A network approach to topic models

Martin Gerlach,[1,2] Tiago P. Peixoto,[3,4] and Eduardo G. Altmann[5,2]

[1] Department of Chemical and Biological Engineering, Northwestern University, Evanston, IL 60208, USA
[2] Max Planck Institute for the Physics of Complex Systems, D-01187 Dresden, Germany
[3] Department of Mathematical Sciences and Centre for Networks and Collective Behaviour, University of Bath, Claverton Down, Bath BA2 7AY, United Kingdom
[4] ISI Foundation, Via Alassio 11/c, 10126 Torino, Italy
[5] School of Mathematics and Statistics, University of Sydney, 2006 NSW, Australia

One of the main computational and scientific challenges in the modern age is to extract useful information from unstructured texts. Topic models are one popular machine-learning approach which infers the latent topical structure of a collection of documents. Despite their success — in particular of its most widely used variant called Latent Dirichlet Allocation (LDA) — and numerous applications in sociology, history, and linguistics, topic models are known to suffer from severe conceptual and practical problems, e.g. a lack of justification for the Bayesian priors, discrepancies with statistical properties of real texts, and the inability to properly choose the number of topics. Here, we approach the problem of identifying topical structures by representing text corpora as bipartite networks of documents and words and using methods from community detection in complex networks, in particular stochastic block models (SBM). We show that our SBM-based approach constitutes a more principled and versatile framework for topic modeling solving the intrinsic limitations of Dirichlet-based models through a more general choice of nonparametric priors. It automatically detects the number of topics and hierarchically clusters both the words and documents. In practice, we demonstrate through the analysis of artificial and real corpora that our approach outperforms LDA in terms of statistical model selection.

I. INTRODUCTION

The accelerating rate of digitization of information increases the importance and number of problems which require automatic organization and classification of written text. Topic models [1] are a flexible and widely used tool which identifies semantically related documents through the topics they address. These methods originated in machine learning [2] and were largely based on heuristic approaches such as singular value decomposition in latent semantic indexing (LSI) [3]. Only a more statistically principled approach, based on the formulation of probabilistic generative models [4], allowed for a deeper theoretical foundation within the framework of Bayesian statistical inference [5]. This, in turn, lead to a series of key developments, in particular probabilistic latent semantic indexing (pLSI) [6] and latent Dirichlet allocation (LDA) [7, 8]. The latter established itself as the state-of-the-art method in topic modeling and has been widely used not only for recommendation and classification [9] but also bibliometrical [10–12], psychological [13], and political [14] analysis. Beyond the scope of natural language, LDA has also been applied in biology [15] (developed independently in this context [16]), image processing [17], and music [18].

However, despite its success and overwhelming popularity, LDA is known to suffer from fundamental flaws in the way it represents text. In particular, it lacks an intrinsic methodology to choose the number of topics, and contains a large number of free parameters that can cause overfitting. Furthermore, there is no principled justification for the use of the Dirichlet prior in the model formulation, which is done simply due to mathematical convenience. In particular, this choice is not designed to be compatible with well-known properties of real text [19], such as Zipf's law [20] for the frequency of words. More recently, consistency problems have also been identified with respect to how planted structures in artificial corpora can be recovered with LDA [21]. A substantial part of the research in topic models focuses on creating more sophisticated and realistic versions of LDA that account for, e.g., syntax [22], correlations between topics [23], meta-information (such as authors) [24], or burstiness [25]. Other approaches consist of post-inference fitting of the number of topics [26, 27] or the hyperparameters [28, 29], or the formulation of nonparametric hierarchical models [30, 31]. While all these approaches lead to demonstrable improvements, they do not provide satisfying solutions to the aforementioned issues because they share the limitations due to the choice of Dirichlet priors and/or they introduce additional heuristic approaches in the optimization of the free parameters.

A seemingly unrelated — but equally popular — problem is the detection of communities in complex networks [32], which have been shown to provide a useful abstraction of many complex systems in nature and society [33]. The idea of community detection is to find large-scale structure, i.e. the identification of groups of nodes with similar connectivity patterns. This is motivated by the fact that these groups describe the heterogeneity (i.e. the nonrandom structure) of the network and may correspond to functional units, giving potential insights on the generative mechanisms behind the network formation. While there is a variety of different

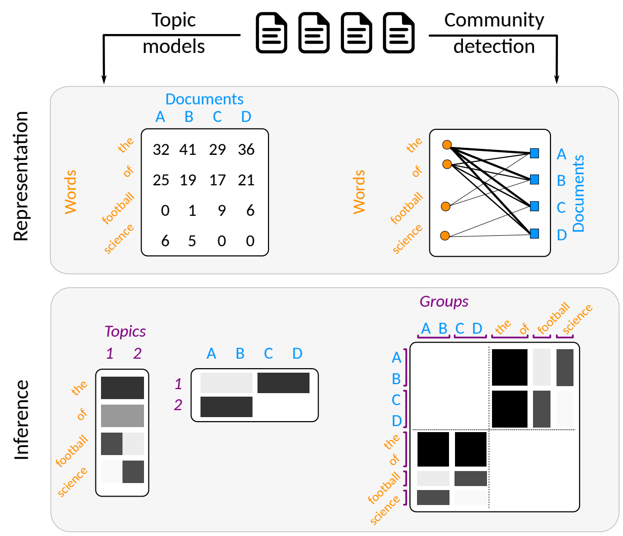

FIG. 1. Two approaches to extract information from collections of texts. Topic models represent the texts as a document-word matrix (how often each word appears in each document) which is then written as a product of two matrices of smaller dimensions with the help of the latent variable topic. The approach we propose here represents texts as a network and infers communities in this network. The nodes consists of documents and words and the strength of the edge between them is given by the number of occurrences of the word in the document, yielding a bi-partite multigraph that is equivalent to the word-document matrix used in topic models.

approaches to community detection [34, 35], most methods are heuristic and optimize a quality function, the most popular being modularity [36]. Modularity suffers from severe conceptual deficiencies, such as its inability to assess statistical significance leading to detection of groups in completely random networks [37], or its incapacity in finding groups below a given size [38]. Methods like modularity maximization are analogous to the pre-pLSI heuristic approaches to topic models, sharing with them many conceptual and practical deficiencies. In an effort to quench these problems, many researchers moved to probabilistic inference approaches, most notably those based on stochastic block models (SBM) [39–41], mirroring the same trend that occurred in topic modelling.

In this paper we propose and apply a unified framework to the fields of topic modeling and community detection. As illustrated in Fig. 1, by representing the word-document matrix as a bipartite network the problem of inferring topics becomes a problem of inferring communities. We use this correspondence, and the mathematical equivalence between pLSI of texts and SBMs of networks [42], to adapt results and methods developed within the community-detection framework to topic models. In particular, we derive a nonparametric Bayesian parametrization of pLSI — adapted from a hierarchical stochastic block model (hSBM) [43–45] — that makes fewer assumptions about the underlying structure of the data. As a consequence, it better matches the statistical properties of real texts and solves many of the intrinsic limitations of LDA. We show that our model outperforms LDA not only in various real corpora but even in artificial corpora generated from LDA itself. Additionally, our non-parametric approach uncovers topical structures on many scales of resolution, automatically determines the number of topics together with the word classification, and its symmetric formulation allows the documents themselves to be clustered into hierarchical categories. The goal of our manuscript is to, first, show that Dirichlet-based topic models should be abandoned in favour of more principled models and, second, to show how the largely independent fields of topic modeling and community detection are better addressed together.

II. COMMUNITY DETECTION FOR TOPIC MODELING

In this section we expose the connection between topic modeling and community detection, as illustrated in Fig. 2. We first revisit how a Bayesian formulation of pLSI assuming Dirichlet priors leads to LDA and how the former can be re-interpreted as a mixed-membership SBM. We then use the latter to derive a more principled approach to topic modeling using non-parametric and hierarchical priors.

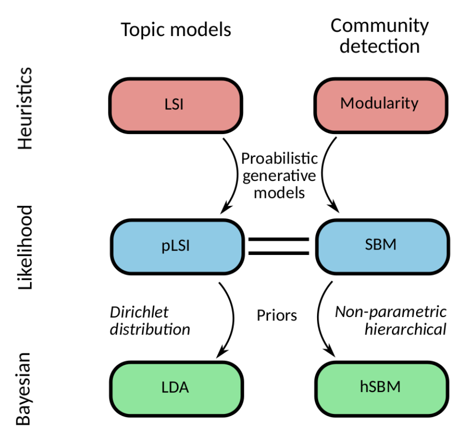

FIG. 2. Paralelism between topic models and community-detection methods. The probabilistic latent semantic index (pLSI) and stochastic block models (SBM) are mathematically equivalent and therefore methods from community detection (e.g., the hSBM we propose in this manuscript) can be used as alternatives to traditional topic models (e.g., LDA).

A. Topic models: pLSI and LDA

PLSI is a model that generates a corpus composed of D documents, where each document d has k_d words [6]. The placement of the words in the documents is done based on the assignment of topic mixtures to both document and words, from a total of K topics. More specifically, one iterates through all D documents, and for each document d one samples $k_d \sim \mathrm{Poi}(\eta_d)$ and for each word token $l \in \{1, k_d\}$, first a topic r is chosen with probability θ_{dr}, and then a word w is chosen from that topic with probability ϕ_{rw}. If n^r_{dw} is the number of occurrences of word w of topic r in document d (summarized as $\boldsymbol{n}$), the probability of a corpus is [46]

$$P(\boldsymbol{n}|\boldsymbol{\eta},\boldsymbol{\theta},\boldsymbol{\phi}) = \prod_d \eta_d^{k_d} e^{-\eta_d} \prod_{wr} \frac{(\phi_{rw}\theta_{dr})^{n^r_{dw}}}{n^r_{dw}!}. \quad (1)$$

For an unknown text, we could simply maximize (1) to obtain the best parameters $\boldsymbol{\eta}$, $\boldsymbol{\theta}$, and $\boldsymbol{\phi}$ which describe the topical structure of the corpus. However, this approach cannot be used directly to model textual data without a significant danger of overfitting. The model possess a large number of parameters, that grows as the number of documents, words, and topics is increased, and hence a maximum likelihood estimate will invariably incorporate a considerable amount of noise. One solution to this problem is to employ a Bayesian formulation, by proposing prior distributions to the parameters, and integrating over them. This is precisely what is done in LDA [7, 8], where one chooses Dirichlet priors $D_d(\boldsymbol{\theta}_d|\boldsymbol{\alpha}_d)$ and $D_r(\boldsymbol{\phi}_r|\boldsymbol{\beta}_r)$ with hyperparameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ for the probabilities $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ above, and one uses instead the marginal likelihood.

$$\begin{aligned}
&P(\boldsymbol{n}|\boldsymbol{\eta},\boldsymbol{\beta},\boldsymbol{\alpha}) \\
&= \int P(\boldsymbol{n}|\boldsymbol{\eta},\boldsymbol{\theta},\boldsymbol{\phi}) \prod_d D_d(\boldsymbol{\theta}_d|\boldsymbol{\alpha}_d) \prod_r D_r(\boldsymbol{\phi}_r|\boldsymbol{\beta}_r)\, \mathrm{d}\boldsymbol{\theta}\mathrm{d}\boldsymbol{\phi}, \\
&= \prod_d \eta_d^{k_d} e^{-\eta_d} \prod_{wr} \frac{1}{n^r_{dw}!} \times \\
&\quad \prod_d \frac{\Gamma(\sum_r \alpha_{dr})}{\Gamma(k_d + \sum_r \alpha_{dr})} \prod_r \frac{\Gamma(\sum_w n^r_{dw} + \alpha_{dr})}{\Gamma(\alpha_{dr})} \times \\
&\quad \prod_r \frac{\Gamma(\sum_w \beta_{rw})}{\Gamma(\sum_{dw} n^r_{dw} + \sum_w \beta_{rw})} \prod_w \frac{\Gamma(\sum_d n^r_{dw} + \beta_{rw})}{\Gamma(\beta_{rw})},
\end{aligned} \quad (2)$$

If one makes a noninformative choice, i.e. $\alpha_{dr} = 1$ and $\beta_{rw} = 1$, inference using Eq. (2) is nonparametric and less susceptible to overfitting. In particular, one can obtain the labeling of word tokens into topics, n^r_{dw}, conditioned only on the observed total frequencies of words in documents, $\sum_r n^r_{dw}$, in addition to the number of topics K itself, simply by maximizing or sampling from the posterior distribution. The weakness of this approach rests in the fact that the Dirichlet prior is a simplistic assumption about the data-generating process: In its noninformative form, every mixture in the model — both of topics in each document as well as words into topics — is assumed to be equally likely, precluding the existence of any form of higher-order structure. This limitation has prompted the widespread practice of inferring using LDA in a parametric way, by maximizing the likelihood with respect to the hyperparameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, which can improve the quality of fit in many cases. But not only this undermines to a large extent the initial purpose of a Bayesian approach — as the number of hyperparameters still increase with the number of documents, words and topics, and hence maximizing over them reintroduces the danger of overfitting — but also it does not sufficiently address the original limitation of the Dirichlet prior. Namely, regardless of the hyperparameter choice, the Dirichlet distribution is *unimodal*, meaning that it generates mixtures which are either concentrated around the mean value, or spread away uniformly from it towards pure components. This means that for any choice of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ the whole corpus is characterized by a single typical mixture of topics into documents, and a single typical mixture of words into topics. This is an extreme level of assumed homogeneity which stands in contradiction to a clustering approach initially designed to capture heterogeneity.

In addition to the above, the use of nonparametric Dirichlet priors is inconsistent with well-known universal statistical properties of real texts; most notably the highly-skewed distribution of word frequencies, which typically follows Zipf's law [20]. In contrast, the noninformative choice of the Dirichlet distribution with hyperparameters $\beta_{rw} = 1$ amounts to an expected *uniform* frequency of words in topics and documents. Although this disagreement can be addressed by choosing appropriate values of β_{rw}, such an approach, as already mentioned, runs contrary to nonparametric inference, and is subject to overfitting.

In the following, we will show how the same original pLSI model can be re-cast as a network model that completely removes the limitations described above, and is capable of uncovering heterogeneity in the data at multiple scales.

B. Topic models and community detection: Equivalence between pLSI and SBM

We show that pLSI is equivalent to a specific form of a mixed membership SBM as proposed by Ball et al. [42].

The SBM is a model that generates a network composed of $i = 1, \ldots, N$ nodes with adjacency matrix A_{ij}, which we will assume without loss of generality to correspond to a multigraph, i.e. $A_{ij} \in \mathbb{N}$. The nodes are placed in a partition composed of B overlapping groups, and the edges between nodes i and j are sampled from a Poisson distribution with average

$$\sum_{rs} \kappa_{ir}\omega_{rs}\kappa_{js}, \quad (3)$$

where ω_{rs} is the expected number of edges between group r and group s, and κ_{ir} is the probability that node i is sampled from group r. The likelihood to observe $\boldsymbol{\mathcal{A}} = \{\mathcal{A}_{ij}^{rs}\}$, i.e. a particular decomposition of A_{ij} into labelled half-edges such that $A_{ij} = \sum_{rs} \mathcal{A}_{ij}^{rs}$, can be written as

$$P(\boldsymbol{\mathcal{A}}|\boldsymbol{\kappa},\boldsymbol{\omega}) = \prod_{i<j}\prod_{rs} \frac{e^{-\kappa_{ir}\omega_{rs}\kappa_{is}}(\kappa_{ir}\omega_{rs}\kappa_{js})^{\mathcal{A}_{ij}^{rs}}}{\mathcal{A}_{ij}^{rs}!} \times \prod_{i}\prod_{rs} \frac{e^{-\kappa_{ir}\omega_{rs}\kappa_{is}/2}(\kappa_{is}\omega_{rs}\kappa_{is}/2)^{\mathcal{A}_{ii}^{rs}/2}}{\mathcal{A}_{ii}^{rs}/2!}, \tag{4}$$

by exploiting the fact that the sum of Poisson variables is also distributed according to a Poisson.

The connection to pLSI can now be made by rewriting the token probabilities in Eq. (1) in a symmetric fashion as

$$\phi_{rw}\theta_{dr} = \eta_w \theta_{dr}\phi'_{wr}, \tag{5}$$

where $\phi'_{wr} \equiv \phi_{rw}/\sum_s \phi_{sw}$ is the probability that the word w belongs to topic r, and $\eta_w \equiv \sum_s \phi_{sw}$ is the overall propensity with which the word w is chosen across all topics. In this manner, the likelihood (1) can be rewritten as

$$P(\boldsymbol{n}|\boldsymbol{\eta},\boldsymbol{\phi}',\boldsymbol{\theta}) = \prod_{dwr} \frac{e^{-\lambda_{dw}^r}(\lambda_{dw}^r)^{n_{dw}^r}}{n_{dw}^r!}, \tag{6}$$

with $\lambda_{dw}^r = \eta_d\eta_w\theta_{dr}\phi'_{wr}$. If we choose to view the counts n_{dw} as the entries of the adjacency matrix of a bipartite multigraph (with documents and words as nodes), the likelihood (6) is equivalent to the likelihood (4) of the SBM, if we assume that each document belongs to its own specific group, $\kappa_{ir} = \delta_{ir}$, with $i = 1,\ldots,D$ for document nodes, and by re-writing $\lambda_{dw}^r = \omega_{dr}\kappa_{rw}$. Therefore, the SBM of Eq. (4) is a generalization of pLSI that allows the words as well as the documents to be clustered into groups, and includes it as a special case when the documents are not clustered.

In the symmetric setting of the SBM, we make no explicit distinction between words and documents, both of which become nodes in different partitions of a bipartite network. We base our Bayesian formulation that follows on this symmetric parametrization.

C. Community detection and the hierarchical SBM

Taking advantage of the above connection between pLSI and SBM, we show how the idea of hierarchical SBMs developed in Refs. [43–45] can be extended such that they can be effectively used for the inference of topical structure in texts.

Like pLSI, the SBM likelihood of Eq. (4) contains a large number of parameters that grows with the number of groups, and therefore cannot be used effectively without knowing the most appropriate dimension of the model beforehand. Analogously to what is done in LDA, this can be addressed by assuming noninformative priors for the parameters $\boldsymbol{\kappa}$ and $\boldsymbol{\omega}$, and computing the marginal likelihood (see SM, Sec. IA, for an explicit expression)

$$P(\boldsymbol{\mathcal{A}}|\bar{\omega}) = \int P(\boldsymbol{\mathcal{A}}|\boldsymbol{\kappa},\boldsymbol{\omega})P(\boldsymbol{\kappa})P(\boldsymbol{\omega}|\bar{\omega})\,\mathrm{d}\boldsymbol{\kappa}\mathrm{d}\boldsymbol{\omega}, \tag{7}$$

where $\bar{\omega}$ is a global parameter determining the overall density of the network. This can be used to infer the labelled adjacency matrix $\{\mathcal{A}_{ij}^{rs}\}$ as done in LDA, with the difference that not only the words but also the documents would be clustered into mixed categories.

However, at this stage the model still shares some disadvantages with LDA. In particular, the noninformative priors make unrealistic assumptions about the data, where the mixture between groups and the distribution of nodes into groups is expected to be unstructured. Among other problems, this leads to a practical obstacle, as this approach possesses a "resolution limit" where at most $O(\sqrt{N})$ groups can be inferred on a sparse network with N nodes [45, 47]. In the following we propose a qualitatively different approach to the choice of priors by re-formulating the above model as an equivalent *microcanonical* model [45] (see SM, Sec. IB, for a proof) such that we can write the marginal likelihood as the joint likelihood of the data and its discrete parameters,

$$P(\boldsymbol{\mathcal{A}}|\bar{\omega}) = P(\boldsymbol{\mathcal{A}},\boldsymbol{k},\boldsymbol{e}|\bar{\omega}) = P(\boldsymbol{\mathcal{A}}|\boldsymbol{k},\boldsymbol{e})P(\boldsymbol{k}|\boldsymbol{e})P(\boldsymbol{e}|\bar{\omega}), \tag{8}$$

with

$$P(\boldsymbol{\mathcal{A}}|\boldsymbol{k},\boldsymbol{e}) = \frac{\prod_{r<s} e_{rs}! \prod_r e_{rr}!! \prod_{ir} k_i^r!}{\prod_{rs}\prod_{i<j}\mathcal{A}_{ij}^{rs}! \prod_i \mathcal{A}_{ii}^{rs}!! \prod_r e_r!} \tag{9}$$

$$P(\boldsymbol{k}|\boldsymbol{e}) = \prod_r \left(\!\!\binom{e_r}{N}\!\!\right)^{-1} \tag{10}$$

$$P(\boldsymbol{e}|\bar{\omega}) = \prod_{r\le s} \frac{\bar{\omega}^{e_{rs}}}{(\bar{\omega}+1)^{e_{rs}+1}} = \frac{\bar{\omega}^E}{(\bar{\omega}+1)^{E+B(B+1)/2}}. \tag{11}$$

where $e_{rs} = \sum_{ij}\mathcal{A}_{ij}^{rs}$ is the total number of edges between groups r and s (we used the shorthand $e_r = \sum_s e_{rs}$ and $k_i^r = \sum_{js}\mathcal{A}_{ij}^{rs}$), $P(\boldsymbol{\mathcal{A}}|\boldsymbol{k},\boldsymbol{e})$ is the probability of a labelled graph $\boldsymbol{\mathcal{A}}$ where the labelled degrees $\boldsymbol{k}$ and edge counts between groups $\boldsymbol{e}$ are constrained to specific values (and not their expectation values), $P(\boldsymbol{k}|\boldsymbol{e})$ is the uniform prior distribution of the labelled degrees constrained by the edge counts $\boldsymbol{e}$, and $P(\boldsymbol{e}|\bar{\omega})$ is the prior distribution of edge counts, given by a mixture of independent geometric distributions with average $\bar{\omega}$.

The main advantage of this alternative model formulation is that it allows us to remove the homogeneous assumptions by replacing the uniform priors $P(\boldsymbol{k}|\boldsymbol{e})$ and $P(\boldsymbol{e}|\bar{\omega})$ by a hierarchy of priors and hyperpriors that incorporate the possibility of higher-order structures. This can be achieved in a tractable manner without the need of solving complicated integrals that would be required by introducing deeper Bayesian hierarchies in Eq. (7) directly.

In a first step, we follow the approach of Ref. [44] and condition the labelled degrees $\boldsymbol{k}$ on an overlapping partition $\boldsymbol{b} = \{b_{ir}\}$, given by

$$b_{ir} = \begin{cases} 1 & \text{if } k_i^r > 0, \\ 0 & \text{otherwise,} \end{cases} \tag{12}$$

such that they are sampled by a distribution

$$P(\boldsymbol{k}|\boldsymbol{e}) = P(\boldsymbol{k}|\boldsymbol{e},\boldsymbol{b})P(\boldsymbol{b}). \tag{13}$$

Importantly, the labeled degree sequence is sampled conditioned on the *frequency of degrees* $\boldsymbol{n}_{\boldsymbol{k}}^{\boldsymbol{b}}$ inside each mixture $\boldsymbol{b}$, which itself is sampled from its own noninformative prior,

$$P(\boldsymbol{k}|\boldsymbol{e},\boldsymbol{b}) = \left[\prod_{\boldsymbol{b}} P(\boldsymbol{k}_{\boldsymbol{b}}|\boldsymbol{n}_{\boldsymbol{k}}^{\boldsymbol{b}})P(\boldsymbol{n}_{\boldsymbol{k}}^{\boldsymbol{b}}|\boldsymbol{e}_{\boldsymbol{b}},\boldsymbol{b})\right] P(\boldsymbol{e}_{\boldsymbol{b}}|\boldsymbol{e},\boldsymbol{b}), \tag{14}$$

where $\boldsymbol{e}_{\boldsymbol{b}}$ is the number of incident edges in each mixture (see SM, Sec. IC, for detailed expressions).

Due to the fact that the frequencies of the mixtures as well as the frequencies of the labelled degrees are treated as latent variables, this model admits group mixtures which are far more heterogeneous than the Dirichlet prior used in LDA. In particular, as was shown in Ref. [45], the expected degrees generated in this manner follow a Bose-Einstein distribution, which is much broader than the exponential distribution obtained with the prior of Eq. (10). More importantly, the asymptotic form of the degree likelihood will approach the true distribution as the prior washes out [45], making it more suitable for skewed empirical frequencies, such as Zip's law or mixtures thereof [48], without requiring specific parameters — such as exponents — to be determined a priori.

In a second step, we follow Refs. [43, 45] and model the prior for the edge counts $\boldsymbol{e}$ between groups by interpreting it as an adjacency matrix itself, i.e. a multigraph where the B groups are the nodes. We then proceed by generating it from another SBM (which has its own partition into groups and matrix of edge counts). Continuing in the same manner yields a hierarchy of nested SBMs, where each level $l = 1, \ldots, L$ clusters the groups of the levels below. This yields a probability (see Ref. [45]) given by

$$P(\boldsymbol{e}|E) = \prod_{l=1}^{L} P(\boldsymbol{e}_l|\boldsymbol{e}_{l+1},\boldsymbol{b}_l)P(\boldsymbol{b}_l) \tag{15}$$

with

$$P(\boldsymbol{e}_l|\boldsymbol{e}_{l+1},\boldsymbol{b}_l) = \prod_{r<s} \left(\!\!\binom{n_r^l n_s^l}{e_{rs}^{l+1}}\!\!\right)^{-1} \prod_r \left(\!\!\binom{n_r^l(n_r^l+1)/2}{e_{rr}^{l+1}/2}\!\!\right)^{-1} \tag{16}$$

$$P(\boldsymbol{b}_l) = \frac{\prod_r n_r^l!}{B_{l-1}!} \binom{B_{l-1}-1}{B_l-1}^{-1} \frac{1}{B_{l-1}}, \tag{17}$$

where the index l refers to the variable of the SBM at a particular level, e.g., n_r^l is the number of nodes in group r at level l.

The use of this hierarchical prior is a strong departure from the noninformative assumption considered previously (although it contains it as a special case when the depth of the hierarchy is $L = 1$). It means that we expect some form of heterogeneity in the data at multiple scales, where groups of nodes are themselves grouped in larger groups forming a hierarchy. Crucially, this removes the "unimodality" inherent in the LDA assumption, as the group mixtures are now modelled by another generative level which admits as much heterogeneity as the original one. Furthermore, it can be shown to significantly alleviate the resolution limit of the noninformative approach, since it enables the detection of at most $O(N/\log N)$ groups in a sparse network with N nodes [43, 45].

Given the above model we can find the best overlapping partitions of the nodes by maximizing the posterior distribution

$$P(\{\boldsymbol{b}_l\}|\boldsymbol{A}) = \frac{P(\boldsymbol{A},\{\boldsymbol{b}_l\})}{P(\boldsymbol{A})}, \tag{18}$$

with

$$P(\boldsymbol{A},\{\boldsymbol{b}_l\}) = P(\boldsymbol{\mathcal{A}}|\boldsymbol{k},\boldsymbol{e}_1,\boldsymbol{b}_0)P(\boldsymbol{k}|\boldsymbol{e}_1,\boldsymbol{b}_0)P(\boldsymbol{b}_0)\times \prod_l P(\boldsymbol{e}_l|\boldsymbol{e}_{l+1},\boldsymbol{b}_l)P(\boldsymbol{b}_l). \tag{19}$$

which can be efficiently inferred using MCMC, as described in Refs. [44, 45]. The nonparametric nature of the model makes it possible to infer the number of groups (both for documents and words) directly from the posterior distribution, without the need for extrinsic methods or supervised approaches to prevent overfitting.[49]

1. *Bipartiteness*

The model above generates arbitrary multigraphs, whereas text is represented as a bipartite network of words and documents. Since the latter is a special case of the former, where words and documents belong to distinct groups, the model can be used as it is, as it will "learn" the bipartite structure during inference. However, a more consistent approach for text is to include this information in the prior, since we should not have to infer what we already know. This can be done via a simple modification of the model, where one replaces the prior for the overlapping partition appearing in Eq. (13) by

$$P(\boldsymbol{b}) = P_{\mathrm{w}}(\boldsymbol{b}^{\mathrm{w}})P_{\mathrm{d}}(\boldsymbol{b}^{\mathrm{d}}), \tag{20}$$

where $P_{\mathrm{w}}(\boldsymbol{b}^{\mathrm{w}})$ and $P_{\mathrm{d}}(\boldsymbol{b}^{\mathrm{d}})$ now correspond to a disjoint overlapping partition of the words and documents, respectively. Likewise, the same must be done at the upper levels of the hierarchy, by replacing (17) with

$$P(\boldsymbol{b}_l) = P_{\mathrm{w}}(\boldsymbol{b}_l^{\mathrm{w}})P_{\mathrm{d}}(\boldsymbol{b}_l^{\mathrm{d}}). \tag{21}$$

In this way, by construction, words and documents will never be placed together in the same group.

III. COMPARING LDA AND HSBM

In this section we demonstrate how well our approach is able to infer the topical structure when compared to LDA. We compare both models based on the description length Σ, where smaller values indicate a better model [50]. We obtain Σ for LDA from Eq. (2) and Σ for hSBM from Eq. (19) as[51]

$$\begin{aligned} \Sigma_{\mathrm{LDA}} &= -\ln P(\boldsymbol{n}|\boldsymbol{\eta},\boldsymbol{\beta},\boldsymbol{\alpha})P(\boldsymbol{\eta}) \\ \Sigma_{\mathrm{hSBM}} &= -\ln P(\boldsymbol{A},\{\boldsymbol{b}_l\}). \end{aligned} \tag{22}$$

The motivation for this approach is two-fold. One the one hand it offers a well-founded approach to unsupervised model selection within the framework of information theory, as it corresponds to the amount of information necessary to describe simultaneously i) the data when the model parameters are known, and ii) the parameters themselves. As the complexity of the model increases, the former will typically decrease, as it fits more closely the data, while at the same time it is compensated by an increase of the latter term, which serves as a penalty that prevents overfitting. On the other hand, in contrast to supervised model selection approaches such as perplexity, it allows for a straightforward comparison without the introduction of further approximations, e.g. in the evaluation of the held-out likelihood [28]. In addition, given data and two models (M_1 and M_2, where we incorporate latent variables into the model description) with description length (Σ_{M_1} and Σ_{M_2}), we can relate the difference $\Delta\Sigma \equiv \Sigma_{M_1} - \Sigma_{M_2}$ to the Bayes' Factor (BF) [52], i.e. how much more likely one model is compared to the other given the data

$$\mathrm{BF} \equiv \frac{P(M_1 \mid \mathrm{data})}{P(M_2 \mid \mathrm{data})} = \frac{P(\mathrm{data} \mid M_1)P(M_1)}{P(\mathrm{data} \mid M_2)P(M_2)} = e^{-\Delta\Sigma}, \tag{23}$$

where we assume that each model is a priori equally likely, i.e. $P(M_1) = P(M_2)$.

A. Artificial corpora sampled from LDA

We consider artificial corpora constructed from the generative process of LDA, incorporating some aspects of real texts, (for details see Materials & Methods, Artificial corpora drawn from LDA). Although LDA is not a good model for real corpora — as the Dirichlet assumption is not realistic — it serves to illustrate that even in a situation that clearly favors LDA, the hSBM frequently provides a better description of the data.

From the generative process we know the true latent variable of each word token. Therefore, we are able to obtain the inferred topical structure from each method by simply assigning the true labels without using approximate numerical optimization methods for the inference. This allows us to separate intrinsic (conceptual) properties of each method from external properties related to the numerical implementation.

In order to allow for a fair comparison between hSBM and LDA, we consider two different choices in the inference of each method, respectively. LDA requires the specification of a set of hyperparameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ used in the inference. While in this particular case we know the *true* hyperparameters that generated the corpus, in general these are unknown. Therefore, in addition to the true values, we also consider a noninformative choice, i.e. $\alpha_{dr} = 1$ and $\beta_{rd} = 1$. For the inference with hSBM, we only use the special case where the hierarchy has a single level (amounting to a noninformative prior). We consider two different parametrizations of the SBM: 1. Each document is assigned to its own group, i.e. they are not clustered and 2. different documents can belong to the same group, i.e. they are clustered. While the former is motivated by the original correspondence between pLSI and SBM, the latter shows the additional advantage offered by the possibility of clustering documents due to its symmetric treatment of words and documents in a bipartite network (see SM, Sec. II, for the explicit expressions of the Σ in each case).

Our main finding, shown in Fig. 3, is that hSBM is consistently better than LDA (shows a shorter description length) for corpora of almost any text length $k_d = m$ ranging over 4 orders of magnitude. These results hold for asymptotically large (in terms of the number of documents) corpora as shown in the inset of Fig. 3, where we observe that the normalized description length of each model converges to a fixed value when increasing the size of the corpus, i.e. increasing the number of documents. These results hold across a wide range of parameter setting (number of topics and values of the hyperparameters, see SM Sec. IV).

The LDA description length Σ_{LDA} does not depend strongly on the considered prior (true or noninformative), as the size of the corpora increases. This is consistent with the typical expectation that in the limit of large data (i.e. large number of documents), the prior "washes out". Note, however, that for smaller corpora (see Inset) the Σ of the non-informative prior is significantly worse than the Σ of the true prior.

In contrast, the hSBM provides much shorter description lengths than LDA for the same data when allowing documents to be clustered as well. The only exception is for very small texts ($m < 10$ tokens) – where we have not converged to the asymptotic limit in the per-word description length. In the limit $D \to \infty$ we expect hSBM to provide a similarly good or better model than LDA for all text lengths. The improvement of the hSBM over LDA in a LDA-generated corpus is counterintuitive because, for sufficient data, we expect the true model to provide a better description for it. However, for a model like LDA the limit of "sufficient data" involves the simul-

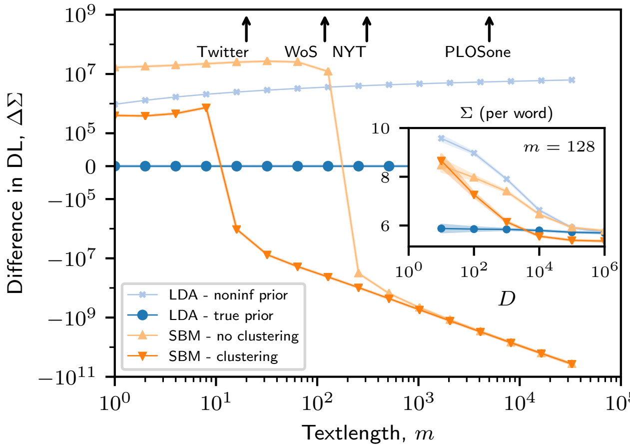

FIG. 3. Comparison between LDA and SBM for artificial corpora drawn from LDA. The description length (DL, Σ) of LDA and hSBM for an artificial corpus of $D = 10^6$ documents with length m drawn from the generative process of LDA with hyperparameters $\alpha_{dr} = \beta_{wr} = 1$. The 4 curves correspond to different choices in the parametrization of the topic models: i) LDA with non-informative priors (light blue - $\times$), ii) LDA with true priors, i.e. the hyperparameters used to generate the artificial corpus (dark blue - $\bullet$), iii) hSBM with without clustering of documents (light orange - $\blacktriangle$), and iv) hSBM with clustering of documents (dark orange - $\blacktriangledown$). We show the difference in description length compared to the first of the models, i.e. $\Delta\Sigma = \Sigma_i - \Sigma_{\text{LDA - true prior}}$. For comparison, the upper axis indicates the typical sizes (measured in number of word tokens) of documents in different platforms: Twitter, Web Of Science (WoS), New York Times (NYT), and the scientific journal PLOSone. Inset: Normalized Σ (per word token) as a function of the number of documents D for fixed textlength $m = 128$.

taneous scaling of the number of documents, words, *and* topics to very high values. In particular, the generative process of LDA requires a large number of documents (topics) to resolve the underlying Dirichlet distribution of the topic-document distribution (word-topic distribution). While the former is realized growing the corpus by adding documents, the latter aspect is non-trivial because the observed size of the vocabulary V is not a free parameter but is determined by the word-frequency distribution and the size of the corpus (in word tokens and, thus, the number of documents) through the so-called Heaps' law [19]. This means that as we grow the corpus by adding more and more documents, initially the vocabulary increases linearly and only at very large corpora it settles into an asymptotic sublinear growth (see SM, Sec. V). This, in turn, requires an ever larger number of topics to resolve the underlying word-topic distribution. Such large number of topics is not feasible in practice because it renders the whole goal and concept of topic models obsolete (which is to compress the information by obtaining an effective, coarse-grained, description of the corpus at a manageable number of topics).

In summary, the limits in which LDA provides a better description are irrelevant in practice (either extremely small texts or very large number of topics) and the observed limitations of LDA are due to the following reasons: i) the finite number of topics used to generate the data always leads to an under-sampling of the Dirichlet distributions, and ii) LDA is redundant in the way it describes the data in this sparse regime. In contrast, the assumptions of the hSBM are better suited for this sparse regime, and hence leads to a more compact description of the data, despite the fact the corpora were in fact generated by LDA.

B. Real corpora

We compare LDA and SBM for a variety of different datasets, as shown in Table I (see Materials & Methods for details on the numerical implementations). When using LDA, we consider both noninformative priors and fitted hyperparameters, for a wide range of numbers of topics. We obtain systematically smaller values for the description length using the hSBM — which can be an order of magnitude smaller for larger corpora. For real corpora, the difference is exacerbated by the fact the hSBM is capable of clustering documents, capitalizing on a source of structure in the data which is completely unavailable to LDA.

As our examples also show, LDA cannot be used in a direct manner to choose the number of topics, as the noninformative choice systematically underfits (Σ_{LDA} increases monotonically with the number of topics), and the parametric approach systematically overfits (Σ_{LDA} decreases monotonically with the number of topics). In practice, users are required to resort to heuristics [53, 54], or more complicated inference approaches based on the computation of the model evidence, which are not only numerically expensive, but can only be done under onerous approximations [8, 28]. In contrast, the hSBM is capable of extracting the appropriate number of topics directly from its posterior distribution, while simultaneously avoiding both under- and overfitting [43, 45].

In addition to these formal aspects, we argue that the hierarchical nature of the hSBM, and the fact that it clusters words *as well* as documents, makes it more useful in interpreting text. We illustrate this with a case study in the next section.

C. Case study: Wikipedia articles

We illustrate the results of the inference with the hSBM for articles taken from the English Wikipedia in Fig. 4, showing the hierarchical clustering of documents and words. To make the visualization clearer, we focus on a small network created from only three scientific disciplines: Chemical Physics (21 articles), Experimental Physics (24 articles), and Computational Biology (18 ar-

Corpus				Σ_{LDA}				Σ_{LDA} (hyperfit)				Σ_{hSBM}	hSBM groups	
	Docs.	Words	Word Tokens	10	50	100	500	10	50	100	500		Doc.	Words
Twitter	10,000	12,258	196,625	1,424,984	1,812,497	2,061,596	2,620,555	1,140,357	1,110,186	1,091,998	1,056,321	963,260	365	359
Reuters	1,000	8,692	117,661	934,897	1,041,081	1,118,719	1,407,650	879,684	876,656	881,107	879,321	341,199	54	55
Web of Science	1,000	11,198	126,313	1,034,132	1,169,207	1,263,202	1,560,443	1,035,555	1,057,491	1,065,584	1,075,433	426,529	16	18
New York Times	1,000	32,415	335,749	2,804,854	3,026,505	3,207,833	3,926,430	2,701,001	2,699,711	2,695,955	2,693,749	1,448,631	124	125
PlosONE	1,000	68,188	5,172,908	50,080,627	50,201,016	50,520,057	52,546,657	49,782,605	49,497,904	49,326,867	48,741,824	8,475,866	897	972

TABLE I. Hierarchical SBM outperforms LDA in real corpora. Each row corresponds to a different dataset (for details, see Material & Methods, Datasets for real corpora). We provide basic statistics of each dataset in columns "Corpus". The models are compared based on their description length Σ, see Eq. 22. Results for LDA with noninformative and fitted hyperparameters are shown in columns "Σ_{LDA}" and "Σ_{LDA} (hyperfit)" for different number of topics $K \in \{10, 50, 100, 500\}$. Result for the hSBM are shown in column "Σ_{hSBM}" and the inferred number of groups (documents and words) in "hSBM groups".

ticles). For clarity, we only consider words that appear more than once, such that we end up with a network of 63 document-nodes, $3,140$ word-nodes, and $39,704$ edges.

The hSBM splits the network into groups on different levels, organized as a hierarchical tree. Note that the number of groups and the number of levels were not specified beforehand but automatically detected in the inference. On the highest level, hSBM reflects the bipartite structure into word- and document-nodes, as is imposed in our model. While we considered articles from three different categories (one category from biology and two categories from physics), the second level in the hierarchy separates documents into only two groups corresponding to articles about biology (e.g. bioinformatics or K-mer) and articles on physics (e.g. Rotating wave approximation or Molecular beam). For lower levels, articles become separated into a larger number of groups, e.g. one group contains two articles on Euler's and Newton's law of motion, respectively. For words, the second level in the hierarchy splits nodes into three separate groups. We find that two groups represent words belonging to physics (e.g. beam, formula, or energy) and biology (assembly, folding, or protein) while the third group represents function words (the, of, or a). The latter group can be clearly identified by measuring the dissemination coefficient, U_D (an extension of the popular tf-idf taking into account the unequal lengths of documents [55]), which quantifies how uneven words are distributed among documents in comparison to a random null model, where $U_D < 1$ ($U_D > 1$) indicate words that are less (more) evenly disseminated than a Poisson process. From this analysis (right side of Fig. 4) we observe that words from the bottom group have a systematically higher coefficient of dissemination (with median$(U_D) \lesssim 1$) than words from the other groups. While typically function words are filtered manually, e.g. by using a list of stop words [9], here, words that are evenly distributed among all documents end up separated in its own branch in the hierarchy of word-groups, rendering such manual interventions unnecessary.

IV. DISCUSSION

The underlying equivalence between pLSI and the overlapping version of the SBM means that the "bag of words" formulation of topical corpora is mathematically equivalent to bipartite networks of words and documents with modular structures. From this we were able to formulate a topic model based on a hierarchical version of the SBM (hSBM) in a fully Bayesian framework alleviating some of the most serious conceptual deficiencies in current approaches to topic modeling such as LDA. In particular, the model formulation is nonparametric, and model complexity aspects such as the number of topics can be inferred directly from the model's posterior distribution. Furthermore, the model is based on a hierarchical clustering of both words and documents — in contrast to LDA which is based on a nonhierarchical clustering of the words alone. This enables the identification of structural patterns in text that is unavailable to LDA, while at the same time allowing for the identification of patterns in multiple scales of resolution.

We have shown that hSBM constitutes a better topic model compared to LDA not only for a diverse set of real corpora but even for artificial corpora generated from LDA itself. It is capable of providing better compression (description length) — as a measure of the quality of fit — as well as a richer interpretation of the data. More importantly, however, the hSBM offers an alternative to the existing paradigm of Dirichlet-priors in virtually any variation of current approaches to topic modeling. It can hardly be overstated that Dirichlet priors are simply motivated by its computational convenience and do not reflect prior knowledge on the actual usage of language. In contrast, our work shows how to formulate and incorporate different (and as we have shown more suitable) priors in a fully Bayesian framework. Furthermore, it also serves as a working example that efficient numerical implementations of non-Dirichlet topic models are feasible and can be applied in practice to large collections of documents. Thus, we believe that our results elucidate the limitations not just of LDA but of topic modeling focusing exclusively on Dirichlet priors in general.

More generally, our results show how the same mathematical ideas can be used to two extremely popular and mostly disconnected problems: the inference of topics in

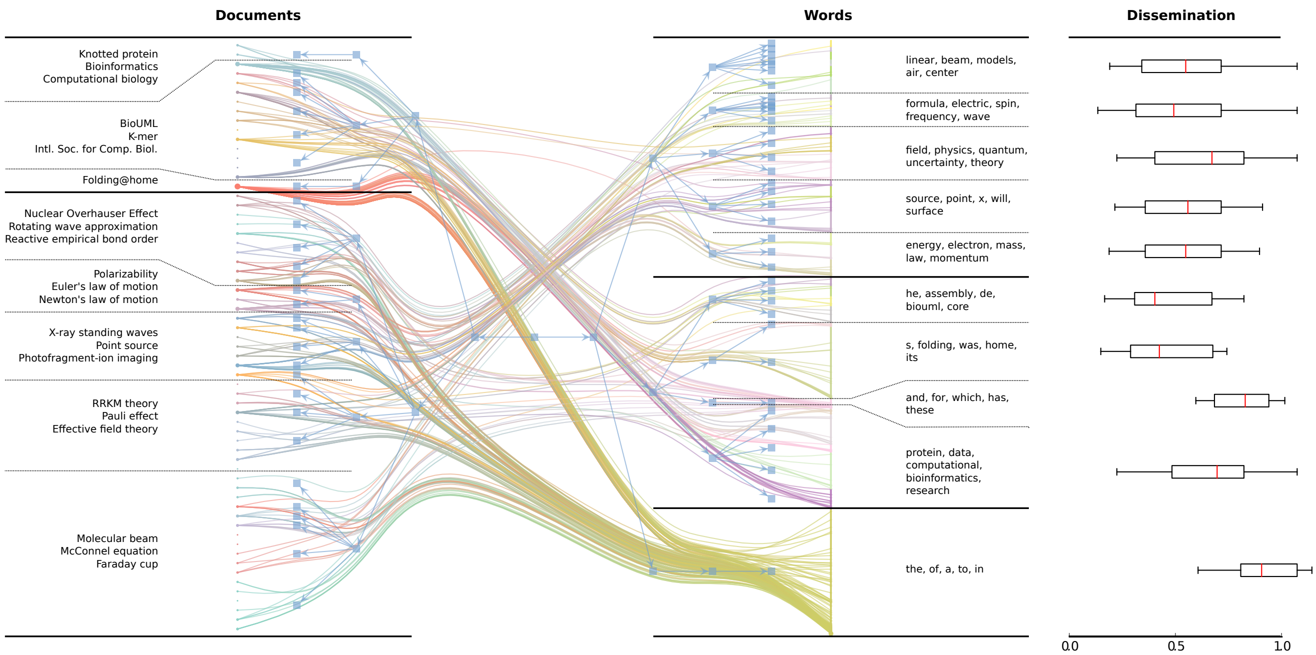

FIG. 4. Inference of hSBM to articles from the Wikipedia. Articles from 3 categories (Chemical Physics, Experimental Physics, and Computational Biology). First hierarchical level separates into documents (left) and words (right). For the docs: 3 (or less) randomly selected articles from groups on the third level of the hierarchy (sparated by dotted lines). For words: 5 most frequent words from groups on the third level of the hierarchy (separated by dotted lines) Groups on third level are grouped in 2nd level of hierarchy (solid lined). We calculate the dissemination coefficient [55] (U_D indicates how uneven words are distributed among documents with respect to a random null model). The boxplots show the $5, 25, 50, 75, 95$-percentile.

corpora and of communities in networks. We used this connection to obtain improved topic models, but there are many additional theoretical results in community detection that should be explored in the topic-model context, e.g., fundamental limits to inference such as the undetectable-detectable phase transition [56] or the analogy to Potts-like spin systems in statistical physics [57]. Another aspect of the connection is that the traditional application of topic models (the analysis of text) leads to classes of networks usually not considered by community-detection algorithms. In particular, the network is bipartite (words-documents), the topics/communities can be overlapping, and the number of links (word tokens) and nodes (word types) are connected to each other through Heaps' law. The latter constitutes only one of numerous statistical laws in language [19] (Zipf's law [20] probably being the most famous even beyond the realm of language). While these regularities are well-studied empirically it remains unclear how to best include them as prior knowledge or how they affect the inference of communities in word-document networks.

V. MATERIALS & METHODS

A. Artificial corpora

For the construction of the artificial corpora, we fix the parameters in the generative process of LDA, i.e. the number of topics K, the hyperparameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, and the length of individual articles m. The $\boldsymbol{\alpha}$ ($\boldsymbol{\beta}$) - hyperparameters determine the distribution of topics (words) in each document (topic).

The generative process of LDA can be decribed in the following way. For each topic $r \in \{1, \ldots, K\}$ we sample a distribution over words $\boldsymbol{\phi}_r$ from a V-dimensional Dirichlet distribution with parameters β_{rw} for $w \in \{1, \ldots, V\}$. For each document $d \in \{1, \ldots, D\}$ we sample a topic mixture $\boldsymbol{\theta}_d$ from a K-dimensional Dirichlet distribution with parameters α_{dr} for $r \in \{1, \ldots, K\}$. For each word-position $l_d \in \{1, \ldots, k_d\}$ (k_d is the length of document d) we first sample a topic $r^* = r_{l_d}$ from a multinomial with parameters $\boldsymbol{\theta}_d$ and then sample a word w from a multinomial with parameters $\boldsymbol{\phi}_{r^*}$.

We assume a parametrization in which i) each document has the same topic-document hyperparameter, i.e. $\alpha_{dr} = \alpha_r$ for $d \in \{1, \ldots, D\}$, and ii) each topic has the same word-topic hyperparameter, i.e. $\beta_{rw} = \beta_w$ for $r \in \{1, \ldots, K\}$. We fix the average probability of occurrence of a topic, p_r, (word, p_w) by a introducing scalar hyperparameters α (β), i.e. $\alpha_{dr} = \alpha K(p_r)$ for

$r \in \{1, \ldots, K\}$ ($\beta_{rw} = \beta V(p_w)$ for $w = 1, \ldots, V$). In our case we choose i) equiprobable topics, i.e. $p_r = 1/K$ and ii) empirically measured word-frequencies from the Wikipedia corpus, i.e. $p_w = p_w^{\mathrm{emp}}$ with $w = 1, \ldots, 95129$, yielding a Zipfian distribution, shown to be universally described by a double power law [48]) (see SM, Sec. III, for details on the distribution of word frequencies).

B. Datasets for real corpora

For the comparison of hSBM and LDA we consider different datasets of written texts varying in genre, time of origin, average text length, number of documents, and language; as well as datasets used in previous works on topic models, e.g. [7, 21, 58, 59]:

1. "Twitter", a sample of Twitter messages [60];
2. "Reuters", a collection of documents from the Reuters financial newswire service denoted as "Reuters-21578, Distribution 1.0" [60];
3. "Web of Science", abstracts from physics papers published in the year 2000
4. "New York Times (NYT)", a collection of newspaper articles [61];
5. "PlosOne", full text of all scientific articles published in 2011 in the journal PLoS One [62].

In all cases we considered a random subset of the documents, as detailed in Table I. For the NYT data we did not employ any additional filtering since the data was already provided in the form of pre-filtered word counts. For the other datasets we employed the following filtering: i) we decapitalized all words, ii) we replaced punctuation and special characters (e.g. ".", ",", or "/") by blank spaces such that we can define a word as any substring between two blank spaces, and iii) keep only those words which consisted of the letters a-z.

C. Numerical Implementations

For inference with LDA we used package mallet [63]. The algorithm for inference with the hSBM presented in this work is implemented in C++ as part of the `graph-tool` Python library [64]. We provide code on how to use hSBM for topic modeling in [65].

ACKNOWLEDGMENTS

We thank M. Palzenberger for help with the "Web of Science"-data. EGA thanks L. Azizi for helpful discussions.

[1] David M. Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77, 2012.

[2] Michael I. Jordan and Tom M. Mitchell. Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260, 2015.

[3] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990.

[4] Zoubin Ghahramani. Probabilistic machine learning and artificial intelligence. *Nature*, 521(7553):452–459, 2015.

[5] Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. *Bayesian Data Analysis.* CRC Press, 2003.

[6] Thomas 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, 1999.

[7] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[8] Thomas L. Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, 101 Suppl:5228–35, 2004.

[9] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schutze. *Introduction to Information Retrieval.* Cambridge University Press, 2008.

[10] Gideon S. Mann, David Mimno, and Andrew McCallum. Bibliometric impact measures leveraging topic analysis. In *Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries - JCDL '06.* ACM Press, 2006.

[11] Kevin W Boyack, David Newman, Russell J Duhon, Richard Klavans, Michael Patek, Joseph R. Biberstine, Bob Schijvenaars, André Skupin, Nianli Ma, and Katy Börner. Clustering more than two million biomedical publications: comparing the accuracies of nine text-based similarity approaches. *PloS one*, 6(3):e18029, 2011.

[12] Leah G. Nichols. A topic model approach to measuring interdisciplinarity at the National Science Foundation. *Scientometrics*, 100(3):741–754, 2014.

[13] Danielle S. McNamara. Computational Methods to Extract Meaning From Text and Advance Theories of Human Cognition. *Topics in Cognitive Science*, 3(1):3–17, 2011.

[14] Justin Grimmer and Brandon M. Stewart. Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, 21(3):267–297, 2013.

[15] B. Liu, Lin Liu, Anna Tsykin, Gregory J. Goodall, Jeffrey E. Green, Min Zhu, Chang Hee Kim, and Jiuyong Li. Identifying functional miRNA-mRNA regulatory modules with correspondence latent dirichlet allocation. *Bioinformatics*, 26(24):3105–3111, 2010.

[16] Jonathan K. Pritchard, Matthew Stephens, and Peter Donnelly. Inference of Population Structure Using Multilocus Genotype Data. *Genetics*, 155(2):945–959, 2000.
[17] Fei-Fei Li and Pietro Perona. A Bayesian Hierarchical Model for Learning Natural Scene Categories. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*. IEEE, 2005.
[18] Katherine Ellis, Emanuele Coviello, Antoni B. Chan, and Gert Lanckriet. A Bag of Systems Representation for Music Auto-Tagging. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(12):2554–2569, 2013.
[19] Eduardo G. Altmann and Martin Gerlach. Statistical Laws in Linguistics. In *Creativity and Universality in Language*. Springer, 2016.
[20] George K. Zipf. *The Psycho-Biology of Language*. Routledge, London, 1936.
[21] Andrea Lancichinetti, M. Irmak Sirer, Jane X Wang, Daniel Acuna, Konrad Körding, and Luìs A. Nunes Amaral. A high-reproducibility and high-accuracy method for automated topic classification. *Physical Review X*, 5:011007, 2015.
[22] Thomas L. Griffiths, M. Steyvers, David M. Blei, and Joshua B. Tenenbaum. Integrating topics and syntax. In *Advances in Neural Information Processing Systems 17*, 2005.
[23] Wei Li and Andrew McCallum. Pachinko allocation. In *Proceedings of the 23rd international conference on Machine learning - ICML '06*. ACM Press, 2006.
[24] Michel Rosen-Zvi, Thomas L. Griffiths, Mark Steyvers, and Padraic Smyth. The author-topic model for authors and documents. In *UAI '04 Proceedings of the 20th conference on Uncertainty in artificial intelligence*, 2004.
[25] Gabriel Doyle and Charles Elkan. Accounting for burstiness in topic models. In *Proceedings of the 26th Annual International Conference on Machine Learning - ICML '09*. ACM Press, 2009.
[26] Jian Tang, Ming Zhang, and Qiaozhu Mei. "Look Ma, No Hands!" A Parameter-Free Topic Model. *arXiv:1409.2993*, 2014.
[27] Weizhong Zhao, James J. Chen, Roger Perkins, Zhichao Liu, Weigong Ge, Yijun Ding, and Wen Zou. A heuristic approach to determine an appropriate number of topics in topic modeling. *BMC Bioinformatics*, 16(13):S8, 2015.
[28] Hanna M. Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. Evaluation methods for topic models. In *Proceedings of the 26th Annual International Conference on Machine Learning - ICML '09*. ACM Press, 2009.
[29] Matthew A. Taddy. On Estimation and Selection for Topic Models. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics, PMLR*, pages 1184–1193, 2012.
[30] Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.
[31] David M. Blei, Thomas L. Griffiths, and Michael I. Jordan. The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. *Journal of the ACM*, 57(2):1–30, 2010.
[32] M. E. J. Newman. Communities, modules and large-scale structure in networks. *Nature Physics*, 8(1):25–31, 2011.
[33] M.E.J. Newman. *Networks: An Introduction*. Oxford University Press, 2010.
[34] Santo Fortunato. Community detection in graphs. *Physics Reports*, 486(3-5):75–174, 2010.
[35] Mason A. Porter, Jukka-Pekka Onnela, and Peter J. Mucha. Communities in Networks. *Notices of the AMS*, 56(9):1164–1166, 2009.
[36] M. E. J. Newman and Michelle Girvan. Finding and evaluating community structure in networks. *Physical Review E*, 69(2):026113, 2004.
[37] Roger Guimerà, Marta Sales-Pardo, and Luís A Nunes Amaral. Modularity from fluctuations in random graphs and complex networks. *Physical Review E*, 70(2):025101, 2004.
[38] Andrea Lancichinetti and Santo Fortunato. Limits of modularity maximization in community detection. *Physical Review E*, 84(6):066122, 2011.
[39] Paul W. Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social Networks*, 5(2):109–137, 1983.
[40] Edoardo M. Airoldi, David M. Blei, Stephen E. Fienberg, and Eric P. Xing. Mixed membership stochastic blockmodels. *Journal of machine learning research : JMLR*, 9:46, 2007.
[41] Brian Karrer and M. E. J. Newman. Stochastic blockmodels and community structure in networks. *Physical Review E*, 83(1):016107, 2011.
[42] Brian Ball, Brian Karrer, and M. E. J. Newman. Efficient and principled method for detecting communities in networks. *Physical Review E*, 84(3):036103, 2011.
[43] Tiago P. Peixoto. Hierarchical Block Structures and High-Resolution Model Selection in Large Networks. *Physical Review X*, 4(1):011047, 2014.
[44] Tiago P. Peixoto. Model selection and hypothesis testing for large-scale network models with overlapping groups. *Physical Review X*, 5(1), 2015.
[45] Tiago P. Peixoto. Nonparametric Bayesian inference of the microcanonical stochastic block model. *Physical Review E*, 95(1):012317, January 2017.
[46] We denote matrices by bold-face symbols, e.g. $\boldsymbol{\theta} = \{\theta_{dr}\}$ with $d = 1, \dots, D$ and $r = 1, \dots, K$ where θ_{dr} is an individual entry. The notation $\boldsymbol{\theta}_d$ refers to the vector $\{\theta_{dr}\}$ with fixed d and $r = 1, \dots, K$.
[47] Tiago P. Peixoto. Parsimonious Module Inference in Large Networks. *Physical Review Letters*, 110(14):148701, April 2013.
[48] Martin Gerlach and Eduardo G Altmann. Stochastic Model for the Vocabulary Growth in Natural Languages. *Physical Review X*, 3(2):021006, 2013.
[49] The latter can be seen interpreting Eq. 19 as a description length, see discussion after Eq. 22.
[50] J. Rissanen. Modeling by shortest data description. *Automatica*, 14(5):465–471, September 1978.
[51] Σ_{LDA} is conditioned on the hyperparameters $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$ and therefore it is exact for noninformative priors ($\alpha_{dr} = 1$ and $\beta_{rd} = 1$) only. Otherwise, Eq. 22 is only a lower bound for Σ_{LDA} because it lacks the terms involving *hyperpriors* for $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$. For simplicity, we ignore this correction in our analysis and therefore we favor LDA.
[52] Robert E. Kass and Adrian E. Raftery. Bayes Factors. *Journal of the American Statistical Association*, 90:773–795, 1995.
[53] R. Arun, V. Suresh, C. E. Veni Madhavan, and M. N. Narasimha Murthy. On Finding the Natural Number of Topics with Latent Dirichlet Allocation: Some Observations. In *Advances in Knowledge Discovery and*

Data Mining, pages 391–402. Springer, Berlin, Heidelberg, June 2010.

[54] Juan Cao, Tian Xia, Jintao Li, Yongdong Zhang, and Sheng Tang. A density-based method for adaptive LDA model selection. *Neurocomputing*, 72(79):1775–1781, March 2009.

[55] Eduardo G. Altmann, Janet B. Pierrehumbert, and Adilson E. Motter. Niche as a determinant of word fate in online groups. *PloS one*, 6(5):e19009, 2011.

[56] Aurelien Decelle, Florent Krzakala, Cristopher Moore, and Lenka Zdeborová. Inference and Phase Transitions in the Detection of Modules in Sparse Networks. *Physical Review Letters*, 107(6):065701, 2011.

[57] Dandan Hu, Peter Ronhovde, and Zohar Nussinov. Phase transitions in random Potts systems and the community detection problem: spin-glass type and dynamic perspectives. *Philosophical Magazine*, 92(4):406–445, 2012.

[58] Hanna M. Wallach, David Mimno, and Andrew McCallum. Rethinking LDA : Why Priors Matter. In *Advances in Neural Information Processing Systems 22 (NIPS 2009)*, 2009.

[59] Arthur Asuncion, Max Welling, Padraic Smyth, and Yee Whye Teh. On Smoothing and Inference for Topic Models. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*, 2009.

[60] Steven Bird, Ewan Klein, and Edward Loper. Natural Language Processing with Python. `http://www.nltk.org/nltk_data/` [Online; accessed 17-July-2015], 2009.

[61] M. Lichman. UCI machine learning repository. `http://archive.ics.uci.edu/ml` [Online; accessed 17-July-2015], 2013.

[62] Plos API. All articles published in the journal plos one. `http://api.plos.org/` [Online; accessed 17-October-2013].

[63] Andrew K. McCallum. Mallet: A machine learning for language toolkit. `http://mallet.cs.umass.edu/`, 2002.

[64] Tiago P. Peixoto. The `graph-tool` python library. *figshare*, 2014. Available at `https://graph-tool.skewed.de`.

[65] Martin Gerlach. Using stochastic block models for topic modeling. `https://github.com/martingerlach/hSBM_Topicmodel`, 2017.

Supplementary Material for the manuscript: "A network approach to topic models"

I. MARGINAL LIKELIHOOD OF THE SBM

A. Non-informative priors

For the labeled network $\boldsymbol{\mathcal{A}}$ considered in the main text, section **Community detection: The hierarchical SBM**, Eq. (4), we have

$$P(\boldsymbol{\mathcal{A}}|\boldsymbol{\kappa},\boldsymbol{\omega}) = \prod_{i<j}\prod_{rs}\frac{e^{-\kappa_{ir}\omega_{rs}\kappa_{is}}(\kappa_{ir}\omega_{rs}\kappa_{js})^{\mathcal{A}_{ij}^{rs}}}{\mathcal{A}_{ij}^{rs}!}\times \prod_{i}\prod_{rs}\frac{e^{-\kappa_{ir}\omega_{rs}\kappa_{is}/2}(\kappa_{is}\omega_{rs}\kappa_{is}/2)^{\mathcal{A}_{ii}^{rs}/2}}{\mathcal{A}_{ii}^{rs}/2!}. \tag{S1}$$

If we now make a non-informative choice for the priors,

$$P(\boldsymbol{\kappa}) = \prod_r (n-1)!\delta(\textstyle\sum_i \kappa_{ir}-1), \tag{S2}$$

$$P(\boldsymbol{\omega}|\bar{\omega}) = \prod_{r\leq s}\frac{e^{-\omega_{rs}/\bar{\omega}}}{\bar{\omega}}, \tag{S3}$$

we can compute the integrated marginal likelihood as

$$\begin{aligned} P(\boldsymbol{\mathcal{A}}|\bar{\omega}) &= \int P(\boldsymbol{\mathcal{A}}|\boldsymbol{\kappa},\boldsymbol{\omega})P(\boldsymbol{\kappa})P(\boldsymbol{\omega}|\bar{\omega})\,\mathrm{d}\boldsymbol{\kappa}\mathrm{d}\boldsymbol{\omega}, \\ &= \frac{\bar{\omega}^E}{(\bar{\omega}+1)^{E+B(B+1)/2}}\frac{\prod_{r<s}e_{rs}!\prod_r e_{rr}!!}{\prod_{rs}\prod_{i<j}\mathcal{A}_{ij}^{rs}!\prod_i \mathcal{A}_{ii}^{rs}!!}\times \\ &\quad \prod_r\frac{(N-1)!}{(e_r+N-1)!}\prod_{ir}k_i^r!. \end{aligned} \tag{S4}$$

B. Equivalence with microcanonical model

As mentioned in the main text, Eq. 7 can be decomposed as

$$P(\boldsymbol{\mathcal{A}}|\bar{\omega}) = P(\boldsymbol{\mathcal{A}},\boldsymbol{k},\boldsymbol{e}|\bar{\omega}) = P(\boldsymbol{\mathcal{A}}|\boldsymbol{k},\boldsymbol{e})P(\boldsymbol{k}|\boldsymbol{e})P(\boldsymbol{e}|\bar{\omega}), \tag{S5}$$

with

$$P(\boldsymbol{\mathcal{A}}|\boldsymbol{k},\boldsymbol{e}) = \frac{\prod_{r<s}e_{rs}!\prod_r e_{rr}!!\prod_{ir}k_i^r!}{\prod_{rs}\prod_{i<j}\mathcal{A}_{ij}^{rs}!\prod_i\mathcal{A}_{ii}^{rs}!!\prod_r e_r!} \tag{S6}$$

$$P(\boldsymbol{k}|\boldsymbol{e}) = \prod_r \left(\!\!\binom{e_r}{N}\!\!\right)^{-1} \tag{S7}$$

$$P(\boldsymbol{e}|\bar{\omega}) = \prod_{r\leq s}\frac{\bar{\omega}^{e_{rs}}}{(\bar{\omega}+1)^{e_{rs}+1}} = \frac{\bar{\omega}^E}{(\bar{\omega}+1)^{E+B(B+1)/2}}. \tag{S8}$$

where $e_{rs} = \sum_{ij}\mathcal{A}_{ij}^{rs}$ is the total number of edges between groups r and s (we used the shorthand $e_r = \sum_s e_{rs}$ and $k_i^r = \sum_{js}\mathcal{A}_{ij}^{rs}$). $P(\boldsymbol{\mathcal{A}}|\boldsymbol{k},\boldsymbol{e})$ is the probability of a labelled graph $\boldsymbol{\mathcal{A}}$ where the labelled degrees $\boldsymbol{k}$ and edge counts between groups $\boldsymbol{e}$ are constrained to specific values. This can be seen by writing

$$P(\boldsymbol{\mathcal{A}}|\boldsymbol{k},\boldsymbol{e}) = \frac{\Xi}{\Omega}, \tag{S9}$$

with

$$\Omega = \frac{\prod_r e_r!}{\prod_{r<s} e_{rs}! \prod_r e_{rr}!!} \tag{S10}$$

being the number of configurations (i.e. half-edge pairings) that are compatible with the constraints, and

$$\Xi = \frac{\prod_{ir} k_i^r!}{\prod_{rs} \prod_{i<j} \mathcal{A}_{ij}^{rs}! \prod_i \mathcal{A}_{ii}^{rs}!!} \tag{S11}$$

is the number of configurations that correspond to the same labelled graph $\{A_{ij}^{rs}\}$. $P(\boldsymbol{k}|\boldsymbol{e})$ is the uniform prior distribution of the labelled degrees constrained by the edge counts $\boldsymbol{e}$, since $\left(\!\binom{e_r}{N}\!\right)$ is the number of ways to distribute e_r indistinguishable items into N distinguishable bins. Furthermore, $P(\boldsymbol{e}|\bar{\omega})$ is the prior distribution of edge counts, given by a mixture of independent geometric distributions with average $\bar{\omega}$.

C. Labelled degrees and overlapping partitions

As described in the main text, section **Community detection: The hierarchical SBM**, Eq. (13), the distribution of labeled degrees is given by

$$P(\boldsymbol{k}|\boldsymbol{e}) = P(\boldsymbol{k}|\boldsymbol{e}, \boldsymbol{b})P(\boldsymbol{b}), \tag{S12}$$

where the overlapping partition is distributed according to

$$P(\boldsymbol{b}) = \left[\prod_q P(\boldsymbol{b}_q|\boldsymbol{n}_{\boldsymbol{b}}^q)P(\boldsymbol{n}_{\boldsymbol{b}}^q|n_q)\right] P(\boldsymbol{q}|\boldsymbol{n})P(\boldsymbol{n}). \tag{S13}$$

Here, $\boldsymbol{b}$ corresponds to a specific set of groups, i.e. a mixture, of size $q = |\boldsymbol{b}|$. The distribution above means that we first sample the frequency of mixture sizes from the distribution

$$P(\boldsymbol{n}) = \left(\!\!\binom{Q}{N}\!\!\right)^{-1}, \tag{S14}$$

where Q is the maximum overlap size (typically $Q = B$, unless we want to force non-overlapping partitions with $Q = 1$). Given the frequencies, the mixture sizes are sampled uniformly on each node

$$P(\boldsymbol{q}|\boldsymbol{n}) = \frac{\prod_q n_q!}{N!}. \tag{S15}$$

We now consider the nodes with a given value of $q_i = q$ separately, and we put each one of them in a specific mixture $\boldsymbol{b}$ of size q. We do so by first sampling the frequencies in each mixture $\boldsymbol{n}_{\boldsymbol{b}}^q$ uniformly

$$P(\boldsymbol{n}_{\boldsymbol{b}}^q|n_q) = \left(\!\!\binom{\binom{B}{q}}{n_q}\!\!\right)^{-1}, \tag{S16}$$

and then we sample the mixtures themselves, conditioned on the frequencies,

$$P(\boldsymbol{b}_q|\boldsymbol{n}_{\boldsymbol{b}}^q) = \frac{\prod_{\boldsymbol{b}} n_{\boldsymbol{b}}^q!}{n_q!}. \tag{S17}$$

The labeled degree sequence is sampled conditioned on this overlapping partition and also on the frequency of degrees $\boldsymbol{n}_{\boldsymbol{k}}^{\boldsymbol{b}}$ inside each mixture $\boldsymbol{b}$,

$$P(\boldsymbol{k}|\boldsymbol{e}, \boldsymbol{b}) = \left[\prod_{\boldsymbol{b}} P(\boldsymbol{k}_{\boldsymbol{b}}|\boldsymbol{n}_{\boldsymbol{k}}^{\boldsymbol{b}})P(\boldsymbol{n}_{\boldsymbol{k}}^{\boldsymbol{b}}|\boldsymbol{e}_{\boldsymbol{b}}, \boldsymbol{b})\right] P(\boldsymbol{e}_{\boldsymbol{b}}|\boldsymbol{e}, \boldsymbol{b}). \tag{S18}$$

Here, $e_{\boldsymbol{b}}^r = \sum_i k_i^r \delta_{b_i^r,1}$ is the sum of the degrees with label r in mixture $\boldsymbol{b}$, which is sampled uniformly according to

$$P(\boldsymbol{e}_{\boldsymbol{b}}|\boldsymbol{e}, \boldsymbol{b}) = \prod_r \left(\!\!\binom{m_r}{e_r}\!\!\right)^{-1}, \tag{S19}$$

where $m_r = \sum_{\boldsymbol{b}} b_r[n_{\boldsymbol{b}} > 0]$ is the number of occupied mixtures that contain component r. Given the degree sums, the frequency of degrees is sampled according to

$$P(\boldsymbol{n}_{\boldsymbol{k}}^{\boldsymbol{b}}|\boldsymbol{e}_{\boldsymbol{b}}, \boldsymbol{b}) = \prod_{r\in\boldsymbol{b}} p(e_{\boldsymbol{b}}^r, n_{\boldsymbol{b}}^r)^{-1}, \tag{S20}$$

where $p(m, n)$ is the number of partitions of the integer m into exactly n parts, which can be pre-computed via the recurrence

$$p(m, n) = p(m - n, n) + p(m - 1, n - 1), \tag{S21}$$

with the boundary conditions $p(0, 0) = 1$ and $p(m, n) = 0$ if $n \leq 0$ or $m \leq 0$, or alternatively via the relation

$$p(m + n, n) = q(m, n) \tag{S22}$$

where $q(m, n)$ is the number of partitions of m into *at most* n parts, and using accurate asymptotic approximations for $q(m, n)$ (see Ref. [45]). Finally, having sampled the frequencies, we sample the labeled degree sequence uniformly in each mixture

$$P(\boldsymbol{k}_{\boldsymbol{b}}|\boldsymbol{n}_{\boldsymbol{k}}^{\boldsymbol{b}}) = \frac{\prod_{\boldsymbol{k}} n_{\boldsymbol{k}}^{\boldsymbol{b}}!}{n_{\boldsymbol{b}}!}. \tag{S23}$$

We refer to Ref. [44] for further details of the above distribution.

II. ARTIFICIAL CORPORA DRAWN FROM LDA

A. Drawing artificial documents from LDA

We specify α_{dr} and β_{rw}, i.e. the hyperparameters used to *generate* the artificial corpus (note that the hyperparameters used in the inference with LDA can be different) and fixing V, K, D, M and proceed in the following way:

- For each topic $r = 1, \ldots, K$:
  - Draw the word-topic distribution ϕ_w^r (frequencies of words conditioned on the topic r) from a V-dimensional Dirichlet:
    $\phi_w^r \sim Dir_V(\beta_{wr})$
- For each document $d = 1, \ldots, D$:
  - Draw the topic-document distribution θ_d^r (frequencies of topics conditioned on the doc d) from a K-dimensional Dirichlet:
    $\theta_d^r \sim Dir_K(\alpha_{dr})$
  - For each token $i_d = 1, \ldots, n_d$ (n_d is the length of each document) in document d:
    - Draw a topic r_{i_d} from the categorical θ_d^r
    - Draw a word-type w_{i_d} from the categorical $\phi_w^{r_{i_d}}$

B. Inference of corpora drawn from LDA

When we draw artificial corpora we obtain the labeled word-document counts n_{wd}^r, i.e. the "true" labels from the generative process of LDA as described in Sec. II A. In the following we describe how to obtain the description length of LDA and SBM when assigning the "true" labels as the result of the inference. In this way, we obtain the best possible inference results from each method. We can, therefore, compare the two models conceptually and avoid the issue of which particular numerical implementation was used.

1. Inference with LDA

In the inference with LDA we simply need the word-topic, $n^r_w = \sum_{d=1}^{D} n^r_{dw}$, the document-topic counts, $n^r_d = \sum_{w=1}^{V} n^r_{dw}$, and the word-document matrix $n_{dw} = \sum_{r=1}^{K} n^r_{dw}$ and use them to obtain the description length for LDA.

Note that for the inference we also have to specify the hyperparameters used in the inference, $\hat{\alpha}_{dr}$ and $\hat{\beta}_{rw}$. One approach is to consider the *true prior* (the same hyperparameter we used to generate the corpus) such that $\hat{\alpha}_{dr} = \alpha_{dr}$ and $\hat{\beta}_{rw} = \beta_{rw}$. In general, however, the data is not generated from LDA such that it is unclear which is the best choice of hyperparameters for inference. Therefore, we also consider the case of a *non-informative prior* in which $\hat{\alpha}_{dr} = 1$ and $\hat{\beta}_{rw} = 1$.

2. Inference with SBM

For the stochastic block model (SBM) we consider texts as a network in which the nodes consist of documents and words and the strength of the edge between them is given by the number of occurrences of the word in the document, yielding a bi-partite multigraph. We consider the case of a degree-corrected, overlapping SBM with only one layer in the hierarchy.

a. No clustering of documents For the SBM we use a particular parametrization starting from the equivalence between the degree-corrected SBM [42] and probabilistic semantic indexing (pLSI) [6], as described in the main text, section **Topic models: pLSI and LDA** Each document-node is put in its own group and the word-nodes are clustered into word-groups. The latter correspond to the topics in LDA (with possible mixtures among those groups) thus giving us a total of $B = D + K$ groups.

b. Clustering of documents Instead of putting each document in a separate group we cluster the documents into K groups as well such that we have $B = 2K$ groups in total. Note that this corresponds to a completely symmetric clustering of the groups in which we choose the indices such that $r = 1, \ldots, K$ are groups for the document-nodes and $r = K+1, \ldots, 2K$ are word-nodes. For a given word-token of word-type w appearing in document d labeled in topic $r = j$, we label the two half-edges as $r_d = j$ (the half-edge on the document-node) and $r_w = K + j$ (the half-edge on the word-node).

III. EMPIRICAL WORD-FREQUENCY DISTRIBUTION

In the comparison of hSBM and LDA for corpora drawn from the generative process of LDA, we parametrize the word-topic hyperparameter as $(\beta_{rw}) = (\beta_w) \equiv \boldsymbol{\beta}$ for $r = 1, \ldots, K$ with $\boldsymbol{\beta} = \beta V p_w$ for $w = 1, \ldots, V$. We use an empirical word-frequency distribution p_w as measured from all articles in the Wikipedia corpus contained in the categories "Scientific Disciplines". In Fig. S1 we show the empirically measured rank-frequency distribution for $V = 95129$ different words and $M = 5,118,442$ word tokens in total. We observe that this distribution is characterized by a heavy-tailed distribution with two power-laws. In Ref. [48] it has been shown that virtually any collection of documents follows such a distribution of word frequencies.

IV. VARYING THE HYPERPARAMETERS AND NUMBER OF TOPICS

In Fig. 3 of the main text we compare LDA and hSBM for corpora drawn from LDA for the case $K = 10$ and $\alpha = \beta = 1.0$. In Figs. (S2, S3) we vary the scalar hyperparameters α and β over four orders of magnitude and the number of topics over two orders of magnitude. While the individual curves for the description length of the different models look different, the qualitative behavior shown in Fig. 3 of the main text remains the same. This shows that hSBM provides a better description of the topical structure even for corpora generated from the generative process of LDA under very general conditions.

V. WORD-DOCUMENT NETWORKS ARE NOT SPARSE

Typically, in community detection it is assumed that networks are sparse, i.e. the number of edges E scales linearly with the number of nodes N, i.e. $E \propto N$ [34] In Fig. S4 we observe a different scaling for word-document networks, i.e. a superlinear scaling $E \propto N^{\delta}$ with $\delta > 1$. This is a direct result of the sublinear growth of the number of the number of different words with the total number of words in the presence of heavy-tailed word-frequency distributions

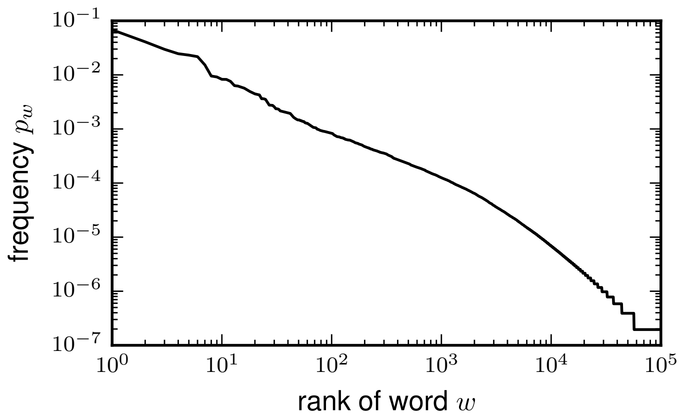

FIG. S1. Empirical rank-frequency distribution. The rank-frequency distribution shows the frequency of each word, $p_w = n_w/M$, ordered according to their rank, where n_w is the number of times word w occurs and $M = \sum_w n_w$ is the total number of words. A word is assigned rank r if it is the r-th most frequent word, i.e. the most frequent word has rank 1.

(known as Heaps' law in quantitative linguistics [19]), which leads to the superlinear growth of the number of edges with the number of nodes. This means that the density, i.e. the average number of edges per node, increases as more documents are added to the corpus.

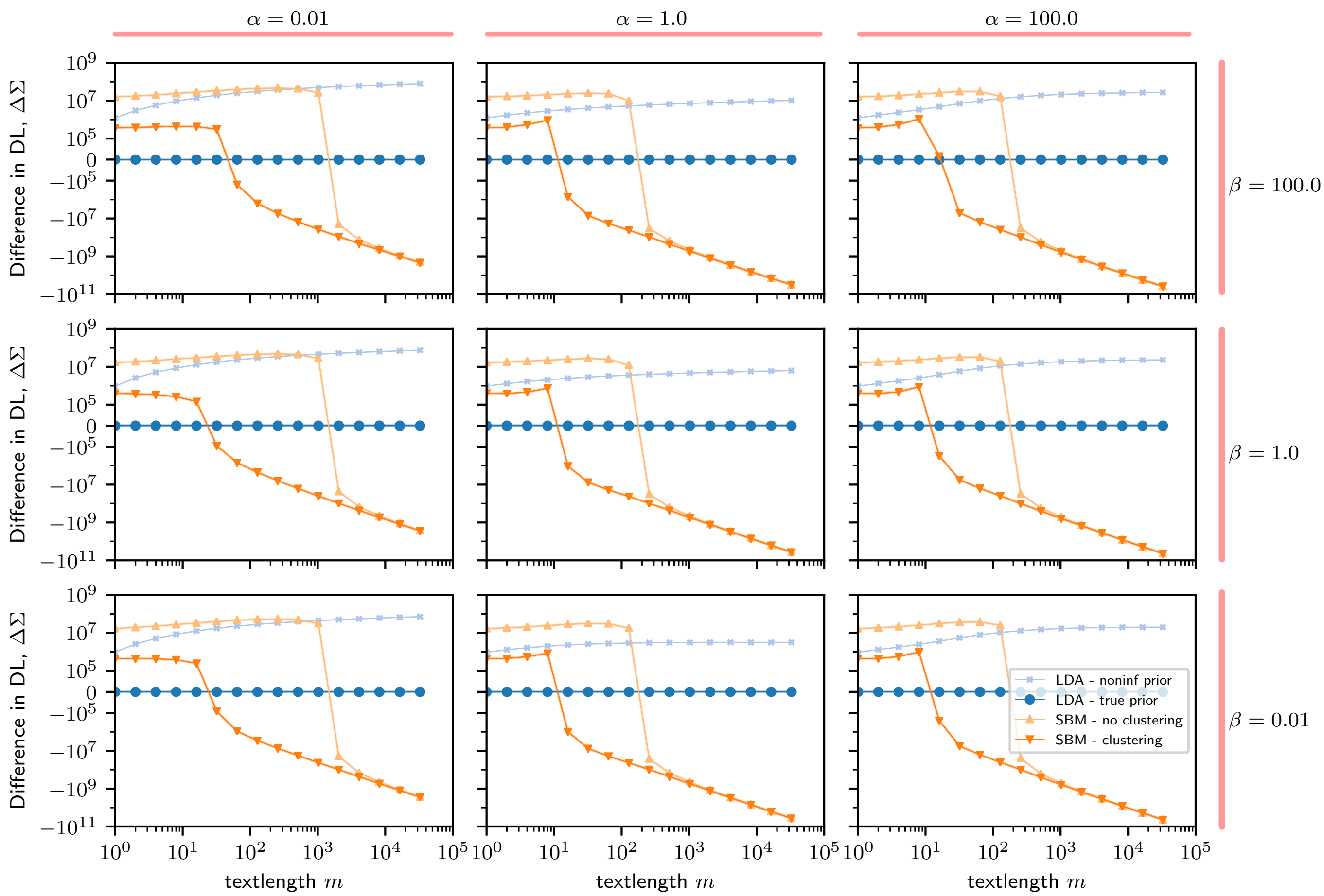

FIG. S2. Varying the hyperparameters α and β in the comparison between LDA and SBM for artificial corpora drawn from LDA. Description length per word for 4 different models: i) LDA with a true (non-informative) prior and ii) hSBM where we (do not) allow for clustering for $D = 10^6$ documents and $K = 10$ topics as a function of the text length m. Note that the panel in the middle corresponds to Fig. 3 in the main text.

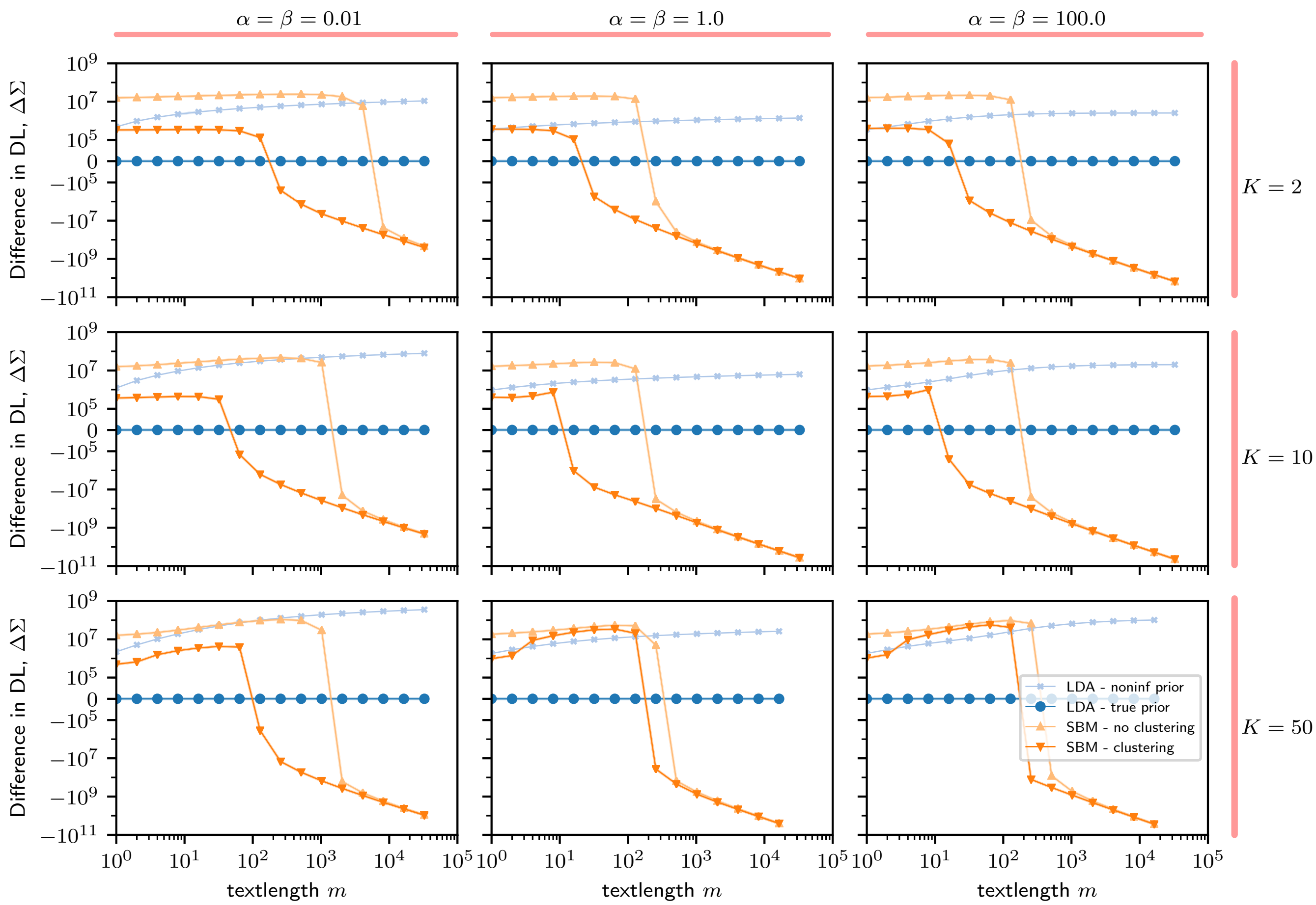

FIG. S3. Varying the number of topics K in the comparison between LDA and SBM for artificial corpora drawn from LDA. Description length per word for 4 different models: i) LDA with a true (non-informative) prior and ii) hSBM where we (do not) allow for clustering for $D = 10^6$ documents and choosing $\alpha = \beta$ for the scalar hyperparameters as a function of the text length m. Note that the panel in the middle corresponds to Fig. 3 in the main text.

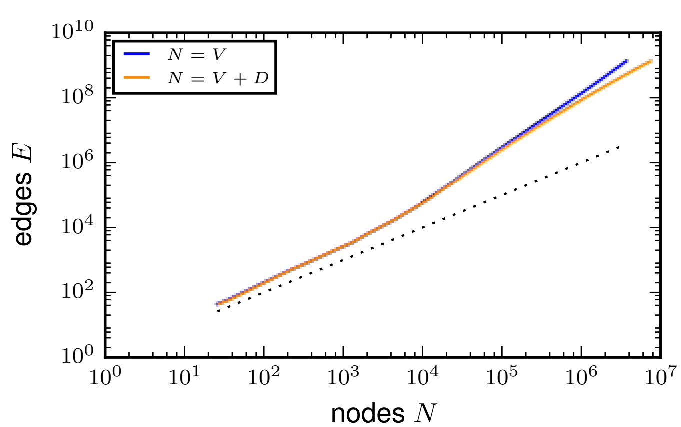

FIG. S4. Word-document networks are not sparse. The number of edges, E, as a function of the number of nodes, N, for the word-document network from the English Wikipedia. The network is grown by adding articles one after another in a randomly chosen order. Shown are the two cases, where i) only the V word-types are counted as nodes ($N = V$) and ii) both the word-types and the documents are counted as nodes ($N = V + D$). For comparison we show the linear relationship $E = N$ (dotted).