Supplementary Information for "A Unified Model of Text and Citations for Topic-Specific Citation Networks: Application to the Supreme Court of the United States"*

ByungKoo Kim[†‡] Saki Kuzushima[†§] Yuki Shiraito[¶]

First draft: July 13, 2022
This draft: February 10, 2025

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.

‡Assistant Professor, KDI School of Public Policy, Sejong, Republic of Korea, Email: kimbk@kdischool.ac.kr.

§Postdoctoral Fellow, Program on U.S.-Japan Relations, Weatherhead Center for International Affairs, Harvard University, Cambridge, Massachusetts, USA. Email: sakikuzushima@fas.harvard.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 Constructing SCOTUS Paragraph-document Citation Network

We construct a new dataset of the SCOTUS opinions that combines text and citation networks. The original data is obtained from the Caselaw Access Project, which allows public access to all official and published opinions at all levels of the US courts (Caselaw Access Project, 2024). The data contains the full text of majority and minority opinions in addition to their metadata, such as decision dates, reporter names, volumes in the reporter, and page numbers. We decided to focus on the text of majority opinions and discard minority opinions since minority opinions rarely receive recognition as legal precedents. In total, the population data contains 24,000 cases with 749,888 paragraphs with the year ranging from 1834 to 2013.

The document networks of the SCOTUS consist of two forms of datasets: text and citation networks. With respect to the text, we construct a "paragraph"-feature matrix based on the population corpus. A paragraph feature matrix is similar to a common document-feature matrix, where a (i, j) element of the matrix corresponds to the number of times a unique feature j appears in a document i. The only difference is that a paragraph-feature matrix uses paragraphs instead of documents as a unit. This is because our proposed model uses paragraphs as a unit of analysis. After tokenizing the corpus, we removed punctuations, symbols, special characters, numbers, and common English stopwords.[1] In addition to the common list of stopwords, we also removed legal terms that are common across the documents in our data such as "court", "state", "law" and, "trial". After removing too frequent words and too rare words, the population paragraph-feature matrix contains 32,644 unique features.

The other component is a citation network. While previous studies have constructed citation networks of the SCOTUS cases (Fowler et al., 2007; Clark and Lauderdale, 2012), their unit of analysis is at the document level while ours is at the paragraph level. In other words, we want to form an adjacency matrix of $G \times N$ where G is the number of paragraphs and N is the number of documents, and the (ip, j) element of the matrix is 1 if paragraph p of document i cites document j, and 0 otherwise. Since such data is not readily available, we constructed our own citation network of the SCOTUS cases by extracting citations from the text via regular expression matching. One of the challenges of this approach is that a citation is recorded by multiple reporters and appears in the paragraph as many times as the number of reporters. To avoid complication, we focused on the citations to the official reporter, *the United States Reports*, because this is the recommended and the most dominant

[1] We used the set of English stopwords provided in `quanteda` package in R (Benoit et al., 2018).

citation method. A citation to a case in the United States Reports typically has a relatively consistent format and thus is easier to be extracted through regular expression matching. For instance, a citation to *Roe v. Wade* is typically written as *Row v. Wade, 410 U.S. 113 (1973)*. Since we focus on the SCOTUS cases only, citations to and from outside of the corpus (e.g. citations to and from the Courts of Appeals and State courts) were discarded. This results in 191,173 citations in total.

In this paper, we focus on a subset of this dataset for our applications. For our application, we decided to focus on cases on the Privacy issue area, which includes decisions about abortion and public disclosure of private information. We chose this as our primary application data since existing literature on citation networks of the SCOTUS cases often focuses on this issue (Fowler et al., 2007; Clark and Lauderdale, 2012). It is also an important application given the recent controversial decision that overruled the landmark case on constitutional rights to abortion. After we subset the data, we performed more preprocessing based on the term frequency within the subset. More details of data pre-processing for each subset are available in the Supplementary Information document, Section A. This subset on the Privacy issue area consists of 106 documents with 4,669 paragraphs, 5,838 unique words, and 452 citations.

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

This section describes the details of the collapsed Gibbs sampler for the proposed model. Our model is as follows.

$$
\begin{aligned}
D_{ipj} &= \begin{cases} 1 \text{ if } D^*_{ipj} \geq 0 \\ 0 \text{ if } D^*_{ipj} < 0 \end{cases} \\
D^*_{ipj} &\sim \mathcal{N}(\boldsymbol{\tau}^T \mathbf{x}_{ipj}, 1) \quad \text{where } \mathbf{x}_{ipj} = [1, \kappa_j^{(i)}, \eta_{j,z_{ip}}] \\
\mathbf{w}_{ip} &\sim \text{Multinomial}(N_{ip}, \boldsymbol{\Psi}_{z_{ip}}) \\
z_{ip} &\sim \text{Multinomial}(1, \text{softmax}(\boldsymbol{\eta}_i)) \\
\boldsymbol{\Psi}_k &\sim \text{Dirichlet}(\boldsymbol{\beta}) \\
\boldsymbol{\eta}_i &\sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\
\boldsymbol{\mu} &\sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) \\
\boldsymbol{\tau} &\sim \mathcal{N}(\boldsymbol{\mu}_\tau, \boldsymbol{\Sigma}_\tau)
\end{aligned}
\tag{1}
$$

The full posterior is denoted as follows.

$$
p(\boldsymbol{\eta}, \boldsymbol{\Psi}, \mathbf{Z}, \boldsymbol{\tau} | \mathbf{W}, \mathbf{D}) \propto p(\boldsymbol{\mu}|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) p(\boldsymbol{\tau}|\boldsymbol{\mu}_\tau, \boldsymbol{\Sigma}_\tau) p(\boldsymbol{\eta}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) p(\boldsymbol{\Psi}|\boldsymbol{\beta}) p(\mathbf{Z}|\boldsymbol{\eta}) p(\mathbf{W}|\boldsymbol{\Psi}, \mathbf{Z}) p(\mathbf{D}|\mathbf{D}^*) p(\mathbf{D}^*|\boldsymbol{\tau}, \boldsymbol{\eta}, \mathbf{Z}, \mathbf{D}) \tag{2}
$$

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

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

B.1 Derivation of the conditional distribution for Z

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

$$p(z_{ip}^k = 1 | \mathbf{Z}_{-ip}, \boldsymbol{\eta}, \mathbf{W}, \mathbf{D}^*) \propto p(z_{ip}^k = 1 | \boldsymbol{\eta}_i) p(\mathbf{W}_{ip} | z_{ip}^k = 1, \mathbf{Z}_{-ip}, \mathbf{W}_{-ip}) \prod_{j=1}^{i-1} p(D_{ipj}^* | z_{ip}^k = 1, \mathbf{Z}_{-ip}, \boldsymbol{\tau}, \boldsymbol{\eta}, \kappa) \tag{4}$$

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

$$\begin{aligned} p(\mathbf{W}|\mathbf{Z}) &= \int_{\boldsymbol{\Psi}} p(\mathbf{W}, \boldsymbol{\Psi}|\mathbf{Z}) d\boldsymbol{\Psi} \\ &= \int_{\boldsymbol{\Psi}} p(\mathbf{W}|\boldsymbol{\Psi}, \mathbf{Z}) p(\boldsymbol{\Psi}|\mathbf{Z}) d\boldsymbol{\Psi} \\ &= \int_{\boldsymbol{\Psi}} p(\mathbf{W}|\boldsymbol{\Psi}, \mathbf{Z}) p(\boldsymbol{\Psi}) d\boldsymbol{\Psi} \end{aligned} \tag{5}$$

for ipth paragraph with kth topic yields the following.

$$\begin{aligned} p(\mathbf{W}_{ip} | z_{ip}^k = 1, \mathbf{Z}_{-ip}, \mathbf{W}_{-ip}) \propto \int_{\boldsymbol{\Psi}_k} \Psi_{k1}^{\beta_1 - 1} \Psi_{k2}^{\beta_2 - 1} ... \Psi_{kV}^{\beta_V - 1} \prod_v \Psi_{kv}^{\sum_{l=1}^{n_{ip}} \mathbb{I}(W_{ipl} = v)} \\ \times \prod_v \prod_{(i',p') \neq (i,p)} \Psi_{kv}^{\sum_{l=1}^{n_{i'p'}} \mathbb{I}(W_{i'p'l} = v) \mathbb{I}(z_{i'p'}^k = 1)} d\boldsymbol{\Psi}_k \end{aligned} \tag{6}$$

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

$$\begin{aligned} p(\mathbf{W}_{ip} | z_{ip}^k = 1, \mathbf{Z}_{-ip}, \mathbf{W}_{-ip}) &\propto \int_{\boldsymbol{\Psi}_k} \Psi_{k1}^{\beta_1 + c_{k,ip}^1 + c_{k,-ip}^1 - 1} \Psi_{k2}^{\beta_2 + c_{k,ip}^2 + c_{k,-ip}^2 - 1} ... \Psi_{kV}^{\beta_V + c_{k,ip}^V + c_{k,-ip}^V - 1} d\boldsymbol{\Psi}_k \\ &= \frac{\prod_v \Gamma(\beta_v + c_{k,ip}^v + c_{k,-ip}^v)}{\Gamma(\sum_v \beta_v + c_{k,ip}^v + c_{k,-ip}^v)} \end{aligned} \tag{7}$$

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

third word. Then

$$p(\mathbf{W}_{ip}|z_{ip}^k = 1, \mathbf{Z}_{-ip}, \mathbf{W}_{-ip}) \propto \frac{\prod_v \Gamma(\beta_v + c_{k,ip}^v + c_{k,-ip}^v)}{\Gamma(\sum_v \beta_v + c_{k,ip}^v + c_{k,-ip}^v)} \tag{8}$$

The numerator is

$$\begin{aligned} &\Gamma(\beta_1 + 2 + c_{k,-ip}^1)\Gamma(\beta_3 + 1 + c_{k,-ip}^3) \times \prod_{v \neq (1,3)} \Gamma(\beta_v + c_{k,-ip}^v) \\ &= (\beta_1 + 1 + c_{k,-ip}^1)(\beta_1 + c_{k,-ip}^1)(\beta_3 + c_{k,-ip}^3) \times \prod_v \Gamma(\beta_v + c_{k,-ip}^v) \end{aligned} \tag{9}$$

In the same sense, the denominator is

$$\Gamma(3 + \sum_v \beta_v + c_{k,-ip}^v) = (2 + \sum_v \beta_v + c_{k,-ip}^v)(1 + \sum_v \beta_v + c_{k,-ip}^v)(\sum_v \beta_v + c_{k,-ip}^v)\Gamma(\sum_v \beta_v + c_{k,-ip}^v) \tag{10}$$

Rearrange the above and we have

$$\frac{(\beta_1 + 1 + c_{k,-ip}^1)(\beta_1 + c_{k,-ip}^1)(\beta_3 + c_{k,-ip}^3)}{(2 + \sum_v \beta_v + c_{k,-ip}^v)(1 + \sum_v \beta_v + c_{k,-ip}^v)(\sum_v \beta_v + c_{k,-ip}^v)} \times \frac{\prod_v \Gamma(\beta_v + c_{k,-ip}^v)}{\Gamma(\sum_v \beta_v + c_{k,-ip}^v)} \tag{11}$$

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

$$p(\mathbf{W}_{ip}|z_{ip}^k = 1, \mathbf{Z}_{-ip}, \mathbf{W}_{-ip}) \propto \frac{(\beta_1 + 1 + c_{k,-ip}^1)(\beta_1 + c_{k,-ip}^1)(\beta_3 + c_{k,-ip}^3)}{(2 + \sum_v \beta_v + c_{k,-ip}^v)(1 + \sum_v \beta_v + c_{k,-ip}^v)(\sum_v \beta_v + c_{k,-ip}^v)} \tag{12}$$

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

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

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

$$\exp\left\{ -\frac{1}{2}\Big(\tau_2^2\eta_{jk}^2 + 2\big(\tau_0\tau_2 + \tau_1\tau_2\kappa_j^{(i)} - \tau_2 D_{ipj}^*\big)\eta_{jk}\Big) \right\} \tag{14}$$

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

$$
\begin{aligned}
p(\boldsymbol{\eta}|\mathbf{Z},\mathbf{W},\mathbf{D}) &= \prod_{i=1}^{N}\Big(\prod_{p=1}^{N_i} p(z_{ip}|\boldsymbol{\eta}_i)\Big)\mathcal{N}(\boldsymbol{\eta}_i|\boldsymbol{\mu},\boldsymbol{\Sigma})\prod_{p=1}^{N_i}\prod_{j=1}^{i-1} p(D^*_{ipj}|\kappa,\boldsymbol{\eta}_i,\mathbf{Z}) \\
&= \prod_{i=1}^{N}\Big(\prod_{p=1}^{N_i}\frac{e^{\eta_{i,z_{ip}}}}{\sum_{j=1}^{K}e^{\eta_{ij}}}\Big)\mathcal{N}(\boldsymbol{\eta}_i|\boldsymbol{\mu},\boldsymbol{\Sigma})\prod_{p=1}^{N_i}\prod_{j=1}^{i-1} p(D^*_{ipj}|\kappa,\boldsymbol{\eta}_i,\mathbf{Z}) \qquad (15)
\end{aligned}
$$

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

$$
\begin{aligned}
\ell(\eta_{ik}|\eta_{i,-k}) &= \prod_{p=1}^{N_i}\Big(\frac{e^{\rho_{ik}}}{1+e^{\rho_{ik}}}\Big)^{z_{ip,k}}\Big(\frac{1}{1+e^{\rho_{ik}}}\Big)^{1-z_{ip,k}} \\
&= \frac{(e^{\rho_{ik}})^{t_{ik}}}{(1+e^{\rho_{ik}})^{N_i}} \qquad (16)
\end{aligned}
$$

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

$$
p(\eta_{ik}|\eta_{i,-k},\mathbf{Z},\mathbf{W},\mathbf{D},\boldsymbol{\tau}) \propto \ell(\eta_{ik}|\eta_{i,-k})\mathcal{N}(\eta_{ik}|\nu_{ik},\sigma_k^2)p(D^*|\boldsymbol{\eta},\boldsymbol{\tau},\mathbf{Z}) \qquad (17)
$$

where

$$
\begin{aligned}
\nu_{ik} &= \mu_k - \Lambda_{kk}^{-1}\boldsymbol{\Lambda}_{k,-k}(\boldsymbol{\eta}_{i,-k} - \boldsymbol{\mu}_{i,-k}) \\
\sigma_k^2 &= \boldsymbol{\Lambda}_{kk}^{-1} \\
\boldsymbol{\Lambda} &= \boldsymbol{\Sigma}^{-1} \qquad (18)
\end{aligned}
$$

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

$$
\begin{aligned}
p(\mathbf{D}^*|\boldsymbol{\eta},\boldsymbol{\tau},\mathbf{Z}) &= \prod_i\prod_p\prod_{j=1}^{i-1}\exp\Big\{-\frac{1}{2}\big(D^*_{ipj} - (\tau_0 + \tau_1\kappa_j^{(i)} + \tau_2\eta_{j,z_{ip}})\big)^2\Big\} \\
&\propto \prod_i\prod_p\prod_{j=1}^{i-1}\exp\Big\{-\frac{1}{2(1/\tau_2^2)}\Big(\eta^2_{j,z_{ip}} - 2\frac{D^*_{ipj}-\tau_0-\tau_1\kappa_j^{(i)}}{\tau_2}\eta_{j,z_{ip}}\Big)\Big\} \\
&\propto \prod_i\prod_p\prod_{j=1}^{i-1}\mathcal{N}(\eta_{j,z_{ip}}|\mu^*_{ipj},\frac{1}{\tau_2^2}) \\
&= \prod_i\prod_p\prod_{j=1}^{i-1}\prod_k\mathcal{N}(\eta_{jk}|\mu^*_{ipj},\frac{1}{\tau_2^2})^{\mathbb{I}(z_{ip}=k)} \qquad (19)
\end{aligned}
$$

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

$$\begin{aligned}
&\prod_k \prod_{s=i+1}^{N} \prod_{p=1}^{N_s} \mathcal{N}(\eta_{ik} | \mu^*_{spi}, \sigma^{2^*})^{\mathbb{I}(z_{sp}=k)} \\
&\equiv \prod_{k=1}^{K} \mathcal{N}(\eta_{ik} | m_{ik}, V_{i,kk})
\end{aligned} \tag{20}$$

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

$$\begin{aligned}
V^{-1}_{i,kk} &= \frac{1}{\sigma^{2^*}} \sum_{s=i+1}^{N} \sum_{p=1}^{N_s} \mathbb{I}(z_{sp} = k) \\
&= \tau_2^2 \sum_{s=i+1}^{N} \sum_{p=1}^{N_s} \mathbb{I}(z_{sp} = k)
\end{aligned} \tag{21}$$

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

$$\begin{aligned}
m_{ik} &= \frac{\tau_2^2 \sum_{s=i+1}^{N} \sum_{p=1}^{N_s} \mu^*_{spi} \mathbb{I}(z_{sp} = k)}{V^{-1}_{i,kk}} \\
&= \frac{\sum_s \sum_p \mu^*_{spi} \mathbb{I}(z_{sp} = k)}{\sum_s \sum_p \mathbb{I}(z_{sp} = k)}
\end{aligned} \tag{22}$$

Then the η conditional is

$$p(\eta_{ik} | \eta_{i,-k}, \mathbf{Z}, \mathbf{W}, \mathbf{D}, \boldsymbol{\tau}) \propto \ell(\eta_{ik} | \eta_{i,-k}) \mathcal{N}(\eta_{ik} | \nu_{ik}, \sigma_k^2) \mathcal{N}(\eta_{ik} | m_{ik}, V_{i,kk}) \tag{23}$$

We now introduce Polya-Gamma augmentation such that

$$\begin{aligned}
p(\eta_{ik} | \eta_{i,-k}, \mathbf{Z}, \mathbf{W}, \mathbf{D}, \boldsymbol{\tau}, \lambda_{ik}) &\propto \exp\{(t_{ik} - \frac{N_i}{2})\rho_{ik} - \frac{\lambda_{ik}}{2}\rho_{ik}^2\} \mathcal{N}(\eta_{ik} | \nu_{ik}, \sigma_k^2) \mathcal{N}(\eta_{ik} | m_{ik}, V_{i,kk}) \\
&\propto \mathcal{N}(\eta_{ik} | \frac{t_{ik} - N_i/2}{\lambda_{ik}} + \log(\sum_{l \neq k} e^{\eta_{il}}), 1/\lambda_{ik}) \mathcal{N}(\eta_{ik} | \nu_{ik}, \sigma_k^2) \mathcal{N}(\eta_{ik} | m_{ik}, V_{i,kk})
\end{aligned} \tag{24}$$

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

$$p(\eta_{ik} | \eta_{i,-k}, \mathbf{Z}, \mathbf{W}, \mathbf{D}, \boldsymbol{\tau}, \lambda_{ik}) \propto \mathcal{N}(\eta_{ik} | \tilde{\mu}_{ik}, \tilde{\sigma}_k^2) \tag{25}$$

where

$$\tilde{\sigma}_k^2 = (\sigma_k^{-2} + \lambda_{ik} + v_{i,kk}^{-1})^{-1}$$
$$\tilde{\mu}_{ik} = \tilde{\sigma}_k^2\big(v_{i,kk}^{-1} m_{ik} + \sigma_k^{-2}\nu_{ik} + t_{ik} - \frac{N_i}{2} + \lambda_{ik}\log(\sum_{l\neq k} e^{\eta_{il}})\big) \tag{26}$$

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 $\boldsymbol{\eta}_i$.

$$p(\lambda_{ik}|\mathbf{Z}, \mathbf{W}, \boldsymbol{\eta}) \propto PG(N_i, \rho_{ik}) \tag{27}$$

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

$$p(D^*_{ipj}|\boldsymbol{\eta}, \mathbf{Z}, \boldsymbol{\tau}, \mathbf{D}) \propto \begin{cases} TN_{(0,\infty)}(\tau_0 + \tau_1\kappa_j^{(i)} + \tau_2\eta_{j,z_{ip}}, 1) & \text{if } D_{ipj} = 1 \\ TN_{(-\infty,0]}(\tau_0 + \tau_1\kappa_j^{(i)} + \tau_2\eta_{j,z_{ip}}, 1) & \text{if } D_{ipj} = 0 \end{cases} \tag{28}$$

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

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

$$\begin{aligned} p(\boldsymbol{\tau}|\boldsymbol{\eta}, \mathbf{Z}, \mathbf{D}^*) &\propto exp\Bigg\{ -\frac{1}{2}\sum_{ipj}\Big(D^*_{ipj} - \mathbf{x}_{ipj}^T\boldsymbol{\tau}\Big)^2\Bigg\} N(\boldsymbol{\mu}_{\boldsymbol{\tau}}, \Sigma_\tau) \\ &\propto N(\tilde{\boldsymbol{\tau}}, \tilde{\Sigma_\tau}) \end{aligned} \tag{29}$$

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

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

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

$$\hat{\Psi}_{kv} = \frac{\sum_i\sum_p\big(\beta_v + \mathbb{I}(z_{ip}^k = 1)\mathrm{W}_{ip,v}\big)}{\sum_i\sum_p\sum_l\big(\beta_l + \mathbb{I}(z_{ip}^k = 1)\mathrm{W}_{ip,l}\big)} \tag{30}$$

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,

$$\begin{aligned} z_{ip}^{(0)} &\sim \text{Categorical}(\hat{\boldsymbol{\theta}}_i) \quad \forall p = 1, 2, ..., N_i \\ \boldsymbol{\eta}_i^{(0)} &= \log(\hat{\boldsymbol{\theta}}_i / \hat{\theta}_{iK}) \end{aligned} \tag{31}$$

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

$$\begin{aligned} \tilde{\tau}_0 &= \frac{1}{2}\log(\text{density}(\mathbf{D})) \\ \tilde{\tau}_1, \tilde{\tau}_2 &\sim \text{unif}(0, 1) \end{aligned} \tag{32}$$

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

$$\begin{aligned} {D^*_{ipj}}^{(0)} &\sim TN_{(-\infty,0)}(\tilde{\tau}_0 + \tilde{\tau}_1\kappa_j^{(i)} + \tilde{\tau}_2\eta^{(0)}_{j,z_{ip}^{(0)}}, 1) \quad \text{if } D_{ipj} = 0 \\ {D^*_{ipj}}^{(0)} &\sim TN_{[0,\infty)}(\tilde{\tau}_0 + \tilde{\tau}_1\kappa_j^{(i)} + \tilde{\tau}_2\eta^{(0)}_{j,z_{ip}^{(0)}}, 1) \quad \text{if } D_{ipj} = 1 \end{aligned} \tag{33}$$

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

$$\boldsymbol{\tau}^{(0)} = (\sum_{ipj} \mathbf{x}^{(0)}_{ipj} {\mathbf{x}^{(0)}_{ipj}}^T)^{-1} (\sum_{ipj} {\mathbf{x}^{(0)}_{ipj}}^T {D^*_{ipj}}^{(0)}) \tag{34}$$

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

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

$$\lambda_i^{(0)} \sim \mathrm{PG}(N_i, \boldsymbol{\eta}_i^{(0)}) \tag{35}$$

D Simulation Results

D.1 MCMC Plots of Key Parameters

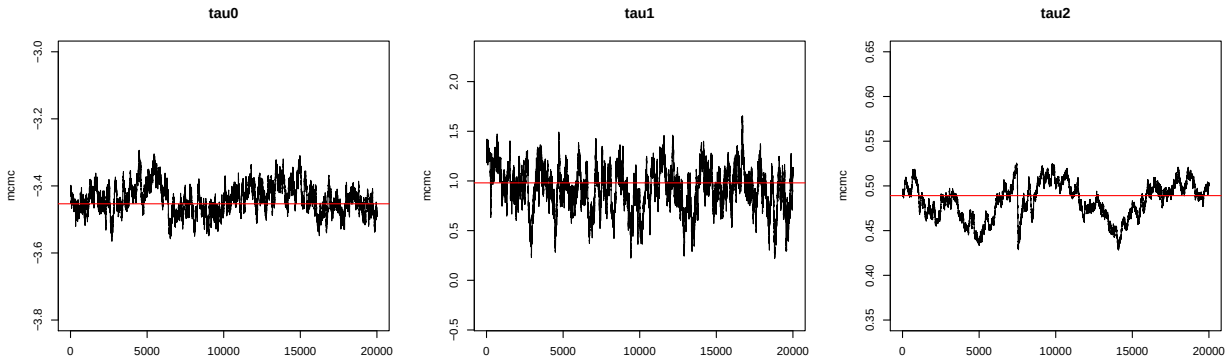


Figure D.1: MCMC convergence of $\boldsymbol{\tau}$ posterior samples in simulation. Horizontal red line indicates the true values of $\boldsymbol{\tau}$.

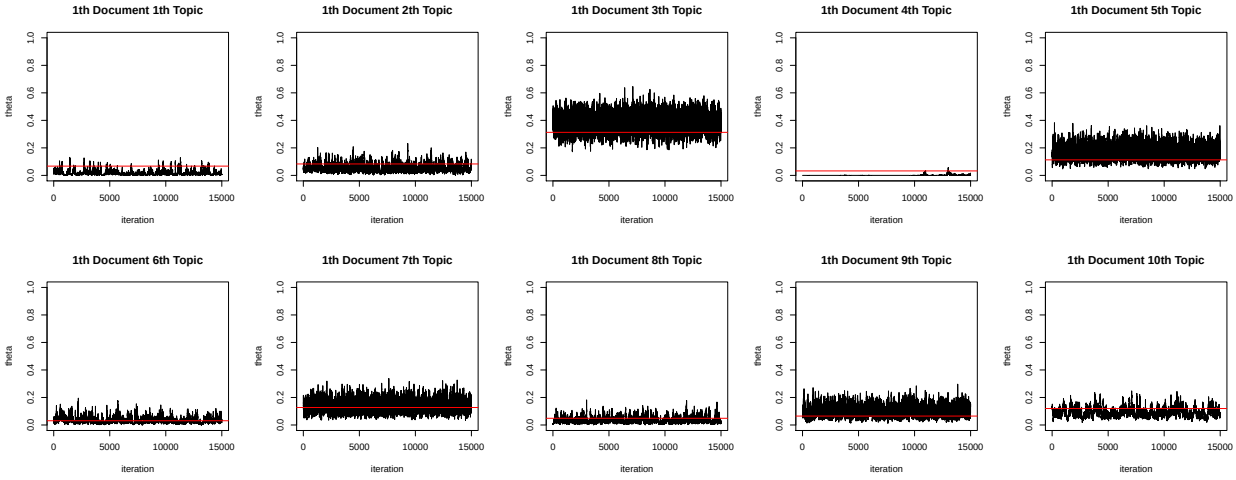


Figure D.2: MCMC convergence of $\boldsymbol{\theta}$ parameters for the first document. $\boldsymbol{\theta}$ values are obtained by transforming the posterior samples of $\boldsymbol{\eta}$ of the corresponding document. Horizontal red line indicates the true values of $\boldsymbol{\theta}$ for the 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 $\boldsymbol{\theta}$.

D.2 Recovery of the True Latent Variables

We generate 100 simulation datasets with similar sizes as our application datasets. Specifically, we set the simulation datasets to have about equal number of documents, paragraphs,

unique words and words.[2] Citations are generated based on the hyperparameters we input, and we set them so that the number of citations will be similar to those in our application data. This exercise gives us some evidence on the validity of our results on the application datasets.

We show that the PCTM can recover the true parameters from random initialization using our Gibbs sampler. We fit the PCTM on one of the simulation datasets while the initial parameters of the paragraph topic, $\mathbf{Z}$, and the distribution of topics, $\boldsymbol{\eta}$, are randomly initialized. Then, we compare the estimated paragraph topics and the distribution of topics with the true values of those parameters.

Figure D.3 plots the posterior samples of paragraph topics against the true paragraph topics. Numbers on the x-axis and y-axis denote topic labels. The darkness of cell colors is proportional to the number of paragraphs in those cells. The cell in the second row and the third column, for example, denotes the number of paragraphs that are assigned topic 2 in posterior samples when the true topic is 3. Darker colors on the diagonal lines suggest that the model recovers true topics correctly, which we see on the right panel of Figure D.3. In comparison, the left panel of Figure D.3 illustrates that the Gibbs sampler was initiated with randomly generated values of paragraph topics.

We conduct a similar exercise with the document-level topic mixture $\boldsymbol{\eta}$. To make the comparison more rooted in conventional topic models, we convert $\boldsymbol{\eta}$ to $\boldsymbol{\theta}$ using softmax in this exercise. In Figure D.4, we plot the mode of posterior samples of $\boldsymbol{\theta}$ against the mode of the true topic mixture. The darker colors indicate a higher number of documents in the corresponding cell. Similar to Figure D.3, we observe evenly spread colors on the left panel as opposed to the concentrated dark colors on the diagonal entries on the right panel. This shows that the PCTM recovers

These two results verify that the PCTM can recover true topics from random initialization when applied to simulation data. This adds to the credibility of the topic estimations in our application since our simulation data resembles our application data.

[2] 106 documents, an average of 44 paragraphs per document, 5838 unique words, and an average of 51 words per paragraph.

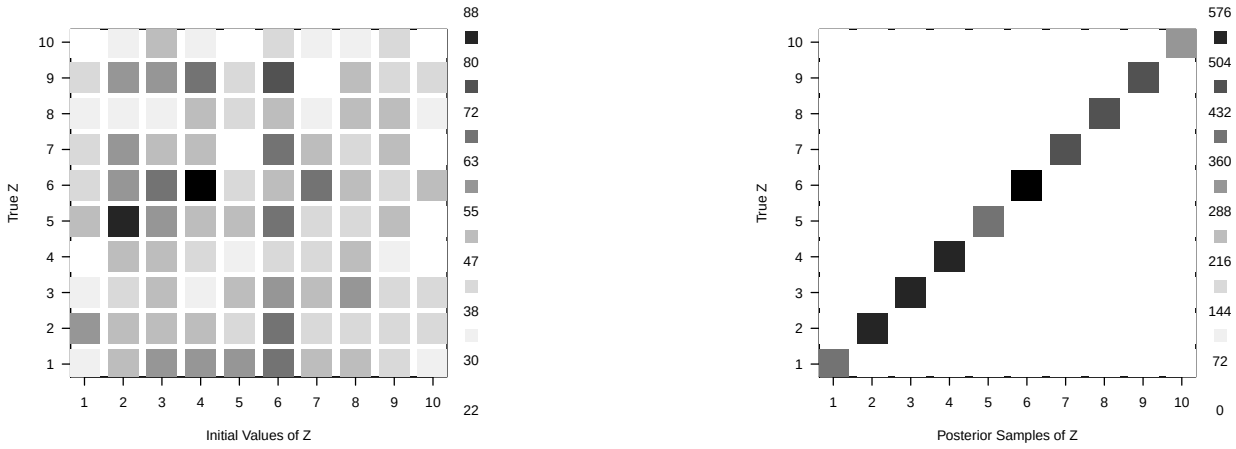


Figure D.3: The comparison of the estimated and the true topics of paragraphs. On the right panel, the (k, l) cell shows the number of paragraphs whose estimated topic is l while the true topic is k. We estimate topics using the paragraph topic parameter, $\mathbf{Z}$, using the last draw from our Gibbs sampler. The cells with darker colors indicate a higher number of paragraphs. The concentration on the diagonal elements means that the topics are estimated correctly. As a comparison, the left panel plots randomly initialized paragraph topics against true paragraph topics. They show that the PCTM can recover the true topics even when the topics are randomly provided at the initialization of our Gibbs sampler.

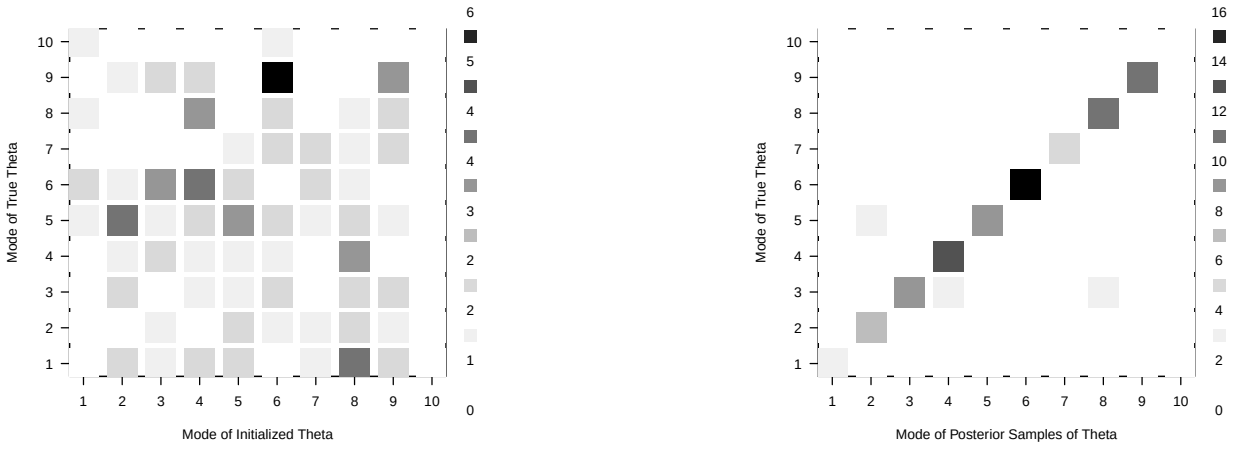


Figure D.4: The comparison of the estimated and the true topic distribution of documents. On the right panel, the (k, l) cell shows the number of documents whose mode of the estimated topic distribution, $\boldsymbol{\theta}$, across K topics is l while the mode of the true topic distribution is k. We obtain $\boldsymbol{\theta}$ by applying the softmax transformation on each draw of $\boldsymbol{\eta}$ in our Gibbs sampler, and then obtain the estimated $\boldsymbol{\theta}$ by their posterior mean. The cells with darker colors mean a higher number of documents are in the cell. The concentration on the diagonal elements means that the modes of the topic distributions are estimated correctly. As a comparison, the left panel plots the mode of randomly initialized $\boldsymbol{\theta}$ against true mode of $\boldsymbol{\theta}$. It shows that the PCTM can recover the true mode of the topic distribution even when the topics are randomly provided at the initialization of our Gibbs sampler.

E Results on the SCOTUS cases on Voting Rights

The SCOTUS documents and citations on voting rights proliferated exponentially since the enactment of Voting Rights Act (VRA) in 1965. A number of sections in VRA were challenged over the course of modern American political history, and the majority of those challenges made their way to the Supreme Court. The Supreme Court database assigns 3 issue codes for opinions related to voting.[3] After examining a subset of documents with these issue codes, we decided to set the number of topics to 4 for PCTM.

Table E.1 presents the 10 words that appear most frequently for each topic. The first

Topic Label	Voter Eligibility	Ballot Access	Preclearance Requirement	Voter Dilution
1	counti	ballot	chang	plan
2	resid	primari	attorney	minor
3	appel	polit	preclear	black
4	school	offic	counti	major
5	properti	counti	practic	polit
6	citi	file	procedur	popul
7	tax	interest	cover	racial
8	board	independ	plan	member
9	citizen	nomin	section	dilut
10	test	burden	object	white

Table E.1: Top 10 words of highest probability for each topic from PCTM.

topic `Voter Eligibility` includes paragraphs that address conditions under which a voter is eligible to register for certain elections. For example, `Allen et al. v. State Board of Elections et al.` (1969) contains a paragraph of the first topic that discusses whether a 31-year-old man was eligible to cast his vote in a local school district election based on his tax records and property ownership in the neighborhood. The second topic `Ballot Access` concerns the issue of candidates' access to ballots. A paragraph of this topic in `Carrington v. Rash et al.` (1965) states that "... the Texas system creates barriers to candidate access to the primary ballot, thereby tending to limit the field of candidates from which voters might choose." Preclearance requirement in Voting Rights Act of 1965 section 5. is the primary issue in the third topic. `Cipriano v. City of Houma et al.` (1969) contains a paragraph of this topic that stipulates "... and unless and until the court enters such judgment no person shall be denied the right to vote for failure to comply with such qualification, prerequisite, standard, practice, or procedure: Provided, That such qualification, prerequisite, standard, practice, or procedure may be enforced without such proceeding if

[3]The three issue codes on voting are voting, Voting Rights Act of 1965, Ballot Access.

the' qualification, prerequisite, standard, practice, or procedure has been submitted by the chief legal officer or other appropriate official ..." The fourth topic, on the other hand, addresses Voting Rights Act of 1965, section 2 that prohibits voting practices that leads to dilution of voting strength of minority groups. For example, `Mcdonald et al. v. Board of Election Commissioners of Chicago et al.` (1969) contains multiple paragraphs of this topic one of which states that "... the Court upheld a constitutional challenge by Negroes and Mexican-Americans to parts of a legislative reapportionment plan adopted by the State of Texas ... ."

The 4 topics that PCTM identified have varying presence in American political history over time. Figure E.1 shows the cumulative count of paragraphs of each topic. The growth

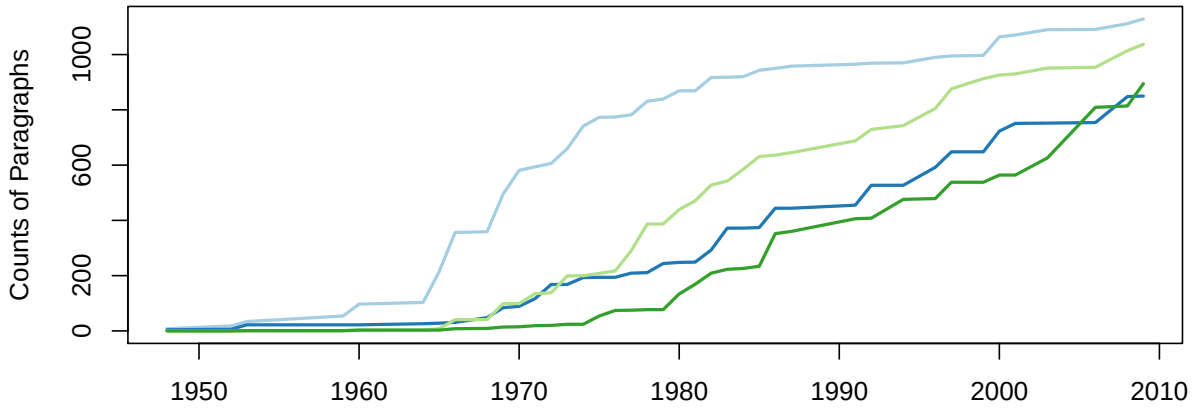
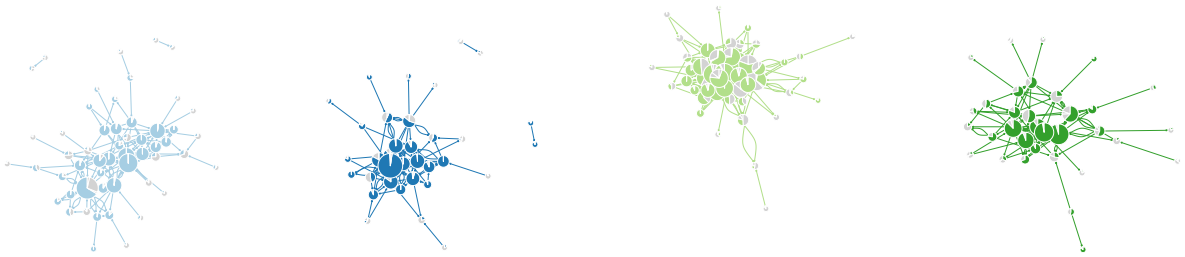


Figure E.1: Cumulative number of topics in Voting Rights subset over time.

of `Voter Eligibility` topic (in light blue) is most evident until the 1980s and the topics on `Preclearance Requirement` (in light green) or `Voter Dilution` (in dark green) become more prevalent in relatively recent periods. This is consistent with Ansolabehere and Snyder (2008) that describes that discourses on malapportionment was more common in earlier periods, and the topics on equal representation and access to vote, especially with respect to race and minority groups, are becoming more prominent issues in modern American politics.

Figure E.2 shows groups of cases that make citations of the given topic. The location of cases on each network is based on their connection patterns such that cases that cite other cases jointly are placed closer to each other. The majority of cases in the third and the fourth panel are located very close to each other, indicating that those cases heavily cite each



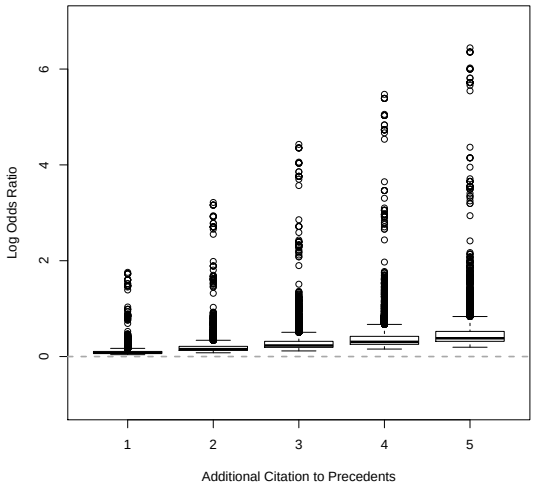
(a) Voter Eligibility (b) Ballot Access (c) Preclearance Requirement (d) Voter Dilution

Figure E.2: The subnetwork specific to each topic. The subnetworks are created by extracting opinions that either send or receive citations of the given topic. The topic-specific subnetworks can be useful in revealing whether and the extent to which topological features of the network varies by topic. For each subnetwork, paragraphs of other topics are all colored in gray for better visualization.

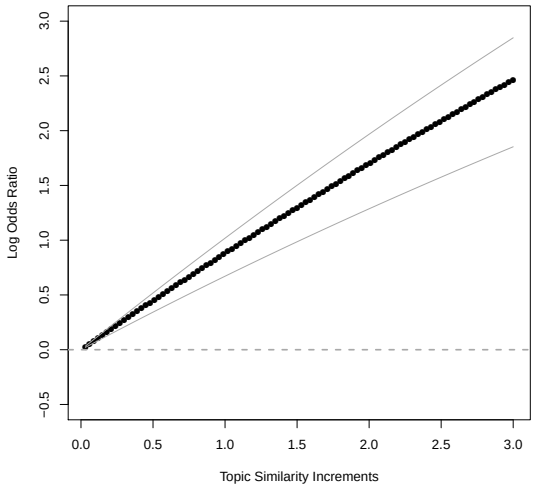
other. On the other hand, the citation subnetwork in the first panel (`Voter Eligibility`) is more spread out in comparison. This reflects the fact that opinions on `Preclearance Requirement` and `Voter Dilution` have proliferated in a shorter period of time, closely building up on past cases of the same topic whereas opinions on `Voter Eligibility` have expanded more independently and incrementally over a longer period of time.

The coefficients in the latent citation propensity for Voting subset also have expected signs, with posterior samples of τ_1 and τ_2 both staying above 0. That is, for the citation decisions of opinions for Voting, the authority as well as the topic similarity of precedents have positive impacts. Moreover, the distribution of all $\boldsymbol{\tau}$ entries stays very similar between the Privacy and the Voting subset, indicating that the citation dynamics do not vary much between different issue areas within the SCOTUS

Similar to the exercise to create Figure F.1, 10,000 randomly drawn pairs of paragraphs and precedents for the Voting subset were used to generate Figure E.3. The left panel of Figure E.3 presents the improvements in the log odds ratio as we increment the authority of the given precedent. For example, if the given precedent had 3 more citations, the odds of the given paragraph citing the given precedent increases by about 25%. The right panel shows changes in log odds ratio as the topic similarity between the given precedent and the given paragraph increases.



(a) Change in Log Odds Ratio by Additional Citations to Precedents

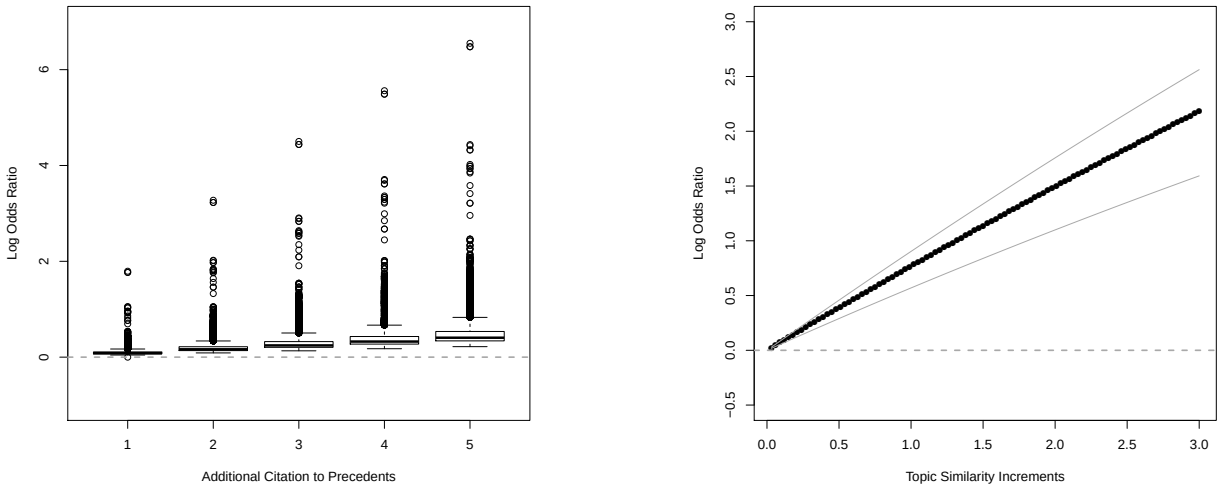


(b) Change in Log Odds Ratio by Increases in $\eta_{j,z_{ip}}$

Figure E.3: Changes in the log odds ratio of citation between a paragraph and a precedent as we increment the authority and the topic similarity of the given precedent. Same exercise used in Figure F.1b is employed to create this figure.

F More Results on the SCOTUS cases on Privacy

F.1 Influence of the In-degree and Topic Similarity on the Probability of Citation



(a) Change in Log Odds Ratio by Additional Citations to Precedents

(b) Change in Log Odds Ratio by Increases in $\eta_{j,z_{ip}}$

Figure F.1: Changes in the log odds ratio of citation between a paragraph and a precedent as we increment the authority and the topic similarity of the given precedent. 10,000 random pairs of paragraphs and precedents were drawn from the data to create this figure. The left panel displays the distribution of improvements in log odds ratio if the given precedent had given additional citations. Each point is one of the 10,000 randomly drawn paragraph-precedent pairs. The right panel shows the improvements in log odds ratio if the given precedent were more topically similar to the given paragraph. The black points represent the average improvements in log odds ratio, and gray lines indicate the 2.5% and 97.5% quantile of log odds improvements respectively.

The $\boldsymbol{\tau}$ coefficients in the latent citation propensity have expected signs. The average value of posterior samples for τ_1 is 0.7 and the 95% credible interval does not include 0, which suggests that the authority of documents has a positive impact on citation likelihood given topics. Similarly, posterior samples for τ_2 stays above 0, suggesting that topic similarity between precedents and the citing paragraphs has a positive impact on citation decisions.

In Figure F.1 we offer one way to interpret coefficients $\boldsymbol{\tau}$ in latent citation propensity.[4] Since the latent citation propensity follows the structure of probit regression, one can employ the conventional approach to interpreting the coefficients where we calculate improvements

[4]For more detailed information on the posterior samples of $\boldsymbol{\tau}$, see Supplementary Information E.

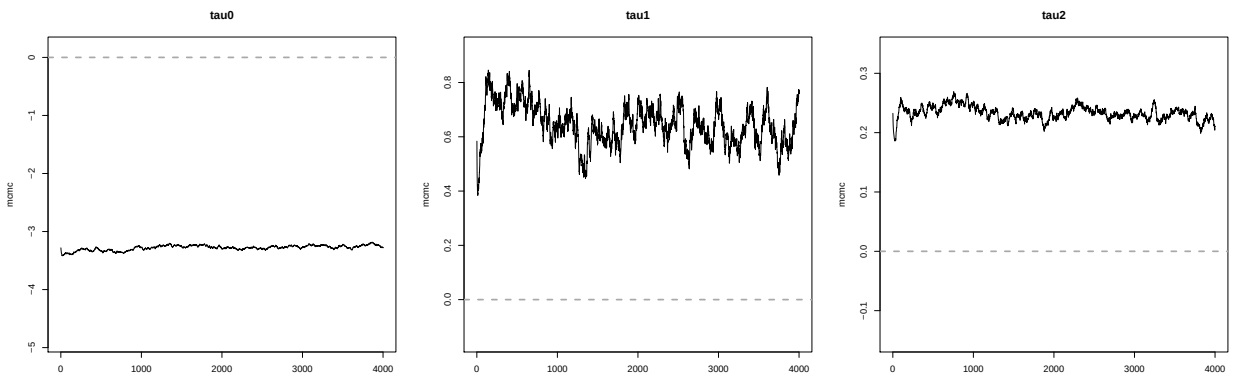


Figure F.2: MCMC convergence of $\boldsymbol{\tau}$ posterior samples for the SCOTUS application on Privacy issue area. Horizontal red line indicates the true values of $\boldsymbol{\tau}$.

in predicted probability as we increment one predictor while fixing other predictors at their means. This approach, however, presents two potential challenges. First, citation networks are usually sparse. Under our modeling framework, the sparse feature of citation networks is more emphasized as paragraphs are the unit that makes citations. The citation network for the Privacy subset contains only 452 citations when the fully connected network would have 243,685 citations. Partly due to such sparsity, improvements in predicted probability can be highly marginal. Second, the authority of a precedent, or the indegree, is known to follow the power-law distribution which is highly skewed to the right (Eom and Fortunato, 2011). When a distribution is highly skewed, the mean is less likely to be the representative value of the distribution.

To address the above two challenges, we examine improvements in log odds ratio rather than predicted probability. Additionally, when incrementing one predictor we follow Hanmer and Ozan Kalkan (2013) and use observed values of other predictors rather than their means. To create Figure F.1 we randomly sampled 10,000 paragraph-precedent pairs from the subset data and computed the extent of improvements in log odds ratio as we increased the authority and topic similarity of the given precedent. The left panel presents the improvements in log odds ratio when the authority of the given precedent is incremented. For example, if the given precedent had 3 more citations, the odds of the given paragraph citing the given precedent increases by about 20%. Similarly, the right panel displays improvements in log odds ratio as the given precedent becomes more topically similar ($\eta_{j,z_{ip}}$) to the given paragraph.

F.2 MCMC Convergence Diagnostics

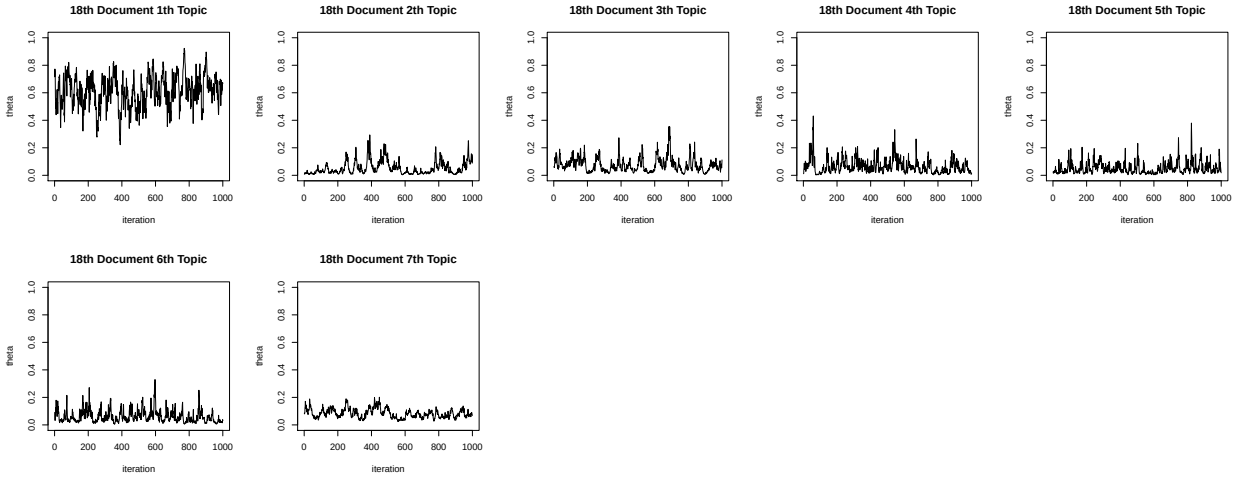


Figure F.3: MCMC convergence of $\boldsymbol{\theta}$ parameters for the 18th document in the subset of Privacy issue area. $\boldsymbol{\theta}$ values are obtained by transforming the posterior samples of $\boldsymbol{\eta}$ of the corresponding document. Horizontal red line indicates the true values 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 $\boldsymbol{\theta}$.

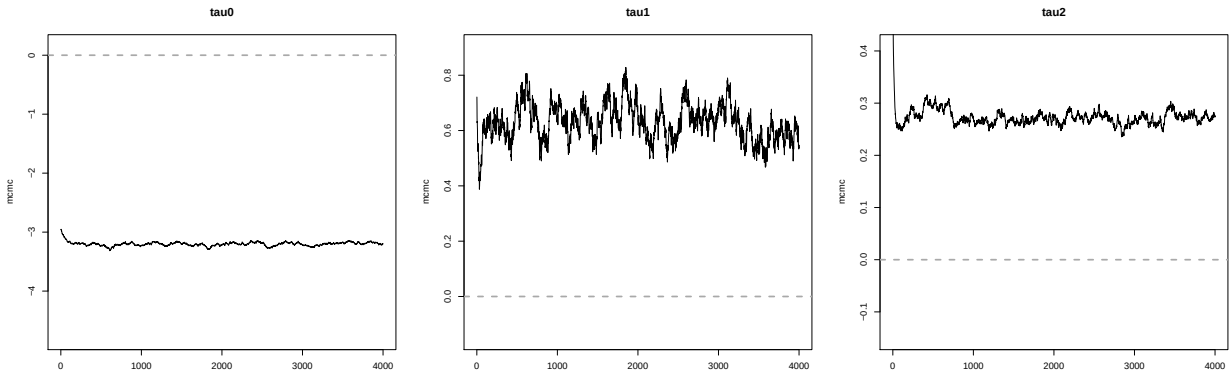


Figure F.4: MCMC convergence of $\boldsymbol{\tau}$ posterior samples for the SCOTUS application on Voting Rights issue area. Horizontal red line indicates the true values of $\boldsymbol{\tau}$.

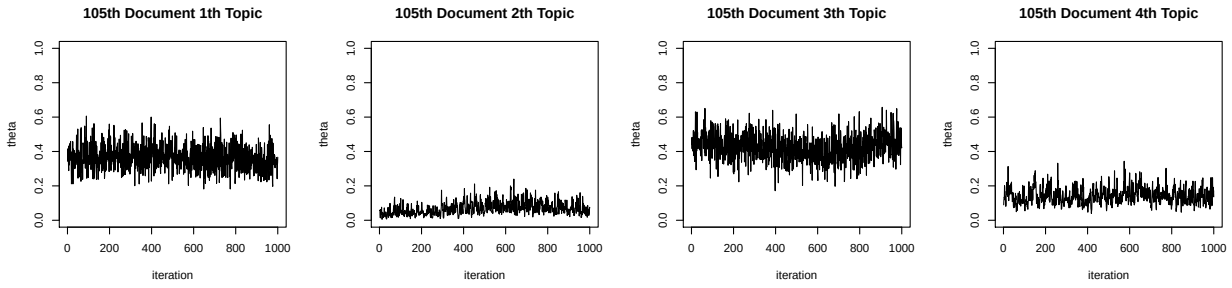


Figure F.5: MCMC convergence of $\boldsymbol{\theta}$ parameters for the 105th document in the subset of Voting Rights issue area. $\boldsymbol{\theta}$ values are obtained by transforming the posterior samples of $\boldsymbol{\eta}$ of the corresponding document. Horizontal red line indicates the true 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 $\boldsymbol{\theta}$.

G Comparison of the Predictive Performance against Existing Methods

In this section, we compare the predictive performance of the PCTM against two alternative models for document networks: the RTM and the LDA combined with Logistic Regression (LDA + Logistics). In both alternative models, citations arise as a function of topic similarity at the word level. We use documents in the Privacy subset for this exercise. We choose paragraphs in `Gonzales v. Carhart` as our test set because `Gonzales v. Carhart` contains a sufficiently large number of citations and words to demonstrate how they contribute to the predictive performance.[5] We discard documents temporally later than `Gonzales v. Carhart`.

Our exercise is essentially a leave-one-out cross-validation for each paragraph in `Gonzales v. Carhart`. Specifically, we take a paragraph in `Gonzales v. Carhart` as test data, and all other paragraphs in `Gonzales v. Carhart` and documents prior to it are assigned to the training data. Then we compute the predictive probability on the paragraph in the test set given our parameters fit on the training data. Note that due to the structure of this exercise, `Gonzales v. Carhart` will appear in both the training set and the test set. The above exercise is repeated for all 88 paragraphs in `Gonzales v. Carhart`.

One challenge in this exercise is that the PCTM assigns topics to each paragraph while the RTM and the LDA assign topics to each word. That is, the RTM and the LDA do not recognize paragraphs in the data. Therefore, we treat the paragraph in the test data as if it is a new version of `Gonzales v. Carhart` when we compute the predictive probability in the RTM and the LDA. In other words, we estimate topics of words in the test data from the topic probability for `Gonzales v. Carhart`.

A formal description of the prediction exercise is as follows. $\mathbf{W}_{iq}$ and $\mathbf{D}_{iq}$ are the data in a paragraph q of a document i. This corresponds to the test paragraph. $\mathbf{W}^{train}, \mathbf{D}^{train}$ are the data in the training set. This includes all paragraphs other than the paragraph q of the document i as well as all documents prior to the document i. The parameters with $\hat{\cdot}$ symbol indicate that they are estimates based on the training data. The following gives the posterior predictive probability for the PCTM.

[5] `Gonzales v. Carhart` contains 12 citations, which is about 94 percentile of the distribution of the number of citations per document. It is the 9th latest document in our corpus.

$$
\begin{aligned}
&p(\mathbf{W}_{iq}, \mathbf{D}_{iq} | \mathbf{W}^{train}, \mathbf{D}^{train}) \\
&= \sum_{k=1}^{K} \Big\{ p(\mathbf{W}_{iq} | z_{iq} = k, \hat{\mathbf{\Psi}}) \\
&\quad \times \prod_{j=1}^{i-1} \mathbb{P}(D^*_{iqj} > 0 | \hat{\boldsymbol{\tau}}, \hat{\boldsymbol{\eta}}, z_{iq} = k)^{\mathbb{I}\{D_{iqj}=1\}} \mathbb{P}(D^*_{iqj} < 0 | \hat{\boldsymbol{\tau}}, \hat{\boldsymbol{\eta}}, z_{iq} = k)^{\mathbb{I}\{D_{iqj}=0\}} \\
&\quad \times p(z_{iq} = k | \hat{\boldsymbol{\eta}}_i) \Big\}
\end{aligned}
\tag{36}
$$

By contrast, the following gives the posterior predictive probability for the RTM and LDA. We follow Chang and Blei (2009) for the notation of parameters. $\boldsymbol{\theta}$ is a $N \times K$ document-topic matrix. $\boldsymbol{\eta}$ is a K-length vector of coefficient and ν is intercept in the regression of citation on the topic.

$$
\begin{aligned}
&p(\mathbf{W}_{iq}, \mathbf{D}_{iq} | \mathbf{W}^{train}, \mathbf{D}^{train}) \\
&= \sum_{\mathbf{z}} \Big\{ p(\mathbf{W}_i | \mathbf{Z}_{iq} = \mathbf{z}, \hat{\mathbf{\Psi}}) \\
&\quad \times \prod_{j=1}^{i-1} \left[\psi \left(\hat{\boldsymbol{\eta}} (\bar{\mathbf{Z}_{iq}} \circ \bar{\hat{\mathbf{Z}}_j}) + \hat{\nu} \right) \right]^{\mathbb{I}\{D_{iqj}=1\}} \left[1 - \psi \left(\hat{\boldsymbol{\eta}} (\bar{\mathbf{Z}_{iq}} \circ \bar{\hat{\mathbf{Z}}_j}) + \hat{\nu} \right) \right]^{\mathbb{I}\{D_{iqj}=0\}} \\
&\quad \times p(\mathbf{Z_{iq}} = \mathbf{z} | \hat{\boldsymbol{\theta}}_i) \Big\}
\end{aligned}
\tag{37}
$$

Note that $\mathbf{Z}_{iq}$ is a vector with its length equal to the number of words in the test paragraph. Since it is infeasible to compute all possible values of $\mathbf{Z}_{iq}$, we use Monte Carlo simulation to approximate its distribution. For LDA+Logistic model, the parameters are estimated by fitting LDA on the training data and then regressing the citation on the topics.

The results are displayed in Figure G.1. Each symbol represents the difference in the log posterior probability between models for each paragraph. The left panel compares the PCTM with the RTM and the right panel compares it with the LDA+Logistic regression. Solid symbols denote the differences in the predictive probabilities for paragraphs without citations and hollow symbols are for ones with citations. The main takeaway is that the PCTM almost always outperforms the other two models. In particular, the improvement in predictive probability becomes greater when the prediction is made on paragraphs with more words. One explanation for this is that the PCTM suffers less from overfitting than the RTM or the LDA does with respect to predictions. Since the RTM and the LDA assign topic parameters to each word, the model complexity for both models increases exponentially as

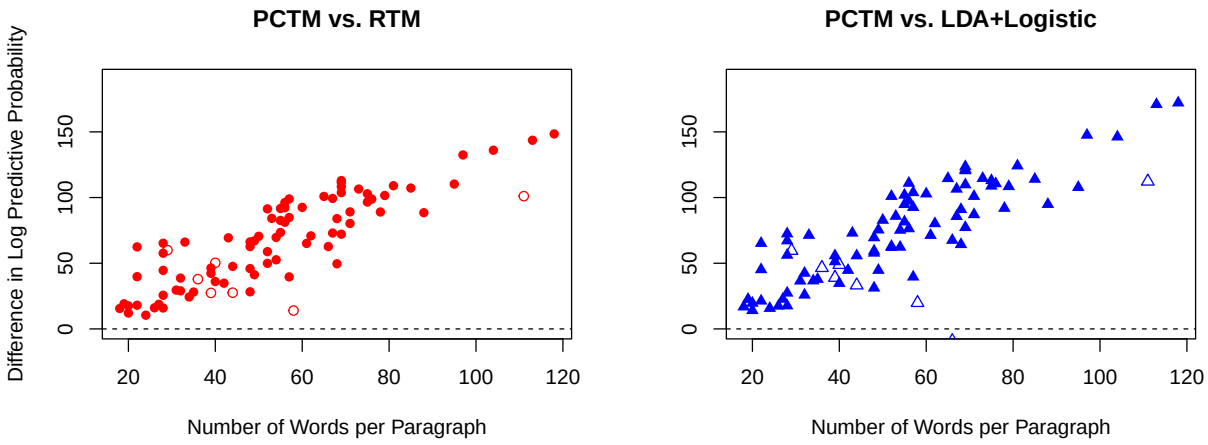


Figure G.1: Difference in Predicted Probability with PCTM, RTM, and LDA + Logistic Regression. The x-axis is the number of words per paragraph. The y-axis is the difference in the log posterior probability between PCTM and other models. The compared models are RTM for the left panel and LDA + Logistic regression for the right panel. Each symbol represents the difference in the log posterior probability between models for each paragraph. Solid symbols are paragraphs without citations and hollow symbols are with citations. The prediction was performed by first fitting the models on a subset of the corpus temporally prior to the test paragraph, and then computing the predictive probability of the test paragraph as if the test paragraph is a new paragraph in the last document of the training corpus. R package `lda` was used to fit the RTM and the LDA. Overall, the PCTM achieves higher posterior predictive probability compared to the RTM and the LDA + Logistic Regression models, particularly when a paragraph contains many words.

more words are included in the document. For the PCTM, on the other hand, increasing the number of words in paragraphs does not significantly impact the model complexity because the topic parameter is for paragraphs, not words.

H 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{38}
$$

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

Ansolabehere, S. and Snyder, J. M. (2008). *The end of inequality: One person, one vote and the transformation of American politics*. WW Norton & Company Incorporated.

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

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., and Matsuo, A. (2018). quanteda: An r package for the quantitative analysis of textual data. *Journal of Open Source Software*, 3(30):774.

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

Caselaw Access Project (2024). Caselaw access project.

Chang, J. and Blei, D. (2009). Relational topic models for document networks. In *Artificial intelligence and statistics*, pages 81–88. PMLR.

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.

Clark, T. S. and Lauderdale, B. E. (2012). The genealogy of law. *Political Analysis*, 20(3):329–350.

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.

Eom, Y.-H. and Fortunato, S. (2011). Characterizing and modeling citation dynamics. *PloS one*, 6(9):e24926.

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.

Hanmer, M. J. and Ozan Kalkan, K. (2013). Behind the curve: Clarifying the best approach to calculating predicted probabilities and marginal effects from limited dependent variable models. *American Journal of Political Science*, 57(1):263–277.

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.