

Probabilistic Models for Context in Social Media

Novel Approaches and Inference Schemes

Christoph Carl Kling

Institute for Web Science and Technologies
University of Koblenz–Landau
ckling@uni-koblenz.de

November 2016

Vom Promotionsausschuss des Fachbereichs 4: Informatik der
Universität Koblenz–Landau zur Verleihung des akademischen Grades

Doktor der Naturwissenschaften (Dr. rer. nat.)

genehmigte Dissertation.

PhD thesis at the University of Koblenz-Landau

Datum der wissenschaftlichen Aussprache:	16.11.2016
Vorsitz des Promotionsausschusses	Prof. Dr. Ralf Lämmel
Berichterstatter:	Prof. Dr. Steffen Staab
Berichterstatter:	Prof. Dr. Markus Strohmaier
Berichterstatter:	Prof. Dr. Lars Schmidt-Thieme

Abstract

This thesis presents novel approaches for integrating context information into probabilistic models.

Data from social media is typically associated with metadata, which includes context information such as timestamps, geographical coordinates or links to user profiles. Previous studies showed the benefits of using such context information in probabilistic models, e.g. improved predictive performance. In practice, probabilistic models which account for context information still play a minor role in data analysis. There are multiple reasons for this. Existing probabilistic models often are complex, the implementation is difficult, implementations are not publicly available, or the parameter estimation is computationally too expensive for large datasets. Additionally, existing models are typically created for a specific type of content and context and lack the flexibility to be applied to other data.

This thesis addresses these problems by introducing a general approach for modelling multiple, arbitrary context variables in probabilistic models and by providing efficient inference schemes and implementations.

In the first half of this thesis, the importance of context and the potential of context information for probabilistic modelling is shown theoretically and in practical examples. In the second half, the example of topic models is employed for introducing a novel approach to context modelling based on document clusters and adjacency relations in the context space. These models allow for the first time the efficient, explicit modelling of arbitrary context variables including cyclic and spherical context (such as temporal cycles or geographical coordinates). Using the novel three-level hierarchical multi-Dirichlet process presented in this thesis, the adjacency of context clusters can be exploited and multiple contexts can be modelled and weighted at the same time. Efficient inference schemes are derived which yield interpretable model parameters that allow analyse the relation between observations and context.

Zusammenfassung

Diese Arbeit stellt neue Verfahren zur Integration von Kontext-Information in probabilistische Modelle vor.

Daten aus sozialen Medien sind oft mit Metadaten assoziiert, wie beispielsweise Uhrzeit, geographische Koordinaten oder Nutzerdaten. Frühere Studien haben die Bedeutung von Kontextinformationen für die Vorhersagekraft von Wahrscheinlichkeitsmodellen gezeigt. In der Praxis spielen Wahrscheinlichkeitsmodelle die Kontextinformationen integrieren dennoch eine geringe Rolle. Die Gründe hierfür sind vielfältig: Vorhandene Wahrscheinlichkeitsmodelle sind oft komplex, die Implementierung ist schwierig, Implementierungen sind nicht frei verfügbar oder die Parameterschätzung ist für größere Datensätze zu aufwendig. Dazu kommt, dass vorhandene Modelle typischerweise für eine spezielle Art von Kontextinformation angepasst sind und nicht ohne Weiteres auf andere Daten angewendet werden können.

Diese Arbeit stellt einen neuen Ansatz zur Modellierung von mehreren, beliebigen Kontext-Variablen in Wahrscheinlichkeitsmodellen vor und präsentiert effiziente Inferenz-Strategien.

In der ersten Hälfte der Arbeit wird die Bedeutung von Kontextinformationen anhand theoretischer und praktischer Beispiele gezeigt. In der zweiten Hälfte werden neue Topic-Modelle vorgestellt, die Kontextinformationen anhand von Dokument-Clustern und Nachbarschaftsbeziehungen im Kontext-Raum modellieren. Diese Topic-Modelle erlauben erstmals die effiziente, explizite Modellierung von zyklischen Kontextvariablen und Kontextvariablen mit Verteilung auf der n -Sphäre (beispielsweise von Zeit-Zyklen und geographisch verteilten Variablen). Mit Hilfe des neuartigen *Hierarchical Multi-Dirichlet Process* (HMDP), der in dieser Arbeit vorgestellt wird, können Nachbarschaftsbeziehungen zwischen Kontext-Clustern ausgenutzt werden und es können mehrere Kontexte gleichzeitig modelliert und gewichtet werden. Es werden effiziente Inferenz-Strategien hergeleitet, die interpretierbare Modell-Parameter liefern, welche die Analyse der Beziehung zwischen Beobachtungen und Kontext-Informationen erlauben.

Acknowledgments

This dissertation would not have been possible without the help and support of many friends and colleagues.

I thank my friend and colleague Jérôme Kunegis who supported me throughout my PhD studies and helped shaping my ideas. I thank my supervisor Steffen Staab for giving me the freedom and time to get familiar with probabilistic topic modelling. Markus Strohmaier supported me in (finally) writing the paper on online democracies and gave valuable feedback. Arnim Bleier helped me a lot by pointing me to related work and by introducing me to online inference techniques. Heinrich Hartmann guided my research on online democracies in the right direction. Sergej Sizov supported me in the early stages of my research and motivated me to do topic modelling.

I thank Lisa Posch, Jérôme Kunegis and Arnim Bleier for proofreading this thesis and for many helpful comments.

Dirk Homscheid provided the dataset of the Linux kernel mailing list, and Damien Fay provided the user profiles of the online network for sexual fetishists.

I thank my colleagues at the Institute WeST of the University of Koblenz–Landau and at the Department of Computational Social Science at GESIS in Cologne for many valuable discussions, especially on the study of online democracies.

Finally, I want to thank my family for their support and encouragement.

Contents

Introduction	3
The Importance of Context	3
Probabilistic Models and Prior Distributions	4
Graphical Models and Topic Models	5
Applications	6
Contributions	7
1 Foundations and Related Work	11
1.1 The Concept of Probability	11
1.2 Basic Concepts and Distributions	12
1.2.1 Bernoulli distribution	13
1.2.2 Likelihood	13
1.2.3 Binomial distribution	14
1.2.4 Marginal probability	14
1.2.5 Maximum likelihood estimation	14
1.3 Bayesian Networks	15
1.3.1 Dependencies in Bayesian networks	17
1.3.2 Plate notation	18
1.4 Priors and Bayesian Inference	18
1.4.1 Beta distribution	19
1.4.2 Maximum a posteriori estimation	20
1.4.3 Bayesian inference	22
1.4.4 Multinomial distribution	23
1.4.5 Dirichlet distribution	24
1.4.6 Maximum a posteriori estimation and Bayesian inference for the multinomial distribution	25
1.4.7 Fisher distribution and approximate inference	25
1.4.8 Latent variables and expectation-maximisation	28
1.5 Topic Models	30
1.5.1 Latent semantic analysis	31
1.5.2 Probabilistic latent semantic analysis	32
1.5.3 Latent Dirichlet allocation	33
1.6 Inference for Complex Graphical Models	34

1.6.1	Expectation-maximisation and maximum a posteriori inference	34
1.6.2	Gibbs sampling	35
1.6.3	Variational inference	37
1.7	Evaluation of Topic Models	40
1.7.1	Human evaluation	40
1.7.2	Perplexity	40
1.8	Non-Parametric Topic Models	41
1.8.1	The Dirichlet process	42
1.8.2	Inference for the HDP	46
1.9	Summary	51
2	Single-Context Voting Models	53
2.1	Problem Setting and Approach	53
2.1.1	Problem setting	54
2.1.2	Approach	55
2.2	Delegative Democracies	55
2.2.1	Democracy platforms	56
2.2.2	Pirate parties	56
2.3	Description of the Dataset	56
2.3.1	LiquidFeedback platform	56
2.3.2	Dataset	57
2.4	Voting Behaviour and Delegations	57
2.4.1	Existence and role of super-voters	57
2.4.2	User approval rates	60
2.4.3	Impact of delegations	62
2.4.4	Temporal analysis of the delegation network	64
2.5	Power in Online Democracies	64
2.5.1	Power indices	65
2.5.2	Probabilistic interpretation of power indices	66
2.5.3	Theoretical (uniform) power indices	67
2.5.4	Empirical power	68
2.5.5	Non-uniform power indices	70
2.5.6	Evaluation	74
2.5.7	Distribution of Power	75
2.6	Discussion	76
2.7	Summary	76
3	Single-Context Topic Models	79
3.1	A Classification of Context Variables	80
3.1.1	Discrete context variables	80
3.1.2	Linear and continuous context variables	80
3.1.3	Cyclic and spherical context variables	80
3.2	Single-Context Topic Models	81

3.2.1	Topic models for discrete context variables	82
3.2.2	Topic models for linear context variables	84
3.2.3	Topic models for cyclic and spherical context variables	84
3.2.4	Categorisation of approaches	87
3.3	Drawbacks of Existing Models	87
3.4	Multi-Dirichlet Process Topic Models	90
3.5	The Basic Model	91
3.5.1	Geographical clustering	91
3.5.2	Topic detection	92
3.6	The Neighbour-Aware Model	94
3.6.1	Definition of a geographical network	95
3.6.2	Advantages of the neighbour-aware model	95
3.7	The MDP Based Geographical Topic Model	97
3.7.1	The multi-Dirichlet process	98
3.7.2	Inference	99
3.7.3	Estimation of scaling parameters for the MDP	100
3.7.4	MDP-based topic model	101
3.7.5	Generative process	101
3.7.6	Generalised estimator of the concentration parameter	103
3.8	Evaluation	106
3.8.1	Datasets	106
3.8.2	Experimental setting	107
3.8.3	Comparison with LGTA	108
3.8.4	Effect of the region parameter	110
3.8.5	User study	110
3.8.6	Runtime comparison	114
3.9	Summary	115
4	Multi-Context Topic Models	119
4.1	Generalisations and Special Cases of MGTM	119
4.1.1	Relation to the hierarchical Dirichlet process	120
4.1.2	Relation to the author topic model	120
4.1.3	Relation to the citation influence model	120
4.1.4	A generalisation of the HMDP for arbitrary contexts	121
4.2	Existing Multi-Context Topic Models	121
4.2.1	Dirichlet-multinomial regression	121
4.2.2	Topic models for context-specific topics	122
4.2.3	Distance-based topic models for multiple contexts	122
4.3	Hierarchical MDPs for Multiple Contexts	123
4.3.1	The HMDP Topic Model	124
4.4	Efficient Inference	127
4.4.1	Collapsed variational Bayes for the HMDP	128
4.4.2	Online inference	134
4.5	Applications of the HMDP	137

4.5.1	Improved inference	137
4.5.2	Modelling user profiles of a social network for fetishists	144
4.5.3	Context selection and analysis with the HMDP – a study of the Linux kernel mailing list	149
4.6	Summary	156
Conclusion		157
	Findings	158
	Outlook	159
Bibliography		169
A Appendix		171
A.1	MLE Estimate for the Binomial Distribution	171
A.2	MAP Estimate for the Binomial Distribution	172
A.3	Details PCSVB for HMDP	173
A.4	ML Estimate Scaling Parameter α_1	175
A.5	Topic Descriptions	175

Introduction

Man kann für eine *große* Klasse von Fällen der Benützung des Wortes “Bedeutung” – wenn auch nicht für *alle* Fälle seiner Benützung – dieses Wort so erklären:
Die Bedeutung eines Wortes ist sein Gebrauch in der Sprache.

Ludwig Wittgenstein, Urfassung Philosophische Untersuchungen, 40

The Importance of Context

While the amount of social media data analysed with probabilistic models is growing, one particular kind of information is still neglected in many data mining applications: *the context, in which data was created.*

The location in which a photography is taken can affect the interpretation of its content. The political system, in which votes are cast, might influence the behaviour of voters and should be accounted for in an analysis of voting systems. The metadata of an email storing the time at which an email was written might give a hint about the topics covered in that email.

It is surprising that on the one hand nearly all social media content is associated with metadata and, on the other hand, most probabilistic models applied to social media data do not account for context information. The reasons for this development include the high complexity of available models, the resulting complexity of the inference together with a lack of freely available implementations, and missing scalability for large datasets. Additionally, for the important class of probabilistic topic models, there is no efficient method for explicitly modelling cyclic or spherical metadata. This is important, as the most-common context information in social media – the timestamp of a document – contains cycles such as the daily 24h cycle, the weekly cycle and the yearly cycle. A large share of data from social media is additionally associated with geographical coordinates.

Since more metadata stemming e.g. from mobile sensors or user profiles is available for an increasing share of documents in the web, the question of how to include this rich information in probabilistic models is pressing.

Problem and Approach

Especially if there exist complex dependencies between context variables and observations, existing methods for including context information into probabilistic models are either not capable of detecting these structures, the resulting models are overly complex so that an interpretation or a plausibility check of parameters is impossible, or the inference costs prevent the application on large datasets.

In this thesis, the problem of including context information in probabilistic models is studied in three parts. **(i) System context models.** The inclusion of a system-context – the properties of a given system such as a online platform – is studied on the example of probabilistic power indices which include the properties of a democratic platform. **(ii) Single-context models.** The problem of including a single context variable into probabilistic models is studied on the example of geographical topic models. **(iii) Multi-context models.** The multi-context case – where multiple context information is available – is studied on the example of probabilistic topic models conditioned on multiple, arbitrary metadata.

For modelling the influence of arbitrary context information on mixture models, a novel class of probabilistic models is presented which uses so-called *Multi-Dirichlet Processes* (MDP). The properties of MDP-based models allow the derivation of efficient inference schemes, which are derived and implemented.

Probabilistic Models

Probability theory provides a powerful framework for creating models which explain real-world observations. Modern philosophers of science like Karl Popper have argued that inductive science – which creates hypotheses based on observations – does not prove hypotheses, but makes them more truth-likely by testing them in experiments [Pop79]. It therefore is possible that a new observation, which was not predicted by existing hypotheses, falsifies a theory, even if the theory was successfully tested before in thousands of experiments. In this case, existing theories can often be extended or updated, to account for novel observations.

Probabilistic models provide the framework to express such uncertainties and provide natural ways for assigning probabilities to previously non-observed events. It even is possible to account for novel observations by updating the parameters of a model or by changing parts of the model structure [KF09]. Therefore, it seems natural to employ probabilistic models for expressing theories about real-world observations.

The mechanisms for including the possibility for previously unseen observations into probabilistic models are based on *prior distributions*, commonly

referred to as *priors*. Prior distributions introduce probability distributions over parameters of a model, expressing beliefs about the probability of parameter settings which can be independent of observations.

Prior distributions also allow to model the influence of context information during parameter inference: The context of an observation can be linked to a belief on expected model parameters. Therefore, based on a given context, prior distributions can be constructed to predict the parameters of new observations. The problem of including context information into probabilistic models in many cases involves the modelling and learning of the right prior distributions to express beliefs about model parameters.

Bayesian Networks and Topic Models

An important class of probabilistic models are Bayesian networks, which are directed graphical models. Graphical models allow for expressing complex causal structures, potentially combining thousands of probability functions, which explain a generative process behind observed data.

Graphical models not only allow for coding complex relations between observed and hidden parameters and variables, but also provide a high flexibility in their application. Significant parts of this thesis are about probabilistic topic models – models which explain co-occurrences of grouped observations, which are typically applied to model words grouped by documents, e.g. words in newspaper articles [BNJ03]. However, the very same model structure can be e.g. employed to group bird species based on their genotype data [PSD00]. And by replacing the multinomial distributions behind topics with other distributions, arbitrary mixture models can be constructed using exactly the same structural properties and inference structure (e.g. Gaussian mixture models).

Therefore, while the presented methods for including context information in probabilistic models are applied for modelling social media data, their field of application is much broader.

Structure

In chapter 1, the foundations of probabilistic modelling are explained with a focus on topic modelling.

In chapter 2 of this thesis, probabilistic power indices for predicting the power of voters in online democracies are presented, which include the context in which votes are cast by modelling system-specific voting bias. Power indices are crucial to assess the distribution of power in novel online democracy systems, which have the potential to change the political processes of the future.

In Chapter 3, a classification of metadata variables into *context classes* is introduced. Using the structural properties of common context classes, probabilistic models which include context information based on mixtures of prior distributions are presented. Additionally, mixtures of Dirichlet processes, called *Multi-Dirichlet Processes (MDP)* are described and a Gibbs-sampling-based inference scheme is derived. Based on adjacency-networks of document clusters, hierarchies of multi-Dirichlet processes are employed to model the influence of geographical metadata on topics in a document collection. The impact of this modelling – a *dynamic smoothing* of context clusters and an improved sharing of topic information between adjacent context clusters – is described and evaluated.

Chapter 4 presents a generalisation of hierarchical multi-Dirichlet process models which allows to model the influence of multiple, arbitrary metadata such as demographic variables, temporal information (including temporal cycles) and geographical locations. Again, a mixture of prior distributions on the document-topic distributions is constructed in a mixture of multi-Dirichlet processes. Using the properties of multi-Dirichlet process and by applying multiple approximations – namely a practical collapsed stochastic variational Bayes inference with a zeroth-order Taylor approximation of counts – efficient inference strategies for multi-Dirichlet processes are derived.

An overview over the structure of this thesis is given in Figure 1.

Applications

Integrating context information (e.g. from metadata) into probabilistic models serves several purposes:

Data analysis. The presented topic models exploit context information for extracting topics of higher quality from document collections. The topics can provide insights into the topic distribution of a corpus or of single documents. Additionally, the relation between context variables and topics can be directly visualised and analysed using the interpretable model parameters.

Context selection. Additionally, the importance of context variables for topic prediction is learned during inference, so that less important variables can be automatically excluded during inference. The weighting of context variables can be visualised and interpreted.

Prediction. The detected relations between context variables and document topics can be used to predict observations (i.e. words in documents) for a given context such as the geographical location, the day of the week

in the weekly cycle or user data. This can be e.g. exploited in recommender systems. Additionally, the metadata of a given document can be predicted.

Information retrieval. As the inference schemes for the presented methods are efficient, allow a distributed computation and support online inference, the models can be used to extract topics of large document collections. The topics then can serve as input for topic-based retrieval algorithms [WC06].

Contributions

All novel methods and implementations presented in this thesis are own contributions.

Jérôme Kunegis, Heinrich Hartmann, Markus Strohmaier and Steffen Staab contributed text for the study on power indices in Chapter 2, which was published at the International Conference on Web and Social Media (ICWSM) 2015 [KKH⁺15]. The temporal network analysis in Section 2.4.4 including Figure 2.6 is by Jérôme Kunegis. Heinrich Hartmann contributed to the introduction of power indices in Section 2.5.1. Jérôme Kunegis, Sergej Sizov and Steffen Staab contributed to the text on the multi-Dirichlet geographical topic model in Chapter 3, which was published at the International Conference on Web Search and Data Mining (WSDM) 2014 [KKSS14].

The following is a list of the main contributions made in this thesis.

In Chapter 2:

- A novel generalisation of the Banzhaf and the Shapley power index is introduced in Sec. 2.5, which allows the inclusion of system-context, i.e. observed voting bias.
- The first evaluation of power indices using large datasets of observed voting behaviour in delegative democracies is presented in Section 2.5.6. Using observed power in real-world data, a probabilistic interpretation of power indices is employed for evaluation.

In Chapter 3:

- The novel concept of geographical networks for modelling complex spatial structures in probabilistic models is presented in Sec. 3.6.1. Using context networks, it is for the first time possible to efficiently model topics with complex distributions on the sphere (e.g. geographically distributed documents).

- Two models for context networks, one based on model selection and one based on a mixture of Dirichlet processes are presented in Sec. 3.6 and Sec. 3.7.4.
- A generalisation of the Dirichlet process for base measures consisting of mixtures of Dirichlet processes is introduced in Sec. 3.7.1 as the multi-Dirichlet process (MDP), which has been developed independently from [LF12].
- A generalised estimator for the concentration parameter of a symmetric Dirichlet distribution with changing dimensionality is derived in Sec. 3.7.6 to learn the hyper-parameter of mixing proportions in a MDP.
- A three-level hierarchical MDP is employed in Sec. 3.7.4 for modelling geographical network structures.
- An implementation of the Gibbs sampler is published as open source.

In Chapter 4:

- A generalisation of the hierarchical multi-Dirichlet process for multiple context variables is presented in Sec. 3.7.1.
- The HMDP topic model is the first topic model which is able to explicitly model multiple cyclic or spherical context variables.
- With the HMDP, it is possible to weight and to eventually select (i.e. remove) context variables during sampling.
- The HMDP topic model also is the first to allow for a direct interpretation of the parameters governing the influence of multiple non-trivial context variables.
- A practical collapsed stochastic variational Bayesian inference (PCSVB) scheme for the three-level HMDP is derived in Sec. 4.4.
- The implementation of PCSVB for the HMDP is published as open source.

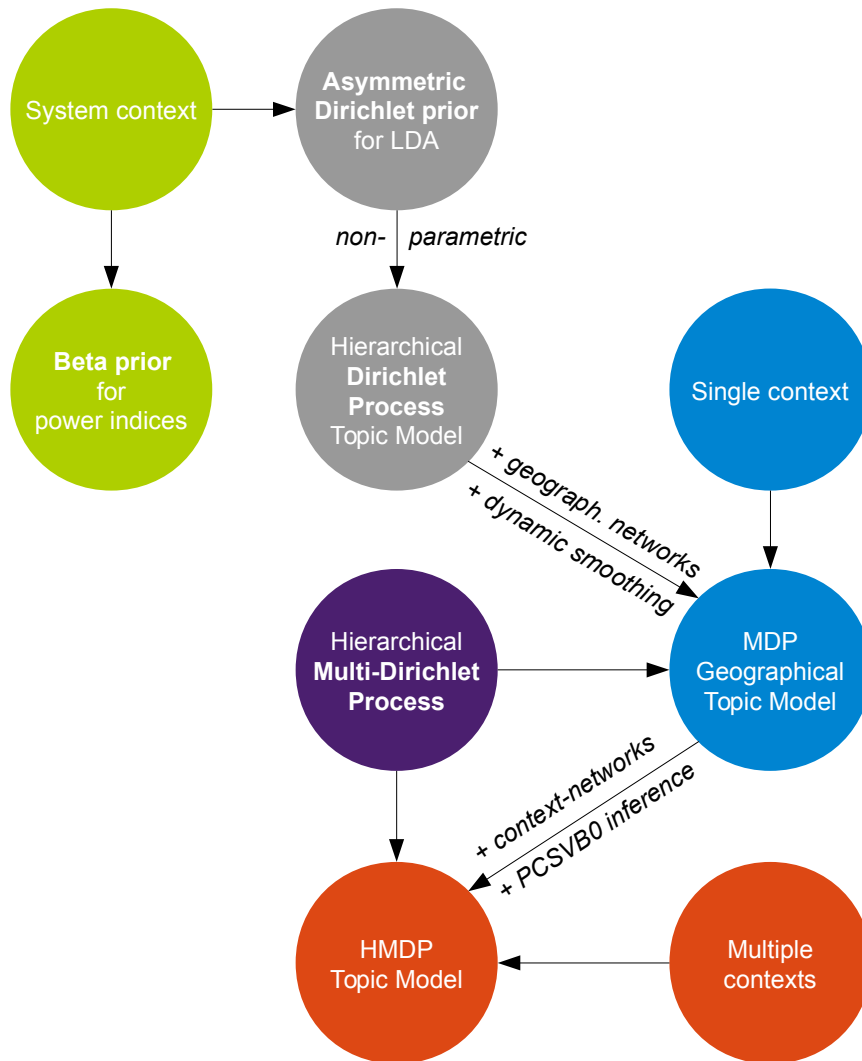


Figure 1: Overview of the structure of the thesis. In chapter 2, a system context model is presented which adds system-specific prior distribution to power indices. In topic modelling, the LDA model [BJ06a] can be extended for a system-specific prior distribution, which yields the asymmetric LDA model by Wallach [WMM09]. Replacing the asymmetric prior distribution of asymmetric LDA with a Dirichlet process yields a non-parametric topic model, the Hierarchical Dirichlet Process topic model (HDP). In chapter 3, a generalisation of the three-level HDP topic model for multinomial mixtures of Dirichlet processes is introduced, the Multi-Dirichlet Process (MDP). Using the MDP, a smoothing of topic distributions based on a geographical network of document clusters is introduced. This model is called the MDP Geographical Topic Model (MGTM). In chapter 4, the notion of geographical networks is generalised to multiple, arbitrary context networks and an efficient inference scheme based on Practical Collapsed Variational Bayesian inference with zeroth-order Taylor approximations (PCSVB0) is derived.

Chapter 1

Foundations and Related Work

In this chapter, a brief introduction to the foundations of probability theory is given. This includes the most important inference techniques for probabilistic models, namely *maximum likelihood estimation*, *maximum a posteriori estimation* and *Bayesian inference*. Then, basic probabilistic topic models are reviewed, including *probabilistic latent semantic indexing* (PLSA), *latent Dirichlet allocation* (LDA) and *hierarchical Dirichlet Process* topic models (HDP). For the latter models, two inference methods for complex probabilistic models are presented: *Gibbs sampling* and *variational inference*. Finally, methods for evaluating probabilistic models are discussed.

The example of a democratic vote. This chapter explains the fundamental concepts of probability with the help of the example of a democratic vote. The example will be used in preparation of the voting models presented in the next chapter.

In a democratic vote, a set of *voters* is voting on a given *proposal*. Every voter can vote with *yes* or *no*, and the sum of positive votes is used to calculate if a proposal is passed. If the share of positive votes exceeds a given *quorum* – e.g. 50% – the proposal will be successful, otherwise it will fail.

1.1 The Concept of Probability

The field of science can be divided into two parts:

- In *deductive* sciences, knowledge is created by logically deducing new facts from a set of known facts using a set of given operators. The derived facts can be proven to be true.

- In *inductive* sciences, hypotheses are created based on observations. Hypotheses are tested in order to increase their probability of being true or their *truthlikeness* as Popper would say [Pop79]. However, concepts such as truth or certain knowledge do not exist in inductive sciences, and all the stochastic models created to describe patterns in observations are potentially falsifiable [Pop79, Die95].

It therefore seems to be a natural choice to employ the concept of probability when developing models for explaining real-world observations. Probabilistic models are able to describe the uncertainty associated with rules derived from observations and provide the flexibility to react and adapt to new observations.

There are two competing notions of probability:

- In the *frequentist* interpretation, probability corresponds to the expected frequency of an event given a number of trials. Probabilities are inferred by dividing the number of times an event was observed by the total number of observations.
- From the *Bayesian* viewpoint, probabilities are beliefs about the likelihood of events. The belief does not have to relate to any observed events, and thus probabilities cannot be directly inferred. [KF09, p. 16]

One could want to describe the probability p_{im} with which a voter i will vote *yes* on proposal m . It was never observed how voter i voted for this very proposal, so a *frequentist* would not be able to make any statement about the probability p_{im} as this particular event was never observed before. Additionally, it would be absurd to conduct an experiment where voter i votes n times on the very same proposal m to learn about the probability with which the voter will agree with proposal m .

The frequentist interpretation of probabilities – while being useful for a multitude of applications (e.g. in descriptive statistics) – plays only a minor role in this thesis. In the models developed later in this thesis, non-observed or hidden variables are employed which can only be interpreted from a Bayesian viewpoint.

1.2 Basic Concepts and Distributions

Probability distributions are defined by probability functions which assign a positive value to measurable events $x \in \Sigma$ where Σ is the set of *measurable events*. Σ is a subset of the *sample space* Ω , the set of possible outcomes. The subset Σ is required to contain the empty set as well as the whole sample space Ω , and it must be closed under the complement and union of its contained sets [KF09]. Such subsets are called a *σ -algebra*. $P(\Omega)$,

the probability assigned to the event of sampling any element of the sample space, is 1.

There are two different types of probability distributions:

- Discrete probability distributions (also referred to as *probability mass functions*) which are defined on a discrete set of measurable events.
- Continuous probability distributions are defined by *probability density functions* (PDF) $f(x) \in \mathbb{R}^+$ with $x \in \Sigma$, where Σ is a continuous space of measurable events. The PDF value itself has no direct semantic interpretation other than as a relative likelihood, which can be calculated for measurable events consisting of a single value from the sample space.

In the setting of a single vote, the sample space is $\Omega = \{0, 1\}$, where 0 indicates a negative and 1 a positive vote. The space of measurable events is $\Sigma = \{\{\}, \{0\}, \{1\}, \{0, 1\}\}$ and $P(\Omega) = p(0) + p(1) = 1$.

1.2.1 Bernoulli distribution

The most simple non-trivial probability distribution is the Bernoulli distribution. It is defined as:

$$\begin{aligned} p(v | p) &= \begin{cases} p & \text{if } v = 1 \\ 1 - p & \text{if } v = 0 \end{cases} \\ &= p^v \cdot (1 - p)^{1-v} \quad \text{with } v \in \{0, 1\}. \end{aligned} \quad (1.1)$$

In the example of a vote, the Bernoulli distribution could define a belief in a voter to vote yes (coded as $v = 1$), which she then does with probability p , according to this very simple model. In the following, this probability p will be called the *approval rate* of the voter.

1.2.2 Likelihood

The *likelihood* of a probabilistic model is given by the probability of the outcomes under the model parameters.

If the example of a vote is extended to describe the voting behaviour of a voter who participates in a set of votes $m \in \{1, 2, \dots, M\}$ (with a constant approval rate of p), the outcomes are the voting decisions $(v_1, \dots, v_M) \in \{0, 1\}^n$ and the parameter is the approval rate p . The resulting likelihood is a product of Bernoulli distributions:

$$\mathcal{L}(p | \mathbf{v}) = p(\mathbf{v} | p) = \prod_{m=1}^M p^{v_m} \cdot (1 - p)^{1-v_m} = p^{n_1} \cdot (1 - p)^{n_0} \quad (1.2)$$

where n_1 is the number of times voter i voted *yes*, and n_0 the number of times she voted *no*.

1.2.3 Binomial distribution

If one does not know the individual votes of the voters but only the number of positive and negative votes, then a binomial distribution describes the probabilities of observed vote counts:

$$p(n_0, n_1 | p) = \frac{(n_0 + n_1)!}{n_0! \cdot n_1!} \cdot p^{n_1} \cdot (1 - p)^{n_0} \quad (1.3)$$

There are $(n_0 + n_1)!$ ways of ordering the votes to obtain the observed counts, but one cannot distinguish the positive votes (which could be ordered $n_1!$ different ways) and the negative votes (which could be ordered $n_0!$ different ways).

1.2.4 Marginal probability

In a probabilistic model with several variables, it sometimes is desirable to focus on some specific variables by factoring out other variables. One can do so by calculating the expected value of the variable of interest given the factored out variable. This process is called *marginalising out* the variable.

For instance, in the voting example one could be interested in the probability of $\mathbf{v} = (v_1, \dots, v_M) \in \{0, 1\}^M$, the binary vector of voting decisions, but not in the agreement rate p . Given that all values for p are equally likely (i.e. p takes on values in $[0, 1]$ with equal probability), the *marginal probability* of \mathbf{v} is:

$$p(\mathbf{v}) = \int_0^1 p^{n_1} \cdot (1 - p)^{n_0} dp = \frac{n_0! \cdot n_1!}{(n_0 + n_1 + 1)!} \quad (1.4)$$

where the parameter p is integrated out and the following equation can be employed:

$$\int_0^1 p(n_0, n_1 | p) dp = \frac{(n_0 + n_1)!}{n_0! \cdot n_1!} \int_0^1 p^{n_1} \cdot (1 - p)^{n_0} dp = \frac{1}{n_0 + n_1 + 1} \quad (1.5)$$

because under the assumption of uniformity every number of positive votes is equally likely and n_1 can take on $n_0 + n_1 + 1 = M + 1$ different values.

1.2.5 Maximum likelihood estimation

The parameters of such a simple probabilistic model can be estimated by maximising the likelihood of the given observations under the parameters, in this case a single parameter p . The maximum likelihood estimation is typically performed on the logarithm of the likelihood to ease the calculation of the derivative. For the example of the voter, the parameter p can be estimated dividing the number of positive votes by the total number of votes:

$$p = \frac{n_1}{n_0 + n_1}. \quad (1.6)$$

The complete derivation is given in Appendix A.1.

1.3 Bayesian Networks

One of the key concepts in probabilistic modelling are probabilistic graphical models (PGM) [KF09]. In this work, *Bayesian networks* are used, which are a subclass of PGMs. Bayesian networks make use of *conditional independence* relations between variables in graphical model to create visual representations of probabilistic models and to derive inference schemes.

Recall that two random variables A and B are *independent* if $P(A \cap B) = P(A) \cdot P(B)$. Two variables A and B are called *conditionally independent given C* if:

$$P(A \cap B \mid C) = P(A \mid C) \cdot P(B \mid C). \quad (1.7)$$

In the example of the voter who draws votes $\mathbf{v} = (v_1, \dots, v_M)$ which are positive with probability p , a conditional independence was observed and exploited: If the probability p is unknown and all votes are observed except for the first vote, the belief in the probability of v_1 depends on all the other observed votes. For instance, if it was observed that the voter voted *yes* in all the votes except for the first vote, where the voting decision is unknown, one would yield a high estimate for the approval rate p and therefore assign a high probability for v_1 being a positive vote as well.

The situation is different if the approval rate p is known. Now the belief about the probability of v_1 to be a positive vote only depends on the approval rate, and it is conditionally independent of all the other votes given the approval rate p . This intuition was used when defining the likelihood of the votes in Equation 1.2:

$$p(v_1, \dots, v_M \mid p) = p(v_1 \mid p) \cdot p(v_2 \mid p) \cdots p(v_M \mid p)$$

where it was implicitly assumed that the voting decisions of the voter are independent given the approval rate (i.e. the voting decision in the first vote does not have an impact on the second voting decision etc.).

This assumption can be expressed with a Bayesian network which codes independence relations between probabilistic variables. In a Bayesian network representation, variables are denoted as nodes and directed edges can be interpreted as causal relations [Mur01, KF09]. The absence of edges in a Bayesian network indicates conditional independence relations between variables.

The Bayesian network for the model of the voting example is shown in Fig. 1.1(a). In the model, votes v_{i1}, \dots, v_{iM} of voter i are caused by the agreement rate p and v_{im} denotes the m th vote. Given the agreement rate p , all votes are conditionally independent.

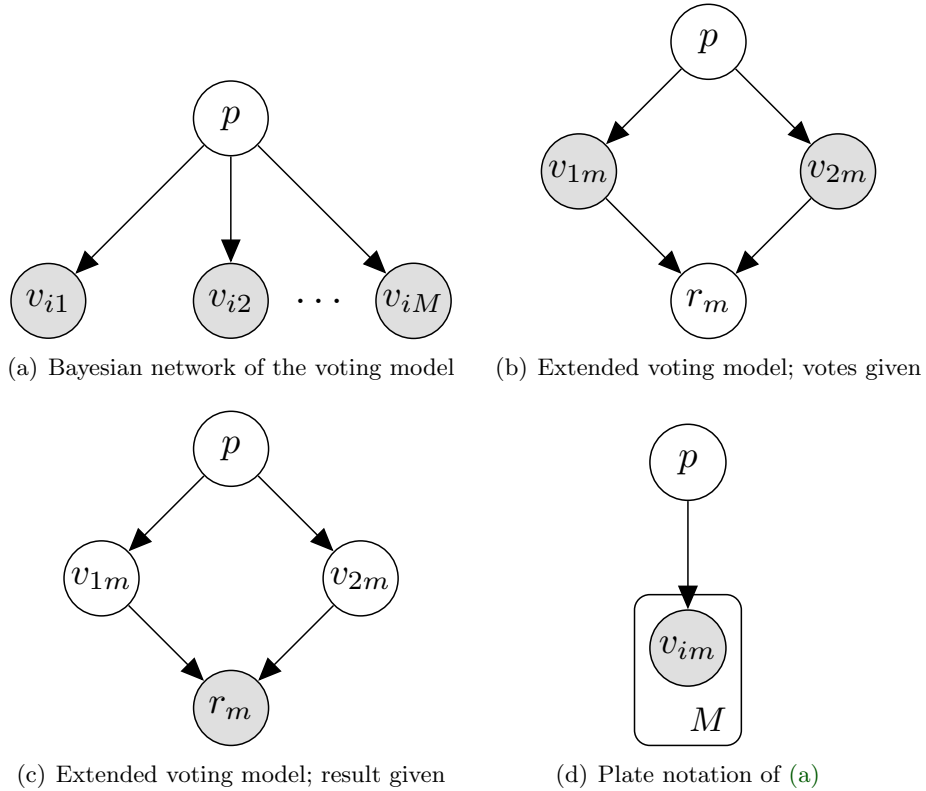


Figure 1.1: (a) Bayesian network representation of M observed variables v_1, \dots, v_M sampled from a Bernoulli distribution with parameter p . Observed variables are highlighted with a grey background. The observations are independent and identically distributed (i.i.d.). (b) Bayesian network for a vote with two voters which both share the same approval rate p and cause the voting result r with their votes, which are observed. (c) Bayesian network for a vote with two voters which both share the same approval rate p and cause the observed voting result r with their votes, which are unobserved. (d) Plate notation of the Bayesian network for the voting model from (a). The plate notation helps to simplify the notation of complex Bayesian networks. Repeating subgraphs are surrounded by plates, a displayed variable denotes the number of repetitions (in this case M).

1.3.1 Dependencies in Bayesian networks

A Bayesian network codes conditional (in)dependence relations based on the notion of *trails* and *d-separation*. *Trails* are undirected, loop-free paths between two variables. A *trail* between two variables u and v is *d-separated* if at least one of the following conditions holds [KF09]:

1. The trail contains a sequence $u \dots \rightarrow x \leftarrow \dots v$ and x is given. E.g. a given parent node of two variables renders them independent in the absence of another trail.
2. The trail contains the sequence $u \dots \rightarrow x \rightarrow \dots v$ or the sequence $u \dots \leftarrow x \leftarrow \dots v$ and x is given. E.g. a variable is independent of descendants of child nodes if these child nodes are given and there is no other trail between the variable and the descendants.
3. The trail contains a sequence $u \dots \leftarrow x \rightarrow \dots v$ and neither x nor any descendant of x is given. E.g. two variables are dependent of the co-parents of a child node if the child node or one of its descendants is given.

Two variables in a Bayesian network are independent if all *trails* between them are *d-separated*.

Rule 1 was already implicitly demonstrated in Fig. 1.1(a): The belief in the first voting decision v_{i1} of voter i becomes independent of all other voting decisions if the parent node p , the agreement rate, is given.

For demonstrating Rule 2, the example is extended. Imagine a setting where there are two voters who together vote on a proposal m . The voting decisions are v_{1m} and v_{2m} . Now the belief about the outcome of the vote can be added to the model, expressed as $r_m \in \{0, 1\}$, indicating the voting result (i.e. if the proposal is approved or not). A proposal needs 50% of the votes to pass the vote. Clearly, this variable is caused by voting decisions v_{1m} and v_{2m} . The corresponding Bayesian network is depicted in Figure 1.1(b). Now, if there is no information given about the variables, the Bayesian network indicates that the voting decisions are caused by the approval rate, and the voting result then is caused by the voting decisions. Thus the approval rate depends on the result (i.e. if the result is positive, then we know that the approval rate is greater than zero). However, given that the voting decisions are known, the approval rate becomes independent of the result, and it is possible to estimate the approval rate by only looking at the voting decisions.

Rule 3 can be explained with the same setting. The Bayesian network is given in Figure 1.1(c). If one does not know the voting result r_m , variables v_{1m} and v_{2m} are independent. However, if it would be known that the vote was passed (i.e. $r_m = 1$), the variables would become dependent. There are three out of four scenarios which result in a passed vote: either the first voter votes yes, or the second – or both. If it now would be known that

the vote was passed, and it would be known that the second voter voted *no*, it would be evident that the first voter voted *yes* with a probability of 100%. Both variables became dependent. In contrast, if the voting result is unknown, the knowledge about the voting decision of the second voter does not affect the belief in the vote of the first voter.

1.3.2 Plate notation

Looking at the Bayesian network of the voting example in Figure 1.1(a), one can see that not all child nodes of the approval rate p are depicted in the figure. In general, for complex graphical models, it is undesirable to draw the complete network of nodes, potentially consisting of thousands of nodes. Instead, repetitions in the network can be utilised to simplify the representation using the so-called *plate notation*. The plate notation of the Bayesian network for the voting model is depicted in Figure 1.1(d).

In plate notation, repeating subgraphs are surrounded by a plate which is labelled with the number of repetitions. Indices are used to indicate the repeating nodes in the subgraph. In the voting example, voting decisions v_{im} are surrounded by a plate with M repetitions (the number of votes) and the index m is used to distinguish between single voting decisions.

1.4 Priors and Bayesian Inference

Parameter estimation for the models presented so far was limited to maximum-likelihood estimation. For instance, in Equation 1.6 the approval rate p in a vote was estimated by dividing the number of positive votes by the total number of votes. For cases where a higher number of observations is available, this method is plausible. However, for a small number of observations, maximum likelihood estimation can yield bad parameter estimates. Consider a scenario where the voting behaviour of a voter is known for only two votes (on two different proposals), where the voter voted *no* in both cases. Maximum likelihood estimation now would estimate that the approval rate should be zero. Obviously, this estimate is unintuitive – there are not enough observations available to rule out the possibility that the voter would never vote *yes*. From a science-theoretic point of view, it is in general not desired to allow probabilities to become zero, as it is impossible to rule out that a previously unobserved event will occur in the future [Pop79]. Additionally, probabilistic models should not assign zero probabilities to events because they limit the set of explainable events and potentially assign a likelihood of zero to new observations. This especially becomes a problem in cases where a probabilistic model is used as an input for machine learning algorithms or when the likelihood is employed to compare the performance of several probabilistic models, e.g. for model selection [KF09].

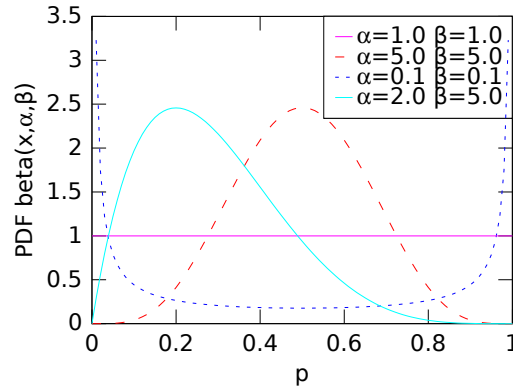


Figure 1.2: Probability density functions of the beta distribution at changing parameters. The expectation is $\alpha/(\alpha + \beta)$. Parameters smaller than 1 assign high densities to extreme values (i.e. 1 or 0) while values greater than 1 lead to densities centred around the expectation.

Prior probabilities help to overcome this problem. Before the approval rate p in the example is estimated, it is already known that p will not be taking on extreme values (i.e. not 0 or 1). It might be even known from existing theories that the approval rate should be e.g. close to 0.5. This information can be coded in a probability distribution over the approval rate.

1.4.1 Beta distribution

The typical prior for the parameter of a binomially distributed variable is the beta distribution, which takes two positive, real-valued parameters α and β :

$$\text{Beta}(x \mid \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha) \cdot \Gamma(\beta)} \cdot x^{(\alpha-1)} \cdot (1-x)^{(\beta-1)} \quad (1.8)$$

where the gamma function $\Gamma(x)$ is used, which is a generalisation of the factorial function for real numbers with

$$\Gamma(x) = (x-1)! \quad \forall x \in \mathbb{N}^+. \quad (1.9)$$

Note that the following equation holds:

$$\frac{\Gamma(x+1)}{\Gamma(x)} = x \quad (1.10)$$

which is frequently exploited in parameter inference for probabilistic models involving beta distributions and related probability density functions. Later in this thesis, in chapter 4, approximations to the gamma function will be used to speed up the parameter inference for complex probabilistic models.

The derivative of the logarithm of the gamma function is the digamma function:

$$\Psi(x) = \frac{d}{dx} \log(\Gamma(x)). \quad (1.11)$$

The digamma function plays a role in parameter inference, e.g. when the logarithm of the probability density function of the beta distribution is maximised.

The first part of the beta distribution is the inverse of the *beta function*, defined as

$$B(\alpha, \beta) = \frac{\Gamma(\alpha) \cdot \Gamma(\beta)}{\Gamma(\alpha + \beta)} \quad (1.12)$$

which ensures that the integral of $\text{Beta}(x | \alpha, \beta)$ over the sample space $[0, 1]$ equals 1:

$$B(\alpha, \beta) = \int_0^1 x^{(\alpha-1)} \cdot (1-x)^{(\beta-1)} dx \quad \forall \alpha, \beta \in \mathbb{R}^+. \quad (1.13)$$

The probability density of the beta distribution at different parameter settings is shown in Figure 1.2. The higher the parameters, the more focussed the distribution gets, indicating a strong prior belief. If the parameters are both set to $\alpha = \beta = 1$, the beta distribution becomes a uniform distribution and for parameters smaller than 1, the distribution becomes *sparse*, assigning high weights on values equal to 1 or 0. Beta priors with $\alpha = \beta$ are commonly referred to as *symmetric* priors.

1.4.2 Maximum a posteriori estimation

For the voting example, one now can multiply the likelihood with the prior probability of the parameter to get the so called *posterior probability*, the probability of the parameters given observations.

Given observations $\mathbf{n} = \{n_0, n_1\}$ and a beta prior with hyperparameters α, β on the parameter p of a binomial distribution, Bayes' theorem yields

$$\underbrace{p(p | \alpha, \beta)}_{\text{posterior}} = \frac{\overbrace{p(\mathbf{n} | p)}^{\text{likelihood}} \cdot \overbrace{p(p | \alpha, \beta)}^{\text{prior}}}{\underbrace{p(\mathbf{n})}_{\text{evidence}}} \quad (1.14)$$

where the *evidence* can be interpreted as a normalisation constant which makes sure that the posterior is a probability:

$$p(\mathbf{n}) = \int_A^B p(\mathbf{n} | p) \cdot p(p | \alpha, \beta) dp \quad (1.15)$$

and the domain of p is $[A, B]$. The evidence is independent of the parameter p and therefore it plays no role in parameter inference and is frequently omitted.

Omitting the evidence, the posterior probability of the vote model is proportional to

$$p(p \mid n_0, n_1, \alpha, \beta) \propto p(n_0, n_1 \mid p) \cdot p(p \mid \alpha, \beta) \propto p^{n_1 + \alpha - 1} (1 - p)^{n_0 + \beta - 1}.$$

which is a beta distribution with parameters $n_1 + \alpha$ and $n_0 + \beta$.

This beta distribution over the parameter p of the model can be maximised to obtain the parameter setting which maximises the likelihood *under the prior*.

The beta distribution is concave for $\alpha > 1$ and $\beta > 1$ (cf. Fig. 1.2) and for these cases, one obtains the parameter which maximises the a posteriori probability by setting the derivative (of the logarithm) to zero. This yields [Hei08]:

$$p = \frac{n_1 + \alpha - 1}{n_1 + n_0 + \alpha + \beta - 2}.$$

For a detailed derivation see Appendix A.2. Instead of only counting the positive votes and dividing by the total number of votes, the beta prior effectively adds so-called *pseudo counts* to the observations. The inference e.g. acts as if $(\alpha - 1)$ positive and $(\beta - 1)$ negative votes were already observed. An exception which is ignored in literature (e.g. [Hei08, AWST12, KF09, p. 754]) is that for sparse parameters $\alpha < 1$ or $\beta < 1$ one gets negative pseudo-counts and the MAP estimate from Eq. 1.4.2 returns inadequate results. The reason is that the posterior stops being concave for counts smaller than one. Therefore different estimates are required for this special case, which are given here:

$$\hat{p}_{MAP} = \begin{cases} \frac{n_1 + \alpha - 1}{n_1 + n_0 + \alpha + \beta - 2} & \text{if } n_1 + \alpha > 1 \wedge n_0 + \beta > 1 \\ 1 & \text{else if } n_1 + \alpha > n_0 + \beta \\ 0 & \text{else if } n_1 + \alpha < n_0 + \beta \\ x, x \sim \text{unif}(0, 1) & \text{else if } n_1 + \alpha = n_0 + \beta = 1 \\ 1 \text{ or } 0 & \text{else if } n_1 + \alpha = n_0 + \beta \end{cases} \quad (1.16)$$

i.e. the probability estimate is either $\hat{p}_{MAP} = 0$ or $\hat{p}_{MAP} = 1$ for counts smaller than one. In this case, the posterior probability is infinite for both $p = 1$ and $p = 0$, and the likelihood is growing faster to the direction of the larger counts. Therefore it is reasonable to set $\hat{p}_{MAP} = 1$ if $n_1 + \alpha$ is larger and $\hat{p}_{MAP} = 0$ if it is smaller than $n_0 + \beta$. Theoretically, there could be situations where the counts are smaller than one and equal, i.e. $n_1 + \alpha = n_0 + \beta$, and in this case the maximum is found by setting \hat{p}_{MAP} to 1 or 0. For cases where $n_1 + \alpha = n_0 + \beta = 1$, the posterior is the uniform distribution

and therefore all parameter settings maximise the posterior. In MAP, non-sparse beta priors smooth the estimate, while sparse beta priors produce more extreme estimates, compared to the maximum likelihood estimate for a multinomial without priors.

For the example of the vote with two observed negative votes, one could decide to set a prior which assigns high probability to approval rates around 50%, like a symmetric beta prior with parameters $\alpha = \beta = 5.0$. While maximum likelihood estimation yields an estimated approval rate of 0, the maximum a posteriori inference would estimate the approval rate as 0.4, by including the prior belief in the approval rate of the voter.

1.4.3 Bayesian inference

A different and more elegant way of using the prior probability over the parameters is Bayesian inference. Both maximum likelihood estimation and maximum a posteriori inference maximise parameters with regard to a continuous probability distribution over the parameter. All information about the shape of this probability distribution is lost, i.e. there is no effect of the existence of other parameter settings with high probability on the parameter estimate. And MAP might return zero probabilities which typically are undesired for probabilistic modelling.

Therefore, it can be advantageous to include the whole information of the distribution of the parameter into the parameter estimate. Bayesian inference estimates parameters as the *expectation of the posterior* over the parameters.

For the example of a binomial distribution with a beta-distributed prior, this yields [Hei08]:

$$\begin{aligned} E[p(p | n_0, n_1, \alpha, \beta)] &= \int_0^1 p \cdot p(p | n_0, n_1, \alpha, \beta) dp \\ &= \int_0^1 p \cdot \frac{p(n_0, n_1 | p) p(p | \alpha, \beta)}{\int_0^1 p(n_0, n_1 | p') p(p' | \alpha, \beta) dp'} dp \\ &= \int_0^1 \frac{p^{1+n_1+\alpha-1} \cdot (1-p)^{n_0+\beta-1}}{\int_0^1 p'^{n_1+\alpha-1} \cdot (1-p')^{n_0+\beta-1} dp'} dp. \end{aligned}$$

The binomial distribution can be moved to the exponent of the numerator and integrals replaced by beta functions using Equation 1.13:

$$= \frac{\int_0^1 p^{n_1+\alpha} \cdot (1-p)^{n_0+\beta-1} dp}{B(\alpha + n_1, \beta + n_0)} = \frac{B(\alpha + n_1 + 1, \beta + n_0)}{B(\alpha + n_1, \beta + n_0)}.$$

Plugging in equation 1.12 yields

$$= \frac{\Gamma(\alpha + n_1 + 1) \cdot \Gamma(\beta + n_0) \cdot \Gamma(\alpha + \beta + n_1 + n_0)}{\Gamma(\alpha + n_1) \cdot \Gamma(\beta + n_0) \cdot \Gamma(\alpha + \beta + n_1 + n_0 + 1)}$$

and using the properties of the gamma function by plugging in Equation 1.10 gives

$$E[p(p | n_0, n_1, \alpha, \beta)] = \frac{n_1 + \alpha}{n_1 + n_0 + \alpha + \beta}. \quad (1.17)$$

Obviously, Bayesian inference yields a different parameter estimate than maximum a posteriori inference (Eq. 1.4.2): The parameters α and β of the beta distribution directly correspond to pseudo counts. In the voting example, this means that the inference acts as if α positive and β negative votes were already observed. As the parameters of the beta distribution are greater than zero, the probability estimates are guaranteed to be non-zero.

For the example of the vote with two observed negative votes and a prior of $\alpha = \beta = 5.0$, maximum likelihood estimation estimated the approval rate as $p = 0$, maximum a posteriori inference as $p = 0.4$. Bayesian inference now would estimate the approval rate as $p = \frac{0+5}{0+2+5+5} = 0.41\bar{6}$. In practice, Bayesian inference often yields more appropriate parameter estimates than maxima-based estimates, as the whole information of the posterior is used instead of a single maximum.

1.4.4 Multinomial distribution

It is straightforward to extend the binomial distribution for multiple (i.e. more than two) categories. For instance, in the voting example one could add *abstention* to the set of possible voting decisions, so that for a single voting decision $v_{im} \in \{0, 1, 2\}$ where the numbers code *no*, *yes* and *abstention*, respectively.

The resulting probability distribution is the multinomial distribution: Given K categories with counts n_1, n_2, \dots, n_K for each category and probabilities p_1, p_2, \dots, p_K , the probability of the observed counts is (c.f. Eq. 1.3):

$$p(n_1, n_2, \dots, n_K | p_1, p_2, \dots, p_K) = \frac{(\sum_{i=1}^K n_i)!}{n_1! \dots n_K!} \cdot p_1^{n_1} \dots p_K^{n_K}. \quad (1.18)$$

The first part of the formula accounts for the possible combinations of events which produce the given numbers of observations in each category. As for the binomial distribution, this term is not relevant for maximum likelihood estimation of the parameters p_1, p_2, \dots, p_K (compare to Eq. 1.6). The right-hand side of the equation can be turned into a binomial distribution by merging categories. In the example of a vote, if one e.g. does not distinguish between negative votes and abstention, one obtains a two-category distribution again which is a binomial with two counts, n_{yes} and $n_{\{no, abstention\}}$. Analogous to Eq. 1.6 the maximum likelihood estimate of the probability of a single category in a multinomial then is given by

$$p_i = \frac{n_i}{\sum_{i=1}^K n_i} \quad \forall i \in \{1, \dots, K\}. \quad (1.19)$$

1.4.5 Dirichlet distribution

For maximum a posteriori estimation and Bayesian inference, the *Dirichlet distribution*, a multi-dimensional generalisation of the beta distribution, can be used as a prior for the parameters of the multinomial distribution. The Dirichlet distribution is defined as (c.f. Eq. 1.8):

$$p(\theta_1, \dots, \theta_K \mid \alpha_1, \dots, \alpha_K) = \frac{\prod_{k=1}^K \Gamma(\alpha_k)}{\Gamma\left(\sum_{k=1}^K \alpha_k\right)} \cdot \prod_{k=1}^K \theta_k^{\alpha_k-1} \quad (1.20)$$

with $\theta_k \in [0, 1]^K \quad \forall k \in \{1, \dots, K\}$ and $\sum_{k=1}^K \theta_k = 1$. The normalisation factor is based on a generalisation of the beta function (see Eq. 1.12):

$$\int \prod_{k=1}^K \theta_k^{\alpha_k-1} d\boldsymbol{\alpha} = \frac{\Gamma\left(\sum_{k=1}^K \alpha_k\right)}{\prod_{k=1}^K \Gamma(\alpha_k)} \quad (1.21)$$

The interpretation of the parameters of the Dirichlet distribution is the same as for the beta distribution. Parameter values smaller than one induce sparsity, while values larger than one lead to a smoothed distribution. If all parameters are set to the same value ($\alpha_i = \alpha_j \quad \forall k \in \{1, \dots, K\}$), the prior is symmetric and assigns the same prior probability to each θ_i . In this case, one might re-write the Dirichlet distribution as

$$p(\theta_1, \dots, \theta_K \mid \alpha) = \frac{\Gamma(\alpha)^K}{\Gamma(K \cdot \alpha)} \cdot \prod_{k=1}^K \theta_k^{\alpha-1}. \quad (1.22)$$

For the special case of $\alpha = 1$, the symmetric Dirichlet distribution equals a uniform distribution. The parameter α commonly is referred to as a *concentration parameter* [KF09] as large values lead to high probabilities around the expected value while small values assign high probability densities to areas distant to the expected value, where e.g. one θ_i virtually takes on all the probability mass while the other output parameters θ are close to zero.

For a better understanding of the Dirichlet distribution, Polya introduced an iterative scheme of drawing coloured balls from urns which yields a Dirichlet distribution in its infinite limit, the *Polya urn scheme*. Imagine that there are balls of K colours, one colour for every category of a given Dirichlet distribution. There further exists an urn which contains α_1 balls of colour 1, α_2 balls of colour 2 and so on. In the Polya urn scheme, a ball is repeatedly drawn from the urn, its colour is noted, the ball returned and an additional ball of the same colour added to the urn. After normalising the counts by the number of balls in the urn, one obtains a probability for each category to be drawn from the urn; and the infinite limit of this scheme will yield Dirichlet distributed probabilities.

1.4.6 Maximum a posteriori estimation and Bayesian inference for the multinomial distribution

Again, the analogy to the binomial distribution can be exploited to obtain parameter estimates for the multinomial distribution with a Dirichlet prior: One can merge all categories except one to turn the distribution into a binomial distribution with a beta prior, for which the parameter estimate is known. The posterior probability then is Dirichlet distributed with $p(p_1, \dots, p_K) = \text{Dir}(n_1 + \alpha_1, \dots, n_K + \alpha_K)$.

A maximum a posteriori estimate of the multinomial distribution with a Dirichlet-distributed prior is (analogous to Eq. 1.4.2) obtained as:

$$\hat{p}_i = \begin{cases} \frac{(n_i + \alpha_i - 1) \cdot [n_k + \alpha_k > 1]}{\sum_{k=1}^K (n_k + \alpha_k - 1) [n_k + \alpha_k > 1]} & \text{if } \exists n_j + \alpha_j \geq 1, j \in \{1, \dots, K\} \\ [n_i + \alpha_i > n_j + \alpha_j \forall j \neq i] & \text{else if } \text{Dir}(\mathbf{n} + \boldsymbol{\alpha}) \text{ is unimodal} \\ \text{randp}() & \text{otherwise} \end{cases} \quad (1.23)$$

where Iverson brackets are used: False statements within the square brackets return 0, true statements return 1, i.e. in the second case the probability of the class with the highest counts (including pseudo-counts) is set to 1. Theoretically, there could be situations where the posterior $\text{Dir}(\mathbf{n} + \boldsymbol{\alpha})$ is multi-modal, i.e. there exist several categories whose counts are maximal. In this case a maximum is found by randomly setting p_i to 1 for one of these categories. If all counts are equal to one, the posterior is uniform and a MAP estimate can be randomly drawn from this uniform distribution. This behaviour of the estimator is denoted by the $\text{randp}()$ function.

The Bayesian inference of parameter p_i of a multinomial distribution with a Dirichlet prior given observed counts n_1, n_2, \dots, n_K is (analogous to Eq. 1.17):

$$p_i = \frac{n_i + \alpha_i}{\sum_{k=1}^K n_k + \alpha_k}. \quad (1.24)$$

1.4.7 Fisher distribution and approximate inference

So far, this chapter presented three discrete probability distributions – the Bernoulli distribution and its generalisations, the binomial and the multinomial distribution. Additionally, two probability density functions, the beta and Dirichlet distribution, were introduced. All these distributions have a closed-form solution for estimating the parameter which maximises the likelihood or the posterior. However, there exist probability distributions where maximum likelihood estimation does not yield a closed-form solution.

As mentioned, the parameter α of the symmetric Dirichlet distribution is commonly referred to as a *concentration parameter*. Another probability

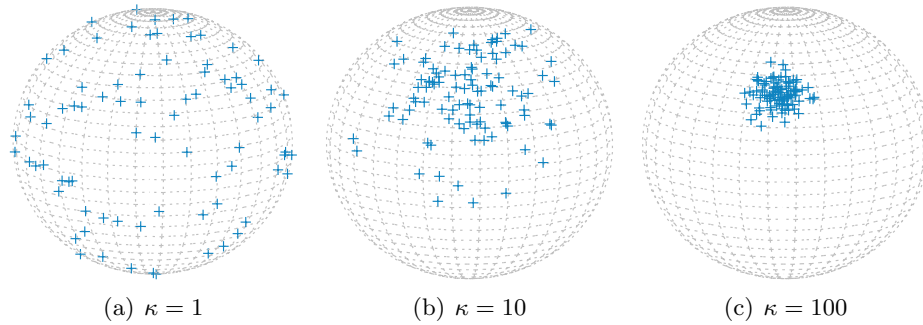


Figure 1.3: The Fisher distribution, a special case of the Mises–Fisher distribution for three dimensions. Each of the figures shows 100 random Fisher-distributed samples on the three-dimensional unit sphere with concentration parameters $\kappa = 1$ (a), $\kappa = 10$ (b) and $\kappa = 100$ (c). Similar as for the symmetric Dirichlet (and Beta) distribution, a higher concentration parameter means a higher probability density around the expected value.

density function with a concentration parameter is the Mises–Fisher distribution.

The Mises–Fisher distribution is a probability density function on an n -sphere, the generalisation of the sphere for arbitrary dimensions. An important special case of the Mises–Fisher distribution is the Fisher distribution on the three-dimensional sphere. In short, the Fisher distribution is comparable to an isotropic Gaussian distribution on the plane (i.e. a “non-directed” Gaussian with equal variance and no correlation between both dimensions) [Fis53, Wat82]. The Fisher distribution is defined as

$$f(x \mid \kappa, \mu) = \frac{\kappa}{4\pi \cdot \sinh(\kappa)} \cdot e^{\kappa \cdot \mu^T x}$$

where μ is the mean, which is located on the unit-sphere, and $\kappa \in \mathbb{R}^+$ is the concentration parameter. As for the Dirichlet distribution, a high concentration parameter implies a high probability mass centred around the expected value. For a low concentration parameter, the probability density function becomes more uniform and the uniform distribution is reached at $\kappa = 0$.

Maximum likelihood estimation of Fisher distributions

Given a set of N observed samples $\{x_1, \dots, x_N\}$ from this distribution, one can apply maximum likelihood estimation for obtaining parameter estimates.

Each observation x_i is a point on the sphere and the estimate for the mean direction is [MJ09, p. 198]:

$$\mu_x = \frac{\sum_i x_i}{\|\sum_i x_i\|_2} \quad (1.25)$$

The estimate of the concentration parameter κ is more complex, because the normalisation factor has to be taken into account. Maximising the likelihood with respect to κ yields an optimum at [MJ09, Sra12]:

$$A_p(\kappa) = \bar{R} \Rightarrow \kappa = A_p^{-1}(\bar{R}) \quad (1.26)$$

with

$$\bar{R} = \frac{\|\sum_i x_i\|_2}{N} \quad (1.27)$$

and

$$A(\kappa) = \frac{I_{\frac{3}{2}}(\kappa)}{I_{\frac{1}{2}}(\kappa)} \quad (1.28)$$

with $I_s(x)$ denoting the Bessel function of the first kind defined as

$$I_s(x) = \sum_{k=0}^{\infty} \frac{(-1)^k}{\Gamma(k+s+1) \cdot k!} \left(\frac{x}{2}\right)^{2k+s}. \quad (1.29)$$

Because of the complex nature of the Bessel functions (which include a summation over fractions involving gamma functions), it is not possible to obtain a closed-form solution for equation 1.26. There exist several alternative approximations for the concentration parameter of the Mises–Fisher distribution [BDGS05, MJ09, TFO⁺07]. A popular approximation is given by Banerjee et al. [BDGS05]:

$$\kappa \approx \frac{3 \cdot \bar{R} - \bar{R}^2}{1 - \bar{R}^2} \quad (1.30)$$

However, more precise approximations exist which estimate κ in an iterative process [TFO⁺07].

It is evident that for some probability distributions, parameter estimates can be significantly harder to compute than for simple distributions such as the multinomial with a Dirichlet prior. Large graphical models involving complex probability distributions amplify the high computational costs. For large datasets, inference would be slowed down significantly, which renders the application of these models practically impossible. On the other hand, one often wants to use the special properties of complex distributions, e.g. the periodicity of the Fisher distribution. These issues are addressed later in the thesis.

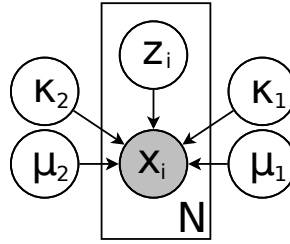


Figure 1.4: Bayesian network representation for a mixture of two Fisher distributions. The observed data $\{x_1, \dots, x_N\}$ (points on the sphere) are drawn from two Fisher distributions with parameters μ_1, κ_1 and μ_2, κ_2 , respectively. Membership variable $z_i \in \{1, 2\}$ indicates from which of the two distributions observation x_i was sampled.

1.4.8 Latent variables and expectation-maximisation

In Section 1.3 Bayesian networks were introduced which included given observations drawn from probability distributions with unknown parameters. In Fig. 1.1(c) a graphical model was shown which includes latent (i.e. unobserved) variables representing the actual votes of two voters, which can be estimated given the known voting result.

The most popular application of latent variables, however, is the inference on generating probability distributions: Given multiple probability distributions and N observed samples \mathbf{x} , find out which observations were sampled from which probability distribution. For the case of two Fisher distributions, the problem can be formalised as follows: Given two Fisher distributions with parameters μ_1, κ_1 and μ_2, κ_2 , which e.g. could correspond to the center and the spread of two neighbouring cities. Additionally, there are latent membership variables \mathbf{z} with $z_i \in \{1, 2\} \forall i \in \{0, \dots, N\}$ which – for every data point x_i – indicate if the data point is sampled from the first or the second Fisher distribution, i.e. an observed photo with a GPS tag is associated with the area of geographic reach of the first or the second city.

Every data point x_i then is sampled from its generating Fisher distribution, known from z_i :

$$x_i \mid z_i \sim f(\kappa_{z_i}, \mu_{z_i}) \quad (1.31)$$

The corresponding graphical model is shown in Fig. 1.4. Models which explain observations with multiple alternative probability distributions, and which learn latent variables assigning each observation to one single probability distribution, are called *single-membership* models.

Parameter inference in the setting of a mixture of two Fisher distributions involves the estimation of two types of variables: the parameters μ and κ of the Fisher distributions and the latent membership variables \mathbf{z} .

This general scheme of modelling data has a broad range of applications: One can choose the number of groups and plug in every desired probability

distribution in the model to explain data by a set of probability distributions with different parameters. Each datapoint is explained by being generated by a single group and thus such models are called *single membership models*. A broad range of data mining methods for detecting a-priori unknown groups of datapoints is based on similarly structured probabilistic models.

Expectation-maximisation

The probability distribution over the latent variables in a model with two Fisher distributions is already quite complex because of the dependencies between the latent variables through the parameters of the Fisher distributions. From Fig. 1.4 one can see that all membership variables \mathbf{z} and parameters $\mu_1, \kappa_1, \mu_2, \kappa_2$ depend on each other through their children, the observations \mathbf{x} , which are given.

If the membership-variables \mathbf{z} would be known, the parameters of each Fisher distribution would only depend on their observations \mathbf{x} and it would be possible to use maximum likelihood estimation to infer the latent parameters.

Additionally, knowing the parameters of the probability distributions, the latent variables \mathbf{z} would become independent of each other and one could use the fact that probability densities are relative likelihoods to infer the membership probabilities for each data point (i.e. it is possible to calculate the probability of an observation being drawn from the first or the second Fisher distribution).

Both facts can be used to learn about the unknown variables in an iterative sampling scheme in which the parameters of the probability distributions are randomly initialised and then the algorithm iteratively *(i) learns about the latent variables* and then *(ii) re-estimates the parameters of each probability distribution*, given the membership of each observed datapoint. The latent variables are estimated as membership-probabilities. For parameter estimation, observations are weighted by the probability of being drawn from the respective probability distribution. This sampling scheme is called *expectation-maximisation* (EM).

For a mixture of Fisher distributions, one can plug in the weighted observations in the formula for maximum likelihood estimation given in Equations 1.25 and 1.30. The expectation-maximisation algorithm then is given by: [BDGS05, GY14]

1. **Initialisation:** Set $\kappa_1 = \kappa_2 = 0$ and set μ_1 and μ_2 to the position of two randomly chosen, distinct observations.
2. **Expectation step:** For every observation x_i , calculate the member-

ship probabilities as:

$$p(z_i = 1) = \frac{f(x_i | \mu_1, \kappa_1)}{f(x_i | \mu_1, \kappa_1) + f(x_i | \mu_2, \kappa_2)}$$

$$p(z_i = 2) = \frac{f(x_i | \mu_2, \kappa_2)}{f(x_i | \mu_1, \kappa_1) + f(x_i | \mu_2, \kappa_2)}$$

3. **Maximisation step:** Re-estimate the parameters of the two Fisher distribution indexed by $t \in \{1, 2\}$ with weighted observations:

$$\mu_t = \frac{\sum_i p(z_i = t) \cdot x_i}{\|(\sum_i p(z_i = t) \cdot x_i)\|_2}$$

$$\bar{R}_t = \frac{\|(\sum_i p(z_i = t) \cdot p(z_i = t) \cdot x_i)\|_2}{N}$$

$$\kappa_t \approx \frac{3 \cdot \bar{R}_t - \bar{R}_t^2}{1 - \bar{R}_t^2}$$

Step 2 and 3 are repeated until the average change in $p(\mathbf{z})$ is smaller than a pre-defined lower limit ϵ .

1.5 Topic Models

The core of this thesis is centred around a special class of probabilistic models which is more complex and more powerful than the models introduced so far. These probabilistic models are called *topic models*, because their most prominent use case is the automatic extraction of *topics* from a corpus of text documents. The motivation behind topic modelling is the processing of large corpora of text documents. For example, in data mining, topic models can be applied to analyse the content of a large corpus by learning which topics are covered and how many documents are available for each topic. In machine learning, processing words of documents is often inefficient. Instead, documents can be *represented* by the topics they cover to reduce the complexity of the input for a prediction algorithm. The historically most important motivation for topic models is information retrieval. Given a search query consisting of several words, it can be computationally expensive to compare the words of the query with the words in each text document of a corpus. Instead, the topics behind the query words can be detected and the search can be limited to documents from the topic of the query. For instance, it could be advantageous to limit the search to text documents from the topic *sports* given the query “*best football players*”.

Generally speaking, topics in topic modelling correspond to sets of frequently co-occurring words in a given corpus of text documents. In the probabilistic setting, topics are multinomial distributions over the known

vocabulary, e.g. all terms appearing in the corpus. This is best explained with an example. Imagine one would do a simple quantitative analysis of a collection of newspaper articles based on word frequencies. For each article, the section is known which can be interpreted as the (very general) topic of the contained news and there are three sections, *politics*, *economy* and *sports*. Looking at the news articles, one could observe that there are typical words which occur frequently in each section. For instance, terms like *financial*, *economy*, *bank* and *crisis* frequently occur in the economy section. For the politics section, one could observe that *election*, *politician* and *vote* occur often and the words *soccer*, *match* and *goal* frequently appear in the sports section. The word frequencies can then be normalised by the total number of words in each newspaper section to obtain a frequentist estimate of the probability of a term in a given section of the newspaper. These multinomial distributions over the vocabulary are simple descriptions of the topic of each news section and *assign high probabilities to semantically related terms*. One could employ such multinomial topic-word distributions to guess the topic of a given text document e.g. by selecting the most probable topic given that the words in the document are independent draws from a topic multinomial.

The target of topic models is to automatically extract topics from a set of documents without requiring any background knowledge about sections or other categorisations of documents, and where documents potentially cover multiple topics. There exist both non-probabilistic and probabilistic approaches for topic modelling.

1.5.1 Latent semantic analysis

The classical method for topic extraction is called *Latent Semantic Analysis* (LSA). For every corpus of text documents there exists a matrix representation D , where rows correspond to documents, columns correspond to the terms in the vocabulary, and entry d_{it} stores the frequency of term t in document i . Note that the ordering of words is lost by this matrix representation of documents. This is the so-called *bag-of-words assumption*. Because of its mathematical implications this assumption is behind all the models presented in this thesis.

LSA is based on a Singular Value Decomposition (SVD) of the document-term matrix D . D can be decomposed as follows:

$$D = U\Sigma V^T \tag{1.32}$$

with Σ being a diagonal matrix containing the singular values of the matrix, and U and V^T being orthogonal matrices, i.e. $UU^T = I$ and $V^TV = I$. V^T can be interpreted as a topic-word matrix, where every row corresponds to a topic vector and the entries give the weight of a term t under topic

k with $v_{kt} \in [-1, 1]$. Terms which are important for a given topic (i.e. co-occur frequently) have a weight close to 1, while terms which co-occur less often than expected with a given topic have negative weights. U can be interpreted as a document-topic matrix, storing the weight of the topics for the given document.

By reordering the matrices Σ , U and V by the size of the singular values and reducing the matrix to the first K singular values, one obtains matrix Σ' . Then U and V are reduced to the first K topics, yielding U' and V' . An approximation to the document-term matrix D then is given by $D' = U'\Sigma'V'^T$. U' has $M \times K$ columns, where M is the number of documents. Given that the vocabulary size V is much larger than K , U' is an effective dimensionality reduction of the document-term matrix D and has especially proven useful for information retrieval tasks.

Because topic vectors may contain negative values and because of undesired behaviour such as the inability to detect polysemy, LSA makes a human interpretation of the detected topics difficult. Additionally, it is not clear how to extend or modify LSA e.g. for documents with metadata.

1.5.2 Probabilistic latent semantic analysis

The most basic non-trivial probabilistic topic model is Probabilistic Latent Semantic Analysis (PLSA) [Hof99]. PLSA models a corpus of M documents $D = \{d_1, \dots, d_M\}$ where documents are sets of words (i.e. the bag-of-words assumption again is employed): $d_i = \{w_{i1}, w_{i2}, \dots, w_{iN_i}\}$ and where N_i is the length of document i . Because the ordering of words is ignored, the probability of observations would be identical for all orderings of observed words in the documents, making the observations exchangeable. De Finetti's theorem tells that any set of exchangeable observations can be modelled by a mixture of probability distributions [BNJ03, Kin78]. In PLSA, the number of topics K is chosen a priori and each word is modelled as the result of the following process:

1. For every of the K topics, draw a topic-word distribution ϕ_k from a uniform distribution.
2. For every document d_i , draw a document-topic distribution θ_i from a uniform distribution.
3. Select a document i from a uniform probability distribution over all documents.

For each word w_{ij} in document d_i , draw a topic z_{ij} from θ_i :

$$z_{ij} \sim \text{Mult}(\theta_i)$$

and then draw a word w_{ij} from $\phi_{z_{ij}}$:

$$w_{ij} \sim \text{Mult}(\phi_{z_{ij}})$$

It is assumed that the number of topics K is known. The resulting graphical model is depicted in Figure 1.5(a).

Topics found by PLSA are multinomial distribution over the set of words and therefore can be interpreted. As for every probabilistic model, extensions to PLSA are straightforward, and numerous variants of PLSA exist in literature. Some of the extensions will be reviewed in a later section.

Inference for PLSA is based on the expectation-maximisation algorithm described in the original paper by Hofmann [Hof99]. First, all probability distributions are randomly initialised. Then, in an iterative scheme until convergence, topic-word assignments z_{in} are calculated and the assignments then are used to re-estimate the topic-word distributions (topics) and document-topic distributions via maximum likelihood estimation.

PLSA – though being an easy-to-implement and popular method for topic extraction – comes with restrictions. First of all, there is no smoothing, meaning that unseen words will lead to zero probabilities in the topic-word multinomials. Second, the PLSA model from the original paper only allows inference for known documents which are selected with equal probability. To infer the topic distribution for new documents within the PLSA framework, the topic-word distributions of existing documents have to be averaged to obtain an empirical topic-word distribution.

1.5.3 Latent Dirichlet allocation

To overcome some of the problems with PLSA, Blei et al. [BNJ03] proposed Latent Dirichlet Allocation (LDA), a probabilistic topic model with a different generative process for documents. As in PLSA, the number of topics K is chosen a priori. The process is as follows:

1. For each document, draw the document length N_i from a Poisson distribution:

$$N_i \sim \text{Poisson}(\lambda)$$

2. For every of the K topics, draw a topic-word distribution ϕ_k from a symmetric Dirichlet distribution with parameter β :

$$\phi_k \sim \text{Dir}(\beta)$$

3. For every document d_i , draw a document-topic distribution θ_i from a symmetric Dirichlet distribution with parameter α :

$$\theta_i \sim \text{Dir}(\alpha)$$

4. Finally, for each word w_{ij} in document d_i , first draw a topic z_{ij} from θ_i and then draw a word w_{ij} from $\phi_{z_{ij}}$:

$$z_{ij} \sim \text{Mult}(\theta_i) \quad w_{ij} \sim \text{Mult}(\phi_{z_{ij}}) \quad (1.33)$$

According to Blei [BNJ03], one key advantage of LDA over PLSA is that LDA is a well-defined model not only for the documents of the training corpus D , but also for unseen documents [GK03]. However, in practice, PLSA can be slightly reformulated to yield a model which permits inference for new documents. It can be easily shown that LDA with uniform priors (i.e. $\alpha = 1$ and $\beta = 1$) is identical to such a corrected PLSA [GK03, AWST12]. The strength of LDA thus lies in the (non-uniform) Dirichlet priors which can be shown to improve the model quality significantly in practice.

Improvements in model quality through Dirichlet priors are due to multiple causes. In Section 1.4.1 and 1.4.4 it was demonstrated that the parameters of a Dirichlet prior serve as pseudo-counts during inference and thus effectively prevent zero probabilities and overfitting. As visualised in Fig. 1.2, symmetric Dirichlet priors (and the beta prior, the two-dimensional case of a Dirichlet) induce sparsity for concentration parameters smaller than 1, i.e. it is very likely to see a small set of categories (or only one) with a high probability while all other categories have probabilities close to zero. Both for the topic-word distributions ϕ and the document-topic distributions θ it can be shown that sparse Dirichlet priors help the topic detection process: First, for the vast majority of datasets, topics only place weight on a small set of terms from the vocabulary [BNJ03]. Second, a single document typically covers only a limited set of topics and therefore a sparse prior for the document-topic distribution is beneficial for many datasets.

However, there are exceptions where documents have a non-sparse distribution over the set of topics, e.g. if the number of topics is small. Additionally, in many cases an *asymmetric* sparse Dirichlet prior over the topic-word distribution can be beneficial for topic quality and perform better than a symmetric prior [WMM09].

1.6 Inference for Complex Graphical Models

Latent Dirichlet Allocation samples words from a mixture of multinomial distributions with Dirichlet priors. Parameter inference for LDA can become more complex and there exists a multitude of different approaches which, however, all turn out to result in similar inference equations [AWST12].

1.6.1 Expectation-maximisation and maximum a posteriori inference

As described in Section 1.5.2, inference for PLSA is based on an expectation-maximisation algorithm where parameters are estimated with maximum likelihood estimation. To account for the Dirichlet priors in LDA, the expectation-maximisation scheme can be modified – instead of maximum likelihood estimation, maximum a posteriori estimation can be employed.

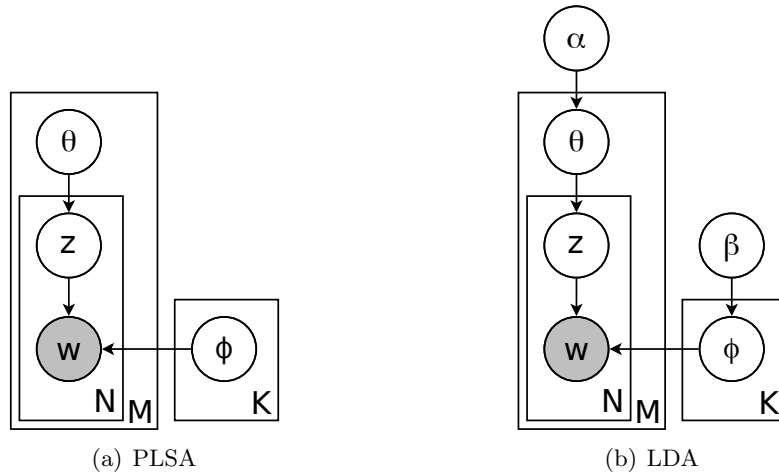


Figure 1.5: Graphical models for (a) Probabilistic Latent Semantic Analysis (PLSA) and (b) Latent Dirichlet Allocation (LDA). While PLSA is able to detect topics as probability distributions over the known vocabulary, in the model by Hofmann [Hof99] both document-topic and topic-word multinomials do not have a prior. LDA adds a Dirichlet prior over the topic-word and the document-topic distribution, which results in a fully generative model that comes with several desired properties, i.e. allowing for sparse document-topic and topic-word priors [BNJ03].

However, in MAP sparse Dirichlet priors might lead to zero probabilities, an unwanted behaviour in probabilistic modelling. MAP also is more likely to yield parameter estimates in local optima because of estimates involving probability estimates of 0 and 1.

It can be shown that MAP parameter estimates lead to lower topic quality compared to alternative inference procedures [AWST12]. Therefore, other inference techniques are employed in practice.

1.6.2 Gibbs sampling

The most popular and easy-to-implement method for inference in complex probabilistic models with latent variables is *Gibbs sampling*.

Probabilistic topic models assume that each observed word is drawn from a topic. The topic of a word is stored in a latent topic assignment variable z_{ij} . In the model of LDA, each topic assignment can take on K different values. For a corpus with $L = \sum_{i=1}^M N_i$ different words, the space of possible settings for the latent variables is huge and has a size of K^L – it is impossible to explore this space with naive methods.

The idea behind Gibbs sampling is to explore the space of possible settings for the latent variables of a probabilistic model by repeatedly re-

estimating a single latent variable based on all the other latent variables, which are kept fixed. Estimates are based on Bayesian inference [Gri02, Hei08].

For LDA, the topic-assignments \mathbf{z} are initialised with random values, i.e. for each topic assignment a topic is drawn from a uniform distribution over the K topics.

The Gibbs sampler then works as follows:

1. A latent topic assignment variable z_{ij} is picked at random, storing the topic of the j th word in document i . Given all the other topic-assignments denoted as \mathbf{z}_{-ij} , one can calculate the probability for each topic $k \in \{1, \dots, K\}$ as $\gamma_{ijk} = p(z_{ij} = k \mid w_{ij}, \mathbf{z}_{-ij}, \mathbf{w}_{-ij}, \theta_i, \phi, \alpha, \beta)$. The j th word in document i is term t , stored in word variable $w_{ij} = t$. This yields the probability [Gri02, Hei08]:

$$\gamma_{ijk} \propto \underbrace{\frac{n_{ik} + \alpha}{N_i + K \cdot \alpha}}_{\text{prob. of topic } k \text{ in document } i} \cdot \underbrace{\frac{n_{kt} + \beta}{n_{k\cdot} + V \cdot \beta}}_{\text{prob. of term } t \text{ under topic } k} \quad (1.34)$$

where n_{ik} stores how often topic k is used in document i

$$n_{ik} = \sum_{j=1}^{N_j} [z_{ij} = k] \quad (1.35)$$

and n_{kt} keeps track of how often term t was assigned to topic k

$$n_{kt} = \sum_{i=1}^M \sum_{j=1}^{N_j} [w_{ij} = t] \cdot [z_{ij} = k] \quad (1.36)$$

where Iverson brackets are employed (i.e. false statements within the square brackets return 0, true statements return 1) and $n_{k\cdot}$ is the total number of words assigned to topic k

$$n_{k\cdot} = \sum_{t=1}^V n_{kt}. \quad (1.37)$$

The topic assignment z_{ij} then is randomly sampled from γ_{ij} :

$$z_{ij} \sim \text{Mult}(\gamma_{ij}) \quad (1.38)$$

i.e. the topic assignment z_{ij} is set to a concrete value which is randomly drawn from the multinomial distribution over the topic probabilities.

2. For each document, the document-topic distribution θ_i can be estimated with:

$$\theta_i = \frac{n_{ik} + \alpha}{N_i + K \cdot \alpha}. \quad (1.39)$$

3. Topic-word distributions are estimated as:

$$\phi_k = \frac{n_{kt} + \beta}{n_{k\cdot} + V \cdot \beta}. \quad (1.40)$$

1.6.3 Variational inference

In the original LDA paper [BNJ03], Blei et al. presented a parameter estimate based on *variational inference*. The idea behind variational inference is simple: Intuitively, one would try to estimate the parameters of a probabilistic model by maximising the posterior (or the marginal distribution) of the model given observations.

However, this is impossible due to the dependencies between the variables in the model [Law01, BNJ03]. One solution to deal with that problem is to assume independence between the variables in an approximate probability distribution and to approximate the optimal parameter setting under that assumption. [XJR12] This is the so-called *mean-field assumption* and the inference is commonly referred to as *variational mean field approximation* [HG09].

Given the likelihood $p(\mathbf{x} \mid \alpha, \beta)$ of the observed words in documents under the hyper-parameters of LDA, it is possible to create a lower bound on the marginal likelihood of the observations and parameters, which is denoted by $q(\mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\phi})$. This lower bound is called the *variational distribution*.

The lower bound is based on Jensen's inequality [JGJS99, BNJ03, KF09]:

$$\begin{aligned} & \log(p(\mathbf{w} \mid \alpha, \beta)) = \\ & \log \int \cdots \int \sum_{\mathbf{z} \in \{1, \dots, K\}^{|\mathcal{D}|}} q(\mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\phi}) \frac{p(\mathbf{w}, \mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\phi} \mid \alpha, \beta)}{q(\mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\phi})} d\theta_1 \cdots \theta_M d\phi_1 \cdots \phi_K \\ & \geq \\ & \int \cdots \int \sum_{\mathbf{z} \in \{1, \dots, K\}^{|\mathcal{D}|}} q(\mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\phi}) \log \frac{p(\mathbf{w}, \mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\phi} \mid \alpha, \beta)}{q(\mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\phi})} d\theta_1 \cdots \theta_M d\phi_1 \cdots \phi_K \\ & = \mathbb{E}_q [\log p(\mathbf{w}, \mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\phi} \mid \alpha, \beta)] - \mathbb{E}_q [\log q(\mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\phi})] \end{aligned} \quad (1.41)$$

where all the parameter settings of \mathbf{z} , $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ are summed over and integrated out in the first step. Looking at the formula, one can see that maximising this lower bound is minimising the Kullback-Leibler (KL) divergence between q and p [JGJS99, KF09] because

$$D_{\text{KL}}(q \parallel p) = \sum_i q(i) \log \left(\frac{q(i)}{p(i)} \right) = - \sum_i q(i) \log \left(\frac{p(i)}{q(i)} \right). \quad (1.42)$$

As mentioned, in variational mean field approximation, the approximate distribution assumes independence between the variables and introduces vari-

ational parameters γ , $\tilde{\alpha}$ and $\tilde{\beta}$ over the variables [JGJS99, BNJ03]:

$$q(\mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\phi}) = \prod_{ij} q(z_{ij} | \gamma_{ij}) \prod_i q(\theta_i | \tilde{\alpha}_i) \prod_k q(\phi_k | \tilde{\beta}_k). \quad (1.43)$$

By iteratively maximising the likelihood for single variational parameters in the lower bound from 1.41, estimates on the latent parameters and variables of the original model are obtained.

The best-performing variational inference scheme for LDA is collapsed variational inference where the multinomial parameters are marginalised out [TNW07, AWST12]. It models the topic-word distributions ϕ and the document-topic distributions θ as dependent variables in the variational approximation, which yields the variational distribution

$$q(\mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\phi}) = q(\boldsymbol{\theta}, \boldsymbol{\phi} | \mathbf{z}) \cdot \prod_{ij} q(z_{ij} | \gamma_{ij}). \quad (1.44)$$

The lower bound on the true likelihood then is given by

$$\log(p(\mathbf{x} | \alpha, \beta)) \geq \mathbb{E}_q[\log p(\mathbf{x}, \mathbf{z} | \alpha, \beta)] - \mathbb{E}_q[\log q(\mathbf{z})] \quad (1.45)$$

and maximising this term with respect to the parameters γ_{ij} of the multinomial distributions over z_{ij} yields an optimum at approximately [AWST12]:

$$\begin{aligned} \gamma_{ijk} &= \frac{\exp\left(\mathbb{E}_q\left[\log(n_{ik}^{-ij} + \alpha) + \log(n_{kt}^{-ij} + \beta) - \log(n_{k\cdot}^{-ij} + V \cdot \beta)\right]\right)}{\sum_{k'=1}^K \exp\left(\mathbb{E}_q\left[\log(n_{ik'}^{-ij} + \alpha) + \log(n_{kt}^{-ij} + \beta) - \log(n_{k'}^{-ij} + V \cdot \beta)\right]\right)} \\ &\approx \underbrace{\frac{\mathbb{E}_q[n_{ik}^{-ij}] + \alpha}{\mathbb{E}_q[n_{i\cdot}^{-ij}] + K \cdot \alpha}}_{\text{prob. of topic } k \text{ in document } i} \cdot \underbrace{\frac{\mathbb{E}_q[n_{kt}^{-ij}] + \beta}{\mathbb{E}_q[n_{k\cdot}^{-ij}] + V \cdot \beta}}_{\text{prob. of term } t \text{ under topic } k} \end{aligned} \quad (1.46)$$

This estimate originally is based on a Taylor expansion of the logarithm of the smoothed, approximately normally distributed counts [TKW08], and in [AWST12] it was mentioned that the variance can be neglected in practice. Interestingly, there is no explicit reason given why the variance can be neglected.

The probabilities of the topic assignments in practice have a sparse Dirichlet prior, i.e. one expects to see either very large or very small probabilities. In this case, the variance of the sum of Bernoulli trials, defined as

$$\text{Var}_q(n_{ik}^{-ij}) = \sum_{l \neq j} \gamma_{ilk} \cdot (1 - \gamma_{ilk}) \quad (1.47)$$

becomes relatively small. The second-order Taylor expansion of the logarithm of smoothed counts at $\alpha + \mathbb{E}_q[n_{ik}^{-ij}]$ is

$$\mathbb{E}_q[\log(\alpha + n_{ik}^{-ij})] \approx \log(\alpha + \mathbb{E}_q[n_{ik}^{-ij}]) - \frac{\text{Var}_q(n_{ik}^{-ij})}{2 \cdot (\alpha + \mathbb{E}_q[n_{ik}^{-ij}])^2} \quad (1.48)$$

Topics with significant counts (i.e. $E_q[n_{ik}^{-ij}] \gg 1$) almost completely determine the topic assignments during inference. Knowing that for such topics $E_q[n_{ik}^{-ij}] \gg \alpha$ under a sparse Dirichlet prior, and given that most topic assignments have a very low probability, the right-hand side of this equation can be approximated as

$$\begin{aligned} \frac{\text{Var}_q(n_{ik}^{-ij})}{2 \cdot (\alpha + E_q[n_{ik}^{-ij}])^2} &\approx \frac{\text{Var}_q(n_{ik}^{-ij})}{2 \cdot (E_q[n_{ik}^{-ij}])^2} = \frac{\sum_{l \neq j} \gamma_{ilk} \cdot (1 - \gamma_{ilk})}{2 \cdot (\sum_{l \neq j} \gamma_{ilk})^2} \\ &\leq \frac{1}{2 \cdot \sum_{l \neq j} \gamma_{ilk}} = \frac{1}{2 \cdot E_q[n_{ik}^{-ij}]} \ll \log(\alpha + E_q[n_{ik}^{-ij}]) \end{aligned} \quad (1.49)$$

which explains why in practice the variance can be neglected for approximating the logarithm of counts.

In the equations given above, expected counts are used, where n^{-ij} indicates that the topic assignment z_{ij} is excluded. $E_q[n_{ik}]$ is the expectation about how often topic k is used in document i :

$$E_q[n_{ik}] = \sum_{j=1}^{N_j} \gamma_{ijk} \quad (1.50)$$

and $E_q[n_{kt}]$ is the expectation of how often term t was assigned to topic k :

$$E_q[n_{kt}] = \sum_{i=1}^M \sum_{j=1}^{N_j} [w_{ij} = t] \cdot \gamma_{ijk}. \quad (1.51)$$

$E_q[n_{k.}]$ denotes the total number of words assigned to topic k :

$$E_q[n_{k.}] = \sum_{t=1}^V E_q[n_{kt}]. \quad (1.52)$$

Equation 1.46 is repeatedly estimated for every topic assignment during inference.

The document-topic distributions and topic-word distributions can be estimated with an equation identical to the estimates of the collapsed Gibbs sampler [SKN12]:

$$E_q[\theta_i] = \frac{E_q[n_{ik}] + \alpha}{N_i + K \cdot \alpha} \quad (1.53)$$

and

$$E_q[\phi_k] = \frac{E_q[n_{kt}] + \beta}{E_q[n_{k.}] + V \cdot \beta} \quad (1.54)$$

respectively.

Because topic assignments follow multinomial distributions parametrised by variational parameters, variational inference for LDA stores more information about the posterior over z during inference and therefore avoids

local optima and converges significantly faster compared to Gibbs sampling [TNW07, AWST12]. However, variational inference techniques in general require the derivation of the maximum of a lower bound on the probability function of a potentially complex model. This maximum often has to be approximated for performance reasons. Additionally, the variational parameters have to be stored and consume additional memory during inference. Due to these differences, in practice Gibbs sampling is often favoured over variational inference.

1.7 Evaluation of Topic Models

There are hundreds of different probabilistic and non-probabilistic models for detecting topics in text corpora. In order to select the optimal topic model for a given collection of documents and to find the optimal parametrisation, one needs measures for evaluating the quality of a trained topic model. Evaluation is based on two fundamentally different approaches. Human evaluation aims at testing the semantic coherence of the top words of each topic in a user study. Theoretical approaches to evaluation calculate a score based on the likelihood of new observations under a trained topic model. The theoretical approaches are limited to probabilistic topic models, while the human evaluation might be applied to non-probabilistic topic models such as LSA.

1.7.1 Human evaluation

Human evaluation of topic models is based on the assumption that the top-words of topics should be semantically coherent. In [CBGW⁺09], Chang et al. propose a *word intrusion* based evaluation, where human raters are presented with the top- N words of a topic ($N = 5$ in the original paper) and an *intruder word* which has a high probability under another topic. The task is to detect the intruder word, which is easier for semantically coherent, i.e. high qualitative topics, and harder for semantically incoherent topics. The rate of correct answers over the raters can be used as a test score, and additionally a box plot can be employed to compare the variance of the performance of raters.

1.7.2 Perplexity

Topic models can be interpreted as hypotheses on the creation of documents. A probabilistic topic model explains documents and their words as the result of a probabilistic process. In the case of LDA, first topic-word distributions and document-topic distributions are sampled from Dirichlet distributions, and then words in documents are created by repeatedly drawing topics from the document-topic distribution of a given document and

subsequently drawing words from the topic-word distribution of the drawn topic.

After the estimation of model parameters on a training corpus using expectation-maximisation, Gibbs sampling or variational inference, it is possible to evaluate the quality of the model based on the likelihood of previously unseen documents. If the hypothesis on the creation of documents and the parameter estimates are good, then the likelihood of new observations will be high. Otherwise, new observations would be unlikely or even impossible – in case the parameters include zero probabilities. The latter can occur either in cases where there are no priors or where sparse priors are given and inference techniques such as maximum a posteriori inference are employed, which might lead to zero probabilities.

The likelihood of observations is a product over the probabilities of each single observation. Therefore, a higher number of observations trivially reduces the likelihood and it is necessary to normalise the likelihood in order to calculate a test score for comparing several topic models.

This can be achieved by normalising the likelihood by the number of words of the previously unseen documents. As the likelihood can take on very small values, the negative log likelihood is used instead and the exponential function is applied after normalisation. This yields the formula for perplexity [BNJ03]:

$$\text{perplexity}(D_{\text{test}}) = \exp \left(- \frac{\sum_{j=1}^M \sum_{i=1}^{N_i} \log p(w_j^i)}{\sum_{j=1}^M N_i} \right) \quad (1.55)$$

where lower perplexity values indicate a higher average log likelihood and therefore a better model performance.

For evaluating a topic model, the parameters are learned on a training corpus D_{train} and calculate the perplexity on a testing corpus D_{test} , learning document-specific parameters for the test documents but keeping the learned topic-word distributions fixed.

If the inference algorithm is non-deterministic, like in Gibbs sampling, an average perplexity is calculated over a number of runs of the sampler.

1.8 Non-Parametric Topic Models

All the probabilistic topic models presented so far require a parameter K for the number of topics to be set. It is theoretically possible to do a human evaluation of the topic quality for different settings of K , but due to high costs this is not feasible in practice. Another approach could be based on perplexity, but unfortunately the perplexity simply decreases (i.e. improves) with an increasing number of topics [BNJ03] even when the topic quality measured by human raters decreases [CBGW⁺09].

One solution for setting the topic count parameter is to set the number of topics to an infinite number and to use a prior over this space of infinite topics which de-facto reduces the number of used topics. The most popular prior over this infinite space of topics is the *Hierarchical Dirichlet Process* (HDP) which is a coupling of *Dirichlet processes*.

1.8.1 The Dirichlet process

The *Dirichlet process* (DP) is defined as a probability density function over a probability measure G_0 with two parameters γ , the *scaling parameter*, and H , the *base measure*:

$$G_0 \sim \text{DP}(\gamma, H). \quad (1.56)$$

In the case of topic models, H is the infinite probability measure on of all possible topic-word distributions, i.e. it contains all possible parameter settings for topic-word multinomials with V categories, where V is the number of terms in the vocabulary. G_0 is a probability measure on the measure on topics which assigns a probability to every single topic from the measure H . The name *Dirichlet process* stems from the property of the DP that for any partition (A_1, \dots, A_N) of the base measure H , the random probability vector of the partitions of G_0 follows a Dirichlet distribution [Fer73, TJBB06]:

$$(G_0(A_1), \dots, G_0(A_N)) \sim \text{Dir}(\gamma H(A_1), \dots, \gamma H(A_N)) \quad (1.57)$$

G_0 both contains information about the topic-word distributions from H and their probabilities.

The stick-breaking process

It is possible to separate the topics and their probability in G_0 by defining the probability vector G_0 as an infinite sum over the drawn topics and their probabilities [TJBB06]. If the k th topic is denoted by ϕ_k , a multinomial topic-word distribution, G_0 is obtained as an infinite weighted sum over topics:

$$G_0 = \sum_{k=1}^{\infty} \beta_k \delta_{\phi_k} \quad (1.58)$$

where β_k are weights drawn from beta distributions with

$$\phi_k \sim H \quad (1.59)$$

$$\tilde{\beta}_k \sim \text{Beta}(1, \gamma) \quad (1.60)$$

$$\beta_k = \tilde{\beta}_k \cdot \prod_{j=1}^{k-1} (1 - \tilde{\beta}_j) \quad (1.61)$$

and δ_{ϕ_k} being a Dirac delta at ϕ_k which assigns the weight 1 to the location of ϕ_k in the topic space.

As the weights β_k can be interpreted as parts of a stick which is repeatedly broken at position $\tilde{\beta}_k$, this process is called the *Stick Breaking Process* (SBP) and the notation is

$$\beta \sim \text{SBP}(\gamma). \quad (1.62)$$

The Chinese restaurant process

Another metaphor for the Dirichlet process is a scheme similar to the Polya urn scheme from Section 1.4.4. Imagine a Chinese restaurant which has infinitely many tables. Customers enter the restaurant and decide to sit on an occupied table with a probability proportional to m_k , the number of customers seated at that table. A customer sits at an unoccupied table with a probability proportional to γ . If z_i is a variable holding the decision for the i th customer, then the customer sits down at table k with probability

$$p(z_i = k) = \begin{cases} \frac{m_k}{(\sum_{j=1}^K m_j) + \gamma} & \text{for existing tables} \\ \frac{\gamma}{(\sum_{j=1}^K m_j) + \gamma} & k = (K + 1) \text{ (open new table)} \end{cases} \quad (1.63)$$

where the counts are updated after every step and K is the number of occupied tables. Repeating this scheme for infinitely many customers yields probabilities for choosing a table proportional to the probabilities of the stick breaking process. This scheme is called the *Chinese Restaurant Process* (CRP) and was introduced by Aldous [Ald85]. It is easy to see that the scaling parameter γ influences the rate of newly created tables – high values lead to a larger number of occupied tables. In the setting of topic models, the number of topics detected is influenced by the scaling parameter.

Hierarchical Dirichlet process topic models

In order to create a probabilistic model which explains words in documents by topics, it is possible to sample the topics of a document as a probability measure $G_d \sim DP(H, \gamma)$ from a Dirichlet process with H as base measure on the topics. Though this procedure would result in a valid probabilistic model, the model would imply that documents draw totally different topics from H , as the probability of two documents to draw the same topic from the infinite measure H over topics would be zero.

Therefore, the document-specific measures on the topic space have to be coupled. In the *Hierarchical Dirichlet Process* (HDP), this is achieved by sampling a common base measure G_0 on the topic space from a Dirichlet process and then sampling document-specific base measures G_m^d from this global base measure. By this, the weight of topics is concentrated at a finite number of topics in G_0 with probability one [BM73] and documents sample

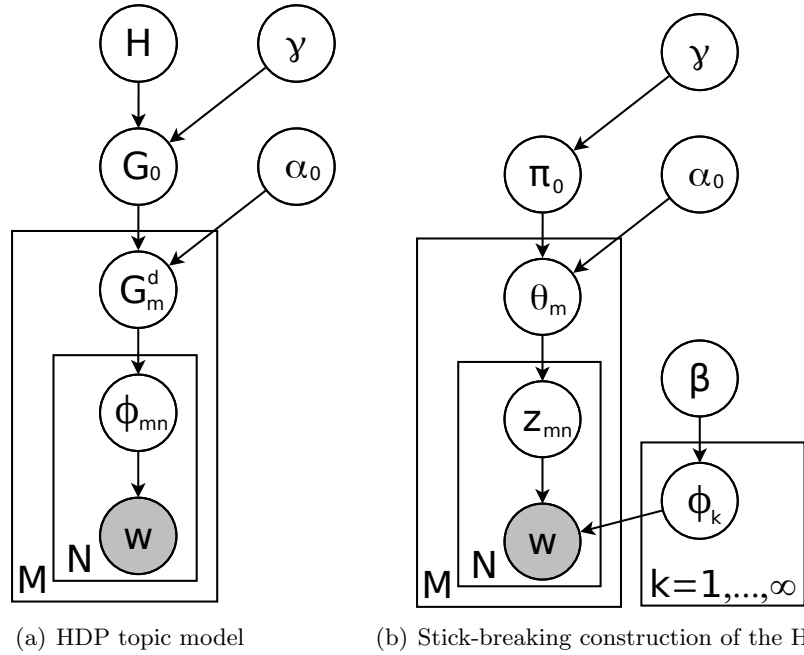


Figure 1.6: The Hierarchical Dirichlet Process (HDP) topic model. (a) Graphical model of the two-level hierarchical Dirichlet process topic model. Topics of documents each are drawn from Dirichlet processes with a shared, global topic distribution as base measure. The global topic distribution is itself a measure drawn from a Dirichlet process with the measure H over all possible topics as base measure. The global topic distribution makes sure that the documents share common topics. The HDP does not require a parameter for the number of topics as PLSA and LDA. ϕ denotes a topic-word distribution drawn from G_m^d , the base measure of document m which gives a document-specific weighting of topics. The global base measure G_0 yields a global weighting of topics. (b) Stick-breaking construction of the HDP topic model. The global topic distribution π_0 is drawn from a Stick-Breaking Process (SBP) which resembles the weights of the Dirichlet process. There are infinitely many topics drawn from a Dirichlet distribution (previously part of H) and the global topic distribution is a distribution over the indices of these topics.

topics from this finite set of topics. The corresponding graphical model is shown in Figure 1.6(a). The generative model of the HDP then is:

1. Sample a global measure on topics from a Dirichlet process with the measure H over topics as base measure and scaling parameter γ :

$$G_0 \sim \text{DP}(\gamma, H) \quad (1.64)$$

with global mixing proportions $\pi_0 \sim \text{SBP}(\gamma)$ and H placing a Dirichlet prior over the topics: $H = \text{Dir}(\beta)$.

2. For every document, sample a document-specific base measure from a Dirichlet process with base measure G_0 and scaling parameter α_0

$$G_m^d \sim \text{DP}(\alpha_0, G_0) \quad (1.65)$$

3. For every word w_{mn} in document m , draw a topic-word distribution ϕ_{mn} from G_0 and draw the word from the multinomial ϕ_{mn} :

$$\phi_{mn} \sim G_m^d \quad (1.66)$$

$$w_{mn} \sim \phi_{mn} \quad (1.67)$$

For simplicity, the different sampled topics ϕ_{mn} are indexed and topic-assignment variables z_{mn} are introduced, holding the index of the topic assigned to word w_{mn} .

The Chinese restaurant franchise

The metaphor of the Chinese restaurant process can be extended for a hierarchical Dirichlet process by introducing the notion of dishes, which are served across several restaurants. In the *Chinese restaurant franchise*, there exist several franchise restaurants (corresponding to document-level Dirichlet processes) in which customers again enter the restaurant as in a CRP. Additionally, when a new table is opened, a dish k is chosen which is served to everybody who will be sitting at that table. The variable m_{jk} stores at how many tables dish k is served in restaurant j and n_{jk} stores how many customers are eating dish k (i.e. sit at tables which serve dish k). The dishes correspond to topics.

When a new table in a franchise-restaurant is opened, the dish is drawn from a global distribution over dishes, which is sampled from a global Dirichlet process. This global Dirichlet process makes sure that customers in different restaurants share the same dishes, as the probability mass generated from a Dirichlet process is discrete with probability one [Fer73]. To stay in the CRP metaphor, the global proportions correspond to customers in a global restaurant sitting at tables. Every time a new table in the global

restaurant is opened, a dish is drawn from H (an infinite measure on all possible topics) and a new topic is created.

In the case of a HDP topic model, the dishes correspond to the topics present in a document, n_{jk} are the topic counts and m_{jk} gives the number of times a topic was drawn from the global topic distribution. This means that the global topic distribution – itself a draw from a DP – behaves like an own restaurant with $\sum_{j=1}^M m_{jk}$ customers sitting at a table serving dish k and every table serves a different dish, as there is no concentration on a discrete set of topics in the measure H on the infinite space of topics.

1.8.2 Inference for the HDP

Inference for models involving infinite probability measures such as the Dirichlet process is different from standard inference techniques. For Gibbs sampling, the infinite parameter space is typically reduced to the parameters actually used in the current sampling step, while the unused parameters (e.g. the infinitely many topics of a HDP topic model) are treated as a single new topic. Variational methods rely on a truncation of the infinite topic space to K dimensions, and there are different flavours of how to implement the resulting sampler. In this section, state-of-the-art methods in Gibbs sampling and variational inference for hierarchical Dirichlet processes are presented.

Gibbs sampler for the HDP

A Gibbs sampler for the HDP is given by [TJBB06, Hei06]. Instead of sampling each topic assignment from a multinomial distribution over a fixed number of topics, the number of topics is given by the number of *used* topics K . Additionally, there is a probability for sampling from the space of unused topics, which is proportional to the scaling parameter α_0 multiplied with the probability of a given word under the base measure H (the measure on the topic space).

The Gibbs sampler is as follows:

- A latent topic assignment variable z_{ij} is picked at random, storing the topic of the j th word in document i . Given all the previous topic-assignments denoted as \mathbf{z}_{-ij} , it is straightforward to calculate the probability for each topic $k \in \{1, \dots, K\}$ as $\gamma_{ijk} = p(z_{ij} = k \mid w_{ij}, \mathbf{z}_{-ij}, \mathbf{w}_{-ij}, \gamma, \alpha_0)$. The j th word in document i is term t , stored in the given word variable $w_{ij} = t$, and the global mixing proportions are $\beta \sim SBP(\alpha_0)$. Then the probability is given by:

$$\gamma_{jmk} \propto \begin{cases} \frac{n_{jk} + \alpha_0 \cdot \pi_{0k}}{N_i + K \cdot \alpha_0} \cdot \frac{n_{kt} + \beta}{n_k + V \cdot \beta} & \text{for previously used topics} \\ \alpha \cdot \pi_{0\text{new}} \cdot \frac{1}{V} & \text{for a new topic} \end{cases} \quad (1.68)$$

where $1/V$ is the probability of a word under a new topic given a vocabulary of size V . H is a symmetric Dirichlet distribution over all possible topic-word multinomials (i.e. there is no preference for any specific term) and thus

$$p(w \mid \phi_{new}, H) = \int p(w \mid \phi_{new}) \cdot p(\phi_{new} \mid H) d\phi_{new} = \frac{1}{V} \quad (1.69)$$

for every word w . Note that the pseudo-counts $\alpha_0 \cdot \pi_{0k}$ correspond to the event of opening a new table and choosing dish k which is already eaten at another table (in this or in another franchise restaurant). The topic assignment is sampled from this $K + 1$ dimensional multinomial:

$$z_{ij} \sim \gamma_{jmk}. \quad (1.70)$$

If a topic is not in use any more (i.e. after an update of a topic-assignment z_{ij} there exists no topic assignment which takes on the topic index), the topic is removed from the list of used topics and K is decreased. If a new topic is sampled, K is increased and the index added to the list of used topics.

The count variables n_{ik} , n_{kt} and n_k are identically estimated as in the Gibbs sampler for LDA given in Section 1.6.2.

- The global topic-proportions π_0 depend on the topic frequencies in the documents. However, the connection is not direct. Recall that in the Chinese restaurant metaphor, words in documents correspond to customers eating a dish, which is drawn from the global topic distribution every time a new table is opened. Therefore, it is necessary to count the number of tables which serve dish k in order to find out about the global distribution of topics.

There are two ways of dealing with that issue. It is possible to keep track of the number of tables during the Gibbs sampling process, which would require a different sampling equation in Eq. 1.68. Alternatively, one can estimate the number of tables using the number of customers, the scaling parameter α_0 and the global mixing proportions π_0 : For a new customer entering the franchise-restaurant, the probability of opening a new table for topic k given that n_k customers are already eating dish k is $\frac{\alpha_0 \cdot \pi_{0k}}{n_k + \alpha_0 \cdot \pi_{0k}}$, and $\frac{n_k}{n_k + \alpha_0 \cdot \pi_{0k}}$ is the probability of a customer to sit down with the n_k other customers at an existing table. Given that there are n_k customers in a franchise-restaurant sitting at m_k tables eating dish k , there exist multiple combinations which result in the observed table counts, and each combination has a well-defined probability.

Accounting for the combinations [Ant74, TJBB06] yields a probability for table counts m_k given observations and parameters:

$$p(m_k = m \mid n_k, \alpha_0, \boldsymbol{\pi}_0) = \frac{\Gamma(\alpha_0 \cdot \pi_{0k})}{\Gamma(n_k + \alpha_0 \cdot \pi_{0k})} \cdot s(n, m) \cdot (\alpha_0 \cdot \pi_{0k})^m \quad (1.71)$$

where $s(n, m)$ are *unsigned Stirling numbers of the first kind* which are recursively defined as [TJBB06]:

$$\begin{aligned} s(0, 0) &= 1 \\ s(n, 0) &= 0 \\ s(n, n) &= 1 \\ s(n, m) &= 0 \quad (\text{for } m > n) \\ s(n+1, m) &= s(n, m-1) + n \cdot s(n, m). \end{aligned} \quad (1.72)$$

Given the table counts, the global topic distribution of the top-level CRP follows a Dirichlet distribution from which it is sampled during inference [TJBB06]:

$$\boldsymbol{\pi}_0 \sim \text{Dir}(m_1, \dots, m_K, \gamma). \quad (1.73)$$

- For each document, the document-topic distribution θ_i can be estimated with:

$$\theta_i = \frac{n_{ik} + \alpha_0 \cdot \pi_{0k}}{N_i + \alpha_0} \quad (1.74)$$

- Topic-word distributions are estimated identically to the Gibbs sampler for LDA:

$$\phi_k = \frac{n_{kt} + \beta}{n_{k\cdot} + V \cdot \beta} \quad (1.75)$$

Variational inference for the HDP

In the Gibbs sampler for the HDP, the fact that topic-assignments take on a single topic index during inference was used to limit the number of topics to the number of topics used in the current state of the sampler. All other topics were treated as a single topic which is representing all new topics.

Variational inference, as introduced in Section 1.6.3, relies on a lower bound on the marginal likelihood and learns variational parameters over the hidden variables. For the topic assignments, it learns a probability distribution over all possible topics. In the case of the HDP, there are infinitely many topics and therefore a distribution over infinitely many topics would be required.

To deal with this issue, the variational distribution can be limited to K topics, where the infinitely many topics from the Dirichlet process are truncated and the last topic takes on the rest of the probability mass [BJ06a].

The most popular inference schemes are based on the stick-breaking representation of the Dirichlet process [BJ06a, TKW08, SKN12] which is depicted in Fig. 1.6(b): The global topic weights $\boldsymbol{\pi}_0$ are drawn from a stick-breaking process, giving probabilities over infinitely many topic-word distributions drawn from a Dirichlet distribution:

$$\boldsymbol{\pi}_0 \sim \text{SBP}(\gamma), \quad \boldsymbol{\theta}_m \sim \text{Dir}(\alpha_0 \boldsymbol{\pi}_0), \quad \phi_k \sim \text{Dir}(\beta); \quad k = 1, \dots, \infty \quad (1.76)$$

and the stick-breaking process makes use of auxiliary variables $\tilde{\pi}_{0k}$ for the stick lengths:

$$\pi_{0k} = \tilde{\pi}_{0k} \prod_{l=1}^{l=k-1} (1 - \tilde{\pi}_{0l}), \quad \tilde{\pi}_{0k} \sim \text{Beta}(1, \gamma). \quad (1.77)$$

Integrating out the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ improves the inference with respect to speed and convergence rate [TKW08]. The resulting marginal distribution over the latent topic assignments and observations then is given by:

$$\begin{aligned} p(\mathbf{w}, \mathbf{z} \mid \alpha_0, \beta, \boldsymbol{\pi}_0) &= \\ & \int \cdots \int \prod_{m=0}^M \left(p(\boldsymbol{\theta}_m \mid \alpha_0 \boldsymbol{\pi}_0) \cdot \prod_{n=1}^{N_m} p(z_{mn} \mid \boldsymbol{\theta}_m) p(w_{mn} \mid \boldsymbol{\phi}_{z_{mn}}) \right) \\ & \cdot \prod_{k=1}^K p(\boldsymbol{\phi}_k \mid \beta) d\boldsymbol{\theta}_1 \cdots d\boldsymbol{\theta}_M d\boldsymbol{\phi}_1 \cdots d\boldsymbol{\phi}_K \\ &= \int \cdots \int \prod_{m=0}^M \left(\frac{\Gamma(\sum_{k=1}^K \alpha_0 \pi_{0k})}{\prod_{k=1}^K \Gamma(\alpha_0 \pi_{0k})} \cdot \prod_{k=1}^K \theta_{mk}^{\alpha_0 \pi_{0k} - 1} \cdot \prod_{n=1}^{N_m} \theta_{m z_{mn}} \phi_{z_{mn} w_{mn}} \right) \\ & \cdot \prod_{k=1}^K \left(\frac{\Gamma(\sum_{t=1}^V \beta)}{\prod_{t=1}^V \Gamma(\beta)} \cdot \prod_{t=1}^V \phi_{kt}^{\beta - 1} \right) d\boldsymbol{\theta}_1 \cdots d\boldsymbol{\theta}_M d\boldsymbol{\phi}_1 \cdots d\boldsymbol{\phi}_K \\ &= \underbrace{\prod_{m=0}^M \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_0 + n_{m\cdot})} \cdot \prod_{k=1}^K \frac{\Gamma(\alpha_0 \pi_{0k} + n_{mk})}{\Gamma(\alpha_0 \pi_{0k})}}_{\text{Probability of topics in documents}} \cdot \underbrace{\prod_{k=1}^K \frac{\Gamma(V \cdot \beta)}{\Gamma(\beta + n_{k\cdot})} \cdot \prod_{t=1}^V \frac{\Gamma(\beta + n_{kt})}{\Gamma(\beta)}}_{\text{Probability of terms in topics}} \\ & \quad (1.78) \end{aligned}$$

where Eq. 1.21 is employed and (in contrast to [TKW08]) it is assumed that the Dirichlet prior on the topic-word distributions is symmetric because of the findings in [WMM09] that asymmetric topic priors are rarely improving the topic quality in practice.

To simplify inference, Teh et al. [TKW08] introduce auxiliary variables to replace the gamma functions on the left-hand side of the formula with the beta distribution (see Eq. 1.8) and the distribution over tables from

Eq. 1.71:

$$\begin{aligned}
p(\mathbf{w}, \mathbf{z} \mid \alpha_0, \beta, \boldsymbol{\pi}_0, \boldsymbol{\eta}, \mathbf{s}) = & \\
\prod_{m=0}^M \frac{\eta_m^{\alpha_0-1} \cdot (1 - \eta_m)^{N_m-1} \prod_{k=1}^K s(n_{mk}, s_{mk}) \cdot (\alpha_0 \pi_{0k})^{s_{mk}}}{\Gamma(n_{m\cdot})} & \\
\cdot \prod_{k=1}^K \frac{\Gamma(\beta)}{\Gamma(\beta + n_{k\cdot})} \cdot \prod_{t=1}^V \frac{\Gamma(\beta + n_{kt})}{\Gamma(V \cdot \beta)}. & \quad (1.79)
\end{aligned}$$

Integrating over the auxiliary variables yields the original formula. One can interpret \mathbf{s}_m as the number of tables in restaurant m .

The variational approximation q assumes independence between all variables to be inferred, and maximising the lower bound with respect to the variational parameters yields the inference equations. As exact inference on the variational distribution – which is itself an approximation – is not required in practice, the approximations by Asuncion et al. [AWST12] and Sato et al. [SKN12] are employed to yield:

$$q(z_{mn} = k) \approx \frac{\mathbb{E}_q[n_{mk}] + \alpha_0 \cdot \pi_{0k}}{n_{m\cdot} + \alpha_0} \frac{\mathbb{E}_q[n_{kt}] + \beta}{\mathbb{E}_q[n_{k\cdot}] + V \cdot \beta} \quad (1.80)$$

where the counts – as in the variational inference scheme for LDA – denote the expected counts under the variational distribution.

For the parameters $\tilde{\boldsymbol{\pi}}_0$ of the stick-breaking process, the approximation by Sato [SKN12] estimates the beta distributed stick lengths as

$$\begin{aligned}
q(\tilde{\pi}_{0k}) &= \text{Beta}(a_k, b_k) \propto \tilde{\pi}_{0k}^{a_k-1} \cdot (1 - \tilde{\pi}_{0k})^{b_k-1} \\
a_k &= 1 + \sum_{m=1}^M \mathbb{E}_q[n_{mk} \geq 1] \\
b_k &= \gamma + \sum_{m=1}^M \sum_{l=k+1}^K \mathbb{E}_q[n_{ml} \geq 1]. \quad (1.81)
\end{aligned}$$

The counts n again denote expected counts, and table counts are estimated as

$$\mathbb{E}_q[n_{mk} \geq 1] = 1 - \prod_{n=1}^{N_m} q(z_{mn} \neq k). \quad (1.82)$$

The hyperparameters α_0 , β and γ can be sampled during inference. A generalised estimator for the parameter of a Dirichlet distribution is given in Chapter 3. Formulas for hyperparameter inference in a HDP are given in Chapter 4.

Finally, the topic-word distributions are estimated as in the variational inference for LDA and the document-topic distributions can be estimated as

in the Gibbs sampler for the HDP, with expected counts instead of counts based on hard topic assignments:

$$\theta_i = \frac{n_{ik} + \alpha_0 \cdot \pi_{0k}}{n_{i\cdot} + \alpha_0}. \quad (1.83)$$

1.9 Summary

In this section, basic concepts behind probabilistic modelling which are used in this thesis were introduced. The binomial and beta distribution will be employed to develop novel probabilistic power indices in Chapter 2. A mixture of Fisher distributions and expectation-maximisation will be used in Chapter 3 for clustering geographically distributed documents. The hierarchical Dirichlet process topic model and the presented inference schemes, Gibbs sampling and variational inference, are the ingredients for developing novel probabilistic models for integrating context information, which are presented in Chapter 3 and Chapter 4.

Chapter 2

Single-Context Voting Models

The focus of this thesis is on novel, high-performance and yet easy-to-use probabilistic models which allow for the inclusion of context information.

In order to demonstrate how context-specific probabilistic models improve model quality (as measured by perplexity), the setting of voting models is employed in this chapter. Specifically, novel generalisations of power indices which account for system-specific voting bias are introduced and their advantages over existing power indices are demonstrated on the largest available voting history of an online delegative democracy [KKH⁺15].

The LiquidFeedback platform of the German Pirate Party is studied with a focus on voting behaviour and the power of voters. The core contribution of the study is the development of a novel probabilistic power index which permits the inclusion of system-specific voting behaviour via prior distributions. It serves as the most basic example on how probabilistic models can be improved using context information.

2.1 Problem Setting and Approach

In the last decade, the World Wide Web has increasingly been adopted for facilitating political processes and conversations [LWBS14]. The Web has also sparked the development of novel voting and democracy platforms impacting both societal and political processes. Today, a wide range of online voting platforms are available, based on different democratic methods such as consensual decision making, liquid democracy [Pau14] or dynamically distributed democracy [TFH14]. These platforms are becoming increasingly popular and political movements and parties have started adopting them to open up and facilitate political coordination. In contrast to experimental data or simulations (e.g. from game theory), the behaviour of voters on these platform is realistic, i.e. voting takes place in a natural environment and the

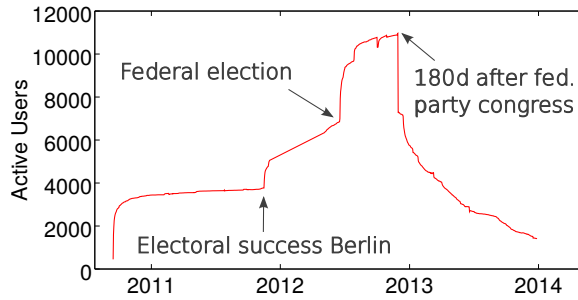


Figure 2.1: User activity. Active users on the LiquidFeedback platform of the German Pirate Party over time. Users are labelled inactive after 180 days without login. Several events led to a rise and decrease in activity.

decisions of voters have a real political impact. Having such a natural setting is crucial for studying voting behaviour in real life political movements and for validating research on voting behaviour and measures of power [Loe99]. Yet, this kind of data has historically been elusive to researchers.

LiquidFeedback represents a popular platform which implements support for *delegative democracy*. A delegative democracy can be described as a mixture of direct and representative democracy. In contrast to a representative democracy, all voters in a delegative democracy in principle are equal, i.e. every voter could directly vote on proposals. Alternatively, each voter can delegate his vote to another voter, raising the voting weight of the delegate by one. The delegate can again delegate his voting weight to a third user and so forth, creating a transitive delegation chain. A key innovation of delegative democracy platforms is the ability of every voter to revoke his delegated votes at any point, preserving full control over his votes and allowing for the emergence of dynamic delegation structures in contrast to representative voting systems. Votes are public and pseudonymous, and therefore both individual and collective voting behaviour can be analysed. A common objection against the use of these platforms is the nature of delegations, as they can potentially give rise to so-called *super-voters*, i.e., powerful users who receive many delegations. It has been asserted in the past that the presence of these *super-voters* undermines the democratic process, and therefore delegative democracy should be avoided.

2.1.1 Problem setting

In order to assess the true potential and limitations of delegative democracy platforms to facilitate political discourse and agenda setting, it is necessary to quantify the distribution of power in delegative democracies and to detect *super-voters* by measuring the power of individual voters. System-specific voting behaviour is expected to influence the power of voters. For instance, a system in which all voters always vote against proposals is unlikely to

assign power to a new voter, as it is impossible to win the necessary majority for passing a vote by coalition. Therefore, the voter behaviour has to be understood and modelled to account for the *system context* in which votes are given. Tapping into the complete voting history and delegation network from world’s largest delegative democracy platform (operated by the German Pirate Party), it is essential to **(i) understand how people vote** in delegative democracy platforms such as LiquidFeedback, and how they delegate votes to *super-voters*. Based on these insights, it is possible to **(ii) model voting power**: how power can be assessed in online democracy systems and how it is used.

2.1.2 Approach

For understanding the behaviour of voters in an online delegative democracy, the voting behaviour of members of the German Pirate Party is analysed over a period of four years from 2009–2013. The German Pirate Party has adopted LiquidFeedback as their online delegative democracy platform of choice. First, the temporal delegation network of users is analysed and the emergence of power structures identified, i.e. the presence of *super-voters* within the party. Next, established power indices from game theory and political science theory – the Shapley and Banzhaf power index [Sha54, Ban65] – are applied to assess the theoretical power of super voters. Then, the predicted power of super voters is compared with their potential as well as their exercised power based on real world voting data. The analysis reveals a *clear gap* between existing theoretical power indices and actual user voting behaviour. As a result, a novel generalisation of power indices is presented that better captures voting behaviour by including the system-specific voting bias. Finally, the proposed power indices are evaluated with data from the LiquidFeedback platform.

2.2 Delegative Democracies

First steps towards the direction of a delegative democracy were published in 1884 by Charles L. Dogson, better known under his pseudonym Lewis Carroll. In his book about the mathematical properties of voting mechanisms, Dogson proposes a voting scheme where elected candidates may delegate their votes to other candidates. The delegated votes then can be further passed to other candidates [Dod84]. A review of further works which influenced the development of the concept of a delegative democracy can be found in [Jab11, Pau14]. Based on these ideas, the novel concept of delegative voting was developed and recently popularised. A formalisation of a delegative democratic system is given in [YYT07]. The implementation of delegative voting systems is non-trivial as loops in the delegation network

have to be detected and resolved and regaining votes potentially can affect a long delegation chain.

2.2.1 Democracy platforms

Existing software implementations of delegative democracy include *LiquidFeedback*¹, *Agora Voting*², *GetOpinionated*³ and *Democracy OS*⁴. This analysis is based on the online voting platform of the German Pirate Party, an instance of LiquidFeedback, a free software that implements an online platform in which votes can be conducted, and users can delegate their vote to other users. LiquidFeedback was adopted by the German Pirate Party in May 2010 [Pau14] and has 13,836 users as of January 2015.

2.2.2 Pirate parties

Pirate parties are an international political movement with roots in Sweden [Fre13], where legal cases related to copyright law led to the formation of a party advocating modern copyright laws and free access to information [MO08]. The scope of the party quickly broadened and nowadays active Pirate parties exist in 63 countries of which 32 are officially registered parties. The German Pirate Party is the largest of all pirate parties with 15,285 members as of January 2016.

2.3 Description of the Dataset

The German Pirate Party maintained the largest installation of LiquidFeedback with 13,836 registered users, and used the software to survey the opinion of members. The German Pirate Party's installation of LiquidFeedback represented the largest online community implementing delegative democracy. This study uses a complete dataset created from daily database dumps of that installation, ranging from August 13 2010 up to November 25 2013, spanning 1,200 days. The data was available to all party members until the system was deactivated in 2015.

2.3.1 LiquidFeedback platform

LiquidFeedback is a complex and powerful implementation of a delegative democracy. A brief overview of the most important processes and policies within the system is given here. More detailed descriptions are available from Jabbusch and from Paulin [Jab11, Pau14].

¹<http://www.liquidfeedback.org>

²<http://www.agoravoting.com>

³<http://www.github.com/getopinionated>

⁴<http://www.democracyos.org>

In LiquidFeedback as used in the German Pirate Party, members can create *initiatives* which are to be voted on to obtain the current opinion of the party members, e.g. for collaboratively developing the party program. Initiatives are grouped into *issues* which group competing initiatives for the same issue. For instance, if a user proposes an initiative to reduce the emission of CO₂ by subsidising the construction of wind turbines, another user could create a competing initiative to subsidise solar fields. Furthermore, issues belong to *areas* which represent main topics such as environmental policies. Each user can create new initiatives, which need a minimum first quorum of supporters for being voted upon. In LiquidFeedback, votes can be delegated to other voters at three levels: At the global level, meaning that all initiatives can be voted for by the delegate on behalf of the delegating user; at the area level, so that delegations are restricted on an area; or at the issue level. The actions of every voter are recorded and public, allowing the control of delegates at the expense of votes not being secret.

2.3.2 Dataset

In total, the dataset includes 499,009 single votes for 6,517 *initiatives* belonging to 3,565 *issues*. Throughout the four-year observation period, a total of 14,964 delegations were made on the global, *area* or *issue* level, constituting the delegation network. The number of active users over the observation period is shown in Fig. 2.1. Usage of LiquidFeedback in the German Pirate Party fluctuates with political events in the party. A strong growth in active users is observed after the electoral success of the Berlin Pirate Party in 2011, where 8.9% of the votes were received. Another point of growth is observed prior to the German federal election in 2012. 180 days after the programmatic federal party congress in 2011, the number of active users drops significantly, when the voting system was used to prepare proposals for the party congress. After the congress, a critical debate on the future role of delegative democracy for the Pirate party started. In a discussion on the effect on *super-voters* – i.e. users with a large share of incoming delegations – the democratic nature of the system was questioned, and many users became inactive.

2.4 Voting Behaviour and Delegations

In the following, different aspects of voting behaviour are studied using the complete voting history and the temporal delegation network.

2.4.1 Existence and role of super-voters

In order to explore whether super-voters exist, and whether they wield an over-proportional influence in the system, we plot the distribution of voting

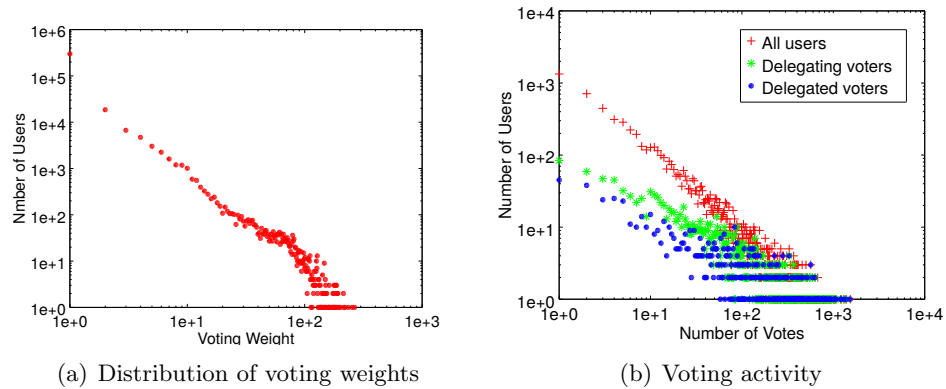


Figure 2.2: Distribution of voting weights (a) and voting activity (b). (a) shows the total number of distinct voters who had the given voting weight at any point in time during the observation period, summing over global, *area* and *issue* delegations. (b) shows the activity distribution of all voters, for delegates and delegating voters, measured by the number of votes cast. The activity distributions are power law like. Delegating voters and delegates vote more frequently, i.e. are more active than other voters.

weights in the delegation network in Fig. 2.2(a), summing over global, *area* and *issue* delegations. Most voters have no delegations (i.e. their voting weight is 1) and a small set of voters possesses a huge voting weight – the *super-voters*. There are only 38 individual voters who received more than 100 delegations in the voting history, and therefore the non-significant statistics for these voters are excluded from the following figures of this study.

The practical power of super-voters does not only depend on their voting weight – it also depends on how often a voter actually participates in votes. One could ask: Are delegates more active than normal users? The overall activity of voters is power law distributed (at a significance level of 0.05) with an exponent of 1.87 and a median of 8. 3,658 members voted more than 10 times, 1,156 voted more than 100 times and 54 members voted more than 1,000 times. The power law exponent of users who received delegations during the observation period is 2.68 with median 64, indicating an increased activity. For controlling this result, the exponent is compared with the exponent of users who delegated their vote at least once to another user. Those users who actively participated in the system have a power law exponent of 2.21 for the number of voted *issues* at a median of 42 – delegates indeed have a increased activity also when compared to active, delegating users.

In order to get an insight in the meaning of delegations, the match of voting decisions between delegates and their delegating voters before the delegation is calculated. The percentage of votes where both users gave

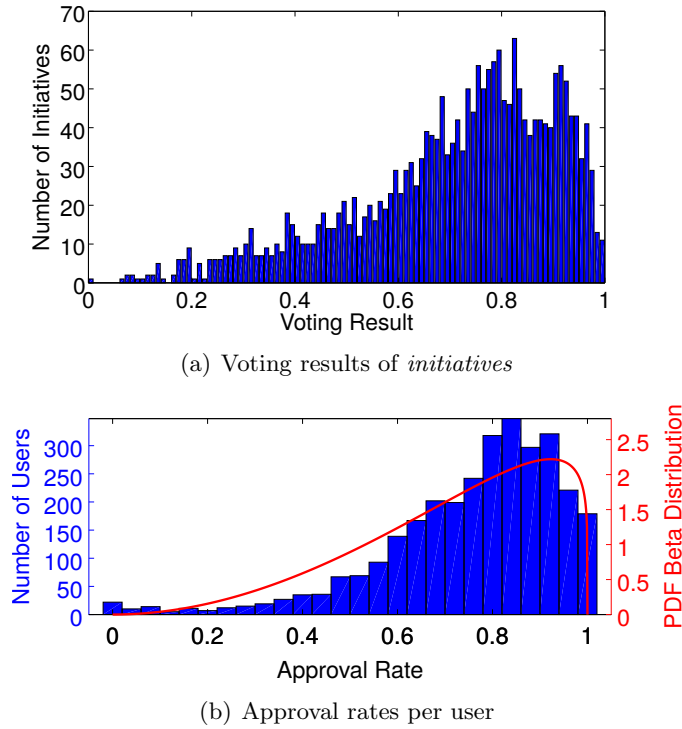


Figure 2.3: Average approval rates per *initiative* and per user.

An approval rate of 1 means maximum approval (all votes have been positive), an approval rate of 0 means minimum approval (all votes have been negative). (a) *Initiatives*. The distribution shows a strong voting bias with a first peak at an approval rate of around 0.75 and a smaller second peak at 0.90. (b) *Users*. Histogram of approval rates of users who voted for more than 10 *initiatives*. It is plausible to approximate the per-user approval rate with a beta distribution. Voters show a strong tendency towards approving initiatives with an expected approval rate of 0.71.

identical ratings (positive/negative) to the same *initiative* is 0.61 whilst any two random voters have an average match of 0.51. As this difference is quite small, delegates do not seem to receive delegations mainly because of shared political views. Instead, they often decide differently than their voters in past votes. This could indicate that delegates in the system then are not expected to represent the opinion of their voters and that they act independently, giving them a high freedom of action.

Another factor in the power of users are voting results. If votes are narrowly decided, even a small weight gives voters the power to decide votes alone. A histogram for the frequency of voting results is shown in Fig. 2.3(a). The distribution is skewed towards positive results with its peak at about 0.8. As the required majority for passing an *initiative* in most cases is 2/3 of

the votes, this means that most of the votes are approved. The distribution of support shows a striking similarity to the distribution of ratings in other online communities as described by Kostakos [Kos09].

2.4.2 User approval rates

In Fig. 2.3(b), the user approval rates, i.e. the percentage of positive votes for each voter, are shown. Users who voted for less than 10 *issues* are excluded to ensure significance. The distribution exhibits a strong bias towards the approval of proposals and reaches the highest numbers at about 0.8 and 0.9. This distribution closely resembles the overall approval of users for *initiatives*. Surprisingly, there is a larger number of “100%-users” (in total 160) who voted yes in **all** of the votes. These users are found to receive a lower number of incoming delegations (1.05 vs. 1.48 on average). One explanation for this behaviour could be that some users only vote for *initiatives* they support and hope that other *initiatives* won’t reach the quorum without their votes. The distribution of user approval can be approximated by a beta distribution which will prove to be useful later for developing novel power indices. Fig. 2.3(b) shows a fitted beta distribution as a dashed line. 100%-users were removed from the data before learning the parameters to obtain a better fit [Min00].

It seems very natural for a democratic voting system without coalitions or party discipline to have a biased distribution of approval rates. As these systems typically include mechanisms to filter out proposals before they reach the voting phase (to prevent an unworkable flood of voting) such as requiring minimum support, the quality of the voted proposals already is relatively high. Due to selection processes, most democratic online systems are likely to exhibit a biased distribution of approval rates.

As the approval distribution is close to the $2/3$ quorum (which is typically required in votes), super voters are expected to have a bigger influence in the voting outcomes.

An interesting observation can be made in the context of the temporal dynamics of approval rates: The average approval rates for the k th vote of all users is shown in Fig. 2.4, illustrating the probability for seeing a positive vote in the first, second etc. vote of a user. Clearly, more experienced users get more critical towards proposals. The learning curve is observed for all users, independent of their activity as measured by the number of voted *issues* – this e.g. can be seen in the approval rates for users of different activity levels depicted in Fig. 2.5(a), which decrease much slower than the learning curve. The negative impact of the number of votes on the approval rate eventually would lead to a stagnation of the system, as the typical quorum of $2/3$ would be reached by hardly any *initiative*.

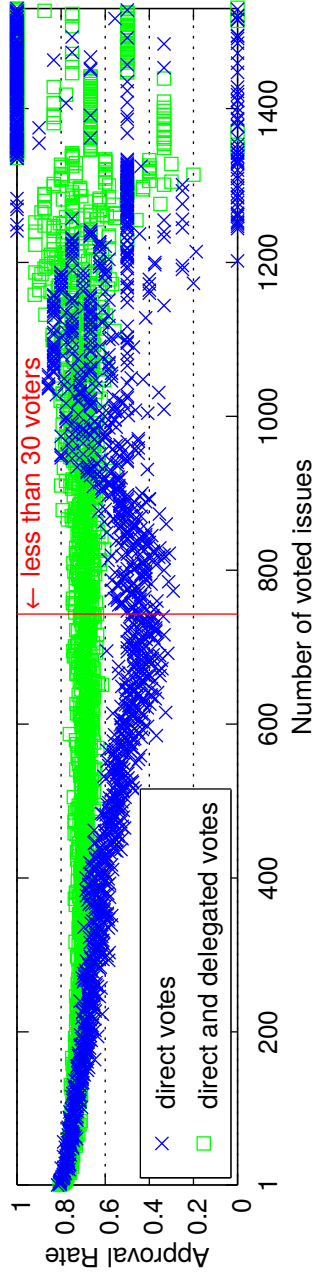
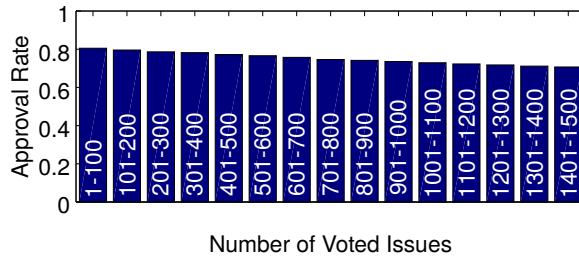
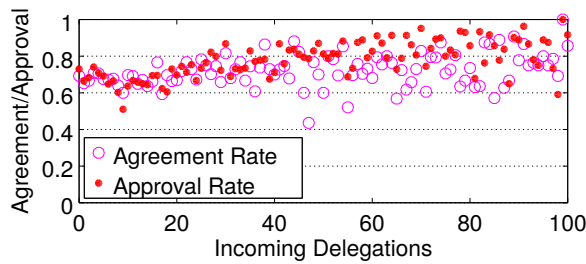


Figure 2.4: Approval rate (Percentage of positive votes) of all voters for the k th voted issue as a function of k . Looking at direct votes only, one can see a decrease of approval rates from 0.8 to below 0.5 with higher voting experience, i.e. voters become more critical. When including votes made by delegates on behalf of voters, only a slight decrease of approval rates is observed and the approval rate quickly stabilises at about 0.7. Delegates therefore have a *stabilising effect on approval rates*. The number of observed votes quickly becomes smaller as the number of voted issues follows a power law. For a low number of observations the estimated approval rate become less precise. The red line marks the point where less than 30 direct voters were observed. The increased approval rate of direct votes around the 750th vote therefore indicates the existence of a very small group of active voters with a high approval rate.



(a) Approval rate vs. activity



(b) Impact of delegations on approval/agreement rates

Figure 2.5: Voting behaviour. (a) Active users as measured by the count of voted issues tend to approve *initiatives* less often. The effect is less pronounced than in Fig. 2.4. (b) Approval rate of votes for given weights. Surprisingly, *super-voters* tend to approve more initiatives (approval rate), and tend to agree more often with the majority compared to normal users (agreement rate). Delegations for authors of *initiatives* were ignored to rule out effects of implicit approval.

2.4.3 Impact of delegations

Surprisingly, such a stagnation cannot be observed in the platform, even in periods when few new users join the system. The effective votes of a user – i.e. all votes including delegated votes made on behalf of a user – are shown in Fig. 2.4. The negative development of approval rates is compensated by delegated votes.

Do these findings imply that *super-voters* are more likely to agree with *initiatives*? And do *super-voters* use their power to turn voting results when voting in favour of initiatives, or do they agree with and vote according to the majority of voters? Fig. 2.5(b) shows the average approval and agreement rate of voters for growing numbers of incoming delegations. The agreement rate is given by the percentage of votes which agree with the majority of voters, excluding delegations. One can observe a positive effect of incoming delegations both on the approval rate and the agreement rate.

In contrast to the intuition that users tend to delegate their votes to users who often vote in favour of proposals, no significant differences were

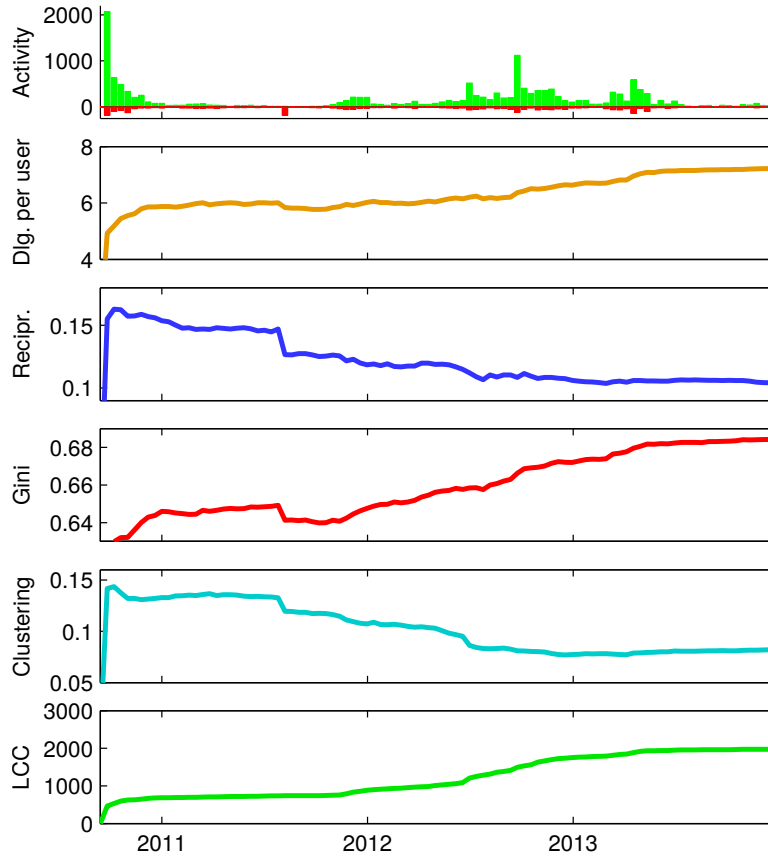


Figure 2.6: Changes in the temporal delegation network over the observation period [KKH⁺15]. The statistics show (from top to bottom): Added and removed delegations, changes in the per-user delegation count, inequality of incoming delegations measured by the Gini coefficient, reciprocity (the proportion of *mutual delegations*), the global clustering coefficient and the size of the largest connected component (LCC). Note that *mutual delegations* are only permitted for distinct areas or *issues*. The observed changes indicate a consolidation of the network, i.e. the emergence of *super-voters* and a stronger concentration of power over time.

found in the approval rates of users with many delegations in their voting history and normal users. However, as soon as users get many incoming delegations, positive votes become more likely. One could hypothesise that voters with many incoming delegations feel social pressure to vote positively and avoid giving a negative vote which would inevitably lead to the failure of a proposal, given the high voting weight. This social control would limit the exercised power of the *super-voters* and stabilise the voting system, effectively preventing political stagnation.

2.4.4 Temporal analysis of the delegation network

Since LiquidFeedback is a novel system, its use is still in an emerging stage, and therefore one would expect its usage patterns to vary over time. Specifically, by analysing the temporal evolution of network-based statistics shown in Fig. 2.6, the following changes are observed [KKH⁺15]:

Changes in the distribution of delegations. While the distribution of received delegations is found to be power law-like, the inequality of this distribution is not constant, as shown by several statistics in Fig. 2.6. In particular, the Gini coefficient of the delegation network's indegree distribution [KP12] is growing, i.e., the inequality of the number of received delegations increases over time. This is consistent with a consolidation of the network, i.e., the emergence of *super-voters* and a stronger concentration of power.

Changes in reciprocity. The reciprocity of the delegation behaviour is measured as the ratio of delegation edges for which a reciprocal delegation edge exists, to the total number of reciprocity edges, and observe that this value decrease over time. This would indicate that the community is going away from a set of small groups of voters that delegate to each other, to a community in which most delegation edges go to *super-voters* who do not delegate back. One must note however that reciprocal delegations are only possible for delegations in different areas, as the set of delegations in a single area must not form cycles.

Changes in clustering. The clustering coefficient gives the probability that two neighbours of a voter are themselves connected, within taking into account edge directions [WS98]. This clustering coefficient decreases over the lifetime of the network while the largest connected component (LCC) is growing, indicating again that the delegation network is slowly becoming less like a set of local groups, and more like a bipartite network of *super-voters* connected to normal voters.

2.5 Power in Online Democracies

The delegation system behind liquid democracy systems creates a complex delegation network. In the previous section it was shown that the network evolves to a state where it assigns high voting weights to a smaller set of voters. This could lead to situations where single voters become very powerful, so that e.g. one voter could decide all votes alone. To test if such *dictators* exist, the distribution of power in the LiquidFeedback system is assessed using power indices.

Power indices are numerical indicators designed to measure the ability of voters to influence voting outcomes. Imagine a vote with n voters with voting weights w_1, w_2, \dots, w_n , i.e. voter 1 has w_1 votes and so on. A typical example is a parliament with several parties, where all the members of a

party are forced to vote the same way and therefore parties act like a single individual voter. In this case, if party 1 won 50 seats in the parliament, it has a voting weight of $w_1 = 50$.

Now one could ask: “how powerful is party 1?”. The answer to that question is not as trivial, as it might seem. First of all, the term *power* has to be defined. Typically, voting power is defined as the ability to change the outcome of a vote. The relation between voting weight and voting power then becomes rather complex. Imagine a setting where there are three voters with $w_1 = 5$, $w_2 = 4$ and $w_3 = 2$. If the voters need 50% to pass a vote, all three voters have the same voting power, as they all need at least to agree with one other voter to reach the majority.

Power indices predict the power of an individual voter given the voting weights of all the participants in a vote. In delegative democracies, power indices can be used to describe the distribution of power among voters in the system and to assess the power of super-voters.

2.5.1 Power indices

The most widely used power indices are the Shapley index [Sha54] and the Banzhaf index [Ban65].

The Shaplex power index

The Shapley index is based on so called *pivotal* voters: For every vote where the majority is in favour of the proposal voted upon, and where the exact order of votes is known, there is a single *pivotal* voter who turns the vote from *rejected* to *approved* with her vote. For a given voter, it is possible to calculate the share of possible outcomes of the vote where she is pivotal.

Simply speaking, if the number of orderings where voter i is pivotal is given by r_i , the formula for the Shapley index is

$$\phi_i = \frac{r_i}{n!} \quad (2.1)$$

where there are n voters and $n!$ gives the number of possible orderings of the votes [Str77].

The Banzhaf power index

The Banzhaf index in contrast ignores the ordering of voters and utilises *winning coalitions* and *swings* instead. *Winning coalitions* are sets of voters W which get the majority of votes. The winning coalitions which would not gain the majority of votes without voter i are stored in the set $S_i = \{s_{i1}, \dots, s_{i|S_i|}\}$, i.e. voter i alone could decide the vote in each of these winning coalitions. A *swing* $s_{ik} \in S_i$ (where k is an index) is a winning coalition from this set.

The size of the set S_i of these winning coalitions which involve voter i divided by the total number of possible voting decisions defines the Banzhaf index [Ban65, Str77]:

$$\beta_i = \frac{|S_i|}{2^{n-1}} \quad (2.2)$$

One can also express the Shapley index using the notion of swings [Str77]:

$$\phi_i = \sum_{s_{ik} \in S_i} \frac{(|s_{ik}| - 1)! \cdot (n - |s_{ik}|)!}{n!} \quad (2.3)$$

where the fact is used that there are $(|s_{ik}| - 1)!$ combinations of the *swing* where voter i voted last and $(n - |s_{ik}|)!$ is the number of possible combinations of additional voters (which are not required to gain the majority), yielding the number of *winning coalitions* so that voter i is pivotal.

Alternative power indices

Both the Banzhaf and the Shapley index are based on game theory and are mostly popular due to their simplicity. Other power indices try to capture the parliamentary reality, e.g. by limiting the index to majorities by minimal coalitions [DJP78, PDJ80]. However, these indices seem not appropriate for delegative democracies, as voting weights change frequently and no fixed coalitions are formed, and thus minimal coalitions are as likely as any other coalition.

Gelman et al. criticised the simplicity of the game-theoretic approaches by suggesting an Ising model for modelling dependencies between voters, e.g. common administrative regions [GKT02]. However, the study lacks the appropriate data for fitting the model as it relies on aggregated voting results and therefore cannot consider decisions at the individual level.

In the following, it will be shown how to utilise user-based voting behaviour to derive adjusted power indices and conduct the first objective evaluation of power indices on large real-world voting data with constantly changing voting weights.

2.5.2 Probabilistic interpretation of power indices

The Banzhaf and the Shapley power index have an interpretation as probabilistic models, as first noted by Straffin [Str77]. If one wants to know the probability that voter i is able to influence the outcome of a vote, and every outcome would be equally likely, then one could count the number of winning coalitions where this is the case and divide by the number of possible voting outcomes. As the number of votes that voter i is able to influence is just the number of swings where voter i voted *yes* plus the number of

swings, where the voter voted *no*, one gets [Str77]:

$$\frac{2 \cdot |S_i|}{2^n} = \frac{|S_i|}{2^{n-1}} = \beta_i \quad (2.4)$$

which shows that the Banzhaf index can be interpreted as a probability.

However, this equation only holds true under the assumption of *independence*: All voters vote *independently* of each other and their probability of voting yes is $p_i = 0.5$ for every voter i . This makes all coalitions and swings equally likely to occur.

As shown in Equation 2.3, the Shapley index can be expressed as a sum over the sets of winning coalitions where voter i is a swing voter. The probability of the swings of voter i is given by their marginal probability under a uniformly distributed p [Str77] as given in Equation 1.4:

$$\begin{aligned} p(S_i^+ \cup S_i^-) &= \sum_{s_{ik} \in S_i^+} \frac{(|s_{ik}|)! \cdot (n - |s_{ik}|)!}{(n+1)!} + \frac{(|s_{ik}| - 1)! \cdot (n - |s_{ik}| + 1)!}{(n+1)!} \\ &= \sum_{s_{ik} \in S_i^+} \frac{(|s_{ik}|)! \cdot (n - |s_{ik}|)! + (|s_{ik}| - 1)! \cdot (n - |s_{ik}| + 1)!}{(n+1)!} \\ &= \sum_{s_{ik} \in S_i^+} \frac{(|s_{ik}| - 1)! \cdot (n - |s_{ik}|)! \cdot (|s_{ik}| + (n - |s_{ik}| + 1))}{(n+1)!} \\ &= \sum_{s_{ik} \in S_i^+} \frac{(|s_{ik}| - 1)! \cdot (n - |s_{ik}|)!}{n!} \end{aligned} \quad (2.5)$$

where S_i^+ denotes the swings where voter i voted positive, S_i^- denotes the swings where voter i voted negative and $n = |S_i^+ \cup S_i^-| = |S_i|$.

2.5.3 Theoretical (uniform) power indices

Both the Banzhaf and the Shapley index can be characterised in probabilistic terms [Str77]. Indeed, assume that the vote of a voter i is drawn randomly with probability p_i for a “yes” and $1 - p_i$ for “no”. The *individual effect* of a voter i is the probability that the voter i makes a difference to the outcome of the entire vote. Of course, the individual effect will depend on the individual probabilities p_i . Typical assumptions behind existing theoretical power indices are:

- Uniformity. Each p_i is chosen from a uniform distribution on $[0, 1]$.
- Independence. Each p_i is chosen independently.
- Homogeneity. All p_i are equal to p , a common agreement rate shared by all voters.

It was shown in [Str77] that the Banzhaf index represents the individual effect of a voter under the assumption of independence and the Shapley index represents the individual effect of a voter under the homogeneity assumption. Both indices rely on the uniformity assumption, and most power indices from literature repeat this assumption [Str77, Str94, GKT02].

2.5.4 Empirical power

Theoretical measures of power are based on simulation. With the large number of observations available from the LiquidFeedback dataset, it is possible to directly measure power in a large, real-world dataset. An important difference to traditional voting data from parliaments is the absence of fixed coalitions, the high number of votes and the relatively stable set of voters. Recall that power in the context of power indices is defined as the ability of a voter to influence voting outcomes.

Potential Power. The ability to decide a vote is calculated with the sum of weights of positive W_m^p and negative W_m^n votes in a voting m , testing if the weight w_{im} of voter i is bigger than the votes needed to reach quorum q_m :

$$\begin{aligned} \gamma_{im}^p = & [w_{im} > \underbrace{q_m \cdot (W_m^p + W_m^n) - W_m^p + w_{im} \cdot v'_{im}}_{\text{votes missing to reach quorum without voter i}}] \\ & \wedge \underbrace{[q_m \cdot (W_m^p + W_m^n) - W_m^p + w_{im} \cdot v'_{im} > 0]}_{\text{quorum not reached without voter i}} \end{aligned} \quad (2.6)$$

where $v'_{im} \in \{0, 1\}$ indicates the decision of voter i in voting m and squared brackets denote Iverson brackets so that $\gamma_{im}^p \in \{0, 1\}$.

Exercised Power. Similarly, one can look at the actual vote of voters and see whether the power actually was used to *reverse the voting result*:

$$\gamma_i^e = \left[\underbrace{\left(\frac{W_m^p - w_{im} \cdot v'_{im}}{W_m^p + W_m^n - w_{im}} > q_m \right)}_{\text{voting result without voter i}} \neq \underbrace{\left(\frac{W_m^p}{W_m^p + W_m^n} > q_m \right)}_{\text{actual voting result}} \right] \quad (2.7)$$

Looking at the voting history, the impact of delegates on voting outcomes can be easily estimated by subtracting delegations from vote counts. Without the delegations, the vast majority of 84.9% of the results remain unchanged – only one in six voting outcomes is not identical to the outcome of a hypothetical direct democratic system.

Does that mean that super-voters are not as powerful as they were thought to be? To answer that question, the potential and the exercised power is shown in Fig. 2.7(a). The ability to decide votes grows approximately

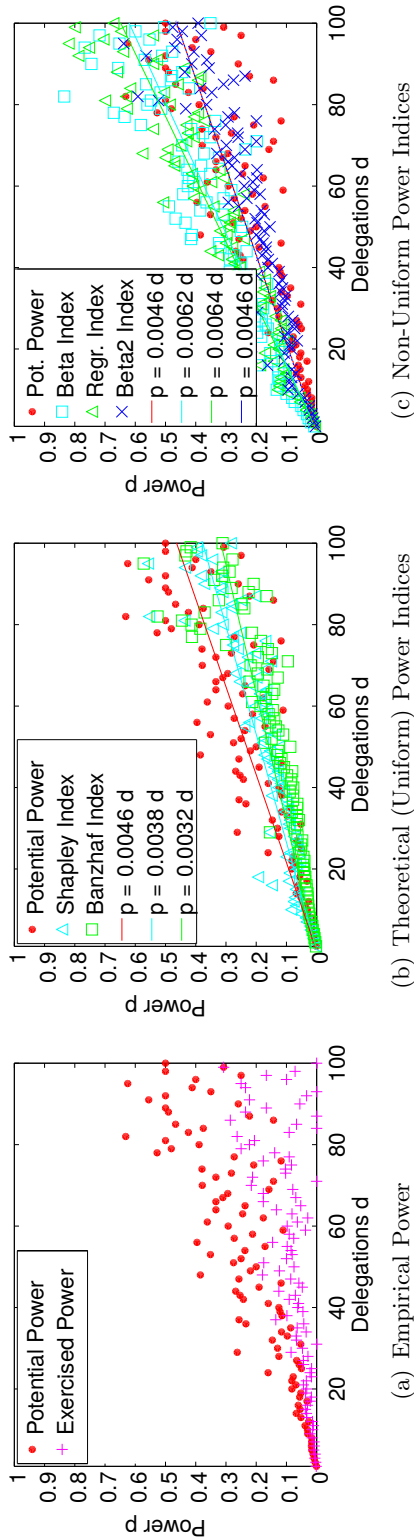


Figure 2.7: Measures of power. (a) Average potential power and exercised power for a given number of delegations. The exercised power grows slower than the practical and theoretical power with a negative correlation of 0.26. This indicates that *super-voters* have a higher tendency to agree with the majority of voters compared to normal voters, or that the decision of *super-voters* influences the voting decision of other voters. (b) Average potential power and average predictions of uniform power indices for given delegation counts. Both uniform indices under -estimate the voting power in the LiquidFeedback data of the German Pirate Party, and the Shapley index yields a relatively better prediction. (c) Averaged potential power and average power index of the beta, Regression and beta2 power index for changing numbers of delegations. The beta and beta2 indices model the observed voting bias in the system, while the regression index predicts the approval rate of a voter based on the number of incoming delegations. Both the beta index and the regression index predict voting results at 0.72, the expected value of the beta distribution with parameters $\alpha = 3$ and $\beta = 1.17$. Therefore, a small number of votes would be sufficient to change the voting result, as the quorum is at $2/3$ for most votes. The beta2 index closely predicts the measured potential power, by modelling the whole range of voting results given by the beta distribution.

linearly with the voting weight and the exercised power measured as the percentage of reversed votes grows slower than the potential power – *super-voters* use their power relatively less often than ordinary delegates. This explains the positive influence of delegations on the majority agreement observed in Fig. 2.5(b). The average ratio between theoretical and empirical power is 0.34 – powerful users reverse the result of a voting in only one of three votes. Potential power and exercised power are weakly negatively correlated with $\rho = -0.26$ ($p < 0.05$).

In theory, power indices are supposed to correspond to the potential power of users. To test this, the Banzhaf and Shapley index are calculated for every vote and the average predicted power is shown in Fig. 2.7(b). On the Pirate Party platform, both theoretical power indices fail to approximate the potential voting power. Instead, the Shapley index and the Banzhaf index understate the potential power of users and predict a growth rate that is lower than the observed growth.

The main focus of this study is on the prediction and recognition of high potential power and the danger of power abuse, and therefore it is not the aim to predict the exercised power. Potential power might not be used at a given time, but there is no reason to assume that this behaviour is stable.

2.5.5 Non-uniform power indices

The limited alignment of existing power indices with observed voting behaviour suggests that some of the fundamental assumptions behind these indices do not hold in real-world voting systems. Existing power indices are based on the uniformity assumption, i.e. that users vote with equal probability in favour or against a proposal. Historically, there was no extensive voting data available to test this assumption. For online platforms such as LiquidFeedback, there is enough data to observe a voting bias [Kos09]. The findings on the distribution of voting results and user approval rates shown in Fig. 2.3 will help to overcome this over-simplifying assumption of uniformity.

In this section, *generalisations of the Banzhaf and Shapley power index* are proposed which allow to model non-uniform distributions of approval rates (as observed in real-world voting data).

The Beta power index

The user approval rate p_j approximately follows a beta distribution. Under a beta distribution, this parameter is sampled from

$$p_j \sim \frac{1}{B(\alpha, \beta)} p_j^{\alpha-1} (1 - p_j)^{\beta-1} = \text{Beta}(p_j | \alpha, \beta). \quad (2.8)$$

For parameter estimation, extreme cases of users with 100% approval are removed and the maximum likelihood estimate for Dirichlet distributions

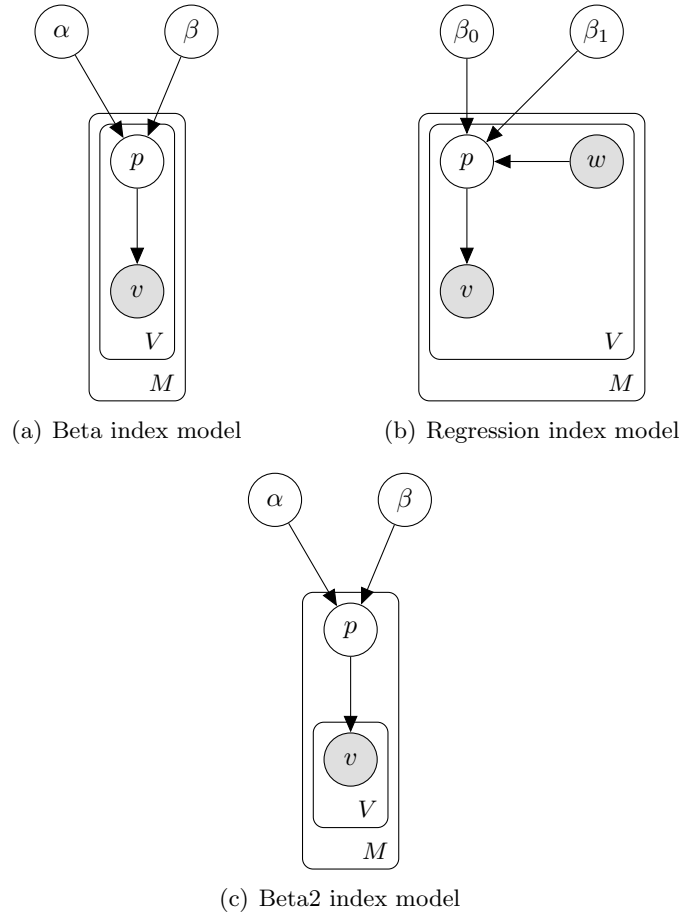


Figure 2.8: Bayesian networks of the beta index, the regression index and the beta2 index. In contrast to the Banzhaf and Shapiro indices, the voting models presented in this thesis assume that the approval rate p of a voter is unknown and drawn from a context-dependent distribution. The *beta index* (a) is based on the assumption that for every of the M votings and for every of the V voters in that voting, there is an approval rate p drawn from a beta distribution with parameters α and β . Each voting decision v is then drawn from a Bernoulli distribution with the approval rate p as parameter. The *regression index* (b) similarly assumes a voting- and voter-specific approval rate, which is this time predicted using the known voting weight w of a voter in a vote, by a logistic regression with parameters β_0 and β_1 . The *beta2 index* (a) finally assumes that for every of the M votings there is only one approval rate p drawn from a beta distribution with parameters α and β . Each of the V voters then draws a voting decision v from this common approval rate.

given by Minka [Min00] is applied to obtain $\alpha = 3.00, \beta = 1.17$. The probability density of the beta distribution is shown in Fig. 2.3(b).

The first novel power index is a generalisation of the Banzhaf index based on the beta distributed p_i . This index is called the *beta power index*. The intuition behind this index is identical to the Banzhaf index: the power of a voter corresponds to the fraction of coalition constellations in which the user is a swing voter. In order to create a non-uniform power index, the permutation of possible coalitions are re-weighted by their probability under beta-distributed p_j for every voter. Every voter $j \in V$ has an assigned probability p_j for approving a proposal and users are independent. For calculating the beta power index β' , all possible winning coalitions in which i is a swing voter are considered and weighted by their probability

$$\beta'_i = \int_0^1 \cdots \int_0^1 \left(\sum_{s_{ik} \in S_i} \prod_{j \in V} p_j^{v_{js_{ik}}} (1 - p_j)^{1 - v_{js_{ik}}} \right) \cdot \prod_{j \in V} \text{Beta}(p_j \mid \alpha, \beta) dp_1 \cdots dp_V \quad (2.9)$$

where S_i denotes the set of possible winning coalitions in which i is a swing voter and $v_{js_{ik}}$ is 1 if voter $j \in V$ of the swing $s_{ik} \in S_i$ voted “yes” and 0 otherwise. The probability of a coalition is given by a multinomial distribution with success probability $\vec{p} = (p_1, \dots, p_V)$, the beta distributed approval rates. By integrating over the approval rates under the beta distributions, the marginal likelihood of each swing is obtained, i.e. the probability of each swing under the biased, beta-distributed approval rate. Beck [Bec75] noted that the probability of a tie is very small under such a model – this finding is trivial and indeed in the whole LiquidFeedback dataset only one *initiative* exhibits a tie. The Bayesian network of the beta index is shown in Fig. 2.8(a).

It is evident that the beta index represents a generalisation of the Banzhaf index – one can choose symmetric beta parameters to retain the original index.

The Regression power index

Another observation besides biased user approval rates is the impact of delegations on the approval rate shown in 2.5(b). To model this influence, a logistic regression can be trained to predict the approval rates for given voting weights which yields an alternative power index. The regression function is given by

$$p_j = \frac{1}{1 + e^{-(\beta_0 + \beta_1 w_j)}} \quad (2.10)$$

where w_j is the voting weight of voter j . Users with 100% approval rate again are removed from the data, which then yields regression parameters $\beta_0 =$

0.7933 and $\beta_1 = 0.0036$. The regression predicts an approval probability p_i of 0.69 at a weight of 1 and 0.76 at a weight of 100.

For obtaining the regression power index ρ_i , the possible coalitions are weighted by the product of all approval rates predicted by logistic regression based on the set of winning coalitions S_i where voter i is a swing voter:

$$\rho_i = \sum_{s_{ik} \in S_i} \prod_{j \in V} p_j^{v_{js_{ik}}} (1 - p_j)^{1 - v_{js_{ik}}} \quad (2.11)$$

where each p_j is calculated using the logistic regression in Eq. 2.10. The graphical model of the regression index is shown in Fig. 2.8(b).

The Beta2 power index

The assumption of independence made by the Banzhaf index implies that voters have inhomogeneous opinions and that there is frequent disagreement in votings, i.e. there exist opposing groups within the party. In contrast, the Shapley index assumes that all voters share a similar ‘‘attitude’’ on a particular *initiative*, i.e. they will approve it with the same probability $p_j = p, \forall j \in V$. However, p_i in the Shapley index is sampled from a uniform distribution.

The Shapley index can be generalised by sampling p from the same beta distribution employed for the beta index: $p \sim \text{Beta}(\alpha, \beta)$ with $\alpha = 3.00, \beta = 1.17$. This index assumes that voters share a homogeneous opinion on *initiatives*, and that there is a positive voting bias to accept proposals. For the overall calculation of the beta2 power index β_i'' , one sums over all possible coalitions where voter i is a swing voter, weighted by their probability:

$$\beta_i'' = \int_0^1 \left[\sum_{s_{ik} \in S_i} \prod_{j \in V} p^{v_{js_{ik}}} (1 - p)^{1 - v_{js_{ik}}} \right] \text{Beta}(p | \alpha, \beta) dp \quad (2.12)$$

where S_i again denotes the set of possible winning coalitions in which i is a swing voter, and $v_{js_{ik}} \in \{0, 1\}$ is the approval of voter $j \in V$ in swing $s_{ik} \in S_i$. As in the beta index, the marginal likelihood of each swing under the beta distributed approval rate is obtained by integrating over the shared approval rate. The Bayesian network of the beta index is shown in Fig. 2.8(c).

Gelman et al. [GKT02] claim that models based on binomial distributions with $p \neq 0.5$ would not be useful because of the small standard deviation of the expected voting result. In the evaluation, it is shown that this interpretation is wrong. A larger variance of voting results is obtained by defining a generative process for votes with approval probability p , where the approval rate is sampled from a beta distribution with a possibly large variance.

2.5.6 Evaluation

For a quantitative comparison of the power indices, the prediction quality is evaluated both at the global and at the local level. At the global level, power indices predict the average power of *super-voters* as in Fig. 2.7. The closeness of the prediction can be measured as the sum of squared errors of the average predicted theoretical power and the average measured potential power for each voting weight $w_i \in [1, 100]$. The squared errors of the power indices are shown in Table 2.1. The biggest deviations are found for the regression, beta and Banzhaf index, indicating that the independence assumption is violated in the voting system. For the Shapley index, a significantly lower value is achieved and the beta2 index provides the closest approximation.

At the local level, the extensive voting history is employed to compare the observed potential power of voters – the ability to decide a vote – to the predicted power index of every user. Following the probabilistic interpretation of power indices [Str77], a power index corresponds with the predicted probability of a voter having potential power. This probability is computed for every voter in each vote. Now, given the measured potential power of a voter, it is straightforward to calculate the log-likelihood of the observed power in the voting history. Formally, the likelihood of a vote is:

$$\log \mathcal{L}(\gamma_m^p | \theta) = \sum_{i \in V_m} \log (p(\gamma_{mi}^p | \theta)) \quad (2.13)$$

where $p(\gamma_{mi}^p | \theta)$ denotes the probability of the observed power under the the tested model with parameters θ , V_m is the set of voters participating in the vote over *initiative m* and γ_{mi}^p indicates the potential power of voter i in voting m .

The likelihood can then be used to calculate the perplexity, a common measure for the predictive quality of a probabilistic model. The per-vote

Table 2.1: Performance of power indices. Perplexity and squared prediction error for the uniform power indices by Banzhaf and Shapley and the non-uniform power indices presented in this thesis, evaluated on the complete voting history of the LiquidFeedback system. Lower perplexity values indicate a better model fit. The beta2 index proposed earlier outperforms existing and other competing power indices.

Model	Squared Error	Perplexity
Shapley [Sha54]	0.903	78.6
Banzhaf [Ban65]	1.320	297.9
Beta power index	2.220	227.8
Regression power index	2.266	232.0
Beta2 power index	0.627	76.6

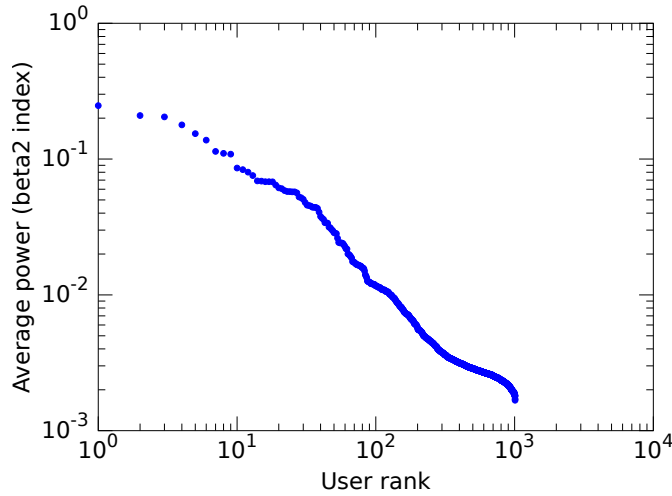


Figure 2.9: Power distribution in the LiquidFeedback system of the German Pirate Party. Users who participated in more than 100 votes are ranked by their average beta2 power index, which is shown on the y-axis. Note that both axes are logarithmic. The distribution is flattened in the beginning compared to a power-law distribution. While the distribution of power is uneven, there are no *dictators* i.e. no voter could decide all votes alone.

perplexity is then given by

$$\text{perplexity} = 2^{-\frac{1}{M} \sum_{i=1}^M \log \mathcal{L}(\gamma_m^p | \theta)} \quad (2.14)$$

where M is the number of voted *initiatives*.

Following the perplexity scores, the beta2 index outperforms all other indices. The Shapley index yields the second best result. The beta index is slightly better than the regression index and the Banzhaf index performs worst.

2.5.7 Distribution of Power

The best performing index, the beta2 index, can be employed to analyse the distribution of the average *power per vote* of voters in the system. Figure 2.9 shows the distribution of power for voters who participated in more than 100 votes, i.e. voters who potentially could have a significant impact on the system. The maximum power index is 0.248, indicating that the most powerful user could decide every fourth vote she or he participated. Looking at the distribution of power, one can see that the distribution does not follow a power-law but is flattened in the beginning. While there exists a group of super-voters with a relatively high power, there are no *dictators* which alone could decide the majority of the votes.

2.6 Discussion

The observed performance of the power indices allows the evaluation of the assumptions behind these models. First, indices based on the independence assumption of voters perform significantly worse than the indices based on the homogeneity assumption, implying that voters share a common attitude towards initiatives, which can be modelled by a common, vote-specific approval rate for all voters. It was found that the integration of the observed positive influence of delegations on the approval rate by the regression index leads to worse performance as measured by squared error and perplexity (see Table 2.1). The effect seems to be more complex and has to be examined in future work. Modelling voters homogeneously – e.g. sampling approval rates independent of voting weights – yields significantly better results.

Including system-specific voting bias in power indices leads to an overall better predictive quality of power indices, measured by lower perplexity. However, only for homogeneous indices a better global prediction was obtained.

The proposed beta2 index, a biased generalisation of the Shapley index, gives a precise prediction of the overall power distribution in a voting system with delegations. It is possible to accurately predict the ability of delegates to decide votes by sampling sets of voters and calculating the beta2 index. The parameters of the beta distribution can be learned from voting history or taken over from similar voting platforms. With these predictions, qualified statements about the distribution of power in voting systems can be made and discussions objectified.

Both the analysis of voting behaviour and the empirical measurement of potential and exercised power exhibit a responsible exercise of power by *super-voters*. This might indicate a responsible selection of delegates, the social control in an enforced public voting and the risk of the immediate loss of voting power by recall of delegations.

2.7 Summary

Online platforms for delegative democracy are likely to gain relevance for political movements and parties in the future. Understanding the voting behaviour and emergence of power in such movements represents an important but open scientific and pressing practical challenge.

The study analysed **(i) how people vote** in online delegative democracy platforms such as LiquidFeedback, and **(ii) how they delegate votes** to *super-voters*. The main objective of this study was to **(iii) better understand the power voters have over voting processes**. In particular, **(iv) the theoretical, potential as well as the exercised power** of super voters in online delegative democracy platforms were examined.

Towards that end, the Banzhaf and Shapley power index were employed, but exhibited conflicts between the assumption of uniformity of voting behaviour made by both indices and the observed voting bias. Thus **(v) a new class of power indices was introduced and evaluated that (a) generalises previous work based on beta distributed voter agreement and (b) achieves significantly better predictions of potential voting power in the evaluation.** To the best of knowledge of the author, the evaluation based on a large voting history represents an innovative objective evaluation of power indices.

By introducing system-specific indices for voting power, this study of a liquid democracy platform is a very basic example of how context-specific priors for probabilistic models can improve the prediction quality. In the following chapter, more complex models – namely probabilistic topic models – will be extended for context information by using the same principal methodology: Instead of relying on e.g. uniformity assumptions, context-specific prior distributions are learned and help to find topics of higher quality.

Chapter 3

Single-Context Topic Models

In this chapter, extensions of mixed-membership models for context variables are presented which – for the first time – allow for the efficient modelling of *cyclic and spherical context variables* such as the daily 24h cycle or geographical coordinates.

Mixed-membership models are based on the assumption that grouped observations are generated by mixtures of multiple latent distributions. The topic models presented here are mixed-membership models, i.e. documents are assumed to be created by a mixture of topics. Though the examples and the evaluation are based on words grouped by documents, the very same models can be applied to arbitrary grouped observations, e.g. attributes of users of a social media platform. Furthermore, it is straightforward to replace the multinomial distribution of topics over terms with any arbitrary probability distribution e.g. to model normal-distributed, grouped measurements.

Context is a broad term and it depends on the dataset which information is perceived as content and which as context. Therefore, context information typically found in social media is categorised in Section 3.1. Important properties, which later are used for the modelling of context variables, are discussed.

Subsequently, context-specific topic models are investigated as a representative class for complex probabilistic models in Section 3.2. Existing approaches for context integration are reviewed, similarities between the approaches are pointed out and weaknesses of the approaches are analysed.

Finally, a novel approach for integrating context information into complex probabilistic models is presented in Section 3.7.4. Based on *multi-Dirichlet processes*, a generalisation of the Dirichlet process employed for coding context information, a novel approach for the modelling of context-dependent topics is developed. The model overcomes weaknesses of existing methods and shows improved performance in modelling real-world datasets from related work.

3.1 A Classification of Context Variables

Most content from social media such as articles, messages or images is associated with metadata. The time of creation, authorship information (potentially linking to a user profile), the location a message was written – all this information can help to understand the *context* in which content was created. Additionally, using information about location and time, more context information such as weather data can be merged in [KSS11]. In this thesis, “*context*” refers to such metadata and derived context information.

In order to exploit this metadata in probabilistic models, the nature of the available context variables has to be understood. The context variables used in this thesis can be classified into four categories: *discrete*, *linear*, *cyclic* and *spherical* context variables.

3.1.1 Discrete context variables

Instances of discrete context variables include information about the source of a document (e.g. the ID of a mailing list where an email was posted), authorship information such as gender or user roles, discrete location information such as the country or state in which a document was written or language information. Discrete context variables can always be coded as a set of binary variables. It often is not clear how discrete context variables relate to each other, e.g. a-priori, it is unclear whether female and male users in a system are similar or dissimilar. Figure 3.1(a) visualises this property. Probabilistic models for purely discrete context variables therefore typically assume independence between the variables. The independence assumption can be replaced with structured information from networks to induce independence. However, networked context information is not the focus of this thesis and there exists a vast amount of literature on how to make use of network information in probabilistic (topic) models [MCZZ08, CB09].

3.1.2 Linear and continuous context variables

A more complex class of context variables are linear metadata such as timestamps or two-dimensional projections of geographical coordinates. Linear context variables relate to each other and can be ordered. This property is visualised in Figure 3.1(b). If a linear variable is continuous, a typical way to model dependencies between documents in the context space is to introduce continuous probability distributions over the context variable – typically on the level of the latent topics.

3.1.3 Cyclic and spherical context variables

A class of context variables which so far received little attention in topic modelling literature is the class of variables which lie on the cycle, sphere or

n-sphere. Geographical locations are common spherical context variables. Cyclic context variables can be found in temporal cycles such as the annual, weekly and daily cycle, which can be directly extracted from timestamps which are available for nearly all social media data. The difference to non-spherical context variables clearly is the absence of an ordering between data points. Every context variable has neighbour relations to other context variables on the n-dimensional sphere, e.g. on a cycle there typically exists a *next* and a *previous* value, as depicted in Figure 3.1(c).

Even though cyclic and spherical context information can be extracted from most datasets, topic models so far only treat two-dimensional projections of cyclic or spherical data. Clearly, it would be better to employ distributions which yield a more realistic model of the data, such as the spherical Mises–Fisher distribution introduced in Section 1.4.7. Unfortunately, for large datasets, parameter inference for models involving Mises–Fisher distributions is expensive as there exists no closed-form solution for inferring the parameters of the distribution. This makes distributions on the n-sphere unattractive for context modelling. The novel models introduced in this thesis are the first topic models to explicitly model cyclic and spherical context variables, overcoming performance issues by using a two-step process and by exploiting neighbour relations between (grouped) cyclic or spherical context variables.

3.2 Single-Context Topic Models

Based on the different structural properties of context variables, different approaches for context-aware probabilistic models were developed in the past.

The topic models presented in Section 1.5 model documents as being generated by a mixture of topics. Every document is associated with a multinomial distribution over the available topics, and topics are distributions over the vocabulary. Words in documents are drawn from the topic-word multinomials of the topics.

In an abstract sense, topics exploit the co-occurrence of words in documents to explain the observed document corpus by latent factors (which are called topics). Knowing contextual information from the metadata of a document – such as the source, authorship, time or the geographic location – can help to improve the detection of topics. Topics then not only exploit the co-occurrence of words in documents, but also the co-occurrence of words in the context space, e.g. within all the texts of an author, within a geographical region or within a time frame.

In the following, existing topic models for different types of context variables are reviewed.

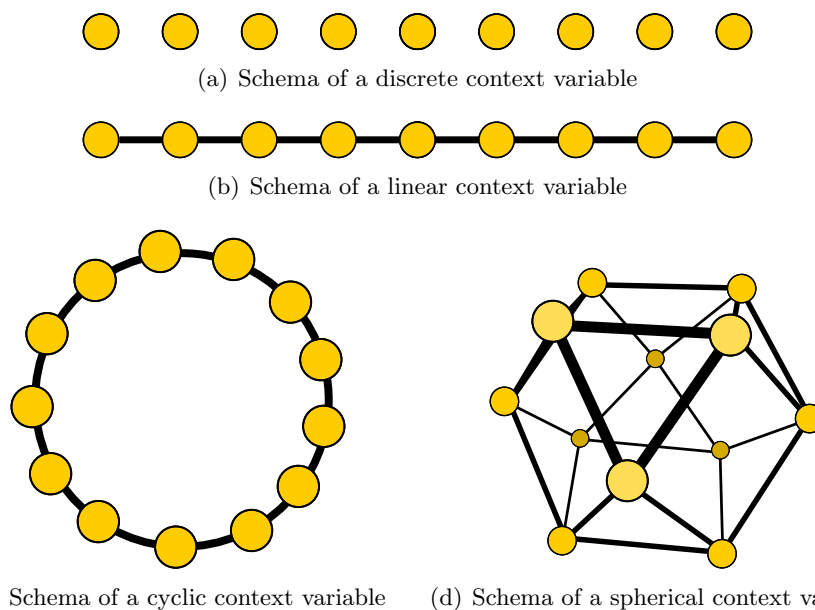


Figure 3.1: Visualisation of the structural properties of context variable types. (a) Discrete context variables typically do not have a pre-defined ordering, while 3.1(b) continuous context variables are comparable and can be ordered. For 3.1(c) cyclic and 3.1(d) spherical context variables, there exists a more complex network of *neighbour relationships*, i.e. on a cycle, there typically exists a predecessor and successor for every data point.

3.2.1 Topic models for discrete context variables

The most simple way of creating a context-specific topic model is to group documents by their source. Given a corpus of documents from one source, it is straightforward to learn the prevalence of topics in the source – for instance for news paper, one could find that there are far more documents about politics than about sports. With that knowledge, ambiguities can be solved easier. Basic LDA does not support the inclusion of such information and assumes a-priori that all topics have the same probability. Wallach et al. [WMM09] put an asymmetric Dirichlet prior over the document-topic distribution which is learned during inference and showed that this yields an improved perplexity. This approach bears similarities to the generalisation of power indices presented in the previous section, where a system-specific prior was introduced which improved the predictive performance of the underlying model. The Bayesian network representation of the model by Wallach et al. is shown in Fig. 3.2(b).

A rather simple alternative for discrete contexts is to directly merge documents from the same context, as done in the author-topic model by Rosen-Zvi et al. [RZGSS04]. In this case, topics are purely learned based

on the co-occurrence of words in the context space (i.e. within the collected works of authors) and the information about co-occurrences of words in documents is lost.

Multiple sources then can be modelled by learning a specific prior for each source. Teh et al. [TJBB06] went one step further, grouping multiple sources by sampling topic distributions from a common Stick-Breaking Process (SBP) in a three-level HDP. That way, topic preferences can be shared across similar sources (e.g. news papers with a similar focus) and the scaling parameter α_0 on the second level DP governs the similarity between different sources. Figure 3.2(c) shows the resulting model. G_0 is a global measure on the topic space, G^c holds the topic probabilities for every context and there are C different values, the context variable can take (e.g. different sources). G^d is the document-specific topic distribution which has the context-specific distribution as a prior.

An interesting variation of this motive is the Citation Influence Model by Dietz et al. [Die06] where citation information is included in a topic model for scientific papers. In the model, words in a document either are drawn from a document-specific topic distribution or from the topic distribution of a cited document. That way, multiple contexts (cited papers are the context in which a paper was written) can be taken into account and the strength of the influence of the different contexts can be estimated (in situations where the topics of the cited documents differ significantly).

Mei et al. [MLSZ06] extend PLSA for spatio-temporal context by mixing the topic distribution of documents with location- and time-specific topic distributions. Locations and timestamps are modelled as discrete sets. In practice, the division of data into location and time intervals results in sparse data.

The models presented so far can be called “upstream” models [MM08], because the context information influence the Dirichlet prior over topics, which is learned from the topic assignment of words.

A different approach of coping with discrete context variables are models where topics not only are each associated with a distribution over the vocabulary, but additionally have a multinomial distribution over the context variable. Mimno et al. call such models “downstream” topic models [MM08].

One instance of such a model is the model for discrete geographical information (e.g. over administrative areas) by Wang et al. [WWXM07]. The model extends LDA in a way that topics are multinomial distributions both over words and a discrete set of locations. The authors are aware of the fact that some related locations, such as locations within a country, are expected to share a similar topic distribution. They therefore suggest to introduce a hierarchy between locations such as countries or cities to share topic information by merging those locations.

A typical problem of such “downstream” models is the strength of the influence of context on topics. Every context information is repeated N_m

times (the length of the document), and this yields the implicit assumption that context information is exactly equally important for finding topics as the words of the documents.

3.2.2 Topic models for linear context variables

A simple trick for modelling continuous linear context variables is to discretise the context space by clustering and to subsequently apply a topic model for discrete context variables.

A temporal clustering of documents can be used to share information between documents in a three level HDP, as for discrete variables. Additionally, cluster-specific topic distributions can be modelled as dependent from the previous cluster in time, as proposed in [RDC08] (for general probabilistic models) and in [ZSZL10] (for topic models). Additionally to the topic preference, the topic-word distributions of topics can evolve over time as modelled in the infinite Dynamic Topic Model (iDTM) [AX12] where topic-word distributions depend on the topics of the preceding temporal cluster.

Alternatively, the “downstream” approach can be applied, extending topics for distributions over the continuous context space. In the popular Topics Over Time (TOT) model by Wang et al. [WM06], topics are extended for beta distributions over the time frame of the document corpus (i.e. all timestamps are normalised to the interval between 0 and 1). A non-parametric extension of TOT based on Dirichlet processes, the npTOT model was proposed in [DHWX13].

A very general approach for modelling the influence of context variables on topics is the Dirichlet-multinomial regression (DMR) topic model by Mimno et al. [MM08]. The DMR topic model places a Dirichlet-multinomial regression over the topic distributions of documents, which takes discrete and continuous context variables as an input. This approach is especially interesting as the model structure remains simple and inference is relatively cheap. However, it is limited to linear variables. For discrete context variables, the approach is similar to a model which learns a topic distribution for every context variable and then mixes the topic distributions with learned weights. Therefore, the DMR topic model can be considered a generalisation of the upstream modelling approach.

3.2.3 Topic models for cyclic and spherical context variables

Topic models which *explicitly* model the properties of cyclic context were not proposed so far. The most commonly modelled spherical context in topic modelling is geographical context. All existing topic models for cyclic or spherical context project the context to a simpler representation before modelling.

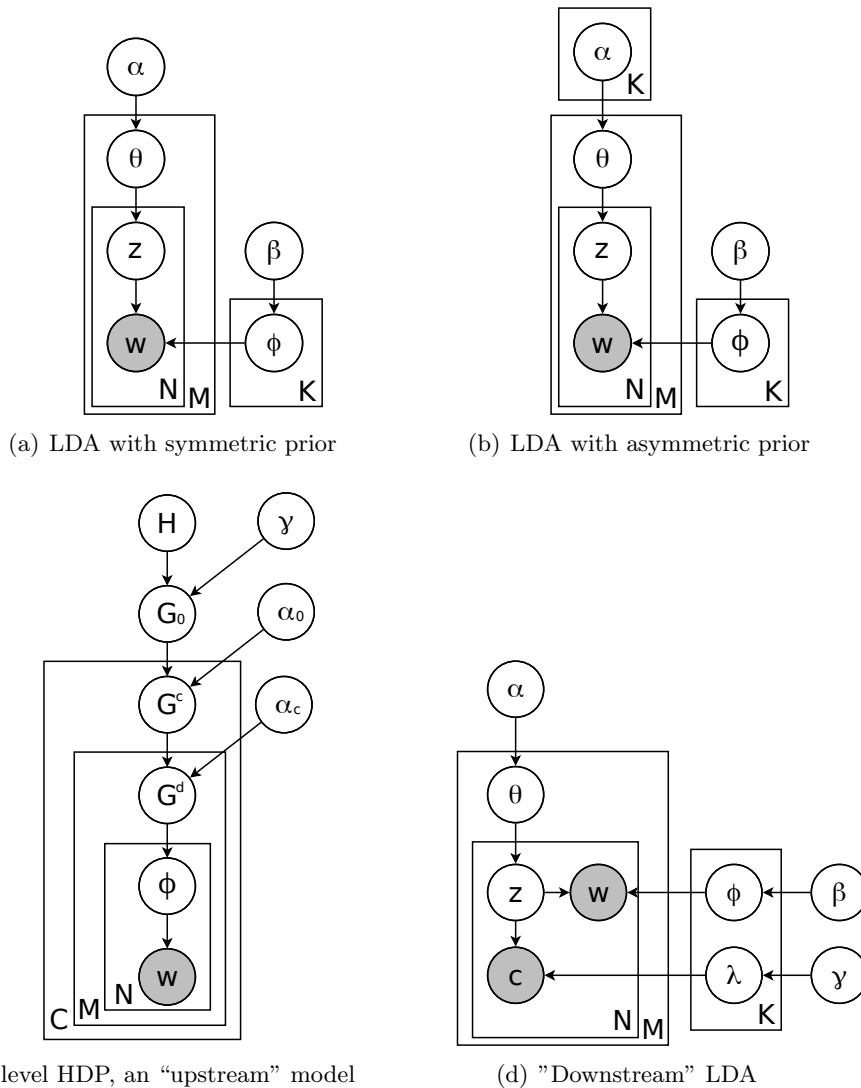


Figure 3.2: Probabilistic topic models for discrete context variables. Figure (a) shows the basic LDA model with a symmetric Dirichlet prior for reference. Figure (b) shows a modified LDA with an asymmetric Dirichlet prior with parameters $\alpha_1 \dots \alpha_K$ over document-topic distributions, which allows to capture topic preferences across similar documents. (c) Groups of documents can be grouped themselves, for instance in a three-level hierarchical Dirichlet process. This enables the sharing of topic preferences between the groups. Both models follow the "upstream" approach, where context information influences the prior over topics. (d) An alternative way of integrating context into probabilistic topic models is to extend the notion of topics by introducing a topic-specific distribution over the context space, additionally to the distribution over the vocabulary. Every word then comes with a copy of the context variable, and both the word and the (copy of the) context variable are drawn from the same topic.

In [YCH⁺11a], the yearly cycle was modelled in a downstream model by mapping it on a timeline and fitting several Gaussian distributions, with means repeating in equal distances. Every topic has an associated mean, giving the position on the periodic interval. However, only a limited number of periodic repetitions can be modelled and thus there is no natural way for adding new documents which appear later in time. Additionally, topics can only have one peak (i.e. one Gaussian distribution) per temporal cycle (i.e. per year).

Yin et al. [YCH⁺11b] propose a clustering of documents based on their location in a preprocessing step which then serves as an input for the Location Driven Model (LDM). Geographically distributed data are clustered in a preprocessing step to obtain a discrete set of locations. In the model, all documents within a spatial cluster share a common topic distribution, similar to the author-topic model [RZGSS04]. The topic distributions of the clusters are modelled as independent variables.

A first approach for modelling geographical topics using Gaussian distributions was proposed by Sizov [Siz10]. GeoFolk first uses the Mercator projection of documents on a two-dimensional map and then makes use of the “downstream” approach. Every topic in GeoFolk has a Gaussian distribution on the coordinates of a document (note that latitude and longitude are modelled as drawn from two independent Gaussians in the original paper). The drawback of this kind of topic modelling clearly is the limited geographical distribution of topics: Every topic has a normal geographical distribution and topic areas that are not normal distributed are split into independent topics.

Yin et al. [YCH⁺11b] therefore introduced *latent geographical topic analysis* (LGTA), an extended version of GeoFolk based on PLSA. Instead of directly assigning normal distributions to topics, in the model of Yin several normal distributions are assigned to *regions* which have a distribution over the set of topics. Clearly, there now can be several Gaussian regions sharing the same topic. Regions now take the role of discrete locations as in the model of Mei. Therefore, the model inherits the problem of merging regions of one kind.

In [AHS13], Ahmed et al. present a hierarchical topic model which models both document and region specific topic distributions and additionally models regional variations of topics. Relations between the Gaussian distributed geographical regions are modelled by assuming a strict hierarchical relation between regions that is learned during inference.

A more general approach for modelling arbitrary, complex context information such as geolocations was introduced by Agovic and Banerjee [AB12]. Given that the similarity between topic distributions of documents directly depends on their respective position in the context space, topic distributions of documents can be sampled from a Gaussian process (GP) prior which encodes the similarity of documents in the context space. However, it

is unclear how to choose the right GP kernel in the geographical scenario, as the similarity of document-topic distributions across the geographical space is typically hard to predict and involves complex structures such as countries or geographical zones.

3.2.4 Categorisation of approaches

The presented methods all share a set of common motifs for integrating context into topic models. A distinction of models based on the context type was used to structure the presentation of topic models. Additionally, the distinction between *upstream* and *downstream* models [MM08] was mentioned. Abstractly speaking, some contextual topic models use a *transformation of context data* in a simpler space, such as discretising continuous variables or mapping geographical data from a three-dimensional sphere on a two-dimensional map. Some models go beyond single distributions (e.g. Gaussians) and model *complex dependencies in the context space* between documents of different geographical regions, e.g. by introducing hierarchical structures or temporal dependencies via prior distributions. While simple models base on PLSA and LDA, advanced models base on Dirichlet processes and thus are *non-parametric*, i.e. the number of topics must not be set in advance. Additionally, the Dirichlet process leads to a coupling of the topic priors of different context variables. A comparison of the methods is shown in Table 3.1.

3.3 Drawbacks of Existing Models

For cyclic and spherical context, existing topic models exhibit weaknesses in modelling complex context structures. As all approaches for explicitly modelling cyclic or spherical context are designed for the geographical context, in the following models for geographic coordinates are examined. However, the problems presented also occur with cyclic context data, e.g. temporal cycles.

Under the bag-of-words assumption, both words and context variables can be modelled as being generated by latent factors in a probabilistic model. Existing approaches for integrating geographical context adopt topic models such as PLSA [Hof99] or LDA [BNJ03] and extend the models by assigning distributions over locations to topics, or by introducing latent geographical regions. In models which extend topics for spatial distributions (such as two-dimensional normal distributions) [Siz10], topics with a complex (i.e non-Gaussian) spatial distribution cannot be detected. In models with latent, Gaussian distributed regions [YCH⁺11b, AHS13], documents within a complex shaped topic area do not influence the topic distribution of distant documents within the same area. Therefore, topics with a complex spatial distribution such as topics distributed along coastlines,

	Downstream / Upstream	Transformation of context	Complex context dependencies	Non-parametric
Discrete context				
Three-level HDP [TJBB06]	U			✓
Citation-Influence [Die06]	U		✓	
Mei et al. [MLSZ06]	U			
Wang et al. [WWXM07]	D		✓	
Linear context				
EvoHDP [ZSZL10]	U	✓	✓	✓
TOT [WM06]	D			
npTOT [DHWX13]	D			✓
iDTM [AX12]	U	✓	✓	✓
Dir-Mult Regress. [MM08]	U		✓	
Cyclic and Spherical context				
LPTA [YCH ⁺ 11a]	D	✓		
LDM [YCH ⁺ 11b]	U	✓		
LGTA [YCH ⁺ 11b]	U	✓		
Agovic et al. [AB12]	U	✓	✓	
GeoFolk [Siz10]	D	✓		
Ahmed et al. [AHS13]	U	✓	✓	✓
MGTM	U	✓	✓	✓

Table 3.1: Comparison of common topic models for discrete and continuous context variables. Upstream models exploit context information using context-dependent priors for document-topic distributions. For complex context variables, it is typical to transform the context variables in order to better model their properties. Advanced models are able to detect complex dependencies beyond simple probability distributions in the context space, e.g. hierarchies. Finally, models based on hierarchical Dirichlet processes share topic information between different context clusters and are non-parametric and do not require a parameter for the number of topics.

rivers or country borders are harder to detect by such methods. More elaborate models introduce artificial assumptions about the structure of geographical distributions by introducing hierarchical structures [AHS13] or by defining Gaussian process kernels [AB12] in advance. Additionally, some approaches [MLSZ06, YCH⁺11b] do not model document-specific topic distributions.

In contrast to existing models, the multi-Dirichlet process (MDP) based geographical topic model (MGTM) presented in this thesis uses a MDP mixture model that groups documents by geographical regions. A geographical network between spatially adjacent regions is used to equalise topic distributions within coherent topic areas. Consequently, it allows for constructing generative models which provide a better data fit than existing approaches.

A geographical topic model is a statistical model of a set of spatially distributed documents that uses word co-occurrences both within texts and within geographical regions.

From an application perspective, location-aware topic models should satisfy the following top-level requirements:

- (1) Modelling document-specific topic distributions: Documents typically cover a small set of topics, an assumption used for prediction (e.g. tag recommendation)
- (2) Recognition of topics with complex (e.g. non-Gaussian) spatial distributions and changing observation densities
- (3) Detection of coherent topic regions that form complex shaped areas of similar characteristics (e.g. countries, seas, mountain ranges, etc.) require prior knowledge for the parameter setting
- (4) The influence of context information on the learned topics should be governed by a parameter which is learned during inference

The existing geographical topic models described in the previous sections have major drawbacks with respect to these requirements. The models of Yin and Hong [HAG⁺12, YCH⁺11b] do not model document-specific topic distributions; the model of Sizov [Siz10] cannot detect topics with a complex spatial distribution; and the model of Wang [WWXM07] supports the merging of semantically related geographical regions but lacks a general merging method. Finally, only the model by Ahmed et al. is parameter-free [AHS13].

Models based on a hierarchical relation between regions such as the model by Wang et al. [WWXM07] and Ahmed et al. [AHS13] have drawbacks, not only in modelling complexity as mentioned in [WWXM07]. Particular hierarchical relations such as *city-state-country* might work for representing geographical topics such as languages or cultural behaviour. However they would be misleading e.g. for topics representing geographical features such as rivers or mountain areas. In most cases, there will be no

hierarchy which fits all topics. Additionally, when introducing a hierarchy of Gaussian distributed regions as in [AHS13], geographical topics which fit into such a hierarchy will be preferred over topics with a non-elliptic shape such as, say, coast lines, which would be poorly approximated by a hierarchy of Gaussian regions. Therefore, introducing a hierarchical relation between regions will prevent the model from properly learning topics whose complex geographical distribution does not fit such a simple hierarchical structure. Table 3.2 summarizes the requirements met by the models presented.

3.4 Multi-Dirichlet Process Topic Models

Consider the general setting of documents consisting of words which are annotated with their geographic location. For topic modelling, words are assumed to be exchangeable within documents, the bag-of-words assumption. Formally, a corpus of size M consists of documents $D = \{d_1, \dots, d_M\}$, and a document d_j consists of a set of N_j words denoted by $\mathbf{w}_j = (w_{j1}, \dots, w_{jN_j})$ and a geographical location, a latitude and longitude pair $\text{loc}_j = (\text{lat}_j, \text{lon}_j)$.

By de Finetti’s theorem [BNJ03], words can be modelled as a mixture of independent and identically distributed random variables generated by latent factors. The document location is generated by L latent factors corresponding to geographical clusters associated with continuous distributions on the geographical space. The K latent factors assigned to words, written as topics ϕ_1, \dots, ϕ_K , are multinomial distributions over the vocabulary of size V .

In the following section, three novel geographical topic models are presented: a basic model using a three-level hierarchical Dirichlet process, an extension that considers neighbour relations between regions by model selec-

Table 3.2: Requirements met by existing models for spherical context, and by the presented model (MGTM). ¹partial fulfilment

Model	Requirements			
	(1)	(2)	(3)	(4)
Mei et al. [MLSZ06]		✓		
Wang et al. [WWXM07]	✓	✓	(✓) ¹	
GeoFolk [Siz10]	✓			
LDM [YCH ⁺ 11b]		✓		
LGTA [YCH ⁺ 11b]		✓		
Agovic et al. [AB12]	✓	✓	(✓) ¹	(✓) ¹
Ahmed et al. [AHS13]	✓	✓	(✓) ¹	✓
MGTM	✓	✓	✓	✓

tion and an improved version based on the multi-Dirichlet process introduced in this thesis.

3.5 The Basic Model

For the basic geographical topic model, locations and words are modelled separately, i.e. the geographical locations are mapped on a set of clusters. This has two reasons:

Detection of coherent topic areas. The separation of spatial clusters and document semantics allows the definition of meaningful neighbour relations between spatially adjacent clusters. In fact, as shown later, the use of these spatial adjacency relations permits the detection of coherent topic areas and significantly improves the topic quality in the final model. Existing models that use continuous document positions do not allow the use of spatial adjacency as a proxy for similarity between the topics of geographical regions. The reason is that in these models the geographical location of a region is influenced both by words and document locations, and thus two geographically adjacent regions can be very dissimilar (i.e. having completely different topic distributions). This prevents the direct modelling of geographically coherent topic areas.

Computational complexity. Probabilistic clustering methods in two- or three-dimensional space usually converge very fast, while samplers for probabilistic topic models usually take many iterations. Integrating both processes would result in a high computational overhead which is unacceptable for large datasets in real-world applications.

The basic topic model takes a set of geographical clusters as input. In order to get a clustering which also is a generative model of document positions, a mixture of Fisher distributions (see Section 1.4.7) is fitted to the data. The clusters are used to group documents in a three-level hierarchical Dirichlet process in order to ensure that documents within a geographical cluster share similar topics. This basic topic model is identical to the topic model for multiple corpora proposed by Teh et al. [TJBB06].

3.5.1 Geographical clustering

Existing approaches for geographical topic modelling rely on a representation of document positions in Euclidean space of latitude and longitude. This causes problems for documents located close to the poles or to the International Date Line. Instead, it is more appropriate to employ the unit sphere as a model for the shape of the Earth. For geographical clustering, it

is reasonable to assume that document locations follow a Fisher distribution, defined in Section 1.4.7.

Given the number of Fisher distributions L and assuming a uniform prior, the expectation-maximisation algorithm presented in Section 1.4.8 is employed for parameter estimation. To construct a non-parametric model where the number of regions is inferred from data, the number of clusters L can be sampled using a Dirichlet process that samples from a space of Fisher distributions [BHO10]. Clearly, the geographical distribution of topics and therefore the topic-word distributions in a trained topic model will depend on the number of regions. In order to ensure comparability between approaches, the number of regions is kept as a fixed parameter in the algorithms evaluated in this chapter.

3.5.2 Topic detection

The choice of the underlying topic model is crucial for the task of geographical topic detection. Existing methods are typically based on PLSA [YCH⁺11b] or LDA [HAG⁺12, Siz10]. The models presented in this theses are based on the hierarchical Dirichlet process (HDP) [TJBB06] instead as it is non-parametric, yields a sound generative model and supports a grouping of documents by external factors such as geographical clusters.

Topic models based on Dirichlet processes always have at least two levels: In order to share topics between documents (see Section 1.8), each document-topic distribution is sampled from a higher-level topic-distribution, e.g. a global topic distribution which is itself a draw from a Dirichlet process [TJBB06].

It is natural to extend the hierarchical scheme by adding layers for document groups with characteristic topic distributions. For the basic model, the three-layer Dirichlet process hierarchy for modelling document corpora proposed in [TJBB06] is applied, but documents are grouped by geographical regions instead using the spatial clustering of documents defined before.

The three-level hierarchical topic model using geographical clusters is then defined as follows: Given a set of L geographical clusters, each cluster is a subset D_l of the document corpus. First, a global probability measure G_0 on the topic space is drawn from a Dirichlet process with base measure H on the continuous topic space:

$$G_0 \sim DP(\gamma, H),$$

where γ is the scaling parameter for the Dirichlet process, influencing the sparsity of the global topic distribution. A symmetric Dirichlet prior is placed over H . The mixture proportions β for the global topic distribution belonging to the base measure G_0 (see Section 1.8.1) are generated by a stick-breaking process $\beta \sim SBP(\gamma)$ [TJBB06]. For every geographical cluster, a

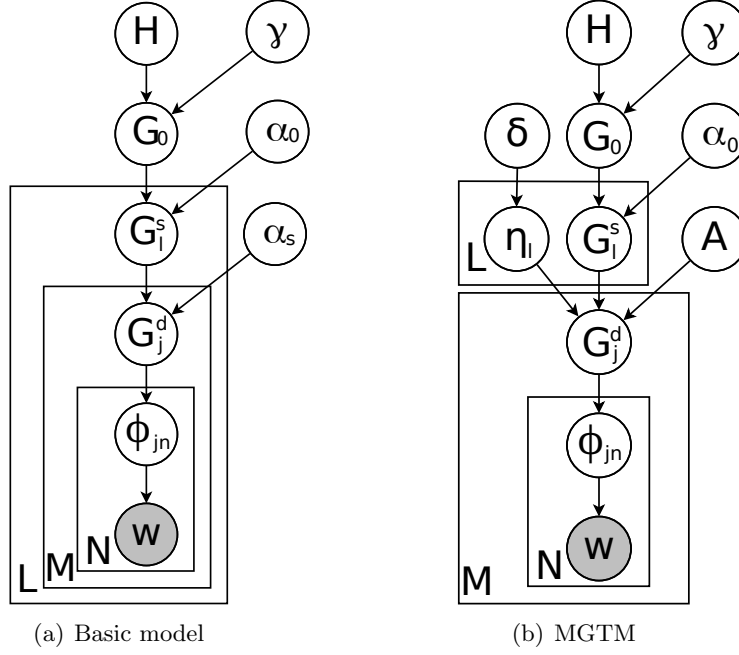


Figure 3.3: Graphical models for (a) the basic model and (b) the multi-Dirichlet process geographical topic model (MGTM). MGTM introduces dependencies between geographical clusters by including adjacency relations between clusters into the model.

region-specific topic distribution G_l^s is drawn from the global measure G_0 on the topic space:

$$G_l^s \sim DP(\alpha_0, G_0), \quad l = 1, \dots, L$$

with mixing proportions β_l^s and scaling parameter α_0 . Finally, the documents from each region-specific document set D_l draw a document-specific topic probability measure from G_l :

$$G_j^d \sim DP(\alpha_s, G_l^s), \quad d_j \in D_l$$

with mixing proportions π_j . All clusters share the common scaling parameter α_s .

Note that superscripts here are used to distinguish the base measures of documents G^d , base measures of geographical clusters G^s and the global base measure G_0 . In the rest of this thesis, **superscripts of base measures or multinomial topic distributions do not denote indices or exponentiations**, but are used to distinguish different levels of hierarchical Dirichlet processes.

The resulting model is given in Figure 3.3(a). A collapsed Gibbs sampler and strategies for hyperparameter inference are given in [TJBB06].

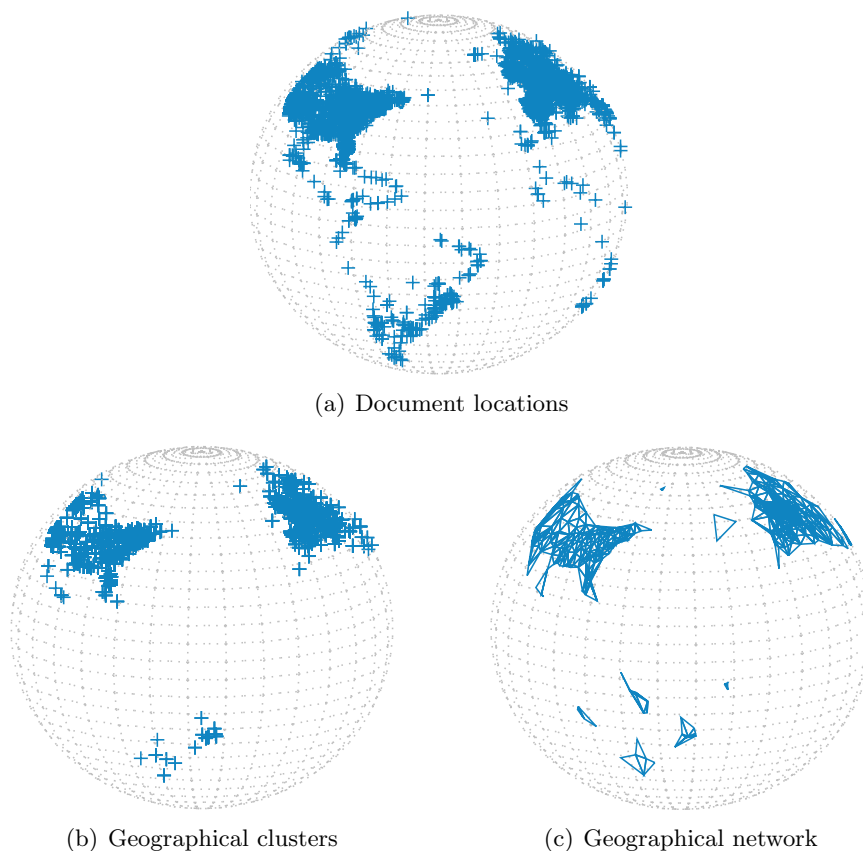


Figure 3.4: Document positions (a), geographical clusters (b) and geographical network (c) for the car dataset presented in Section 3.8.1. The vertices of the geographical network correspond to the means of Fisher distributions fit on the data and the edges are based on a Delaunay triangulation of mean locations. The geographical network is used in the neighbour-aware model (NAM) and in the multi-Dirichlet process topic model (MGTM) to smooth the topic distributions of adjacent clusters.

3.6 The Neighbour-Aware Model

In order to exploit and model complex structures in the context space, the basic model is extended to include *geographical network* information. Geographical networks are constructed based on adjacency relations between geographical clusters, and therefore the resulting model is called the *neighbour-aware model* (NAM). The application of geographical networks is a novelty in geographical topic modelling.

3.6.1 Definition of a geographical network

For cyclic and spherical context variables, a network of clusters can be defined by exploiting adjacency relations between clusters. There are multiple ways of defining spatial adjacency relations. For the models presented here, the Delaunay triangulation is applied, obtained as the dual of the Voronoi diagram [Aur91]. Calculating the Voronoi diagram for points on the sphere (i.e. creating cells around every geographical cluster which contain all points for which the given cluster centroid is closest), and connecting clusters of adjacent Voronoi cells yields the Delaunay triangulation. Adjacency relations defined by the Delaunay triangulation are intuitive and do not require any parameters and thus permit the creation of parameter-free methods.

For the studies in this thesis, triangles with side lengths greater than one eighth of the earth radius are discarded, as it is not expected to find such large structures in data. An example for the resulting geographical network is shown in Figure 3.4.

Note that other adjacency definitions such as k-nearest-neighbour (KNN) could be used as well. However, one has to keep in mind that in real-world data, observation densities typically vary significantly, e.g. for some regions on earth there will be far less observations than on others. Applying naive adjacency definition such as KNN can lead to a large number of network components which prevents the detection of complex coherent structures in the context space.

3.6.2 Advantages of the neighbour-aware model

Using geographical neighbour relations has several advantages over the basic model:

Exploiting similarity for smoothing. Geographical clusters adjacent in space often are similar in their topic distribution. Most geographical topics cannot be approximated by a simple spatial probability distribution such as a Gaussian or Fisher distribution and for these complex topic areas, coherent sets of multiple spatial distributions are a reasonable approximation. Therefore adjacent regions may smooth their topic distributions to increase the probability of detecting such coherent topic areas.

Coping with sparsity. Geographical proximity seems to be a natural factor for smoothing the topic distributions of geographical clusters: The closer two geographical clusters are, the more similar their topic distributions should be. However, as mentioned before, the observation density in the geographical space is often inhomogeneous. Geographical proximity in practice needs to be re-weighted for areas of sparse or dense observations [BFC98]. This re-weighting requires an additional parameter and

would be rather arbitrary. Basing the smoothing on a geographical network of adjacent clusters avoids these issues and is parameter-free.

Sharing emerging topics. In the basic model, new topics emerge locally, first on the document level, then on cluster level and finally on the global level. Under the assumption that adjacent clusters are likely to be similar, new topics should be actively shared with neighbour clusters. Sharing topics through a network of adjacent clusters during Gibbs sampling has an interesting analogy in the evolutionary process of memes [Daw06]. Topic assignments of documents can be seen as limited resources which are occupied by topics. Topics are competing with each other in occupying these resources, and topics which have a better fit (i.e. assign higher likelihoods to observed words) have an evolutionary advantage against other topics. At the same time, topics evolve to better fit to the occupied words (i.e. the topic-word multinomial is constantly updated based on topic assignments). The sharing of topics through a geographical network resembles the transmission of memes through social interaction. Eventually, strong topics, which describe observed documents well, will survive, while poor topics will perish during the sampling process. A schematic representation of this process is depicted in Figure 3.5.

The idea behind the extended topic model is to include an uncertainty over the cluster membership of documents in clusters in order to more strongly connect topic distributions of adjacent geographical clusters. Each document topic distribution is assumed to be drawn either from the topic distribution of its geographical cluster or from one of the adjacent cluster distributions. P_l is the union of the cluster index l and the set of neighbour cluster indices, and λ_j indicates from which cluster-specific topic distribution β_r^s the document d_j was sampled. The set of topic distributions β^s can be used for Bayesian model selection: Given a uniform prior over the probability for a document d_j to be sampled from G_r^s with $r \in P_l$, the sampling equation for λ_j is

$$p(\lambda_j = r \mid \mathbf{z}, \mathbf{m}, \beta^s) \propto \prod_{k=1}^K (\alpha_s \beta_{rk}^s)^{m_{jk}} \quad (r \in P_l) \quad (3.1)$$

where m_{jk} gives the number of times a topic was drawn from the global topic distribution. This equation is identical to the sampling equation in [CG11], except that the weights of the document-specific topic distribution π_j are integrated out. The model structure then is sampled during Gibbs sampling and the rest of the sampler remains the same as for the basic model.

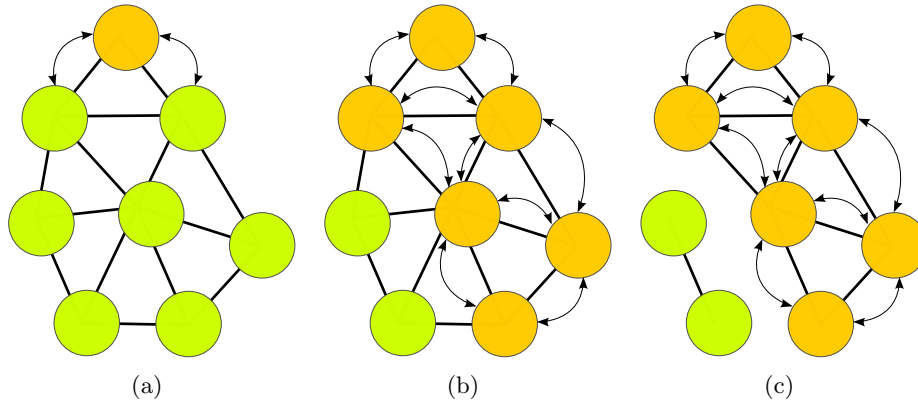


Figure 3.5: Schematic representation of the sharing of topic information between adjacent clusters in the neighbour-aware model (NAM) and in the multi-Dirichlet process topic model (MGTM). (a) During Gibbs sampling, a newly detected topic (denoted by a yellow node) will be offered to adjacent geographical clusters via a common topic prior. (b) If some of the words in the documents of the adjacent clusters are better explained by this new topic, the priors of the adjacent clusters will assign a high probability to the novel topic as well. (c) If an adjacent cluster does not take on the topics of the neighbour clusters, the parameter for topic exchange is lowered, ultimately deactivating the topic exchange between dissimilar adjacent clusters.

3.7 The MDP Based Geographical Topic Model

The neighbour-aware topic model clearly leads to an interaction between adjacent cluster-topic distributions. However, in some cases this interaction does not yield the intended smoothing.

Consider the example of two adjacent geographical clusters which both have a high probability for two topics, while other geographical cluster-topic distributions assign very low probabilities to both of the topics. Now, the probability of this model would clearly be maximised if one of the two clusters has a very high probability for the first topic, and the other cluster for the second topic. This (unwanted) effect occurs in cases where there are only few adjacent clusters with high probabilities for a small set of topics. In practice, this is often the case as data are sparse and the number of geographical clusters is small.

To overcome this apparent drawback, a *dynamic smoothing* technique is introduced. This smoothing is based on the *multi-Dirichlet process* (MDP), a generalisation of the Dirichlet process that combines multiple base measures into a single mixing distribution over the space of the base measures.

Note that a Dirichlet-distributed mixture of Dirichlet processes was in-

independently developed by Lin et al. [LF12] and applied in two-level mixtures of uncoupled MDPs.

In the following, the Multi-Dirichlet process is defined and an inference scheme is derived. Finally, a MDP-based topic model is presented which allows for an improved smoothing of topic distributions of geographical clusters.

3.7.1 The multi-Dirichlet process

The *multi-Dirichlet process* (MDP) is defined using a notation similar to that used in [TJBB06]. Let G_1, \dots, G_P be probability measures on a standard Borel space (Φ, \mathcal{B}) associated with positive real parameters $\alpha_1, \dots, \alpha_P$. Then the multi-Dirichlet process $MDP(\alpha_1, \dots, \alpha_P, G_1, \dots, G_P)$ is defined as a probability measure G over (Φ, \mathcal{B}) , which for every finite measurable partition (A_1, \dots, A_r) of Φ yields a Dirichlet-distributed random vector, denoted $(G(A_1), \dots, G(A_r))$, with:

$$(G(A_1), \dots, G(A_r)) \sim \text{Dir} \left(\sum_{p=1}^P \alpha_p G_p(A_1), \dots, \sum_{p=1}^P \alpha_p G_p(A_r) \right) \quad (3.2)$$

In the following, the base measures are referred to as *parent distributions* of the MDP. An alternative notation of the concentration parameters $\alpha_1, \dots, \alpha_P$ is given by

$$A = \sum_{p=1}^P \alpha_p \quad \eta_p = \frac{\alpha_p}{A}, \quad p \in \{1, \dots, P\} \quad (3.3)$$

which gives a convenient parametrisation for the MDP:

$$MDP(A, \eta_1, \dots, \eta_P, G_1, \dots, G_P).$$

Using the alternative notation, the MDP can be understood as a Dirichlet process with base distribution $G_0 = \sum_{p=1}^P \eta_p G_p$, the weighted sum of parent distributions, and scaling parameter A . Given a set of observed samples from G , $\phi_1, \dots, \phi_{i-1}$, the probability of a factor $\phi_i \in \Phi$ to be sampled from G can be estimated by integrating out G using the properties of the Dirichlet distributed partitions [Nea00] and replacing the base measure with the weighted sum of parent distributions:

$$\phi_i \mid \phi_1, \dots, \phi_{i-1} \sim \frac{1}{i-1+A} \sum_{j=1}^{i-1} \delta(\phi_j) + A \sum_{p=1}^P \frac{\eta_p}{i-1+A} G_p \quad (3.4)$$

with $\delta(\phi_j)$ being the Dirac delta, giving weight to a single point ϕ_j . One immediately can see that a MDP with a single parent distribution yields a standard Dirichlet process.

3.7.2 Inference

Topic assignments can be sampled by extending the inference strategies using the Chinese restaurant franchise representation by Teh et al. [TJBB06] introduced in Section 1.8.1: For a given two-level hierarchical Dirichlet process, global “dishes” are introduced, corresponding to the Dirichlet distributed random variables on the first level from which the Dirichlet process of the second level samples factors ϕ_j . For factor sampling, customers corresponding to the factors ϕ_j form Dirichlet distributed groups sitting at tables in a restaurant and all customers at a table share the same dish. The number of customers at the i th table is given by m_i and the tables are samples from the Dirichlet distributed base distribution of dishes. A detailed explanation of the Chinese restaurant process and its parameters is given in [TJBB06]. The Gibbs sampling equation for topic assignment z_{ji} of word w_{ji} in document d_j is:

$$p(z_{ji}=k \mid \mathbf{z}_{-ji}, \mathbf{m}, \boldsymbol{\beta}^s) \propto (m_{jk} + \sum_{p \in P_l} \alpha_p \beta_{pk}^s) f_k^{-x_{ji}}(x_{ji}) \quad (3.5)$$

for topics already sampled, and

$$p(z_{ji}=k^{\text{new}} \mid \mathbf{z}_{-ji}, \mathbf{m}, \boldsymbol{\beta}^s) \propto \left(\sum_{p \in P_l} \alpha_p \beta_{pk}^s \right) f_{k^{\text{new}}}^{-x_{ji}}(x_{ji}) \quad (3.6)$$

for new topics where β_p^s are mixing proportions of the parent distributions, $f_k^{-x_{ji}}(x_{ji})$ is a topic-specific probability function with parameters from the parent distribution and \mathbf{z}_{-ji} denotes the set of all topic assignments except for z_{ji} . The number of customers in document d_j assigned to the k th factor is given by m_{jk} .

For sampling the number of components, the MDP can be interpreted as a multinomial mixture of stick-breaking weights of multiple Dirichlet processes. This becomes apparent if one makes use of the alternative representation from Eq. 3.3. Substituting α_0 by (a sum of) $A\eta_{lp}$ in Equation 40 from [TJBB06] gives:

$$p(m_{jk}=\mathbf{m} \mid \mathbf{z}, \mathbf{m}_{-jk}) = \frac{\Gamma(\sum_{p \in P_l} A\eta_{lp} \beta_{pk}^s)}{\Gamma((\sum_{p \in P_l} A\eta_{lp} \beta_{pk}^s) + n_{jk})} s(n_{jk}, m_{jk}) \cdot A^{m_{jk}} \binom{m_{jk}}{m_{jk1}, \dots, m_{jkP}} \prod_{p \in P_l} (\eta_{lp} \beta_{pk}^s)^{m_{jkp}} \quad (3.7)$$

and where l is the index of the MDP of document d_j and P_l denotes the set of parent distribution indices. The function $s(n, m)$ denotes the unsigned Stirling numbers of the first kind introduced in Section 1.8.2 and $\binom{m_{jk}}{m_{jk1}, \dots, m_{jkP}}$ the multinomial coefficient. The number of tables m_{jkp} is tracked per topic

and parent and sampled simultaneously for all parent distributions. Variable m_{jk} denotes the sum of tables over all parents, n_{jk} is the number of customers (topic assignments) for a given document and topic. For sampling the tables, Gamma functions are dropped as they do not depend on \mathbf{m} , the sum of tables m_{jk} is sampled per topic and then the parent specific table counts m_{jkp} are the result of m_{jk} draws from a multinomial with normalised parameters $\eta_{lp}\beta_{pk}$.

Sampling the weights β_p for each G_p is done using $m_{.kp}$, the sum over all tables of topic k and parent p from documents with parent distribution G_p . If G_p is sampled from a parent Dirichlet process with scaling parameter α_0 and weights β , then

$$\beta_p \sim \text{Dir}(m_{.1p} + \alpha_0\beta_1, \dots, m_{.Kp} + \alpha_0\beta_K, \alpha_0\beta_u) \quad (3.8)$$

where β_k denotes the weight of topic k in the parent Dirichlet process and β_u is the weight of the previously unseen topics.

3.7.3 Estimation of scaling parameters for the MDP

Sampling for scaling parameters α_p is similar to the sampling for Dirichlet processes as described in [TJBB06]. Instead of directly sampling the concentration parameters α_p , first A is sampled and then η . The probability of the total table counts for all documents in the MDP is given by:

$$p(\mathbf{m}_l | \mathbf{n}, \mathbf{m}, \boldsymbol{\eta}, A) = \prod_{j \in D_l} \frac{\Gamma(A)}{\Gamma(A + n_{j.})} s(n_{j.}, m_{j.}) A^{m_{j.}} \cdot \binom{m_{j.}}{m_{j1}, \dots, m_{jP}} \prod_{p \in P_l} \eta_{lp}^{m_{jp}} \quad (3.9)$$

where D_l is the set of documents which is sampled from the MDP with index l . The left part of the equation is identical to Equation 44 in [TJBB06] with parameter A as concentration parameter. Therefore sampling for A is identical as for a normal DP. The document specific table counts $m_{j.}$ are obtained by summing over the sampling results from Equation 3.7. Obviously, the right side of Equation 3.9 is a multinomial again. As $\boldsymbol{\eta}$ governs the influence of parent distributions, it is possible to introduce a symmetric Dirichlet prior over the sampling parameters:

$$\eta_l \sim \text{Dir}(\delta_l) \quad (3.10)$$

For a MDP with index l , Bayesian inference for the multinomial parameters η_l then yields an estimate based on the table counts of the parent DPs:

$$\hat{\eta}_{lp} = \frac{m_{.p} + \delta}{m_{..} + |P_l|\delta} \quad (3.11)$$

3.7.4 MDP-based topic model

The extension of NAM for the multi-Dirichlet process, the multi-Dirichlet process geographical topic model (MGTM), is obtained by replacing the model selection for uncertain cluster memberships by multi-Dirichlet processes. Instead of sampling for document memberships from the set of potential parent distributions P_l , the potential parent distributions of the NAM are used as indices of parent base distributions in a MDP. Every document has one or more parent base distributions G_r^s (the topic distribution of a spatial cluster) with $r \in P_l$, holding the indices of the region of the document and the adjacent regions. A schematic representation of the resulting dependencies is shown in Figure 3.6. The weight of region r in the MDP is given by η_{lr} and is estimated during inference. With a concentration parameter $\delta > 1$ the cluster weights of parent distributions are smoothed, which is explicitly coding the assumption that adjacent clusters are similar. The concentration parameter δ can be estimated from data as well, allowing insights into the similarity of regions.

As the number of adjacent regions varies across regions, and the dimensionality of η_l is changing, standard inference schemes from literature cannot be applied. Therefore, a generalised estimator for the concentration parameter of a Dirichlet distribution with changing dimensionality is derived in Section 3.7.6.

3.7.5 Generative process

The generative process of MGTM is:

1. Draw a global topic measure

$$G_0 \sim \text{DP}(\gamma, H)$$

2. Draw cluster-specific topic measures:

$$G_l^s \sim \text{DP}(\alpha, G_0) \quad (3.12)$$

3. For every geographical cluster, draw a weighting over the parent cluster-specific topic distributions:

$$\boldsymbol{\eta}_l \sim \text{Dir}(\delta) \quad (3.13)$$

4. For every document of every geographical cluster l with N_l adjacent clusters, draw a topic measure from a multi-Dirichlet process with:

$$G_j^d \sim \text{MDP}(A, \eta_{l1}, \dots, \eta_{lN_l}, G_{P_{l1}}^s, \dots, G_{P_{lN_l}}^s) \quad (3.14)$$

5. For every word n in document j , draw a topic-word distribution and a word w_{jn} from this distribution

$$\phi_{jn} \sim G_j^d \quad w_{jn} \sim \phi_{jn}$$

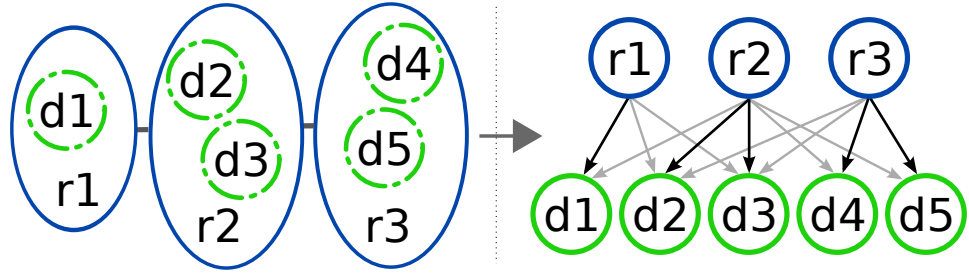


Figure 3.6: Modelling of adjacency relations in MGTM. The geographical adjacency of regions (left) is used in the model to derive dependencies of document-specific topic distributions from the topic distributions of regions (right). Dependencies from regions adjacent to the region of a document are shown in grey.

The resulting model is shown in Figure 3.3(b). The dashed arrow connecting the cluster specific topic distributions G_l^s and the document specific distributions G_j^d indicates that not every cluster specific distribution is a parent base distribution of each MDP.

In MGTM, all documents of a given region share the same MDP and thus the same weights η_l for the parent topic distributions of the region and its neighbour regions. Each region stores the influence of its adjacent regions on the topic distribution of the contained documents in the multinomial parameters $\boldsymbol{\eta}_l$ which is adjusted during inference. Given P_l , the union of the region index l and its neighbour region indices, $\boldsymbol{\eta}_l$ assigns a probability to every region to be chosen as a base topic distribution by the documents D_l in region l , and thus η_{lr} is an indicator of similarity between the l th region and its r th neighbour.

As mentioned, it is possible to smooth the influence of adjacent regions by setting a Dirichlet prior over η . In contrast, a prior for the model selection of NAM (Eq. 3.1) only could re-weight but not smooth the probability for cluster memberships. Using the MDP, one obtains a flexible and stable framework for sharing information across context clusters. The influence of adjacent context clusters is learned during inference, and the MDP allows for a smoothing of topic distributions of adjacent clusters via a Dirichlet prior.

The resulting model has an important advantage: For the neighbour-aware model, it was assumed that adjacent spatial clusters are similar and therefore their topic distributions should be smoothed. However, this assumption could not be explicitly modelled in the NAM. The MDP ensures that the model in fact creates homogeneous topic distributions for similar adjacent regions and at the same time prevents a smoothing of dissimilar regions by adjusting the influence parameter $\boldsymbol{\eta}$ during the sampling process and by re-estimating the concentration parameter δ , leading to a *dynamic*

smoothing of topic regions.

An MDP-based topic model also could be used to create a model in which all documents are connected to all base measures of the second-level Dirichlet process. This topic model would learn a clustering of documents and similarity relations between clusters. MGTM restricts documents to base measures of their geographical clusters to make sure that the learned patterns correspond to geographical structures.

3.7.6 Generalised estimator of the concentration parameter

For the MDP based model, a dynamic smoothing of topic distributions between adjacent geographical clusters based on a multinomial distribution with a symmetric Dirichlet prior was introduced.

Unlike in standard applications, the dimensionality of the symmetric Dirichlet distributions potentially changes for different MDPs in the model. The reason behind this is the changing number of adjacent clusters in the underlying geographical network. For each cluster, the multinomial distribution drawn from the Dirichlet prior has $P_i \in \mathbb{N}^+$ dimensions, corresponding to the probability of drawing a topic from cluster $p \in \{1, \dots, P\}$. Despite the changing dimensionality, the consistent semantics of the concentration parameter (effectively working as pseudo-counts during inference) makes it desirable to learn a single symmetric parameter δ which is shared by all Dirichlet distributions regardless of their dimensionality. In this section, the inference method by Minka [Min00] is reviewed and a novel, generalised inference scheme for estimating the shared concentration parameter of multiple Dirichlet distributions with changing dimensionality is presented.

The widely-used parameter estimate presented by Minka [Min00] is based on a maximum likelihood estimation employing a lower bound on the likelihood [Hua05]. The likelihood of observations $D = \{x_1, x_2, \dots, x_M\}$ with counts m_{np} for the p th observation of the multinomial in the n th trial and asymmetric Dirichlet prior $\vec{\alpha} = \{\alpha_1, \dots, \alpha_P\}$ is given by

$$p(D | \vec{\alpha}) = \prod_{n=1}^M p(x_n | \vec{\alpha}) = \prod_{n=1}^M \left(\frac{\Gamma(\sum_{p=1}^P \alpha_p)}{\Gamma(m_n + \sum_{p=1}^P \alpha_p)} \prod_{p=1}^P \frac{\Gamma(m_{np} + \alpha_p)}{\Gamma(\alpha_p)} \right). \quad (3.15)$$

Now the straightforward inference approach would be to maximise the likelihood directly. However, the derivative with respect to α_p contains digamma functions of α_k and there exists no closed form solution for the optimisation. One can find the parameters maximising the likelihood using the Newton-Raphson method as for instance done in Appendix 4.2 in [BNJ03].

Minka instead proposes to optimise a lower bound on the likelihood

which is based on two inequalities [Min00]:

$$\frac{\Gamma(x)}{\Gamma(n+x)} \geq \frac{\Gamma(\hat{x}) \cdot e^{(\hat{x}-x) \cdot b}}{\Gamma(n+\hat{x})} \quad (3.16)$$

with

$$b = \Psi(n+\hat{x}) - \Psi(\hat{x}) \quad (3.17)$$

where $\hat{x} \in \mathbb{R}^+$ can take on arbitrary values;

and (for $n \geq 1$):

$$\frac{\Gamma(n+x)}{\Gamma(x)} \geq c \cdot x^a \quad (3.18)$$

with

$$a = (\Psi(n+\hat{x}) - \Psi(\hat{x})) \cdot \hat{x} \quad (3.19)$$

$$c = \frac{\Gamma(n+\hat{x})}{\Gamma(\hat{x})} \cdot \hat{x}^{-a} \quad (3.20)$$

where $\hat{x} \in \mathbb{R}^+$ again. The first inequality is intuitive for $x > 1$, because in this case the digamma function can be approximated as $\Psi(x) \approx \log(x - 0.5)$ [AWST12] so that for Equation 3.16

$$\frac{\Gamma(\hat{x})e^{(x-\hat{x}) \cdot b}}{\Gamma(n+\hat{x})} \approx \frac{\Gamma(\hat{x})}{\Gamma(n+\hat{x})} \cdot \left(\frac{(n+\hat{x}-0.5)}{(\hat{x}-0.5)} \right)^{(\hat{x}-x)} \quad (3.21)$$

where the right-hand side of the formula creates a lower bound on the gamma function by multiplying and dividing by $(\hat{x} - 0.5)$ and $(n + \hat{x} - 0.5)$ respectively, instead of counting up until x and $n + x$.

Plugging both inequalities into Equation 3.15 yields:

$$p(D | \vec{\alpha}) \geq \prod_{n=1}^M \left(\frac{\Gamma(\sum_{p=1}^P \hat{\alpha}_p) \cdot \exp((\hat{\alpha}_p - \alpha_p) \cdot b_p)}{\Gamma(m_n + \sum_{p=1}^P \hat{\alpha}_p)} \prod_{p=1}^P c_{pn} \cdot \alpha_p^{a_{np}} \right) \quad (3.22)$$

and taking the logarithm gives:

$$\begin{aligned} \log p(D | \vec{\alpha}) \geq & \sum_{n=1}^M \left(\log(\Gamma(\sum_{p=1}^P \hat{\alpha}_p)) + \sum_{p=1}^P \hat{\alpha}_p \cdot b_n - \sum_{p=1}^P \alpha_p \cdot b_n \right. \\ & \left. - \log \left(\Gamma(m_n + \sum_{p=1}^P \hat{\alpha}_p) \right) + \sum_{p=1}^P c + a_{np} \cdot \log(\alpha_p) \right). \end{aligned} \quad (3.23)$$

For maximising the likelihood with respect to parameter α_p , one first sorts out those summands of the formula which do not depend on α_p :

$$\log p(D | \vec{\alpha}) \geq - \sum_{p=1}^P \alpha_p \cdot \sum_{n=1}^M b_n + \sum_{n=1}^M \sum_{p=1}^P a_{np} \cdot \log(\alpha_p) + (\text{const.}) \quad (3.24)$$

Note that the corresponding Equation 130 in the original paper by Minka gives an incorrect log-likelihood estimate which fortunately does not affect the result of the differentiation [Min00].

Finally, setting the derivative of the log-likelihood equal to zero leads to a fixed-point iteration which in every step optimises $\alpha_p = \alpha_p^{new}$ in order to get a tighter lower bound based on the old estimates $\hat{\alpha}_p = \alpha_p^{old}$:

$$\alpha_p^{new} = \alpha_p^{old} \frac{\sum_{n=1}^M (\Psi(m_{np} + \alpha_p^{old}) - \Psi(\alpha_p^{old}))}{\sum_{n=1}^M (\Psi(m_n + \sum_{p=0}^P \alpha_p^{old}) - \Psi(\sum_{p=0}^P \alpha_p^{old}))} \quad (3.25)$$

For the symmetric prior with changing number of outcomes, the likelihood of multinomially-distributed observations $D = \{x_1, x_2, \dots, x_M\}$ with counts m_{np} (in MGTm corresponding to the tables of parent p of the n th document) is

$$p(D | \alpha) = \prod_{n=1}^M p(x_n | \alpha) = \prod_{n=1}^M \left(\frac{\Gamma(P_n \cdot \alpha)}{\Gamma(m_n + P_n \cdot \alpha)} \prod_{p=1}^P \frac{\Gamma(m_{np} + \alpha)}{\Gamma(\alpha)} \right) \quad (3.26)$$

with P_n being the number of possible outcomes (in MGTm the number of connected clusters in the geographical network) for the n th document. A lower bound then is given by

$$\log p(D | \hat{\alpha}) \geq \sum_{n=1}^M \left(\log(\Gamma(P_n \cdot \hat{\alpha})) + P_n \cdot \hat{\alpha} \cdot b_n - P_n \cdot \alpha \cdot b_n - \log(\Gamma(m_n + P_n \cdot \hat{\alpha})) + \sum_{p=1}^P (c + a_{np} \cdot \log(\hat{\alpha})) \right) \quad (3.27)$$

$$= \sum_{n=1}^M (-P_n \cdot \alpha \cdot b_n) + \sum_{n=1}^M \sum_{p=1}^P a_{np} \cdot \log(\alpha) + (\text{const.}) \quad (3.28)$$

Setting the derivative to zero yields:

$$\begin{aligned} \sum_{n=1}^M (-P_n \cdot b_n) + \sum_{n=1}^M \sum_{p=1}^P a_{np} \cdot \frac{1}{\alpha} &\stackrel{!}{=} 0 \\ \Leftrightarrow \alpha &= \frac{\sum_{n=1}^M \sum_{p=1}^P a_{np}}{\sum_{n=1}^M P_n \cdot b_n} \end{aligned}$$

so that

$$\alpha^{new} = \alpha^{old} \cdot \frac{\sum_{n=1}^M (\sum_{p=1}^P \Psi(m_{np} + \alpha^{old}) - P_n \cdot \Psi(\alpha^{old}))}{\sum_{n=1}^M (P_n \cdot \Psi(m_n + P_n \cdot \alpha^{old}) - \Psi(P_n \cdot \alpha^{old}))}. \quad (3.29)$$

In practice, the computation of alpha can be sped up by using a random sample of observed counts, as the calculation of digamma functions is relatively expensive and the required accuracy, e.g. for topic models, is typically small. However, for the multi-Dirichlet process topic model, the number of regions often is small enough to calculate the concentration parameter on all observed table counts.

3.8 Evaluation

In this section, the ability of MGTM to improve the quality of topics by detecting more accurate, coherent topic areas is demonstrated. The evaluation is in four parts: First, the basic model, the neighbour-aware model, the multi-Dirichlet process model and a state-of-the-art model for geographical topic detection, LGTA by Yin et al. are compared, using the datasets and parameters given in [YCH⁺11b]. Second, the influence of raised region parameters on topic quality is evaluated for all four methods. Then, the runtime of the presented methods is compared on the largest dataset for larger region parameters. Finally, the topic quality of LGTA and MGTM is evaluated in a user study, based on topics trained on the largest of the datasets.

3.8.1 Datasets

As the evaluation of topic models is heavily dependent on the datasets used for comparison, it is crucial to use existing datasets in order to guarantee a fair comparison. Therefore, the evaluation is based on the datasets from [YCH⁺11b], created for the evaluation of the LGTA model. The datasets consist of photographs with geographic coordinates and text tags from the photo sharing service Flickr. The landscape dataset contains 5,791 photos, tagged by “landscape” together with the terms *mountains*, *mountain*, *beach*, *ocean*, *coast*, *desert* from within the US. Topic models should recognise separated topics for mountain regions, coastal regions and desert areas as they belong to different, almost mutually exclusive geographical landscapes. The activities dataset contains 1,931 images, taken within the US and tagged by “hiking” and “surfing”. These two activities should be recognised as separate topics of behaviour. The car dataset contains 34,707 globally distributed images annotated with *chevrolet*, *pontiac*, *cadillac*, *gmc*, *buick*, *audi*, *bmw*, *mercedesbenz*, *fiat*, *peugeot*, *citroen* or *renault*, filtered for event-like images tagged with *autoshow*, *show*, *race*, *racing*. Only the tags from the set of car brands were kept. Concerning the geographical topics, American, German and French car brands are expected to be detected. The Manhattan dataset consists of images from New York containing the tag *manhattan*. Different parts of Manhattan should be detected. For the food dataset, Yin et al. filtered geotagged photos containing the tags *cuisine*,

food, gourmet, restaurant, restaurants, breakfast, lunch, dinner, appetizer, entree, dessert and kept 278 co-occurring tags. Cultural food patterns such as national cuisines are latent topics hidden in the data [YCH⁺11b]. An overview of the data is given in Table 3.3.

3.8.2 Experimental setting

In order to test the generalisation performance of geographical topic models, the word perplexity (see Section 1.7.2) is calculated. Lower perplexity values indicate a better model fit. Yin et al. used the word perplexity in their evaluation of LGTA which they showed to be superior to GeoFolk [Siz10] and a set of basic geographical topic models. Models that outperform LGTA in perplexity therefore also outperform GeoFolk and the basic methods evaluated by Yin [YCH⁺11b].

The comparison between the basic Dirichlet process-based model and its extensions is needed to test the effect of including the additional information of the geographical network in the model and to compare the multi-Dirichlet process with a smoothing mechanism based on model selection.

For each perplexity calculation, a random 80% / 20% split is used to create a training set D_{train} and a test set D_{test} . As explained in Section 3.4, each document d_j is represented by a word set \mathbf{w}_j . The likelihood of words in held-out documents is calculated using the location of documents, the set of topics and other parameters sampled from the training dataset.

For the hierarchical Dirichlet process-based models, the probability of a document is given by

$$p(\mathbf{w}_j) = \prod_{w_i \in \mathbf{w}_j} \sum_{k=1}^{K+1} \phi_{kw_i} \pi_{jk}$$

where ϕ_{kw_i} is the probability of word w_i under topic k and π_{jk} is the document-topic distribution for topic k . $K + 1$ denotes the index of a previously unseen topic and the topic-word probability ϕ_{K+1, w_i} is given by

Table 3.3: Collection period, document count (M) and vocabulary size (V) of the datasets used for comparison [YCH⁺11b]

Dataset	Collection period	M	V
Landscape	09/01/2009 – 09/01/2010	5.791	1.143
Activity	09/01/2009 – 09/01/2010	1.931	408
Manhattan	09/01/2009 – 09/01/2010	28.922	868
Car	01/01/2006 – 09/01/2010	34.707	12
Food	01/01/2006 – 09/01/2010	151.747	278

$\phi_{K+1,t} = 1/V$, $t \in \{1, \dots, V\}$ as new topics are drawn from a symmetric Dirichlet prior over the topic space.

For convenience, the parameters for the (multi-)Dirichlet processes are set to the values used for the evaluation of three-level HDPs in [TJBB06]. A *Gamma*(1, 0.1) prior is assigned to γ and a *Gamma*(1, 1) prior to α_0 , α_s and A . The concentration parameters are initialised to 1. For the multi-Dirichlet process, the weights η of parent distributions are initialised to $1/P$ and the concentration parameter is set to $\delta = 10$. The base measure H is a symmetric Dirichlet distribution with concentration parameter 0.5, except for the car dataset where the parameter is set to 5 for smooth topic-word distributions, as all car brands are expected to appear in all topic areas. The number of iterations of the Gibbs sampler is set to a low value of 200. The source code for MGTm is available from: <https://github.com/ckling/mgtm>.

For the evaluation of LGTA, the parameters from the original paper by Yin [YCH⁺11b] were used. The stopping criterion is set to a change in log-likelihood lower than 0.0001 and the background model weight is set to 0.1. LGTA requires a parameter for the number of normally distributed regions which is analogous to the number of geographical clusters (Fisher distributions) in the models presented in this section. For comparison, identical numbers of regions are used. The setting depends on the dataset and is taken from the LGTA paper.

As the number of detected topics varies for models based on the hierarchical Dirichlet process, each of the HDP-based methods is run 100 times on each dataset and the resulting perplexity is averaged for topic counts with at least ten samples for all three models. The perplexity of LGTA is calculated for the same number of topics by averaging over ten runs.

3.8.3 Comparison with LGTA

Resulting perplexity scores for each model are given in Figure 3.7 (a), (c), (e), (g) and (h). The experiments show that the base model, NAM and MGTm are superior to LGTA for all datasets. This finding can be explained by the ability of the models to model document-specific topic distributions that cannot be detected by LGTA. However, the performance of the HDP-based models differs. For the globally distributed datasets (car and food), MGTm performs significantly better than the base model and NAM. In contrast, for local datasets with a small number of regions, all HDP-based methods perform comparably well.

The dynamic smoothing by MGTm helps to detect coherent topic regions and can effectively improve the topic quality for large, complex structured data while for simple datasets the basic model performs similar or even better.

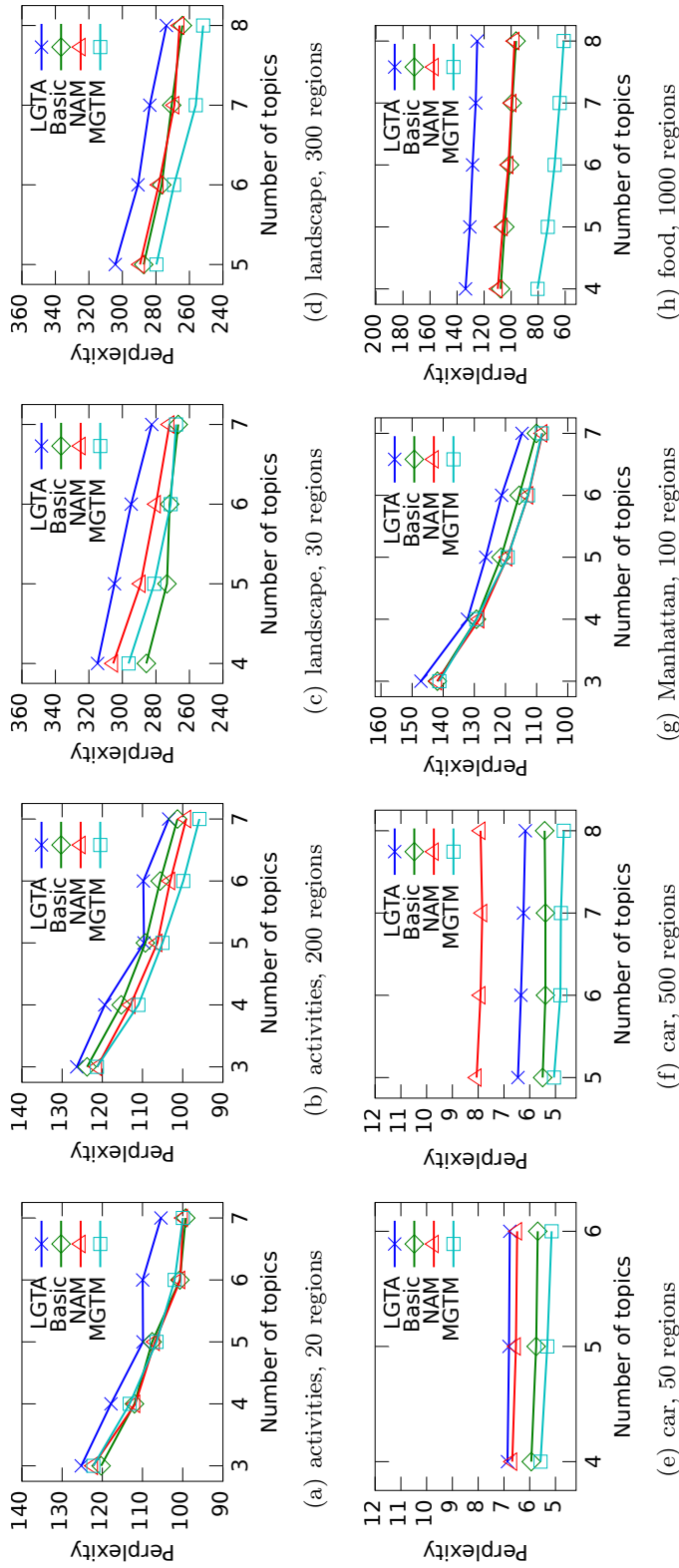


Figure 3.7: Comparison of average word perplexity for LGTA, the basic model, NAM and MGTM for given topic and region counts. Lower values are better. The HDP-based models sample the number of topics, therefore the average perplexity is shown for topic counts that occur in at least 10 out of 100 runs for each method. MGTM is able to improve the topic prediction for a higher number of geographical regions, corresponding to the learning of a more detailed geographical distribution of topics. LGTA, the basic model and NAM are unable to cope with the sparsity of regions, while MGTM learns which adjacent geographical regions are similar and smooths the topic distributions of these regions.

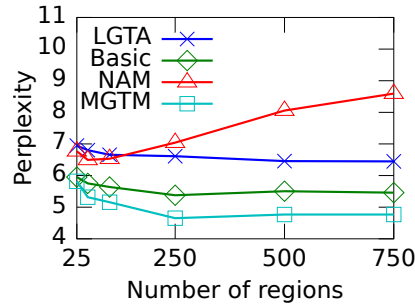


Figure 3.8: Comparison of average per-word perplexity for LGTA, the basic model, NAM and MGTM on the car dataset with five topics for growing region counts. The plot shows the undesired behaviour of NAM to optimise the posterior by setting the topic probabilities of sparse regions to extreme values, leading to an increased perplexity for high region counts.

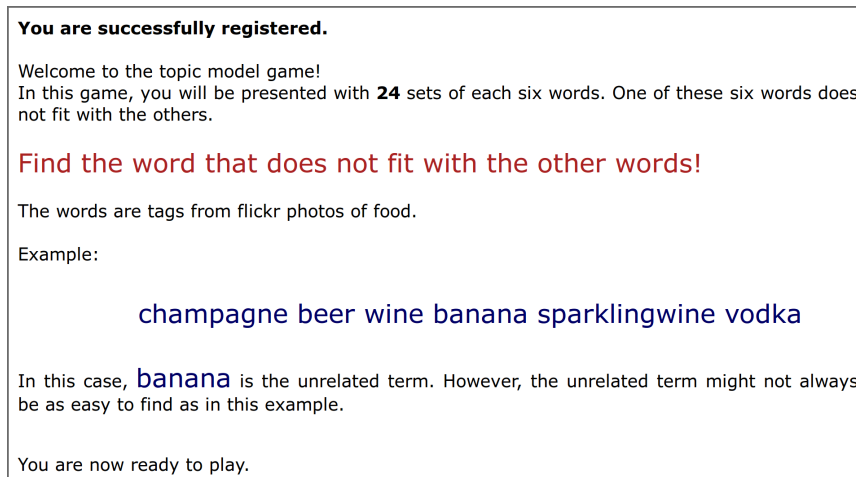
3.8.4 Effect of the region parameter

To further investigate the behaviour of MGTM for complex structured regions, the experiments are repeated for the three datasets with the smallest number of regions increasing the region parameter by a factor of ten. The results are given in Figure 3.7 (b), (d) and (f). For an increased number of regions, MGTM shows an improved perplexity for all three datasets and outperforms the basic and neighbour-aware model, demonstrating its ability to effectively exploit the adjacency relation between regions for sharing topic information.

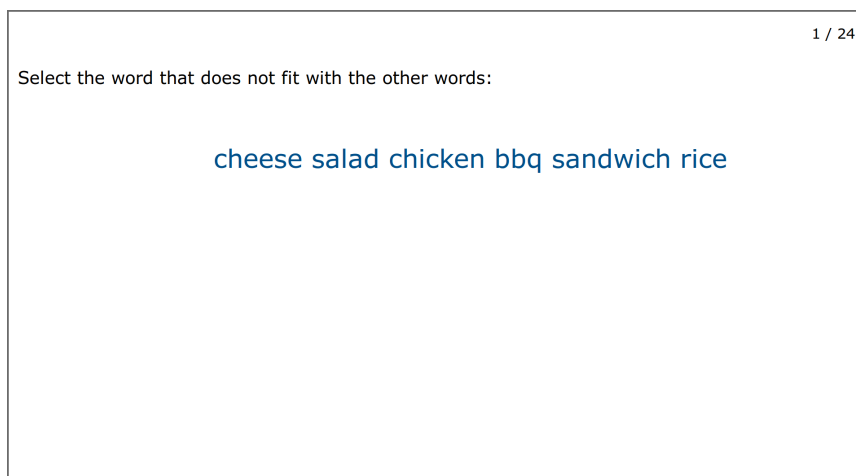
The effect of a growing number of regions for the car dataset at a fixed number of five topics is plotted in Figure 3.8. The car dataset is adequate to demonstrate the usage of geographic information in the topic models as the documents contain only a single word, meaning that intra-document co-occurrences of words do not contribute to the model creation. One can observe that LGTA, the basic model and MGTM improve the topic quality for increased region parameters, but the improvement of MGTM is considerably larger than for LGTA and the basic model. The dynamic smoothing based on the MDP helps to improve the topic quality by exchanging topic information between similar adjacent clusters. In contrast, the perplexity of NAM dramatically gets worse for larger region counts due to the instability of the naive smoothing mechanism based on model selection.

3.8.5 User study

A lower perplexity does not necessarily indicate an improved topic quality in terms of semantic coherence [CBGW⁺09]. Therefore a user study was



(a) Introduction of the topic model game



(b) Game for human topic evaluation

Figure 3.9: The topic model game. (a) Screenshot from the introduction of the topic model game implemented for the evaluation of MGTM which allows a human evaluation of the semantic coherence of the top-N words of topics. The game implements the evaluation proposed in [CBGW⁺09]. (b) The player is presented with six words, consisting of the top-5 words of a topic and one intruder word from a different topic. In order to gain points, the player has to select the intruder word. For high qualitative, semantically coherent topics, this task should be easier than for incoherent topics of low quality.

conducted to evaluate the semantic coherence of words within the topics detected by LGTA, the basic model, NAM and MGTM for the food dataset with 1000 regions at 4, 6 and 8 topics. Figure 3.5 shows the words with the highest probability for the topics detected by LGTA and MGTM at eight topics. Participants performed the “word intrusion” task introduced by Chang et al. [CBGW⁺09]: For evaluating a topic, users are presented with a set of six words, which consists of the five words with the highest probability under the topic and a word from another topic from the same model. The user’s task is to “find the word which does not fit with the other words”. In case of semantically coherent topic words, the intruder can be easily found. To additionally test the interaction between topics, the intruding word was chosen from a set of words which had a low probability (not in the top 25 words) in the evaluated topic and a high probability (top 5 of the remaining words) in another topic. The study was conducted with 31 users which were presented with word sets in a random order of models and topics. Only one word set per model-topic combination was shown to the user and a total of 1,446 of word sets were rated. Screenshots of the implementation of the intrusion detection game are shown in Figure 3.9.

In order to measure the quality of a model, the overall model precision is calculated (the percentage of intruders detected by participants) and the per-topic precisions within a given model.

Table 3.4 shows the average precision and the median of the per-topic precisions for all four models. Clearly, MGTM performs considerably better with both an average model precision and median model precision of around 0.8 compared to about 0.6 for LGTA. Only for the case of 4 topics, the neighbour-aware model shows a comparable precision. However, for 6 topics the precision is worse compared to LGTA and for 8 topics it is only slightly better. Similarly, the basic model is worse than LGTA for 4 topics and only slightly better for 6 and 8.

To analyse the distribution of the per-topic precision, the corresponding box-and-whisker plot for the case of 8 topics is given in Figure 3.10. Clearly, the quality of the topics detected by the basic model and NAM is mixed

Table 3.4: Model precision and median of per-topic precisions for LGTA, the basic model, NAM and MGTM on the food dataset with 1000 regions.

	4 topics		6 topics		8 topics	
	avg	median	avg	median	avg	median
LGTA	0.67	0.64	0.57	0.57	0.60	0.58
Basic	0.45	0.57	0.63	0.61	0.64	0.58
NAM	0.79	0.75	0.51	0.48	0.64	0.60
MGTM	0.79	0.80	0.82	0.81	0.78	0.75

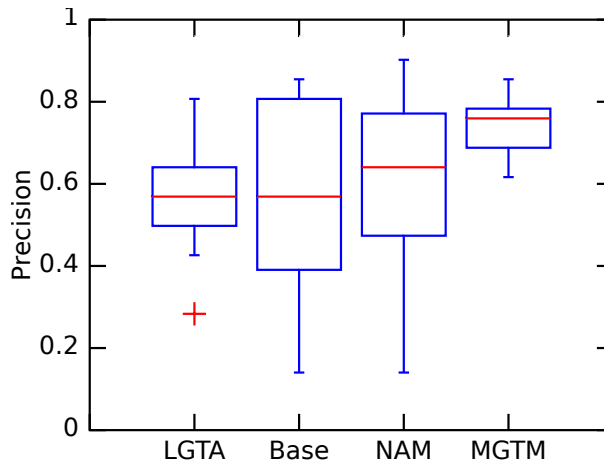


Figure 3.10: Results of the human evaluation. The Boxplots show the model precision for LGTA, the basic model, NAM and MGTM on the food dataset with 1000 regions, 8 topics. Higher is better.

– the per-topic precision ranges from very low values of about 0.2 to high values greater than 0.8. The precision of LGTA is more consistent, as the topic precision is closer distributed around the median with only one outlier. Finally, the per-topic precision of MGTM is high for all topics and most homogeneous among all models.

The human evaluation supports the findings of the perplexity comparison – indeed, MGTM detects semantically more coherent topic-word distributions by exploiting the spatial structure of topics using dynamic smoothing. The differences between LGTA and MGTM can be explained by taking the topics from Table 3.5 as an example.

Topic quality. The key difference between LGTA and MGTM is the semantic coherence of the topic-word distributions. Topic 2 of LGTA assigns high probabilities to the terms *chocolate*, *cheese*, *bread* and *fish*. By contrast, the most similar topic of MGTM contains the semantically related words *chocolate*, *icecream*, *strawberry* and *baking* – all related to desserts. The incoherent word-selection of LGTA is due to the fact that these tags often occur within a small region and repetitions of similar word combinations in adjacent regions are not sufficiently taken into account.

Globality. The first topic from LGTA and the corresponding topic from MGTM mostly contain terms related to seafood. Clearly, the words *rice* and *chicken* from the seafood topic of LGTA do not fit – they often occur in Asia, where many photos of seafood are located. The seafood topic from MGTM is more coherent – it assigns a high probability to the word *wine*, as it is often consumed together with fish across Europe. From this exam-

ple, one can see that the topics of LGTA are strongly influenced by local, region-specific patterns in tag co-occurrences whereas MGTM is more influenced by intra-document co-occurrences of tags and the global distribution of topics.

The reason is that LGTA does not model document-specific topic distributions. Instead, all documents within a region share the same topic distribution and therefore individual deviations from the regional topics are not recognised in the model.

In contrast, MGTM allows for document-specific topic distributions and permits deviations from the regional topic distribution. By detecting single documents fitting the topic of seafood in coastal regions all over the world, and by exchanging this topic information over the network of adjacent regions, a global topic of seafood is established.

Support for non-compact topics. Some of the topics detected in the food dataset are expected to exhibit a complex spatial distribution. As mentioned before, MGTM is able to detect such complex spatial structures. To give an example, the maps in Table 3.5 show the geographical distribution of documents with a higher-than-average probability for the seafood topic (Topic 1) as detected by LGTA and MGTM. One would expect Topic 1 to have a distribution along coastlines. This is the case for Topic 1 of both LGTA and MGTM, which covers both countries where fish is regularly eaten (such as the UK and the Netherlands) and countries where the seafood topic mainly appears at the coast (e.g. Spain, France). However, the geographical distribution of the seafood topic of LGTA has a large gap on the coast between Spain and France and is not detected in Denmark or on mainland Italy. The reason is that there are not many photos showing seafood in these areas and therefore the evidence is not sufficient for LGTA. Due to the dynamic smoothing of adjacent areas, MGTM still is able to detect such topics and thus correctly detects the seafood topic in documents along the whole coastline as seen on the map.

3.8.6 Runtime comparison

Another advantage of the HDP-based models is the separation of geographical clustering and the topic sampling step. By excluding the distance calculation between every document and every region centre from the slowly converging topic sampling process, the runtime is expected to be reduced significantly.

For comparing the runtime of the distinct methods, the implementation of LGTA provided by Yin was optimised before the runtime was measured on the largest dataset for different settings of the region parameter on a 2.8GHz CPU with 72GB of RAM using a single core. The topic parameter of LGTA is set to 7. The runtime in seconds is given in Figure 3.11. The runtime of LGTA linearly grows with a higher region count, as in every iteration every

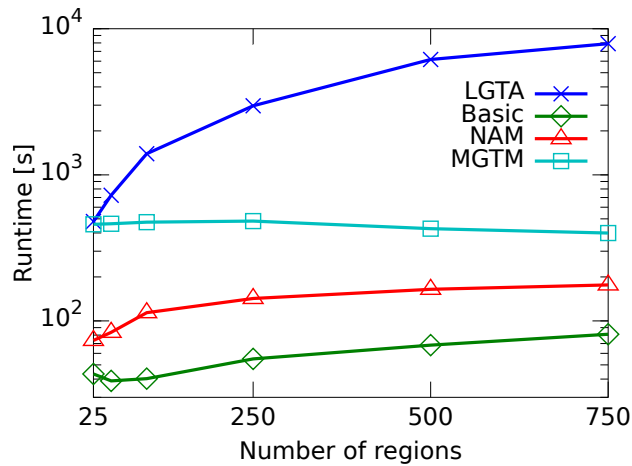


Figure 3.11: Average runtimes (in seconds) of LGTA, the basic model, NAM and MGTM for the food dataset. The runtime was measured for different numbers of regions. Note that the y-axis is a log scale.

document has to be compared with every region for sampling its membership probability. The Base model, NAM and MGTM use a separate geographical clustering step that can be efficiently implemented and takes only a fraction of the total runtime. The topic sampling is practically not influenced by the number of regions as it only creates additional region-specific topic distributions. MGTM shows a higher runtime compared to the basic model as the sampling of region-topic distributions in the multi-Dirichlet process is more expensive than in a normal hierarchical Dirichlet process. In return, for a larger number of regions, MGTM detects and merges topics with a coherent spatial distribution which results in a lower number of detected topics and a slightly decreased runtime. MGTM thus has a significantly reduced runtime and can be applied to much larger datasets.

Furthermore, it is straightforward to implement a distributed algorithm for MGTM as the distributed Gibbs sampling equations for hierarchical Dirichlet processes from [NASW09] can be directly applied to multi-Dirichlet processes and region-specific topic distributions can be shared across processors with dependent document-topic distributions using the same technique as for sharing topics across processors.

3.9 Summary

The results from the user study and extensive quantitative evaluation of MGTM show a clear improvement in topic quality compared to state-of-the-art methods in topic modelling. This means that **(i) MGTM detects**

topics of a higher quality as measured by per-word perplexity and by precision in user experiments. Additionally, the runtime analysis demonstrates that **(ii) modelling context with three-level multi-Dirichlet processes is highly efficient and thus suitable for large datasets.** After the preprocessing step, the runtime of the model is almost independent of the number of modelled geographical regions. **(iii) The presented model is the first to make use of adjacency relations between groups of documents for a dynamic smoothing of topic distributions.** Finally, the improved performance of MGTM at higher numbers of regions shows that **(iv) in real-world datasets, many geographical topics have a complex, non-Gaussian spatial distribution and that their detection can be supported.**

The presented topic model is just one example of how to use the MDP for dynamic smoothing. In the following chapter, a generalisation of MGTM and the underlying three-level hierarchical multi-Dirichlet process for multiple contexts is presented. Additionally, efficient online inference schemes based on variational inference are derived.

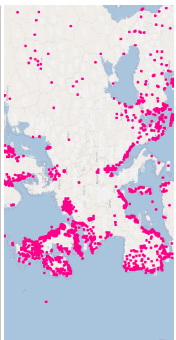
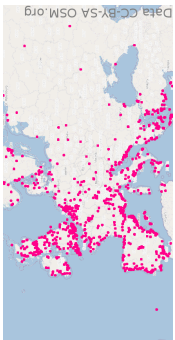
	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Map of Topic 1
LGTA	fish seafood rice shrimp crab lobster chicken	chocolate cheese bread fish wine tapas orange	japanese sushi ramen fish noodle sashimi noodles	vegetarian vegan chocolate baking bread cheese bacon	wine italian coffee french pizza chocolate bakery	chinese chicken noodles soup rice vietnamese dimsum	mexican bbq chicken burger sandwich fries hamburger	sushi thai korean japanese salmon rice tuna	
MGTM	seafood fish lobster shrimp crab wine salmon	chocolate icecream strawberry baking cream coffee pie	japanese sushi fish ramen sashimi rice salmon	salad cheese tomato bread chicken fish vegetarian	wine pizza coffee italian pasta cheese french	chinese thai chicken rice soup noodles korean	mexican tacos taco salsa burrito chicken chips	bbq burger fries hamburger grill chicken sandwich	

Table 3.5: Topic descriptions for the food dataset detected by LGTA and MGTM. The topics of MGTM were reordered to match the topics of LGTA. The maps show the positions of documents with an above average probability for Topic 1 as detected by LGTA (top) and MGTM (bottom).

Chapter 4

Multi-Context Topic Models

In this chapter, a novel class of topic models for multiple context variables is presented. The derived topic models can treat both discrete and continuous context data, and the latter can be linear, cyclic or spherical.

Having a large set of potentially important context variables, it is necessary to weight contexts by their impact on the learned topics to detect and remove irrelevant context variables and to exploit valuable context information [KS96]. The proposed class of topic models fulfils this requirement by learning a (potentially sparse) weighting of context-specific topic distributions in a three-level multi-Dirichlet process.

Including multiple context variables leads to increased memory consumption, a larger parameter space and therefore a slower convergence of the topic parameters during inference. To address this issue, an approximate inference procedure for three-level multi-Dirichlet processes is derived, extending and combining inference schemes for hierarchical Dirichlet processes.

4.1 Generalisations and Special Cases of MGTM

The Multi-Dirichlet process Geographical Topic Model (MGTM) presented in the previous chapter was applied to geographically distributed documents. In a preprocessing step, documents were clustered based on their geographical locations. Adjacent clusters were connected in a three-level Hierarchical Multi-Dirichlet Process (HMDP) where documents draw their topic distributions from a cluster-specific mixture of cluster-topic distributions. Cluster-topic distributions were drawn from a global Dirichlet process, ensuring that adjacent clusters exchange topic information and documents of different clusters share the same discrete set of topics.

In this section, the relation of three-level hierarchical multi-Dirichlet process models to existing topic models is shown in order to demonstrate the flexibility of the model structure of the HMDP and to show how it can be used for modelling different types of context variables.

The model structure behind the MGTM can be interpreted as a generalisation of several well-known topic models. In fact it is straightforward to show the equivalence of the HMDP model structure to several existing topic models under special parameter settings.

4.1.1 Relation to the hierarchical Dirichlet process

The novelty of the hierarchical multi-Dirichlet process lies in the introduction of multi-Dirichlet processes, which are employed to code a network of context clusters based on adjacency relations. If there is only one context variable and the context clusters are completely unconnected, the model reduces to a three-level hierarchical Dirichlet process [TJBB06], the structure behind the basic model presented in Section 3.5.

4.1.2 Relation to the author topic model

The model further simplifies in cases where not only all context clusters are unconnected, but documents consist of single words only. In this case, the table counts of documents in the Chinese restaurant franchise are identical to the topic counts (being 1 for the topic assigned to a document and 0 otherwise). This means that the words of documents within one cluster directly influence the topic distribution of the cluster – which is equivalent to merging all the documents in a cluster to form a single big document. The same effect can be achieved by letting α_1 (the scaling parameter of the second-level (multi-)Dirichlet process) go to infinity. This increases the probability of opening a new table for every topic-assignment and thus also leads to a de-facto merging of documents.

Therefore, the HMDP is identical to models which merge documents, such as the author-topic model by Rosen-Zvi et al. [RZGSS04], if all context clusters are unconnected and documents consist of single words only, or if α_1 is set to a very high value.

4.1.3 Relation to the citation influence model

A mixture of topic distributions was also used in the Citation Influence Model (CIM) by Dietz et al. [Die06] for modelling the influence of cited papers in a corpus of scientific publications. In contrast to the three-level MDP mixture model of the HMDP topic model, CIM directly mixes the topic distributions of cited papers and the citing paper to yield a document-topic distribution for a given paper. Cited papers in CIM are similar to the context clusters in MGTM, but while CIM mixes document-topic distributions, MGTM is mixing context-specific distributions to create a prior for the topic distributions of documents. However, if the scaling parameter α_1 is set to a large value, documents will directly be sampled from the

context-topic distribution. By creating a context cluster for each single document and connecting the context clusters of cited documents, one obtains a non-parametric version of the CIM. The influence of each cited paper then is governed by the multinomial parameters η . Additionally, the concentration parameter δ of the Dirichlet prior over the influence of context clusters (corresponding to citations) now can be learned using the generalised MLE estimator derived in Section 3.7.6.

The contextual focussed topic model [CZC12] is similar to the CIM, as it creates document-specific topic distributions by mixing the topic distribution of the author, the venue and a unique topic distribution of a document. Replacing the notion of authors and venues with cited papers, the CIM is obtained, and thus the HMDP can be used to model the influence of authors and venues on the content of documents.

4.1.4 A generalisation of the HMDP for arbitrary contexts

A hierarchical multi-Dirichlet process turns into a standard three-level hierarchical Dirichlet process if the context clusters are unconnected. Thus, the HMDP is also applicable for coding the influence of discrete context variables.

For modelling non-spherical, continuous contexts, documents can be clustered in the context space and adjacent clusters can be connected, e.g. based on adjacency on the timeline. Thus, the HMDP is a flexible model and can model a variety of different context variables.

In practice, real-world corpora from social media have multiple context variables, such as information from user profiles, temporal information or geographical information altogether. An ideal model would be able to treat multiple contexts simultaneously.

4.2 Existing Multi-Context Topic Models

In this section, existing topic models for modelling multiple context variables are briefly reviewed before it is shown how to model multiple context variables with the HMDP in the next section.

4.2.1 Dirichlet-multinomial regression

In the Dirichlet-multinomial regression topic model by [MM08], the influence of multiple context variables was modelled by learning a Dirichlet-multinomial regression to predict the document-topic distribution of a document given its context. The big advantage of this approach is that multiple different factors can be taken into account for modelling the dependence of topics on context, while keeping the model simple. A drawback is the limited

flexibility of the regression model – only discrete and non-spherical continuous variables can be treated, and only linear or similarly simple dependencies can be modelled.

4.2.2 Topic models for context-specific topics

There also exists a regression-based approach for multiple context variables which introduces context-specific variants of topics, i.e. given the context of a document, some topic-specific words will be more likely than in a global topic distribution [RB14]. The approach is somewhat similar to [AHS13], where context-specific topic hierarchies were detected. However, such models are not suitable for all applications: Context-specific topics can make it difficult to compare the topic distribution across documents, as local topics can deviate significantly from the global topics. This can lead to situations where two documents of different contexts have a high probability for the same topic, but the topic distributions of the context-specific interpretations of this topic are very different, meaning that the two documents are actually unrelated. Therefore, such models are limited to applications where such deviations are acceptable, and models for context-specific topics are not the scope of this thesis.

4.2.3 Distance-based topic models for multiple contexts

More complex approaches based on a transformation of context data, such as the model by Agovic and Banerjee [AB12] which is relying on a Gaussian process prior on the document-topic distributions, require the definition of a kernel function, which would have to be trained during inference, in order to model complex dependencies and structures in the context space (e.g. borders or coastlines in geographical data). This makes the model more complex and computationally expensive, as iterative distance calculations on real numbers would be required.

Distance-based approaches first define a distance function between documents and then learn topic distributions by mixing the topics of documents based on their distances [BF11]. These approaches require the iterative calculation of distances between documents, which is expensive. Additionally, they would require an iterative re-adjustment of the distance measure and cannot cope with sparse regions in the context space.

Therefore, a model which is both able to model complex structures in multiple context spaces and at the same time remains computationally tractable, even on big data sets, is required.

4.3 Hierarchical MDPs for Multiple Contexts

It is natural to extend hierarchical multi-Dirichlet processes for multiple contexts variables. In MGTM, a mixture of adjacent context clusters was introduced to construct a dependency of documents on both the topic distributions of their own cluster and on the topic distributions of adjacent clusters at the same time.

The same idea can be employed to include multiple contexts: Documents do not only depend on one context-specific topic distribution, but on a *multi-Dirichlet process mixture of multiple context distributions*.

To clarify the terms used in the description of the model, a definition of context spaces and context variables is given and the notion of *context groups* and *context clusters* is introduced.

Context spaces and context variables. Documents in social media typically are associated with metadata describing the context in which a document is created. A *context space* is a space to which documents are mapped based on their context information. A *context variable* stores the location of each document in the context space. In the multi-context case, there are multiple context spaces and context variables. Context variables are assumed to be independent.

For instance, for a corpus of geo-tagged documents, the unit-sphere could be a context space to which all documents are mapped based on their latitude and longitude. Each document then has an entry in a context variable which stores the coordinates of the documents on the unit sphere. If documents additionally have a timestamp, a second context space can be added, for instance the timeline. Alternatively, time and space can be combined to yield one spatio-temporal context space, preserving dependencies between time and space.

Context groups. The HMDP topic model for multiple contexts requires a preprocessing step, in which documents with similar context variables of one context space are grouped into *context groups*. Every document has multiple *context groups*, one for every context space. A *context group* links a set of documents which share the same set of parent *context-clusters* in the multi-Dirichlet process of the HMDP topic model.

For instance, in MGTM, all documents which were assigned to the same geographical cluster in the preprocessing step belong to the same *context group*.

Context clusters. *Context clusters* each have a distribution over the set of topics, and serve as base measures for the multi-Dirichlet process. The *context cluster* memberships are given by the *context group* of a document.

In the geographical setting, *context clusters* were associated with the index of a geographical cluster. The *context group* of documents defined that documents are drawn from the topic distributions of their own geographical cluster and from the topic distributions of adjacent geographical clusters.

For the multi-context HMDP, *context clusters* are more abstract concepts: *Context clusters* are associated with topic distributions which can be arbitrarily mixed to obtain the prior for the documents of a *context group*. The mixing proportions are group-specific. This further increases the flexibility of the model. However, for the applications of the HMDP topic model presented in this chapter, the usage of *context clusters* is simple. For discrete context variables, *context groups* and *context clusters* are directly related, so that a context group has one parent *context cluster* of the same index. For linear or cyclic context variables, the parents of a *context group* are the *context cluster* of the same index and its adjacent *context clusters*.

4.3.1 The HMDP Topic Model

Using the notion of *context groups* and *context clusters*, the generative process of the multi-Dirichlet process for arbitrary contexts is as follows:

1. A global topic distribution G_0 is drawn from a DP with a symmetric Dirichlet distribution H over the topic space as base measure:

$$G_0 \sim DP(\gamma, H) \quad H = Dir(\beta_0) \quad (4.1)$$

2. For every context space f of the F context spaces and every of the C_f context clusters of the context space, draw a topic distribution for each context cluster j :

$$G_j^c \sim DP(\alpha_0, G_0) \quad (4.2)$$

3. The strength of the influence of a context space on the document topics is stored in ζ , which is drawn from a Dirichlet distribution:

$$\zeta \sim Dir(\varepsilon) \quad (4.3)$$

4. Within a given context space, there exist *context clusters* and for a given *context group*, there exists a set of parent *context clusters* with indices P_{fg} from which the documents of that group generate their topic prior. The influence of each *context cluster* within a group is given by η_{fi} , a multinomial drawn from a symmetric Dirichlet distribution with parameter δ_f :

$$\eta_{fi} \sim Dir(\delta_f) \quad (4.4)$$

5. Document-specific topic distributions are sampled from a multi-Dirichlet process. For every *context space* f , the *context group* membership g_{mf} of document m determines from which parent *context clusters* the topic prior of a document is created. The number of parent *context clusters* for *context group* g in *context space* f is given by L_{fg} . The mixing parameter η_{fi} governs the influence of *context clusters* within a *context space*, and feature weights ζ weight the influence of each *context space* within the MDP.

For every document m , a document-specific topic distribution is drawn from a multi-Dirichlet process with *context clusters* as parent distributions, group-specific mixing proportions δ and *context space* weights η (the parent *context clusters* are known from the *context group* memberships \mathbf{g}_m of the document).

This can be denoted with a simplified parametrisation of the multi-Dirichlet process, where the parameter $\alpha_1 \zeta \eta$ is a short-hand notation for the vector of mixing proportions and the second parameter \mathbf{G}^c is

Table 4.1: Variables of the hierarchical multi-Dirichlet process topic model.

M	Number of documents
N_m	Number of words in document m
F	Number of context spaces
C_f	Number of <i>context clusters</i> for context space f
A_f	Number of <i>context groups</i> with common parents in context space f
β_0	Concentration parameter of the topic prior
H	Space of all possible topic multinomials, with Dirichlet prior β_0
G_0	Global measure on the topic space (includes topics and their weight)
G_{fj}^c	Measure on the topic space for cluster j of context space f
G_m^d	Document-specific measure on the topic space
γ	Scaling parameter of the global Dirichlet process (DP)
α_0	Scaling parameter for the <i>context cluster</i> DPs
α_1	Scaling parameter for the document MDPs
ϕ_{mn}	Topic drawn for w_{mn}
w_{mn}	Word n of document m
g_{mf}	<i>Context group</i> of document m in context space f
ζ	Mixing weights for the context spaces
η_{fi}	Mixing weights for group i in context space f
ε	Concentration parameter of the Dirichlet prior on ζ
δ_f	Concentration parameter of the Dirichlet prior on η in context space f

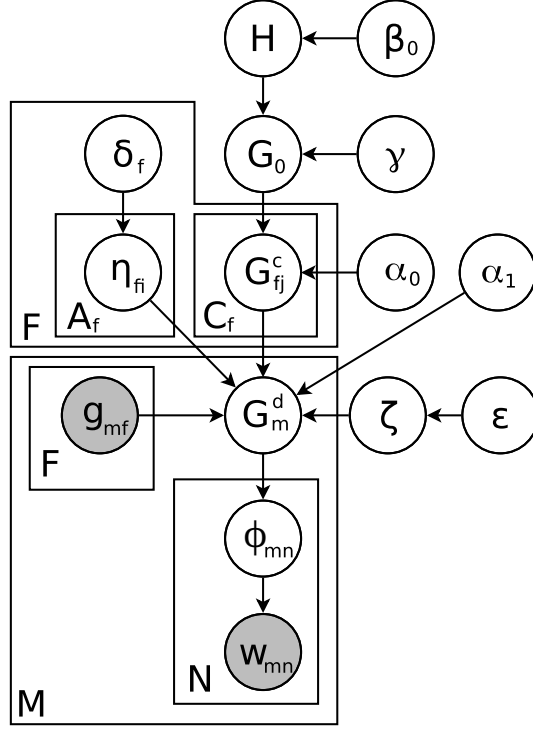


Figure 4.1: Graphical model of the **Hierarchical multi-Dirichlet Process (HMDP)** topic model for arbitrary contexts. In the HMDP topic model, document-topic distributions are sampled from a multi-Dirichlet process with several context-specific topic distributions of *context clusters* as base measures.

the matrix of cluster-specific base measures of parent clusters:

$$G_m^d \sim MDP(\alpha_1 \zeta \eta, G^c)$$

A schematic visualisation of the dependencies between the base measures in this hierarchical multi-Dirichlet process is given in Figure 4.2.

6. Finally, for every of the N_m words of document m , a topic-word multinomial ϕ_{mn} is drawn from G_m^d (the document-specific measure on the infinite topic space) and a word w_{mn} is drawn from this multinomial:

$$\phi_{mn} \sim G_m^d \quad w_{mn} \sim \phi_{mn} \quad (4.5)$$

The graphical model for the hierarchical multi-Dirichlet process is given in Figure 4.1. For every context space, the given *context group* g_{mf} of document m governs on which parent nodes G_m^d depends.

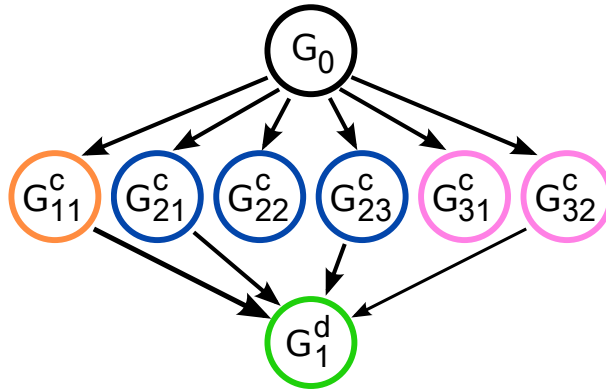


Figure 4.2: Schematic representation of the influence of the topic distributions of context clusters on a document-topic distribution.

On the top level, G_0 represents the global base measure on topics, from which for each *context cluster* a measure on the topic space is drawn in a Dirichlet process. Each context space (indicated by colour) has a set of *context clusters*. Typical context spaces include the timeline or the geographical space. Clusters can be obtained by applying appropriate clustering methods in the context space. Each context cluster has an own measure G^c on the infinite space of topics. For each document, a measure on the topic space G^d is obtained by drawing from a Dirichlet process with a mixture of the measures of the different contexts as base measure. The mixing proportions are given by multinomial parameters ζ , indicated by the thickness of arrows in the figure. For each context space f , the *context group* $i = g_{fm}$ determines from which of the *context clusters* the topic-prior is generated and the group-specific influence of each *context cluster* j is governed by a multinomial parameter η_{fij} . Groups of different context spaces are modelled as independent. Note that the superscripts do not correspond to indices or exponentiations, but are used to distinguish the topic distributions of the different levels of the hierarchical multi-Dirichlet process.

4.4 Efficient Inference

Inference based on Gibbs sampling has a relatively slow convergence rate compared to variational inference [AWST12]. The larger the parameter space, the more likely it is that the Gibbs sampler will take a long time for traversing the areas with a low likelihood. Additionally, a larger number of parameters might lead to local optima, in which the Gibbs sampler might get stuck.

For MGTM, the parameter space was already significantly larger compared to a standard two-level HDP, as there are additional counts required to estimate the topic distribution of regions, which (in case of the collapsed Gibbs sampler) depend on the choice of parent distributions stored in an

extra variable.

By modelling the influence of multiple context spaces on the topic distribution of documents, the parameter space grows substantially, and Gibbs sampling becomes more and more inefficient. For large datasets, the inference scheme should have the following characteristics:

Fast convergence rate. The per-word perplexity on held-out data should decrease (i.e. improve) quickly as the parameter space is traversed and areas of a high likelihood are reached.

Low memory consumption. Storing the *context group* assignments per word would result in an increased memory consumption, which is problematic for large datasets.

Online inference. Ideally, the model should be able to process streams of documents and to update the model parameters without requiring iterative runs on all documents of a potentially quickly growing dataset.

Feature selection. Treating a multitude of context information from rich document metadata as found in social media is a challenging problem. Some of the context information will be more valuable than others for modelling topics, and some context variables will have virtually no significant influence on topics. An ideal inference scheme should be able to select relevant context information and to exclude irrelevant context spaces during inference.

In the following, an efficient online variational inference for the HMDP with additional approximations is derived, which fulfils these requirements. The inference scheme is based on the Practical Collapsed Stochastic Variational Bayesian inference (PCSVB) for the two-level HDP proposed by Bleier [Ble13].

4.4.1 Collapsed variational Bayes for the HMDP

To develop a variational inference scheme, the inference strategy presented in [TKW08] is adapted and the stick-breaking representation of the MHDP is employed by introducing π_0, π_j^c and π_m^d , which are SBP distributed probabilities on topic indices corresponding to the topic weights in G_0, G_j^c and G_m^d , respectively. The indexed topic-word multinomials over the vocabulary are stored in vectors ϕ_k , where k is the index of the topic. For every word n in document m , a topic index z_{mn} is drawn from π_m^d and a word w_{mn} drawn

from $\phi_{z_{mn}}$. Formally:

$$\begin{aligned}
\pi_0 &\sim SBP(\gamma) \\
\zeta_f &\sim Dir(\varepsilon) \\
\eta_{fi} &\sim Dir(\delta_{fi}) \\
\pi_{fj}^c &\sim DP(\alpha_0, \pi_0) \\
\pi_m^d &\sim MDP(\alpha_1 \zeta \eta, \pi_{P_g}^c) \\
\phi_k &\sim Dir(\beta_0) \\
z_{mn} &\sim \pi_m^d \\
w_{mn} &\sim \phi_{z_{mn}}
\end{aligned}$$

where the notation of the MDP is simplified by using a short-hand notation for the topic distributions $\pi_{P_g}^c$ of the parent *context clusters* and for the group-specific mixing proportions $\alpha_1 \zeta \eta$. Figure 4.3 shows the graphical model of the stick-breaking construction and Table 4.2 gives an overview of the parameters used in the model.

The joint distribution of the HMDP model over observations and parameters then is given by:

$$\begin{aligned}
&p(\mathbf{w}, \mathbf{z}, \boldsymbol{\pi}^d, \boldsymbol{\pi}^c, \boldsymbol{\pi}_0, \boldsymbol{\zeta}, \boldsymbol{\eta}, \boldsymbol{\phi} \mid \beta_0, \gamma, \alpha_0, \alpha_1, \varepsilon, \boldsymbol{\delta}, \mathbf{g}) \\
&= \prod_{m=1}^M \prod_{n=1}^N \left[p(w_{mn} \mid \phi_k, z_{mn}) p(z_{mn} \mid \pi_m^d) \right] p(\pi_m^d \mid \alpha_1, \boldsymbol{\eta}, \boldsymbol{\zeta}, \boldsymbol{\pi}^c, \mathbf{g}_m) \\
&\quad \prod_{f=1}^F \left[p(\zeta_f \mid \varepsilon) \prod_{i=1}^{A_f} p(\eta_{fi} \mid \delta_f) \prod_{j=1}^{C_f} p(\pi_{fj}^c \mid \pi_0, \alpha_0) \right] p(\pi_0 \mid \gamma) \prod_{k=1}^{\infty} p(\phi_k \mid \beta_0).
\end{aligned} \tag{4.6}$$

Collapsed representation

To simplify inference and to yield a more exact variational approximation, the topic-word distributions and the document-topic distributions can be integrated out as in [TNW07, TKW08]. Additionally, the cluster-topic distributions of *context clusters* and the multinomial distributions over the mixing proportions ζ and η of *context clusters* in the MDP are integrated out. For permitting a variational approximation, the collapsed variational inference scheme introduced in Section 1.8.2 is applied. The number of topics used by the top-level Dirichlet process is truncated at K topics [BJ06b].

The resulting marginal distribution over words and topic assignments is

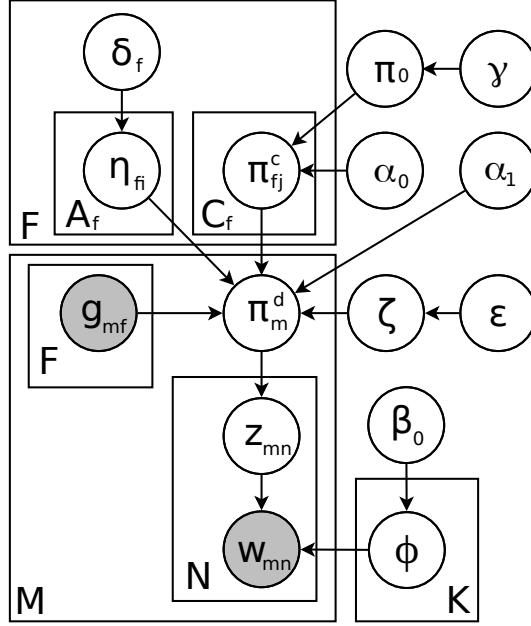


Figure 4.3: Stick-breaking construction of the hierarchical multi-Dirichlet process model for arbitrary contexts. A truncation at K topics is used in preparation of the variational inference scheme.

given by

$$\begin{aligned}
 & p(\mathbf{w}, \mathbf{z} \mid \beta_0, \boldsymbol{\pi}_0, \alpha_0, \alpha_1, \varepsilon, \boldsymbol{\delta}, \mathbf{g}) \\
 &= \prod_{m=0}^M \frac{\Gamma(\alpha_1)}{\Gamma(\alpha_1 + N_m)} \cdot \prod_{k=1}^K \frac{\Gamma(\alpha_1 \pi_{mk}^s + n_{m \cdot k})}{\Gamma(\alpha_1 \pi_{mk}^s)} \\
 & \cdot \prod_{k=1}^K \frac{\Gamma(V \cdot \beta_0)}{\Gamma(V \cdot \beta_0 + n_{k \cdot})} \cdot \prod_{t=1}^V \frac{\Gamma(\beta_0 + n_{kt})}{\Gamma(\beta_0)}
 \end{aligned} \tag{4.7}$$

where π_{mk}^s denotes the document-specific topic prior for topic k defined later in Eq. 4.9, which is a weighted mixture of the multinomial topic distributions of the *context clusters*. $n_{m \cdot k}$ are the topic counts of a document, indicating how often a topic was assigned to its words.

Practical approximation

The gamma functions make inference difficult. One way to simplify the equation is to introduce auxiliary variables corresponding to the tables in the Chinese restaurant process as shown in Equation 1.79. However, there exists an alternative approximation by Sato et al., which does not require the sampling of table counts: For small scaling parameters or short documents, it is unlikely to observe table counts larger than one [SKN12, Ble13].

The expected table counts then simplify to a binary variable indicating the presence of a topic, so that e.g. the table counts of a document reduce to $[n_{m..k} > 0]$ using Iverson brackets. Inference based on this simplification is called *practical inference* [SKN12]. In the case of the HMDP, the joint distribution further simplifies to:

$$\begin{aligned}
& p(\mathbf{w}, \mathbf{z} \mid \beta_0, \boldsymbol{\pi}_0, \alpha_0, \alpha_1, \varepsilon, \boldsymbol{\delta}, \gamma, \mathbf{g}) \\
& \approx \prod_{m=0}^M \frac{\Gamma(\alpha_1)}{\Gamma(\alpha_1 + N_m)} \cdot \prod_{k=1}^K (\Gamma(n_{m..k}) \alpha_1 \pi_{mk}^s)^{[n_{m..k} > 0]} \\
& \quad \cdot \prod_{k=1}^K \frac{\Gamma(V \cdot \beta_0)}{\Gamma(V \cdot \beta_0 + n_{k.})} \cdot \prod_{t=1}^V (\Gamma(n_{kt}) \beta_0)^{[n_{kt} > 0]}. \tag{4.8}
\end{aligned}$$

The expected table counts per topic then are obtained as the probability of seeing a topic in a given document.

As mentioned before, the document-topic-prior π_{mk}^s can be integrated out. Given observations, the topic prior of an individual document is a mixture of the estimated topic distributions π_{fj}^c of parent *context clusters*

Table 4.2: Variables of the SBP representation of the hierarchical multi-Dirichlet process topic model. The stick-breaking representation is employed to derive the variational inference scheme.

M	Number of documents
N_m	Number of words in document m
F	Number of context spaces
C_f	Number of <i>context clusters</i> in context space f
A_f	Number of <i>context groups</i> for context space f
g_{mf}	<i>Context group</i> ID of document m for context space f
π_0	Global topic distribution
π_{fj}^c	Topic distribution for cluster j of context space f
π_m^d	Document-specific topic distribution
ϕ_k	Topic-word distribution of topic k
β_0	Topic prior: Symmetric concentration parameter
γ	Scaling parameter of the global SBP
α_0	Scaling parameter for the <i>context cluster</i> DPs
α_1	Scaling parameter for the document MDPs
z_{mn}	Topic assignment of document m , word n
K	Truncation level of topics
ζ	Mixing weights for the context spaces
η_{fi}	Mixing weights for group i in context space f
ε	Concentration parameter of the Dirichlet prior on ζ
δ_f	Concentration parameter of the Dirichlet prior on η in context space f

from different *context spaces* with expected probability

$$\pi_m^s = \sum_{f=1}^F \sum_{j \in P_{fg'}} \frac{n'_{f..} + \varepsilon}{n'_{...} + F \cdot \varepsilon} \cdot \frac{n''_{fg'j} + \delta_f}{n''_{fg'.} + L_{fg'} \cdot \delta_f} \cdot \pi_{fj}^c \quad (4.9)$$

where mixing proportions ζ and η are integrated out and $g' = g_{mf}$ is a short-hand notation for the context-specific index of the *context group* of document m in *context space* f . L_{fg} gives the number of parent *context cluster* IDs for group g in context space f and P_{fg} is the set of parent *context cluster* IDs for group g in context space f . The variable $n'_{f..}$ denotes the number of topics drawn from clusters of *context space* f , $n'_{...}$ is the total number of topics drawn from all context clusters, $n''_{fg'j}$ is the number of topics drawn from the *context group* g' , and $n''_{fg'.$ is the total number of topics drawn from the *context clusters* belonging to the *context group* g' of the document in *context space* f .

The expected value of the probability of topic k in *context cluster* j of *context space* f can be approximated using the approximation for the table counts of documents:

$$\pi_{fjk}^c \approx \frac{\sum_{m \in G_{fj}} [n_{mfjk} > 0] + \alpha_0 \pi_{0k}}{\sum_{k=1}^K \sum_{m \in G_{fj}} [n_{mfjk} > 0] + \alpha_0} \quad (4.10)$$

where n_{mfjk} stores the counts of how often topic k from cluster j in context space f was assigned to a word in the document, G_{fj} is the set of document IDs which are drawing topics from cluster j of context space f , and $[n_{mfjk} > 0]$ is the practical estimate of the number of tables drawn from group j of context space f and topic k for document m . Table counts can then be calculated during inference by multiplying the table estimate $[n_{m..k} > 0]$ with the topic probabilities of the *context clusters* of *context space* f of document m .

Variational approximation

Two variables were not integrated out: The topic assignments \mathbf{z} and π_0 , the global topic prior. To break the dependencies between these variables, variational mean field approximation is employed, which factorises the probabilities:

$$q(\mathbf{z}, \pi_0) = \prod_{m=1}^M q(z_{mn}) \cdot q(\pi_0). \quad (4.11)$$

A practical variational lower bound on the marginal distribution (c.f. Eq. 1.41) then is given by

$$\begin{aligned} & \log p(\mathbf{w}, \mathbf{z} \mid \pi_0, \beta_0, \alpha_0, \alpha_1, \varepsilon, \delta, \gamma, \mathbf{g}) \\ & \geq \mathbb{E}_q[p(\mathbf{w}, \mathbf{z} \mid \beta_0, \alpha_0, \alpha_1, \varepsilon, \delta, \mathbf{g}, \pi_0) \cdot p(\pi_0 \mid \gamma)] - \mathbb{E}_q[\log q(\mathbf{z}, \pi_0)]. \end{aligned} \quad (4.12)$$

By taking the derivative with respect to $q(z_{mn} = k)$ and following the approximations by Asuncion [AWST12] and Sato [SKN12], given in Equation 1.80, the variational parameter for a single topic assignment can be obtained. Note that the inference additionally assigns a context and a cluster to every word from which the selected topic stems. However, the assignment can be made after learning the variational distribution over the topic assignment. Details on the inference are given in the appendix A.3. The resulting variational distribution over topic assignments is:

$$q(z_{mn} = k) \approx \frac{\mathbb{E}_q[n_{m..k}] + \alpha_1 \pi_{mk}^s}{N_m + \alpha_1} \frac{\mathbb{E}_q[n_{kt}] + \beta}{\mathbb{E}_q[n_{k.}] + V \cdot \beta} \quad (4.13)$$

where the counts ($n_{m..k}$, n_{kt} and n_k) are expected counts under the variational distribution and π_{mk}^s is the estimated topic weight from Eq. 4.9. Counts are estimated by summing over $q(z_{mn})$ for all the other words of the corpus as in the variational inference scheme for LDA given in Equation 1.46.

In the practical approximation, table counts are obtained by summing over topic assignment probabilities, which are calculated as the inverse of the probability of not seeing a topic in a document:

$$\mathbb{E}_q[n_{m..k} > 0] = 1 - \prod_{n=1}^{N_m} q(z_{mn} \neq k). \quad (4.14)$$

The estimated table counts of the documents are used to infer the topic distribution of *context clusters*. As every document has only one table per topic, the probability of a topic in a document is shared across clusters based on the topic weight in each cluster. For a given cluster j of context space f the table count n'_{fjk} for topic k is approximated as

$$\mathbb{E}_q[n'_{fjk}] = \sum_{m \in G_{fj}} \mathbb{E}_q[n_{mfjk} > 0] \approx \sum_{m \in G_{fj}} \frac{\mathbb{E}_q[n_{m..k} > 0] \cdot \mathbb{E}_q[\pi_{fjk}^c]}{\sum_{f'=1}^F \sum_{h \in P_{f'g_{mf}}} \mathbb{E}_q[\pi_{f'hk}^c]} \quad (4.15)$$

using the practical approximation, i.e. the table counts of a document are assumed to be binary per topic. G_{fj} is the set of indices of documents depending on cluster j in context space f , P_{fg} is the set of parent *context cluster* IDs for group g in context space f . For the global topic distribution, the estimate is obtained as in Eq. 1.81 using expected global table counts $m..k$ to derive stick lengths in a SBP representation. Because the number of *context clusters* is relatively small compared to the number of documents, and because clusters might contain thousands of tables of documents (using the Chinese restaurant process metaphor), a practical approximation of global table counts $m..k$ is not meaningful. Instead, the number of tables is sampled as in Eq. 1.71:

$$q(m_{fjk} = m) \propto s(n'_{fjk}, m) \cdot (\alpha_0 \cdot \pi_{0k})^m \quad (4.16)$$

where n'_{fjk} denotes the sums of document tables over all documents in cluster j of context f , and $s(n, m)$ again denotes the unsigned Stirling numbers of the first kind which account for the possible combinations and orderings in which tables might be occupied in the Chinese restaurant process. For estimating the global table count per topic, the sum over all table counts of a given topic is calculated as $m_{..k} = \sum_{f=1}^F \sum_{j=1}^{C_f} m_{fjk}$. Using the global table counts, the global topic distribution π_0 is obtained similar to Eq. 1.81 with:

$$\begin{aligned} q(\tilde{\pi}_{0k}) &= \text{Beta}(a_k, b_k) \propto \tilde{\pi}_{0k}^{a_k-1} \cdot (1 - \tilde{\pi}_{0k})^{b_k-1} \\ a_k &= 1 + \mathbb{E}_q[m_{..k}] \\ b_k &= \gamma + \sum_{l=k+1}^K \mathbb{E}_q[m_{..l}] \end{aligned} \quad (4.17)$$

which is used to obtain the global topic distribution as

$$\mathbb{E}_q[\pi_{0k}] = \mathbb{E}_q[\tilde{\pi}_{0k}] \prod_{l=1}^{l=k-1} (1 - \mathbb{E}_q[\tilde{\pi}_{0l}]). \quad (4.18)$$

4.4.2 Online inference

The inference scheme described so far requires a large memory to store variational parameters for the topic assignment of each single word in the corpus. Additionally, the corpus is assumed to be static, and it is not clear how to efficiently include novel documents during inference.

One solution to this problem is the practical collapsed stochastic variational inference scheme for the HDP given in [Ble13]. Stochastic inference consists in updating the global topic distribution of a topic model with parameters learned on mini-batches of documents using a decreasing learning rate [HBB10]. The learning rate ρ follows a function

$$\rho_t = \frac{s}{(\tau + t)^\kappa} \quad (4.19)$$

where t is the step during sampling, τ is an offset to prevent too big steps in the beginning and κ influences the slope of the learning curve.

The parameters of the model – i.e. the topic-word distribution ϕ_k of a given topic – can then be updated using

$$\phi_k = (1 - \rho_t) \cdot \phi_k + \rho_t \cdot \tilde{\phi}_k \quad (4.20)$$

where $\tilde{\phi}_k$ is the estimated parameter based on documents in a batch. A typical parameter setting is $s = 1$, $\tau = 64$ and $\kappa = 0.5$.

Foulds et al. [FBD⁺13] extended this idea by updating the counts of a collapsed variational inference scheme for LDA instead of the parameters.

Every batch is assumed to be representative of the whole corpus, and given that the corpus consists of C words, topic-word counts n_{kt} can be updated as

$$n_{kt} = (1 - \rho_t) \cdot n_{kt} + \rho_t \cdot \frac{C}{B} \cdot \tilde{n}_{kt} \quad (4.21)$$

where B is the batch size and \tilde{n}_{kt} denotes the estimated count of term t for topic k in the batch. In the HMDP, the same approach can be applied to update the topic and table counts per document, the topic and table counts per cluster, the table counts on the top-level Dirichlet process [Ble13], the table counts of all clusters of a context space and the table counts per *context cluster* in a *context group* as given in 4.15. For the *context groups*, a batch size of B^c is used, i.e. there have to be B^c documents seen in a *context group* before an update is performed on the counts associated with the *context clusters* belonging to that group.

The scaling parameters of the HMDP can be updated using a sample of documents and the equation from [SKN12]:

$$\alpha_1 = \frac{\sum_{m \in S} \sum_{k=1}^K \mathbb{E}_q [n_{m \cdot k} \geq 1]}{\sum_{m \in S} (\Psi(N_m + \alpha_1^{\text{old}}) - \Psi(\alpha_1^{\text{old}}))} \quad (4.22)$$

where S denotes the document indices of the sample and α_1 is the previous value of the scaling parameter. In practice, a subsample of documents will be sufficient for a precise estimate. For the experiments in the following sections, every document is included in the sample. An alternative estimate for the concentration parameter which does not require auxiliary variables (i.e. table counts) and yields more plausible results for larger values of α_1 is given in the appendix A.4. The stochastic update proposed in [Ble13] might be necessary if the sample size is small or documents are very heterogeneous. As the counts already are updated with a learning rate and the equation is a fixed-point iteration, it seems not necessary to employ an additional learning rate in the case of the HMDP.

Similarly, the scaling parameter for cluster-topic distributions can be estimated as

$$\alpha_0 = \frac{\sum_{f=1}^F \sum_{j=1}^{C_f} \sum_{k=1}^K \mathbb{E}_q [m_{fjk}]}{\sum_{f=1}^F \sum_{j=1}^{C_f} (\Psi(\sum_{k=1}^K \mathbb{E}_q [n'_{fjk}] + \alpha_0^{\text{old}}) - \Psi(\alpha_0^{\text{old}}))} \quad (4.23)$$

where m_{fjk} are the table counts from Eq. 4.16, n'_{fjk} is the sum of document-table counts for a given cluster and all clusters are included in the parameter estimate. And following [SKN12], the parameter of the topic Dirichlet prior can be estimated as

$$\beta_0 = \frac{\sum_{k=1}^K \sum_{t=1}^V \mathbb{E}_q [n_{kt} > 0]}{\sum_{k=1}^K (\Psi(\mathbb{E}_q [n_{k \cdot}] + \beta_0^{\text{old}}) - \Psi(\beta_0^{\text{old}}))} \quad (4.24)$$

which can be executed after seeing all documents (for a fixed corpus) or after seeing enough documents to yield a meaningful estimate.

The online inference scheme reduces the memory consumption of the inference algorithm dramatically. If the corpus contains M documents which contain C words, the memory used for the topic assignments and the parent assignments in the MDP is saved, which reduces the memory consumption by at least $C \times K$ values for the topic assignments and $M \times F$ values for the parent assignments.

Initialisation

In [BJ06a, TKW08, FBD⁺13], a random initialisation of parameters was suggested as a requirement for variational inference schemes. Indeed, variational inference needs an initial imbalance to distribute probability mass on the topic distributions in an uneven fashion. On the practical side, the Dirichlet process already introduces an imbalance by sampling the global topic distribution from a stick-breaking process. Therefore, variational inference schemes for hierarchical Dirichlet processes *do not require a random parameter initialisation* and also can find topics if parameters initially set to zero. This yields a deterministic inference procedure, which is desired e.g. for studies which require the reproducibility of results. The cost of a non-random initialisation is a slower convergence of the sampler, and thus a zero initialisation is deprecated for large datasets.

Implementation

The code of the practical collapsed stochastic variational Bayesian inference scheme for the hierarchical multi-Dirichlet process topic model is a completely new implementation and released as open source under GPL v3 licence¹.

Documents together with their *context group* memberships are given as input. Additionally, the parent *context clusters* of documents in a *context group* have to be specified. Finally, a truncation level K has to be chosen by the user.

Optional parameters include the batch size and the parameters τ , κ and s for the learning rate. A prior on the *context space* weights and a context-specific smoothing of the weights between connected context clusters can be specified. All hyper-parameters are automatically learned during inference.

As the definition of context clusters can be done using simple heuristics or standard data mining software, the application can be used without prior knowledge of topic models.

¹<https://github.com/ckling/hmdp>

4.5 Applications of the HMDP

In the following, the practical collapsed stochastic variational HMDP topic model is applied on different datasets to demonstrate its ability to model a variety of different context information.

The HMDP has unique properties. Most importantly, it allows to simultaneously model the influence of *multiple, arbitrary context information*, to *weight the importance of the context information* and to *exclude irrelevant context information*. In contrast to related topic models [MM08, AB12], the HMDP topic model does not require a function which links context spaces to topics, and thus can detect *arbitrary non-linear patterns in the context space, including cyclic and spherical patterns*. Additionally, the weights of different context spaces have a *natural interpretation as probabilities*, unlike in e.g. kernel- or regression-based models.

The special abilities of the HMDP and the generality of the model are demonstrated on three very different corpora: First, the impact of practical collapsed stochastic variational Bayesian inference on performance and topic quality is examined on the largest dataset from Section 3.8.1, the geographically distributed and tagged photos of food. The second dataset is a set of tagged user profiles from a social network of sexual fetishists, with context information from users profiles. The third corpus consists of messages of the Linux kernel mailing list, observed over a period of almost twenty years. Context information involves all aspects of time, such as daily, weekly and yearly cycles – which for the first time can be explicitly modelled in a topic model using the HMDP.

4.5.1 Improved inference

The hierarchical multi-Dirichlet process (HMDP) for multiple context variables is a generalisation of the multi-Dirichlet geographical topic model (MGTM). As such, the Gibbs sampler for MGTM can be employed to evaluate the impact of the practical collapsed stochastic variational Bayesian inference (PCSVB) derived in the previous section.

Differences between PCSVB and Gibbs sampling

For a single context space, in which documents are geographically clustered and where adjacent clusters are dependent via context groups, the HMDP topic model is identical to MGTM.

Still, the parameters learned during inference are expected to differ due to the different inference technique. The original inference scheme of MGTM is a Gibbs sampler. In theory, the Gibbs sampler would converge to a parameter estimate which optimises the exact posterior of the model. In practice, Gibbs samplers for models with many parameters are prone to get stuck in local optima [LJT07].

The PCSVB inference scheme presented with the HMDP topic model involves several approximations:

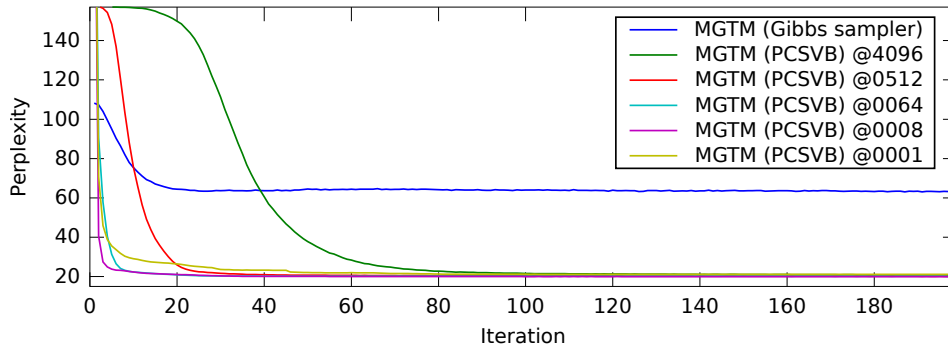
- **Variational mean field approximation.** Instead of maximising the posterior, a lower bound is maximised, which assumes independence between the inferred variables.
- **Approximation of expected counts.** The variance of the expected counts is neglected.
- **Practical approximation.** Table counts in the Chinese restaurant franchise are approximated as binary counts.
- **Stochastic inference.** Instead of directly updating the parameters during inference, observations are processed in mini-batches. Parameters are updated with a learning rate instead of exactly updating the counts, assuming that all documents in the corpus behave like the observations in the batch.

To summarise, the Gibbs sampler in theory maximises the exact posterior, but the inferred parameters could be a local optimum, while the PCSVB inference traverses very fast through the parameter space to find an optimum, but it is unclear if the update direction is reasonable, due to the various approximations.

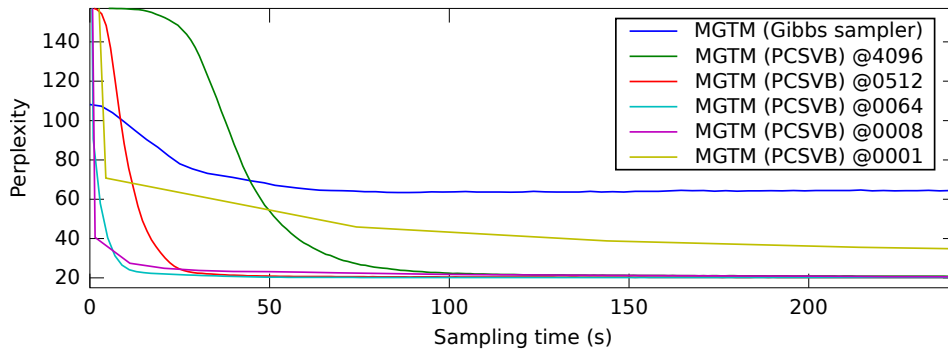
Convergence and perplexity

A standard way of comparing the performance of two different sampling strategies is to calculate the per-word perplexity of the learned models. For the comparison, the largest dataset from the evaluation of MGTm from Section 3.8.1 is used: The food dataset, consisting of 151,747 tagged photographs of food, where the location the photo was taken is given as GPS coordinates in the metadata. The number of distinct tags is very low with 278 tags, allowing for a potentially very precise prediction. As in the evaluation of MGTm, the topics and the region-specific priors are given as fixed parameters and the document-specific distribution is sampled before the perplexity is calculated.

Figure 4.4(a) shows the perplexity after a given number of sampling iterations, for the Gibbs sampler and PCSVB at changing batch sizes (groups and documents share the same batch size). The number of truncated topics for PCSVB was set to 8 to match the number of topics used by the Gibbs sampler. The parameter for the learning rate κ was set to 0.5 with $\tau = 64$ as in [HBB10]. Interestingly, PCSVB yields a dramatically improved perplexity compared to the Gibbs sampler. Regardless of the batch size, the perplexity converges to a value of about 20, an extremely low value. For a small batch size of 8, PCSVB shows the fastest convergence. Small batch sizes



(a) Perplexity of MGTM versus sampling iteration for various inference settings



(b) Perplexity of MGTM versus computing time for various inference settings

Figure 4.4: Comparison of convergence rates measured by per-word perplexity for the Gibbs sampler and practical collapsed stochastic variational Bayes (PCSVB) for the HMDP topic model on the food dataset (see Section 3.8.1). The Gibbs sampler detected 8 topics, so PCSVB was set to a truncation level of 8. The perplexity is calculated after a full iteration of the sampler and plotted against (a) the number of sampling iterations and (b) the sampling time. PCSVB yields a drastically better prediction than the Gibbs sampler. The optimal perplexity is obtained at a batch size of 8 with a perplexity of 19.7. For a batch size of 64, the sampler converges earliest in time at a perplexity of 20.2. This batch size thus is to be preferred for settings where resources are limited. After convergence, the differences in perplexity for varying batch sizes are marginal. Very small batch sizes lead to a slow convergence rate given the computing time.

also mean higher computation costs, as the counts have to be updated after every batch. Therefore, the inference scheme was executed on a single core of a 2.6 GHz processor and the sampling time was recorded. The perplexity for a given sampling time, plotted in Figure 4.4(b), shows that a batch size of 64 yields the earliest convergence in time. Since the perplexity converges to

very good values for all batch sizes, a batch size of 64 would be the preferred setting.

Topic quality

To better understand the phenomenon of the perplexity improvement, the topics detected by the Gibbs sampler and PCSVB are shown in Table 4.3. Topics detected by PCSVB were reordered to match the topics detected by the Gibbs sampler if possible. *Topic 1* of the Gibbs sampler is about seafood, while PCSVB is mixing seafood with vegan and vegetarian food, which yields an inconsistent topic. Similarly, all the topics detected by PCSVB are mixtures of several underlying topics. Therefore the quality of the topics detected by PCSVB would be low if semantic coherence would be the measure of topic quality (e.g. in a human evaluation).

However, this behaviour is optimal for optimising the predictive quality of the model. PCSVB is able to identify *more* than just eight topics. In order to maximise the posterior, the words of the different topics are distributed over the available eight topic slots.

One could ask why the *very same model* produces different topic estimates. The answer lies in the way topics are created. As mentioned in Section 3.6.2, the Gibbs sampler resembles the evolutionary process behind the creation of memes [Daw06]. All documents start with one topic, which is supposed to explain all the document words. Within a document, a novel way of explaining the observations (the words) in the document can be created during inference, resulting in a new topic. All the documents in the region of this document and the adjacent regions get to know the topic via a shared topic prior. Additionally, with a small probability, the topic is offered to all documents in the corpus via a global topic prior. If the topic is a better explanation for the observed words and documents, it will eventually be taken up by a large enough number of documents to be sustained; otherwise, it will perish. Gibbs sampling therefore starts with one topic and slowly creates new topics during inference which have to occupy sufficient resources in order to persist. The newly created topics are likely to be pure if the probability for creating a new topic is reasonably high. Words which cannot be explained by existing topics will typically create new topics instead of being mixed into existing topics.

In contrast, PCSVB starts with a total mixture of K topics – every word and every document has the same probability for all topics. During inference, small imbalances in the global topic distribution lead to non-uniform topic word distributions. The latter then lead to imbalances in the document-topic distributions and the cluster-topic distributions – which due to the sparse Dirichlet priors strengthen the imbalances of topic-word distributions. This feedback process is repeated until convergence. Because topics are mixtures from the very beginning, it is likely that a topic is a

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8
Gibbs sampler	seafood	chocolate	japanese	salad	wine	chinese	mexican	bbq
	fish	icecream	sushi	cheese	pizza	thai	tacos	burger
	lobster	strawberry	fish	tomato	coffee	chicken	taco	fries
	shrimp	baking	ramen	bread	italian	rice	salsa	hamburger
	crab	cream	sashimi	chicken	pasta	soup	burrito	grill
	wine	coffee	rice	fish	cheese	noodles	chicken	chicken
	salmon	pie	salmon	vegetarian	french	korean	chips	sandwich
	fish	chocolate	japanese	chicken	cheese	orange	italian	burger
PCSVB	seafood	mexican	chinese	salad	coffee	pasta	bbq	french
	vegetarian	cream	sushi	soup	wine	tea	sandwich	fries
	vegan	strawberry	rice	tomato	pizza	steak	icecream	hamburger
	shrimp	baking	thai	pork	bread	chips	bacon	bakery
	indian	pie	noodles	beef	love	potatoes	grill	pastry
	salmon	lemon	curry	potato	pancakes	beans	sausage	tapas

Table 4.3: Topic descriptions for the food dataset detected by MGTm using a Gibbs sampler and practical collapsed stochastic variational Bayesian inference (PCSVB). The topics detected using PCSVB were reordered to match the topics detected by the Gibbs sampler. Though the perplexity of the model learned by PCSVB is significantly better than the model from the Gibbs sampler (cf. Figure 4.4), the topics detected with PCSVB seem semantically incoherent. The Gibbs sampler is better in detecting a low number of topics and automatically infers the number of topics during sampling. For a too-small number of topics, PCSVB mixes different semantic topics (e.g. vegan and seafood) of a dataset, which maximises the posterior but leads to rather meaningless topics.

mixture of multiple latent groups of co-occurring words.

To test this behaviour, the number of truncated topics is increased to $K = 25$ and topics are re-calculated. Table 4.4 shows the detected topics. Looking at *Topic 24*, PCSVB now is able to correctly detect the seafood topic. Additionally, the cuisines of ten different countries are separated from each other: there is each a topic for the Chinese (1), Spanish (4), Korean (6), French (13), Japanese (14), Thai (15), Italian (16), Indian (17), Vietnamese (19) and Mexican (20) cuisine.

A comparison of the semantic coherence of topics between the different sampling schemes is impossible: The Gibbs sampler is non-parametric, and it is very unlikely to detect 25 topics in the food dataset. Only for parametric models such as LGTA it is possible to explicitly set the number of topics. In Table A.1 of the appendix, the food topics detected by latent geographical topic analysis (LGTA) [YCH⁺11b] for 25 topics are given to show that MGTM is superior in detecting coherent complex-shaped regions, even if the number of topics is increased and the inference scheme is switched to PCSVB. In contrast to MGTM, LGTA has problems to detect coherent geographical topics if the underlying distribution is non-Gaussian, which can be seen by looking at the topics for the Italian, Japanese and Indian cuisines, which are split into several distinct topics.

Discussion of results on the food dataset

With the detected topics on the food dataset, it seems that there is not *one* sampling scheme which works best. Both the Gibbs sampler and variational inference have strengths and weaknesses. The Gibbs sampler clearly is to be preferred if the number of topics should be automatically detected and topics are required to be semantically coherent. This is a desired property in applications for creating descriptions of datasets. Users have to be aware of the fact that the number of topics is potentially underestimated by the Gibbs sampler.

The number of topics can be inferred by PCSVB, as unused topics have a small probability in the global topic prior π_0 . However, for a low truncation level, all available topics under the truncation level K are used, because they improve the posterior of the model [AWST12]. Therefore, for a low K , PCSVB yields a parametric model, similar to a multi-level version of asymmetric LDA [WMM09] with asymmetric document-topic priors.

PCSVB yields semantically coherent topics only if the truncation level is correctly adjusted, for instance by conducting a human evaluation of topic quality. For a too-low truncation level, the topics are superior in word prediction, making PCSVB the inference of choice for prediction problems. If a high number of topics is desired, PCSVB potentially is able to detect a higher number of semantically coherent topics than the Gibbs sampler.

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
chinese	vegan	rice	tapas	salad
dimsum	vegetarian	chicken	spanish	soup
duck	tofu	fish	paella	pasta
noodles		pork	pescado	potato
hotpot		shrimp	olives	bread
chopsticks		beef	octopus	salmon
tofu		mango	tortilla	mushroom
Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
korean	ramen	bbq	hamburger	baking
noodles	noodle	barbecue	hotdog	butter
soup	curry	barbeque	cheeseburger	berries
beef	soba	grill	burgers	strawberries
noodle	udon		hotdogs	vanilla
tofu	fish		deli	peppers
pork	noodles		burger	cinnamon
Topic 11	Topic 12	Topic 13	Topic 14	Topic 15
bacon	orange	french	japanese	thai
sausage	grill	bistro	sushi	stickyrice
burger	chocolate	pastry	sashimi	padthai
pork	icecream	resto	tofu	curry
steak	love	patisserie	tuna	
chips	steak	croissant	miso	
beef	bread	tart	chopsticks	
Topic 16	Topic 17	Topic 18	Topic 19	Topic 20
italian	fish	coffee	sushi	mexican
pizza	tea	bakery	vietnamese	tacos
pasta	indian	cream	roll	taco
pizzeria	bread	cocktail	pho	salsa
toscana	curry	latte	salmon	margarita
spaghetti	chips	pancake	wasabi	guacamole
gelato	coffee	butter	cocktails	cajun
Topic 21	Topic 22	Topic 23	Topic 24	Topic 25
sandwich	chocolate	cheese	seafood	wine
pizza	cookie	tomato	fish	cheese
fries	bento	onion	lobster	chocolate
chicken	strawberry	corn	crab	bread
cheese	cheese	garlic	shrimp	fish
burger	tomato	lettuce	oyster	orange
pancakes	peach	chili	crabs	pizza

Table 4.4: Topics detected by the PCSVB inference for MGTm on the food dataset. Parameters are set to 25 topics and 1000 regions at a batch size of 8 both for groups and documents. Without any prior information, local topics for the Chinese (1), Spanish (4), Korean (6), French (13), Japanese (14), Thai (15), Italian (16), Indian (17), Vietnamese (19) and Mexican (20) cuisine are detected. Words with a probability < 0.01 are not displayed.

4.5.2 Modelling user profiles of a social network for fetishists

So far, the inference technique of the HMDP topic model was evaluated against a Gibbs sampler. It could be shown that PCSVB is able to detect a high number of semantically coherent topics. The core of the HMDP model however is the ability to include multiple context information and to weight them. Due to the intuitive structure of the model, *all the parameters of the HMDP have a natural interpretation as weights and pseudo-counts*. This makes the HMDP a valuable method for understanding the influence of context information on topics.

To demonstrate the ability of HMDP to weight context information, a special dataset is employed: Tagged user profiles of FetLife, the largest online social network for sexual fetishists.

Dataset description

On the online fetish platform, users create profiles where they state their *gender* (there are eleven different genders, e.g. transvestites, trans-males etc.), *age*, *relationship status*, if their relationship is a *lifetime relationship*, their *sexual orientation*, *sexual role* (e.g. master or slave) and the information if *fetish events* are attended. A detailed description of the demographic context is given in [FHS⁺15]. Additionally to the demographic variables – which in the following will be interpreted as the context information – users describe their fetishes with freely chosen tags (which might be copied from other users). The tags allow the fetishists to be found by other users with similar interests, and thus many users create a rich profile description.

The dataset used for this study consists of 126,408 tagged user profiles crawled in 2013. 2,140 tags are used more than 100 times and were included in the analysis. The gender and age distributions are shown in Figure 4.5. 70% of the users are male, 24% are female and 6% chose a different gender. The age distribution is heavily skewed towards young users, most fetishists on the website are in their twenties or thirties. It seems that some users do not want to reveal their real age and leave the pre-defined year of birth (1977) in the registration form unchanged, resulting in a peak at 36 years. Other users chose the lowest option, 1920, which yields a second peak at 93 years. Further minor peaks at birth years of 1980, 1970 and 1960 indicate users who do not want to state their precise age.

Detected topics

For topic detection, the categorial context variables were directly translated into clusters. In order to code the continuous age variable, every age (from 18 to 93) was coded as a single cluster and adjacent age clusters were connected to share topic information across clusters. Additionally, a cluster for missing values was created for every context variable. PCSVB for the HMDP then

Topic 1 slave master blindfolds oral sex talking dirty whips	Topic 2 mind fucks crying fear interrogation wrestling sadomasochism	Topic 3 spreader bars orgasm denial begging blindfolds ben wa balls eye contact r... ¹	Topic 4 anal training anal stretching anal beads anal hooks strap ass play
Topic 5 creampie lactation breastfeeding breeding impregnation ... ³ incest play	Topic 6 nipple torture bondage leather whips high heels humiliation	Topic 7 photography art erotica bondage art writing erotica tantra hot oil massages	Topic 8 making home m... ² gangbangs pain dildos exhibitionism talking dirty
Topic 9 feminization sissification sissy training forced femini... ⁷ dollification cross dressing	Topic 10 anonymous enc... ⁴ group sex outdoor sex sex with stra... ⁸ sex in public public play	Topic 11 bbw bbw bondage forced mastur... ⁶ forced nudity forced submission nipple play	Topic 12 cock and ball... ⁵ candle wax electrotorture pain masochism humiliation
Topic 13 kissing fingering light bondage caressing handjobs blow jobs	Topic 14 exhibitionism erotic photog... ⁹ piercings voyeurism candle wax nipple torture	Topic 15 sexual slavery degradation public humili... ¹⁰ slavery total power e... ¹¹ objectification	Topic 16 orgasm control orgasm denial teasing forced orgasms obedience tra... ¹² edge play
Topic 17 caning belt spanking whipping corporal puni... ¹⁵ belt whippings padding	Topic 18 double penetr... ¹³ fucking machines vaginal stret... ¹⁴ fisting triple penetr... ¹⁶ pussy pumping	Topic 19 ass worship pussy worship body worship facesitting queening oral servitude	Topic 20 cuckold forced deepthroat strapon female domination deepthroat forced bi

Table 4.5: Description of the first 20 topics detected in the fetish community dataset. The full topic description can be found in Table A.2 of the Appendix. ¹eye contact restrictions ²making home movies ³impregnation fantasy ⁴anonymous encounters ⁵cock and ball torture ⁶forced masturbation ⁷forced feminization ⁸sex with strangers ⁹erotic photography ¹⁰public humiliation ¹¹total power exchange ¹²obedience training ¹³double penetration ¹⁴vaginal stretching ¹⁵corporal punishment ¹⁶triple penetration

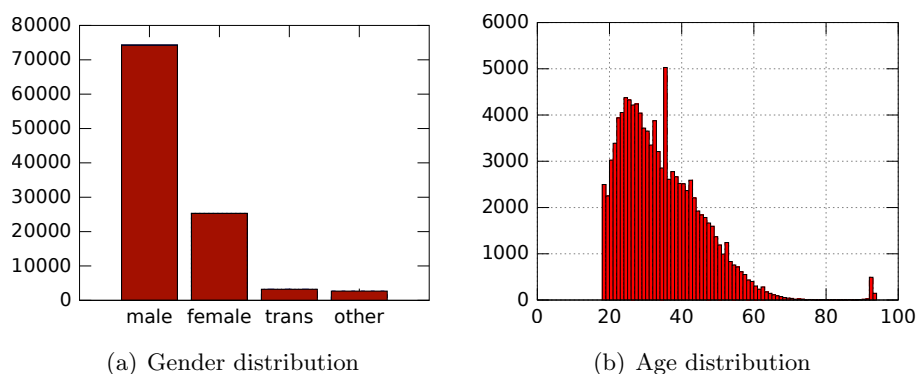


Figure 4.5: Age and gender distribution in the fetish dataset. (a) About 24% of the users identify themselves with a female gender, 70% stated a male gender, 3% identified as transvestites and 3% stated other gender information (there are eleven options in total). (b) The age distribution shows that users are at least 18 years old. It peaks at 25, and shows a second peak at 36 years, which is the predefined age on the account creation web page. Some members do not want to reveal their age and thus leave the predefined age unchanged or pick the lowest available year of birth (1920), which was chosen by about 700 users and results in a peak at 90+ years.

was run with a truncation level of 25, 50 and 100, at a batch size of 8 and a learning rate parameter of $\kappa = 0.5$. Even at a high number of topics, the topics seem semantically coherent, i.e. synonyms are grouped together. The first 20 topics for $K = 100$ are shown in Table 4.5, and the full 100 topics are given in Table A.2 of the Appendix. It is hard to rate the quality of the detected topics without a reference. Therefore, the topics detected by standard LDA with $K = 100$ and identical parameters as learned by the HMDP during inference (i.e. the topic-word and the document-topic prior) are given in Table A.9. LDA has less sparse topics (though the priors were identical), which is visualised by greying out words with a probability below one percent. Additionally, very long tags (consisting of several words) have a higher probability in LDA.

However, even though LDA and HMDP detect different topics, the top words of the topics seem semantically coherent for both and it is left to future work to evaluate the topic quality of both models in a user study.

Context weighting

The unique feature of HMDP is the weighting of the context spaces. During inference, the mixing proportions ζ for weighting the influence of different context spaces in the multi-Dirichlet process are constantly updated. If a the topic distributions of the clusters of a context space predict the topics

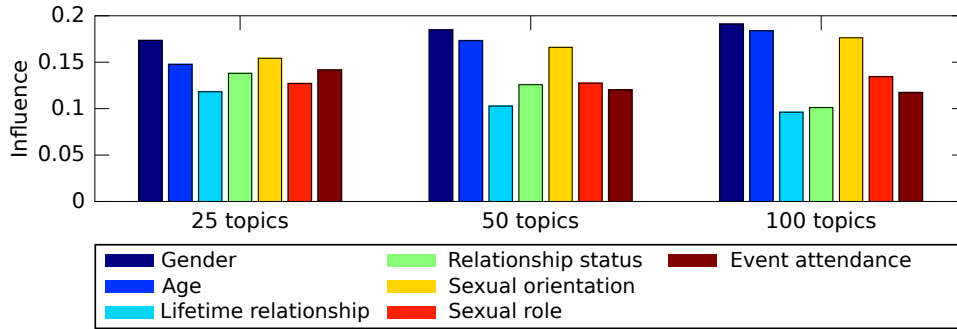


Figure 4.6: Weighting of context spaces (corresponding to demographic information) for the fetish dataset at changing numbers of topics. The ordering is stable for changing numbers of topics, but a higher number of topics allows for a better discrimination between the influence of demographic variables on the observed topics.

of documents well, the weight of that context space will be raised.

Every context weight in the HMDP can be directly interpreted as the probability that a document is explained by the priors of the clusters in a context space. Context-specific priors are weighted relative to each other, i.e. the weight of the priors of context clusters in a context space is decreased if there exists a different context space which predicts topics better. The weight of a context space also is relative in a sense that combinations of context variables are taken into account. If two context spaces together explain the topic distribution of documents well, the context spaces of both variables will receive higher weights. It thus might be that a single context space alone does not have a high predictive quality, despite having a high weight in the HMDP model. A context space which does not contribute to the topics of a document – or which is redundant given the other context spaces – can even be completely explained away [KS96]. Unlike in regression models, it is not necessary to check for high correlations between context variables, as the model would automatically down-rate one of the context spaces of the context variables, using minimal imbalances in the initial setting of the inference algorithm.

For the fetish dataset, there are seven demographic variables. Figure 4.6 shows their learned weight ζ of context spaces (corresponding to demographic information) in the HMDP for a truncation level of 25, 50 and 100 topics. One can see that gender, age and sexual orientation are the most important variables for predicting the fetishes of a user. This observation is stable for different truncation levels. The sexual role and the participation in events are the next important factors according to HMDP with 100 topics. However, for 50 and 25 topics, this observation does not hold. The reason why the number of topics affects the weights of context spaces is a

mutual dependence between context space weights and topics. For a low number of topics, the HMDP topic model will be unable to detect some of the context-dependent fetish topics (e.g. topics which are more popular for younger fetishists). Thus the weighting of context spaces is complicated for a low number of topics (with the extreme case of one topic, where no statement about the importance of context variables can be made). Raising the truncation level will result in a better analysis on the importance of context spaces.

However, for a too-high number of topics, the detected topics would no longer be semantically coherent, making an interpretation of the model parameters meaningless. The number of topics is automatically determined by the hierarchical multi-Dirichlet process. To make sure that the MDP detected the right number of topics, the maximum number of meaningful topics can be empirically evaluated before statements on the importance of context variables are made.

Note that the weighting of context variables cannot be achieved by correlating context information with topics learned by a standard topic model, as there is a mutual dependence between context and topics: If the topics were different, the context space weighting would be different, too. And if the context weighting was different, the topic prior of each document would change, effectively changing the topic assignments and the detected topics. One example for this behaviour would be the car dataset used to evaluate the MGTM topic model in Section 3.8.1: Because documents in this dataset contain only a single word, a normal topic model would not be able to detect any topics. If the dataset would contain timestamps for the photos, a weighting of the impact of temporal information and geographical information would only be possible if meaningful topics were detected first, which requires the use of geographical and temporal information.

The HMDP is the first topic model which allows the integration and human-interpretable weighting of multiple arbitrary context variables at the same time. Because interactions between context variables are taken into account, the weighting of context spaces differs from the ranking of variables e.g. by forward selection techniques [KS96] which have the advantage of being universally applicable at the cost of failing to detect complex dependencies between variables.

In order to demonstrate this behaviour, one could use the perplexity as quality measure and select the single context space which optimises the perplexity of the model. Applying this technique for the fetish dataset at 100 topics, one finds that the information if a user is in a relationship yields a per-word perplexity of 17.88 after 200 iterations. The context space which alone yields the worst prediction is the gender, with a perplexity of 19.37. However, given all the context variables, the gender of a user is the most important context information for topic prediction – showing the importance of including dependencies between context variables.

For the given dataset, the detected weighting of context spaces cannot be generalised in the sense that statements about the relation between demographic variables and fetishes could be made. However, by modelling representative data with the HMDP topic model, insights in the connection between context variables and topics can be gained, which makes the HMDP a useful tool for data analysis.

4.5.3 Context selection and analysis with the HMDP – a study of the Linux kernel mailing list

The HMDP not only allows a weighting of context information (i.e. the influence of topic distributions of clusters of each context space on the topics of the documents), but also can be used to select (i.e. remove unnecessary) context information. Additionally, the context-specific topic distributions can be used to learn about the specific relation between context information and single topics.

For demonstrating the ability of the HMDP topic model to select and describe context information, a third dataset is employed, which is substantially different to the food dataset and the fetish dataset: the archived messages of the Linux kernel mailing list (LKML) collected over a period of almost twenty years.

Dataset description

The Linux kernel mailing list is a central communication channel in the development of the Linux kernel, where bugs are reported and technical topics are discussed [MS02, HKS15]. It brings together professional contributors from companies, hobbyists, members of universities as well as public and private research institutes [HKS15].

A complete dump of messages was downloaded from the LKML archive, consisting of emails archived from 1995 until the end of 2014. In total 3,381,285 messages were retrieved. Figure 4.7 shows the number of messages posted in the LKML over the collection period. The activity in the LKML is growing every year.

Five context spaces can be constructed using the metadata from the LKML dataset. Every email has an associated timestamp, which yields information about

- the timeline,
- the yearly cycle,
- the weekly cycle,
- and the daily cycle.

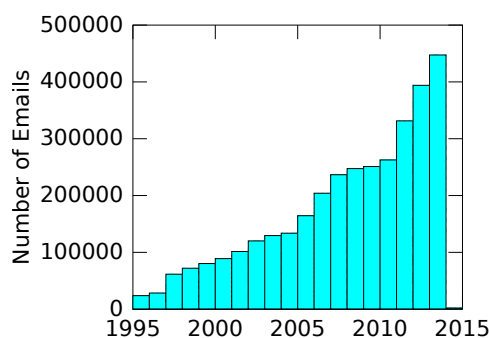


Figure 4.7: Emails per year in the Linux kernel mailing list over the observation period from 1995 to 2014. The activity is increasing every year: In 1996, on average 63 mails were posted per day and in 2014, 1226 messages were posted per day on the mailing list.

In addition, the LKML has several mailing lists for subtopics. The mailing list ID on which a mail was published is stored, and over the years in total 1730 different mailing lists were created. If an email was posted on multiple mailing lists, it was duplicated in the dataset.

To the best of the knowledge of the author, explicitly modelling temporal cycles in a topic model is a novelty. This is important, because information about temporal cycles is available for almost all documents from social media, as the only information required is a timestamp (and, in some cases, the time zone of a user). In the following, it will be shown that including temporal cycles in a topic model significantly contributes to topic prediction and can allow to gain valuable insights into the semantics of topics.

Detected topics

In order to detect topics on this larger dataset, messages were stemmed and stopwords removed using the Snowball stemmer and stopword list [Por01]. Rare words which occurred less than 1,000 times in the 20 years of the observation period were excluded for performance reasons. The batch size was set to a large value of 4096, and the parameters of the learning rate function were set to $\kappa = 0.5$ and $\tau = 64$, which is identical to the optimal parameters empirically found in [HBB10].

For modelling the context information, the mailing list ID can be directly used as context variable. The three temporal cycles and the timeline were split in 1,000 equally-sized clusters each, by sorting messages by time and creating chunks. This resulted in a total number of 5,730 context clusters.

The truncation level of topics was set to 50. This is a rather small value given the size of the dataset, but already lead to interpretable topics. The complete topic description is given in Table 4.6 and Table 4.7. Most topics

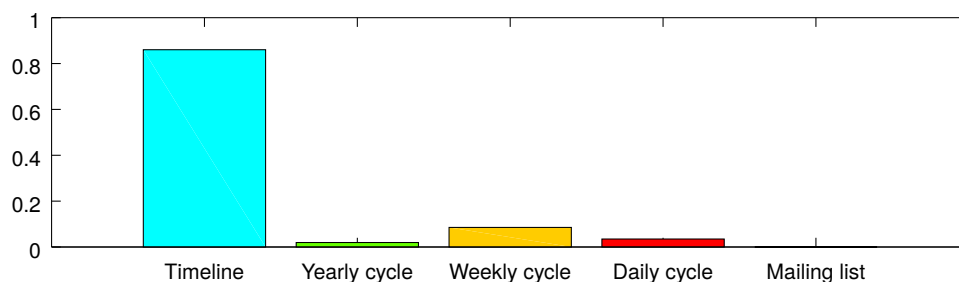


Figure 4.8: Weighting of context information for the Linux kernel mailing list. The timestamp of a mail is most predictive, followed by the position on the week cycle (e.g. indicating if a message was written on a work day), the daily cycle and the weekly cycle. Knowing the mailing list a email was posted on does not have any impact on the topic prediction and thus the probability is set to a value close to zero during inference. Using this mechanism, unnecessary context information can be removed.

cover technical aspects of kernel development, such as “device, driver, pci” (*Topic 2*) or “packet, network, connect” (*Topic 3*), while some topics such as “thank, patch, appli” (*Topic 1*) cover the communication between users who search and find help.

Context analysis

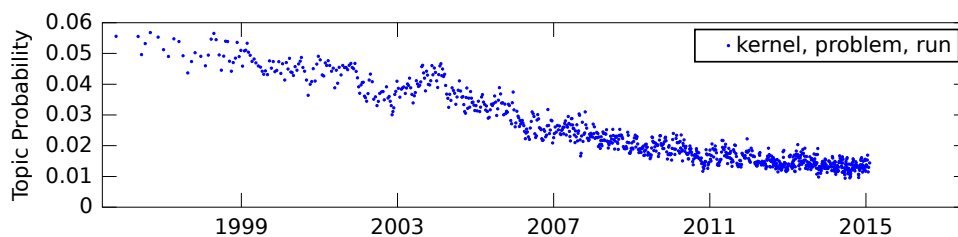
The weighting of context spaces given in parameter ζ for the mixing proportions of context spaces in the MDP is given in Figure 4.8. The time an email is created has the largest predictive power on the topics of a document – more than 80% of the topics of documents are explained by the location of documents on the timeline. The timeline is followed by the weekly cycle, which explains about 10% of the topics used in the messages. Finally, the daily and yearly cycle share the rest of the probability. Knowing the mailing list on which an email was published does not contribute to topic prediction and is virtually set to a probability of zero. A possible explanation for this is that the number of topics is too low to detect the subtopics which are discussed on the various mailing lists. However, the fact that one of the context spaces has an assigned weight of virtually zero demonstrates the ability of the HMDP to exclude unimportant context spaces during inference. This is important for datasets with a large number of context variables. One natural extension of the PCSVB sampling scheme would be to explicitly prune context spaces if their weight falls below a predefined threshold, as in [EB15]. In addition, context variables could be added during inference. Variable sets where interactions between variables are assumed to be present should be evaluated simultaneously. That way, large sets of context variables could be efficiently evaluated during online inference.

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
thank	devic	packet	problem	use
patch	driver	network	patch	size
appli	pci	connect	bug	number
test	usb	socket	issu	valu
greg	bus	server	fix	structur
send	port	receiv	report	pointer
Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
block	use	fix	case	int
data	support	chang	way	struct
buffer	work	add	differ	diff
read	fix	use	think	return
disk	user	remov	use	static
file	patch	cleanup	reason	unsign
Topic 11	Topic 12	Topic 13	Topic 14	Topic 15
messag	time	architectur	check	regard
list	perform	implement	return	best
mail	run	use	call	sorri
email	test	support	code	hello
question	result	code	fail	pgp
help	system	arch	case	thank
Topic 16	Topic 17	Topic 18	Topic 19	Topic 20
tree	code	acpi	version	state
merg	system	pci	cheer	power
linus	make	kernel	releas	devic
patch	devic	irq	avail	suspend
fix	issu	tabl	git	alan
branch	file	apic	linux	system
Topic 21	Topic 22	Topic 23	Topic 24	Topic 25
kernel	just	chang	set	memori
problem	like	patch	option	page
run	see	move	enabl	address
machin	seem	one	depend	map
boot	someth	two	use	tabl
system	look	comment	default	use

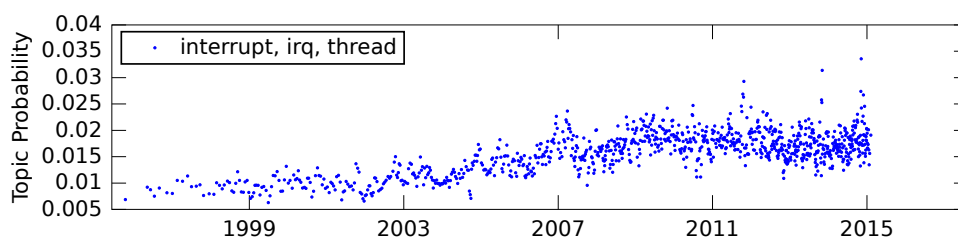
Table 4.6: Topics detected by the HMDP topic model in the Linux kernel mailing list. Part 1/2. For every topic, the six most-probable terms are displayed. The ordering of topics is arbitrary. All displayed words have a probability greater than 1% in the respective topic.

Topic 26	Topic 27	Topic 28	Topic 29	Topic 30
error	kernel	linux	time	think
function	trace	softwar	also	yes
warn	oop	busi	call	good
type	call	develop	current	right
declar	code	free	look	make
refer	jan	project	still	sure
Topic 31	Topic 32	Topic 33	Topic 34	Topic 35
kernel	need	control	kernel	add
modul	well	intel	need	mode
compil	still	bridg	new	support
build	correct	status	driver	creat
make	alradi	memori	set	patch
load	bit	capabl	just	chang
Topic 36	Topic 37	Topic 38	Topic 39	Topic 40
pleas	page	lock	file	driver
know	alloc	call	mount	clock
let	memori	path	directori	control
patch	free	hold	filesystem	regist
review	swap	loop	root	devic
anyon	cach	race	inod	gpio
Topic 41	Topic 42	Topic 43	Topic 44	Topic 45
event	thing	interrupt	cpu	use
perf	realli	irq	task	new
debug	peopl	thread	node	driver
output	just	timer	cpus	function
trace	even	handler	memori	instead
use	want	queue	schedul	interfac
Topic 46	Topic 47	Topic 48	Topic 49	Topic 50
user	get	get	kernel	remov
process	think	tri	scsi	function
use	list	time	drive	line
program	two	now	devic	name
userspac	chang	work	error	defin
access	sinc	back	ide	use

Table 4.7: Topics detected by the HMDP topic model in the Linux kernel mailing list. Part 2/2. Words with a probability below 1% are greyed out. Some topics model technical aspects of the kernel development, such as *Topic 27*, other topics cover language aspects such as questions which are covered in *Topic 36*.



(a) Example of a topic of decreasing importance over time



(b) Example of a topic of increasing importance over time

Figure 4.9: Probability of (a) *Topic 21* (“kernel, problem”) and (b) *Topic 43* (“interrupt”) to appear in messages on the Linux kernel mailing list, over time. While the share of messages about kernel problems is decreasing, the share of messages mentioning the topic of interrupts is growing over time. Note that the interpretation of the cluster-specific topic distributions as topic probabilities is only possible due to the practical inference approximation.

To understand the importance of the timeline and the weekly cycle on topic prediction, the topic probabilities of context clusters can be analysed. The cluster-specific topic distributions in the HMDP topic model have a natural interpretation if parameters are learned via practical inference.

In the Gibbs sampler, mixing proportions (between adjacent geographical clusters) were estimated by assigning the tables of the Chinese restaurant franchise to their parent distribution. Every document can have multiple tables per topic, and thus long documents might have a bigger impact on the cluster-specific topic distribution than short documents. Therefore, it is hard to interpret the weights of the prior distributions associated with context clusters.

Practical inference changes this setting by assuming that table counts are binary, i.e. there is only one table or no table, and the estimate of table counts reduces to the question if a topic appears in a given document or not. As the context-specific topic distributions in practical inference schemes are estimated as sums of these table counts (see Equation 4.10), the context-specific topic distributions directly correspond to the probability that a topic occurs in a document, given the context information.

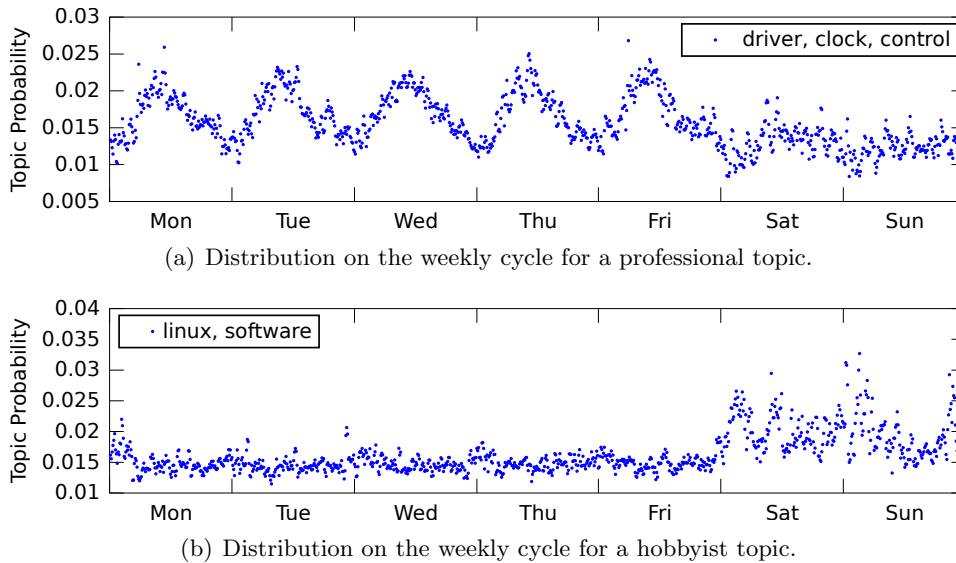


Figure 4.10: Probability of (a) *Topic 40* (“driver”) and (b) *Topic 28* (“linux, software”) to appear in messages on the Linux kernel mailing list over the weekly cycle. Messages discussing drivers have a twice-as-high probability of 2% during the working hours, indicating that the topic is mostly discussed by professionals. In contrast, messages mentioning the topic of linux software are twice as likely on the weekend than during the week, indicating a hobbyist topic. These two patterns on the weekly cycle are typical for the topics of professionals and hobbyists and most of the topics in the dataset exhibit similar patterns.

Because of the high importance of the timeline on topic prediction, it is expected that the topic probability of several topics changes over time. The temporal development of the topic probabilities of *Topic 21* (“kernel problem”) and *Topic 43* (“interrupt”) on the timeline shown in Figure 4.9 confirm this intuition. While in the beginning of the LKML almost 6% of the messages covered *Topic 21*, in 2015 less than 2% of the messages covered that topic. On the other hand, the share of *Topic 43* developed from less than 1% of the messages to about 2% in 2015.

Similarly, the topic probabilities on the weekly cycle can be examined. The topic probabilities for *Topic 40* (“driver”) and *Topic 28* (“linux, software”) are shown in Figure 4.10. *Topic 40* is exemplary for a professional topic: it is significantly more often discussed during the working hours than during the night or on the weekend. In contrast, the topic of linux software seems to be more popular with hobbyists and regularly raises in importance on the weekend.

Practical inference for the HMDP allows for a direct interpretation of the

topic weights of context clusters as topic probabilities. Including the location of documents on the weekly cycle can be beneficial both for predicting topics and for understanding the meaning of a topic, whether it is mainly used by professionals or hobbyists.

4.6 Summary

The comparison of the Gibbs sampler and PCSVB on the food dataset provided several interesting insights. The results indicate that Gibbs sampling for the parameters of the HMDP topic model is a better method for detecting a low number of semantically coherent topics. PCSVB improves the predictive performance of HMDP, detects a higher number of semantically coherent topics and yields a faster convergence of the sampler.

The application of the HMDP topic model on the fetish dataset and on the messages of the Linux kernel mailing list demonstrated the special properties of the HMDP:

(i) The HMDP is able to model multiple context variables. Context variables can include discrete, continuous, cyclic and spherical context. Using efficient PCSVB online inference, **(ii) the HMDP can be employed to model large text corpora.** As the structure of the model remains relatively simple, standard approaches for distributed computing such as [NASW09] can be applied. As demonstrated by the use case of the Linux kernel mailing list with nearly 6000 context clusters, **(iii) PCSVB for the HMDP can cope with a large number of context clusters.** Context clusters are a required input and can be created by simple heuristics, e.g. by sorting data and creating equally-sized blocks. During inference, non-relevant context spaces are weighted down and can be excluded, which means that **(iv) the HMDP can be employed to select context spaces relevant for topic prediction.** Additionally, **(v) the parameters of the HMDP have a natural interpretation** as probabilities. Specifically, the topic weights of context-variables correspond to the probability that a document in the context cluster includes the given topic. The mixing weights of the context clusters in the MDP correspond to the probability that a given context space explains the topic of a document in the corpus.

Conclusion

In this thesis, a novel way of integrating context information in probabilistic models was presented using the example of topic models.

First, the benefit of including system-specific context was demonstrated on the example of power indices. Using a probabilistic interpretation, an evaluation of existing power indices was conducted on real-world observations from the delegative democracy platform of the German Pirate Party. It could be shown that novel generalisations of the Banzhaf and Shapley index, which are able to model system-specific voting bias, lead to an improved prediction of voting power.

Then, geographical networks based on adjacency relations between clusters of geographically distributed documents were introduced as an instance of a network representation of context variables. The multi-Dirichlet process and three-level hierarchical multi-Dirichlet processes (HMDP) were presented as a novel way of modelling context information using context networks. It could be shown that using a HMDP topic model, it is possible to detect topics with a complex distribution in the context space by modelling dependencies between adjacent clusters of the context network.

Finally, a generalisation of hierarchical multi-Dirichlet processes for multiple context variables was presented, which exploits multiple context networks based on adjacent document clusters in arbitrary context spaces. An efficient online inference scheme based on practical stochastic variational Bayesian inference (PCSVB) was derived which significantly improves the convergence rate as well as the predictive quality of the model, and which allows to detect a larger number of semantically coherent topics. The unique ability of the HMDP topic model to simultaneously model multiple complex structures in the context space was demonstrated. It was shown that the parameters governing the connection between the context space and the detected topics have a natural interpretation as probabilities. Both the implementation of the Gibbs sampler for the geographical topic model based on the HMDP² and the PCSVB inference scheme for the HMDP³ are published as open source.

²<https://github.com/ckling/mgtm>

³<https://github.com/ckling/hmdp>

Findings

The main findings of this thesis are:

Delegative democracies for evaluating power indices. It could be shown that voting data of large delegative democracies can be used to evaluate power indices. In order to do so, the probabilistic interpretation of power indices by [Str77] is employed and the perplexity of observed voting power is calculated.

System-specific power indices. A probabilistic interpretation of power indices allows the integration of context information. Specifically, observed voting bias can be integrated in existing power indices via system-specific parameters. It could be shown that power indices that take voting bias into account yield a better prediction.

Context networks for modelling spherical context variables in probabilistic topic models. The novel concept of modelling spherical (e.g. geographically distributed) variables with context networks in probabilistic models was introduced. Context networks are based on a clustering of data in the context space and on adjacency relations between clusters. The network structure can be implicitly modelled using model selection, but is most efficiently modelled with the three-level hierarchical multi-Dirichlet process presented in this thesis. Including this network structure in a probabilistic topic model can significantly improve the topic quality, which was shown for geographically distributed context variables.

Context networks for modelling multiple, arbitrary context variables. Using hierarchical multi-Dirichlet processes (HMDP), multiple discrete, linear, cyclic or spherical context variables (such as temporal or geographical context variables) can be included in a probabilistic model. In the case of topic models, the context variables can be described in terms of topic probabilities.

HMDP for weighting and selecting multiple context variables in probabilistic models. Context variables can be weighted and ultimately removed using learned parameters for context influence in the HMDP. It is possible to place Dirichlet distributions over the weights of context spaces, allowing to include prior beliefs on the importance of context spaces.

Efficient inference for hierarchical multi-Dirichlet processes. A practical collapsed stochastic variational Bayesian inference (PCSVB) [Ble13] scheme for the parameters of the three level HMDP was presented in this

thesis. Due to the dramatically reduced memory consumption and a faster convergence rate, the inference scheme allows for the online processing of large corpora of documents with metadata. Additionally, it allows to detect more semantically coherent topics than a Gibbs sampler, indicating convergence problems of the Gibbs sampler in the high dimensional context space.

Outlook

In future work, the impact of modelling context information on the quality of detected topics will be evaluated in detail. The potential of the HMDP for answering research questions in the social sciences (e.g. about the influence of context variables on observations) will be demonstrated on appropriate data, which allow to relate findings to existing theory.

The network structure between context clusters in the HMDP allows for the a-priori modelling of complex dependencies in the context space such as dependencies between different context variables. A systematic comparison of different network structures for modelling dependencies between temporal context variables will be conducted to develop best practices in modelling temporal context.

The development of the implementation of PCSVB for the HMDP will be continued to simplify the use, e.g. by automatically creating context clusters for typical context variables such as timestamps, geo-coordinates and discrete variables from given data. Additionally, a distributed implementation based on Apache Hadoop will be released for modelling large corpora with metadata.

Finally, the HMDP has a wider range of applications than topic models or mixed-membership models of multinomial distributions. It can be applied to include structural information in a multitude of probabilistic models. The application of the HMDP in other probabilistic models will be the subject of future research.

Bibliography

- [AB12] A. Agovic and A. Banerjee. Gaussian process topic models. *ArXiv e-prints*, March 2012.
- [AHS13] Amr Ahmed, Liangjie Hong, and Alex Smola. Hierarchical geographical modeling of user locations from social media posts. In *WWW*, 2013.
- [Ald85] D.J. Aldous. Exchangeability and related topics. In *École d'Été St Flour 1983*, pages 1–198. Springer-Verlag, 1985. Lecture Notes in Math. 1117.
- [Ant74] C. Antoniak. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, 2:1152–1174, 1974.
- [Aur91] Franz Aurenhammer. Voronoi diagrams – a survey of a fundamental geometric data structure. *ACM Comput. Surv.*, 23(3):345–405, 1991.
- [AWST12] Arthur U. Asuncion, Max Welling, Padhraic Smyth, and Yee Whye Teh. On smoothing and inference for topic models. *CoRR*, abs/1205.2662, 2012.
- [AX12] A. Ahmed and E. P. Xing. Timeline: A dynamic hierarchical Dirichlet process model for recovering birth/death and evolution of topics in text stream. *ArXiv e-prints*, March 2012.
- [Ban65] John Banzhaf. Weighted voting does not work: a mathematical analysis. *Rutgers Law Review*, 19:317–343, 1965.
- [BDGS05] Arindam Banerjee, Inderjit S. Dhillon, Joydeep Ghosh, and Suvrit Sra. Clustering on the unit hypersphere using von mises-fisher distributions. *Journal of Machine Learning Research*, 6:1345–1382, 2005.
- [Bec75] Nathaniel Beck. A note on the probability of a tied election. *Public Choice*, 23(1):75–79, 1975.

- [BF11] David M. Blei and Peter I. Frazier. Distance dependent Chinese restaurant processes. *Journal of Machine Learning Research*, 12:2461–2488, 2011.
- [BFC98] Chris Brunson, Stewart Fotheringham, and Martin Charlton. Geographically weighted regression. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 47(3):431–443, 1998.
- [BHO10] Mark Bangert, Philipp Hennig, and Uwe Oelfke. Using an infinite von Mises–Fisher mixture model to cluster treatment beam directions in external radiation therapy. In *ICMLA*, pages 746–751. IEEE Computer Society, 2010.
- [BJ06a] David M. Blei and Michael I. Jordan. Variational inference for Dirichlet process mixtures. *Bayesian Analysis*, 1(1):121–144, 2006.
- [BJ06b] David M. Blei and Michael I. Jordan. Variational methods for the Dirichlet process. In *Proc. ICML*, 2006.
- [Ble13] A. Bleier. Practical collapsed stochastic variational inference for the HDP. *ArXiv e-prints*, December 2013.
- [BM73] D. Blackwell and J. MacQueen. Ferguson distributions via Polya urn schemes. *The Annals of Statistics*, 1:353–355, 1973.
- [BNJ03] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003.
- [CB09] Jonathan Chang and David M. Blei. Relational topic models for document networks. In David A. Van Dyk and Max Welling, editors, *AISTATS*, volume 5 of *JMLR Proceedings*, pages 81–88. JMLR.org, 2009.
- [CBGW⁺09] Jonathan Chang, Jordan Boyd-Graber, Chong Wang, Sean Gerrish, and David M. Blei. Reading tea leaves: How humans interpret topic models. In *Neural Information Processing Systems*, 2009.
- [CG11] Kevin Robert Canini and Thomas L. Griffiths. A nonparametric Bayesian model of multi-level category learning. In *Proc. AAAI Conf. on Artificial Intelligence*, 2011.
- [CZC12] Xu Chen, Mingyuan Zhou, and Lawrence Carin. The contextual focused topic model. In Qiang Yang 0001, Deepak Agarwal, and Jian Pei, editors, *KDD*, pages 96–104. ACM, 2012.

- [Daw06] Richard Dawkins. *The selfish gene*. Oxford University Press, 2006.
- [DHWX13] Avinava Dubey, Ahmed Hefny, Sinead Williamson, and Eric P. Xing. A nonparametric mixture model for topic modeling over time. In *SDM*, pages 530–538. SIAM, 2013.
- [Die95] Andreas Diekmann. *Empirische Sozialforschung. Grundlagen, Methoden, Anwendungen*. Rowohlt, Reinbek bei Hamburg, 1995.
- [Die06] Laura Dietz. Exploring social topic networks with the author-topic model. In *Workshop “Semantic Network Analysis” at European Semantic Web Conference ’06*, pages 54–60, Budva, Montenegro, June 2006.
- [DJP78] John Deegan Jr and Edward W Packel. A new index of power for simple n -person games. *Int. J. of Game Theory*, 7(2), 1978.
- [Dod84] Charles Lutwidge Dodgson. *The principles of parliamentary representation*. Harrison and Sons, 1884.
- [EB15] Tarek Elguebaly and Nizar Bouguila. Simultaneous high-dimensional clustering and feature selection using asymmetric Gaussian mixture models. *Image Vision Comput.*, 34:27–41, 2015.
- [FBD⁺13] J. Foulds, L. Boyles, C. Dubois, P. Smyth, and M. Welling. Stochastic collapsed variational Bayesian inference for latent Dirichlet allocation. *ArXiv e-prints*, May 2013.
- [Fer73] T. Ferguson. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1:209–230, 1973.
- [FHS⁺15] Damien Fay, Hamed Haddadi, Michael C. Seto, Han Wang, and Christoph Carl Kling. An exploration of fetish social networks and communities. *NetSci-X*, abs/1511.01436, 2015.
- [Fis53] Ronald Fisher. Dispersion on a sphere. *Proc. of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 217(1130), 1953.
- [Fre13] Martin Fredriksson. An open source project for politics. In James Arvanitakis and Ingrid Matthews, editors, *The Citizen in the 21st Century*. Inter-Disciplinary Press, 2013.
- [GK03] M. Girolami and A. Kaban. On an equivalence between PLSI and LDA. In *Proc. of ACM SIGIR*, 2003.

- [GKT02] A. Gelman, J.N. Katz, and F. Tuerlinckx. The mathematics and statistics of voting power. *Statistical Science*, 17(4):420–435, 2002.
- [Gri02] Tom Griffiths. Gibbs sampling in the generative model of Latent Dirichlet Allocation. Technical report, Stanford University, 2002.
- [GY14] Siddharth Gopal and Yiming Yang. Von Mises-Fisher clustering models. In *ICML*, volume 32 of *JMLR Proceedings*, pages 154–162. JMLR.org, 2014.
- [HAG⁺12] Liangjie Hong, Amr Ahmed, Siva Gurumurthy, Alexander J. Smola, and Kostas Tsioutsoulouklis. Discovering geographical topics in the Twitter stream. In *Proc. World Wide Web Conf.*, pages 769–778, 2012.
- [HBB10] Matthew D. Hoffman, David M. Blei, and Francis R. Bach. Online learning for latent Dirichlet allocation. In John D. Lafferty, Christopher K. I. Williams, John Shawe-Taylor, Richard S. Zemel, and Aron Culotta, editors, *NIPS*, pages 856–864. Curran Associates, Inc., 2010.
- [Hei06] Gregor Heinrich. Infinite LDA. Estimation of the inherent number of topics using Dirichlet process prior and exploiting sparsity of conditional topic distributions, August 2006.
- [Hei08] Gregor Heinrich. Parameter estimation for text analysis. Technical note version 2 (1: 2005), vsonix GmbH and University of Leipzig, February 2008.
- [HG09] Gregor Heinrich and Michael Goesele. Variational Bayes for generic topic models. In *Proc. 32nd (German) Annual Conference on Artificial Intelligence*, 2009.
- [HKS15] Dirk Homscheid, Jérôme Kunegis, and Mario Schaarschmidt. Private-collective innovation and open source software: Longitudinal insights from Linux kernel development. In Marijn Janssen, Matti Mäntymäki, Jan Hidders, Bram Klievink, Winfried Lamersdorf, Bastiaan van Loenen, and Anneke Zuiderwijk, editors, *I3E*, volume 9373 of *Lecture Notes in Computer Science*, pages 299–313. Springer, 2015.
- [Hof99] Thomas Hofmann. Probabilistic latent semantic analysis. In *Proc. Uncertainty in Artificial Intelligence*, pages 289–296, 1999.

- [Hua05] Jonathan Huang. Maximum likelihood estimation of Dirichlet distribution parameters. *CMU Technique Report*, 2005.
- [Jab11] Sebastian Jabbusch. Liquid democracy in der Piratenpartei. Master's thesis, University of Greifswald, 2011.
- [JGJS99] Michael I. Jordan, Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul. An introduction to variational methods for graphical models. *Mach. Learn.*, 37(2):183–233, 1999.
- [KF09] Daphne Koller and Nir Friedman. *Probabilistic Graphical Models – Principles and Techniques*. MIT Press, 2009.
- [Kin78] John F. C. Kingman. Uses of exchangeability. *The Annals of Probability*, 6(2):183–197, 1978.
- [KKH⁺15] Christoph Carl Kling, Jérôme Kunegis, Heinrich Hartmann, Markus Strohmaier, and Steffen Staab. Voting behaviour and power in online democracy: A study of LiquidFeedback in Germany's Pirate Party. In *Proc. Int. Conf. on Weblogs and Social Media*, 2015.
- [KKSS14] Christoph C. Kling, Jérôme Kunegis, Sergej Sizov, and Steffen Staab. Detecting non-gaussian geographical topics in tagged photo collections. In *WSDM'14: Proceedings of the 7th International Conference on Web Search and Data Mining*, 2014.
- [Kos09] Vassilis Kostakos. Is the crowd's wisdom biased? A quantitative analysis of three online communities. In *Proc. Int. Conf. on Computational Science and Engineering*, pages 251–255, 2009.
- [KP12] Jérôme Kunegis and Julia Preusse. Fairness on the web: Alternatives to the power law. In *Proc. Web Science Conf.*, 2012.
- [KS96] Daphne Koller and Mehran Sahami. Toward optimal feature selection. In *Proceedings of the 13th International Conference on Machine Learning (ICML'96)*, pages 284–292, 1996.
- [KSS11] Christoph Carl Kling, Sergej Sizov, and Steffen Staab. Virtual field research with social media: A pilot case of biometeorology. In *ACM WebSci'11*, pages 1–2, June 2011. WebSci Conference 2011.
- [Law01] Neil David Lawrence. *Variational inference in probabilistic models*. PhD thesis, University of Cambridge, 2001.

- [LF12] Dahua Lin and John W. Fisher. Coupling nonparametric mixtures via latent Dirichlet processes. In Peter L. Bartlett, Fernando C. N. Pereira, Christopher J. C. Burges, Léon Bottou, and Kilian Q. Weinberger, editors, *NIPS*, pages 55–63, 2012.
- [LJT07] Percy Liang, Michael I. Jordan, and Ben Taskar. A permutation-augmented sampler for dp mixture models. In *ICML '07: Proceedings of the 24th international conference on Machine learning*, pages 545–552, New York, NY, USA, 2007. ACM.
- [Loe99] George Loewenstein. Experimental economics from the vantage-point of behavioural economics. *The Economic J.*, 109(453):25–34, 1999.
- [LWBS14] Haiko Lietz, Claudia Wagner, Arnim Bleier, and Markus Strohmaier. When politicians talk: Assessing online conversational practices of political parties on Twitter. In *International AAAI Conference on Weblogs and Social Media (ICWSM2014)*, Ann Arbor, MI, USA, June 2-4, 2014.
- [MCZZ08] Qiaozhu Mei, Deng Cai, Duo Zhang, and ChengXiang Zhai. Topic modeling with network regularization. In Jinpeng Huai, Robin Chen, Hsiao-Wuen Hon, Yunhao Liu, Wei-Ying Ma, Andrew Tomkins, and Xiaodong Zhang, editors, *WWW*, pages 101–110. ACM, 2008.
- [Min00] Thomas Minka. Estimating a Dirichlet distribution. Technical report, MIT, 2000.
- [MJ09] Kanti V. Mardia and Peter E. Jupp. *Directional statistics*, volume 494. John Wiley & Sons, 2009.
- [MLSZ06] Qiaozhu Mei, Chao Liu, Hang Su, and ChengXiang Zhai. A probabilistic approach to spatiotemporal theme pattern mining on weblogs. In *Proc. World Wide Web Conf.*, pages 533–542, 2006.
- [MM08] David M. Mimno and Andrew McCallum. Topic models conditioned on arbitrary features with Dirichlet-multinomial regression. In David A. McAllester and Petri Myllymäki, editors, *UAI*, pages 411–418. AUAI Press, 2008.
- [MO08] Fredrik Miegel and Tobias Olsson. From pirates to politicians. *Democracy, journalism and technology: New developments in an enlarged Europe*, pages 203–16, 2008.

- [MS02] Jae Yun Moon and Lee Sproull. Essence of distributed work: The case of the linux kernel. In Pamela J. Hinds and Sara Kiesler, editors, *Distributed Work*, pages 381–404. MIT Press, Cambridge, MA, 2002.
- [Mur01] Kevin P. Murphy. An introduction to graphical models. Technical report, Intel Research, 2001.
- [NASW09] David Newman, Arthur U. Asuncion, Padhraic Smyth, and Max Welling. Distributed algorithms for topic models. *JMLR*, 10:1801–1828, 2009.
- [Nea00] Radford M. Neal. Markov chain sampling methods for Dirichlet process mixture models. *J. of Computational and Graphical Statistics*, 9(2):249–265, 2000.
- [Pau14] Alois Paulin. Through liquid democracy to sustainable non-bureaucratic government. In *Proc. Int. Conf. for E-Democracy and Open Government*, pages 205–217, 2014.
- [PDJ80] Edward W. Packel and John Deegan Jr. An axiomated family of power indices for simple n -person games. *Public Choice*, 35(2):229–239, 1980.
- [Pop79] Karl R. Popper. *Objective Knowledge: An evolutionary approach*. Clarendon Press, Oxford, revised edition, 1979.
- [Por01] Martin F. Porter. Snowball: A language for stemming algorithms. Published online, October 2001. Accessed 11.03.2008, 15.00h.
- [PSD00] J. K. Pritchard, M. Stephens, and P. Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155:945–955, 2000.
- [RB14] Maxim Rabinovich and David M. Blei. The inverse regression topic model. In *ICML*, volume 32 of *JMLR Proceedings*, pages 199–207. JMLR.org, 2014.
- [RDC08] Lu Ren, David B. Dunson, and Lawrence Carin. The dynamic hierarchical Dirichlet process. In William W. Cohen, Andrew McCallum, and Sam T. Roweis, editors, *ICML*, volume 307 of *ACM International Conference Proceeding Series*, pages 824–831. ACM, 2008.
- [RZGSS04] Michal Rosen-Zvi, Tom Griffiths, Mark Steyvers, and Padhraic Smyth. The author-topic model for authors and documents. In *20th Conference on Uncertainty in Artificial Intelligence*, volume 21, Banff Park Lodge, Banff, Canada, July 2004.

- [Sha54] Lloyd S. Shapley. A method for evaluating the distribution of power in a committee situation. *Am. Polit. Sci. Rev.*, 48:787–792, 1954.
- [Siz10] Sergej Sizov. GeoFolk: latent spatial semantics in Web 2.0 social media. In *Proc. Int. Conf. on Web Search and Data Mining*, pages 281–290, 2010.
- [SKN12] Issei Sato, Kenichi Kurihara, and Hiroshi Nakagawa. Practical collapsed variational Bayes inference for hierarchical Dirichlet process. In Qiang Yang 0001, Deepak Agarwal, and Jian Pei, editors, *KDD*, pages 105–113. ACM, 2012.
- [Sra12] Suvrit Sra. A short note on parameter approximation for von Mises-Fisher distributions: and a fast implementation of $I_s(x)$. *Computational Statistics*, 27(1):177–190, 2012.
- [Str77] Philip D. Straffin. Homogeneity, independence and power indices. *Public Choice*, 30:107–118, 1977.
- [Str94] Philip D. Straffin. Power and stability in politics. *Handbook of Game Theory with Economic Applications*, 2, 1994.
- [TFH14] Antonio Tenorio-Fornés and Samer Hassan. Towards an agent-supported online assembly. In *Proc. Int. Conf. on Advanced Collaborative Networks, Systems and Applications*, pages 72–77, 2014.
- [TFO⁺07] Akihiro Tanabe, Kenji Fukumizu, Shigeyuki Oba, Takashi Takenouchi, and Shin Ishii. Parameter estimation for von Mises-Fisher distributions. *Computational Statistics*, 22(1):145–157, 2007.
- [TJBB06] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *J. of the American Statistical Association*, 101:1566–1581, 2006.
- [TKW08] Yee Whye Teh, Kenichi Kurihara, and Max Welling. Collapsed variational inference for HDP. In *Advances in Neural Information Processing Systems (NIPS)*, 2008.
- [TNW07] Y. W. Teh, D. Newman, and M. Welling. A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation. In *Advances in Neural Information Processing Systems*, volume 19, 2007.
- [Wat82] Geoffrey S Watson. Distributions on the circle and sphere. *Journal of Applied Probability*, pages 265–280, 1982.

- [WC06] Xing Wei and W. Bruce Croft. LDA-based document models for ad hoc retrieval. In *Proc. SIGIR*, 2006.
- [WM06] Xuerui Wang and Andrew McCallum. Topics over time: a non-Markov continuous-time model of topical trends. In Tina Eliassi-Rad, Lyle H. Ungar, Mark Craven, and Dimitrios Gunopulos, editors, *KDD*, pages 424–433. ACM, 2006.
- [WMM09] Hanna M. Wallach, David M. Mimno, and Andrew McCallum. Rethinking lda: Why priors matter. In Yoshua Bengio, Dale Schuurmans, John D. Lafferty, Christopher K. I. Williams, and Aron Culotta, editors, *NIPS*, pages 1973–1981. Curran Associates, Inc., 2009.
- [WS98] Duncan J. Watts and Steven H. Strogatz. Collective dynamics of ‘small-world’ networks. *Nature*, 393(1):440–442, 1998.
- [WWXM07] Chong Wang, Jinggang Wang, Xing Xie, and Wei-Ying Ma. Mining geographic knowledge using location aware topic model. In *Proc. Workshop on Geographic Information Retrieval*, pages 65–70, 2007.
- [XJR12] Eric P. Xing, Michael I. Jordan, and Stuart J. Russell. A generalized mean field algorithm for variational inference in exponential families. *CoRR*, abs/1212.2512, 2012.
- [YCH⁺11a] Zhijun Yin, Liangliang Cao, Jiawei Han, Chengxiang Zhai, and Thomas Huang. LPTA: A probabilistic model for latent periodic topic analysis. In *Data Mining (ICDM), 2011 IEEE 11th International Conference on*, pages 904–913. IEEE, 2011.
- [YCH⁺11b] Zhijun Yin, Liangliang Cao, Jiawei Han, Chengxiang Zhai, and Thomas S. Huang. Geographical topic discovery and comparison. In *Proc. World Wide Web Conf.*, pages 247–256, 2011.
- [YYT07] Hiroshi Yamakawa, Michiko Yoshida, and Motohiro Tsuchiya. Toward delegated democracy: Vote by yourself, or trust your network. *Int. J. of Human and Social Sciences*, 1(2), 2007.
- [ZSZL10] Jianwen Zhang, Yangqiu Song, Changshui Zhang, and Shixia Liu. Evolutionary hierarchical Dirichlet processes for multiple correlated time-varying corpora. In Bharat Rao, Balaji Krishnapuram, Andrew Tomkins, and Qiang Yang 0001, editors, *KDD*, pages 1079–1088. ACM, 2010.

Appendix A

Appendix

A.1 MLE Estimate for the Binomial Distribution

The complete derivation of the maximum likelihood estimate for the binomial distribution with n_0 negative and n_1 positive observations is:

$$\begin{aligned} \frac{\partial}{\partial p} \log(\mathcal{L}(p | \mathbf{v})) &\stackrel{!}{=} 0 \\ \Leftrightarrow \frac{\partial}{\partial p} \log(p(\mathbf{v} | p)) &= 0 \\ \Leftrightarrow \frac{\partial}{\partial p} \log\left(\prod_{j=1}^n p^{v_j} \cdot (1-p)^{1-v_j}\right) &= 0 \\ \Leftrightarrow \frac{\partial}{\partial p} \sum_{j=1}^n \log(p^{v_j} \cdot (1-p)^{1-v_j}) &= 0 \\ \Leftrightarrow \frac{\partial}{\partial p} (n_1 \cdot \log(p) + n_0 \cdot \log(1-p)) &= 0 \\ \Leftrightarrow n_1 \cdot \frac{\partial}{\partial p} (\log(p)) + n_0 \cdot \frac{\partial}{\partial p} (\log((1-p))) &= 0 \\ \Leftrightarrow n_1 \cdot \frac{1}{p} + n_0 \cdot \frac{1}{1-p} \cdot (-1) &= 0 \\ \Leftrightarrow n_1 \cdot \frac{1}{p} = n_0 \cdot \frac{1}{1-p} &\Leftrightarrow n_1 \cdot (1-p) = n_0 \cdot p \\ \Leftrightarrow n_1 - n_1 \cdot p = n_0 \cdot p &\Leftrightarrow n_1 = (n_0 + n_1) \cdot p \\ \Leftrightarrow \frac{n_1}{n_0 + n_1} = p & \tag{A.1} \end{aligned}$$

A.2 MAP Estimate for the Binomial Distribution

The derivation of the MAP estimate for the binomial distribution with a beta-distributed prior is given by:

$$\begin{aligned}
 \frac{\partial}{\partial p} \log(p(p | n_0, n_1, \alpha, \beta)) &= \frac{\partial}{\partial p} \log \text{Beta}(p, n_1 + \alpha, n_0 + \beta) \stackrel{!}{=} 0 \\
 \Leftrightarrow \frac{\partial}{\partial p} \log(p^{n_1 + \alpha - 1} \cdot (1 - p)^{n_0 + \beta - 1}) &= 0 \\
 \Leftrightarrow (n_1 + \alpha - 1) \cdot \frac{1}{p} + (n_0 + \beta - 1) \cdot \frac{1}{1 - p} \cdot (-1) &= 0 \\
 \Leftrightarrow (n_1 + \alpha - 1) \cdot (1 - p) &= (n_0 + \beta - 1) \cdot p \\
 \Leftrightarrow p = \frac{n_1 + \alpha - 1}{n_1 + n_0 + \alpha + \beta - 2} & \tag{A.2}
 \end{aligned}$$

A.3 Details PCSVB for HMDP

The practical stochastic variational inference scheme for the hierarchical multi-Dirichlet process is based on several approximations.

In contrast to a conventional collapsed sampler, the inference presented in this thesis samples more complex topic assignments of words: Each word is simultaneously assigned to a topic, a context space and to a cluster in that context space.

The marginal distribution over those extended topic assignments, the words and the global topic distribution π_0 (using the stick-breaking construction for the latter) under the practical approximation (i.e. table counts in the CRP representation are either zero or one) is given by:

$$\begin{aligned}
& p(\mathbf{w}, \mathbf{z}, \tilde{\boldsymbol{\pi}}_0, \mathbf{m} \mid \beta_0, \gamma, \alpha_0, \alpha_1, \varepsilon, \boldsymbol{\delta}, \mathbf{g}) \\
&= \prod_{m=0}^M \frac{\Gamma(\alpha_1)}{\Gamma(\alpha_1 + N_m)} \cdot \prod_{k=1}^K \prod_{f=1}^F \prod_{C_f} \Gamma(n_{mfjk}) \\
&\cdot \alpha_1^{n'_1} \cdot \frac{\Gamma(F \cdot \varepsilon)}{\Gamma(n'_{1..} + F \cdot \varepsilon)} \prod_{f=1}^F \frac{\Gamma(n'_{f..} + \varepsilon)}{\Gamma(\varepsilon)} \cdot \prod_{i=1}^{A_f} \frac{\Gamma(|P_i| \cdot \delta_f)}{\Gamma(n''_{f.i} + |P_i| \cdot \delta_f)} \prod_{p \in P_i} \frac{\Gamma(n''_{fip} + \delta_f)}{\Gamma(\delta_f)} \\
&\cdot \prod_{f=1}^F \prod_{j=0}^{C_f} \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_0 + n'_{fj})} \cdot \prod_{k=1}^K s'_{fjk} m_{fjk} \left(\alpha_0 \tilde{\pi}_{0k} \prod_{l=1}^{k-1} (1 - \tilde{\pi}_{0l}) \right)^{m_{fjk}} \\
&\cdot \prod_{k=1}^{K-1} \gamma \cdot (1 - \tilde{\pi}_{0k})^{\gamma-1} \\
&\cdot \prod_{k=1}^K \frac{\Gamma(V \cdot \beta_0)}{\Gamma(V \cdot \beta_0 + n_k)} \cdot \prod_{t=1}^V \frac{\Gamma(\beta_0 + n_{kt})}{\Gamma(\beta_0)}
\end{aligned} \tag{A.3}$$

where helper variables $\tilde{\boldsymbol{\pi}}_0$ and \mathbf{m} (global table counts) are introduced for the global topic distribution. n'_{fjk} are table counts

for topic k in cluster j of context space f , n''_{fgj} denotes the table counts for the j th cluster of group g' in context space f . $\pi_{01} = \tilde{\pi}_{01}$ and $\pi_{0k} = \tilde{\pi}_{0k} \prod_{l=1}^{k-1} (1 - \tilde{\pi}_{0l})$ for $k > 1$.

Using

$$E_q \left[s(n'_{fjk}, m_{fjk}) \left(\alpha_0 \tilde{\pi}_{0k} \prod_{l=1}^{k-1} (1 - \tilde{\pi}_{0l}) \right)^{n_{fjk}} \right] = \frac{\Gamma(n'_{fjk} + \alpha_0 \pi_{0k})}{\Gamma(\alpha_0)} \quad (\text{A.4})$$

and employing the practical approximation ($\Gamma(n + \alpha) \approx \Gamma(n) \forall (n \geq 1)$) and the zeroth-order Taylor approximation for the logarithm of approximately normally distributed counts ($\exp(E_q[\log(x)]) \approx x$) yields an update equation for the variational distribution over topic assignments (which at the same time give the cluster from the prior distribution from which the topic was sampled):

$$q(z_{mm} = (k, f, j)) \propto \left(n_{m,fjk} + \alpha_1 \cdot (n'_{fj} + \varepsilon) \cdot \frac{n''_{fgj} + \delta f}{n''_{fg'} + L_{fg'} \cdot \delta f} \cdot \frac{n'_{fjk} + \alpha_0 \pi_{0k}}{n'_{fj} + \alpha_0} \right) \cdot \left(\frac{\beta_0 + n_{kt}}{V \cdot \beta_0 + n_k} \right). \quad (\text{A.5})$$

where all parameters and counts are expected counts under the variational distribution.

Summing over all clusters j and contexts f yields the update equation for assigning a topic to a word, and because under the practical approximation a topic is assigned at most once to a document, the cluster-topic counts can be estimated from the binary document-table counts of the topics after sampling the table counts of a document.

A.4 ML Estimate Scaling Parameter α_1

The scaling parameters of the HMDP topic model alternatively can be updated using Newtons method on the exact (i.e. non-approximated) joint distribution of the truncated HMDP process model. The marginal joint distribution over words and topic assignments is:

$$\begin{aligned}
 p(\mathbf{w}, \mathbf{z} \mid \beta_0, \boldsymbol{\pi}_0, \alpha_0, \alpha_1, \varepsilon, \boldsymbol{\delta}) = & \\
 \prod_{m=0}^M \frac{\Gamma(\alpha_1)}{\Gamma(\alpha_1 + N_m)} \cdot \prod_{k=1}^K \frac{\Gamma(\alpha_1 \pi_{mk}^s + n_{m \cdot k})}{\Gamma(\alpha_1 \pi_{mk}^s)} & \\
 \cdot \prod_{k=1}^K \frac{\Gamma(V \cdot \beta_0)}{\Gamma(V \cdot \beta_0 + n_{k \cdot})} \cdot \prod_{t=1}^V \frac{\Gamma(\beta_0 + n_{kt})}{\Gamma(\beta_0)}. & \quad (\text{A.6})
 \end{aligned}$$

Taking the derivative with respect to α_1 yields

$$\begin{aligned}
 \frac{d \log p(\mathbf{w}, \mathbf{z} \mid \beta_0, \boldsymbol{\pi}_0, \alpha_0, \alpha_1, \varepsilon, \boldsymbol{\delta})}{d \alpha_1} = & \\
 M \cdot \Psi(\alpha_1) - \sum_{m=0}^M \Psi(\alpha_1 + N_m) + \sum_{k=1}^K \Psi(\alpha_1 \pi_{mk}^s + n_{m \cdot k}) \pi_{mk}^s - \Psi(\alpha_1 \pi_{mk}^s) \pi_{mk}^s & \quad (\text{A.7})
 \end{aligned}$$

using digamma functions. The second derivative involving trigamma functions is given by

$$\begin{aligned}
 \frac{d \log p(\mathbf{w}, \mathbf{z} \mid \beta_0, \boldsymbol{\pi}_0, \alpha_0, \alpha_1, \varepsilon, \boldsymbol{\delta})}{d \alpha_1^2} = & \\
 M \cdot \psi_1(\alpha_1) - \sum_{m=0}^M \psi_1(\alpha_1 + N_m) & \\
 + \sum_{k=1}^K \psi_1(\alpha_1 \pi_{mk}^s + n_{m \cdot k}) \cdot \pi_{mk}^s{}^2 - \psi_1(\alpha_1 \pi_{mk}^s) \cdot \pi_{mk}^s{}^2 & \quad (\text{A.8})
 \end{aligned}$$

and an iteration of the Newton method is then calculated as:

$$\alpha_1^{\text{new}} = \alpha_1^{\text{old}} - \frac{\log p(\mathbf{w}, \mathbf{z} \mid \beta_0, \boldsymbol{\pi}_0, \alpha_0, \alpha_1, \varepsilon, \boldsymbol{\delta})'}{\log p(\mathbf{w}, \mathbf{z} \mid \beta_0, \boldsymbol{\pi}_0, \alpha_0, \alpha_1, \varepsilon, \boldsymbol{\delta})''}. \quad (\text{A.9})$$

A.5 Topic Descriptions

The tables on the following pages contain extended and additional topic descriptions from Chapter 4.

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
sandwich	japanese	dimsum	tapas	sausage
pizza	sushi	vietnamese	spanish	bacon
salad	fish	chinese	paella	pork
cheese	sashimi	rice	chocolate	beef
italian	seafood	noodles	fish	steak
deli	tuna	soup	wine	bbq
bacon	rice	seafood	seafood	beans
Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
sushi	chicken	japanese	fish	bbq
japanese	rice	korean	mediterranean	potatoes
salmon	soup	ramen	bread	coffee
tuna	potato	noodle	salad	chocolate
shrimp	mushroom	soba	orange	barbeque
roll	beef	noodles	pasta	chicken
avocado	cheese	rice	icecream	toscana
Topic 11	Topic 12	Topic 13	Topic 14	Topic 15
baking	cheese	indian	italian	love
chocolate	bento	chips	wine	chocolate
bread	tomato	bread	pizza	strawberry
butter	lettuce	tea	pasta	strawberries
cheese	bacon	fish	coffee	pie
cookie	salad	salad	pizzeria	wine
orange	chicken	chicken	turkish	cheesecake
Topic 16	Topic 17	Topic 18	Topic 19	Topic 20
soup	coffee	seafood	french	bbq
tea	wine	fish	wine	barbecue
salad	chocolate	shrimp	cheese	grill
tofu	bakery	lobster	chocolate	chicken
chicken	pastry	crab	bread	chili
bread	icecream	oyster	bistro	onion
beef	cream	chicken	orange	orange
Topic 21	Topic 22	Topic 23	Topic 24	Topic 25
vegetarian	chinese	thai	grill	mexican
vegan	noodles	rice	pizza	burger
tofu	chicken	curry	hotdog	hamburger
salad	fish	fish	icecream	fries
indian	duck	indian	hotdogs	cheeseburger
rice	pork	seafood	fries	tacos
beans	chopsticks	noodles	bbq	chicken

Table A.1: Topics detected by LGTA [YCH⁺11b] on the food dataset with 25 topics and 1000 regions. Multiple local topics for Italian (1, 14), Japanese (2, 6 and 8), Spanish (4), Indian (8, 21 and 23), French (19), Chinese (22), Thai (23) and Mexican (20) cuisine are detected. Words with a probability < 0.01 are not displayed. Because of the independence of regions, the Japanese, Indian and Italian cuisines are split into several distinct topics and in Topic 8, 23 and 25 multiple cuisines are mixed. This is due to the independence of geographical clusters.

Topic 1 slave master blindfolds oral sex talking dirty whips ass play hair pulling bondage	Topic 2 mind fucks crying fear interrogation wrestling somasochism tears emotional sadism violence	Topic 3 spreader bars orgasm denial begging blindfolds ben wa balls eye contact r... ¹ butt plugs crawling bare bottom s... ²	Topic 4 anal training anal stretching anal beads anal hooks strap ass play dildos anal fisting
Topic 5 creampie lactation breastfeeding breeding impregnation ... ⁴ incest play taboo bareback milking	Topic 6 nipple torture bondage leather whips high heels humiliation lingerie chains spanking	Topic 7 photography art erotica bondage art writing erotica tantra hot oil massages body paint chakra energy... ⁵ food play	Topic 8 making home m... ³ gangbangs pain dildos exhibitionism talking dirty ass play handcuffs chains
Topic 9 feminization sissification sissy training forced femini... ⁹ dollification cross dressing transformation human doll maid	Topic 10 anonymous enc... ⁶ group sex outdoor sex sex with stra... ¹⁰ sex in public public play orgy glory hole swinging	Topic 11 bbw bbw bondage forced mastur... ⁸ forced nudity forced submission nipple play blow jobs forced exhibi... ¹¹ guided mastur... ¹²	Topic 12 cock and ball... ⁷ candle wax electrotorture pain masochism humiliation strap high heels dildos
Topic 13 kissing fingering light bondage caressing handjobs blow jobs sex multiple orgasms sleepy sex	Topic 14 exhibitionism erotic photog... ¹³ piercings voyeurism candle wax nipple torture nipples bondage whips	Topic 15 sexual slavery degradation public humili... ¹⁴ slavery total power e... ¹⁵ objectification sexual object... ¹⁷ female humili... ¹⁸ humiliation	Topic 16 orgasm control orgasm denial teasing forced orgasms obedience tra... ¹⁶ edge play tease and denial begging cock milking

Table A.2: Fetish Topics by HMDP 1/7 ¹eye contact restrictions

²bare bottom spanking ³making home movies ⁴impregnation fantasy ⁵chakra energy play
⁶anonymous encounters ⁷cock and ball torture ⁸forced masturbation ⁹forced feminization
¹⁰sex with strangers ¹¹forced exhibitionism ¹²guided masturbation ¹³erotic photography
¹⁴public humiliation ¹⁵total power exchange ¹⁶obedience training ¹⁷sexual objectification
¹⁸female humiliation

Topic 17 caning belt spanking whipping corporal puni... ²¹ belt whippings padding riding crops discipline bullwhips	Topic 18 double penetr... ¹⁹ fucking machines vaginal stret... ²⁰ fisting triple penetr... ²² pussy pumping vacuum pumping glass dildos speculums	Topic 19 ass worship pussy worship body worship facesitting queening oral servitude boot worship foot worship female supremacy	Topic 20 cuckold forced deepthroat strapon female domination deepthroat forced bi cuckold humil... ²³ throat fucking strapon dildos
Topic 21 threesomes tattoos multiple orgasms double penetr... ²⁴ outdoor sex rough sex porn kissing cuddles	Topic 22 chastity devices chastity bondage discipline chains humiliation enforced chastity high heels male chastity	Topic 23 suspension rope bondage shibari japanese bondage kinbaku bondage art suspension bo... ²⁵ outdoor bondage breast bondage	Topic 24 bondage oral sex anal sex blindfolds role play candle wax diaper handcuffs hair pulling
Topic 25 master pain blow jobs multiple orgasms exhibitionism anal training spreader bars bare handed s... ²⁸ face fucking	Topic 26 catsuits spandex lycra zentai latex wam vacuum bed tights wet and messy	Topic 27 gagging cocksucking face fucking cock worship bukake cum swallowing blow jobs double penetr... ²⁹	Topic 28 spanking erotic photog... ²⁶ oral sex blindfolds talking dirty mutual mastur... ²⁷ voyeurism lingerie bondage
Topic 29 stockings panty hose mutual mastur... ³⁰ dildos high heels lingerie vibrators spanking erotic photog... ³⁵	Topic 30 me love moaning you a brilliant mind cuddles hugs cuddling and ... ³² screaming	Topic 31 cuckolding golden showers ball kicking chastity devices whipping cross dressing electrotorture sissy maid tr... ³³ cbt	Topic 32 cunnilingus bisexuality lesbian domin... ³¹ kissing switching face fucking butt plugs female ejacul... ³⁴ bondage

Table A.3: Fetish Topics by HMDP 2/7 ¹⁹double penetration ²⁰vaginal stretching ²¹corporal punishment ²²triple penetration ²³cuckold humiliation ²⁴double penetration ²⁵suspension bondage ²⁶erotic photography ²⁷mutual masturbation ²⁸bare handed spanking ²⁹double penetration ³⁰mutual masturbation ³¹lesbian domination ³²cuddling and conversation ³³sissy maid training ³⁴female ejaculation ³⁵erotic photography

Topic 33 female ejacul... ³⁶ squirting oral sex tongue sucking anal sex cum talking dirty tongues rough sex	Topic 34 ball gags collars gags strap spreader bars orgasm control sensory depri... ³⁸ nipple torture hair pulling	Topic 35 shemales transsexual transgender gay sex bimbofication gender play gay bondage strap bimbofication	Topic 36 femdom female supremacy male submission humiliation small penis h... ³⁷ cfnm cunt worship domestic serv... ³⁹ feminization
Topic 37 skirts with n... ⁴⁰ eating pussy doggy style r... ⁴¹ facial blowjobs pussy eating blow jobs deep throat blowjob	Topic 38 cbt ballbusting trampling ball stretching ball kicking cock slapping cock and ball... ⁴² genital torture castration	Topic 39 cross dressing transvestism latex lingerie high heels chastity devices rubber enemas leather	Topic 40 ass play rimming watersports toys oral sex anal sex gangbangs snowballing strap
Topic 41 big tits tit fucking breasts black women older women interracial sex asian sluts blow jobs	Topic 42 sensual domin... ⁴³ flirting humor intelligence sensual play seduction snuggling teasing touching	Topic 43 body modification piercings needle play tattoos blood play biting blood genital piercings knife play	Topic 44 cock milking prostate massage prostate milking medical play urethral sounds catheters strap shaving urethral fucking
Topic 45 oral sex masturbation strap talking dirty role play biting discipline lingerie ass play	Topic 46 power exchange discipline handcuffs humiliation bondage breath play sensory depri... ⁴⁴ pain role play	Topic 47 deep throating rough sex blow jobs play rape face fucking hair pulling talking dirty cmnf bare handed s... ⁴⁵	Topic 48 domination submission male authority discipline rough sex high heels begging talking dirty handcuffs

Table A.4: Fetish Topics by HMDP 3/7 ³⁶female ejaculation ³⁷small penis humiliation ³⁸sensory deprivation ³⁹domestic servitude ⁴⁰skirts with no panties ⁴¹doggy style rough and hard ⁴²cock and ball torture ⁴³sensual domination ⁴⁴sensory deprivation ⁴⁵bare handed spanking

Topic 49 shoes boot licking boot worship gloves boot smoking boots fetish wear high heels	Topic 50 spanking hair pulling role play sex in public oral sex discipline mutual mastur... ⁴⁶ biting piercings	Topic 51 masks master chains slave vibrators chastity devices role play sensory depri... ⁴⁷ electrotorture	Topic 52 foot worship foot massage feet toes and feet barefoot legs socks high heels foot
Topic 53 feet foot high heels humiliation lingerie leather forniphilia having my hai... ⁵¹ cock and ball... ⁵²	Topic 54 electrical play candle wax violet wand wartenberg pi... ⁴⁸ flogging needle play subspace rope clamps and clips	Topic 55 ass to mouth human toilet toilet slave spitting toilet armpits human ashtray face farting watersports	Topic 56 orgasm control play rape gags obedience tra... ⁴⁹ wax abduction play suspension bo... ⁵⁰ service degradation
Topic 57 sex in public masturbation biting lingerie voyeurism oral sex high heels mutual mastur... ⁵⁷ dildos	Topic 58 slave mistress femdom domination lesbian domin... ⁵⁵ yes male submission slavery cock and ball... ⁵⁸	Topic 59 blindfolds spanking breath play sensory depri... ⁵⁴ whips discipline handcuffs candle wax bondage	Topic 60 bare handed s... ⁵³ control hair pulling candle wax bare bottom s... ⁵⁶ slapping biting bondage attention
Topic 61 sadism pain masochism pinching whips nipple torture chains breath play spanking	Topic 62 intelligence consensual no... ⁵⁹ begging biting polyamory blindfolds sapiosexuality hair pulling talking dirty	Topic 63 red heads geeks cosplay hentai small tits anime glasses freckles goth	Topic 64 latex rubber masks leather fetish wear pvc gas masks vinyl gasmask

Table A.5: Fetish Topics by HMDP 4/7 ⁴⁶mutual masturbation ⁴⁷sensory deprivation ⁴⁸wartenberg pinwheels ⁴⁹obedience training ⁵⁰suspension bondage ⁵¹having my hair played with ⁵²cock and ball torture ⁵³bare handed spanking ⁵⁴sensory deprivation ⁵⁵lesbian domination ⁵⁶bare bottom spanking ⁵⁷mutual masturbation ⁵⁸cock and ball torture ⁵⁹consensual nonconsent

Topic 65 flogging wax bondage caning bdsm pain collar and leash breast and ni... ⁶² plugs	Topic 66 bondage corsets ball gags bare bottom s... ⁶⁰ chains whips bdsm butt plugs domination	Topic 67 piss pissing anal fisting enema ass licking golden showers vaginal fisting ass butt plug	Topic 68 you belong to me respect lust genuine and d... ⁶¹ yes when i want it no how i want it treat her lik... ⁶³
Topic 69 obedience tra... ⁶⁴ domestic serv... ⁶⁵ behavior modi... ⁶⁷ service bathroom use ... ⁶⁸ collars kneeling discipline eye contact r... ⁷¹	Topic 70 daddy daddy daughte... ⁶⁶ age play babygirl fuck me schoolgirl schoolgirl un... ⁷⁰ good girl incest play	Topic 71 biting scratching ice cubes leaving marks tearing off c... ⁶⁹ rough sex bruises teasing candle wax	Topic 72 diapers diaper diaper lover infantilism wearing diapers abdl plastic pants diaper punishment adult baby
Topic 73 anal butt plugs rimming anal sex blow jobs ass to mouth bdsm fingering bondage	Topic 74 breast bondage breast spanking clit spanking clit torture female humili... ⁷² tit slapping breast whipping nipple torture cunt torture	Topic 75 shaving clothespins clamps and clips fisting figging nipple torture face slapping caning breath play	Topic 76 watersports golden showers bald girls gay facesitting w... ⁷³ pee head shaving golden shower wetting
Topic 77 goth vampires victorian lif... ⁷⁴ accents sex in the ce... ⁷⁵ music fur martial arts muscles	Topic 78 age play role play ass play strap biting breath play erotic photog... ⁷⁶ fisting nipple torture	Topic 79 massages erotic literature cuddles tickling nudity outdoor sex foot tickling tickle torture blow jobs	Topic 80 pegging androgyny panties strap on pet play crossdressing hair cunnilingus blow jobs

Table A.6: Fetish Topics by HMDP 5/7 ⁶⁰bare bottom spanking ⁶¹genuine and deep submission ⁶²breast and nipple torture ⁶³treat her like a lady ⁶⁴obedience training ⁶⁵domestic servitude ⁶⁶daddy daughter roleplay ⁶⁷behavior modification ⁶⁸bathroom use control ⁶⁹tearing off clothing ⁷⁰schoolgirl uniform ⁷¹eye contact restrictions ⁷²female humiliation ⁷³facesitting watersports ⁷⁴victorian lifestyles ⁷⁵sex in the cemetery ⁷⁶erotic photography

Topic 81 handcuffs latex anal sex leather high heels blindfolds hair pulling bondage breath play	Topic 82 schoolgirl un... ⁷⁷ uniforms maid uniforms business suits french maids schoolgirl military uniforms stockings cheerleading ... ⁸¹	Topic 83 bare bottom s... ⁷⁸ otk spanking bondage hairbrush spa... ⁷⁹ role play hair pulling spanking candle wax cornertime	Topic 84 humiliation enemas anal sex ass play strap watersports discipline mutual mastur... ⁸⁰ chastity devices
Topic 85 anal sex nipple torture spanking oral sex talking dirty blindfolds ass play diaper rough sex	Topic 86 corsets burlesque high heels lingerie fishnets vintage lingerie skirts corsetry lace	Topic 87 nipples mutual mastur... ⁸² masturbation discipline lingerie dildos chains oral sex handcuffs	Topic 88 vibrators slave bisexuality sensual domin... ⁸³ spanking blindfolds swallowing oral sex talking dirty
Topic 89 kidnapping ro... ⁸⁴ abduction play bondage duct tape gags handcuffs damsels in di... ⁸⁶ pursuit ball gags	Topic 90 lipstick handjobs stockings skirts fingering flirting porn tit fucking breasts	Topic 91 mummification bondage equipment bondage tape armbinders vacuum bed suspension bo... ⁸⁵ hoods bondage posture collars	Topic 92 petplay puppy play hypnosis mind control pony play erotic hypnosis kitten petplay mental bondage human doll
Topic 93 toys strap vibrators dildos lingerie high heels mutual mastur... ⁸⁸ spanking masturbation	Topic 94 face slapping choking asphyxiaphilia rough sex talking dirty gagging play rape humiliation belt spanking	Topic 95 smothering face sitting cunt worship cunnilingus pussy worship strap golden showers rimming spitting	Topic 96 bdsm restraints handcuffs blindfolds domination obedience tra... ⁸⁷ collars leather submission

Table A.7: Fetish Topics by HMDP 6/7 ⁷⁷*schoolgirl uniform* ⁷⁸*bare bottom spanking* ⁷⁹*hairbrush spanking* ⁸⁰*mutual masturbation* ⁸¹*cheerleading uniforms* ⁸²*mutual masturbation* ⁸³*sensual domination* ⁸⁴*kidnapping roleplay* ⁸⁵*suspension bondage* ⁸⁶*damsels in distress* ⁸⁷*obedience training* ⁸⁸*mutual masturbation*

Topic 97	Topic 98	Topic 99	Topic 100
licking	fisting	cyber sex	slut
massage	anal sex	webcams	letting her up for air
being tied up	high heels	online play	and then doing it again
sucking	sex in public	sex online	being held down and fucked
roleplay	rimming	phone sex	skirt up
kissing	erotic photography	webcam	being used as a slut
silk	lingerie	blackmail	shut the fuck up and bend over
nibbling	humiliation	skype sex	good girl
blindfolded	talking dirty	online domination	forced deepthroat

Table A.8: Fetish Topics by HMDP 7/7. Terms with a probability smaller than 0.01 were greyed out. The topic distribution has a sparse Dirichlet prior which places a high probability on the first words of a topic and very low probabilities on the rest of the words. Therefore, words with a very low probability often are not meaningful for describing a topic.

Topic 1 high heels corsets latex erotic photog... ³ lingerie burlesque stockings tattoos art erotica	Topic 2 bondage shibari rope bondage/... ² japanese bondage blindfolds candle wax ball gags bondage art bdsm	Topic 3 blow jobs cocksucking deep throating tit fucking face fucking cum handjobs cyber sex fingering	Topic 4 shibari rope bondage/... ¹ japanese bondage bondage art suspension bo... ⁴ outdoor bondage spreader bars remote-contro... ⁵ kinbaku
Topic 5 bondage spanking blindfolds handcuffs collar and le... ⁹ discipline hair pulling master/slave oral sex	Topic 6 oral sex masturbation sex in public mutual mastur... ⁸ dildos anal sex vibrators voyeurism exhibitionism	Topic 7 lingerie masturbation erotic photog... ⁷ role play sex in public making home m... ¹⁰ pantyhose/sto... ¹¹ mutual mastur... ¹² talking dirty	Topic 8 pain breast/nipple... ⁶ sadism whips discipline candle wax breath play collar and le... ¹³ masochism
Topic 9 ball gags gags rope bondage/... ¹⁴ restraints spreader bars bondage equipment bondage tape duct tape outdoor bondage	Topic 10 anal sex ass play rimming oral sex watersports fisting strap-ons dildos toys	Topic 11 latex strap-ons cock and ball... ¹⁵ cross dressing chastity devices high heels humiliation leather foot/feet	Topic 12 sadism leather piercings masks electrotorture nipples pain masochism pinching
Topic 13 masturbation sex in public oral sex mutual mastur... ¹⁸ exhibitionism anal sex vibrators toys dildos	Topic 14 toys oral sex vibrators masturbation dildos anal sex spanking bondage talking dirty	Topic 15 ballbusting cock and ball... ¹⁷ femdom cbt ball kicking boot licking mistress/slave female supremacy cock slapping	Topic 16 breast/nipple... ¹⁶ fisting flogging deep throating figging face slapping clothespins biting whips

Table A.9: Fetish topics by LDA 1/7 ¹rope bondage/suspension
²rope bondage/suspension ³erotic photography ⁴suspension bondage ⁵remote-control de-
vices ⁶breast/nipple torture ⁷erotic photography ⁸mutual masturbation ⁹collar and
lead/leash ¹⁰making home movies ¹¹pantyhose/stockings ¹²mutual masturbation ¹³collar
and lead/leash ¹⁴rope bondage/suspension ¹⁵cock and ball torture ¹⁶breast/nipple torture
¹⁷cock and ball torture ¹⁸mutual masturbation

Topic 17 pain candle wax sadism breast/nipple... ¹⁹ whips breath play discipline biting collar and le... ²⁴	Topic 18 masturbation oral sex sex in public erotic photog... ²⁰ mutual mastur... ²² talking dirty lingerie voyeurism exhibitionism	Topic 19 intelligence tantra art erotica chakra energy... ²¹ sensual domin... ²³ caressing spiritual bdsm humor kissing	Topic 20 oral sex toys blindfolds vibrators masturbation spanking handcuffs anal sex lingerie
Topic 21 bare bottom s... ²⁵ spanking bare handed s... ²⁶ otk spanking belt spanking caning discipline paddling hairbrush spa... ²⁹	Topic 22 diapers age play humiliation watersports enemas spanking diaper bathroom use ... ²⁸ bondage	Topic 23 foot/feet high heels feet foot massage pantyhose/sto... ²⁷ barefoot toes and feet tickling kissing	Topic 24 bondage bdsm d/s flogging pain caning oral sex wax collar and leash
Topic 25 anal sex butt plugs anal ass to mouth fisting ass play rimming anal stretching anal training	Topic 26 face slapping face fucking verbal humili... ³⁰ rough sex gagging/choke... ³² deep throating humiliation hair pulling choking	Topic 27 orgasm control orgasm denial forced orgasms teasing bring-them-to... ³³ forced mastur... ³⁴ multiple orgasms remote-contro... ³⁵ tease and denial	Topic 28 cross dressing latex strap-ons pantyhose/sto... ³¹ high heels lingerie transvestism ass play leather
Topic 29 cock and ball... ³⁶ cbt ball stretching cock milking prostate milking prostate massage urethral sounds electrotorture electrical play	Topic 30 anal sex ass play rimming watersports oral sex fisting strap-ons age play master/slave	Topic 31 being more co... ³⁷ causing peopl... ³⁸ finding out i... ³⁹ intelligence cuddles spelling and punctuation kissing girls with hi... ⁴⁰	Topic 32 lingerie bondage toys oral sex handcuffs spanking vibrators high heels blindfolds

Table A.10: Fetish topics by LDA 2/7 ¹⁹breast/nipple torture ²⁰erotic photography ²¹chakra energy play ²²mutual masturbation ²³sensual domination ²⁴collar and lead/leash ²⁵bare bottom spanking ²⁶bare handed spanking ²⁷pantyhose/stockings ²⁸bathroom use control ²⁹hairbrush spanking ³⁰verbal humiliation and degradation ³¹pantyhose/stockings ³²gagging/choked by cock ³³bring-them-to-the-edge-of-orgasm-but-don't-let-them-cum-for-a-while ³⁴forced masturbation ³⁵remote-control devices ³⁶cock and ball torture ³⁷being more complex than an anonymous list of fetishes could show ³⁸causing people to have to actually converse with me ³⁹finding out if they like *me* and not just what gets me off. ⁴⁰girls with high iqs and low morals

Topic 33 tattoos piercings goth vampires body modification biting genital piercings bisexuality sex in the ce... ⁴⁴	Topic 34 foot worship foot/feet foot massage face sitting/... ⁴² ass worship femdom body worship boot licking boot worship	Topic 35 bdsm bondage domination collar and le... ⁴³ collars ball gags restraints submission obedience tra... ⁴⁵	Topic 36 intelligence sensual domin... ⁴¹ kissing domination orgasm control teasing erotic literature light bondage rough sex
Topic 37 intelligence biting scratching corsets polyamory sapiosexuality sensory depri... ⁴⁶ leaving marks shibari	Topic 38 face slapping crying cutting fear violence bruises asphyxiaphilia choking leaving marks	Topic 39 bondage oral sex spanking blindfolds hair pulling biting anal sex handcuffs high heels	Topic 40 spanking bondage blindfolds discipline anal sex oral sex hair pulling collar and le... ⁴⁷ breast/nipple... ⁴⁸
Topic 41 collar and le... ⁴⁹ caging/confin... ⁵¹ chastity devices humiliation obedience tra... ⁵⁴ discipline master/slave bondage 24/7	Topic 42 masturbation sex in public mutual mastur... ⁵² oral sex voyeurism exhibitionism erotic photog... ⁵⁶ toys making home m... ⁵⁸	Topic 43 schoolgirl un... ⁵⁰ teacher/student schoolgirl role play cosplay kitten petplay anime hentai petplay	Topic 44 whips chastity devices discipline breast/nipple... ⁵³ leather caging/confin... ⁵⁵ cock and ball... ⁵⁷ latex bondage
Topic 45 making you do... ⁵⁹ when i want it how i want it and you will ... ⁶⁰ moaning screaming rough sex groaning and ... ⁶³ scratches	Topic 46 vintage lingerie burlesque corsets lace costumes/dres... ⁶¹ stockings lipstick fishnets lingerie	Topic 47 teasing kissing cuddles caressing snuggling ice cubes massages submission touching	Topic 48 anal sex ass play bondage spanking humiliation fisting breast/nipple... ⁶² oral sex rimming

Table A.11: Fetish topics by LDA 3/7 ⁴¹sensual domination ⁴²face sitting/smothering ⁴³collar and lead/leash ⁴⁴sex in the cemetery ⁴⁵obedience training ⁴⁶sensory deprivation ⁴⁷collar and lead/leash ⁴⁸breast/nipple torture ⁴⁹collar and lead/leash ⁵⁰schoolgirl uniform ⁵¹caging/confinement ⁵²mutual masturbation ⁵³breast/nipple torture ⁵⁴obedience training ⁵⁵caging/confinement ⁵⁶erotic photography ⁵⁷cock and ball torture ⁵⁸making home movies ⁵⁹making you do whatever the fuck i feel like ⁶⁰and you will like it because i like it ⁶¹costumes/dressing-up ⁶²breast/nipple torture ⁶³groaning and other sounds of pleasure and pain

Topic 49 pain whips breast/nipple... ⁶⁶ sadism candle wax chains discipline sensory depri... ⁷⁶ masochism	Topic 50 corsets leather gloves high heels uniforms stockings riding crops costumes/dres... ⁷⁷ domination	Topic 51 being treated... ⁶⁴ “cum for me being fucked ... ⁶⁷ a fistful of ... ⁶⁹ being told i’... ⁷¹ slut” “you belong t... ⁷⁴ writhing in h... ⁷⁸ being told “y... ⁷⁹	Topic 52 aren’t you?” “you’re a fil... ⁶⁵ treat her lik... ⁶⁸ fuck her like... ⁷⁰ being treated... ⁷² ”skirts with ... ⁷³ i’m going to ... ⁷⁵ “if i catch you being pushed ... ⁸⁰
Topic 53 candle wax spanking bondage hair pulling bare handed s... ⁸¹ breath play tickling blindfolds biting	Topic 54 clit spanking breast bondage breast spanking clit torture tit slapping clit pumping breast whipping cunt torture female ejacul... ⁸⁴	Topic 55 doctor/nurse play masks foot/feet rubber cross dressing scent enemas cock and ball... ⁸² pantyhose/sto... ⁸⁵	Topic 56 threesomes group sex sex in public exhibitionism outdoor sex voyeurism bisexuality double penetr... ⁸³ erotic photog... ⁸⁶
Topic 57 face sitting/... ⁸⁷ ass worship mistresses wi... ⁸⁸ femdom cunnilingus pussy worship cunt worship strap-ons mistress/slave	Topic 58 latex rubber mummification ball gags ballet boots/... ⁸⁹ vacuum bed hoods gas masks masks	Topic 59 bondage chastity devices spanking high heels discipline corsets golden showers whipping caning	Topic 60 rough sex hair pulling play rape choking face slapping domination blow jobs kissing deep throating
Topic 61 strap-ons cross dressing mistresses wi... ⁹¹ shemales bisexuality transgender transexual transvestism pegging	Topic 62 rough sex deep throating face fucking blow jobs gagging/choke... ⁹² anal sex play rape double penetr... ⁹⁴ threesomes	Topic 63 bondage spanking blindfolds discipline collar and le... ⁹³ candle wax master/slave handcuffs hair pulling	Topic 64 female ejacul... ⁹⁰ squirting cunnilingus deep throating blow jobs multiple orgasms rough sex anal threesomes

Table A.12: Fetish topics by LDA 4/7 ⁶⁴being treated like a beautiful princess but fucked like a dirty little whore ⁶⁵“you’re a filthy little slut ⁶⁶breast/nipple torture ⁶⁷being fucked with a hand on my throat and threats being whispered into my ear ⁶⁸treat her like a lady ⁶⁹a fistful of hair and a long passionate kiss ⁷⁰fuck her like a slut ⁷¹being told i’m a good girl ⁷²being treated like a beautiful princess but fucked like a dirty little whore ⁷³“skirts with no panties” ⁷⁴“you belong to me” whispered in my ear ⁷⁵i’m going to fuck you” ⁷⁶sensory deprivation ⁷⁷costumes/dressing-up ⁷⁸writhing in his arms and struggling as he whispers everything he’s going to do in my ear ⁷⁹being told “you’re mine” ⁸⁰being pushed up against a wall in a passionate kiss ⁸¹bare handed spanking ⁸²cock and ball torture ⁸³double penetration ⁸⁴female ejaculation ⁸⁵pantyhose/stockings ⁸⁶erotic photography ⁸⁷face sitting/smothering ⁸⁸mistresses with strap-ons ⁸⁹ballet boots/shoes ⁹⁰female ejaculation ⁹¹mistresses with strap-ons ⁹²gagging/choked by cock ⁹³collar and lead/leash ⁹⁴double penetration

Topic 65 rough sex bare bottom s... ⁹⁶ bare handed s... ⁹⁷ hair pulling skirt up play rape panties down belt spanking “sit the fuck... ¹⁰³	Topic 66 verbal humili... ⁹⁵ humiliation public humili... ⁹⁸ degradation obedience tra... ¹⁰⁰ sexual slavery exhibitionism objectification female humili... ¹⁰⁴	Topic 67 bukake gangbangs group sex sex with stra... ⁹⁹ double penetr... ¹⁰¹ cocksucking creampie cum glory hole	Topic 68 latex leather bondage masks rubber collar and le... ¹⁰² high heels handcuffs chains
Topic 69 sex in public anal sex licking threesomes rough sex oral sex fingering “skirts with ... ¹⁰⁸ kissing	Topic 70 domination orgasm control ball gags collars bdsm obedience tra... ¹⁰⁷ butt plugs bare bottom s... ¹⁰⁹ restraints	Topic 71 obedience tra... ¹⁰⁵ mind control orgasm control mind fucks mental bondage hypnosis d/s behavior modi... ¹¹⁰ control	Topic 72 femdom mistress/slave mistresses wi... ¹⁰⁶ strap-ons chastity devices male submission chastity cuckold face sitting/... ¹¹¹
Topic 73 obedience tra... ¹¹² domestic serv... ¹¹³ slavery 24/7 sexual slavery petplay verbal humili... ¹¹⁹ service-orient... ¹²¹ behavior modi... ¹²³	Topic 74 pursuit take-down & ... ¹¹⁴ abduction play kidnapping ro... ¹¹⁵ consensual no... ¹¹⁸ play rape predator/prey interrogation tearing off c... ¹²⁴	Topic 75 sex in public masturbation voyeurism erotic photog... ¹¹⁶ exhibitionism oral sex mutual mastur... ¹²⁰ making home m... ¹²² talking dirty	Topic 76 caning flogging whipping wartenberg pi... ¹¹⁷ riding crops single tail whips electrical play violet wand bullwhips
Topic 77 cross dressing feminization sissification transvestism forced femini... ¹²⁹ maid uniforms mistresses wi... ¹³² french maids costumes/dres... ¹³⁷	Topic 78 a collar cuddles being pushed ... ¹²⁵ two wrist cuffs two ankle cuf... ¹³⁰ hugs “trust me you will be c... ¹³⁵ being hugged ... ¹³⁸	Topic 79 mutual respect kissing cuddles trust and com... ¹²⁷ caressing love sensual domin... ¹³³ massages intelligence	Topic 80 respect lust the heartfelt... ¹²⁶ and mental an... ¹²⁸ genuine and d... ¹³¹ a brilliant mind a creative pl... ¹³⁴ intelligent c... ¹³⁶ subtlety

Table A.13: Fetish topics by LDA 5/7 ⁹⁵verbal humiliation and degradation ⁹⁶bare bottom spanking ⁹⁷bare handed spanking ⁹⁸public humiliation ⁹⁹sex with strangers ¹⁰⁰obedience training ¹⁰¹double penetration ¹⁰²collar and lead/leash ¹⁰³“sit the fuck down ¹⁰⁴female humiliation ¹⁰⁵obedience training ¹⁰⁶mistresses with strap-ons ¹⁰⁷obedience training ¹⁰⁸“skirts with no panties” ¹⁰⁹bare bottom spanking ¹¹⁰behavior modification ¹¹¹face sitting/smothering ¹¹²obedience training ¹¹³domestic servitude ¹¹⁴take-down & capture ¹¹⁵kidnapping roleplay ¹¹⁶erotic photography ¹¹⁷wartenberg pinwheels ¹¹⁸consensual non-consent ¹¹⁹verbal humiliation and degradation ¹²⁰mutual masturbation ¹²¹service-oriented submission ¹²²making home movies ¹²³behavior modification ¹²⁴tearing off clothing ¹²⁵being pushed up against a wall in a passionate kiss ¹²⁶the heartfelt kind that comes from trust ¹²⁷trust and communication. ¹²⁸and mental and emotional connection ¹²⁹forced feminization ¹³⁰two ankle cuffs and a smile ¹³¹genuine and deep submission ¹³²mistresses with strap-ons ¹³³sensual domination ¹³⁴a creative player and imaginative lover. ¹³⁵you will be crying your eyes out long before this spanking is over.” ¹³⁶intelligent conversation ¹³⁷costumes/dressing-up ¹³⁸being hugged by strong safe arms

Topic 81 bondage hair pulling spanking blindfolds discipline handcuffs whips pain collar and le... ¹⁴⁶	Topic 82 lingerie high heels schoolgirl un... ¹⁴⁰ stockings pantyhose/sto... ¹⁴² corsets costumes/dres... ¹⁴⁴ blow jobs butt plugs	Topic 83 bare bottom s... ¹³⁹ belt spanking butt plugs breast spanking bare handed s... ¹⁴³ ball gags breast/nipple... ¹⁴⁵ clit spanking breast bondage	Topic 84 kissing cunnilingus massages female ejacul... ¹⁴¹ blow jobs fingering multiple orgasms cuddles oral sex
Topic 85 breast/nipple... ¹⁴⁷ pain electrotorture chastity devices whips cock and ball... ¹⁵⁰ caging/confin... ¹⁵² sadism chains	Topic 86 age play daddy/girl daddy daughte... ¹⁴⁸ bare bottom s... ¹⁴⁹ role play schoolgirl un... ¹⁵¹ spanking anal blow jobs	Topic 87 needle play knife play biting blood play breath play scratching fire play candle wax piercings	Topic 88 watersports golden showers human toilet rimming toilet slave ass to mouth ass worship face sitting/... ¹⁵³ spitting
Topic 89 bondage blindfolds spanking handcuffs oral sex master/slave discipline hair pulling collar and le... ¹⁵⁸	Topic 90 “i’m not asking i’m telling.” “look me in t... ¹⁵⁴ not an option being a priority you think you... ¹⁵⁵ why are you s... ¹⁵⁶ thats so cute talking in a ... ¹⁵⁹	Topic 91 biting tattoos kissing cuddles rough sex teasing massages scratching corsets	Topic 92 forced depththroat depththroat cocksucking cum whore slut being fucked ... ¹⁵⁷ anal fisting “skirts with ... ¹⁶⁰
Topic 93 face fucking deep throating gagging/choke... ¹⁶¹ forcing her d... ¹⁶³ letting her u... ¹⁶⁵ and then doin... ¹⁶⁷ face slapping forced depththroat blow jobs	Topic 94 anal stretching anal training anal beads anal hooks butt plugs fisting anal enemas vaginal stret... ¹⁶⁸	Topic 95 hair pulling rough sex bare bottom s... ¹⁶² consensual no... ¹⁶⁴ bare handed s... ¹⁶⁶ play rape begging intelligence belt spanking	Topic 96 latex rubber leather pvc masks catsuits fetish wear high heels gas masks

Table A.14: Fetish topics by LDA 6/7 ¹³⁹*bare bottom spanking* ¹⁴⁰*schoolgirl uniform* ¹⁴¹*female ejaculation* ¹⁴²*pantyhose/stockings* ¹⁴³*bare handed spanking* ¹⁴⁴*costumes/dressing-up* ¹⁴⁵*breast/nipple torture* ¹⁴⁶*collar and lead/leash* ¹⁴⁷*breast/nipple torture* ¹⁴⁸*daddy daughter roleplay* ¹⁴⁹*bare bottom spanking* ¹⁵⁰*cock and ball torture* ¹⁵¹*schoolgirl uniform* ¹⁵²*caging/confinement* ¹⁵³*face sitting/smothering* ¹⁵⁴*“look me in the eye while i’m hurting you”* ¹⁵⁵*you think you have a choice* ¹⁵⁶*why are you so wet?* ¹⁵⁷*being fucked in the ass by a girl* ¹⁵⁸*collar and lead/leash* ¹⁵⁹*talking in a sweet voice while you’re doing something really mean* ¹⁶⁰*“skirts with no panties”* ¹⁶¹*gagging/choked by cock* ¹⁶²*bare bottom spanking* ¹⁶³*forcing her down on your cock til she gags* ¹⁶⁴*consensual nonconsent* ¹⁶⁵*letting her up for air* ¹⁶⁶*bare handed spanking* ¹⁶⁷*and then doing it again* ¹⁶⁸*vaginal stretching*

Topic 97	Topic 98	Topic 99	Topic 100
bondage	breast/nipple... ¹⁶⁹	red heads	bondage
spanking	pain	cyber sex	discipline
blindfolds	spanking	webcams	spanking
oral sex	whips	small tits	blindfolds
lingerie	candle wax	online play	humiliation
handcuffs	toys	tit fucking	collar and le... ¹⁷⁰
high heels	bondage	big tits	handcuffs
hair pulling	discipline	older women	exhibitionism
anal sex	nipples	sex online	master/slave

Table A.15: Fetish topics by LDA 7/7 The topics differ from the topics detected by the HMDP topic model. ¹⁶⁹*breast/nipple torture* ¹⁷⁰*collar and lead/leash*

CURRICULUM VITAE

CONTACT INFORMATION

Name | Christoph Carl Kling
Email | ckling at uni-koblenz.de

SCIENTIFIC INTERESTS

Probabilistic models with a focus on topic models
Context-dependent data mining and applications
Computational sociology

EDUCATION AND PROFESSIONAL EXPERIENCE

since 2015 | PhD candidate, Department for Computational Social Science, GESIS - Leibniz Institute for the Social Sciences, Cologne, Germany
2011-2015 | PhD candidate, Institute for Web Science and Technologies, University of Koblenz-Landau, Germany
2005 – 2010 | Diplom "with distinction" (grade 1.1) in Computer Science with a minor in Business Informatics, University of Koblenz-Landau, Germany
2005 | Abitur (grade 2.1), Rabanus-Maurus-Gymnasium, Mainz, Germany
2004 – 2008 | Freelancer in web development

AWARDS

2015 | Honourable Mention Award, *ICWSM 2015, Oxford*

PUBLICATIONS

2016 | Christoph Carl Kling, Lisa Posch, Arnim Bleier, and Laura Dietz. Topic model tutorial: A basic introduction on latent Dirichlet allocation and extensions for web scientists. In *Proceedings of the 8th ACM Conference on Web Science, WebSci '16*, pages 10–10, New York, NY, USA, 2016. ACM
2016 | D. Fay, H. Haddadi, M. C. Seto, H. Wang, and C. C. Kling. An exploration of fetish social networks and communities. In *NetSciX'16*, January 2016
2015 | Christoph C. Kling, Jérôme Kunegis, Heinrich Hartmann, and Markus Strohmaier. Voting behaviour and power in online democracy: A study of liquidfeedback in germany's pirate party. In *ICWSM'15*, 2015
2014 | Christoph C. Kling, Jerome Kunegis, Sergej Sizov, and Steffen Staab. Detecting non-gaussian geographical topics in tagged photo collections. In *WSDM*, 2014
2011 | Christoph Carl Kling, Sergej Sizov, and Steffen Staab. Virtual field research with social media: A pilot case of biometeorology. In *ACM WebSci'11*, pages 1–2, June 2011. WebSci Conference 2011
2011 | Christoph Carl Kling and Thomas Gottron. Detecting culture in coordinates: Cultural areas in social media. In *DETECT'11: Proceedings of the International Workshop on DETecting and Exploiting Cultural diversity on the Social Web*, 2011

SKILLS

Teaching | Tutorials for machine learning and data mining, data science, information retrieval
Programming | Java, Octave / Matlab, Python, R, PHP
Data Analysis | Experience in developing and implementing data mining applications and statistical methods
Social | Intercultural experience
Organizational | Basic knowledge in project management, corporate management and leadership

INVITED TALKS

2015 | Voting Behaviour and Power in Online Democracy: A Study of LiquidFeedback in Germany's Pirate Party, *Chair of Systems Design, ETH Zurich*
2012 | Memetic Topic Models, *Research Group on Web Science, University of Freiburg*
2012 | Topic Detection for Networked Data, *AI on the Web 2012, Workshop at the 35th Annual German Conference on Artificial Intelligence (KI 2012)*
2012 | Topic Detection Using Context Information, *GESIS – Leibniz Institute for the Social Sciences, Cologne*

LANGUAGES

German		Native
English		Fluent
Ancient Greek		<i>Graecum</i> certificate
Latin		<i>Latinum</i> certificate

EXTRACURRICULAR ACTIVITIES

2014		Candidate of the Pirate Party (second list position) in the election for the city council Mainz
2013		Founding member of Intaktiv e.V. (association for genital integrity of minors)
2011		Founding member of GBS Mainz / Rhenish Hesse e.V. (a humanist philosophy club)

