Paragraph-citation Topic Model 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

Abstract

Topic modeling is one of the most popular approaches to statistical text analysis in many fields, especially in the social sciences. An important feature of text data in social sciences is that many corpora consist of document networks in which documents cite other documents. However, existing topic models either ignore the network structure or make simplifying assumptions that do not reflect the structural properties of actual citation networks. In this paper, we propose a topic model that jointly analyzes both text and citations. In the proposed paragraph-citation topic model(PCTM), topics are assigned to paragraphs rather than tokens. The topic of a paragraph then shapes both the distribution of words and the likelihood of citations emanating from that paragraph to other documents. To model the likelihood of citations to other documents, we introduce a latent citation propensity variable that incorporates two stylized facts about citation networks: the authority and the topic similarity of the documents. We demonstrate the utility of our model by applying it to two subsets of majority opinions of the Supreme Court of the United States: all opinions on Privacy and Voting Rights issues. We then use the results of the first application on all Supreme Court opinions on Privacy issues to predict the topic structure of the most recent case on reproductive rights, `Dobbs v. Jackson Women's Health Organization`.

[*] 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`.

1 Introduction

Topic models are widely used to explore semantic context from a large corpus in an unsupervised way. These models, such as Latent Dirichlet Allocation (LDA) (Blei et al., 2003) and its extensions (Blei and Lafferty, 2006, 2007; Roberts et al., 2014), discover latent clusters of words that share semantic meanings from the co-occurrence of words across documents.

One common assumption in topic models is that documents have no explicit connections. However, we often find a reference structure between documents in text data (i.e. citations in legal opinions, citations in journal articles, retweets in twitter, etc.). For example, LDA is employed to analyze Twitter data without taking into account the fact that many of the tweets are retweets of previous tweets (Barry et al., 2018; Yang and Zhang, 2018; Sommer et al., 2012; Hidayatullah and Ma'arif, 2017). Likewise, topic analyses of academic journals in Daenekindt and Huisman (2020), De Battisti et al. (2015) and Blei and Lafferty (2007) are conducted disregarding citations that reference other articles. Rice (2017) uses LDA to analyze opinions in the Supreme Court of the United States(SCOTUS) and demonstrate that justices have strategic motivations to alter topic contents in dissenting opinions from topics of majority opinions. Similar to other studies in above, the citation network in SCOTUS was not utilized in Rice (2017). While texts offer rich information on the topic structure of documents, citations in document networks can also help discover topics better as citations are more likely to occur between documents of similar semantic context. In this sense, we can infer that documents connected with citations are likely to be addressing similar topics, compared to documents that are completely disconnected.

In a similar vein, existing studies of network analysis on citation networks tend to not utilize texts. For instance, Fowler et al. (2007) constructs a document-level score of legal importance in the citation network of the SCOTUS opinions from 1791 to 2005. This study ranks 26,681 majority opinions of the SCOTUS according to their legal importance scores. While the legal importance score offers a useful summary of the legal value of a given case, the scores can be more meaningful by addressing the semantic or issue differences between opinions. For example, opinions on voting rights are unlikely to cite opinions on search and seizure simply because they pertain to different legal domains. However, since there are more opinions and citations on search and seizure, they can be ranked higher than opinions on voting rights in terms of legal importance. That is, not addressing semantic context can result in a misleading conclusion of considering a legally important opinion in voting rights as less valuable than a legally important opinion in search and seizure. Other studies use hand-labeled issue codes to obtain semantic coherence. Clark and Lauderdale (2012) focuses on the reproductive rights opinions of the SCOTUS and constructs a "family tree of law" that succinctly summarizes the historical development of legal doctrines. Lupu and Voeten (2012) analyzes the citation network of the European Court of Human Rights and examines how international courts justify their rulings in the areas of human rights. While human coding can provide semantic consistency at a broader level, utilizing texts in documents allows researchers to obtain semantic contexts refined at the level of their research purposes. In our application, our model identifies 7 distinct topics for opinions on privacy issues and 4 topics for opinions on voting rights.

In this paper, we propose a new topic model that allows researchers to analyze text and citations jointly. Our proposed model, which we call the paragraph-citation topic model (PCTM), augments a topic model by adding a latent citation propensity to model network processes of citation formation. Because the PCTM assigns topics to not only text but also citation links, topic assignments in a document can be informed by those in other documents through citations.

A key feature of the PCTM is that it uses paragraphs, instead of tokens or words, as the unit of topic assignment. Substantively, this is a reasonable assumption when each paragraph contains a coherent topic, as is the case for carefully written documents with lengths such as legal opinions and academic papers. Technically, this assumption enables the model to attach words to citations through a common topic. In existing topic models for document networks, citations are conditioned directly on the document-level topic mixture (Chang and Blei, 2009), or citation topics drawn from the topic mixture are independent of words (Nallapati et al., 2008). As a result, these models are unable to estimate which words are associated with the context from which a citation arises. By contrast, since the PCTM assumes that words and citations within the same paragraph are generated from a common topic, not only the words within the same paragraph but also the words in other paragraphs of the same topic provide information on the semantic context of a citation. In addition, by uncovering the heterogeneity of topics across paragraphs, the PCTM allows us to identify if and to what extent citations within the same document are in different semantic contexts. These advantages are demonstrated in our applications.

We use a corpus of the SCOTUS's majority opinions for the application in this paper. While we believe that the PCTM is applicable to many other datasets, the SCOTUS decisions have been extensively studied in political science using either text analysis or network analysis. Compared to these existing studies, the PCTM's distinct feature poses a unique challenge in constructing a dataset, because the data need to record the location of citations. The format of citations in the SCOTUS decisions allows us to overcome this challenge with string matching.

The remainder of the paper is organized as follows. Section 2 introduces a new dataset of text and citation network of the SCOTUS opinions we constructed for this project. Section 3 describes our model, the PCTM, and its inference. Section 4 shows simulation results to verify the performance of the PCTM. Section 5 presents results of applying the PCTM on the SCOTUS opinions on privacy and voting rights. We conclude and provide remarks on future research in Section 6.

2 Application: The United States Supreme Court Opinions

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[1], which allows public access to all official and published opinions at all levels of the US courts. 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. See Section 3 for more information about our proposed model. After tokenizing the corpus, we removed punctuations, symbols, special characters, numbers, and common English stopwords.[2] 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 our model incorporates paragraph structures. In other words, we want to form an adjacency matrix of $NP \times N$ where NP is the number of paragraphs and N is the number of documents. The (ip, j) element of the matrix is 1 if

[1] https://case.law

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

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, because the same case is recorded by multiple reporters, there are several possible ways to cite the same case. 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 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 two subsets of this dataset for our applications. We subset documents by their issue areas defined by Supreme Court Database (Spaeth et al., 2020) and perform further pruning based on their frequencies within and across documents. This is because words that are specific to a subset of documents in the entire corpus may turn out to be common terms in an issue area. For example, the word "taxation" is not commonly shared by documents in the entire corpus, but it appears in almost all documents of *Federal Taxation* issue area. More details of data pre-processing for each subset are available in the Supplementary Information document, section A. The first subset consists of cases classified as *Privacy* issue area, which includes decisions about abortion, public disclosure of private information and etc. 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. The subset about *Privacy* consists of 106 documents with 4,669 paragraphs, 5,838 unique words, and 452 citations. The other subset is cases dedicated to issue codes on *Voting Rights*. The subset on *Voting Rights* consists of 105 documents with 3,911 paragraphs, 3,836 unique words, and 618 citations.

3 The Proposed Model

Our proposed model is built on a topic model, a popular model to discover latent clusters or "topics" of documents (Blei et al., 2003; Blei and Lafferty, 2007). A topic model that analyzes documents with citation networks must address the following questions. First, what gives rise to citations? In other words, by what process do authors of a document decide to cite another document? Second, how does the topic structure enter into citation decisions? and how do citations shape topic structure of citing and cited documents?

To address these questions, we augment a topic model by latent citation propensity to model authors' decisions to make citations in relation to the topic structure. The latent citation propensity is shaped by a regression model that reflects the known factors of strategic citation behavior such as the authority (or popularity) of the cited document (Larsson et al., 2017; Lupu and Voeten, 2012; Lupu and Fowler, 2013; Pelc, 2014) as well as the similarity of topics between citing and cited documents.

Additionally, we propose to use paragraphs as the unit for the topic assignment. The Relational Topic Model(RTM) by Chang and Blei (2009) views citations as undirected connection between documents with similar topic mixture. We view citations as the directed reference from a paragraph to another document. The advantage of this perspective is that it reflects a more realistic data-generating process. A paragraph is often the vehicle of one coherent topic, and citations within that paragraph are likely to be referring to documents of very similar, if not the same, topic prevalence. For example, an opinion in the SCOTUS typically identifies multiple legal doctrines that apply to a given case and addresses them in different paragraphs. Therefore, citations within one paragraph are likely to be pointing to a collection of opinions that address the same legal doctrine. In other words, citations in paragraphs of different topics are likely to be references to different legal contexts, even if they are from the same document. We believe such characteristics are not limited to legal documents of the SCOTUS, but a general feature of any document network, and they should be reflected in the process of uncovering topic structure. Below, we delineate our modeling strategy that addresses the above questions in detail.

3.1 Paragraph-citation Topic Model

Let N, NP and V be the total number of documents, total number of paragraphs and total number of unique words respectively. D_{ipj} is a binary indicator that denotes the existence of a citation from pth paragraph in ith document towards jth document. The data generating process is modeled as follows.

```
For each document $i$
  Draw topic proportion $\boldsymbol{\eta}_i \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
  For each paragraph $p$:
      Draw topic assignment $z_{ip} \sim \text{Mult}(1, \text{softmax}(\boldsymbol{\eta}_i))$
      Draw word $w_{ip} \sim \text{Mult}(N_{ip}, \boldsymbol{\Psi}_{z_{ip}})$
```

```
For all documents prior to $i$, $j$:
    Draw latent citation propensity
    $D^*_{ipj} \sim \mathcal{N}(\boldsymbol{\tau}^T \mathbf{x}_{ipj}, 1)$
    Draw citation
    $D_{ipj} = 1$ if $D^*_{ipj} \geq 0$ and 0 otherwise
```

where $\mathbf{x}_{ipj}$ is a vector of covariates that shape the latent citation propensity for pth paragraph in document i to cite j document. $\mathbf{x}_{ipj}$ consists of 3 terms – the intercept, indegree, and $\eta_{j,z_{ip}}$. The intercept in $\mathbf{x}_{ipj}$ is to capture the overall sparsity of the citation network. Since networks in the real world are generally very sparse, we expect the intercept τ_0 to be negative. The indegree of a precedent is included to capture the authority. This follows existing studies of strategic citation that commonly point to the importance of the authority of a precedent as one of the major attracting factors of citations (Hansford and Spriggs, 2006; Lupu and Voeten, 2012; Lupu and Fowler, 2013). This is also consistent with a well-known dynamic in social networks called as "rich-get-richer" or more technically as "preferential attachment" where popular individuals become more popular (Newman, 2001; Wang et al., 2008). The indegree term is denoted $\kappa_j^{(i)}$, with superscript (i) to indicate the authority of the jth document at the time of i's writing. We expect its coefficient τ_1 to be positive. Finally, $\eta_{j,z_{ip}}$ is added to capture the topic similarity between the citing paragraph ip and document j. Since we expect that citations are more likely to occur between documents of similar topics, we expect its coefficient τ_2 to be positive.

While we currently include 3 document-level covariates in $\mathbf{x}$, researchers can add other covariates that fit their research purposes. For instance, the political ideology of judges in a precedent and a citing case can be an important factor in citation decisions (Lupu and Fowler, 2013). Then researchers can include a binary copartisanship indicator in $\mathbf{x}_{ipj}$ that takes 1 if the author of opinion i and the author of opinion j are appointed by the same president and 0 otherwise. Given the data (words and citations, $\mathbf{W}, \mathbf{D}$), our posterior probability is

$$p(\boldsymbol{\eta}, \boldsymbol{\Psi}, \mathbf{Z}, \boldsymbol{\tau} | \mathbf{W}, \mathbf{D}) \propto p(\boldsymbol{\mu} | \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) p(\boldsymbol{\tau} | \boldsymbol{\mu}_\tau, \boldsymbol{\Sigma}_\tau) p(\boldsymbol{\eta} | \boldsymbol{\mu}, \boldsymbol{\Sigma}) p(\boldsymbol{\Psi} | \boldsymbol{\beta}) p(\mathbf{Z} | \boldsymbol{\eta}) p(\mathbf{W} | \boldsymbol{\Psi}, \mathbf{Z}) p(\mathbf{D} | \mathbf{D}^*) p(\mathbf{D}^* | \boldsymbol{\tau}, \boldsymbol{\eta}, \mathbf{Z}, \mathbf{D}) \tag{1}$$

3.2 Related Models

Three features distinguish the PCTM from existing models for document networks. First, the PCTM recognizes the direction of citations. To our knowledge, existing models for document networks discount the fact that connections (or hyperlinks) between documents are directed (Chang and Blei, 2009; Liu et al., 2009). This may result in information loss or misleading conclusions, especially when directions are an important structural aspect of the data. In citation networks direction of citations show the temporal ordering of the documents which can be a critical feature if one is interested in diffusion dynamics, for example. A few models reflected the directionality of citations. The Author-recipient topic model (ART) by McCallum et al. (2007) incorporates the direction of document connections, but only indirectly by modeling topic mixtures to be a function of author and recipient characteristics. The pairwise citation LDA by Nallapati et al. (2008) considers all pairs of documents and thereby recognizes directed citations, but it models them as bidirectional which includes citations that cannot occur systematically (e.g. past documents citing future documents).

Second, the PCTM allows a document to send multiple citations – possibly of different topics – to another document. Most recent models for document networks such as the RTM (Chang and Blei, 2009), Topic-link LDA (Liu et al., 2009) and pairwise citation LDA (Nallapati et al., 2008) commonly model connections between documents as binary process – whether the given pair of documents are connected or not. On the other hand, in the PCTM a paragraph is the unit where citations arise, and a document typically consists of multiple paragraphs. This allows a document to cite another document as many as the number of its paragraphs. The number of citations between the given two documents can contain rich information such as the strength of their topic similarity. In addition, since citations share the topic of the paragraphs in which they occur, the topic composition of citations can convey useful information on the semantic context of the linkages between documents.

Third, the PCTM incorporates regression structure to model a paragraph's latent citation propensity over all precedents and thereby offers flexibility for researchers to model strategic citation dynamics as they theorize. Existing models for document networks commonly model citations as a simple process of topic similarity at the word level. The RTM for example employs the word topics averaged at the document level to model citations between documents (Chang and Blei, 2009). While topically similar documents are more likely to connect to each other by intuition, past studies have emphasized that there can be more social and political processes involved to whether and how often documents receive citations (Hansford and Spriggs, 2006; Lupu and Fowler, 2013; Pelc, 2014). The PCTM in the current form includes the authority of the precedent in addition to the topic similarity between the citing paragraph and the cited document. On top of this, researchers can add any variable at the paragraph, document, and paragraph-document dyad level in the regression to model strategic citations without disrupting the Bayesian inference we introduce below.

3.3 Bayesian Inference

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. Furthermore, variational inference is an optimization method that outputs point estimates. This requires additional steps to obtain measure uncertainty in estimation. Quantifying uncertainty in variational inference is often done through bootstrapping (Chen et al., 2018; Imai et al., 2016). However, obtaining bootstrap samples representative of the pseudo population can be highly challenging for network data since observations are connected (Chen et al., 2019; Levin and Levina, 2019). It often requires block sampling which entails computing other network quantities (i.e. geodesic distance in Raftery et al. (2012)) but these additional processes could defeat the advantage of the computational efficiency of using variational inference.

To remedy this problem, we follow the recent advances in the inference of Correlated Topic Models(CTM) that adopts partial collapsing (Held and Holmes, 2006; Chen et al., 2013; Linderman et al., 2015). We first partially collapse the posterior distribution by integrating out the topic-word probability parameter $\boldsymbol{\Psi}$. 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

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

where $p(\mathbf{W}|\mathbf{Z})$ results from collapsing $\boldsymbol{\Psi}$ as follows.

$$\begin{aligned} p(\mathbf{W}|\mathbf{Z}) &= \int_{\boldsymbol{\Psi}} p(\mathbf{W}, \boldsymbol{\Psi}|\mathbf{Z})d\boldsymbol{\Psi} \\ &= \int_{\boldsymbol{\Psi}} p(\mathbf{W}|\boldsymbol{\Psi}, \mathbf{Z})p(\boldsymbol{\Psi}|\mathbf{Z})d\boldsymbol{\Psi} \\ &= \int_{\boldsymbol{\Psi}} p(\mathbf{W}|\boldsymbol{\Psi}, \mathbf{Z})p(\boldsymbol{\Psi})d\boldsymbol{\Psi} \end{aligned} \tag{3}$$

The above takes the form of Dirichlet-multinomial distribution which enters in the conditional posterior distribution of $\mathbf{Z}$ below. The conditional posterior distribution of $\mathbf{Z}$ for ipth paragraph is

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

where

$$\begin{aligned} \pi_{ipj,k} = \exp\Bigg\{ &\eta_{ik} + \log\prod_v \Gamma(\beta_v + c^v_{k,ip} + c^v_{k,-ip}) - \log\Gamma(\sum_v \beta_v + c^v_{k,ip} + c^v_{k,-ip}) \\ &- \frac{1}{2}\Big(\tau_2^2\eta_{jk}^2 + 2\big(\tau_0\tau_2 + \tau_1\tau_2\kappa_j^{(i)} - \tau_2 D^*_{ipj}\big)\eta_{jk}\Big)\Bigg\} \end{aligned} \tag{5}$$

Here, $c^v_{k,ip}$ denotes the total number of times the vth word appears in paragraph ip of topic k such that $c^v_{k,ip} = \sum_{l=1}^{n_{ip}} \mathbb{I}(W_{ipl} = v)\mathbb{I}(z^k_{ip} = 1)$. Likewise, $c^v_{k,-ip}$ is the total number of times the vth term appears in paragraphs with kth topic except for ip. The form of the conditional posterior for the ipth paragraph-level topic z^k_{ip} offers a convenient interpretation on the *source of information*. The first part $p(z^k_{ip} = 1|\boldsymbol{\eta}_i)$ displays the topic information from document-level topic prevalence. The second part represents topic information from the words in ipth paragraph. The third part $\prod_{j=1}^{i-1} p(D^*_{ipj}|z^k_{ip} = 1, \mathbf{Z}_{-ip}, \boldsymbol{\tau}, \boldsymbol{\eta}, \kappa)$ is equivalent to the total amount of topic information from citations.

The conditional posterior distribution of $\boldsymbol{\eta}$ for ith document is jointly defined with the augmenting Polya-Gamma distribution for $\boldsymbol{\lambda}$. The conditional posterior distribution for λ_{ik} is

$$p(\lambda_{ik}|\mathbf{Z}, \mathbf{W}, \boldsymbol{\eta}) \propto PG(N_i, \rho_{ik}) \tag{6}$$

where $\rho_{ik} = \eta_{ik} - \log(\sum_{l \neq k} e^{\eta_{il}})$.

With λ_{ik}, we can obtain the conditional posterior of $\boldsymbol{\eta}$ for ith document as follows.

$$p(\eta_{ik}|\eta_{i,-k}, \mathbf{Z}, \mathbf{W}, \mathbf{D}, \boldsymbol{\tau}, \lambda_{ik}) \propto \mathcal{N}(\eta_{ik}|\tilde{\mu}_{ik}, \tilde{\sigma}^2_k) \tag{7}$$

where

$$\begin{aligned} \tilde{\sigma}^2_k &= (\sigma_k^{-2} + \lambda_{ik} + v^{-1}_{i,kk})^{-1} \\ \tilde{\mu}_{ik} &= \tilde{\sigma}^2_k\big(v^{-1}_{i,kk}m_{ik} + \sigma_k^{-2}\nu_{ik} + t_{ik} - \frac{N_i}{2} + \lambda_{ik}\log(\sum_{l \neq k} e^{\eta_{il}})\big) \end{aligned} \tag{8}$$

For the definition of $v_{i,kk}$, m_{ik}, ν_{ik}, and t_{ik} as well as the detailed derivation, see our Supplementary Information document, section B.2.

The conditional posterior for latent citation propensity parameter $\mathbf{D}^*$ is

$$p(D^*_{ipj}|\boldsymbol{\eta}, \mathbf{Z}, \boldsymbol{\tau}, \mathbf{D}) \propto \begin{cases} TN_{[0,\infty)}(\tau_0 + \tau_1\kappa_j^{(i)} + \tau_2\eta_{j,z_{ip}}, 1) & \text{if } D_{ipj} = 1 \\ TN_{(-\infty,0]}(\tau_0 + \tau_1\kappa_j^{(i)} + \tau_2\eta_{j,z_{ip}}, 1) & \text{if } D_{ipj} = 0 \end{cases} \tag{9}$$

The conditional posterior for $\boldsymbol{\tau}$ follows the following distribution. Let $\mathbf{x}_{ipj} = [1, \kappa_j^{(i)}, \eta_{j,z_{ip}}]^T$ and $\boldsymbol{\tau} = [\tau_0, \tau_1, \tau_2]^T$

$$\begin{aligned} p(\boldsymbol{\tau}|\boldsymbol{\eta}, \mathbf{Z}, \mathbf{D}^*) &\propto exp\Bigg\{ -\frac{1}{2}\sum_{ipj}\Big(D^*_{ipj} - \mathbf{x}^T_{ipj}\boldsymbol{\tau}\Big)^2\Bigg\} N(\boldsymbol{\mu_\tau}, \boldsymbol{\Sigma_\tau}) \\ &\propto N(\tilde{\boldsymbol{\tau}}, \tilde{\boldsymbol{\Sigma_\tau}}) \end{aligned} \tag{10}$$

where $\tilde{\boldsymbol{\Sigma_\tau}} = \Bigg(\Big(\sum_{ipj}\mathbf{x}_{ipj}\mathbf{x}^T_{ipj}\Big) + \boldsymbol{\Sigma}_\tau^{-1}\Bigg)^{-1}$ and $\tilde{\boldsymbol{\tau}} = \tilde{\boldsymbol{\Sigma_\tau}}\Bigg(\Big(\sum_{ipj}\mathbf{x}^T_{ipj}D^*_{ipj}\Big) + \boldsymbol{\Sigma}_\tau^{-1}\boldsymbol{\mu_\tau}\Bigg)$

4 Simulation Results

We validate the performance of the PCTM using simulation. First, we show that the PCTM can recover the true topics from randomly initialized topics. Second, we show that the PCTM fits simulation data better than the existing models for document networks.

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.[3] 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.

First, 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 1 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

[3] 106 documents, an average of 44 paragraphs per document, 5838 unique words, and an average of 51 words per paragraph.

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 1. In comparison, the left panel of Figure 1 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 2, 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 1, 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.

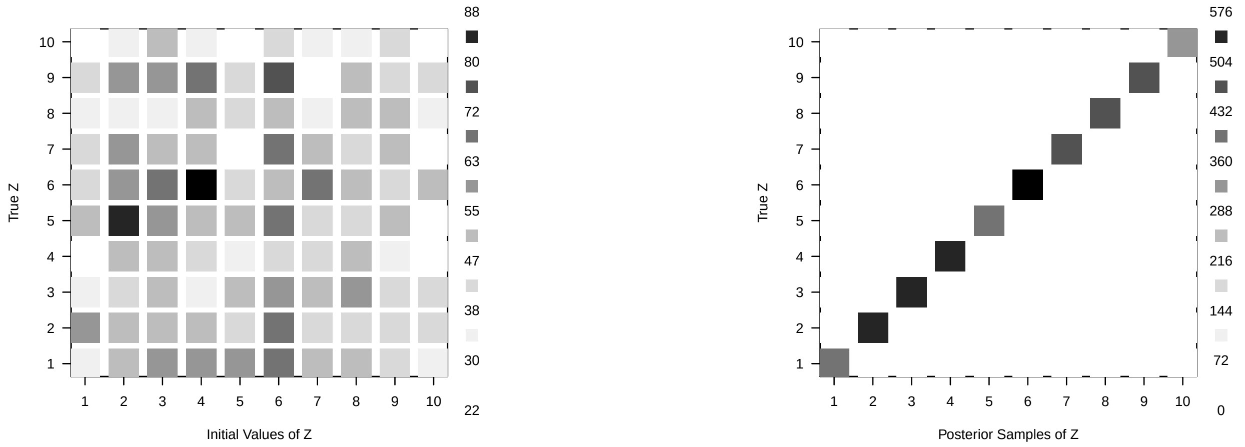


Figure 1: The comparison of the estimated and the true topics of paragraphs. On the right panel, the (k, l) cell shows the number of paragraphs whose estimated topic is l while the true topic is k. We estimate topics using the paragraph topic parameter, $\mathbf{Z}$, using the last draw from our Gibbs 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.

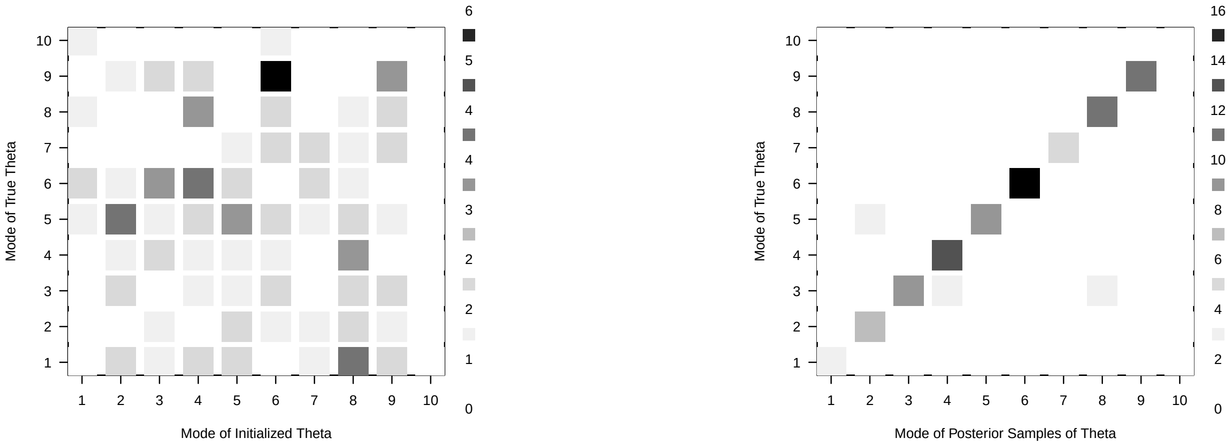


Figure 2: The comparison of the estimated and the true topic distribution of documents. On the right panel, the (k, l) cell shows the number of documents whose mode of the estimated topic distribution, $\boldsymbol{\theta}$, across K topics is l while the mode of the true topic distribution is k. We 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.

Next, we validate the performance of the PCTM by comparing posterior predictive probabilities of a new document with two existing models for document networks: (1) Relational Topic Model (RTM) and (2) Latent Dirichlet Allocation (LDA) combined with Logistic Regression. Both models assume that a pair of documents with similar topic distribution is more likely to cite each other. The difference between RTM and LDA + Logistic Regression is that the latter uses a two-step approach to model the generation of text and citations separately while the former jointly models text and citations (Chang and Blei, 2009).[4] The goal is to compare the posterior predictive probability of a new document given past documents across the three models. The higher predictive probability indicates a better model fit.

The following is the procedure of the simulation exercise on predictive probability. We fit the three models, the PCTM, RTM, and LDA + Logistic Regression on simulation data, using all the documents except the last document. We chose the last document as the test document because our corpus has a temporal order. Then, compute the posterior predictive probabilities of the words and the citations in each paragraph of the last document. We then

[4] We fit LDA and RTM using an `R` package, `lda`.

take the average across paragraphs to obtain the average posterior predictive probabilities of a paragraph in the last document. We repeat this process for 100 simulation datasets and compare the predictive probabilities across models. The following gives the posterior predictive probability for the PCTM. $\mathbf{W}_{iq}$ and $\mathbf{D}_{iq}$ are the data in a paragraph q of a document i. $\mathbf{W}^{train}, \mathbf{D}^{train}$ are the data in documents other than document i. The parameters with $\hat{}$ symbol indicate that they are samples from the posterior distributions from the model trained with $\mathbf{W}^{train}, \mathbf{D}^{train}$. We draw 1000 samples from the posterior for those parameters and compute the average to obtain the final predictive probability.

$$\begin{aligned} & p(\mathbf{W}_{iq}, \mathbf{D}_{iq} | \mathbf{W}^{train}, \mathbf{D}^{train}) \\ & = \sum_{k=1}^{K} \Big\{ p(\mathbf{W}_{iq} | z_{iq} = k, \hat{\mathbf{\Psi}}) \\ & \quad \times \prod_{j=1}^{i-1} \mathbb{P}(D^*_{iqj} > 0 | \hat{\boldsymbol{\tau}}, \hat{\boldsymbol{\eta}}, z_{iq} = k)^{\mathbb{I}\{D_{iqj}=1\}} \mathbb{P}(D^*_{iqj} < 0 | \hat{\boldsymbol{\tau}}, \hat{\boldsymbol{\eta}}, z_{iq} = k)^{\mathbb{I}\{D_{iqj}=0\}} \\ & \quad \times p(z_{iq} = k | \hat{\boldsymbol{\eta}}) \Big\} \end{aligned} \tag{11}$$

Figure 3 shows the histogram of the predictive probabilities of the last document across RTM, LDA + Logistic Regression, and the PCTM. It shows that the predictive probabilities of the PCTM are almost always higher than the other two models. This means that the PCTM fits the type of data we use in our application better than those existing models.

5 Empirical Results

This section presents the results of applying the PCTM to the SCOTUS dataset, and compares it with the results from two existing models: LDA, which does not use citation information at all, and RTM, which uses both text and citation information, but assumes that edges are undirected and does not care where in a document a citation is made. We focus on the two subsets of the dataset: Privacy and Voting rights.

5.1 Privacy

Figure 4 displays the results of LDA, RTM, and the PCTM on the entire SCOTUS opinions on Privacy issue area. LDA assigns topics based on words without reference to how documents are connected. RTM incorporates the networked structure of documents but assumes that connections between documents are undirected and binary. In addition, RTM is agnostic about the semantic context of the network because it does not take into account where in

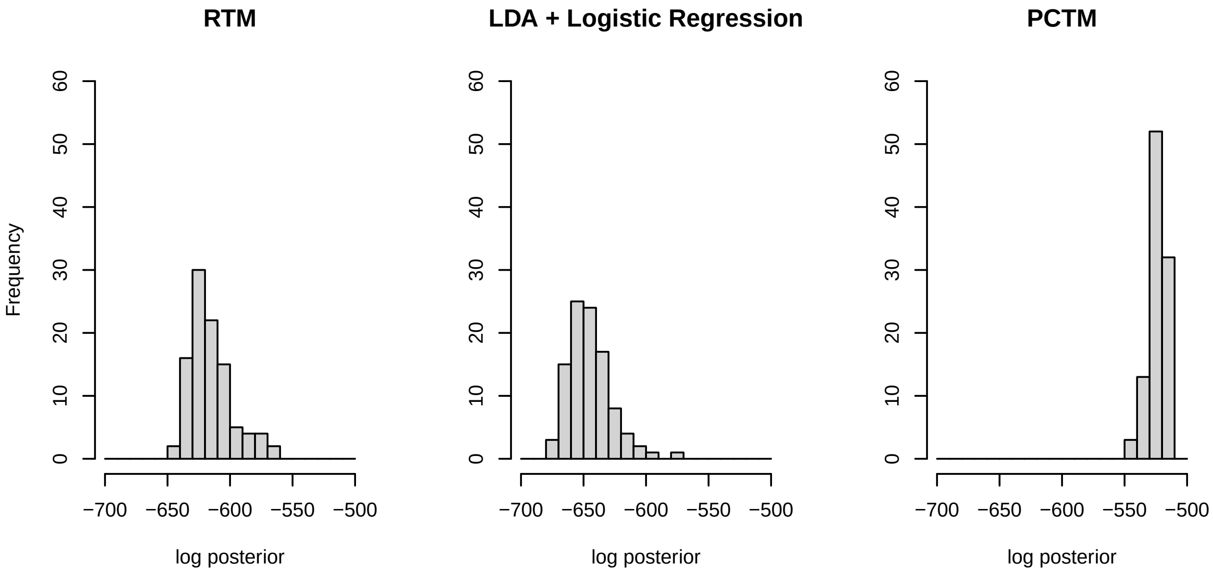


Figure 3: Comparison of the predictive probabilities of Relational Topic Model (RTM), Latent Dirichlet Allocation (LDA) + Logistic Regression, and the PCTM (from left to right). Distribution of the log posterior predictive probabilities across 100 simulation datasets. Overall, the PCTM has higher posterior predictive probabilities over RTM or LDA + Logistic Regression.

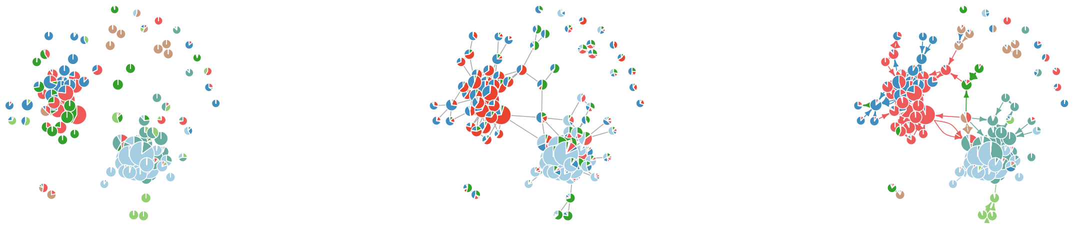
a document a citation is made (thus, all edge colors are in gray). While RTM's assumption on the simple network – undirected and binary edges – helps uncover the topic structure of connected documents, it falls short of reflecting the structural properties of citation networks where edges are directed, acyclic and documents often cite another document multiple times.

As seen in (c) of Figure 4, the PCTM assigns topics to citations, recognizes the direction of the connection, and allows a document to connect to another document more than once. A document referencing another document multiple times can provide a rich information about the topic structure of the two documents as well as the semantic context of the linkage between them. In (c) a document(in red color) at the center of the network makes two citations of red color to another document, suggesting that the citing document is primarily addressing legal doctrines addressed by the red paragraphs of the cited document.

Words most frequent for each topic in the PCTM are given in Table 1. The Supreme Court Database assigns 4 issue codes to opinions of Privacy issue area[5], but we identify 7 distinct topics in PCTM. The labels in the table are provided by the authors.

The first and the third topic both address abortion as the substantive case in point but differ in the context in which abortion is addressed. Paragraphs of the first topic illuminate

[5] The 4 issue codes are privacy, abortion, right to die and Freedom of Information Act.



(a) LDA

(b) RTM

(c) PCTM

Figure 4: The result of three topic models, LDA, RTM, and PCTM from (a) to (c), on the US Supreme Court opinions of *Privacy* issue area. A node represents an opinion, and an edge represents a citation between opinions. The nodes are colored according to the proportion of paragraphs with the same estimated topics. The colors of an edge is based on the estimated topic of the paragraph where the citation is made. Note that the topic spaces of the three models are not exactly the same. Same colors are assigned to topics that share the top 5 most frequent words between the three models. (a) LDA estimates topic structure of documents without reference to the citation network. (b) RTM takes into account the linkage between documents for the estimation of topics, but assumes that edges are undirected and remains agnostic about the topics of citations. (c) PCTM recognizes the directions of edges and estimates the topic structure of both documents and citations. PCTM offers a semantic context over how documents are connected by identifying the topic of the paragraph in which a citation is made.

Topic Label	Regulation of Abortion Procedure	Procedural Posture	Const. Rights to Abortion	Speech & Protest	Damage to Privacy	Privacy vs Govnt. Interest	Public Disclosure of Private Information
1	abort	appeal	right	clinic	damag	drug	inform
2	parent	district	abort	injunct	act	act	agenc
3	minor	board	constitu	right	actual	test	exmpt
4	physician	ani	protect	public	congress	student	disclosur
5	perform	order	medic	speech	person	school	record
6	woman	agency	amend	petition	privaci	respond	public
7	medic	document	decis	protest	right	use	govern
8	interest	rule	person	zone	ani	ani	act
9	health	unit	interest	interest	general	district	congress
10	consent	act	life	person	doe	petition	foia

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

abortion as woman's right and discuss the conditions in which the decision can be restricted or unrestricted such as woman's health, being a minor, or ill-informed by her physician and etc. The third topic addresses it in a broader context of a person's right to life and death (e.g. is the right to birth control limited to married couples). The second topic addresses the processes involving lower and higher courts, which we believe to be a byproduct of having paragraphs as the unit for topic assignments. Almost all majority opinions in the SCOTUS have at least one paragraph discussing how the case was appealed from the lower court to higher courts. Since the set of vocabulary and citations in those paragraphs are generally distinct from other paragraphs, the PCTM tends to assign a topic for this category. Paragraphs of the fourth topic mostly concern public protests and speeches surrounding (anti) abortion decisions in courts. The fifth topic addresses what constitutes damage to privacy under the Privacy Act of 1974. The sixth and seventh topics both concern the public disclosure of private information. The sixth topic, which we label as `Privacy vs Government Interest`, mainly addresses the access to private information such as the history of drug abuse that might disrupt the operations of government agencies. The seventh topic, on the other hand, concerns whether the way private information is recorded constitutes a violation of Privacy Act of 1974.

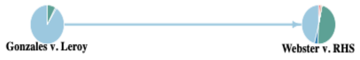
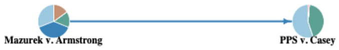
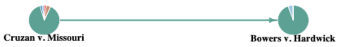
Regulation of Abortion Procedures	Procedural Posture	Const. Rights to Abortion
		
... The law need not give abortion doctors unfettered choice in the course of their medical practice, nor should it elevate their status above other physicians in the medical community. In Casey the controlling opinion held an informed-consent requirement in the abortion context was "no different from a requirement that a doctor give certain specific information about any medical procedure." 505 U. S., at 884 (joint opinion). The opinion stated "the doctor-patient relation here is entitled to the same solicitude it receives in other contexts." Ibid.; see also **Webster v. Reproductive Health Services, 492 U. S. 490, 518-519 (1989)** ...	...The District Court denied respondents' motion for a preliminary injunction, finding that they had not established any likelihood of prevailing on their claim that the law imposed an "undue burden" within the meaning of **Planned Parenthood of Southeastern Pa. v. Casey, 505 U. S. 833 (1992). 906 F. Supp. 561, 567 (Mont. 1995)**. The Court of Appeals for the Ninth Circuit vacated the District Court's judgment ...	Although many state courts have held that a right to refuse treatment is encompassed by a generalized constitutional right of privacy, we have never so held. We believe this issue is more properly analyzed in terms of a Fourteenth Amendment liberty interest. See **Bowers v. Hardwick, 478 U. S. 186, 194-195 (1986)**

Table 2: Paragraphs containing citations of Topics 1,2, and 3. The top row displays two opinions and a citation with color-coded topics. The second row for each topic contains the paragraph that contains the citation between the two opinions in the first row.

Note that `NASA v. Nelson` and `US v. RCFP` in the second and third columns of Table 3 both cite `Whalen v. Roe`, but the context of the citations vary. For `NASA v. Nelson`, the focus was on whether the employer(NASA) should have access to private information(history of drug abuse) of its employees whereas for `US v. RCFP`, `Whalen v. Roe` was mainly about

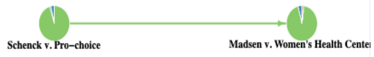
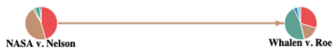
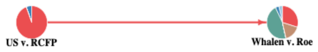
Speech & Protest	Privacy vs Govnt. Interest	Public Disclosure of Private Information
		
Petitioners, two individual defendants, appealed to Court of Appeals for the Second Circuit. While the case was on appeal, we decided **Madsen v. Women's Health Center, Inc., 512 U. S. 753 (1994)**, a case which also involved the effect of an injunction on the expressive activities of antiabortion protesters. (We discuss Madsen in greater depth in Part II-A, infra.) We held that "our standard time, place, and manner analysis is not sufficiently rigorous" when it comes to evaluating content-neutral injunctions that restrict speech. The test instead, we held, is "whether the challenged provisions of the injunction burden no more speech than necessary to serve a significant government interest." 512 U. S., at 765.	With these interests in view, we conclude that the challenged portions of both SF-85 and Form 42 consist of reasonable, employment-related inquiries that further the Government's interests in managing its internal operations. See Engquist, 553 U. S., at 598-599; **Whalen v. Roe, 429 U. S.**, at 597-598. As to SF-85, the only part of the form challenged here is its request for information about "any treatment or counseling received" for illegal-drug use within the previous year. ... The Government has good reason to ask employees about their recent illegal-drug use. Like any employer, the Government is entitled to have its projects staffed by reliable, law-abiding persons who will " 'efficiently and effectively'" discharge their duties.	... Here, the former interest, "in avoiding disclosure of personal matters," is implicated. Because events summarized in a rap sheet have been previously disclosed to the public, respondents contend that Medico's privacy interest in avoiding disclosure of a federal compilation of these events approaches zero. We reject respondents' cramped notion of personal privacy ... We have also recognized the privacy interest in keeping personal facts away from the public eye. In **Whalen v. Roe, 429 U. S. 589 (1977)**, we held that "the State of New York may record, in a centralized computer file, the names and addresses of all persons who have obtained, pursuant to a doctor's prescription, certain drugs for which there is both a lawful and an unlawful market." Id., at 591. In holding only that the Federal Constitution does not prohibit such a compilation, we recognized that such a centralized computer file posed a "threat to privacy":

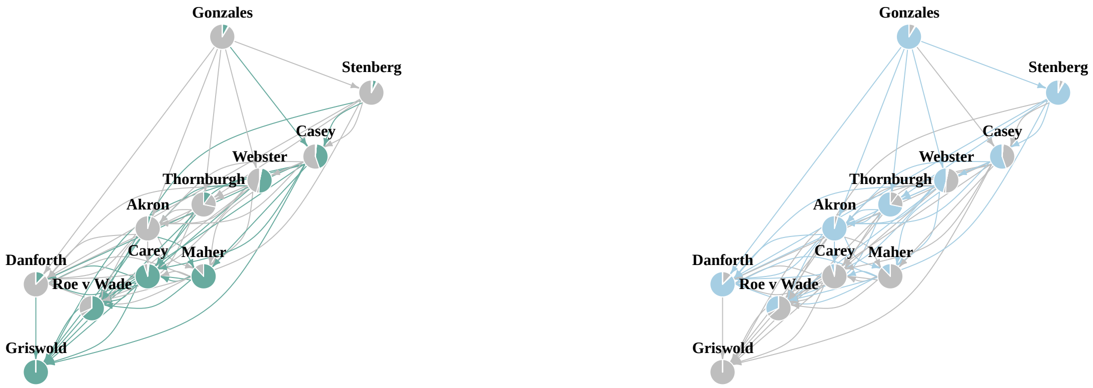
Table 3: Paragraphs containing citations of Topics 4,6, and 7. The top row displays two opinions and a citation with color-coded topics. The second row for each topic contains the paragraph that contains the citation between the two opinions in the first row.

the record-keeping of private information (in rap sheet in `US v. RCFP` and in computer files in `Whalen v. Roe`) and the consequent public disclosure of that information. This highlights that the semantic context of citations may differ even when the given citations refer to the same document.

Another advantage of the PCTM is that the temporal ordering of the documents are directly incorporated in the model. To emphasize this aspect, we show 11 selected opinions on Reproductive rights in Figure 5. [6]

Figure 5 displays the topic structure of the 11 selected opinions on reproductive rights. We observe that the topic structure of the subnetwork is governed mostly by two topics – `Regulation of Abortion Procedures` or `Constitutional Rights to Abortion`. More precisely, earlier opinions mostly consist of `Right to Abortion` topic while more recent opinions show a greater prevalence of `Regulation of Abortion Procedures`. This is consistent with Clark and Lauderdale (2012) that explains that the discourse on abortion in the Supreme Court was on person's constitutional right to birth control in earlier cases such as `Griswold v. Connecticut` (1965), and cases afterward subsequently focused on the details of how abortion procedures should be regulated. Later cases also make more explicit references to abortion and woman's right such that `Planned Parenthood v. Casey` (1992), for

[6] The 11 opinions on reproductive rights are selected based on Figure 4 of Clark and Lauderdale (2012).



(a) Constitutional Rights to Abortion

(b) Regulation of Abortion Procedures

Figure 5: The citation network of 11 selected opinions on reproductive rights. The opinions are part of the SCOTUS subset on Privacy issue area. The left panel highlights the paragraphs and citations of `Const. Rights to Abortion` topic (in teal). The right panel colors the paragraphs and citations of `Regulation of Abortion Procedures` topic. The y-axis represents chronological order such that opinions placed lower indicate older in time and opinions placed in the upper part of the figure are more recent documents.

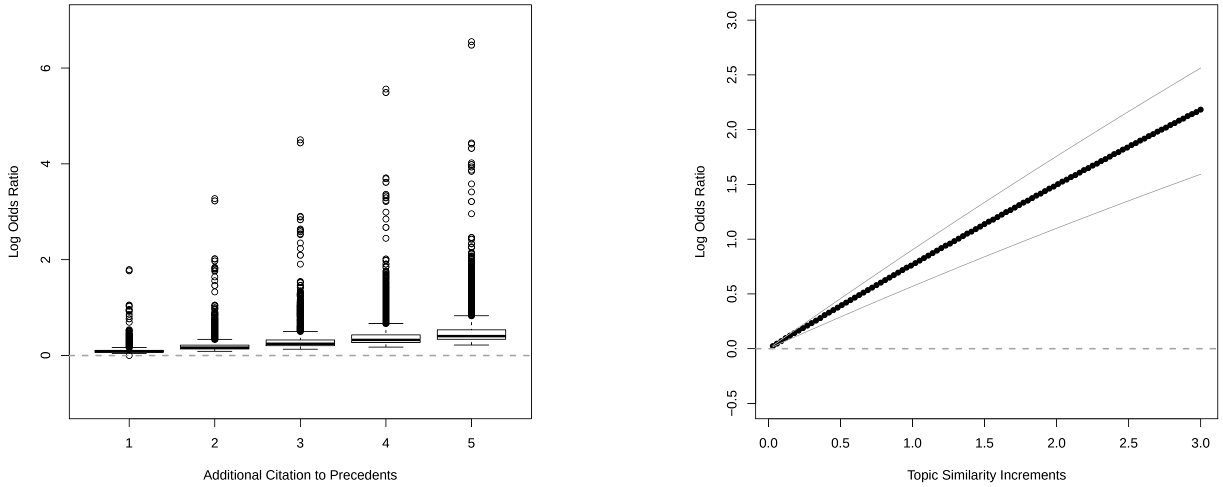
instance, states that "The ability of women to participate equally in the economic and social life of the Nation has been facilitated by their ability to control their reproductive lives."[7]

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 6 we offer one way to interpret coefficients $\boldsymbol{\tau}$ in latent citation propensity.[8] 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 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

[7] `https://reproductiverights.org/our-work/landmark-cases/`

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



(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 6: 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.

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 6 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.

5.2 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 majority of those challenges made their way to the Supreme Court. The Supreme Court database assigns 3 issue codes for opinions related to voting.[9] After examining a subset of documents with these issue codes, we decided to set the number of topics to 4 for PCTM.

Table 4 presents the 10 words that appear most frequently for each topic. The first topic

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 4: Top 10 words of highest probability for each topic from PCTM.

`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

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

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 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 7 shows the cumulative count of paragraphs of each topic. The growth

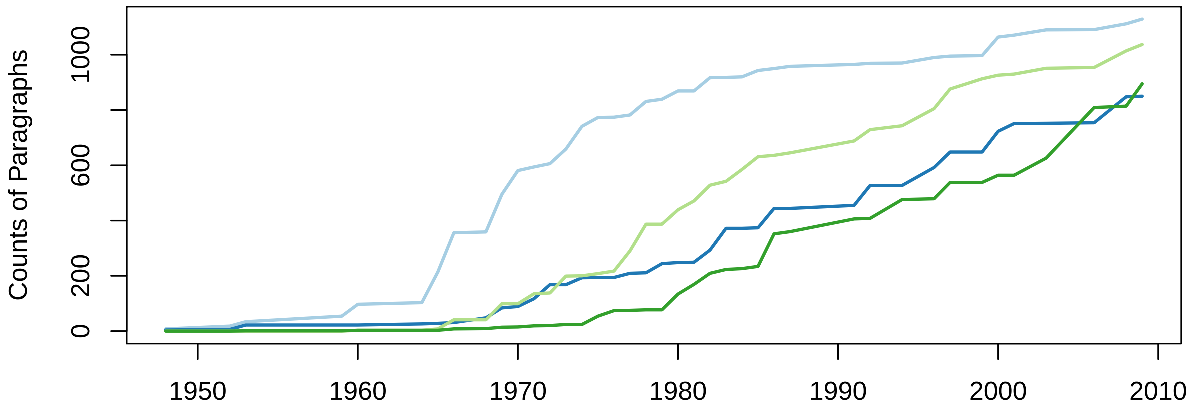
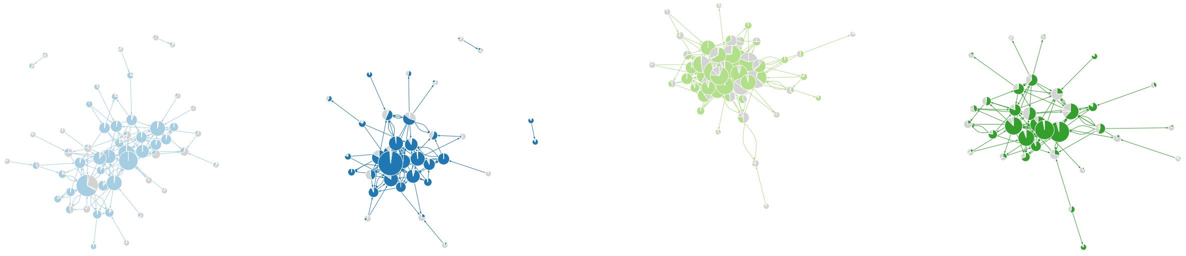


Figure 7: Cumulative number of topics in Voting Rights subset over time.

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 8 shows groups of cases that make citations of the given topic. The location



(a) Voter Eligibility

(b) Ballot Access

(c) Preclearance Requirement

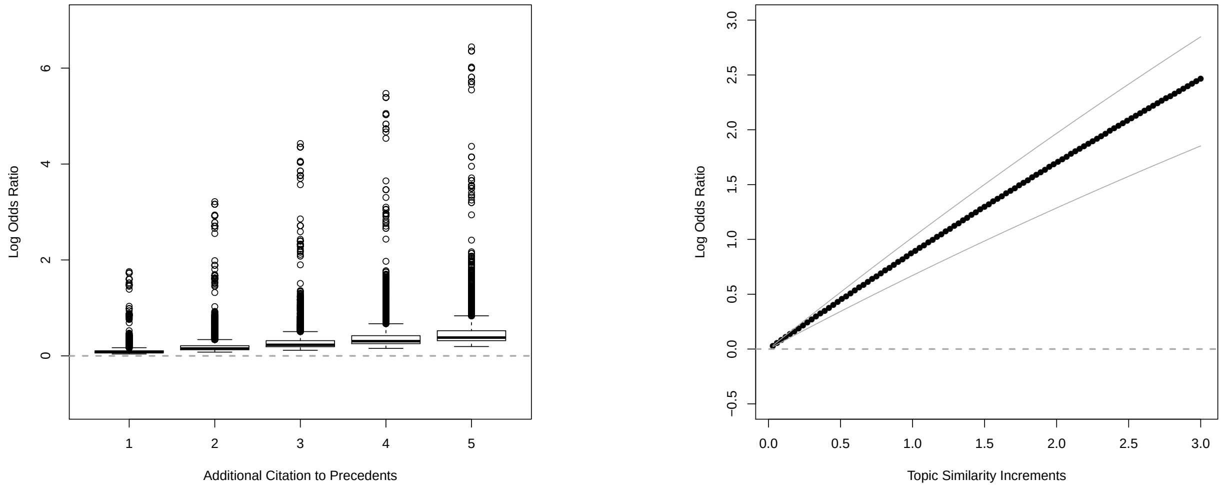
(d) Voter Dilution

Figure 8: 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.

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 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 3 $\boldsymbol{\tau}$ 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 6, 10,000 randomly drawn pairs of paragraphs and precedents for the Voting subset were used to generate Figure 9. The left panel of Figure 9 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 9: 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 6b is employed to create this figure.

5.3 Application on a New Case in Abortion

This section presents additional results on a new controversial case regarding abortion. On June 24 2022, the Supreme Court made a landmark decision on abortion that invoked a nationwide controversy. In the case, `Dobbs v. Jackson Women's Health Organization`, the SCOTUS held that abortion is not a part of constitutional rights, and it conferred individual states the right to ban abortion. This case overturned both `Roe v. Wade` and `Planned Parenthood v. Casey`, the landmark precedents that have served as the legal basis for the constitutional rights to abortion. While qualitative reading of `Dobbs v. Jackson Women's Health Organization` suggests that this case is a clear deviation from the recent trends in abortion rulings in many ways, it is difficult to demonstrate the deviations in a quantitative way.

Using the PCTM, we examine how `Dobbs v. Jackson Women's Health Organization` differs from the recent rulings on abortion. To do so, we computed the predicted probability of topics of the paragraphs in `Dobbs v. Jackson Women's Health Organization`. We first train the PCTM on the abortion corpus used in the above analysis and then computed the posterior predictive distribution of topics. The exact procedure to obtain the posterior predictive probability is in Appendix.

To establish the face validity of the topics in `Dobbs v. Jackson Women's Health`

Organization, Table 5 presents two paragraphs that cite the same precedent, Planned Parenthood v. Casey (505 U.S., 878), but with different estimated topics. The left paragraph has the estimated topic constitutional rights to abortion while the right paragraph has the topic regulation of abortion procedure. The left paragraph is an introductory paragraph of the judges' criticism of Casey's argument that abortion is a part of the liberty protected by the Fourteenth Amendment. This is clearly related to whether abortion is a part of constitutional rights or not. By contrast, the right paragraph criticizes the "undue burden" test that Casey decides. Undue burden test offers criteria about what kind of state regulations on abortion is prohibited. Therefore, we can infer that this paragraph discusses a more specific issue about how states regulate abortions. By reading these paragraphs, we can verify that our estimated topics match our interpretations of the topics.

How do the topics in Dobbs v. Jackson Women's Health Organization differ from the recent rulings on abortion? For comparison, we also computed the predicted probability of the topics for the two recent precedents about abortion in our corpus: Gonzales v. Carhard and Stenberg v. Carhard. Figure 10 shows the predicted probability of topics for each paragraph for the three cases on abortion, Gonzales v. Carhard, Stenberg v. Carhard, and Dobbs v. Jackson Women's Health Organization, from top to bottom. Each vertical bar represents a paragraph, and each bar is colored according to the predicted probability of topics. Since we want to focus on the difference in the legal discourse regarding abortion, we focus our analysis on the two topics relevant to abortion: constitutional rights to abortion or regulation of abortion procedure. While more than 90% of the paragraphs of both Gonzales and Stenberg are assigned with regulation of abortion procedure topic, only 28% of the paragraphs in Dobbs v. Jackson are assigned with the regulation topic and 67% of the paragraphs are assigned with constitutional rights to abortion. This accurately reflects the fact that Dobbs v. Jackson Women's Health Organization is distinct from the current trend in the abortion rulings in our corpus.

6 Concluding Remarks

In this paper, we developed and applied a new topic model for jointly analyzing text and citations. Many corpora in social sciences, such as the SCOTUS decisions, consist of a document network in which documents are interconnected through citations of each other. On the one hand, previous application studies using such datasets tend to use either the network data or the text data, but not both. On the other hand, existing models for document networks are unable to estimate topics for citations associated with words, often because they assume that words and citations are independent given the topic mixture of

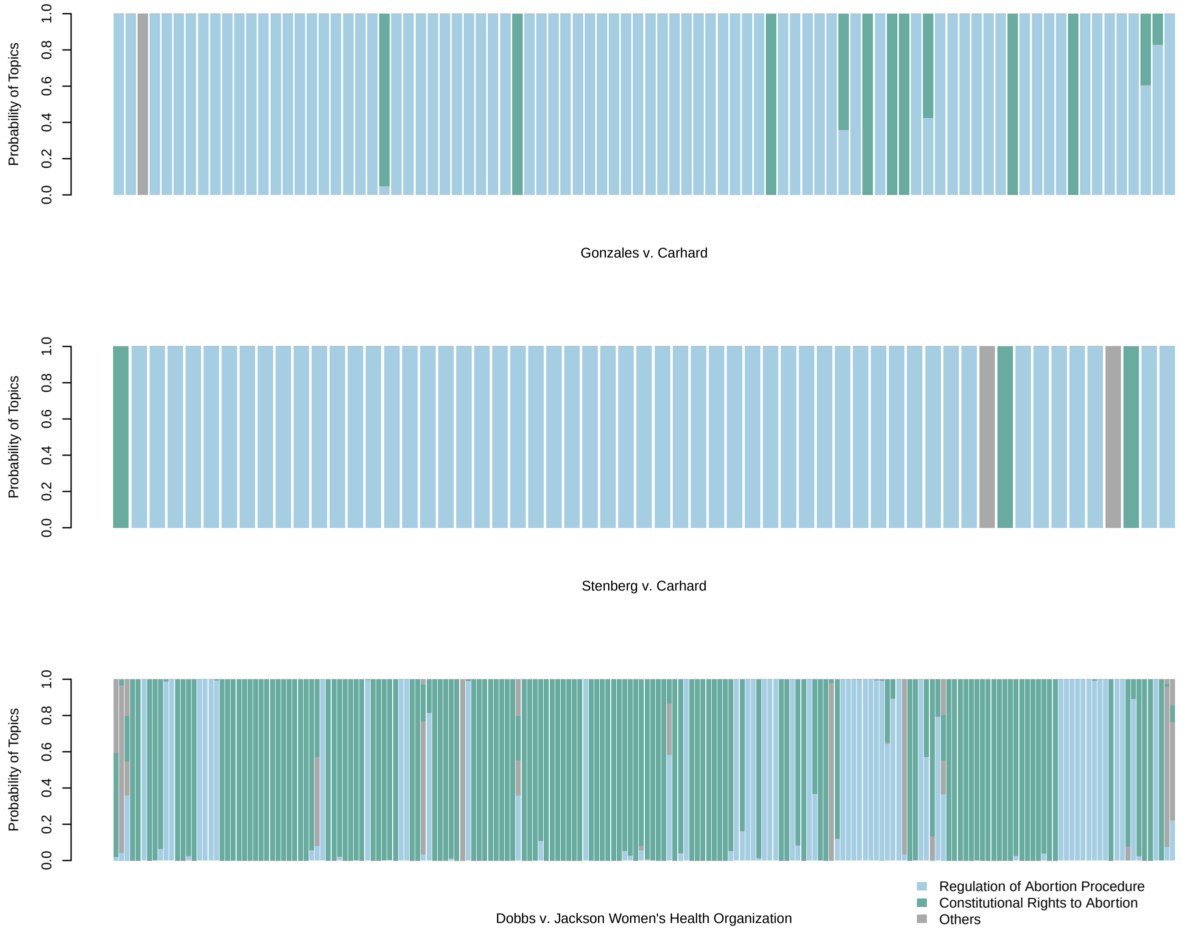


Figure 10: Predicted Probability of Topics for the Paragraphs of Dobbs v. Jackson. Each vertical bar represents a paragraph. Each paragraph is colored according to the predicted probability of topics. We focus on two topics related to abortion: constitutional rights to abortion and regulation of abortion procedure. The case are Gonzales v. Cargard, Stenberg v. Carhard, and Dobbs v. Jackson Women's Health Organization, from top to bottom. Dobbs v. Jackson Women's Health Organization case have more paragraphs with constitutional rights to abortion topic rather than regulation of abortion procedure topic while the two recent precedents in our corpus, Gonzales v. Carhard and Stenberg v. Carhard, are the opposite. This shows that Dobbs v. Jackson Women's Health Organization goes against the recent trend in the abortion cases in our corpus, where stronger emphasis is placed on how abortion can be regulated by the states instead of whether abortion is a part of the constitutional rights, as shown in Gonzales v. Carhard and Stenberg v. Carhard.

Constitutional Rights to Abortion	Regulation of Abortion Procedures
We turn to Casey's bold assertion that the abortion right is an aspect of the "liberty" protected by the Due Process Clause of the Fourteenth Amendment. **505 U.S., at 846**	The Casey plurality tried to put meaning into the "undue burden" test by setting out three subsidiary rules [...] The first rule is that "a provision of law is invalid, if its purpose or effect is to place a substantial obstacle in the path of a woman seeking an abortion before the fetus attains viability." **505 U.S., at 878**

Table 5: Comparison of Paragraphs in `Dobbs v. Jackson` with Different Estimated Topics on Abortion.

their document.

Our proposed model overcomes this limitation by modeling paragraphs as the unit of topic assignment. This allows us to assign topics to citations, and we utilized this new property in our applications to the citation networks of the SCOTUS opinions. The PCTM can also model the formation of citation networks with the latent citation propensity. In our application, we included a precedent's authority and topic similarity to a citing paragraph in the latent citation propensity. The results are consistent with existing studies of strategic citation behavior in the SCOTUS (Hansford and Spriggs, 2006) such that we observe the positive impact of a document's authority on citation formations in the SCOTUS.

While we focused on the SCOTUS opinions for our application, the applications of the PCTM need not be limited to citation networks of legal documents. We hope that our model will help address a number of important research questions in the analysis of document networks. For example, a researcher can focus on the latent citation propensity part of our model to understand the role of authors' gender in citation making in academic journals. Since academic articles address diverse scholarly subjects, capturing semantic contexts in the analysis of citation formation is critical, and can be properly addressed in our model. We can also imagine studies uncovering the topic structure of document networks with information obtained from both text and networks. Finally, we emphasize that our model can yield topic-specific subnetworks which then can be used together with established measures of networks, such as legal importance scores in Fowler et al. (2007), to produce better academic insights.

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.

Barry, A. E., Valdez, D., Padon, A. A., and Russell, A. M. (2018). Alcohol advertising on twitter—a topic model. *American Journal of Health Education*, 49(4):256–263.

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. (2006). Dynamic topic models. In *International Conference on Machine Learning, ACM*.

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

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

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.

Chen, Y., Gel, Y. R., Lyubchich, V., and Nezafati, K. (2019). Snowboot: bootstrap methods for network inference. *arXiv preprint arXiv:1902.09029*.

Chen, Y.-C., Wang, Y. S., and Erosheva, E. A. (2018). On the use of bootstrap with variational inference: Theory, interpretation, and a two-sample test example. *The Annals of Applied Statistics*, 12(2):846–876.

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

Daenekindt, S. and Huisman, J. (2020). Mapping the scattered field of research on higher education. a correlated topic model of 17,000 articles, 1991–2018. *Higher Education*, 80(3):571–587.

De Battisti, F., Ferrara, A., and Salini, S. (2015). A decade of research in statistics: A topic model approach. *Scientometrics*, 103(2):413–433.

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.

Hansford, T. G. and Spriggs, J. F. (2006). *The politics of precedent on the US Supreme Court*. Princeton University Press.

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

Hidayatullah, A. F. and Ma'arif, M. R. (2017). Road traffic topic modeling on twitter using latent dirichlet allocation. In *2017 international conference on sustainable information engineering and technology (SIET)*, pages 47–52. IEEE.

Imai, K., Lo, J., and Olmsted, J. (2016). Fast estimation of ideal points with massive data. *American Political Science Review*, 110(4):631–656.

Larsson, O., Naurin, D., Derlén, M., and Lindholm, J. (2017). Speaking law to power: the strategic use of precedent of the court of justice of the european union. *Comparative Political Studies*, 50(7):879–907.

Levin, K. and Levina, E. (2019). Bootstrapping networks with latent space structure. *arXiv preprint arXiv:1907.10821*.

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.

Liu, Y., Niculescu-Mizil, A., and Gryc, W. (2009). Topic-link lda: joint models of topic and author community. In *proceedings of the 26th annual international conference on machine learning*, pages 665–672.

Lupu, Y. and Fowler, J. H. (2013). Strategic citations to precedent on the us supreme court. *The Journal of Legal Studies*, 42(1):151–186.

Lupu, Y. and Voeten, E. (2012). Precedent in international courts: A network analysis of case citations by the european court of human rights. *British Journal of Political Science*, 42(2):413–439.

McCallum, A., Wang, X., and Corrada-Emmanuel, A. (2007). Topic and role discovery in social networks with experiments on enron and academic email. *Journal of Artificial Intelligence Research*, 30:249–272.

Nallapati, R. M., Ahmed, A., Xing, E. P., and Cohen, W. W. (2008). Joint latent topic models for text and citations. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 542–550. ACM.

Newman, M. E. (2001). Clustering and preferential attachment in growing networks. *Physical review E*, 64(2):025102.

Pelc, K. J. (2014). The politics of precedent in international law: A social network application. *American Political Science Review*, 108(03):547–564.

Raftery, A. E., Niu, X., Hoff, P. D., and Yeung, K. Y. (2012). Fast inference for the latent space network model using a case-control approximate likelihood. *Journal of computational and graphical statistics*, 21(4):901–919.

Rice, D. R. (2017). Issue divisions and us supreme court decision making. *The Journal of Politics*, 79(1):210–222.

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., Albertson, B., and Rand, D. G. (2014). Structural topic models for open-ended survey responses. *American Journal of Political Science*, 58(4):1064–1082.

Sommer, S., Schieber, A., Heinrich, K., and Hilbert, A. (2012). What is the conversation about?: A topic-model-based approach for analyzing customer sentiments in twitter. *International Journal of Intelligent Information Technologies (IJIIT)*, 8(1):10–25.

Spaeth, H. J., Epstein, L., Martin, A. D., Segal, J. A., Ruger, T. J., and Benesh, S. C. (2020). Supreme court database, version 2021 release 01.

Wang, M., Yu, G., and Yu, D. (2008). Measuring the preferential attachment mechanism in citation networks. *Physica A: Statistical Mechanics and its Applications*, 387(18):4692–4698.

Yang, S. and Zhang, H. (2018). Text mining of twitter data using a latent dirichlet allocation topic model and sentiment analysis. *International Journal of Computer and Information Engineering*, 12(7):525–529.