Science Advances
AAAS

advances.sciencemag.org/cgi/content/full/4/7/eaaq1360/DC1

Supplementary Materials for

A network approach to topic models: Finding topics through community detection in word-document networks

Martin Gerlach*, Tiago P. Peixoto, Eduardo G. Altmann

*Corresponding author. Email: martin.gerlach@northwestern.edu

Published 18 July 2018, *Sci. Adv.* **4**, eaaq1360 (2018)
DOI: 10.1126/sciadv.aaq1360

This PDF file includes:

Section S1. Marginal likelihood of the SBM

1.1 Noninformative priors

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

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

$$\prod_{i}\prod_{rs}\frac{e^{-\kappa_{ir}\omega_{rs}\kappa_{is}/2}(\kappa_{is}\omega_{rs}\kappa_{is}/2)^{\mathcal{A}_{ii}^{rs}/2}}{\mathcal{A}_{ii}^{rs}/2!} \tag{S1}$$

If we now make a noninformative 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

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

1.2 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 $\bar{\omega}$.

1.3 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 nonoverlapping 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. (*42*)). Finally, having sampled the frequencies, we sample the labeled degree sequence uniformly in each mixture

$$P(\boldsymbol{k_b}|\boldsymbol{n_k^b}) = \frac{\prod_{\boldsymbol{k}} n_{\boldsymbol{k}}^{\boldsymbol{b}}!}{n_{\boldsymbol{b}}!} \tag{S23}$$

We refer to Ref. (*41*) for further details of the above distribution.

Section S2. Artificial corpora drawn from LDA

2.1 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 $\phi_w^{r_{i_d}}$

2.2 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 above. 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.

2.2.1 Inference with LDA

In the inference with LDA we simply need the word-topic, $n_w^r = \sum_{d=1}^D n_{dw}^r$, the document-topic counts, $n_d^r = \sum_{w=1}^V n_{dw}^r$, and the word-document matrix $n_{dw} = \sum_{r=1}^K n_{dw}^r$ 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 *noninformative prior* in which $\hat{\alpha}_{dr} = 1$ and $\hat{\beta}_{rw} = 1$.

2.2.2 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 bipartite multigraph. We consider the case of a degree-corrected, overlapping SBM with only one layer in the hierarchy.

No clustering of documents For the SBM we use a particular parametrization starting from the equivalence between the degree-corrected SBM (*33*) and probabilistic semantic indexing (pLSI) (*4*), 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.

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).

Section S3. Varying the hyperparameters and number of topics

In Fig. 4 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. (S1, S2, S3) we show that these results hold under very general conditions by varying i) the values of the scalar hyperparameters; ii) the number of topics; and iii) the base measure of the vector-valued hyperparameters $\vec{\alpha}$ and $\vec{\beta}$ (symmetric or asymmetric following the approach in Ref. (*58*)). While the individual curves for the description length of the different models look different, the qualitative behavior shown in Fig. 4 of the

main text remains the same. In all cases, the hSBM performs better than the LDA with noninformative priors; and only in few cases the hSBM has a larger description length than LDA with the true hyperparameters which actually generated the data. Note that the latter case constitutes an exception because i) the generating hyperparameters are unknown in practice; and ii) as the hyperparameters deviate from the noninformative choice, the LDA description length computed ceases to be complete, becoming only a lower bound to the complete one which involves integration over the hyperparameters (as is thus intractable).

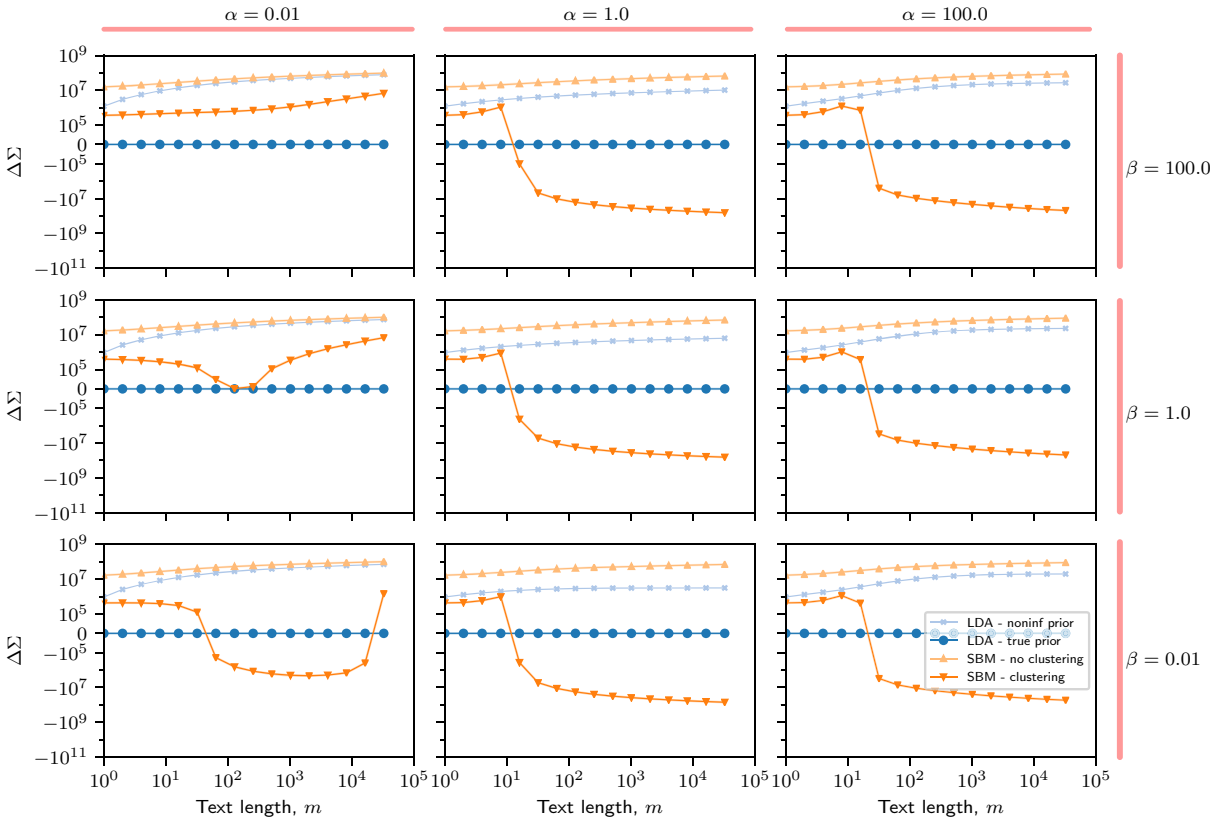

Fig. S1. Varying the hyperparameters α and β in the comparison between LDA and SBM for artificial corpora drawn from LDA. Same as in Fig. 4A (main text) with different values $\alpha \in \{0.01, 1.0, 100.0\}$ and $\beta \in \{0.01, 1.0, 100.0\}$. Note that the panel in the middle corresponds to Fig. 4 in the main text.

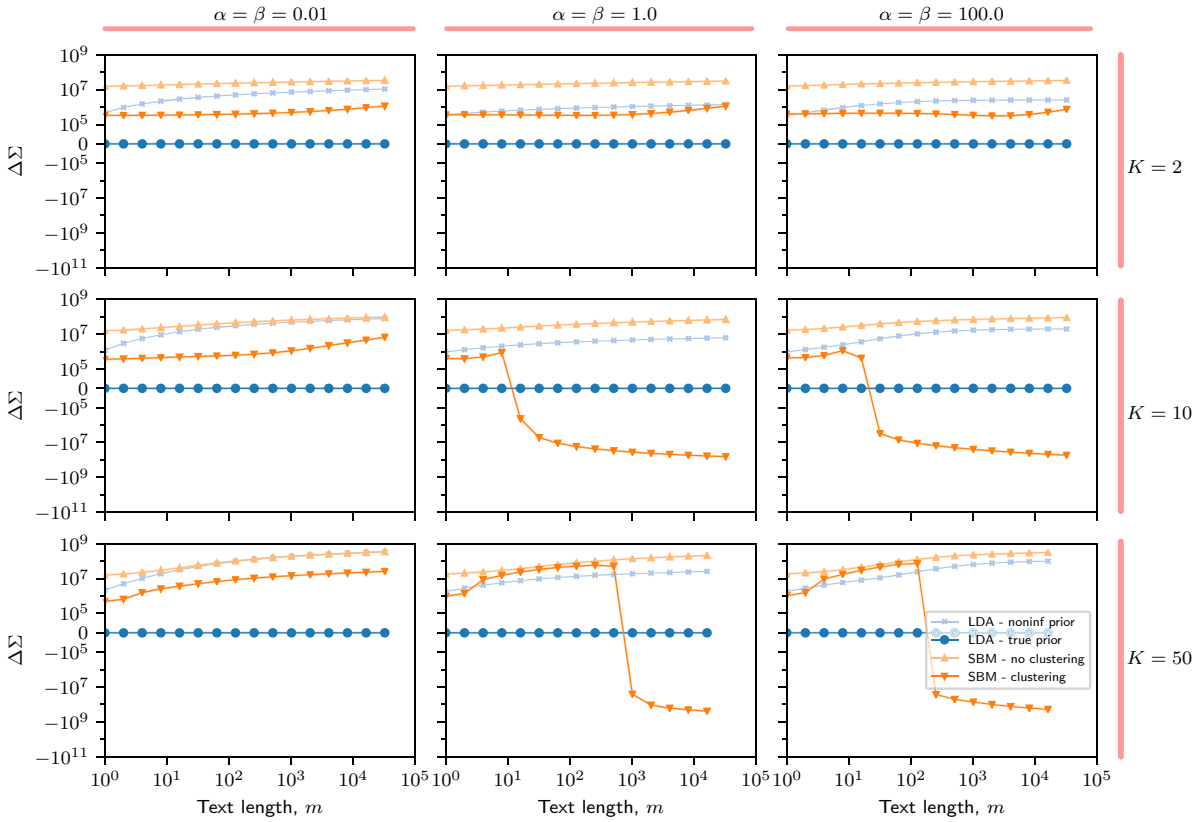

Fig. S2. Varying the number of topics *K* in the comparison between LDA and SBM for artificial corpora drawn from LDA. Same as in Fig. 4A (main text) with different values $K \in \{2, 10, 100\}$ and $(\alpha, \beta) \in \{(0.01, 0.01), (1.0, 1.0), (100.0, 100.0)\}$. Note that the panel in the middle corresponds to Fig. 4 in the main text.

Section S4. 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. $E \propto N$ (*31*). 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 (known as Heaps' law in quantitative linguistics (*14*)), which leads to the superlinear growth of

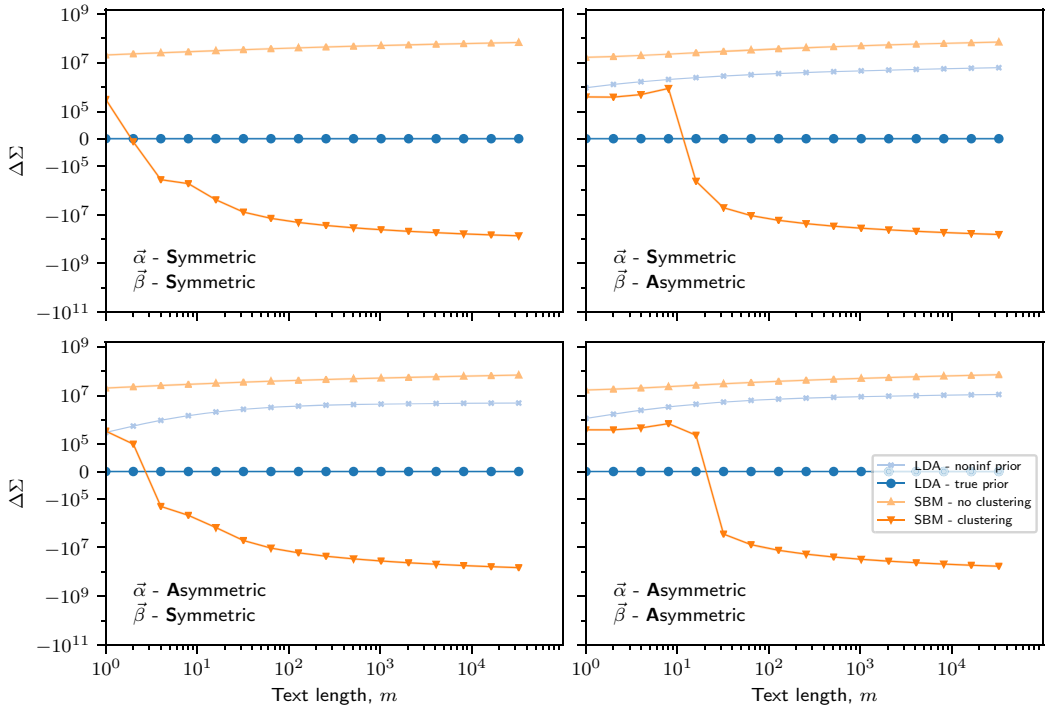

Fig. S3. Varying the base measure of the hyperparameters α and β in the comparison between LDA and SBM for artificial corpora drawn from LDA. Same as in Fig. 4A (main text) with different symmetric and asymmetric $\vec{\alpha}$ and $\vec{\beta}$. For $\vec{\alpha}$, the symmetric case is given by $\alpha_{dr} = \alpha$ and the asymmetric case is given by $\alpha_{dr} = \alpha \times K \times p_r$ with $p_r \propto r^{-1}$ for $r = 1, \ldots, K$ and $\sum_r p_r = 1$. For $\vec{\beta}$, the symmetric case is given by $\alpha_{wr} = \beta$ and the asymmetric case is given by $\beta_{wr} = \beta \times V \times p_w$ with p_w empirically measured in fig. S5 and V is the number of word-types.

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.

Section S5. 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

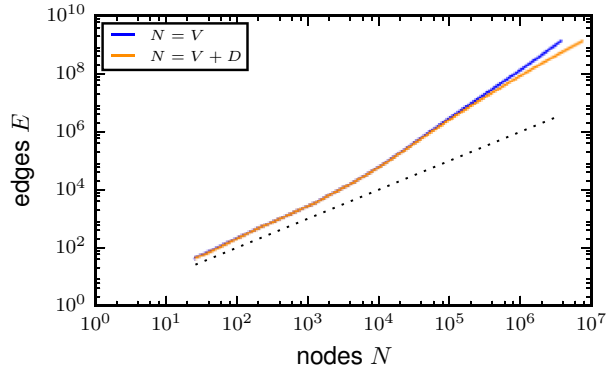

Fig. S4. Word-document networks are not sparse. The number of edges, E, as a function of the number of nodes, N, for the word-document network from the English Wikipedia. The network is grown by adding articles one after another in a randomly chosen order. Shown are the two cases, where i) only the V 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). Figure adapted from Ref. (*61*).

Figure S5 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. (*44*) it has been shown that virtually any collection of documents follows such a distribution of word frequencies.

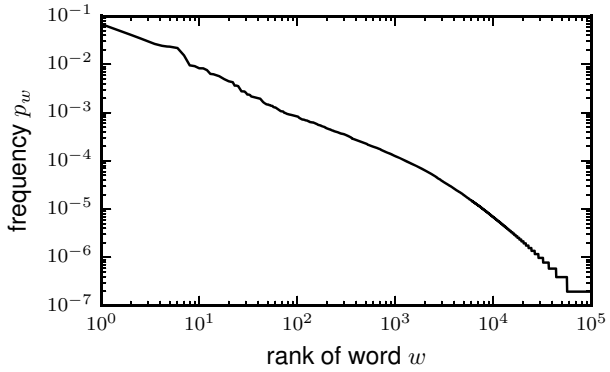

Fig. S5. Empirical rank-frequency distribution. The rank-frequency distribution shows the frequency of each word, $p_w = n_w/M$, ordered according to their rank, where n_w is the number of times word w occurs and $M = \sum_w n_w$ is the total number of words. A word is assigned rank r if it is the r-th most frequent word, i.e. the most frequent word has rank 1.