Building the United States Supreme Court Disposition Corpus 1791-2009

MICHAEL JAMES BOMMARITO II
DANIEL MARTIN KATZ

UNIVERSITY OF MICHIGAN - ANN ARBOR
CENTER FOR THE STUDY OF COMPLEX SYSTEMS
DEPARTMENT OF POLITICAL SCIENCE

PRESENTATION @ UPENN LDC
JANUARY 2010

Broad Overview

- Introduction the United States Supreme Court
- Goals: Why are we creating the data set?
- Dataset Construction: How are we creating the data set?
- Research: How are we trying to use the data set?
- Future: What do you want to do with the data set?
  What are we doing wrong or right?

Supreme Court of the United States (SCOTUS)

- 1791 - 2009
- 111 Justices
- 30K + written opinions
- 300K + citations
- 50M+ words

Supreme Court of the United States (SCOTUS)

- Highest court in the United States
- Renders dispositions in a wide class of disputes
- Opinions / Dispositions recorded in U.S. Reports and other sources
- Jurisdiction and norms have changed over time

Getting to the Court

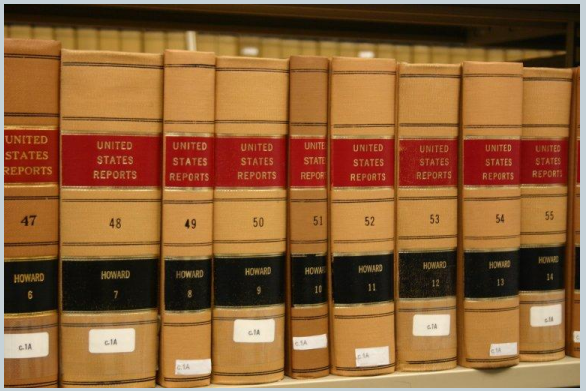
- Original jurisdiction v. Appellate jurisdiction
  - Original jurisdiction
    - Often disputes between states involving boundaries, etc.
  - Appellate jurisdiction
    - Typically via writ of certiorari

      (1) Cert. grant (may or may not include text/reasoning)

      (2) Cert. denied (may or may not include text/reasoning)

Data Background – SCOTUS Corpus

- While other scholars might subdivide differently…
- In reviewing the full corpus/citation network we would divide it as follows:
  - Early Years: 1791-1816
  - Developing Years: 1817 – Civil War
  - Reconstruction - Judge's Bill (1925)
  - Judge's Bill (1925) – Reagan Era
  - Reagan Era - Present



Goals- Create Comprehensive Records

- Justices:
  - How did Justice Rehnquist's language change over time?
  - How did Justice Warren's citation practices change over time?
- Cases
  - What text did *Roe v. Wade* contain?
  - What sources did *Marbury v. Madison* **cite**?
  - Which source has *Marbury v. Madison been* ***cited by***?
- Concepts
  - Changing conceptions of the 4th Amendment, etc.
  - When was the principle of X, Y or Z first used?

Goals – Aid Social Science Research

- Novel marriage of:
  - Votes
  - Citations
  - Opinion Content
- Potential applications:
  - Training prediction models
  - Understanding judicial behavior
  - Evaluating judicial fidelity

Data- Dispositions / Opinion Units

- Dispositions are the Superset
  - (1) Cert. grant & (2) Cert Denied
  - (3) Other Motions, etc. [Stays of Execution …]
- Opinion units are a Subset of Dispositions
  - (4) Majority opinion & (5) Concurrence & (6) Dissent

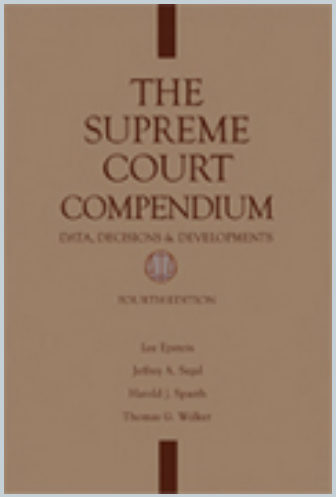
Data- Dispositions / Opinion Units

- Cases can feature multiple dimensions, e.g.:
  - Jurisdiction + Freedom of Religion?

- Justices can carve out preferred pronouncement
  - Craft an opinion to distinguish between dimensions
  - …or even "Concur in part, Dissent in part" !

KENNEDY, J., delivered the opinion of the Court, in which ROBERTS, C. J., and SCALIA and ALITO, JJ., joined, in which THOMAS, J., joined as to all but Part IV, and in which STEVENS, GINSBURG, BREYER, and SOTOMAYOR, JJ., joined as to Part IV. ROBERTS, C. J., filed a concurring opinion, in which ALITO, J., joined. SCALIA, J., filed a concurring opinion, in which ALITO, J., joined, and in which THOMAS, J., joined in part. STEVENS, J., filed an opinion concurring in part and dissenting in part, in which GINSBURG, BREYER, and SOTOMAYOR, JJ., joined. THOMAS, J., filed an opinion concurring in part and dissenting in part.

Data– Current Best Authority



Epstein, et al.

CHAPTER TWO
THE SUPREME COURT'S REVIEW PROCESS, CASELOAD, AND CASES

I. Introduction to Chapter Two

II. Tables

We wanted to build on this and other related work!

Data–Sources

- Official Report :
  - U.S. Reporter (____ U.S. ____)
- Major Subscription Reporters:
  - Lawyers' Edition (____ L. Ed. ___)
  - Supreme Court Reporter (____ S. Ct. ____)

Data - Process

- Acquire complete digital copies of:
  - Lawyers' Edition - LexisNexis
  - U.S. Reporter – bulk.resource.org
  - Justia, Oyez, USSC+
  - Other Sources
- Build parsers for both sources that extract:
  - Case "name", e.g., Plaintiff v. Defendant
  - Case citations, e.g., 544 U.S. 300
  - Date (of decision, hopefully)
  - "Opinion units" with authorship
- Cross-check!

Data – Process, ctd.

- Parsers are not easy!
- Want to capture **all Supreme Court dispositions.**
- Practices and language change over time
  - Reporter citations.
  - Shared case appendices.
  - Number of terms per year.
  - Date reported.
  - Norms on authorship and public dissent.
  - Varying autonomy of clerks.
- These dynamics are themselves often worth studying.

Data– Simple Statistics

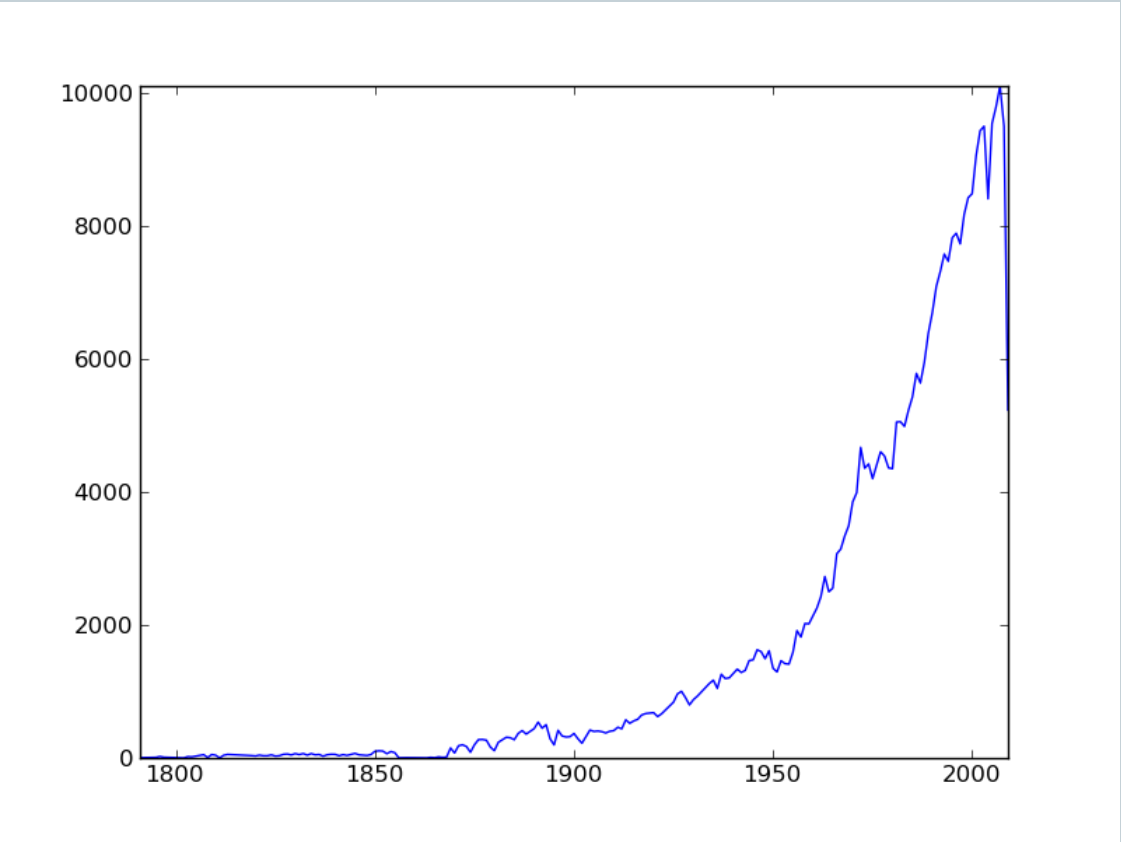


Figure: Dispositions {~caseload} parsed per year, 1791-2009

(Note: Data Coverage for 1852-1866 & 2009 Still Being Perfected)

Data - Classification

NAME: Clarence R. Allen, Petitioner v. Steven W. Ornoski, Acting Warden.
DATE: 2006-01-16 00:00:00
CITATIONS: 546 U.S. 1136;126 S. Ct. 1139;163 L. Ed. 2d 944;2006 U.S. LEXIS 763;74 U.S.L.W. 3405
AUTHOR: Breyer
OPINIONTYPE: dissent

Justice Breyer, dissenting.

Petitioner is 76 years old, is blind, suffers from diabetes, is confined to a wheelchair, and has been on death row for 23 years. I believe that in the circumstances he raises a significant question as to whether his execution would constitute "cruel and unusual punishmen[t]." U. S. Const., Amdt. 8. See Knight v. Florida, 528 U. S. 990, 993, 120 S. Ct. 459, 145 L. Ed. 2d 370 (1999) (Breyer , J., dissenting from denial of certiorari); Elledge v. Florida, 525 U. S. 944, 119 S. Ct. 366, 142 L. Ed. 2d 303 (1998) (same); Lackey v. Texas, 514 U. S. 1045, 115 S. Ct. 1421, 131 L. Ed. 2d 304 (1995) (Stevens , J., respecting denial of certiorari). I would grant the application for stay of execution.

Sample from actual corpus.

Data - Classification

How do we identify ***substantive*** dispositions?

1. Remove stopwords
2. Stem the tokens (Porter)
3. Remove dispositions with high proportions of “problem” stems
4. Remove dispositions without at least 30 unique stems

```
# Count some simple statistics
textLength = len(fileText)
textWords = [w.lower() for w in nltk.word_tokenize(fileText) if w.lower() not in stopwords]

# Skip the empties...
if len(textWords) == 0:
    continue

textAvgWord = sum(map(len,textWords))/float(len(textWords))

# Tokenize the text
textStems = [porter.stem(w.lower()) for w in nltk.word_tokenize(fileText) if w.lower() not in stopwords and len(w) > 3]
textStemCount = dict([(s,textStems.count(s)) for s in set(textStems)])

# Count the number of stems in the writ set
writStems = ['petit', 'writ', 'certiorari', 'deni','rehear','forma','pauperi']
writStemCount = sum([textStems.count(s) for s in textStemCount.keys() if s in writStems])

'''
Ignore dispositions with more than 10% of their unique stems in the problem set.
'''
if writStemCount > 0.1 * len(textStems):
    print 'stemprob', fileName

'''
Ingnore dispositions without at least 30 unique stems.
'''
if len(textStemCount) < 30:
    print 'stemcount', fileName
```

Build a coded sample to train a decision tree classifier.

Data– Simple Statistics

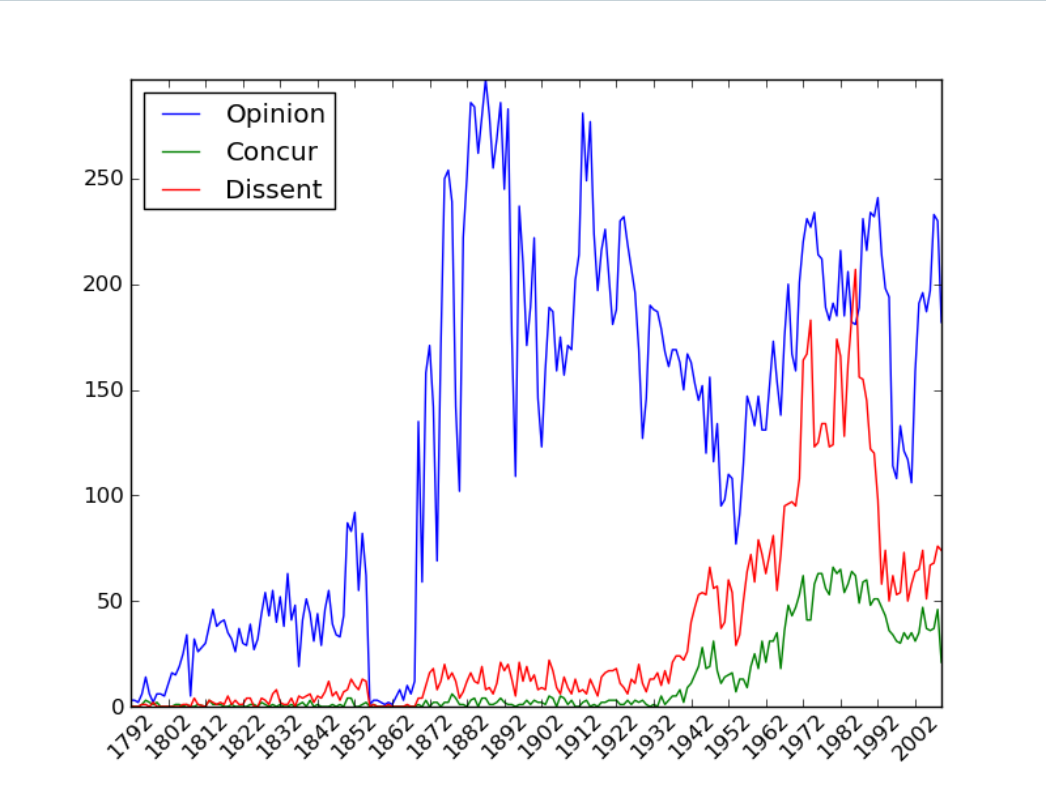


Figure: Number of substantive dispositions by type, 1791-2009

(Note: Data Coverage for 1852-1866 & 2009 Still Being Perfected)

Research – Citation Networks

Examples Realized in the Text

1. ___ Dallas ____
2. ___ Cranch ____
3. ___ Wheat. ____
4. ___ Peters ____
5. ___ Howard ____
6. ___ Black ____
7. ___ Wall. ____
8. ____ U.S. ____
9. ____ L. Ed. (2d) ____
10. ____ S. Ct. ____
11. Case name
12. Docket
13. Ante at page ____
14. 5 U.S. (1 Reporter) 137

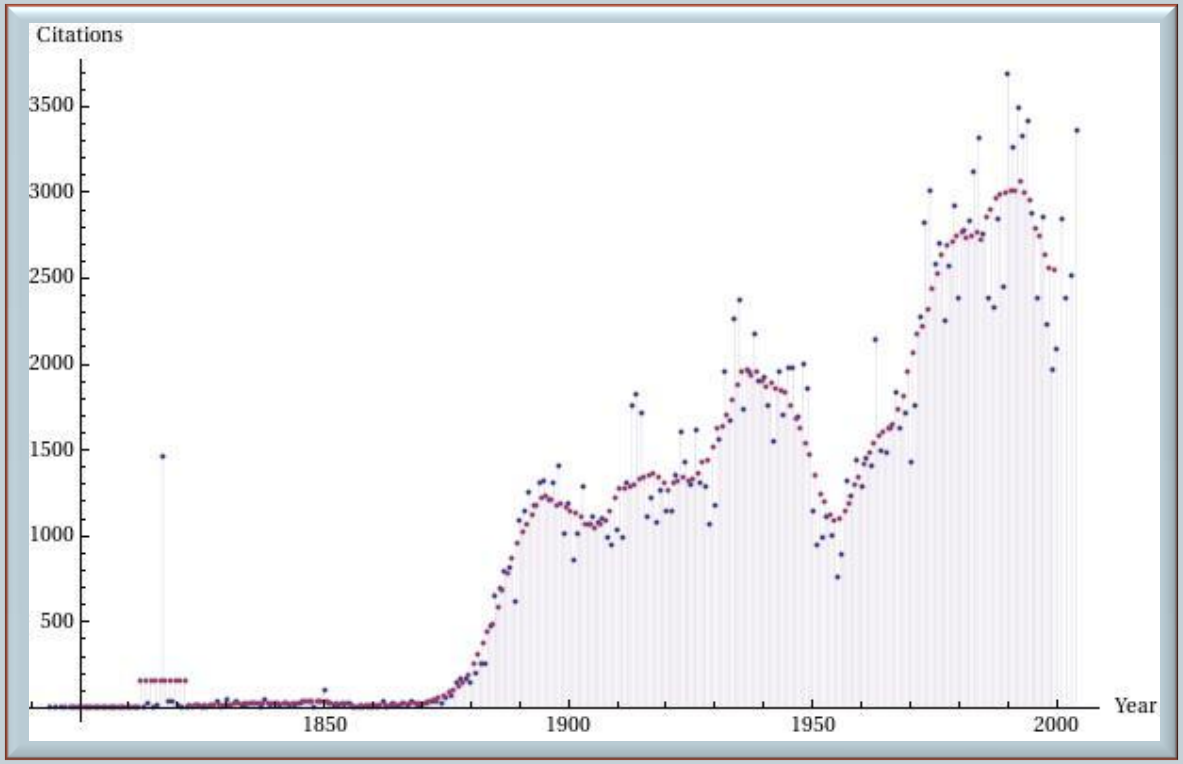
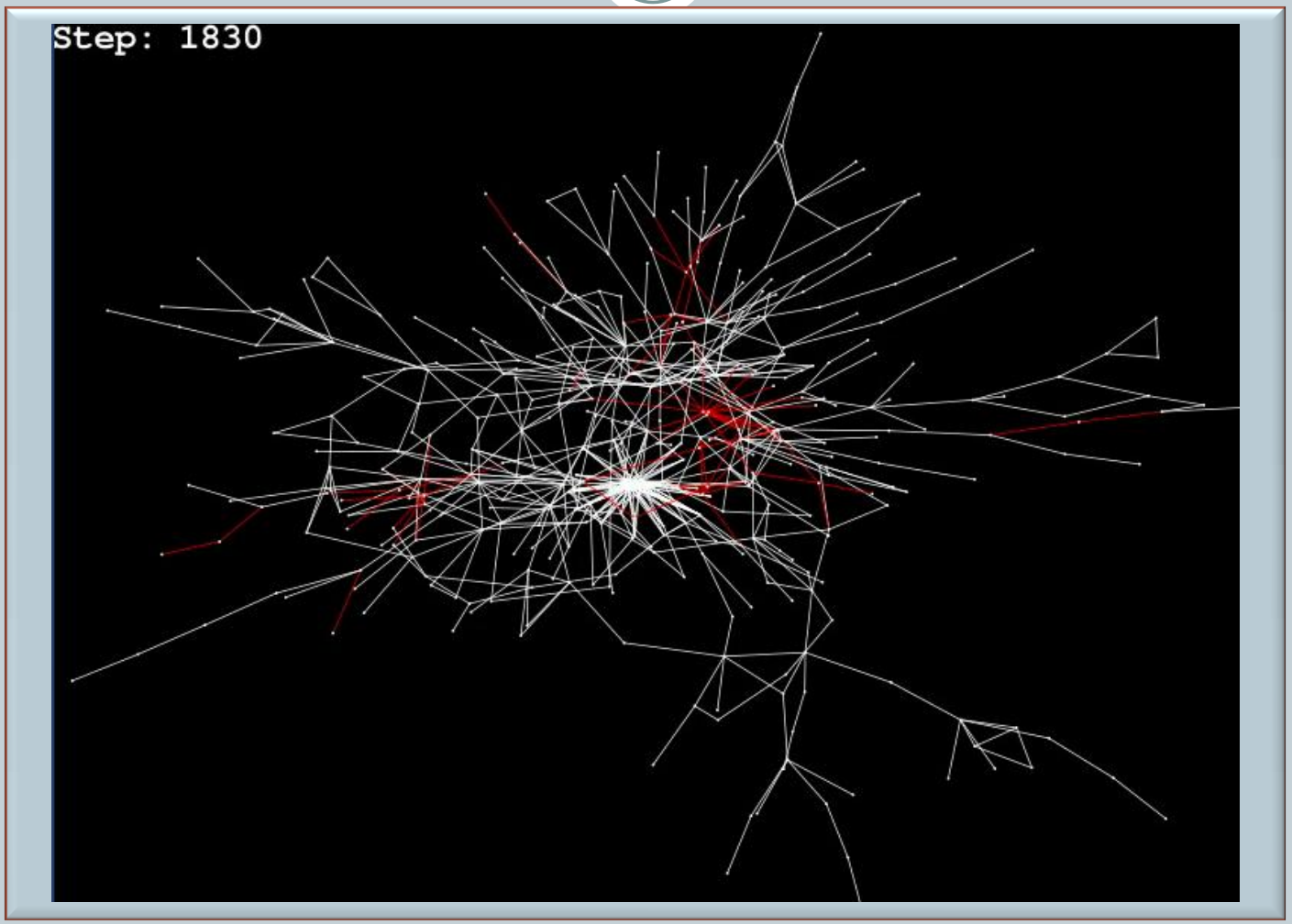


Figure: Number of Supreme Court to Supreme Court citations.

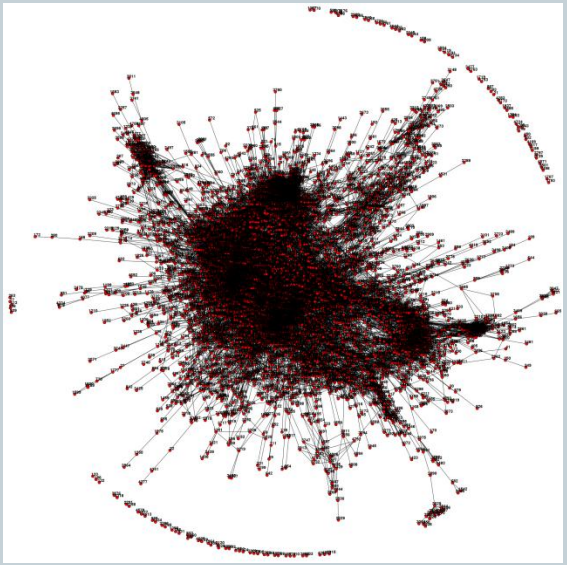
Research – Citation Networks



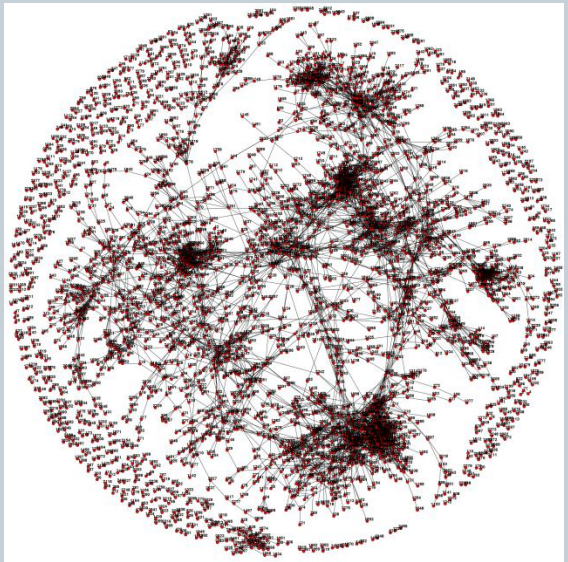
Supreme Court Citation Network Movie

Research – Semantic Networks

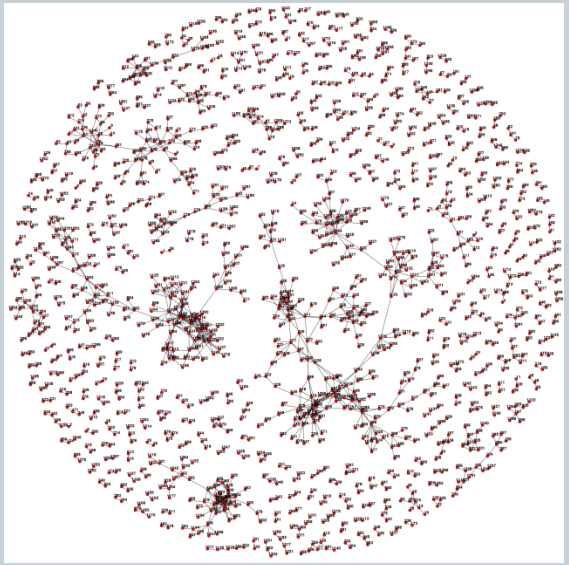
$$\sigma(A,B) = \frac{\#\{topWords(A) \cap topWords(B)\}}{N}$$



\sigma > 20%



\sigma > 30%



\sigma > 40%

Supreme Court Semantic Networks, 1865.

Law as a Seamless Web? Comparison of Various Network Representations of the United States Supreme Court Corpus (1791-2005)
Bommarito, Katz & Zelner

Research – Citation Networks

- What dynamics drive this network?
  - Topical citations – Citations driven by case topic
  - Strategic citations – Citations driven by policy preference (Lupu & Fowler 2010)
  - Temporal citations – Citations driven by recent cases (Leicht, et al 2007)
  - Analogical citations – Citations driven by analogical reasoning

Research – Citation Networks

- We want to integrate these dynamics into a model.
- Need data:
  - Topics: LDA (Blei 2003) or CTM (Blei 2006) ?
  - Author recognition: not explicit, van Halteren 2004?
  - Voting data: not explicit, 1937-present (Spaeth, SCDB)
  - Detect analogical reasoning: Any ideas?
  - Detecting textual entailment with citation
- RTM (Chang 2009) models both topics and links.
  - Do you have any experience with implementation?

Research – Citation Networks

- Dynamic Acyclic Labeled Weighted Multidigraph!
  - Dynamic: Answers have to make sense today & tomorrow
  - Acyclic: Citations must obey direction of time
  - Digraph: Cases assert asymmetric relationships
  - Weighted: Citations may be negative or positive
  - Multidigraph: Formalize conception of “dimensionality”
- Problem:
  - Most methods for undirected unweighted graphs

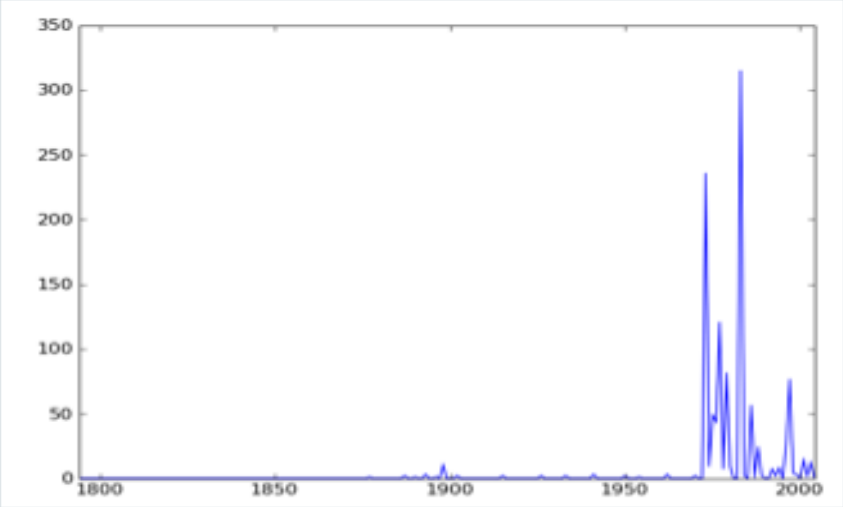
Research – Citation Networks

- Dynamics:
  - On the Stability of Community Detection Algorithms on Longitudinal Citation Data. Bommarito, Katz, Zelner.
  - Experimental study of "stability" of canonical community detection methods.

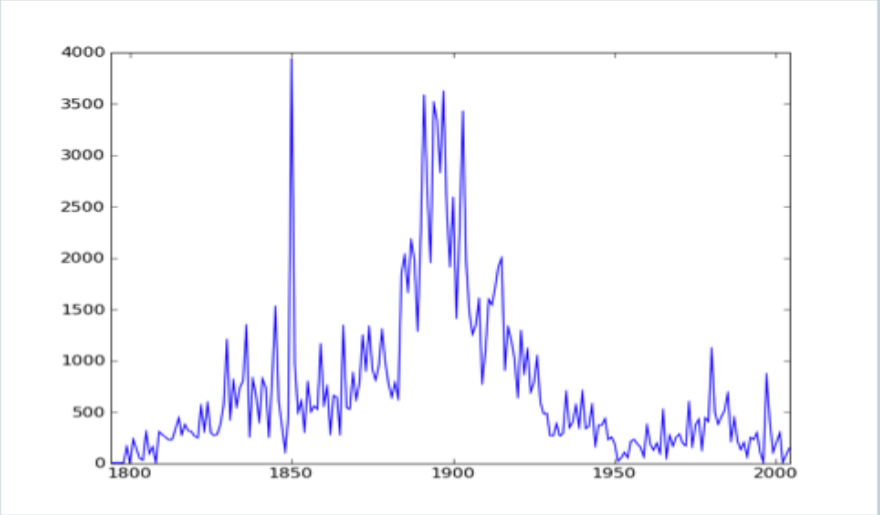
- Acyclic multidigraph:
  - Distance Measures for Dynamic Citation Networks. Bommarito, Katz, Zelner, Fowler.
  - Introduce a family of distance measures that have very attractive properties relative to previously existing.

Research – Word Usage

Abortion



Property



Note: the Selection operator.

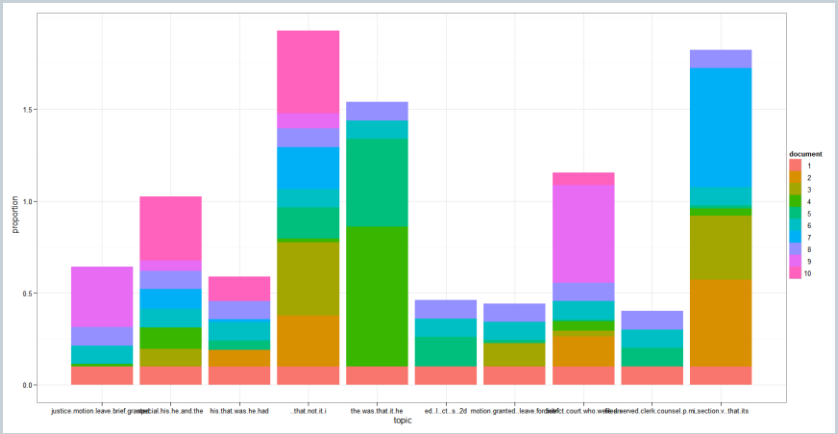
Projects

Our Current Projects:

1.Finish collecting & coding data, e.g., 1853-1866

2.Train & apply a disposition classifier to determine case outcome.

3.Apply Chang & Blei 2009 relational topic model



Sample LDA Topic Distribution

Projects

What Others Have Done:

1. Andrew Martin & Kevin Quinn: Prediction competition ***without*** textual/citation data.

2. Jiahong Yuan and Mark Liberman: Author recognition on Supreme Court ***audio*** @ Oyez

3. Wayne MacIntosh, et al.: Working to incorporate materials from the Supreme Court ***briefs***.

Bibliography

L. Epstein, J. Segal, H. Spaeth, T. Walker. The Supreme Court Compendium. 2006.

M.F. Porter. An algorithm for suffix stripping. 1980

Y. Lupu, J. Fowler. The Strategic Content Model of Supreme Court Opinion Writing. 2010

J. Fowler, S. Jeon. The Authority of Supreme Court Precedent. 2008

D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. 2003

D. Blei, J. Lafferty. A correlated topic model of Science. 2007

J. Chang, D. Blei. Hierarchical relational models for document networks. 2009

H. van Halteren. Linguistic profiling for author recognition and verification. 2004

SES-0921869, SES-0923665, SES-0919149, SES-0918613, SES-0751966, SES-9910535, SES-9614000, SES-9211452, SES-8313773. NSF grants supporting the Supreme Court Database at Wash. U.
http://scdb.wustl.edu/about.php

M. Evans, W. McIntosh, C. Cates, J. Lin. Recounting the Courts? Toward a Text-Centered Computational Approach to Understanding the Dynamics of the Judicial System. 2007

Jiahong Yuan and Mark Liberman, Speaker Identification on the SCOTUS Corpus:
http://www.ling.upenn.edu/~jiahong/publications/asa08.pdf