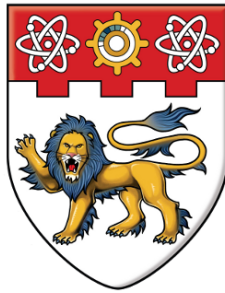


Topic Models for the Inference of Human Mobility



Daniel Rugeles

Supervisor: Asst. Professor Gao Cong

Dr. Shonali Krishnaswamy

School of Computer Engineering

Nanyang Technological University

Report Submitted for the Confirmation for the Admission to the Degree of
Doctor of Philosophy

December 2015

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains fewer than 65,000 words including appendices, bibliography, footnotes, tables and equations and has fewer than 150 figures.

Daniel Rugeles

December 2015

Acknowledgements

I express my gratefulness for being part of the teams under the direction of Prof. Gao Cong at Nanyang Technological University, and Dr. Shonali Krishnaswamy in the Institute for Infocomm Research, A*Star. They have helped me to shape my research skills including finding the right scientific resources to follow, developing critical thinking to produce state of the art contributions and scientifically conducting experiments. It is my pleasure to study and research under their supervision.

Also, I would like to thank some of our team members. Without their help, it would not have been possible to evolve in my research project. I specially thank Dr. Quan Yuan and Kaiqi Zhao from the DISCO lab, and Dr. Manoranjan Dash, Dr. Joao Gomes, Dr. Nguyen Minh Nhut and Koo Kee Kiat from A*star.

Abstract

Latent Dirichlet Allocation (LDA) models the relationship between documents and words by representing a document as the collection of words written in the document, such configuration allows the extraction of topics represented by distributions over words (word-topics). Alternatively, we may think of switching the role between documents and words. i.e. representing a word as the collection of documents that include the word. This novel configuration allows the extraction of topics represented by distributions over documents (document-topics). Both configurations use the same data to extract different but complementary information. We propose combining the two configurations into one model by modeling the mutual reinforcement relationship between the word-topics and the document-topics. Our experiments show that our model gives a better fit than other topic models on held-out test-sets for several data-sets. We also compare topic extraction and we study biclustering and reviewer recommendation as applications of our model.

Table of contents

List of figures	ix
List of tables	xii
1 Introduction	1
1.1 Background	1
1.2 Motivation	3
1.3 Aim	4
1.4 Objectives	5
1.5 Challenges	5
1.6 Organization	6
2 Literature Review	7
2.1 Topic Models and its applications in Human Mobility Field	8
2.1.1 N-gram Topic Models	9
2.1.2 Bayesian non-parametric topic models	10
2.1.3 Correlated Topic Models	11
2.1.4 Temporal Dynamics of the Topic Models	11
2.1.5 Extending LDA with additional variables	13
2.2 Improving the Inference of Topic Models	13
2.2.1 Model-based Inference methods	14

2.2.2	Family-based Inference methods	15
2.3	Biclustering	17
2.3.1	Biclustering Structures	18
2.3.2	Biclustering Types	19
2.3.3	Related biclustering Algorithms	19
3	Yin Yang Latent Dirichlet Allocation	22
3.1	Notation and Terminology	22
3.1.1	Users and Places.	22
3.1.2	Mathematical Notation	23
3.2	Designing Yin Yang Latent Dirichlet Allocation Model	23
3.2.1	Categorical Mixture Model	23
3.2.2	Latent Dirichlet Allocation	24
3.3	Yin Yang Latent Dirichlet Allocation (Y^2 -LDA)	26
3.3.1	Intuitions	26
3.3.2	Using CMM to implement intuitions 1 and 2	27
3.3.3	Defining the Bivariate Categorical Distribution to implement intuition 3	28
3.3.4	Model Representation	29
3.4	Generative Process	32
3.5	Inference Process	32
3.6	Derivation of the Inference Equation using Collapsed Gibbs Sampling	34
3.6.1	Posterior Derivation given the conditions of the model	34
3.6.2	Posterior Derivation for each record	38
3.7	Relationship with Latent Dirichlet Allocation Model	40
3.7.1	Inference of the Latent Dirichlet Allocation	40
3.7.2	Comparison with the inference equation of Y^2 -LDA	40

3.8	Bicluster Extraction using Y^2 -LDA	41
3.9	Why does Y^2 -LDA works?	43
3.9.1	Topic influence between a pair of records	43
3.9.2	Records in the same biclique have the strongest mutual influence . .	44
4	Applications and Experimental Evaluation	46
4.1	Datasets	46
4.2	Modeling Visit Records	47
4.2.1	Application	47
4.2.2	Experimental Settings	47
4.2.3	Results and Analysis	48
4.3	Reviewer Recommendation.	49
4.3.1	Application	49
4.3.2	Experimental Settings	50
4.3.3	Results and Analysis	50
4.4	User and Place Classification	51
4.4.1	Application	51
4.4.2	Experimental Settings	52
4.4.3	Results and Analysis	52
4.5	Biclustering	54
4.5.1	Application	54
4.5.2	Extracting biclusters from Y^2 -LDA	54
4.5.3	Experimental Settings	54
4.5.4	Results and Analysis	55
4.6	Dimensionality Reduction	57
4.6.1	Application	57
4.6.2	Experimental Settings	57

Table of contents	viii
4.6.3 Results and Analysis	58
4.7 Topic Visualization	59
5 Future Work	63
5.1 Topic Models for Automatic Place Labeling	63
5.2 Modeling Human Mobility	65
5.3 Fast Inference of Topic Models using Distributed Computing	66
References	68
Appendix A Perplexity of an Individual Topic when Using Y^2-LDA	75

List of figures

b	LDA approach to modeling	4
a	Relation given by the visit of users to places	4
c	Dual LDA approach to modeling	4
1.1	Modeling approach of LDA and the Dual LDA.	4
2.1	Expanded graphical representation of the LDA model. The red arrows indicate the main independence assumptions about the LDA model. a) The order of Documents is irrelevant b) The order of words is irrelevant c) Topic dependency is only given by a Dirichlet prior distribution parameterized by α d)The number of topics must be given a priori.	8
2.2	A simple topic model and an approximate topic model assuming independences between all parameters in the model	16
2.3	Bicluster Structures. Figure modified from [46]	18
3.1	Information extracted by an LDA (left) and a Dual LDA (right). The width of the arrow represent the preference of a topic, the ovals represents the topics.	24
3.2	Information extracted by an LDA (left) and a Dual LDA (right). The width of the arrow represent the preference of a topic, the ovals represents the topics.	26

3.3	Information extracted by Y^2 -LDA. The ovals represents the topics and the width of the arrows represents the strength of the relationship between two topics.	27
3.4	Implementation of intuitions 1 and 2.	27
3.5	Function used to define the <i>Categorical</i> ² distribution. The visualization indicates the mapping from the parameters to the support set.	30
3.6	Yin Yang Latent Dirichlet Allocation graphical model	30
3.7	Demonstration of how Y^2 -LDA extracts bicliques.	45
4.1	Reviewer recommendation	50
4.2	Comparison for place-topic extraction between LDA and Y^2 -LDA. Y^2 -LDA was initialized with the results of the LDA. The numbers represent the number of samples allocated to a topic	53
4.3	Recovery and Relevance results for bicluster classification. The datasets are visualized with a matrix whose color represent the number of visits of users (rows) to places (columns).	56
4.4	Different initialization methods for extracting biclusters with Y^2 -LDA . . .	57
4.5	Results of table 4.3 for every user, we sort the users and places by their average Battacharyya distance to facilitate the visualization of the results. .	59
4.6	On top, we show four movie-topics represented by a) Most representative movies and its year of release. b) Histogram of the movie genres weighted by the number of instances that a movie was allocated to a topic. On the middle we represent the correlation between topics using a parallel sets visualization. On the bottom, we show the user-topics represented by a) Histogram of the user's profession b) Histogram of the user's age. Both histograms are weighted by the number of instances that a user was allocated to a topic. . .	60

4.7	Author-topics, word-topics and its interrelation. The width of the box surrounding each topic represents the total amount of samples allocated to the topic.	62
5.1	Research Plan	63

List of tables

1.1	LDA and Dual LDA representation in several scientific fields.	2
3.1	Variable definitions for the Yin Yang Latent Dirichlet Allocation model. . .	31
4.1	Average perplexity computed over twenty trials for four different data sets. .	48
4.2	Average misclassification error for ten random initializations. The initializa- tion method is specified in the headers of the table. The asterisk represents those cases where the models completely confused two topics.	51
4.3	On the left side of the table, we show the average Battacharyya distance between the approximated distributions $\hat{P}_{LDA}(\mathcal{P} \mathcal{U})$, $\hat{P}_{Y^2-LDA}(\mathcal{P} \mathcal{U})$ and the estimated distribution $P(\mathcal{P} \mathcal{U})$. We assume the maximum likelihood estimation of $P(\mathcal{P} \mathcal{U})$ as ground truth. On the right side, we display the Battacharyya distance between the approximated distributions $\hat{P}_{LDA}(\mathcal{U} \mathcal{P})$, $\hat{P}_{Y^2-LDA}(\mathcal{U} \mathcal{P})$ and the estimated distribution $P(\mathcal{U} \mathcal{P})$. We assume the maximum likelihood estimation of $P(\mathcal{U} \mathcal{P})$ as ground truth.	58

Chapter 1

Introduction

1.1 Background

Latent Dirichlet Allocation (LDA) is a hierarchical bayesian model of a discrete pair of variables such as document and words or user and places. LDA defines one of the variables as a collection of instances of the other variable [9]. For example, in the text data, the documents are defined to be a collection of words. This definition is modeled by the hierarchical representation of the LDA. The top of the hierarchy models the hyper-parameters at the corpus level, the middle level models the parameters of the documents at the document-level and the words are represented at the bottom of the hierarchy. This hierarchical structure facilitates modeling the probability of observing a word given the set of documents, but it hinders the calculation of the probability of observing a document given the set of words. This is a disadvantage because the latter probability represents useful information for applications in document recommendation and information retrieval.

Alternatively, we could switch the role of the variables. That is, we could define the words as the collection of documents in which the word appear. To differentiate this approach from the aforementioned use of LDA, we label this model as *Dual LDA*. The Dual LDA models the hyper-parameters at the top of the hierarchy, the parameters of a word at the

Scientific Field	Variables	LDA Representation	Dual LDA Representation
Text Mining	Documents	A document is represented by the set of words written in it.	A word is represented by the set of documents where the word has been written.
	Words		
Computer Vision	Images	An image is represented by a set of features.	A feature is represented by the set of images where the feature is observed.
	Features		
Bioinformatics	Genes	A gene is represented by the biological samples where the gene is active.	A biological Sample is represented by the set Genes that are active in the sample.
	Samples		
Human Mobility	Users	A user is represented by the places that he visits	A place is represented by the set of users that visit the place.
	Places		

Table 1.1 LDA and Dual LDA representation in several scientific fields.

middle level and the documents at the bottom of the hierarchy. In contrast with the LDA model, The Dual LDA facilitates the calculation of the probability of observing a document given the set of words at the cost of hindering the calculation of the probability of observing a word given the set of documents. Obtaining a Dual LDA is possible because the cardinality between the variables is a many-to-many relationship. This is the case for the scientific fields where LDA has been applied. e.g. Text mining [53, 73, 48], Computer vision [61, 40, 34], Bioinformatics [3], and the Human Mobility field [21, 44]. In table 1.1 we show how LDA has been represented in these fields and how the Dual LDA would produce a different representation for the same pair of variables.

Without loss of generality, we will use users and places to illustrate the modeling approaches of an LDA and a Dual LDA. In figure 1.1 a), we use a bipartite graph to represent a sample check-in dataset. Every vertex in the graph represents a user visiting one place. For simplicity, we omit the case where users visit the same place several times. In Figure 1.1 b) we present how LDA models the sample dataset by defining a user as the set of places visited by the user. Thus, we color one set to represent one user. In this approach to modeling, we

can group the places by measuring the co-occurrence of places visited among all users. For example, the places p_1 and p_2 are both visited by the same users and so they have a high likelihood of being grouped together. These groups commonly emerge as topics. A formal definition of topics will be presented in section 3.2.1. We refer to LDA's modeling approach as an *individual-based* or *user-based* approach because we look at the places visited by each *user* to obtain groups of representative places. In figure 1.1 c), we present the modeling approach of the Dual LDA. In this model, each place is defined as the set of users that visited the place and so we use sets to represent the places in the dataset. In this case, the co-occurrence of users among the sets will emerge as a set of topics. For example, users u_1 and u_2 visited half of the places in the dataset and so they are likely to appear together in a topic. We label these topics as user-topics. We refer to the topics extracted by the Dual LDA as user-topics because they represent a grouping over the set of users. Naturally, we will refer to the topics extracted by the LDA as place-topics since they represent a grouping over the set of places. We label the Dual LDA's modeling approach as an *location-based* or *place-based* approach because we look at the users who visit each of the places with the objective of obtaining groups of representative users.

In both approaches, we use the same data to obtain complementary information [80]. Mutual reinforcement of location-based models and individual-based models has shown improvement over a single approach in non-probabilistic clustering in the user-place domain [1] and the document-word domain [14]. We expect that a mutual reinforcement of the LDA and its dual will yield better topic assignments.

1.2 Motivation

In order to design a model that benefits from the information extracted by the LDA and the dual LDA, we can design a new topic model including the extraction of the place-topics and the extraction of user-topics. In addition, we can model the dependency between the

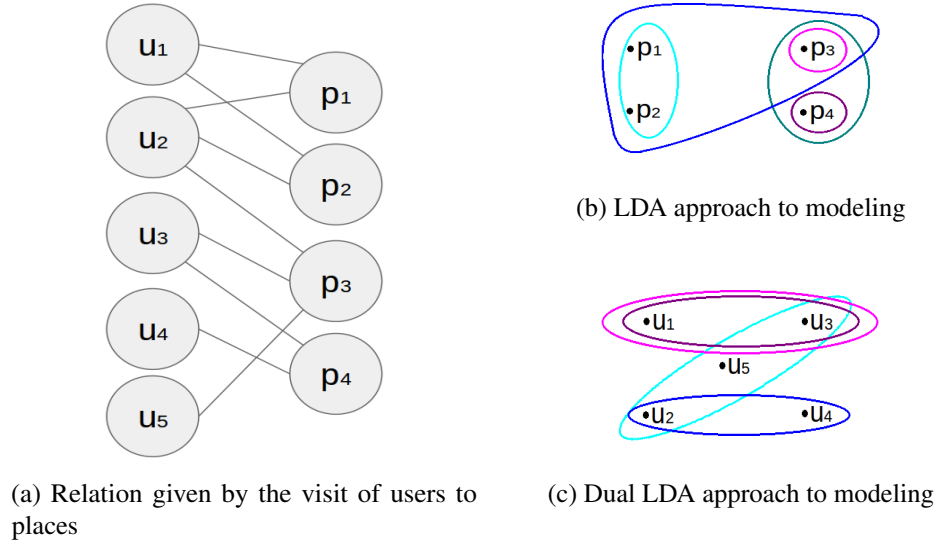


Fig. 1.1 Modeling approach of LDA and the Dual LDA.

user-topics and the place-topics. The information of the dependency between the topics brings the following benefits: First, we can improve the quality of LDA's samples obtained according to the generative process and measured by the log-likelihood or perplexity. Second, the information of the dependency between the topics represents the information about *Which groups of users visit which groups of places*. This is the same information that can be obtained by using a family of algorithms called *Biclustering Algorithms*. As consequence, our topic model can solve the biclustering problem. Third, we will have a novel approach to modeling mutual dependency between latent random variables.

1.3 Aim

The aim of our research is to design, and implement topic models that can be used to solve the following questions about the field of human mobility:

- Which POIs are most popular given a group of users?

- Which are the most common groups of users that visit a group of similar places?
- Which POI would a user be likely to visit next?
- When do users prefer visiting a specific type of places?
- What is relationship between urban planning areas of Singapore in terms of mobility?

Y^2 -LDA is able to solve the first two questions mentioned above, we will target on solving the other questions in our Future work as explained in section 5.2

1.4 Objectives

The steps that we will take to achieve our aims are:

1. To model, derive and implement a topic model for the generation of users and places.
2. To design and implement a probabilistic model for the automatic labeling of location data obtained from cellular towers.
3. To model, derive and implement a topic model for simulating human mobility.
4. Formulate and implement an architecture for the inference of a family of topic models using distributed computing.

In this report we explained with details how we accomplish the first objective. We explain how do we plan to accomplish objective two in section 5.1, objective three 5.2 and objective four in section 5.3

1.5 Challenges

Building a model that captures the interrelation between topics has two main challenges. First, we must model the mutual dependency between topics representing distributions over

different variable. Existing work has explored modeling the correlation between the same kind of topics [41, 7]. The proposed approaches use the parameters of the covariance matrix of the Multivariate Log-Normal distribution to store the covariance between each pair of topics and the variance of each topic. However, having different kinds of topics means having to compare two distributions over different variables. However, Since our objective is to extract relations between kinds of topics, we cannot use the covariance matrix encoding. As solution, we propose a new probability distribution created from transforming the categorical distribution into a distribution whose parameters are mutually dependent, every parameter identifies the relationship between a pair of topics.

As a second challenge, we must solve the intractability of the inference process for our model. Because our model uses the definition of a new probability distribution, we must derive analytically the posterior probability of observing the topics given the conditions described by our model. To solve this challenge we have found a mathematical relationship between our defined distribution and the categorical distribution. Using this relationship we can add a prior distribution to our model and we can also approximate the posterior probability of observing the topics given using the collapsed gibbs sampling formulation.

1.6 Organization

The rest of the report is organized as follows. In Chapter 3, we show the formal definitions of the Yin Yang Latent Dirichlet Allocation as well as the inference updating equations derived used Collapsed Gibbs Sampling. In addition, we demonstrate how LDA corresponds to a particular case of Y^2 -LDA; Chapter 4 explains the applications of Y^2 -LDA: a) Data Generation b) Individual topic extraction and c) Correlation of topics as a biclustering application d) Dimensionality reduction, in addition, we present experimental evaluation for each of the applications; Finally, Chapter 6 concludes our current work and discusses future research directions and plans.

Chapter 2

Literature Review

In this chapter, we review several existing proposals related to modeling human mobility using topic models. We divide these proposals into two categories. First, we introduce the proposals on how to model topic models and how these models have been applied in the Human mobility field. Then, we review the research work on how to solve the intractability of the inference of topic models. Understanding the existing inference algorithms is important because the inference process is the main bottleneck in topic modeling development.

By reviewing the existing work, we find that the existing topic models have improved Latent Dirichlet Allocation by either relaxing its assumptions or extending the model to include additional variables. Though some of these models capture the dependency between topics of the same kind (intra-topic dependency). None of these models have considered capturing the topic dependencies between different kind of topics (inter-topic dependency).

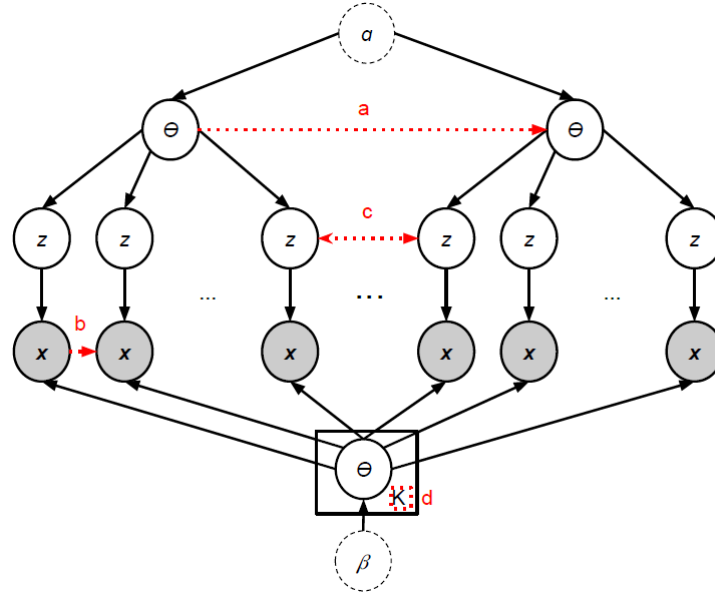


Fig. 2.1 Expanded graphical representation of the LDA model. The red arrows indicate the main independence assumptions about the LDA model. a) The order of Documents is irrelevant b) The order of words is irrelevant c) Topic dependency is only given by a Dirichlet prior distribution parameterized by α d) The number of topics must be given a priori.

2.1 Topic Models and its applications in Human Mobility Field

Topic models are commonly used to organize documents by using the set of words written in them [4, 63], some researchers studying human mobility behavior adopted some of these models with the objective of extracting information about the mobility of users within places. Topic models started with the introduction on the Latent Dirichlet Allocation model (LDA) [9], this model has been adapted to the human mobility field by some authors with the intention of extracting place-topics z by modeling a user Θ as a document and the set of places x visited by the user as the set of words in the document [21, 44, 70]. The topics are parameterized by a multinomial distribution ϕ with a Dirichlet distribution parameterized by β , Each document is represented by a multinomial distribution Θ with a different Dirichlet distribution parameterized by α . In figure 2.1, we show an extended model of LDA pointing

at the main assumptions made by the model. Reducing these assumptions has shown a positive impact in the application of topic models in several fields, we will now review how researchers have relaxed these assumptions and how these improved models are being used to model human mobility.

2.1.1 N-gram Topic Models

LDA is a bag of words model, this assumption implies that the ordering of the words inside a document is irrelevant [5]. Some researchers have addressed this problem by conditioning the generation of a word on the previous n -words. Hanna Wallach proposed a bigram topic model, where topics would be extracted as pairs of words [67]. This idea was extended by the Distant N-Gram Topic Model [72], where the authors focused in extending the bigram models by considering the dependencies between a word and the n -th previous words. The N-gram topic model tends to create an exponential number of parameters as N -increases. One possible solution is to use a Hidden Markov Model to model the dependency between words. This idea was developed by Griffiths et al, their composite model switches between LDA and a standard HMM so that the dependencies between words are governed by the HMM and the content of the words is governed by the LDA [29]. An alternative modeling relaxing the bag of words assumption is the Hidden Topic Markov Model (HTMM) [30]. HTMM considers the relationship between the topic assigned to the current word and the topics assigned to several previous words instead of considering the relationship between the current word and the previous word. In the field of human mobility, N-gram topic models have made possible the extraction of place patterns by considering the relationship between the current place and the n -th previous places [20, 19].

2.1.2 Bayesian non-parametric topic models

LDA assumes that the number of topics must be known a priori. This assumption has been relaxed with the introduction of the hierarchical Dirichlet process (HDP) that allows the extraction of the number of topics of an LDA model without a explicit specification [6]. HDP operates by assuming that there exists a finite number of groups sharing the same set of mixture components with different mixing proportion. To describe the mixing proportions of each group, one probability measure is assigned to each of the groups using a Dirichlet Process. The base distribution shared as a parameter among the Dirichlet Processes cannot control the trade-off between the desired property of sharing clusters among groups and allowing all possible mixing combinations of clusters in a given group. As solution, Teh et al considered using a prior over this base distribution. Because of the benefits of using conjugate priors, the authors propose using another Dirichlet Process as prior distribution, hence the name of the Hierarchical Dirichlet Process [65]. The number of topics is modeled as infinite and the relevance of each topic is represented by the sampling from Dirichlet process. One sample of the Dirichlet process will yield a discrete distribution whose support has finite dimensionality, since every dimension represents a topic, the extracted number of topics will be finite as well. The Dirichlet process has several construction definitions being the stick-breaking construction the simplest known [36].

Several human mobility models have been based on the [49, 50] Some human mobility models have also used Dirichlet Processes as to mine mobility patterns. J.Joseph et al, proposed a bayesian non parametric approach to model mobility patterns justified by the fact that Dirichlet process extract overlapping routes where Markov models will normally get confused [37], however they did not use topics in their work [50].

2.1.3 Correlated Topic Models

Another assumption of LDA is the near independence among the topics, the Dirichlet hyper prior used in LDA as prior knowledge about the topic distributions can always be replaced by one independent Gamma distribution per topic. The Correlated topic model [7] has been designed to relax this assumption by changing the Dirichlet hyper-prior for a Multivariate Log-normal hyper-prior, this distribution models the correlation between the topics by using its parametrization given by a covariance matrix and a mean vector. Every parameter of the covariance matrix is used to encode the correlation between each pair of topics, while the mean vector represents the importance of the topics. The correlated topic model has been applied to the modeling of human mobility behavior by modeling a document as a geographical region and the set of words as the set of places, the topics found correspond to geographically related place topics [77].

As an alternative for modeling the correlation between topics, Li et al proposed the Pachinko Allocation Topic Model (PAM) which uses a directed acyclic graph to represent the topics with each non-leaf node and the words with the leaf nodes [41]. Hence, The relationship between all nodes represents the distributions of topics over topics and distributions of topics over words. PAM is a generalization of LDA since a graph where all nodes are connected to only one node will correspond to the standard LDA. The correlation between topics is given by the hierarchical relations existent between topics. PAM has an advantage over the correlated topic model given by the possibility of a sparse parametrization of the topic intra-correlation. In the Correlated topic models the correlation between all topics must be parameterized.

2.1.4 Temporal Dynamics of the Topic Models

A final assumption of LDA is the exchangeability of the documents. i.e. The ordering of documents is irrelevant when doing the inference of the topics and the topics will not exhibit

changes in time. A set of topic models represent the temporal dynamics of the data by modeling the latent topics as changing throughout time.

Blei and Lafferty presented the Dynamic Topic Model (DTM) that uses a Kalman filter to model the temporal alignment among topics across collections of documents [8]. As consequence, DTM uses the Markov assumption that the future state of the topics only depends on the inference values of the present state of the topics. Several models have proposed different ways to improve DTM. In [74], the authors proposed the Dynamic Mixture Model to force temporal dependence between documents rather than the temporal dependence between collections of documents, this allows a finer modeling of the evolution of the per document topic distributions. Similar approaches are the Sequential LDA (SeqLDA) [15] and the Evolutionary HDP (EvoHDP) [79]. Both models propose changing the Dirichlet prior of DMM. SeqLDA uses a Pitman-Yor process to improve of the space and computational complexity of the inference process, while EvoHDP uses a Hierarchical Dirichlet Process to automatically extract the number of topics including emergence and disappearance.

Also, In Topics over time model, the authors relaxes the Markov assumption of DTM by treating time as an observed continuous variable such as the topics [71]. This helps avoiding a Markov model's risk of inappropriately dividing a topic in two when there is a brief gap in its appearance. However, with all topics are drawn from one distribution, TOT and DMM are not appropriate for identifying topics that change in time.

The Dynamic Topic Model remains as the only topic model that models topics and its changes of topics over time. However DTM, also named as the Discrete Time Dynamic Topic Model (dDTM) splits the data into groups based on temporal discretization and thus, it requires to choose a discretization value. Because of computational limitations, the discretization value is chosen based on the trade off between accuracy and computational complexity, nonetheless choosing the right discretization value should be a decision based on the data. To solve this issue, Wong and Blei proposed the Continuous Time Dynamic Topic

Model (cDTM) which corresponds to the infinitesimal discretization of the dDTM. cDTM uses a Brownian motion to model the evolution of the topics in time [69].

2.1.5 Extending LDA with additional variables

Aside from solving the assumptions of LDA, other researchers have studied extending an LDA model by including additional variables in order to be able to use topic models for particular supervised learning applications such place recommendation or next place prediction. For geographic-textual data, researchers have studied different variables as user, geographical and functional region, time and text [76, 2, 51, 78]. For check-in data researchers have focused on users, time, item preference, geographical and functional regions [42]. All the mentioned models represent a user as a document, hence we recognize these models as individual-based approaches.

Other extensions of the LDA model have allowed the exploration of non-personalized models. Non-personalized models are important, as they provide essential information in recommendation tasks [10]. The author topic model (ATM) is an extension of LDA where each word in a document is assumed to be written by a specific user, and a document can be written by several authors [58]. This model has been used in the human mobility behavior field by modeling the document as time of the day and the authors as the users. As result, ATM model can extract place-topics of interest for a set of users in a particular time of the day [18]. The author topic model has also been used for aggregating locations instead of users in the place-recommendation task [43].

2.2 Improving the Inference of Topic Models

Modeling topic models requires more than understanding the relationships between the problem variables and relaxing assumptions to create more realistic models. The computational

complexity of LDA-based topic models such all the ones mentioned in section 2.2.2 has been proved to be NP-hard [62]. Additionally, different inference algorithms have different advantages and there is no solution on how to generalize the inference process in topic modeling. As consequence, deep understanding of inference methods is necessary in order to design and apply new topic models to large data-sets. In this section, We review the approximation algorithms that have been proposed in the literature to solve the intractability problem of topic models. We classify these algorithms in two categories: The model-based inference methods and the family-based inference methods.

2.2.1 Model-based Inference methods

Model-based methods correspond to those techniques that must be tailored for each particular topic model. To solve the intractability of topic models, Hoffman et al proposed using an Expectation-Maximization algorithm (EM) [33] for an LDA model with no priors also known as probabilistic Latent Semantic Analysis. The algorithm aims to find the corresponding values of the latent variables and the given parameters of the model based on the conditional dependencies implied by the model. The same approach works for LDA using Dirichlet priors [52]. A pragmatic improvement over EM has been achieved by using the variational EM framework [9] in which a new simplified tractable model is used to find the parameters of the intractable model. The simplified model makes the assumption that the latent variables and the parameters are mutually independent. Rather than solving the simplified model, the variational bayesian inference (VB) method solves the simplified model whose Kullback-Leibler Divergence to the original model is minimized. This problem is equivalent to the problem of maximizing the free variational energy which is a known framework for solving the inference of probabilistic models. The equivalence between the problems can be shown using Jensen's Inequality [9] or by manipulation of the equations [22].

In [28], Griffiths and Steyvers proposed an MCMC approach based on collapsing or integrating out the parameters of the model which is analytically feasible since the prior distributions are conjugate with the distributions over the variables of the model, the model was labeled Collapsed Gibbs Sampling (CGS). The advantage of CGS over VB is the existence of theoretical bounds over the convergence of the inference process, however CGS requires conjugate priors in order to collapse the parameters of the model. Teh et al proposed Collapsed Variational Bayesian Inference [?]. CVB is a variational bayesian algorithm that uses the Collapsed Gibbs Sampling approach of collapsing the parameters by integrating them out from the joint probability distribution of the model. So instead of approximating the model with a simplified version of the model, CVB proposes collapsing the parameters of the model and then solving the collapsed version of the model using variational inference. As result CVB is able relax the independence assumptions between the parameters used in VB, though the latent variables are still assumed to be mutually independent.

2.2.2 Family-based Inference methods

More recent algorithms aim to develop inference techniques for a family of probabilistic models. Blei et al proposed Stochastic Variational Inference [32], an algorithm to approximate the posterior distribution using subsets of the data rather than the entire dataset. Their method works for a family of models with i) Conjugate priors ii) The generative process follows a conditional chain. e.g. In LDA, Θ is condition by α , Φ is condition by β and w is conditioned by Φ_z iii) the distributions assigned to the random variables must belong to the exponential family. The works by algorithm can be applied to any model a complete conditional is the set of conditional distributions of a single hidden variable in terms of all other hidden and non-hidden variables in the model.

The algorithm works by recognizing that all models that follow conditions i,ii, and iii can be expressed in terms of a simple topic model as displayed in Figure 2.2a. The simple topic

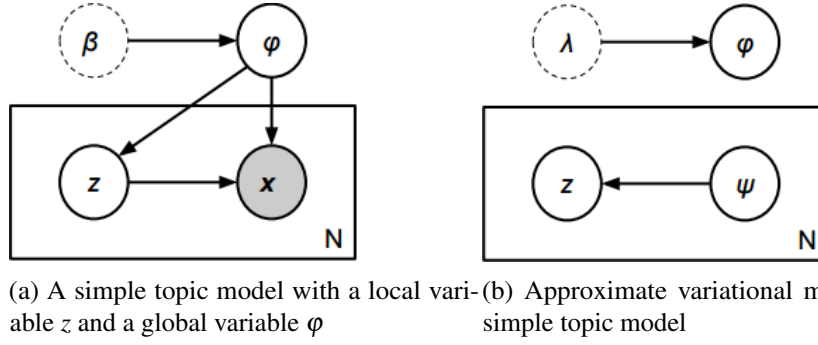


Fig. 2.2 A simple topic model and an approximate topic model assuming independences between all parameters in the model

model p is composed of a local latent variable α linked to every instance of the data and global latent variables ϕ governing all instances of the local latent variables. Being this a Bayesian formulation, β corresponds to the given hyperparameters of ϕ . Solving the simple model will be enough to solve a family of topic models. The authors of this work have decided to solve the inference of the simple topic model using Variational inference. The approximate model q is given in figure 2.2b. Ideally we want the model \hat{q} that is more similar to the original model given by p , that is $\hat{q} = \operatorname{argmin}_q KL(q(z, \phi) || p(z, \phi | x))$. Using Jensen's inequality, the authors show how this problem is equivalent to maximizing the evidence lower bound (ELBO). i.e. Maximizing the right hand side of equation 2.2

$$\log p(x) \geq E_q[\log p(x, z, \phi)] - E_q[\log q(z, \phi)] = L(q) \quad (2.1)$$

Because this equation depends on p , solving the inference problem will depend on p having conjugate priors and using distributions in the exponential family (conditions i and iii). This model can be applied to models such Bayesian mixture models [26], hidden Markov models [27, 24], Kalman filters [38, 23], probabilistic factor analysis/matrix factorization models [55, 13] and some Bayesian nonparametric mixture models [17, 66].

More recently, Black Box Variational Inference [57] was introduced as a generalization of Stochastic Variational Inference by removing conditions i) and iii). The relaxation is

possible since the objective function $L(q)$ is solved using Stochastic Optimization and Monte Carlo estimation. Stochastic optimization is used to find the maximum value of $L(q)$ using noisy estimates of $\nabla_{\varphi}L$ in a framework provided by the Robbins-Monro. The framework guarantees convergence if the following iterative process is followed:

$$q(\varphi)_{t+1} \leftarrow q(\varphi)_t + \rho h_t(q) \quad (2.2)$$

where ρ corresponds to the learning rate and $h_t(q)$ is a distribution whose expectation approximates $\nabla_{\varphi}L$. The authors use Monte Carlo estimation to approximate $\nabla_{\varphi}L$ as follows $h_t(q) = \frac{1}{S} \sum_{s=1}^S \nabla_{\varphi} \log q(z_s|\varphi) (\log p(x, z_s) - \log q(z_s|\varphi))$ where z_s corresponds to a sample from the variational distribution as given by $z_s \sim q(z|\varphi)$

2.3 Biclustering

Biclustering algorithms have been used to cluster simultaneously two variables [31]. While it is understandable that clustering one variable is beneficial for organizing or summarizing information, it is not intuitive why do we need to extract Biclusters. To see why extracting biclusters can be useful, let us consider a simple example:

Tom is planning a party for his new four-room flat. Each room has a separate sound system, so he wants to play a movie in each room. As a host, Bob wants everyone to enjoy the movie. Therefore, he needs to distribute movies and guests to each room in order to ensure that each guest watch their favorite movie. Tom has invited forty guests, and he owns twenty movies. He sends out a survey to each guest asking to rate from 1 to 5 each movie. After receiving the guest's responses, Tom collects the data into a 40×20 matrix A, where $A_{i,j}$ stores the ranking of j – th movie given by the i – th guest.

A Biclustering algorithm would take as an input the matrix A and it would output a list of biclusters where each bicluster is defined by a subset of the guests and a subset of the movies.

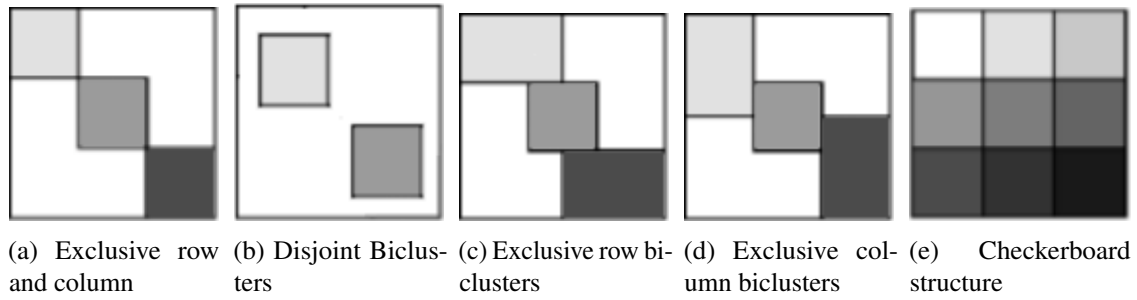


Fig. 2.3 Bicluster Structures. Figure modified from [46]

After biclustering, the rows and columns of the data matrix may be reordered to show the assignment of guests and movies to the rooms in the flat.

Biclustering can be applied to any pair of variables whose type of relationship is described as *many-to-many*. For example, users and movies, documents and words, or users and webpages. We can categorize the families of algorithms by structure and by type [46].

2.3.1 Biclustering Structures

The Biclustering structure represents the shape of the biclusters that can be extracted from a matrix. The different bicluster structures are shown in figure 2.3. Their description is given below. :

Exclusive row and column Every user and every place is assigned to exactly one bicluster.

Disjoint Biclusters Every user and every place is assigned to at most one bicluster.

Exclusive row biclusters Every user is assigned to one bicluster, but every place is assigned to at least one bicluster.

Exclusive columns biclusters Every user is assigned to at least one bicluster, but every place is assigned to one bicluster.

Checkerboard structure Every user, place pair is assigned to exactly one bicluster.

2.3.2 Biclustering Types

The biclustering type refers to the relationship in the matrix that forces the rows and the columns to form a bicluster. There are two main types of Bicluster: Biclusters with constant values and biclusters with coherent values. The first type correspond to those biclusters whose elements have the same or a similar value. There exists several subtypes. For example, biclusters with same values on the rows or biclusters with same values on its columns.

The second type corresponds to those biclusters whose elements have some coherence. For example, either the rows, or the columns are dependent according to some mathematical condition or the values in the bicluster follow a known distribution. In particular, when the coherence is defined to extract biclusters whose rows and columns have a high-frequency of occurring in the dataset, the biclusters obtained corresponds to the bicliques of the dataset.

2.3.3 Related biclustering Algorithms

In this section, we briefly review some of the biclustering algorithms that best performed on tested data with varying conditions, including varying noise and varying numbers of biclusters [16, 46].

Cheng and Church

Cheng and Church proposed the first biclustering algorithm. Their algorithm is famous for having the theoretical guarantee of finding the bicluster where its rows and columns are shifted versions of each other. i.e. The bicluster is represented by a matrix with rank equals to one [12]. Their algorithm is the most common baseline among biclustering algorithms [46]. The algorithm operates by assuming that all the data corresponds to a bicluster, and then iteratively it removes the row or the column with lowest similarity score to the rest of the bicluster. At each iteration, the algorithm evaluates the candidate bicluster using a function score and when the bicluster reaches its smallest size (last iteration); it retrieves

the bicluster with highest score. If a second bicluster needs to be found, the data from the first algorithm is replaced by noisy data and the algorithm is restarted. In Short, Cheng and Church biclustering algorithm extracts a bicluster at a time.

xMOTIF

xMOTIF [54] aims to find subset of rows that occur simultaneously across the columns under a set of conditions. e.g. a linear order across the columns. xMOTIF retrieves The largest bicluster that contains the maximum number of conserved rows. Similarly to Cheng and Church algorithm, xMOTIF extracts one bicluster at a time.

QUBIC

While the methods explained above identify patterns under parameterized conditions such a linear scaling among the rows or columns. Quantitative Biclustering (QUBIC) attempts to find similar patterns based on positive and negative correlation, so that the algorithm is less sensitive to outliers in the data [45]. The algorithm uses a combinatorial technique that has demonstrated an improvement over other combinatorial techniques such SAMBA [64], ISA [35], Bimax [56]. The discovery of biclusters follows the same procedure as [12] and [54], that is, biclusters are discovered one at a time.

Spectral Bipartite Graph

This method proposes a novel approach to solve the biclustering problem by transforming the matrix model into a bipartite graph where every row is now part of one set of elements, the columns are another other set of elements and the values in the matrix corresponds to the weight of the relationship between the elements in both sets. The cliques in the graph will correspond to biclusters. However, because in most cases there are not cliques in the

Algorithm	Type	Structure	Discovery	Approach
Block Clustering	Constant	a	One at a time	Divide and Conquer
Bimax [56]	Coherent Values	b	One at a time	Greedy
FLOC [47]	Coherent Values	a,b	Simultaneous	Greedy
PLAID [39]	Coherent Values	a,b	One at a time	Probabilistic
CTWC [11]	Coherent Values	b	One at a time	Combinatorial
ITWC [25]	Coherent Values	b	One at a time	Combinatorial
Spectral [14]	Coherent Values	a	Simultaneous	Optimization
QUBIC [45]	Coherent Values	a,b,c	One at a time	Optimization
xMotif [54]	Coherent Values	a,b,c	Simultaneous	Greedy
SAMBA [64]	Coherent Evolution	a,b,c	One at a time	Combinatorial
Bayesian [60]	Coherent Values	e	Simultaneous	Probabilistic
Cheng [12]	Constant	a,b,c	One at a time	Iterative
Mondrian [59]	Coherent Values	e	Simultaneous	Probabilistic

graph, Dhillon proposes an optimization function taking into account the weight between the vertices and the sizes of the clique to extract groups of vertices that better fit a clique.

Chapter 3

Yin Yang Latent Dirichlet Allocation

In section 1.1 we have introduced the modeling approaches of the LDA and the Dual LDA. In particular, We have shown how for the same data, LDA extracts place-topics in an individual-based model while the Dual LDA extract place-topics in a location-based model. In this chapter we present a model that extracts user-topics, place-topics and the dependency between both types of topics. We explain how we design the model, the generative process and the inference equation that describes how we find the parameters of the model given a training dataset. Once we present the inference process of our model, we can present the proof that our model corresponds to a generalization of LDA. Finally, we will explain why the model is able to extract the information corresponding to *Which groups of users visit which groups of places?*.

3.1 Notation and Terminology

3.1.1 Users and Places.

Without loss of generality, we will use users and places to illustrate the modeling approach of our model. We model a set of users U indexed by $u \in 1 : |U|$, a set of places P indexed by

$p \in 1 : |P|$, and a relation V indexed by $v \doteq \{u, p\} \in 1 : |V|$ representing every time an user visit a place.

3.1.2 Mathematical Notation

For the mathematical definitions and derivations in the document, we will use the following notation:

- \mathbb{R}^m denotes the m-dimensional vector space of real numbers.
- $\mathbb{N}_{[m:n]}$ denotes the set of natural numbers ranging from m to n.
- $\mathbb{1}$ denotes the indicator function

Other terminology such the parameter names and representations will be defined as we explain our model.

3.2 Designing Yin Yang Latent Dirichlet Allocation Model

In this subsection, we present how we designed Yin Yang Latent Dirichlet Allocation model based on the constructions of the Categorical Mixture Model (CMM) and the Latent Dirichlet Allocation model (LDA). We start by describing how CMM uses a probabilistic framework to extract topics. Then, we show how LDA uses the CMM to extract the topics for each document in a corpus. In the next section, we will explain how we use two CMM models to extract both user-topics and place-topics in a single model.

3.2.1 Categorical Mixture Model

Mixture models represents the observations of one random variable by combining a finite number of probability distributions. In the case of the CMM, the objective is to represent

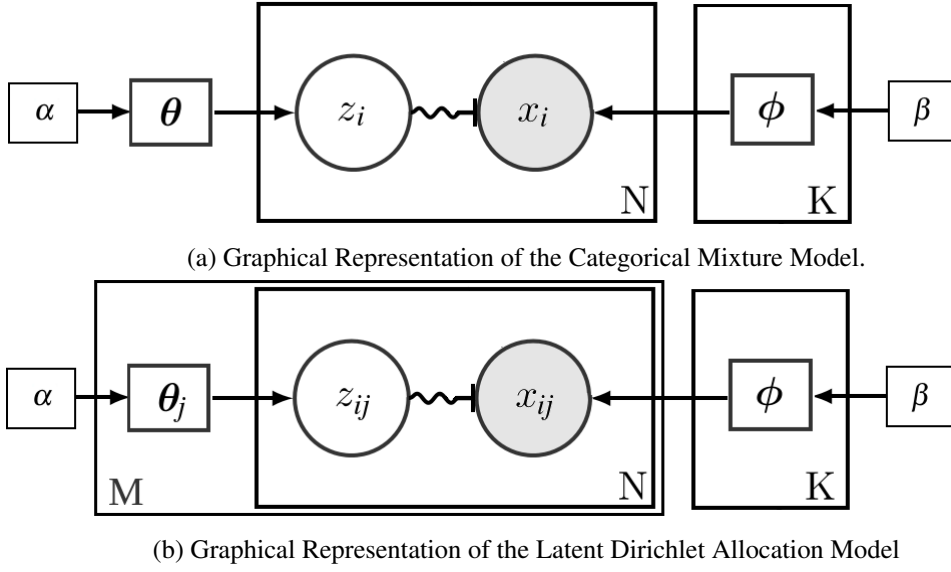


Fig. 3.1 Information extracted by an LDA (left) and a Dual LDA (right). The width of the arrow represent the preference of a topic, the ovals represents the topics.

how N observations x_i are generated by mixing K categorical distributions parameterized by ϕ where the mixing proportions are given by another categorical distribution with parameters θ . The topics in this model correspond to the information extracted by the distributions parameterized by ϕ . The term *topic* originated when this model was applied to a set of words because in that scenario, the parameters ϕ represent distributions over the vocabulary of words. We depict this information in figure 3.1a.

In the Mobility field, when we use the CMM to obtain a mixture of distributions over the set of places, we will refer to information described by the parameters ϕ as *place-topics*. Similarly, when we use the CMM to obtain a mixture of distributions over the set of users, we refer to the information encoded by the parameters ϕ as *user-topics*.

3.2.2 Latent Dirichlet Allocation

The Latent Dirichlet Allocation model extends the Categorical Mixture Model by dividing the words into M different documents. Instead of parameterizing the entire set of words with

one categorical distribution, it parameterizes every document with a categorical distribution with parameters Θ_j . We depict this information in figure 3.1b.

In the mobility field we can use an LDA to split the set of places into $|U|$ different users. As consequence, every user is defined as the places that he visit. In this setting, LDA's topics represent distributions over the set of places and therefore we obtain a explicit grouping of the set of places as given by the topics. In addition LDA gives a direct representation of the distribution of places visited by any user u :

$$P(p|u) = P(p|\Theta_u) = \sum_{Z_p=1}^{K_p} P(p|\Phi_{Z_p})P(Z_p|\Theta_u) \quad (3.1)$$

However, LDA hinders the representation of the distribution of users visiting any of the places p ($P(u|p)$). In addition, the model does not extract explicit information about grouping the set of users.

The above-mentioned limitations of an LDA are solved using a Dual LDA. The Dual LDA is also an LDA model that interchanges the role of the users and places. In the Dual LDA, instead of splitting the set of places, we split the set of users into $|P|$ different places. As consequence, every place is defined as the users that visited the place. In this case, the topics represent distributions over the set of users and the distribution of users visiting any of the places p ($P(u|\Theta_p)$) is directly represented by the model. A graphical representation of the information extracted by an LDA and its Dual LDA is presented in figure 3.2.

In summary, we can express the relationship between users and places into two approaches.

i) LDA or individual based-approach:

- A user is defined by the places that he visited.
- We obtain topics represented by distributions over the set of places.

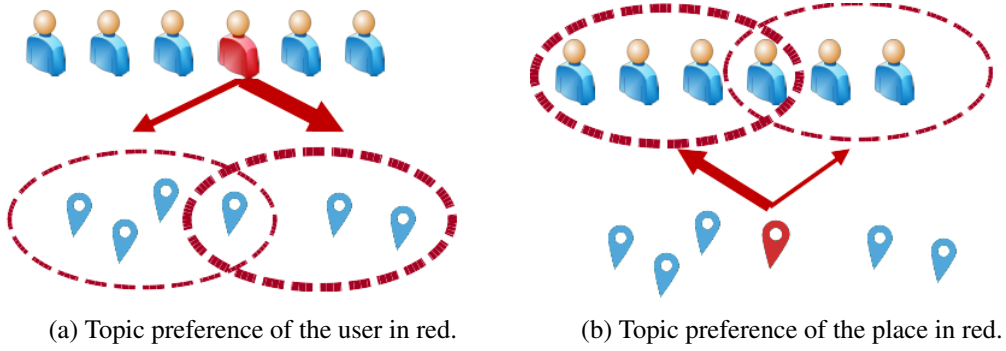


Fig. 3.2 Information extracted by an LDA (left) and a Dual LDA (right). The width of the arrow represent the preference of a topic, the ovals represents the topics.

- We can represent the distribution of the places for a given user.

ii) Dual LDA or location based-approach:

- A place is defined by the users who visit the place.
- We obtain topics represented by distributions over the set of users.
- We can represent the distribution of the users for a given place.

3.3 Yin Yang Latent Dirichlet Allocation (Y²-LDA)

3.3.1 Intuitions

We propose a new way of modeling the relationship between the two variables. Instead of describing the distribution of places given a user ($P(p|u)$), or the distribution of users given a place ($P(u|p)$). We propose a description of the joint distribution of the two variables ($P(p, u)$). If we obtain a description of the joint distribution, we can quickly derive the conditional distributions modeled by the LDA and the dual LDA. In our model, we want to extract place-topics as done by the LDA and we want to extract user-topics as done by the Dual LDA. As a novelty, we also extract the mutual dependency between the topics. This

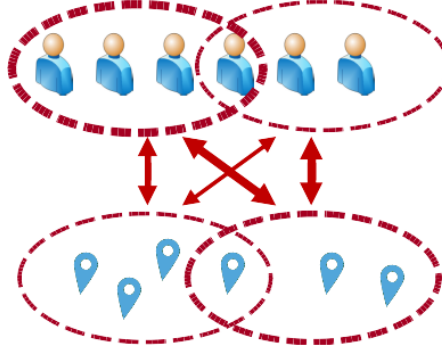


Fig. 3.3 Information extracted by Y²-LDA. The ovals represents the topics and the width of the arrows represents the strength of the relationship between two topics.

dependency will provide the information about which user-topics are related with which place-topics (Figure 3.3). In conclusion, we obtain the following intuitions:

1. Group the users into user-topics.
2. Group the places into place-topics.
3. Extract the mutual dependency between the place-topics and the user-topics.

3.3.2 Using CMM to implement intuitions 1 and 2

We construct the model by implementing the three intuitions. For the implementation of intuitions 1 and 2, we use the Categorical Mixture model presented in subsection 3.2.1. We know that the CMM we can extracts topics of a discrete variable. Therefore, we use one CMM to extract user-topics and another CMM for extracting place-topics. A graphical representation of the model imple-

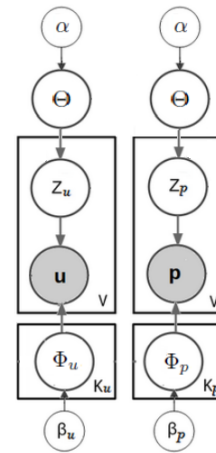


Fig. 3.4 Implementation of intuitions 1 and 2.

menting only intuitions 1 and 2 is presented in figure 3.4.

3.3.3 Defining the Bivariate Categorical Distribution to implement intuition 3

Describing the dependency between two kinds of topics presents a challenge because it has not been studied by Topic Models. We address this challenge by defining a probability distribution to measure the probability of observing a user-topic and a place-topic. Because we require a probability distribution to describe the simultaneous generation of two variables, we will define a Bivariate Probability distribution. Additionally, it would be ideal to use a categorical distribution because categorical distributions is typically used to capture the mixing components of the topics. With a categorical distribution, we can use the parameter $\Theta_{i,j}$ to capture the relationship from the i – th user-topic to the j – th place-topic. Given these conditions, we define the *Bivariate Categorical Distribution*.

In order to define a probability distribution, we need to specify the support, parameters and probability mass function of the distribution. The support of a probability distribution represents the set of possible values of a random variable having that distribution, the parameters corresponds to the set of variables that describe the probability mass function and the probability mass function represents how we assign a probability to every element in the support of the distribution. Using this components, we present the following description of the *Bivariate Categorical Distribution*:

Support The distribution has a two dimensional support since we require simultaneous sampling of two random variables.

Parameters The distribution uses $K_u \cdot K_p$ parameters. Each parameter represents the relationship between one place-topic and one user-topic.

Probability Mass Function $f_{\Theta}(\vec{x}) = \sum_i^{K_u} \sum_j^{K_p} \mathbb{1}[x_0 = i \wedge x_1 = j] \Theta_{i,j}$ This function represents the probability of observing every user-topic, place-topic pair (\vec{x}) using a parameter.

Now we present the formal definition of the Bivariate Categorical Distribution:

Definition 1. A Bivariate Categorical Distribution $Categorical^2(\Theta, K_u, K_p)$ is a discrete probability distribution parameterized by $\Theta \in \mathbb{R}_{[0,1]}^{K_u \cdot K_p}$ and $K_u, K_p \in \mathbb{N}_{>0}$. The support of the distribution is $X \in \mathbb{N}_{[1:K_u]} \times \mathbb{N}_{[1:K_p]}$. The probability mass function is given by $f_{\Theta}(\vec{x}) = \sum_i^{K_u} \sum_j^{K_p} \mathbb{1}[x_0 = i \wedge x_1 = j] \Theta_{i,j}$

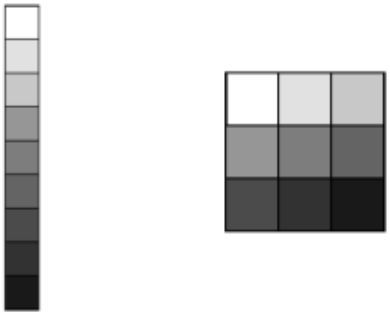
Proof. We proof that the Bivariate Categorical Distribution is a valid distribution by showing how we can represent the Bivariate Categorical Distribution in terms of the Categorical distribution:

Given a random variable $x \sim Categorical^2(\Theta, K_u, K_p)$. We need to use a mapping g so that $y = g(x) \sim Categorical(\Theta)$. This mapping is given by $g : \mathbb{N}_{[1:k]} \rightarrow \times \mathbb{N}_{[1:K_u]} \times \mathbb{N}_{[1:K_p]}$ where $g(x) = (x \div K_p, x \bmod K_p)$. We present in figure 3.5 a visualization of the mapping of g . Because g is a bijection, we are guaranteed that we can always represent a Bivariate Categorical Distribution in terms of the Categorical Distribution and viceversa.

■

3.3.4 Model Representation

Connecting the implementation of the above-mentioned intuitions, we obtain the Yin Yang Latent Dirichlet Allocation (Y²-LDA) model. Y²-LDA assumes a known dimensionality K_u of user-topics and K_p of place-topics. In figure 3.6, we present the graphical model that encodes the dependencies between the random variables used in Y²-LDA. We also present the formal definition of all the variables used by our model in table 3.6.

$$g : \mathbb{N}_{[1:k]} \rightarrow \mathbb{N}_{[1:K_u]} \times \mathbb{N}_{[1:K_p]}$$


$$g(x) = (x \div K_p, x \bmod K_p)$$

Fig. 3.5 Function used to define the *Categorical*² distribution. The visualization indicates the mapping from the parameters to the support set.

From the structure of the model, we observe that we can obtain the joint probability of observing the users and places as given by:

$$P(u, p) = \sum_{Z_u=1}^{K_u} \sum_{Z_p=1}^{K_p} P(u|\Phi_{Z_u})P(p|\Phi_{Z_p})P(Z_u, Z_p|\Theta) \quad (3.2)$$

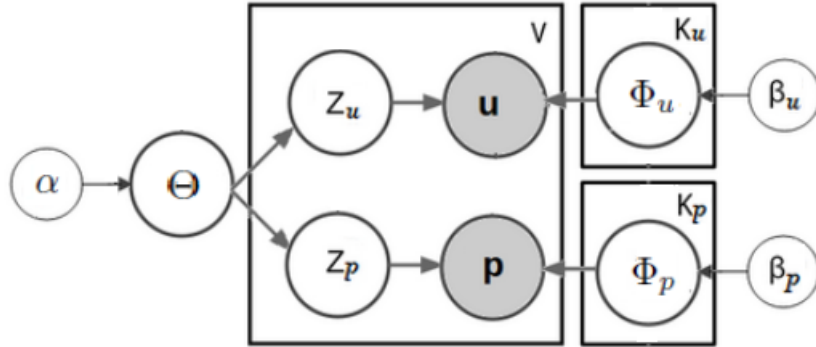


Fig. 3.6 Yin Yang Latent Dirichlet Allocation graphical model

Variable	Dimension	Description	Distribution
v	1	Index to a visit record	NA
V	1	Number of places	NA
K_p	1	Number of place-topics	NA
K_u	1	Number of user-topics	NA
U	1	Number of Users	NA
P	1	Number of Places	NA
Φ_p	$P \times K_p$	Place per place-topic distribution	$Dirichlet(\beta_p)$
β_p	$P \times 1$	Hyper-parameter of Φ_p	NA
p_v	1	Observed place in the v -th visit record	$Categorical(\Phi_p, K_p)$
Φ_u	$U \times K_u$	User per user-topic distribution	$Dirichlet(\beta_u)$
β_u	$U \times 1$	Hyper-parameter of Φ_u	NA
u_v	1	Observed user in the v -th visit record	$Categorical(\Phi_u, K_u)$
θ	$K_u \times K_p$	Joint distribution of user-topic and place-topics	$Dirichlet(\alpha)$
α	$K_u \times K_p$	Hyper-parameter of θ	NA
$z_{u(v)}, z_{p(v)}$	1,1	User-topic and Place-topic assigned to the v -th visit record	$Categorical^2(\Theta, K_p, K_u)$

Table 3.1 Variable definitions for the Yin Yang Latent Dirichlet Allocation model.

3.4 Generative Process

The generative process uses the intuitive idea of LDA that users are drawn from user-topics. From the Dual LDA, it uses the intuitive idea that places are drawn from place-topics. To connect both models, Y^2 -LDA uses an extra latent variable (Θ) from which a place-topic and a user-topic are sampled simultaneously. This latent variable contains the information of the relationship between every user-topic and every place-topic. In algorithm 1, we present the algorithmic description of the generative process.

Algorithm 1 Y^2 -LDA Generative Process

```

1: procedure YY-LDASAMPLE
2:   Draw a distribution over topics  $\Theta \sim \text{Dirichlet}(\alpha)$ 
3:   for  $i$  in  $1 \dots K_u$  do
4:     Draw a distribution over the users  $\Phi_i \sim \text{Dirichlet}(\beta_u)$ 
5:   end for
6:   for  $j$  in  $1..K_p$  do
7:     Draw a distribution over the places  $\Phi_j \sim \text{Dirichlet}(\beta_p)$ 
8:   end for
9:   for  $v$  in  $1..|V|$  do
10:    Draw simultaneously a user-topic and a place-topic  $z_{p(v)}, z_{u(v)} \sim \text{Categorical}^2(\Theta)$ 
11:    Draw a user  $u \sim \text{Categorical}(\Phi_{z_{u(v)}})$ 
12:    Draw a place  $p \sim \text{Categorical}(\Phi_{z_{p(v)}})$ 
13:   end for
14: end procedure

```

3.5 Inference Process

Another challenge of the Y^2 -LDA is the inference of the posterior distribution $P(z_u, z_p | u, p)$. This distribution is intractable to compute due to the fact that the topics are dependent on each other. Furthermore, applying commonly used inference techniques such Variational Inference or Collapsed Gibbs Sampling is not viable since the generative process that we

describe requires a simultaneous sampling. This type of sampling has not been studied in topic modeling.

Equation 3.3 correspond to the posterior distribution that we use to update the user-topic $z_{u(v)}$ and place-topic $z_{p(v)}$ assigned to the v -th visit record. For each visit record, we must sample from this distribution to update its topic assignments, the posterior distribution must be recomputed after each assignment. Updating all the records corresponds to one iteration. After several iterations, the posterior distribution will converge, and we have found the topic assignment for each record. All details of the posterior derivation can be found in section 3.6.

$$P(z_{u(v)}, z_{p(v)} | z_{u(-v)}, z_{p(-v)}, u, p, \alpha, \beta_u, \beta_p) \propto \frac{\mathcal{Z} \mathcal{P}(v) \mathcal{U}(v)}{\sum_{j=1}^{|P|} \mathcal{P}(j) \sum_{i=1}^{|U|} \mathcal{U}(i)} \quad (3.3)$$

$$\begin{aligned} \mathcal{Z} &= \sum_{m=1}^{|U|} \sum_{n=1}^{|P|} c(z_{p(v)}, z_{u(v)}, m, n)^{(-v)} + \alpha_{z_{u(v)}, z_{p(v)}} \\ \mathcal{P}(i) &= \sum_{y=1}^{K_p} \sum_{n=1}^{|P|} c(z_{p(v)}, y, p(i), n)^{(-v)} + \beta_{p(i)} \\ \mathcal{U}(i) &= \sum_{x=1}^{K_u} \sum_{m=1}^{|U|} c(x, z_{u(v)}, m, u(i))^{(-v)} + \beta_{u(i)} \\ c(x, y, m, n) &= \sum_{j=1}^{|V|} I(z_{u(j)} = x, z_{p(j)} = y, u(j) = m, p(j) = n) \end{aligned}$$

The inference equation depends on three types of counting. First, a normalized counting of the user assignment to the user-topics \mathcal{U} capturing user co-occurrence. Second, a normalized counting of the place assignment to the place-topics \mathcal{P} capturing place co-occurrence. And finally a counting over the current relationship between user-topics and place-topics \mathcal{Z} .

3.6 Derivation of the Inference Equation using Collapsed Gibbs Sampling

In this section we derive the posterior distribution used to assign the topics. The derivation includes two main sections. First, we apply collapsed gibbs sampling to derive analytically the posterior distribution given the conditions of the model. Second, we use the fact that every record is independent and identically distributed so that we can find the posterior derivation for each particular record in the dataset.

3.6.1 Posterior Derivation given the conditions of the model

First we write down the target distribution that we want to derive:

$$P(z_{u,v}, z_{p,v} | z_{u,-v}, z_{p,-v}, u, p, \alpha, \beta_u, \beta_p) \quad (3.4)$$

The idea is to find the updating topics z_p and z_u for a given visit record v , given the training data. The definition of the variables used in this derivation can be found in figure 3.6.

Now, we start by formulating the joint distribution over all the variables in the model.

$$P(u, p, z_u, z_p, \Phi_u, \Phi_p, \Theta | \alpha, \beta_u, \beta_p) = \quad (3.5)$$

To obtain the target distribution using collapsed gibbs sampling, we need to integrate out the unknown parameters Φ_u, Φ_p and Θ .

$$\int_{\Theta} \int_{\Phi_u} \int_{\Phi_p} P(u, p, z_u, z_p, \Phi_u, \Phi_p, \Theta | \alpha, \beta_u, \beta_p) d\Phi_p d\Phi_u d\Theta = \quad (3.6)$$

Next, we apply Y²-LDA's independence assumptions.

$$\int_{\Theta} \int_{\Phi_u} \int_{\Phi_p} P(\Theta | \alpha) P(z_p | \Theta) P(z_u | \Theta) P(p | \Phi_p) P(\Phi_p | \beta_p) P(u | \Phi_u) P(\Phi_u | \beta_u) d\Phi_p d\Phi_u d\Theta \quad (3.7)$$

We split the integrals into three independent cases:

Case 1 (Θ):

$$\int_{\Theta} P(\Theta | \alpha) P(z_p | \Theta) P(z_u | \Theta) d\Theta = \quad (3.8)$$

We use the template to describe the V number of visit records:

$$\begin{aligned} \int_{\Theta} P(\Theta | \alpha) \prod_{v=1}^V (P(z_{p,v} | \Theta) P(z_{u,v} | \Theta)) d\Theta = \\ \int_{\Theta} P(\Theta | \alpha) \prod_{v=1}^V (P(z_{p,v}, z_{u,v} | \Theta)) d\Theta = \end{aligned} \quad (3.9)$$

Θ is a $K_u \times K_p$ dimensional variable. As demonstrated in section ??, Θ_k can be transformed into $\Theta_{i,j}$ using a bijection function g where $i \in \mathbb{N}_{1:K_u}$, $j \in \mathbb{N}_{1:K_p}$.

$$\int_{\Theta_k} P(\Theta_k | \alpha) \prod_{v=1}^V (P(z_{p,v}, z_{u,v} | \Theta_{i,j})) d\Theta_k = \quad (3.10)$$

We replace the probability distribution with the definition of the Dirichlet distribution with equal number of parameters α as the dimensionality of Θ .

$$\iint_{\Theta_{i,j}} \frac{\Gamma\left(\sum_{k=1}^{K_u \times K_p} (\alpha_k)\right)}{\prod_{k=1}^{K_u \times K_p} (\Gamma(\alpha_k))} \prod_{k=1}^{K_u \times K_p} \left(\Theta_k^{\alpha_k - 1}\right) \prod_{v=1}^V (\Theta_{i,j}) \, d\Theta_{i,j} = \quad (3.11)$$

For simplicity, We define the counting function $c(a, b, c, d)$ using an SQL query:

```
Select count(*)
From VisitRecords
Where usertopic=a
      and placetopic=b
      and user=c
      and place=d
```

If a, b, c or d are equal to *, then the correspondent filter is removed from the *Where* clause.

$$\int_{\Theta_{i,j}} \frac{\Gamma\left(\sum_{k_u, k_p=1}^{K_u \times K_p} (\alpha_{k_u, k_p})\right)}{\prod_{k_u, k_p=1}^{K_u \times K_p} (\Gamma(\alpha_{k_u, k_p}))} \prod_{k_u, k_p=1}^{K_u \times K_p} \left(\Theta_{k_u, k_p}^{\alpha_{k_u, k_p} - 1}\right) \prod_{k_u, k_p=1}^{K_u \times K_p} \left(\Theta_{k_u, k_p}^{c(k_p, k_u, *, *)}\right) \, d\Theta_{k_p} \propto \quad (3.12)$$

Now, we arrange the variables so that we can describe a Dirichlet distribution which will integrate to one. This method has been used to derive the updating topics LDA¹. After the integration we find the following:

¹<http://lingpipe-blog.com/2010/07/13/collapsed-gibbs-sampling-for-lda-bayesian-naive-bayes/>

$$\frac{\prod_{k_u, k_p=1}^{K_u \times K_p} (\Gamma(c(k_p, k_u, *, *) + \alpha_{ku, kp}))}{\Gamma\left(\sum_{k_u, k_p=1}^{K_u \times K_p} (c(k_p, k_u, *, *) + \alpha_{ku, kp})\right)} \quad (3.13)$$

Case 2 (Φ_u):

Following a similar formulation that the one used for case 1.

$$\int_{\Phi_u} P(u|\Phi_u) P(\Phi_u|\beta_u) d\Phi_u \quad (3.14)$$

$$\prod_{k_u=1}^{K_u} \left(\int_{\Phi_u} P(u|\Phi_u) P(\Phi_u|\beta_u) d\Phi_u \right) \quad (3.15)$$

We obtain:

$$\prod_{k_u=1}^{K_u} \left(\frac{\prod_{v=1}^V (\Gamma(c(*, k_u, *, u_v) + \beta_{u,v}))}{\Gamma(\sum_{v=1}^V (c(*, k_u, *, u_v) + \beta_{u,v}))} \right) \quad (3.16)$$

Case 3 (Φ_p):

Following the same formulation that the one used for case 2.

$$\int_{\Phi_p} P(p|\Phi_p) P(\Phi_p|\beta_p) d\Phi_p = \quad (3.17)$$

$$\prod_{k_p=1}^{K_p} \left(\int_{\Phi_p} P(p|\Phi_p) P(\Phi_p|\beta_p) d\Phi_p \right) = \quad (3.18)$$

We obtain:

$$\prod_{k_p=1}^{K_p} \left(\frac{\prod_{v=1}^V (\Gamma(c(k_p, *, p_v, *) + \beta_{p,v}))}{\Gamma(\sum_{v=1}^V (c(k_p, *, p_v, *) + \beta_{p,v}))} \right) \quad (3.19)$$

In Summary:

$$P(z_u, z_p, u, p | \alpha, \beta_u, \beta_p) \propto$$

$$\frac{\prod_{k_u, k_p=1}^{K_u \times K_p} (\Gamma(c(k_p, k_u, *, *) + \alpha_{ku, kp}))}{\Gamma(\sum_{k_u, k_p=1}^{K_u \times K_p} (c(k_p, k_u, *, *) + \alpha_{ku, kp}))} \prod_{k_u=1}^{K_u} \left(\frac{\prod_{v=1}^V (\Gamma(c(*, k_u, *, u_v) + \beta_{u,v}))}{\Gamma(\sum_{v=1}^V (c(*, k_u, *, u_v) + \beta_{u,v}))} \right) \prod_{k_p=1}^{K_p} \left(\frac{\prod_{v=1}^V (\Gamma(c(k_p, *, p_v, *) + \beta_{p,v}))}{\Gamma(\sum_{v=1}^V (c(k_p, *, p_v, *) + \beta_{p,v}))} \right) \quad (3.20)$$

3.6.2 Posterior Derivation for each record

Since every visit record is considered to be independent, we can update every single record's user-topic and place-topic instead, we refine our objective probability function to the following:

$$P(Z_u^{(v)}, Z_p^{(v)} | Z_u^{(-v)}, Z_p^{(-v)}, u_v, p_v, \alpha, \beta_u, \beta_p) \quad (3.21)$$

Where $Z_u^{(v)}, Z_p^{(v)}$ represents the user and place topic for the v -th record and $Z_u^{(-v)}, Z_p^{(-v)}$ represent the user and place topic for all other records. Once again, we split equation 3.20 into three cases.

Case 1 (Θ): Given $u_v, p_v, Z_u^{(-v)}, Z_p^{(-v)}$:

$$\frac{\prod_{k_u, k_p=1}^{K_u \times K_p} (\Gamma(c(k_p, k_u, *, *)^{-v} + \alpha_{ku, kp}))}{\Gamma\left(\sum_{k_u, k_p=1}^{K_u \times K_p} (c(k_p, k_u, *, *)^{-v} + \alpha_{ku, kp})\right)} \quad (3.22)$$

We use the property of the gamma distribution $\Gamma(x+1) = \Gamma(x)x$ in order to extract the components that correspond to our v -th record of interest.

$$\frac{\prod_{k_u, k_p \neq z_{p,v}, z_{u,v}}^{K_u \times K_p} (\Gamma(c(k_p, k_u, *, *)^{-v} + \alpha_{ku, kp}) \Gamma(c(z_{p,v}, z_{u,v}, *, *)^{-v} + \alpha_{zu_v, zp_v}))}{\Gamma\left(1 + \sum_{k_u, k_p=1}^{K_u \times K_p} (c(k_p, k_u, *, *)^{-v} + \alpha_{ku, kp})\right)} \frac{c(z_{p,v}, z_{u,v}, *, *)^{-v} + \alpha_{zu_v, zp_v}}{\Gamma\left(1 + \sum_{k_u, k_p=1}^{K_u \times K_p} (c(k_p, k_u, *, *)^{-v} + \alpha_{ku, kp})\right)} \propto$$

Finally, after simplifying all terms that are not proportional to the topics $Z_u^{(v)}, Z_p^{(v)}$

$$c(z_{p,v}, z_{u,v}, *, *)^{-v} + \alpha_{zu_v, zp_v} \quad (3.23)$$

We use the same formulation for the other cases and we obtain the updating equation for $Z_u^{(v)}, Z_p^{(v)}$

$$P\left(Z_u^{(v)}, Z_p^{(v)} | Z_u^{(-v)}, Z_p^{(-v)}, u_v, p_v, \alpha, \beta_u, \beta_p\right) \propto$$

$$\frac{\left(c(z_{p(v)}, z_{u(v)}, *, *)^{(-v)} + \alpha_{zu(v), zp(v)}\right) \left(c(z_{p(v)}, *, p_{(v)}, *)^{(-v)} + \beta_{p(v)}\right) \left(c(*, z_{u(v)}, *, u_{(v)})^{(-v)} + \beta_{u(v)}\right)}{\sum_{j=1}^{|P|} \left(c(z_{p(j)}, *, p_{(j)}, *)^{(-v)} + \beta_{p(j)}\right) \sum_{i=1}^{|U|} \left(c(*, z_{u(v)}, *, u_{(i)})^{(-v)} + \beta_{u(i)}\right)} \quad (3.24)$$

3.7 Relationship with Latent Dirichlet Allocation Model

We start by presenting the inference equation used to update the topics in the LDA model. Then, we compare with the inference equation obtained by the Y^2 -LDA model. In conclusion, we find that Y^2 -LDA generalizes LDA.

3.7.1 Inference of the Latent Dirichlet Allocation

The topic assignment to the v -th visit record $\{u(v), p(v)\}$ is given by sampling from the following derived posterior probability [9]:

$$P(z_{p(v)} | z_{p(-v)}, u, p, \alpha_u, \beta_p) \propto \frac{\bar{\mathcal{L}} \mathcal{P}(v)}{\sum_{j=1}^{|P|} \mathcal{P}(j)}$$

where

$$\bar{\mathcal{L}} = \sum_{n=1}^{|P|} c(z_{p(v)}, u_{(v)}, n)^{(-v)} + \alpha_{z_{p(v)}}$$

$$\mathcal{P}(i) = \sum_{m=1}^{|U|} c(z_{p(v)}, m, p_{(i)})^{(-v)} + \beta_{p(i)}$$

$$c(x, m, n) = \sum_{j=1}^{|V|} \mathbb{1}(z_{u(j)} = x, u_{(j)} = m, p_{(j)} = n)$$

Since the function c corresponds to sums of indicator functions, it is easy to interpret the two factors responsible for the topic assignment:

- i) $\bar{\mathcal{L}}$: Counting of topics assigned to user u_v .
- ii) $\frac{\mathcal{P}(v)}{\sum_{j=1}^{|P|} \mathcal{P}(j)}$: Probability of observing the place p_v given a topic.

3.7.2 Comparison with the inference equation of Y^2 -LDA

We show that LDA is a particularization of Y^2 -LDA by proving that LDA's inference equation can be reduced to the inference equation of Y^2 -LDA when we apply the condition that each

user is assigned to a single and unique user-topic. i.e. users and user-topics are assigned in a *one-to-one* mapping

Proof. Because we are given the user-topic assignments for every user, the inference of Y²-LDA only depends on $z_p(v)$. So, recalling equation 3.3, we begin by analyzing the term $\frac{\mathcal{U}(v)}{\sum_{i=1}^{|U|} \mathcal{U}(i)}$. Because $\mathcal{U}(v)$ does not depend on $z_p(v)$, we can replace $\mathcal{U}(v)$ with a constant value. Then, we analyze the term \mathcal{Z} . Since $z_u = u \forall u \in 1 : U$, we can replace z_u with u and as result, we end up with the LDA inference equation. That is:

$$P(z_{p(v)} | z_u, z_{p(-v)}, u, p, \alpha, \beta_u, \beta_p) =$$

$$P(z_{p(v)} | z_{p(-v)}, u, p, \alpha_u, \beta_p) \propto \frac{\mathcal{Z} \mathcal{P}(v)}{\sum_{j=1}^{|P|} \mathcal{P}(j)} \text{ where}$$

$$\mathcal{Z} = \sum_{m=1}^{|U|} \sum_{n=1}^{|P|} c(z_{p(v)}, u(v), m, n)^{(-v)} + \alpha_{u(v), z_{p(v)}}$$

and α_u corresponds to a K_p dimensional vector conditioned on the user u . In LDA, the hyper-parameter of the Dirichlet prior on the per-user topic distributions $\bar{\alpha}$ is independent of the given user to avoid over-fitting. We can break the dependence of α and the given user by setting α to the same value for all users. ■

Finally, we note that if we start off by assigning a *one-to-one* mapping between the places and the place-topics, we would end up with the inference updating equation of the Dual LDA.

3.8 Bicluster Extraction using Y²-LDA

If we do not implement intuition three (Subsection 3.3.3), Y²-LDA would obtain a clustering of the set of users as given by the user-topics. Independently, Y²-LDA would obtain a

clustering model of the set of places as given by the place-topics. By enforcing a mutual dependency between the user-topics and the place-topics, we obtain an algorithm capable of identifying relationships between subsets of the set of places and subsets of the set of users. The relationship between a group of users and a group of places corresponds to a Bicluster.

In section 2.3, we have introduced the problem of biclustering as well as the categorization of the biclustering algorithms by types and structures. Using Y^2 -LDA, we don't fit the structure of the biclusters and so we can extract the structures presented in figures 2.3a, 2.3b, 2.3c, 2.3d. With respect to the biclustering types, Y^2 -LDA can extract those biclusters that approximate to the bicliques of the dataset. We will give a formal explanation on how does the model extract bicliques in section 3.9.

We obtain the biclusters from the parameters of the Y^2 -LDA model. First, we can obtain the groups of users from the parameter Φ_u . Second, we can obtain the groups of places from the parameter Φ_p . Third, we can obtain the biclusters from the parameter Θ , because Θ represents the relationship between every place-topic and every user-topic.

Because the parameters ϕ_u represent $|K_u|$ distributions over the whole set of users and ϕ_p represents $|K_p|$ distributions over the whole set of places. Then, we need to establish thresholds in order to decide which user's are more relevant in a user-topic, which places are more relevant in a place-topic and which relationships are the strongest relationships between user-topics and place-topics.

We extract the subset of users in the m -th user-topic from $U^m = \{u \in \mathbb{U} | P(u | \phi_{u,(m)}) > Q_{75} \forall m \in 1 : K_u\}$, and subset of places in the n -th place-topic from $P^m = \{p \in \mathbb{P} | P(p | \phi_{p,(n)}) > Q_{75} \forall n \in 1 : K_p\}$. Once we partition the groups of users and groups of places, we need to find the top K strongest relationships between the groups of users and groups of places. We can do that by selecting the top K parameters $\Theta_{i,j}$ with the largest value. Alternatively, we can select the top K parameters so that we don't need to specify the number of biclusters that we want to extract. We can extract the list of the most relevant biclusters from $\mathcal{B} =$

$\{U^m, P^n | P(z_u = m, z_p = n | \Theta, \alpha) > Q_{75} \forall m \in 1 : K_u, \forall n \in 1 : K_p\}$. Where Q_{75} is the 75-th percentile.

3.9 Why does Y²-LDA works?

We show that Y²-LDA is able to extract bicliques by analyzing which records influence the topic assignment of which other records. We discover that the records belonging to the same biclique have the highest mutual influence and so they are most likely to be assigned to the same user-topic and the same place-topic.

3.9.1 Topic influence between a pair of records

We start by rewriting the inference equation (Eq. 3.24) in terms of three factors:

$$P(z^u, z^p | u, p, \alpha, \beta) = P(z^u, z^p | \Theta) P(u | \Phi^u) P(p | \Phi^p)$$

The topic influence from the record $\langle u, p \rangle$ to the record $\langle \hat{u}, \hat{p} \rangle$ is given by the following equation:

$$P(z^u, z^p | \hat{u}, \hat{p}, \alpha, \beta) = P(z^u, z^p | u, p, \alpha, \beta) \frac{P(p | \Phi^p) P(u | \Phi^u)}{P(\hat{p} | \Phi^p) P(\hat{u} | \Phi^u)} \quad (3.25)$$

We know that the influence is maximal if $\frac{P(p | \Phi^p) P(u | \Phi^u)}{P(\hat{p} | \Phi^p) P(\hat{u} | \Phi^u)} = 1$ because the topic assignment of $\langle u, p \rangle$ will follow the same equation as the topic assignment for the record $\langle \hat{u}, \hat{p} \rangle$.

If the two records belong to the same biclique, then it must follow that either i) The user in both records is the same ($u = \hat{u}$) ii) The place in both records is the same ($p = \hat{p}$) iii) Trivially, the user and the place is the same in both records.

In these three cases, we can simplify equation 3.25 because either $\frac{P(u|\Phi^u)}{P(\hat{u}|\Phi^u)} = 1$ or $\frac{P(p|\Phi^p)}{P(\hat{p}|\Phi^p)} = 1$. It is only in the case when two records are not in the same biclique that equation 3.25 will not be simplified. If equation 3.25 cannot be simplified, both records will not have a strong influence in the topic assignment of each other. In conclusion, those records that share the same user or the same place have the highest influence in the topic assignment of each other.

3.9.2 Records in the same biclique have the strongest mutual influence

We have proved that a pair of records in the same biclique will have higher influence when compared with a pair of records that do not belong to the same biclique. Therefore, all records in a biclique will mutually reinforce the topic assignment of each other and so the user-topic, place-topic assignment will tend to be the same for all these records. For a better understanding of this fact, we provide a graphical example in figure 3.7.

First, we represent the visit records in a graph so that each vertex represents the visit of a user to a place (Figure 3.7a). Second, we visualize with a colored set those records that influence the topic assignment of the record $\langle u_1, p_1 \rangle$. In this case, we note that the records $\langle u_2, p_1 \rangle$ and $\langle u_1, p_2 \rangle$ will have higher influence on $\langle u_1, p_1 \rangle$ because those records have either the same place or the same user (Figure 3.7b). Third, we visualize the influence of those records in the biclique composed by $u_1, u_2; p_1, p_2$. We observe how all the records in the biclique mutually influence each other (Figure 3.7c). Fourth, we display an example of two records who do not belong to the same biclique. The records $\langle u_1, p_2 \rangle$ and $\langle u_4, p_4 \rangle$ influence the topic assignment of the record $\langle p_2, u_4 \rangle$. But they will not have a mutual influence (Figure 3.7d). Fifth, we display all the influence sets generated by the records in the dataset. We observe that from the aggregation of all the influence sets, the records including $u_1, u_2; p_1, p_2$ are likely to be assigned to the same user-topic, place-topic pair. We conclude the same for the records including $u_4, u_5; p_3, p_4$ (Figure 3.7e).

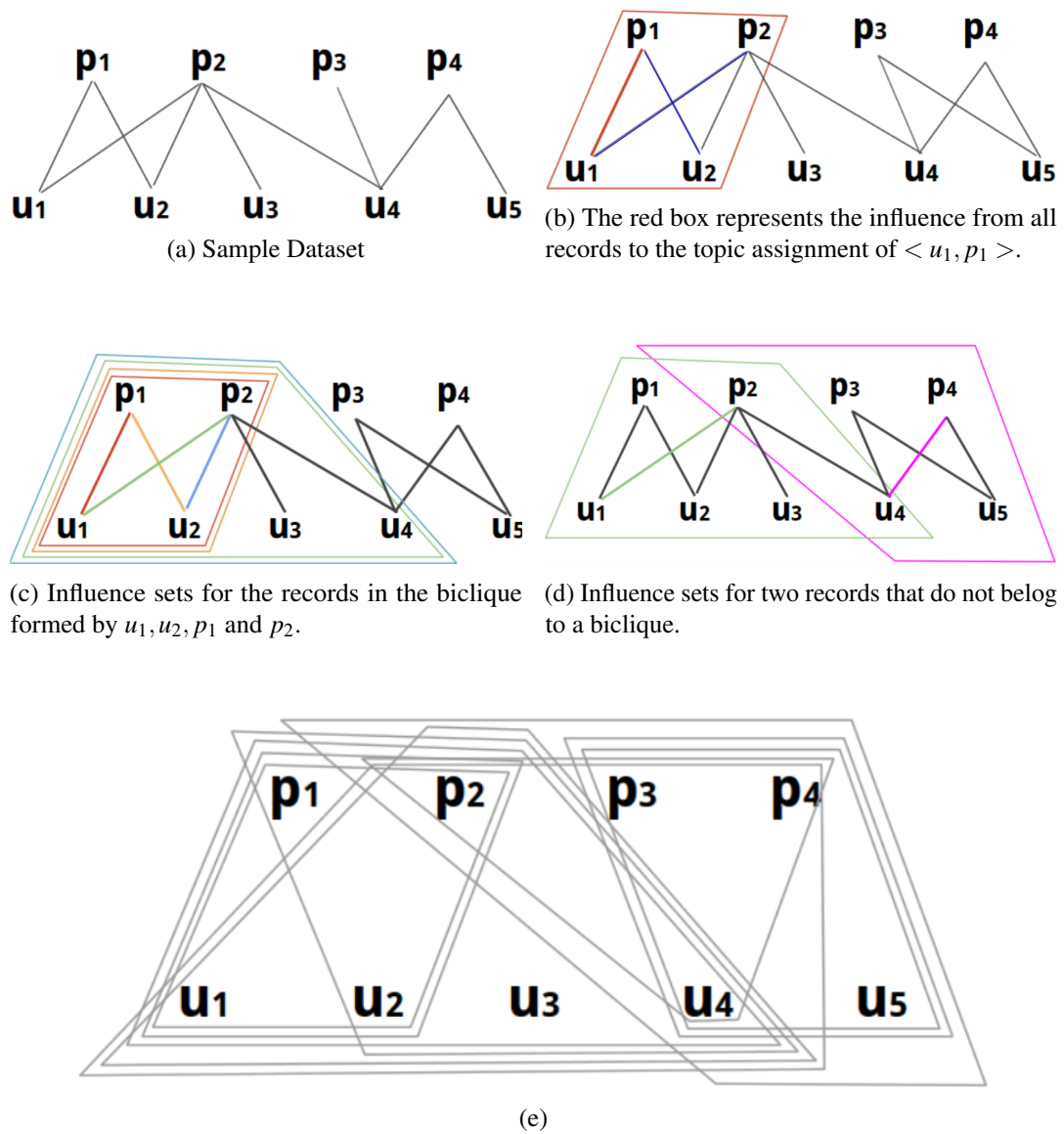


Fig. 3.7 Demonstration of how Y^2 -LDA extracts bicliques.

Chapter 4

Applications and Experimental Evaluation

4.1 Datasets

Foursquare. The first type of data contains check-in data. Every check-in record can be represented as a $\langle \text{user_id}, \text{place_id}, \text{category} \rangle$ tuple. The records include 227428 check-ins from 1082 users into 38333 venues collected from Apr 2012 to Feb 2013 in New York [75], 573703 check-ins recorded from 2293 users into 61858 venues collected from Apr 2012 to Feb 2013 in Tokyo [75], and 342850 check-ins from 2341 users into 44173 venues collected from Aug 2010 to Jul 2011 in Singapore.

Singtel Cell Location. The second data set contains cell-tower location records which can be represented as a $\langle \text{user_id}, \text{cell_tower lat}, \text{cell_tower lon} \rangle$ tuple. We used 6880128 records from 20830 Singapore's residents and visitors into 4328 cell-towers located across Singapore collected during 2013. This records were provided by Singtel .

⁰<http://info.singtel.com/>

NIPS. The third dataset is the text data obtained from papers accepted by NIPS from 1988 until 2003 [?]. The data set contains 40552970 records generated by 2865 authors who wrote 2483 documents using a vocabulary of 14036 words.

4.2 Modeling Visit Records

4.2.1 Application

Y^2 -LDA can generate new data samples for two variables. To simulate new data using Y^2 -LDA we need to run the inference process to extract the topics and then we can use the generative process described in section 3.4 to sample instances of new records.

4.2.2 Experimental Settings

To evaluate Y^2 -LDA's ability to model data, we can compute the perplexity on a held-out test set. The perplexity on a test set is a commonly used metric for the predictability of the given test set. The following equation is used to find the perplexity for a set of records \mathcal{W} :

$$\text{Perplexity } (\mathcal{W}) = \exp \left(- \frac{\sum_{v \in \mathcal{W}} \log P(\mathcal{W}_v | \Phi, \beta)}{|\mathcal{W}|} \right)$$

We compute the perplexity for evaluating the predictability of words, authors, users and places using the NIPS and Foursquare datasets. We compare the perplexity of Y^2 -LDA with an LDA model, its dual formulation and the perplexity at the initialization. Since Y^2 -LDA generates both word and authors simultaneously, we require to extract independently the perplexity of predicting users given by $P(u | \Phi_u, \alpha)$ and the perplexity of predicting places as given by $P(p | \Phi_p, \alpha)$. In the appendix we show how to compute these probabilities when using Y^2 -LDA.

Initialization	Model	Fsq New York		Fsq Tokyo		Fsq S'pore		NIPS	
		User	Place	User	Place	User	Place	Document	Word
Uniform	K-means	1559	992	4341	1922	2674	11810	3302	2290
K-means	LDA	1411	820	3994	1532	2457	11746	2583	1529
K-means	Y ² -LDA	1411	788	4002	1412	2395	11731	2514	1452
K-means LDA	Y ² -LDA	1411	783	4005	1349	2381	11730	2543	1496

Table 4.1 Average perplexity computed over twenty trials for four different data sets.

For LDA, we set the number of places-topics and user-topics to fifty, we select $\alpha_i = 0.5 \forall i \in 1 : K$ and $\beta_j = 0.1 \forall j \in W$. Where W corresponds to the vocabulary set. In Y²-LDA models, We set $\alpha_i = 0.05 \forall i \in K_u \times K_p$, all other parameters are set identically. For foursquare's data, we split the dataset by time, so that the 40% more recent records are held-out for testing. For the NIPS dataset, we select 20% of the documents at random.

We repeat the experiment with different initialization methods. We use uniform sampling as a naive initialization method, then we use K-means and finally we explore the initialization of Y²-LDA using the extracted results of LDA. The latter method will help us check if there is an improvement in the topic extraction given by the mutual reinforcement of location-based models and individual-based models.

4.2.3 Results and Analysis

We present the results in table 4.1. We observe that Y²-LDA converges to the best results. When we use the results of an LDA and its dual LDA to initialize Y²-LDA, Y²-LDA yields overall acceptable results and substantial faster convergence. The improvement is relatively larger for the evaluation of place topics when compared to the user topic, this is a consequence of the larger sparsity in the set of places. This effect is also observed on the experiments on the NIPS dataset.

4.3 Reviewer Recommendation.

4.3.1 Application

We evaluate Y^2 -LDA in an application where we require to make use of the Dual LDA. As suggested by Rosen-Zvi et al, Automated Reviewer Recommendations is an application that requires extracting author similarity [58]. We can capture this similarity by using an LDA so that we obtain author-topics. We propose a modified version of the application where we generate a list of likely authors who may have written a document given the collection of words in a document. Our assumption is that if we can correctly identify the authors of a document, then we should be able to identify similar authors who may have written the document and hence they could serve as reviewers for the document. Since the application doesn't require any information linking the document to the authors, we could recommend reviewers to documents with a single author or documents written by an author not present in the corpus.

More formally, given a document \mathcal{D} represented by a set of words w , we can use Y^2 -LDA to extract the probability of an author a to write a word w from equation 4.1

$$P(a, w) = \sum_{z_a} \sum_{z_w} P(a | \Phi_{z_w}, \beta_w) P(a | \Phi_{z_a}, \beta_a) P(z_a, z_w | \Theta) \quad (4.1)$$

Then, we use $\prod_{w \in \mathcal{D}} P(a | w)$ to find the probability of an author to have written the document \mathcal{D} . For comparative purposes, we also extract $P(a | w)$ using an LDA by representing the document level using words and the word-level using the set of all authors in the corpus. i.e. we use an LDA to represent a word as the set of authors who have written this word. In this model, the probability of an author a to write a word w is calculated from equation 4.2.

$$P(a | w) = \sum_{z_a} P(a | \Phi_{z_a}) P(z_a | \Theta_w) \quad (4.2)$$

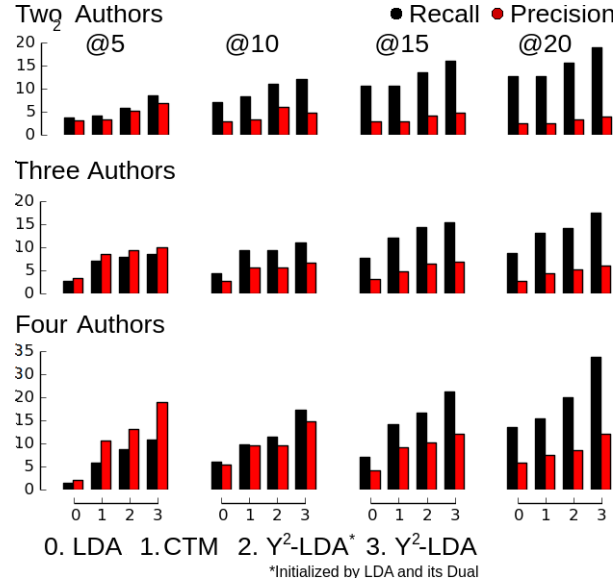


Fig. 4.1 Reviewer recommendation

We use a similar formulation to estimate $P(a|w)$ using the Correlated Topic Model (CTM) [7].

4.3.2 Experimental Settings

For this experiment, we split the papers in the NIPS corpus into a training set and a testing set, after deleting all papers in the test set with one author and those papers where the authors are not in the training set, we obtain as the test set 525 papers with two authors, 305 with three authors and 111 with four authors. We evaluate the precision and recall at n by selecting the top n authors from $P(a|\mathcal{D})$. In all the models we set the number of topics equal to fifty. In the case of Y²-LDA both author-topics and word-topics are set to fifty as well.

4.3.3 Results and Analysis

We present the results in figure 4.1. We observe an improvement of Y²-LDA over CTM and an improvement of CTM over LDA. This is expected as both CTM and Y²-LDA generalize LDA. Overall we can see how papers with larger number of authors have a better performance

Trial	Uniform			Kmeans		
	LDA	Y ² -LDA	LDA+Y ² -LDA	LDA	Y ² -LDA	LDA+Y ² -LDA
1	0.032	0.032	<u>0.016</u>	0.104	0.024	<u>0.008</u>
2	0.032	0.272*	<u>0.016</u>	<u>0.296*</u>	0.304*	<u>0.296*</u>
3	<u>0.008</u>	0.272*	<u>0.008</u>	0.208	0.336	<u>0.000</u>
4	<u>0.016</u>	0.016	<u>0.016</u>	0.448	0.416	<u>0.208</u>
5	0.040	0.024	<u>0.016</u>	0.224	0.112	<u>0.008</u>
6	0.312*	0.288*	<u>0.264*</u>	0.288*	<u>0.144</u>	0.272*
7	0.176	0.016	<u>0.008</u>	0.112	0.128	<u>0.048</u>
8	<u>0.000</u>	0.152	<u>0.000</u>	<u>0.016</u>	<u>0.016</u>	<u>0.016</u>
9	0.304*	<u>0.272*</u>	0.304	0.024	0.016	<u>0.008</u>
10	0.184	<u>0.000</u>	0.016	0.344*	0.304*	<u>0.160</u>
Average	0.1104	0.1344	0.0664	0.2064	0.1800	0.1024

Table 4.2 Average misclassification error for ten random initializations. The initialization method is specified in the headers of the table. The asterisk represents those cases where the models completely confused two topics.

when trying to identify the authors of the paper. As expected, the precision decreases as we increase the list of recommended authors and in contrast, the recall increases. Finally, we note that initializing GLDA with a uniform distribution leads to better performance than initializing it with the result of an LDA and its dual LDA, though the convergence time is about an order of magnitude larger.

4.4 User and Place Classification

4.4.1 Application

Y²-LDA can also be used to classify data using the latent topics as the classes that represent the data. IN a hard classification, We can assign class z to user u_i using $\hat{z}_u = \arg \max_z P(u = u_i | \Phi_u, z_u)$ by using the extracted topics. In a soft-classification mode, one could extract $P(z_u | u = u_i) = \frac{P(u = u_i | \Phi_u, z_u) P(\Phi_u, z_u)}{P(u = u_i)}$ where $P(u = u_i)$ can be found by counting from the

given data and $P(\Phi_{u,z_u}) = \frac{\prod_{i=1}^K \Gamma(\beta_i)}{\Gamma(\sum_{i=1}^K \beta_i)} \prod_{i=1}^K \Phi_{u,i}^{\beta_i-1}$ as specified by the Dirichlet prior with parameter β

4.4.2 Experimental Settings

In this experiment we test Y²-LDA's ability to represent individual user-topics and individual place-topics. We selectively filter the cell-tower data so that we can obtain clear topics of users and places. We select all the visit records of five groups of 250 users whose home and work correspond to the same cell tower location. We assume as ground truth that there exists five user-topics, one topic representing a group of 250 users. For every user u , we assign the user-topic \hat{z}_u using $\hat{z}_u = \arg \max_{z_u} P(u = u_i | \Phi_u, z_u)$. For the evaluation, we compare the average misclassification error of Y²-LDA against an LDA. We initialize the models uniformly and using K-Means. In addition, we also used the results of an LDA and its dual LDA to initialize the Y²-LDA.

4.4.3 Results and Analysis

The results of the experiment are presented in table 4.2. As in experiment 4.2, we observe consistent improvement of Y²-LDA model when initialized using K-means and the independent LDA models. However, in those cases where K-means initialization is not so convenient such in trials 7 and 8, we cannot expect Y²-LDA to solve the confusion between the extracted topics of the independent LDA models. We also observe that LDA is a better way to initialize Y²-LDA model. We attribute the improvement of Y²-LDA over LDA to the fact that Y²-LDA models an iterative inference of both user-based models and location-based model. In contrast, the LDA is a location-based model.

Additionally, we used the New York Foursquare's dataset to compare the extracted place-topics of LDA and Y²-LDA when initialized with the results of independent LDA models. In figure 4.2 We point out all differences that we consider to be an improvement. Typically,

LDA

Topic 1

Bar 15361.0
~~Gym / Fitness Center 4301.0~~
 Coffee Shop 3192.0
 American Restaurant 2874.0
 Hotel 2821.0
 Airport 2449.0
 Park 2255.0

Topic 2

Residential Building 3957.0
~~Bus Station 3672.0~~
 College Academic Bui 3460.0
 Road 2669.0
 Train Station 2140.0
 Other Great Outdoors 1913.0
~~Building 1624.0~~

Topic 3

Home (private) 14256.0
 Office 11490.0
 Gym / Fitness Center 4868.0
~~Food & Drink Shop 703.0~~
~~Bar 450.0~~
 Government Building 399.0
 Other Great Outdoors 327.0

Topic 4

Food & Drink Shop 4901.0
 Coffee Shop 3990.0
~~Medical Center 3097.0~~
 Deli / Bodega 2627.0
 Clothing Store 1757.0
 Drugstore / Pharmacy 1581.0
 Bank 1401.0

Topic 5

Subway 9347.0
 Neighborhood 4027.0
 Train Station 3658.0
 Deli / Bodega 1507.0
 Other Great Outdoors 1258.0
 Park 1130.0
 Bridge 1074.0

GLDA

Topic 1

Bar 15901.0
 Coffee Shop 3399.0
 Hotel 2845.0
 American Restaurant 2829.0
 Park 2647.0
 Airport 2487.0
 Mexican Restaurant 1706.0

Topic 2

Residential Building 4155.0
 College Academic Bui 3479.0
 Home (private) 2924.0
 Medical Center 2898.0
 Deli / Bodega 2627.0
 Road 2568.0
 Other Great Outdoors 2140.0

Topic 3

Office 12493.0
 Home (private) 9714.0
 Gym / Fitness Center 8669.0
 Building 1323.0
 Government Building 727.0
 American Restaurant 383.0
 Automotive Shop 284.0

Topic 4

Food & Drink Shop 6224.0
 Coffee Shop 2068.0
 Clothing Store 1916.0
 Pizza Place 1617.0
 Drugstore / Pharmacy 1611.0
 Chinese Restaurant 1430.0
 Salon / Barbershop 1379.0

Topic 5

Subway 9342.0
 Train Station 5469.0
 Neighborhood 4252.0
 Home (private) 2740.0
 Bus Station 2432.0
 Deli / Bodega 1522.0
 Park 1491.0

Fig. 4.2 Comparison for place-topic extraction between LDA and Y²-LDA. Y²-LDA was initialized with the results of the LDA. The numbers represent the number of samples allocated to a topic

Y^2 -LDA is able to solve some confusion between the places represented by two topics. e.g. After convergence of the LDA model, *Gym* appeared in topic 1 and 3. However, when we used Y^2 -LDA to improve the convergence of LDA, *Gym* was moved to topic 3 where it is better represented. Our model also improved the ranking of the places inside a topic. e.g. In topic 5, Y^2 -LDA allocated more instances to *Train Station* in a topic that clearly corresponds to places related to transportation.

4.5 Biclustering

4.5.1 Application

The inter-topic correlation can be interpreted as the relationship between subsets of users and subsets of places. Semantically, this correlation represents the information of "Which sets of users visit which sets of places", or "Which sets of documents are described by which sets of words" Extracting this relationship corresponds to the same objective of biclustering [31]. Formally, A Biclustering algorithm takes as an input a matrix A with dimensionality $\mathbb{I} \times \mathbb{J}$ and it outputs a list of biclusters \mathcal{B} where each bicluster $B_{I,J}$ is defined by a subset of the rows $I \subseteq \mathbb{I}$ and a subset of the columns $J \subseteq \mathbb{J}$. In the user-place space $\mathbb{U} \times \mathbb{P}$, $A_{i,j} = 5$ represents five visits from the i -th user to the j -th location. A bicluster B_{U^m, P^n} is defined by a subset of users $U^m \subseteq U$ and a subset of the places $P^n \subseteq P$.

4.5.2 Extracting biclusters from Y^2 -LDA

4.5.3 Experimental Settings

We want to evaluate Y^2 -LDA's ability to extract high-frequency biclusters under the structures presented in figures 2.3a, 2.3b, 2.3c, 2.3d because this information represents the solution to the question "Which are the most common groups of users that visit a group of similar

places?". Since we do not have ground truth labels for the answers to our question. We propose an experiment where we use three synthetic datasets. The first dataset assumes that every user and every place must belong to exactly one bicluster 2.3a. The second dataset assumes that every user and every place belongs to at least one bicluster 2.3c, and 2.3d. The third dataset assumes that every user must belong to at most one bicluster 2.3b. We use a normal distribution with a variance σ^2 to represent the visit frequency for a subset of users and a subset of places. All datasets are set to have 100000 users and 100000 places.

As the evaluation metric, we use recovery and relevance [16] as given by $S(\mathcal{B}_1, \mathcal{B}_2) = \frac{1}{|\mathcal{B}_1|} \sum_{B_1 \in \mathcal{B}_1} \max_{B_2 \in \mathcal{B}_2} |B_1 \cap B_2|$ Where \mathcal{E} is the list of expected biclusters. Relevance is defined as $S(\mathcal{B}, \mathcal{E})$ and Recovery is defined as $S(\mathcal{E}, \mathcal{B})$. Intuitively, these metrics are an analogy to Recall and Precision.

We compare results with bicluster algorithms that best performed on tested data with varying conditions, including varying noise and varying numbers of biclusters [16],[46]. The implementation of all algorithms were taken from the author's website.

4.5.4 Results and Analysis

In figure 4.4, we display the results of this experiment. In the first dataset, we observe an improvement of the spectral bipartite graph over other models, that is because Spectral bipartite graph assumes that all places and all users must be assigned to exactly one bicluster and so the algorithm is fitted for the assumptions of the first dataset, when we relax the assumption of dataset 1 in datasets 2 and 3, the results of the spectral bipartite graph decrease when compared against Y^2 -LDA, overall Y^2 -LDA yields better results when initialized using Y^2 -LDA.

The spectral bipartite graph [14] algorithm performs much better than other baseline models because the spectral bipartite graph algorithm models the relation between users and places with a bipartite graph weighted with the number of visits. Therefore, this algorithm

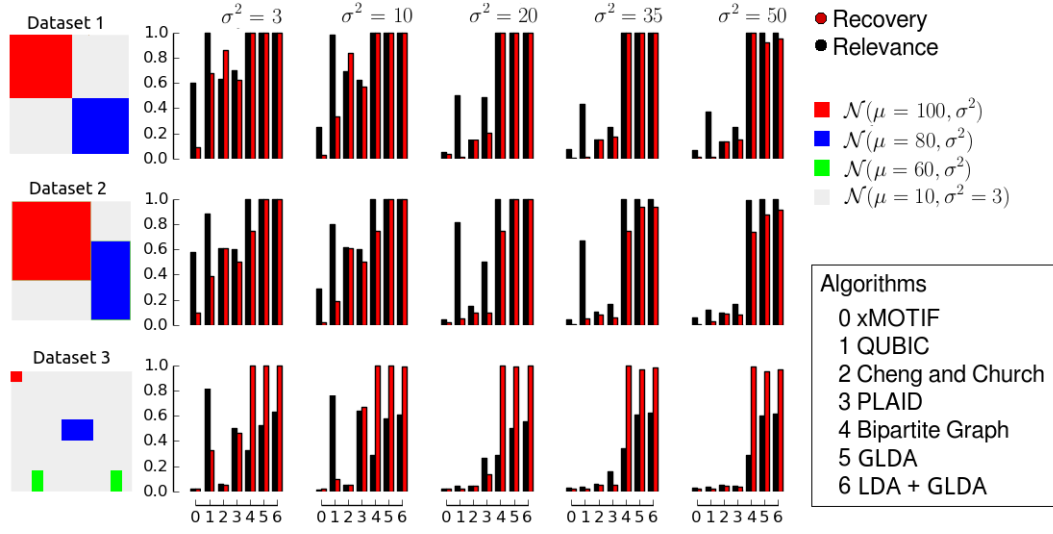


Fig. 4.3 Recovery and Relevance results for bicluster classification. The datasets are visualized with a matrix whose color represent the number of visits of users (rows) to places (columns).

is designed for extracting high-frequency biclusters. Other biclustering algorithms aim to find similarities between rows and columns without taking into account the absolute value of their relationship.

In addition, we visualize one sample of the biclusters extracted by Y^2 -LDA to compare the difference between initializing Y^2 -LDA using two LDA models and Y^2 -LDA using two k-means algorithms. The visualization is shown in figure 4.3, every color represent a different bicluster, on the right we can see the input dataset where the number of visits from user i -th to place j -th is given by a value between 0 and 200.

As result, we observe that initializing Y^2 -LDA with two independent LDAs (LDA- Y^2 -LDA) is a better way to find more relevant biclusters for two reasons. First, the number of extracted biclusters is more precise since LDA- Y^2 -LDA adds more weight to the topic-correlation that represents the ground truth biclusters. i.e. LDA- Y^2 -LDA has a better representation of the topic inter-correlation. Second, LDA- Y^2 -LDA extracts biclusters that better fit the shape of the ground truth biclusters. This is because LDA- Y^2 -LDA extracts

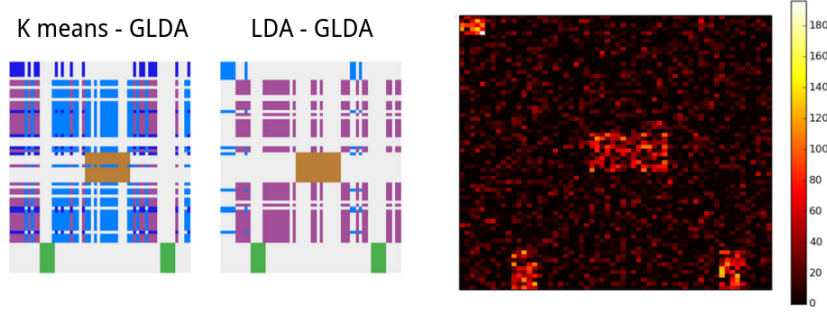


Fig. 4.4 Different initialization methods for extracting biclusters with Y^2 -LDA

topics that better represent the subsets of users and subset of places that we expect as ground truth.

4.6 Dimensionality Reduction

4.6.1 Application

Topic models are commonly used for dimensionality reduction [9]. The new dimensionality can be reduced by using the inferred topics. In the LDA model, for the user-place domain, every user (\mathcal{U}) will be assigned a mixture of topics and every topic will correspond to a mixture over the set of places (\mathcal{P}). i.e. We are describing $\hat{P}(\mathcal{P}|\mathcal{U})$ in a lower dimension. An optimal representation of the lower dimension conditional distribution would be able to get close to the truth distribution $P(\mathcal{P}|\mathcal{U})$.

4.6.2 Experimental Settings

In this experiment, we assume the ground truth to be the maximum likelihood estimation of $P(\mathcal{P}|\mathcal{U})$, we further assume that \mathcal{P} and \mathcal{U} are multinomial random variables. For the evaluation, we extract the estimated value of $\hat{P}_{lda}(\mathcal{P}|\mathcal{U}; z)$ using the LDA model as $P_{lda}(\mathcal{U}|\mathcal{P}) = \sum_{z_u}^{K_u} P(z_u|\Theta, \alpha)P(\mathcal{U}|\Phi_{z_u}\beta_u)$. Similarly using the dual LDA Formulation, we extract $P(\mathcal{U}|\mathcal{P}; z)$. For the Y^2 -LDA model we approximate the conditional distribution by first

N° Topics	$P(\mathcal{P} \mathcal{U})$			$P(\mathcal{U} \mathcal{P})$		
	Y ² -LDA	LDA	Improvement	Y ² -LDA	LDA	Improvement
3	.261 (.003)	.262 (.005)	44.7%	.644 (.006)	.647 (.005)	58.2%
6	.227 (.002)	.245 (.006)	76.3%	.592 (.006)	.621 (.005)	69.9%
9	.214 (.003)	.242 (.005)	78.1%	.565 (.007)	.623 (.005)	89.4%
12	.204 (.004)	.559 (.008)	99.7%	.550 (.007)	1.00 (.004)	99.8%
15	.193 (.004)	.528 (.011)	99.9%	.523 (.006)	1.01 (.005)	99.9%

Table 4.3 On the left side of the table, we show the average Battacharyya distance between the approximated distributions $\hat{P}_{LDA}(\mathcal{P}|\mathcal{U})$, $\hat{P}_{Y^2-LDA}(\mathcal{P}|\mathcal{U})$ and the estimated distribution $P(\mathcal{P}|\mathcal{U})$. We assume the maximum likelihood estimation of $P(\mathcal{P}|\mathcal{U})$ as ground truth. On the right side, we display the Battacharyya distance between the approximated distributions $\hat{P}_{LDA}(\mathcal{U}|\mathcal{P})$, $\hat{P}_{Y^2-LDA}(\mathcal{U}|\mathcal{P})$ and the estimated distribution $P(\mathcal{U}|\mathcal{P})$. We assume the maximum likelihood estimation of $P(\mathcal{U}|\mathcal{P})$ as ground truth.

finding the join distribution $P(\mathcal{U}, \mathcal{P}; z_u, z_p) = \sum_{z_u} \sum_{z_p}^{K_u K_p} P(z_u, z_p | \Theta, \alpha) P(\mathcal{U} | \Phi_{z_u} \beta_u) P(\mathcal{P} | \Phi_{z_p} \beta_p)$ and then we extract the conditional distributions by normalizing the joint distribution. We use the Bachattaryya distance to measure the distance between distributions, the Bachattaryya distance is a metric commonly used to measure the distance for discrete distributions [?]. In the case of multinomial distributions the Bachattaryya distance between p and q corresponds to $Battacharyya(p, q) = -\ln \left(\sum_x \sqrt{p(x), q(x)} \right)$. We report results based on averaging the results of running the experiment 50 times.

4.6.3 Results and Analysis

In table 4.3, we observe the average Battacharyya distance between the approximated distributions $\hat{P}_{LDA}(\mathcal{P}|\mathcal{U})$, $\hat{P}_{Y^2-LDA}(\mathcal{P}|\mathcal{U})$ and the estimated ground truth distribution $P(\mathcal{P}|\mathcal{U})$. We also show the percentage of the users whose average distance to the ground truth distribution is improved as well as the percentage of places improved by the Y²-LDA model.

Y²-LDA improves the place-distribution for around 75% of all the users and 70% of the user-distributions per place. These results hold independently of the number of topics. However, when we use a small number of topics, the performance of Y²-LDA and LDA tend

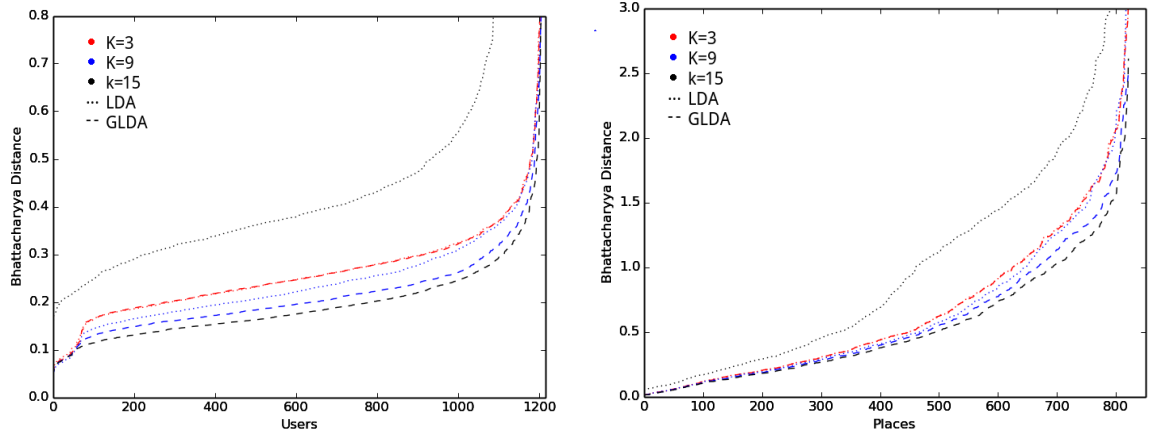


Fig. 4.5 Results of table 4.3 for every user, we sort the users and places by their average Battacharyya distance to facilitate the visualization of the results.

to converge. Naturally, as we increase the number of topics, we obtain smaller distances to the truth distributions. When we surpass 10 topics, LDA decreases its performance dramatically. When we use LDA to extract larger number of topics, several topics end up being repeated. This is a consequence of LDA modeling the correlation of user to topics without considering the relationship between users. In figure 4.5 we visualize the distance of the place-distribution to the ground truth for every individual and the distance of the user-distribution to the ground truth for every place. The figure shows how changing the number of topics reflects an homogeneous improvement or decrease of performance for all the users.

4.7 Topic Visualization

Finally, we created a visualization to show the extracted topics and their correlation. We tried this visualization on two domains. Firstly, we used a Movie rating dataset using one million ratings from 6000 users on 4000 movies ¹. We selected this dataset since we also have available user demographic information that can be used to explain the user topics.

¹<http://grouplens.org/datasets/movielens/>

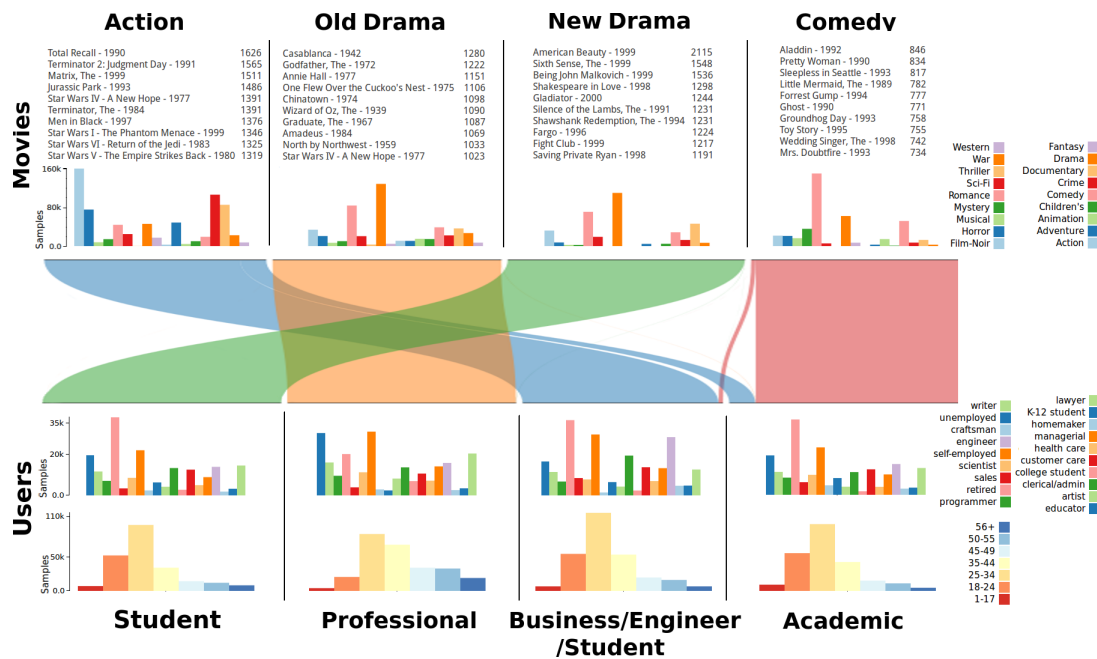


Fig. 4.6 On top, we show four movie-topics represented by a) Most representative movies and its year of release. b) Histogram of the movie genres weighted by the number of instances that a movie was allocated to a topic. On the middle we represent the correlation between topics using a parallel sets visualization. On the bottom, we show the user-topics represented by a) Histogram of the user's profession b) Histogram of the user's age. Both histograms are weighted by the number of instances that a user was allocated to a topic.

We applied Y^2 -LDA to the user-movie domain in order to extract *Which groups of users watch which types of movies*. The results are displayed in figure 4.6 We use the parallel sets visualization to display the correlation of topics ². In this experiment, We obtained a one to one relation between the user-topics and the movie-topics. The label of topics is based on our judgments. Our main findings indicate that users related to academia have higher preference towards drama and comedy movies, while users related to the fields of business and engineering have a higher interest in action movies. As expected, Professionals and older users naturally prefer to watch older movies.

Secondly, we tried this visualization on the NIPS dataset, we filtered those words that don't exhibit a particular meaning e.g. data, figure, image. The results are displayed in figure 4.7. In this experiment, we extracted six author topics and ten words topics to be able to compare with NIPS technical ³. As result, we observe high correlation between the extracted word topics and NIPS technical areas as the fields of Hardware Technologies, Learning theory, Neuroscience and Speech Recognition emerge as topics. Different from the expected NIPS Technical Areas, we find a topic categorization that falls into solving approaches rather than Technical Areas. For instance, probability, optimization and artificial neural networks emerge each as topics, while in NIPS they are all considered to be part of the Algorithms and Architectures technical area.

³<https://nips.cc/Conferences/2008/CallForPapers>

³<https://www.jasondavies.com/parallel-sets/>

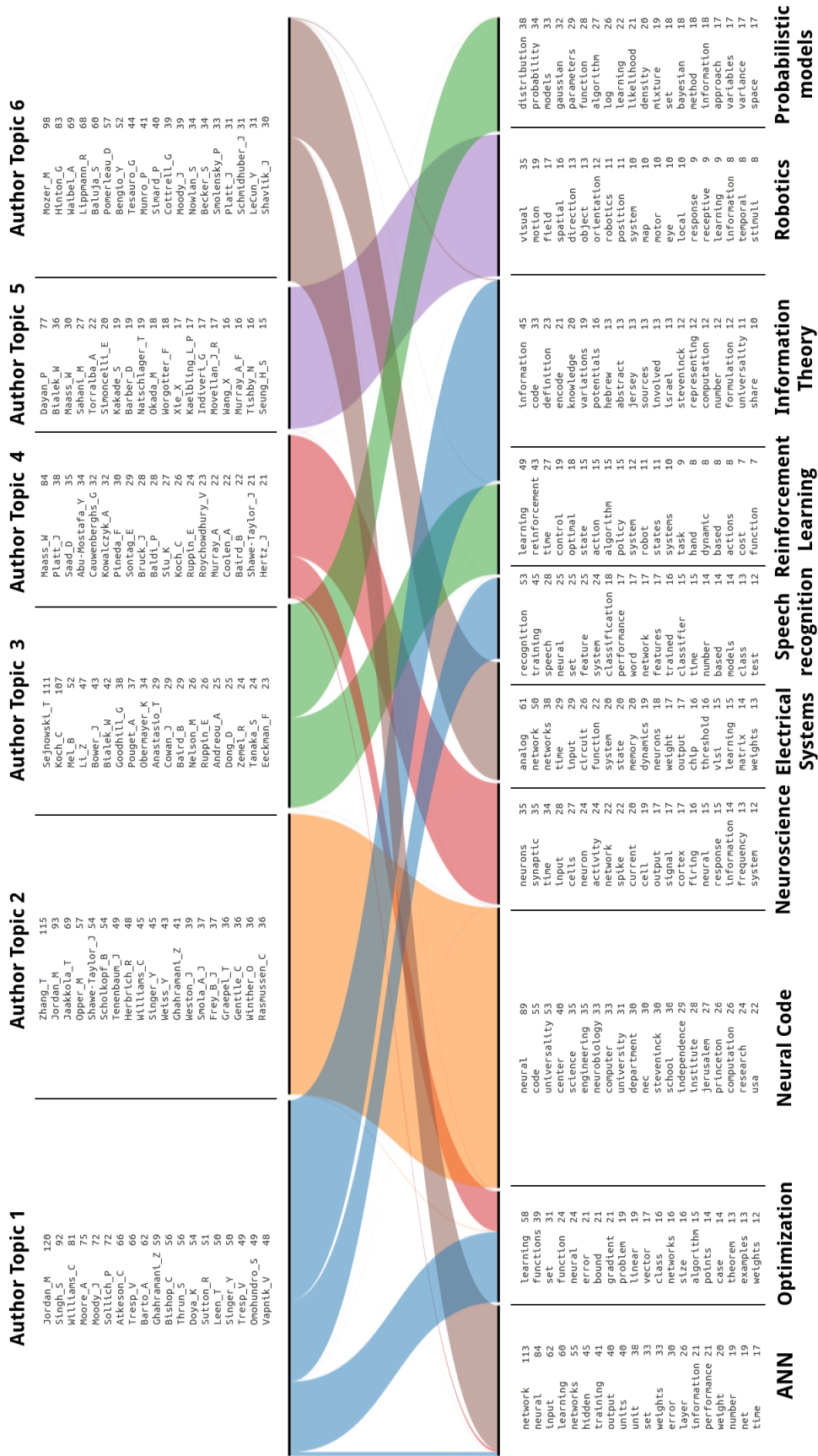


Fig. 4.7 Author-topics, word-topics and its interrelation. The width of the box surrounding each topic represents the total amount of samples allocated to the topic.

Chapter 5

Future Work

In this chapter, we discuss the future lines of research for my current work. In figure 5.1, we display a brief research plan with the expected allocated time for each of the future lines of work.

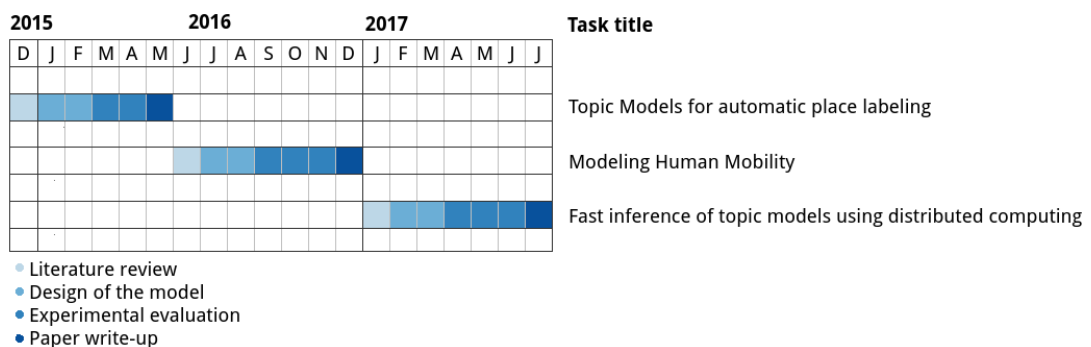


Fig. 5.1 Research Plan

5.1 Topic Models for Automatic Place Labeling

In recent years, researchers have been studying the problem of extracting semantic information from GPS data. i.e. labeling latitude and longitude tuples with semantics such home, lab, Mom's house, movie theater or bus station. Unfortunately, collecting GPS data is costly

in terms of energy consumption, Furthermore, in practice, users tend to use GPS sensors for brief moments of time.

In contrast, cellular networks can be used to collect location data with a higher sampling frequency and at a cheaper cost. Specially, since every connection between the phone and a cell tower automatically generates a new location record. As a disadvantage, the estimated user's location is about one order of magnitude worse than the estimated user's location using GPS.

In the literature, most of the approaches propose algorithms for extracting the semantics from GPS data. However, a few publications have been made towards extracting semantics from cellular tower records: From AT&T labs, we found an algorithm for extracting home and work from cellular data. Also, from Telefonica Research, we found a model for trajectory classification leading to the discovery of the popular roads used by the Spanish population.

We plan to develop a probabilistic model for labeling semantic information to the set of places obtained by using cellular tower records. This is a challenging task because of the high uncertainty of the individual's location. However, we believe that combining non-personalized approaches with personalized approaches will give a considerable improvement over the current baselines. From non-personalized approaches, we can consider the place popularity conditioned on the user preference as well as the probabilistic relationship between cellular towers to the set of labeled places. From a personalized perspective, we can use the user's historical visits to find a distribution over places given the user, current time and distance to the previous location. Such distribution is sparse and we will need to address that problem as well. Finally, we will combine these approaches into a unified model, this is also a difficult task given the lack of models in the literature that combine personalized models with non-personalized models.

5.2 Modeling Human Mobility

Singtel Dataset comprises data collected from more than 3 million users at an average sampling rate of approximately 12 samples per hour. Given that information, we have the possibility of modeling human mobility in Singapore. The problem consists in identifying the latent factors that drive the residents of Singapore to move inside the city. This problem is analogous to the language model, in which researchers try to identify the latent factors that writers use when redacting a document. In fact, most human mobility models are inspired by language models. For instance, LDA model is a unigram model that has been used for modeling the distribution of places given an user, the main assumption of this model is the fact that the order of places is irrelevant when selecting a location for a user. To relax this assumption, Ferrari et al, considered adapting an N-gram topic model, in their adaptation, the place selected for the user depends on the previous location as well as the places of interest as represent by the distribution of topics for the given user. However, this javascript:void(0);model does not consider temporal features. More recently, Hsung-Ping Hsieh et al have considered a T-Gram model to extend the N-gram models by considering time features. Other alternatives to model human mobility include probabilistic automata, transfer learning and non probabilistic methods.

In our model, we want to extend the topic models proposed in the state of the art by including additional features that have not been considered yet. For example, current models do not constrain the mobility of an individual by activity level, clearly every individual has different activity levels, depending on the individual, there will always be a different expected number of places that he might visit. In addition, current models do not consider the social relationships that take place between individuals or the transportation and leisure preference of an individual. By extending the state of the art with this information, we expect our model to be able to generate data that has higher similitude with the records obtained by location data providers.

5.3 Fast Inference of Topic Models using Distributed Computing

In our research, we have proposed a model for extracting user-topics, place-topics and their inter-relationship. Also, we plan to develop a topic model for labeling POIs, and another topic model for simulating human mobility. Though, topic models in general have been proven to be highly applicable, the inference of topic models is intractable and as consequence its analytic derivation as well as computational complexity are a known challenge.

As solution, Blei et al first proposed Variational Inference as a method for solving the analytic inference problem of the topic models, this method has been proved to be efficient in terms of speed, although there is no theoretical guarantees of convergence. Later, Griffiths and Steyvers proposed Collapsed Gibbs Sampling approach to solve the inference of topic models with a theoretical bounding on the approximation to the truth inference values but slower convergence. Despite other inference methods have been proposed Collapsed Gibbs Sampling is the most common approach used in practice. With respect to the computational efficiency, there exists contributions regarding parallel implementations of topic models using distributed computing. However, all the work remains tailored to a particular topic model. In order to advance faster in this research field, it might be beneficial to consider studying topic models as families of models rather than individual models. Teh et al. have proposed an inference method for the families of topic models that consist of discrete random variables with Dirichlet priors on the parameters, however there is no existing work extending their work towards a distributed approach.

Because the topic models proposed in our research belong to this family of topic models, we propose an architecture for scalable inference of the family of topic models consisting of discrete random variables with Dirichlet priors on categorical parameters. The benefits include the scalability of the topic models used in our research as well as common topic

models such as the author-topic model, the Pachinko allocation model, and the Mixed membership stochastic block models for relational data.

References

- [1] Bao, J., Zheng, Y., and Mokbel, M. F. (2012). Location-based and preference-aware recommendation using sparse geo-social networking data. In *Proceedings of the 20th International Conference on Advances in Geographic Information Systems, SIGSPATIAL '12*, pages 199–208, New York, NY, USA. ACM.
- [2] Bauer, S., Noulas, A., Seaghdha, D. O., Clark, S., and Mascolo, C. (2012). Talking places: Modelling and analysing linguistic content in foursquare. In *SocialCom/PASSAT*, pages 348–357. IEEE.
- [3] Bicego, M., Lovato, P., Ferrarini, A., and Delledonne, M. (2010). Biclustering of expression microarray data with topic models. In *Pattern Recognition (ICPR), 2010 20th International Conference on*, pages 2728–2731.
- [4] Blei, D. M. (2012a). Probabilistic topic models. *Commun. ACM*, 55(4):77–84.
- [5] Blei, D. M. (2012b). Probabilistic topic models. *Commun. ACM*, 55(4):77–84.
- [6] Blei, D. M., Griffiths, T. L., and Jordan, M. I. (2010). The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. *J. ACM*, 57(2):7:1–7:30.
- [7] Blei, D. M. and Lafferty, J. D. (2006a). Correlated topic models. In *In Proceedings of the 23rd International Conference on Machine Learning*, pages 113–120. MIT Press.
- [8] Blei, D. M. and Lafferty, J. D. (2006b). Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning, ICML '06*, pages 113–120, New York, NY, USA. ACM.
- [9] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022.
- [10] Bobadilla, J., Ortega, F., Hernando, A., and Gutiérrez, A. (2013). Recommender systems survey. *Know.-Based Syst.*, 46:109–132.
- [11] Chandra, B., Shanker, S., and Mishra, S. (2006). A new approach: Interrelated two-way clustering of gene expression data. *Statistical Methodology*. Bioinformatics.
- [12] Cheng, Y. and Church, G. M. (2000). Biclustering of expression data. In *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology*, pages 93–103. AAAI Press.

- [13] Collins, M., Dasgupta, S., and Schapire, R. E. (2002). A generalization of principal components analysis to the exponential family. In Dietterich, T., Becker, S., and Ghahramani, Z., editors, *Advances in Neural Information Processing Systems 14*, pages 617–624. MIT Press.
- [14] Dhillon, I. S. (2001). Co-clustering documents and words using bipartite spectral graph partitioning. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '01, pages 269–274, New York, NY, USA. ACM.
- [15] Du, L., Buntine, W., and Jin, H. (2010). Sequential latent dirichlet allocation: Discover underlying topic structures within a document. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, pages 148–157.
- [16] Eren, K., Deveci, M., Küçüktunç, O., and Çatalyürek, Ü. V. (2013). A comparative analysis of biclustering algorithms for gene expression data. *Briefings in Bioinformatics*, 14(3):279–292.
- [17] Escobar, M. D. and West, M. (1994). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90:577–588.
- [18] Farrahi, K. and Gatica-Perez, D. (2011). Discovering routines from large-scale human locations using probabilistic topic models. *ACM Trans. Intell. Syst. Technol.*, 2(1):3:1–3:27.
- [19] Farrahi, K. and Gatica-Perez, D. (2012). Extracting mobile behavioral patterns with the distant n-gram topic model. In *Wearable Computers (ISWC), 2012 16th International Symposium on*, pages 1–8.
- [20] Farrahi, K. and Gatica-Perez, D. (2014). A probabilistic approach to mining mobile phone data sequences. *Personal Ubiquitous Comput.*, 18(1):223–238.
- [21] Ferrari, L., Rosi, A., Mamei, M., and Zambonelli, F. (2011). Extracting urban patterns from location-based social networks. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks*, LBSN '11, pages 9–16, New York, NY, USA. ACM.
- [22] Fox, C. and Roberts, S. (2012). A tutorial on variational Bayesian inference. *Artificial Intelligence Review*, 38(2):85–95.
- [23] Fox, E., Sudderth, E., Jordan, M., and Willsky, A. (2011). Bayesian nonparametric inference of switching dynamic linear models. *Signal Processing, IEEE Transactions on*, 59(4):1569–1585.
- [24] Fox, E. B., Sudderth, E. B., Jordan, M. I., and Willsky, A. S. (2008). An hdp-hmm for systems with state persistence. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pages 312–319, New York, NY, USA. ACM.
- [25] Getz, G., Levine, E., and Domany, E. (2000). Coupled two-way clustering analysis of gene microarray data. *Proceedings of the National Academy of Sciences*, (22):12079–12084.

- [26] Ghahramani, Z. and Beal, M. J. (2000). Variational inference for bayesian mixtures of factor analysers. In *In Advances in Neural Information Processing Systems 12*, pages 449–455. MIT Press.
- [27] Ghahramani, Z. and Jordan, M. I. (1997). Factorial hidden markov models. *Mach. Learn.*, 29(2-3):245–273.
- [28] Griffiths, T. L. and Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(Suppl. 1):5228–5235.
- [29] Griffiths, T. L., Steyvers, M., Blei, D. M., and Tenenbaum, J. B. (2005). Integrating topics and syntax. In Saul, L., Weiss, Y., and Bottou, L., editors, *Advances in Neural Information Processing Systems 17*, pages 537–544. MIT Press.
- [30] Gruber, A., Weiss, Y., and Rosen-Zvi, M. (2007). Hidden topic markov models. In *AISTATS*, volume 2 of *JMLR Proceedings*, pages 163–170. JMLR.org.
- [31] Hartigan, J. A. (1972). Direct clustering of a data matrix. 67(337):123–129.
- [32] Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. (2013). Stochastic variational inference. *J. Mach. Learn. Res.*, 14(1):1303–1347.
- [33] Hofmann, T. (1999). Probabilistic latent semantic analysis. In *In Proc. of Uncertainty in Artificial Intelligence, UAI’99*, pages 289–296.
- [34] Hörster, E., Lienhart, R., and Slaney, M. (2007). Image retrieval on large-scale image databases. In *Proceedings of the 6th ACM International Conference on Image and Video Retrieval, CIVR ’07*, pages 17–24, New York, NY, USA. ACM.
- [35] Ihmels, J., Bergmann, S., and Barkai, N. (2004). Defining transcription modules using large-scale gene expression data. *Bioinformatics*, 20:2004.
- [36] Ishwaran, H. and James, L. F. (2001). Gibbs Sampling Methods for Stick-Breaking Priors. *Journal of the American Statistical Association*, 96(453):161–173.
- [37] Joseph, J., Doshi-Velez, F., Huang, A. S., and Roy, N. (2011). A bayesian nonparametric approach to modeling motion patterns. *Auton. Robots*, 31(4):383–400.
- [38] Kalman, R. E. (1960). A New Approach to Linear Filtering and Prediction Problems. *Transactions of the ASME – Journal of Basic Engineering*, (82 (Series D)):35–45.
- [39] Lazzeroni, L. and Owen, A. (2000). Plaid models for gene expression data. *Statistica Sinica*, 12:61–86.
- [40] Li, F.-F. and Perona, P. (2005). A bayesian hierarchical model for learning natural scene categories. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05) - Volume 2 - Volume 02*, CVPR ’05, pages 524–531, Washington, DC, USA. IEEE Computer Society.
- [41] Li, W. and McCallum, A. (2006). Pachinko allocation: Dag-structured mixture models of topic correlations. In *Proceedings of the 23rd International Conference on Machine Learning, ICML ’06*, pages 577–584, New York, NY, USA. ACM.

- [42] Liu, B., Fu, Y., Yao, Z., and Xiong, H. (2013). Learning geographical preferences for point-of-interest recommendation. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, pages 1043–1051, New York, NY, USA. ACM.
- [43] Liu, B. and Xiong, H. (2013). Point-of-interest recommendation in location based social networks with topic and location awareness. In *Proceedings of the 13th SIAM International Conference on Data Mining, May 2-4, 2013. Austin, Texas, USA.*, pages 396–404.
- [44] Long, X., Jin, L., and Joshi, J. (2012). Exploring trajectory-driven local geographic topics in foursquare. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, UbiComp '12, pages 927–934, New York, NY, USA. ACM.
- [45] Ma, Qin M, T. H. T. P. A. Y. X. (2009). Qubic: a qualitative biclustering algorithm for analyses of gene expression data. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*.
- [46] Madeira, S. and Oliveira, A. (2004). Biclustering algorithms for biological data analysis: a survey. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, 1(1):24–45.
- [47] Mahfouz, M. and Ismail, M. (2012). Soft flexible overlapping biclustering utilizing hybrid search strategies. In Hassanien, A., Salem, A.-B., Ramadan, R., and Kim, T.-h., editors, *Advanced Machine Learning Technologies and Applications*, volume 322 of *Communications in Computer and Information Science*, pages 315–326. Springer Berlin Heidelberg.
- [48] Marlin, B. (2003). Modeling user rating profiles for collaborative filtering. In *In NIPS*17*. MIT Press.
- [49] McInerney, J., Zheng, J., Rogers, A., and Jennings, N. R. (2013a). Modelling heterogeneous location habits in human populations for location prediction under data sparsity. In *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '13, pages 469–478, New York, NY, USA. ACM.
- [50] McInerney, J., Zheng, J., Rogers, A., and Jennings, N. R. (2013b). Modelling heterogeneous location habits in human populations for location prediction under data sparsity. In *International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp)*, pages 469–478.
- [51] Mei, Q., Liu, C., Su, H., and Zhai, C. (2006). A probabilistic approach to spatiotemporal theme pattern mining on weblogs. In *Proceedings of the 15th International Conference on World Wide Web*, WWW '06, pages 533–542, New York, NY, USA. ACM.
- [52] Minka, T. and Lafferty, J. (2002). Expectation-propagation for the generative aspect model. In *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence*, UAI'02, pages 352–359, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

- [53] Misra, H., Cappé, O., and Yvon, F. (2008). Using lda to detect semantically incoherent documents. In *Proceedings of the Twelfth Conference on Computational Natural Language Learning, CoNLL '08*, pages 41–48, Stroudsburg, PA, USA. Association for Computational Linguistics.
- [54] Murali, T. M. and Kasif, S. (2003). Extracting conserved gene expression motifs from gene expression data. In *Pac. Symp. Biocomput.*, pages 77–88.
- [55] Paisley, J. and Carin, L. (2009). Nonparametric factor analysis with beta process priors. In *Proceedings of the 26th Annual International Conference on Machine Learning, ICML '09*, pages 777–784, New York, NY, USA. ACM.
- [56] Prelić, A., Bleuler, S., Zimmermann, P., Wille, A., Bühlmann, P., Gruissem, W., Hennig, L., Thiele, L., and Zitzler, E. (2006). A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics*, 22(9):1122–1129.
- [57] Ranganath, R., Gerrish, S., and Blei, D. M. (2014). Black box variational inference. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics, AISTATS 2014, Reykjavik, Iceland, April 22-25, 2014*, pages 814–822.
- [58] Rosen-Zvi, M., Griffiths, T., Steyvers, M., and Smyth, P. (2004). The author-topic model for authors and documents. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence, UAI '04*, pages 487–494, Arlington, Virginia, United States. AUAI Press.
- [59] Roy, D. M. and Teh, Y. W. (2009). The mondrian process. In Koller, D., Schuurmans, D., Bengio, Y., and Bottou, L., editors, *Advances in Neural Information Processing Systems 21*, pages 1377–1384.
- [60] Shan, H. and Banerjee, A. (2008). Bayesian co-clustering. In *Data Mining, 2008. ICDM '08. Eighth IEEE International Conference on*, pages 530–539.
- [61] Sivic, J., Russell, B. C., Efros, A. A., Zisserman, A., and Freeman, W. T. (2005). Discovering object categories in image collections. In *Proceedings of the International Conference on Computer Vision*.
- [62] Sontag, D. and Roy, D. (2011). Complexity of inference in latent dirichlet allocation. In Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira, F., and Weinberger, K., editors, *Advances in Neural Information Processing Systems 24*, pages 1008–1016. Curran Associates, Inc.
- [63] Steyvers, M. and Griffiths, T. (2007). *Probabilistic Topic Models*. Lawrence Erlbaum Associates.
- [64] Tanay, A., Sharan, R., and Shamir, R. (2002). Discovering statistically significant biclusters in gene expression data. In *ISMB*, pages 136–144.
- [65] Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2004). Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101.
- [66] Teh, Y. W., Newman, D., and Welling, M. (2007). A collapsed variational bayesian inference algorithm for latent dirichlet allocation. In Schölkopf, B., Platt, J., and Hoffman, T., editors, *Advances in Neural Information Processing Systems 19*, pages 1353–1360. MIT Press.

- [67] Wallach, H. M. (2006). Topic modeling: Beyond bag-of-words. In *Proceedings of the 23rd International Conference on Machine Learning, ICML '06*, pages 977–984, New York, NY, USA. ACM.
- [68] Wallach, H. M., Murray, I., Salakhutdinov, R., and Mimno, D. (2009). Evaluation methods for topic models. In *Proceedings of the 26th Annual International Conference on Machine Learning, ICML '09*, pages 1105–1112, New York, NY, USA. ACM.
- [69] Wang, C., Blei, D. M., and Heckerman, D. (2008). Continuous time dynamic topic models. In McAllester, D. A. and Myllymäki, P., editors, *UAI*, pages 579–586. AUAI Press.
- [70] Wang, C., Wang, J., Xie, X., and Ma, W.-Y. (2007a). Mining geographic knowledge using location aware topic model. In *Proceedings of the 4th ACM Workshop on Geographical Information Retrieval, GIR '07*, pages 65–70, New York, NY, USA. ACM.
- [71] Wang, X. and McCallum, A. (2006). Topics over time: A non-markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '06*, pages 424–433, New York, NY, USA. ACM.
- [72] Wang, X., McCallum, A., and Wei, X. (2007b). Topical n-grams: Phrase and topic discovery, with an application to information retrieval. In *Proceedings of the 2007 Seventh IEEE International Conference on Data Mining, ICDM '07*, pages 697–702, Washington, DC, USA. IEEE Computer Society.
- [73] Wei, X. and Croft, W. B. (2006). Lda-based document models for ad-hoc retrieval. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '06*, pages 178–185, New York, NY, USA. ACM.
- [74] Wei, X., Sun, J., and Wang, X. (2007). Dynamic mixture models for multiple time series. In *Proceedings of the International Joint Conference on Artificial Intelligence*, volume 459, pages 2909–2914.
- [75] Yang, D., Zhang, D., Zheng, V., and Yu, Z. (2015). Modeling user activity preference by leveraging user spatial temporal characteristics in lbsns. *Systems, Man, and Cybernetics: Systems, IEEE Transactions on*, 45(1):129–142.
- [76] Yin, Z., Cao, L., Han, J., Zhai, C., and Huang, T. (2011). Geographical topic discovery and comparison. In *Proceedings of the 20th International Conference on World Wide Web, WWW '11*, pages 247–256, New York, NY, USA. ACM.
- [77] Yuan, J., Zheng, Y., and Xie, X. (2012). 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, KDD '12*, pages 186–194, New York, NY, USA. ACM.
- [78] Yuan, Q., Cong, G., Ma, Z., Sun, A., and Thalmann, N. M. (2013). Who, where, when and what: Discover spatio-temporal topics for twitter users. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '13*, pages 605–613, New York, NY, USA. ACM.

-
- [79] Zhang, J., Song, Y., Zhang, C., and Liu, S. (2010). Evolutionary hierarchical dirichlet processes for multiple correlated time-varying corpora. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '10, pages 1079–1088, New York, NY, USA. ACM.
 - [80] Zheng, Y., Zhang, L., Xie, X., and Ma, W.-Y. (2009). Mining interesting locations and travel sequences from gps trajectories. In *Proceedings of the 18th International Conference on World Wide Web*, WWW '09, pages 791–800, New York, NY, USA. ACM.

Appendix A

Perplexity of an Individual Topic when Using Y^2 -LDA

To compare GLDA's perplexity with an LDA model, we need to extract the user-topics $P(u|\Phi_u, \alpha)$ and the place-topics $P(p|\Phi_p, \alpha)$ separately. Because these distributions are intractable, we use Chib-Estimation method [68] to obtain an approximate result. The estimation method requires that we derive analytically $P(u, z_u|\Phi_u, \alpha)$, and $P(p, z_p|\Phi_p, \alpha)$.

We can get the former distribution since $P(u, z_u|\Phi_u, \alpha) = P(u|\Phi_u, z_u)P(z_u|\alpha)$. We can get $P(u|\Phi_u, z_u)$ from the inference process. With respect to $P(z_u|\alpha)$, we can start by

writing out $P(z_u, z_p|\alpha) = \int_{\Theta} P(z_u, z_p|\theta)P(\theta|\alpha)$ which can be re-written as $P(z_u, z_p|\alpha) = \int_{\Theta_i} \int_{\Theta_j} P(z_u|\Theta_j)P(\Theta_j|\alpha)P(z_p|\Theta_i)P(\Theta_i|\alpha)$. Finally, we to obtain $P(z_u|\alpha) = \int_{\Theta_j} P(z_u|\Theta_j)P(\Theta_j|\alpha) \propto$

$$\frac{\prod_{i=1}^{K_p} (\Gamma(c(i, z_u, *, *)^{-v} + \alpha_{z_u \times K_p + i}))}{\Gamma(\sum_{i=1}^{K_u} (c(i, k_u, *, *)^{-v} + \alpha_{z_u \times K_p + i}))}$$