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

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

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

Abstract

Social scientists analyze citation networks to study how documents influence subsequent work across various domains such as judicial politics and international relations. However, conventional approaches that summarize document attributes in citation networks often overlook the diverse semantic contexts in which citations occur. This paper develops the paragraph-citation topic model (PCTM), which analyzes citation networks and document texts jointly. The PCTM extends conventional topic models by assigning topics to paragraphs of citing documents, allowing citations to share topics with their embedding paragraphs. Our empirical analysis of U.S. Supreme Court opinions in the privacy issue domain, which includes cases on reproductive rights, demonstrates that citations within individual documents frequently span multiple substantive areas, and citations to individual documents show considerable topical diversity.

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

†These authors have contributed equally to this work.

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

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

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

1 Introduction

Social scientists often use citation network data to study the influence of documents, such as academic articles, books, laws, and court opinions. Research in judicial politics has analyzed the citation networks of the United States Supreme Court (SCOTUS) opinions, revealing how some cases exert significant influence on future rulings (Clark and Lauderdale, 2012; Fowler et al., 2007). Similarly, in international relations, scholars explore how citations shape power dynamics in areas like trade (Pelc, 2014), human rights (Lupu and Voeten, 2012), and jurisdictional conflicts (Larsson et al., 2017).

Conventional approaches seek to summarize document attributes within a network, but often overlook the diverse semantic contexts in which citations occur. Since the semantic content of documents influences citation network structures (Bai et al., 2018; Chang and Blei, 2010; Zhang and Lauw, 2022), accounting for semantic heterogeneity in document networks can reveal information that might otherwise remain hidden. The measures of precedential importance for various courts of law, for instance, implicitly treat the absence of citations as a reflection of limited precedential value rather than a potential semantic disconnect between documents (Fowler et al., 2007; Lupu and Voeten, 2012; Pelc, 2014). However, a high volume of citations to a court case may indicate its status as a landmark case, the popularity of legal topics it addresses, or both.

Recognizing the importance of semantic heterogeneity in document networks, previous studies have used human-coded topics to ensure semantic coherence, restricting their analyses to documents within discrete semantic domains such as criminal justice (Olsen and Küçüksu, 2017) or reproductive rights (Clark and Lauderdale, 2012). However, human coding often captures broad categories, leaving significant semantic variation within these groups unaddressed. Also, researchers may wish to automatically detect semantic heterogeneity at the granularity that fits their research purpose, or the semantic context itself can be of research interest rather than an object to control for.

This paper develops a Bayesian topic model that systematically integrates citation network and document text. Our proposed model, the paragraph-citation topic model (PCTM), extends conventional topic models by assigning a topic to each paragraph of the citing document, allowing citations to share topics with text of the paragraphs that they are in. This marks a departure from other topic models for document networks (i.e. Relational Topic Models) by allowing citations in one document to have heterogeneous topics. Our empirical analysis demonstrates that citations within individual documents frequently span multiple substantive areas. Moreover, our findings reveal considerable topical diversity in citations to individual documents, illustrating how a single opinion can intersect multiple domains of legal discourse (i.e., *Roe v. Wade* engages with various legal issue areas, including civil procedure, constitutional law, healthcare policy, privacy rights, and beyond).

1.1 Related Models

A growing body of scholarship has developed models for joint analysis of texts and citation networks (Chang and Blei, 2010; Liu et al., 2009; Bai et al., 2018; Le and Lauw, 2014; Zhang and Lauw, 2020). Early LDA-based approaches leverage citations to improve topic estimation, with semantically similar documents more likely to be connected through citations (Chang and Blei, 2010; Liu et al., 2009; Nallapati et al., 2008). More recent advances employ deep learning techniques to represent texts and citations in lower-dimensional latent embedding spaces (Bai et al., 2018; Zhang and Lauw, 2022). The PCTM extends this growing literature in three substantive ways.

First, the PCTM assigns topics to paragraphs rather than individual tokens. This modeling strategy stems partly from the observation that paragraphs written by trained professionals often represent coherent units of idea, but more importantly, it is the modeling choice that allows researchers to identify the semantic context of each citation by finding a topic (i.e., a distribution of words) within which the citation is embedded. Existing models, by contrast, do not have direct connections between a citation and words around it. In Chang and Blei (2010) and Liu et al. (2009), the generative process of citations is based on the mixture of topics in the entire document, rather than assigning topics to individual citations. The Pairwise Link-LDA model and the Link-PLSA-LDA model developed by Nallapati et al. (2008) assign topics to individual citations, but these topics are conditionally independent of the topics assigned to words given the document-level parameters. The PCTM is unique in that it explicitly takes into account the proximity of citations to words in the same paragraph, allowing for a more nuanced understanding of the semantic context of citations.

Second, the PCTM allows a document to send multiple citations—possibly of different topics—to another document. Past research has focused on topic estimation in document networks, treating citations primarily as binary linkages between documents. Consequently, the semantic context of individual citations has remained largely unaddressed in existing models. While we build on previous work by utilizing citations to enhance topic estimation, our approach differs by explicitly modeling the semantic context of each citation. Specifically, the PCTM assigns topics to each paragraph and its embedded citations, allowing citations within the same document to represent distinct topics.

Finally, the PCTM models paragraph-level citation propensities through a regression framework, offering researchers flexibility in modeling strategic citation dynamics. This approach aligns with social science studies that emphasize how social and political processes influence citation patterns and frequency (Hansford and Spriggs, 2006; Lupu and Fowler, 2013; Pelc, 2014). In its current form, our model's regression layer incorporates both precedential authority and topic similarity between citing paragraphs and cited documents. Researchers can include any variables at the paragraph, document, or paragraph-document dyadic level to model strategic citation behavior.[1]

2 The United States Supreme Court Opinions

The SCOTUS as the highest judicial authority in the United States holds a pivotal role in social science studies with implications for social norms, public policy, and individual rights. At the center of the SCOTUS ruling is the principle of *stare decisis* in which a decision must "stand by things decided." The *stare decisis* establishes that precedents take crucial importance in the SCOTUS as they exert varying levels of influence on future rulings. In particular, landmark cases such as *Roe v. Wade* are a crucial subject of study in social sciences. With the *stare decisis*, a SCOTUS ruling is not just about the case at hand, but also about how to interpret the relevant precedents that together shape the boundary of social norms and behavior.

Due to the significance of precedents in the SCOTUS, many social science studies have been dedicated to exploring various aspects of precedents. Many scholars have focused on the political processes involved in the choice and the representation of precedents in the SCOTUS majority opinions (Hansford and Spriggs, 2006; Bailey and Maltzman, 2008; Clark and Lauderdale, 2010). How precedents are treated by future cases and eventually fade away was another focal point of research (Black and Spriggs, 2013; Broughman and Widiss, 2017).

Mapping the SCOTUS cases and citations into a network, past studies employed network analysis to measure the structural properties of precedents in the citation network. Clark and Lauderdale (2012), for instance, fits the latent tree model to the SCOTUS citation network and uncovers the hierarchy of precedents as an estimation of the evolution of legal doctrine. Another strand of research highlights the positions precedents take in the citation network (Fowler and Jeon, 2005; Fowler et al., 2007). Fowler et al. (2007) and Fowler and Jeon (2008) propose a variation of the eigenvector centrality score to gauge how legally "central" a case is for the SCOTUS at a given point in time.

[1] In this sense, our model is similar to the Structural Topic Model (STM) by Roberts et al. (2014) where exogenous covariates shape the topic prevalence of documents through a generalized linear model. One can imagine our model as a variation of STM where the regression layer includes endogenous processes of citation formation.

While recognizing the usefulness of the network analysis for the SCOTUS citation network, we find that existing approaches commonly overlook the topic heterogeneity of the citation network. Network analysis of the SCOTUS citation network treats presence and absence of citations as informative signals. In Fowler et al. (2007) and Fowler and Jeon (2008), for example, precedents that attract many citations are likely to be structurally central and precedents without many citations are considered to be peripheral. While the presence of citations can be an informative signal for the importance of a case, the absence of citations may simply be due to topic inconsistency rather than its importance. That is, we do not expect a case to cite a precedent if the given precedent addresses completely distinct legal topics. When the network analytic methods are applied to the universe of cases without special attention to the topic differences between them, one may mistakenly interpret the topic differences as indicators of importance.

Another point we highlight is that the topic space of an opinion is multidimensional. For example, *Roe v. Wade* is mostly known for the right to privacy in abortion, but it also addresses other legal topics such as substantive due process, end-of-life decisions, and legislative restraints. A citation to *Roe v. Wade* can be concerning the right to privacy, but it could also be about other topics such as legislative restraints. This suggests that subsetting down to a broad legal category of cases for network analysis, such as seen in Clark and Lauderdale (2010) where authors limit their scope to search and seizure and the freedom of religion opinions, may not be sufficient to capture nuanced legal topics that a case touches upon.

To address the above key challenge, we propose to incorporate the text of the SCOTUS majority opinions with the citation network. In the following sections, we propose a model that incorporates both the text and the citation network of the SCOTUS majority opinions. Our model can uncover the topic structure of the majority opinions with topic model while utilizing the network linkage in the citation network. We apply our model to all privacy opinions in the SCOTUS and demonstrate that the resulting topic-homogenous citation subnetwork can be used for further network analysis.

For the application of our model, we obtain the universe of the SCOTUS majority opinions on the privacy issue area from the Caselaw Access Project[2]. Subsetting follows the issue area categorization provided by Supreme Court Database (Spaeth et al., 2020). The privacy issue area is chosen for our application because 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 Privacy opinions subset consists of

[2] `https://case.law`

106 documents with 4,669 paragraphs, 5,838 unique words, and 452 citations. More details of data pre-processing for each subset are available in Supplementary Information, Section A.

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: By what process do authors decide to cite another document? How does the topic structure enter into citation decisions, and conversely, how do citations help determine the 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. 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 refer 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 point 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

First, we introduce the notation. Let N, G, V, and K be the total number of documents, total number of paragraphs, and total number of unique words, and the number of topics, respectively. We use N_{ip} to denote the number of words in paragraph p of document i. Our data consist of words, $\mathbf{W}$, and citations, $\mathbf{D}$. $\mathbf{W}$ is a matrix of size $G \times V$ where each row is $\mathbf{w}_{ip}$, a vector of length V that represents the number of times each unique word appears in a paragraph p of document i. $\mathbf{D}$ is a matrix of size $G \times N$ where each element, D_{ipj} is a binary variable that indicates the existence of a citation from pth paragraph in ith document towards jth document. $\mathbf{D}^*$ is a matrix of size $G \times N$ and its element, D^*_{ipj}, is a latent variable that represents the latent citation propensity of pth paragraph in ith document to cite jth document. We have another latent variable $\mathbf{Z}$, a vector of length G where each element is z_{ip}, a scalar that takes a value from $\{1, \ldots, K\}$, and it represents the topic assignment of pth paragraph in ith document. We have three main parameters to estimate: $\boldsymbol{\eta}$, $\boldsymbol{\Psi}$, and $\boldsymbol{\tau}$. $\boldsymbol{\eta}$ is a matrix of size $N \times K$ where each row is $\boldsymbol{\eta}_i$, a vector of length K that represents the topic proportion of document i, generated from a multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. $\boldsymbol{\mu}$ is further generated from a normal distribution with mean $\boldsymbol{\mu}_0$ and covariance $\boldsymbol{\Sigma}_0$. $\boldsymbol{\Psi}$ is a matrix of size $K \times V$ where each row is $\boldsymbol{\Psi}_k$, a vector of length V that represents the word distribution of topic k. $\boldsymbol{\Psi}_k$ is generated from a Dirichlet distribution with parameter $\boldsymbol{\beta}$. $\boldsymbol{\tau}$ is a vector that represents the coefficients of the regression model that shapes the latent citation propensity, generated from a multivariate normal distribution with mean $\boldsymbol{\mu}_\tau$ and covariance $\boldsymbol{\Sigma}_\tau$.

The data-generating process is modeled as follows.

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

where $\mathbf{x}_{ipj}$ is a vector of covariates that shape the latent citation propensity for pth paragraph in document i to cite document j. $\mathbf{x}_{ipj}$ consists of 3 terms – the intercept, indegree, and $\eta_{j,z_{ip}}$, and $\boldsymbol{\tau} = [\tau_0, \tau_1, \tau_2]$ is a vector of coefficients. 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 "rich-get-richer" or, more technically, "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 presidents of the same party and 0 otherwise.

Given words and citations, $\mathbf{W}$ and $\mathbf{D}$, our posterior probability is

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

3.2 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 the additional advantage of computational speed. However, obtained parameters are for the variational distribution which is an approximation to the target posterior. The quality of the approximation is often not sufficiently explored. Furthermore, the variational inference is an optimization method that outputs point estimates. This requires additional steps to obtain a measure of 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 Model (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 $\mathbf{\Psi}$. 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 $\mathbf{\Psi}$ integrated out, our new posterior is proportional to

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

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

$$\begin{aligned} p(\mathbf{W}|\mathbf{Z}) &= \int_{\mathbf{\Psi}} p(\mathbf{W}, \mathbf{\Psi}|\mathbf{Z})d\mathbf{\Psi} \\ &= \int_{\mathbf{\Psi}} p(\mathbf{W}|\mathbf{\Psi}, \mathbf{Z})p(\mathbf{\Psi}|\mathbf{Z})d\mathbf{\Psi} \\ &= \int_{\mathbf{\Psi}} p(\mathbf{W}|\mathbf{\Psi}, \mathbf{Z})p(\mathbf{\Psi})d\mathbf{\Psi} \end{aligned} \tag{4}$$

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

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

where

$$\begin{aligned} \pi_{ipj,k} = \exp\Bigg\{ &\eta_{ik} + \log\prod_v \Gamma(\beta_v + c_{k,ip}^v + c_{k,-ip}^v) - \log\Gamma(\sum_v \beta_v + c_{k,ip}^v + c_{k,-ip}^v) \\ &-\frac{1}{2}\Big(\tau_2^2\eta_{jk}^2 + 2\big(\tau_0\tau_2 + \tau_1\tau_2\kappa_j^{(i)} - \tau_2 D_{ipj}^*\big)\eta_{jk}\Big)\Bigg\} \end{aligned} \tag{6}$$

We use $\mathbf{Z}_{-ip}$ and $\mathbf{W}_{-ip}$ to denote the set of all topic assignments and words except for the ipth paragraph, respectively. Here, $c_{k,ip}^v$ denotes the total number of times the vth 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 vth term appears in paragraphs with kth topic except for ip. The form of the conditional posterior for the ipth paragraph-level topic z_{ip}^k offers a convenient interpretation on the *source of information*. The first part $p(z_{ip}^k = 1|\boldsymbol{\eta}_i)$ displays the topic information from document-level topic prevalence. The second part represents topic information from the words in ipth paragraph. The third part $\prod_{j=1}^{i-1} p(D_{ipj}^*|z_{ip}^k = 1, \mathbf{Z}_{-ip}, \boldsymbol{\tau}, \boldsymbol{\eta}, \kappa)$ 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

$$p(\lambda_{ik}|\mathbf{Z}, \mathbf{W}, \boldsymbol{\eta}) \propto PG(N_i, \rho_{ik}) \tag{7}$$

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.

$$p(\eta_{ik}|\eta_{i,-k}, \mathbf{Z}, \mathbf{W}, \mathbf{D}, \boldsymbol{\tau}, \lambda_{ik}) \propto \mathcal{N}(\eta_{ik}|\tilde{\mu}_{ik}, \tilde{\sigma}_k^2) \tag{8}$$

where

$$\begin{aligned} \tilde{\sigma}_k^2 &= (\sigma_k^{-2} + \lambda_{ik} + v_{i,kk}^{-1})^{-1} \\ \tilde{\mu}_{ik} &= \tilde{\sigma}_k^2 \big(v_{i,kk}^{-1} m_{ik} + \sigma_k^{-2} \nu_{ik} + t_{ik} - \frac{N_i}{2} + \lambda_{ik} \log(\sum_{l \neq k} e^{\eta_{il}}) \big) \end{aligned} \tag{9}$$

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

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

$$p(D_{ipj}^*|\boldsymbol{\eta}, \mathbf{Z}, \boldsymbol{\tau}, \mathbf{D}) \propto \begin{cases} TN_{[0,\infty)}(\tau_0 + \tau_1 \kappa_j^{(i)} + \tau_2 \eta_{j,z_{ip}}, 1) & \text{if } D_{ipj} = 1 \\ TN_{(-\infty,0]}(\tau_0 + \tau_1 \kappa_j^{(i)} + \tau_2 \eta_{j,z_{ip}}, 1) & \text{if } D_{ipj} = 0 \end{cases} \tag{10}$$

where $TN_{[a,b)}(\mu, \sigma^2)$ denotes the truncated normal distribution with mean μ and variance σ^2 truncated to the interval $[a, b)$. The conditional posterior for $\boldsymbol{\tau}$ follows the following distribution. Let $\mathbf{x}_{ipj} = [1, \kappa_j^{(i)}, \eta_{j,z_{ip}}]^T$ and $\boldsymbol{\tau} = [\tau_0, \tau_1, \tau_2]^T$

$$\begin{aligned} p(\boldsymbol{\tau}|\boldsymbol{\eta}, \mathbf{Z}, \mathbf{D}^*) &\propto exp\bigg\{ -\frac{1}{2} \sum_{ipj} \Big(D_{ipj}^* - \mathbf{x}_{ipj}^T \boldsymbol{\tau} \Big)^2 \bigg\} N(\boldsymbol{\mu}_{\boldsymbol{\tau}}, \boldsymbol{\Sigma}_{\boldsymbol{\tau}}) \\ &\propto N(\tilde{\boldsymbol{\tau}}, \tilde{\boldsymbol{\Sigma}_{\boldsymbol{\tau}}}) \end{aligned} \tag{11}$$

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

Using simulation data, we confirm that the proposed Gibbs sampler recovers the true latent topics from random initialization. Our discussion about the initialization of the Gibbs sampler is presented in Supplementary Information, Section C, and the results of the simulation studies are presented in Supplementary Information, Section D.

4 Empirical Results

This section presents the results of applying the PCTM to the SCOTUS dataset, focusing on the privacy issue area.[3] We present three main results. First, we fit the PCTM and existing alternatives, LDA and RTM, to the SCOTUS opinions on the privacy issue area and discuss the advantages of the PCTM over the existing models.[4] We find that the main advantage of the PCTM is its ability to use paragraph-level topics to extract informative topic-specific subset of the citation network. Second, we utilize these topic-specific subnetworks to measure the importance of cases within each topic, following the methodological framework of Fowler et al. (2007). Our analysis reveals that case importance varies substantially across topic domains. Third, we conduct the predictive analysis of the topic structure of *Dobbs v. Jackson Women's Health Organization*, the recent case that overruled *Roe v. Wade*, based on words and citations in its paragraphs. We find that the predicted topics of *Dobbs v. Jackson Women's Health Organization* address abortion in markedly different ways from post-*Roe v. Wade* cases, but in ways reminiscent of pre-*Roe v. Wade* cases. Together, these results demonstrate the advantage of the PCTM in uncovering valuable insights from the text and citation data of the SCOTUS opinions.

4.1 Topic Composition of SCOTUS Opinions on Privacy

Table 1 displays the top 10 most frequent words for each topic estimated in the PCTM. The Supreme Court Database assigns 4 issue codes to opinions of the privacy issue area, but we identify 7 distinct topics in the PCTM.[5] The labels in the table are provided by the authors.

[3] We also present additional results with a dataset on voting rights issue area in Supplementary Information, Section E.

[4] The convergence diagnostics of the PCTM and discussions about the parameters not discussed in this section are provided in Supplementary Information, Section F

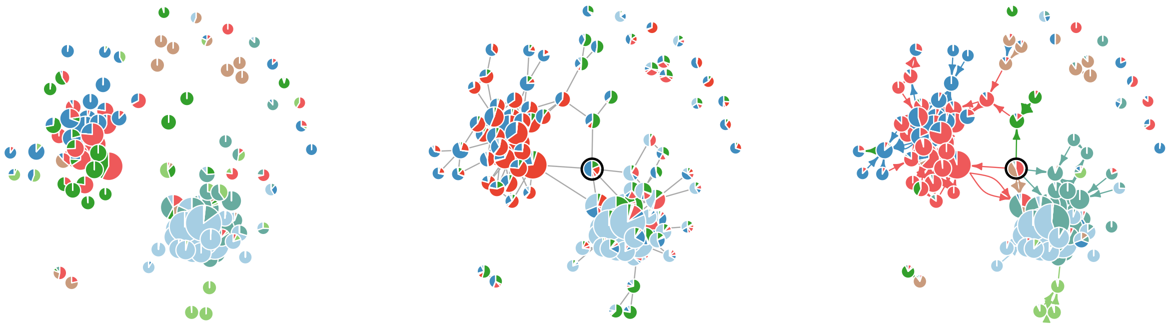
[5] The four issue codes identified by the Supreme Court Database are privacy, abortion, right to die and Freedom of Information Act. To determine the optimal number of topics for our analysis, we implemented an iterative approach, beginning with a 4-topics specification and systematically increasing the number of topics up to 15. We ultimately selected a 7-topics model as it provided the most coherent representation of legally salient themes within the privacy issue area, based on our substantive knowledge of constitutional law and privacy jurisprudence.

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 the PCTM.

The first and the third topics both address abortion as the substantive case in point but differ in the context in how abortion is addressed. Paragraphs of the first topic illuminate abortion as a woman's right and discuss the conditions in which the decision can be restricted or unrestricted, such as a woman's health, being a minor, or being ill-informed by her physician, 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 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.

Next, we compare the results of LDA, RTM, and PCTM on the privacy issue area of the SCOTUS opinions. Figure 1 displays the results of LDA, RTM, and the PCTM on the entire SCOTUS opinions on the privacy issue area. LDA assigns topics based on words without



(a) LDA (b) RTM (c) PCTM

Figure 1: The result of three topic models, LDA, RTM, and PCTM from (a) to (c), on the US Supreme Court opinions of the privacy issue area. A node represents an opinion, and an edge represents a citation between opinions. The color composition of a node follows the topic proportion of words (LDA, RTM) or paragraphs (PCTM) in the given opinion. The color 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.

reference to how documents are connected. RTM incorporates the networked structure of documents but assumes that connections between documents are undirected and binary. Moreover, RTM remains agnostic to the semantic context of citations since it does not consider their location within documents, which is reflected in the uniformly gray edges shown in Figure 1b.

By contrast, in Figure 1c, the PCTM assigns topics to paragraphs, which allows citations within the same document to have different topics. For example, focus on the case, *NASA vs. Nelson*, represented by the node at the center of the network highlighted by a black circle in Figure 1b and Figure 1c. In Figure 1c, it has six out-going edges colored differently according to the PCTM, which implies that the citations are made in the paragraphs addressing different topics. By contrast, the same case in Figure 1b has six edges colored gray, which means that RTM does not differentiate the topics of the citations. This showcases the advantage in the PCTM can provide a richer insight into the topic structure of the citations by identifying the topic of the paragraph in which a citation is made.

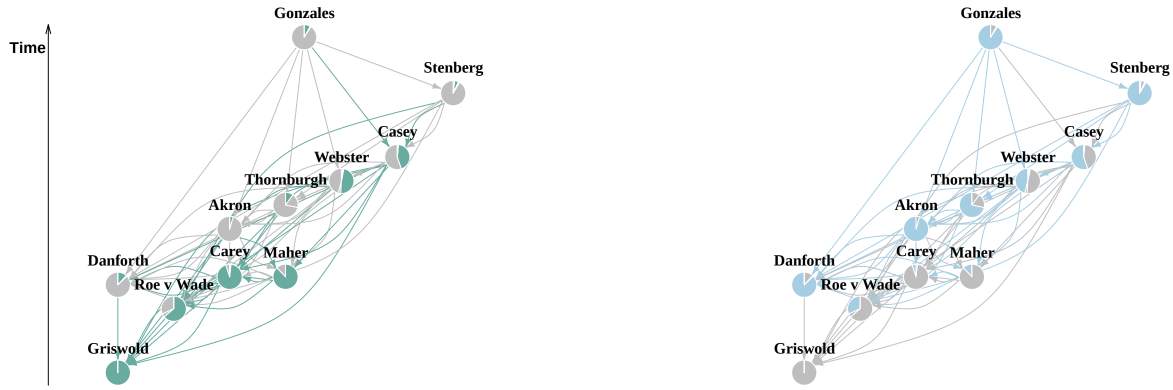
To highlight the advantage of the PCTM in finding heterogeneous semantic context

Privacy vs Govnt. Interest	Public Disclosure of Private Information
	
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 2: Paragraphs containing the same citations but assigned with different topics, `Privacy vs Government Interest` and `Public Disclosure of Information`. The top row displays a pair of opinions and a citation between the two color-coded by topics, and the left node is the citing opinion and the right node is the cited opinion. The second row for each topic contains the text of the paragraph where the citation is made in the two citing opinions in the first row.

around citations, we provide example paragraphs containing citations to the same case but with different topics in Table 2. Since Supreme Court cases typically address multiple legal domains, subsequent citations to these cases often engage with distinct aspects of their jurisprudence. For instance, *NASA v. Nelson* and *US v. RCFP* in Table 2 both cite *Whalen v. Roe*, but in distinct substantive contexts. 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*, citing *Whalen v. Roe* was mainly about the form of 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 demonstrates the semantic heterogeneity of citations even when they refer to the same document, the nuance that the PCTM can capture with paragraph-level topic assignment.

The PCTM also allows us to visualize the evolution of topics over time by extracting topic-specific subnetworks. We show how the topics on abortion (`Regulation of Abortion Procedure` and `Constitutional Rights to Abortion`) have changed over time. To emphasize this aspect, we extract from our citation network 11 selected opinions on Reproductive



(a) Constitutional Rights to Abortion

(b) Regulation of Abortion Procedures

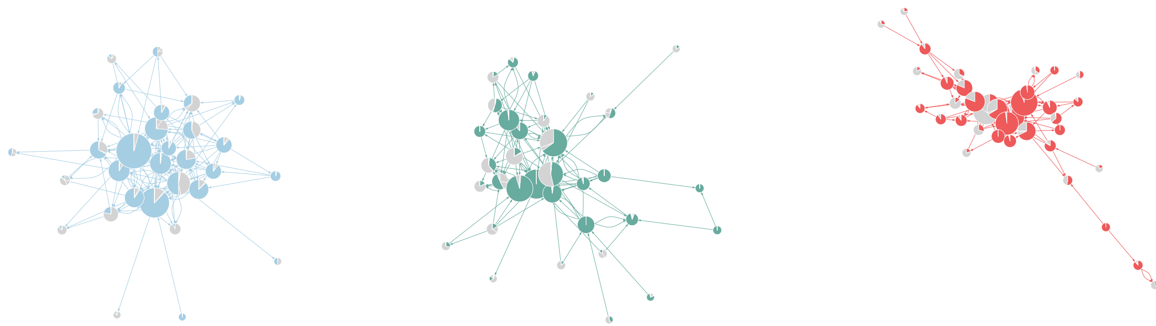
Figure 2: The citation network of 11 selected opinions on reproductive rights. The opinions are part of the SCOTUS subset on the privacy issue area. The left panel highlights the paragraphs and citations of `Constitutional Rights to Abortion` topic. The right panel colors the paragraphs and citations 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.

rights in Figure 2.[6]

Figure 2 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`. Earlier opinions predominantly focus on the `Constitutional Rights to Abortion` topic, establishing the constitutional foundations through cases like *Griswold v. Connecticut* (1965), which centered on privacy rights and reproductive autonomy. Later cases shifted toward the `Regulation of Abortion Procedures`, addressing specific implementation questions such as viability standards and the undue burden test. This evolution is exemplified in *Planned Parenthood v. Casey* (1992), which both reaffirmed constitutional protections and established new regulatory frameworks, stating 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."

While the discussion so far has focused on the substantive implications the PCTM can provide, we also provide discussion about the advantage of the PCTM in predicting new

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



(a) Regulation of Abortion (b) Const. Right Abortion (c) Disclosure of Private Info

Figure 3: Subnetworks 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.

words and citations compared to existing models in Supplementary Information, Section G.

4.2 Document-importance in Topic-specific Citation Networks

The PCTM's ability to assign topics to citations enables extraction of topic-specific subnetworks. We construct these subnetworks by including opinions that either send or receive citations of topic k. Figure 3 displays the resulting subnetworks for three topics: `Regulation of Abortion Procedures`, `Constitutional Rights to Abortion`, and `Public Disclosure of Private Information`.

Topic-specific subnetworks represent citation patterns within distinct semantic domains, enabling the application of established network analysis methods to semantically coherent subsets of citations. These methods include the "family tree of law" approach developed by Clark and Lauderdale (2012) and the importance score proposed in Fowler et al. (2007). Here, we focus on Fowler et al.'s importance scores, which measure an opinion's precedential significance and predict its likelihood of future citations. Recognizing semantic differences, however, is critical when computing importance scores because the absence of a citation to a precedent could have two different meanings: that the given precedent does not carry much legal weight or that the given precedent addresses a completely distinct legal issue. To demonstrate the significance of semantic context in citation analysis, we compare importance scores computed on the complete network with those derived from topic-specific subnetworks.

The importance score has two parts based on their citation directions. The outward

	Top 1 Inward-relevant	Top 2 Inward-relevant	Top 3 Inward-relevant
All Topics	Planned Parenthood v. Danforth	Roe v. Wade	Griswold v. Connecticut
Reg. Abortion	Planned Parenthood v. Danforth	Colautti v. Franklin	Bellotti v. Baird
Proc. Posture	Renegotiation Board v. Bannercraft	Hickman v. Taylor	EPA v. Mink
Const. Abortion	Griswold v. Connecticut	Roe v. Wade	Eisenstadt v. Baird
Speech & Protest	Schenck v. Pro-choice Network	Madsen v. Women's Health Center	Roe v. Wade
Damage to Privacy	Doe v. Chao	US ex rel. Touhy v. Ragen	US v. Reynolds
Privacy v. Govnt.	Vernonia v. Wayne	Chandler v. Miller	Whalen v. Roe
Pub. Disclosure	EPA v. Mink	Air Force v. Rose	NLRB v. Sears

Table 3: Top 3 most inward-relevant cases by topics. The inward relevance scores are computed following Fowler et al. (2007).

relevance score is based on the number of citations an opinion makes, evaluating the opinion's weight in referencing pertinent legal questions. An opinion with high outward relevance score cites many other opinions that are also deemed important and legally relevant. The inward relevance score is based on the number of citations an opinion receives from other opinions, gauging the extent to which it serves as the integral part of the law as a precedent. An opinion with high inward relevance score is cited by many other important and influential opinions. Since these scores are computed using eigenvectors, they are invariant to scales. In this light, Fowler et al. (2007) suggests using ranks of inward and outward relevance scores as the measure of importance for opinions as precedents.

In Table 3, none of the topic-specific top 3 inward-relevant cases exactly match those that are from the entire citation network of the privacy cases. The top 3 inward-relevant for all topics (row 1) seem to be drawing information from two topics – `Regulation of Abortion` and `Constitutional Rights to Abortion`. If one is interested in `Speech & Protest`, for example, *Schenck v. Pro-choice Network* is the most inward-relevant. *Schenck v. Pro-choice Network* is an influential case that draws the line between public safety and free speech. In *Schenck v. Pro-choice Network*, the SCOTUS concluded that the fifteen feet buffer zone between anti-abortion protestors and abortion clinics was constitutional, but deemed unconstitutional fifteen feet buffer zone between protestors and people seeking entrance to clinics. For `Public Disclosure of Information` topic, *EPA v. Mink* is the most inward-relevant. The case addresses the disclosure of secret documents prepared for a scheduled underground nuclear test, gauging the balance between the Freedom of Information Act (1966) and national security matters. Both examples show that one can draw very different conclusions on which case is most inward-relevant, depending on the legal context and area.

Table 4 shows that while the top three outward-relevant cases in the complete citation network primarily reflect rankings from the `Regulation of Abortion` and `Constitutional Rights to Abortion` topics, different patterns emerge when examining specific topics. For instance, in the `Public Disclosure of Information` topic, *Department of Justice v. Reporters Committee for the Freedom of Press* is the most outward-relevant case. The given

	Top 1 Outward-relevant	Top 2 Outward-relevant	Top 3 Outward-relevant
All Topics	Hodgson v. Minnesota	Akron v. Akron Center	Webster v. Reproductive Health
Reg. Abortion	Akron v. Akron Center	Hodgson v. Minnesota	Webster v. Reproductive Health
Proc. Posture	NLRB v. Sears	US v. Weber	DOI v. KWUPA
Const. Abortion	Carey v. Population Services Int.	Planned Parenthood v. Casey	Hodgson v. Minnesota
Speech & Protest	Hill v. Colorado	Schenck v. Pro-choice Network	Roe v. Wade
Damage to Privacy	Federal Aviation Admin. v. Cooper	NASA v. Nelson	US ex rel. Touhy v. Ragen
Privacy v. Govnt.	Board of Education v. Earls	Chandler v. Miller	Whalen v. Roe
Pub. Disclosure	DOJ v. Reporters Comm.	FBI v. Abramson	DOJ v. Tax Analysts

Table 4: Top 3 most outward-relevant cases by topics. The outward relevance scores are computed following Fowler et al. (2007).

case addresses whether the FBI should disclose criminal records to media outlets in the interest of public knowledge and safety. Together with Table 3, Table 4 shows that legal context can be heterogeneous within the privacy issue area, and such semantic heterogeneity can lead to varying conclusions on the precedential importance of cases.

4.3 Topic Prediction for a New Abortion Case

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 the topic structure of *Dobbs v. Jackson Women's Health Organization* differs from the recent rulings on abortion in our corpus. To do so, we computed the predicted probability of topics of the paragraphs in *Dobbs v. Jackson Women's Health Organization* using the model fitted on our abortion corpus. 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 formula to obtain the posterior predictive probability is in Supplementary Information, Section H.

To validate that the meaning of the topics is consistent in the new case, *Dobbs v. Jackson Women's Health Organization*, we provide a qualitative analysis of the estimated topics by focusing on the paragraphs that cite the same precedent. Table 5 presents two paragraphs that cite the same precedent, *Planned Parenthood v. Casey* (505 U.S., 878), but with different

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.
Both paragraphs cite the same precedent, *Planned Parenthood v. Casey* (505 U.S., 878), but with different estimated topics.

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. The 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 the interpretation of the topics in this new case match our interpretations of the topics in the abortion corpus.

How do the topic structure in *Dobbs v. Jackson Women's Health Organization* differ from the recent landmark cases in our corpus? 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*, two recent landmark cases in abortion in our corpus. Figure 4 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 of Abortion Procedure` topic and 67% of the paragraphs are assigned with `Constitutional Rights to Abortion`.

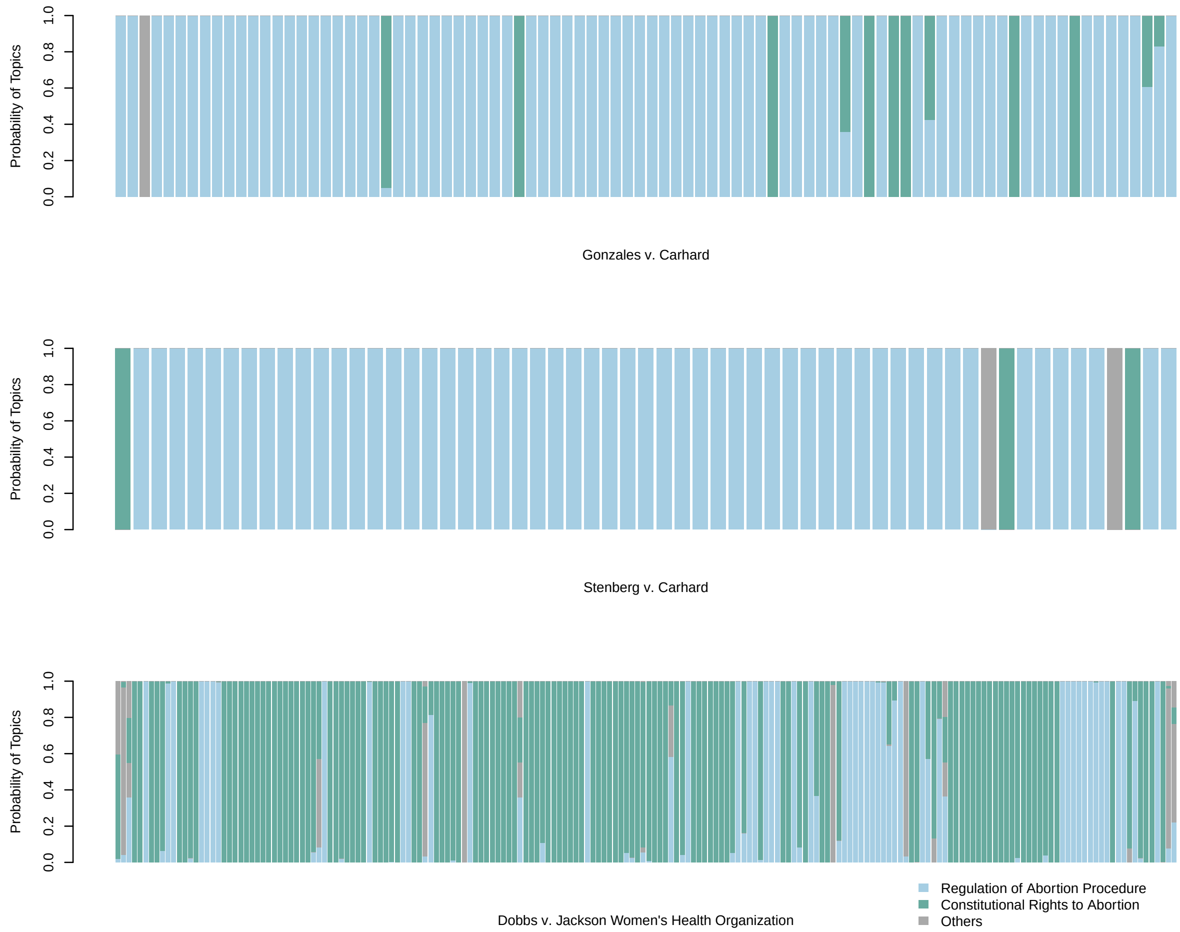


Figure 4: Predicted Probability of Topics for the Paragraphs of Dobbs v. Jackson. Each vertical bar represents a paragraph. Each paragraph is colored according to the predicted probability of topics. We focus on two topics related to abortion: `Constitutional Rights to Abortion` and `Regulation of Abortion 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 the 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*.

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.

5 Concluding Remarks

Social scientists often use citation networks to study how documents influence following documents in various domains, such as political science, international relations, and legal studies. However, conventional approaches to analyzing citation networks often overlook the semantic context in which citations occur. While existing studies use document-level labels to find the context of citations, this approach assumes that all citations within a document are made under the same context, which may lead to misunderstanding of how citations reflect the influence of documents. To address this challenge, this paper proposes a novel joint model of text and citations, the paragraph-citation topic model. The key innovation of the PCTM is to assign topics to paragraphs, which allows citations in different paragraphs to be associated with different topics. After deriving a collapsed Gibbs sampler for inference, we applied the PCTM to the SCOTUS opinions on privacy issues to highlight the diversity of topics of citations within each document. Also, the model uncovered informative subnetworks of the judicial opinions that shared citations with the same topic.

The applications of the PCTM need not be limited to citation networks of legal documents. The model will help address a number of important research questions in the analysis of document networks. For example, a researcher can use the latent citation propensity in the PCTM 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. Moreover, as the PCTM estimates topic-specific subnetworks of citations using information from both text and networks in a unified framework, it 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

Bai, H., Chen, Z., Lyu, M. R., King, I., and Xu, Z. (2018). Neural relational topic models for scientific article analysis. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 27–36.

Bailey, M. A. and Maltzman, F. (2008). Does legal doctrine matter? unpacking law and policy preferences on the us supreme court. *American Political Science Review*, 102(3):369–384.

Black, R. C. and Spriggs, J. F. (2013). The citation and depreciation of us supreme court precedent. *Journal of Empirical Legal Studies*, 10(2):325–358.

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.

Broughman, B. J. and Widiss, D. A. (2017). After the override: An empirical analysis of shadow precedent. *The Journal of Legal Studies*, 46(1):51–92.

Chang, J. and Blei, D. M. (2010). Hierarchical relational models for document networks. *The Annals of Applied Statistics*, 4(1):124 – 150.

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. (2010). Locating supreme court opinions in doctrine space. *American Journal of Political Science*, 54(4):871–890.

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. and Jeon, S. (2008). The authority of supreme court precedent. *Social networks*, 30(1):16–30.

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.

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.

Le, T. M. and Lauw, H. W. (2014). Probabilistic latent document network embedding. In *2014 IEEE International Conference on Data Mining*, pages 270–279. IEEE.

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.

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.

Olsen, H. P. and Küçüksu, A. (2017). Finding hidden patterns in ecthr's case law: On how citation network analysis can improve our knowledge of ecthr's article 14 practice. *International Journal of Discrimination and the Law*, 17(1):4–22.

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.

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.

Zhang, C. and Lauw, H. W. (2020). Topic modeling on document networks with adjacent-encoder. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 6737–6745.

Zhang, D. C. and Lauw, H. (2022). Dynamic topic models for temporal document networks. In *International Conference on Machine Learning*, pages 26281–26292. PMLR.