Detecting Communities of Commuters: Graph Based Techniques vs Generative Models

Ashish Dandekar, Stéphane Bressan, Talel Abdessalem, Huayu Wu, Wee Siong Ng

September 7, 2016

Motivation

Card Number	In-Timestamp	Out-timestamp	In-ID	Out-ID
c530524	yyyy-dd-mm;07:22:49.0	yyyy-dd-mm;07:28:50.0	2383	1467
c530545	yyyy-dd-mm;12:09:40.0	yyyy-dd-mm;12:29:40.0	1464	8
c630568	yyyy-dd-mm;13:10:30.0	yyyy-dd-mm;13:40:50.0	2413	99
c534554	yyyy-dd-mm;20:08:12.0	yyyy-dd-mm;20:28:07.0	2384	2
c837483	yyyy-dd-mm;16:02:10.0	yyyy-dd-mm;16:34:33.0	1467	185
c254234	yyyy-dd-mm;09:09:43.0	yyyy-dd-mm;09:19:23.0	1899	99

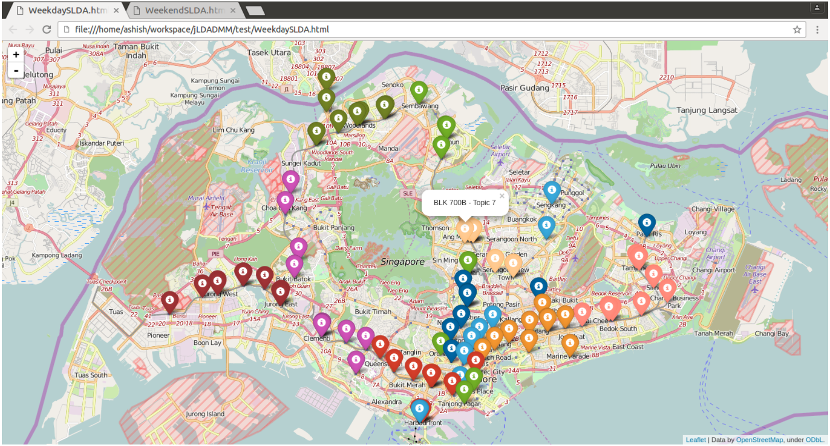
...

...

...

Millions of such records!

Motivation



NUS
National University of Singapore

Introduction

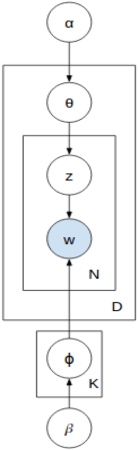
- Community detection by using overlaps in mobility
- Exisiting Techniques
  - Traditional Data Mining Techniques
  - Graph based techniques
- Generative Model
  - Statistical modelling
  - Bayesian approach
  - Generative process

Problem - Are generative models more effective than graph based techniques?

Related Work

- Urban Computing [19]
  - Reducing waiting time of commuters [5]
  - Travelling behaviour analysis [12, 11, 13]
  - Identifying tourists from daily commuters [16]
- Graph based techniques [6]
  - Divisive algorithm [7]
  - Modularity optimization [2, 4]
- Generative Models
  - Finding communities in LBSN data using LDA [14, 10, 3]
  - Extending LDA to handle geolocations [15, 9]
  - Extending LDA to handle spatio-temporal events [17, 18]

Latent Dirichlet Allocation - LDA[1]



Notation

- N : Vocabulary size
- D : Total number of Documents
- K : Total number of Topics

Intuition

- Bag of Words assumption
- A document is a distribution over topics
  - $\bar{\theta}_m \rightarrow K$-dim vector; $m \in [1...D]$
- A topic is a distribution over words
  - $\bar{\phi}_k \rightarrow N$-dim vector; $k \in [1...K]$

Adopting LDA to Spatio-Temporal Data

What does LDA require?

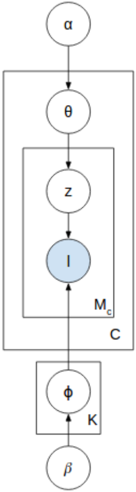
Bags of words!

Analogy

- LBSN: Users and their checkins
- Taxi: Taxis and their GPS positions
- Public Transport Data: Commuters and bus/train stops

SLDA - Spatial LDA

- Document → Commuter
- Words → Spatial mobility of a commuter
- Topics → Spatial mobility patterns

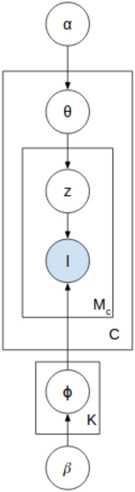


SLDA

SLDA - Spatial LDA

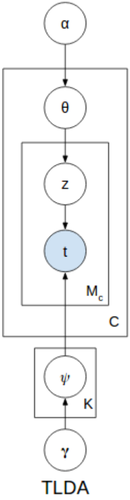
- Document → Commuter
- Words → Spatial mobility of a commuter
- Topics → Spatial mobility patterns

What about time?



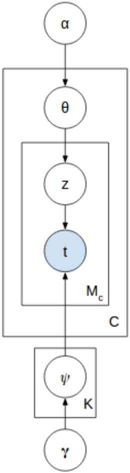
SLDA

TLDA - Temporal LDA



- Document → Commuter
- Words → Temporal mobility of a commuter
- Topics → Temporal mobility patterns

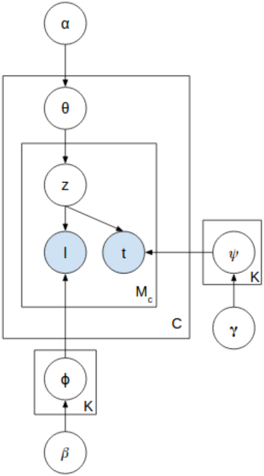
TLDA - Temporal LDA



- Document → Commuter
- Words → Temporal mobility of a commuter
- Topics → Temporal mobility patterns

Can we consider both space and time simultaneously?

STLDA - Spatio-Temporal LDA



- Document → Commuter
- Words → Spatio-temporal events
- Topics → Spatial and temporal mobility patterns

Inference

Inference[8]

Algorithm 1 Gibbs Sampling Interation

for all commuters $c \in \mathcal{C}$ **do**
    for all visits $v \in \mathcal{M}$ **do**
        $K \leftarrow$ topic assigned to v
        Decrement counts $\phi_{k,v}, \theta_k$
        $Z \leftarrow$ sample new topic
        Increment counts $\phi_{z,v}, \theta_z$
    end for
end for

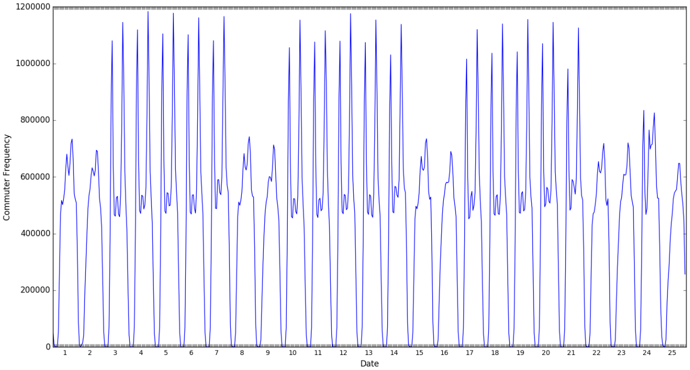
Experiments

EZ-link Data

Field	Description
Card_Number_E	ID of the EZ-link card
Transport_Mode	Bus, MRT or LRT
Entry_Date	Date of the tap-in
Entry_Time	Time of the tap-in
Exit_Date	Date of the tap-out
Exit_Time	Time of the tap-out
Payment_Mode	Mode of the payment
Commuter_Category	Category of the card
Origin_Location_ID	Location ID of the tap-in
Destination_Location_ID	Location ID of the tap-out

Table: Dataset Schema

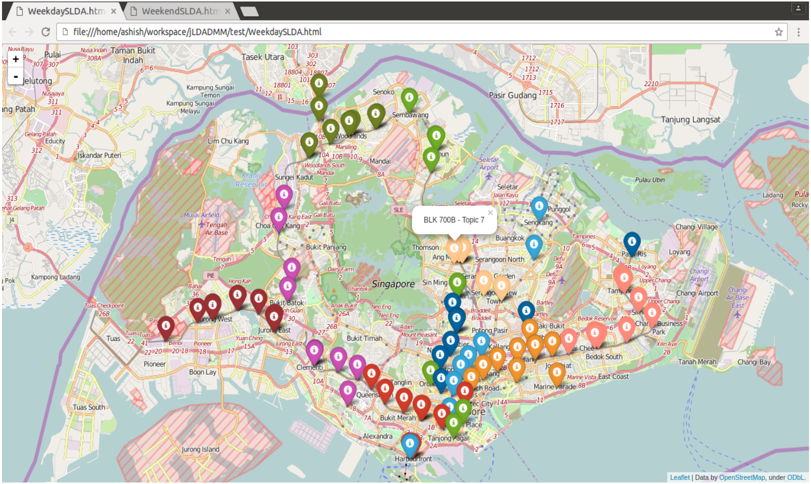
EZ-link Data


Commuter Frequency
1200000
1000000
800000
600000
400000
200000
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
Date

EZ-link Data

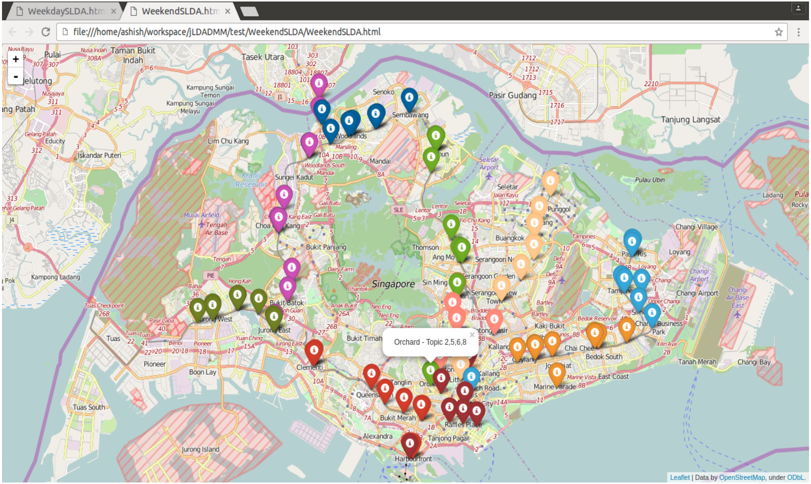
- Filtered two weekdays and two weekends
- Sampled 40,000 regular commuters

EZ-link Data: Weekday Topics (SLDA)



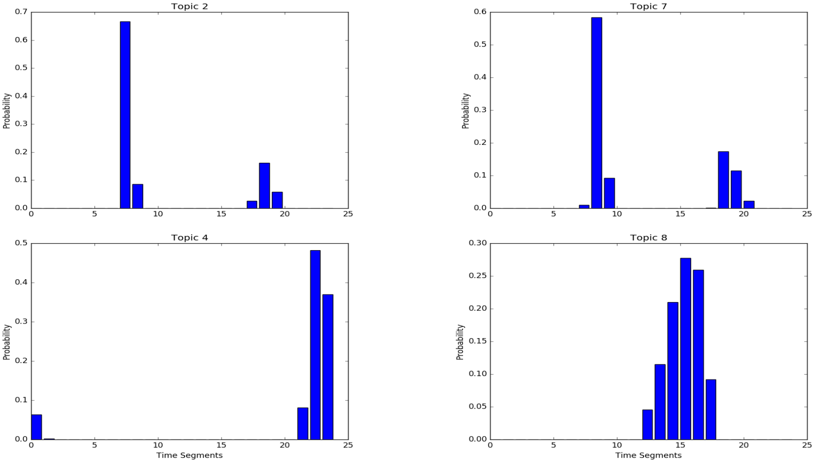
NUS
National University of Singapore

EZ-link Data: Weekend Topics (SLDA)



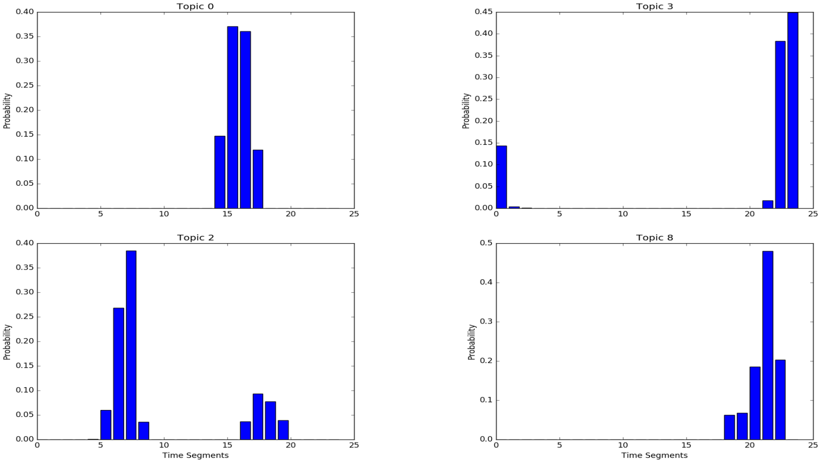
NUS
National University of Singapore

EZ-link Data: Weekday Clusters (TLDA)

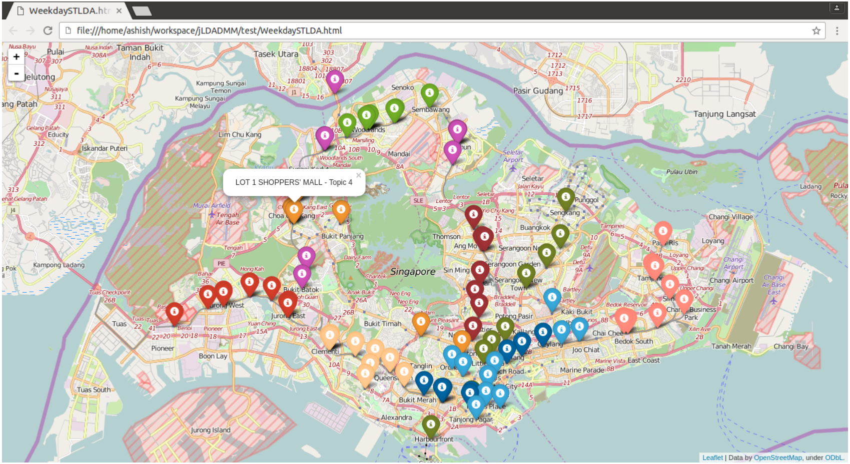


NUS National University of Singapore

EZ-link Data: Weekend Clusters (TLDA)

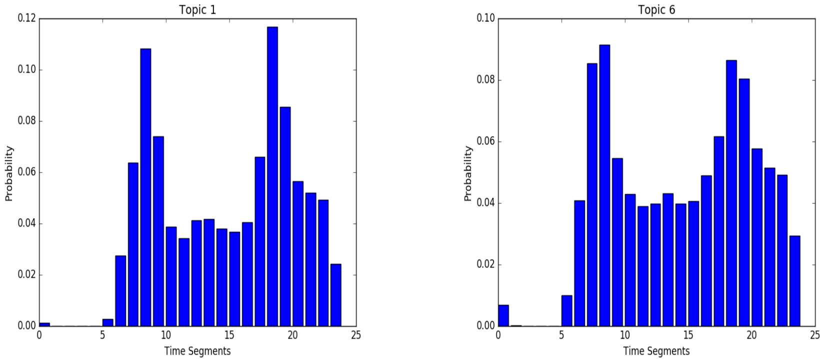


EZ-link Data: Weekday Topics (STLDA)



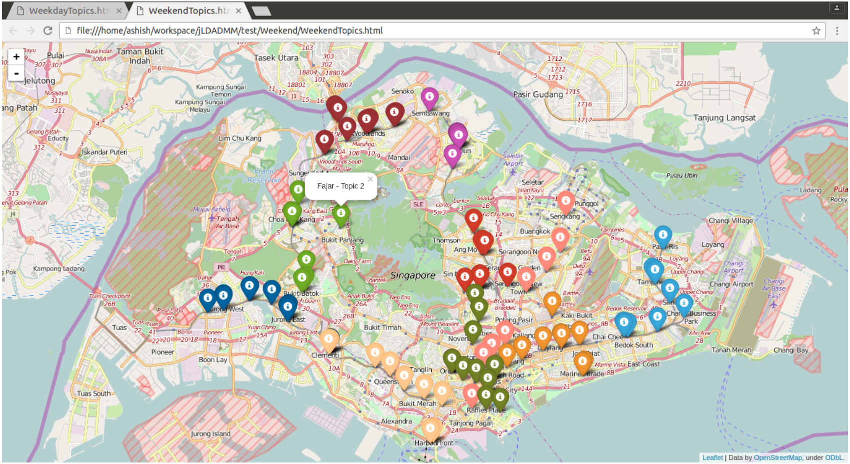
Spatial Part

EZ-link Data: Weekday Clusters (STLDA)



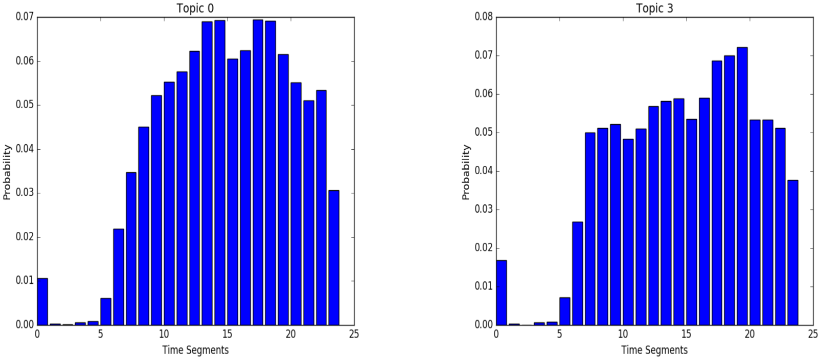
Temporal Part

EZ-link Data: Weekend Topics (STLDA)



Spatial Part

EZ-link Data: Weekend Clusters (STLDA)



Temporal Part

Comparison

Can we compare results with graph based technique?

- No groundtruth
- Multiple sparse and small communities

Comparison

Can we compare results with graph based technique?

- No groundtruth
- Multiple sparse and small communities

Generate synthetic yet realistic data!

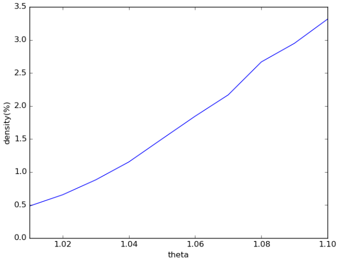
Synthetic Data: Generation

Documents Generation

- Choose distributions
  - visits per commuter $\rightarrow$ Gamma distribution
  - each community $\rightarrow$ Zipf distribution over locations
- Use generative process for the model

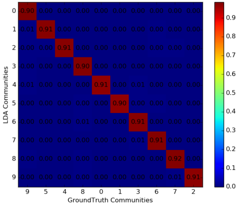
Graph Generation

- Add an edge between two commuters if mobilities have non-empty intersection
- Weigh the edge by the cardinality of overlap

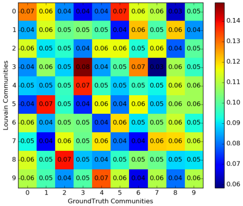


Result Analysis

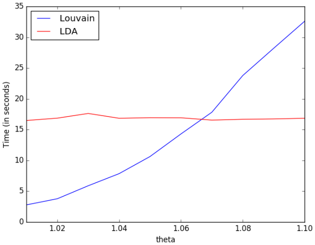
LDA vs Groundtruth



Lovain vs Groundtruth



Efficiency



Why are Graph algorithms less effective?

An Example

- Pairs of commuters A-B and C-D co-occur 5 times

Why are Graph algorithms less effective?

An Example

- Pairs of commuters A-B and C-D co-occur 5 times
- A-B co-occur 5 times at one place
- C-D co-occur 5 times at different places

NUS National University of Singapore

Why are Graph algorithms less effective?

An Example

- Pairs of commuters A-B and C-D co-occur 5 times
- A-B co-occur 5 times at one place
- C-D co-occur 5 times at different places

Loss of information in graph generation!

Conclusion

- Proposed sptio-temporal model for communitites of commuters
- Conducted experiments on real-world data
- Extended experiments to synthetic data so as to have fair quantitative comparison
- Reasoned why generative model is more effective than graph based techniques

NUS National University of Singapore

Thank You!

References I

D. M. Blei, A. Y. Ng, and M. I. Jordan.
Latent dirichlet allocation.
the Journal of machine Learning research, pages 993–1022, 2003.

V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre.
Fast unfolding of communities in large networks.
Journal of statistical mechanics: theory and experiment, page P10008, 2008.

Y.-S. Cho, G. Ver Steeg, and A. Galstyan.
Socially relevant venue clustering from check-in data.
In *11th Workshop on Mining and Learning with Graphs, MLG–2013*, 2013.

A. Clauset, M. E. Newman, and C. Moore.
Finding community structure in very large networks.
Physical review E, page 066111, 2004.

B. Ferris, K. Watkins, and A. Borning.
Onebusaway: results from providing real-time arrival information for public transit.
In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1807–1816. ACM, 2010.

S. Fortunato.
Community detection in graphs.
Physics reports, pages 75–174, 2010.

M. Girvan and M. E. Newman.
Community structure in social and biological networks.
Proceedings of the national academy of sciences, pages 7821–7826, 2002.

References II

T. L. Griffiths and M. Steyvers.
Finding scientific topics.
Proceedings of the National academy of Sciences, (suppl 1):5228–5235, 2004.

B. Hu and M. Ester.
Spatial topic modeling in online social media for location recommendation.
In *Proceedings of the 7th ACM conference on Recommender systems*, pages 25–32. ACM, 2013.

K. Joseph, C. H. Tan, and K. M. Carley.
Beyond local, categories and friends: clustering foursquare users with latent topics.
In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, pages 919–926. ACM, 2012.

N. Lathia and L. Capra.
How smart is your smartcard?: measuring travel behaviours, perceptions, and incentives.
In *Proceedings of the 13th international conference on Ubiquitous computing*, pages 291–300. ACM, 2011.

N. Lathia and L. Capra.
Mining mobility data to minimise travellers' spending on public transport.
In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1181–1189. ACM, 2011.

N. Lathia, D. Quercia, and J. Crowcroft.
The hidden image of the city: sensing community well-being from urban mobility.
In *Pervasive computing*, pages 91–98. Springer, 2012.

X. Long, L. Jin, and J. Joshi.
Exploring trajectory-driven local geographic topics in foursquare.
In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, pages 927–934. ACM, 2012.

References III

S. Sizov.
Geofolk: latent spatial semantics in web 2.0 social media.
In *Proceedings of the third ACM international conference on Web search and data mining*, pages 281–290. ACM, 2010.

M. Xue, H. Wu, W. Chen, W. S. Ng, and G. H. Goh.
Identifying tourists from public transport commuters.
In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1779–1788. ACM, 2014.

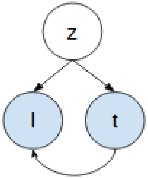
H. Yin, B. Cui, Z. Huang, W. Wang, X. Wu, and X. Zhou.
Joint modeling of users' interests and mobility patterns for point-of-interest recommendation.
In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 819–822. ACM, 2015.

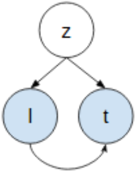
H. Yin, X. Zhou, Y. Shao, H. Wang, and S. Sadiq.
Joint modeling of user check-in behaviors for point-of-interest recommendation.
In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 1631–1640. ACM, 2015.

Y. Zheng, L. Capra, O. Wolfson, and H. Yang.
Urban computing: concepts, methodologies, and applications.
ACM Transactions on Intelligent Systems and Technology (TIST), 5(3):38, 2014.

Work underway!

Space-time Interdependence





Bag of Words Assumption

How much valid is it?