2020

Mining the Harvard Caselaw Access Project

Felix B. Chang
University of Cincinnati College of Law, felix.chang@uc.edu

Erin McCabe
University of Cincinnati - Main Campus

James Lee
University of Cincinnati - Main Campus

Mining the Harvard Caselaw Access Project

FELIX B. CHANG
ERIN MCCABE
JAMES LEE †

This Article illustrates how machine learning ("ML") can advance antitrust scholarship through the extraction and analysis of big data. We have built a ML platform that analyzes large datasets through topic modeling, an algorithm that maps the statistical relationships among words. The platform creates visualizations that illuminate linguistic patterns in antitrust decisions extracted from Harvard Law Library's Caselaw Access Project, which has recently digitized almost all published decisions in the U.S.

Topic modeling provides new perspectives on how courts tackle two thorny question in antitrust: the measure of market power and the balance between antitrust and regulation. Our visualizations depict how thousands of

† Felix Chang is a Professor of Law at the University of Cincinnati ("UC") College of Law. Erin McCabe is a UC Digital Scholarship Fellow. James Lee is the Academic Director of the UC Digital Scholarship Center. We thank the Andrew W. Mellon Foundation for financial support. Thanks, too, to Rosa Abrantes-Metz, Josh Beckelhimer, Ned Cavanagh, David Donald, Harry First, Eleanor Fox, James Grimmelmann, Scott Hemphill, Michael Livermore, Zhaowei Ren, Danny Sokol, and Adam Ziegler. This article benefitted from the Next Generation of Antitrust Scholars Conference at NYU, the Online Workshop on the Computational Analysis of Law at Virginia, and the Machine Lawyering Conference at the Chinese University of Hong Kong.

antitrust cases cluster around specific terms—as well as how these clusters have evolved over time. We present these visualizations as a new suite of tools to assess the weighty policy arguments that currently dominate antitrust.

CONTENTS

I. INTRODUCTION

Is legal scholarship driven mainly by ideas or by tools?[1] Decades ago, empirical methods revolutionized legal research, eventually gaining widespread acceptability despite concerns about experimental design.[2] More recently, scholars and judges have begun adopting the methods of corpus linguistics, which map word frequency and collocation, to discern the ordinary meaning of phrases in a statute or the Constitution.[3] These techniques are among the advances of computational legal analysis ("CLA"), which unleashes quantitative empirical techniques such as machine learning and natural language processing upon legal texts.[4]

The newest tool to gain a following in CLA is topic modeling, a form of natural language processing that depicts the probability distribution of terms over a corpus of texts.[5] Heralded for its propensity to analyze large, unstructured datasets, topic modeling has already illuminated patterns in judicial opinions,[6] loan agreements,[7] and national

[1] On the origin of this question in science, see Freeman J. Dyson, *Is Science Mostly Driven by Ideas or by Tools?*, 338 SCIENCE 1426 (2012).

[2] *See* Daniel E. Ho & Larry Kramer, *The Empirical Revolution in Law*, 65 STAN. L. REV. 1195 (2013). *See also* Joshua Angrist & Jörn-Steffen Pischke, *The Credibility Revolution in Empirical Economics: How Better Research Design is Taking the Con out of Econometrics*, 24 J. ECON. PERSP. 3 (2010).

[3] Corpus linguistics studies language through its usage within a body of texts. TONY MCENRY & ANDREW WILSON, CORPUS LINGUISTICS 1 (2001). For examples of its application in legal scholarship, see Stefan Th. Gries & Brian G. Slocum, *Ordinary Meaning and Corpus Linguistics*, 2017 BYU L. REV. 1417; Thomas R. Lee & Stephen C. Mouritsen, *Judging Ordinary Meaning*, 127 YALE L.J. 788 (2018).

[4] Michael A. Livermore & Daniel N. Rockmore, *Introduction: From Analogue to Digital Legal Scholarship*, *in* LAW AS DATA: COMPUTATION, TEXT, & THE FUTURE OF LEGAL ANALYSIS xvii (Michael A. Livermore & Daniel N. Rockmore eds., 2019).

[5] *See* Michael A. Livermore et al., *The Supreme Court and the Judicial Genre*, 59 Ariz. L. Rev. 837, 841–42 (2017); David M. Blei et al., *Latent Dirichlet allocation*,3 J. MACHINE LEARNING RES. 993 (2003).

[6] *See id.*

[7] Bernhard Ganglmair & Malcolm Wardlaw, Complexity, Standardization, and the Design of Loan Agreements (2017), *available at* https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2952567

constitutions.[8] Yet the tool has not drawn the level of scrutiny of previous empirical methods. To date, topic modeling enthusiasts in law have sidestepped basic questions such as (i) how do disembodied terms, whatever their interrelation, represent legal doctrine and (ii) should legal texts be spliced and read in this way?[9]

A reflexive embrace of topic modeling and, more generally, CLA risks giving machine learning too quick a pass,[10] without vetting the underlying algorithms.[11] Word frequencies "without regard to position, syntax, content, and semantics" should not comprise the basis for bold claims.[12] Unmoored, CLA resembles what the philosopher Richard Rorty characterized of certain kinds of literary criticism as "imposing a vocabulary . . . a 'grid' . . . on the text which may have nothing to do with any vocabulary used in the text or by its author, and seeing what happens."[13]

We aim to correct the oversight by engaging with critiques of machine learning from areas outside law. For topic modeling in particular, although its sheen is still fresh in legal circles, the technique has circulated for years in digital humanities ("DH"), the branch of traditional humanities that incorporates machine-driven computation into its analysis.[14] DH is a collaborative endeavor at its core, so when legal scholars borrow its tools without considering assessments of

[8] *See* David S. Law, *Constitutional Archetypes*, 95 TEX. L. REV. 153 (2016).

[9] Exceptions include Stanley Fish, *If You Count It, They Will Come*, 12 N.Y.U. J. L. & LIBERTY 333 (2019); Michael A. Livermore & Daniel N. Rockmore, *Distant Reading the Law*, *in* LAW AS DATA, *supra* note 4.

[10] For an explanation of machine learning, see Ryan Copus et al., *Credible Prediction: Big Data, Machine Learning and the Credibility Revolution*, *in* LAW AS DATA, *supra* note 4 ("Machine learning is not a specific research tool; it is a catch-all term that refers to any method that features *learning* by a *machine* about quantitative data.").

[11] An algorithm is a set of instructions to perform a task, given a specific input.

[12] Nan Z. Da, *The Computational Case against Computational Literary Studies*, 45 CRIT. INQUIRY 601 (2019).

[13] RICHARD RORTY, CONSEQUENCES OF PRAGMATISM 151 (1982).

[14] *See* ANNE BURDICK ET AL., DIGITAL_HUMANITIES 3 (2012); Matthew G. Kirschenbaum, *What Is Digital humanities and What's It Doing in English Departments?*, 150 ADE BULL. 1 (2010).

machine analysis from the humanities and computer science, we abandon the spirit from which we draw inspiration.

More fundamentally, the technical and computational abilities of machine learning evolve at a startling pace. If legal scholars do not slow down now to reflect upon the viability of the methodologies—or to reset our collaboration with statisticians, humanists, and computer scientists—then the likelihood of confronting essential questions grows ever distant.

Rather than merely reciting the diverse critiques of CLA, we incorporate them to improve machine learning algorithms for legal research. We have built a machine learning platform that analyzes large datasets through variations on topic modeling. In the most novel variation, we aggregate—or embed—six levels of topic modeling into a single set of visualizations. Using aggregated modeling, the platform reveals linguistic patterns within a corpus of cases extracted from Harvard Law School's Caselaw Access Project, which has recently digitized almost all published decisions in the U.S.[15]

Through our modifications, topic modeling can create metadata, similar to the headnotes of commercial legal databases, that make legal research more efficient. Our central contribution to the growing field of CLA is to take analysis from the level of words and phrases to the level of topics and documents, providing greater contextualization. The ensuing visualizations, apart from their immense beauty, translate topic modeling into intuitive models that law scholars can interpret without statistical or empirical training.

We are mindful that our solution to flawed machine learning is *more* machine learning—or at least *better* machine learning. Yet virtually all criticisms of algorithmic data extraction and processing can be distilled to one theme: the

[15] *See About*, CASELAW ACCESS PROJECT, https://case.law/about/ (last accessed July 29, 2019).

need to provide greater contextualization.[16] In response, we have not abandoned DH methods but sought to improve them.

As a test, we have compiled a large pool of federal antitrust cases to see what our algorithms reveal of two thorny doctrinal questions: the measure of market power and the balance between antitrust and regulation. Our platform's visualizations depict how thousands of market power and antitrust–regulation cases cluster around different terms—as well as how these clusters have evolved over time. We have chosen to start with market power and the antitrust–regulation balance because doctrinal ambiguities leave interpretation in these areas wide open, thereby broadening the terms that courts engage.

Because our platform's analysis of antitrust cases occurs through machines, it is bound by neither legal precedent nor economic theory. Thus, our project addresses not the normative question of how *should* courts gauge market power but the empirical question of how *do* courts gauge market power. While algorithmic processing has its limits,[17] machine-generated visualizations can provide a fresh take on thousands of cases.[18] Concomitantly, legal doctrines around market power and the antitrust–regulation balance serve as a back-end check on the precision of aggregated modeling.

---

[16] *See, e.g.* Evan C. Zoldan, *Corpus Linguistics and the Dream of Objectivity*, 50 SETON HALL L. REV. 401, 406 (2019) ("Rather than simply serving as another "tool in the toolbox" of statutory interpretation, corpus linguistics is different from traditional tools of statutory interpretation because it leads to interpretations that are radically acontextual."); Da, *supra* note 12.

[17] *See* SAFIYA UMOJA NOBLE, ALGORITHMS OF OPPRESSION: HOW SEARCH ENGINES REINFORCE RACISM (2018); Nathan Newman, *How Big Data Enables Economic Harm to Consumers, Especially to Low-Income and Other Vulnerable Sectors of the Population*, 18 J. INTERNET L. 11 (2014).

[18] *See* Law, *supra* note 6, at 164–65 ("Topic modeling is well suited to the analysis of large numbers of complex, varied documents . . . because it is capable of identifying verbal patterns and substantive topics in raw text without any need for time-consuming and potentially erroneous hand-coding of the text"); Elliott Ash & Daniel L. Chen, *Case Vectors: Spatial Representations of the Law Using Document Embeddings*, *in* LAW AS DATA, *supra* note 4, at 314 ("[Topic modeling] algorithms have provided a window to the relations between documents at scale.").

Our second contribution is to antitrust itself, which is also at an inflection point in the oscillation between ideas and tools. More than any time since the rise of the Chicago school, antitrust today is dominated by ideas. From the new Brandeis school (sometimes dismissed as "hipster antitrust")[19] to the multisided platform debate,[20] weighty ideas on inequality and big tech are driving the conversations in antitrust.[21] Often, these conversations unfold without a rigorous methodology to quantify their claims. We see in topic modeling a new suite of tools to hone the doctrinal and policy arguments, just as the Harvard school of antitrust refined the Chicago school's brash theoretical pronouncements decades ago.[22]

Aggregated modeling excels in presenting high-level summaries of caselaw. For instance, market power cases splinter into a few large categories: patent cases, health care cases, telecommunications cases, tying cases, banking and financial cases, and cases delving deeply into civil and evidentiary procedure.[23] Similarly, from the antitrust–regulation corpus, we see that cases pertaining to the Interstate Commerce Commission were supplanted over time by telecommunications cases, a pattern consistent with deregulation.[24] Doctrinally, these inferences are not necessarily novel, but they do confirm the conjectures of other antitrust

[19] *See* Maurice E. Stucke & Ariel Ezrachi, *The Rise, Fall, and Rebirth of the U.S. Antitrust Movement*, HARV. BUS. REV. (Dec. 15, 2017); Lina Khan & Sandeep Vaheesan, *Market Power and Inequality: The Antitrust Counterrevolution and Its Discontents*, 11 HARV. L. & POL'Y REV. 235 (2017).

[20] *See* DAVID S. EVANS & RICHARD SCHMALENSEE, MATCHMAKERS: THE NEW ECONOMICS OF MULTISIDED PLATFORMS (2016).

[21] *See* Elizabeth Warren, *Here's How We Can Break Up Big Tech*, MEDIUM, Mar. 8, 2019; *Corporate Accountability an Democracy*, BERNIE SANDERS, https://berniesanders.com/issues/corporate-accountability-and-democracy/; TIM WU, THE CURSE OF BIGNESS: ANTITRUST IN THE NEW GILDED AGE (2018).

[22] *See* William E. Kovacic, *The Intellectual DNA of Modern U.S. Competition Law for Dominant Firm Conduct: The Chicago/Harvard Double Helix*, 2007 COLUM. BUS. L. REV. 1.

[23] *See infra* Section IV.

[24] Joseph D. Kearney & Thomas W. Merrill, *The Great Transformation of Regulated Industries Law*, 98 COLUM. L. REV. 1323, 1330-34 (1998).

scholars who had theorized from narrower samplings of caselaw.[25]

Our results are more provocative, however, for what they suggest about caselaw research. Currently, scholars and practitioners rely heavily on proprietary databases such as Westlaw and Lexis to identify the most relevant cases for a research question. A search in Westlaw for federal cases bearing the terms "antitrust" and "market power,'" for example, yields top results that include *Eastman Kodak*,[26] *Jefferson Parish*,[27] *Grinnell*,[28] *Microsoft*,[29] and *du Pont*,[30] all of them classic cases on market power.[31] Curiously, however, these classic cases do not tend to show up in our visualizations, whether as top terms or as top cases within a topic.[32] By contrast, the top (or most relevant) cases identified by topic modeling are not prioritized by Westlaw or Lexis, but they are influential nonetheless among practitioner circles within a particular circuit.[33]

These results question how Westlaw and Lexis execute their searches, a process that is notoriously opaque.[34] For example, how do the commercial databases differ from widely accepted statistical algorithms in defining what constitutes *relevant* caselaw? In publicizing our algorithms, we hope to nudge the commercial databases toward greater transparency.

---

[25] Narrower sampling is often a feature of qualitative research, and doctrinal research is often qualitative.

[26] Eastman Kodak Co. v. Image Technical Services, Inc., 504 U.S. 451 (1992).

[27] Jefferson Parish Hosp. Dist. No. 2 v. Hyde, 466 U.S. 2 (1994).

[28] U.S. v. Grinnell Corp., 384 U.S. 563 (1966).

[29] U.S. v. Microsoft Corp., 253 F.3d 34 (D.C. Cir. 2001).

[30] U. S. v. E. I. du Pont de Nemours & Co., 351 U.S. 377 (1956).

[31] The other cases in the top 11 were *In re Copper Market Antitrust Litig.*, 200 F.R.D. 213 (S.D.N.Y. 2001) (which did not even include the term "market power" or consider the concept), *Illinois Tool Works Inc. v. Independent Inc., Inc.*, 547 U.S. 28 (2006), *In re Aggrenox Antitrust Litig.*, 199 F. Supp. 3d 662 (D. Conn. 2016), *Sentry Data Systems, Inc. v. CVS Health*, 379 F. Supp. 3d 1320 (S.D. Fla. 2019), *Datel Holdings Ltd. v. Microsoft Corp.*, 712 F. Supp. 2d 974 (N.D. Cal. 2010), and *Rebel Oil Co., Inc. v. Auto Flite Oil Co., Inc.*, 51 F.3d 1421 (9th Cir. 1995). The search was performed on Mar. 20, 2020.

[32] The only exception being *Microsoft*, 253 F.3d.

[33] *See infra* Section IV.

[34] *See* Susan Nevelow Mart, *The Algorithm As a Human Artifact: Implications for Legal [Re]Search*, 109 LAW LIBR. J. 387, 389 (2017).

Finally, the source of our data, the Caselaw Access Project, portends a sea change in information retrieval. In recent years, freely available legal repositories have debuted, promising to democratize legal research. Nonetheless, technical and financial barriers to data extraction and analysis persist. Insurgent challengers to Westlaw and Lexis are pledging to harness innovations in information technology to deliver "faster and smarter" legal research.[35] Yet it is not clear whether these gatekeepers also intend for research to be cost-effective, especially for academic and nonprofit communities.

We see our project as a leap in algorithmic topic modeling for legal research, especially as a complement to commercial databases. Ultimately, we hope that our project will prompt other collaborations between DH and law, while pressing information technology insurgents to keep legal research open and cost-effective.

The remainder of this Article unfolds as follows: Section II canvases critiques of CLA methods and tinkers with improvements to topic modeling. Section III introduces our platform and summarizes our methodology. Section IV presents preliminary findings and draws inferences that both affirm and complicate previous antitrust research. Section V hazards predictions for the way forward. Section VI concludes.

II. TOPIC MODELING LEGAL TEXTS IN THE ERA OF BIG DATA

Machine learning abounds in finance, policing, employment, politics, and health services,[36] but as a research technique, it is just gaining traction in legal academia.[37] Law

[35] *What is Fastcase?*, FASTCASE, https://www.fastcase.com/about/ (last accessed Feb. 3, 2020).

[36] *See, e.g.*, VIRGINIA EUBANKS, AUTOMATING INEQUALITY HOW HIGH TECH TOOLS PROFILE POLICE & PUNISH THE POOR (2018).

[37] One exception is the application of corpus linguistics to statutory interpretation to discern the ordinary meaning of language. *See, e.g.*, Stefan Th. Gries & Brian G. Slocum, *Ordinary Meaning and Corpus Linguistics*, 2017 B.Y.U. L. REV. 1417. Recently, BigML also started to provide machine

scholars are quick to castigate the use of machine learning and, more broadly, artificial intelligence in law—yet slow to employ them in legal research. This is an odd conundrum. After all, in our era of big data, data is king.[38] And in law, no data is bigger than legal texts. Applied to a corpus of case law, machines can uncover explicit and latent linguistic and semantic patterns, bringing out significant word clusters "that the eye cannot see."[39] The proliferation of free, open-source legal databases and the explosion in data processing capabilities makes our era a truly exciting one for legal research.

Nonetheless, these technical advances do little to address the reservations that legal scholars harbor toward CLA. The tools of corpus linguistics, or instance, have come under scrutiny for their tendency to decontextualize settings.[40] These are variations of DH practices known as "deformance" and "tampering" at their most extreme, rearranging texts in the fashion of what post-structuralists call "a new cut."[41]

This Section offers topic modeling as a viable tool for legal research. In many ways more nuanced than word frequency and collocation, topic modeling is beginning to gain traction within CLA, so it is not wholly unfamiliar to law scholars. Yet the technique has certain vulnerabilities, as digital humanists and computer scientists have previously pointed out. This Section therefore reintroduces topic modeling, especially in triangulation with corpus linguistics and other empirical methods, which are more familiar. It also summarizes the criticisms of topic modeling, as a preview to our improvements to traditional topic modeling algorithms.

learning services to academics. *See* BigML, https://bigml.com/ (last accessed Jan. 15, 2020).

[38] For a definition of big data, see Svetlana Sicular, *Gartner's Big Data Definition Consists of Three Parts, Not to Be Confused with Three "V"*s, FORBES (Mar. 27, 2013, 8:00 AM), http://www.forbes.com/sites/gartnergroup/2013/03/27/gartners-big-data-definition-consists-of-three-parts-not-to-be-confused-with-three-vs/.

[39] Lauren Klein, *Distant Reading after Moretti*, https://lklein.com/digital-humanities/distant-reading-after-moretti/ (Jan. 10, 2018).

[40] *See, e.g.*, Carissa Byrne Hessick, *Corpus Linguistics and the Criminal Law*, 2017 BYU L. REV. 1503 (2017); Zoldan, *supra* note 16.

[41] Fish, *supra* note 9, at 303–04.

A. *A Primer on Topic Modeling*

Topic modeling illustrates the probable distribution of terms and their co-occurrence within a dataset, a process that uncovers latent, or hidden, patterns within the dataset.[42] These patterns are presented as "topics," where each topic is comprised of the terms most likely to appear together.[43] Topic modeling builds upon the general concept of modeling, which creates representations of data patterns in a statistically or logically coherent form.[44] While models abound in legal research, topic modeling is performed through machine learning to amplify processing power.[45] The tool enables researchers to analyze tomes of data without having to manually code them first, as is custom in traditional empirical methods.[46]

Topic modeling is particularly useful in text-intensive projects because of its propensity to uncover language patterns. For instance, researchers in one discipline—say, statistics—may cite influential papers within their discipline but miss relevant papers in another discipline—e.g., economics or biology.[47] If topic modeling is run on papers from statistics, economics, and biology, it can cut through citation biases to identify the terms and topics common to all three fields, resulting in more useful literature recommendations.

The pervading topic modeling algorithm is latent Dirichlet allocation ("LDA"), which reveals the Dirichlet

---

[42] *See* Jason Chuang et al., *Interpretation and Trust: Designing Model-Driven Visualizations for Text Analysis*, *in* CHI '12: PROCEEDINGS OF THE SIGCHI CONFERENCE ON HUMAN FACTORS IN COMPUTING SYSTEMS (2012).

[43] Chong Wang & David M. Blei, *Collaborative Topic Modeling for Recommending Scientific Articles*, *in* KDD '11: PROCEEDINGS OF THE 17TH ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING (2011).

[44] *See* KEVIN D. ASHLEY, ARTIFICIAL INTELLIGENCE AND LEGAL ANALYTICS: NEW TOOLS FOR LAW PRACTICE IN THE DIGITAL AGE 234–35 (2017).

[4545] *See id.* at 77 (case-based legal reasoning models), 131 (legal argument models), 234 (machine learning models).

[46] Livermore et al., *supra* note 5, at 842.

[47] *See* Wang & Blei, *supra* note 43.

allocation, or multivariable probability distributions, of topics over a fixed vocabulary within a dataset.[48] True to form, LDA was deployed early on in projects such as the Stanford Dissertation Browser, an interactive tool that shows the commonalities across dissertations written at Stanford from 1993 to 2008, and an algorithm to recommend scientific articles to researchers.[49]

Two features of topic modeling—its ability to sift through large volumes of texts and to uncover hidden connections within those texts—makes it tantalizing for legal research. While the tool remains new to law scholars,[50] a growing number of researchers are adopting it for projects on loan agreements,[51] constitutions around the world,[52] Supreme Court legal opinions,[53] and control rights in union contracts.[54]

While not wholly analogous to topic modeling, corpus linguistics is in some ways an apt comparator for its trajectory from linguistics into law. Corpus linguistics takes an empirical approach to the study of language by gauging ordinary meaning through the usage of words and phrases in a corpus, or body, of natural language texts.[55] Its advocates in law, including most prominently Justice Thomas Lee on the Utah Supreme Court, argue that its methods can elucidate ordinary

[48] *See* Blei et al., *supra* note 5.

[49] *See An Experiment in Document Exploration*, STANFORD DISSERTATION BROWER, https://nlp.stanford.edu/projects/dissertations/ (last accessed Feb. 2, 2020); Chuang et al., *supra* note 42; Wang & Blei, *supra* note 43.

[50] In 2016, David Law identified only two instances of topic modeling in legal research. *See* Law, *supra* note 6, at n.31.

[51] Bernhard Ganglmair & Malcolm Wardlaw, Complexity, Standardization, and the Design of Loan Agreements (2017), *available at* https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2952567.

[52] Law, *supra* note 6.

[53] Greg Leibon et al., *Bending the Law: Geometric Tools for Quantifying Influence in the Multinetwork of Legal Opinions*, 26 ARTIFICIAL INTELLIGENCE & L. 145 (2018); Livermore et al., *supra* note 5.

[54] Elliott Ash et al., The Language of Contract: Promises and Power in Union Collective Bargaining Agreements (2019), *available at* https://extranet.sioe.org/uploads/sioe2019/ash_macleod_naidu.pdf.

[55] For a concise summary with direct applicability to law, see Lee & Mouritsen, *supra* note 3, at 828–80. *See also* MCENRY & WILSON, *supra* note 3, at 1.

meaning of words and phrases in statutory interpretation.[56] Among law schools, Brigham Young University has invested most heavily in the marriage of corpus linguistics and law, building out two corpora of texts—a database of 5.2 billion words from web-based newspapers and magazines since 2010 and a historical database of over 400 million words from the 1810s to the 2000s—that formed the basis for some of Justice Lee's work.[57] Researchers can perform functions that count word frequency, identify other words located in close proximity, and display concordance—or a key word in the context of its surroundings. These capabilities help piece together the ordinary meaning and semantic contexts of key words. This approach has caught on as a new form of empirical analysis, possibly even hewing close to originalism.[58]

Understandably, corpus linguistics has provoked denunciation. Legal scholars have decried as "radically acontextual" the separation of statutory language "from its distinctly legal context."[59] Word frequency and collocation crowdsource for meaning by scanning random corpora of natural language, including sources of dubious judicial value such as Urban Dictionary.[60] All in all, as critics point out, the faith of corpus linguistics adherents in *technique* seems to derive from a mistrust of judicial *discretion*, as if word frequency conveys a more objective, verifiable truth than the intuition of judges.[61] In countering that judges may be more accountable for their decisions than machine learning technocrats,[62] critics echo a broader skepticism of artificial intelligence as an unelected,

[56] *See, e.g.*, Thomas R. Lee & James C. Phillips, *Data-Driven Originalism*, 167 U. PA. L. REV. 261 (2019); Lee & Mouritsen, *supra* note 3.
[57] *See NOW Corpus (News on the Web)*, BYU, http://corpus.byu.edu/now [http://perma.cc/UTD2-BC86].
[58] *See Law & Corpus Linguistics*, BYU Law, https://lawcorpus.byu.edu/ (last accessed Feb. 13, 2020). Other recent examples of its application include Jennifer Mascott, Who Are "Officers of the United States"?, 70 STAN. L. REV. 443, 564 (2018); Josh Blackman and James C. Phillips, *Corpus Linguistics and the Second Amendment* (Harvard Law Review Blog, Aug. 7, 2018).
[59] Zoldan, *supra* note 16, at 447.
[60] *Id.* at 417.
[61] *See* Hessick, *supra* note 40, at 1512.
[62] *See id.* at 1516–17.

unaccountable decision-maker that is incapable of empathy.[63] For these scholars, corpus linguistics may offer an impartiality that is too simply elusive to stand in place human analysis and judgment.

We can reach even further back to find a comparable antecedent for topic modeling in empirical legal studies ("ELS"), which approaches law through empirical methods.[64] ELS has a rich history,[65] one that cannot be adequately summarized here, but in the course of that history, it has had to contend with two criticisms that are relevant to the rise of topic modeling. The first is that empirical research has suffered a "credibility revolution" in its use of observed data to make causal inferences.[66] Starting in economics, this revolution forced empirical researchers in all fields to root out bias through better research design.[67] Related to this point about rigor is a second critique—that ELS lacks a theoretical framework. This charge manifests as different variations: that ELS scholarship is too data-driven,[68] that it fetishizes technique at the expense of the bigger picture.[69]

Topic modeling, of course, is distinct from both the techniques of corpus linguistics and the approach of ELS. When legal texts are taken as the datasource, topic modeling avoids the corpus linguistics pitfall of looking to irrelevant sources. Corpus linguistics, by contrast, is usually deployed in the hunt for ordinary meaning as part of statutory interpretation, which

[63] *See* Kiel Brennan-Marquez & Stephen E. Henderson, *Artificial Intelligence and Role Reversible Judgment*, 109 J. CRIM. L. & CRIMINOL. 137 (2019).

[64] For a more fundamental summary of empirical legal studies, see Theodore Eisenberg, *The Origins, Nature, and Promise of Empirical Legal Studies and a Response to Concerns*, 2011 U. ILL. L. REV. 1713, 1720 ("a core principle [of empirical legal studies] seems indisputable: it is better to have more systematic knowledge of how the legal system works rather than less, regardless of the normative implications of that knowledge").

[65] *See, e.g.*, *id.*; Ho & Kramer, *supra* note 2.

[66] Copus, *supra* note 8, at 21.

[67] Angrist & Pischke, *supra* note 2; Copus, *supra* note 8, at 21.

[68] Eisenberg, *supra* note 64, at 1732–33.

[69] Brian Leiter, *On So-Called "Empirical Legal Studies" and Its Problems*, BRIAN LEITER'S L. SCH. REP. (July 6, 2010, 6:41 AM), http://leiterlawschool.typepad.com/leiter/2010/07/on-socalled-empirical-legal-studies.html.

justifies departing from legal texts.[70] The analogy to ELS also breaks down if topic modeling is not being used for predictive purposes. After all, topic modeling was invented by computer scientists as an information retrieval mechanism, even though it has since been adopted as a predictive mechanism.[71] We, too, employ topic modeling to gather information and verify doctrinal claims rather than make predictions. Functionally, our use of the tool diverges with one of the primary goals of ELS and the subject of its denigration.[72]

Nonetheless, topic modeling is still vulnerable to the same reproach of overreliance on disembodied words that plagues corpus linguistics.[73] More specifically, how can we account for context in performing statistical analysis (fundamentally a quantitative endeavor)? We anticipate questions from ELS as well. How can we ensure that topic modeling does not merely dazzle with its technical prowess but shows us something significant? Put differently, why should we care about these results? And if the method is so important, what steps have we taken to guarantee its rigor?

These questions will be answered in turn in the following sections.

B. *Criticisms from Digital Humanities and Computer Science*

In theory, topic modeling illuminates patterns that cannot be seen by the human eye, at least not with traditional close readings of text. It is a form of distant reading, which considers texts "from afar, using statistics to support large-scale claims."[74] While distant reading appears to belie the close textual analysis that underpins legal research, especially qualitative doctrinal research, the reality is that law scholars

[70] For a summary of this hunt for ordinary meaning, see Lee & Mouritsen, *supra* note 3, at 796–800.
[71] *See* Benjamin M. Schmidt, *Words Alone: Dismantling Topic Models in the Humanities*, 2 J. DIGITAL HUM. 49 (2012).
[72] *See* Copus, *supra* note 8.
[73] *See* Da, *supra* note 12.
[74] Michael A. Livermore & Daniel N. Rockmore, *Distant Reading and the Law*, *in* LAW AS DATA, *supra* note 4, at 4. *See also* Klein, *supra* note 36.

have implemented the quantitative methods of social sciences for decades, and CLA methods are merely the latest development. Distant reading can spur interesting collaborations on legal research, particularly in formulating the type of systematic review that can vet the claims of doctrinal work.[75]

DH and computer science, however, have lived with topic modeling far longer; there, criticisms of the tool are well-developed. Detractors of the computational approach to reading charge that it is "prone to fallacious overclaims or misinterpretations of statistical results because it often places itself in a position of making claims based purely on word frequencies without regard to position, syntax, context, and semantics."[76] More pointedly, the excitement around topic modeling merely stems from the fact that it seems to work better than other "rearrangement algorithms"; without the proper supervision, the tool resembles a "bad research assistant" that produces inexplicable and misleading results as much as "flickers of deeper truths."[77]

Context is therefore central to the viability of topic modeling. Robust visualizations must be able to show the texts from which the words are drawn—or, with legal texts, the cases that are statistically most likely to be comprised of the words that constitute a topic. Relatedly, it is possible to focus too much on a few discrete topics and lose the forest for the trees, so topics must be surveyed as a whole rather than in isolation.[78] The opposite is also true: topic modeling can overwhelm users as much by the grandness of its topics (i.e., too many topics) as by the exquisiteness of its detail (i.e., too many terms within a topic). To borrow from *Gulliver's Travels*, the eighteenth century satire at the outset of the Enlightenment's scientific discoveries,

[75] *See id.* at 16; William Baude et al., *Making Doctrinal Work More Rigorous: Lessons from Systematic Reviews*, 84 U. CHI. L. REV. 37 (2017).
[76] Da, *supra* note 12, at 611.
[77] Schmidt, *supra* note 71.
[78] Andrew Goldstone & Ted Underwood, What *Can Topic Models of PMLA Teach Us About the History of Literary Scholarship?*, 2 J. DIG. HUM. (2012), http://journalofdigitalhumanities.org/2-1/what-can-topic-models-of-pmla-teach-us-by-ted-underwood-and-andrew-goldstone/.

topic modeling creates both a gargantuan world and a miniscule world, and the user may be adrift at both extremes.[79] For this reason, a topic modeling interface must simultaneously be able to break topics down to their constituent words and aggregate them into networks.[80] We respond to these critiques by building visualizations that can do both, as presented in the next Subsection.

Contextual shifts can also come from words themselves. Over time, for instance, usage evolves; spellings change, registers shift, and terms assume ironic connotations.[81] Topics must capture all the dynamic ranges of a word to be comprehensive. To cite a more specific example from antitrust, the prevailing paradigm of market power is first to define the relevant product and geographic markets and then to calculate the market shares of the defendant within those markets.[82] Our algorithms capture several topics where the term "relevant" is featured prominently. As a robustness check, we review the cases within those topics to ensure that "relevant" refers to market definition rather than the relevance of a legal or factual argument.[83]

Beyond decontextualization, DH and computer science identify other deficiencies of topic modeling. Some of them are relevant to legal scholarship and can be addressed; others may be relevant but cannot be programmed around. In the first camp is the argument that the computer scientists who created LDA intended topic modeling to perform functions quite different than what DH scholars have made them do.[84] David

[79] *See* JONATHAN SWIFT, GULLIVER'S TRAVELS (1726).
[80] *See* Schmidt, *supra* note 71.
[8181] *See, e.g.*, *id.* ("In any 150-year topic model, for example, the spelling of "any one" will change to "anyone," "sneaked" to "snuck", and so forth. The model is going to have to account for those changes somehow, either by simply forcing all topics to occupy narrow bands of time, or by assuming that the vocabulary of (say) chemistry did not change from 1930 to 1980.").
[82] *See* HERBERT HOVENKAMP, FEDERAL ANTITRUST POLICY: THE LAW OF COMPETITION AND ITS PRACTICE § 6.4 (5th ed. 2011).
[83] *See infra* Section IV.
[84] Schmidt, *supra* note 71 ("New ways of reading the composition of topics are necessary, because humanists seem to want to do slightly different things

Blei, one of the pioneers of LDA, had envisioned topic modeling as an information retrieval algorithm that made "large collections of text browsable by giving useful tags to the documents," a function similar to Westlaw's insertion of headnotes.[85] Precursors of LDA, including most prominently latent semantic analysis from the 1990s, were designed for information retrieval and indexing as well.[86] However enticing it may be to harness topic modeling and other CLA techniques for prediction of, say, litigation outcomes, these tools might be better restricted to discrete retrieval, indexing, and archival functions in law, at least until legal scholars have a better grasp of their capabilities.[87] Those functions, as we shall argue later, include tagging caselaw with helpful metadata to enable more efficient browsing, rather than to make predictions about how a case might come out.

When empirical techniques are used for prediction, they draw scrutiny instantly. CLA methods are no different.[88] Yet even if topic modeling is not used to forecast outcomes, it can fail simple robustness and reproducibility checks. Scholars have shown that if a corpus of text is changed slightly (e.g., 1% of the original sample removed), the ensuing topics are entirely different.[89] Similarly, the modeling sampled in prominent DH papers have not always withstood reproduction by others.[90] These methodological concerns question whether topic modeling is stable and verifiable.

with topic models than the computer scientists who invented them and know them best.").

[85] *Id.*

[86] *See* Scott Deerwester et al., *Indexing by Latent Semantic Analysis*, 41 J. AM. SOC. INFO. SCI. 391 (1990).

[87] To some extent, this inclination is understandable. The predictive possibilities of text analytics draw grant funding and industry–university partnerships. For an interesting account at Georgia State University, see Charlotte S. Alexander, *Using Text Analytics to Predict Litigation Outcomes*, *in* LAW AS DATA, *supra* note 4.

[88] *See* Copus, *supra* note 8. For a discussion of how these discussions may hamper our imagination of what CLA can do, see Allen Riddell, *Prediction Before Inference*, *in* LAW AS DATA, *supra* note 4.

[89] Da, *supra* note 12, at 628.

[90] *Id.* at 628–29.

Finally, because humanities scholars have rebuked the digitization of their field in ways that have some applicability to legal research. These include institutional and political economy criticisms that DH replicates a Silicon Valley ethos of disruption for disruption's sake while masking a neoliberal takeover of university research functions.[91] Computational analysis saps institutions of traditional scholarly research and writing, replacing these functions instead with grants-dependent research labs. Additionally, the corpora from which documents are extracted and the programmers coding the algorithms often do not accommodate diverse perspectives.[92] These shortcomings are important to bear in mind as CLA moves forward, even if they are not completely within the control of law scholars.

C. *Aggregated Modeling*

We take seriously the criticisms levied at topic modeling from DH and computer science. Accordingly, we have constructed a way to aggregate up to six different LDA topic models in one iteration. In this way, we create "model of models" that addresses some of the contextualization, robustness, and reproducibility concerns surrounding the tool. This Subsection introduces the features of our aggregated modeling; we believe it still must be paired with other modeling tools for caselaw research to be comprehensive, and we leave the next Section to fully describe our methodology. We argue that the full suite of these topic modeling tools can streamline caselaw research by adding metadata, comprised of topics and their constituent terms, to signal relevance to a user's research questions. Because topic modeling is still rather novel in law, we would also restrict them to information retrieval rather than predictive functions for now.

[91] *See* Danielle Allington et al., *Neoliberal Tools (and Archives): A Political History of Digital Humanities*, LOS ANGELES REV. BOOKS, May 1, 2016.
[92] *See id.*; Klein, *supra* note 36.

Several improvements to traditional topic models flow from their aggregation. First, our visualizations place topics in both large and small contexts.

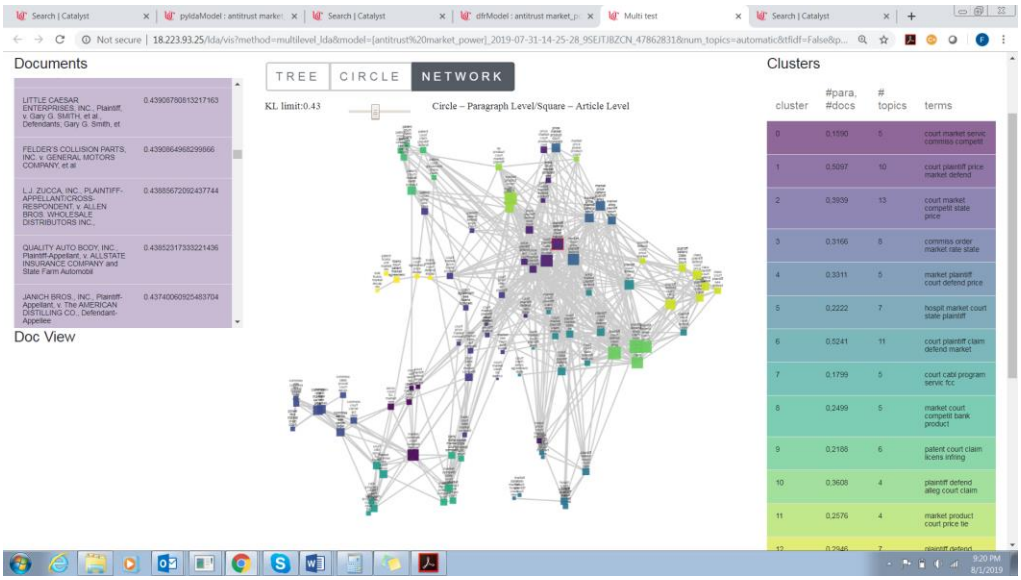

Figure 1: Network View of Market Power Cases in Model of Models

Figure 1 shows a network of antitrust market power cases distributed as topic clusters across space. A topic cluster is an aggregation of multiple topics, where each topic is a collection of terms that are statistically most likely to appear together. The right-hand panel lists each cluster as a distinct shade of color; the clusters are also numbered. In addition, each cluster displays the number of topics that comprise the cluster as well as the top words in the topics. The central graphic depicts the relationship among the clusters. The left-hand panel lists the top "documents," or cases, within a topic as well as the relevant metadata (e.g., case name). It also enables the retrieval of cases.

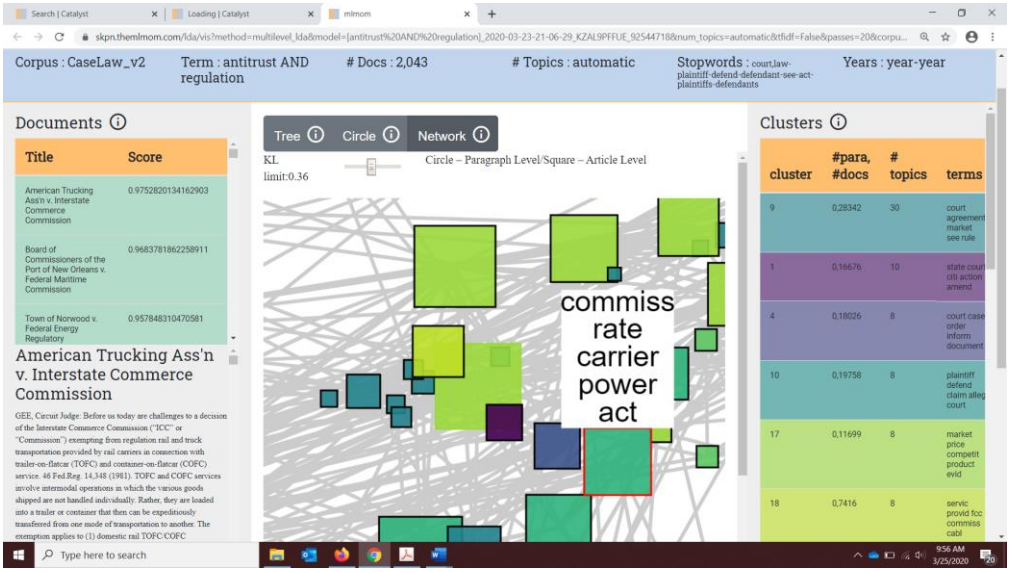

Figure 2: Close-Up View of Antitrust–Regulation Cases with Document Retrieval

Figure 2 demonstrates the case retrieval function on a corpus of antitrust–regulation cases: The highlighted topic cluster in the center encompasses topics with the terms "commiss[ion]," "rate," "carrier," "power," and "act," while the document retrieval feature enables the user to pull up specific cases. Here we have chosen to highlight *American Trucking Associations, Inc. v. I.C.C.*, the top case in the cluster.[93] Note the "top" case means that case that is cross-listed in the most topics.

Our aggregated modeling presents two levels of information: cluster networks show the connections among the topics, while the document retrieval interface shows the specific cases that contribute to each topic. In this way, topics are contextualized at both the macro- and the microscopic levels. The two scales of analysis allow us to see the full complexity of the corpus as a spatial arrangement of how terms are scattered across the cases that comprise the network.

The visualizations employ vector space modeling, with topic clusters are strewn across space. In classic vector space models, such as Google's Word2Vec, algorithms process the conceptual relations between words and depict each word as a

[93] 656 F.2d 1115 (5th Cir. 1981).

vector, or dimension, in space.[94] The angle between two vectors, or their cosine, portrays the magnitude of difference between those words. The dimension reduction approach of Word2Vec aids in interpretability, portraying related words as crowding together and dissimilar words as far-flung. This intuition, that related words congregate, informs our visualizations of topic clusters in the neural network architecture, where topics congregate.

To bolster model stability, a feature that traditional topic modeling sometimes lacks,[95] our algorithms run topic models at least twenty times for each query. As with any empirical project based on copious amounts of data, topic modeling is subject to margins of error, or "wobble." As the next Section details, we run variations of more traditional topic modeling as comparators for each query. In comparison, model aggregation reduces the wobble significantly because the process only picks up the most stable and persistent topics across multiple iterations.

The frequency of iterations also helps to present topics more coherently. Insignificant topic clusters are removed on multiple runs, so the aggregation ensures that visualizations present larger networks that have picked up truly significant term repetitions, rather than statistically aberrant patterns.

In the end, we deploy topic modeling not so much for its predictive ability or even its insight into the meaning of words in themselves but for its indexing and information retrieval capabilities. In creating a case retrieval function, we permit users to pull up the texts which showcase the words of a topic model in their original context. All the additional information presented in the visualizations—from network connections to topic clusters—can be taken as metadata that signal the relevance of antitrust cases to particular words and topics. This,

[94] *See* Thomas Mikolov, *Distributed Representations of Words and Phrases and their Compositionality*, *in* ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 26 (C.J.C. Burges et al. eds., 2013); Elliott Ash & Daniel L. Chen, *Case Vectors: Spatial Representations of the Law Using Document Embeddings*, *in* LAW AS DATA, *supra* note 4, at 315–7.

[95] *See, e.g.*, Da, *supra* note 12, at 625.

in effect, is the same functions that subscribers pay lavish fees to Westlaw and Lexis for. It is, as we shall argue, a necessary check to the proprietary databases, which are notoriously opaque about their algorithms.

III. METHODOLOGY

Big data caselaw research is often hindered by two primary obstacles. First, only a few repositories hold a corpus of easily extractable caselaw.[96] Second, even if cases could be easily extracted, their interpretation is limited by modeling that can translate machine analysis into intuitive visualizations.[97]

This Section details how we are using recent technical advances to overcome the hurdles to data extraction and data interpretation. In October 2018, Harvard Law School unveiled its Caselaw Access Project ("CAP"), which had digitized all book-published U.S. case law between 1658 and 2018, some 40 million pages.[98] We have created two pools of cases out of the CAP dataset: 36,000 federal cases bearing the word "antitrust" and 305,000 federal cases bearing the word "regulation." We whittle the first pool down to 2,591 cases with the term "market power" (the "Market Power Corpus") and the second pool down to 7,308 with the term "antitrust" (the "Antitrust–Regulation Corpus").

As for data interpretation, we adjust traditional topic modeling algorithms to generate visualizations of both pools of cases. We pair our aggregated modeling with open-source topic modeling algorithms, so the composites reflect the various dimensions of the corpora. The open-source visualizations are fairly easy to replicate: they incorporate the work of programmers and DH scholars who have made the tools freely

[96] The leading commercial databases, Westlaw and Lexis, are not conducive to high-volume data mining because they require licenses and complicated APIs. Other platforms, such as the U.S. Securities and Exchange Commission's EDGAR filing system or the U.S. Federal Register, do not hold cases. Despite the proclivity of law for natural language text mining, easy access to copious amounts of case law is limited.

[97] *See* Chuang et al., *supra* note 42.

[98] *About*, CASELAW ACCESS PROJECT, *supra* note 15.

available.[99] While we have selected this suite of topic modeling algorithms for fit to one another, we have also done so out of the interests of transparency and reproducibility. Our hope is that data interpretation will be as open as data extraction.

This Section begins by introducing the CAP. Then it explains our data and access procedures, before concluding with our modeling and visualization processes.

A. *The Caselaw Access Project*

CAP, a partnership among Harvard Law School's Library Innovation Lab, its Berkman-Klein Center, and the legal research company Ravel Law, spent over three years to simply digitize all court decisions published in the 40,000 bound volumes in the Harvard Law School Library.[100] The resulting database is the most comprehensive of its kind outside of the Library of Congress.[101] CAP's cases span some 360 years and all federal and state courts, as well as territorial courts in American Samoa, Dakota Territory, Guam, Native American Courts, Navajo Nation, and the Northern Mariana Islands.[102]

The great advantage of the CAP dataset is that cases are provided in a clean, digestible form, so users need not write application programming interfaces ("APIs") to pull data. Texts are presented in machine-readable format, greatly simplifying big data projects. Cases can be extracted through either APIs or bulk downloads.[103] By contrast, commercial legal databases

[99] *See* The topic browser visualization is adapted from Andrew Goldstone's dfr-browser project. See Andrew Goldstone, *DFR-Browser: Take a MALLETT to Disciplinary History*, https://agoldst.github.io/dfr-browser/ (last accessed Feb. 27, 2020); Ben Mabey, *Welcome to PyLDAvis's Documentation*, https://pyldavis.readthedocs.io/en/latest/index.html (last accessed Feb. 27, 2020).

[100] *See About*, *supra* note 15.

[101] Jason Tashea, *Caselaw Access Project Gives Free Access to 360 Years of American Court Cases*, ABA J., Oct. 30, 2018.

[102] *About*, *supra* note 15.

[103] *Id.*; Tashea, *supra* note 101.

require users to utilize their own APIs to pull large volumes of cases, as well as the execution of license agreements.[104]

CAP will disrupt legal research. By making freely available all published decisions in nearly every U.S. jurisdiction, it threatens the Westlaw and Lexis paywalls, greatly expanding legal access for anyone with an Internet connection. The database is free for the public, though LexisNexis, which now owns Ravel Law, controls commercial use.[105]

Apart from comprehensiveness, CAP is also run on a versatile interface that has shared stock APIs for software developers.[106] It also includes basic searching, browsing, and downloading functions, as well as the ability to explore historical trends in the caselaw.[107] Given CAP's flexibility and ease of use, law scholars have already begun using it for big data projects.[108]

CAP does have limitations. Notably, it excludes cases published after June 2018 and cases not designated as officially published, such as some lower court decisions. The scope restrictions also leave out unpublished trial documents, such as filings and exhibits. Nonetheless, the corpus is large enough to compile rich models and graphs.

B. *Data and Access*

Data for our project was made available through CAP, which contains 6.7 million unique cases (and over 1.7 million federal cases). Having applied for and obtained researcher access from CAP, we gathered data by writing python-based

[104] We spent close to a year negotiating license agreements with Westlaw and Lexis and also tinkering with APIs—until CAP went live.
[105] Tashea, *supra* note 101.
[106] *See id*.
[107] *See Tools*, CASELAW ACCESS PROJECT, https://case.law/tools/ (last accessed Mar. 26, 2020).
[108] *See, e.g.*, Jaromir Savelka et al., *Improving Sentence Retrieval from Case Law for Statutory Interpretation*, ICAIL (2019); Jonathan H. Choi, *An Empirical Study of Statutory Interpretation in Tax Law* (2019), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3460962.

calls to its API. CAP's own APIs feature tools that permit searching through all text in selected cases (as opposed to searches using tags or other metadata). We created two pools of cases: all federal cases with the word "antitrust," a total of approximately 36,000 cases; and all federal cases with the word "regulation," a total of approximately 305,000 cases. These serve as the bases for our Market Power Corpus of 2,591 cases from the "antitrust" pool and our Antitrust–Regulation Corpus of 7,308 cases from the "regulation" pool.

At first glance, these numbers seemed small to us, particularly the count of 36,000 for all federal antitrust cases. However, two limitations help explain the variance: first, CAP stops in 2018 at Volume 281 of the third series of the Federal Supplement and Volume 881 of the third series of the Federal Reporter, omitting approximately two years of recent cases.[109] Second, CAP excludes unpublished decisions.

We verified the case counts in the Market Power Corpus and the Antitrust–Regulation Corpus in several ways. A Westlaw search and subsequent filter for reported federal cases with the terms "antitrust and 'market power'" returned 2,732 cases; for reported federal cases with the terms "antitrust and regulation," this number was 9,775. We also utilized CAP's historical trends interface for verification. CAP has a little over 1.7 million unique federal cases in its corpus, and a search in historical trends reveals that antitrust cases have comprised a low of about 0.1% to a high of almost 4% of all federal cases, with a median roughly short of 2% (or about 34,000 cases).[110] Overall, we have more than a robust sampling for federal antitrust cases.

Manual assessment quickly becomes impracticable when examining a corpus as extensive as CAP. Thus, the application of machine learning provides a more manageable

[109] E-mail from Adam Ziegler, CAP, to F. Chang, on Feb. 10, 2020.

[110] A simple search using CAP's historical trends function reveals that antitrust cases rose to a high of 4% of all federal cases in the 1980s. *See Historical Trends*, CASELAW ACCESS PROJECT, https://case.law/trends/ (search for "us: antitrust"). We also verified CAP's count of federal antitrust cases, which was roughly 32,000.

approach. We use LDA as the baseline algorithm to sort through each case's natural language and produce models of topics based on the clustering of frequently recurring words.[111] LDA proceeds in two steps: first, the algorithm groups words that have a high probability of co-occurrence into word clusters, or topics; then, it identifies the decisions where each topic is most likely to appear. This computational approach to language allows us to see certain trends through topics generated from the case law documents' own semantic and syntactic structures, rather than applying human data and metadata structures to a dataset. Put differently, machine learning has the potential to provide a neutral way of ordering this volume of case law, devoid of human—and doctrinal—preconceptions.

C. *Modeling and Visualizations*

Using Elasticsearch (a full-text search and analytics engine)[112] and the python Gensim package,[113] we built a web-based platform. The platform performs topic modeling by using the unsupervised machine learning clustering algorithm LDA to sift through cases. LDA models are generated based on the distribution of latent topics in a document and the distribution of words in those topics.[114] Each topic is constructed based on a probability distribution of words.[115] For instance, one topic

[111] *See* Blei et al., *supra* note 5.

[112] *Elasticsearch: The Heart of the Elastic Stack*, ELASTIC, https://www.elastic.co/products/elasticsearch (last accessed Sept. 14, 2019).

[113] *Gensim 3.8.1*, PYTHON PACKAGE INDEX, https://pypi.org/project/gensim/ (Sept. 26, 2019 data release) (last accessed Oct. 20, 2019).

[114] Blei et al., *supra* note 5.

[115] For a more detailed explanation, see Chuang et al., *supra* note 42:

> Given as input a desired number of topics K and a set of documents containing words from a vocabulary V, LDA derives K topics β_k, each a multinomial distribution over words V. For example, a "physics" topic may contain with high probability words such as "optical," "quantum,"

might feature the term "market" with high probability, whereas its association with another topic will not be as strong. Similarly, one document might have a high presence of topic 1, pertaining to procedural and evidentiary matters, whereas that same topic only features faintly in another document. In this project, we have defined a "document" as an individual case from our dataset.

As with any empirical project based on copious amounts of data, term relevance and topic modeling are subject to margins of error, which we affectionately call the "wobble." We have found that the wobble is slight for two of the three types of visualizations (topic browser and pyLDAvis) and virtually negligible for the third (aggregated). As discussed in the prior Section, aggregated modeling minimizes variance by running any query up to twenty times.

The models provide visualizations of cases grouped by recurring terms, depicting both the relationships among terms and the relationships among groups of cases. We rely on three types of visualizations, all built around topic modeling. In totality, the visualizations capture the full nuances of the topics. The remainder of this Subsection explores all three types, using the Market Power Corpus as the dataset.

The first set of visualizations are generated by our unique aggregated modeling algorithms. These create "multilevel" or "model-of-models" visualizations that provide a hierarchical view of topics and topic clusters in three different formats—tree, circle, and network (see Figures 3–6).

> "frequency," "laser," etc. Simultaneously, LDA recovers the per-document mixture of topics θ_d that best describes each document. For example, a document about using lasers to measure biological activity might be modeled as a mixture of words from a "physics" topic and a "biology" topic.

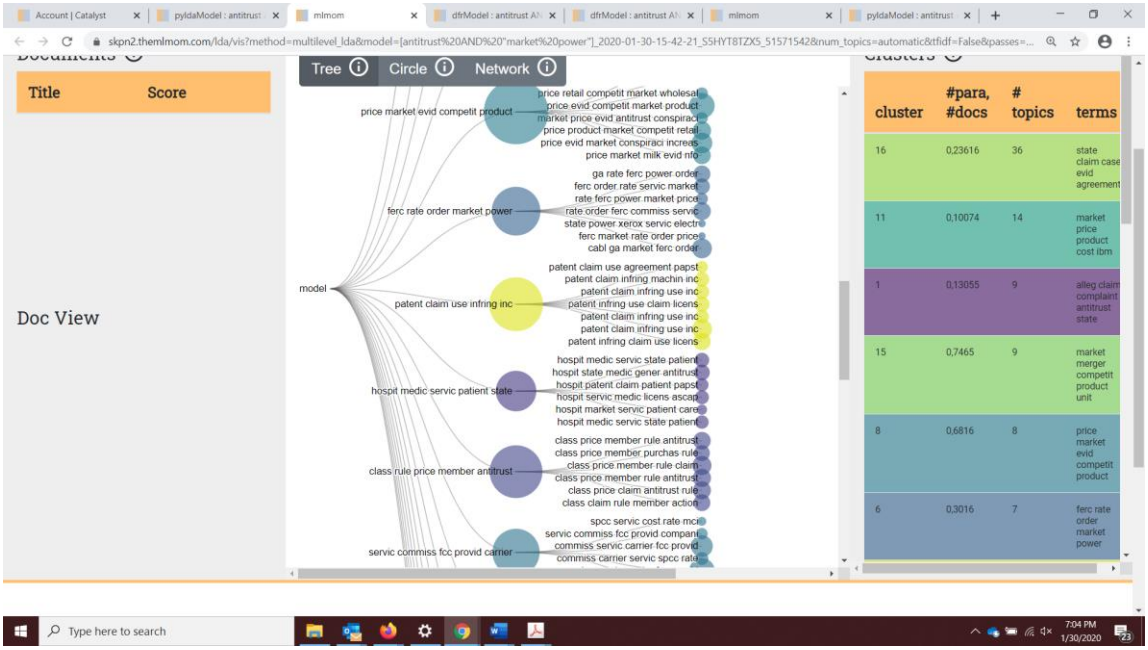

Figure 3: Multilevel Visualization of Market Power Cases in Tree Format

In the tree format of Figure 3, the smaller nodes on the right represent topics (e.g., machine-grouped terms "price," "retail," "competit[ion]," "market," and "wholesal[e]"), while the larger nodes represent clusters of topics (e.g., a cluster with "price," "market," "evid[ence]," "competit[ion]," and "product"). The size of each cluster node or topic node represents the significance of the cluster or topic to the overall corpus. The right-hand bar shows the number of topics within each cluster (thereby functioning as a proxy for the cluster's diversity), and the left-hand bar lists the top cases in each topic.

Circle view presents the same information, but in a format that more clearly conveys the topics where each word appears. Clicking on a specific word pulls up how it is shared across topic clusters. For example, Figure 4 (below) shows the recurrence of the term "market" within all topics. In contrast, network view constructs a spatial representation where each topic comprises a vector in space (see Figure 1 above). It is

adapted from the neural network architecture Word2Vec, where each word represents a vector.[116]

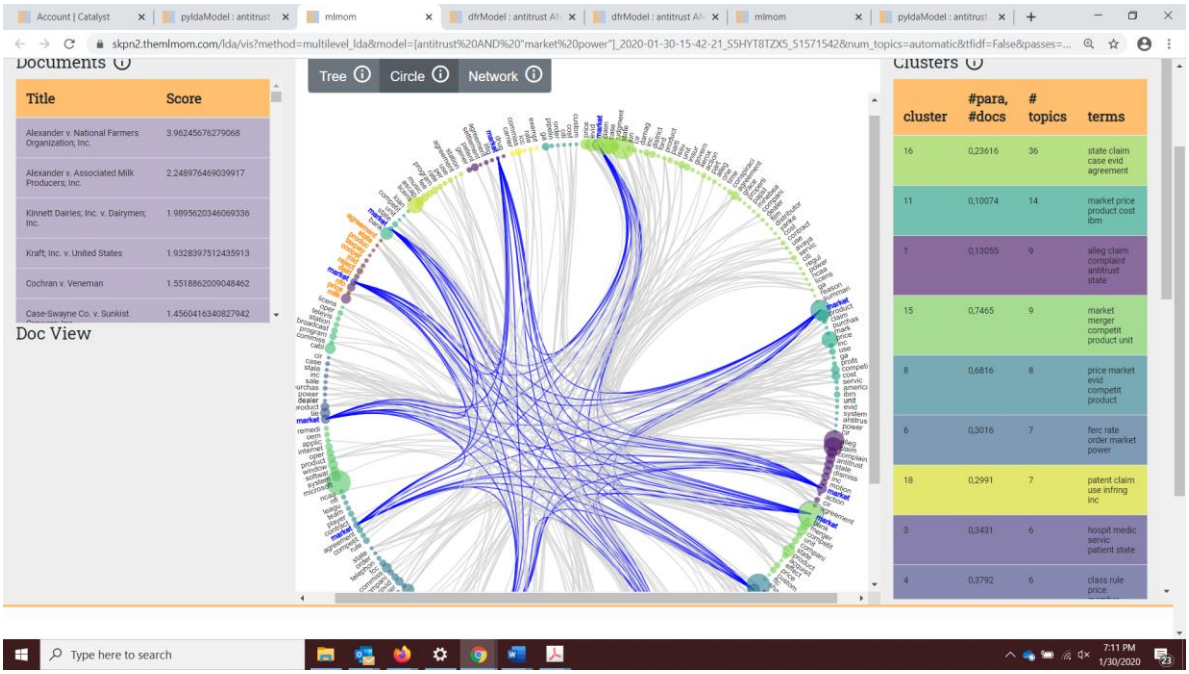

Figure 4: Multilevel Visualization Showing the Recurrence of the Term "Market"

The second set of visualizations, "topic browser," are generated from the DFR framework of Andrew Goldstone, a DH scholar. Topic browser visualizations organize cases into topics, enabling detailed analyses of where (i.e., in what topics) certain terms recur (see Figures 5 and 6).

[116] Mikolov, *supra* note 94.

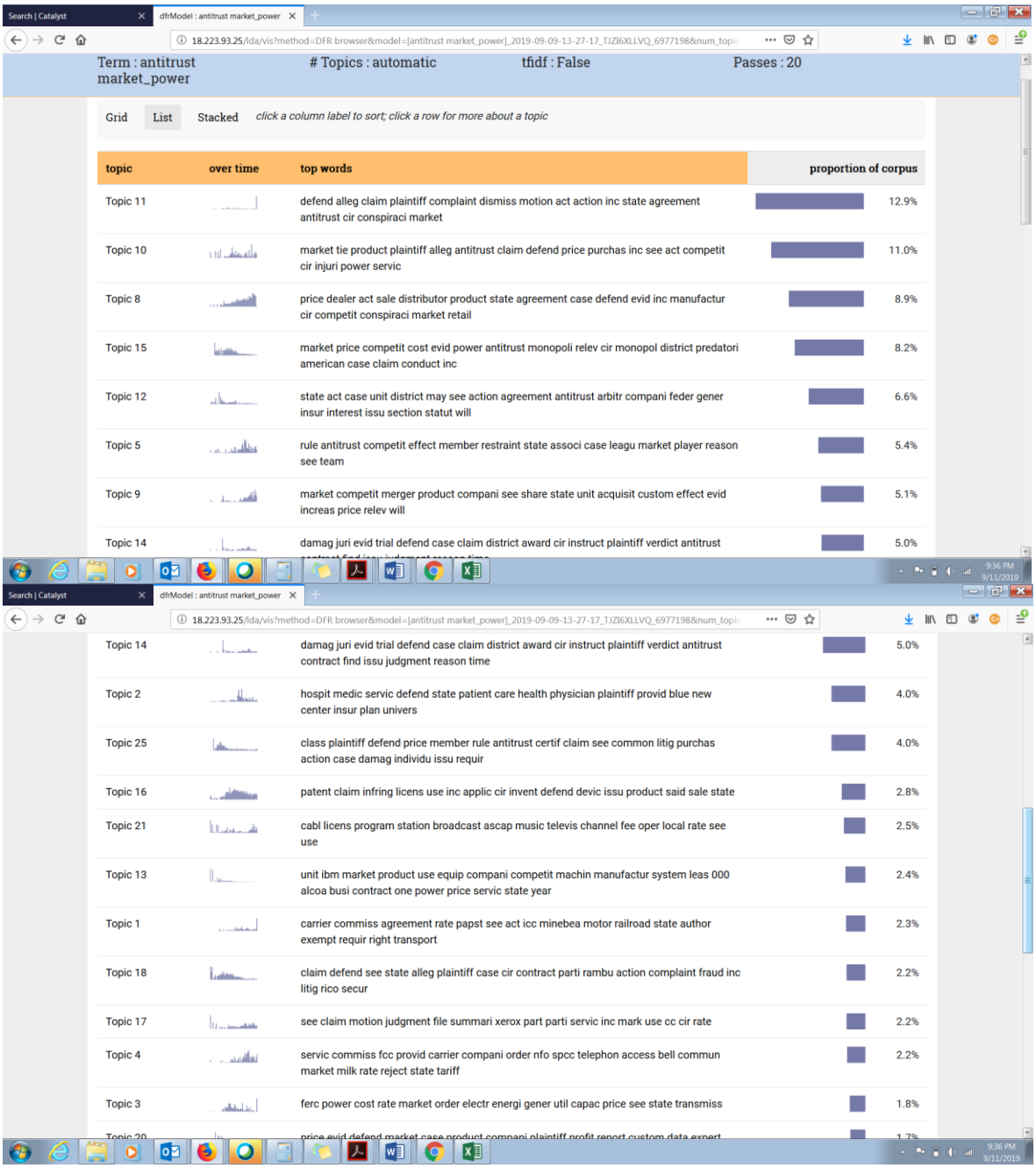

Figure 5: Topic Browser Visualization of Market Power Cases in List Format

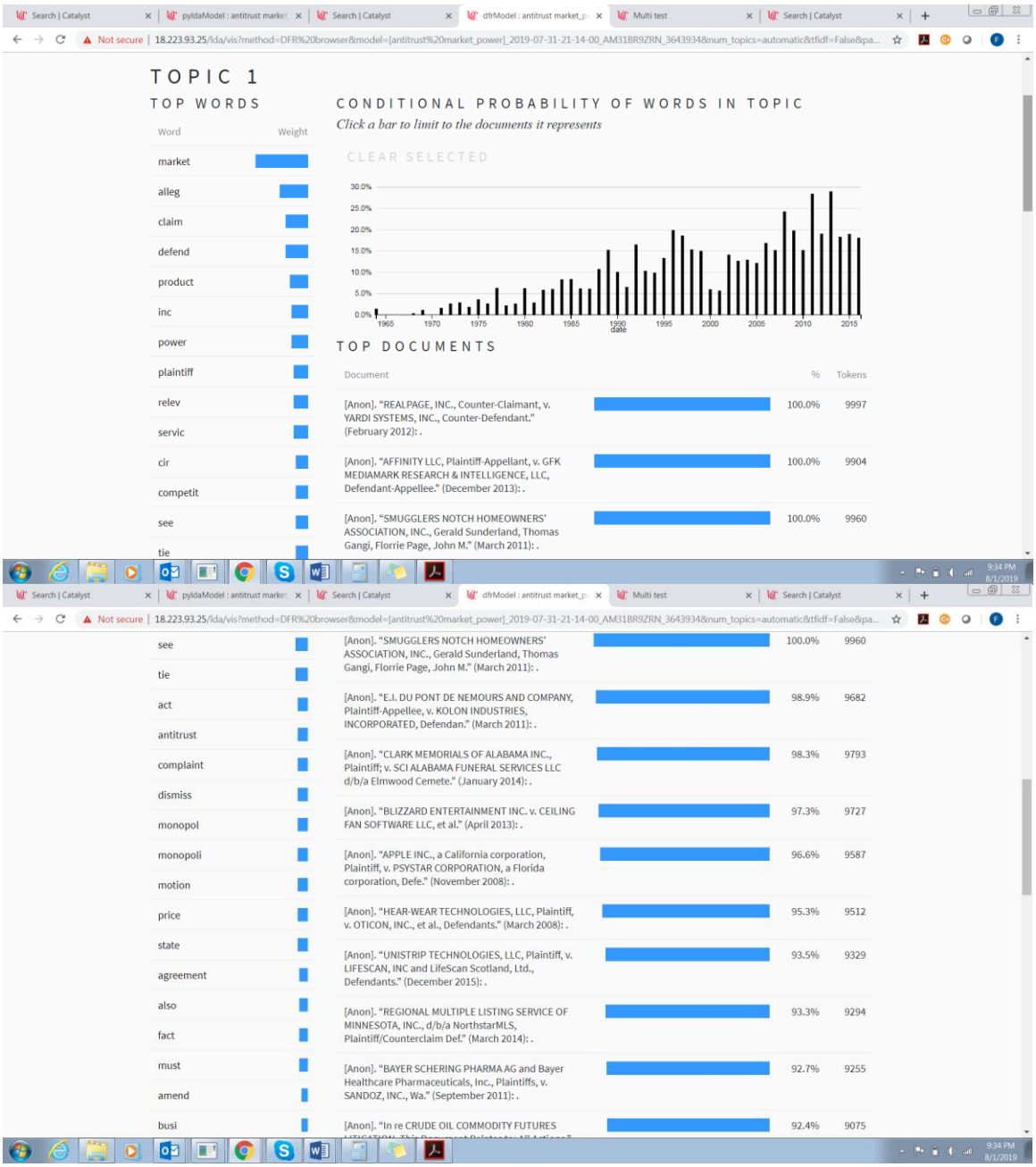

Figure 6: Breakdown of Terms and Cases within a Topic in Topic Browser View

From the overview in Figure 5, the user can browse a specific topic by clicking on it, which brings up the topic's top terms and cases as shown in Figure 6. Both the overview and

single-topic view display histograms on the time periods when certain topics were more prevalent.

Clicking on each term pulls up the topics where the term appears.[117] For instance, Figure 11 (below) shows that "relev" (as in relevant market, which would come up in market definition) appears in only three topics—a slightly surprising result in a corpus of cases dealing with market power.

The third set of visualizations, python-based LDA visualizations ("pyLDAvis"), is built from the framework of the programmer Ben Mabey. PyLDAvis depicts the distance between topics, in a format that most closely resembles the Word2Vec architecture (see Figure 7).[118] Word2Vec is a two-layer neural network devised by Google that assigns each term onto a vector in space. The totality of such a graph represents the entire corpus and can have hundreds of vectors, each corresponding to a term, thereby illustrating the proximity and distance among terms.[119] In the pyLDAvis adaptation, the size of each topic bubble represents the weight of that topic. When a topic is highlighted, the platform pulls up the top probable words contained in that topic.[120]

[117] The topic browser visualization is adapted from Andrew Goldstone's dfr-browser project. *See* Goldstone, *supra* note 99.

[118] pyLDAvis is adapted from package led by Ben Mabey. *See* Mabey, *supra* note 99.

[119] *See* Mikolov, *supra* note 94. For an illustration of Word2Vec, see Jay Alammar, *The Illustrated Word2Vec*, GITHUB, Mar. 29, 2019, https://jalammar.github.io/illustrated-word2vec/. For a criticism from the legal perspective, see Levendowski, *supra* note 124.

[120] For a mathematical expression of probability, one of the key concepts in this statistical analysis, see Carson Sievert & Kenneth E. Shirley, *LDA vis: A Method for Visualizing and Interpreting Topics*, *in* PROCEEDINGS OF THE WORKSHOP ON INTERACTIVE LANGUAGE LEARNING, VISUALIZATION, AND INTERFACES 63, 66 (Jason Chuang et al. eds. 2014). The probability of any term within a topic is its *relevance* within that topic. Relevance can be expressed as $r(w,k) \mid = \lambda \log(\varphi_{kw}) + (1 - \lambda) \log (\varphi_{kw} / p_w)$, where λ is the weight of the probability of term w under topic k relative to its lift.

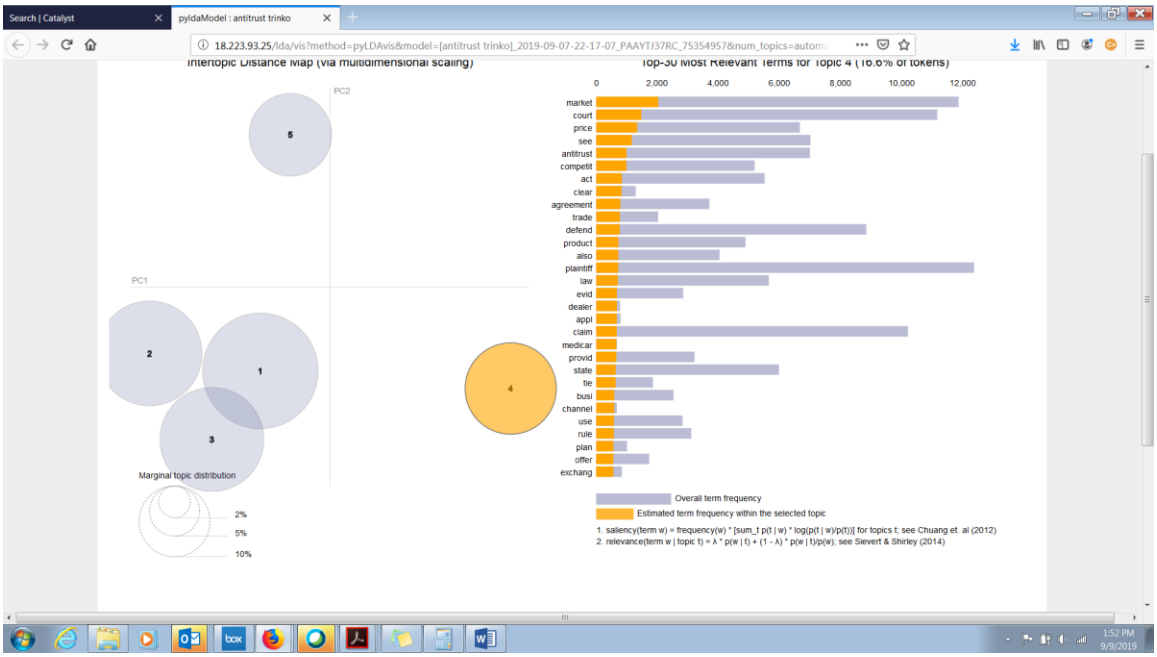

Figure 7: pyLDAvis View of Antitrust Cases Containing "Trinko"

Figure 7 shows how our algorithms have sorted antitrust cases with the word "Trinko" into four topics.[121] In the screen shot, topic 4 is highlighted, bringing up its top terms. With pyLDAvis and the other visualizations, the platform user can set the number of topics manually. Here, the model is comprised of five topic bubbles.

Two additional points are notable. First, generic words such as "court," "see," "claim," and "plaintiff" are prevalent in the initial results. Although their presence renders the topics more generic, their appearance validates our machine learning because antitrust cases are replete with these words—words that algorithms are not trained to filter out.[122] We can refine the results by excluding generic words from the visualizations.[123]

[121] *See* Verizon Communications Inc. v. Law Offices of Curtis V. Trinko, 540 U.S. 398 (2004). *Trinko* reset the balance between antitrust and regulation while also gutting the essential facilities doctrine.

[122] Our platform has the capacity to exclude these generic terms in the construction of visualizations.

[123] Excluded words are tagged as "stop words." At this point, the platform can only filter out up to nine stop words.

Second, these three types of visualizations are different than Word2Vec, which has been the visualization of choice on many legal research projects so far.[124] From a methodological perspective, our project therefore pushes machine learning in legal scholarship beyond word-level analysis, by building topic and even meta-topic models.

Finally, we have begun to read the top cases within each topic to see how courts think through market power. For example, we reviewed cases within the topics highlighted in Figure 11 below where "relev[ant]" was a top word; then we read cases in other topics, where "relev[ant]" had a lower probability distribution, presumably because the relevant product and geographic markets were not defined. (In each topic, cases are ordered by the probability score of that topic's appearance in the case.)

IV. RESULTS

To test our modifications to LDA, we analyze large numbers of federal antitrust cases up to 2018, which we extracted from CAP. The machine-generated visualizations shed light on two vexing areas of antitrust law: market power and the balance between antitrust and regulation. Because law is a text-heavy field, topic modeling is particularly appropriate as an analytical tool. And because antitrust concepts are open-ended and resolved through the deliberation of associated terms, antitrust is an apt place to start.

Our results fall into three categories. The first category consists of big-picture observations that flow from the macroscopic perspective of topic models. These observations validate certain doctrinal views articulated in prior scholarship on matters such as deregulation and market power. The second category is comprised of observations that challenge straightforward interpretation. In these results, the cases do not

[124] For a description of Word2Vec, see Amanda Levendowski, *How Copyright Law Can Fix Artificial Intelligence's Implicit Bias Problem*, 93 WASH. L. REV. 579 (2018).

seem to fit with their categorization in a topic, which raises questions about the sensitivity and accuracy of the algorithms. The third category consists of results that raise questions about traditional caselaw research. These questions include how commercial legal databases execute their searches and what constitutes good precedent in antitrust.

We understand that legal scholars are often skeptical of algorithmic processing and, except for those in the CLA camp, have generally refrained from employing them in research. For all their utopian promises, algorithms in society seem to amplify rather than eliminate human biases.[125] Accordingly, because we rely so heavily on algorithms for this project, we have tried to be cautious in their use and in our conclusions. Therefore, rather than disrupting for disruption's sake, we offer topic modeling as a way to affirm—but also to complicate—traditional research and prior conclusions on antitrust doctrine.

The remainder of this Section offers a doctrinal primer on market power and the antitrust–regulation balance. Then it categorizes our observations.

A. *A Doctrinal Primer*

1. Market power

Market power is a concept fraught with controversy. Conceptually, it is easy to grasp: market power means the ability to set price above a producer's marginal cost.[126] Practically, however, it is difficult to prove. Direct evidence, such as of anticompetitive effects, if often too hard to come by. Hence, courts must abide by circumstantial evidence of market power, which uses market share as a proxy.

---

[125] *See, e.g.*, Amanda Levendowski, *How Copyright Law Can Fix Artificial Intelligence's Implicit Bias Problem*, 93 WASH. L. REV. 579 (2018); Jack Balkin, *The Three Laws of Robotics in the Age of Big Data*, 78 OHIO ST. L.J. 1217 (2017); Dan L. Burk, *Algorithmic Fair Use*, 86 U. CHI. L. REV. 283 (2019); Kiel Brennan-Marquez & Stephen E. Henderson, *Artificial Intelligence and Role Reversible Judgment*, 109 J. CRIM. L. & CRIMINOL. 137 (2019).

[126] William M. Landes & Richard A. Posner, *Market Power in Antitrust Cases*, 94 HARV. L. REV. 937, 939 (1981).

This paradigm—market definition/market share—has become both the prevailing way of gauging market power and, simultaneously, the target of generations of fierce criticism. In the first step of the paradigm, a relevant product market is defined, enabling the subsequent calculation of a defendant's relevant market share.[127] The product market is drawn, in technical terms, as the smallest grouping of sales where the elasticities of demand and supply are low enough that a monopolist controlling the grouping could reduce output and increase price substantially above marginal cost.[128] Then the relevant geographical market is defined along similar lines and the defendant's geographic market share is also calculated.

Market definition has come under fire from scholars for decades because of its imprecision.[129] The controversy stretches back to one of the first major market power cases, *U.S. v. du Pont*,[130] where the Supreme Court accepted a test of market power that came to be so disparaged, the case became the namesake for the error: the cellophane fallacy. In *du Pont*, the Court accepted the defendant's definition of the market as all flexible wrapping materials, including products like wax paper and aluminum foil, rather than cellophane itself[131]—even though these substitutes were able to compete with cellophane precisely because du Pont had been underpricing it.[132] In short, the Court conflated the elasticity of demand for a product with the cross-elasticity, or reasonable interchangeability, of the product and its substitute. For this and other reasons, commentators have condemned market definition for its incoherence.[133]

---

[127] HOVENKAMP, *supra* note 82, at 92.
[128] *Id.* at 93 n.2.
[129] *See, e.g.*, Louis Kaplow, *Why (Ever) Define Markets?*, 124 HARV. L. REV. 437, 440 (2010); Herbert Hovenkamp, *Markets in Merger Analysis*, 57 ANTITRUST BULL. 887, 891, 894–95 (2012); Landes & Posner, *supra* note 125.
[130] 351 U.S. 377 (1956).
[131] *Id.* at 399–400.
[132] *See* Jonathan B. Baker, *Market Definition: An Analytical Overview*, 74 ANTITRUST L.J. 129 (2007).
[133] *See* Kaplow, *supra* note 126. For a reply, see Gregory J. Werden, *Why (Ever) Define Markets? An Answer to Professor Kaplow*, 78 ANTITRUST L.J. 729, 740 (2013).

In dynamic markets, which today consist primarily of Internet markets, circumstantial evidence of market power is less important.[134] Reliance on market definition/market share can even lead to erroneous results—most notably, the inclusion of both merchant and consumer interfaces into a two-sided platform where a complaint alleges harm only to one side.[135]

Nonetheless, examinations of collusion and exclusion are seldom complete without market power analysis of the constituent markets. Market power is the very first step, for instance, in a monopolization action under Section 2 of the Sherman Act,[136] the basis for many of the charges against tech firms.[137] It is therefore a hugely important yet open-ended issue that is assuming even greater urgency.

2. *Balancing antitrust and regulation*

Another contested issue in antitrust is how courts approach competition in regulated industries such as finance, telecommunications, and health care. In the 1960s, cases on the balance between antitrust and regulation such as *Silver v. New York Stock Exchange* followed a "plain repugnancy" standard, where courts strived to permit the cohabitation of regulation and antitrust, precluding the latter only where the former clearly pre-empted it.[138] In the next decade, plain repugnancy became simply repugnancy," under which antitrust was to defer if there was just the potential for conflict with regulation.[139] Significantly, this body of law came in contexts where the statutes in question did not contain an express antitrust savings clause that preserved antitrust actions, so courts were dealing with *implied* antitrust immunity. In 2004,

[134] Howard A. Shelanski, *Information, Innovation, and Competition Policy for the Internet*, 161 U. PA. L. REV. 1663, 1674 (2013).
[135] *See* Ohio v. American Express Co., 585 U.S. ___ (2018). For criticisms, see John M. Newman, *Antitrust in Digital Markets*, 72 VAND. L. REV. 1497 (2019).
[136] This is the "power plus conduct" framework of *Grinnell*, 384 U.S. 563 (1966), and *U.S. v. Aluminum Co. of America*, 148 F.2d 416 (2d Cir. 1945).
[137] CHRIS SAGERS, UNITED STATES V. APPLE: COMPETITION *IN* AMERICA (2019).
[138] *See*, e.g., Silver v. N.Y. Stock Exch., 373 U.S. 341, 357 (1963).
[139] *See* Gordon v. New York Stock Exchange, Inc., 422 U.S. 659 (1975).

however, the Court in *Trinko* found that even a statute with an antitrust savings clause—namely, the Telecommunications Act of 1990—could preclude the application of antitrust laws because of the *potential* for conflict.[140]

Over the last half century, then, the doctrine balancing antitrust and regulation has conferred federal courts greater discretion to dismiss private actions over conduct that might be regulated by administrative agencies. In moving from plain repugnancy to simple repugnancy to presumed repugnancy, this doctrine now requires antitrust to defer when regulation has spoken, however quietly. Concomitantly, however, regulators have undergone a paradigm shift in the last half century, moving away from the filed rate doctrine, whereby natural monopolies had to abide by rates filed with the Interstate Commerce Commission ("ICC").[141] With the gutting and eventual abolition of the ICC, this intrusive regulation was replaced with a framework that prioritizes market transactions, with regulators merely setting the baselines for competition, a trend commonly but inaccurately called *deregulation*.[142]

The consequences of these shifts are grave. Where regulators have promulgated—and then rescinded—rules to pre-empt anticompetitive effects,[143] federal courts might not step in to fill the void as a consequence of presumed repugnancy. In bowing to regulators, courts can foster anticompetitive effects, which hampers innovation and cheats consumers. Since in *Trinko*, academics have offered a flurry of proposals to overhaul the balance between antitrust and

[140] *See* Verizon Communications Inc. v. Law Offices of Curtis V. Trinko, 540 U.S. 398 (2004).
[141] *See* Kearney & Merrill, *supra* note 24, at 1330–34.
[142] *Id.* at 1324–25, 1336–37.
[143] *Compare* Protecting and Promoting the Open Internet, FCC 15-24, GN Docket No. 14-28, Report and Order on Remand, Declaratory Ruling, and Order (Mar. 12, 2015) (promulgating net neutrality rules), *with* Restoring Internet Freedom, FCC 17-108, Declaratory Ruling, Report and Order, and Order (Dec. 14, 2017) (repealing net neutrality rules).

regulation.[144] In our era of regulatory abdication, scholars are looking to antitrust to step into the voids.[145] Whether those proposals materialize depends in large part on how courts strike that balance.

B. *Observations and Inferences*

In this Subsection, we present and categorize our observations from topic modeling and, where possible, draw preliminary inferences—recognizing that some inferences may be premature and require further research. Each of the three sets of visualizations we employ—multilevel, pyLDAvis, and topic browser—comes with its own advantages and drawbacks. Consequently, we approach modeling algorithms as an iterative process, adjusting where possible.

1. Macrotrends

a. Diversification of market power cases

Topic modeling is adept at highlighting macrotrends. To harness that power, we incorporated a histogram function into topic browser view that shows the relative proportion of each topic in the corpus as time progresses. In running histograms, we can immediately see how the Market Power and Antitrust–Regulation corpora have changed over the decades (see Figure 8 below).

---

[144] *See* Brett Frischmann & Spencer Weber Waller, *Revitalizing Essential Facilities*, 75 ANTITRUST L.J. 1 (2008); Howard A. Shelanski, *The Case for Rebalancing Antitrust and Regulation*, 109 MICH. L. REV. 683 (2011); Adam Candeub, *Trinko and Re-Grounding the Refusal to Deal Doctrine*, 66 U. PITT. L. REV. 821 (2005).

[145] *See, eg.*, Samuel N. Weinstein, *Financial Regulation in the (Receding) Shadow of Antitrust*, 91 TEMP. L. REV. 447 (2019); Tim Wu, *Antitrust via Rulemaking: Competition Catalysis*, 16 COLO. TECH. L. J. 33 (2017).

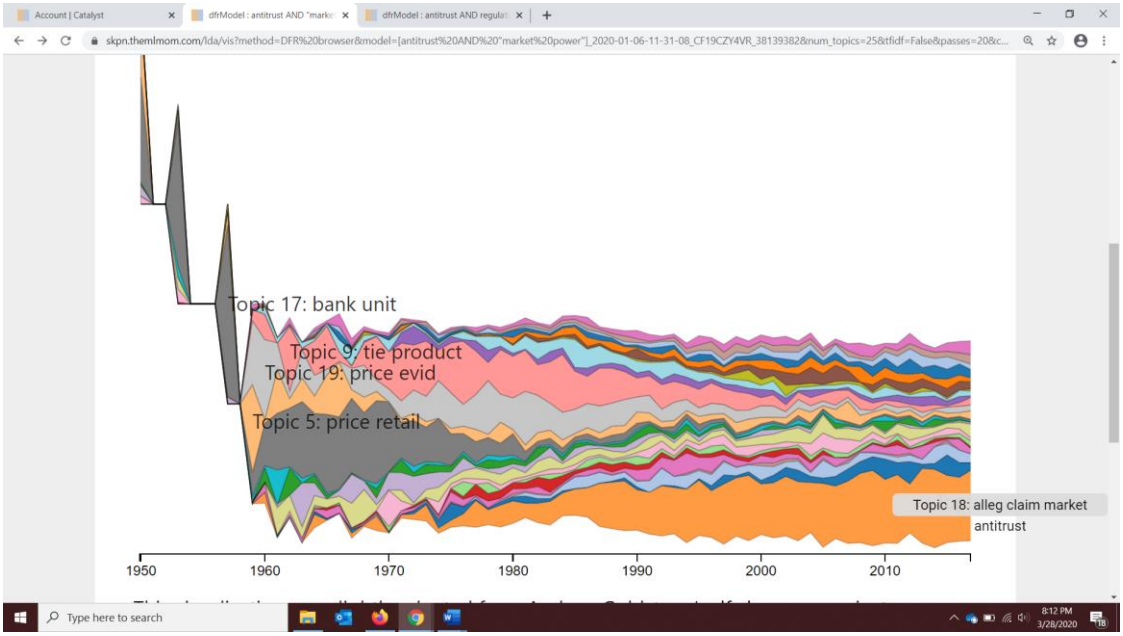

Figure 8: Topic Browser Stacked View with Histogram of Market Power Corpus

Starting in the late 1950s, market power cases exploded. Initially these cases were concentrated in the banking sector, where a slew of mergers were stayed by regulators and taken to court. Bank merger cases are unique enough to comprise a topic of their own, Topic 17, where several of the top terms are financial (e.g., "bank," "market," "compani[es]," "competit[ion]," "merger," "area," and "loan").[146] Starting in the 1970s, however, the number of cases in this topic declines rapidly, both in absolute numbers and as a proportion of the entire Market Power Corpus.[147]

[146] Here the top cases are U.S. v. Connecticut National Bank, 362 F.Supp. 240 (D. Conn. 1973), vacated by U.S. v. Connecticut National Bank, 418 U.S. 656 (1974); U.S. v. Phillipsburg Nat. Bank & Trust Co., 306 F.Supp. 645 (D.N.J. 1969), vacated by U.S. v. Phillipsburg Nat. Bank & Trust Co. 399 U.S. 350 (1970).

[147] The total number of market power cases can be verified on CAP's historical trends tracker. A search on CAP for federal cases with "antitrust" and "market power," for instance, shows that while antitrust cases have increased dramatically, market power caes have held steady. *See Historical Trends*, CASELAW ACCESS PROJECT, https://case.law/trends/ (search for "us: antitrust, market power").

The only other topic to undergo such a drastic decline is Topic 9, which includes mostly tying cases. Under the Sherman and Clayton acts' tying prohibitions, a seller cannot condition the availability of one item (the desired product) on the buyer's purchase of another item (the tied product). Tying cases are among antitrust's most complicated because courts and scholars have never agreed precisely on whether the practice merits per se treatment or rule of reason review. According to the leverage theory, if a monopolist dominates the desired product market, then the monopolist can leverage its way into dominance in the tied product market by conditioning the availability of the desired product on the purchase of the tied product.[148] Afterward, the monopolist can extract two sets of monopoly rents. The Chicago school, however, has succeeded in advancing its single monopoly profit theory, which holds that a true monopolist does not need to leverage its way into a tied product market because it can already extract rents in the desired product market. [149] Even though the Supreme Court continued to treat tying as per se illegal,[150] scholars have backed away from an unequivocal per se stance for decades. Recent work by economists and law scholars has vindicated parts of the leverage theory.[151]

In place of tying and bank merger cases, litigation-related topics have assumed greater prominence. These include

[148] *See* HOEVENKAMP, *supra* note 80, at 459.

[149] *See, e.g.*, Ward S. Bowman, Jr., *Tying Arrangements and the Leverage Problem*, 67 YALE L.J. 19 (1957); Richard S. Markovits, *Tie-ins, Reciprocity, and the Leverage Theory*, 76 YALE L.J. 1397 (1967); Richard A. Posner, *The Chicago School of Antitrust Analysis*, 127 U. PA. L. REV. 925 (1979).

[150] *See* Jefferson Parish Hosp. Dist. No. 2 v. Hyde, 466 U.S. 2 (1984); Eastman Kodak Co. v. Image Technical Servs., Inc., 504 U.S. 451 (1992).

[151] *See, e.g.*, *See* Patrick Rey & Jean Tirole, *A Primer on Foreclosure* 1, in HANDBOOK OF INDUSTRIAL ORGANIZATION III (Mark Armstrong & Rob Porter eds., 2006); Einer Elhauge, *Tying, Bundled Discounts, and the Death of the Single Monopoly Profit Theory*, 123 HARV. L. REV. 397 (2009). *See also* Thomas G. Krattenmaker & Steven C. Salop, *Anticompetitive Exclusion: Raising Rivals' Costs to Achieve Power over Price*, 96 YALE L.J. 209, 242 (1986).

a general litigation topic (Topic 18),[152] a general merger topic (Topic 15),[153] and a class actions topic that took off in 1998.[154] The trajectory is one of market power cases diversifying over time, spanning different types of claims and procedural strategies, such as class actions. As for the topics that declined in influence, the fall of bank merger cases is consistent with the increasing permissiveness of antitrust and financial regulators; rather than suing to block bank mergers, regulators were content to let the financial services industry consolidate after the 1970s.[155] This was especially pronounced as financial markets began to internationalize, which led to broader relevant geographic markets being defined more broadly and U.S. regulators easing up on consolidation to give domestic industries an advantage in cross-border competition. As for the waning of the tying topic, this coincided with the years the Supreme Court decided two seminal tying cases, *Jefferson Parish Hospital* in 1984 and *Eastman Kodak* in 1992.[156] However, ours is not a controlled study, and there may be confounding factors. Tying cases started to abate, for instance, when the Chicago school became ascendant.

---

[152] In Topic 18, the top terms are "alleg[ation]," "claim," "market", "antitrust," and "complaint." The top cases are *Wagner v. Circle W. Mastiffs*, 732 F. Supp. 2d 792 (S.D. Ohio 2010); *Bushnell Corp v. ITT*, 973 F. Supp. 1276 (D. Kan. 1997); and *Wolf Concept SARL v. Eber Bros Wine & Liquor Corp*, 736 F. Supp. 2d 661 (W.D.N.Y. 2010).

[153] In Topic 15, the top terms are "market," "merger," "custom[er]," "product," "compet[ition]," "FTC," "price," "injunc[tion]," and "relev[ant]." The top cases are *FTC v. Whole Foods Market, Inc.*, 502 F. Supp. 2d 1 (D.D.C. 2007), reversed by *F.T.C. v. Whole Foods Market, Inc*., 548 F.3d 1028 (D.C.Cir. 2008); *FTC v. CCC Holdings Inc.*, 605 F.Supp.2d 26 (D.D.C. 2009); and *FTC v Whole Foods Market, Inc.*, 548 F.3d.

[154] In Topic 1, the top terms include "class," "price," "member," "certif[ication]," "claim," and "common." The top cases are *In re Processed Egg Products Antitrust Litig.*, 312 F.R.D. 124 (E.D. Pa. 2015); *In re Titanium Dioxide Antitrust Litig.*, 959 284 F.R.D. 328 (D. Md. 2012).

[155] *See* Arthur E. Wilmarth, Jr., *The Transformation of the U.S. Financial Services Industry, 1975–2000: Competition, Consolidation, and Increased Risks*, 2002 U. ILL. L. REV. 215.

[156] Jefferson Parish Hosp. Dist. No. 2 v. Hyde, 466 U.S. 2 (1984); Eastman Kodak Co. v. Image Technical Servs., Inc., 504 U.S. 451 (1992).

b. Deregulation

The Antitrust–Regulation Corpus, too, exhibited diversification over time, with cases spanning various industries and regulatory schemes. In this vein, the decline of two topics is notable: a regulated industries topic (Topic 12)[157] and a banking and telecommunications topic (Topic 3).[158] Coinciding with their decline, general antitrust litigation topics rose sharply.[159]

These swings cohere with a broader pattern that scholars have previously noticed, where cases pertaining to the Interstate Commerce Commission ("ICC") were supplanted by telecommunications cases and other garden variety antitrust litigation. The ICC has its roots in the Interstate Commerce Act of 1887, which formulated the strict rate-setting rules of the *filed rate doctrine*, pursuant to which regulated entities were to file their rates with the commission.[160] The dwindling of ICC cases portends a shift away from public utility-style regulation and toward a framework where regulators simply set ground rules designed to maximize competition within an industry, such as

[157] In Topic 12, the top terms include "commiss[ion]," "rate," "carrier," "order," "file," "power," "regul[ation]," "rule," "tariff," "transport[ation]," "agenc[y]," "FERC," and "ICC." The top cases are *American Trucking Ass'n v. ICC*, 467 U.S. 354 (1984); *Brizendine v. Cotter & Co.*, 4 F.3d 457 (7th Cir. 1993), vacated by *Cotter & Co. v. Brizendine*, 511 U.S. 1103 (1994), in consideration of *Security Services, Inc. v. Kmart Corp.*, 511 U.S. 431 (1994); and *Security Services, Inc. v. Kmart Corp.*, 511 U.S. 431 (1994). Note that the topic cites *American Trucking* from September 1981, but no such case exists.

[158] In Topic 3, the top terms include "bank," "cabl[e]," "broadcast," "televise[ion]," "program," "station," "competit[ion," licens[e]," and "commiss[ion]." The top cases are *U.S. v. Marine Bancorporation, Inc.*, 418 U.S. 602 (1974); and *Satellite Broadcasting & Commun. Ass'n v. FCC*, 275 F.3d 337 (4th Cir. 2001). Interestingly, the early cases are bank merger cases, but the later cases are cable company cases. Both types of cases engage with similar vocabularies.

[159] This includes Topic 19, whose top terms are "claim," "alleg[ation," "complaint," "dismiss," "motion," "state," and "action." The top cases are *Caraang v. PNC Mortgage*, 795 F. Supp. 2d 1098 (D. Haw. 2011); and *Sonterra Caital Master Fund Ltd. v. Credit Suisse Group AG*, 277 F. Supp. 3d 521 (S.D.N.Y. 2017).

[160] Joseph D. Kearney & Thomas W. Merrill, *The Great Transformation of Regulated Industries Law*, 98 COLUM. L. REV. 1323, 1330-34 (1998).

the Telecommunications Act of 1996, a trend commonly (though not altogether accurately) called *deregulation.* [161]

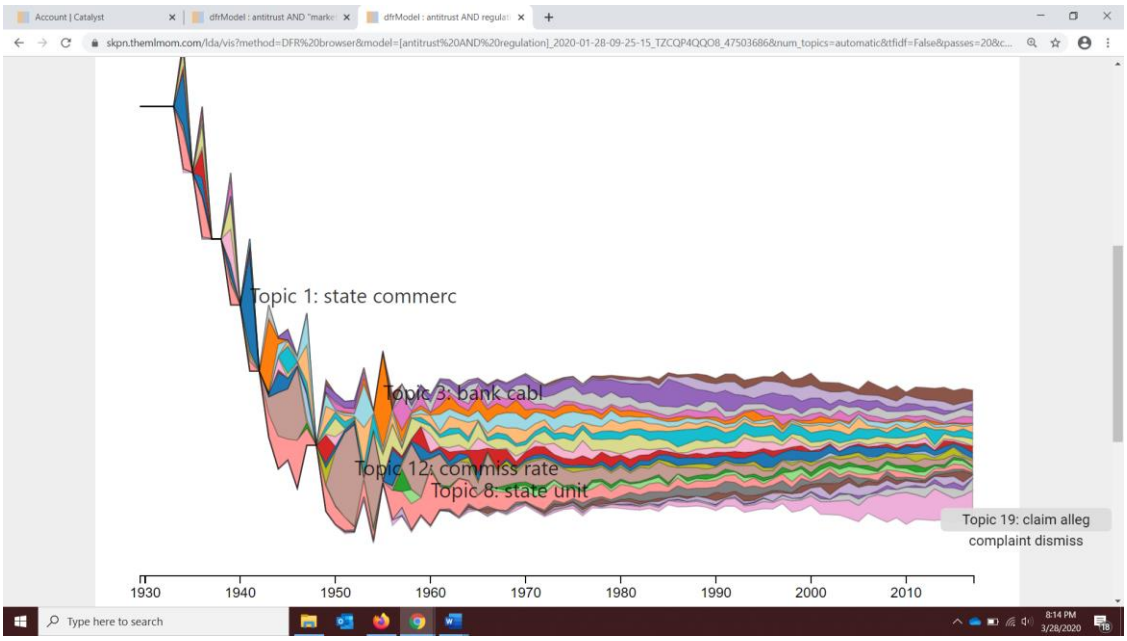

Figure 9: Topic Browser Stacked View with Histogram of Antitrust–Regulation Corpus

Topic browser histograms are a good starting point for historical trends. However, because topic browser view lists dozens of terms for each topic, the details can quickly overwhelm. As a supplement, then, we use the multilevel visualizations of aggregated modeling to eliminate the "noise" and scale up to a higher level of abstraction: topic clusters. This type of visualization can reveal the clusters that now make up the Antitrust–Regulation Corpus, giving a snapshot of how cases and topics have splintered (see Figure 9).

[161] *Id.*

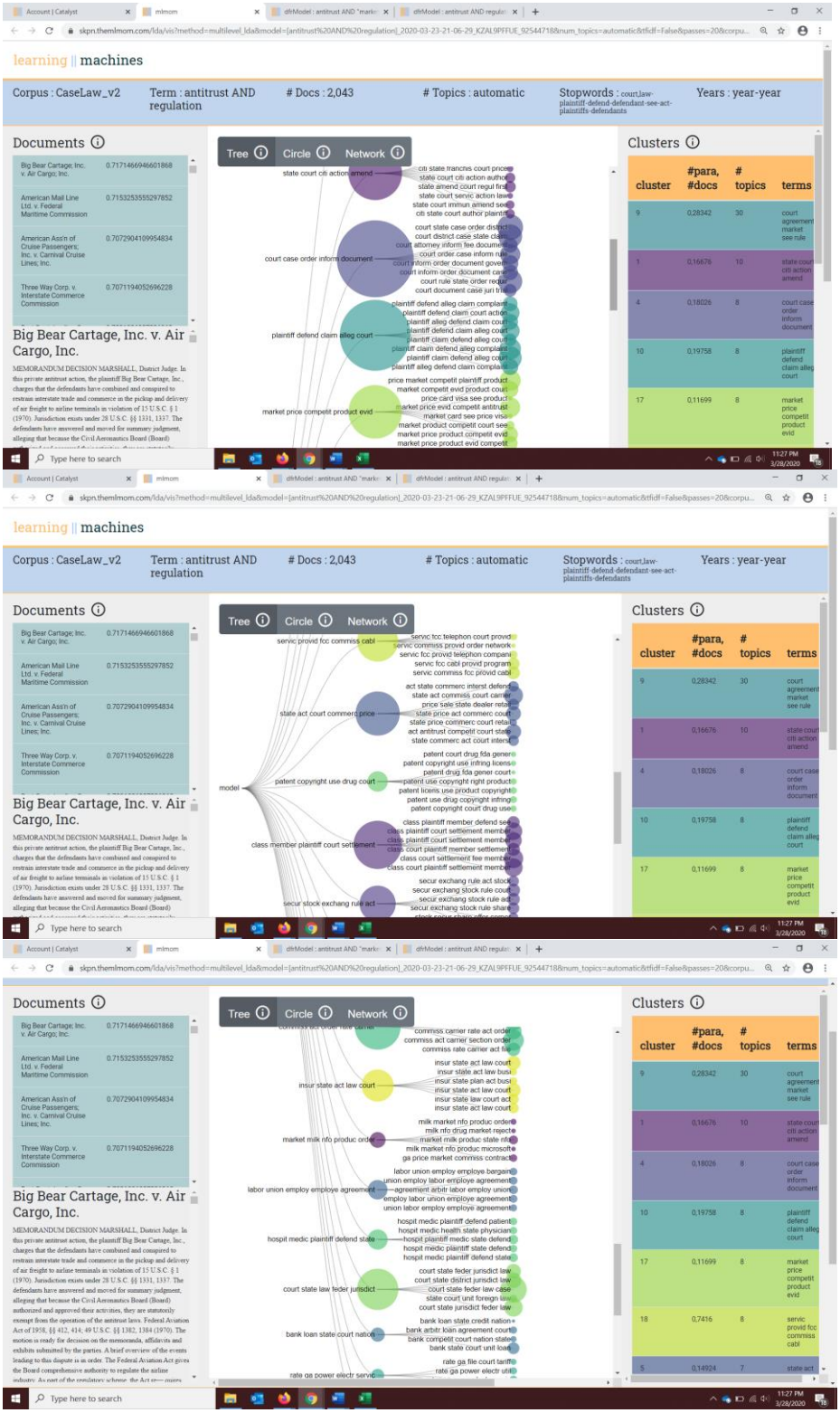
learning || machines
Corpus : CaseLaw_v2
Term : antitrust AND regulation
Docs : 2,043
Topics : automatic
Stopwords : court,law-plaintiff-defend-defendant-see-act-plaintiffs-defendants
Years : year-year
Documents
Tree
Circle
Network
Clusters
cluster
#para, #docs
topics
terms
Big Bear Cartage, Inc. v. Air Cargo, Inc.
American Mail Line Ltd. v. Federal Maritime Commission
American Ass'n of Cruise Passengers; Inc. v. Carnival Cruise Lines; Inc.
Three Way Corp. v. Interstate Commerce Commission
state court citi action amend
court case order inform document
plaintiff defend claim alleg court
market price competit product evid
model
patent copyright use drug court
class member plaintiff court settlement
secur stock exchang rule act
insur state act law court
market milk nfo produc order
labor union employ employe agreement
hospit medic plaintiff defend state
court state law feder jurisdict
bank loan state court nation
Type here to search

Figure 10: Multilevel Visualization of Antitrust–Regulation Corpus

In Figure 10, the Antitrust–Regulation Corpus is broken down into twenty-one clusters that correspond to the regulatory frameworks where antitrust litigation frequently arises. These include patent,[162] health care,[163] telecommunications,[164] securities and stock exchanges,[165] insurance,[166] labor,[167] power and electricity service,[168] banking,[169] and milk powder regulations.[170] Significantly, Cluster 9, the largest topic cluster, covers 30 topics that share the term "agreement." These topics cover multiple industries, including transportation, health care, technology, sports, credit cards, telecommunications, and airlines.

The prevalence of "agreement" in Cluster 9 suggests that a plaintiff's framing of the defendants' actions as a conspiracy, contract, or other agreement is the most common strategy. The per se illegality of conspiracies under antitrust obviates the need to gather additional evidence if a plaintiff can successfully couch the defendant's conduct as an agreement in violation of the Sherman or Clayton Act.[171] Indeed, collusive acts such as price-fixing and market division are often viewed as the core of antitrust prohibitions.[172] In the difficult instances where defendants mirror one another in conduct, factors that lead to the inference of agreement can move a case from one of

[162] Cluster 15.
[163] Cluster 14.
[164] Cluster 18.
[165] Cluster 3.
[166] Cluster 19.
[167] Cluster 6.
[168] Cluster 11.
[169] Cluster 7.
[170] Cluster 0.
[171] The antitrust cases standing for the proposition that agreement cannot be inferred from ambiguous evidence are also the classic summary judgment cases. *See, e.g.*, *Matsushita Electrical Industrial Co. v. Zenith Radio Corp*., 475 U.S. 574 (1986).
[172] *See* Jonathan B. Baker, *Exclusion as a Core Competition Concern*, 78 ANTITRUST L.J. 527, 545 (2013).

conscious parallelism to coordination.[173] Against this doctrinal backdrop, many of the cases in Cluster 9 feature agreements permitted by a regulatory frameworks but nonetheless charged by plaintiffs as anticompetitive.[174]

c. Industrial change

The histograms also tell an intriguing story about industrial change. In both corpora, there are declines in topics where "manufacturing" and "dealer" are among the top terms. This decline is particularly notable as a counterpoint to the finding of Steven Salop and Lawrence White over thirty years ago that manufacturing was overrepresented in private antitrust suits.[175] In their seminal article analyzing data from the Georgetown Private Antitrust Litigation Study (the "Georgetown Study"), Professors Salop and White found that 44.3 percent of defendants and 24.1 percent of plaintiffs hailed from the manufacturing sector.[176] These results correlate with the types of claims that predominated in the Georgetown dataset: refusals to deal, horizontal price fixing, tying or exclusive dealing, and price discrimination—claims reflecting disputes between retailers or wholesalers and their suppliers.[177]

---

[173] The antitrust literature on parallelism is rich. *See, e.g.*, Richard A. Posner, *Oligopoly and the Antitrust Laws: A Suggested Approach*, 21 STAN. L. REV. 1562 (1969); Donald F. Turner, *The Definition of Agreement Under the Sherman Act: Conscious Parallelism and Refusals To Deal*, 75 HARV. L. REV. 655 (1962); C. Scott Hemphill & Tim Wu, *Parallel Exclusion*, 122 YALE L.J. 1182 (2013). For an illustration, see *In re Text Messaging Antitrust Litigation*, 630 F.3d 622 (7th Cir. 2010); *In re Text Messaging Antitrust Litigation*, 782 F.3d 867 (7th Cir. 2015).

[174] *See, e.g.*, Kartell v. Blue Shield of Massachusetts, 542 F.Supp. 782 (D. Mass. 1982); Board of Com'rs of Port of New Orleans v. Federal maritime Commission, 440 F.2d 1312 (5th Cir. 1971); Metropolitan Intercollegiate Basketball Ass'n v. National Collegiate Athletics Ass'n, 337 F.Supp.2d 563 (S.D.N.Y. 2004).

[175] Steven C. Salop & Lawrence J. White, *Economic Analysis of Private Antitrust Litigation*, 74 GEO. L.J. 1001, 1004–05 (1986). For more on the Georgetown Project, see Lawrence J. White, *The Georgetown Study of Private Antitrust Litigation*, 543 ANTITRUST L.J. 59 (1985).

[176] *Id.*

[177] *Id.* at 1005.

Data from the Georgetown Study end in 1983, but from our corpora, we can infer that in the following decades, there is a decline in manufacturing and dealer cases but a rise in health care and patent cases, as a proportion of antitrust decisions overall.[178] To the extent these patterns reveal a shift in antitrust litigation, they may also betray a supplanting of manufacturing by health care, intellectual property, and other sectors. At a doctrinal level, we know, too, that tying and refusals to deal have been pared back by the courts.[179] And at a procedural level, we can perceive a marked rise in antitrust class actions. Altogether, these trends appear to confirm the waning of American manufacturing and, as a corollary, the demise of antitrust litigation between retailers and suppliers.

2. Inference challenges

Our visualizations do present challenges for drawing inferences. For a variety of reasons, some top cases in a topic wind up being aberrant upon review. The frequent examples are cases that do not engage substantively with market power.[180] We can partially pre-empt such results by screening for cases where query words (e.g., "market power") appear more than a desired number of times (e.g., 10 times). In this way, the visualizations will be compiled only out of those cases.

---

[178] The comparison of the Georgetown Study and our corpora is not an apples-to-apples comparison. The Georgetown Study was the joint effort of many attorneys reviewing and hand coding 2,350 antitrust cases from 1973 to 1983 in five federal districts. By contrast, our dataset is every federal antitrust decision up to late 2018—some 35,000 cases. Our dataset is both broader and narrower than the Georgetown dataset. While Salop and White covered settled cases, we can only look at cases that resulted in a judicial opinion. But our timelines and jurisdictions are broader, and we can also delve more deeply into the language of the cases.

[179] *See, e.g.*, Verizon Communications Inc. v. Law Offices of Curtis V. Trinko, 540 U.S. 398 (2004).

[180] *E.g.*, Wagner v. Circle W. Mastiffs, 732 F. Supp. 2d 792 (S.D. Ohio 2010) (the top case in the general litigation Topic 18, featuring virtually no discussion of market power because it was a price fixing case); Bushnell Corp v. ITT, 973 F. Supp. 1276 (D. Kan. 1997) (second top case in Topic 18, with no consideration of market power, where Sherman Act § 1 and § 2 claims were dismissed because the plaintiff presented insufficient evidence).

However, the interest for precise results must be balanced against the ability of machine learning to create visualizations that portray the corpora in new ways. A corpus can be restricted algorithmically, for instance, by excluding generic words (e.g., "court," "law," "plaintiff," and "defendant") or by collecting cases that mention key words more than a threshold number of times. Yet at some point, this strips away a key benefit of the topic modeling: to discern relationships among terms that we might otherwise gloss over.[181]

a. Aberrant results

Users of topic modeling must bear in mind that the algorithm constructs topics out of the terms that are most *statistically* likely to appear together. Thus, a case may be pushed to the forefront of a tying topic even though the opinion mentions tying only once—if the rest of the opinion contains all the other terms associated with the topic.[182] This is another common spurious result—one that, at this point, can only be identified by reading individual cases. Of course, the user of commercial databases must vet search results as well, so the requirement to actually read cases is not unique to topic modeling.

By way of comparison, in the Georgetown Study, Professors Salop and White quantified cases where antitrust was not the central issue but ancillary to a contract or tort claim ("noncentral cases") at 21.6 percent of the corpus, a fairly sizeable proportion.[183] Relatedly, 2.4 percent of the Georgetown corpus featured cases where an antitrust action was appended

[181] For now, we have also chosen to restrict our analysis to more general queries so as to quickly identify the algorithms' deficiencies.

[182] *See, e.g.*, Smith v. Mobil Oil Corp., 667 F. Supp. 1314 (W.D. Mo. 1987) (top case in tying Topic 9, with no consideration of market power). *But see R & G Affiliates, Inc. v. Knoll International, Inc.*, 589 F. Supp. 1395 (S.D.N.Y. 1984) (the second top case in Topic 9, which engaged in a substantive analysis of market power).

[183] Salop & White, *supra* note 171, at 1048–49.

as a counterclaim.[184] The specter of treble damages under private antitrust litigation would give any counterparty pause. In some areas of law, such as the Bank Holding Company Act's anti-tying provisions,[185]antitrust counterclaims are almost induced by their quasi-per se treatment.[186] It is therefore little surprise that noncentral decisions lurk in our corpora as well.

b. Machine versus human associations

In harnessing machine learning as a means of distant reading, we are essentially replacing *human* associations of words and meaning with *statistical* associations. This, too, can frustrate inferences. The terms in a topic might carry strong doctrinal associations. For instance, in the Antitrust–Regulation Corpus, "immunity" figures prominently in Topic 9 (from topic browser visualizations), suggesting on a cursory perusal that this Topic may bear upon the repugnancy doctrine and the balance between antitrust and regulation. In reviewing the cases and other terms, we discover that this is actually a Parker immunity topic concerning antitrust immunity for state action, as opposed to antitrust immunity through regulatory pre-emption.[187] Parker immunity, or the antitrust state action doctrine, covers certain state and local regulations that affect competition, exempting them from federal antitrust laws. It is a variation on federalism questions more typically encountered in constitutional law. By contrast, antitrust immunity in regulatory setting is usually implicates the question of whether

[184] *Id.* at 1048.
[185] 12 U.S.C. § 1972.
[186] For more on the bank tying provisions, see Felix B. Chang, *Death to Credit as Leverage: Using the Bank Anti-Tying Provision to Curb Financial Risk*, 9 NYU J. L. & BUS. 851 (2013). Tying counterclaims are often found in cases where a lender moves against a defaulting borrower.
[187] See Parker v. Brown, 317 U.S. 341 (1943). The other top terms in Topic 9 are "state," "citi[es]," "action," "power," "municip[al]," "district," "noerr" [after the Noerr-Pennington Doctrine], and "parker." The top cases are *Snake River Valley Electric Ass'n v. PacifiCorp*, 228 F.3d 972 (9th Cir. 2000), *Snake River Valley Electric Ass'n v. PacifiCorp*, superseded by (9th Cir. 2001); *Town of Hallie v. City of Eau Claire*, 700 F.2d 376 (7th Cir. 1983); and *Town of Hallie v. City of Eau Claire*, 471 U.S. 34 (1985).

regulation displaces antitrust—and the extent to which an antitrust savings clause resuscitates private antitrust litigation from regulatory pre-emption.

We can confirm that *Trinko*[188] and the old cases on repugnancy such as *Silver*[189] and *Gordon*[190] do appear in Topic 9—they are just not among the top results.[191] In fact, *Trinko* has a closer association with other topics (i.e., pertaining to telecommunications, federal legislation, and antitrust procedure) than with Parker immunity.[192] Here again, the result is not altogether surprising: *Trinko* comes up under commercial database searches for federal antitrust cases dealing with the Telecommunications Act of 1990, the essential facilities doctrine, and antitrust immunity.[193] Put differently, a case can constitute precedent in a number of areas.

Altogether, these instances of imprecision in topic modeling—at least what the human eye perceives as intuitively imprecise—complicate the ability to efficiently test research questions. As a more tangible example, we might infer something about how frequently courts engage in market definition from the fact that the term "relev[ant]" does not appear across even half of the topics in the Market Power Corpus (see Figure 11).

[188] Verizon Communications Inc. v. Law Offices of Curtis V. Trinko, 540 U.S. 398 (2004).
[189] Silver v. N.Y. Stock Exch., 373 U.S. 341, 357 (1963).
[190] Gordon v. New York Stock Exchange, Inc., 422 U.S. 659 (1975).
[191] The platform has a "bibliography" feature that lists all cases.
[192] *Trinko* has a 48.3% association with Topic 11 (top words "service[e]," "fcc," and "commiss[ion"), a 13.8% association with Topic 10 (top words "market," "claim," "competit[ion]," "antitrust," and "evid[ence]"), and a 11.7% association with Topic 8 (top words "state," "unit," "congress," and "statut[e]"). It has only a 5.8% association with Topic 9.
[193] Interestingly, *Trinko* is not among the top 20 results in Westlaw under a search for "antitrust /p regulation /p immunity." Notably, *Billing v. Credit Suisse First Boston Ltd.*, 426 F.3d 130 (2d Cir. 2005), the lower court decision of a Supreme Court opinion closely associated with *Trinko*, does appear as the seventh result.

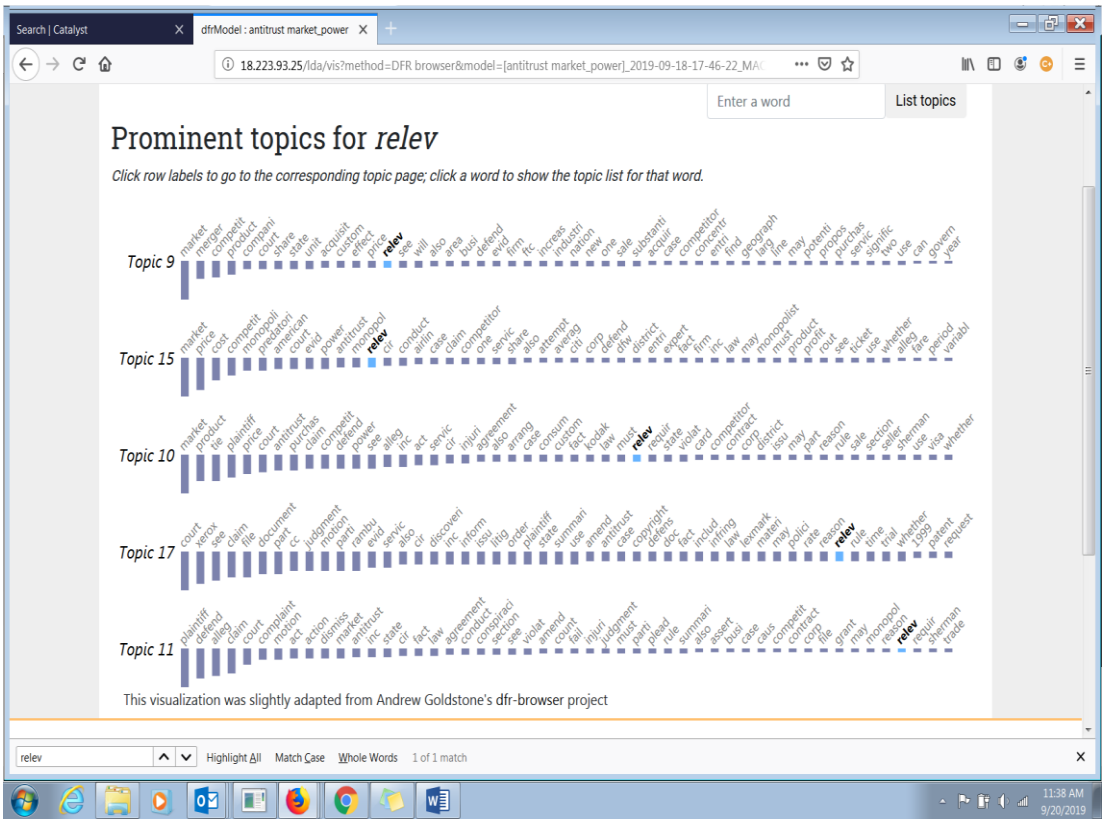

Figure 11: Topic Browser View of Topics Containing "Relev[ant]"

We might reasonably attribute this to two possibilities: either a court has accepted one party's market definition, or a court directly finds market power because there is evidence of anticompetitive effects. Yet "effect" also does not appear across many topics (see Figure 11), which is hardly surprising, since anticompetitive effects are difficult enough for economists to measure and even harder for courts to articulate. Significantly, the terms "relev[ant]" and "effect" do not overlap in topics, so we might also postulate that courts are using them as alternative proxies for market power.

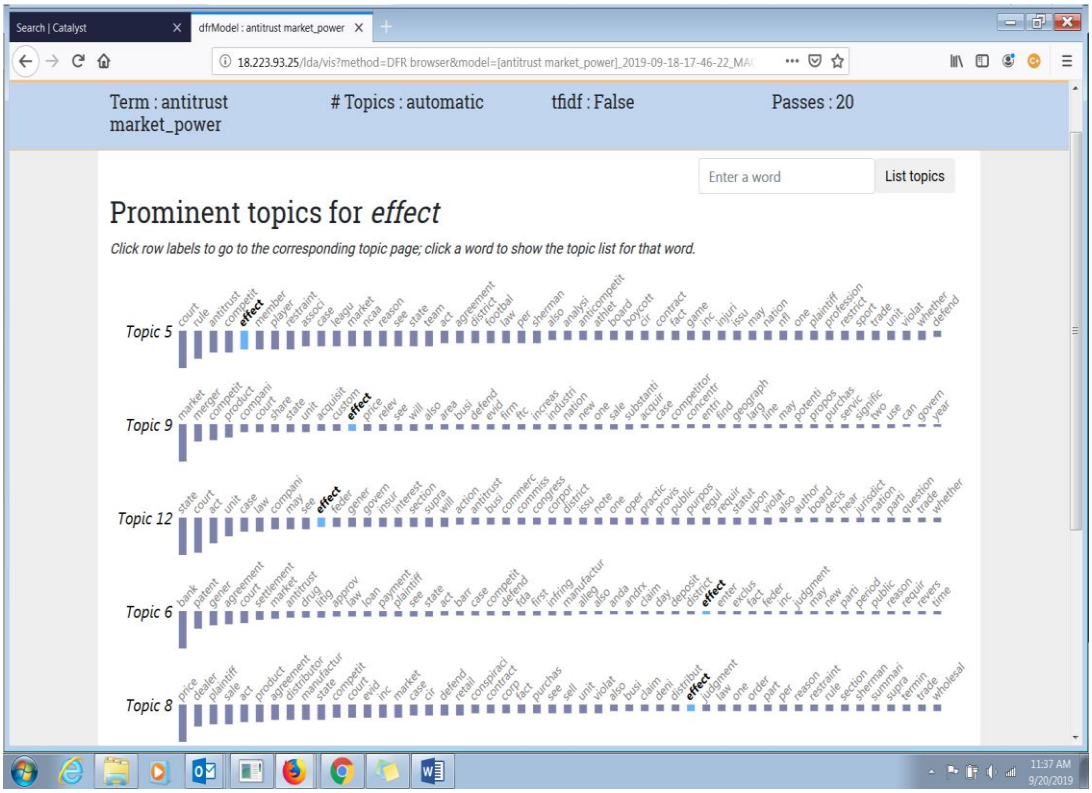

Figure 12: Topic Browser View of Topics Containing "Effect"

As we read the cases in the topics, however, we see that these inferences must be cautiously drawn. For instance, even within topics where "relev" is not highlighted as among the words (each topic lists approximately 50 top words), we find that courts often do take up the relevant product market, even if in cursory form. There simply may have been 50 other words that show much more frequently in the topic than "relev."[194] Again, however, we should not resort to filtering out too many terms that we consider generic, lest we sacrifice the fresh perspective of machine learning.

[194] *See* Chuang et al., *supra* note 42 ("In-depth analyses may require more than inspection of individual words. Analysts may want additional context in order to verify observed patterns and trust that their interpretation is accurate.").

V. SUPPLEMENTING TRADITIONAL RESEARCH

Nearly a decade ago, scholars in computer science, the field where topic modeling was invented, noted that model-driven visual analytics can suffer from problems of interpretation and trust. They defined *interpretation* as "the facility with which an analyst makes inferences about the underlying data" and *trust* as "the actual and perceived accuracy of an analyst's inferences."[195] Today, topic modeling has entered legal scholarship, and we hold out aggregated modeling as an improvement. Clearly, though, problems with interpretation linger—not to mention trust.

This Section addresses problems of interpretation and trust with topic modeling, extending the analysis to legal research more generally. In doing so, it suggests how the technique can both complicate and supplement traditional research.

A. *A Modest Proposal*

As noted above, there are impediments to drawing neat, clean inferences from our models. We acknowledge that, left unaddressed, these impediments can snowball into problems of trust. Hence, we have pursued modifications that shore up topic modeling's interpretative facilities at a basic level, which bolsters our scholarly community's receptivity toward—or trust of—the tool. Embedding a document reader feature in both multilevel and topic browser views enables our platform's users to pull up every case in a cluster or topic. In turn, cases can be read more thoroughly to check their conformity with their respective topics. This feature allows us to vet how topic modeling's information retrieval function scales to law.

In its early years, topic modeling was deployed to recommend scientific articles in a way that broke down

[195] *Id.* at 2.

disciplinary silos and cut through citation biases.[196] Some of the first computer scientists to experiment with collaborative topic modeling, for example, realized that researchers rely on citations to discover articles similar to one they have encountered, which reinforces the bias toward heavily cited papers.[197] Consequently, a scholar will tend to cite others within her discipline, at the expense of finding relevant literature in another field.[198] Topic modeling was devised as a powerful alternative, to catch the interdisciplinary linkages that might otherwise be overlooked. Staying true to this legacy, we argue that the best use of topic modeling—for now—might well be its ability to suggest areas of overlooked scholarship or doctrine.[199]

As a more concrete example, when we see that an immunity topic contains a high number of state action cases along with classic decisions on antitrust repugnancy,[200] we would read this as a suggestion for scholars interested in antitrust savings clauses to look into Parker immunity. A narrow search for savings clauses, focusing on landmark cases such as *Trinko* and *Credit Suisse*,[201] might otherwise miss this connection, directing the researcher simply to the antitrust–regulation balance. A few scholars writing on the antitrust immunity have already observed the connections between regulatory pre-emption and state action,[202] as has at least one

[196] *See* Wang & Blei, *supra* note 43. *See also supra* notes 49 and 50 and accompanying discussion.
[197] Wang & Blei, *supra* note 43.
[198] *Id.*
[199] To quote a critic of topic modeling, its utility may well be was a "content-based recommendation [system] (such as Facebook advertising products to its users)." Da, *supra* note 12, at 625.
[200] E.g., Snake River Valley Elec. Ass'n v. PacifiCorp, 228 F.3d 972 (9th Cir. 2000); Town of Hallie v. City of Eau Claire, 700 F.2d 376 (7th Cir. 1983). These are the top two cases in Topic 9 in the Antitrust–Regulation Corpus.
[201] Credit Suisse Securities (USA) LLC v. Billing, 551 U.S. 264 (2007).
[202] *See* Darren Bush, *Mission Creep: Antitrust Exemptions and Immunities As Applied to Deregulated Industries*, 2006 UTAH L. REV. 761; Daniel F. Spulber & Christopher S. Yoo, *Mandating Access to Telecom and the Internet: The Hidden Side of* Trinko, 107 COLUM. L. REV. 1822 (2007). *See also* HOVENKAMP, *supra* note 82, at § 19.3c.

court.[203] Yet this is not an intuitive connection to make; for the body of writings on state action and regulatory immunity have existed without much reference to one another.

As we tinker further with topic modeling, we can make instant improvements to sharpen the platform's interpretive precision. One upgrade is extending the numerical filters to individual terms, rather than a combination of all terms. As of now, we can screen for pertinent results by running visualizations on decisions where "antitrust" and "regulation" occur over a threshold number in each document. However, that threshold only runs on the combination of search terms. Thus, in a query for documents where search terms appear 50 times or more, the algorithms return decisions where "regulation" may appear 49 times in a document but "antitrust" appears only once. Several of the top decisions in Topic 14 of the Antitrust–Regulation Corpus, for instance, features the term "antitrust" only in the context of quoting antitrust cases as precedent on injunctions.[204] What initially appears to be an antitrust and first amendment topic ends up, at least from the top documents, as a constitutional law topic with antitrust caselaw cited for procedural guidance. This is not altogether surprising, since many foundational civil procedure decisions spun out of antitrust litigation.[205] An easy improvement, however, is to extend the numerical filters to both "antitrust" and "regulation."

The results from Topic discussed above, where the top results are noncentral cases, also suggest that the proportion of aberrant results in our two corpora might be quite different. At the very least, aberrant results arise for different reasons. In the Antitrust–Regulation Corpus, the top decisions in some topics only feature the term "antitrust" in the context of discussions of

[203] *See* American Agriculture Movement v. Board of Trade, City of Chicago, 977 F.2d 1147 (7th Cir. 1992).

[204] *See, e.g.*, .g., Kimberlin v. Quinlan, 6 F.3d 789 (D.C. Cir. 1993); Real Truth About Obama, Inc. v. FEC, 575 F.3d 342 (4th Cir. 2009); Kiser v. Kamdar, 831 F.3d 784 (6th Cir. 2016); O Centro Espirita Beneficiente Uniao Do Vegetal v. Ashcroft, 389 F.3d 973 (10th Cir. 2004).

[205] *See, e.g.*, Bell Atlantic Corp. v. Twombly, 550 U.S. 544 (2007); Matsushita v. Zenith Radio Corp., 475 U.S. 574 (1986).

procedural precedent. In the Market Power Corpus, by contrast, antitrust issues arise in some decisions only as counterclaims or ancillary actions, where they are summarily dismissed. Noncentral or aberrant antitrust decisions emerge more regularly in the Antitrust–Regulation Corpus. Indeed, the sizes of the two corpora, with Antitrust–Regulation being roughly three times the size of Market Power, appears to corroborate this thesis.

Overall, it is premature to draw any firm conclusion about the relevance of results, just as it is too early to aggressively filter out stop words. At this point, because our aim is to deploy topic modeling for its ability to suggest unexplored connections to other areas, we should refrain from steering unsupervised machine learning with too heavy a human hand. Thus, we currently see the greatest value in topic modeling's ability to distant-read an unstructured dataset and reveal the latent connections.

For a tool as transformative as topic modeling, its usage as a sort of glorified document retrieval mechanism may seem to be a modest proposal. However, cross-doctrinal extrapolation is one of the most common ways that legal scholarship has advanced.[206] Law scholars are fond of arguing by analogy; topic modeling gives us a better framework for doing so by drawing attention to shared vocabularies. Fortifying the algorithm's interpretive precision is one of the most important tasks before it gains more widespread usage. Harnessing the algorithm's information retrieval prowess also tests the robustness of its results. If we can prove that aberrant results are minimal, or at least within the range of prior studies, then we will have also built a foundation for our community's trust.

---

[206] *See, e.g.*, Cheryl I. Harris, *Whiteness As Property*, 106 HARV. L. REV. 1709 (1993); Samuel Issacharoff & Richard H. Pildes, *Politics as Markets: Partisan Lockups of the Democratic Process*, 50 STAN. L. REV. 643 (1998); Jonathan R. Macey & James P. Holdcroft, Jr., *Failure Is An Option: An Ersatz-Antitrust Approach to Financial Regulation*, 120 YALE L.J. 1368 (2011); Darrell A.H. Miller, *Text, History, and Tradition: What the Seventh Amendment Can Teach Us About the Second*, 122 YALE L.J. 852 (2013).

B. *A Bolder Proposal*

Combing through topic modeling visualizations raises interesting questions about the way we read cases and understand precedent. In each of the datasets, case names hardly ever surface as top terms. For instance, *Lorain Journal*,[207] *Alcoa*,[208] *Grinnell*,[209] and *DuPont*,[210] all of them classic market power cases, do not appear as terms in the Market Power Corpus.[211] In a narrowly focused topic—say, on tying—landmark cases such as *Eastman Kodak*[212] and *Jefferson Parish*[213] do not materialize as terms either. (The notable exception is *Microsoft*,[214] which shows up more frequently, even being picked up as a term in multilevel view).[215] This suggests that courts may be relying less on cases and more on a range of terms and concepts to figure out market power.

Often, cases that appear as the top results are only infrequently cited by legal scholars. These cases are not understood to be precedent-setting, though they can be heavily cited in practitioners' manuals or by other courts within a federal district or circuit.[216] Another discrepancy from commercial databases is that topic modeling occasionally

[207] Lorain Journal Co. v. U.S., 342 U.S. 143 (1951).
[208] U.S. v. Alcoa, 148 F.2d 416 (2d Cir. 1945).
[209] U.S. v. Grinnell Corp., 384 U.S. 563 (1966).
[210] U.S. v. E. I. du Pont de Nemours & Co., 351 U.S. 377 (1956).
[211] Again, we know these cases are part of the corpus because they appear in the bibliography. *See supra* note 173.
[212] Eastman Kodak Co. v. Image Technical Services, Inc., 504 U.S. 451 (1992).
[213] Jefferson Parish Hosp. Dist. No. 2 v. Hyde, 466 U.S. 2 (1994).
[214] U.S. v. Microsoft Corp., 253 F.3d 34 (D.C. Cir. 2001).
[215] In the antitrust-regulation cases, *Trinko* does not appear as a top term. Yet we can confirm that this case is picked up in the topic modeling, because there is a "bibliography" function on the platform that lists all the cases. This may simply be because *Trinko* is still relatively recent and has not been cited by other cases incorporated into the modeling.
[216] *See supra* note 145, and search in Westlaw's "citing references" function for *Wagner v. Circle W. Mastiffs*, 732 F. Supp. 2d 792 (S.D. Ohio 2010); *Bushnell Corp v. ITT*, 973 F. Supp. 1276 (D. Kan. 1997); and *Wolf Concept SARL v. Eber Bros Wine & Liquor Corp*, 736 F. Supp. 2d 661 (W.D.N.Y. 2010). By contrast, *see supra* note 169 and search in "citing references" for *Town of Hallie v. City of Eau Claire*, 471 U.S. 34 (1985), a seminal Parker immunity case.

returns decisions that have been overturned or vacated.[217] These results might not be troubling in topics dealing with arcane doctrine (e.g., old ICC cases). For their part, commercial databases, too, can lead readers to overturned decisions. Nonetheless, the ability of Lexis and Westlaw to flag a decision's precedential value in its metadata is helpful and cannot yet be replicated by topic modeling.

More fundamentally, the disparity between the top results from topic modeling and top results from commercial databases calls for reconciliation, but this is virtually impossible because users know so little about the search algorithms that Westlaw and Lexis employ. This opacity is a stark problem. Surveying search results across six different platforms, Susan Mart has found astonishing little overlap in the top cases when a query is run.[218] As Professor Mart notes, these inexplicable results are frustrating because the platform operators reveal virtually nothing about their algorithms. On a different level, algorithms compound human biases, and society is urgently re-evaluating the use of artificial intelligence for predictive purposes.[219] The lack of "algorithmic accountability" on the part of commercial databases is a detriment to research and the legal profession.[220]

An accountability deficit plagues not just incumbent databases but insurgent ones as well. Newcomers Casetext, Fastcase, Ravel (now owned by Lexis), and to some extent Google are challenging Westlaw and Lexis for the legal research market.[221] They promise to harness innovations in information

[217] *See, e.g.,* Brizendine v. Cotter & Co., 4 F.3d 457 (7th Cir. 1993) (judgment vacated by *Cotter & Co. v. Brizendine*, 114 S.Ct. 2095 (1994)).

[218] Mart, *supra* note 34, at 390 (noting "hardly any overlap in the cases that appear in the top ten results returned by" Castext, Fastcase, Google Scholar, Lexis Advance, Ravel, and Westlaw).

[219] *See* Cade Metz & Adam Satariano, *The Algorithm That Grants Freedom, or Takes It Away*, N.Y. TIMES, Feb. 6, 2020.

[220] *See* Mart, *supra* note 34, at 389 ("Algorithmic accountability in legal databases will help assure researchers of the reliability of their search results and will allow researchers greater flexibility in mining the rich information in legal databases. If researchers know generally what a search algorithm is privileging in its results, they will be better researchers.").

[221] For an empirical comparison of legal research providers, see id.

technology to deliver "faster and smarter" legal research.[222] Questions remain nevertheless. Do the insurgents' marketing slogans also encompass "cheaper," especially for academic and nonprofit communities? And given freely available tools such as CAP and topic modeling, how relevant are for-profit providers?

In pairing CAP with topic modeling, we are not attempting to dethrone the incumbents. Rather, our goals here are modest—at this stage, as we continue to fine tune aggregated modeling, we simply seek to supplement traditional doctrinal research.

However, we would advance a bolder proposal as well: by being transparent with topic modeling's weaknesses and how we are trying to overcome them, we intend to force legal research providers to be more forthcoming with their algorithms. This market is seeing more competition than it has in a long time. Powered by data analytics, upstarts are entering the market flaunting ever bolder claims. As they encroach upon Lexis and Westlaw's market shares, and as the incumbents defend their positions, both sides will have to justify why users should opt for their services.

Entering this fray, we have shown that, armed with a free dataset and some open-source algorithms, lawyers can replicate some of the search functionalities hiding behind paywalls. Admittedly, cobbling these functions together requires technical skills and often financial backing; however, homemade machine learning will put increasing pressure on for-profit legal research providers. Because consumers have more options than ever before, the operators of those paywalls must make the case for their products, including how they are superior. And when divergent results arise, as they inevitably do,[223] we anticipate that users will press providers for an explanation.

[222] *What is Fastcase?*, FASTCASE, https://www.fastcase.com/about/ (last accessed Feb. 3, 2020).
[223] *See* Mart, *supra* note 34, at 390.

VI. CONCLUSION

Topic modeling algorithms can be modified to address the criticisms of its detractors by providing greater context at the micro- and macroscopic levels. We have found that aggregating topic modeling over many iterations helps to eliminate aberrant results while providing contextualization. Simultaneously, our adjustments also highlight details that can serve as metadata to streamline doctrinal research.

There is still much to be done with our platform and visualizations. Looking ahead, we plan to improve the platform's capability to eliminate more generic words. As this happens, the visualizations will be more informative, and the cases will be grouped more accurately. Of course, we must exclude terms with care, lest we comprise the function of uncovering patters that the machine's algorithms illuminate.

The source of our dataset, CAP, also raises novel issues. For instance, the availability of data promises to democratize legal research, but there are still technical and financial barriers to data extraction and analysis. As alternatives to large commercial databases emerge, a pitched battle will unfold to capture the legal research and analytics market.

We see our project as a step in the use of algorithmic topic modeling in legal research, especially as a complement to commercial databases. Ultimately, we hope that our project will prompt other collaborations between DH and law, while pressing information technology insurgents to keep legal research open and cost-effective. In the near term, however, we can utilize topic modeling for discrete, mundane tasks such as recommending cases to help scholars and practitioners argue by analogy. Given the advances of CAP and topic modeling, we are living in one of the most exciting eras for legal research.

VII. APPENDIX

This Appendix lists the top 12 topics in the Market Power Corpus (2,591 total decisions) and the Antitrust–Regulation Corpus (7,308 total decisions) from topic browser view. In addition, it provides the proportion of the corpus occupied by each topic, as well as the top terms and decisions in each topic. Recall that topics are statistical distributions over terms.

Given its size, Antitrust–Regulation Corpus was filtered down to decisions where the key terms ("antitrust" and "regulation") occur more than approximately 20 times in each decision, resulting in a total of 3,527 documents.

For both corpora, we excluded the stop words "court," "law," "plaintiff," "defend," "defendant," "see," "act," "plaintiffs," and "defendants" from the visualizations.

Market Power Corpus		
Topic Number/ Proportion of Corpus	*Top Terms*	*Top Decisions*
18 (23.4%)	alleg[ation], claim, market, antitrust, complaint, inc, cir, motion, dismiss, state, competit[ion], injuri[es], action, agreement, conduct, fact, must, relev[ant]	Wagner v. Circle M. Mastiffs (Aug. 2010); Bushnell Corp. v. ITT Corp. (July 1997); Wolf Concept S.A.R.L. v. Eber Brox. Wine & Liquor Corp. (Aug. 2010); Full Draw Productions v. Easton Sports, Inc. (Dec. 1997); JES Properties, Inc. v. USA Equestrian, Inc. (Mar. 2003)
9 (10.7%)	tie, product, market, inc, power, purchas[e], state, case, cir, contract, arrang[ement], claim, dealer, sale, corp[orate], district,	Smith v. Mobil Oil Corp. (July 1987); R & G Affiliates, Inc. v. Knoll Int'l, Inc. (June 1984); Mozart Co. v. Mercedes-Benz of North America, Inc. (Sept. 1984); Siegel v. Chicken Delight, Inc.

	evid[ence], fact, franchis[e]	(Apr. 1970); Anderson v. Home Style Stores, Inc. (Apr. 1973)
19 (7.7%)	price, evid[ence], conspiraci[es], juri[sdiction], case, market, state, agreement, alleg[ation], claim, damag[es], manufactur[er], antitrust, busi[ness], cir, compani[es], competit[ion], dealer	Rossi v. Standard Roofing, Inc. (Sept. 1998); Zenith Radio Corp. v. Matsushita Electric Industrial Co. (Mar. 1981); Rossi v. Standard Roofing, Inc. (Mar. 1997); Muenster Butane, Inc. v. Stewart Co. (July 1981); Sunkist Growers, Inc. v. Winckler & Smith Citrus Products Co. (Sept. 1960)
25 (4.2%)	commiss[ion], carrier, rate, regul[ation], requir[e], state, author, decis[ion], exempt, icc, railroad, agenc[y], agreement, case, competit[ion], congress	American Trucking Ass'n v. ICC (Sept. 1981); Water Transport Ass'n v. ICC (June 1987); Burlington Northern Railroad v. United Transportation Union Int'l (June 1988); Regular Common Carrier Conference v. U.S. (June 1987); Central & Southern Motor Freight Tariff Ass'n v. U.S. (Mar. 1985)
21 (4.1%)	state, district, case, unit, claim, jurisdict[ion], right, action, arbitr[ation], feder[al], issu[e], parti[es], also, antitrust, appeal, applic[ation], effect	Lockyer v. .Mirant Corp. (Feb. 2005); United Phosphorus, Ltd. v. Angus chemical Co. (Mar. 2003); Monsanto Co. v. McFarling (Aug. 2002); Rohm & Haas Co. v. Dawson Chemical Co. (Jan. 1983)
12 (4.0%)	hospit[al], state, medic[al/ine], gener[al/ic], agreement, antitrust, new, patient, physician, univers[al],	Daniel v. Am. Bd. of Emergency Medicine (Nov. 1997); Islami v. Covenant Medical Center, Inc. (Dec. 1992); Friedman v. Delaware County Memorial Hosp. (Oct. 1987); Ezpeleta v. Sisters of Mercy

	action, case, center, claim	Health Corp. (Aug. 1986); Reddy v. Good Samaritan Hosp. & Health Center (Sept. 2000)
5 (3.9%)	price, retail, competit[ion], sale, wholesal[e], cost, discount, discrimen[ate], market, purchas[e], case, product, competitor, evid[ence], patman, robinson	Hoover Color Corp. v. Bayer Corp. (Dec. 1999); Smith Wholesale Co. v. R.J. Reynolds Tobacco Co. (Feb. 2007); Hoover Color Corp. v. Bayer Corp. (July 1998); Boise Cascade Corp. v. FTC (Jan. 1988); Lewis v. Philip Morris Inc. (Jan. 2004)
17 (3.7%)	bank, unit, state, market, compani[es], merger, corpor[ation], effect, area, busi[ness], case, loan, may, nation	U.S. v. Connecticut Nat'l Bank (June 1973); U.S. v. Philipsburg Nat'l Bank & Trust Co. (Oct. 1969); U.S. v. First Nat'l Bank (June 1969); U.S. v. First Nat'l Bank of Maryland (Jan. 1970); U.S. v. Manufacturers Hanover Trust Co. (Mar. 1965)
1 (3.2%)	class, price, member, purchas[e], certify[ication], claim, common, milk, nfo, rule, action, antitrust, damag[es]	In re Processed Egg Products Antitrust Litig. (11/2015); In re Titanium Dioxide Antitrust Litig. (8/12); In re Processed Egg Products Antitrust Litig. (11/15); In re Cathode Ray Tube (CRT) Antitrust Litig. (7/15); In re Graphics Processing Units Antitrust Litig. (7/08)
24 (3.0%)	cabl[e], servic[e], commiss[ion], broadcast, fcc, oper[ate], program, commun[ication], local, market, provid[er], station,	Time Warner Entertainment Co. v. U.S. (May 2000); Turner Broadcasting System, Inc. v. FCC (June 1994); Turner Broadcasting System, Inc. v. FCC (Mar. 1997); Turner Broadcasting System, Inc.

	compani[es], interest, must, public, regul[ate], state, televis[ion]	v. FCC (Apr. 1993); Cincinnati Bell Tellephone Co. v. FCC (Nov. 1995)
4 (3.0%)	market, card, rule, visa, agreement, compet[ition], restraint, effect, fee, per, reason, analysi[s], associ[ation], bank, case, member, merchant	U.S. v. Visa U.S.A. Inc. (10/01); U.S. v. Visa U.S.A. Inc. (9/03); In re ATM Fee Anttrust Litig. (3/08); Affinion Benefits Group, LLC v. Econ-O-Check Corp (3/11); U.S. v. American Express Co. (Sept. 2016)
16 (3.0%)	patent, claim, infring[e], licens[e], use, inc, applic[ation], cir, invent, issu[e], said, art, devic[e], evid[ence], fed[eral], justment, mean, motion, prior, product, royalti[es]	Semiconductor Energy Lab. Co. v. Chi Mei Optoelectronics Corp. (June 2007); Engel Indus., Inc. v. Lockformer Co. (Sept. 1996); Engineered Products Co. v. Donaldson Co. (Apr. 2004); VAE Nortrak North America, Inc. V. Progress Rail Services Corp. (Oct. 2006); Nystrom v. Trex Co. (June 2004)
Antitrust–Regulation Corpus		
Topic Number/ Proportion of Corpus	*Top Terms*	*Top Decisions*
9 (8.9%)	state, antitrust, citi[es], action, immun[e/ity], author[ity], competit[ion], power, activ[ity], conduct, law, Sherman, alleg[ation], case, municip[ality], privat[e]	Snake River Valley Elec. Ass'n v. PacifiCorp. (Oct. 2000); Town of Hallie v. City of Eau Claire (Feb. 1983); Town of Hallie v. City of Eau Claire (Mar. 1985); Bright v. Ogden City (Dec. 1985); Independent Taxicab Drivers' Employees v. Greater Houston Transportation Co. (May 1985)

12 (7.6%)	commiss[ion], rate, carrier, order, file, power, regul[ate/ation/ator], agreement, author[ity], case, contract, reason, requir[e], rule, section, tariff, transport	American Trucking Ass'n v. ICC (Sept. 1981); Brizendine v. Cotter (Aug. 1993); Security Services, Inc. v. Kmart Corp. (May 1994); Southern Motor Carriers Rate Conference v. U.S. (Oct. 1985); American Short Line Railroad v. U.S. (Dec. 1984)
19 (6.8%)	claim, alleg[ation], complaint, dismiss, motion, state, action, cir, inc, violat[e/ion], antitrust, requir[e], rule, also, amend, argu[ment/e], conduct, contract, count, fact, fail, fraud, injuri[es], must, plead	Caraang v. PNC Mortgage (June 2011); Sonterra Capital Master Fund Ltd. v. Credit Suisse Group AG (Sept. 2017); Mincey v. World Savings Bank (Aug. 2008); Young v. Wells Fargo & Co. (Oct. 2009); In re Packaged Seafood Products Antitrust Litig. (Mar. 2017)
24 (6.2%)	state, feder[al], claim, action, jurisdict[ion], case, right, cir, appeal, dismiss, issu[e], properti[es], provid[e], amend, author, court, decis[ion], determin[e], govern[ment], grant, judgment, motion, order	Hillis Motors, Inc. v. Hawaii Automobile Dealers' Ass'n (June 1993); Haydo v. Amerikohl Mining, Inc. (Oct. 1987); Florida Agency for Health Care Admin. v. Bayou Shores SNF, LLC (July 2016); Florida Agency for Health Care Admin. v. Bayou Shores SNF, LLC (In re Bayou Shores SNF, LLC) (July 2016); McGuire v. U.S. (Feb. 2013)
23 (5.9%)	price, evid[ence], product, market, sale, competit[ion], conspiraci[es], case, dealer, distributor,	Package Shop, Inc. v. Anheuser-Busch, Inc. (Oct. 1987); Pearl Brewing Co. v. Anheuser-Busch, Inc. (Feb. 1972); J.F. Feeser, Inc. v. Serv-A-Portion, Inc. (Aug. 1990);

	retail, agreement, alleg[e/ation], busi[ness], fact, inc, manufactur[e/er], may, purchas[e]	Zenith Radio Corp. v. Matsushita Elec. Indus. Co. (Mar. 1981); Beermart, Inc. v. Stroh Brewery Co. (Apr. 1986)
8 (5.8%)	state, unit, congress, statut[e], case, section, legisl[ation], govern[ment], regul[ation], author[ity], feder[al], foreign, gener	K Mart Corp. v. Cartier, Inc. (May 1988); Hart v. U.S. (Oct. 1978); Coalition to Preserve the Integrity of American Trademarks v. U.S. (May 1986); U.S. v. Mersey (Feb. 1960); Vivitar Corp. v. U.S. (Aug. 1984)
20 (5.6%)	agenc[y/ies], order, rule, inform, regul[ation/ator/ate], requir[e], govern[ment], issu[e], proceed, review, administr[ative/ator], applic[ation], case, decis[ion]	Deering Milliken, Inc. v. FTC (July 1978); Shell Oil Co. v. DOE (Aug. 1979); In re FTC Corporate Patterns Report Litig. (Apr. 1977); Nat'l Union Fire Ins. Co. of Pittsburgh v. Midland Bancor, Inc. (Dec. 1994); FTC v. Atlantic Richfield Co. July 1977)
1 (4.6%)	state, commerc[e/ial], regul[ation/ator/ate], interest, claus[e], feder[al], statut[e], wast, author[ity], citi[es], congress, local	Environmental Tech. Council v. Sierra Club (Oct. 1996); Tocher v. City of Santa Ana (July 2000); Ben Oehrleins & Sons & Daughter, Inc. v. Hennepin County (June 1997); CSX Transportation, Inc. v. City of Plymouth (Apr. 2000); Stucky v. City of San Antonio (July 2001)
4 (4.6%)	patent, antitrust, damag[es], juri[es/sdiction], claim, gener, trial, cir, district, agreement, inc, judgment, use, action, corp,	Valley Drug Co. v. Geneva Pharmaceuticals, Inc. (Sept. 2003); In re Terazosin Hydrochloride Antitrust Litig. (Aug. 2004); In re Yarn Process Patent Validity & Anti-Trust Litig. (Apr. 1974); Kearney & Trecker Corp. v. Cincinnati Milacron, Inc. (Oct.

	evid[ence], infring[e/ement]	1975); In re Terazosin Hydrochloride Antitrust Litig. (Jan. 2005)
10 (4.5%)	market, claim, competit[ion], antitrust, evid[ence], product, relev[ant], servic[e], inc, monopol[y], power	AD/SAT, Div. of Skylight, Inc. v. Associated Press (June 1999); Hendricks Music co. v. Steinway, Inc. (June 1988); AD/SAT v. Associated Pres (Feb. 1996); Allen-Myland, Inc. v. IBM Corp. (July 1988); Creative Copier Services v. Xerox Corp. (Feb. 2000)
14 (4.3%)	amend, state, first, regul[ation/ator/ate], speech, case, cir, claim, govern[ment], interest, protect, right, also, constitut[e/ion]	Kimberlin v. Quinlan (Oct. 1993); Real Truth About Obama, Inc. v. FEC (Aug. 2009); Kiser v. Kamdar (Aug. 2016); O Centro Espirita Beneficiente Uniao Do Vegetal v. Ashcroft (Nov. 2004)
25 (3.8%)	secur[ity/ities], exchang[e], stock, bank, compani[es], corpor[ate/ation], offer, issu[e], loan, rule, sec, share, action, busi[ness], case, interest, invest	Koppers Co. v. American Express Co. (Apr. 1988); Revlon, Inc. v. Pantry Pride, Inc. (Sept. 1985); Stonehill v. Security Nat'l Bank (June 1975); SEC v. Falstaff Brewing Corp. (May 1980); U.S. v. Morgan (Oct. 1953)