Detecting communities of commuters: Graph Based Techniques versus Generative Models

Ashish Dandekar[1], Stéphane Bressan[1], Talel Abdessalem[1,2,3], Huayu Wu[4], and Wee Siong Ng[4]

[1] National University of Singapore, Singapore
ashishdandekar@u.nus.edu
steph@nus.edu.sg
[2] Telecom ParisTech, Paris-Saclay University, Paris, France
talel.abdessalem@telecom-paristech.fr
[3] IPAL, CNRS
[4] Institute of Infocomm Research, A*STAR, Singapore
huwu@i2r.a-star.edu.sg
wsng@i2r.a-star.edu.sg

Abstract. The main stage for a new generation of cooperative information systems are smart communities such as smart cities and smart nations. In the smart city context in which we position our work, urban planning, development and management authorities and stakeholders need to understand and take into account the mobility patterns of urban dwellers in order to manage the sociological, economic and environmental issues created by the continuing growth of cities and urban population. In this paper, we address the issue of the detection of communities of commuters which is one of the crucial aspects of smart community analysis.

A community of commuters is a group of users of a public transportation network who share similar mobility patterns. Existing techniques for mobility patterns analysis, based on spatio-temporal data clustering, are generally based on geometric similarity metrics such as Euclidean distance, cosine similarity or variations of edit distance. They fail to capture the intuition of mobility patterns, based on recurring visitation sequences, which are more complex than simple trajectories with start and end points.

In this work, we look at visitations as observations for generative models and we explain the mobility patterns in terms of mixtures of communities defined as latent topics which are seen as independent distributions over locations and time. We devise generative models that match and extend Latent Dirichlet Allocation (LDA) model to capture the mobility patterns. We show that our approach, using generative models, is more efficient and effective in detecting mobility patterns than traditional community detection techniques.

Keywords: Urban Computing, Smart cities, LDA, Community Detection, Human mobility

1 Introduction

Urban planning, development and management authorities and stakeholders need to understand the mobility of urban dwellers in order to manage the sociological, economic and environmental issues created by the continuing growth of cities and urban population. We propose a novel approach to the detection of communities of commuters in an effort to assist urban development authorities and other stakeholders to analyze transportation demand and needs and provide customized and adaptive public transport services and ancillary services.

The study of human mobility concerns with answering the question of where was a particular user(who) was at what time and for which purpose[1]. Such a kind of data about urban dwellers is available for study via automated fare collection (AFC) system which smart cities employ for their public transportation networks. The system facilitates commuters with the cards equipped with electronic chips which track the timestamped origin and destination location of the commuters and accordingly calculates the fare for the journey. This sequence of tappings on a card of a particular commuter defines the mobility of the commuter in the city in terms of where the commuter was at what time. A community of commuters is a group of users of a public transportation network who share similar mobility patterns as recorded by their automated fare collection cards.

Existing methods of community discovery are either graph based or similarity based. On the one hand, graph based techniques [2–5] use the connectivity information to find cohesive communities. There are two major concerns related to graph based community detection techniques. Firstly, user graph is not always available due to privacy concerns. Secondly, these techniques do not scale efficiently as the size and density of the graph increases. On the other hand, similarity based techniques [6–8] require a metric to compute the similarity between two objects. The techniques used for clustering spatio-temporal data and trajectories are generally based on geometric similarity metrics such as Euclidean distance [6], cosine similarity or variations of edit distance [9]. They fail to capture the intuition of mobility patterns, which are recurring traveling patterns, are more complex than simple trajectories with start and end points.

In this paper we propose an approach based on generative models to find communities of commuters from the spatio-temporal information stored on automated fare collection cards. We consider timestamped visits of the commuters as observations and form communities based on the latent topics, spatial and temporal, found by the generative model. We show that such a method of community discovery is not only efficient but also effective compared to community detection techniques.

The rest of the paper is organized as follows. Section 2 delineates the related work. Section 3 presents the intuition along with the formal terminology. In Section 4 describes the proposed method of generative models. We present the experiments and the evaluation in Section 5. Additionally, Section 5 describes the dataset and the data preprocessing steps. We conclude the paper by discussing the work underway in Section 6.

2 Related Work

Related work spans three different domains of research, namely *urban computing*, *Latent Dirichlet Allocation (LDA)* and *community detection*, as well as their mutual overlap.

Urban Computing [10] is a process of acquisition, integration, and analysis of big and heterogeneous data generated by diverse sources in urban spaces, such as sensors, devices, vehicles, buildings, and humans, to tackle the major issues that cities face. It also helps understanding the latent trends and foresee the development of the city. Public transport data has been studied by authors for the betterment of the mobility of the citizens inside the cities. In [11–13], London public transport has been widely studied for exploring traveling behaviors, minimizing traveling time and finding communities of citizens. Ferris et al. [14] have developed an app *OneBusAway* to reduce the waiting time of commuters by providing real-time bus arrival information in King county, Washington. To identify tourists from the daily commuters of the public transportation services Xue et al. [15] have devised a method and tested it on the data from Singapore. Montis et al. [16] have found communities of commuters to help in demarcation of the sub-regions in Sardinia. In [17], authors have used to spatio-temporal data generated from social networks to find mobility patterns in urban area.

Blei et al. [18] have proposed **Latent Dirichlet Allocation** (LDA) - a soft clustering technique used for finding latent topics by intuitively capturing the co-occurrence of the words in the textual corpora. Till the date, it is a widely used technique for topic discovery. In [19, 20], authors have altered the original model to handle geospatial data. LDA has been used to find the group of places in cities using the check-in data from LBSN Foursquare users [21–23]. They have shown that LDA has the ability to cluster the places based on the hidden user interests than their geographical proximity. In the GeoFolk model proposed by Sizov [24], author has handled the spatial aspect by considering longitude and latitude to be sampled independently from uniform distributions. Hu et al. [25] have jointly modeled longitude and latitude as bivariate Gaussian distribution. Yin et al. [26, 27] have extended joint spatial modeling by incorporating time of visit and textual discription of Point of Interests into the model. Liu et al. [28] propose two models to explore dependence of time on spatial visits of users. Unlike these works, the present work focuses on the comparative study of generative models and graph based techiniques of community discovery. Further, we use the generative process of the proposed model to produce synthetic data.

Community detection in graphs is a widely studied topic. There are different ways in which these algorithms are classified. Girvan-Newmann Algorithm [2] is a divisive algorithm which finds communities by removing the edges with high *betweenness*. Louvain [4] and fastgreedy [3] algorithms belong to a class of community detection algorithms which optimizes the *modularity*. There are spectral as well as randomized algorithms for community discovery. Please refer to [5] for an extensive survey of existing techniques.

To the best of our knowledge, we are the first ones to propose generative model for detecting communities of commuters in the public transportation networks.

3 Preliminaries

This section defines some useful terminology which we use throughout the paper and some primary ideas on the concept of community as we consider it in this work.

3.1 Terminology

Although a single commuter may possess two or more cards which she can use interchangeably, we treat each unique card as a single **commuter**. Let $\mathcal{C}$ denotes the set of all commuters, i.e. a set of distinct card numbers in the dataset. A commuter can visit any location l from a finite set of locations $\mathcal{L}$ at any time t from a set of time segments $\mathcal{T}$.

The list of spatio-temporal visits done by a commuter as recorded on the corresponding card is defined as **mobility** of the commuter. For instace mobility of a commuter c, denoted as $\mathcal{M}_c$, is a multiset $\{l-t|l \in \mathcal{L}, t \in \mathcal{T}\}$. Each element of $\mathcal{M}_c$ is called as a **visit** of a commuter c. Sometimes, we want to focus on either spatial or temporal mobility of a commuter instead of her spatio-temporal visits. We denote spatial mobility of a commuter c as $\mathcal{M}_c^l$ which is a multiset $\{l|l \in \mathcal{L}\}$. Similarly, temporal mobility of a commuter c is denoted as $\mathcal{M}_c^t$ which is a multiset $\{t|t \in \mathcal{T}\}$.

Table 1 lists the notation which are used throughout the paper.

3.2 Communities of Commuters

The central problem of the current work is to find communities of commuters using their mobility data. A community is a group of commuters who share *similar* mobility patterns. So, the notion of similarity is crucial in the task of classification. Intuitively, two commuters are to be placed in the same community if their mobilities overlap to a significant extent. Given the spatio-temporal nature of the data, one can envisage the similarity in mobility in multiple ways: as an overlap among the recurrent visits to locations or as an overlap among the time of travels or as an overlap among *spatio-temporal* visits. In our approach, we devise generative models which are able to capture these different intuitions of similarity.

4 Generative Models

The probabilistic graphical model, LDA [18], finds latent topics in the text corpora. LDA assumes every document as a distribution over K topics, K being

Symbol	Description
$\mathcal{C}$	Set of commuters
$\mathcal{L}$	Set of locations
$\mathcal{T}$	Set of time segments
K	Total number of topics
$\mathcal{M}_c$	Mobility of a commuter c
$\mathcal{M}_c^t$	Temporal Mobility of a commuter c
$\mathcal{M}_c^l$	Spatial Mobility of a commuter c
α	Dirichlet prior over topics
β	Dirichlet prior over locations
γ	Dirichlet prior over time-segments
θ_c	Topic distribution of a commuter c
ϕ_k	Location distribution of a topic k
ψ_k	Time distribution of a topic k
z	Latent topic
t	Timestamp
l	Location

Table 1: Notations

an input to the model, and each of the K topics as the distribution over words. This makes LDA a probabilistic clustering technique that considers overlapping clusters. The assumption of LDA that a document is a probability distribution of over the topics befits the problem at our hand. Intuitively, a commuter belongs to different social communities in the city. So, a commuter with his mobility can be seen as a distribution over K communities. Similarly, each community can be seen as distribution over the visits.

Inspired by this analogy, we adopt and extend LDA to different generative models.

4.1 SLDA and TLDA

In LDA, there is only one observed variable *viz.* word in the document. So if we consider document as either spatial or temporal visits of a commuter, we can directly adopt LDA to find latent topics. Considering, every commuter as a document, a bag of locations which she has visited, we can find communities of commuters who share an overlap among the locations they visit. Therefore intuitively, this *spatial* adoption of LDA - SLDA, shown in the Figure 1, captures spatially cohesive topics. Similarly, if we consider every commuter as a document, a bag timestamps at which she reaches certain locations, we can find communities who share an overlap among the periods of time at which commuters travel. This *temporal* adoption of LDA, TLDA, is also shown in Figure 1. Depending on the segmentation of time variable, topics found by TLDA carry different semantics. If we fold time on hourly basis everyday, then we find daily temporal patterns of commuters such as people travelling in the afternoon, working people who go

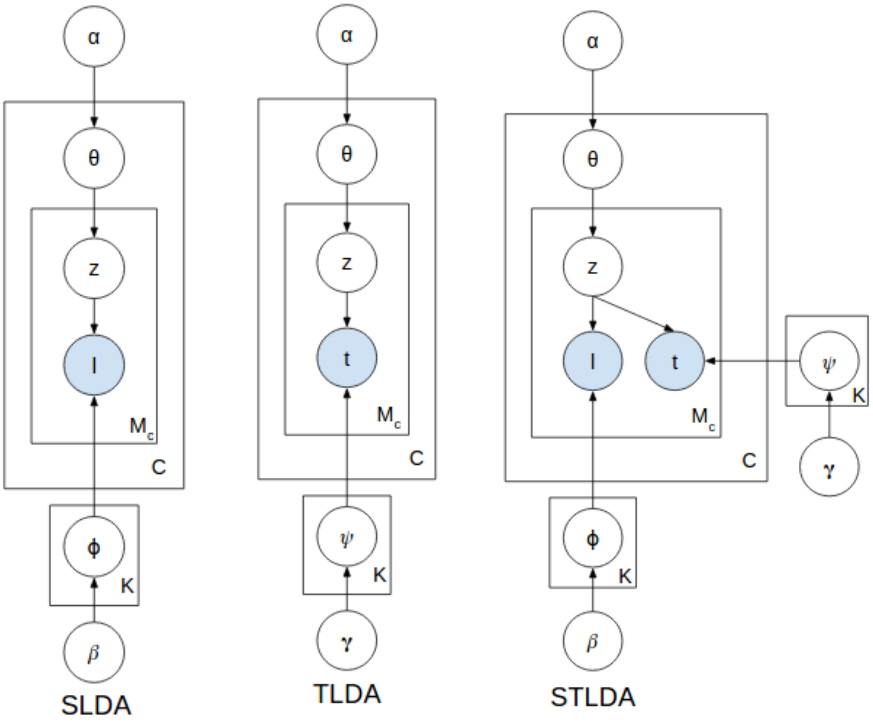


Fig. 1: Generative Models

to work in the morning and get back to home in the evening, etc. If we fold time on weekly basis, accordingly, we find weekly temporal patterns.

The generative process for SLDA for a commuter c is as follows:

1. For every topic k:
   (a) Sample $\phi_k \sim Dirichlet(\beta)$
2. For every commuter $c \in \mathcal{C}$:
   (a) Sample $\theta \sim Dirichlet(\alpha)$
   (b) For every visit v by c:
      i. Sample $z \sim Multinomial(\theta)$
      ii. Sample $l \sim Multinomial(\phi_z)$

In case of TLDA, we have similar generative process except that we sample time $t \sim Mult(\psi_z)$ where ψ_z is the distribution over time unlike ϕ_z which is the distribution over locations.

4.2 STLDA

As the words are the only observables in the original LDA model, we can not directly use it to find the communities with spatial and temporal overlap. We

extend the LDA model, to spatio-temporal LDA (STLDA), by assuming independent sampling distributions for both location and time for every visit of a commuter. Commuters go to a certain place at a certain time for a definite purpose. For instance, if location l_1 is office place of a commuter c then it is highly improbable that he visits it at night or after working hours. With this intuition, we assign a single topic to each spatio-temporal visit of a commuter which leads to the assignment of the same topic to both location and time of travel in the visit. Figure 1 shows the plate diagram for STLDA.

The generative process for STLDA for a commuter c is as follows:

1. For every topic k:
   (a) Sample $\phi_k \sim Dirichlet(\beta)$
   (b) Sample $\psi_k \sim Dirichlet(\gamma)$
2. For every commuter $c \in \mathcal{C}$:
   (a) Sample $\theta \sim Dirichlet(\alpha)$
   (b) For every visit v by c:
      i. Sample $z \sim Multinomial(\theta)$
      ii. Sample $l \sim Multinomial(\phi_z)$
      iii. Sample $t \sim Multinomial(\psi_z)$

4.3 Inference

For learning parameters of our models, we use Gibbs sampling based inference of LDA as presented in [29]. The backbone framework for Gibbs sampling for all of the proposed models remain the same except the update equation which changes for every model. Algorithm 1 shows Gibbs sampling based inference of STLDA. SLDA and TLDA follow the same algorithm except that we ignore timestamp and location in their sampling procedures respectively.

Update equations are given as follows:

$$P(z_i = k|\bar{z}_{\neg i}, \bar{l}) \propto \frac{n^{(l)}_{kl,\neg i} + \beta_l}{\sum_{l=1}^{\mathcal{L}} n^{(l)}_{kl,\neg i} + \beta_l} \left(n^{(k)}_{c,\neg i} + \alpha_k\right) \quad \text{(SLDA)}$$

$$P(z_i = k|\bar{z}_{\neg i}, \bar{t}) \propto \frac{n^{(t)}_{kt,\neg i} + \gamma_t}{\sum_{t=1}^{\mathcal{T}} n^{(t)}_{kt,\neg i} + \gamma_t} \left(n^{(k)}_{c,\neg i} + \alpha_k\right) \quad \text{(TLDA)}$$

$$P(z_i = k|\bar{z}_{\neg i}, \bar{l}, \bar{t}) \propto \frac{n^{(l)}_{kl,\neg i} + \beta_l}{\sum_{l=1}^{\mathcal{L}} n^{(l)}_{kl,\neg i} + \beta_l} \frac{n^{(t)}_{kt,\neg i} + \gamma_t}{\sum_{t=1}^{\mathcal{T}} n^{(t)}_{kt,\neg i} + \gamma_t} \left(n^{(k)}_{c,\neg i} + \alpha_k\right) \quad \text{(STLDA)} \tag{1}$$

where $n^{(z)}_c$ denotes the number of times a visit with a topic z is observed for a commuter c, $n^{(l)}_{zl}$ denotes the number of times a topic z has been assigned to a location l and $n^{(t)}_{zt}$ denoted the number of times a topic z has been assigned to a time segment t. $\neg i$ denotes removal of i^{th} visit from the mobility. Parameters

Algorithm 1 Gibbs sampling based inference

//Initialization
zero all count variables $n_c^{(z)}, n_c, n_{zl}^{(l)}, n_{zl}, n_{zt}^{(t)}, n_{zt}$
for all commuters $c \in [1, 2, ..., \mathcal{C}]$ **do**
  for all spatio-temporal words $w \in [1, 2, ..., \mathcal{M}_c]$ **do**
    Sample topic $z \sim Mult(1/K)$
    Increment document-topic count $n_c^{(z)} + 1$
    Increment document-topic sum $n_c + 1$
    Increment topic-location count $n_{zl}^{(w_l)} + 1$
    Increment topic-location sum $n_{zl} + 1$
    Increment topic-time count $n_{zt}^{(w_t)} + 1$
    Increment topic-time $n_{zt} + 1$
  end for
end for
//Gibbs Sampling
while not converged **do**
  for all commuters $c \in [1, 2, ..., \mathcal{C}]$ **do**
    for all spatio-temporal words $w \in [1, 2, ..., \mathcal{M}_c]$ **do**
      k = currently assigned topic for w
      Decrement counts $n_m^{(k)}, n_m, n_{kl}^{(w_l)}, n_{kl}, n_{kt}^{(w_t)}, n_{kt}$ by 1
      Z = sample new topic according to Equation 1
      Increment counts $n_m^{(z)}, n_m, n_{zl}^{(l)}, n_{zl}, n_{zt}^{(t)}, n_{zt}$ by 1
    end for
  end for
end while
Output ϕ, θ, ψ according to Equation 2

of the generative model are given by:

$$\begin{aligned} \theta_{c,k} &= \frac{n_c^k + \alpha_k}{\sum_{k=1}^{K} n_c^k + \alpha_k} \\ \phi_{k,l} &= \frac{n_{zl}^k + \beta_l}{\sum_{l=1}^{\mathcal{L}} n_{zl}^k + \beta_l} \\ \psi_{k,t} &= \frac{n_{zt}^k + \gamma_t}{\sum_{t=1}^{\mathcal{T}} n_{zt}^k + \gamma_t} \end{aligned} \tag{2}$$

5 Experimental Evaluation and Analysis

In this section we evaluate the performance of the proposed method on both synthetic and real dataset. The synthetic data is constructed in a realistic manner based on experts' experiences with real datasets. The real dataset is a snapshot of Singapore public transportation automated fare system. We compare the effectiveness and efficiency of the proposed method with one of the widely used graph

based community detection algorithms. We present the performance evaluation along with quantitative and qualitative analysis of the proposed technique.

All programs are run on a Linux machine with quad core 2.40GHz Intel® Core i7™processor and 8GB memory. Python® 2.7.6 is used as a scripting language. We have used the Java library, *jLDADMM*, developed by Dat et al. [30] for finding the latent topics using LDA[5] and modified the source to adapt to proposed extension of LDA. For the graph computation, we use Python wrapper for an highly-efficient *igraph*[6] library written in C.

5.1 Qualitative Evaluation on Real Dataset

We now qualitatively evaluate the performance of our method on a snapshot of automated fare collection system data of Singapore public transportation which comprises of buses, Mass Rapid Transport (MRT) and Light Rail Transit (LRT).

Dataset Description The dataset comprises of tappings made by commuters of public transportation system consisting of MRT/LRT stations and bus stops, identified by the cards' unique identifier. Please refer to Table 2 for the detailed dataset schema. Figure 2 shows statistics of commuters over 25 days of a month. One can identify recurrent pattern in the travels. We expect the majority of commuters to travel between home and work place during weekdays and between home and shopping or leisure centres during weekends. The pattern seems to break on 25^{th} which is a public holiday in Singapore.

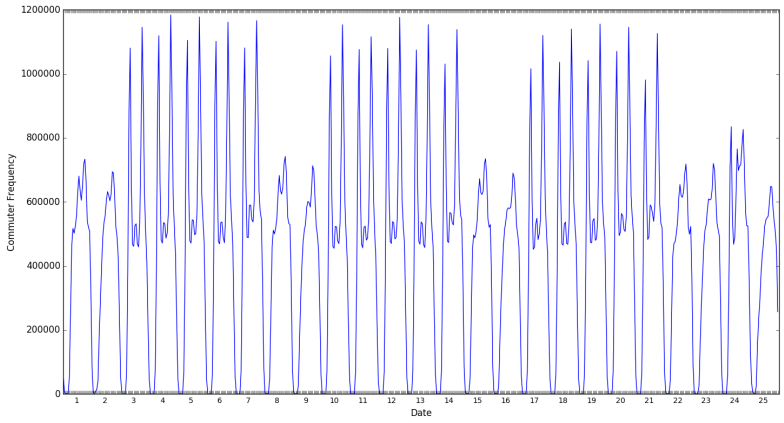


Fig. 2: Statistics of commuters over a 25 consecutive days

[5] http://jldadmm.sourceforge.net/
[6] http://igraph.org/redirect.html

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 2: Dataset Schema

We prepare two datasets comprising 41000 commuters over 4985 locations, The first dataset spans over weekdays in a typical week. The second dataset spans over one weekend. Time is considered at the granularity of an hour on every day. In this way, a document after these preprocessing is a bag of geospatial timestamps. A typical word looks like - "56-20"- which means location 56 at time 20:00.

We perform 2000 cycles of Gibbs sampling iteration to obtain the communities by LDA. Every commuter is assigned to the community with the maximum probability.

Result Analysis Figure 3 shows the topics for weekdays and Figure 4 for weekends when one ignores the temporal dimension. In the map, top 10 places for each topic are shown. The lines in the map correspond to the MRT lines in Singapore. Although we start without any assumption related to geographic proximity, the observed clusters (topics) correspond to some segments of the MRT lines.

A closer observation of Figure 3 shows that the topics are concentrated towards the southern part of Singapore, known as the Central Business District (CBD), which hosts numerous workplaces. For instance, topics 0, 1, 5, 7, which appear at the periphery of Singapore, correspond to the segments of the MRT lines passing through residential areas whereas topics 2 and 4 show the transition from residential areas to business district. Typically, if we look at topic 4 it links the area with high-tech companies and dissolves in the business district. Topic 3 exclusively captures the CBD area in Singapore. So, it asserts the hypothesis that on weekdays, people commute between the residential areas to the work places.

Compared to Figure 3, topics in Figure 4 represent segments of the MRT line in the peripheral parts of Singapore, mostly near the shopping malls close to residential areas. For instance, for people living in western residential areas, topic 3 and 6, Jurong East, which appears at the intersection of two topics, is a popular weekend destination. Orchard, which is situated in CBD, is one of the

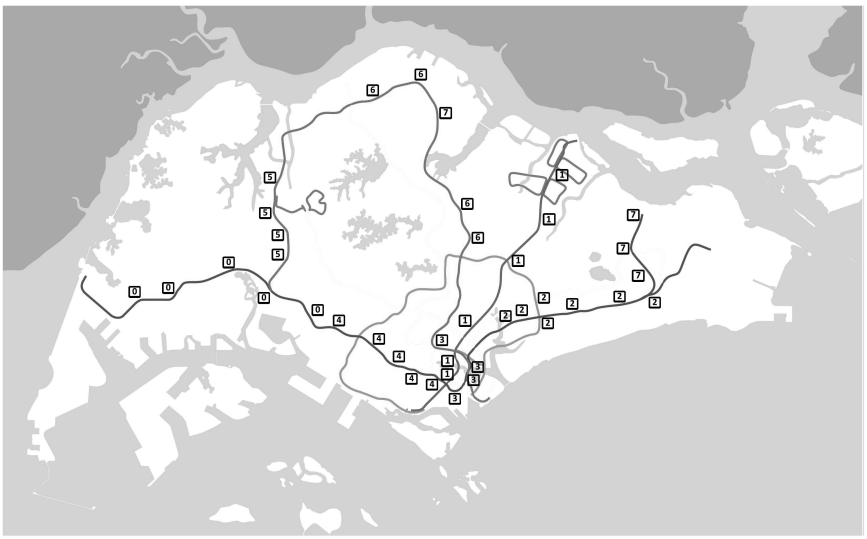


Fig. 3: Weekday topics for SLDA

most popular shopping place in Singapore and it appears in Topic 0, 2, 4, 5 in Figure 4. Closer observation tells that all the topics which appear at Orchard are spatially closer to CBD. So, it asserts the hypothesis that on weekends people travel from residential areas to the shopping or leisure centers in their vicinity.

Figure 5 shows the topics found by TLDA model. We observe topics in both the weekend and the weekdays dataset capture similar temporal trends. In Figure 5, we show a selection of them: Topics 4, 5 and 7 are from weekdays dataset and topic 2 is from weekend dataset. We observe that such model captures temporal patterns of travel. We clearly see that in topic 7 which is peaked around 9 a.m. and 6 p.m. captures the community of the working people. By observing topic 2, we see that there is a community of people who work on weekends. Although both of them denote the communities of working class, there is a subtle difference between them. On weekdays, the distribution is sharply peaked around 9 a.m. whereas in topic 2, which is from weekend dataset, we see the distribution to be flattened. On the weekdays, the working professionals, going to business district, need to follow strict working hours. On the weekends, shops open at different times and so the workers travel at different times. Aside from these, we have topics which capture the people travelling trends in the night (topic 4) as well as in the afternoon (topic 5).

Figure 6 and Figure 7 show the topics for weekday when we consider both spatial and temporal dimension independently for each other. All the topics

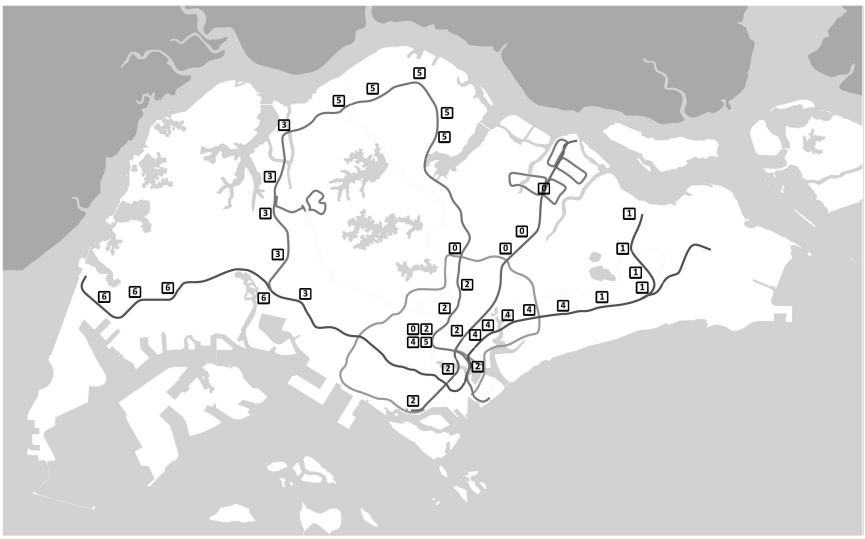


Fig. 4: Weekend topics for SLDA

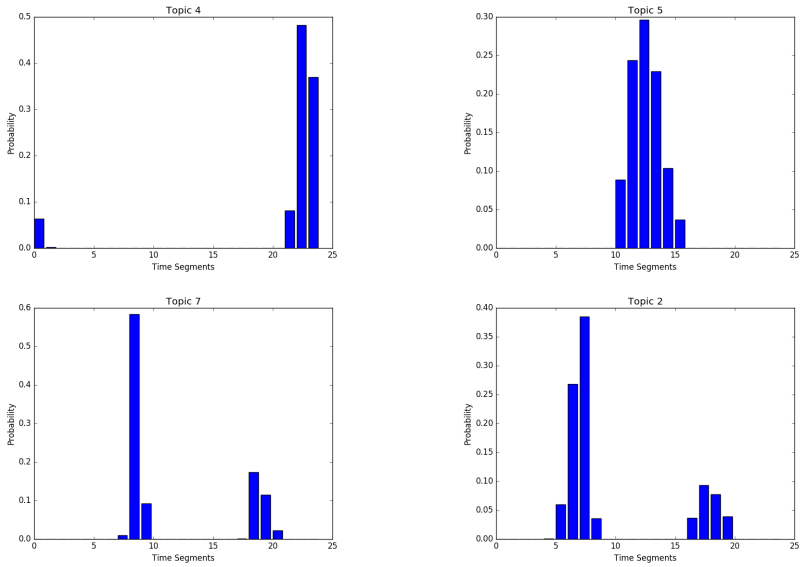


Fig. 5: Topics for TLDA

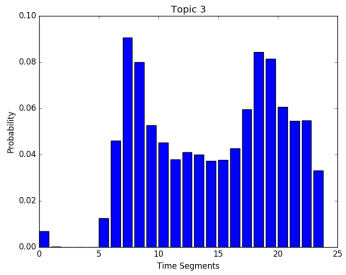


Fig. 6: Weekday temporal topic (STLDA)

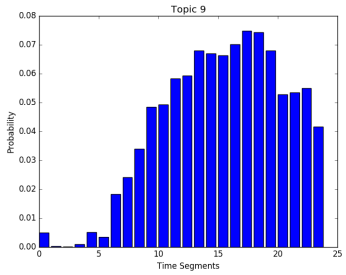


Fig. 7: Weekend temporal topic (STLDA)

found on the weekday follow the typical bimodal pattern shown in Figure 6. Similarly, all of the topics found on the weekend follow the pattern in which we do not observe high traffic in the morning before 9 a.m. as shown in Figure 7. We observe that when one considers location and time simultaneously, the temporal topics not only follow the pattern of global statistics shown in Figure 2 but also do not significantly differ among themselves except a few fluctuations within a window of an hour or two.

The reasons are attributed to the area of Singapore and connectivity among different regions in it. Singapore is a city nation where majority of the population lives in the suburbs and travel to downtown for working. Singapore is highly connected with an efficient public transport system[7]. So, a commuter in Singapore does not take more than 90 minutes to reach one place from other place.

Comparison with Graph Based Technique Since we do not have groundtruth communities for the real dataset, we can not objectively perform comparative evaluation of the effectiveness of the proposed models with graph based techniques. So as to have fairness and reproducibility in evaluation, we extend our experiments by conducting evaluation on a synthetic dataset, generated in a realistic manner. The results are presented in the next section.

The experiments with one of the widely used graph based technique (Louvain algorihtm) on the real dataset finds many sparse communities of commuters for both weekend and weekdays, out of which less than a dozen contains more than 3 commuter. Louvain algorithm on sparse graph fails to find large cohesive communities and instead finds many small communities. An attempt to visualize them does not lead to valuable qualitative insight.

[7] http://www.worldcitiessummit.com.sg/sites/sites2.globalsignin.com.2.wcs-2014/files/Smart_Mobility_Innovative_Solutions.pdf

5.2 Quantitative Evaluation on Synthetic Dataset

In this section, we illustrate the data generation process and show that the proposed models are more effective than the Louvain algorihtm. We further reason why one observes the fall in the effectiveness of any graph based algorithm compared to the generative models.

Dataset Description We generate a synthetic yet realistic dataset. Following our findings with the real dataset, we consider 10 communities with 1000 different commuters in each one. We consider 10,000 different locations. From a statistical analysis of the real data, we observe that every community has a skewed distribution over the places. Very few places have a high probability to be visited by commuters in the community while the majority of remaining places have negligible probability to be visited. Agreeing to these observations, we distribute places over communities. Each community is characterized by a probability distribution of the places visited by commuters of the community. Additionally, we observe, in the real dataset, that a commuter generally visits up to 8 places. For every commuter, we sample the number of her visits from a Gamma distribution[8]. We then sample her actual visits from the distribution of places in her community. This distribution is given by a Zipf distribution for an order of places following the probability distribution that characterizes her community.

Comparison with Graph Based Technique We compare our method with a state of the art graph community detection method: Louvain algorithm [4]. It is a graph partitioning technique that provides a computationally efficient approach to detect communities through greedy optimization of *modularity*. In order to compare the proposed method with a graph clustering method, we construct corresponding *commuter graph* in which commuters are the vertices of the graph. We add an edge between two vertices of the graph whenever the corresponding two commuters share a common visit. The graph has no self-loops. Edges are weighted by the number of common visits between the two commuters that the two vertices represent.

We have verified that the parameter θ of the Zipf distribution controls the density of the graph. The higher the value of θ, more skewed the distribution and denser the corresponding graph. As the value of θ increases, fewer places have high probability to be visited while the probability of the remaining places to be visited drops off more quickly. With high values of θ, commuters in a community are more likely to visit the same few places with high probability. Hence, the density of the graph increases. Figure 8 shows the almost linear increase of density of the graph with the increase of the value of parameter θ.

Results Analysis We evaluate the **effectiveness** of the proposed method comparatively to the Louvain algorithm against the groundtruth which comprises

[8] shape parameter is set to 6 and scale parameter is set to 1.2

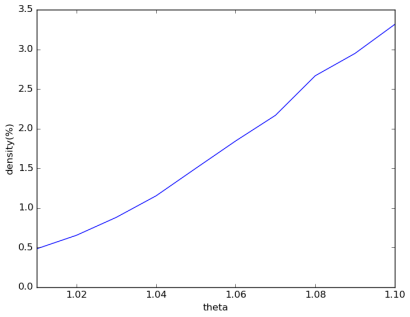


Fig. 8: Density variation with parameter theta

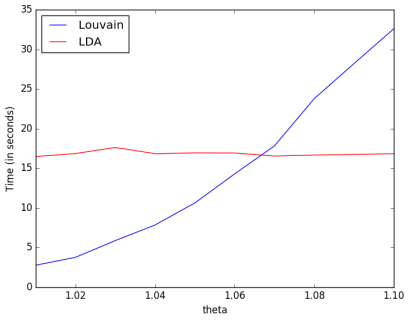
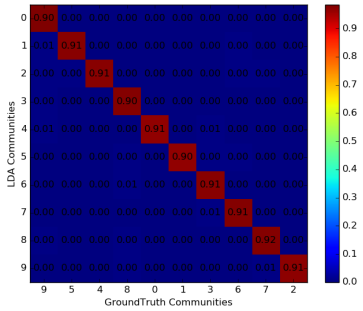
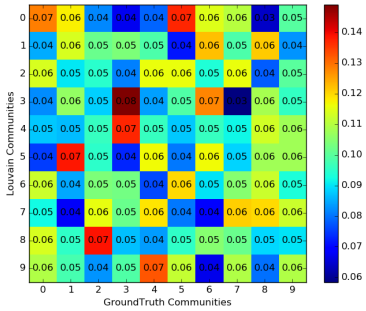


Fig. 9: Comparative Efficiency Evaluation



(a) LDA with $K = 10$



(b) Louvain with top 10 communities

Fig. 10: Comparison of LDA and Louvain with ground truth

of 10 communities as described above. In all ten cases, Louvain algorithm finds communities whose count far exceeds than the number of communities in the groundtruth. However, the top-10 found communities contain more than 97% of the commuters. We consider these top-10 communities for evaluation. We run LDA for K equals 10 and 2000 Gibbs sampling iterations. LDA is able to find a quasi one-to-one mapping to the ground truth. Figure 10 shows the confusion matrices. Each cell in the confusion matrix is annotated by Jaccard similarity between corresponding communities. Due to lack of space, we show the confusion matrix for only one of the datasets ($\theta = 1.1$). The confusion matrices clearly show that LDA is more effective than Louvain. From a closer inspection, Jaccard similarity values in Louvain comparison are very small and suggest a very poor effectiveness for the problem at hand. Results for other values of density which are not presented here in details due to space limitation, confirm that the observation remains consistent.

We observe a drastic drop in the effectiveness in case of Louvain algorithm in Figure 10b. Reasons for such an observation are rooted at the inner details of the two techniques. LDA uses iterative Gibbs sampling for inference which changes the topics *viz.* the distribution over visits in every iteration. Accordingly, the allottment of commuters, corresponding topic also changes in every iteration. The cohesion of topics goes on improving over the iteration until convergence. In contrast, when we weigh the edges in the graph by co-occurrences, we lose this dimension of topics defined over visits. We do no more have freedom to probe whether the co-occurring places on the edge (which we have taken into account by assigning weight) belong to a same topic or not. Once the graph is constructed, the network structure becomes agnostic to the qualitative notion of topics as clusters of visits. This loss of degree of freedom hampers the effectiveness of graph based community detection techniques.

We comparatively evaluate the **efficiency** of the proposed algorithm and that of the Louvain algorithm on ten synthetic datasets for varying values of the parameter θ. Figure 9 shows the running time for the two algorithms. We see that Louvain algorithm is more efficient for sparser graphs. We get this efficiency at the cost of effectiveness since for sparse graphs we get large number of small communities.

Thus, the use of generative model that we propose is qualitatively effective and its algorithm with Gibbs sampling practically efficient.

6 Discussion and Future work

We propose a generative model to the detection of communities of commuters from data on automated fare collection cards. A community of commuters is a group of users of a public transportation network who share similar mobility patterns. We argue that both trajectory clustering techniques and graph-based community detection techniques are inadequate. The former are inadequate because of the metrics they rely on that cannot cater for the complexity of the mobility of commuters. We empirically show that the latter are less efficient and less effective for practical purposes than the method we propose. We consider mobility as a sequence of observations for a Latent Dirichlet Analysis and we explain the mobility patterns in terms of mixtures of communities defined as latent topics in this generative statistical model.

We confirm empirically, using a synthetic but realistic series of datasets, that our method is effective and efficient. We also confirm empirically, using a real dataset from Singapore, that our method yields qualitatively meaningful results. It is able to reconstitute recognizable typical mobility patterns of the Singapore population.

As future work, we are extending our approach to visits that cannot be characterized by a known set of fixed locations. This is the case of the trajectories of moving objects, commuters and vehicles reporting their positions by means of devices equipped with positioning systems. For this we are considering to learn the similarity spatio-temporal model rather than rely on identified locations and

time intervals. We are now refining the Latent Dirichlet Allocation model constructed from the real data in order to use as a generative model for simulation. Such a model would give us the ability to generate highly realistic data, yet synthetic data, thus addressing the issue of privacy in big data analytics.

Acknowldgement

This research is funded by research grant R-252-000-622-114 by Singapore Ministry of Education Academic Research Fund (project 251RES1607 - Janus: Effective, Efficient and Fair Algorithms for Spatio-temporal Crowdsourcing) and is a collaboration between the National University of Singapore, Télécom ParisTech and Singapore Agency for Science, Technology and Research.

References

1. Gao, H., Liu, H.: Mining human mobility in location-based social networks. Synthesis Lectures on Data Mining and Knowledge Discovery **7**(2) (2015) 1–115
2. Girvan, M., Newman, M.E.: Community structure in social and biological networks. Proceedings of the national academy of sciences (2002) 7821–7826
3. Clauset, A., Newman, M.E., Moore, C.: Finding community structure in very large networks. Physical review E (2004) 066111
4. Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. Journal of statistical mechanics: theory and experiment (2008) P10008
5. Fortunato, S.: Community detection in graphs. Physics reports (2010) 75–174
6. Agrawal, R., Faloutsos, C., Swami, A.: Efficient similarity search in sequence databases. Springer (1993)
7. Vlachos, M., Kollios, G., Gunopulos, D.: Discovering similar multidimensional trajectories. In: Data Engineering, 2002. Proceedings. 18th International Conference on, IEEE (2002) 673–684
8. Wang, H., Su, H., Zheng, K., Sadiq, S., Zhou, X.: An effectiveness study on trajectory similarity measures. In: Proceedings of the Twenty-Fourth Australasian Database Conference-Volume 137, Australian Computer Society, Inc. (2013) 13–22
9. Chen, L., Özsu, M.T., Oria, V.: Robust and fast similarity search for moving object trajectories. In: Proceedings of the 2005 ACM SIGMOD international conference on Management of data, ACM (2005) 491–502
10. Zheng, Y., Capra, L., Wolfson, O., Yang, H.: Urban computing: concepts, methodologies, and applications. ACM Transactions on Intelligent Systems and Technology (TIST) **5**(3) (2014) 38
11. Lathia, N., Capra, L.: 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, ACM (2011) 1181–1189
12. Lathia, N., Capra, L.: How smart is your smartcard?: measuring travel behaviours, perceptions, and incentives. In: Proceedings of the 13th international conference on Ubiquitous computing, ACM (2011) 291–300
13. Lathia, N., Quercia, D., Crowcroft, J.: The hidden image of the city: sensing community well-being from urban mobility. In: Pervasive computing. Springer (2012) 91–98

14. Ferris, B., Watkins, K., Borning, A.: Onebusaway: results from providing real-time arrival information for public transit. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM (2010) 1807–1816
15. Xue, M., Wu, H., Chen, W., Ng, W.S., Goh, G.H.: Identifying tourists from public transport commuters. In: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM (2014) 1779–1788
16. De Montis, A., Caschili, S., Chessa, A.: Commuter networks and community detection: a method for planning sub regional areas. The European Physical Journal Special Topics (2013) 75–91
17. Al-Ghossein, M., Abdessalem, T.: Somap: Dynamic clustering and ranking of geotagged posts. In: Proceedings of the 25th International Conference Companion on World Wide Web, International World Wide Web Conferences Steering Committee (2016) 151–154
18. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. the Journal of machine Learning research (2003) 993–1022
19. Yuan, J., Zheng, Y., Xie, X.: Discovering regions of different functions in a city using human mobility and pois. In: Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM (2012) 186–194
20. Hu, B., Jamali, M., Ester, M.: Spatio-temporal topic modeling in mobile social media for location recommendation. In: Data Mining (ICDM), 2013 IEEE 13th International Conference on, IEEE (2013) 1073–1078
21. Long, X., Jin, L., Joshi, J.: Exploring trajectory-driven local geographic topics in foursquare. In: Proceedings of the 2012 ACM Conference on Ubiquitous Computing, ACM (2012) 927–934
22. Joseph, K., Tan, C.H., Carley, K.M.: Beyond local, categories and friends: clustering foursquare users with latent topics. In: Proceedings of the 2012 ACM Conference on Ubiquitous Computing, ACM (2012) 919–926
23. Cho, Y.S., Ver Steeg, G., Galstyan, A.: Socially relevant venue clustering from check-in data. In: 11th Workshop on Mining and Learning with Graphs, MLG–2013. (2013)
24. Sizov, S.: Geofolk: latent spatial semantics in web 2.0 social media. In: Proceedings of the third ACM international conference on Web search and data mining, ACM (2010) 281–290
25. Hu, B., Ester, M.: Spatial topic modeling in online social media for location recommendation. In: Proceedings of the 7th ACM conference on Recommender systems, ACM (2013) 25–32
26. Yin, H., Cui, B., Huang, Z., Wang, W., Wu, X., Zhou, X.: Joint modeling of users' interests and mobility patterns for point-of-interest recommendation. In: Proceedings of the 23rd ACM international conference on Multimedia, ACM (2015) 819–822
27. Yin, H., Zhou, X., Shao, Y., Wang, H., Sadiq, S.: 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, ACM (2015) 1631–1640
28. Liu, Y., Ester, M., Hu, B., Cheung, D.W.: Spatio-temporal topic models for check-in data. In: Data Mining (ICDM), 2015 IEEE International Conference on, IEEE (2015) 889–894
29. Griffiths, T.L., Steyvers, M.: Finding scientific topics. Proceedings of the National academy of Sciences (suppl 1) (2004) 5228–5235
30. Nguyen, D.Q.: jLDADMM: A Java package for the LDA and DMM topic models. http://jldadmm.sourceforge.net/ (2015)