Latent Dirichlet Allocation and its Application to Temporal Decomposition of Mobility Data

Yaoyang Zhang
SID: 3032114788
Department of Civil and Environmental Engineeirng
UC Berkeley
yaoyangzhang@berkeley.edu

Latent Dirichlet allocation (LDA) is a three-level hierarchical Bayesian model. It is first introduced by Blei, Ng, and Jordan[1], and is mainly used for modeling discrete data such as text corpora. In this report, the LDA model is reviewed and compared with other text modeling models. Two inference methods, variational inference and Gibbs sampling, are then presented. An application of LDA to inference of trip purpose and temporal decomposition of the mobility data is presented at the end.

1 Latent Dirichlet Allocation (LDA): The Model

1.1 Notation

LDA is a generative probabilistic model for collection of discrete data and is usually used in text modeling. So the three entities of the three-level hierarchical model is referred to "words", "documents", and "corpus" in the original paper[1]. They are formally defined as follows:

word A word is the basic unit of discrete data, defined to be an item from a vocabulary indexed by $\{1,...,V\}$. The vth word in the vocabulary is represented by a V-vector w such that $w^v = 1$ and $w^u = 0$ for $u \neq v$.

document A document is a sequence of N words denoted by $\mathbf{w} = (w_1,...,w_N)$, where w_n is the nth word in the sequence.

corpus A corpus is a collection of M words denoted by $D = (\mathbf{w_1},...,\mathbf{w_M})$

1.2 The Model

The basic idea of LDA is that the documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words.

For each document $\mathbf{w_i}$ in the corpus D, the words are assumed to be generated by the following process:

1. For each document $\mathbf{w_i}$ we have a multinomial distribution over T topics, with parameters θ, that is, $P(z_i = j) = \theta_j$, where $z_i = j$ means the topic j is chosen in document $\mathbf{w_i}$. Dirichlet distribution is conjugate to multinomial, so we assume a Dirichlet prior on θ with parameter α, that is, $\theta \sim \text{Dir}(\alpha)$

2. The jth topic is also multinomial over the N words. with parameters ϕ, that is, $P(w_i|z_i = j) = \phi^j_{w_i}$. Similarly, a Dirichlet prior is also assumed on ϕ with parameter β, that is, $\phi \sim \text{Dir}(\beta)$.

3. The distribution over the words is thus

$$P(w_i) = \sum_{j=1}^{T} P(w_i|z_i = j)P(z_i = j)$$

A graphical model representation of the model is in figure 1, where the boxes are "plates" representing replicates. The outer plate represents documents, which are sampled M times, and the inner plate represents the repeated choice of topics and words within a document, which are sampled N times for each document.

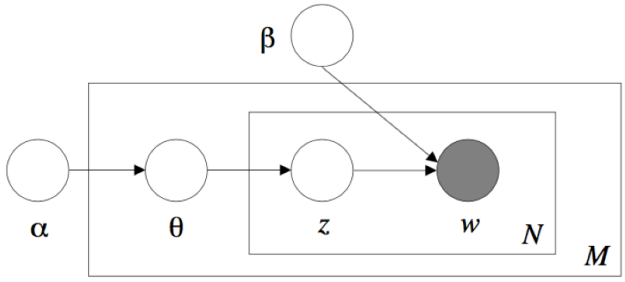


Fig. 1. Graphical model for latent Dirichlet allocation

The Dirichlet distribution is a distribution over distribution. A k-dimensional Dirichlet random variable θ can take values in the $(k-1)$-simplex (that is, $\theta_i \geq 0$, $\sum_{i=1}^{k} \theta_i = 1$),

and has the following probability density on this simplex:

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^{k}\alpha_i)}{\prod_{i=1}^{k}\Gamma(\alpha_i)}\theta_i^{\alpha_i-1}\ldots\theta_k^{\alpha_k-1}$$

where α is a k-vector with components $\alpha_i \geq 0$, and where $\Gamma(x)$ is the Gamma function. The Dirichlet is a convenient because it is in the exponential family, has finite dimensional sufficient statistics and is conjugate to multinomial distribution.

Several probability distributions of interest can thus be derived. The joint distribution of topic mixture θ, N topics $\mathbf{z}$ and N words $\mathbf{w}$ is given by

$$p(\theta,\mathbf{z},\mathbf{w}|\alpha,\beta) = p(\theta|\alpha)\prod_{i=1}^{N}p(z_i|\theta)p(w_i|z_i,\beta)$$

Integrating over θ, we obtain the joint distribution of N topics $\mathbf{z}$ and N words $\mathbf{w}$

$$p(\mathbf{z},\mathbf{w}|\alpha,\beta) = \int p(\theta|\alpha)\prod_{i=1}^{N}p(z_i|\theta)p(w_i|z_i,\beta)d\theta$$

Summing over $\mathbf{z}$, we obtain the marginal distribution of a document $\mathbf{w}$

$$p(\mathbf{w}|\alpha,\beta) = \int p(\theta|\alpha)\prod_{i=1}^{N}\sum_{z_i}p(z_i|\theta)p(w_i|z_i,\beta)d\theta$$

1.3 Relationship with Other Models

1.3.1 Unigram Model

The unigram model assumes that the words of every document are drawn independently from a single multinomial distribution

$$p(\mathbf{w}) = \prod_{i=1}^{N}p(w_i)$$

This model is illustrated in figure 2(a).

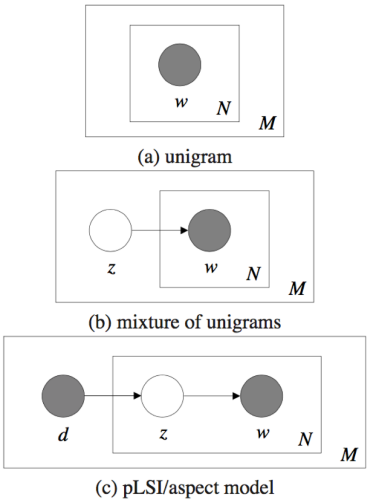


Fig. 2. (a) Graphical model for unigram model (b) Graphical model for mixture of unigrams (c) Graphical model for probabilistic latent semantic indexing

1.3.2 Mixture of Unigrams

If we add latent topics to the unigram model, we obtain the mixture of unigrams model[2]. The probability density of a document is

$$p(\mathbf{w}) = \sum_{z}p(z)\prod_{i=1}^{N}p(w_i|z)$$

This model assumes that each document exhibits exactly one topic, which is a quite limited assumption for practice. This model is illustrated in figure 2(b).

1.3.3 Probabilistic Latent Semantic Indexing (pLSI)

The probabilistic latent semantic indexing (pLSI) is a widely used document model [3]. It assumes that a document label d and a word w_i are conditionally independent given an unobserved topic z

$$p(d,w_i) = p(d)\sum_{z}p(w_i|z)p(z|d)$$

pLSI does take into consideration the probability that a document can contain more than one topics. However d is a multinomial random variable with as many possible values as there are training documents and the model learns the topic mixtures $p(z|d)$ only for those documents on which it is trained. For this reason, pLSI is not a well-defined generative model of documents; there is no natural way to use it to assign probability to a previously unseen document. Moreover, number of parameters which must be estimated grows linearly with the number of training documents and the model is prone to overfitting.

This model is illustrated in figure 2(c).

2. Inference Methods

The inference problem is to compute the posterior distribution of the hidden variables given a document

$$p(\theta, \mathbf{z}|\mathbf{w}, \alpha, \beta) = \frac{p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta)}{p(\mathbf{w}|\alpha, \beta)}$$

However this distribution is intractable to compute.[3] Thus we have to resort to approximate inference algorithms. We will present two methods, the variational inference method, which is presented in the original paper[1], and the Gibbs sampling method[2],[3].

2.1 Variational Method

The basic idea of variational methods is to make use of Jensen?s inequality to obtain an adjustable lower bound on the log likelihood. A simple way to do this is to remove some of the vertices and edges of the original graph. In figure 3, we drop the vertex $\mathbf{w}$ and the edges between $\theta, \mathbf{z}, \mathbf{w}$. We thus obtain a family of distribution on the latent variables, which is characterized by the following variational distribution

$$q(\theta, \mathbf{z}|\gamma, \phi) = q(\theta|\gamma)\prod_{i=1}^{N} q(z_i|\phi_i)$$

Next step would be formulating a optimization problem that determines the value of γ and ϕ. The problem is given as follow

$$(\gamma^*, \phi^*) = \arg\min_{(\gamma,\phi)} D(q(\theta, \mathbf{z}|\gamma, \phi)||p(\theta, \mathbf{z}|\mathbf{w}, \alpha, \beta))$$

where D is the KL divergence between q and the true posterior p. The minimization can be done via an iterative fixed-point method and the update is

$$\phi_{ij} \propto \beta_{jw_i} \exp\{E_q(\log(\theta_i)|\gamma)\}$$

$$\gamma_j = \alpha_j + \sum_{i=1}^{N} \phi_{ij}$$

The conditional expectation $E_q(\log(\theta_i)|\gamma)$ can be computed via Taylor approximation. The details of the derivation can be found in[1]. It is worth noting that the variational distribution is a conditional distribution that is a function of $\mathbf{w}$, and thus the optimizing parameters (γ^*, ϕ^*) are also functions of $\mathbf{w}$.

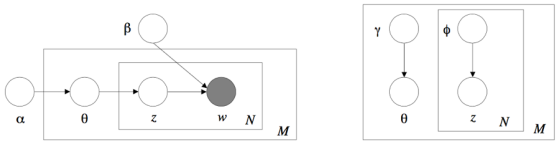


Fig. 3. LDA and its variational representation

2.2 Gibbs Sampling

If we use only symmetric Dirichlet priors for both the topic distribution θ and word distribution ϕ, that is, the Dirichlet parameters α and β are uniform, a Markov Chain Monte Carlo (MCMC) method can be used for inference. An advantage of MCMC is that we do not need to explicitly represent the model parameters[4].

We thus use Gibbs sampling, a MCMC method, to perform inference. In Gibbs sampling, the next state is reached by sequentially sampling all variables from their distribution when conditioned on the current values of all other variables and the data.

The conditional posterior distribution of z_i is given by

$$p(z_i = j|\mathbf{z_{-i}}, \mathbf{w}) \propto P(w_i|z_i = j, \mathbf{z_{-i}}, \mathbf{w})P(z_i = j|\mathbf{z_{-i}}) \quad (1)$$

where $\mathbf{z_{-i}}$ is the assignment of all the z_k where $k \neq i$. Recall that we use the following notation

$$P(z_i = j) = \theta_j$$

$$P(w_i|z_i = j) = \phi^j_{w_i}$$

The first term of (1) can be written as

$$P(w_i|z_i = j, \mathbf{z_{-i}}, \mathbf{w}) = \int P(w_i|z_i = j, \phi^j)P(\phi^j|\mathbf{z_{-i}}, \mathbf{w})d\phi^j \quad (2)$$

The rightmost term of (2) can be further written as

$$P(\phi^j|\mathbf{z_{-i}}, \mathbf{w}) \propto P(\mathbf{w}|\mathbf{z_{-i}}, \phi^j)P(\phi^j) \quad (3)$$

Since $P(\phi^j)$ is Dirichlet(β) and it is conjugate to multinomial distribution $P(\mathbf{w}|\mathbf{z_{-i}}, \phi^j)$, $P(\phi^j|\mathbf{z_{-i}}, \mathbf{w})$ will be Dirichlet($\beta + n^w_{-i,j}$), where $n^w_{-i,j}$ is the number of instances of word w assigned to topic j, not including the current word. The first term in (2) is simply $\phi^j_{w_i}$. Thus the integral (2) yields

$$P(w_i|z_i = j, \mathbf{z_{-i}}, \mathbf{w}) = \frac{n^w_{-i,j} + \beta}{n^{(.)}_{-i,j} + N\beta} \quad (4)$$

where $n^{(.)}_{-i,j}$ denotes the total number of words assigned to topic j, not including the current one and N is the total number of words. Do the same procedure for the second term in (1) we get the following conditional probability

$$p(z_i = j|\mathbf{z_{-i}}, \mathbf{w}) \propto \frac{n^w_{-i,j} + \beta}{n^{(.)}_{-i,j} + N\beta} \frac{n^{d_i} + \alpha}{n^{d_i}_{-i} + T\alpha} \quad (5)$$

With (5) Gibbs sampling is straightforward. Initialize z_i to a value between 1 and T, then run a certain number of iterations where at each iteration z_i is drawn from the distribution specified by (5). After enough number of iterations, the Markov Chain will converge to the true distribution. The implementation of LDA we are going to use for the later application is PLDA+[5], which is also based on this Gibbs sampling method.

3. Application to Temporal Decomposition of Mobility Data

3.1 Overview

Mobility data has grown tremendously for the past decades and is becoming quite ubiquitous. A good understanding of the structure of the mobility flow is essential and crucial in transportation planning, vehicle fleet management, and related policy making.

A very common and basic type of mobility data is the OD pair which encodes the origin (O) and the destination (D) of a trip. OD tables have been studied in the field of transportation engineering for decades and and play an important role in traffic flow prediction, transportation simulation and planning. Trip purpose is one of the important traits of the mobility data, it tells about the structure of the trip flow in the system and categorize the passengers into different groups such as commuters, leisure traveller or tourists. Understanding the trip purpose is important and rewarding in optimizing the transportation system and proving reliable mobility services. However, it is not straightforward to infer trip purpose from the OD tables and most of the trip purpose can only be obtained by expensive and time-consuming surveys.

While directly obtaining trip purposes is hard, machine learning methods from text modeling provide us with a novel way to tackle this problem. It has been shown[4] that by treating each OD pair as a "word", the text modeling methods can be used to extract latent "topics" from the collection of "words" and the "topics" can be viewed as different trip purposes. In the rest of this report, we will implement LDA to extract trip purposes from a large amount of OD pairs in the Bay Area Rapid Transit (BART) system over a 3-week period and perform temporal decomposition of the data with regard to different trip purposes.

The procedure of implementing LDA to perform temporal decomposition of the mobility data is described in the following figure

3.2 The Data

The data we are going to use is the CLIPPER card transaction information during a three-week period in September 2014. There are over 2 million transaction records affiliated with BART. Each transaction entry represents a trip and records both the tag-on time and the tag-off time as well as the tag-on location and tag-off location. There are 44 stations in the system and each trip is from a station i to a station j. For privacy issues, we are only given randomized ids for each trip and two ids are identical only if the same card holder takes these two trips within one day.

3.3 Data Exploration

Before performing LDA, it will be interesting to explore the data and see what information we can extract from the data. We first plot the total counts of trip during a one-week period in figure. It is very obvious that during the weekday, there are two sharp peaks in each day and on weekend, there is only one small hump. There seems to be a certain underlying pattern of the trips and it differs from weekdays to weekends.

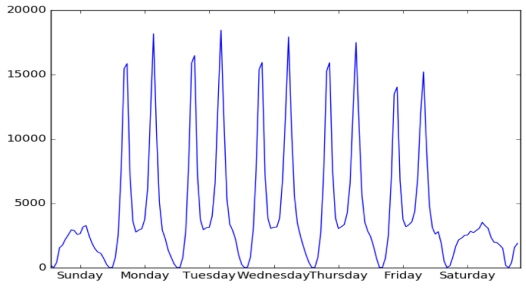


Fig. 4. Total counts of trip during one week

We also plot the counts of trips that start from or end certain stations over a one-week period. As shown in figure, certain stations (e.g. Montgomery St.) show only one sharp peak for tag-off counts in the morning and only one sharp peak for tag-on counts in the evening during the weekdays. In contrast, certain stations (e.g. Downtown Berkeley), show two peaks both for tag-off and tag-on counts during the weekdays. Stations like Montgomery St. with one peak for both tag-on and tag-off counts may be associated with work locations and stations like Downtown Berkeley may be associated with both work and home locations.

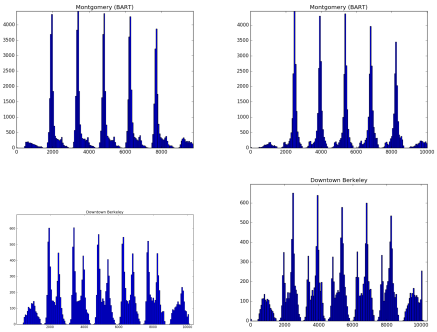


Fig. 5. Top: counts of tag-off (left) and tag-on (right) at Montgomery St. station. Bottom: counts of tag-off (left) and tag-on (right) at Downtown Berkeley station

3.4 Document Composition

In order to implement LDA to perform temporal decomposition of the mobility data, we have to preprocess the OD pairs so that they become "readable" words. A very straightforward way to do that is to make a trip from i to j simply a word "$i\_$ to $\_j$", which makes it a unique readable word.

Based on the intuition that the same trips can have different purposes at different time, we would like to divide the whole corpus into different documents based on time, where in the same document, we assume that all the trips from the same origin and to the same destination would have the same conditional probabilities given a topic z. To choose such an appropriate time interval we have to make a good balance between the number of samples within the interval and the range of the interval. That is, it could be neither too long nor too short. A good practical choice of such interval would be one hour.

So the entire corpus is composed as follows: first we turn every trip into readable word, then for each hour we make all the trips that happens within that hour into a documents, and we combine all the documents together to form a corpus.

3.5 Decomposition Results

The LDA implementation used in this report is a parallel implementation of LDA[5]. It is mainly based on the Gibbs sampling method described in section 3.4. The decomposition result for a one-week period is displayed in figure 4, where we choose the number of topics $T = 4$, and $\alpha = 10, \beta = 0.01$. Several interesting patterns can be spotted from the plot. During the weekdays, there are two peaks, one in the morning and the other in the evening, each of which is represented by a single topic. These two topics, however, disappear on weekends. The other two peaks, though much weaker than the first two, appears every day throughout the week and do not vary too much. A closer look at the decomposition within one day is shown in figure 5 and 6, where figure 5 is for a typical weekday (Monday) and figure 6 is for a typical weekend (Sunday).

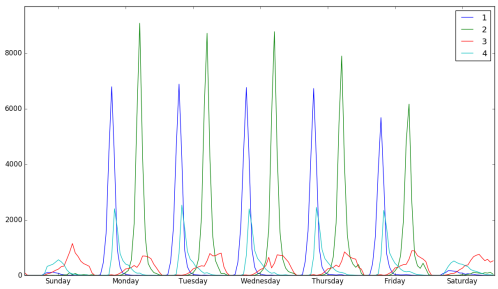


Fig. 6. Temporal decomposition of the mobility data into 4 topics during a 1-week period. With parameter $\alpha = 10$, $\beta = 0.01$

It can be observed that the first peak during the weekdays starts at 6:00 am and ends at 8:00 while the second the peak during the weekdays starts at 4:00pm and ends at 6:00pm. It seems very natural to interpret these weekday peaks as morning commute and evening commute. On weekends, there are also two different peaks but the first peak starts at around 8:00am and ends at around 2:00pm while the second peak starts at 2:00pm and does not end until 9:00pm. These two peaks behave differently than the weekday peaks because they start later and last longer. An interpretation of morning and evening leisure activities might be appropriate for these two topics since we do not have further knowledge of the data.

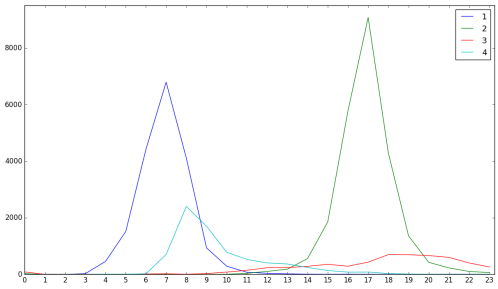


Fig. 7. A typical weekday pattern (Monday)

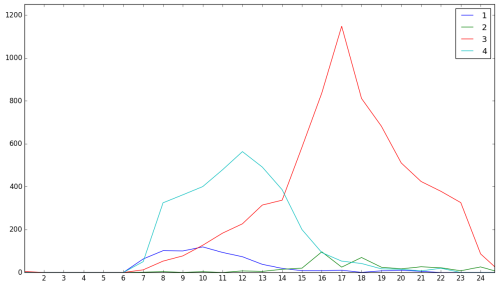


Fig. 8. A typical weekend pattern (Sunday)

3.6 Some Analysis

It should be noted that the pre-process of turning the OD pairs into documents implicitly hides the time information because all we know about the documents is that one document contains trips within one hour, but we have no idea of the temporal relationship between any two documents. However, LDA managed to detect consistent patterns throughout the week even if we do not know the temporal relationship between different documents.

Another thing we can explore is the influence of the parameters α and β. Recall that α is the parameter of the Dirichlet prior on the probabilities of the latent topics being chosen and β is the parameter of the Dirichlet prior on the probabilities of the words being chosen given a latent topic. Under the assumption that the Dirichlet priors are symmetric, a higher Dirichlet parameter would cause the distribution to be more concentrated in the center on the simplex while a lower Dirichlet parameter would cause the distribution to gather at the vertices of the simplex, see figure 7. An intuitive

interpretation of a high α would be that we want the document to have more similar topics while a low α indicates that a document tends to choose less topics. The same logic applies to β. Several decomposition results are shown in figure 8 with different α with a fixed $\beta = 0.01$ and a fixed number of topics of 4. We can observe that as α decreases, the morning commute peak and the morning leisure peak become indistinguishable during the weekdays. An good empirical choice of α would be $T/50$.

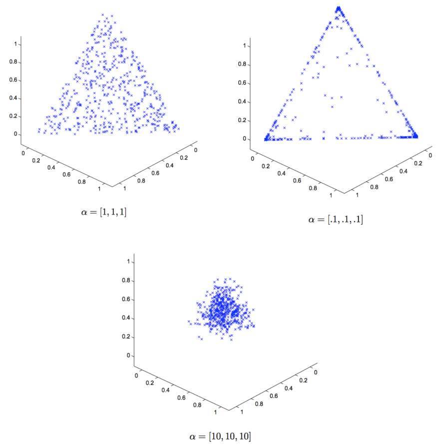
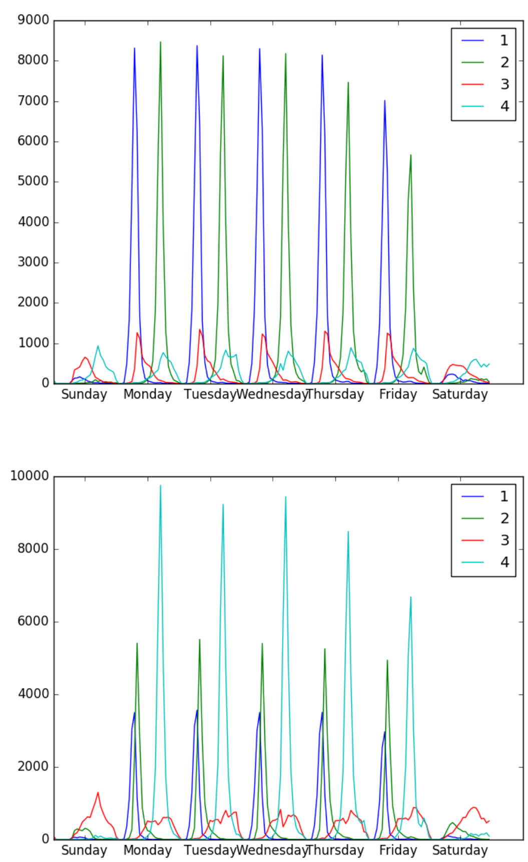


Fig. 9. Sampling from distribution of θ on the 2D simplex for different choices of α



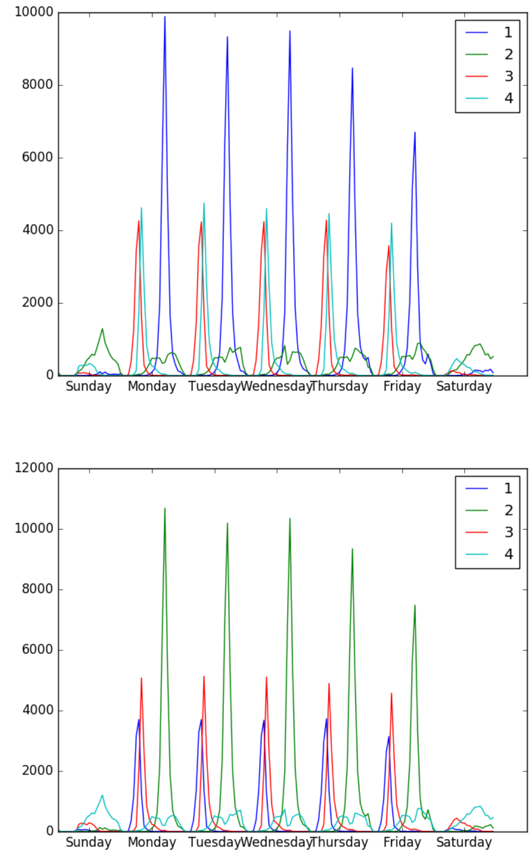


Fig. 10. Different decomposition results with $\beta = 0.01, T = 4$. From the top to the bottom: $\alpha = 10, \alpha = 1$, $\alpha = 0.1$, $\alpha = 0.01$

3.7 Spatial Visualization

So far we have seen that LDA has done a good job in performing temporal decomposition of the mobility data. However, there is one important thing we did not take into consideration, the spatial property of the trips. Because we translated every trip into a word, namely a string, we lost all the spatial information related to the trips. Now we want to recover the trips from the decomposition results and visualize it spatially under different topics to see whether the decomposition is reasonable.

Because of the flows in the system can go either way, we will use two directed graph that are complement to each other to represent all the flows in the system. The BART transit map is shown as follow and we will make the Embarcadero station in downtown San Francisco as a benchmark. We define all the flows that head towards it would be consider in-bound flows and all the flows that head away from it would be consider out-bound flow. The visualization for the morning peak of the ”morning commute” topic and the evening peak of the ”evening commute” topic is shown in figure. An interesting observation is that the ”morning commute” and the ”evening commute” are both one-way flows – that is, the morning commute is mostly made up of in-bound flows and the evening commute is mostly made up of

out-bound flows. This indicates that commuter tend to go towards downtown San Francisco to work in the morning and leave the city in the evening. Although we do not have further information about the trip purposes from the data, LDA somehow provides reasonable results.

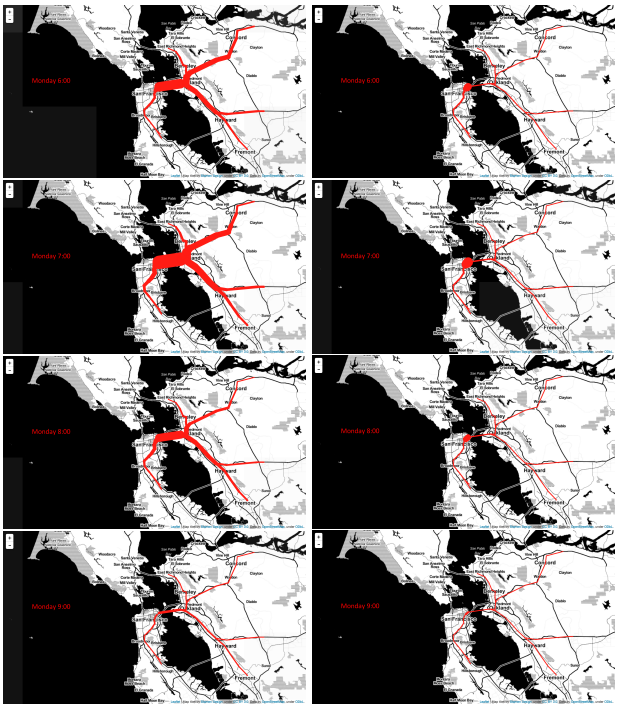


Fig. 11. "Morning commute" topic from 6am to 9am. Left: inbound flow, Right: outbound flow

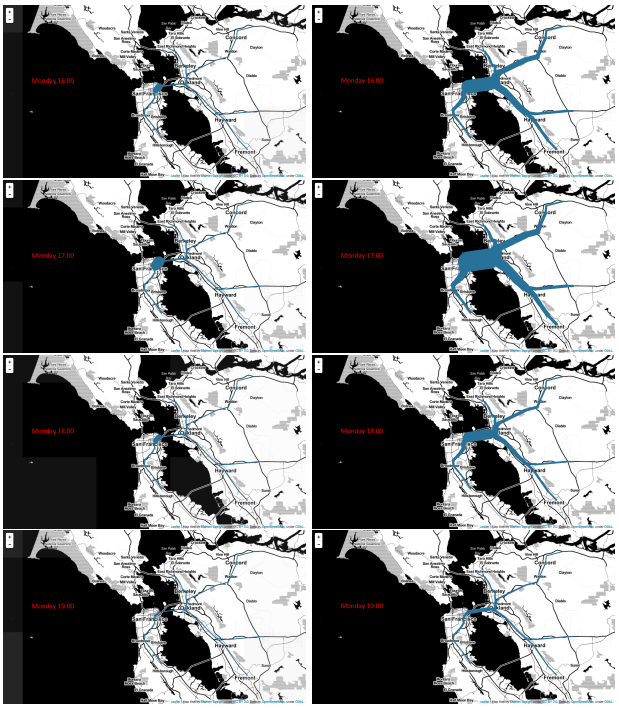


Fig. 12. "Evening commute" topic from 4pm to 7pm. Left: inbound flow, Right: outbound flow

3.8 Validation with Survey Results

In order to validate the decomposition result obtained from LDA, we use the survey results conducted by BART in 2015. Details about the survey can be found in[6]. However, the survey results did not specify any temporal attributes of the trip purposes, we therefore compare the ratios of total home-work and work-home OD pairs obtained by LDA to the ratios provided by BART. We only calculate the ration of trips that take place during the weekday because the survey was performed only during weekdays. The first column is the ration obtained by survey, and the second to the fifth column is the ration obtained by LDA with $\alpha = 10, 1, 0.1, 0.01$ respectively. It is clear that the choice of $\alpha = 10$ provides the closest estimate of Home-Work trips and Work-Home trips to the results from the survey.

	BART	$\alpha = 10$	$\alpha = 1$	$\alpha = 0.1$	$\alpha = 0.01$
Home-Work	0.389	0.321	0.191	0.212	0.197
Work-Home	0.345	0.331	0.359	0.361	0.383

Table 1. LDA results compared to the survey results from BART

However, it should be noted that the ratio did not take the temporal attributes into consideration as we pointed out before, and the survey was conducted in 2015, which may not be ideal for validation since our data was acquired in 2014.

3.9 Conclusion

In this report, the application of LDA to temporal decomposition is performed. This is a novel way to infer trip purpose solely based on the trip location and time. We compare the decomposition results from the choices of different parameter α. We then use a survey result conducted by BART in 2015 to validate our result. LDA provides close estimation of Home-Work and Work-Home trips when we choose $\alpha = 10$. As mobility data is becoming quite ubiquitous, an cheap and efficient inference of the trip purposes can benefit the entire transportation system and enable much smarter mobility services.

References

[1] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." Journal of machine Learning research 3.Jan (2003): 993-1022.

[2] Griffiths, Thomas L., and Mark Steyvers. "Finding scientific topics." Proceedings of the National academy of Sciences 101.suppl 1 (2004): 5228-5235.

[3] Griffiths, Tom. "Gibbs sampling in the generative model of latent dirichlet allocation." (2002).

[4] Coffey, Cathal, and Alexei Pozdnoukhov. "Temporal decomposition and semantic enrichment of mobility flows."

Proceedings of the 6th ACM SIGSPATIAL International Workshop on Location-Based Social Networks. ACM, 2013.

[5] Liu, Zhiyuan, et al. "Plda+: Parallel latent dirichlet allocation with data placement and pipeline processing." ACM Transactions on Intelligent Systems and Technology (TIST) 2.3 (2011): 26.

[6] BART Station Profile Study, 2015, http://www.bart.gov/about/reports/profile