A new evaluation framework for topic modeling algorithms based on synthetic corpora

Hanyu Shi [1,†], **Martin Gerlach**[1,†], **Isabel Diersen**[1], **Doug Downey**[2], **Luís A. N. Amaral**[1,*]
[1]Department of Chemical and Biological Engineering,
[2]Department of Electrical Engineering and Computer Science
Northwestern University, Evanston, Illinois 60208, USA
[*]amaral@northwestern.edu, [†]Both authors contributed equally to this manuscript.

Abstract

Topic models are in widespread use in natural language processing and beyond. Here, we propose a new framework for the evaluation of probabilistic topic modeling algorithms based on synthetic corpora containing an unambiguously defined ground truth topic structure. The major innovation of our approach is the ability to quantify the agreement between the planted and inferred topic structures by comparing the assigned topic labels at the level of the tokens. In experiments, our approach yields novel insights about the relative strengths of topic models as corpus characteristics vary, and the first evidence of an "undetectable phase" for topic models when the planted structure is weak. We also establish the practical relevance of the insights gained for synthetic corpora by predicting the performance of topic modeling algorithms in classification tasks in real-world corpora.

1 Introduction

Topic modeling is a powerful natural language processing tool for the unsupervised inference of the latent topics of a collection of texts (Blei, 2012; Crain et al., 2012). A variety of topic modeling algorithms have been proposed to cope with a broad set of technical challenges and diverse types of written documents (Blei et al., 2003; Griffiths and Steyvers, 2004; Blei and Lafferty, 2007; Buntine and Mishra, 2014; Lancichinetti et al., 2015). Due to the large number of topic models in the literature and their widespread use, it is crucial to benchmark available algorithms. The need for such approaches is exacerbated by the increase of topic modeling applications in computational social science, where the purpose of the models is not to predict documents (in which case held-out likelihood would suffice) but instead to help understand the corpus, which requires an evaluation of the inferred topics themselves (Boyd-Graber et al., 2017).

Our analysis is grounded on the assumption that a hidden topic structure exists in the texts (i.e. a latent variable leading to deviations from the random usage of words). Under this assumption, a topic modeling algorithm can be viewed as an instrument for the measurement of the hidden structures. Crucial to measurement is the existence of a standard that provides ground truth (Bandalos, 2018; Allen and Yen, 2001). For example, the use of synthetic datasets has become standard in order to probe machine learning algorithms in fields such as clustering (Jain, 2010) or community detection (Lancichinetti et al., 2008).

Currently employed evaluation methods for topic models are often subjective, and can lack theoretical justification. Indeed, the debate is ongoing as to which evaluation method is best (Wallach et al., 2009b; Chang et al., 2009; Röder et al., 2015). From a practical perspective, the relative performance of topic modeling algorithms varies substantially across different corpora with different characteristics (see e.g. Fig. 1, which compares several topic modeling algorithms on classification tasks). While we would expect that certain algorithms or settings are better suited to particular document characteristics (e.g., corpus size, document length, number of topics, burstiness, etc.), it remains unclear how such properties affect the performance of topic modeling algorithms, beyond a certain measure of machine learning "folklore" (Tang et al., 2014).

In this work, we present a new framework for topic

Proceedings of the 22$^{\text{nd}}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89.

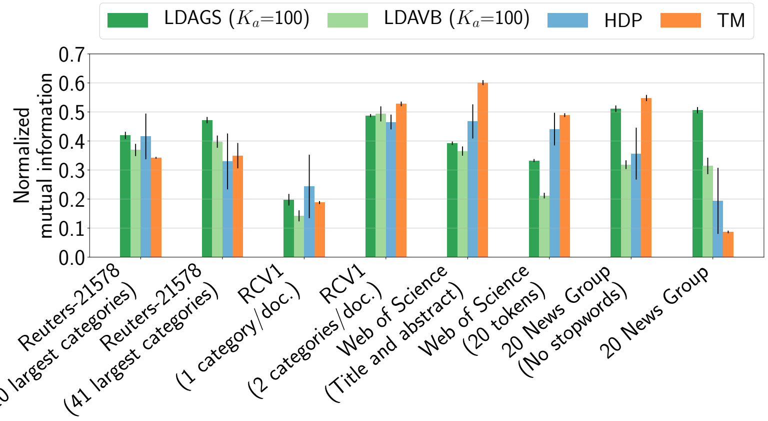

Figure 1: **Performance of topic models is inconsistent across diverse real-world corpora.** Normalized mutual information of four topic modeling algorithms in unsupervised document classification for 8 real-world corpora. See *Supplementary Material*, Secs. S1, S2 and S4 for details on the corpora (and the pre-processing steps), the topic modeling algorithms, and the comparison metric, respectively.

model evaluation relying on generating a synthetic corpus containing an unambiguous ground truth. First, we propose a novel way to generate synthetic corpora that generalizes upon previous approaches. Our approach allows us to isolate the impact of various corpus characteristics, such as size, number of topics, the signal-to-noise ratio, burstiness, or fraction of stopwords, which in real-world corpora are either unknown or impossible to tone. Second, we propose a new evaluation metric based on the normalized mutual information that compares the agreement between planted and inferred topics on the level of individual word tokens. Our approach yields an absolute measure of topic modeling accuracy, eliminating the need for post-inference heuristics such as "topic matching" (Lancichinetti et al., 2015). While synthetic ground truth has been used for topic model evaluation in the past, ours is the first framework for evaluating how well topic modeling algorithms perform the key task of inferring per-token topic assignments. Altogether, the formalization of synthetic corpora allows us to probe more accurately the ability of different topic modeling algorithms to resolve a wide range of topic structures, beyond simplistic assumptions of LDA. We present experiments showing how different popular topic modeling algorithms fare as these characteristics change, for one type of synthetic corpus. We show how our measurement framework leads to new insights, including evidence of an "undetectable region" for sufficiently weak topic structures, or how the choice of hyperparameters can bias the inference result. Finally, we show that our approach is predictive of the performance of topic modeling algorithms in classification tasks in real-world corpora.

2 Background

A popular approach for evaluating topic models is to inspect their output manually (Murakami et al., 2017), but this approach is expensive and subjective. The most common quantitative approaches to evaluate topic modeling algorithms rely on intrinsic evaluation methods, such as held-out likelihood (Wallach et al., 2009b), and topic coherence (Newman et al., 2010; Mimno et al., 2011), or on extrinsic tasks such as document classification (Lu et al., 2011; Xie and Xing, 2013) and information retrieval (Schütze et al., 2008; Wei and Croft, 2006). However, these approaches allow only for limited insights into why topic modeling algorithms fail or succeed. For example, perplexity and topic coherence can only provide a relative measure of performance: how well does a topic model do *in relation* to another model? In contrast, extrinsic evaluation tasks allow for the formulation of absolute measures based on the prediction of document metadata, often considered as "ground truth" labels in the literature. However, extrinsic evaluation approaches, and the latter identification in particular, are also problematic because: (*i*) manual labeling is subjective and error prone; (*ii*) they evaluate the topic structure only indirectly (e.g. via the fraction of correctly classified documents); and (*iii*) they implicitly assume that the manually generated labels are truly encoded in the topic structure of the documents. The latter assumption has been shown to be surprisingly unsupported in other domains (Hric et al., 2014; Peel et al., 2017).

It has been recently shown that topic modeling can be formally mapped to the problem of community detection in networks (Karrer and Newman, 2011; Gerlach et al., 2017). The formulation of benchmark corpora pursued here follows the idea of benchmark graphs in community detection. There, the basic approach is to build synthetic networks with known (planted) community structure and evaluate an algorithm by comparing the overlap between the planted and the inferred community structures (Lancichinetti et al., 2008; Girvan and Newman, 2002; Danon et al., 2005; Sales-Pardo et al., 2007; Sawardecker et al., 2009; Lancichinetti and Fortunato, 2009; Guimerà et al., 2007). This approach allowed researchers to gain new insights into community detection algorithms such as (*i*) the spurious appearance of large values of modularity in random networks (Guimerà et al., 2004); (*ii*) the existence of a resolution limit concerning the minimum size of the groups that can be inferred (Fortunato and Barthelemy, 2007); or (*iii*) the existence of an undetectable phase in which no algorithm is able to infer a structure (Decelle et al., 2011).

The use of synthetic corpora has appeared sporadically in the context of topic modeling (see Supplementary Materials, Table S3). In most cases, the synthetic data comes from the generative process of LDA and is tested only on intrinsic evaluation methods such as held-out likelihood (Wallach et al., 2009b) or topic coherence (AlSumait et al., 2009). Comparison between planted and inferred structure is usually done by visual inspection (Griffiths and Steyvers, 2004; Andrzejewski et al., 2009), focuses only on either the word-topic or topic-document distribution requiring "matching of topics" (Taddy, 2012; Arora et al., 2013; Lancichinetti et al., 2015), or evaluate very specific hypothesis of the fitted model (such as the independence of words and documents in individual topics (Mimno et al., 2011)). Our work formalizes and generalizes these ideas: *(i)* by developing a framework to investigate a wide range of topical structures and including a number of realistic features that might be of interest to practitioners; and *(ii)* proposing a measure that compares the planted and inferred structure (i.e. the topic labels) on the level of individual word tokens.

3 Evaluating topic modeling algorithms using synthetic corpora

Our approach to comparing the performance of topic modeling algorithms using synthetic corpora consists of two main steps (Fig. 2) [1]. First, we generate a synthetic benchmark with a planted ground truth structure; and second, we quantify the overlap between the planted and inferred structures.

3.1 Generating synthetic corpora

Our approach to generating synthetic corpora with a planted structure is based on the formulation of the generative process employed by probabilistic topic models (Blei, 2012; Crain et al., 2012). Consider a corpus of $d = 1, \ldots, D$ documents each with length m_d (and $N = \sum_d m_d$ words in total) generated from K topics and V unique words defining the vocabulary $\mathcal{V}$. The statistical characteristics of the corpus are determined by two sets of conditional probabilities: $P(t|d)$, indicating the probability of topic t within document d; and $P(w|t)$, indicating the probability with which word w is used by topic t. Specifically, for each token $w(i_d)$, defined as the word at position $i_d = 1, \ldots, m_d$ in document d, we first draw a topic $z(i_d) = t$ with probability $P(t|d)$ and then a word $w(i_d) = w$ is chosen with probability $P(w|t = z(i_d))$. Typically, one makes assumptions about these probabilities in the form of prior distributions. For example, in the case of Latent Dirichlet allocation (LDA), it is assumed that $P(t|d)$ and $P(w|t)$ are drawn from Dirichlet distributions with hyperparameters α and β, respectively. Given an observed corpus, the aim in topic modeling is then to determine the most likely distributions $\hat{P}(t|d)$ and $\hat{P}(w|t)$ by inferring the latent topic variables $\hat{z}(i_d)$ (Fig. 2A).

Here, we take the inverse approach by *a priori* fixing the distributions $P(t|d)$ and $P(w|t)$ and using the generative process to produce a synthetic corpus. Formally, our generation process includes the following steps.

First, we assign each word $w \in \mathcal{V}$ from the vocabulary to either the stopwords set $\mathcal{V}_S$ ($V_S \equiv |\mathcal{V}_S|$) or topical word set $\mathcal{V}_T$ ($V_T \equiv |\mathcal{V}_T|$) such that $V = V_S + V_T$.

Second, we fix the global word distribution $P(w)$ ($\sum_w P(w) = 1$) and the number of topical words assigned to each topic V_t ($\sum_t V_t = V_T$). Here, we consider a uniform or power-law functional form for their distributions.

Third, we assume that each word w belongs uniquely to one topic t denoted by t_w assigned randomly (such that we have V_t words in topic t). This assignment determines the topic distribution $P(t)$ over the entire corpus

$$P(t) = \frac{\sum\limits_{w \in \mathcal{V}_T} \delta_{t_w,t} \cdot P(w)}{\sum\limits_{w \in \mathcal{V}_T} P(w)}, \qquad (1)$$

where $\delta_{i,j}$ is Kronecker delta function, i.e., $\delta_{i,j} = 1$ only if $i = j$. Assuming that each document d belongs uniquely to one topic denoted by t_d which is randomly assigned with probability $P(t)$.

Fourth, we define the word-topic distribution $P(w|t)$

[1] Code to generate synthetic corpora is available at: https://github.com/amarallab/synthetic_benchmark_topic_model

with structure parameter c_w as

$$P(w|t) = \begin{cases} c_w \ \delta_{t_w,t} \dfrac{P(w)}{P(t)} + (1-c_w)\ P(w), \text{ if } w \in \mathcal{V}_T \\ P(w), \text{ if } w \in \mathcal{V}_S \end{cases}. \quad (2)$$

While the topical words ($w \in \mathcal{V}_T$) are characterized by a linear combination of a structured term and a random, unstructured term, the stopwords ($w \in \mathcal{V}_S$) appear randomly in all topics. Similarly, we define the topic-document distribution $P(t|d)$ with structure parameter c_d as

$$P(t|d) = c_d \ \delta_{t_d,t} + (1-c_d)\ P(t), \quad (3)$$

where the first term is the structured part and the second is the random, unstructured part.

The resulting synthetic corpus contains a fully known planted structure since we know the topic label $z(i_d)$ of each individual token $w(i_d)$ (Fig. 2A). The general formulation not only allows us to investigate a wide range of topical structures, but also to incorporate statistical laws observed in real-world corpora (Altmann and Gerlach, 2016), such as Zipfian word-frequency distribution, stopwords, or burstiness (Fig. 2B-E), see *Supplementary Material* Sec. S5.

3.2 Comparing planted and inferred structure

Typically, the results of topic modeling algorithms are evaluated either at the level of the topic-document distribution $P(t|d)$ in applications such as document classification, or at the level of the word-topic distribution $P(w|t)$ to judge the topic quality such as in topic coherence (Bhatia et al., 2017). Here, we propose a new approach by quantifying the overlap between the planted and the inferred structure based on the comparison of the topic labels of each individual token.

Specifically, for each token $w(i_d)$ we record the planted topic label as $z^{\mathrm{pl}}(i_d)$ and the inferred topic label as $z^{\mathrm{inf}}(i_d)$ and construct a confusion matrix $p_{t,t'}$, which counts the fraction of tokens having a planted topic label t and an inferred topic label t'

$$p_{t,t'} = \frac{1}{N} \cdot \sum_{d=1}^{D} \sum_{i_d=1}^{m_d} \delta_{z^{\mathrm{pl}}(i_d),t} \cdot \delta_{z^{\mathrm{inf}}(i_d),t'}. \quad (4)$$

From this we calculate the normalized mutual information, $\hat{I}$, a commonly used metric to quantify the overlap between different partitions (Danon et al., 2005) defined as:

$$\hat{I} = \frac{2I}{H+H'}, \quad (5)$$

where I is the mutual information and H (and H') are the respective entropies

$$\begin{aligned} I &= \sum_t \sum_{t'} p_{t,t'} \log \frac{p_{t,t'}}{p_t p_{t'}}, \\ H &= -\sum_t p_t \log p_t, \ \ H' = -\sum_{t'} p_{t'} \log p_{t'}. \end{aligned} \quad (6)$$

We thus obtain a measure between $\hat{I} = 0$ indicating no overlap, and $\hat{I} = 1$ indicating perfect overlap. Note that $\hat{I}$ takes into account that the number of topics in the inference results does not have to match the number of planted classes (Fig. S1). The major advantage of the NMI is its easy interpretability: it quantifies the average amount of information one gains about the planted label of a token upon learning its inferred topic label. Furthermore, $\hat{I}$ is invariant with respect to permutation of the topic labels; thus we avoid the issue of finding the "best match" between planted and inferred topics, typically addressed by non-trivial heuristic approaches (Lancichinetti et al., 2015) (See Fortunato (2010) for advantages of $\hat{I}$ over other measures, such as Jaccard index).

This measure is related to the Variation of Information proposed in Schofield and Mimno (2016), i.e. $VOI = const. \times (1-\hat{I})$; however, while Schofield and Mimno (2016) compare different outputs of a topic modeling algorithm under different pre-processing steps, here we use the measure to compare the planted ground truth against the output of the topic modeling algorithm.

4 Results

We report three different experiments that illustrate how synthetic corpora can yield new insights on topic modeling algorithms. As a representative sample, we evaluate four topic modeling algorithms on these corpora: LDA using Gibbs sampling (LDAGS) (Griffiths and Steyvers, 2004; McCallum, 2002), LDA using variational inference (LDAVB) (Blei et al., 2003; Řehůřek and Sojka, 2010), Hierarchical Dirichlet Processes (HDP) (Teh et al., 2006; Wang, 2010), and TopicMapping (TM) (Lancichinetti et al., 2015; Lancichinetti, 2016) (see *Supplementary Material* Sec. S2 for details) using default parameter settings of the corresponding implementations unless stated otherwise.

4.1 Degree of structure

Our first experiment evaluates how modeling accuracy varies with the degree of topic structure in the synthetic corpus. Here, we consider a simple version of the synthetic corpus described in Sec. 3.1 with a single parameter for the degree of structure $c = c_w = c_d$ such that we can vary between a trivial ($c = 1$) and

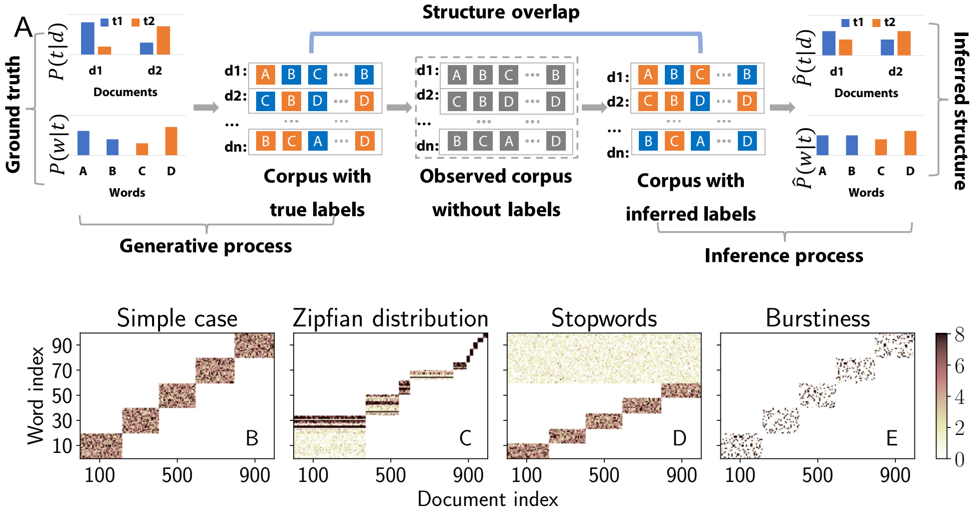

Figure 2: **Proposed framework for the evaluation of topic models based on synthetic corpora. (A)** Evaluation framework. **(B-E)** Examples of synthetic corpora with different statistical features observed in real-world corpora showing the number of occurrences of each word in each document with $D = 1000$ and $V = 100$.

an impossible ($c = 0$) inference problem (as shown in Fig. S2). More specifically, a smaller value of c corresponds to a higher level of noise in the synthetic corpus. In addition, we fix that there are no stopwords ($V_s = 0$), and that the global word-distribution and the topic-size distribution are uniform; $P(w) = 1/V$ and $V_t = V/K$.

In Fig. 3 we compare the overlap between planted and inferred structure as a function of c for synthetic corpora with $K = 10$ planted topics.

In general, the performance of all algorithms increases non-linearly with the degree of structure c (Fig. 3A). We observe substantial differences between algorithms for both the mean (identifying TM as a systematically more accurate algorithm) and the standard deviation (identifying HDP as a systematically less reproducible algorithm). We also observe a region ($c < c^*$ with $c^* \approx 0.3$), where none of algorithms are able to recover any structure ($\hat{I} = 0$) despite the fact that the synthetic corpus contains some small degree of structure ($c > 0$). The latter suggests the existence of an "undetectable phase", a phenomenon recently reported in the context of community detection (Decelle et al., 2011).

For the LDA algorithms in Fig. 3A, we assume the number of topics (a parameter which has to be specified *a priori* in LDA) is $K_a = 100$, which is a common choice for real-world corpora in the literature (Wallach et al., 2009b; Wei and Croft, 2006; Aletras et al., 2017; Steyvers and Griffiths, 2007). Not surprisingly, in Fig. 3B we observe a substantial improvement in performance when considering the unlikely case of guessing the correct number of topics ($K_a = K = 10$). We find that the performance of LDA algorithms is typically reduced by choosing both too many or too few topics highlighting how uninformed modeling assumptions can strongly affect performance (Fig. S3).

We further investigate how accurately non-parametric topic models such as HDP and TopicMapping can infer the number of topics (Fig. 3C). TopicMapping finds the correct number of topics even for only moderately structured corpora, but it completely fails for very unstructured corpora by overfitting the data reflecting the intrinsic difficulty when the signal-to-noise ratio is low. In contrast, HDP tends to overestimate the number of topics in this experiment, even more so as the degree of structure becomes large. This suggests that the model is arbitrarily splitting ground truth topics into distinct topics, a hypothesis that is corroborated by the relatively low reproducibility of the method (in terms of the average overlap between two different inferred solutions on the same data, as shown in Fig. S4). Thus, in this experiment we do not find the number of topics inferred by HDP to be reliable.

4.2 Impact of LDA-implementation and hyperparameter values

Despite the considerable advances in our understanding of LDA since its original formulation (Blei et al., 2003), we still lack a systematic understanding of the impact of different approximation tech-

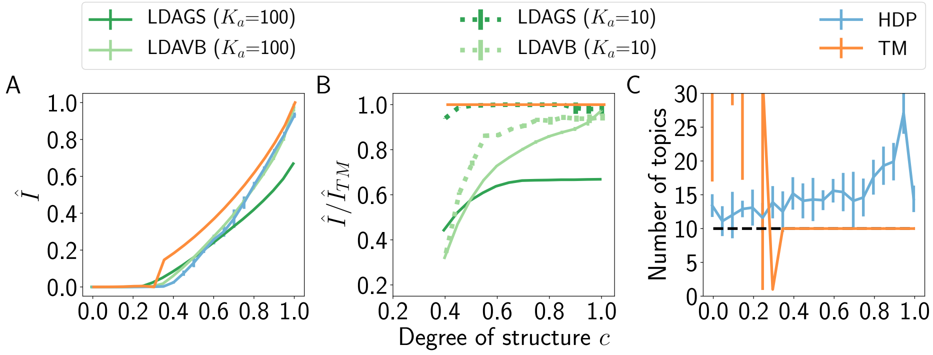

Figure 3: **Performance of topic modeling algorithms in synthetic corpora with varying degree of structure.** **(A)** Normalized mutual information, $\hat{I}$, between planted and inferred structures for different topic modeling algorithms as a function of the structure parameter c. **(B)** Relative performance of different topic modeling algorithms against TopicMapping, the best performing algorithm in **A**. **(C)** Number of inferred topics for non-parametric topic modeling algorithms. Synthetic corpora were generated with $K = 10$ topics, $D = 10^4$ documents, document length $m = 100$, and vocabulary size $V = 10^3$. The lines (error bars) denote averages (one standard deviation) estimated from 10 realizations.

niques on the performance (Zhang et al., 2016). While some groups have investigated the advantages of Collapsed Variational Bayes over mean-field Variational Bayes (Mukherjee and Blei, 2009) or the effect of hyperparameter choice (Asuncion et al., 2009; Wallach et al., 2009a), to our knowledge there have been no systematic studies exploring the inferred solutions in terms of the corresponding topic distributions and how they depend on the hyperparameters or inference algorithms.

In order to understand the differences between the Variational Bayes (VB) and Gibbs Sampling (GS) implementations of LDA observed in Fig. 3, we investigate in detail the planted and inferred $P(t|d)$ and $P(w|t)$ for both algorithms (Fig. 4). We find that neither can accurately infer the ground truth topic distributions endowed with a mixed structure in both $P(t|d)$ and $P(w|t)$. With default hyperparameters, the GS implementation infers a pure word-topic distribution and places the fluctuations almost exclusively on $P(t|d)$ (Fig. 4, 1st column). In contrast, the VB implementation infers a pure topic-document distribution and places the fluctuation mainly on $P(w|t)$ (Fig. 4, 2nd column). However, these differences can be explained, in part, by different default values for the hyperparameters. Assuming the correct number of topics ($K_a = 10$) and using the same hyperparameters (default values from VB implementation) for both the GS and the VB inference, we obtain almost identical results from the two LDA algorithms (Fig. 4, 2nd & 4th columns). In contrast, the VB implementation is virtually unable to infer any meaningful structure when using the default hyperparameters of Gibbs Sampling (Fig. 4, 5th column). Interestingly, when the true number of topics is unknown, we observe substantial differences in *how* the two algorithms overfit the ground truth structure (Fig. S5).

To ensure the reliability of these findings, we repeated our analyses increasing the number of iterations 10-fold for each algorithm, obtaining identical results (Figs. S6, S7).

These results confirm that the choice of default hyperparameters can bias the output of topic modeling algorithms. More generally, however, they show how our approach can reveal intricate differences in performance which are inaccessible in standard evaluation approaches such as document classification, where only partial information on the inferred structure is used, e.g., the maximum in the topic-document distribution $P(t|d)$ (Fig. S8).

4.3 Insights on real world corpora

The synthetic corpora discussed earlier constitute a simplified abstraction of the topic structure of real-world corpora. Thus, it may not be obvious that the insights drawn from synthetic corpora will be generalizable to real-world corpora. Therefore, we next investigate two examples supporting the hypothesis that despite its simplicity the synthetic corpus not only allows to make predictions on the performance of topic modeling algorithms in similar real-world corpora, but it also provides additional insights as to why different algorithms perform differently on distinct corpora (Fig. 5).

We measure performance in real-world corpora in an unsupervised classification task using human-assigned

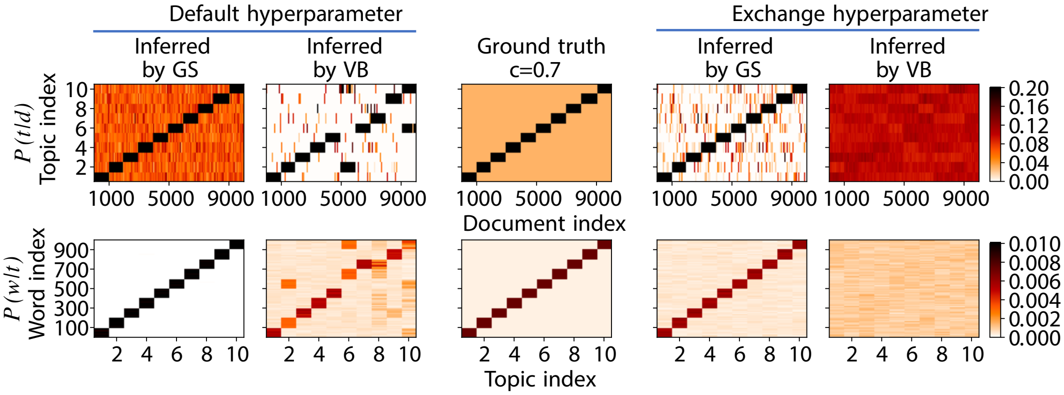

Figure 4: **Default hyperparameter settings bias the inferred topic structure of different LDA implementations.** Comparison of topic distributions $P(t|d)$ (top row) and $P(w|t)$ (bottom row) from the planted and inferred structure from LDAGS and LDAVB using two different sets of hyperparameters: original defaults as defined in each implementation (left panels) and defaults from the other implementation, respectively (right panels). Ground truth is displayed in the middle column. Same parameters as in Fig. 3 fixing $c = 0.7$ and using $K_a = 10$.

document labels as a ground truth proxy. In analogy to the approach in Eqs. (4,5) we quantify the correspondence between external and inferred document labels using the normalized mutual information (see *Supplementary Material* Sec.S4 for details).

Stopwords. While many practitioners remove stopwords from corpora prior to analysis, there is no consensus on the effect of stopwords on the performance of topic modeling algorithms (Zaman et al., 2011; Schofield et al., 2017). We thus investigate the effect of stopwords using the 20 News Group (20NG) dataset motivated by the fact that it exhibited the strongest dependence of performance on stopword shown in Fig. 1. Using the English stopword list from MALLET (McCallum, 2002), we estimate that about 43% of word tokens in the 20NG corpus are stopwords. For our analysis, we remove at random a given fraction of these tokens. We find that performance of topic modeling algorithms varies but generally increases as we decrease the fraction of stopwords (Fig. 5A).

We construct a synthetic corpus with $c = 0.7$ (we obtain similar results with different values, Fig. S9), $K = 40$, and a varying fraction P_s of stopwords. Measuring performance by unsupervised document classification, we find the same pattern as for the real corpus (Fig. 5B). In contrast, measuring performance as the overlap between planted and inferred structure yields substantial differences, which reflect the additional detail provided by the structure overlap (Fig. 5C). Considering the inferred topic distributions (Fig. S10), we find that LDAVB infers a pure topic-document distribution, assigning most of the uncertainty to the word-topic distribution and correctly identifying most of the stopwords, while LDAGS assigns most of the uncertainty to the topic-document distribution, trying to infer a pure word-topic distribution resulting in overfitting the stopwords and assigning them to inferred topics. In document classification, most of this information remains invisible, leading to indistinguishable results for the two algorithms.

Document length. It has been reported that topic models have low performance on corpora of short document, such as Twitter posts (Hong and Davison, 2010). However, the effect of document length on the performance of topic models is still not well characterized (Tang et al., 2014). We thus investigate the effect of text length by considering only the first m_d words of each document in the Web of Science (WOS) dataset, a collection of 40,526 scientific articles (title and abstract) from 7 academic areas. Prior to analysis, we removed all stopwords (using the stopword list from MALLET (McCallum, 2002)). We find that performance improves with increasing document length (Fig. 5D); yet, the ranking of the models' performance remains virtually unchanged.

We construct a synthetic corpus with similar properties fixing $c = 0.7$ (we obtain similar results with different values, Fig. S11) and $K = 10$ and varying the length m_d of each document. For both measures of performance, classification (Fig. 5E) and structure overlap (Fig. 5F) we qualitatively reproduce the findings on the real corpus. In particular, we recover the same ranking for the performance of topic models.

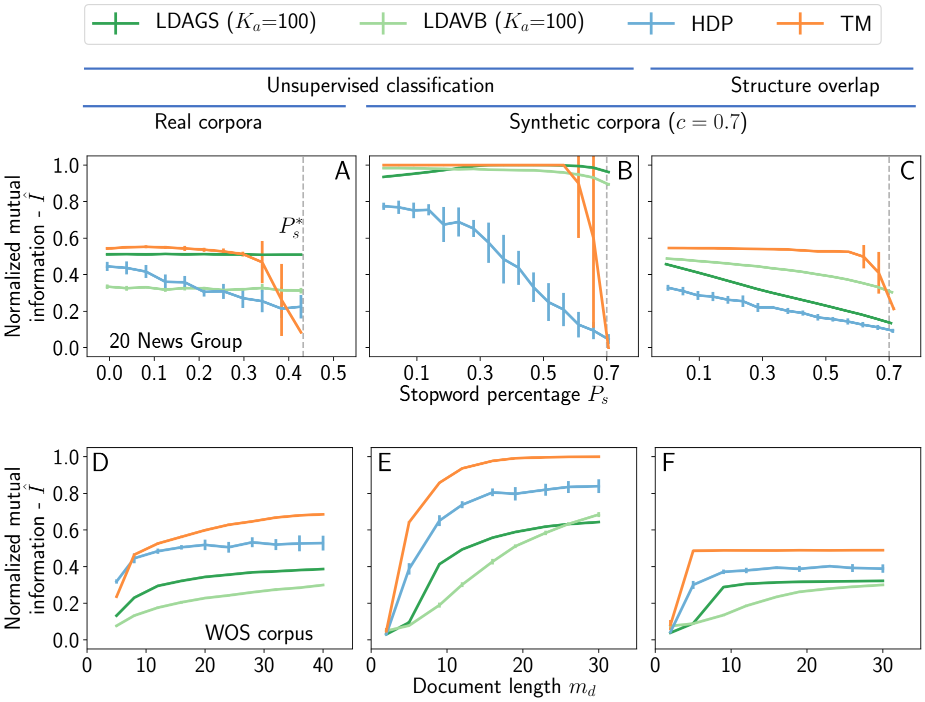

Figure 5: **Performance in synthetic corpora is strongly correlated to performance in real-world corpora.** Comparison between 20 News Group data and synthetic corpora with $K = 40$, $D = 10^4$ $m_d = 100$, $V = 10^3$, $c = 0.7$ varying the fraction of stopwords P_s (top row) using $K_a = 100$, and WOS data and a synthetic corpus with $K = 10$, $D = 10^4$, $V = 10^3$, $c = 0.7$ varying the document length m_d (bottom row). (**A**, **D**) NMI from unsupervised document classification in real-world corpora. (**B**, **E**) NMI from unsupervised document classification in synthetic corpora. (**C**, **F**) NMI from structure overlap in synthetic corpora. While each case measures NMI (in bits), panels (A,B,D,E) compare labels of documents and panels (C,F) compare labels of word tokens.

5 Discussion

Our study illustrates how the use of synthetic corpora can lead to new insights on topic model performance unattainable when only studying real-world corpora. Our approach allows us to systematically investigate the effect of both individual properties of the corpus (document length, stopwords, etc.) and parameters of the topic modeling algorithms (assumed number of topics, hyperparameters, etc.). For example, our analysis reveals that (*i*) the number of topics determined by popular non-parametric approaches (such as HDP) cannot be relied upon; (*ii*) there exist fundamental limits to algorithms' ability to infer a topic structure. and (*iii*) the default hyperparameter settings induce a substantial bias in the inferred solutions of different implementations of the same topic model. Most importantly, we demonstrate the practical relevance of our approach by showing that relative performance in synthetic corpora predicts relative performance in real-world corpora.

While these results raise more questions than they can answer, we believe that our proposed framework offers a complimentary approach to gain a better understanding of topic modeling algorithms. In particular, it allows us to systematically identify strengths and weaknesses of topic modeling algorithms in different applications and under different conditions allowing for more informed choices among a large number of available algorithms.

Unarguably, the presented synthetic corpora are far from the complexity of real-world corpora. However, our framework provides enough flexibility to accommodate different features such as burstiness, syntax, or structures beyond the bag-of-words model (phrases, sentences, etc.) in future studies with increasing complexity of the synthetic corpora.

Acknowledgements

L.A.N.A. thanks the John and Leslie McQuown Gift and support from Department of Defense Army Research Office (Grant Number W911NF-14-1-0259).

References

N. Aletras, T. Baldwin, J. H. Lau, and M. Stevenson. Evaluating topic representations for exploring document collections. *Journal of the Association for Information Science and Technology*, 68(1):154–167, 2017. ISSN 2330-1643.

M. J. Allen and W. M. Yen. *Introduction to Measurement Theory*. Waveland Press, 2001.

L. AlSumait, D. Barbará, J. Gentle, and C. Domeniconi. Topic significance ranking of LDA generative models. In *Machine Learning and Knowledge Discovery in Databases*, pages 67–82. Springer Berlin Heidelberg, 2009.

E. G. Altmann and M. Gerlach. Statistical laws in linguistics. In *Creativity and Universality in Language*, pages 7–26. Springer International Publishing, 2016. ISBN 978-3-319-24403-7.

D. Andrzejewski, X. Zhu, and M. Craven. Incorporating domain knowledge into topic modeling via dirichlet forest priors. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 25–32. ACM, 2009.

S. Arora, R. Ge, Y. Halpern, D. Mimno, A. Moitra, D. Sontag, Y. Wu, and M. Zhu. A practical algorithm for topic modeling with provable guarantees. In *Proceedings of the 30th International Conference on Machine Learning*, pages 280–288. PMLR, 2013.

A. Asuncion, M. Welling, P. Smyth, and Y. W. Teh. On smoothing and inference for topic models. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 27–34. AUAI Press, 2009. ISBN 978-0-9749039-5-8.

D. L. Bandalos. *Measurement Theory and Applications for the Social Sciences*. Guilford Publications, 2018.

S. Bhatia, J. H. Lau, and T. Baldwin. An automatic approach for document-level topic model evaluation. *arXiv:1706.05140*, 2017.

D. M. Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77, 2012.

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

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

J. Boyd-Graber, Y. Hu, and D. Mimno. Applications of topic models. *Foundations and Trends in Information Retrieval*, 11:143–296, 2017.

W. L. Buntine and S. Mishra. Experiments with non-parametric topic models. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 881–890. ACM Press, 2014. ISBN 9781450329569.

J. Chang, S. Gerrish, C. Wang, J. Boyd-graber, and D. M. Blei. Reading tea leaves: How humans interpret topic models. In *Advances in Neural Information Processing Systems*, pages 288–296. Curran Associates, Inc., 2009.

S. P. Crain, K. Zhou, S.-H. Yang, and H. Zha. Dimensionality reduction and topic modeling: From Latent Semantic Indexing to Latent Dirichlet allocation and beyond. In *Mining Text Data*, chapter 4, pages 1–522. Springer US, 2012.

L. Danon, A. Daz-Guilera, J. Duch, and A. Arenas. Comparing community structure identification. *Journal of Statistical Mechanics: Theory and Experiment*, 2005(09):P09008, 2005.

A. Decelle, F. Krzakala, C. Moore, and L. Zdeborová. Inference and phase transitions in the detection of modules in sparse networks. *Physical Review Letters*, 107(6):065701, 2011.

S. Fortunato. Community detection in graphs. *Physics Reports*, 486:75–174, 2010.

S. Fortunato and M. Barthelemy. Resolution limit in community detection. *Proceedings of the National Academy of Sciences*, 104(1):36–41, 2007.

M. Gerlach, T. P. Peixoto, and E. G. Altmann. A network approach to topic models. *arXiv: 1708.01677*, pages 1–19, 2017.

M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826, 2002.

T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101:5228–5235, 2004.

R. Guimerà, M. Sales-Pardo, and L. A. N. Amaral. Modularity from fluctuations in random graphs and complex networks. *Physical Review E*, 70(2):025101, 2004.

R. Guimerà, M. Sales-Pardo, and L. A. N. Amaral. Module identification in bipartite and directed networks. *Physical Review E*, 76:036102, 2007.

L. Hong and B. D. Davison. Empirical study of topic modeling in twitter. In *Proceedings of the First Workshop on Social Media Analytics*, pages 80–88. ACM, 2010. ISBN 978-1-4503-0217-3.

D. Hric, R. K. Darst, and S. Fortunato. Community detection in networks: Structural communities versus ground truth. *Physical Review E*, 90(6):062805, 2014.

A. K. Jain. Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 31:651–666, 2010.

B. Karrer and M. E. J. Newman. Stochastic blockmodels and community structure in networks. *Physical Review E*, 83(1):016107, 2011.

A. Lancichinetti. Topicmapping. *https://bitbucket.org/andrealanci/topicmapping*, 2016.

A. Lancichinetti and S. Fortunato. Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities. *Physical Review E*, 80:016118, 2009.

A. Lancichinetti, S. Fortunato, and F. Radicchi. Benchmark graphs for testing community detection algorithms. *Physical Review E*, 78:046110, 2008.

A. Lancichinetti, M. Irmak Sirer, J. X. Wang, D. Acuna, K. Kording, and L. A. N. Amaral. High-reproducibility and high-accuracy method for automated topic classification. *Physical Review X*, 5(1): 011007, 2015.

Y. Lu, Q. Mei, and C. Zhai. Investigating task performance of probabilistic topic models: An empirical study of PLSA and LDA. *Information Retrieval*, 14 (2):178–203, 2011.

A. K. McCallum. Mallet: A machine learning for language toolkit. *http://mallet.cs.umass.edu*, 2002.

D. Mimno, H. M. Wallach, E. Talley, M. Leenders, and A. McCallum. Optimizing semantic coherence in topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 262–272. Association for Computational Linguistics, 2011. ISBN 978-1-937284-11-4.

I. Mukherjee and D. M. Blei. Relative performance guarantees for approximate inference in latent Dirichlet allocation. In *Advances in Neural Information Processing Systems*, pages 1129–1136. Curran Associates, Inc., 2009.

A. Murakami, P. Thompson, S. Hunston, and D. Vajn. What is this corpus about?': Using topic modelling to explore a specialised corpus. *Corpora*, 12(2):243–277, 2017.

D. Newman, J. H. 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*, pages 100–108. Association for Computational Linguistics, 2010. ISBN 1-932432-65-5.

L. Peel, D. B. Larremore, and A. Clauset. The ground truth about metadata and community detection in networks. *Science Advances*, 3(5):e1602548, 2017.

R. Řehůřek and P. Sojka. Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50. ELRA, 2010.

M. Röder, A. Both, and A. Hinneburg. Exploring the space of topic coherence measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, pages 399–408. ACM, 2015. ISBN 978-1-4503-3317-7.

M. Sales-Pardo, R. Guimerà, A. A. Moreira, and L. A. N. Amaral. Extracting the hierarchical organization of complex systems. *Proceedings of the National Academy of Sciences*, 104(39):15224–15229, 2007. ISSN 0027-8424.

E. N. Sawardecker, M. Sales-Pardo, and L. A. N. Amaral. Detection of node group membership in networks with group overlap. *The European Physical Journal B*, 67(3):277–284, 2009.

A. Schofield and D. Mimno. Comparing apples to apple: The effects of stemmers on topic models. *Transactions of the Association for Computational Linguistics*, 4:287–300, 2016.

A. Schofield, M. Magnusson, and D. Mimno. Pulling out the stops: Rethinking stopword removal for topic models. In *Proceedings of Conference of the European Chapter of the Association for Computational Linguistics*, pages 432–436. Association for Computational Linguistics, 2017.

H. Schütze, C. D. Manning, and P. Raghavan. *Introduction to Information Retrieval.* Cambridge University Press, 2008.

M. Steyvers and T. Griffiths. Probabilistic topic models. *Handbook of latent semantic analysis*, 427(7): 424–440, 2007.

M. Taddy. On estimation and selection for topic models. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, pages 1184–1193. PMLR, 2012.

J. Tang, Z. Meng, X. Nguyen, Q. Mei, and M. Zhang. Understanding the limiting factors of topic modeling via posterior contraction analysis. In *Proceedings of the 31st International Conference on Machine Learning*, pages 190–198. PMLR, 2014.

Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476): 1566–1581, 2006.

H. M. Wallach, D. Mimno, and A. McCallum. Rethinking LDA: Why priors matter. In *Advances in Neural Information Processing Systems*, 2009a.

H. M. Wallach, I. Murray, R. Salakhutdinov, and D. Mimno. Evaluation methods for topic models. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1105–1112. ACM Press, 2009b.

C. Wang. Hierarchical Dirichlet process. *https://github.com/blei-lab/hdp*, 2010.

X. Wei and W. B. Croft. LDA-based document models for ad-hoc retrieval. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 178–185. ACM, 2006. ISBN 1-59593-369-7.

P. Xie and E. P. Xing. Integrating document clustering and topic modeling. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, pages 694–703. AUAI Press, 2013.

A. N. K. Zaman, P. Matsakis, and C. Brown. Evaluation of stop word lists in text retrieval using Latent Semantic Indexing. In *International Conference on Digital Information Management*, pages 133–136. IEEE, 2011.

J. Zhang, J. Zeng, M. Yuan, W. Rao, and J. Yan. LDA Revisited. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pages 1763–1772. ACM Press, 2016.