FLDA: Latent Dirichlet Allocation Based Unsteady Flow Analysis

Fan Hong, Chufan Lai, Hanqi Guo, *Student Member, IEEE*, Enya Shen, Xiaoru Yuan, *Member, IEEE*, Sikun Li

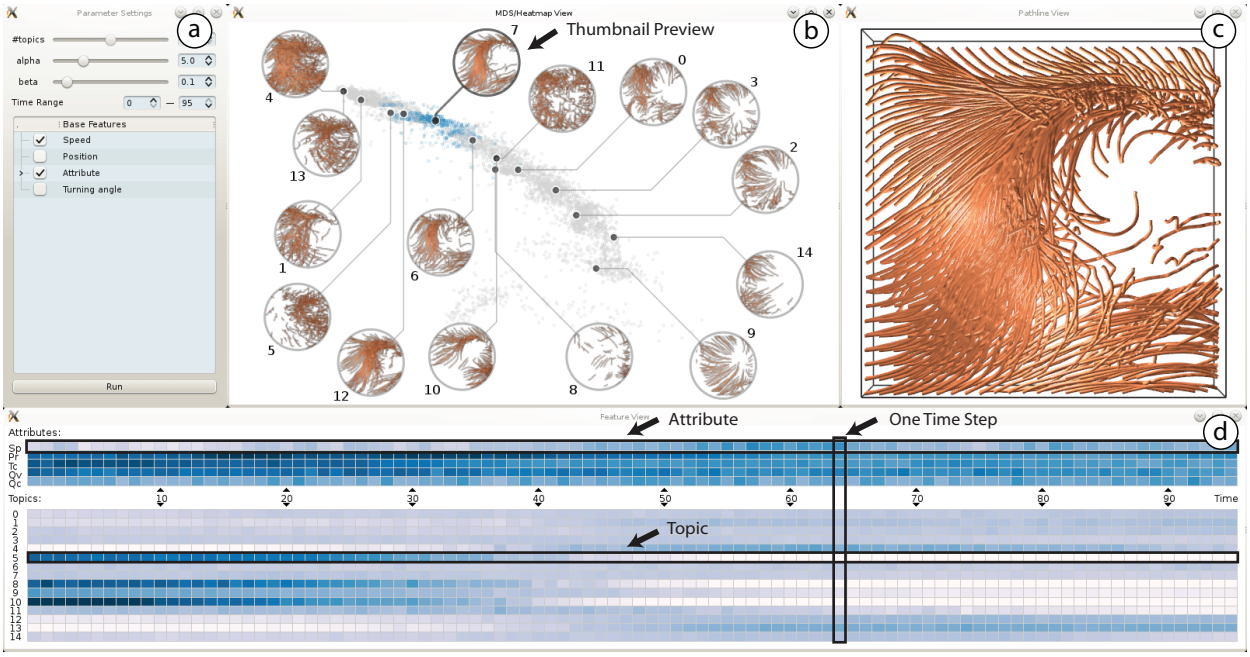


Fig. 1. The user interface of our proposed FLDA system, including: (a) parameter setting panel for adjusting parameters in LDA model; (b) MDS/Heatmap view with thumbnails of pathline previews embedded; (c) pathline view for rendering pathlines in 3D space; (d) feature view. Projection in (b) is rendered as a density map in pseudo color, with distribution of the selected topic highlighted. (d) in the pixel-oriented style consists of two parts: the attribute-time view at the top and the topic-time view at the bottom. Each column represents one time step, while each row denotes a topic or an attribute.

Abstract—In this paper, we present a novel feature extraction approach called FLDA for unsteady flow fields based on Latent Dirichlet allocation (LDA) model. Analogous to topic modeling in text analysis, in our approach, pathlines and features in a given flow field are defined as documents and words respectively. Flow topics are then extracted based on Latent Dirichlet allocation. Different from other feature extraction methods, our approach clusters pathlines with probabilistic assignment , and aggregates features to meaningful topics at the same time. We build a prototype system to support exploration of unsteady flow field with our proposed LDA-based method. Interactive techniques are also developed to explore the extracted topics and to gain insight from the data. We conduct case studies to demonstrate the effectiveness of our proposed approach.

Index Terms—Flow visualization, Topic model, Latent Dirichlet allocation (LDA)

1 Introduction

In recent years, there is an increasing demand on effective visualization of multivariate unsteady flow field data, especially in areas such as climate research, ocean research, air pollution research, etc. Visualization and analysis for such data is quite difficult and challenging, due to the complexity of the data and visualization tasks. In general, such flow visualization involves exploring attribute space, geometry space and advection simultaneously.

Flow field data has been long studied in visualization community for years. Traditional flow visualizations including texture-based [30] and geometry-based [36] methods seldom consider the associated attribute information. They may also suffer from visual clutter problem when dealing with 3D flow field.

Now more visualization methods are developed to detect, identify and extract interesting features from the flow field. One genre is to manually define features at first, after which the system will try to detect corresponding features from flow field [28, 46, 39, 8]. Salzbrunn et al. [39] has shown that any suitable set of pathline predicates can be interpreted as features in unsteady flow structures. However, insightful features are hard to define when scientists do not have enough prior knowledge about the data, especially when multivariate information involved. Another genre takes clusters as features. The flow field data is first transformed to a space easier for exploration, then automatic algorithm or manual exploration is conducted to identify and extract

- *Fan Hong and Chufan Lai are with Key Laboratory of Machine Perception (Ministry of Education), School of EECS, Peking University. E-mail: {fan.hong,chufan.lai}@pku.edu.cn.*
- *Hanqi Guo and Xiaoru Yuan are with Key Laboratory of Machine Perception (Ministry of Education), School of EECS, and Center for Computational Science and Engineering, Peking University, E-mail: {hanqi.guo,xiaoru.yuan}@pku.edu.cn.*
- *Enya Shen is with School of Computer Science, National University of Defense Technology, Changsha, China, E-mail: shenenya.nudt@gmail.com.*
- *Sikun Li is with School of Computer Science, National University of Defense Technology, Changsha, China, E-mail: sikunli@126.com.*

Manuscript received 31 Mar. 2014; accepted 1 Aug. 2014. Date of publication 11 Aug. 2014; date of current version 9 Nov. 2014.
For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org.
Digital Object Identifier 10.1109/TVCG.2014.2346416

clusters. Projection methods are more effective in revealing the intrinsic features of flow field data by mapping data to lower dimensions. Daniels et al. [22] proposed to project samples of 2D flow into 2D space to extract important spatial structures. Streamlines can also be projected according to their geometry distances [38]. When it comes to multivariate unsteady flow field, Lagrangian-based Attribute Space Projection (LASP) proposed by Guo et al. [17] is capable of projecting massive pathlines based on their distances in the attribute space. However, these methods often suffer from the gap between the data space and feature space. The identified clusters lack semantic meaning to help understand these intrinsic structures.

In this paper, we propose a novel unsteady flow analysis method based on Latent Dirichlet allocation (LDA) model, named the *Flow LDA*, or FLDA in short. LDA model is widely used in text analysis to study topics from a large corpus of documents which are naturally decomposed into words. In FLDA, we define *pathlines* as documents and *features* as words respectively. After estimating the underlying model, we obtain a list of topics which are fused from features defined, and simultaneously pathlines are clustered into topics with probabilistic assignment. The *topics* can serve as connectors between the collection of pathlines and the features.

FLDA can be regarded as a mixture of the two types of feature extraction methods mentioned before. On the one hand, features are defined at first to describe expected facets of the flow behaviors. FLDA model will identify features containing those facets, but in a way of fused topics, rather than as separated features. On the other hand, FLDA can be treated as a probabilistic clustering method. We are able to obtain semantic implications of clusters from LDA-derived topics, which reflect intrinsic patterns hidden in the flow field data. Moreover, the probabilistic assignment gives a consecutive measurement of the pathline-topic relationship, which could provide more insights into clusters.

Although LDA model has been successfully applied to many research fields besides text analysis, including computer graphics, computer vision [42, 9, 31] and even recently traffic trajectory analysis [12], it has not been systematically tested for flow field visualization. It is nontrivial to apply LDA model to flow field data. The key problem is how to define features in FLDA, which is corresponding to words in LDA model, and how to explain topics and their relationship with pathlines and features. In the original LDA model for text analysis, words are naturally obtained from documents. When applied in computer vision or computer graphics for segmentation, classification or pattern recognition, spatial regions are usually used as words. However, in flow field analysis, it is not sufficient to consider only spatial positions. Especially for multivariate unsteady flow field, we expect to involve multivariate features. Therefore, we need to carefully define *words* in our FLDA model, in order to obtain a good description of the expected facets. We have tackled the feature definition problem and apply the FLDA model for complexity flow visualization, as a new approach to understand and explore complex flow fields. We also developed a set of interactive techniques for better understanding the extracted topics and their relationship with features and pathlines.

The remainder of this paper is organized as follows. We describe the background of FLDA model and flow field visualization and analysis techniques in Section 2. In Section 3, we introduce our method on the basis of LDA model, and describe the pipeline of our prototype system. We give more details about the parameter settings of FLDA model, and visualization techniques in Section 4. Case studies are shown in Section 5 to demonstrate the effectiveness of our method. In Section 6, we compare the proposed FLDA method with other approaches. At last, we conclude this paper with a brief review and discuss about the future work.

2 Background

Our work has mainly three relevant fields: multivariate flow field visualization, LDA analysis methods, and flow exploration methods. In this section, we briefly review literature in these fields.

2.1 Multivariate Flow Field Visualization

Multivariate flow field visualization is a challenging task due to the complexity of the data. Using linked multi-dimensional visualizations for feature extraction and rendering is a major methodology to tackle this problem.

One of the most popular multi-dimensional data visualization tools is Parallel Coordinate Plot (PCP) [23]. PCP-based flow field visualizations had been proposed in recent years [2, 4, 47]. Akiba et al. [1, 2] proposed a tri-space visualization combining time histogram, PCP and the spatial rendering for identifying features in temporal, attribute and spatial domain. Similarly, Blass et al. [4] used PCP for multivariate analysis in a time-varying background, where features are spatially rendered as isosurfaces. For multivariate particle data, Jones et al. [27] deployed PCP to select particles for pathline generation.

Scatterplots coupled with dimension reduction techniques or derived statistic information are also widely used. Chen et al. [11] proposed a method to embed DTI fibers into 2D Multi-Dimensional Scaling (MDS) based on their mutual mean distances, so as to alleviate the clutter problems in 3D space. Guo et al. [18] proposed a seamless integration between PCP and MDS plots, which provides high efficient feature extraction by avoiding context switching. Such approach is further developed by Zhao et al. [47], combining Locally Linear Embedding (LLE) with PCP for an easier edit on transfer functions. Helmut proposed the SimVis system [15] that makes use of scatterplots with linked brushes to select interesting features in particle simulations. Jañicke et al. [25] transformed the attribute space into 2D point clouds where special features could be distinguished without clutter or obscurity. More recently, Maciejewski et al. [35] encoded attribute relationships instead of projections in the scatterplot for better guidance in 2D transfer function design. For vector field of flow data, dimension projection techniques give more insight by embedding field lines into lower dimension spaces. For streamline embedding, Hausdorff distances are used as the distance metric [38] between seed points so that the projection plot reveals spatial similarity of corresponding streamlines. To further consider multivariate behavior of unsteady flow field, LASP [17] extended the geometry space distance to attribute space distance for traced pathlines. That method is capable of extracting multivariate features in the Langragian perspective for unsteady flow data. Facing a similar data complexity problem, we adopt the MDS to exhibit the relationship of pathlines in the attribute space, which provides a carrier to visualize flow topics from the multivariate facet. For a better performance, we use the Pivot MDS [6] to reduce the computational complexity.

Besides PCP and scatterplot, other approaches are also used to show the relationships among variables. Sauber et al. [40] proposed multifield-graphs, in which variables are hierarchically grouped for a descendant correlation display in the spatial domain. Woodring and Shen [44] used a spreadsheet and a tree map for showing the comparison relationship between attributes. Concerning data clustering in the attribute space, Linsen et al. [33] also used the tree structure to present the hierarchical structure, aligned with cluster contours shown in the star coordinate context. From the information-theoretic aspect, Wang et al. [41] adopted circular graph layout to present information transfer between variables. Chen et al. [10] showed the static correlations between samples extracted from the volume using scatterplot matrix. Bruckner and Möller [7] split the simulation data into clusters, with a star glyph to present the multivariate feature of each cluster.

2.2 Topic Model and LDA Analysis

Topic model is widely used in text analysis. Landauer et al. [29] proposed the concept of Latent Semantic Analysis in 1988. Latent Semantic Analysis add a latent semantic layer between documents and words. Latent semantics are extracted from the relationship among words to construct semantic space, where documents are then projected to obtain a sparse representation. pLSI/pLSA [19, 20] introduced statistic analysis and generative model based an LSA. pLSA solves synonyms and polysemy problem, but suffers from overfitting. Blei et al. [5] proposed the concept of topic model and related LDA model. LDA is a multi-layer Beyasian model, including three layers, e.g. words, top-

ics, and documents. Every topic is a mixture of words, while every document is a mixture of topics. By introducing Dirichlet distribution, LDA model is able to avoid overfitting which pLSA suffers. Among these two most popular topic models, pLSA is actually a special case of LDA. After LDA model, lots of variations raise.

In text analysis, lots of visualization are proposed for the results derived by LDA model. Termite [13] utilized a 2D table to represent the distribution between keywords and topics, using size of circles to indicate the probabilities. Documents can be projected to lower dimensional space by considering not only the distances between words, but also the distances between latent semantic topics [21, 24]. To visualize the evolution of topics along time, TIARA [43] encoded the hotness of topics using the width of rivers in ThemeRiver. TextFlow [14] further used the metaphor of rivers to indicate the emerging, vanishing, merging and splitting events in topic models. LeadLine is also a river-like visualization, but more emphasized on the bursting to topics hotness. iVisClustering [32] not only provides various visualization techniques for LDA model, but also enables users steering of LDA process. Besides applications in text analysis field, LDA model has been adopted in computer graphics and computer vision field [42, 9, 31] for various purposes, such as segmentation, classification, pattern recognition, etc. In visualization of traffic data, Chu et al. [12] use LDA model to discover hidden themes from trajectories data.

In this paper, we build a topic model for flow field data to describe the relationship between pathlines and features by introducing a latent layer. Our prototype system provides several visualizations for LDA results. We focus more on revealing the insight of topics, since topics in our method have richer meaning.

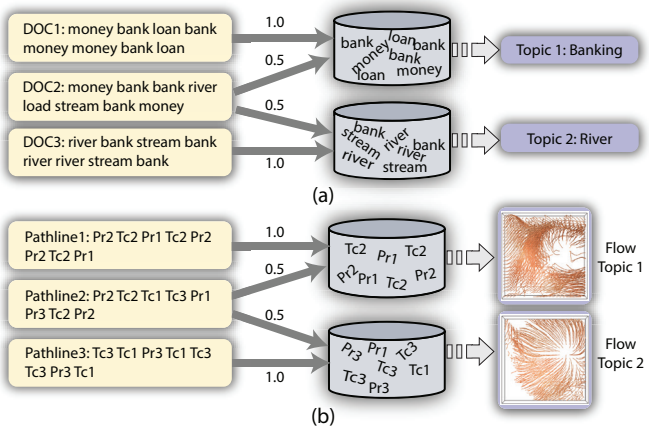


Fig. 2. Illustration of (a) typical LDA model and (b) our flow LDA model. Pathlines and features are defined to be equivalent to documents and words in topic model respectively. Tc and Pr stands for features of temperature and pressure respectively.

2.3 Exploration on Flow Field Data

Traditional flow visualization methods, such as texture-based [30] and geometry-based [36] methods, usually involve little user interaction. More recently, interactive feature extraction techniques have been developed to help users explore flow field. Parallel coordinates are often used to visualize the scalar field in flow data, where users are able to brush numerical ranges to extract features [1, 2, 4]. Alternative to parallel coordinates, projection approaches are able to reveal inherent structures from data. 2D scalar samples are projected into 2D space to extract important spatial structures. Streamlines and pathlines can also be embeded into lower-dimensional space advection information based on the distances in geometry space [38] and attribute space [17]. Users are able to extract rich features from projection plots.

Recently, graph-based visualization is emerging for flow exploration. TransGraph [16] is proposed to visualize information transition along time between blocks, where users can select different levels of hierarchy. Ma et al. [34] proposed FlowGraph to explore the dual relationships between streamlines and blocks with rich interaction and query techniques. Jänicke et al. [26] extract local flow patterns as nodes in graph, and their transitions as edges where users can track features over time. There are some work visualizing the attribute relationship of scalar field using graph-like form [40, 3].

In our work, based on a novel perspective, FLDA, we provide a pixel-wised feature view, together with preview-facilitated MDS/Heatmap view to help explore flow field data.

3 Overview

In this work, we innovate a flow field analysis method which is based on LDA model, and implement a prototype system to support LDA-based exploration on flow field. In this section, we will first introduce basics of LDA model and its usage in the topic analysis field. Then we give our equivalent definition of LDA concepts in the flow field scenario. At last, we describe the pipeline of our prototype system.

3.1 Basics of LDA Model

Latent Dirichlet allocation (LDA) was first proposed by Blei et al. [5] to explain why documents are similar from the latent topic level instead of the word level. We first introduce some basics of typical LDA model. The symbols used in this paper are listed in Table 1.

Typical LDA	D	Number of documents.
	K	Number of topics.
	$\mathcal{V}$	Vocabulary.
	d_j	The j^{th} document.
	N_j	Number of words in d_j.
	N	Total number of words in all documents, $\sum_{j=1}^{D} N_j$.
	w_{ij}	The i^{th} word in d_j document.
	z_{ij}	Topic assignment for word w_{ij}.
	θ_j	Probability of topics in document d_j.
	ϕ_k	Probability of words in topic k.
	α	Dirichlet prior for θ.
	β	Dirichlet prior for ϕ.
Flow LDA	x	Facet of flow field, $x \in \{\text{Speed, Attribute, Angle}, \dots\}$.
	$\mathcal{F}_x$	Flow feature set defined on facet x.
	$\mathcal{V}_f$	Vocabulary of flow features, $\bigcup_x \mathcal{F}_x$.

Table 1. Symbols used in this paper.

The LDA model is typically used to analyze topics in the corpus of documents. The underlying generative process is that any document d_j is modeled as a mixture of K topics, while any topic k is characterized by a multinomial distribution ϕ_k over vocabulary $\mathcal{V}$. Among all variables, only w_{ij} is observable, while others like z_{ij}, θ_j, and ϕ_k are latent variables. LDA model generates observations of latent variables using the following process:

1. For every document d_j, draw a topic distribution θ_j from a Dirichlet prior with parameter α, i.e. $\theta_j \sim Dir(\alpha)$, where $j \in \{1, \cdots, D\}$.
2. For every topic k, draw a word distribution ϕ_k from a Dirichlet prior with parameter β, i.e. $\phi_k \sim Dir(\beta)$, where $k \in \{1, \cdots, K\}$.
3. For word position i in j^{th} document, where $i \in \{1, \cdots, N_j\}$, and $j \in \{1, \cdots, D\}$, choose a topic for this position $z_{ij} = k \sim Multinomial(\theta_j)$, and then choose a word from the chosen topic $w_{ij} \sim Multinomial(\phi_{z_{ij}})$.

After the generative process is defined, the total probability of the model can be described as:

$$P(\boldsymbol{W}, \boldsymbol{Z}, \boldsymbol{\theta}, \boldsymbol{\phi}; \alpha, \beta) = \prod_{i=1}^{K} P(\phi_i; \beta) \prod_{j=1}^{M} P(\theta_j; \alpha) \prod_{t=1}^{N} P(Z_{j,t}|\theta j) P(W_{j,t}|\phi_{Z_{j,t}}),$$

where $\boldsymbol{W}, \boldsymbol{Z}, \boldsymbol{\theta}, \boldsymbol{\phi}$ denotes the vector version of w_{ij}, z_{ij}, θ_j, and ϕ_k respectively. The model estimation process is to maximize the likelihood function by Bayesian inference with parameters α and β. The original model [5] uses variational inference method, while in our work, we adopt an implementation using Gibbs sampling [37]. The time complexity of one iteration is $O(KDM)$, where M is the average length of documents. It can be simplified to $O(KN)$, where N donates the total number of words in all documents.

The LDA model can be applied as a document clustering method. The K topics could be treated as clusters, and the topic distribution θ_j for the document d_j denotes the probabilities of membership to every cluster. A more careful approach is to treat θ_j as a lower-dimensional feature vector for every document, and conduct another clustering routine, such as k-means. However, in FLDA model, we choose the former approach, because we think the topics and distributions act as better connectors between the data space and feature space.

3.2 FLDA Model

To employ LDA model in flow field analysis, we need to define equivalent LDA concepts at the very beginning (Figure 2). Without loss of generality, we give our definition on pathlines, which could be easily extended to other kinds of fieldlines. In our FLDA model, we consider *pathlines* as the central subjects, which play a similar role as documents do in the topic model. We then consider *features* as words, with pathlines being *bags of features*, which is analogous to the concept of documents being *bags of words* in topic model. For any facet x of the flow field, such as spatial, geometric, multivariate or temporal facet, we can define a feature set $\mathcal{F}_x$ to describe different behaviors of the selected facet. For example, we can define features as blocks in spatial domain to represent similar spatial behaviors, or as bins of the attribute value to represent similar attribute behaviors, etc. In this way, any pathlines sharing similar behaviors have common features. All feature sets $\mathcal{F}_x$ can then be united to an overall feature vocabulary $\mathcal{V}_f$. In our prototype system, we provide a predefined vocabulary containing feature sets of commonly concerned facets. Users are free to choose a subset of predefined vocabulary or add their own feature sets to the existing vocabulary.

After defining the feature vocabulary $\mathcal{V}_f$, we pick out features for every pathline to make its *feature bag*. The FLDA method uses the feature bags as input to estimate the underlying topic model. As results, topics, distribution of topics per pathline, and distribution of features per topic are generated. Similar to the LDA model, every derived topic can be treated as a mixture of features. Meanwhile, pathlines are assigned to topics with probabilities.

The generated topics are depicted from two aspects, the feature side and the pathline side. On the one hand, in order to provide good semantic explanations for topics, it is important to choose appropriate definitions of features. Easy as it seems in static data, the problem is tricky in the time-varying background. In our prototype system, we prefer to define features as similar behaviors at a single timestep, rather than over consecutive timesteps or the entire time span. In this way, it is easier to decompose a topic along the time dimension to inspect the evolution of corresponding behaviors. On the other hand, topics can be treated as fuzzy clusters constructed by pathlines based on the behaviors they shared. Pathlines convey rich multivariate information, geometric information and temporal information, which makes it possible to extract topics from some facets and then explain them from the others. While in text analysis, the documents only contain textual information, thus the topics can only be extracted and explained from the same aspect. From this point of view, topics in flow field are more explainable than those in text analysis.

3.3 Features Definition

The input of LDA model is the bags of features, which is analogue to the the bags of words in topic model. In topic model, documents can be naturally decomposed into words. The order of words in one document does not influence the results. However, a simple copy of the definition doesn't work in the flow field scenario. First of all, dividing pathlines based on spatial location doesn't make much sense, since they rarely share the same sample points. Such a trivial definition will lead to meaningless results. More importantly, the advection information of pathlines will be lost if we do not involve the order of sample points into the FLDA model.

In our approach, we define various kinds of features as words to describe flow behaviors from different facets. The predefined feature sets for every sample point on a pathline are listed below. All these feature sets are assembled to a huge vocabulary $\mathcal{V}_f$.

- Speed magnitude ($\mathcal{F}_{\text{Speed}}$).
- Attribute value ($\mathcal{F}_{\text{Attr}}$).
- Turning angle ($\mathcal{F}_{\text{Angle}}$).
- Spatial positions ($\mathcal{F}_{\text{Block}}$).

Words in topic models are always discrete, while features we defined in FLDA are usually from continuous numerical ranges (or spaces). We divide the whole range or space into several bins or regions, each of which corresponds to one feature. At the same time, we add temporal information to distinguish between similar behaviors but appeared at different moments. Even similar bins or regions in one facet but at different timesteps correspond different features, which is a major difference from the word definition in topic model. In this way, the generated topics could be decomposed along time dimension to give more useful explanation for unsteady flow field. There is a possibility that feaures are defined as behaviors across consecutive timesteps. However, this type of features definitions will depend on users' priori knowledge on data.

Take the facet of speed magnitude for example, we can generate a feature set as follows: At timestep t, the speed magnitude of all sample points falls into an interval $[a_t, b_t]$. We can draw a finite sequence on the interval

$$a_t = x_{t,0} < x_{t,1} < x_{t,2} < \cdots < x_{t,n_t} = b_t,$$

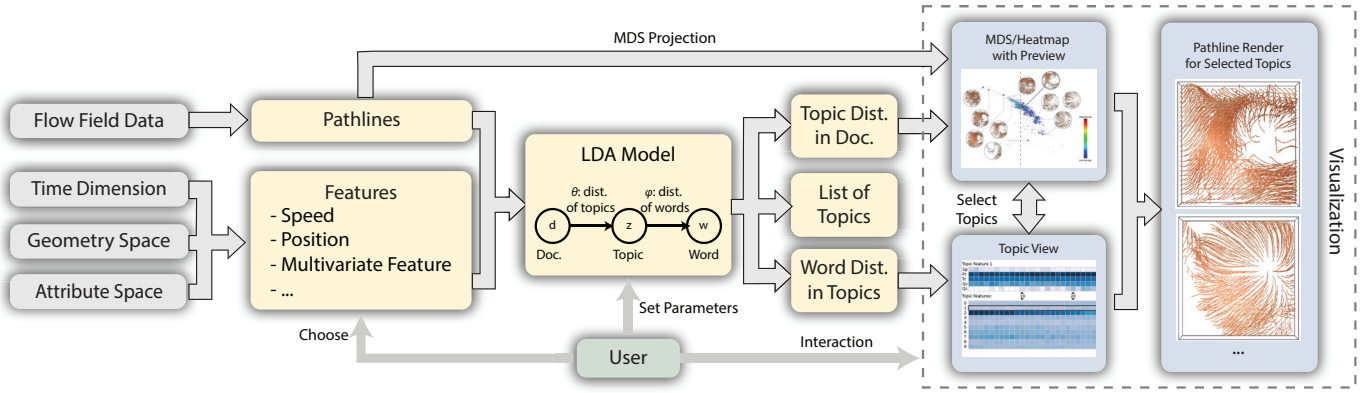


Fig. 3. Pipeline of our system. Pathlines are extracted from flow field in the preprocess step. After users choose the predefined feature sets to make feature vocabulary, every pathline is treated as a bag of features, which serves as the input of the LDA model. Topics, topic distributions per pathline, and feature distributions per topic are generated by LDA model. Multiple interactive views are created to visualize the results.

which partitions the numerical range into n_t equal-range subintervals. Then we define a speed magnitude feature $s_{(t,\cdot)}$ corresponding to each subinterval, and collect all features at timestep t, i.e.

$$\begin{aligned} s_{t,i} &\equiv [x_{t,i}, x_{t,i+1}], \text{where } 0 \leq i < n_t, \\ \mathcal{S}_t &= \{s_{t,0}, \cdots, s_{t,n_t-1}\}. \end{aligned} \tag{1}$$

By collecting all speed magnitude features $\mathcal{S}_t$ at every timestep, we obtain a feature set for speed magnitude facet,

$$\mathcal{F}_{\text{Speed}} \equiv \mathcal{S} = \bigcup_{t=0}^{T-1} \mathcal{S}_t, \text{where } 0 \leq t < T.$$

We can define feature sets for other facets similarly. In this way, a feature vocabulary is obtained by uniting all predefined feature sets,

$$\mathcal{V}_f = \mathcal{F}_{\text{Speed}} \cup \mathcal{F}_{\text{Attr}} \cup \mathcal{F}_{\text{Angle}} \cup \mathcal{F}_{\text{Block}} \cup \cdots$$

We can also decompose the feature sets and the feature vocabulary along the time dimension:

$$\mathcal{V}_f(t) = \mathcal{F}_{\text{Speed}}(t) \cup \mathcal{F}_{\text{Attr}}(t) \cup \mathcal{F}_{\text{Angle}}(t) \cup \mathcal{F}_{\text{Block}}(t) \cup \cdots,$$

where t donates one timestep.

The definition of feature vocabulary $\mathcal{V}_f$ is flexible. On the one hand, users could select one or multiple feature sets to construct a new feature vocabulary to meet their requirements. On the other hand, simple feature definitions are not useless in our approach. Even with simple features, FLDA model could generate meaningful topics by mixing features to describe complicated flow behaviors.

3.4 System Pipeline

Based on FLDA model, we implement a prototype system to support the visualization and exploration on multivariate flow field data (Figure 3).

In our preprocessing step, we first extract pathlines from an unsteady flow field, together with multivariate information. Pathlines are generated using adaptive Runge-Kunta method, and the resampled with a fixed time interval. At the same time, we employ our previous work, LASP [17], to project the multivariate pathlines into a 2D space based on their distances in the attribute space along the advection, so that intrinsic multivariate structures are revealed.

The FLDA model starts from desired facets chosen by users from the system presets. Feature sets are generated from these definitions according to the previous description. For every pathline, the system then picks up those features whose corresponding behaviors are observed. After that, each pathline can be regarded as a bag of features, which serves as the input of LDA model. The underlying latent topic model is then estimated, where topics, topic distributions per pathline, and feature distributions per topic are generated.

In the visualization part, we create the *MDS/Heatmap view* and the *feature view* to show the results. The MDS/Heatmap view gives an overview of the topic distribution in the projection space, together with thumbnail previews showing the spatial distribution of each topic. The feature view visualizes temporal distributions of features for every topic. Users are able to find out when and in which facets the pathlines in the same topic resemble each other. Topic highlighting is supported in both views. The MDS/Heatmap view can highlight pathlines who have high degree of memberships in the selected topic. The feature view then visualizes the temporal distribution of features for every facet, where users are able to observe which facets of similar behaviors are dominated in this topic, and when this happens. Apart from the two views, in the pathline view, we render those pathlines with high degree of membership.

In the system, users are able to adjust the LDA parameters, including feature vocabulary, the number of topic number, Dirichlet prior α, β, the number of iterations, etc. Among them, feature definition is significantly responsible for the results. Besides the several presets provided by the system, users can also extend the vocabulary by adding their own feature sets on demand.

4 System Details

We implement a prototype system which employs FLDA for multivariate flow field analysis. We first introduce our system interface, where users can explore the results of FLDA from various aspects. We then give details on experiments under different parameter settings.

4.1 User Interface

The interface includes four parts: parameter setting panel, pathline view, preview with MDS/Heatmap and feature view, as shown in Figure 1.

4.1.1 Parameter Setting Panel

In the parameters setting panel (Figure 1(a)), users are able to choose different feature sets to compose the feature vocabulary $\mathcal{V}_f$ and set necessary parameters. Users then launch FLDA analysis. Results generated are visualized in the three views.

4.1.2 Pathline View

The pathline view (Figure 1(c)) is set to visualize the spatial distribution of pathlines which have a high degree of membership to a specific topic. The probability $\theta_{j,k}$ is considered as the degree of membership of document d_j to topic k. Interactions like rotating, zooming and panning are provided to enable an elaborative observation from different viewports. By inspecting the spatial shapes of pathlines, users can estimate the behavior coherence within a topic, so as to evaluate the rationality of topics.

4.1.3 Preview with MDS/Heatmap

In this view, we mainly focus on the topic distribution per document $\theta_{(\cdot,\cdot)}$ generated by FLDA. The MDS/Heatmap view (Figure 1(b)) consists of two parts, namely the MDS projection of the original pathlines and the spatial previews of topics. This view mainly serves three purposes:

1. Reveal the intrinsic multivariate structures of attribute space.
2. Provide an intuitive overview of all topics by embedding their spatial distributions as previews.
3. Reveal the correlation of different topics by comparing the multivariate distributions of their members.

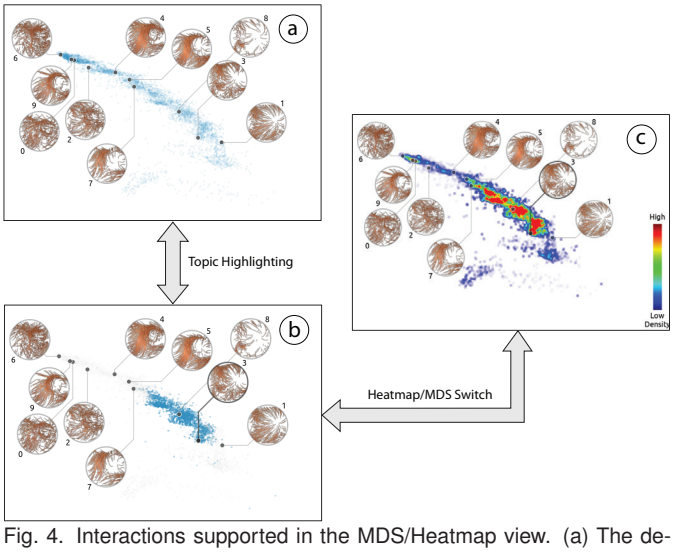


Fig. 4. Interactions supported in the MDS/Heatmap view. (a) The default MDS projection is rendered as a scatterplot. (b) When a topic is selected, samples with a high degree of membership to this topic is highlighted using focus+context technique. (c) Users can switch the projection to a heatmap style where the density information could be more clearly seen.

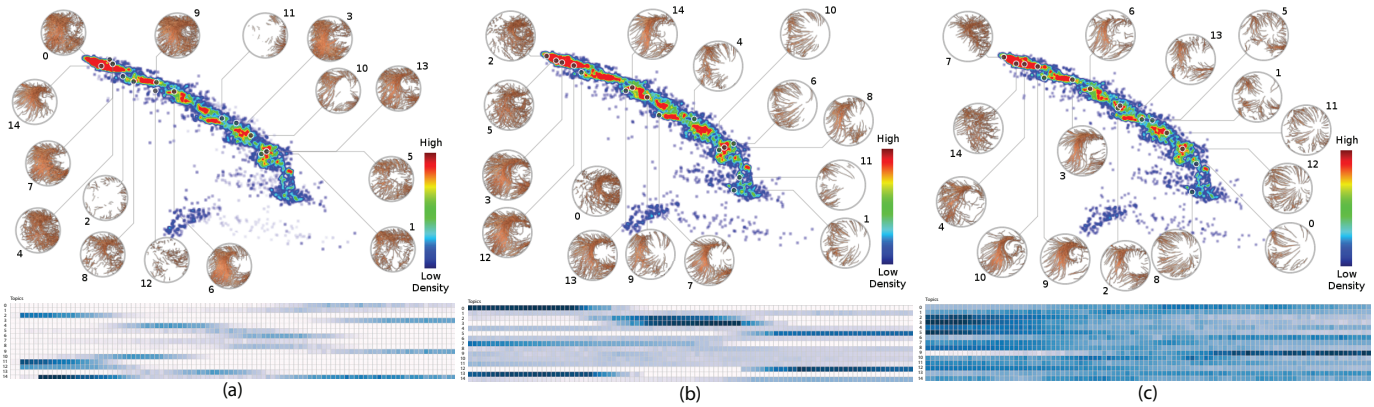


Fig. 5. Visualization of FLDA results under different feature set definitions. (a) Feature sets of SPEED and Turning Angle; (b) Feature set of P only; (c) Feature set of TC only.

For the first purpose, we provide a LASP projection plot [17] to present the distribution of pathlines. The distance of two pathlines is defined in the attribute space with Lagrangian specification, which is calculated by accumulating the differences of attribute values at all corresponding sample positions of pathlines. Two rendering styles are provided: the traditional scatterplot style and the heatmap style (Figure 4(b)). Users are allowed to change the projection style, depending on whether individual data samples or density of the distribution is the major concern. In the heatmap style, users can adjust the color mapping to emphasize areas with different densities, so as to recover details that would be hidden with an improper color scale.

For the second purpose, we introduce previews in the projection plot to indicate the spatial distributions of topics. Each preview contains a small snapshot of the pathline rendering result of the corresponding topic. Pathlines in previews are kept the same temporal and spatial range as in pathline view. To avoid clutter in previews, only a subset of pathlines are rendered in the preview. The number of the subset comes from a power function of original number of pathlines in the topic. For every topic k, the corresponding preview is anchored at its central sample, which minimizes the distance to all samples weighted by probability $\theta_{j,k}$. A leader then emits from the anchor position to its corresponding thumbnail. The previews are carefully arranged to avoid overlaps, occlusions and line crossings, using a heuristic zone-based labeling algorithm proposed by Wu et al. [45], which combines genetic and greedy strategies to minimize the overall length of lines. Users can select a topic by clicking on its preview, which results in highlighting in the pathline view and the projection view.

For the last purpose, topic highlighting is introduced (Figure 4(b)). In the scatterplot style, we adopt the focus+context strategy and highlight those samples with a high degree membership while fading out others. When in the heatmap style, we modulate the density by the weights $\theta_{j,k}$ for every sample to emphasize distribution of the chosen topic k (Figure 4(c)). By viewing the distribution of the topic, users can see clearly how its members resemble each other in the attribute space. By switching between focuses, users can learn the differences between topics.

Providing spatial previews greatly enhances users' comprehension of the results. Firstly, embedding previews in the projection closely relates the spatial and multivariate distributions of a topic without the trouble to switch between contexts. Secondly, by viewing the snapshots before they drill down to details, users may gain an intuitive perception about the spatial features of topics, which may cast a light on the following exploration. At last, the small multiple strategy enables users to compare different topics without switching between contexts.

4.1.4 Feature View

While the MDS/Heatmap view reveals correlations between pathlines and topics $\theta_{(\cdot,\cdot)}$ in the attribute space, the feature view (Figure 1(d)) visualizes the relationships between features and topics $\phi_{(\cdot,\cdot)}$ in an time-varying context. The feature view contains three parts: the topic view, facet view for one selected topic, and time histogram for one facet of one selected topic, all of which are of pixel-oriented style. These three views enable a progressive exploration process for topic-feature relationships.

In the topic view, each row shows the feature distribution ϕ_k for one topic k, while each column indicates one timestep. The color in every cell encodes the accumulated probability of all features at the timestep t for the topic k, i.e.

$$P(t,k) \equiv \sum_{x} \sum_{w \in \mathcal{F}_x(t)} \phi_{w,k}.$$

Users can perform topic selection and time span selection in the topic view. Topic selection is linked to the MDS/Heatmap view and pathline view, which gives detailed information for the chosen topic. It will also trigger the attribute view to refresh for the selected topic.

When a topic k is selected, the attribute view visualizes the distribution ϕ_k for every facet x. Each row corresponds to one feature set $\mathcal{F}_x$, and each column still indicates one timestep. For feature set $\mathcal{F}_x$, each cell encodes the accumulated probability alone for this topic, i.e.

$$P_k(x,t) \equiv \sum_{w \in \mathcal{F}_x(t)} \phi_{w,k}.$$

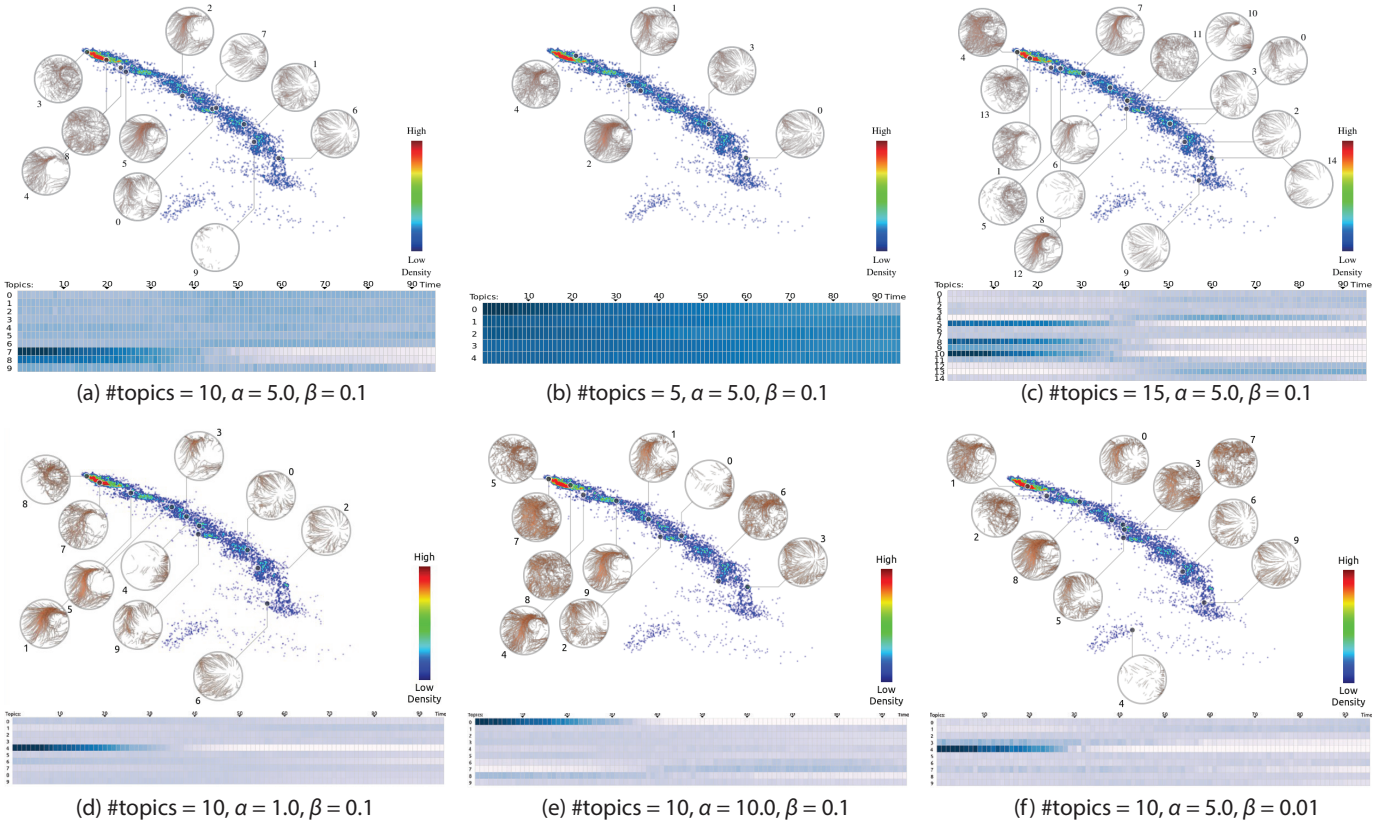
By decomposing the ϕ_k from a whole feature vocabulary into several feature sets, users are able to observe what kind of similar behaviors aggregate more in this topic, and when this happens.

Since our features are defined based on discretization of value range, there is a demand to investigate the change of the features' corresponding values in one topic. We further create a pop-up time histogram to visualize the change of values when a topic k and a feature set $\mathcal{F}_x$ is selected. Every column is a histogram at corresponding timestep t which comes directly from the probability of features defined in Equation 1.

With the three parts in the feature view, $\theta_{(\cdot,\cdot)}$ is visualized at different aggregation levels, which provides fruitful explanation for topics. It is helpful for understanding the data on the topic-word level, and in the temporal context. On the one hand, the topic view provides a good access for users to observe and compare the feature (word) distributions of topics, which could gives insight into how the topic coherency varies across time. On the other hand, it reveals the multivariate time-varying features of topics, indicating *when and in which facets* pathlines (documents) resemble each other.

4.2 FLDA Parameter Setting

In the typical LDA model, a set of parameters can be adjusted to tune the results, including number of topics K, Dirichlete prior on the per-document topic distributions α, Dirichlete prior on the per-topic word distribution β, and number of iterations. In our FLDA model, predefined feature vocabulary is also provided as parameters.



(a) #topics = 10, $\alpha = 5.0, \beta = 0.1$ (b) #topics = 5, $\alpha = 5.0, \beta = 0.1$ (c) #topics = 15, $\alpha = 5.0, \beta = 0.1$

(d) #topics = 10, $\alpha = 1.0, \beta = 0.1$ (e) #topics = 10, $\alpha = 10.0, \beta = 0.1$ (f) #topics = 10, $\alpha = 5.0, \beta = 0.01$

Fig. 6. Visualization of FLDA results of different parameter combinations under a fixed vocabulary of attribute feature sets. $\#\text{topics} = 10, \alpha = 5.0, \beta = 0.1$ is chosen as default setting. From the results, the number of topics shows great influence over the results, while α and β show small sensitivity.

The topic number K is difficult to decide since it greatly depends on the actual data. Too few topics will miss some important patterns, while too many topics may produce redundant, meaningless or trivial results. The Dirichlet prior α and β influence the topic distribution per document and the word distribution per topic respectively. Smaller values make the distribution more concentrated, and vice versa. However, as the α and β values in our work, we do not observe significant differences in the results. The number of iterations also affects the quality of results, but we found that the output converges quickly within a hundred iterations. In our case study, we fix this value to 100, which is a balance between time efficiency and the quality of results.

In our experiments, we first investigate the FLDA output under different definitions of features. In Figure 5, we show the topic view and MDS/Heatmap view of the results side by side. We can observe some topics have significantly uneven temporal distribution of features in configuration (a) and (b) from topic view, while in configuration (c), the temporal distribution for topics are more even. By careful comparison, we are able to find lot of differences between topics from the previews.

We then tested combinations of other parameters under a fixed feature vocabulary which contains only the 5 scalar attributes of Isabel data. We use $\#\text{topics} = 10, \alpha = 5.0$, and $\beta = 0.1$ as the default setting. From the results shown in Figure 6, we can see that as the number of topics increases, the temporal distribution of features has significant changes. For some topics, the accumulated distribution of their features becomes more concentrated on a small time range instead of being evenly spread over the whole time span. From the Heatmap with previews, we are able to observe that the topics are merging and splitting as K increases. However, for Dirichlet prior α, β, our test shows that these two parameters have relatively small sensitivity in our FLDA model.

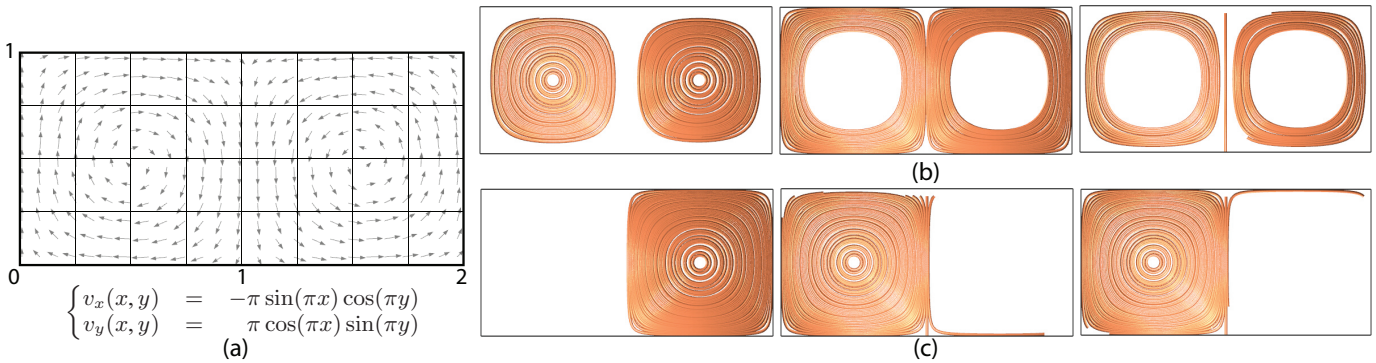


Fig. 7. Visualization results for Double Gyre dataset (a). We extract 3 topics from Double Gyre dataset with feature sets of turning angles (b) and spatial positions (c) respectively. For turning angles features, we use histograms of 256 bins to discretize angle values. For spatial positions features, we parition the space domain into 8×4 blocks.

5 Case Study

As case studies, we tested our FLDA method on the Double Gyre Data and the Hurricane Isabel data. In this section, we present the corresponding results, and discuss the effectiveness of our system.

5.1 Double Gyre Data

We use a simple dataset, Double Gyre, to demonstrate the basic usage of FLDA. The vector field is given by

$$\begin{cases} v_x(x,y) &= -\pi \sin(\pi x)\cos(\pi y) \\ v_y(x,y) &= \pi \cos(\pi x)\sin(\pi y) \end{cases}$$

over the region $[0,2] \times [0,1]$. The flow field consists of two symmetric vortexes which are shown in Figure 7(a). We choose this data set in order to give an intuitive understanding of the relationship between the extracted topics and observed flow properties. Although this data is time-independent, there is no barrier to apply our method to streamlines.

We construct feature vocabulary from facets of turning angle and spatial blocks respectively. Three topics are extracted with α set to 0.01, β set to 0.001, and the number of iterations being 1000. Results are shown in Figure 7(b)(c). When we choose turning angles as words, the streamlines are clustered by their distances to the center of the corresponding vortex, to which the turning angle behaviors of streamlines is strongly related. For the feature vocabulary of spatial blocks, we partition the domain into 8×4 grids to obtain words. From the results, we can observe there are mainly two topics of streamlines, which are basically constructed from the left and right vortex, and another topic is almost duplicated in this setting. These extracted topics match our observation.

In this case, the relationship between topics and feature vocabularies is easily obtained, since we only involve one facet in the feature vocabulary and do not consider attribute information. When it comes to complicated scenarios, more explorations are required to get the insight into topics.

5.2 Isabel Data

Hurricane Isabel data comes from an atmospheric simulation. The spatial resolution of this data set is $500 \times 500 \times 100$, covering a physical space of $2,139$ km $\times$ $2,004$ km $\times$ 19.8 km. The data has 48 timesteps corresponding to 48 hours. As for attributes, we consider the wind speed vector filed (U, V, and W), and five scalar fields, including wind speed magnitude (Sp), pressure (Pr), temperature (Tc), the water vapour mixing ratio (Qv), and total cloud moisture mixing ratio (Qc), which are suggested by domain experts as important attributes for hurricane analysis. In the preprocessing step, we extracted 5,768 pathlines which are traced from time 0 with 4 samples per hour.

For this data, we only consider facets of the 5 scalar attributes mentioned above. We use the FLDA model to extract 15 topics from pathlines with α set to 5.0, β set to 0.1, and the iteration count to100. An overview of all topics is presented in Figure 8(a)(b), including the MDS/heatmap view and the feature view. Among the 15 extracted topics, there are three topics, the 9^{th}, 12^{th}, and 14^{th}, which show some interesting spatial behaviors as displayed in Figure 8(c)-(e). Topic (c) contains pathlines advecting from the hurricane eye to the outside in the low altitude region, while topic (d) contains pathlines that travel from outside to inside in an anti-clockwise direction with a higher altitude. Pathlines in topic (e) are advecting at the periphery of the hurricane, which also inhabits in the low altitude region. Figure 8(f)-(h) shows the attribute view, and the time histogram of pressure (Pr) and temperature (Tc) for each topic. Besides the geometric patterns, they can also be treated as clusters from multivariate facet. We can observe that pathlines in topic (c) have more similar multivariate behaviors in the first half of the advection, while the similar phenomenon appear in the last half of advection for topic (d). For pathlines in topic (e), the similarity of attributes is roughly stable through all the advection time. From the time histogram, explicit attribute changes of the topics could be more clearly observed. Pathlines in topic (c) and (e)

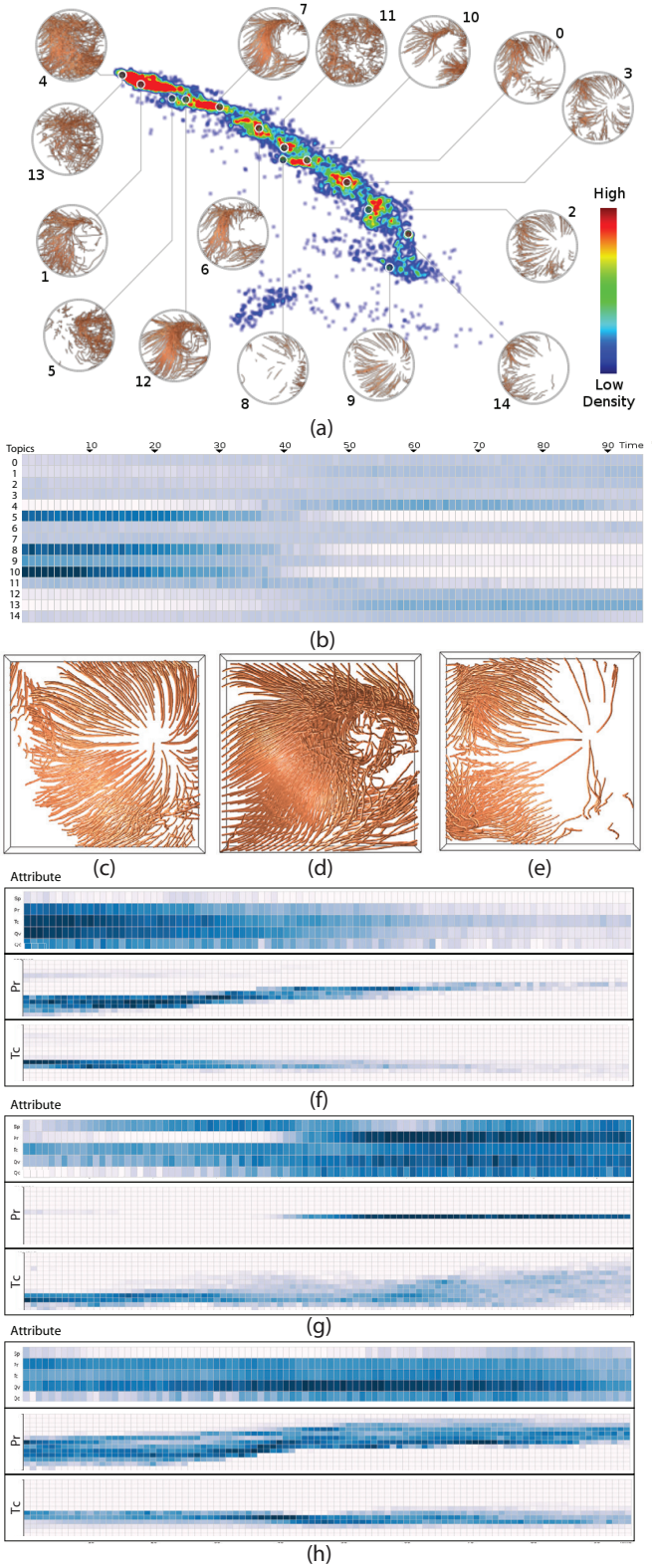


Fig. 8. We extract 15 topics from Hurricane Isabel data by considering the feature sets of wind speed magnitude (Sp), pressure (Pr), temperature (Tc), the water vapour mixing ratio (Qv), and total cloud moisture mixing ratio (Qc). The overview of topics are visualized in (a) heatmap view and (b) topic view. The 9^{th}, 12^{th}, and 14^{th} topic are selected for further investigation. These three topics have very different spatial behaviors as shown in (c)-(e). The attribute view of each topic and time histogram of Pr and Tc are shown in (f)-(h).

have an increasing pressure and decreasing temperature in the advection process, which indicates lower pressure and higher temperature in the hurricane eye than the periphery. For pathlines in topic (d), the temperature generally goes up when they move towards the center, but the pressure always keep stable. Those discoveries could be explained by their advection, since the sample points in the hurricane eye have more similar properties, while those outside the hurricane have more diverging behaviors.

6 Discussions

In this part, we compare our FLDA model to other methods from two aspects: flow exploration and pathline clustering.

6.1 Comparison to Flow Analysis Methods

The FLDA model brings a novel perspective to explore flow filed data. Previous flow exploration methods include texture-based and geometry-based methods, parallel coordinates, projection methods, interactive graph exploration, etc. Compared to these methods, our method have two major differences. 1) FLDA model not only clusters pathlines using the fuzzy assignment, but also produces meaningful multi-facet topics by incorporating simple features. These topics reveal complex inherent flow behaviors, which may be difficult to define for detection and extraction without prior knowledge. 2) It's easy for our approach to fuse features from various facets of the flow field data by treating every pathline as bags of features. Moreover, the feature components could be totally heterogeneous from very different fields, which enables users to explore the data in a more flexible way. While these complex flow behaviors are often difficult to find without priori knowledge.

6.2 Comparison to Cluster Algorithms

We also conducted a comparison between K-Means clustering algorithm and our flow LDA model on the Isabel data. The K-Means algorithm calculates the distance matrix of pathlines using the accumulation of sample-wise distances in the attribute space, which is the same with LASP [17]. Since only the distance matrix instead of the original high-dimensional data is available, we actually use a variation of K-Means, named K-Medoids. Our method also uses all 5 attributes to define the feature vocabulary. The cluster (topic) number is set to 5 for a simpler and intuitive comparison. The clustering results are compared side by side in Figure 9. We can observe fairly close results except the 2nd cluster (topic). These clusters have very similar spatial behaviors and distributions in the projection space, which could prove the effectiveness of our method as a clustering algorithm. However, the FLDA model excels the K-Means in that it provides a fuzzy, rather than binary description on the cluster distributions. In this way, it's also less sensitive to the value of K, since samples are not exclusive in the clustering process. For the 2nd cluster (topic), because of deterministic assignment of K-Means, pathlines in this cluster are isolated from other parts. While in FLDA model, the probabilistic assignment not only relieves this problem, but also reveals additional compensatory information by providing another topic. Besides the projected results, users are also able to perceive when and in which attributes are the pathlines more similar or more diverged in the attribute space from the feature distribution view.

Apart from more flexible and explainable results, there is a better time complexity of FLDA model when a small number of feature sets is chosen to construct vocabulary $\mathcal{V}_f$. In K-Means algorithm, the precomputation of distance matrix costs $O(D^2T)$ and one iteration costs $O(D^2)$, while one iteration in LDA model costs $O(KDM)$. M denotes the average size of features in pathlines, which equals T times the number of features chosen in our definition. In a common scenario, the number of features in vocabulary and the number of topics is often a small number compared to D, which makes FLDA model generally faster than the K-Means algorithm.

7 Conclusions and Future Work

In this paper, we introduce a novel LDA-based flow analysis method, *Flow LDA* (FLDA). We extend the traditional LDA model to flow field scenario with a quite different definition of words. A prototype system is developed to extract flow features, as well as to explain the semantics of topics. Our case studies demonstrate the effectiveness of the FLDA model.

In the future, we would like to develop our work in different ways. In our experiments, the parameters in FLDA models show different sensitivity behaviors from the topic model. The effects of parameters to LDA results can be studied thoroughly to provide an exploration guidance. The FLDA can also be further introduced into ensemble scenario to give a comparative analysis and visualization to ensemble simulation data.

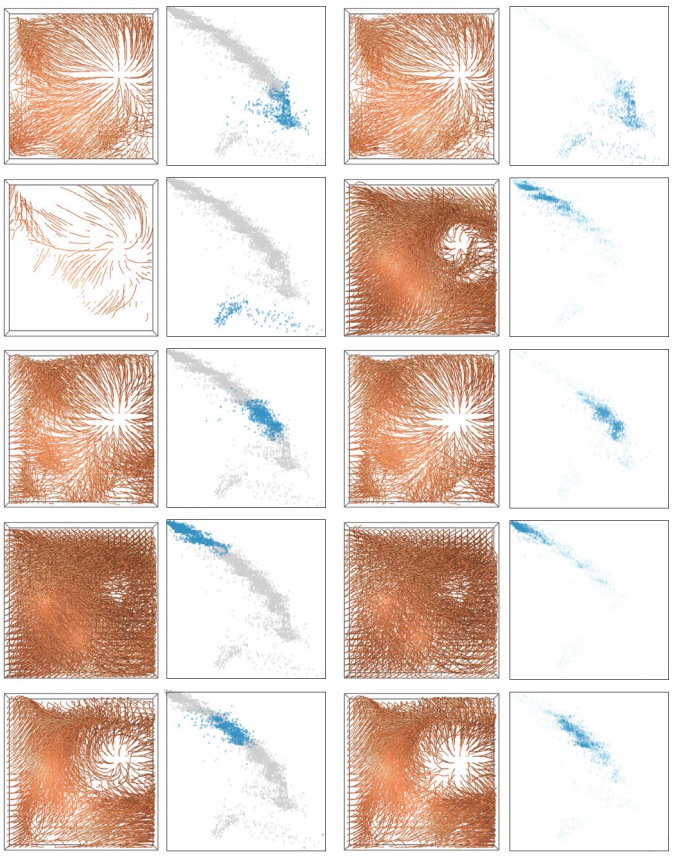


Fig. 9. Cluster results comparison between K-Means algorithm (a) and FLDA model (b). The number of clusters (topics), K, is set to 5.

Acknowledgments

This work is supported by NSFC No. 61170204. This work is also partially supported by NSFC Key Project No. 61232012 and the Strategic Priority Research Program - Climate Change: Carbon Budget and Relevant Issues of the Chinese Academy of Sciences Grant No. XDA05040205.

References

[1] H. Akiba and K.-L. Ma. A tri-space visualization interface for analyzing time-varying multivariate volume data. In *Proceedings of the 9th Joint Eurographics/IEEE VGTC conference on Visualization*, pages 115–122, 2007.

[2] H. Akiba, K.-L. Ma, J. H. Chen, and E. R. Hawkes. Visualizing multivariate volume data from turbulent combustion simulations. *Computing in Science and Engineering*, 9(2):76–83, 2007.

[3] A. Biswas, S. Dutta, H.-W. Shen, and J. Woodring. An information-aware framework for exploring multivariate data sets. *IEEE Trans. Vis. Comput. Graph.*, 19(12):2683–2692, 2013.

[4] J. Blaas, C. P. Botha, and F. H. Post. Extensions of parallel coordinates for interactive exploration of large multi-timepoint data sets. *IEEE Trans. Vis. Comput. Graph.*, 14(6):1436–1451, 2008.

[5] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[6] U. Brandes and C. Pich. Eigensolver methods for progressive multidimensional scaling of large data. In *Graph Drawing*, pages 42–53, 2006.

[7] S. Bruckner and T. Möller. Result-driven exploration of simulation parameter spaces for visual effects design. *IEEE Trans. Vis. Comput. Graph.*, 16(6):1468–1476, 2010.

[8] R. Bujack, I. Hotz, G. Scheuermann, and E. Hitzer. Moment invariants for 2d flow fields using normalization. In *Proc. IEEE Pacific Visualization Symposium 2014*, pages 33–40. IEEE, 2014.

[9] L. Cao and F.-F. Li. Spatially coherent latent topic model for concurrent segmentation and classification of objects and scenes. In *Proc. of International Conf. on Computer Vision*, pages 1–8, 2007.

[10] C.-K. Chen, C. Wang, K.-L. Ma, and A. T. Wittenberg. Static correlation visualization for large time-varying volume data. In *Proc. IEEE Pacific Visualization Symposium 2011*, pages 27–34, 2011.

[11] W. Chen, Z. Ding, S. Zhang, A. MacKay-Brandt, S. Correia, H. Qu, J. A. Crow, D. F. Tate, Z. Yan, and Q. Peng. A novel interface for interactive exploration of dti fibers. *IEEE Trans. Vis. Comput. Graph.*, 15(6):1433–1440, 2009.

[12] D. Chu, D. Sheets, Y. Zhao, Y. Wu, J. Yang, M. Zheng, and G. Chen. Visualizing hidden themes of taxi movement with semantic transformation. In *Proc. IEEE Pacific Visualization Symposium 2014*, pages 137–146. IEEE, 2014.

[13] J. Chuang, C. D. Manning, and J. Heer. Termite: visualization techniques for assessing textual topic models. In *Proceedings of the International Working Conference on Advanced Visual Interfaces*, pages 74–77, 2012.

[14] W. Cui, S. Liu, L. Tan, C. Shi, Y. Song, Z. Gao, H. Qu, and X. Tong. Textflow: Towards better understanding of evolving topics in text. *IEEE Trans. Vis. Comput. Graph.*, 17(12):2412–2421, 2011.

[15] H. Doleisch. Simvis: interactive visual analysis of large and time-dependent 3d simulation data. In *Winter Simulation Conference*, pages 712–720, 2007.

[16] Y. Gu and C. Wang. Transgraph: Hierarchical exploration of transition relationships in time-varying volumetric data. *IEEE Trans. Vis. Comput. Graph.*, 17(12):2015–2024, 2011.

[17] H. Guo, F. Hong, Q. Shu, J. Zhang, J. Huang, and X. Yuan. Scalable lagrangian-based attribute space projection for multivariate unsteady flow data. In *Proc. IEEE Pacific Visualization Symposium 2014*, pages 33–40. IEEE, 2014.

[18] H. Guo, H. Xiao, and X. Yuan. Scalable multivariate volume visualization and analysis based on dimension projection and parallel coordinates. *IEEE Trans. Vis. Comput. Graph.*, 18(9):1397–1410, 2012.

[19] T. Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57, 1999.

[20] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42(1/2):177–196, 2001.

[21] B. W. H. II, E. M. Talley, G. A. P. C. Burns, D. Newman, and G. LaRowe. The nih visual browser: An interactive visualization of biomedical research. In *Proceedings of the 13th International Conference Information Visualization*, pages 505–509, 2009.

[22] J. D. II, E. W. Anderson, L. G. Nonato, and C. T. Silva. Interactive vector field feature identification. *IEEE Trans. Vis. Comput. Graph.*, 16(6):1560–1568, 2010.

[23] A. Inselberg. The plane with parallel coordinates. *The Visual Computer*, 1(2):69–91, 1985.

[24] T. Iwata, T. Yamada, and N. Ueda. Probabilistic latent semantic visualization: topic model for visualizing documents. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 363–371, 2008.

[25] H. Jänicke, M. Böttinger, and G. Scheuermann. Brushing of attribute clouds for the visualization of multivariate data. *IEEE Trans. Vis. Comput. Graph.*, 14(6):1459–1466, 2008.

[26] H. Jänicke and G. Scheuermann. Visual analysis of flow features using information theory. *IEEE Computer Graphics and Applications*, 30(1):40–49, 2010.

[27] C. Jones, K.-L. Ma, S. Ethier, and W.-L. Lee. An integrated exploration approach to visualizing multivariate particle data. *Computing in Science and Engineering*, 10(4):20–29, 2008.

[28] W. Kendall, J. Huang, and T. Peterka. Geometric quantification of features in large flow fields. *IEEE Computer Graphics and Applications*, 32(4):46–54, 2012.

[29] T. K. Landauer, P. W. Foltz, and D. Laham. An introduction to latent semantic analysis. *Discourse Processes*, 25:259–284, 1998.

[30] R. S. Laramee, H. Hauser, H. Doleisch, B. Vrolijk, F. H. Post, and D. Weiskopf. The state of the art in flow visualization: Dense and texture-based techniques. *Comput. Graph. Forum*, 23(2):203–222, 2004.

[31] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2006*, pages 2169–2178, 2006.

[32] H. Lee, J. Kihm, J. Choo, J. T. Stasko, and H. Park. ivisclustering: An interactive visual document clustering via topic modeling. *Comput. Graph. Forum*, 31(3):1155–1164, 2012.

[33] L. Linsen, T. V. Long, P. Rosenthal, and S. Rosswog. Surface extraction from multi-field particle volume data using multi-dimensional cluster visualization. *IEEE Trans. Vis. Comput. Graph.*, 14(6):1483–1490, 2008.

[34] J. Ma, C. Wang, and C.-K. Shene. Flowgraph: A compound hierarchical graph for flow field exploration. In *Visualization Symposium (PacificVis), 2013 IEEE Pacific*, pages 233–240, Feb 2013.

[35] R. Maciejewski, Y. Jang, I. Woo, H. Jänicke, K. P. Gaither, and D. S. Ebert. Abstracting attribute space for transfer function exploration and design. *IEEE Trans. Vis. Comput. Graph.*, 19(1):94–107, 2013.

[36] T. McLoughlin, R. S. Laramee, R. Peikert, F. H. Post, and M. Chen. Over two decades of integration-based, geometric flow visualization. *Comput. Graph. Forum*, 29(6):1807–1829, 2010.

[37] X.-H. Phan and C.-T. Nguyen. Gibbslda++: A c/c++ implementation of latent dirichlet allocation (lda), 2007.

[38] C. Rössl and H. Theisel. Streamline embedding for 3d vector field exploration. *IEEE Trans. Vis. Comput. Graph.*, 18(3):407–420, 2012.

[39] T. Salzbrunn, C. Garth, G. Scheuermann, and J. Meyer. Pathline predicates and unsteady flow structures. *The Visual Computer*, 24(12):1039–1051, 2008.

[40] N. Sauber, H. Theisel, and H.-P. Seidel. Multifield-graphs: An approach to visualizing correlations in multifield scalar data. *IEEE Trans. Vis. Comput. Graph.*, 12(5):917–924, 2006.

[41] C. Wang, H. Yu, R. W. Grout, K.-L. Ma, and J. H. Chen. Analyzing information transfer in time-varying multivariate data. In *Proc. IEEE Pacific Visualization Symposium 2011*, pages 99–106, 2011.

[42] X. Wang and E. Grimson. Spatial latent dirichlet allocation. In *Proceedings of Neural Information Processing Systems Conference (NIPS) 2007*, 2007.

[43] F. Wei, S. Liu, Y. Song, S. Pan, M. X. Zhou, W. Qian, L. Shi, L. Tan, and Q. Zhang. Tiara: a visual exploratory text analytic system. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 153–162, 2010.

[44] J. Woodring and H.-W. Shen. Multi-variate, time varying, and comparative visualization with contextual cues. *IEEE Trans. Vis. Comput. Graph.*, 12(5):909–916, 2006.

[45] H.-Y. Wu, S. Takahashi, C.-C. Lin, and H.-C. Yen. A zone-based approach for placing annotation labels on metro maps. In *Smart Graphics*, pages 91–102, 2011.

[46] L. Xu, T.-Y. Lee, and H.-W. Shen. An information-theoretic framework for flow visualization. *IEEE Trans. Vis. Comput. Graph.*, 16(6):1216–1224, 2010.

[47] X. Zhao and A. E. Kaufman. Multi-dimensional reduction and transfer function design using parallel coordinates. In *Proceedings of the 8th IEEE/EG international conference on Volume Graphics*, pages 69–76, 2010.