Paragraph-Citation Topic Model for Networked Text Data

Byungkoo Kim, Saki Kuzushima, and Yuki Shiraito University of Michigan

kimbk@umich.edu,skuzushi@umich.edu,shiraito@umich.edu Click the title to watch our short video summary!

Abstract

- Social scientists often analyze document networks, e.g., court decisions, academic articles, etc.
- Need measurement of each citation's context
- Existing topic models for document networks:
  - Assign the same topic to all citations in a document
  - No relationships between citation topics and word topics
- Paragraph-citation topic models:
  - Words and citations in a paragraph share a topic
  - Topic assignment for each citation
  - Common interpretation of word and citation topics
- Apply to the U.S. Supreme Court opinions

Model Overview

- PCTM jointly models document texts and network structure
- **Paragraph** as a coherent unit of analysis
- Citations within a paragraph have a shared topic
- Citations establish linkage between topics of citing paragraph and cited document

Paragraph-citation Topic Model

For each document i
  Draw topic proportion $\boldsymbol{\eta}_i \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$
  For each paragraph p:
    Draw assignment $z_{ip} \sim Mult(1, softmax(\boldsymbol{\eta}_i))$
    Draw word $w_{ip} \sim Mult(N_{ip}, \boldsymbol{\Psi}_{z_{ip}})$
    For all documents prior to i, j:
      Draw latent citation utility
      $D^*_{ipj} \sim \mathcal{N}(\boldsymbol{\tau}\mathbf{x}_{ipj}, 1)$
      Draw citation
      $D_{ipj} = 1$ if $D^*_{ipj} \geq 0$ and 0 otherwise
with $\mathbf{x}_{ipj} = [Intercept, Topic_j, Indegree_j, \cdots]$

Variational Inference

Variational Distribution for Topic probability, Topic assignment, Citation:

$$q(\boldsymbol{\eta}) = \prod_{i=1}^{N} \prod_{k=1}^{K} \mathcal{N}(\lambda_{ik}, \nu^2_{ik})$$

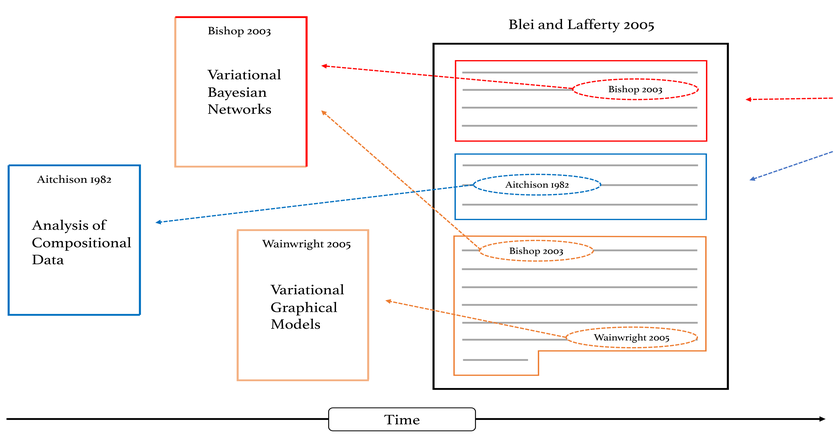
$$q(\mathbf{Z}) = \prod_{i=1}^{N} \prod_{p=1}^{N_i} Mult(\boldsymbol{\phi}_{ip})$$

$$q(\mathbf{D}^*) = \prod_{i=1}^{N} \prod_{p=1}^{N_i} \prod_{j=1}^{i-1} TN(\boldsymbol{\tau}\mathbf{x}_{ipj}, 1)$$

Conditioning on all the other parameters,

- E-step: maximize with respect to variational parameters $\boldsymbol{\lambda}, \boldsymbol{\nu}, \boldsymbol{\phi}$
- M-step: maximize with respect to model parameters $\boldsymbol{\mu}, \Sigma, \boldsymbol{\tau}, \boldsymbol{\Psi}$

Repeat until convergence.



Document Citation Network with Paragraph-level Topics

Simulation Analysis

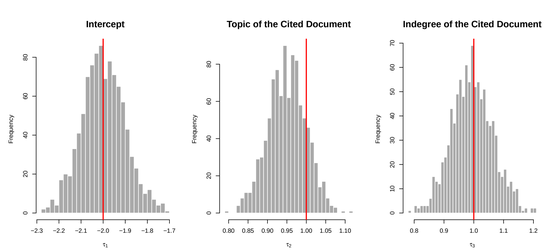
- 200 Documents, 10 paragraphs per document, 5 citations per paragraphs
- 1000 runs with random initialization
- $\mathbf{x}_{ipj} = [Intercept, Topic_j, Indegree_j]$

Result 1: Paragraph Topic Estimation Accuracy

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
Topic 1	1				
Topic 2	0	0.999			
Topic 3	0.017	0.003	0.988		
Topic 4	0	0.024	0.026	0.957	
Topic 5	0	0.017	0	0	0.983

- Diagonal entries show the fraction of topic match between true and estimated topics.
- Our model correctly estimate the topic of the simulation data.

Result 2: Simulated τ Values



- True τ values marked as vertical red lines.
- The estimates for each τ coefficient are distributed around the true τ values with very small variance.

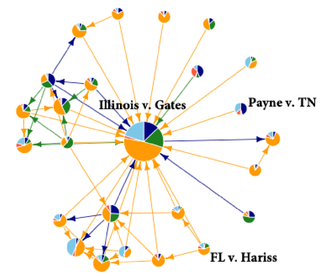
Application

- US Supreme Court cases since Reagan Administration (1981-2018) obtained from Caselaw Access Project
- Used a subset of criminal procedure cases identified by The Supreme Court Database
- Total of 42,428 paragraphs and 889 documents
- Number of topics set to 5
- $\mathbf{x}_{ipj} = [Intercept, Topic_j, Indegree_j]$

Result 1: Top 7 Words by Topics

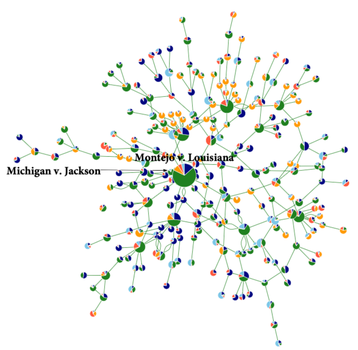
Topic					
Label	Prisoner Petitions	Corrections Sentencing	Search Seizure	Right to Counsel	Capital Punishment
1	prison	sentence	search	trial	sentence
2	petition	offense	office	rule	juri
3	respond	convict	amend	appeal	death
4	sentence	crime	reason	defend	constitution
5	time	petition	police	right	trial
6	trial	juri	fourth	claim	rule
7	statement	defend	warrent	counsel	circumstance

Result 2: Citation Network with Illinois v. Gates case



- Nodes represent documents and edges represent citations. Colors represent the paragraph-level topics and they correspond to words of the same color in Result 1. Size of the nodes are proportional to their degree centrality.
- **Illinois v. Gates** is a known monumental case in Search & Seizure. This is captured by orange-colored citations recognizing it.
- It also shows that **Illinois v. Gates** contains paragraphs that address topics other than Search & Seizure which then get cited by later documents of different topic compositions.

Result 3: Citation Network with Right to Counsel Topic



- The above is a topic-specific (Right to Counsel) slice of citation network. **Montejo v. Louisiana** is a monumental case on Right to Counsel because it reverses the doctrine established in **Michigan v. Jackson** which is another monumental case.
- Later cases chose to cite **Montejo v. Louisiana** over **Michigan v. Jackson**, making **Montejo v. Louisianan** node grow bigger in size. Hence, the above figure provides a glimpse of the evolution of legal doctrine in US Supreme Court on Right to Counsel.