Supplementary Information for "Paragraph-citation Topic Models for Corpora with Citations: An Application to the United States Supreme Court"*

ByungKoo Kim†‡ Saki Kuzushima†§ Yuki Shiraito¶

First draft: July 13, 2022
This draft: September 23, 2022

Contents

*We thank Kevin Quinn and Stuart Benjamin for their comments on the draft. We also thank Christopher Lucas, Max Goplerud and the audience of the 39th annual summer meeting of the Society for Political Methodology for their constructive comments.

†These authors have contributed equally to this work.

‡Ph.D. Candidate, Department of Political Science, University of Michigan, Ann Arbor, Michigan, USA. Email: `kimbk@umich.edu`.

§Ph.D. Candidate, Department of Political Science, University of Michigan, Ann Arbor, Michigan, USA. Email: `skuzushi@umich.edu`.

¶Assistant Professor, Department of Political Science, University of Michigan. Center for Political Studies, 4259 Institute for Social Research, 426 Thompson Street, Ann Arbor, MI 48104-2321. Phone: 734-615-5165, Email: `shiraito@umich.edu`, URL: `shiraito.github.io`.

A Data preprocessing

Results of topic models can be highly sensitive to how data is preprocessed (Denny and Spirling, 2018). In addition to the simple preprocessing steps we introduced in Section 2, we removed words that appear very commonly across documents. The list of these words are "Statue","Supp","Ann","Rev","Stat","Judgment","Reverse","Follow","Certiorari" and "Opinion". While words such as "Follow" or "Reverse" could convey certain contexts, in legal opinions they are typically used to define how the drafted opinion stands in relation to precedents, and we believe they do not contain useful information with respect to topic discovery. In addition, words such as "Supp" or "Ann" are short words for Supplementary and Annex, which are specific collection of legal documents and thus removed for a better detection of topics.

Since common terms can vary by different subsets, we made additional preprocessing for each subset we used for application of our model. For each subset, we removed terms that appear too frequently as well as terms that appear too infrequently. Terms too common across documents for Privacy subset include "agent", "month","level" and "unfair" and for Voting Rights subset the removed words include"Vote", "Voter","Elect" and "Candid". For both subsets, terms that were too uncommon turned out to be simple typos or names of people or institutions such as "Rawlinson". The above process removed about 40% of the terms.

B Model inference: collapsed Gibbs sampler

$$
\begin{aligned}
\boldsymbol{\eta}_i &\sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\
z_{ip} &\sim \text{Multinom}(1, \text{softmax}(\boldsymbol{\eta}_i)) \\
\boldsymbol{\Psi}_k &\sim \text{Dirichlet}(\boldsymbol{\beta}) \\
\mathbf{w}_{ip} &\sim \text{Multinom}(N_{ip}, \boldsymbol{\Psi}_{z_{ip}}) \\
D^*_{ipj} &\sim \mathcal{N}(\boldsymbol{\tau}^T \mathbf{x}_{ipj}, 1) \\
D_{ipj} &= \begin{cases} 1 \text{ if } D^*_{ipj} \geq 0 \\ 0 \text{ if } D^*_{ipj} < 0 \end{cases}
\end{aligned} \tag{1}
$$

where $\mathbf{x}_{ipj} = \{1, \kappa_j^{(i)}, \eta_{j,z_{ip}}\}$

$\boldsymbol{\mu}$ and $\boldsymbol{\tau}$ each have hyperpriors assigned as

$$
\begin{aligned}
\boldsymbol{\mu} &\sim \text{MVN}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) \\
\boldsymbol{\tau} &\sim \text{MVN}(\boldsymbol{\mu}_\tau, \boldsymbol{\Sigma}_\tau)
\end{aligned} \tag{2}
$$

The full posterior is denoted as follows.

$$
p(\boldsymbol{\eta}, \boldsymbol{\Psi}, \mathbf{Z}, \boldsymbol{\tau} | \mathbf{W}, \mathbf{D}) \propto p(\boldsymbol{\mu}|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) p(\boldsymbol{\tau}|\boldsymbol{\mu}_\tau, \boldsymbol{\Sigma}_\tau) p(\boldsymbol{\eta}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) p(\boldsymbol{\Psi}|\boldsymbol{\beta}) p(\mathbf{Z}|\boldsymbol{\eta}) p(\mathbf{W}|\boldsymbol{\Psi}, \mathbf{Z}) p(\mathbf{D}|\mathbf{D}^*) p(\mathbf{D}^*|\boldsymbol{\tau}, \boldsymbol{\eta}, \mathbf{Z}, \mathbf{D}) \tag{3}
$$

Unfortunately, the inference of the given posterior distribution is hard due to the non-conjugacy between normal prior for $\boldsymbol{\eta}$ and the logistic transformation function (Blei and Lafferty, 2007). Variational inference is the most frequently employed tool to address this problem, with an additional advantage of computational speed. However, obtained parameters are for the variational distribution which is an approximation to the target posterior. Moreover, the quality of the approximation is often not sufficiently explored (Add citations here).

To remedy this problem, we follow the recent advances in the inference of CTM models (Held and Holmes, 2006; Chen et al., 2013; Linderman et al., 2015). We first partially collapse the posterior distribution by integrating out $\boldsymbol{\Psi}$. Then we introduce an auxiliary Polya-Gamma variable $\boldsymbol{\lambda}$ and augment the collapsed posterior. Partial collapsing and data augmentation enables us to use Gibbs sampling which is known to produce samples that converge to the exact posterior.

With $\boldsymbol{\Psi}$ integrated out, our new posterior is proportional to

$$\int_{\boldsymbol{\Psi}} p(\boldsymbol{\eta}, \boldsymbol{\Psi}, \mathbf{Z}, \boldsymbol{\tau} | \mathbf{W}, \mathbf{D}) \propto p(\boldsymbol{\mu}|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) p(\boldsymbol{\tau}|\boldsymbol{\mu}_\tau, \boldsymbol{\Sigma}_\tau) p(\boldsymbol{\eta}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) p(\mathbf{Z}|\boldsymbol{\eta}) p(\mathbf{W}|\mathbf{Z}) p(\mathbf{D}|\mathbf{D}^*) p(\mathbf{D}^*|\boldsymbol{\tau}, \boldsymbol{\eta}, \mathbf{Z}, \mathbf{D}) \tag{4}$$

B.1 Derivation of the conditional distribution for Z

For ipth paragraph, the conditional distribution of z_{ip} is

$$p(z_{ip}^k = 1 | \mathbf{Z}_{-ip}, \boldsymbol{\eta}, \mathbf{W}, \mathbf{D}^*) \propto p(z_{ip}^k = 1 | \boldsymbol{\eta}_i) p(\mathbf{W}_{ip} | z_{ip}^k = 1, \mathbf{Z}_{-ip}, \mathbf{W}_{-ip}) \prod_{j=1}^{i-1} p(D_{ipj}^* | z_{ip}^k = 1, \mathbf{Z}_{-ip}, \boldsymbol{\tau}, \boldsymbol{\eta}, \kappa) \tag{5}$$

The first term is $\frac{\mathrm{e}^{\eta_{ik}}}{\sum_l \mathrm{e}^{\eta_{il}}}$ which is proportional to $\mathrm{e}^{\eta_{ik}}$.

The form of second term warrants further elaboration. Integrating out $\boldsymbol{\Psi}$ as

$$\begin{aligned} p(\mathbf{W}|\mathbf{Z}) &= \int_{\boldsymbol{\Psi}} p(\mathbf{W}, \boldsymbol{\Psi} | \mathbf{Z}) d\boldsymbol{\Psi} \\ &= \int_{\boldsymbol{\Psi}} p(\mathbf{W}|\boldsymbol{\Psi}, \mathbf{Z}) p(\boldsymbol{\Psi}|\mathbf{Z}) d\boldsymbol{\Psi} \\ &= \int_{\boldsymbol{\Psi}} p(\mathbf{W}|\boldsymbol{\Psi}, \mathbf{Z}) p(\boldsymbol{\Psi}) d\boldsymbol{\Psi} \end{aligned} \tag{6}$$

for ipth paragraph with kth topic yields the following.

$$\begin{aligned} p(\mathbf{W}_{ip} | z_{ip}^k = 1, \mathbf{Z}_{-ip}, \mathbf{W}_{-ip}) \propto \int_{\boldsymbol{\Psi}_k} & \Psi_{k1}^{\beta_1 - 1} \Psi_{k2}^{\beta_2 - 1} ... \Psi_{kV}^{\beta_V - 1} \prod_v \Psi_{kv}^{\sum_{l=1}^{n_{ip}} \mathbb{I}(W_{ipl} = v)} \\ & \times \prod_v \prod_{(i', p') \neq (i, p)} \Psi_{kv}^{\sum_{l=1}^{n_{i'p'}} \mathbb{I}(W_{i'p'l} = v) \mathbb{I}(z_{i'p'}^k = 1)} d\boldsymbol{\Psi}_k \end{aligned} \tag{7}$$

Here, N_{ip} denotes the total number of words in ipth paragraph, and n_{ip} denotes the total number of unique words in ipth paragraph. Let $C_k^v = \sum_{i=1}^N \sum_{p=1}^{N_{ip}} \sum_{l=1}^{n_{ip}} \mathbb{I}(W_{ipl} = v) \mathbb{I}(z_{ip}^k = 1)$, and $c_{k,ip}^v = \sum_{l=1}^{n_{ip}} \mathbb{I}(W_{ipl} = v) \mathbb{I}(z_{ip}^k = 1)$ then the above can be simplified as

$$\begin{aligned} p(\mathbf{W}_{ip} | z_{ip}^k = 1, \mathbf{Z}_{-ip}, \mathbf{W}_{-ip}) &\propto \int_{\boldsymbol{\Psi}_k} \Psi_{k1}^{\beta_1 + c_{k,ip}^1 + c_{k,-ip}^1 - 1} \Psi_{k2}^{\beta_2 + c_{k,ip}^2 + c_{k,-ip}^2 - 1} ... \Psi_{kV}^{\beta_V + c_{k,ip}^V + c_{k,-ip}^V - 1} d\boldsymbol{\Psi}_k \\ &= \frac{\prod_v \Gamma(\beta_v + c_{k,ip}^v + c_{k,-ip}^v)}{\Gamma(\sum_v \beta_v + c_{k,ip}^v + c_{k,-ip}^v)} \end{aligned} \tag{8}$$

Imagine a paragraph of 3 words $\mathbf{W}_{ip} = \{1, 1, 3\}$, two of the first word and one of the third word. Then

$$p(\mathbf{W}_{ip}|z_{ip}^k=1,\mathbf{Z}_{-ip},\mathbf{W}_{-ip}) \propto \frac{\prod_v \Gamma(\beta_v + c_{k,ip}^v + c_{k,-ip}^v)}{\Gamma(\sum_v \beta_v + c_{k,ip}^v + c_{k,-ip}^v)} \tag{9}$$

The numerator is

$$\begin{aligned} &\Gamma(\beta_1 + 2 + c_{k,-ip}^1)\Gamma(\beta_3 + 1 + c_{k,-ip}^3) \times \prod_{v\neq(1,3)} \Gamma(\beta_v + c_{k,-ip}^v) \\ &= (\beta_1 + 1 + c_{k,-ip}^1)(\beta_1 + c_{k,-ip}^1)(\beta_3 + c_{k,-ip}^3) \times \prod_v \Gamma(\beta_v + c_{k,-ip}^v) \end{aligned} \tag{10}$$

In the same sense, the denominator is

$$\Gamma(3+\sum_v \beta_v + c_{k,-ip}^v) = (2+\sum_v \beta_v + c_{k,-ip}^v)(1+\sum_v \beta_v + c_{k,-ip}^v)(\sum_v \beta_v + c_{k,-ip}^v)\Gamma(\sum_v \beta_v + c_{k,-ip}^v) \tag{11}$$

Rearrange the above and we have

$$\frac{(\beta_1 + 1 + c_{k,-ip}^1)(\beta_1 + c_{k,-ip}^1)(\beta_3 + c_{k,-ip}^3)}{(2+\sum_v \beta_v + c_{k,-ip}^v)(1+\sum_v \beta_v + c_{k,-ip}^v)(\sum_v \beta_v + c_{k,-ip}^v)} \times \frac{\prod_v \Gamma(\beta_v + c_{k,-ip}^v)}{\Gamma(\sum_v \beta_v + c_{k,-ip}^v)} \tag{12}$$

The second term does not depend on z_{ip}^k. Then for $\mathbf{W}_{ip} = \{1, 1, 3\}$, we have

$$p(\mathbf{W}_{ip}|z_{ip}^k=1,\mathbf{Z}_{-ip},\mathbf{W}_{-ip}) \propto \frac{(\beta_1 + 1 + c_{k,-ip}^1)(\beta_1 + c_{k,-ip}^1)(\beta_3 + c_{k,-ip}^3)}{(2+\sum_v \beta_v + c_{k,-ip}^v)(1+\sum_v \beta_v + c_{k,-ip}^v)(\sum_v \beta_v + c_{k,-ip}^v)} \tag{13}$$

If a paragraph consists of only one word such that $W_{ip} = l$, the above changes to

$$p(\mathbf{W}_{ip}|z_{ip}^k=1,\mathbf{Z}_{-ip},\mathbf{W}_{-ip}) \propto \frac{\beta_l + c_{k,-ip}^l}{\sum_v \beta_v + c_{k,-ip}^v} \tag{14}$$

which matches with the form for the equivalent part in collapsed Gibbs for LDA (Porteous et al., 2008; Xiao and Stibor, 2010; Asuncion et al., 2012).

The third term $p(D_{ipj}^*|z_{ip}^k = 1, \mathbf{Z}_{-ip}, \boldsymbol{\tau}, \boldsymbol{\eta}, \boldsymbol{\kappa}) = \exp\{-\frac{1}{2}\big(D_{ipj}^* - (\tau_0 + \tau_1\kappa_j^{(i)} + \tau_2\eta_{j,z_{ip}})\big)^2\}$ is proportional to

$$\exp\bigg\{-\frac{1}{2}\Big(\tau_2^2\eta_{jk}^2 + 2\big(\tau_0\tau_2 + \tau_1\tau_2\kappa_j^{(i)} - \tau_2 D_{ipj}^*\big)\eta_{jk}\Big)\bigg\} \tag{15}$$

B.2 Derivation of the conditional distribution for $\boldsymbol{\eta}$

$$
\begin{aligned}
p(\boldsymbol{\eta}|\mathbf{Z},\mathbf{W},\mathbf{D}) &= \prod_{i=1}^{N}\Big(\prod_{p=1}^{N_i} p(z_{ip}|\boldsymbol{\eta}_i)\Big)\mathcal{N}(\boldsymbol{\eta}_i|\boldsymbol{\mu},\boldsymbol{\Sigma})\prod_{p=1}^{N_i}\prod_{j=1}^{i-1} p(D^*_{ipj}|\kappa,\boldsymbol{\eta}_i,\mathbf{Z}) \\
&= \prod_{i=1}^{N}\Big(\prod_{p=1}^{N_i}\frac{e^{\eta_{i,z_{ip}}}}{\sum_{j=1}^{K} e^{\eta_{ij}}}\Big)\mathcal{N}(\boldsymbol{\eta}_i|\boldsymbol{\mu},\boldsymbol{\Sigma})\prod_{p=1}^{N_i}\prod_{j=1}^{i-1} p(D^*_{ipj}|\kappa,\boldsymbol{\eta}_i,\mathbf{Z})
\end{aligned}
\tag{16}
$$

Following Held and Holmes (2006), the likelihood for η_{ik} conditioned on $\eta_{i,-k}$ is

$$
\begin{aligned}
\ell(\eta_{ik}|\eta_{i,-k}) &= \prod_{p=1}^{N_i}\Big(\frac{e^{\rho_{ik}}}{1+e^{\rho_{ik}}}\Big)^{z_{ip,k}}\Big(\frac{1}{1+e^{\rho_{ik}}}\Big)^{1-z_{ip,k}} \\
&= \frac{(e^{\rho_{ik}})^{t_{ik}}}{(1+e^{\rho_{ik}})^{N_i}}
\end{aligned}
\tag{17}
$$

where $\rho_{ik} = \eta_{ik} - \log(\sum_{l\neq k} e^{\eta_{il}})$ and $t_{ik} = \sum_{p=1}^{N_i}\mathbb{I}(z_{ip}=k)$.

Then

$$
p(\eta_{ik}|\eta_{i,-k},\mathbf{Z},\mathbf{W},\mathbf{D},\boldsymbol{\tau}) \propto \ell(\eta_{ik}|\eta_{i,-k})\mathcal{N}(\eta_{ik}|\nu_{ik},\sigma_k^2)p(D^*|\boldsymbol{\eta},\boldsymbol{\tau},\mathbf{Z})
\tag{18}
$$

where

$$
\begin{aligned}
\nu_{ik} &= \mu_k - \Lambda_{kk}^{-1}\boldsymbol{\Lambda}_{k,-k}(\boldsymbol{\eta}_{i,-k}-\boldsymbol{\mu}_{i,-k}) \\
\sigma_k^2 &= \boldsymbol{\Lambda}_{kk}^{-1} \\
\boldsymbol{\Lambda} &= \boldsymbol{\Sigma}^{-1}
\end{aligned}
\tag{19}
$$

The third term can be rewritten with respect to $\boldsymbol{\eta}$ as

$$
\begin{aligned}
p(\mathbf{D}^*|\boldsymbol{\eta},\boldsymbol{\tau},\mathbf{Z}) &= \prod_i\prod_p\prod_{j=1}^{i-1}\exp\Big\{-\frac{1}{2}\big(D^*_{ipj}-(\tau_0+\tau_1\kappa_j^{(i)}+\tau_2\eta_{j,z_{ip}})\big)^2\Big\} \\
&\propto \prod_i\prod_p\prod_{j=1}^{i-1}\exp\Big\{-\frac{1}{2(1/\tau_2^2)}\Big(\eta_{j,z_{ip}}^2 - 2\frac{D^*_{ipj}-\tau_0-\tau_1\kappa_j^{(i)}}{\tau_2}\eta_{j,z_{ip}}\Big)\Big\} \\
&\propto \prod_i\prod_p\prod_{j=1}^{i-1}\mathcal{N}(\eta_{j,z_{ip}}|\mu^*_{ipj},\frac{1}{\tau_2^2}) \\
&= \prod_i\prod_p\prod_{j=1}^{i-1}\prod_k\mathcal{N}(\eta_{jk}|\mu^*_{ipj},\frac{1}{\tau_2^2})^{\mathbb{I}(z_{ip}=k)}
\end{aligned}
\tag{20}
$$

where $\mu^*_{ipj} = \frac{D^*_{ipj}-\tau_0-\tau_1\kappa_j^{(i)}}{\tau_2}$. We notice that the above can be rewritten as a product of univariate normal distributions such that

$$\prod_{k} \prod_{s=i+1}^{N} \prod_{p=1}^{N_s} \mathcal{N}(\eta_{ik}|\mu^*_{spi}, \sigma^{2*})^{\mathbb{I}(z_{sp}=k)}$$
$$\equiv \prod_{k=1}^{K} \mathcal{N}(\eta_{ik}|m_{ik}, V_{i,kk}) \tag{21}$$

$\mathbf{V}_i$ is a diagnoal matrix with the kth diagonal entry of the inverse of $\mathbf{V}_i$ (or $\mathbf{V}_i^{-1}$) as

$$V^{-1}_{i,kk} = \frac{1}{\sigma^{2*}} \sum_{s=i+1}^{N} \sum_{p=1}^{N_s} \mathbb{I}(z_{sp} = k)$$
$$= \tau_2^2 \sum_{s=i+1}^{N} \sum_{p=1}^{N_s} \mathbb{I}(z_{sp} = k) \tag{22}$$

The kth entry of $\mathbf{m}_i$ then is

$$m_{ik} = \frac{\tau_2^2 \sum_{s=i+1}^{N} \sum_{p=1}^{N_s} \mu^*_{spi} \mathbb{I}(z_{sp} = k)}{V^{-1}_{i,kk}}$$
$$= \frac{\sum_s \sum_p \mu^*_{spi} \mathbb{I}(z_{sp} = k)}{\sum_s \sum_p \mathbb{I}(z_{sp} = k)} \tag{23}$$

Then the η conditional is

$$p(\eta_{ik}|\eta_{i,-k}, \mathbf{Z}, \mathbf{W}, \mathbf{D}, \boldsymbol{\tau}) \propto \ell(\eta_{ik}|\eta_{i,-k}) \mathcal{N}(\eta_{ik}|\nu_{ik}, \sigma_k^2) \mathcal{N}(\eta_{ik}|m_{ik}, V_{i,kk}) \tag{24}$$

We now introduce Polya-Gamma augmentation such that

$$p(\eta_{ik}|\eta_{i,-k}, \mathbf{Z}, \mathbf{W}, \mathbf{D}, \boldsymbol{\tau}, \lambda_{ik}) \propto \exp\{(t_{ik} - \frac{N_i}{2})\rho_{ik} - \frac{\lambda_{ik}}{2}\rho_{ik}^2\} \mathcal{N}(\eta_{ik}|\nu_{ik}, \sigma_k^2) \mathcal{N}(\eta_{ik}|m_{ik}, V_{i,kk})$$
$$\propto \mathcal{N}(\eta_{ik}| \frac{t_{ik} - N_i/2}{\lambda_{ik}} + \log(\sum_{l \neq k} e^{\eta_{il}}), 1/\lambda_{ik}) \mathcal{N}(\eta_{ik}|\nu_{ik}, \sigma_k^2) \mathcal{N}(\eta_{ik}|m_{ik}, V_{i,kk}) \tag{25}$$

Summing all of the above, the conditional distribution of η_{ik} is

$$p(\eta_{ik}|\eta_{i,-k}, \mathbf{Z}, \mathbf{W}, \mathbf{D}, \boldsymbol{\tau}, \lambda_{ik}) \propto \mathcal{N}(\eta_{ik}|\tilde{\mu}_{ik}, \tilde{\sigma}_k^2) \tag{26}$$

where

$$\tilde{\sigma}_k^2 = (\sigma_k^{-2} + \lambda_{ik} + v^{-1}_{i,kk})^{-1}$$
$$\tilde{\mu}_{ik} = \tilde{\sigma}_k^2 \big(v^{-1}_{i,kk} m_{ik} + \sigma_k^{-2} \nu_{ik} + t_{ik} - \frac{N_i}{2} + \lambda_{ik} \log(\sum_{l \neq k} e^{\eta_{il}}) \big) \tag{27}$$

Derivation of conditional distribution for $\boldsymbol{\lambda}$

The Gibbs sampling for the augmentation variable $\boldsymbol{\lambda}$ is obtained by collecting terms that include $\boldsymbol{\lambda}_i$ in the joint of $\boldsymbol{z}_i$ and $\boldsymbol{\eta}_i$.

$$p(\lambda_{ik}|\mathbf{Z},\mathbf{W},\boldsymbol{\eta}) \propto PG(N_i,\rho_{ik}) \tag{28}$$

B.3 Derivation of conditional distribution for $\mathbf{D}^*$

$$p(D^*_{ipj}|\boldsymbol{\eta},\mathbf{Z},\boldsymbol{\tau},\mathbf{D}) \propto \begin{cases} TN_{(0,\infty)}(\tau_0+\tau_1\kappa_j^{(i)}+\tau_2\eta_{j,z_{ip}},1) & \text{if } D_{ipj}=1 \\ TN_{(-\infty,0]}(\tau_0+\tau_1\kappa_j^{(i)}+\tau_2\eta_{j,z_{ip}},1) & \text{if } D_{ipj}=0 \end{cases} \tag{29}$$

B.4 Derivation of conditional distribution for $\boldsymbol{\tau}$

Let $\mathbf{x}_{ipj}=[1,\kappa_j^{(i)},\eta_{j,z_{ip}}]^T$ and $\boldsymbol{\tau}=[\tau_0,\tau_1,\tau_2]^T$

$$\begin{aligned} p(\boldsymbol{\tau}|\boldsymbol{\eta},\mathbf{Z},\mathbf{D}^*) &\propto exp\Bigg\{-\frac{1}{2}\sum_{ipj}\Big(D^*_{ipj}-\mathbf{x}_{ipj}^T\boldsymbol{\tau}\Big)^2\Bigg\}N(\boldsymbol{\mu_\tau},\Sigma_\tau) \\ &\propto N(\tilde{\boldsymbol{\tau}},\tilde{\Sigma_\tau}) \end{aligned} \tag{30}$$

where $\tilde{\Sigma_\tau}=\left(\left(\sum_{ipj}\mathbf{x}_{ipj}\mathbf{x}_{ipj}^T\right)+\Sigma_\tau^{-1}\right)^{-1}$ and $\tilde{\tau}=\tilde{\Sigma_\tau}\left(\left(\sum_{ipj}\mathbf{x}_{ipj}^T D^*_{ipj}\right)+\Sigma_\tau^{-1}\boldsymbol{\mu_\tau}\right)$

B.5 Recovering $\boldsymbol{\Psi}$

We estimate the integrated out parameter $\boldsymbol{\Psi}$ from our posterior samples as follows.

$$\hat{\Psi}_{kv}=\frac{\sum_i\sum_p\left(\beta_v+\mathbb{I}(z_{ip}^k=1)\mathrm{W}_{ip,v}\right)}{\sum_i\sum_p\sum_l\left(\beta_l+\mathbb{I}(z_{ip}^k=1)\mathrm{W}_{ip,l}\right)} \tag{31}$$

C Initialization strategy for collapsed Gibbs sampler

Similar to other topic models, the PCTM contains a number of parameters for an estimation which increases the concern for multi-modality of the parameter space. Bad initial values can negatively impact the convergence of mcmc chains to the posterior distribution. Initial values distant from the global mode of the parameter space results in slow convergence. Also, for models with high dimensional parameter space, such as LDA or PCTM, bad initial values increase the possibility of the mcmc chain being stuck at local modes that offer suboptimal interpretations at best. To address these concerns, we propose to fit LDA with variational EM to obtain reasonable initial values for $\boldsymbol{\eta}$, then use them to generate reasonable initial values for other parameters $(\mathbf{Z}, \boldsymbol{\lambda}, \mathbf{D}^*, \boldsymbol{\tau})$.

We first fit LDA with variational EM on document-level document-feature matrix to obtain $\hat{\boldsymbol{\theta}}$. For ith document,

$$
\begin{aligned}
z_{ip}^{(0)} &\sim \text{Categorical}(\hat{\boldsymbol{\theta}}_i) \quad \forall p = 1, 2, ..., N_i \\
\boldsymbol{\eta}_i^{(0)} &= \log(\hat{\boldsymbol{\theta}}_i / \hat{\theta}_{iK})
\end{aligned} \tag{32}
$$

Set $\tilde{\tau}_0$, or the sparsity parameter, using the observed density of the citation matrix and randomly draw the other two parameters as

$$
\begin{aligned}
\tilde{\tau}_0 &= \frac{1}{2}\log(\text{density}(\mathbf{D})) \\
\tilde{\tau}_1, \tilde{\tau}_2 &\sim \text{unif}(0, 1)
\end{aligned} \tag{33}
$$

Sample $\mathbf{D}^*$ using the above parameters

$$
\begin{aligned}
{D^*_{ipj}}^{(0)} &\sim TN_{(-\infty,0)}(\tilde{\tau}_0 + \tilde{\tau}_1 \kappa_j^{(i)} + \tilde{\tau}_2 \eta^{(0)}_{j, z_{ip}^{(0)}}, 1) \quad \text{if } D_{ipj} = 0 \\
{D^*_{ipj}}^{(0)} &\sim TN_{[0,\infty)}(\tilde{\tau}_0 + \tilde{\tau}_1 \kappa_j^{(i)} + \tilde{\tau}_2 \eta^{(0)}_{j, z_{ip}^{(0)}}, 1) \quad \text{if } D_{ipj} = 1
\end{aligned} \tag{34}
$$

Then set $\boldsymbol{\tau}^{(0)}$ again using MLE

$$
\boldsymbol{\tau}^{(0)} = (\sum_{ipj} \mathbf{x}_{ipj}^{(0)} {\mathbf{x}_{ipj}^{(0)}}^T)^{-1} (\sum_{ipj} {\mathbf{x}_{ipj}^{(0)}}^T {D^*_{ipj}}^{(0)}) \tag{35}
$$

where $\mathbf{x}_{ipj}^{(0)} = \{1, \kappa_j^{(i)}, \eta^{(0)}_{j, z_{ip}^{(0)}}\}$

Finally, set the values of $\boldsymbol{\lambda}^{(0)}$ by

$$
\lambda_i^{(0)} \sim \text{PG}(N_i, \boldsymbol{\eta}_i^{(0)}) \tag{36}
$$

D More results on simulation and SCOTUS application

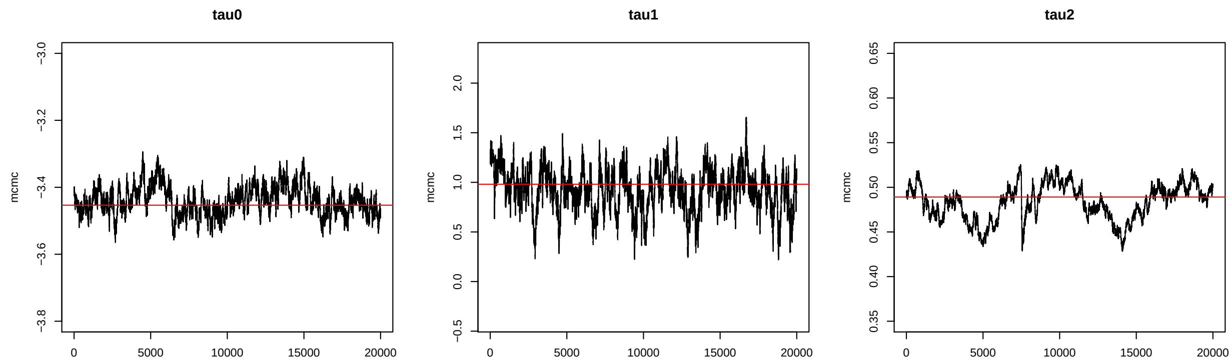

Figure D.1: MCMC convergence of $\boldsymbol{\tau}$ posterior samples in simulation. Horizontal red line indicates the true values of $\boldsymbol{\tau}$.

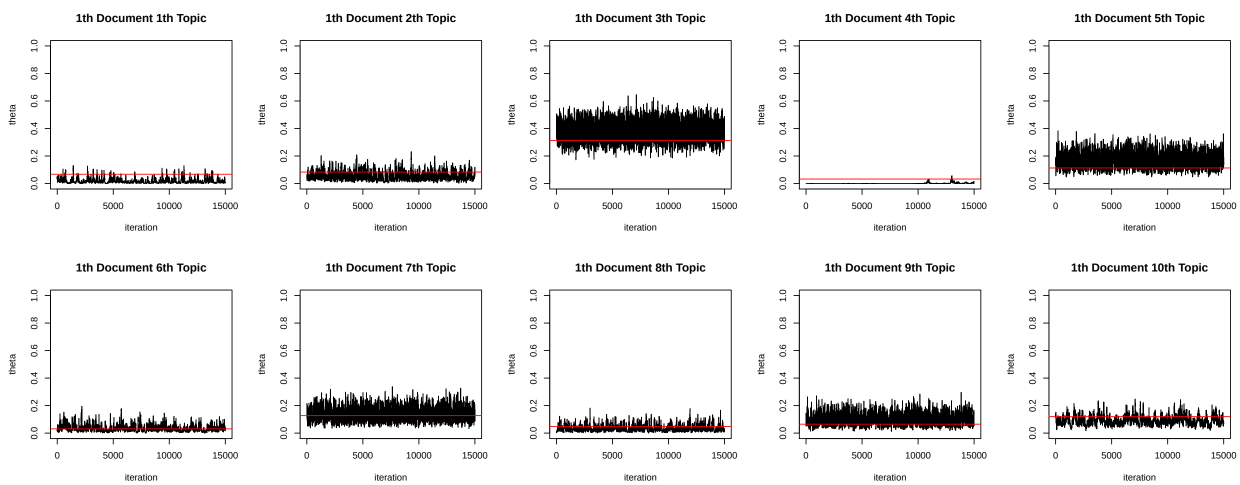

Figure D.2: MCMC convergence of $\boldsymbol{\theta}$ parameters for the first document. $\boldsymbol{\theta}$ values are obtained by transforming the posterior samples of $\boldsymbol{\eta}$ of the corresponding document. Horizontal red line indicates the true values of $\boldsymbol{\theta}$ for the first document for each topic. We do not display the MCMC convergence for other documents, but all documents show similar level of convergence to the true value of $\boldsymbol{\theta}$.

	All Data	Voter Eligibility	Ballot Access	Preclearance Requirement	Voter Dilution
Allen v. State Board of Elections	1	40.5	n/a	1	23.5
Perkins v. Matthews	2	n/a	n/a	2	21
South Carolina v. Katzenbach	3	12	n/a	4	n/a
US v. Sheffield	4	40.5	n/a	3	n/a
City of Richmond v. US	5	n/a	n/a	6	14
Georgia v. US	6	n/a	n/a	5	26
Beer v. US et al.	7	n/a	n/a	8	12
City of Mobile v. Bolden	8	25	20	32	7
City of Lockhart v. US	9	n/a	28.5	9	15.5
City of Rome v. US	10	24	n/a	19	17

Table D.1: The ranks of inward relevance scores computed following Fowler et al. (2007). The displayed cases are ranked in the top 10 with all data, but their ranks vary when we look into topic-specific subnetworks.

	All Data	Voter Eligibility	Ballot Access	Preclearance Requirement	Voter Dilution
Morse v. Rep Party of Virginia	1	18	3	3	25
City of Rome v. US	2	19	n/a	1	22
McCain v. Lybrand	3	n/a	n/a	2	16.5
Riley v. Kennedy	4	45.5	28.5	4	n/a
Dougherty v. White	5	n/a	16	5	n/a
Hathorn v. Lovorn	6	n/a	n/a	6	n/a
NAACP v. Hampton County	7	n/a	n/a	7	n/a
US v. Sheffield	8	22	n/a	13	n/a
Young v. Fordice	9	n/a	n/a	8	n/a
Chisom v. Roemer	10	n/a	n/a	24	4

Table D.2: The ranks of outward relevance scores computed following Fowler et al. (2007). The displayed cases are ranked in the top 10 with all data, but their ranks vary when we look into topic-specific subnetworks.

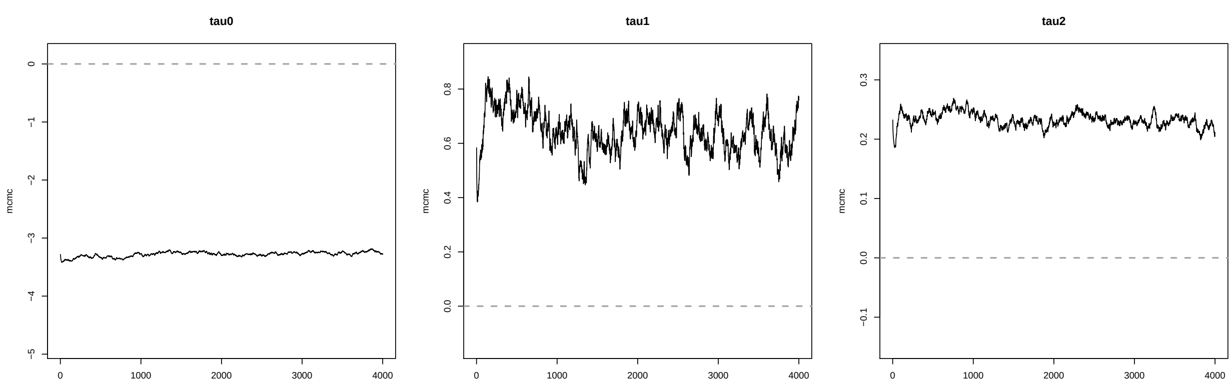

Figure D.3: MCMC convergence of $\boldsymbol{\tau}$ posterior samples for the SCOTUS application on Privacy issue area. Horizontal red line indicates the true values of $\boldsymbol{\tau}$.

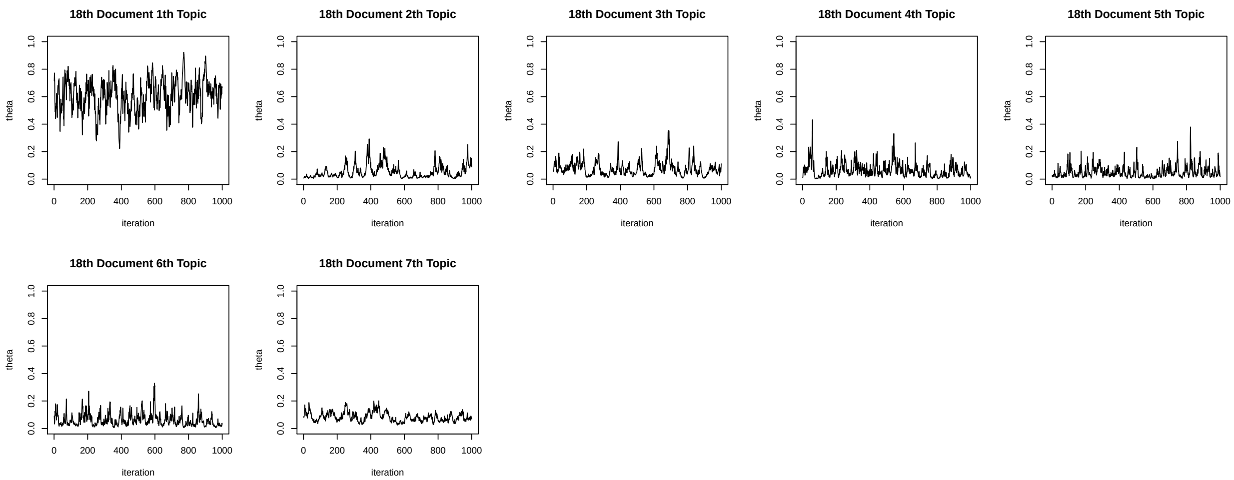

Figure D.4: MCMC convergence of $\boldsymbol{\theta}$ parameters for the 18th document in the subset of Privacy issue area. $\boldsymbol{\theta}$ values are obtained by transforming the posterior samples of $\boldsymbol{\eta}$ of the corresponding document. Horizontal red line indicates the true values of $\boldsymbol{\theta}$ for the 18th document for each topic. We do not display the MCMC convergence for other documents, but all documents show similar level of convergence to the true value of $\boldsymbol{\theta}$.

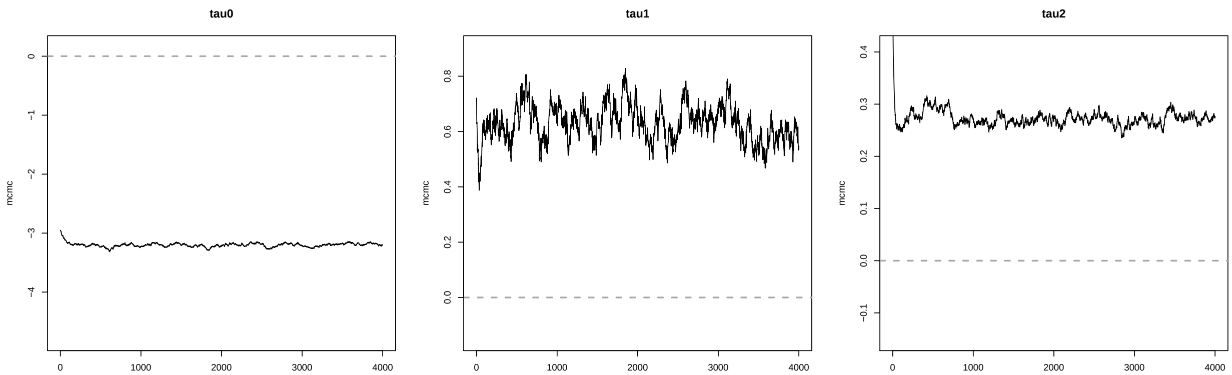

Figure D.5: MCMC convergence of $\boldsymbol{\tau}$ posterior samples for the SCOTUS application on Voting Rights issue area. Horizontal red line indicates the true values of $\boldsymbol{\tau}$.

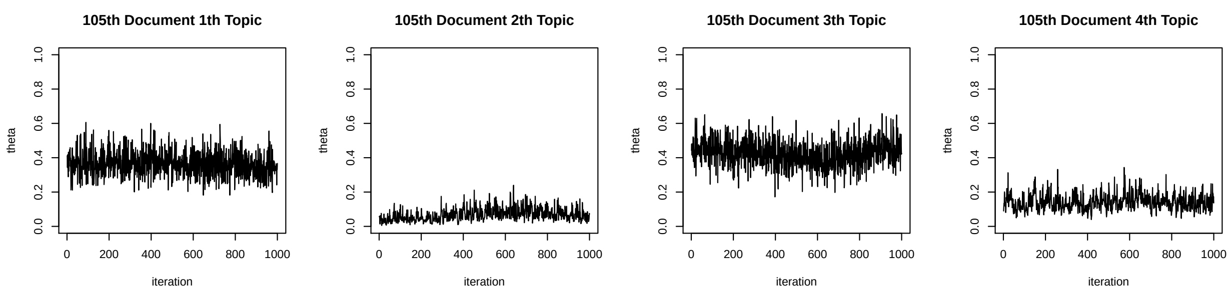

Figure D.6: MCMC convergence of $\boldsymbol{\theta}$ parameters for the 105th document in the subset of Voting Rights issue area. $\boldsymbol{\theta}$ values are obtained by transforming the posterior samples of $\boldsymbol{\eta}$ of the corresponding document. Horizontal red line indicates the true values of $\boldsymbol{\theta}$ for the 105th document for each topic. We do not display the MCMC convergence for other documents, but all documents show similar level of convergence to the true value of $\boldsymbol{\theta}$.

E Posterior Predictive Probability

The posterior probability of words and citations in a paragraph p in a document i can be computed by the following formula.

$$
\begin{aligned}
& p(\mathbf{W}_{ip}, \mathbf{D}_{ip} | \mathbf{W}^{train}, \mathbf{D}^{train}) \\
& \propto \int_{\boldsymbol{\eta}, \boldsymbol{\Psi}, \boldsymbol{\tau}} \sum_{\mathbf{Z}} p(\mathbf{W}_{ip}, \mathbf{D}_{ip} | \mathbf{Z}, \boldsymbol{\eta}, \boldsymbol{\Psi}, \boldsymbol{\tau}) \times p(\mathbf{Z}, \boldsymbol{\eta}, \boldsymbol{\Psi}, \boldsymbol{\tau}, | \mathbf{W}^{train}, \mathbf{D}^{train}) d\boldsymbol{\eta} d\boldsymbol{\Psi} d\boldsymbol{\tau} \\
& \propto \int_{\boldsymbol{\eta}, \boldsymbol{\Psi}, \boldsymbol{\tau}} \sum_{\mathbf{Z}} p(\mathbf{W}_{ip}, \mathbf{D}_{ip} | \mathbf{Z}, \boldsymbol{\eta}, \boldsymbol{\Psi}, \boldsymbol{\tau}) \times p(\mathbf{Z} | \boldsymbol{\eta}, \boldsymbol{\Psi}, \boldsymbol{\tau}, \mathbf{W}^{train}, \mathbf{D}^{train}) p(\boldsymbol{\eta}, \boldsymbol{\Psi}, \boldsymbol{\tau} | \mathbf{W}^{train}, \mathbf{D}^{train}) d\boldsymbol{\eta} d\boldsymbol{\Psi} d\boldsymbol{\tau} \\
& \approx \sum_{k=1}^{K} \Big\{ p(\mathbf{W}_{ip}, \mathbf{D}_{ip} | z_{ip}^{k} = 1, \hat{\mathbf{Z}}^{train}, \hat{\boldsymbol{\eta}}, \hat{\boldsymbol{\Psi}}, \hat{\boldsymbol{\tau}}) \times \mathbb{P}(z_{ip}^{k} = 1 | \hat{\boldsymbol{\eta}}) \Big\} \\
& = \sum_{k=1}^{K} \Big\{ p(\mathbf{W}_{ip} | z_{ip}^{k} = 1, \hat{\boldsymbol{\Psi}}) \times \prod_{j=1}^{i-1} p(D_{ipj} | \hat{\boldsymbol{\tau}}, \hat{\boldsymbol{\eta}}, z_{ip}^{k} = 1) \times \mathbb{P}(z_{ip}^{k} = 1 | \hat{\boldsymbol{\eta}}) \Big\} \\
& = \sum_{k=1}^{K} \Big\{ p(\mathbf{W}_{ip} | z_{ip}^{k} = 1, \hat{\boldsymbol{\Psi}}) \times \prod_{j=1}^{i-1} p(D_{ipj}^{*} > 0 | \hat{\boldsymbol{\tau}}, \hat{\boldsymbol{\eta}}, z_{ip}^{k} = 1)^{\mathbb{I}\{D_{ipj}=1\}} p(D_{ipj}^{*} < 0 | \hat{\boldsymbol{\tau}}, \hat{\boldsymbol{\eta}}, z_{ip}^{k} = 1)^{\mathbb{I}\{D_{ipj}=0\}} \\
& \quad \times \mathbb{P}(z_{ip}^{k} = 1 | \hat{\boldsymbol{\eta}}) \Big\} \\
& \propto \sum_{k=1}^{K} \Bigg\{ \prod_{v=1}^{V} \boldsymbol{\Psi}_{vk}^{W_{ipv}} \times \prod_{j=1}^{i-1} \Big[\int_{t=0}^{\infty} p(D_{ipj}^{*} = t | \boldsymbol{\tau}_0 + \boldsymbol{\tau}_1 \kappa_j^{(i)} + \boldsymbol{\tau}_2 \boldsymbol{\eta}_{jk}) dt \Big]^{\mathbb{I}\{D_{ipj}=1\}} \\
& \quad \times \Big[\int_{t=-\infty}^{0} p(D_{ipj}^{*} = t | \boldsymbol{\tau}_0 + \boldsymbol{\tau}_1 \kappa_j^{(i)} + \boldsymbol{\tau}_2 \boldsymbol{\eta}_{jk}) dt \Big]^{\mathbb{I}\{D_{ipj}=0\}} \times \frac{\exp(\boldsymbol{\eta}_{ik})}{\sum_{k'=1}^{K} \exp(\boldsymbol{\eta}_{ik'})} \Bigg\}
\end{aligned}
\tag{37}
$$

In the third line, we approximate the integral over $\boldsymbol{\eta}$, $\boldsymbol{\Psi}$, and $\boldsymbol{\tau}$ as well as the summation over $\mathbf{Z}$ in the training data. We draw samples of these parameters from the posterior of the model fit on the training data for $\boldsymbol{\eta}$, $\boldsymbol{\tau}$, and $\mathbf{Z}$ in the training data, and we use an MLE estimate for $\boldsymbol{\Psi}$ (see Appendix B.5). The integrals in the last line can be easily computed because D_{ipj}^{*} follows normal distributions with unit variance. We can also see that the posterior probability of a paragraph p in a document i being topic k is proportional to the components inside the summation over k.

References

Asuncion, A., Welling, M., Smyth, P., and Teh, Y. W. (2012). On smoothing and inference for topic models. *arXiv preprint arXiv:1205.2662*.

Blei, D. M. and Lafferty, J. D. (2007). A correlated topic model of science. *The annals of applied statistics*, 1(1):17–35.

Chen, J., Zhu, J., Wang, Z., Zheng, X., and Zhang, B. (2013). Scalable inference for logistic-normal topic models. *Advances in neural information processing systems*, 26.

Denny, M. J. and Spirling, A. (2018). Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it. *Political Analysis*, 26(2):168–189.

Fowler, J. H., Johnson, T. R., Spriggs, J. F., Jeon, S., and Wahlbeck, P. J. (2007). Network analysis and the law: Measuring the legal importance of precedents at the us supreme court. *Political Analysis*, 15(3):324–346.

Held, L. and Holmes, C. C. (2006). Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian analysis*, 1(1):145–168.

Linderman, S., Johnson, M. J., and Adams, R. P. (2015). Dependent multinomial models made easy: Stick-breaking with the pólya-gamma augmentation. *Advances in Neural Information Processing Systems*, 28.

Porteous, I., Newman, D., Ihler, A., Asuncion, A., Smyth, P., and Welling, M. (2008). Fast collapsed gibbs sampling for latent dirichlet allocation. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 569–577.

Xiao, H. and Stibor, T. (2010). Efficient collapsed gibbs sampling for latent dirichlet allocation. In *Proceedings of 2nd asian conference on machine learning*, pages 63–78. JMLR Workshop and Conference Proceedings.