International Journal of Semantic Computing

Exploring Symmetrical and Asymmetrical Dirichlet Priors for Latent Dirichlet Allocation

Shaheen Syed

Department of Information and Computing Sciences, Utrecht University, Princetonplein 5, 3584 CC Utrecht, Netherlands
s.a.s.syed@uu.nl

Marco Spruit

Department of Information and Computing Sciences, Utrecht University, Princetonplein 5, 3584 CC Utrecht, Netherlands
m.r.spruit@uu.nl

Received (Day Month Year)
Revised (Day Month Year)
Accepted (Day Month Year)

Latent Dirichlet Allocation (LDA) has gained much attention from researchers and is increasingly being applied to uncover underlying semantic structures from a variety of corpora. However, nearly all researchers use symmetrical Dirichlet priors, often unaware of the underlying practical implications that they bear. This research is the first to explore symmetrical and asymmetrical Dirichlet priors on topic coherence and human topic ranking when uncovering latent semantic structures from scientific research articles. More specifically, we examine the practical effects of several classes of Dirichlet priors on 2000 LDA models created from abstract and full-text research articles. Our results show that symmetrical or asymmetrical priors on the document–topic distribution or the topic–word distribution for full-text data have little effect on topic coherence scores and human topic ranking. In contrast, asymmetrical priors on the document–topic distribution for abstract data show a significant increase in topic coherence scores and improved human topic ranking compared to a symmetrical prior. Symmetrical or asymmetrical priors on the topic–word distribution show no real benefits for both abstract and full-text data.

Keywords: Topic models; coherence scores; human topic ranking

1. Introduction

Global research efforts have led to an ever-increasing amount of scientific output. Combined with the digitalization of scientific archives, this increase is threatening to overwhelm today's scientists trying to keep track of and identify relevant literature [1]. Consequently, scientists need new tools and algorithms for browsing these collections in a structured way, particularly as topics within articles, which are the ideas contained within articles that can be shared among similar articles, cannot always be detected through traditional keyword searches [2]. Probabilistic

topic models such as latent Dirichlet allocation (LDA) [3] and probabilistic latent semantic indexing (pLSI) [4] are machine-learning algorithms used to automatically uncover underlying semantic structures, such as themes or topics, in large collections of documents. These underlying semantic structures can subsequently be used to categorize, summarize, and annotate large document collections in a purely unsupervised fashion.

LDA, although the simplest topic model, has received much attention from machine-learning researchers and has been adopted and extended in many ways. LDA is a generative probabilistic topic model that aims to uncover hidden or latent thematic structures from large collections of documents. LDA is a three-level hierarchical Bayesian model that models documents as discrete distributions over K latent topics, and every topic is modeled as a multinomial distribution over the fixed vocabulary. Uncovering latent thematic structures proceeds through posterior inference of the latent variables given the observed words. Apart from its applicability to text, LDA has proven useful to other types of data, such as image [5], video [6], and audio [7].

As a conjugate prior to the multinomial distribution, LDA uses a Dirichlet prior to simplify posterior inference. Typically, these priors and related hyperparameters are set to be symmetrical, assuming that *a priori* all topics have equal probability to be assigned to a document and all words have an equal chance to be assigned to a topic. The reasons for choosing symmetrical priors, compared to asymmetrical priors, are not explicitly stated and are often implicitly assumed to have little or no practical effect [8]. However, hyperparameters can have a significant effect on the achieved accuracy for various inference techniques, such as Gibbs sampling, variational Bayes, or collapsed variational Bayes [9]. In fact, inference methods have relatively similar predictive performance when the hyperparameters are optimized, thereby explaining away most differences between them.

Little research has examined the effects of Dirichlet priors on the quality of generated topics. Among the few, Wallach *et. al.* [8] demonstrated that using an asymmetric Dirichlet prior on the document–topic distribution shows significant performance gains concerning the likelihood of held-out documents. However, the likelihood correlates negatively with human interpretability [10], which is often considered the gold standard for topic quality. Consequently, researchers have proposed topic coherence measures [11, 12, 13, 14], a proxy for topic quality that shows improved correlation with human topic ranking data. The underlying idea of topic coherence is rooted in the distributional hypothesis of linguistics [15]—namely, words with similar meanings tend to occur in similar contexts. This paper is the first to explore the practical effects of several classes of Dirichlet priors on the coherence of generated topics. More specifically, we study topic coherence for the combinations of symmetrical and asymmetrical priors on the document–topic distribution, as well as the topic–word distribution, when uncovering latent topics with LDA. In addition, topics are ranked by a domain expert on interpretability, providing a qual-

itative analysis of topic quality for different classes of Dirichlet priors in addition to a quantitative measure. Such analyses can provide valuable guidance to researchers utilizing LDA tools such as Mallet and Gensim [16] to uncover topical structures from scientific articles [17, 18, 19, 20, 21] and unknowingly leaving hyperparameters set to default (i.e. symmetrical).

2. Background

2.1. *Latent Dirichlet Allocation*

LDA is a generative probabilistic topic model that aims to uncover latent semantic structures from a set of documents, D. The latent semantic structures can subsequently be used to organize, categorize, and annotate documents without the need of prior human labeling or annotation. LDA models documents as discrete distributions over K latent topics, and every topic is modeled as a discrete distribution over the fixed vocabulary. As a result, LDA captures the heterogeneity of ideas prevailing in a document collection and can be viewed as a mixed membership model [22]. The underlying latent semantic structure is expressed by topics β, topic proportions θ, and topic assignments z and includes hidden variables that LDA posits into the corpus. However, β, θ, and z are unobserved, and the goal is to determine them from the observed variables (i.e. the words within the documents). LDA's structure allows the observed variables to interact with structured distributions of a hidden variable model [23]. Learning the hidden variables (i.e. the underlying semantic structure) can be achieved by inferring the posterior distribution of the latent variables given the observed documents. The interaction between latent and observed variables is manifested in the generative process behind LDA, the imaginary random process in which we assume the documents come from and are based on probabilistic sampling rules. The generative process is described as follows:

1. For every topic $k = \{1, ..., K\}$
   (1) draw a distribution over the vocabulary V, $\beta_k \sim \text{Dir}(\eta)$
2. For every document d
   (1) draw a distribution over topics, $\theta_d \sim \text{Dir}(\alpha)$ (i.e. per-document topic proportion)
   (2) for each word w within document d
      i. draw a topic assignment, $z_{d,n} \sim \text{Mult}(\theta_d)$, where $z_{d,n} \in \{1, ..., K\}$ (i.e. per-word topic assignment)
      ii. draw a word $w_{d,n} \sim \text{Mult}(\beta_{z_{d,n}})$, where $w_{d,n} \in \{1, ..., V\}$

Where K is the numbers of topics, V is the vocabulary size, and α and η are the Dirichlet hyperparameters that affect the smoothing of topic proportions within documents and words within topics, respectively. The joint distribution of all the hidden and observed variables becomes:

$$p(\beta_K, \theta_D, z_D, w_D|\alpha, \eta) = \prod_{k=1}^{K} p(\beta_K|\eta) \prod_{d=1}^{D} p(\theta_d|\alpha) \prod_{n=1}^{N} p(z_{d,n}|\theta_d) p(w_{d,n}|z_{d,n}, \beta_{d,k}) \tag{1}$$

To learn the distribution of the hidden variables, we invert the generative process and fit the hidden variables onto the observed words. The hidden structure is thus described by the posterior distribution of the latent variables given the observed words:

$$p(\beta_K, \theta_D, z_D|w_D, \alpha, \eta) = \frac{p(\beta_K, \theta_D, z_D, w_D|\alpha, \eta)}{p(w_D|\alpha, \eta)} \tag{2}$$

$$p(w_D|\alpha, \eta) = \int_{\beta_K} \int_{\theta_D} p(w_d|\alpha, \eta) \tag{3}$$

However, the posterior is intractable to compute [3] due to the evidence as expressed in (3). The solution is to approximate the posterior using inference techniques. Two main posterior inference techniques can be discerned: (i) sampling-based algorithms [24, 25] and (ii) variational- or optimization-based algorithms [26, 27, 28]. Sampling-based algorithms, such as Markov Chain Monte Carlo (MCMC) sampling, sample from the posterior—usually one variable at a time—while fixing the other variables. Repeating this process for several iterations causes the process to converge, in which the sample values have the same distribution as if they came from the true posterior. Variational inference aims to find a simplified parametric distribution that is closest to the true posterior measured in the Kullback-Leibler (KL) divergence. Once inference is complete, the posterior distribution reveals the latent structure of the documents expressed by topics β, topic proportions θ, and topic assignments z.

One way to think about LDA is to imagine a document in which one highlights words with colored markers. Words that relate to one topic are colored blue, words that relate to another topic are colored red, and so on. After all of the words have been colored (excluding words such as "the", "a"), all the words with the same color are the topics, and the article will blend the colors in different proportions. Different documents will have different blends of colors, and we could use the proportion of the various colors to situate this specific document in a document collection (e.g. documents addressing mainly the blue topic). Moreover, documents with the same blend of colors discuss the topics in similar proportion and are considered closely related from a topical perspective. Technically, documents with similar topic distributions are close in Kullback-Leibler divergence, a measure to calculate the distance between two probability distributions. LDA as a statistical model captures this intuition. We refer the interested reader to [29] for a concise introduction to LDA.

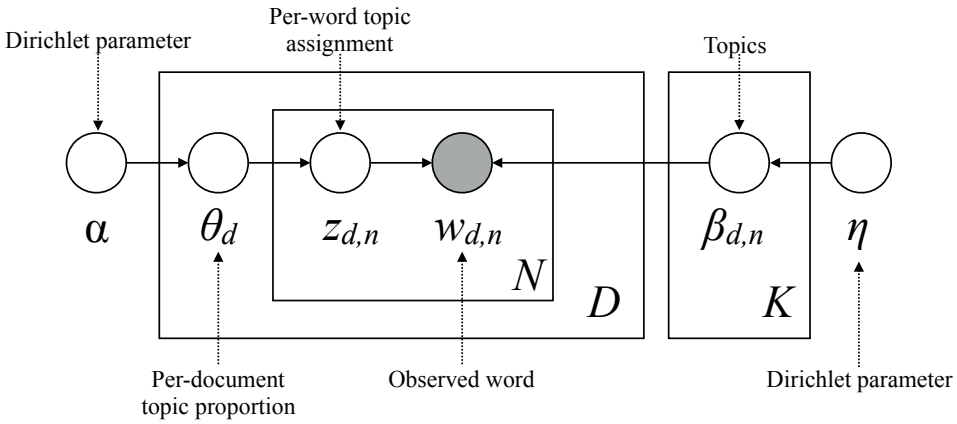


Fig. 1: LDA represented as a graphical model in which the nodes denote the random variables and the edges the dependencies between them. Unshaded nodes are unobserved or hidden variables, and the shaded nodes represent the observed random variables. The boxes, called plates, indicate replication.

2.2. *Research Utilizing LDA*

Topic modeling algorithms, and specifically LDA, have been helpful in elucidating the key ideas within a set of documents, such as articles published in the journal PNAS [30], political science texts [31] or data-driven journalism [32]. Moreover, it is considered that this approach could provide insight into the development of a scientific field and changes in research priorities [33], and do so with greater speed and quantitative rigor than would otherwise be possible through traditional narrative reviews [31]. As such, LDA has been applied, for example, in the domain of transportation research [18], computer science [34, 28, 20], fisheries science [35, 36], conservation science [19], and the fields of operations research and management science [17].

2.3. *Coherence Scores*

Measures such as predictive likelihood on held-out data [37] have been proposed to evaluate the quality of generated topics. However, such measures correlate negatively with human interpretability [10], making topics with high predictive likelihood less coherent from a human perspective. High-quality or coherent latent topics are of particular importance when they are used to browse document collections or understand the trends and development within a particular research field. As a result, researchers have proposed topic coherence measures, which are a qualitative approach to automatically uncover the coherence of topics[11, 12, 13, 14]. Topics are

considered to be coherent if all or most of the words (e.g. a topic's top-N words) are related. Topic coherence measures aim to find measures that correlate highly with human topic evaluation, such as topic ranking data obtained by, for example, word and topic intrusion tests [10]. Human topic ranking data are often considered the gold standard and, consequently, a measure that correlates well is a good indicator for topic interpretability. A recent study by Röder *et. al.* [14] systematically and empirically explored the multitude of topic coherence measures and their correlation with available human topic ranking data; new coherence measures obtained by combining existing elementary elements were also examined. The researchers' systematic approach revealed a new unexplored coherence measure, which they labeled C_V, to achieve the highest correlation with all available human topic ranking data. This study adopts the C_V coherence measure for calculating topic coherence, with a detailed description of the calculations behind this measure described below.

The calculation of C_V starts with the segmentation of the topic's top-N words into pairs of word subsets, $S_i = (W', W^*)$, where $W' \in W$, $W^* \in W$, and W consists of the topic's top-N most probable words. More formally, a pair S is defined as:

$$S = \{(W', W^*) | W' = \{w_i\}; w_i \in W; W^* = W\} \tag{4}$$

For example, if $W = \{w_1, w_2, w_3\}$, then one pair might be $S_i = (W' = w_1), (W^* = w_1, w_2, w_3)$. Such segmentation measures the extent to which the subset W^* supports or conversely undermines the subset W' [38]. The support between word subsets of a pair $S_i = (W', W^*)$ is calculated with a confirmation measure ϕ. C_V uses an indirect confirmation measure that considers not only the words within a pair, but also all words in W. A direct confirmation measure, such as difference, ratio, and likelihood measure, could place a low probability on high-support but low-frequency pairs. An indirect confirmation measure overcomes this by pairing every subset with W, thereby increasing the semantic support of supporting pairs. Word subsets are now represented as context vectors [11], such as $\vec{v}(W')$ by pairing them to all words in W, as exemplified in (5). The relatedness between context vectors and words in W is calculated by normalized pointwise mutual information (NPMI), as shown in (6).

$$\vec{v}(W') = \left\{ \sum_{w_i \in W'} \text{NPMI}(w_i, w_j)^\gamma \right\}_{j=1,\ldots,|W|} \tag{5}$$

$$\text{NPMI}(w_i, w_j)^\gamma = \left(\frac{\log \frac{P(w_i, w_j) + \epsilon}{P(w_i) \cdot P(w_j)}}{-\log(P(w_i, w_j) + \epsilon)} \right)^\gamma \tag{6}$$

In contrast to pointwise mutual information, NPMI achieves a higher correlation with human topic ranking data [11], which is generally a result of reducing the impact of low-frequency counts in word co-occurrences [39]. Given our

running example of $W = \{w_1, w_2, w_3\}$, we obtain the context vector for w_1 as $\vec{w_1} = \{\text{NPMI}(w_1, w_1)^\gamma, \text{NPMI}(w_1, w_2)^\gamma, \text{NPMI}(w_1, w_3)^\gamma\}$, with the constant ϵ to prevent logarithms of zero, and γ to place more weight on higher NPMI values.

Probabilities of single words $p(w_i)$ or the joint probability of two words $p(w_i, w_j)$ can be estimated using a Boolean document calculation—that is, the number of documents in which (w_i) or (w_i, w_j) occurs, divided by the total number of documents. The Boolean document calculation, however, ignores the frequencies and distances of words. C_V incorporates a Boolean sliding window calculation in which a new virtual document is created for every window of size s when sliding over the document at a rate of one word token per step. For example, document d_1 with words w results in virtual documents $d'_1 = \{w_1, ..., w_s\}$, $d'_2 = \{w_2, ..., w_{s+1}\}$, and so on. The probabilities $p(w_i)$ and $p(w_i, w_j)$ are subsequently calculated from the total number of virtual documents. In contrast to the Boolean document calculation, the Boolean sliding window calculation tries to capture the word token proximity to some degree.

The indirect confirmation measure $\phi_{S_i}(\vec{u}, \vec{w})$ is obtained by calculating the cosine vector similarity between all context vectors $\vec{v}(W') \in \vec{u}$ and $\vec{v}(W^*) \in \vec{w}$ of a pair $S_i = (W', W^*)$, as shown in (7).

$$\phi_{S_i}(\vec{u}, \vec{w}) = \frac{\sum_{i=1}^{|W|} u_i \cdot w_i}{\|\vec{u}\|_2 \cdot \|\vec{w}\|_2} \tag{7}$$

Finally, the arithmetic mean of individual confirmation measures is used to arrive at an overall topic coherence score.

3. Methods

3.1. *Dataset*

We compare the influence of Dirichlet hyperparameters on two datasets containing scientific research articles related to the domain of *fisheries*. The first dataset, DS_1, contains all full-text research articles published by the journal *Canadian Journal of Fisheries and Aquatic Sciences* and the journal *ICES Journal of Marine Science* from 1996 to 2016, with $D = 8,012$ documents, vocabulary size of $V = 203,248$, a total of $N = 29,469,919$ words, and on average 3,678 words per document. The second dataset, DS_2, contains only abstract data from the journal *Canadian Journal of Fisheries and Aquatic Sciences*, with $D = 4,417$, $V = 14,643$, $N = 481,168$, and 109 words on average per document. Both journals are domain-specific (i.e. fisheries) journals, but employ a wide scope of research directives related to the biological, ecological, and socio-ecological aspects of fisheries.

The domain of fisheries includes a multitude of knowledge production approaches, from mono- to transdisciplinary. Biologists, oceanographers, mathematicians, computer scientists, anthropologists, sociologists, political scientists, economists, and researchers from many other disciplines contribute to the body

of knowledge of fisheries, together with non-academic participants such as decision makers and stakeholders. Within the domain of fisheries, research into text analytics techniques has only been applied in a number of cases (e.g. [40, 41]).

These journals were chosen for several reasons. First, a fisheries domain expert was available to rank the topics manually. Second, domain-specific journals, in contrast to generic journals such as *Nature*, *Science*, or *PLOS ONE*, increase generalizability to other domain-specific journals that are often the subject of study when uncovering topical structures from scientific publications, such as research performed within the field of computational linguistics [34] or neural information processing systems (NIPS) [20], thereby making our results applicable to such approaches. Finally, the two journals have the highest frequency of publication output within the analyzed period compared to all other fisheries journals.

Words that were part of a standard list of stop words ($n = 153$), single-occurrence words, and words occurring in $\geq 90\%$ of the documents (e.g. *fish*, *analysis*, *research*) were removed. The removal of the top 90% of words serves as an estimate to prevent frequently occurring words from dominating all topics. All documents were tokenized and represented as bag-of-word features. Apart from grouping lowercase and uppercase words, no normalization method (e.g. stemming or lemmatization) was applied to reduce inflectional and derivational forms of words to a common base form. Stemming algorithms can be overly aggressive and could result in unrecognizable words that reduce interpretability when labeling the topics. Stemming might also lead to another problem: It cannot be deduced whether a stemmed word comes from a verb or a noun [42]. As human topic ranking was part of our topic quality evaluation, interpretability was considered to be highly important.

3.2. *Dirichlet Hyperparameters*

Hyperparameter α controls the shape of the document–topic distribution, whereas η controls the shape of the topic–word distribution. A large α leads to documents containing many topics, and a large η leads to topics with many words. In contrast, small values for α and η result in sparse distributions: documents containing a small number of topics and topics with a small number of words. In essence, the hyperparameters α and η have a smoothing effect on the multinomial variables θ and β, respectively. Four different classes or combinations of Dirichlet priors are explored, as listed in Table 1, in which we follow a similar notation (i.e. AA, AS, SA, SS) as described in [8].

Symmetrical priors are often the default setting for LDA tools such as Mallet and Gensim and assume a priori that each of the K topics has an equal probability of being assigned to a document while each word has an equal chance of being assigned to a topic. For the symmetrical prior α, the hyperparameter is a vector with the value $1/K$, where K is the number of topics. The symmetrical prior η has a scalar parameter with the value $1/V$, where V is the size of the vocabulary (full-text data

Table 1: Notation of Dirichlet classes. α = document-topic distribution, η = topic-word distribution

Abbreviation	α	η
AA	Asymmetric	Asymmetric
AS	Asymmetric	Symmetric
SA	Symmetric	Asymmetric
SS	Symmetric	Symmetric

$DS1 = 203,248$, and abstract data $DS_2 = 14,643$). For the asymmetrical priors, we utilize an iterative learning process to approximate the hyperparameters from the data; estimation is required as no exact closed form solution exists. Estimating hyperparameters can be used to increase model quality, and their values can reveal specific properties of the corpus: α for the distinctiveness in underlying semantic structures and η for the group size of commonly co-occurring words [43]. Several methods for hyperparameter estimation exist, such as gradient ascent, fixed point iteration, and Newton-Raphson method. Estimating the Dirichlet parameter α aims to maximize $p(D|\alpha)$ by maximizing the log likelihood function of the data D, with $\log \bar{p}_k$ being the observed sufficient statistics (the following is analogous to that of η).

$$F_{(\alpha)} = \log p(D|\alpha) = N \log\Gamma(\sum_k a_k) - N \sum_k \log\Gamma(a_k) + N \sum_k (a_k - 1) \log \bar{p}_k$$
$$\text{with} \log \bar{p}_k = \frac{1}{N} \log p_{i,k} \tag{8}$$

This study adopts the Newton-Rapson [44] method that provides a quadratic converging method for parameter estimation. Given an initial value for α, parameters are iteratively updated to arrive at an asymmetrical Dirichlet distribution learned from the data. The update is given in (9), with ∇F being the gradient descent, iteratively stepping along a positive gradient to maximize or converge the log-likelihood function F (8).

$$
\begin{aligned}
\alpha_k^{new} &= \alpha_k^{old} - \frac{(\nabla F)_k - b}{q_{kk}} \\
\nabla F &= \frac{\partial F}{\partial \alpha_k} = N\left(\Psi\left(\sum_k \alpha_k\right) - \Psi(\alpha_k) + \log \bar{p}_k\right) \\
b &= \frac{\sum_j (\nabla F)_j / q_{jj}}{1/c + \sum_j 1/q_{jj}} \\
c &= N\Psi'\left(\sum_k \alpha_k\right) \\
q_{jk} &= -N\Psi'(\alpha_k)
\end{aligned}
\tag{9}
$$

3.3. *Creating LDA Models*

LDA models were created for four different classes of priors on α and η, as listed in Table 1. For each class of priors, LDA models were produced by varying the number of topics parameter $K = \{1, ..., 50\}$ and repeating the process five times; one class resulted in 250 LDA models. The same approach was performed on both datasets: DS_1 for full-text data and DS_2 for abstract data. A total of 2000 different LDA models were created. Given that our datasets focus on fisheries only, making them homogeneous in nature, a small number of topics is expected—typically around 10 to 20 given the scope and aims of the selected journals.

The Python library Gensim [16] was used to create LDA models. Posterior inference approximation is performed with online variational Bayes (VB) as proposed by Hoffman *et. al.* [45]. Online VB is based on an online stochastic optimization process and produces similar or improved [45] and faster [46] LDA models compared to its batch variant. The Newton-Raphson process of iteratively learning asymmetrical Dirichlet priors can conveniently be incorporated into online LDA in linear time. The convergence iteration parameter for the expectation step (i.e. E-step) is set to 100, where per-document parameters are fit for the variational distributions [see Algorithm 2 in [45]].

3.4. *Topic Coherence*

The coherence of topics was calculated using the C_V coherence measure as described in detail in Section 2.3; C_V has been shown to obtain the highest correlation with all available human topic ranking data. The segmentation of the topic's top-N words and subsequent calculation of confirmation are calculated for $N = 15$, pairing every top 15 word with every other top 15 word and calculating their semantic support within the corpus. $N = 15$ was chosen, in contrast to, for example, $N = 10$ [11], as no stemming or lemmatization was applied; with $N = 10$, several words with the same base form were among the top 10 words (e.g. *sample*, *sampling*), so analyzing the top 10 words would effectively mean analyzing fewer than 10 distinct words.

The constant ϵ for NPMI calculations (see (6)) avoids logarithms of zero and acts as a smoothing factor. This value is set to a very small number, 10^{-12}, as proposed by Stevens *et. al.* [12]; the coherence measure is highly dependent on the smoothing constant, and a very small value significantly reduces the scores for unrelated words compared to, for example, $\epsilon = 1$ [47]. The γ constant for NPMI calculations is set to 1 (see (5)) to place equal weights on all NPMI values. In contrast to $\gamma = 2$ [11], $\gamma = 1$ produced a higher correlation with human topic ranking data [14]. The sliding window s for the Boolean sliding window calculation is set to 110 [14].

3.5. *Human Topic Ranking*

A fisheries domain expert manually ranked a selection of topics by inspecting the topic's top 15 most probable words and a selection of document titles and content. The domain expert is affiliated with the leading competence institution for fishery and aquaculture in Norway. As topic coherence scores are also obtained from the topic's top 15 words, the manual ranking of the top 15 words allows for equal comparison between the two proposed assessments. The domain expert was asked to provide a label for each topic that best captures the semantics of the top 15 words. In addition, the domain expert was asked to rank the topics concerning semantically correct or, conversely, incorrect words. An incorrect word could be a wrong fisheries domain-related word that does not match the topic label and, thus, does not fit with the semantics of the majority of right words. For example, in cases where most of the topic words resemble the fish species *cod*, an incorrect domain-related word might refer to a different kind of species. Furthermore, incorrect terms may refer to noise terms (i.e. words that serve a grammatical or syntactical purpose only). Topics are subsequently ranked by the number of right terms concerning all of the top 15 words. High-quality topics have $\geq 90\%$ correct words, medium-quality topics have $\geq 80\%$ but $< 90\%$ correct words, and low-quality topics have $< 80\%$ correct words.

3.6. *Relaxing LDA assumptions*

At the time of writing, the original LDA method proposed by Blei and colleagues [3] has over 22,000 citations. The technique has received much attention from machine learning researchers and other scholars and has been adopted and extended in a variety of ways. More concretely, relaxing the assumptions behind LDA can result in richer representations of the underlying semantic structures. The bag of words assumption has been relaxed by conditioning words on the previous words (i.e. Markovian structure) [48]; the document exchangeability assumption (i.e., the order in which documents are analyzed), relaxed by the dynamic topic model [49], and the Bayesian non-parametric model can be utilized to automatically uncover the number of topics [50]. Furthermore, LDA has been extended in various ways. Topics might correlate as a topic about "cars" is more likely to also be about "emission" than it is about "diseases". The Dirichlet distribution is implicitly independent

and a more flexible distribution, such as the logistic normal, is a more appropriate distribution to capture covariance between topics. The correlated topic model aids in this task [51]. Other examples extending LDA include the author-topic model [52], the relational topic model [53], the spherical topic model [54], the sparse topic model [55], and the bursty topic model [56]. Topic models that relax or extend the original LDA model bring additional computational complexity and their own sets of limitations and challenges; nevertheless, it would be interesting to explore these models in future research.

4. Results

4.1. *Topic Coherence*

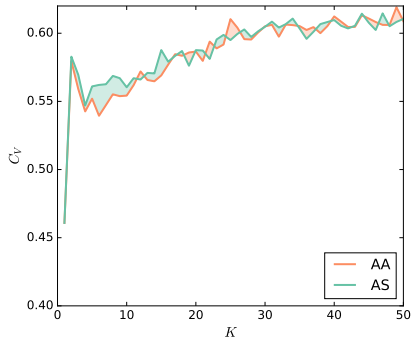
The coherence scores for the prior classes AA, AS, SA, and SS obtained from 8,012 full-text research articles (DS_1) are shown in Fig. 2. Additionally, the coherence scores for prior classes obtained from 4,417 abstracts (DS_2) are shown in Fig. 3. The coherence score represents the mean coherence score from all five runs for each value of k.

A visual inspection of Figs. 2a–2f (full-text data) shows that similar coherence scores are obtained for AA and AS (Fig. 2a), with both sharing an asymmetrical prior over α but a different prior over η. Similar results are obtained when comparing SA and SS (Fig. 2f), sharing a symmetrical prior over α and a different prior over η. Thus, varying η, while maintaining a similar prior over α, shows no real difference in obtained coherence score. A slightly increased coherence is obtained for an asymmetrical prior over α (e.g., Fig. 2d) for $k > 20$. Other combinations explored (e.g. AA–SA, AA–SS, and AS–SS) show similar results: a slight increase in coherence for an asymmetrical prior over α, with η showing no real benefits on topic coherence.

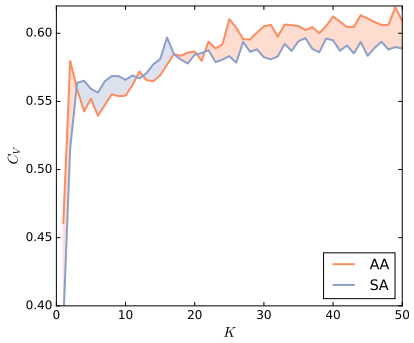
Figs. 3a–3f show coherence scores for LDA models obtained from abstract data (DS_2). AA–AS (Fig. 3a) show that different priors over η, while maintaining the same asymmetrical prior over α, result in similar coherence scores. Similarly, a symmetrical prior over α (Fig. 3f) with different priors over η shows no real differences in topic coherence. However, a large difference in coherence is obtained when varying the priors over α (Fig. 3d), with an asymmetrical α showing improved coherence over a symmetrical α. For DS_2, priors over α, in contrast to results from DS_1, show higher coherence scores for all values of k. Moreover, varying priors over η for DS_1 and DS_2 have a negligible effect on obtained coherence scores.

Table 2 shows the coherence score values obtained from DS_1 for $k = \{2, ..., 50\}$, with $\bar{X}$ representing the mean coherence over 5 runs, s the standard deviation, and f and p the one-way ANOVA F-value and p-value, respectively. The last six columns show the post hoc significance thresholds for all six comparison of Dirichlet priors.

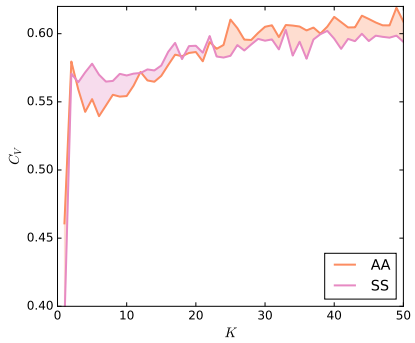
Table 2 reveals that significant differences are obtained starting from $k \geq 25$, although this does not hold for every $k \geq 25$. For $k < 25$, except for $k = 6$, no significant differences are obtained for combinations of priors; asymmetrical or sym-



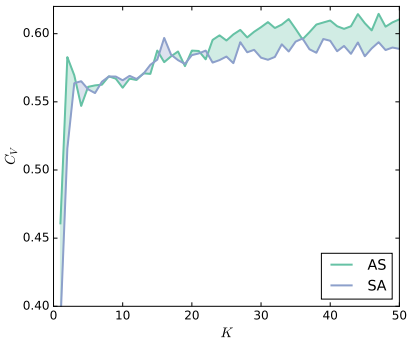
(a) AA - AS (DS_1)



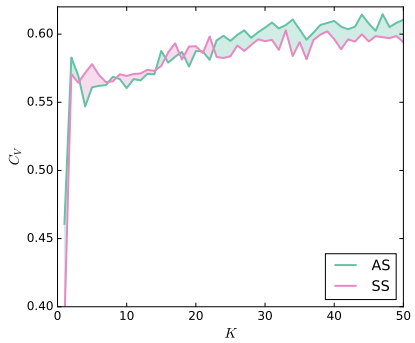
(b) AA - SA (DS_1)



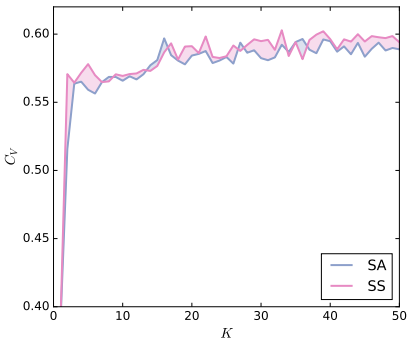
(c) AA - SS (DS_1)



(d) AS - SA (DS_1)

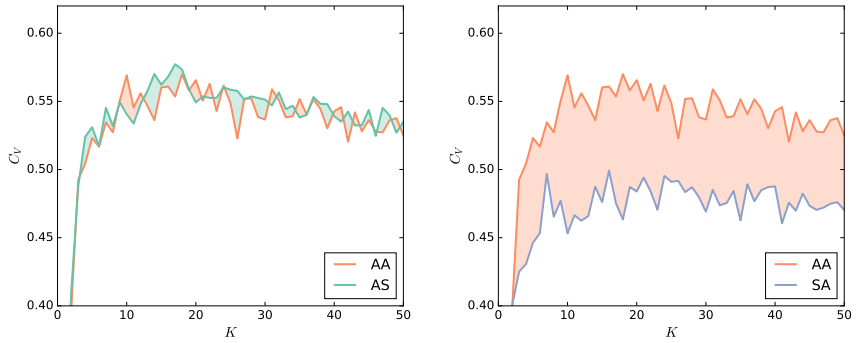


(e) AS - SS (DS_1)



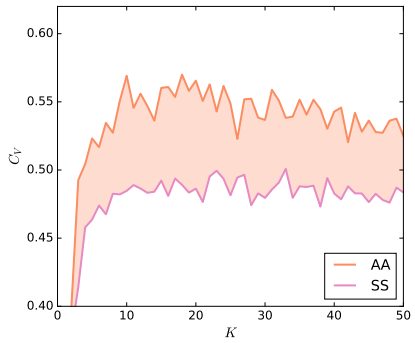
(f) SA - SS (DS_1)

Fig. 2: A comparison of calculated C_V topic coherence scores for all classes of priors (i.e. AA, AS, SA, SS). Coherence scores represent mean scores from five runs for $K = \{1, ..., 50\}$. $DS_1 = 8,012$ full-text articles.

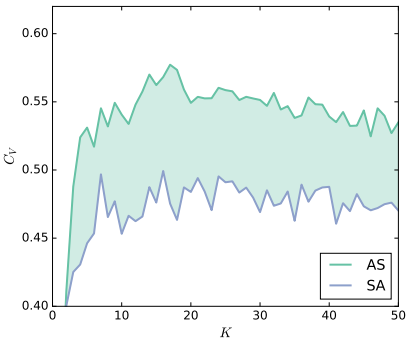


(a) AA - AS (DS_2)

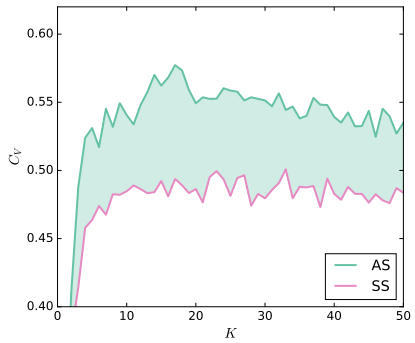
(b) AA - SA (DS_2)



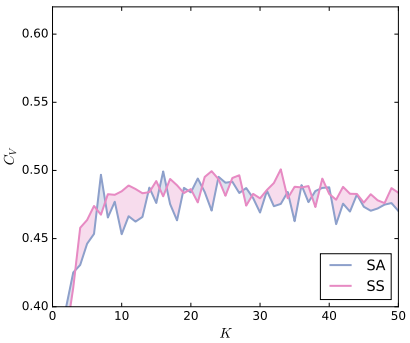
(c) AA - SS (DS_2)



(d) AS - SA (DS_2)



(e) AS - SS (DS_2)



(f) SA - SS (DS_2)

Fig. 3: A comparison of calculated C_V topic coherence scores for all classes of priors (i.e. AA, AS, SA, SS). Coherence scores represent mean scores from five runs for $K = \{1, ..., 50\}$. $DS_2 = 4,417$ abstracts.

metrical priors over α and η have no significant effect on topic coherence. However, the coherence score values for $k < 25$ show slightly higher values (shown in bold) for a symmetrical prior over α compared to an asymmetrical prior. In contrast, for $k \geq 25$, an asymmetrical prior over α shows higher coherence values compared to a symmetrical prior. For all k, where p is significant, SA–SS show no significance and AA–AS show significance only for $k = 6$ and $k = 47$; indicating the marginal importance of symmetrical or asymmetrical priors over η.

Table 3 shows the coherence score values and ANOVA statistics for DS_2. For all $k > 2$, the difference is significant ($p < 0.001$). These significant differences are caused by using an asymmetrical prior over α compared to a symmetrical prior. Where DS_1 shows mixed results between different priors over α, for DS_2, every combination of asymmetrical priors over α outperforms symmetrical priors over α. The post hoc tests for comparisons between different priors over η are almost in all cases not significant, following a similar trend with DS_1.

Table 2: Coherence score values and one-way ANOVA test statistics for DS_1 for $K = \{2, ..., 50\}$. $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

	Mean				**Std. dev.**				**ANOVA Statistics**							
K	$\bar{X}_{AA}$	$\bar{X}_{AS}$	$\bar{X}_{SA}$	$\bar{X}_{SS}$	s_{AA}	s_{AS}	s_{SA}	s_{SS}	f	p	$AA-AS$	$AA-SA$	$AA-SS$	$AS-SA$	$AS-SS$	$SA-SS$
2	0.580	**0.583**	0.516	0.571	0.037	0.021	0.055	0.004	3.237	0.0501						
3	0.559	**0.570**	0.564	0.564	0.026	0.020	0.021	0.035	0.108	0.9543						
4	0.543	0.547	0.565	**0.572**	0.014	0.016	0.020	0.020	2.554	0.0918						
5	0.552	0.561	0.559	**0.578**	0.013	0.010	0.025	0.013	1.788	0.1899						
6	0.539	0.562	0.556	**0.570**	0.014	0.011	0.021	0.007	3.298	0.0475*	**		**			
7	0.547	0.563	0.565	**0.565**	0.010	0.017	0.013	0.018	1.308	0.3063						
8	0.555	**0.569**	0.569	0.565	0.025	0.009	0.012	0.012	0.682	0.5761						
9	0.554	0.567	0.569	**0.571**	0.025	0.019	0.018	0.018	0.571	0.6419						
10	0.554	0.560	0.566	**0.569**	0.011	0.017	0.008	0.018	0.878	0.4732						
11	0.562	0.567	0.569	**0.571**	0.012	0.009	0.022	0.016	0.242	0.8660						
12	**0.572**	0.566	0.567	0.571	0.006	0.005	0.005	0.021	0.270	0.8462						
13	0.566	0.571	0.571	**0.574**	0.014	0.012	0.013	0.005	0.334	0.8006						
14	0.565	0.570	**0.577**	0.573	0.014	0.017	0.007	0.010	0.670	0.5828						
15	0.569	**0.588**	0.581	0.577	0.014	0.016	0.012	0.013	1.309	0.3061						
16	0.577	0.579	**0.597**	0.587	0.011	0.017	0.014	0.011	1.816	0.1848						
17	0.585	0.583	0.585	**0.593**	0.008	0.010	0.008	0.012	0.926	0.4508						
18	0.583	**0.587**	0.581	0.581	0.015	0.009	0.005	0.006	0.359	0.7837						
19	0.586	0.576	0.578	**0.591**	0.010	0.009	0.010	0.014	1.639	0.2200						
20	0.587	0.588	0.584	**0.591**	0.010	0.017	0.008	0.012	0.216	0.8840						
21	0.580	**0.587**	0.586	0.586	0.004	0.012	0.005	0.010	0.659	0.5890						
22	0.594	0.581	0.588	**0.598**	0.010	0.010	0.007	0.020	1.385	0.2834						
23	0.589	**0.595**	0.579	0.583	0.004	0.011	0.012	0.009	2.377	0.1082						
24	0.592	**0.599**	0.581	0.582	0.008	0.019	0.009	0.012	1.730	0.2010						
25	**0.610**	0.595	0.583	0.584	0.009	0.011	0.011	0.007	6.739	0.0038**		**	**			
26	**0.604**	0.599	0.578	0.592	0.008	0.015	0.014	0.013	3.162	0.0534						
27	0.596	**0.603**	0.594	0.588	0.009	0.013	0.008	0.010	1.471	0.2601						
28	0.595	**0.597**	0.586	0.592	0.008	0.008	0.009	0.015	0.812	0.5056						
29	0.601	**0.601**	0.588	0.596	0.008	0.005	0.014	0.017	1.038	0.4024						
30	**0.605**	0.605	0.582	0.595	0.012	0.007	0.007	0.006	6.342	0.0049**		**		**		
31	0.606	**0.608**	0.581	0.596	0.007	0.011	0.004	0.013	7.172	0.0029**		***		**		
32	0.597	**0.604**	0.583	0.588	0.010	0.004	0.005	0.009	6.359	0.0048**		**		***	**	
33	0.606	**0.607**	0.592	0.603	0.006	0.013	0.010	0.017	1.237	0.3290						
34	0.606	**0.611**	0.587	0.584	0.007	0.009	0.009	0.016	5.993	0.0061**		**	**	**	**	
35	**0.605**	0.603	0.594	0.594	0.014	0.010	0.005	0.011	1.269	0.3185						
36	**0.602**	0.596	0.596	0.582	0.007	0.012	0.004	0.006	5.398	0.0093**			**			
37	**0.604**	0.601	0.589	0.596	0.005	0.011	0.010	0.013	1.864	0.1764						
38	0.600	**0.607**	0.586	0.600	0.004	0.004	0.016	0.011	2.937	0.0650						
39	0.605	**0.608**	0.596	0.602	0.004	0.008	0.013	0.006	1.529	0.2453						
40	**0.612**	0.610	0.595	0.596	0.009	0.013	0.009	0.010	3.006	0.0612						
41	**0.608**	0.605	0.587	0.589	0.007	0.008	0.007	0.013	5.712	0.0074**		**	**	**		
42	**0.605**	0.604	0.591	0.596	0.005	0.010	0.015	0.007	1.605	0.2274						
43	0.605	**0.605**	0.585	0.594	0.015	0.010	0.006	0.014	2.664	0.0831						
44	0.613	**0.614**	0.594	0.600	0.009	0.010	0.007	0.011	4.600	0.0167*		**		**		
45	**0.611**	0.608	0.583	0.595	0.011	0.009	0.009	0.017	4.402	0.0194*		**		**		
46	**0.608**	0.602	0.589	0.599	0.009	0.009	0.009	0.007	3.502	0.0400*		**				
47	0.606	**0.615**	0.594	0.598	0.004	0.006	0.011	0.006	6.280	0.0051**	**		**	**	**	
48	**0.606**	0.605	0.588	0.597	0.007	0.007	0.007	0.008	5.384	0.0094**		**		**		
49	**0.619**	0.608	0.590	0.599	0.009	0.008	0.006	0.005	12.376	0.0002***		***	**	**		
50	0.609	**0.611**	0.589	0.594	0.006	0.007	0.010	0.010	6.225	0.0053**		**	**	**	**	

Table 3: Coherence score values and one-way ANOVA test statistics for DS_2 for $K = \{2, ..., 50\}$. $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

	Mean				**Std. dev.**				**ANOVA Statistics**							
K	$\bar{X}_{AA}$	$\bar{X}_{AS}$	$\bar{X}_{SA}$	$\bar{X}_{SS}$	s_{AA}	s_{AS}	s_{SA}	s_{SS}	f	p	$AA-AS$	$AA-SA$	$AA-SS$	$AS-SA$	$AS-SS$	$SA-SS$
2	0.399	**0.411**	0.400	0.380	0.017	0.011	0.021	0.037	1.207	0.3390						
3	**0.493**	0.488	0.425	0.414	0.013	0.026	0.025	0.031	11.187	0.0003***		**	**	**	**	
4	0.504	**0.524**	0.431	0.458	0.010	0.023	0.025	0.018	18.234	0.0000***		***	**	***	**	
5	0.523	**0.531**	0.446	0.464	0.032	0.015	0.030	0.024	10.511	0.0005***		**	**	***	**	
6	0.517	**0.517**	0.453	0.474	0.013	0.015	0.017	0.014	17.836	0.0000***		***	**	***	**	
7	0.535	**0.545**	0.497	0.468	0.009	0.010	0.029	0.013	17.407	0.0000***		**	***	**	***	
8	0.527	**0.532**	0.465	0.483	0.009	0.029	0.020	0.029	7.939	0.0018**		***	**	**	**	
9	**0.551**	0.549	0.477	0.482	0.008	0.016	0.025	0.022	18.741	0.0000***		***	***	**	**	
10	**0.569**	0.541	0.453	0.485	0.003	0.018	0.012	0.012	72.727	0.0000***	**	***	***	***	***	
11	**0.546**	0.534	0.466	0.489	0.019	0.017	0.013	0.026	14.533	0.0001***		***	**	***	**	
12	**0.556**	0.548	0.462	0.486	0.026	0.026	0.026	0.018	14.272	0.0001***		***	**	**	**	
13	0.547	**0.558**	0.466	0.483	0.027	0.022	0.021	0.031	12.797	0.0002***		**	**	***	**	
14	0.536	**0.570**	0.487	0.484	0.011	0.016	0.008	0.019	33.262	0.0000***	**	***	**	***	***	
15	0.560	**0.562**	0.476	0.492	0.017	0.014	0.022	0.015	26.740	0.0000***		***	***	***	***	
16	0.561	**0.568**	0.499	0.481	0.016	0.014	0.023	0.023	19.807	0.0000***		**	***	***	***	
17	0.554	**0.577**	0.475	0.494	0.015	0.011	0.020	0.013	41.061	0.0000***	**	***	***	***	***	
18	0.570	**0.573**	0.463	0.489	0.018	0.016	0.019	0.020	38.081	0.0000***		***	***	***	***	
19	0.558	**0.559**	0.487	0.483	0.005	0.011	0.022	0.016	33.195	0.0000***		***	***	***	***	
20	**0.566**	0.549	0.484	0.486	0.009	0.017	0.014	0.022	27.914	0.0000***		***	***	***	**	
21	0.551	**0.554**	0.494	0.477	0.016	0.018	0.021	0.014	19.513	0.0000***		**	***	**	***	
22	**0.563**	0.553	0.484	0.495	0.024	0.015	0.022	0.012	17.750	0.0000***		**	**	***	***	
23	0.543	**0.553**	0.471	0.499	0.021	0.021	0.013	0.019	16.624	0.0000***		***	**	***	**	
24	**0.562**	0.560	0.495	0.494	0.022	0.014	0.012	0.025	15.922	0.0000***		***	**	***	**	
25	0.549	**0.559**	0.491	0.481	0.011	0.009	0.022	0.007	33.981	0.0000***		**	***	***	***	
26	0.523	**0.558**	0.492	0.494	0.016	0.011	0.019	0.005	19.576	0.0000***	**	**	**	***	***	
27	**0.552**	0.551	0.483	0.496	0.014	0.013	0.013	0.016	27.074	0.0000***		***	***	***	***	
28	0.552	**0.554**	0.487	0.474	0.022	0.010	0.019	0.010	27.501	0.0000***		**	***	***	***	
29	0.538	**0.552**	0.480	0.483	0.015	0.013	0.021	0.010	24.880	0.0000***		**	***	***	***	
30	0.537	**0.551**	0.469	0.480	0.013	0.009	0.015	0.019	30.982	0.0000***		***	**	***	***	
31	**0.559**	0.547	0.485	0.486	0.022	0.018	0.016	0.021	16.533	0.0000***		***	**	***	**	
32	0.551	**0.557**	0.474	0.491	0.017	0.021	0.009	0.013	28.414	0.0000***		***	***	***	***	
33	0.538	**0.544**	0.475	0.501	0.009	0.007	0.013	0.013	36.003	0.0000***		***	**	***	***	
34	0.539	**0.547**	0.484	0.480	0.014	0.006	0.019	0.018	22.547	0.0000***		**	***	***	***	
35	**0.552**	0.538	0.463	0.488	0.013	0.011	0.012	0.015	43.139	0.0000***		***	***	***	***	
36	**0.541**	0.540	0.489	0.488	0.011	0.019	0.016	0.012	15.847	0.0000***		***	***	**	**	
37	0.552	**0.553**	0.477	0.489	0.023	0.007	0.006	0.007	39.560	0.0000***		***	***	***	***	
38	0.545	**0.548**	0.485	0.473	0.014	0.011	0.008	0.017	36.334	0.0000***		***	***	***	***	
39	0.530	**0.548**	0.487	0.494	0.014	0.005	0.011	0.008	32.693	0.0000***	**	**	**	***	***	
40	**0.543**	0.539	0.488	0.483	0.017	0.023	0.010	0.013	15.567	0.0001***		***	***	**	**	
41	**0.546**	0.535	0.461	0.478	0.012	0.009	0.011	0.015	48.701	0.0000***		***	***	***	***	
42	0.520	**0.543**	0.476	0.488	0.021	0.019	0.018	0.017	10.426	0.0005***		**	**	***	**	
43	**0.542**	0.532	0.470	0.483	0.015	0.015	0.015	0.013	23.777	0.0000***		***	***	***	**	
44	0.528	**0.533**	0.482	0.483	0.008	0.016	0.010	0.011	22.455	0.0000***		***	***	***	***	
45	0.536	**0.544**	0.473	0.476	0.006	0.008	0.012	0.010	66.934	0.0000***		***	***	***	***	
46	**0.528**	0.525	0.470	0.483	0.014	0.019	0.007	0.016	16.132	0.0000***		***	**	***	**	
47	0.527	**0.545**	0.472	0.478	0.005	0.016	0.018	0.009	29.467	0.0000***		***	***	***	***	
48	0.536	**0.540**	0.475	0.476	0.009	0.005	0.017	0.010	41.643	0.0000***		***	***	***	***	
49	**0.538**	0.527	0.476	0.487	0.014	0.015	0.009	0.022	14.562	0.0001***		***	**	***	**	
50	0.525	**0.535**	0.470	0.484	0.018	0.013	0.008	0.014	21.727	0.0000***		***	**	***	***	

Table 4: Human topic ranking for DS_2 (abstract) on $k = 17$ LDA model

Class	High-quality	Medium-quality	Low-quality
AA	15/17 (88%)	2/17 (12%)	0/17 (0%)
AS	15/17 (88%)	2/17 (12%)	0/17 (0%)
SA	12/17 (70.5%)	4/17 (23.6%)	1/17 (5.9%)
SS	11/17 (64.7%)	6/17 (35.3%)	0/17 (0%)

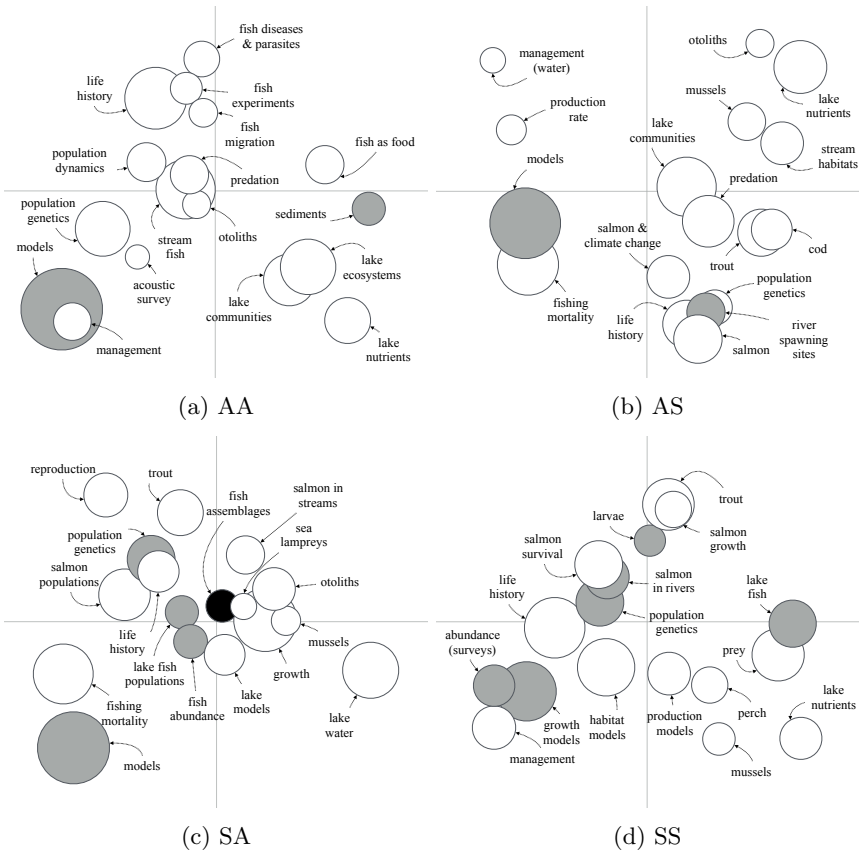
4.2. *Human Topic Ranking*

The results of the fisheries domain expert's human topic ranking are shown in Table 4 (see Section 3.5 for the classification method). Human topic ranking was performed on DS_2 for $k = 17$ LDA models, which is the k-value that shows the best coherence score (via elbow method) and, simultaneously, the k-value with the largest difference amongst all prior classes (ANOVA $f = 41.06$). A similar pattern as found for topic coherence scores can be identified (Figs. 3a–3f): AA and AS with an asymmetrical prior over α result in more high-quality (88%) topics compared to SA and SS with a symmetrical prior over α (70.5% and 64.7% high-quality topics). Both AA and AS perform similarly, indicating that priors over η have no effect on human topic ranking. Furthermore, SA and SS show similar lower human topic ranking, with three topics differently classified: SS has 77.5% of high-quality topics compared to 64.7% for SS, but simultaneously one low-quality topic. A two-dimensional inter-topic distance map for DS_2 with $k = 17$ is displayed in Fig. 4 for all classes of priors. This two-dimensional representation is obtained by computing the distance between topics [57] and applying multidimensional scaling [58]. It displays the similarity between topics concerning their probability distribution over words (i.e. β). In addition, a topic label that best captures the semantics of the top 15 words is attached.

We omitted human topic ranking results for DS_1 as they show an equal number of high-quality and medium-quality topics for all classes of priors and for several arbitrarily chosen k-values ($k < 25$). These results are in line with topic coherence scores that show similar scores for all prior classes (see Figs. 2a–2f). An inspection of $k \geq 25$ LDA models (the point where significant differences between prior classes start) shows an increasing number of incorrect terms for LDA models with a symmetrical prior over α (SA and SS), compared to models with an asymmetrical prior over α.

5. Discussion and Conclusion

Our results show that an asymmetrical prior over α indicates increased topic coherence and topic ranking compared to a symmetrical prior. However, this particularly holds for the DS_2 dataset, the collection of 4,417 abstracts, and not necessarily for



(a) AA (b) AS

(c) SA (d) SS

Fig. 4: A two-dimensional inter-topic distance map (via multidimensional scaling) for all classes of priors for DS_2 with $k = 17$. The surface of the node indicates the overall topic prevalence within the corpus. Color coding is used to indicate human topic ranking classification: white = high-quality, grey = medium-quality, and black = low-quality.

the DS_1 dataset, the collection of 8,012 full-text documents. Thus, selecting a different prior on α has large practical implications for datasets containing a smaller vocabulary size and being homogenous in nature. Symmetrical or asymmetrical priors over η show no real benefits regarding topic coherence and human topic ranking.

The results on DS_2 are in line with research performed by Wallach *et. al.* [8], which found that an asymmetrical prior over α shows improved likelihood of held-out data and that different priors over η show no real differences; the vocabulary size of the DS_2 data set can be compared to the vocabulary size used in their research. However, our results on DS_1, the full-text dataset with significantly higher vocabulary size and an average number of words per document, found no difference

for combinations of priors over α and η, making topic coherence and manual topic ranking less influenced by full-text data.

A symmetrical prior over α assumes that all topics have an equal probability of being assigned to a document. Such an assumption ignores that certain topics are more prominent in a document collection and, consequently, would logically have a higher probability to be assigned to a document. Conversely, specific topics are less common and, thus, not appropriately reflected with a symmetrical prior distribution. Logically speaking, an asymmetrical prior over α would capture this intuition and would, therefore, be the preferred choice. We have empirically shown that this intuition indeed results in significantly higher topic coherence and a better topic ranking for DS_2 and DS_1 for $k \geq 25$. For DS_1 with $k < 25$, the differences are not significant, although human topic ranking shows slightly better topics for the classes with an asymmetrical prior over α.

Concerning priors over η, we naturally want topic–word distributions to be different from each other so as to avoid conflicts between them. A symmetrical prior over η will reflect the power-law usage of words (i.e. some words occur in all topics) while simultaneously resolving ambiguity between topics with a few distinct word co-occurrences [8]. Therefore, symmetrical priors over η are the preferred choice. Although our empirical results indicate no real benefits when varying priors on η, the symmetrical prior shows slight, but still very marginal, overall improved coherence and ranking results for both datasets.

Table 5: A selection of modeling topics for $k = 17$. Terms in bold are considered incorrect words.

Dataset	Class	Label	Top 15 words	Ranking
DS_1	AA	Models (population)	model, cod, stock, mortality, population, recruitment, year, models, estimates, fisheries, abundance, biomass, size, spawning, years	High
DS_1	AS	Models (abundance)	model, length, estimates, uncertainty, abundance, models, values, distribution, gaussian, survey, simulated, simulation, acoustic, parameters, pollock	High
DS_1	SA	Models	model, stock, fishing, catch, fisheries, recruitment, fishery, year, estimates, models, mortality, management, assessment, biomass, effort	High
DS_1	SS	Models (fishing)	model, fishing, catch, stock, fishery, fisheries, mortality, effort, management, estimates, year, models, parameters, population, assessment	High
DS_2	AA	Models	model, data, models, **used**, estimates, **using**, stock, management, fish, recruitment, mortality, population, method, approach, **based**	Medium
DS_2	AS	Models	data, model, models, **used**, **using**, fish, estimates, method, approach, sampling, analysis, based, methods, estimate, **use**	Medium
DS_2	SA	Models	model, data, models, management, **used**, species, estimates, **using**, fish, fisheries, abundance, stock, habitat, analysis, recruitment	Medium
DS_2	SS	Models (growth)	model, growth, data, estimates, fish, **using**, **used**, models, habitat, size, mortality, parameters, length, method, population	Medium

Human topic ranking was based on the presence of incorrect terms being part of the topic's 15 most probable words. A closer look into the reasons why topics are ranked lower reveals that all topics contain correct domain-related terms, but are only ranked lower due to the presence of so-called noise terms (e.g. *used*, *using*, *two*, *among*, *total*, *higher*, *within*, *great*, *large*, *high*, *significantly*). Lower-ranked topics contain a higher number of such terms and, as such, are classified as medium- or low-quality topics. Interestingly, none of the topics have incorrect domain-related terms that could refer to, for example, the biological, ecological, socio-ecological, or social aspects of fisheries. A selection of topics and incorrect terms is shown in Table 5. Topics uncovered from abstract data, combined with a symmetrical prior over α, are more prone to contain such noise terms. For full-text data, all classes of priors show an equal but low number of noise words.

A growing amount of research is utilizing LDA to uncover latent semantic structures from scientific research articles as a mean to discover topical trends and developments within a particular research area [17, 18, 19, 20, 21]. These approaches are often characterized by (i) exploring one or several domain-specific journals (e.g. journals related to transportation research, operations research and management science), (ii) using abstract data, and (iii) using an open source tool (e.g. Mallet, Gensim) to perform LDA. Our approach touches upon all three characteristics; thus, we would recommend an asymmetrical prior over α and a symmetrical prior over η for optimal topic coherence and topic ranking.

Fig. 4 shows a visual representation of 17 latent topics for DS_2. We identify several overlapping topics (e.g. *otoliths*, *population genetics*) and several semantically related topics (e.g. *population dynamics* and *population genetics*; *lake nutrients* and *lake waters*). At the same time, we find several topics occurring in one of the prior classes that are absent in other prior classes. For instance, *fish diseases and parasites* occurs only in AA (Fig. 4a). One reason might be that manual topic labeling is limited by the subjectivity inherent in human interpretation [59]; indeed, an analysis of the topics by another domain expert could yield contradictory results. Another reason might be due to the probabilistic nature of LDA, where differences are merely a result of differences in sampling. Although such analysis is outside the scope of this research, it is an interesting directive for future research. Furthermore, the research performed by Wallach *et. al.* was applied on corpora related to patent, newsgroup and news data, whereas this paper analyzed scientific research articles. Future research might focus on different types of scientific articles, more broadly oriented journals, or other unexplored forms of textual data to gain more insights into the practical effects Dirichlet priors have on LDA's latent topics.

Acknowledgment

This research was funded by the project SAF21 - Social science aspects of fisheries for the 21^{st} Century. SAF21 is a project financed under the EU Horizon 2020 Marie Skłodowska-Curie (MSC) ITN - ETN program (project 642080).

References

[1] P. O. Larsen and M. von Ins, The rate of growth in scientific publication and the decline in coverage provided by Science Citation Index, *Scientometrics* **84** 575–603 (sep 2010).

[2] A. Srivastava and M. Sahami, *Text mining: Classification, clustering, and applications* (CRC Press, 2009).

[3] D. M. Blei, A. Y. Ng and M. I. Jordan, Latent Dirichlet Allocation, *Journal of Machine Learning Research* **3** 993–1022 (2003).

[4] T. Hofmann, Probabilistic latent semantic indexing, in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '99* (ACM Press, New York, New York, USA, 1999), pp. 50–57.

[5] R. Fergus, L. Fei-Fei, P. Perona and A. Zisserman, Learning object categories from Google's image search, in *Tenth IEEE International Conference on Computer Vision (ICCV'05)* (IEEE, Beijing, China, 2005), pp. 1816–1823.

[6] R. Mehran, A. Oyama and M. Shah, Abnormal crowd behavior detection using social force model, in *2009 IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, Miami, FL, USA, jun 2009), pp. 935–942.

[7] S. Kim, S. Narayanan and S. Sundaram, Acoustic topic model for audio information retrieval, in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (IEEE, New Paltz, NY, USA, 2009), pp. 37–40.

[8] H. M. Wallach, D. Mimno and A. Mccallum, Rethinking LDA : Why Priors Matter, in *NIPS'09 Proceedings of the 22nd International Conference on Neural Information Processing Systems* (Curran Associates Inc., Vancouver, British Columbia, Canada, 2009), pp. 1973–1981.

[9] A. Asuncion, M. Welling, P. Smyth and Y. W. Teh, On Smoothing and Inference for Topic Models, in *UAI '09 Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence* (AUAI Press Arlington, Montreal, Quebec, Canada, may 2012), pp. 27–34.

[10] J. Chang, S. Gerrish, C. Wang and D. M. Blei, Reading Tea Leaves: How Humans Interpret Topic Models, in *NIPS'09 Proceedings of the 22nd International Conference on Neural Information Processing Systems* (Curran Associates Inc., Vancouver, British Columbia, Canada, 2009), pp. 288–296.

[11] N. Aletras and M. Stevenson, Evaluating topic coherence using distributional semantics, in *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)* (Association for Computational Linguistics, Potsdam, Germany, 2013), pp. 13–22.

[12] K. Stevens, P. Kegelmeyer, D. Andrzejewski and D. Buttler, Exploring Topic Coherence over Many Models and Many Topics, in *EMNLP-CoNLL '12 Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (Association for Computational Linguistics, Jeju Island, Korea, 2012), pp. 952–961.

[13] D. Newman, J. Lau, K. Grieser and T. Baldwin, Automatic evaluation of topic coherence, in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (June), (Association for Computational Linguistics, Stroudsburg, PA, USA, 2010), pp. 100–108.

[14] M. Röder, A. Both and A. Hinneburg, Exploring the Space of Topic Coherence Measures, in *WSDM '15 Proceedings of the Eighth ACM International Conference on Web Search and Data Mining* (ACM Press, Shanghai, China, 2015), pp. 399–408.

[15] Z. S. Harris, Distributional Structure, *WORD* **10** 146–162 (aug 1954).

[16] R. Rehurek and P. Sojka, Software Framework for Topic Modelling with Large Cor-

pora, in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks* (European Language Resources Association (ELRA), Valletta, Malta, 2010), pp. 45–50.

[17] C. J. Gatti, J. D. Brooks and S. G. Nurre, A Historical Analysis of the Field of OR/MS using Topic Models, *arXiv.org* **stat.ML** (oct 2015).

[18] L. Sun and Y. Yin, Discovering themes and trends in transportation research using topic modeling, *Transportation Research Part C: Emerging Technologies* **77** 49–66 (apr 2017).

[19] M. J. Westgate, P. S. Barton, J. C. Pierson and D. B. Lindenmayer, Text analysis tools for identification of emerging topics and research gaps in conservation science, *Conservation Biology* **29** 1606–1614 (dec 2015).

[20] X. Wang and A. McCallum, Topics over time: a non-Markov continuous-time model of topical trends, in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '06* (ACM Press, Philadelphia, PA, USA, 2006), pp. 424–433.

[21] J. M. Alston and P. G. Pardey, Six decades of agricultural and resource economics in Australia: an analysis of trends in topics, authorship and collaboration, *Australian Journal of Agricultural and Resource Economics* **60** 554–568 (oct 2016).

[22] E. Erosheva, S. Fienberg and J. Lafferty, Mixed-membership models of scientific publications., *Proceedings of the National Academy of Sciences of the United States of America* **101 Suppl** 5220–5227 (apr 2004).

[23] D. M. Blei and J. D. Lafferty, Topic Models, *Text Mining: Classification, Clustering, and Applications* 71–89 (2009).

[24] D. Newman, A. Asuncion, P. Smyth and M. Welling, Distributed inference for latent dirichlet allocation, in *Proceedings of the 20th International Conference on Neural Information Processing Systems* (Vancouver, British Columbia, Canada, 2007), pp. 1081–1088.

[25] I. Porteous, D. Newman, A. Ihler, A. Asuncion, P. Smyth and M. Welling, Fast Collapsed Gibbs Sampling For Latent Dirichlet Allocation, in *KDD '08 Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining* (ACM Press, Las Vegas, Nevada, USA, 2008), pp. 569–577.

[26] D. M. Blei and M. I. Jordan, Variational inference for Dirichlet process mixtures, *Bayesian Analysis* **1** 121–143 (mar 2006).

[27] Y. W. Teh, D. Newman, M. Welling and D. Neaman, A Collapsed Variational Bayesian Inference Algorithm for Latent Dirichlet Allocation, in *NIPS'06 Proceedings of the 19th International Conference on Neural Information Processing Systems* (MIT Press Cambridge, MA, USA, Vancouver, British Columbia, Canada, 2006), pp. 1353–1360.

[28] C. Wang, J. Paisley and D. M. Blei, Online Variational Inference for the Hierarchical Dirichlet Process, in *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS)* **15**, (PMLR, Fort Lauderdale, FL, USA, 2011), pp. 752–760.

[29] D. M. Blei, Probabilistic topic models, *Communications of the ACM* **55** 77–84 (apr 2012).

[30] T. L. Griffiths and M. Steyvers, Finding scientific topics, *Proceedings of the National Academy of Sciences* **101** 5228–5235 (apr 2004).

[31] J. Grimmer and B. M. Stewart, Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts, *Political Analysis* **21** 267–297 (jan 2013).

[32] T. Rusch, P. Hofmarcher, R. Hatzinger and K. Hornik, Model trees with topic

model preprocessing: An approach for data journalism illustrated with the WikiLeaks Afghanistan war logs, *The Annals of Applied Statistics* **7** 613–639 (jun 2013).

[33] M. W. Neff and E. A. Corley, 35 years and 160,000 articles: A bibliometric exploration of the evolution of ecology, *Scientometrics* **80** 657–682 (sep 2009).

[34] D. Hall, D. Jurafsky and C. D. Manning, Studying the history of ideas using topic models, in *EMNLP '08 Proceedings of the Conference on Empirical Methods in Natural Language Processing* (Association for Computational Linguistics, Honolulu, Hawaii, 2008), pp. 363–371.

[35] S. Syed and C. T. Weber, Using Machine Learning to Uncover Latent Research Topics in Fishery Models, *Reviews in Fisheries Science & Aquaculture* **26**(3) 319–336 (2018).

[36] S. Syed, M. Borit and M. Spruit, Narrow lenses for capturing the complexity of fisheries: A topic analysis of fisheries science from 1990 to 2016, *Fish and Fisheries* **00** 1–19 (apr 2018).

[37] D. Wallach, Hanna M., Murray, Iain, Salakhutdinov, Ruslan and Mimno, Evaluation Methods for Topic Models, in *ICML '09 Proceedings of the 26th Annual International Conference on Machine Learning* 2009, pp. 1105–1112.

[38] I. Douven and W. Meijs, Measuring coherence, *Synthese* **156**(3) 405–425 (2007).

[39] G. Bouma, Normalized (Pointwise) Mutual Information in Collocation Extraction, in *Proceedings of German Society for Computational Linguistics (GSCL 2009)* (GSCL, Potsdam, Germany, 2009), pp. 31–40.

[40] S. Syed, M. Spruit and M. Borit, Bootstrapping a Semantic Lexicon on Verb Similarities, in *Proceedings of the 8th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management* **1**(Ic3k), (SCITEPRESS - Science and Technology Publications, 2016), pp. 189–196.

[41] S. Syed and M. Spruit, Full-Text or Abstract? Examining Topic Coherence Scores Using Latent Dirichlet Allocation, in *4th IEEE International Conference on Data Science and Advanced Analytics (DSAA)* (IEEE, Tokyo, Japan, oct 2017), pp. 165–174.

[42] N. Evangelopoulos, X. Zhang and V. R. Prybutok, Latent Semantic Analysis: five methodological recommendations, *European Journal of Information Systems* **21** 70–86 (jan 2012).

[43] G. Heinrich, Parameter estimation for text analysis, *Bernoulli* **35** 1–31 (2005).

[44] J. Huang, Maximum Likelihood Estimation of Dirichlet Distribution Parameters, tech. rep., Carnegie Mellon University (Pittsburgh, Pennsylvania, USA, 2005).

[45] M. D. Hoffman, D. M. Blei and F. Bach, Online Learning for Latent Dirichlet Allocation, in *NIPS'10 Proceedings of the 23rd International Conference on Neural Information Processing Systems - Volume 1* (Curran Associates Inc., Vancouver, British Columbia, Canada, 2010), pp. 856–864.

[46] L. Bottou and O. Bousquet, The Tradeoffs of Large Scale Learning, in *Proceedings of the 20th International Conference on Neural Information Processing Systems* **20**2007, pp. 161–168.

[47] D. Mimno, H. M. Wallach, E. Talley, M. Leenders and A. McCallum, Optimizing semantic coherence in topic models, in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing* (2)2011, pp. 262–272.

[48] H. M. Wallach, Topic modeling: beyond bag-of-words, in *Proceedings of the 23rd international conference on Machine learning - ICML '06* (1), (ACM Press, Pittsburgh, Pennsylvania, USA, 2006), pp. 977–984.

[49] D. M. Blei and J. D. Lafferty, Dynamic topic models, in *Proceedings of the 23rd international conference on Machine learning - ICML '06* (ACM Press, New York, USA, 2006), pp. 113–120.

[50] Y. Whye Teh, M. I. Jordan, M. J. Beal and D. M. Blei, Sharing clusters among related groups: Hierarchical Dirichlet processes, in *NIPS'04 Proceedings of the 17th International Conference on Neural Information Processing Systems* (MIT Press Cambridge, Vancouver, British Columbia, Canada, 2004), pp. 1385–1392.

[51] D. M. Blei and J. D. Lafferty, A correlated topic model of Science, *The Annals of Applied Statistics* **1** 17–35 (jun 2007).

[52] M. Rosen-Zvi, T. Griffiths, M. Steyvers and P. Smyth, The author-topic model for authors and documents, in *Proceedings of the 20th conference on Uncertainty in artificial intelligence* (AUAI Press Arlington, Banff, Canada, 2004), pp. 487–494.

[53] J. Chang and D. M. Blei, Hierarchical relational models for document networks, *The Annals of Applied Statistics* **4** 124–150 (mar 2010).

[54] J. Reisinger, A. Waters, B. Silverthorn and R. J. Mooney, Spherical Topic Models, in *Proceedings of the 27th International Conference on Machine Learning* (International Machine Learning Society (IMLS), Haifa, Israel, 2010), pp. 903–910.

[55] C. Wang and D. M. Blei, Decoupling Sparsity and Smoothness in the Discrete Hierarchical Dirichlet Process, in *NIPS'09 Proceedings of the 22nd International Conference on Neural Information Processing Systems* (Curran Associates Inc., Vancouver, British Columbia, Canada, 2009), pp. 1982–1989.

[56] G. Doyle and C. Elkan, Accounting for burstiness in topic models, in *Proceedings of the 26th Annual International Conference on Machine Learning - ICML '09* (ACM Press, Montreal, QC, Canada, 2009), pp. 281–288.

[57] J. Chuang, D. Ramage, C. Manning and J. Heer, Interpretation and trust, in *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems - CHI '12* (ACM Press, Austin, TX, USA, 2012), pp. 443–452.

[58] C. Sievert and K. Shirley, LDAvis: A method for visualizing and interpreting topics, in *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces* (Association for Computational Linguistics, Baltimore, Maryland, USA, 2014), pp. 63–70.

[59] C. Urquhart, An Encounter with Grounded Theory : Tackling the Practical and Philosophical Issues, *Qualitative Research in Information Systems: Issues and Trends* (1991) 104–140 (2001).