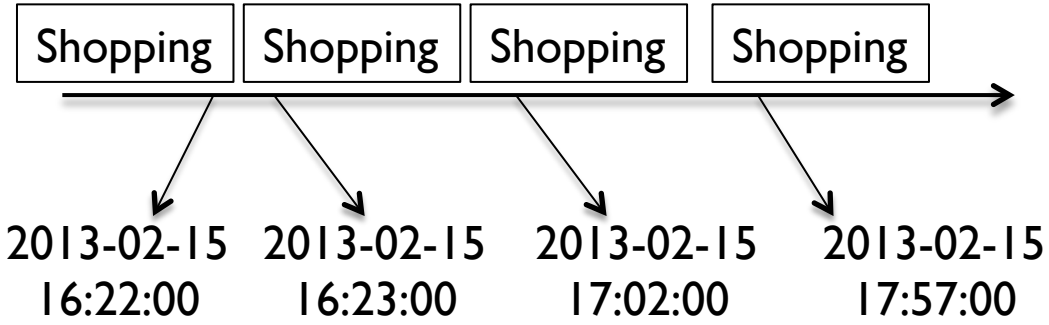
	Slot 1	Slot 2	Slot 3	Slot 4	Slot 5	Slot 6	Slot 7
Activity 1	5	3		4			
Activity 2	2		5		1		

  - There are many blanks in the matrix
- Solution: Activity Transition Model

Time-Activity Correlation (Cont.)



- Why are there many blanks in the matrix?



- 2 time slots

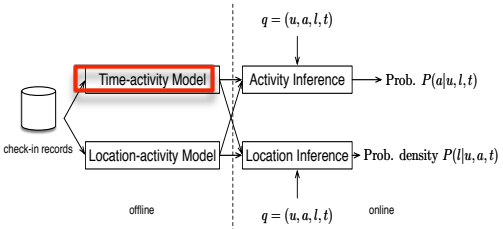
	Slot 1: 16:00~16:59	Slot 2: 17:00~17:59
Shopping	2	2

- 4 time slots

	Slot 1: 16:00~16:29	Slot 2: 16:30~16:59	Slot 3: 17:00~17:29	Slot 4: 17:30~17:59
Shopping	2		1	1

More slots, more blanks!

Activity Transition Model

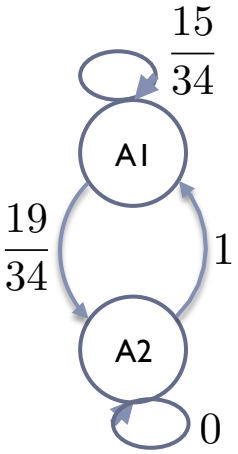


- Goal: To find the transition probability between activities

$$M(a_p, a_f) = \frac{\Sigma_{i=1:24n} F(a_p, s_{i-1}) \times F(a_f, s_i)}{\Sigma_{a_f \in ACT} \Sigma_{i=1:24n} F(a_p, s_{i-1}) \times F(a_f, s_i)}$$

- Example

	Slot 1	Slot 2	Slot 3	Slot 4	Slot 5	Slot 6	Slot 7
Activity 1	5	3	0	4	0	0	0
Activity 2	2	0	5	0	1	0	0



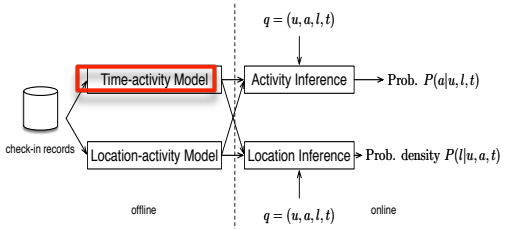
$$M(A1, A1) = 5 \times 3 + 3 \times 0 + 0 \times 4 + 4 \times 0 + 0 \times 0 + 0 \times 0 = 15 / (15 + 19)$$

$$M(A1, A2) = 5 \times 0 + 3 \times 5 + 0 \times 0 + 4 \times 1 + 0 \times 0 + 0 \times 0 = 19 / (15 + 19)$$

$$M(A2, A1) = 1$$

$$M(A2, A2) = 0$$

Time-Activity Correlation



- Goal: Fulfill the blanks

$$C(a_f, s_i) = \frac{F(a_f, s_i) + \Sigma_{a_p \in ACT} F(a_p, s_{i-1}) \times M(a_p, a_f)}{\Sigma_{a_f \in ACT}(F(a_f, s_i) + \Sigma_{a_p \in ACT} F(a_p, s_{i-1}) \times M(a_p, a_f))}$$

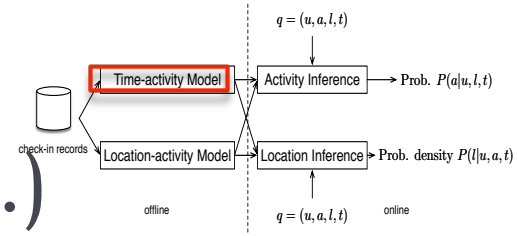
- Example

	Slot 1	Slot 2	Slot 3	Slot 4	Slot 5	Slot 6	Slot 7
Activity 1	5	3		4			
Activity 2	2		5		1		

Step 1: Normalize

	Slot 1	Slot 2	Slot 3	Slot 4	Slot 5	Slot 6	Slot 7
Activity 1	5/7	3/3	0	4/4	0	0	0
Activity 2	2/7	0	5/5	0	1/1	0	0

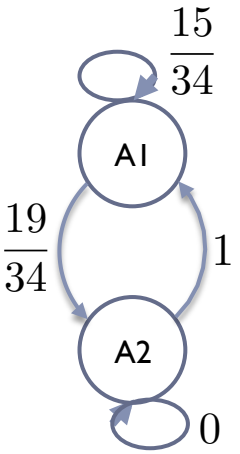
Time-Activity Correlation (cont.)



- Example (cont.)

$$C(a_f, s_i) = \frac{F(a_f, s_i) + \Sigma_{a_p \in ACT} F(a_p, s_{i-1}) \times M(a_p, a_f)}{\Sigma_{a_f \in ACT}(F(a_f, s_i) + \Sigma_{a_p \in ACT} F(a_p, s_{i-1}) \times M(a_p, a_f))}$$

	Slot 1	Slot 2	Slot 3	Slot 4	Slot 5	Slot 6	Slot 7
Activity 1	5/7	1	0	1	0	0	0
Activity 2	2/7	0	1	0	1	0	0



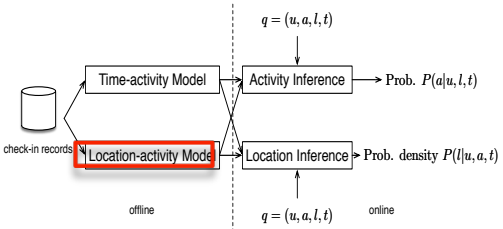
Step 2: Revise the original blanks / Fulfill the blanks

$$C(A1, S2) = \frac{1 + \frac{5}{7} \times \frac{15}{34} + \frac{2}{7} \times 1}{(1 + \frac{5}{7} \times \frac{15}{34} + \frac{2}{7} \times 1) + (0 + \frac{5}{7} \times \frac{19}{34} + \frac{2}{7} \times 0)}$$

$$C(A2, S2) = \frac{0 + \frac{5}{7} \times \frac{19}{34} + \frac{2}{7} \times 0}{(1 + \frac{5}{7} \times \frac{15}{34} + \frac{2}{7} \times 1) + (0 + \frac{5}{7} \times \frac{19}{34} + \frac{2}{7} \times 0)}$$

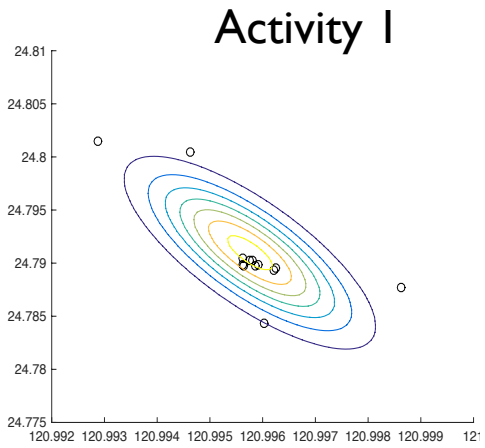
	Slot 1	Slot 2	Slot 3	Slot 4	Slot 5	Slot 6	Slot 7
Activity 1	0.71	0.80	0.28	0.91	0.25	0.86	0.52
Activity 2	0.29	0.20	0.72	0.09	0.75	0.14	0.48

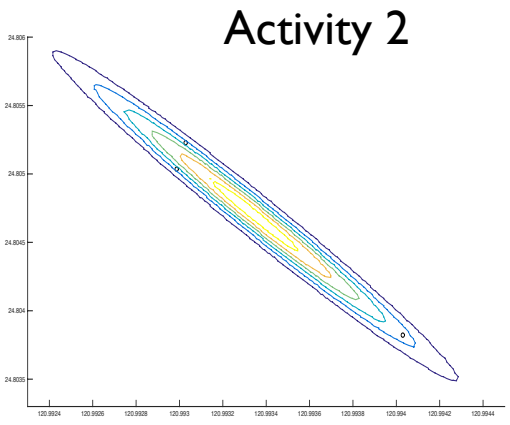
Location-Activity Correlation

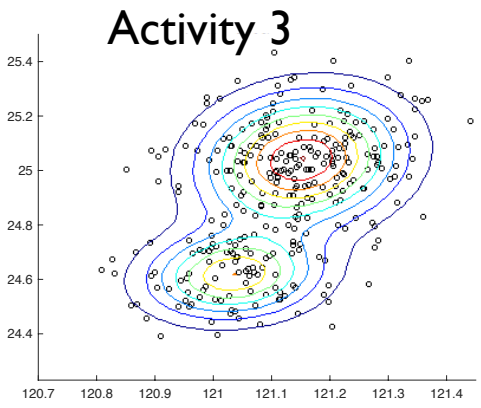


- Issue
  - Data Sparse Problem
  - Functionality

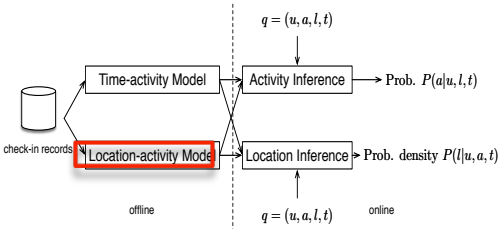
- Solution: Gaussian Mixture Models for each activity type





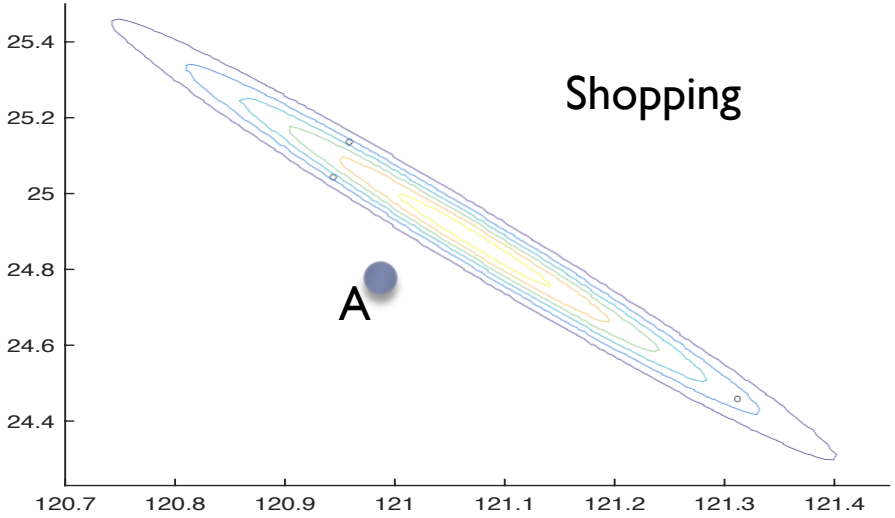


Location-Activity Correlation



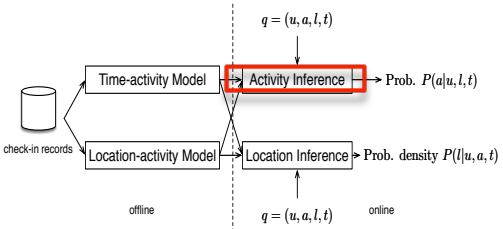
$$P(\ell|a,u) = \Sigma_{i=1}^{k} \lambda_i \mathcal{N}(\mu_i, \Sigma_i)$$

- Example



$$P(L = A|ACT = "Transportation") = \Sigma_{i=1}^{1} \lambda_i \mathcal{N}(A|\mu_{T_i}, \Sigma_{T_i}) = 0.385$$
$$P(L = A|ACT = "Shopping") = \Sigma_{i=1}^{1} \lambda_i \mathcal{N}(A|\mu_{SH_i}, \Sigma_{SH_i}) = 0.013$$

Network Based Activity Inference Model



$$P(a|u,l,t) \approx P(l|a,u)P(a|t,u)$$

- Example (cont.)
  - Time-Activity Correlation

	Slot 1	Slot 2	Slot 3	Slot 4	Slot 5	Slot 6	Slot 7
Activity 1	0.71	0.80	0.28	0.91	0.25	0.86	0.52
Activity 2	0.29	0.20	0.72	0.09	0.75	0.14	0.48

  - Location-Activity Correlation

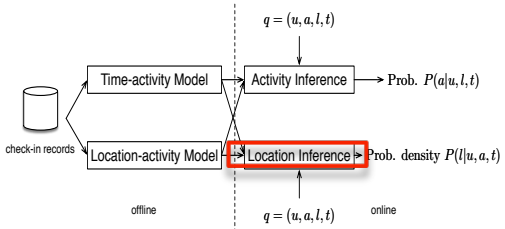
$$P(L = A|ACT = "Activity1") = 0.385$$
$$P(L = A|ACT = "Activity2") = 0.013$$

  - A user at location A in time slot 3…

$$P(ACT = Activity1|L = A, S = Slot3) = 0.385 \times 0.28 = 0.1078$$

$$P(ACT = Activity2|L = A, S = Slot3) = 0.013 \times 0.72 = 0.00936$$

Mobility Inference Model



$$P(\ell|a,t,u) = \frac{P(\ell,a,t|u)}{P(a,t|u)}$$

$$\approx \frac{P(\ell|a,u)P(a|t,u)P(t|u)}{\int_{\ell'} P(\ell'|a,u)P(a|t,u)P(t|u)\mathrm{d}\ell'}$$

$$= \frac{P(\ell|a,u)}{\int_{\ell'} P(\ell'|a,u)\mathrm{d}\ell'} = P(\ell|a,u)$$

Location-Activity Correlation

$$P(\ell|a,u) = \Sigma_{i=1}^{k} \lambda_i \mathcal{N}(\mu_i, \Sigma_i)$$

Discussion Summary

- How to build a model inferring both user activity and mobility?
  - Solution: Consider features as a Bayesian network
    - Temporal condition will trigger users doing some activity
    - What activity the user doing will make the user visits different location
- Data Insufficient Problem
  - Temporal
    - User activity may be effected by the previous one
    - Solution: Build an Activity Transition Model to fulfill the blanks
    - Pros
      - Take the order of activities into consideration
      - Closely describe user activity
  - Geographical
    - The functionality of a location are different to users
    - Solution: Build Gaussian Mixture Model
    - Pros
      - Be able to handle new location

Datasets

	GeoText[1]
# of total users	9,475
# of check-ins	377,616
# of distinct locations	46,320
Max # of check-ins of a user	301
Min # of check-ins of a user	17
Period	2010/3

[1]The dataset is also adopted by Q. Yuan et al. Who, where, when and what: discover spatio-temporal topics for twitter users. In ACM KDD, 2013.
http://www.ark.cs.cmu.edu/GeoText/

Datasets – Activity Distribution

Activity Type[1]	GeoText
Arts & Entertainment (AE)	15,892
Colleges & Universities (CU)	16,697
Events (E)	106
Food (F)	46,332
Nightlife Spots (NS)	16,207
Outdoors & Recreation (OR)	39,832
Professional & Other Places (PO)	110,108
Residences (R)	24,387
Shops & Services (SS)	86,526
Travel & Transport (TT)	21,520
	377,607

[1]https://api.foursquare.com/v2/venues/search

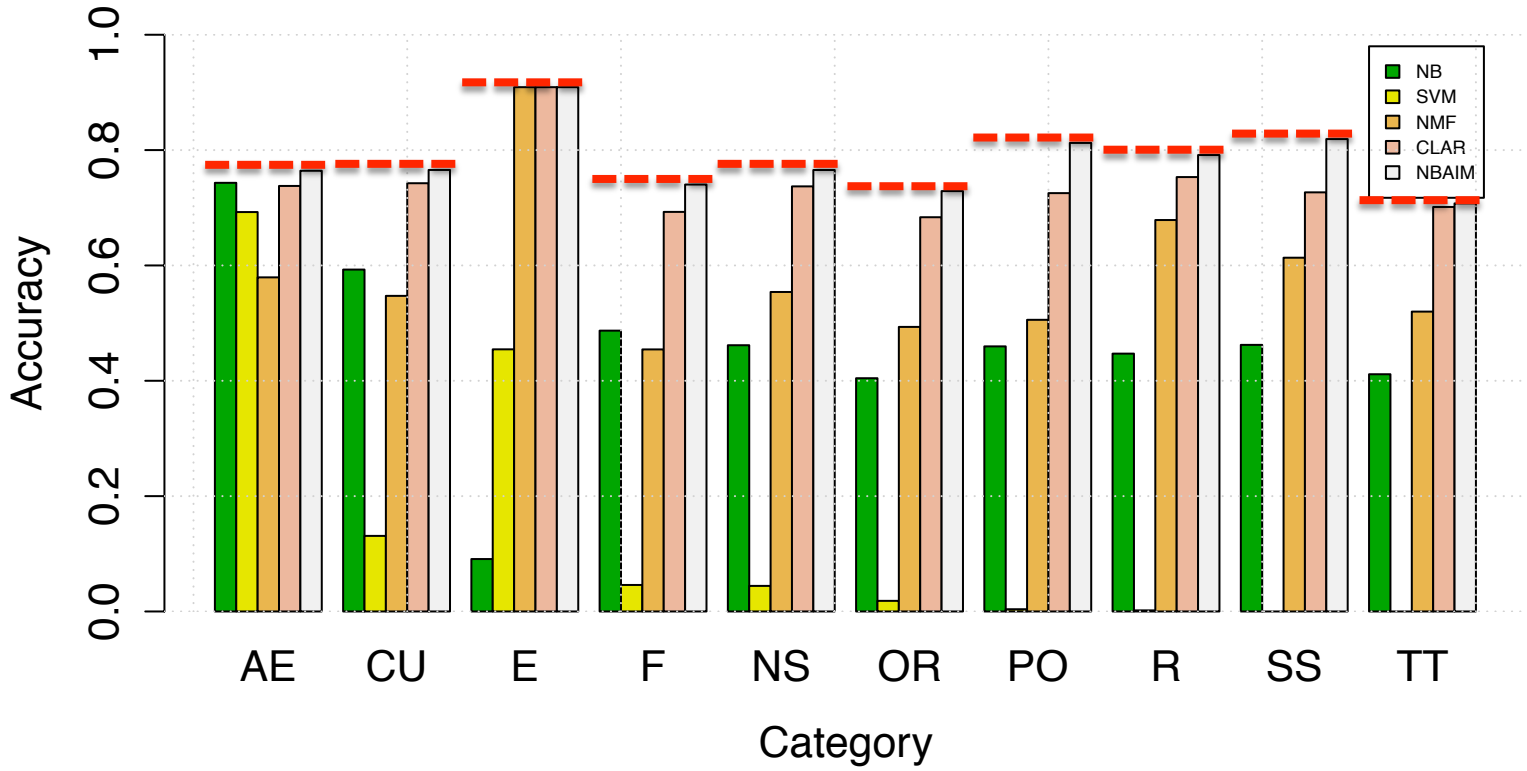
Settings – Activity Inference

- Setting
  - 80% training, 20% testing
  - Users: the users who have visited at least 5 distinct location
  - Time slot length: 1~ 4 hours
- Baseline Methods
  - Naïve Bayesian(NB)
  - Support Vector Machine(SVM)
  - Non-negative Matrix Factorization(NMF)
  - Collaborative Location Activity Recommendation(CLAR)[4]
- Measurement
  - Accuracy

[4]Vincent W. Zheng et al. Collaborative location and activity recommendations with GPS history data (WWW '10)

Single Activity Accuracy

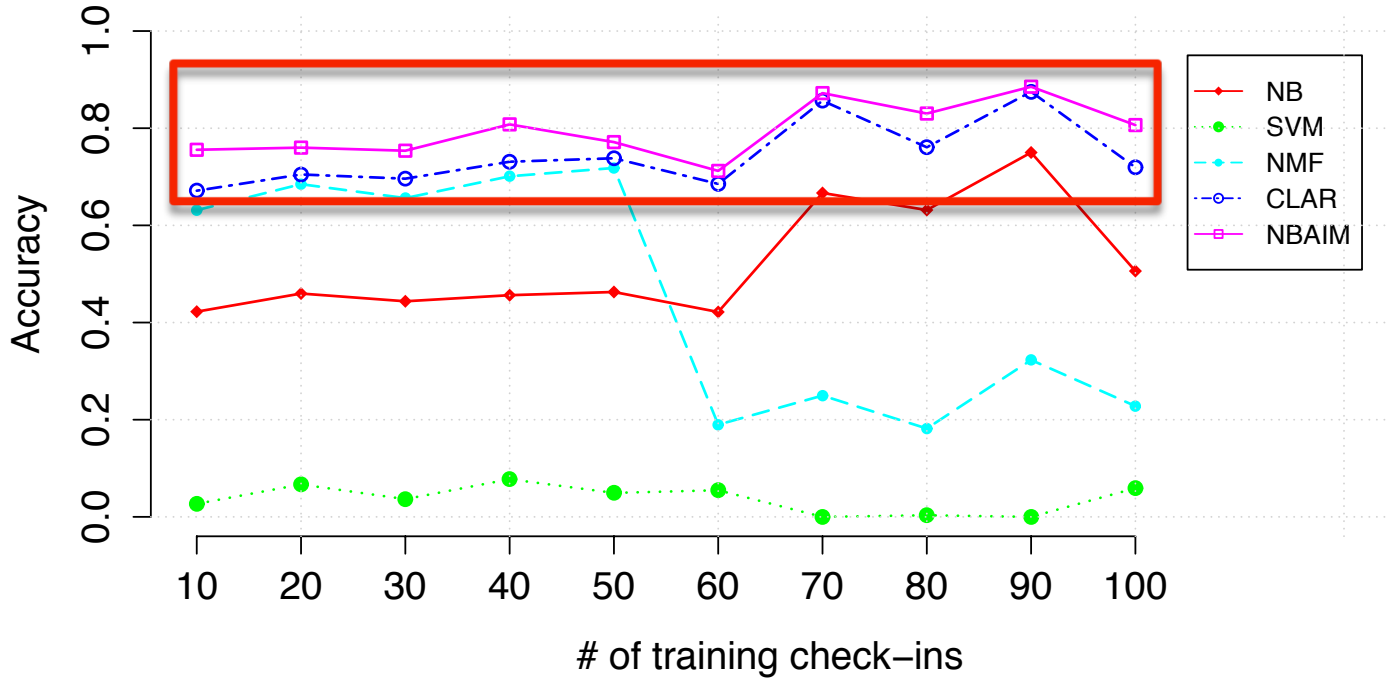
- Category vs. Accuracy



NBAIM outperform the state-of-the-art methods for each category

Compare Methods for Activity Inference

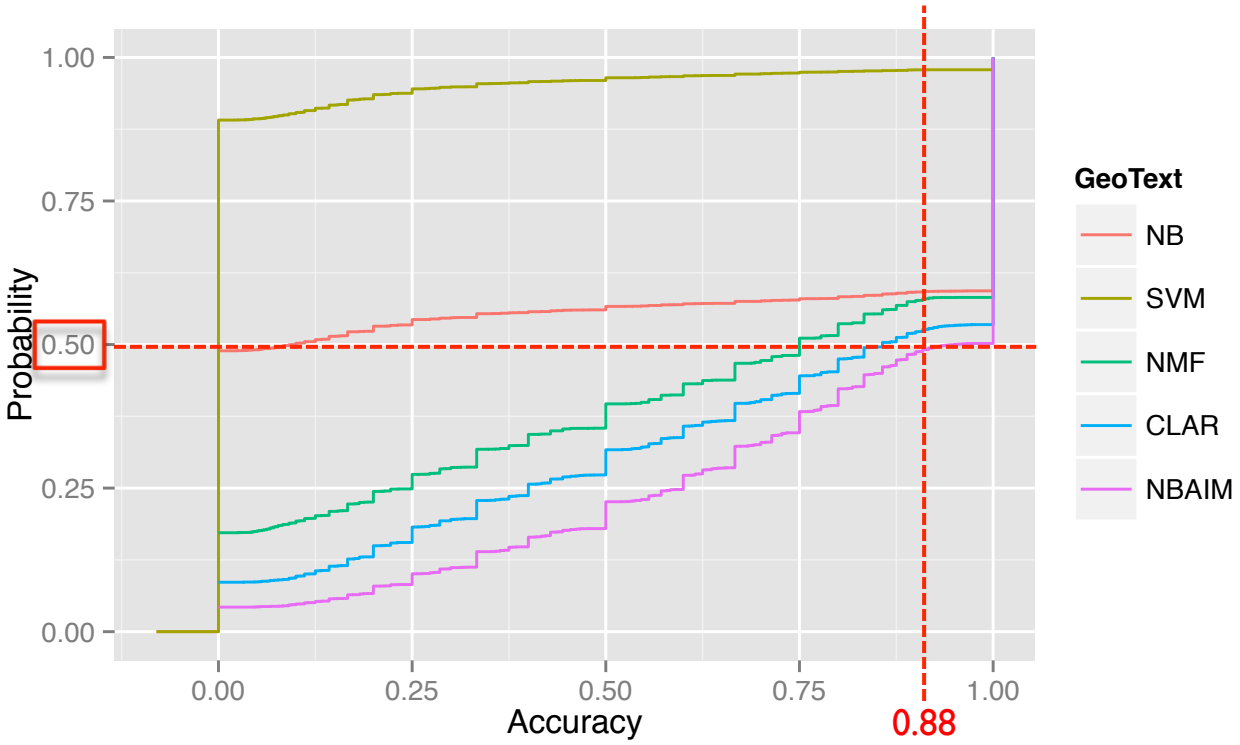
- Training Date Size vs. Accuracy



NBAIM have higher accuracy even when there are only 10 training data

Compare Methods for Activity Inference

- Cumulative Probability Distribution vs. Accuracy

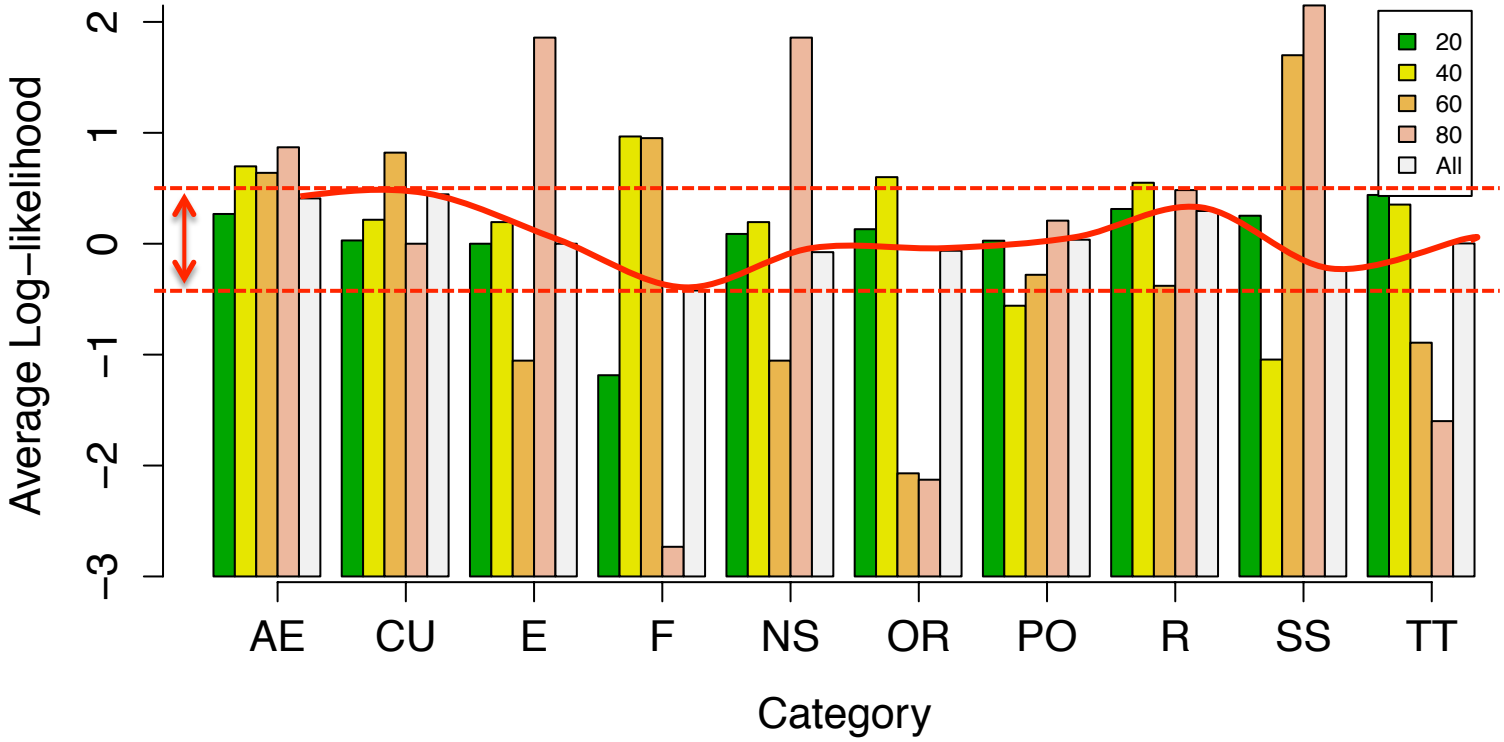


There are more than 50% users have 88% accuracy or more in NBAIM

Settings – Mobility Inference

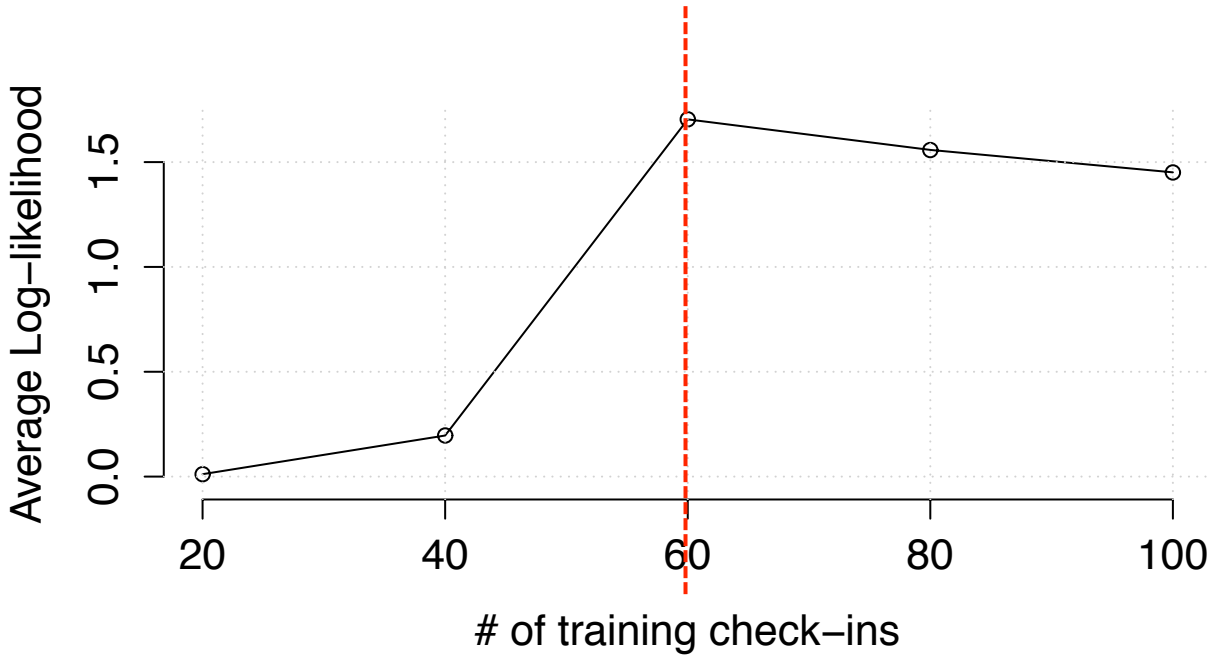
- Setting
  - 80% training, 20% testing
  - Users: the users who have visited at least 5 distinct location
  - k value: 1 ~ 10 centers
- Measurement
  - Log-likelihood

Category vs. Average Log-likelihood



The average log-likelihoods of all training check-ins are in the range

Training Size vs. Average Log-likelihood



NBAIM have have higher avg. log-likelihood when training size is 60 or more

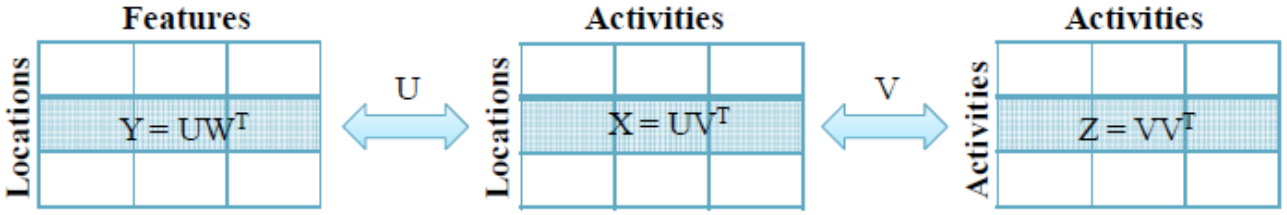
Conclusions

- Formulate the individual activity and mobility inference problem for personalized location-based service in LBSNs
- Utilize Bayesian network to describe the relations among time, location and activity
  - Time-Activity Model: Activity Transition Model
  - Location-Activity Model: GMM
- Show that our proposed methods outperform the state-of-the-art approaches on two real datasets

Thank You

Related Works (cont.)

- Problem: Recommend location or activity
  - Input: GPS, comment
  - Output: location or activity



Non Personal

Lack of handling new locations

[4]Vincent W. Zheng et al. Collaborative location and activity recommendations with GPS history data (WWW '10)

Issue: Factorization

- Original Frequent

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]	[,10]	[,11]	[,12]
[1,]	0	0	0	0	0	0	0	0	0	0	0	0
[2,]	0	0	0	0	1	0	0	0	0	0	0	0
[3,]	1	0	0	0	0	1	0	0	0	0	0	0
[4,]	0	0	0	0	0	0	0	0	0	0	0	0
[5,]	0	0	0	0	0	0	0	0	0	0	0	0
[6,]	0	0	0	0	0	0	0	0	0	0	0	0

- Factorization填空過後結果

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]	[,10]	[,11]	[,12]
[1,]	1	2	2	1	1	1	1	2	2	1	2	1
[2,]	2	2	2	2	1	2	2	2	2	2	2	2
[3,]	1	1	1	1	1	1	1	2	1	1	2	1
[4,]	2	2	2	1	1	2	2	2	2	2	2	3
[5,]	1	1	2	1	1	1	1	2	1	1	1	2
[6,]	1	1	1	1	1	0	1	1	1	1	2	1

- For userA
  - [1,]~[6,]: Dining, Entertainment, Shopping, Sporting, Transportation, Working
  - [,1]~[,24]: Hourly time slots

Issue: Factorization (cont.)

- Original Frequent

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]	[,10]	[,11]	[,12]
[1,]	0	0	0	0	0	0	0	0	0	0	0	0
[2,]	0	0	0	0	1	0	0	0	0	0	0	0
[3,]	1	0	0	0	0	1	0	0	0	0	0	0
[4,]	0	0	0	0	0	0	0	0	0	0	0	0
[5,]	0	0	0	0	0	0	0	0	0	0	0	0
[6,]	0	0	0	0	0	0	0	0	0	0	0	0

- Order-1 transition model

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]	[,10]	[,11]	[,12]
[1,]	0	0.1395349	0.1121800	0.1066645	0.05271323	0.04263442	0.1203737	0.1079042	0.1054192	0.1049075	0.1047254	0.1046436
[2,]	0	0.4186047	0.4908162	0.5143914	0.76243584	0.29227231	0.4863058	0.5170081	0.5263359	0.5304288	0.5323312	0.5332173
[3,]	1	0.1627907	0.1913672	0.2071419	0.10624930	0.62745783	0.1957284	0.2062341	0.2128872	0.2150391	0.2159222	0.2163246
[4,]	0	0.2790698	0.2056367	0.1718022	0.07860163	0.03763544	0.1975921	0.1688536	0.1553577	0.1496246	0.1470212	0.1458145
[5,]	0	0.0000000	0.0000000	0.0000000	0.00000000	0.00000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
[6,]	0	0.0000000	0.0000000	0.0000000	0.00000000	0.00000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000

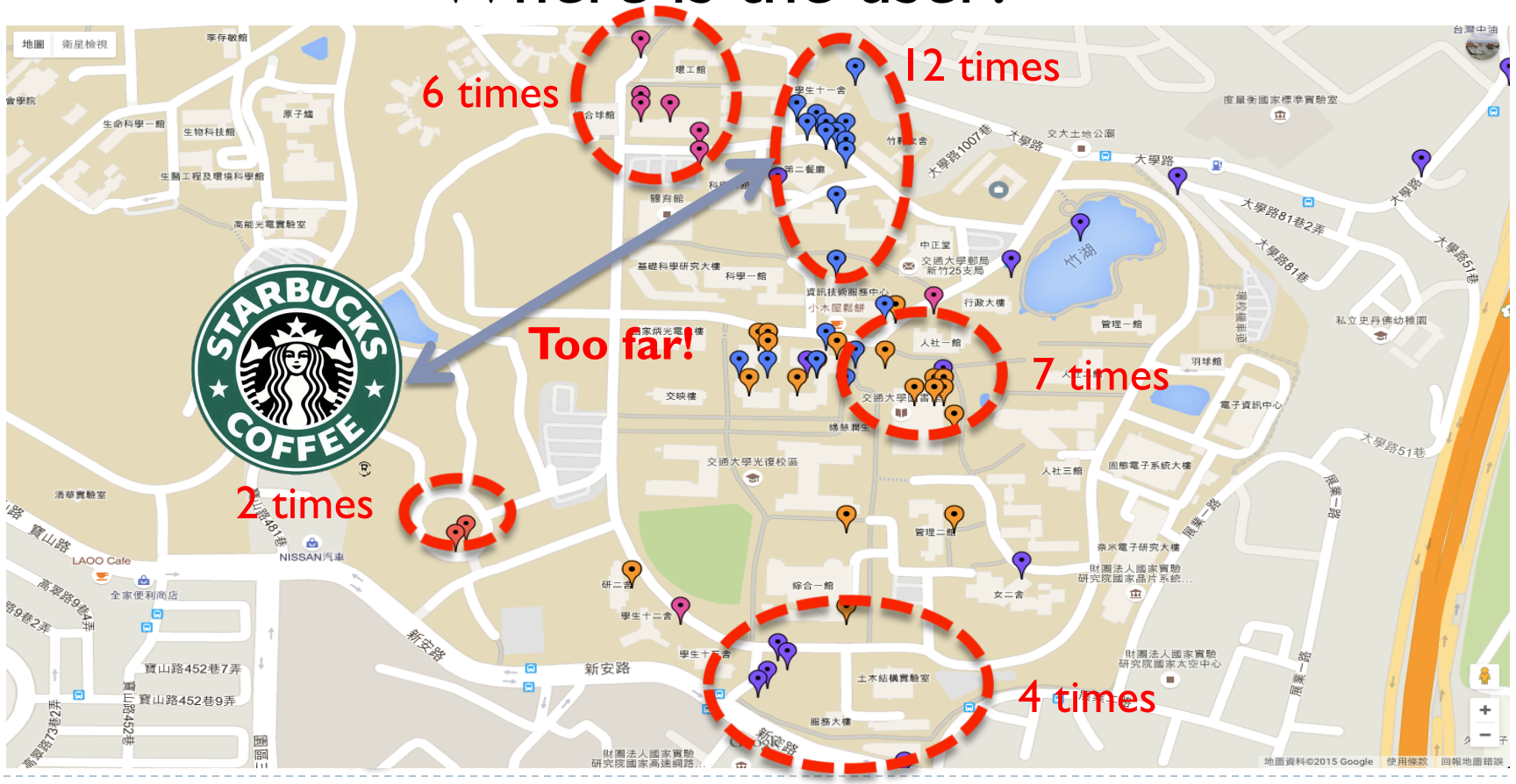
- Factorization Method

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]	[,10]	[,11]	[,12]
[1,]	0.1111111	0.125	0.1111111	0.125	0.125	0.125	0.1111111	0.125	0.125	0.1666667	0.125	0.09090909
[2,]	0.1111111	0.125	0.1111111	0.125	0.125	0.125	0.1111111	0.125	0.125	0.1666667	0.125	0.09090909
[3,]	0.1111111	0.125	0.1111111	0.125	0.125	0.125	0.1111111	0.125	0.125	0.1666667	0.125	0.09090909
[4,]	0.2222222	0.250	0.2222222	0.250	0.250	0.250	0.2222222	0.250	0.250	0.1666667	0.250	0.27272727
[5,]	0.3333333	0.250	0.2222222	0.250	0.250	0.250	0.2222222	0.250	0.250	0.1666667	0.250	0.27272727
[6,]	0.1111111	0.125	0.2222222	0.125	0.125	0.125	0.2222222	0.125	0.125	0.1666667	0.125	0.18181818

Possible Solution - Naïve Bayesian

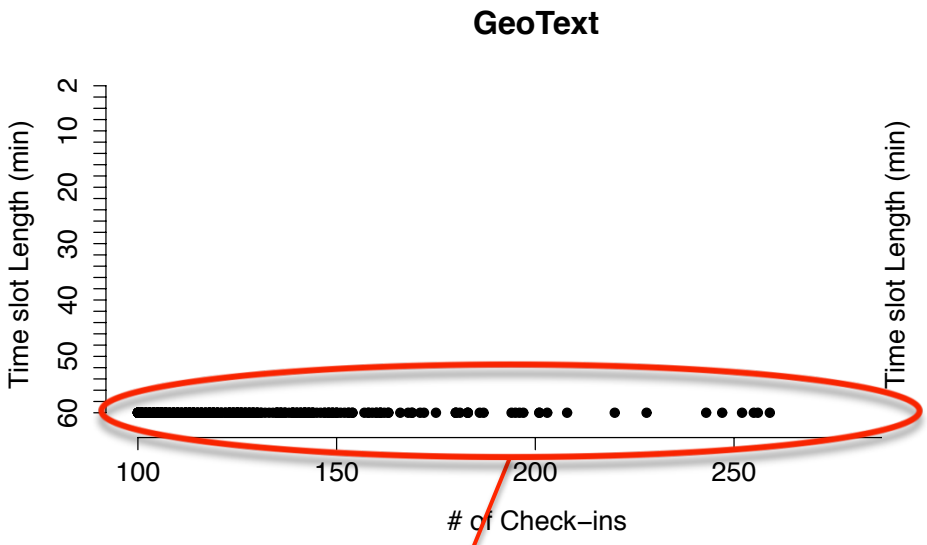
What is the user doing?

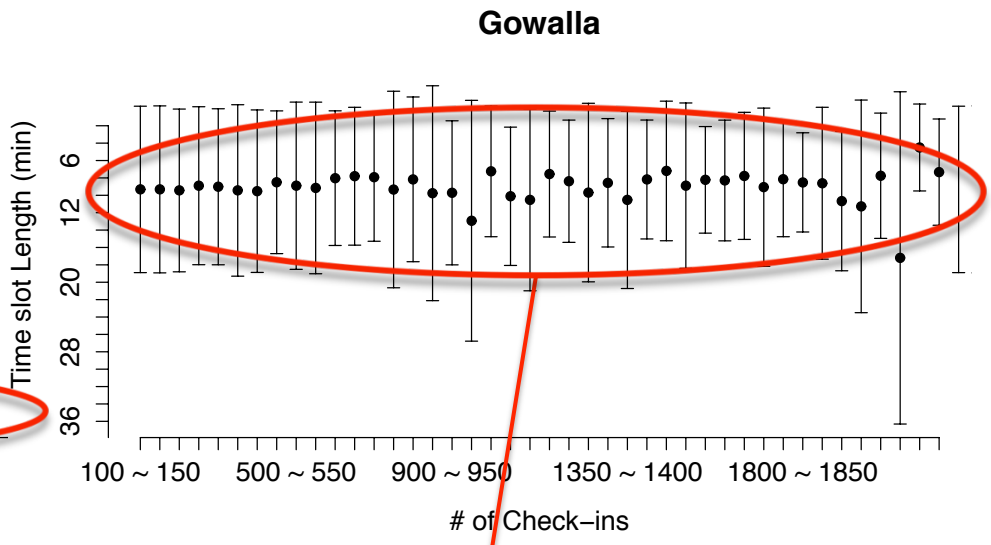
Where is the user?



Cons: Prediction Error

Impact of Number of Time Slot Length

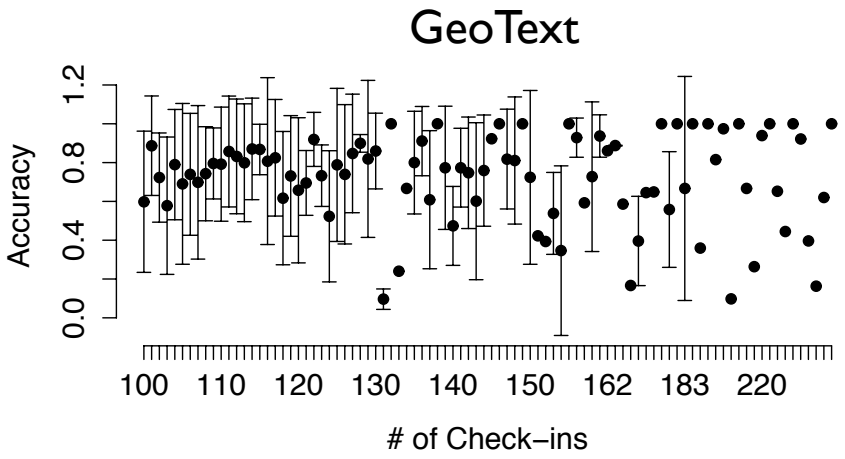


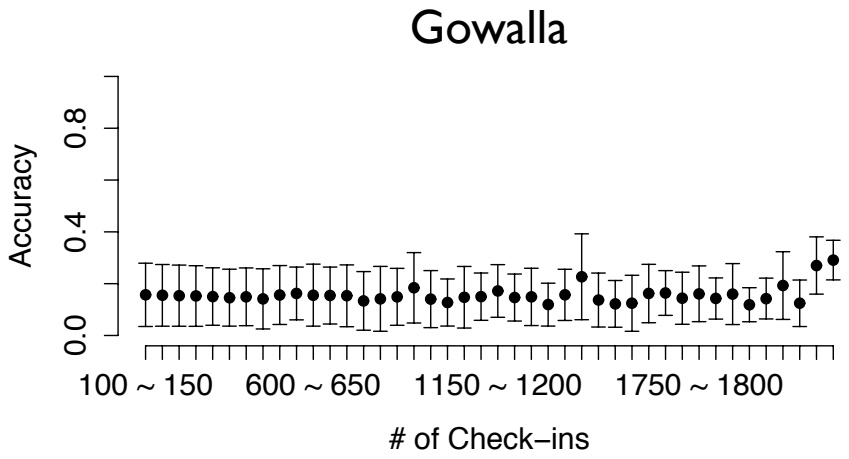


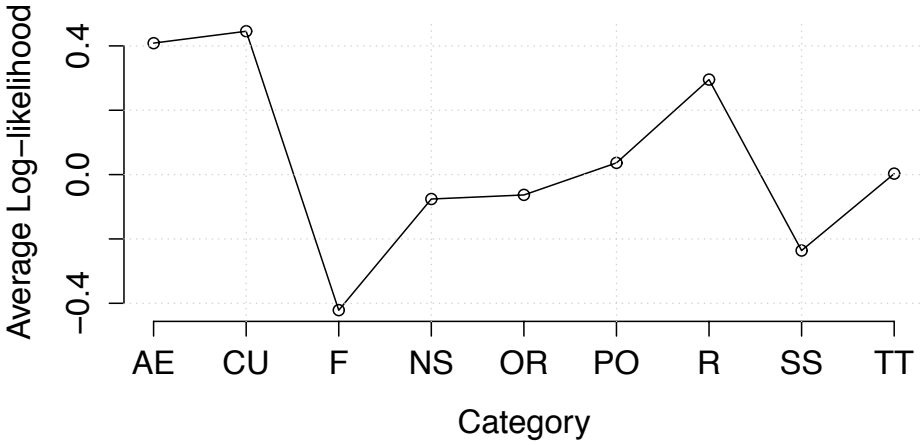
An hour is the best.

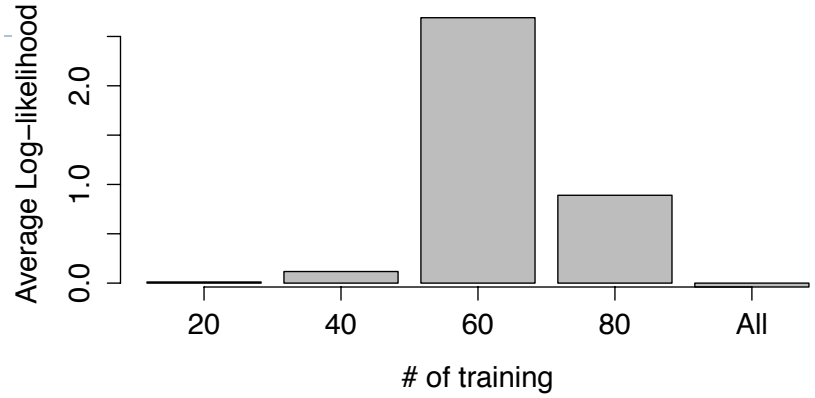
About 10 minutes is better.

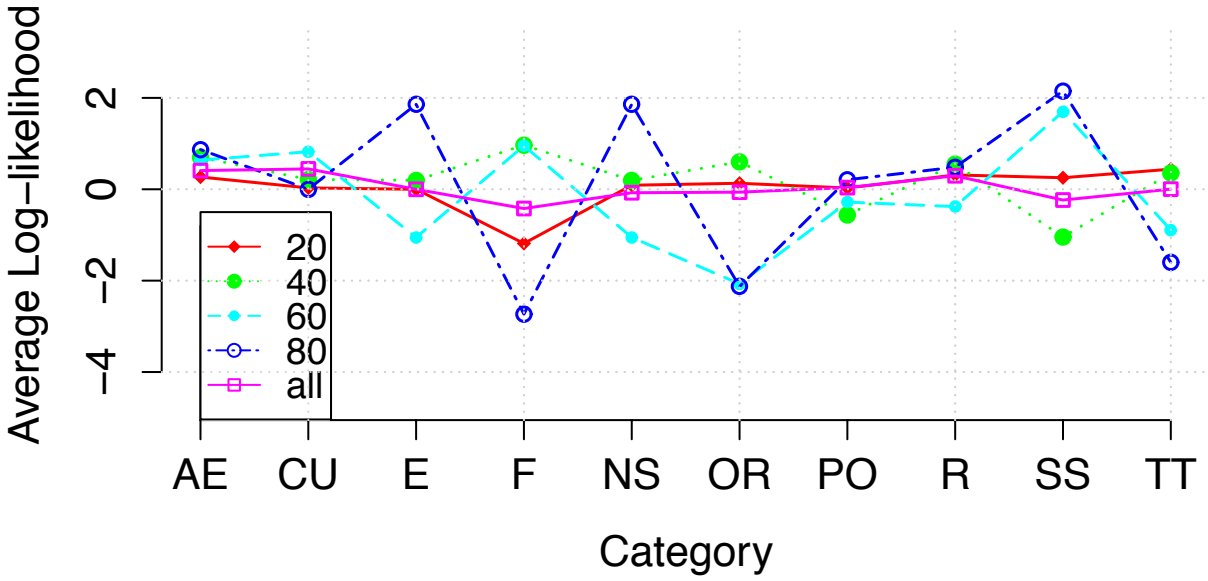
Impact of Number of Check-in Frequency






Average Log−likelihood
−0.4
0.0
0.4
AE
CU
F
NS
OR
PO
R
SS
TT
Category


Average Log−likelihood
0.0
1.0
2.0
20
40
60
80
All
# of training


Average Log−likelihood
−4
−2
0
2
20
40
60
80
all
AE
CU
E
F
NS
OR
PO
R
SS
TT
Category