Institute for
Infocomm Research
A*STAR

Identifying Tourists from Public Transport Commuters

Mingqiang Xue ^
Huayu Wu ^
Wei Chen ^
Wee Siong Ng ^
Gin Howe Goh #

^ Institute for Infocomm Research
# Land Transport Authority of Singapore

Outline

- Introduction
- Backgroud
- Our Approach
  - Station Ranking
  - Label Inference
- Experiments
- Case Study
- Related Work
- Conclusions

Introduction

- Tourism industry, a key economic driver for Singapore:
  - 15 million foreign visitors a year
  - 23 billion Singapore Dollar receipts in 2012

- Understanding tourists travelling behaviors is important:
  - Where do they go?
  - How they travel from one place to another?
  - Where do they stay?

- Useful to stake holders:
  - **Government** (tourism board, city planning, public transport): better planning, improve existing services
  - **Private** (travel agencies, taxis, hotels, restuarants, advertising etc): better or new business

Introduction

- A highly efficient transport system in Singapore
  - Buses, MRTs, LRTs
  - Payment mostly with commuter card (EZ-link)
  - Trajectories (partially) recorded

- Utilized by both locals, business travellers, and tourists in Singapore

- **Who Are the Tourists Among the Commuters?**

Introduction

Public transport Data

bus and train riding records

Tourists traveling Records

Who are tourists?

Tourists Traveling Patterns

How do they travel and where do they go to?

Introduction

Public transport Data

bus and train riding records

Tourists traveling Records

Who are tourists?

Tourists Traveling Patterns

How do they travel and where do they go to?

Main focus

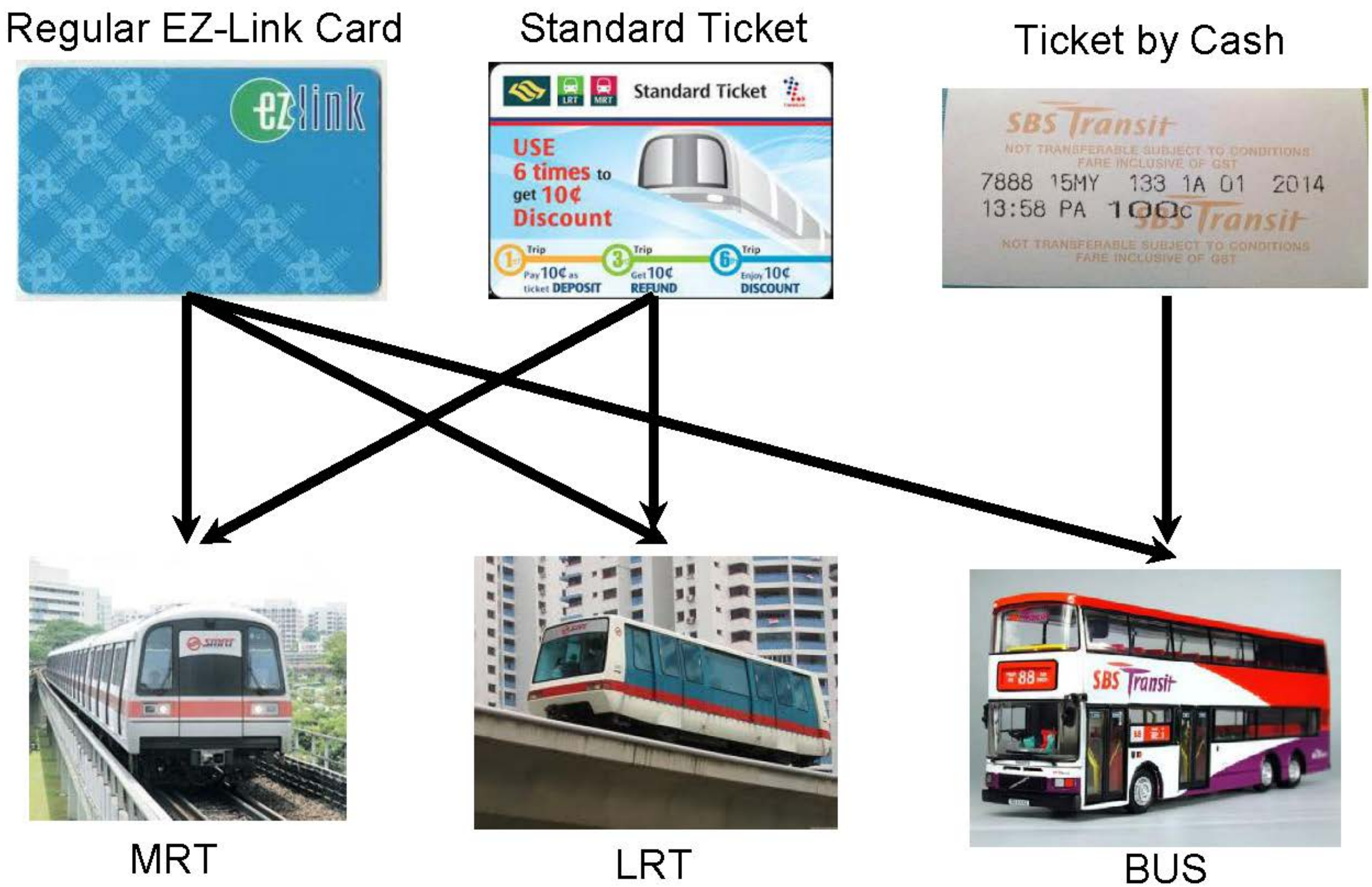
Background – public transport

- The public transport system
  - MRT, similar to the subway in NYC

  - LRT, short distance neighborhood railway transport

  - Bus

Background – ticketing & Payment



Background – travel record

Field	Description
Card_Number_E	Card ID for this ride
Transport_Mode	BUS, LRT, or MRT
Entry_Date	Date when ride started
Entry_Time	Time when ride started
Exit_Date	Date when ride ended
Exit_Time	Time when ride ended
Payment_Mode	Method of payment
Origin_Location_ID	Starting location of the ride
Destination_Location_ID	Ending location of the ride

The travel record Schema

Background

- Many tourists use standard tickets to travel around

- Tourists travelling patterns from standard tickets records
  - Problem: discontinued trajectories, no bus records, size could be small

- Our goal: identify tourists from **regular EZ-link card** users

Our Approach

- A Two staged processs:
  - Stage 1: Initialization
    - Score each MRT/LRT station based on the attractiveness to tourists
  - Stage 2: Iterative Refinement
    - Update the scores for both MRT/LRT stations and tourists in a graph
    - Classify one as a tourist/non-tourist after the final iteration

Our Approach – Stage 1

- t - a tourist commuter
- m_i - an event that a commuter has visited station i

- We solve for each station:

$$\text{Score } s_{m_i} \sim \Pr(t|m_i)$$

Our Approach – Stage 1

- t - a tourist commuter
- m_i - an event that a commuter has visited station i

- We solve for each station:

$$\text{Score } s_{m_i} \sim \Pr(t|m_i) = \Pr(t) \cdot \frac{\Pr(m_i|t)}{\Pr(m_i)}$$

Our Approach – Stage 1

- t - a tourist commuter
- m_i - an event that a commuter has visited station i

$$\text{Score } s_{m_i} \sim \Pr(t|m_i) = \Pr(t) \cdot \boxed{\frac{\Pr(m_i|t)}{\Pr(m_i)}}$$

Our Approach – Stage 1

- t - a tourist commuter
- m_i - an event that a commuter has visited station i
- n_i^s number of trips with standard tickets at station i
- n_i^r number of trips with regular EZ-link card at station i
- n_i^t number of trips from tourists with standard tickets at station i

Score $s_{m_i} \sim \Pr(t|m_i) = \Pr(t) \cdot \boxed{\frac{\Pr(m_i|t)}{\Pr(m_i)}}$

The estimation of $\Pr(m_i|t)$:

- Idea: standard tickets records, but isolate the effects of locals
- $\hat{\theta}$ is the probability that a local uses a standard ticket

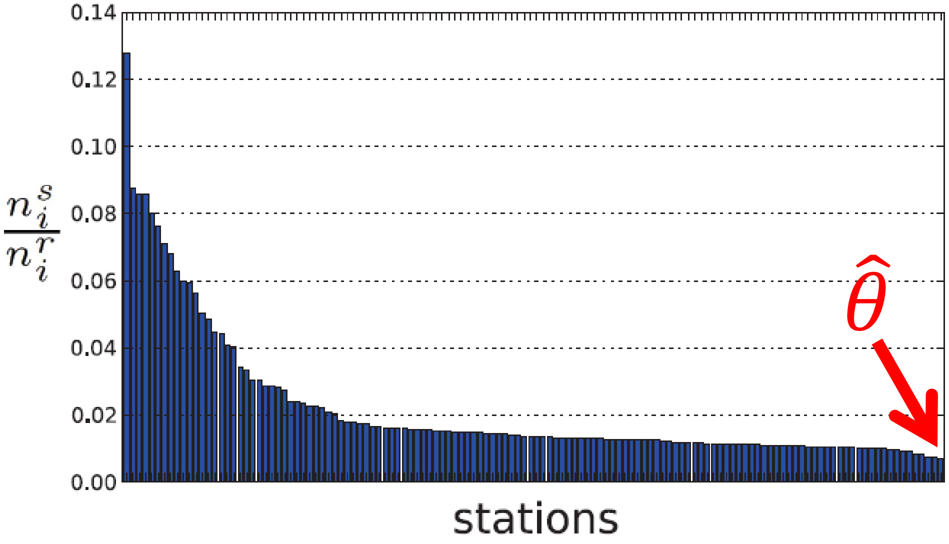
$$\hat{\Pr}(m_i|t) = \frac{n_i^t}{\sum_i n_i^t} \text{ where } n_i^t = n_i^s - n_i^r \cdot \hat{\theta}$$

Institute for Infocomm Research (I[2]R)

Confidential

Our Approach - Stage 1

- The estimation of $\hat{\theta}$:



Name	n_i^s	n_i^r	$\frac{n_i^s}{n_i^r}$
Marymount	6218	629435	0.009879
Yio Chu Kang	20361	2067636	0.009847
Cove	1817	189873	0.009570
Buangkok	7454	787463	0.009466
Layar	345	37211	0.00927
Oasis	489	53696	0.009107
Labrador Park	2473	292858	0.008444
Tongkang	1295	158299	0.008181
Compassvale	2705	358175	0.007552
Dover	8963	1247247	0.007186

$\hat{\theta}$

Dover surroundings: - An isolated educational institution
- No closeby residences

Our Approach – Stage 1

- t - a tourist commuter
- m_i - an event that a commuter has visited station i
- n_i^s number of trips with standard tickets at station i
- n_i^r number of trips with regular EZ-link card at station i
- n_i^t number of trips from tourists with standard tickets at station i

$$\text{Score } s_{m_i} \sim \Pr(t|m_i) = \Pr(t)\frac{\Pr(m_i|t)}{\boxed{\Pr(m_i)}}$$

The estimation of $\Pr(m_i)$:

$$\hat{\Pr}(m_i) = \frac{n_i^s + n_i^r}{\sum_i n_i^s + n_i^r}$$

Our Approach – Stage 1

- t - a tourist commuter
- m_i - an event that a commuter has visited station i
- n_i^s number of trips with standard tickets at station i
- n_i^r number of trips with regular EZ-link card at station i
- n_i^t number of trips from tourists with standard tickets at station i

$$\text{Score } s_{m_i} = \Pr(t) \cdot \frac{\Pr(m_i|t)}{\Pr(m_i)}$$

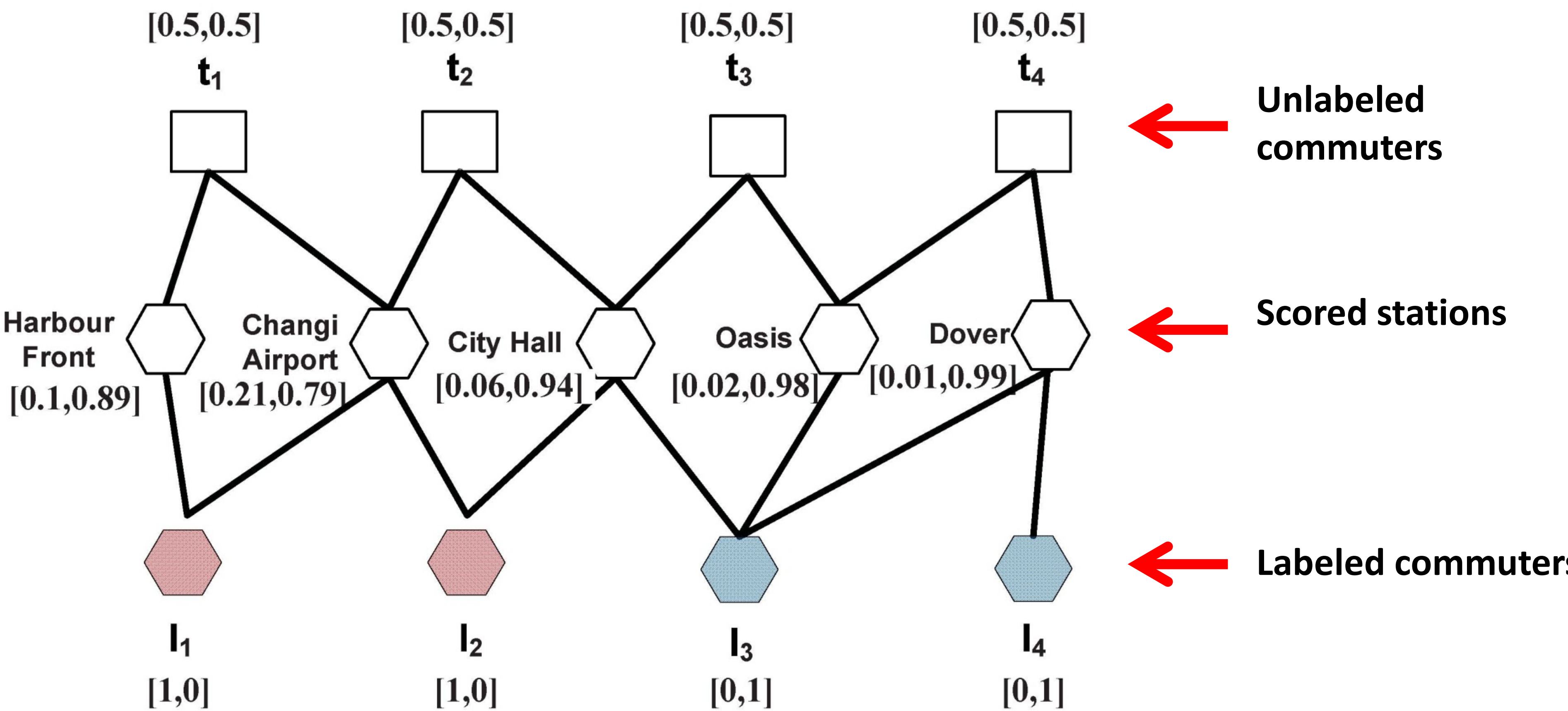
where $$\hat{\Pr}(t) = \frac{\sum_i 2n_i^t}{\sum_i n_i^s + n_i^r}$$

Our Approach – Stage 1

Name	s_{m_i}
Changi Airport	0.213668
Marina Bay	0.145012
Clarke Quay	0.144702
Bayfront	0.128008
Little India	0.118879
Chinatown	0.113837
HarbourFront	0.106443
Bras Basah	0.104787
Esplanade	0.099637
Orchard	0.098623
Lavender	0.093104
Farrer Park	0.081844
Promenade	0.079080
Bugis	0.070973
City Hall	0.064815

Top Ranked stations based on attractiveness

Our Approach – Stage 2

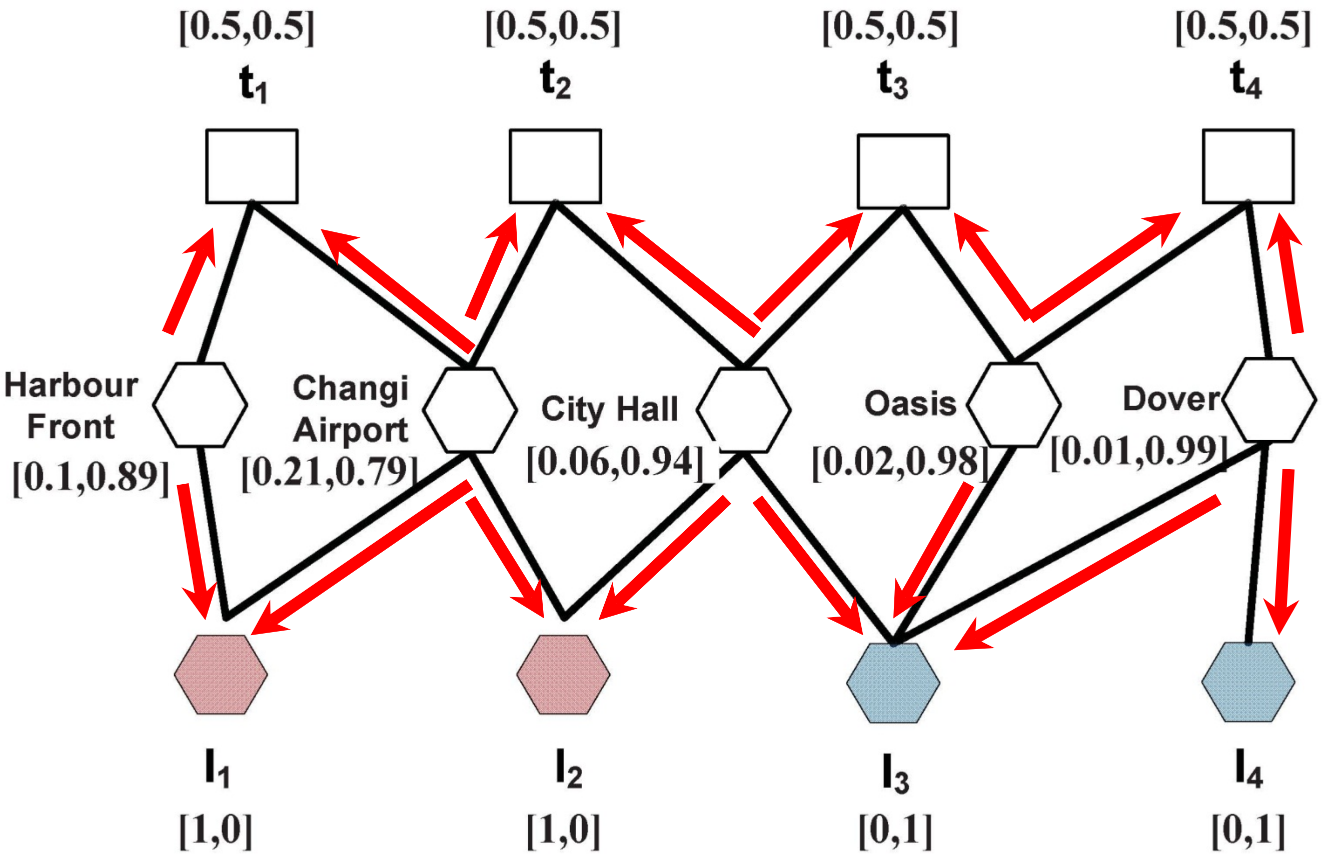


A toy Station-Commuter Relationship graph

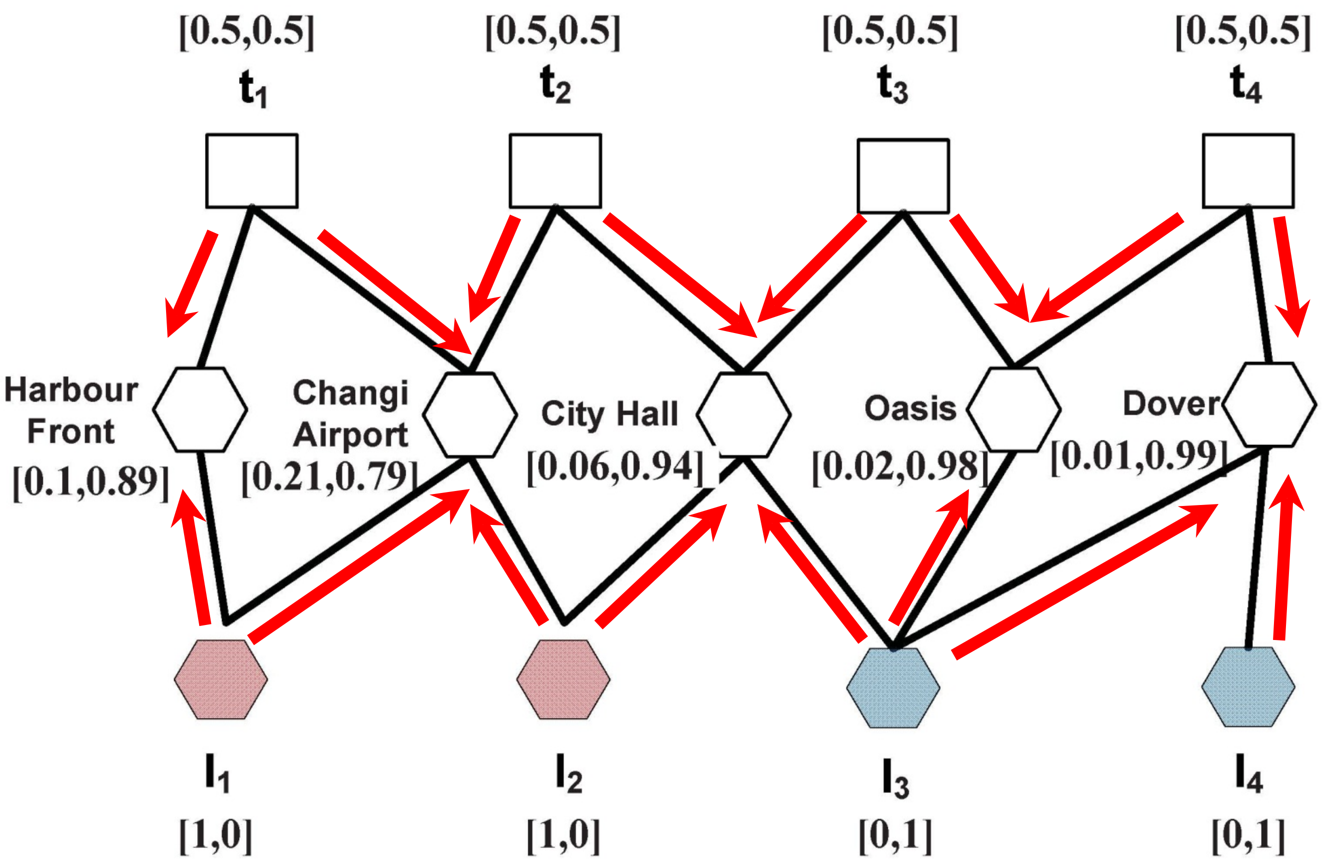
Our Approach – Stage 2

- While # of iterations < predefined threshold (e.g 150) :
  - Update the class distribution of each commuter based on its current class distribution and the class distributions of stations that they visited
  - Update the class distribution of each station based on its current disribution and the class distributions of commuters who visit them

Our Approach – Stage 2



Our Approach – Stage 2



Our Approach – Stage 2

- Updating functions:

$$\left.\begin{aligned}
\phi_{l_i}^{k} &\leftarrow \alpha \cdot \phi_{l_i}^{k-1} + (1-\alpha) \cdot \frac{\sum_{m \in N(l_i)} w_{l_i m} \cdot \phi_m^k}{\sum_{m \in N(l_i)} w_{l_i m}} \\
\phi_{t_i}^{k} &\leftarrow \beta \cdot \phi_{t_i}^{k-1} + (1-\beta) \cdot \frac{\sum_{m \in N(t_i)} w_{t_i m} \cdot \phi_m^k}{\sum_{m \in N(t_i)} w_{t_i m}}
\end{aligned}\right\} \text{Update for commuters}$$

$$\left.\phi_{m_i}^{k} \leftarrow \gamma \cdot \phi_{m_i}^{k-1} + (1-\gamma) \cdot \frac{\sum_{u \in N(m_i)} w_{u m_i} \cdot \phi_{m_i}^k}{\sum_{u \in N(m_i)} w_{u m_i}}\right\} \text{Update for stations}$$

Our Approach – Stage 2

- Final class assignment:

$$\hat{C} = \underset{c}{argmax} \frac{P(t_i|c)}{P(t_i)} = \underset{c}{argmax} \frac{P(c|t_i)}{P(c)}$$

For $c \in \{\text{Tourist}, \text{ Non}-\text{Tourist}\}$

Experiments

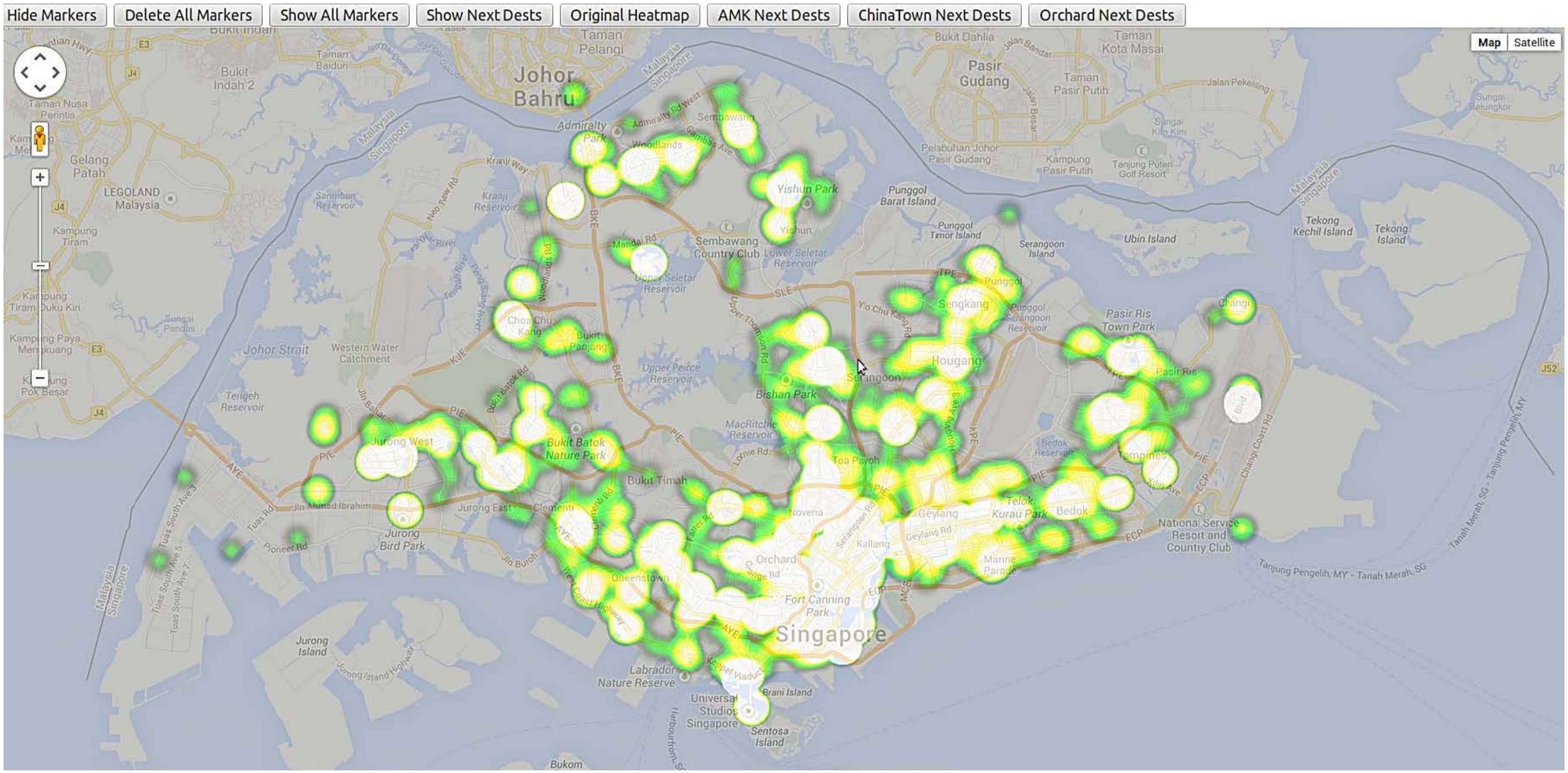
- One-month EZ-link records from LTA
- Preprocessing:
  - Exclude commuters with less than 6 records
- Data description:
  - 1.7 million commuters
  - 49.5 million records
  - Training set: 1000 tourists and 250,000 locals
- Competitors:
  - FTF (Fast Transversal Filter): a state-of-the-art iterative inference algorithm
  - SVM
- Evaluation metric:
  - F1 score: $F1 = \frac{2 \times Precision \times Recall}{Recall + Precision}$

Experiments

	SVM		FTF		I^2	
$p\%$	Macro F1	Micro F1	Macro F1	Micro F1	Macro F1	Micro F1
5%	0.57984	0.8415	0.6109	0.8419	**0.6267**	**0.8504**
10%	0.5917	0.8420	0.6263	0.8464	**0.6572**	**0.8538**
15%	0.6144	0.8411	0.6441	0.8433	**0.6677**	**0.8560**
20%	0.6199	0.8480	0.6758	0.8504	**0.6962**	**0.8575**
25%	0.6286	0.8402	0.6956	0.8459	**0.7154**	**0.8549**

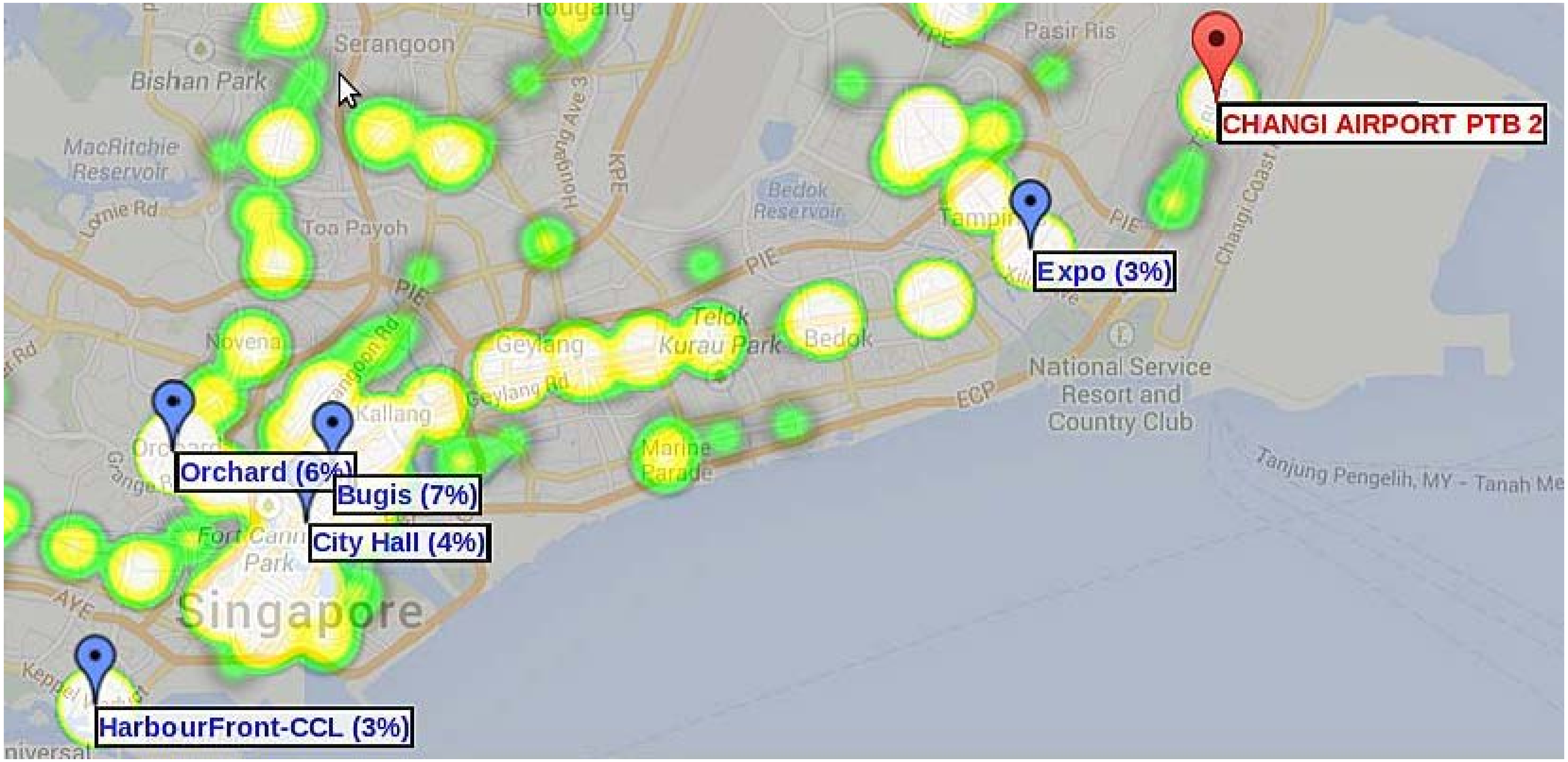
Comparison Results

Case Study



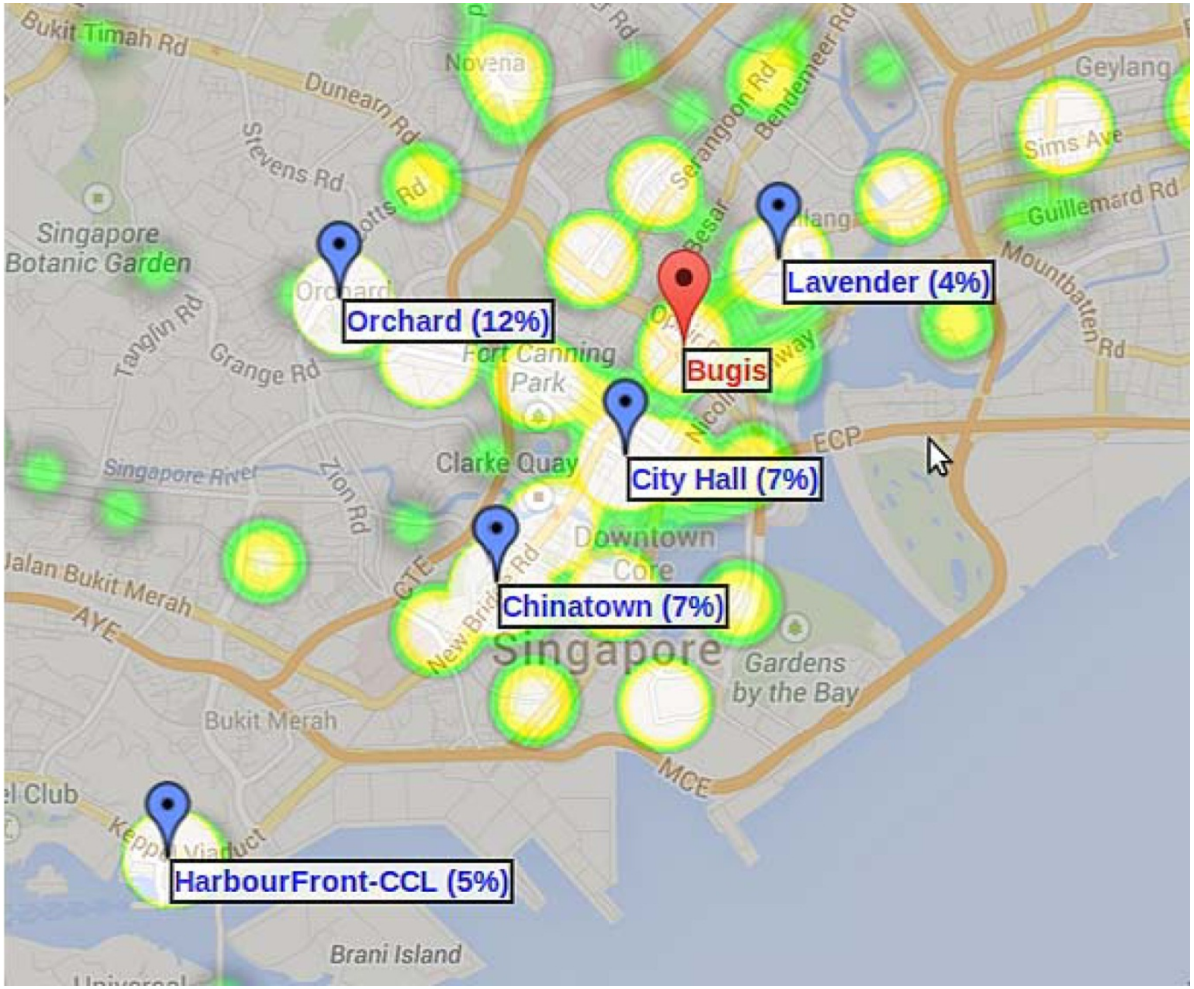
Places visited by tourists by popularity

Case Study



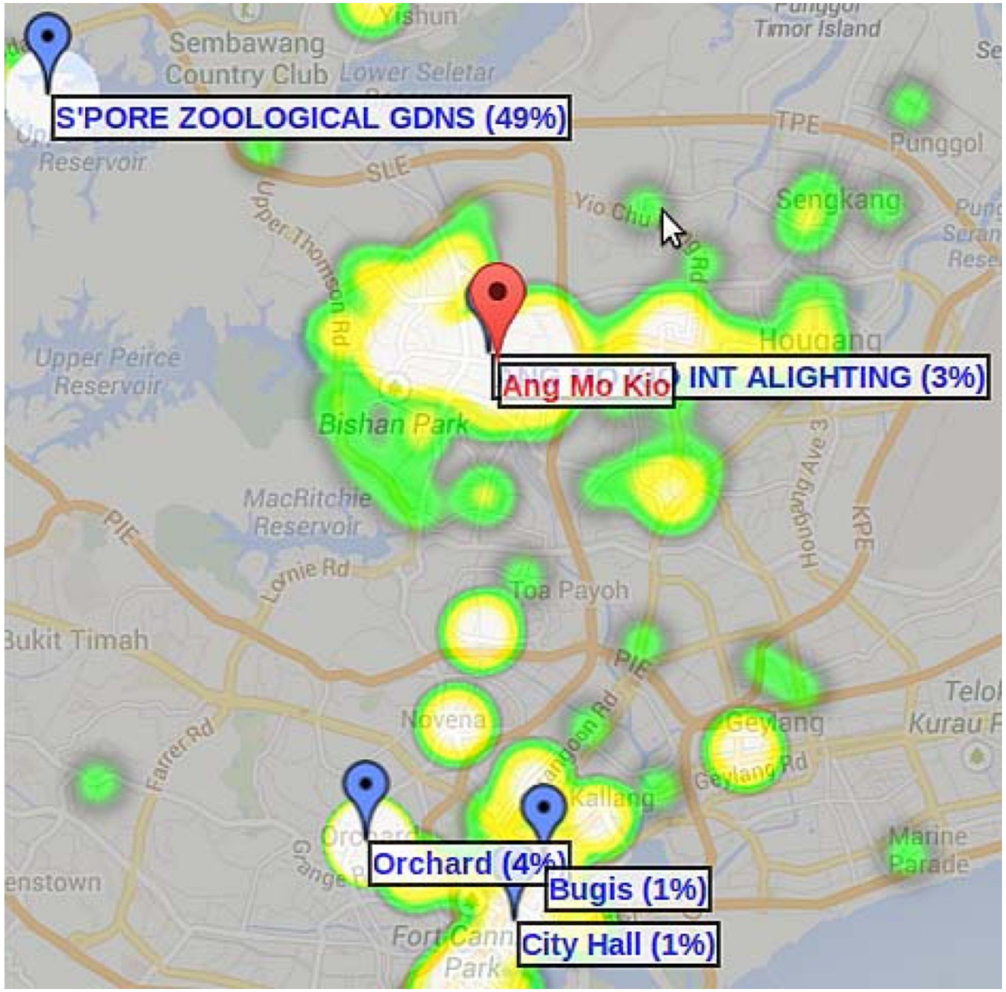
Where do tourists go from the airport?

Case Study



Where do tourists go from bugis?

Cast Study



Why do tourists visit Ang Mo Kio?

Related Work

- Mining public transport data
  - Improve public transport in a city
  - Behaviors of populations (what's the popular shopping places)
  - Behaviors of individuals (what's one's home, work place)

- Mining tourists data
  - Travelling patterns of tourists (e.g based on Geo-tagged images)

Conclusions

- Extract tourists records from public transport data
  - Meaningful to stakeholders, both private and government
- Proposed an algorithm based on:
  - Station scoring and iterative score refinement
- Verified findings with experiments
- Hope to attract interest to solve similar problems in other cities, e.g. Hong Kong, NYC, London etc.

Thank you