

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

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

This draft: July 13, 2022

Abstract

Topic modeling is one of the most popular approaches to statistical text analysis in many fields, especially in social sciences. An important feature of text data in social sciences is that many corpora consist of document networks in which documents cite other ones. We develop a new topic model of both text and citations, which allows researchers to analyze semantic context from which citations arise. In the proposed paragraph-citation topic model, topics are assigned to paragraphs, instead of tokens, and the topic of a paragraph determines the distribution of words and the probability of previous documents being cited in the paragraph. We demonstrate the model by applying to a corpus of the majority opinions of the Supreme Court of the United States. The model distinguishes topic-based subnetworks of citations, suggesting that the reason why a court cases are cited changed over time.

*Prepared for presentation at the 39th annual summer meeting of the Society for Political Methodology held at Washington University in St. Louis, July 2022. We thank Kevin Quinn and Stuart Benjamin for their comments on the draft.

[†]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](https://github.io/shiraito).

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. In political science, for example, Rice (2017) uses LDA to analyze the United States Supreme Court (USSC) opinions and demonstrates that the model identifies known legal topics, such as search and seizure, right to counsel, and death penalty.

One limitation of the existing models is the assumption that documents are unconnected despite the fact that documents in many text data sets are connected through a citation network. Citation networks carry rich information about the content of the documents, because a document is likely to cite another document with similar topics and with higher authority. For instance, the USSC opinions cite previous cases to apply, distinguish from, or overturn precedents, and the importance of citation networks in the USSC is widely recognized in existing studies. Clark and Lauderdale (2012) uses the citation network to construct a “genealogy of law” that succinctly summarises the development of legal doctrines. Fowler and Jeon (2005) and Fowler et al. (2007) construct a measure of important precedents from the citation network. These studies using statistical network analysis suggest the potential for using citation networks to identify latent topics of connected documents. However, the analysis of text data has been disconnected from the analysis of network data in existing research.

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 utility that models known factors that increase the probability of citations, such as the topical similarity between citing and cited documents and the authority of cited documents. Because the PCTM assigns topics to not only text but also citation links, topic assignments in a document may be informed by those in other documents through citations. Also, by analyzing the citation topics, researchers can explore sub-networks each of which is associated with a set of words representing its semantic context.

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 longer documents 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 the other paragraphs that are assigned to the same topic provide information on the semantic context of a citation. In addition, by uncovering the heterogeneity of topics across paragraphs, PCTM allows us to identify the different contexts around citations within the same document. These advantages are demonstrated in an application.

We use a corpus of the USSC’s majority opinions for the application in this paper. While we believe that the PCTM is applicable to many other data sets, the USSC 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 data set, because the data need to record the location of citations. The form of citations in the USSC decisions allows us to overcome this challenge by string matching.

The remainder of the paper is organized as follows. Section 2 introduces a new dataset of text and citation network of the USSC opinions we constructed for this project. Section 3 describes our model, PCTM, and its inference. Section 4 shows simulation results to verify the performance of PCTM. Section 5 presents results of applying PCTM on the USSC 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 USSC opinions that combine text and citation networks. The original data is obtained from the Caselaw Access Project¹, which allows public access to all the official and published US case law, including all state courts and federal courts. The data contains full text of the opinions, both majority and dissenting 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 only and discard dissenting opinions, because the majority opinions are the most important as legal precedents. In total, the population data contains 24,000 cases with 749,888 paragraphs with the year ranges from

¹<https://case.law>

1834 to 2013.

The document networks of the USSC consist of two forms of datasets: text and citation networks. Regarding 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 ended up containing 32,644 unique features.

The other component is a citation network. While previous studies have constructed citation networks of the USSC 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 a paragraph p of a document i cites a document j , and 0 otherwise. Since such data is not readily available, we constructed our own citation network of the USSC 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 in multiple reporters, there are several possible ways to cite the same cases. 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. Luckily, 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 USSC cases only, the citations towards and from outside of the corpus (e.g. citations to and from court of appeals and states courts) were discarded. This results in 191,173 citations in total.

In this paper, we focus on subsets of this datasets. 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

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

in the entire corpus, but it appears in almost all documents of *Federal Taxation* issue area. As a result, we obtain two subsets we use for our applications. The first subset consists of cases classified as *Privacy* issue area, which includes decisions about abortion. We chose this as our primary application data because existing literature of citation networks of the USSC cases often focus 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 and effectively ends 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 about *Voting Rights*. The subset about *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 what process do the 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 utility to model authors’ decision to make citations in relation to the topic structure. The latent citation utility 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 topic assignment. 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 is likely to be referring documents of very similar, if not the same, topic prevalence. For example, an opinion in the USSC typically identifies multiple legal doctrines that apply to a given case, and address them in different paragraphs. Therefore, citations within one paragraph are likely to be a collection of opinions that address the same legal doctrine. In other words, citations in different paragraphs are likely to be references of different legal contexts. We believe such characteristics are not limited to legal documents

of the USSC, but a general feature of any document networks, and they should be reflected in the process of uncovering *topic structure*. Below, we delineate our modeling strategy that addresses the above questions in details.

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 p th paragraph in i th document towards j th 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 utility

$$D_{ipj}^* \sim \mathcal{N}(\boldsymbol{\tau}^T \mathbf{x}_{ipj}, 1)$$

Draw citation

$$D_{ipj} = 1 \text{ if } D_{ipj}^* \geq 0 \text{ and } 0 \text{ otherwise}$$

where \mathbf{x}_{ipj} is a vector of covariates that shape the latent citation utility for p th paragraph in document i to cite j document. Existing studies of strategic citation of precedents commonly point to the importance of authority of an opinion as one of the major attracting factor of citations (Hansford and Spriggs, 2006; Lupu and Voeten, 2012; Lupu and Fowler, 2013). This is consistent with a well-known dynamics in social networks called “preferential attachment” (Newman, 2001; Wang et al., 2008). Following the literature of strategic citation of precedents, we include in \mathbf{x}_{ipj} the indegree of j th document $\kappa_j^{(i)}$ to represent the authority of the j th document at the time of i ’s writing (Hansford and Spriggs, 2006; Lupu and Voeten, 2012; Lupu and Fowler, 2013). Since indegree values of a document continues to change as more documents enter the network, we introduce the superscript (i) in $\kappa_j^{(i)}$ to denote that the given value is the observed indegree when document i was being written. We also include $\eta_{j, z_{ip}}$ to capture the topic prevalence of j th document to gauge topic distance between citing paragraph and cited document. If the preferential attachment is prevalent in the network, we expect τ_1 to be positive. Likewise, if citations are more likely to appear between docu-

ments of similar topics, we expect τ_2 to be positive. Finally, we include the sparsity term τ_0 as an intercept of the regression, which is expected to be negative in general. While we currently include 3 document level covariates in \mathbf{x} , researchers can add other covariates that are considered relevant for citation decisions in \mathbf{x} .

Given the data (words and citations, \mathbf{W}, \mathbf{D}), our posterior probability is

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

3.2 Related Models

Three features distinguish PCTM from existing models for document networks. First, 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, especially when directions are important structural aspect of the data. In citation networks direction of citations show temporal ordering of the documents. 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 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, PCTM allows a document to send multiple citations – possibly of different topics – to another document. Most recent models for document networks such as RTM (Chang and Blei, 2009), Topic-link LDA (Liu et al., 2009) and pairwise citation LDA (Nallapati et al., 2008) commonly model linkages between documents as binary process – whether the given pair of documents are connected or not. On the other hand, in 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 extent 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, PCTM incorporates regression structure to model a paragraph’s latent citation utility over all precedents and thereby offers flexibility for researchers to model strategic

citation dynamics they theorize simply by adding variables of their interest to the regression. PCTM in the current form includes two factors by default: authority of the precedent, and the topic similarity between the citing paragraph and the cited document. Existing models for document networks commonly model citations as solely the function of topic defined at the word level. 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 political processes involved to whether and how documents receives citations and how often.

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 (Add citations here).

To remedy this problem, we follow the recent advances in the inference of CTM models 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

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

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

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

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 ip th 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

$$\begin{aligned}
\pi_{ipj,k} = \exp \left\{ \eta_{ik} + \log \prod_v \Gamma(\beta_v + c_{k,ip}^v + c_{k,-ip}^v) - \log \Gamma\left(\sum_v \beta_v + c_{k,ip}^v + c_{k,-ip}^v\right) \right. \\
\left. - \frac{1}{2} \left(\tau_2^2 \eta_{jk}^2 + 2(\tau_0 \tau_2 + \tau_1 \tau_2 \kappa_j^{(i)} - \tau_2 D_{ipj}^*) \eta_{jk} \right) \right\}
\end{aligned} \tag{5}$$

Here, $c_{k,ip}^v$ denotes the total number of times the v th word appears in paragraph ip of topic k such that $c_{k,ip}^v = \sum_{l=1}^{n_{ip}} \mathbb{I}(W_{ipl} = v) \mathbb{I}(z_{ip}^k = 1)$. Likewise, $c_{k,-ip}^v$ is the total number of times the v th term appears in paragraphs with k th topic except for ip .

The conditional posterior distribution of $\boldsymbol{\eta}$ for i th document is jointly defined with the augmenting Polya-Gamma distribution with $\boldsymbol{\lambda}$. 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 i th document as follows.

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

where

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

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

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

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

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

$$\begin{aligned}p(\boldsymbol{\tau} | \boldsymbol{\eta}, \mathbf{Z}, \mathbf{D}^*) &\propto \exp \left\{ -\frac{1}{2} \sum_{ipj} \left(D_{ipj}^* - \mathbf{x}_{ipj}^T \boldsymbol{\tau} \right)^2 \right\} N(\boldsymbol{\mu}_{\boldsymbol{\tau}}, \boldsymbol{\Sigma}_{\boldsymbol{\tau}}) \\ &\propto N(\tilde{\boldsymbol{\tau}}, \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{\tau}})\end{aligned}\quad (10)$$

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

3.4 Initialization Strategy

Similarly with other topic models, PCTM contains a number of parameters for 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 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 $(\mathbf{Z}, \boldsymbol{\lambda}, \mathbf{D}^*, \boldsymbol{\tau})$. The details of our strategy to generate initial values of the parameters are available in appendix C.

4 Simulation Results

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

We generate 100 simulation data similar to our application data so that we can add confidence to the performance of PCTM on our application data. The hyperparameters of PCTM are chosen so that the resulting simulation data resemble our main application data: the USSC opinion on the privacy issue. This allows us to examine how well PCTM performs compared to existing models, when it is applied to such datasets.

Specifically, the following hyperparameters are identical between the simulation datasets and the application data: the number of documents (106), the number of paragraphs per document (min=1, max=150, mean=44), the number of unique words (5838), the total number of words in the entire corpus (238467), the number of words per paragraph (min=15, max=237, mean=51). By contrast, the number of citations in the simulation data is not always identical to the one of the application data. This is because the number of citations in the simulation data varies across datasets since it is not one of the hyperparameters we can impose to the model. However, we set the priors parameters (prior of τ , in particular) so that the simulation data becomes close to the application data. As a result, the number of citations in the application data is 452 while the average number of citations across simulation datasets is 454.

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

Figure 1 shows the comparison of the estimated and the true paragraph topic, \mathbf{Z} . 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 color mean the higher number of paragraphs. The concentration on the diagonal elements means that the topics are estimated correctly. As a comparison, the left panels replaces the estimated topics with the topics at the initialization. They show that PCTM can recover the true topics even when the topics are randomly provided at the initialization of our Gibbs sampler.

Figure 2 shows the comparison of the estimated and the true topic distribution of documents, $\boldsymbol{\theta}$. On the right panel, the (k, l) cell shows the number of documents whose mode

of the estimated topic distribution, θ , across K topics is l while the mode of the true topic distribution is k . We obtain θ by applying logistic transformation on each draws of η in our Gibbs sampler, and then obtain the estimated θ by their posterior mean. The cells with darker color mean the higher number of documents are at the cell. The concentration on the diagonal elements means that the modes of the topic distributions are estimated correctly. As a comparison, the left panels replaces the estimated θ with those at the initialization. It shows that PCTM can recover the true mode of the topic distribution even when the topics are randomly provided at the initialization of our Gibbs sampler.

These two results verify that 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.

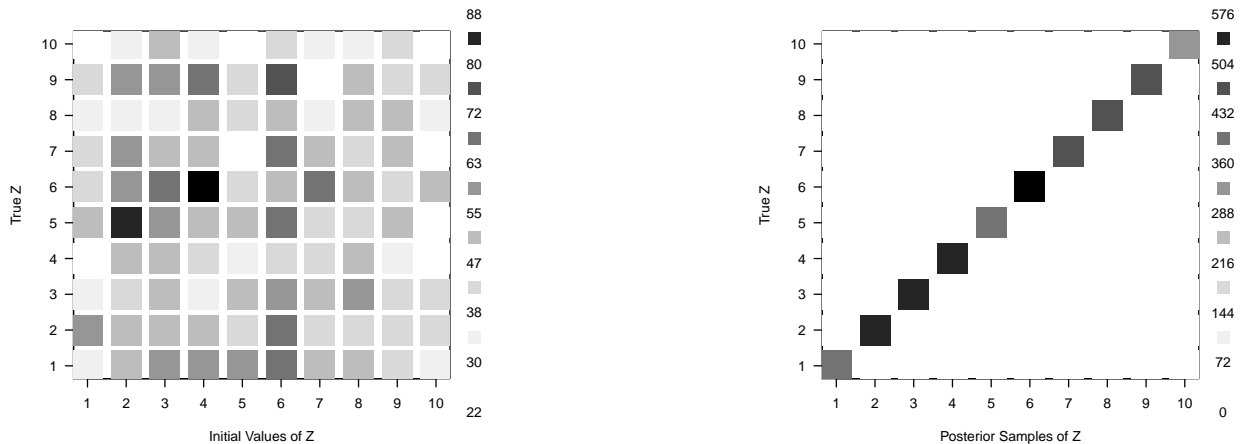


Figure 1: The comparison of the estimated and the true topics of paragraphs. On the right panel, the (k, l) cell shows the number of paragraphs whose estimated topic is l while the true topic is k . We estimate topics using the paragraph topic parameter, \mathbf{Z} , using the last draw from our Gibbs sampler. The cells with darker color mean the higher number of paragraphs. The concentration on the diagonal elements means that the topics are estimated correctly. As a comparison, the left panels replaces the estimated topics with the topics at the initialization. They show that PCTM can recover the true topics even when the topics are randomly provided at the initialization of our Gibbs sampler.

Next, we validate the performance of 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) + 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 two-step approach to model the generation of text and citations separately while the

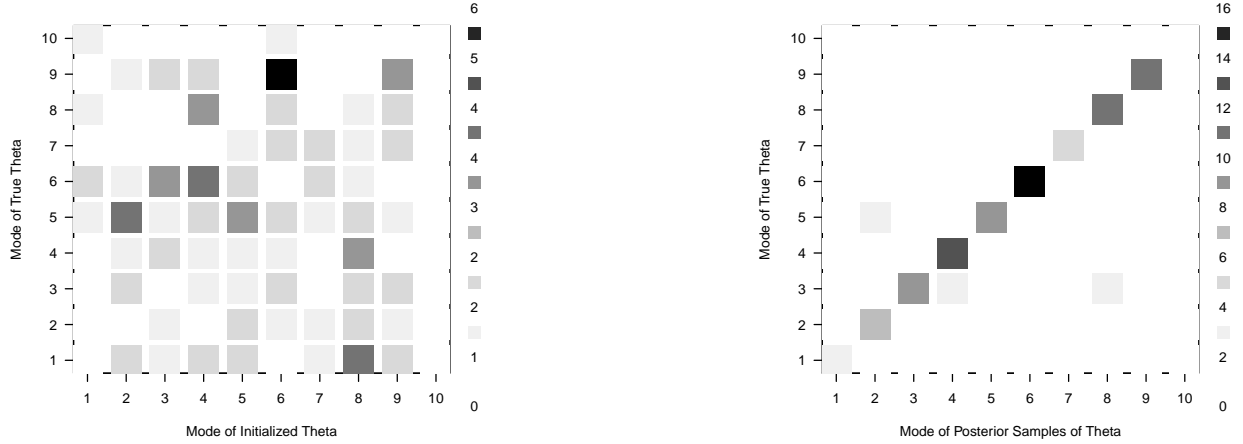


Figure 2: The comparison of the estimated and the true topic distribution of documents. On the right panel, the (k, l) cell shows the number of documents whose mode of the estimated topic distribution, θ , across K topics is l while the mode of the true topic distribution is k . We obtain θ by applying logistic transformation on each draws of η in our Gibbs sampler, and then obtain the estimated θ by their posterior mean. The cells with darker color mean the higher number of documents are at the cell. The concentration on the diagonal elements means that the modes of the topic distributions are estimated correctly. As a comparison, the left panels replaces the estimated θ with those at the initialization. It shows that PCTM can recover the true mode of the topic distribution even when the topics are randomly provided at the initialization of our Gibbs sampler.

former jointly models text and citations (Chang and Blei, 2009).³ The goal is to compare the posterior predictive probability of a new document given past documents across the three models. The higher predictive probability indicate a better model fit.

The following is the procedure of the simulation exercise. We fit the three models, PCTM, RTM, and LDA + Logistic Regression on a 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 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 datasets, and compare the predictive probabilities across models. The following gives the posterior predictive probability for 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 $\hat{\cdot}$ symbol means that they are samples from the posterior distributions from the model trained with $\mathbf{W}^{train}, \mathbf{D}^{train}$. We

³We fit LDA and RTM using an R package, `lda`.

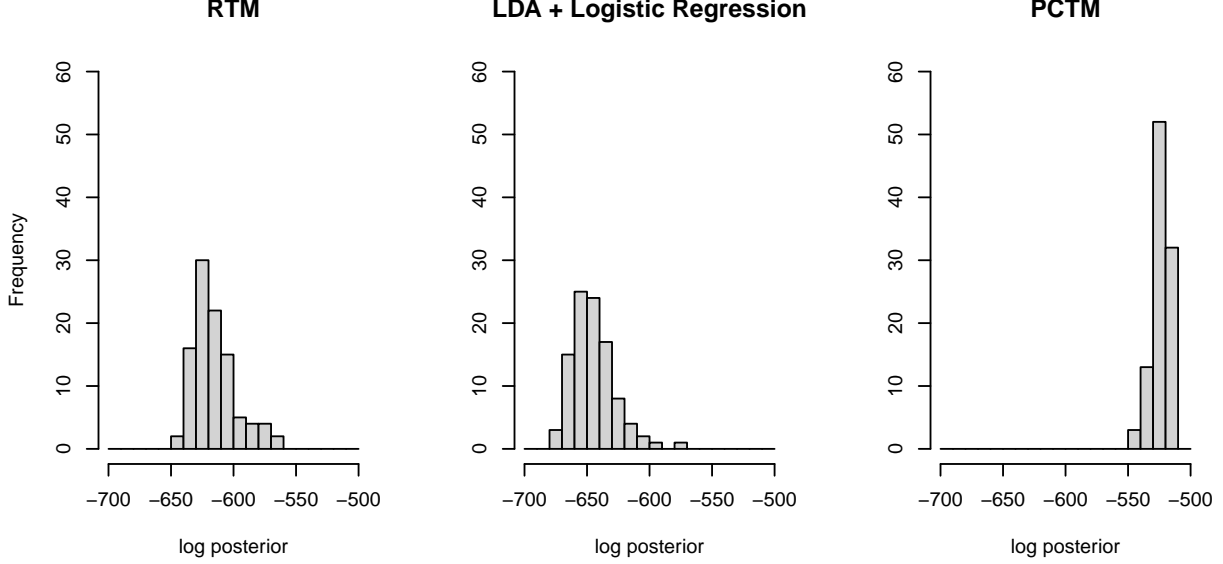


Figure 3: Comparison of the predictive probabilities of Relational Topic Model (RTM), Latent Dirichlet Allocation (LDA) + Logistic Regression, and PCTM (from left to right). Distribution of the log posterior predictive probabilities across 100 simulation datasets. Overall, PCTM has higher posterior predictive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

draw 1000 samples from the posterior for those parameters, and compute the average to obtain the final predictive probability.

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

Figure 3 shows the histogram of the predictive probabilities of the last document across RTM, LDA + Logistic Regression, and PCTM. It shows that the predictive probabilities of PCTM is almost always higher than the other two models. This means that PCTM fits to the type of data we use in our application better than those existing models. One caveat of this exercise is that PCTM is advantageous by its setup because the simulation data is generated following the generative process of PCTM. A more robust test should perform the same exercise but using the application data.

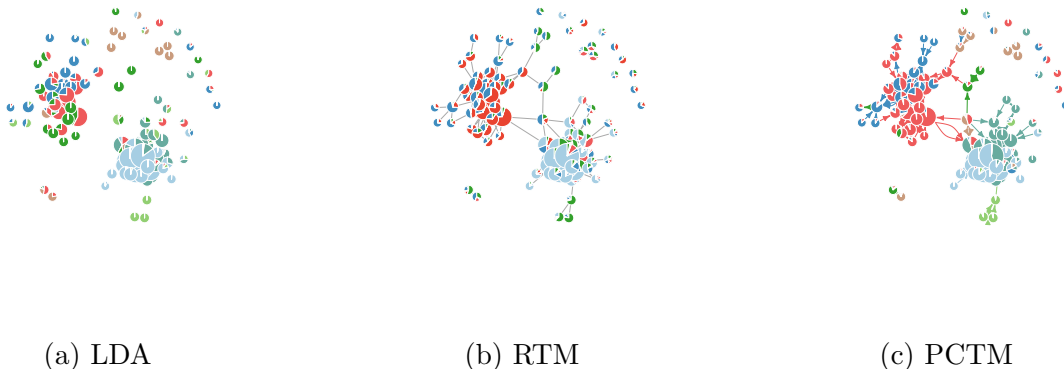


Figure 4: The result of three topic models, LDA, RTM, and PCTM from (a) to (c), on the US Supreme Court opinions of *Privacy* issue area. A node represents an opinion, and an edge represents a citation between opinions. The nodes are colored according to the proportion of paragraphs with the same estimated 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.

5 Empirical Results

This section presents the results of applying PCTM on the USSC dataset, and compare 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 PCTM on the entire USSC 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 connection 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 a

document a citation is made (thus, all edge colors are in gray). While RTM’s assumption on 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 directed, acyclic and documents often cite another document multiple times.

As seen in (c) of Figure 4, 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 PCTM are given in Table 1. The Supreme Court Database assigns 4 issue codes to opinions of Privacy issue area⁴, but we identify 7 distinct topics in PCTM. The labels in the table are provided by the authors.

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.

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 abortion as woman’s right and discusses 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

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

paragraphs as the unit for topic assignment. Almost all majority opinions in USSC 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, PCTM tends to assign a topic for this category. Paragraphs of the fourth topic mostly concerns 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 the seventh topic both concern the public disclosure of private information. The sixth topic, which we label as **Privacy vs Government Interest**, mainly address the access to private information such as the history of drug abuse that might disrupt operations of government agencies. The seventh topic, on the other hand, concerns whether the way private information is recorded constitute violation of Privacy Act of 1974.




Regulation of Abortion Procedures	Procedural Posture	Const. Rights to Abortion
		
<p>... 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) ...</p>	<p>...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 ...</p>	<p>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)</p>

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 personal information(history of drug abuse) of its employees whereas for **US v. RCFP**, **Whalen v. Roe** was mainly about 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 those information. This highlights that the semantic context of citations may differ even when the given citations refer to the same document.

Another advantage of 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




Speech & Protest	Privacy vs Govnt. Interest	Public Disclosure of Private Information
		
<p>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 anti-abortion protesters. (We discuss Madsen in greater depth in Part II-A, <i>infra</i>.) 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.</p>	<p>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 <i>Engquist</i>, 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.</p>	<p>... 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." <i>Id.</i>, 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":</p>

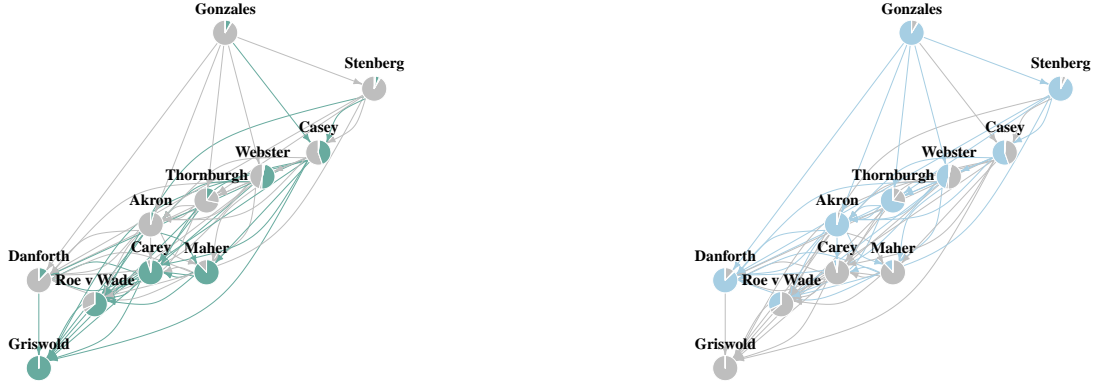
Table 3: Paragraphs containing citations of Topics 4,6, and 7. The top row displays two opinions and a citation with color-coded topics. The second row for each topic contains the paragraph that contains the citation between the two opinions in the first row.

Reproductive rights in Figure 5.⁵

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 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 afterwards subsequently focus 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 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."⁶

⁵The 11 opinions on reproductive rights are selected based on Figure 4 of Clark and Lauderdale (2012).

⁶<https://reproductiverights.org/our-work/landmark-cases/>



(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 USSC 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 indicates older in time and opinions placed in the upper part of the figure are more recent documents.

5.2 Voting Rights

The USSC 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.⁷ 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 **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 is eligible to cast his vote on 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

⁷The three issue codes on voting are voting, Voting Rights Act of 1965, Ballot Access.

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.

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 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 6 shows the cumulative count of paragraphs of each topic. The growth 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 7 show groups of cases that make citations of the given topic. The location of cases

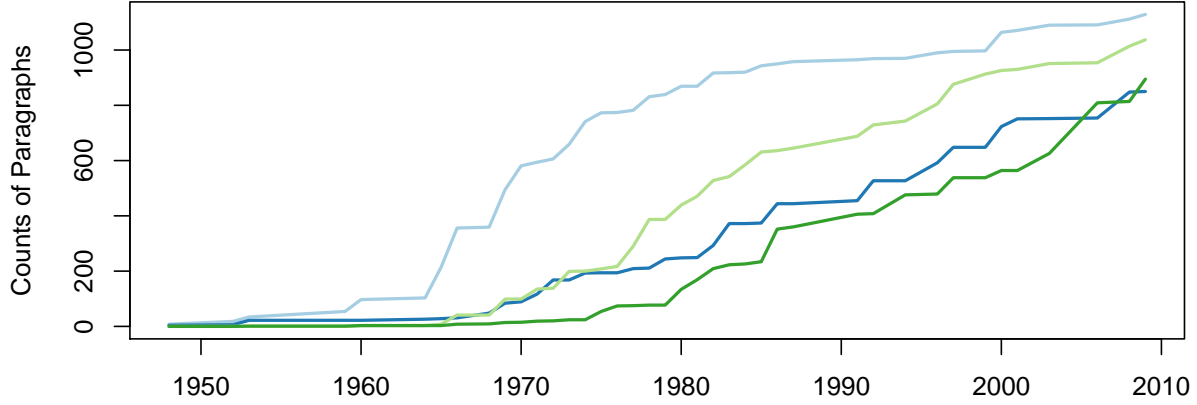


Figure 6: Cumulative number of topics in Voting Rights subset over time.

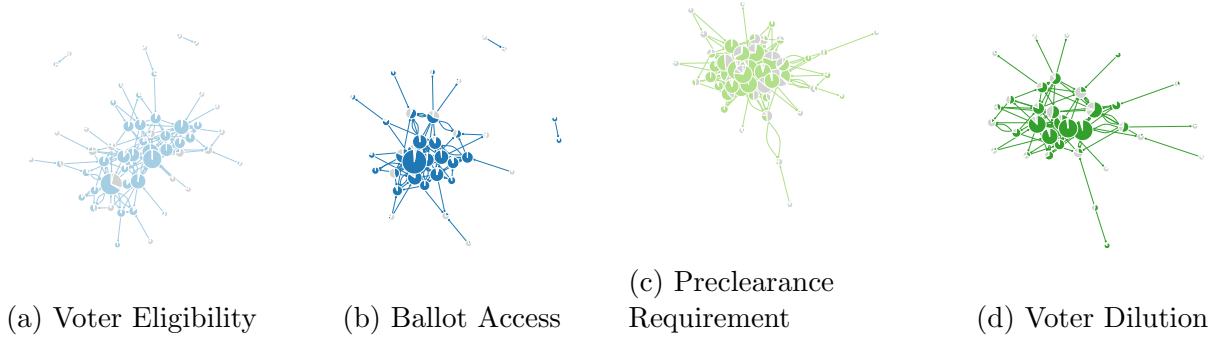


Figure 7: The subnetwork specific to each topic. The subnetworks are created by extracting opinions that either send or receive citations of the given topic. The topic-specific subnetworks can be useful in revealing whether and the extent to which topological features of the network varies by topic. For each subnetwork, paragraphs of other topics are all colored in gray for better visualization.

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 time, closely building up on past cases of the same topic whereas opinions on **Voter Eligibility** have ex-

panded more independently and incrementally over a longer period of time.

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 USSC 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 data sets 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 topic words associated with each citation, since words in the same document do not provide any additional information about citations conditional on the document’s topic mixture in these models. Our proposed PCTM overcomes this limitation by modeling paragraphs as the unit of topic assignment, and we applied the model to the USSC decisions. The proposed modeling strategy allows us to analyze words that represent the semantic context of citations. Our application demonstrates this advantage of the PCTM by finding two sub-networks of citations among cases on reproductive rights.

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.
- 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.
- Clark, T. S. and Lauderdale, B. E. (2012). The genealogy of law. *Political Analysis*, 20(3):329–350.
- Fowler, J. H. and Jeon, S. (2005). The authority of supreme court precedent: a network analysis. *Preprint as of June*, 29:2005.
- 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.
- 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.
- 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.
- 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.
- 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.
- 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.