MODELING THE CASELAW ACCESS PROJECT: LESSONS FOR MARKET POWER AND THE ANTITRUST–REGULATION BALANCE

Felix B. Chang,[*] Erin McCabe[**] & James Lee[***]

This Article applies computational legal methods to antitrust caselaw. We introduce modifications to topic modeling that streamline big data legal analysis. Topic modeling, an algorithm that maps the statistical relationships among words, is quickly gaining traction in law. Yet its adherents have overlooked criticisms of the tool from computer science and digital humanities, especially around decontextualization.

We utilize aggregated modeling, an improvement over traditional topic models, to analyze two corpora of antitrust decisions extracted from the Caselaw Access Project. One corpus covers market power cases, while the other features the terms "antitrust" and "regulation."

Though technical advances have made the extraction of legal data much easier, the development of analytical tools—especially for researchers without statistical training—lags markedly. Our aggregated modeling provides greater contextualization of terms and topics. In publicizing our code and methodology, we hope that our results will prompt legal research providers to be more forthcoming with their search algorithms, which are notoriously opaque.

[*] Professor of Law and Co-Director, Corporate Law Center, University of Cincinnati College of Law. Affiliated Fellow, Thurman Arnold Project, Yale School of Management. E-mail: felix.chang@uc.edu. We thank the Andrew W. Mellon Foundation for financial support. Thanks, too, to Rosa Abrantes-Metz, Pinar Akman, Jonathan Baker, Josh Beckelhimer, Ned Cavanagh, David Donald, Harry First, Kelly Fitzpatrick, Eleanor Fox, Jens Frankenreiter, Peter Grajzl, James Grimmelmann, Scott Hemphill, Jyh-An Lee, Michael Livermore, Ephrat Livni, Fiona Scott Morton, Julian Nyarko, Menesh Patel, Zhaowei Ren, Thibault Schrepel, Danny Sokol, Gijs van Dijck, Sam Weinstein, Ramsi Woodcock, and Adam Ziegler. This Article benefitted from the Next Generation of Antitrust Scholars Conference at NYU, the antitrust series workshop at Yale's Thurman Arnold Project, the Online Workshop on the Computational Analysis of Law at Virginia, the Computational Antitrust 3.0 Conference at Stanford, and the Machine Lawyering Conference at the Chinese University of Hong Kong. We thank Madeleine Coles, Victor Kang, Wendy Antebi, Caitlin Pyatt, Courtney Sinagra, Samuel Holt, Robert Lemus, and the rest of the *Nevada Law Journal* for their careful editing.

[**] Library Fellow, Digital Scholarship Center, University of Cincinnati.

[***] Assistant Professor, Department of English, and Academic Director, Digital Scholarship Center, University of Cincinnati.

TABLE OF CONTENTS

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

[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, *Introduction: 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).

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 (NLP) upon legal texts.[4]

The newest tool to gain a following in CLA is topic modeling, a form of NLP that depicts the probability distribution of terms over a corpus of texts.[5] Because of its propensity to analyze large, unstructured datasets, topic modeling has been utilized in a variety of legal research projects.[6] While the tool is beginning to draw the level of scrutiny of previous empirical methods, topic modeling enthusiasts in law sometimes sidestep 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?[7]

A reflexive embrace of topic modeling and, more generally, CLA risks giving machine learning too quick a pass,[8] without vetting the underlying algorithms.[9] Word frequencies "without regard to position, syntax, content, and semantics" should not comprise the basis for bold claims.[10] Unmoored, CLA

[3] Corpus linguistics studies language through its usage within a body of texts. TONY MCENERY & 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 (2017); 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 xiii, 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. MACH. LEARNING RES. 993 (2003).

[6] *See* Livermore et al., *supra* note 5 (judicial opinions); Bernhard Ganglmair & Malcolm Wardlaw, Complexity, Standardization, and the Design of Loan Agreements (April 13, 2017) (working paper), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2952567 [https://perma.cc/7HXV-2SQG] (loan agreements); David S. Law, *Constitutional Archetypes*, 95 TEX. L. REV. 153 (2016) (national constitutions).

[7] 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 and the Law*, *in* LAW AS DATA, *supra* note 4, at 3, 3–19.

[8] 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, at 21, 23–24 ("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.").

[9] An algorithm is a set of instructions to perform a task, given a specific input. For a recent example of legal research that harnesses the power of algorithms, see David E. Pozen et al., *A Computational Analysis of Constitutional Polarization*, 105 CORNELL L. REV. 1 (2019) (measuring the partisan polarization of constitutional terms by gauging how easily a machine learning algorithm can predict a speaker's ideology or political affiliation).

[10] Nan Z. Da, *The Computational Case Against Computational Literary Studies*, 45 CRIT. INQUIRY 601, 611 (2019). For a series of spirited responses, see Mark Andrew Algee-Hewitt, et al., *Computational Literary Studies: A Critical Inquiry Online Forum*, CRIT. INQUIRY (Mar. 31, 2019, 5:26 PM), https://critinq.wordpress.com/2019/03/31/computational-literary-studies-

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."[11]

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.[12] DH is a collaborative endeavor at its core, so when legal scholars borrow its tools without considering assessments of 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 incorporates variations on topic modeling to sift through large datasets. In the most novel variation, we aggregate multiple layers of topic modeling into a single set of visualizations.[13] The platform has already been used on a variety of datasets, including archaeological publications,[14]

a-critical-inquiry-online-forum/ [https://perma.cc/JKM8-ZXH4]; Taylor Arnold & Duncan Buell, *More Responses to "The Computational Case Against Computational Literary Studies,"* CRIT. INQUIRY (Apr. 12, 2019, 11:33 AM), https://critinq.wordpress.com/2019/04/12/more-responses-to-the-computational-case-against-computational-literary-studies/ [https://perma.cc/DL6Z-RTPY].

[11] RICHARD RORTY, *Nineteenth-Century Idealism and Twentieth-Century Textualism*, *in* CONSEQUENCES OF PRAGMATISM 139, 151 (1982).

[12] *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. 55 (2010).

[13] For a prior example of aggregating topic models, see Peter Grajzl & Peter Murrell, *A Machine-Learning History of English Caselaw and Legal Ideas Prior to the Industrial Revolution I: Generating and Interpreting the Estimates*, 17 J. INST. ECON. 1 (2021) [hereinafter Grajzl & Murrell, *Generating and Interpreting the Estimates*]; Peter Grajzl & Peter Murrell, *A Machine-Learning History of English Caselaw and Legal Ideas Prior to the Industrial Revolution II: Applications*, 17 J. INST. ECON. 201 (2021) [hereinafter Grajzl & Murrell, *Applications*].

[14] Sarah E. Jackson et al., *Archaeological Reflexivity: Examining the Substance of Bone*, UNIV. OF CIN. DIGIT. SCHOLARSHIP CTR. (Nov. 3, 2020, 8:02 PM), https://sites.libraries.uc.edu/dsc/research/projects/archaeological-reflexivity-examining-the-substance-of-bone [https://perma.cc/AX7B-897X] (to identify patterns in language regarding the substance of bone).

Renaissance-era English texts,[15] and U.S. census data.[16] In the project at hand, we have applied aggregated modeling to uncover latent patterns in antitrust decisions extracted from Harvard Law School's Caselaw Access Project (CAP). Unveiled in 2018, CAP has digitized virtually every published decision in virtually every U.S. jurisdiction.[17]

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 beauty, translate topic modeling into intuitive models that law scholars can interpret with little 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 most criticisms of algorithmic data extraction and processing can be distilled to one theme: the need to provide greater contextualization.[18] In response, we have not abandoned DH methods, but sought to improve them.

We have compiled two pools of federal antitrust cases from the CAP corpus: those featuring the term "market power" and those featuring "regulation." Our rationale is that topic modeling visualizations can reveal how courts think through market power and how they balance antitrust when regulation is implicated. 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 employ.

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,[19] machine-generated visualizations can provide a

[15] *Mapping the Global Renaissance*, UNIV. OF CIN. DIGIT. SCHOLARSHIP CTR. (May 27, 2020, 6:24 PM), https://sites.libraries.uc.edu/dsc/research/projects/mapping-the-global-renaissance [https://perma.cc/R2XK-D7VT] (to examine accounts of race during Shakespeare's time).

[16] Robert Gioielli, *CO2 Emissions Racial Study*, UNIV. OF CIN. DIGIT. SCHOLARSHIP CTR. (May 27, 2020, 5:48 PM), https://sites.libraries.uc.edu/dsc/research/projects/co2-emissions-racial-study [https://perma.cc/PGN5-4G2J] (to explore links between race and CO2 emissions in urban areas). For a full list of projects, see *Projects*, UNIV. OF CIN. DIGIT. SCHOLARSHIP CTR., https://sites.libraries.uc.edu/dsc/home [https://perma.cc/BJP9-KRBW].

[17] *See* Harvard L. Sch., *About*, CASELAW ACCESS PROJECT, https://case.law/about/ [https://perma.cc/PKL9-ZNG8] [hereinafter *About the Caselaw Access Project*].

[18] *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 10.

[19] *See* SAFIYA UMOJA NOBLE, ALGORITHMS OF OPPRESSION: HOW SEARCH ENGINES REINFORCE RACISM (2018); Nathan Newman, *How Big Data Enables Economic Harm to Consumers,*

fresh take on thousands of cases.[20] Concomitantly, legal doctrines around market power and the antitrust–regulation balance can help us vet the precision of the topic modeling algorithms.

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[21] to the multi-sided platform debate,[22] weighty ideas on inequality and big tech are driving the conversations in antitrust.[23] 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.[24] Indeed, sensing the promise of CLA, the Department of Justice and Stanford Law School recently launched the Computational Antitrust Project to promote efficiency and consistency in antitrust analyses and procedures.[25]

Aggregated modeling excels in presenting high-level summaries of caselaw.[26] For instance, market power cases splinter into several large categories:

Especially to Low-Income and Other Vulnerable Sectors of the Population, 18 J. INTERNET L. 11 (2014).

[20] *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 313, 314 ("[Topic modeling] algorithms have provided a window to the relations between documents at scale.").

[21] *See* Maurice E. Stucke & Ariel Ezrachi, *The Rise, Fall, and Rebirth of the U.S. Antitrust Movement*, HARV. BUS. REV. (Dec. 15, 2017), https://hbr.org/2017/12/the-rise-fall-and-rebirth-of-the-u-s-antitrust-movement [https://perma.cc/2VKG-AUB6]; Lina Khan & Sandeep Vaheesan, *Market Power and Inequality: The Antitrust Counterrevolution and Its Discontents*, 11 HARV. L. & POL'Y REV. 235 (2017).

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

[23] *See* Elizabeth Warren, *Here's How We Can Break Up Big Tech*, MEDIUM (Mar. 8, 2019), https://medium.com/@teamwarren/heres-how-we-can-break-up-big-tech [https://perma.cc/5EBM-LST7]; *Corporate Accountability and Democracy*, BERNIE SANDERS, https://berniesanders.com/issues/corporate-accountability-and-democracy/ [https://perma.cc/4LMS-53CH]; TIM WU, THE CURSE OF BIGNESS: ANTITRUST IN THE NEW GILDED AGE (2018).

[24] *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 (2007).

[25] *Justice Department Joins Computational Antitrust Project at Stanford Law School*, U.S. DEP'T OF JUST. (Jan. 19, 2021), https://www.justice.gov/opa/pr/justice-department-joins-computational-antitrust-project-stanford-law-school [https://perma.cc/28N5-ZSJV]; *Computational Antitrust*, STAN. L. SCH., https://law.stanford.edu/projects/computational-antitrust/ [https://perma.cc/2QMT-CBY5].

[26] *See* Grajzl & Murrell, *Generating and Interpreting the Estimates*, *supra* note 13; Grajzl & Murrell, *Applications*, *supra* note 13.

patent, tying, general trial, telecommunications, merger, sports, class action, energy, software, milk, and licensing cases, as well as cases delving deeply into civil and evidentiary procedure.[27] 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.[28] Doctrinally, these inferences are not necessarily novel, but they do confirm the conjectures of other antitrust scholars who had theorized from narrower samplings of caselaw.[29]

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 Co. v. Image Technical Services, Inc.*,[30] *Jefferson Parish Hospital District No. 2 v. Hyde*,[31] *United States v. Grinnell Corp.*,[32] *United States v. Microsoft Corp.*,[33] and *United States v. E.I. du Pont de Nemours & Co.*,[34] all of them classic cases on market power.[35] 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.[36] 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.[37]

These results question how Westlaw and Lexis execute their searches, a process that is notoriously opaque.[38] For example, how do the commercial databases differ from widely accepted statistical algorithms in defining what constitutes

[27] *See infra* Part VI.

[28] Joseph D. Kearney & Thomas W. Merrill, *The Great Transformation of Regulated Industries Law*, 98 COLUM. L. REV. 1323, 1330–35 (1998).

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

[30] Eastman Kodak Co. v. Image Tech. Servs., Inc., 504 U.S. 451 (1992).

[31] Jefferson Parish Hosp. Dist. No. 2 v. Hyde, 466 U.S. 2 (1984).

[32] United States v. Grinnell Corp., 384 U.S. 563 (1966).

[33] United States v. Microsoft Corp., 253 F.3d 34 (D.C. Cir. 2001).

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

[35] The other cases in the top eleven 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. Indep. Ink, Inc.*, 547 U.S. 28 (2006), *In re Aggrenox Antitrust Litig.*, 199 F. Supp. 3d 662 (D. Conn. 2016), *Sentry Data Sys., 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. v. Atl. Richfield Co.*, 51 F.3d 1421 (9th Cir. 1995). The search was performed on March 20, 2020.

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

[37] *See infra* Part IV.

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

relevant caselaw? In publicizing our algorithms, we hope to nudge the commercial databases toward greater transparency.

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.[39] 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.[40] 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. The next steps will be to test a series of hypotheses, against the topic models. 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 will unfold as follows: Part I will canvas critiques of CLA methods and tinker with improvements to topic modeling. Part II will introduce our platform and summarize our methodology. Part III will present preliminary findings and will draw inferences that both affirm and complicate previous antitrust research. Part IV will hazard predictions for the way forward.

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

Machine learning abounds in finance, policing, employment, politics, and health services,[41] but as a research technique, it is just gaining traction in legal academia.[42] Legal 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.[43] And in law, no data is bigger than legal texts. Applied to a corpus of caselaw, machines can uncover explicit and latent linguistic and semantic

[39] *See* Gabriella Capone, *Democratizing the Law with Open Data*, MIT COMPUTATIONAL L. REP. (Nov. 20, 2020), https://law.mit.edu/pub/democratizingthelawwithopendata/release/2 [https://perma.cc/VZ5R-GRS9].

[40] *What Is Fastcase?*, FASTCASE, https://www.fastcase.com/about/ [https://perma.cc/4X4R-3ZJM].

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

[42] One exception is the application of corpus linguistics to statutory interpretation to discern the ordinary meaning of language. *See, e.g.*, Gries & Slocum, *supra* note 3, at 1417. Recently, BigML also started to provide machine learning services to academics. *See* BIGML, https://bigml.com/ [https://perma.cc/2WYY-67L5].

[43] 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/ [https://perma.cc/A2QU-7WTS].

patterns, bringing out significant word clusters "that the eye cannot see."[44] 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, for instance, have come under scrutiny for their tendency to decontextualize settings.[45] 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."[46]

This Part 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.[47] Yet the technique has its deficiencies, as digital humanists and computer scientists have pointed out.[48] This Part therefore reintroduces topic modeling, especially in triangulation with corpus linguistics and other empirical methods, which are more familiar. It also summarizes the criticisms of the tool, so as to set the stage for our improvements.

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.[49] These patterns are presented as "topics," where each topic is comprised of the terms most likely to appear together.[50] Topic modeling builds upon the general concept of modeling, which creates representations of data patterns in a statistically or logically coherent form.[51] While models abound in legal research, topic modeling is performed through machine learning to amplify processing power.[52] The tool enables researchers to analyze tomes of data without

[44] Lauren F. Klein, Distant Reading After Moretti, Address at the MLA Annual Convention (Jan. 5, 2018) (transcript available at https://lklein.com/digital-humanities/distant-reading-after-moretti/ [https://perma.cc/73TY-FND8]).

[45] *See, e.g.*, Carissa Byrne Hessick, *Corpus Linguistics and the Criminal Law*, 2017 BYU L. REV. 1503 (2017); Zoldan, *supra* note 18.

[46] Fish, *supra* note 7, at 339–40.

[47] *See, e.g.*, Michael A. Livermore et al., *Computationally Assisted Regulatory Participation*, 93 NOTRE DAME L. REV. 977 (2018); *see also supra* notes 5–6.

[48] *See* Da, *supra* note 10.

[49] *See* Jason Chuang et al., Interpretation and Trust: Designing Model-Driven Visualizations for Text Analysis (May 7, 2012), *in* CHI '12: PROCEEDINGS OF THE SIGCHI CONFERENCE ON HUMAN FACTORS IN COMPUTING SYSTEMS 443, 443 (2012).

[50] Chong Wang & David M. Blei, Collaborative Topic Modeling for Recommending Scientific Articles (Aug. 21, 2011), *in* KDD '11: PROCEEDINGS OF THE 17TH ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING 448, 450 (2011).

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

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

having to manually code them first, as is custom in traditional empirical methods.[53]

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—for example, economics or biology.[54] 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 allocation, or multivariable probability distributions, of topics over a fixed vocabulary within a dataset.[55] 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.[56]

Two features of topic modeling—its ability to sift through large volumes of texts and to uncover hidden connections within those texts—make it tantalizing for legal research. While the tool remains new to law scholars,[57] a growing number of researchers are adopting it for projects on loan agreements,[58] constitutions around the world,[59] Supreme Court legal opinions,[60] control rights in union contracts,[61] and English legal concepts prior to the Industrial Revolution.[62]

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.[63] Its advocates in law, including, most prominently, Justice Thomas Lee on the Utah Supreme Court, argue that its methods can elucidate

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

[54] *See* Wang & Blei, *supra* note 50, at 448.

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

[56] *See Stanford Dissertation Browser: An Experiment in Document Exploration*, STAN. NAT. LANGUAGE PROCESSING GRP., https://nlp.stanford.edu/projects/dissertations/ [https://perma.cc/2KWK-PC8U]; Chuang et al., *supra* note 49, at 446; Wang & Blei, *supra* note 50, at 448.

[57] In 2016, David Law identified only two instances of topic modeling in legal research. *See* Law, *supra* note 6, at 164 n.31.

[58] Ganglmair & Wardlaw, *supra* note 6.

[59] Law, *supra* note 6.

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

[61] Elliott Ash et al., The Language of Contract: Promises and Power in Union Collective Bargaining Agreements (Mar. 30, 2019) (unpublished manuscript) (available at https://extranet.sioe.org/uploads/sioe2019/ash_macleod_naidu.pdf).

[62] *See* Grajzl & Murrell, *Generating and Interpreting the Estimates*, *supra* note 13.

[63] 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.

ordinary meaning of words and phrases in statutory interpretation.[64] 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.[65] 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.[66] 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.[67]

Understandably, corpus linguistics has provoked denunciation. Legal scholars have decried as "radically acontextual" the separation of statutory language "from its distinctly legal context."[68] Word frequency and collocation crowdsource for meaning by scanning random corpora of natural language, including sources of dubious judicial value such as Urban Dictionary.[69] 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.[70] In countering that judges may be more accountable for their decisions than machine learning technocrats,[71] critics echo a broader skepticism of artificial intelligence as an unelected, unaccountable decision-maker that is incapable of empathy.[72] For these scholars, corpus linguistics may offer an impartiality that is simply too elusive to stand in place of 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

[64] *See generally* Thomas R. Lee & James C. Phillips, *Data-Driven Originalism*, 167 U. PA. L. REV. 261 (2019); Lee & Mouritsen, *supra* note 3.

[65] *See NOW Corpus (News on the Web)*, ENGLISH-CORPORA.ORG, https://www.english-corpora.org/now/ [https://perma.cc/X2GD-X39P].

[66] *See* Mark Davies, *English-Corpora.org: A Guided Tour*, ENGLISH-CORPORA.ORG (Nov. 2020), https://www.english-corpora.org/pdf/english-corpora.pdf [https://perma.cc/4MZE-98VV].

[67] *See* BYU L.: L. & CORPUS LINGUISTICS, https://lawcorpus.byu.edu/ [https://perma.cc/6Z3U-UWGA]. 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 & James C. Phillips, *Corpus Linguistics and the Second Amendment*, HARV. L. REV. BLOG (Aug. 7, 2018), https://blog.harvardlawreview.org/corpus-linguistics-and-the-second-amendment/ [https://perma.cc/6VED-DL7N].

[68] Zoldan, *supra* note 18, at 447.

[69] *Id.* at 417.

[70] *See* Hessick, *supra* note 45, at 1512.

[71] *See id.* at 1517–18.

[72] *See* Kiel Brennan-Marquez & Stephen E. Henderson, *Artificial Intelligence and Role-Reversible Judgment*, 109 J. CRIM. L. & CRIMINOLOGY 137, 146 (2019).

empirical methods.[73] ELS has a rich history,[74] 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.[75] Starting in economics, this revolution forced empirical researchers in all fields to root out bias through better research design.[76] 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,[77] that it fetishizes technique at the expense of the bigger picture.[78]

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 data source, 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 justifies departing from legal texts.[79] 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.[80] 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.[81]

Nonetheless, topic modeling is still vulnerable to the same reproach of overreliance on disembodied words that plagues corpus linguistics.[82] 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

[73] 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 (2011) ("[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.").

[74] *See, e.g.*, *id.* at 1713–14; Ho & Kramer, *supra* note 2.

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

[76] Angrist & Pischke, *supra* note 2, at 14; Copus, *supra* note 8, at 21.

[77] Eisenberg, *supra* note 73, at 1733.

[78] Brian Leiter, *On So-Called "Empirical Legal Studies" and Its Problems*, BRIAN LEITER'S L. SCH. REP. (July 6, 2010), http://leiterlawschool.typepad.com/leiter/2010/07/on-socalled-empirical-legal-studies.html [https://perma.cc/JUA5-A7K2].

[79] For a summary of this hunt for ordinary meaning, see Lee & Mouritsen, *supra* note 3, at 796–800.

[80] *See* Benjamin M. Schmidt, *Words Alone: Dismantling Topic Models in the Humanities*, 2 J. DIGIT. HUMANS. 49, 51 (Winter 2012).

[81] *See* Copus, *supra* note 8, at 24–25, 29.

[82] *See* Da, *supra* note 10, at 635.

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 of Topic Modeling*

In introducing our community of legal scholars and antitrust practitioners to aggregated modeling, we have written previously on reservations about this tool within digital humanities and computer science.[83] The primary critique is that it rearranges terms into topics at the expense of context.[84] Decontextualization can lead to overclaims and misinterpretation.[85] Given its dexterity, topic modeling can lead users into a labyrinth of terms and topics, so that they lose perspective either through the grand sweep of the topics in totality (that is, too much macroscale) or through the number of terms within a topic (that is, too much microscale). To borrow from *Gulliver's Travels*, the eighteenth-century satire at the outset of the Enlightenment's scientific discoveries, topic modeling creates both a gargantuan world and a miniscule world, and the user may be adrift at both extremes.[86]

This propensity to decontextualize—or, more appropriately, recontextualize—is an inherent feature of topic modeling. In fact, recontextualization is one of the tool's major draws. After all, topic modeling is a form of distant reading that illuminates patterns in texts—especially large volumes of texts—that cannot be seen by the human eye in close readings.[87] While distant reading appears to belie the close textual analysis that underpins legal research, especially qualitative doctrinal research, the reality is that law scholars have implemented the quantitative methods of social sciences for decades, and CLA methods are merely the latest development. The challenge, then, is to temper both the errors and disorientation of topic modeling, while setting topics and terms in their proper context. More specifically, topic modeling visualizations must be able to

[83] *See, e.g.*, Felix B. Chang et al., *Doctrinal Implications of Computational Antitrust*, 1 STAN. J. COMPUTATIONAL ANTITRUST 117, 120–21 (2021) [hereinafter Chang et al., *Computational Antitrust*]; Felix B. Chang, *The Topic Modeling Revolution*, CPI ANTITRUST CHRON. (forthcoming 2021) (manuscript at 8–9) (on file with author).

[84] *See, e.g.*, Schmidt, *supra* note 80, at 50–51 (panning topic modeling as little more than a "rearrangement algorithm" and, without proper supervision, a "bad research assistant").

[85] Da, *supra* note 10, at 611; *see also* Nan Z. Da, *Computational Literary Studies: Participant Forum Responses, Day 2*, IN THE MOMENT (Apr. 2, 2019, 1:08 PM), https://critinq.wordpress.com/2019/04/02/computational-literary-studies-participant-forum-responses-day-2-2/ [https://perma.cc/5CWT-79QF] ("First, there is statistically rigorous [computational literary studies] work that cannot actually answer the question it sets out to answer or doesn't ask an interesting question at all. Second, there is work that seems to deliver interesting results but is either nonrobust or logically confused . . . . If the conclusions we find in CLS corroborate or disprove existing knowledge, this is not a sign that they are correct but that they are tautological at best, merely superficial at worst.").

[86] *See generally* JONATHAN SWIFT, GULLIVER'S TRAVELS (1726).

[87] Livermore & Rockmore, *supra* note 7, at 4; *see also* Klein, *supra* note 44.

break topics down to their constituent words and aggregate all topics into networks—to depict the relationships among topics just as they illustrate the relationships among terms.

Notably, contextual shifts can also come from words themselves. Over time, for instance, usage evolves; spellings change, registers shift, and terms assume ironic connotations.[88] 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.[89] Our visualizations capture several topics where the term "relevant" is featured prominently. As a 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.[90]

Digital humanists and computer scientists have pointed out other shortcomings of topic modeling. Some of them are relevant to legal scholarship and can be addressed; others may be relevant but cannot be programmed around.[91] In the first camp is the argument that LDA had been created to perform information retrieval functions rather than the analytical and predictive functions that have come to be associated with the algorithm.[92] For example, the computer scientist David Blei thought topic modeling would make "large collections of text browsable by giving useful tags to the documents," a function similar to Westlaw's insertion of headnotes.[93] Those functions, as we shall show later, include tagging caselaw with helpful metadata to enable more efficient browsing, rather than to make predictions about how a case might come out.

Finally, humanities scholars have rebuked the digitization of their field in ways that have some applicability to legal research.[94] These include institutional

[88] *See* Schmidt, *supra* note 80, at 57 ("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.").

[89] *See* HERBERT HOVENKAMP, FEDERAL ANTITRUST POLICY: THE LAW OF COMPETITION AND ITS PRACTICE § 6.4 (5th ed. 2016).

[90] *See infra* Part IV.

[91] *See* Danielle Allington et al., *Neoliberal Tools (and Archives): A Political History of Digital Humanities*, L.A. REV. OF BOOKS (May 1, 2016), https://lareviewofbooks.org/article/neoliberal-tools-archives-political-history-digital-humanities/ [https://perma.cc/CG33-D4JP] ("[T]he unparalleled level of material support that Digital Humanities has received suggests that its most significant contribution to academic politics may lie in its (perhaps unintentional) facilitation of the neoliberal takeover of the university.").

[92] Schmidt, *supra* note 80, at 50 ("New ways of reading the composition of topics are necessary, because humanists seem to want to do slightly different things with topic models than the computer scientists who invented them and know them best.").

[93] *Id.* Indeed, precursors of LDA such as latent semantic analysis had been designed for indexing and information retrieval. *See* Scott Deerwester et al., *Indexing by Latent Semantic Analysis*, 41 J. AM. SOC. INFO. SCI. 391, 391 (1990).

[94] *See* Allington et al., *supra* note 91.

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.[95] 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.[96] 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

This Section introduces the features of our aggregated modeling, which aggregates—or embeds—up to six different LDA topic models in one iteration. Aggregation of topics into clusters has been done by other scholars to achieve a bird's-eye perspective on major themes.[97] At this point, we believe it still should be paired with other tools for caselaw research to be comprehensive, and we leave the next Part 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 have tried to deploy them more for information retrieval than predictive functions.

Beyond achieving a bird's-eye perspective, aggregation also generates a multilevel "model of models" that assuages some contextualization, robustness, and reproducibility concerns. Our platform has the capacity to render clusters in three different types of visualizations. In network view, topic clusters situate topics in both large and small contexts through vector space modeling, which strews clusters across space. Figure 1 depicts how topics in the Antitrust–Regulation corpus cluster, while Figure 2 presents a close-up of Cluster 3, featuring "rate," "gas," "commission," "rates," and "ferc" [F.E.R.C.] as the top terms.

[95] *See id.*

[96] *See id.*; Klein, *supra* note 44.

[97] See Grajzl & Murrell, *Applications*, *supra* note 13, for an example of manual aggregation.

FIGURE 1: NETWORK VIEW OF ANTITRUST–REGULATION CASES IN MODEL OF MODELS

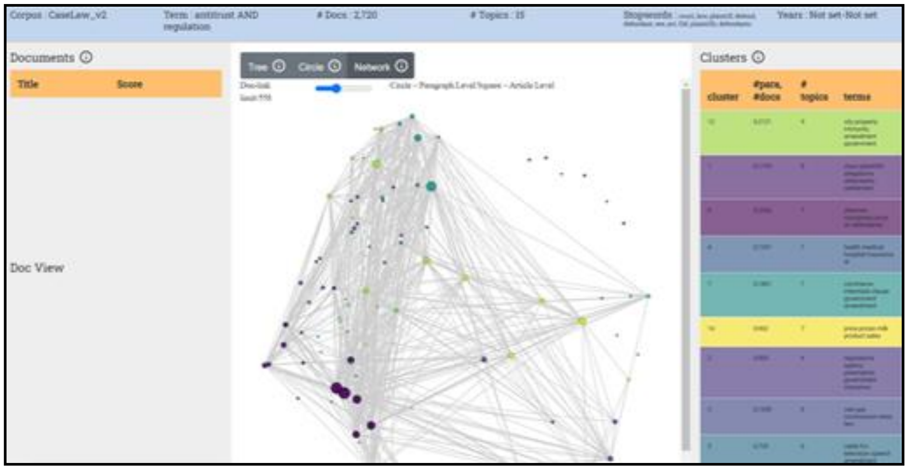
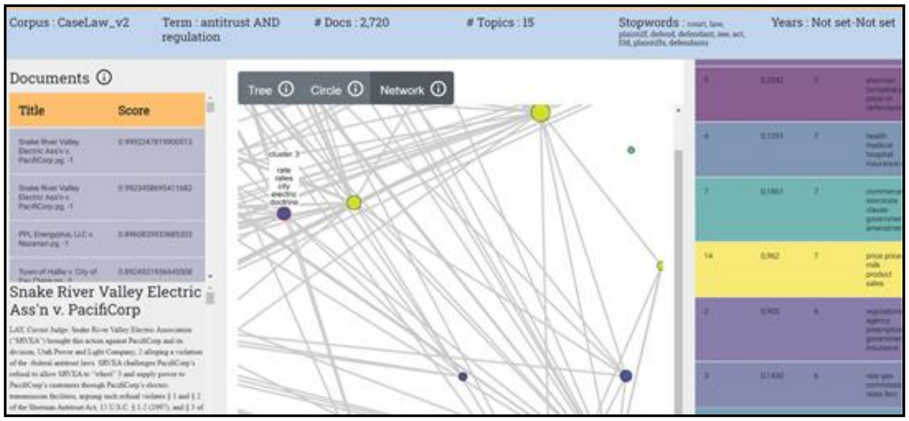

Corpus : CaseLaw_v2 Term : antitrust AND regulation # Docs : 2,720 # Topics : 15 Stopwords : [illegible] Years : Not set-Not set

FIGURE 2: CLOSE-UP VIEW OF ANTITRUST–REGULATION CASES WITH DOCUMENT RETRIEVAL


Corpus : CaseLaw_v2 Term : antitrust AND regulation # Docs : 2,720 # Topics : 15 Stopwords : court, law, plaintiff, defend, defendant, see, act, [illegible] plaintiffs, defendants Years : Not set-Not set

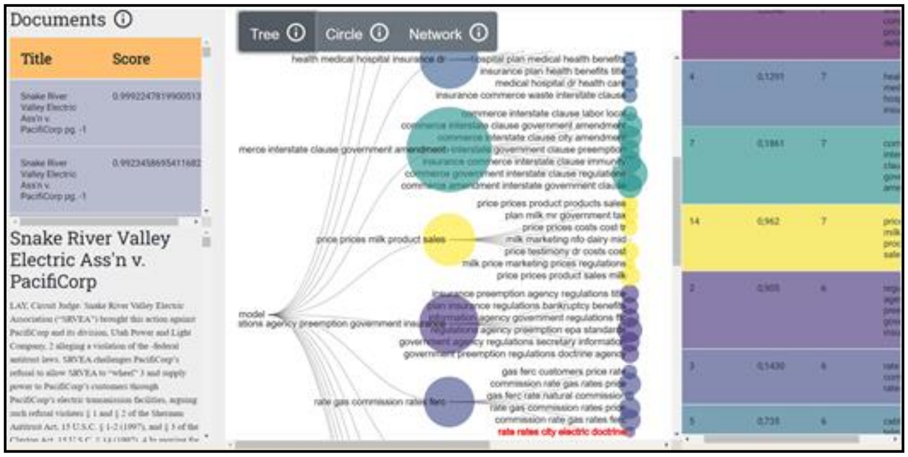
Snake River Valley Electric Ass'n v. PacifiCorp

In classic vector space models, such as Google's Word2Vec, algorithms process the conceptual relations between words and depict each word as a vector, or dimension, in space.[98] 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 network visualization of the models' vector spacing so that we can represent and calculate distance and similarity between topic vectors.

[98] *See* Thomas Mikolov et al., Distributed Representations of Words and Phrases and Their Compositionality, *in* 1 NIPS '13: PROCEEDINGS OF THE 26TH INTERNATIONAL CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS 3136 (C.J.C. Burges et al. eds., 2013); Ash & Chen, *supra* note 20, at 315–17.

The hierarchies of topics and cluster topics is even clearer in "tree" view. The graphic at the center of Figure 3 has grouped topics in the Antitrust–Regulation corpus into fifteen different clusters (numbered 0–14), represented by the larger nodes. Each large cluster is in turn broken down into constituent topics, represented by smaller nodes. Clicking on a topic brings up the top terms. Here, the energy-themed Cluster 3 is comprised of six topics, and the highlighted topic features the top terms "rate," "rates," "city," "electric," "doctrine," "energy" (hidden), "price" (hidden), "utility" (hidden), "regulatory" (hidden), and "wholesale" (hidden).[99] Clicking on the cluster brings up the top documents, or decisions, on the left-hand panel (here, the top case is *Snake River Valley Electric Ass'n v. PacifiCorp*).[100] The right-hand panel shows each cluster and the number of decisions (#docs), the number of topics (#topics), and the top terms in that cluster. The size of each cluster or topic node represents the number of the constituent topics or documents, respectively—or, put differently, the significance of the cluster or topic to the overall corpus.

FIGURE 3: MULTILEVEL VISUALIZATION OF ANTITRUST–REGULATION CASES IN TREE FORMAT



Documents

Title	Score
Snake River Valley Electric Ass'n v. PacifiCorp pg. -1	0.9992247819900513
Snake River Valley Electric Ass'n v. PacifiCorp pg. -1	0.9923458695411682

Snake River Valley Electric Ass'n v. PacifiCorp

LAY, Circuit Judge: Snake River Valley Electric Association ("SRVEA") brought this action against PacifiCorp and its division, Utah Power and Light Company, 2 alleging a violation of the federal antitrust laws. SRVEA challenges PacifiCorp's refusal to allow SRVEA to "wheel" 3 and supply power to PacifiCorp's customers through PacifiCorp's electric transmission facilities, arguing such refusal violates § 1 and § 2 of the Sherman Antitrust Act, 15 U.S.C. § 1-2 (1997), and § 3 of the Clayton Act, 15 U.S.C. § 14 (1997), 4 by requiring the

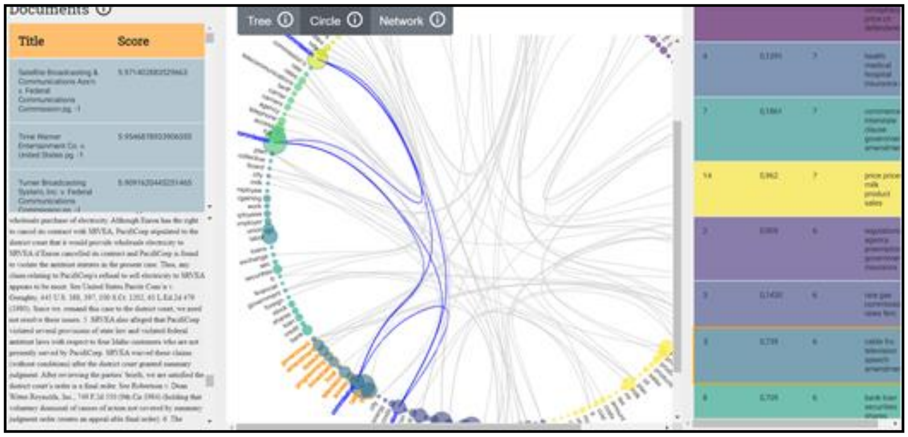
Tree
Circle
Network
health medical hospital insurance dr
hospital plan medical health benefits
insurance plan health benefits title
medical hospital dr health care
insurance commerce waste interstate clause
merce interstate clause government amendment
commerce interstate clause labor local
commerce interstate clause government amendment
commerce interstate clause city amendment
commerce interstate government clause preemption
insurance commerce interstate clause immunity
commerce government interstate clause regulations
commerce amendment interstate government clause
price prices milk product sales
price prices product products sales
plan milk mr government tax
price prices costs cost tr
milk marketing nfo dairy mid
price testimony dr costs cost
milk price marketing prices regulations
price prices product sales milk
model
ations agency preemption government insurance
insurance preemption agency regulations title
plan insurance regulations bankruptcy benefits
information agency government regulations fi
regulations agency preemption epa standards
government agency regulations secretary information
government preemption regulations doctrine agency
rate gas commission rates ferc
gas ferc customers price rate
commission rate gas rates price
gas ferc rate natural commission
rate gas commission rates price
commission rate gas rates ferc
rate rates city electric doctrine
4 0.1291 7
7 0.1861 7
14 0.962 7
2 0.905 6
3 0.1430 6
5 0.735 6

Finally, "circle" view highlights the topics where certain terms appear. Figure 4 highlights the recurrence of the term "commission" across topics pertaining to telecommunications, transportation, energy, and the ICC.

[99] Note that the stop words "court," "law," "plaintiff," "defend," "defendant," "see," "act," "f3d," "plaintiffs," and "defendants" have been excluded.

[100] Snake River Valley Elec. Ass'n v. PacifiCorp, 238 F.3d 1189 (9th Cir. 2001).

FIGURE 4: CIRCLE VIEW SHOWING THE RECURRENCE OF THE TERM "COMMISSION"



Title	Score
Satellite Broadcasting & Communications Ass'n v. Federal Communications Commission pg. -1	[illegible]
Time Warner Entertainment Co. v. United States pg. -1	[illegible]
Turner Broadcasting System, Inc. v. Federal Communications Commission pg. -1	[illegible]

Model aggregation produces several improvements to traditional topic modeling. Aggregation enhances model stability. Whereas traditional topic models may yield different results on different runs,[101] our algorithms run topic models at least twenty times for each query, an iterative process that strips away the insignificant topics. The resulting clusters are comprised of the most stable and persistent topics over multiple iterations. Additionally, aggregation also generates visualizations that set topics in both large and small contexts. Clusters reveal the interconnections among topics, represented in a spatial arrangement that also depicts their semantic difference from one another. Simultaneously, the document retrieval interface allows users to pull up the relevant cases for each topic.

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. The case retrieval function showcases the words of a topic model in their original context in each decision. 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, 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.

II. METHODOLOGY

Big data caselaw research is often hindered by two primary obstacles. First, only a few repositories hold a corpus of easily extractable caselaw.[102] Second,

[101] *See, e.g.*, Da, *supra* note 10, at 625.

[102] 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

even if cases could be easily extracted, their interpretation is limited by modeling that can translate machine analysis into intuitive visualizations.[103]

This Part begins by showing how we tried to overcome both obstacles. 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 available.[104] 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 Part begins by introducing the CAP. Then it explains our data and processing procedures, before concluding with how we validated models through coherence and stability checks.

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.[105] The resulting database is the most comprehensive of its kind outside of the Library of Congress.[106] 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.[107]

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.[108] Texts are presented in machine-readable format, greatly

Register, do not hold cases. Despite the proclivity of law for natural language text mining, easy access to copious amounts of caselaw is limited.

[103] *See generally* Chuang et al., *supra* note 49, at 443–44.

[104] The topic browser visualization is adapted from Andrew Goldstone's dfr-browser project. *See* Andrew Goldstone, *DFR-Browser: Take a MALLETT to Disciplinary History*, ANDREW GOLDSTONE, https://agoldst.github.io/dfr-browser/ [https://perma.cc/8P27-SDHT]; Ben Mabey, *Welcome to PyLDAvis's Documentation!*, PYLDAVIS, https://pyldavis.readthedocs.io/en/latest/index.html [https://perma.cc/CR4B-2SBP].

[105] *See About the Caselaw Access Project*, *supra* note 17.

[106] Jason Tashea, *Caselaw Access Project Gives Free Access to 360 Years of American Court Cases*, ABA J. (Oct. 30, 2018, 7:10 AM), https://www.abajournal.com/news/article/caselaw_access_project_gives_free_access_to_360_years_of_american_court_cas [https://perma.cc/674H-N8E3].

[107] *About the Caselaw Access Project*, *supra* note 17.

[108] *See* Harvard L. Sch., *Tools*, CASELAW ACCESS PROJECT, https://case.law/tools/ [https://perma.cc/6JZ2-9XSG] [hereinafter *Caselaw Access Project Tools*] ("Our downloads

simplifying big data projects.[109] Cases can be extracted through either APIs or bulk downloads.[110] By contrast, commercial legal databases require users to utilize their own APIs to pull large volumes of cases, as well as the execution of license agreements.[111]

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.[112]

Apart from comprehensiveness, CAP is also run on a versatile interface that has shared stock APIs for software developers.[113] It also includes basic searching, browsing, and downloading functions, as well as the ability to explore historical trends in the caselaw.[114] Given CAP's flexibility and ease of use, law scholars have already begun using it for big data projects.[115]

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 Processing*

Data for our project was made available through CAP, which contains 6.9 million unique cases (and over 1.8 million federal cases).[116] Having applied for and obtained researcher access from CAP, we gathered data by writing python-based 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

directory includes derivative datasets, bulk exports, and summaries from the Caselaw Access Project.").

[109] *See* Harvard L. Sch., *Download Files*, CASELAW ACCESS PROJECT, https://case.law/download/ [https://perma.cc/MC3X-SZDA].

[110] *Id.*; Tashea, *supra* note 106.

[111] We spent close to a year negotiating license agreements with Westlaw and Lexis and also tinkering with APIs—until CAP went live.

[112] Tashea, *supra* note 106.

[113] *See id.*

[114] *See Caselaw Access Project Tools*, *supra* note 108.

[115] *See, e.g.*, Jaromir Savelka et al., *Improving Sentence Retrieval from Case Law for Statutory Interpretation*, *in* PROCEEDINGS OF THE SEVENTEENTH INTERNATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE AND LAW 113, 113–22 (2019) (Can.); Jonathan H. Choi, *An Empirical Study of Statutory Interpretation in Tax Law*, 95 N.Y.U. L. REV. 363, 382 (2020).

[116] *See* Harvard L. Sch., *Our Data*, CASELAW ACCESS PROJECT, https://case.law [https://perma.cc/9HJL-XK67]. Note that at the time of writing, CAP had approximately 6.7 million unique federal and state cases and 1.8 million federal cases.

"regulation," a total of approximately 305,000 cases. These serve as the bases for further refinements to the corpora: we filtered the first pool for all decisions with the term "market power" (the "Market Power Corpus"), totaling 2,591; we also filtered the second pool for all decisions with the term "antitrust" (the "Antitrust–Regulation Corpus"), totaling 7,308.

Two limitations in CAP data help explain the relatively small number of cases. 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. Second, CAP excludes unpublished decisions.[117]

Manual assessment quickly becomes impracticable when examining corpora of these sizes. Thus, the application of machine learning provides a more manageable 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.[118] 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 caselaw 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 caselaw, devoid of human—and doctrinal—preconceptions.

Using the text search and analytics engine Elasticsearch (a full-text search and analytics engine)[119] and the python Gensim package,[120] we built a web-based platform. The platform performs topic modeling by using the unsupervised machine learning clustering algorithm LDA to sift through and index decisions. For preprocessing, we utilized Porter Stemmer to excise the suffixes of every word of the corpora.[121]

C. *Triangulation of Models*

Although multilevel visualizations generated from model aggregation are our greatest technical contribution, in analyzing the corpora we often employ two other variations on topic models to bring additional perspectives. One such set of visualizations, "topic browser," organizes the corpora into topics grouped by

[117] For additional verification of this number, including on Westlaw and the CAP platform, see Chang et al., *Computational Antitrust*, *supra* note 83, at 123–24.

[118] *See* Blei et al., *supra* note 5, at 993, 996–97.

[119] *The Heart of the Free and Open Elastic Stack*, ELASTIC, https://www.elastic.co/products/elasticsearch [https://perma.cc/C2GL-N65T].

[120] *Gensim 4.1.2*, PYPI, https://pypi.org/project/gensim/ [https://perma.cc/M8XH-XHEZ].

[121] Martin Porter, *The Porter Stemming Algorithm*, TARTARUS.ORG (2006), https://tartarus.org/martin/PorterStemmer [https://perma.cc/JN72-DHEA].

recurring terms.[122] Topic browser visualizations are particularly useful for illustrating the changes in the significance of all topics over time—that is, what topics comprise a greater proportion of the corpora over time and how its constituent decisions ebb and flow in number (see Figures 5 and 6).

FIGURE 5: MARKET POWER CASES THROUGH THE DECADES (TOPIC BROWSER VIEW)

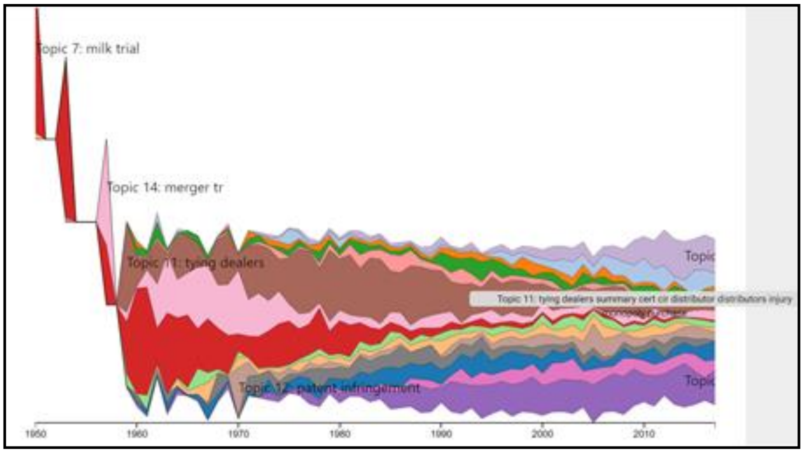
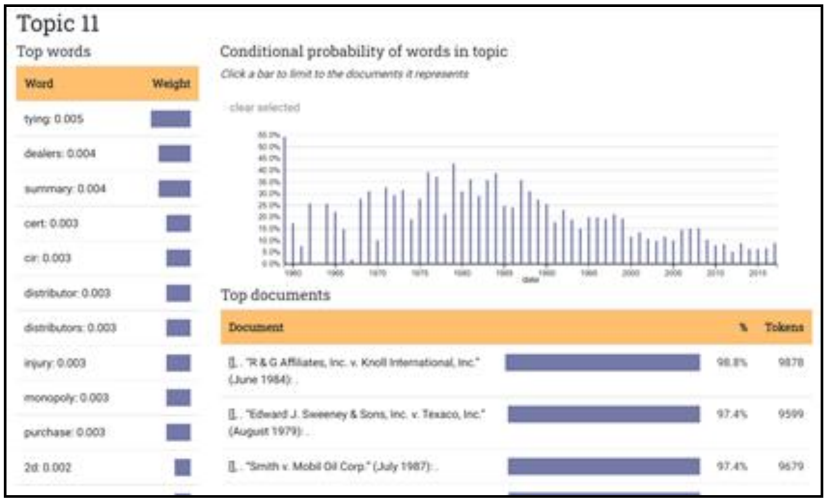


FIGURE 6: TOP TERMS AND CASES IN TOPIC 11 (TOPIC BROWSER VIEW)



Topic 11

Top words

Word	Weight
tying: 0.005	▇
dealers: 0.004	▇
summary: 0.004	▇
cert: 0.003	▇
cir: 0.003	▇
distributor: 0.003	▇
distributors: 0.003	▇
injury: 0.003	▇
monopoly: 0.003	▇
purchase: 0.003	▇
2d: 0.002	▇

Conditional probability of words in topic

Click a bar to limit to the documents it represents

Top documents

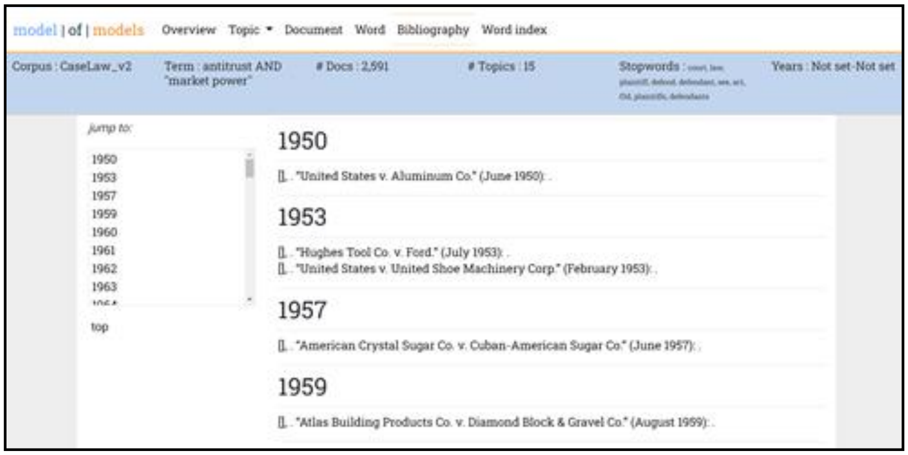
Document		%	Tokens
[], "R & G Affiliates, Inc. v. Knoll International, Inc." (June 1984): .	▇	98.8%	9878
[], "Edward J. Sweeney & Sons, Inc. v. Texaco, Inc." (August 1979): .	▇	97.4%	9599
[], "Smith v. Mobil Oil Corp." (July 1987): .	▇	97.4%	9679

For ease of document retrieval and term cross-referencing, we have engineered topic browser views to enable a user to click on a topic to pull up its top terms and cases. Clicking on a term then creates listing of all topics where the

[122] Topic browser visualizations are adapted from the work of the DH scholar Andrew Goldstone. *See* Goldstone, *supra* note 104.

term recurs.[123] Likewise, clicking on a decision pulls up the topics where the decision can be found. Every decision in the corpus can also be found in the bibliography section of the platform, indexed by year (see Figure 7).

FIGURE 7: BEGINNING OF THE BIBLIOGRAPHY FOR THE MARKET POWER CORPUS



Corpus : CaseLaw_v2 Term : antitrust AND "market power" # Docs : 2,591 # Topics : 15 Stopwords : court, law, plaintiff, defend, defendant, see, art, Cal, plaintiffs, defendants Years : Not set-Not set

1950

[] . "United States v. Aluminum Co." (June 1950): .

1953

[] . "Hughes Tool Co. v. Ford." (July 1953): .
[] . "United States v. United Shoe Machinery Corp." (February 1953): .

1957

[] . "American Crystal Sugar Co. v. Cuban-American Sugar Co." (June 1957): .

1959

[] . "Atlas Building Products Co. v. Diamond Block & Gravel Co." (August 1959): .

In another set of visualizations, python-based LDA visualizations (pyLDAvis), the user sees the corpus in four quadrants, with the distance between topics (represented by bubbles) serving as a proxy for their semantic distance.[124] This depiction resembles the Word2Vec architecture.[125] In our adaptation, the size of each topic bubble represents the significance of that topic to the corpus. Clicking on a topic pulls up its top terms.[126]

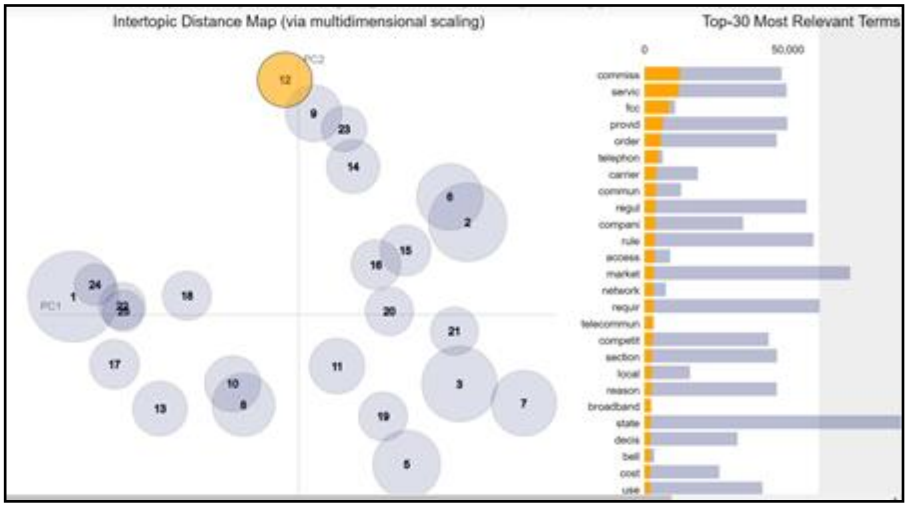
[123] For illustrations, see Chang et al., *Computational Antitrust*, *supra* note 83.

[124] pyLDAvis is derived from the framework of the programmer Ben Mabey. *See* Mabey, *supra* note 104.

[125] *See* Jay Alammar, *The Illustrated Word2Vec*, JAY ALAMMAR (Mar. 27, 2019), https://jalammar.github.io/illustrated-word2vec/ [https://perma.cc/45VT-9BEH]. For a criticism of Word2Vec from the legal perspective, see Amanda Levendowski, *How Copyright Law Can Fix Artificial Intelligence's Implicit Bias Problem*, 93 WASH. L. REV. 579 (2018).

[126] *See* Carson Sievert & Kenneth E. Shirley, LDAvis: 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).

FIGURE 8: PYLDAVIS VIEW OF ANTITRUST–REGULATION CASES



In totality, multilevel, topic browser, and pyLDAvis models triangulate and better contextualize the results. However, as with any empirical project based on copious amounts of data, term relevance and topic modeling are subject to margins of error, or instability (sometimes called "wobbling").[127] 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.

D. *Model Validation*

As in prior collaborations between DSC and other academic units,[128] we have undertaken a multiprong approach to validating our aggregated models, through a combination of qualitative and quantitative methods. This approach includes (i) coherence testing to vet semantic integrity in the topics, (ii) aggregation itself to stabilize the models, and (iii) review of the top cases in a cluster to ensure conformity with the cluster's themes. All in all, this opens the black box of unsupervised ML for scrutiny.

[127] *See* Margaret E. Roberts et al., *Navigating the Local Modes of Big Data: The Case of Topic Models*, *in* COMPUTATIONAL SOCIAL SCIENCE: DISCOVERY AND PREDICTION 51 (R. Michael Alvarez ed., 2016).

[128] *See, e.g.*, Margaret V. Powers-Fletcher et al., *Convergence in Viral Outbreak Research: Using Natural Language Processing to Define Network Bridges in the Bench-Bedside-Population Paradigm*, HARV. DATA SCI. REV., Winter 2021, https://hdsr.mitpress.mit.edu/pub/xhht1a9s/release/2 [https://perma.cc/C88D-XEQG].

Coherence measures the degree of semantic similarity among top terms within a topic.[129] We gauge coherence by testing three variables across topics from several models (alphas, which represent the initial topic probability distributions; betas, which represent term density within a topic; and number of topics in a model, from 10 to 40). (We hold constant one variable and then calculate coherence while varying the other two variables.) We scored coherence on the Market Power Corpus, with stop words excluded; as a baseline, we also ran random coherence on 10,000 documents from the Caselaw Access Project.

For the Market Power Corpus, the minimum and maximum coherence scores were 0.40858 and 0.60601, respectively.[130] We found coherence to be higher when running fifteen topics; hence, we feature the details for this set of multilevel models in the Appendix. For the 10,000 documents, these figures were 0.36367 and 0.49062, respectively.[131] The relatively tight range of coherence across variables, particularly in the 10,000-document corpus, most likely reflects the regimented vocabulary of the CAP dataset. In other words, legal decisions tend to use a fairly standardized vocabulary.

Ideally, we would be aiming for coherence at 0.6 or above. However, these metrics are part of a broader debate over modeling in the data science community.[132] For our part, we lean on the subject matter expertise of our transdisciplinary team to further assess the reliability of our models. Reviewing the top cases in the clusters to check for relevance and conformity with antitrust doctrine allows us to assess the dependability of the results.

To enhance model stability, we have designed the platform to run six models from random seeds of a corpus, which are then integrated into one visualization.[133] The aggregation of topics into clusters emphasizes the overlapping terms and topics from parallel runs while omitting the insignificant ones.

Finally, we have made our code available for transparency and ease of replication.[134]

[129] Shashank Kapadia, *Evaluate Topic Models: Latent Dirichlet Allocation (LDA): A Step-by-Step Guide to Building Interpretable Topic Models*, TOWARDS DATA SCI. (Aug. 19, 2019), https://towardsdatascience.com/evaluate-topic-model-in-python-latent-dirichlet-allocation-lda-7d57484bb5d0 [https://perma.cc/8445-N92N].

[130] *See* CaseLaw Hyperparameter Evaluation 91521 (accessible at https://docs.google.com/spreadsheets/d/1y-qpwlyjQ8WghsTzSjNccjgjM3qGpmcdkInUotbv334/edit#gid=273508978 [https://perma.cc/Z6J7-GLCX] (tab 2)).

[131] *Id.* at tab 1.

[132] *See, e.g.*, Chuang et al., *supra* note 49.

[133] For an extensive discussion of our methodology and model validation, see Chang et al., *Computational Antitrust*, *supra* note 83.

[134] *See* Ezra Edgerton, *Covid Network Bridges Paper Code*, GITHUB: UCDSCENTER (Nov. 2, 2020), https://github.com/ucdscenter/Covid_Network_Bridges_code [https://perma.cc/M2BZ-4YA6]. The modeling source code is the same for our project, even though this page has the title of a different collaboration.

III. RESULTS

To test our modifications to LDA, we analyzed 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 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.[135] 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 Part 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.[136] Practically, however, it is difficult to prove. Direct evidence, such as of

[135] *See, e.g.*, Jack M. Balkin, *The Three Laws of Robotics in the Age of Big Data*, 78 OHIO STATE L.J. 1217 (2017); Dan L. Burk, *Algorithmic Fair Use*, 86 U. CHI. L. REV. 283 (2019); Brennan-Marquez & Henderson, *supra* note 72, at 137; Levendowski, *supra* note 125.

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

anticompetitive effects,[137] is often too hard to come by.[138] Hence, courts must abide by circumstantial evidence of market power, which uses market share as a proxy.[139]

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.[140] In the first step of the paradigm, a relevant product market is defined, enabling the subsequent calculation of a defendant's relevant market share.[141] The product market is drawn, in technical terms, as "the smallest grouping of sales for which the elasticity of demand and supply" are low enough that a monopolist controlling the grouping "could profitably reduce output and increase price substantially above marginal cost."[142] Then, the relevant geographical market is defined along similar lines, and the defendant's geographic market share is also calculated.[143]

Market definition has come under fire from scholars for decades because of its imprecision.[144] The controversy stretches back to one of the first major market power cases, *United States v. du Pont*,[145] 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[146]—even though these substitutes were able to compete with cellophane precisely because du Pont had been underpricing it.[147] 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.[148]

[137] *See* FTC v. Ind. Fed'n of Dentists, 476 U.S. 447, 460–61 (1986) ("Since the purpose of the inquiries into market definition and market power is to determine whether an arrangement has the potential for genuine adverse effects on competition, proof of actual detrimental effects, such as a reduction of output, can obviate the need for an inquiry into market power, which is but a surrogate for detrimental effects." (internal quotations omitted) (quoting 7 PHILLIP AREEDA, ANTITRUST LAW ¶ 1511, at 429 (1986))).

[138] HOVENKAMP, *supra* note 89, § 3.1b, at 106.

[139] *Id.*

[140] *See, e.g.*, Louis Kaplow, *Why (Ever) Define Markets?*, 124 HARV. L. REV. 437, 440 (2010).

[141] HOVENKAMP, *supra* note 89, § 3.1d, at 111.

[142] *Id.* § 3.2, at 112.

[143] *Id.* § 3.1d, at 111.

[144] *See, e.g.*, Herbert Hovenkamp, *Markets in Merger Analysis*, 57 ANTITRUST BULL. 887, 891, 894–95 (2012); *see also* Landes & Posner, *supra* note 136, at 960–63; Kaplow, *supra* note 140.

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

[146] *Id.* at 399–400.

[147] *See* Jonathan B. Baker, *Market Definition: An Analytical Overview*, 74 ANTITRUST L.J. 129 (2007).

[148] *See* Kaplow, *supra* note 140. 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.[149] 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.[150] Market definition is also complicated by a plethora of tests and no systematic way of choosing the right test for each situation at hand.[151]

Nonetheless, examinations of collusion and exclusion are seldom complete without market power analysis of the constituent markets.[152] Market power is the very first step, for instance, in a monopolization action under Section 2 of the Sherman Act,[153] the basis for many of the charges against tech firms.[154] 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 preempted it.[155] 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.[156] 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.[157] In 2004, however, the Court in *Verizon Communications, Inc. v. Trinko* found that even a statute with an antitrust savings

[149] Howard A. Shelanski, *Information, Innovation, and Competition Policy for the Internet*, 161 U. PA. L. REV. 1663, 1674 (2013).

[150] *See* Ohio v. Am. Express Co., 138 S. Ct. 2274 (2018). For criticisms, see John M. Newman, *Antitrust in Digital Markets*, 72 VAND. L. REV. 1497 (2019).

[151] *See generally* Sean P. Sullivan, *Modular Market Definition*, 55 UC DAVIS L. REV. 1091 (2021).

[152] *See* David Glasner & Sean P. Sullivan, *The Logic of Market Definition*, 83 ANTITRUST L.J. 293, 295–97 (2020).

[153] This is the "power plus conduct" framework of *United States v. Grinnell Corp.*, 384 U.S. 563 (1966), and *United States v. Aluminum Co. of Am.* (*Alcoa*), 148 F.2d 416 (2d Cir. 1945). *See also* Sherman Antitrust Act of 1890, 15 U.S.C. § 2.

[154] CHRIS SAGERS, UNITED STATES V. APPLE: COMPETITION IN AMERICA (2019).

[155] *See, e.g.*, Silver v. N.Y. Stock Exch., 373 U.S. 341, 357 (1963).

[156] *See* Gordon v. N.Y. Stock Exch., 422 U.S. 659 (1975).

[157] *See* Howard A. Shelanski, *The Case for Rebalancing Antitrust and Regulation*, 109 MICH. L. REV. 683, 687–89 (2011).

clause—namely, the Telecommunications Act of 1990—could preclude the application of antitrust laws because of the *potential* for conflict.[158]

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.[159] 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).[160] 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*.[161]

The consequences of these shifts are grave. Where regulators have promulgated—and then rescinded—rules to preempt anticompetitive effects,[162] 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 *Trinko*, academics have offered a flurry of proposals to overhaul the balance between antitrust and regulation.[163] In our era of regulatory abdication, scholars are looking to antitrust to step into the voids.[164] Whether those proposals materialize depends in large part on how courts strike that balance.

Most recently, the complaints against Facebook filed by the Federal Trade Commission and state attorneys general and against Google filed by the Justice Department illustrate the stakes for getting this balance right.[165] The government is pursuing remedies, such as divestiture of Instagram and WhatsApp, that go

[158] *See* Verizon Commc'ns Inc. v. L. Offs. of Curtis V. Trinko, 540 U.S. 398 (2004).

[159] Shelanski, *supra* note 157, at 702.

[160] *See* Kearney & Merrill, *supra* note 28, at 1330–34.

[161] *Id.* at 1324–25, 1336–37.

[162] *Compare* Protecting and Promoting the Open Internet, 80 Fed. Reg. 19,737 (Fed. Commc'n Comm'n Apr. 13, 2015) (promulgating net neutrality rules), *with* Restoring Internet Freedom, 33 FCC Rcd. 311 (Dec. 14, 2017) (repealing net neutrality rules).

[163] *See* Brett Frischmann & Spencer Weber Waller, *Revitalizing Essential Facilities*, 75 ANTITRUST L.J. 1 (2008); Adam Candeub, Trinko *and Re-Grounding the Refusal to Deal Doctrine*, 66 U. PITT. L. REV. 821 (2005); Shelanski, *supra* note 157, at 729–31.

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

[165] Complaint for Injunctive and Other Equitable Relief, FTC v. Facebook, Inc., No. 1:20-cv-03590 (D.D.C. filed Dec. 9, 2020) [hereinafter FTC v. Facebook Complaint]; Complaint, New York v. Facebook, Inc., No. 1:20-cv-03589 (D.D.C. filed Dec. 9, 2020); Complaint, United States v. Google LLC, No. 1:20-cv-03010 (D.D.C. filed Oct. 20, 2020).

much further than prayers for relief in prior actions.[166] Because no sectoral regulation currently exists to constrain tech platforms, courts may face intense pressure to devise remedies for problems as varied as privacy, concentration of political power, and distortionary effect on truth, in addition to straightforward antitrust considerations such as monopoly power and anticompetitive effects.

B. *Observations and Inferences*

In this Section, 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 5 above).

Starting in the late 1950s, market power cases exploded.[167] Initially, these cases were concentrated in the banking sector, where a slew of mergers were stayed by regulators and taken to court.[168] Bank merger cases are unique enough to comprise a topic mostly of their own, Topic 14, where several of the top terms are financial (e.g., "merger," "tr[ial]," "acquisition," "bank," "area," "customers," "ftc [F.T.C.]," "share," "banks," and "clayton [Clayton Act]").[169] 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.[170]

[166] *See, e.g.*, FTC v. Facebook Complaint, s*upra* note 165, at 51 (praying for relief such as "divestiture of assets, divestiture or reconstruction of businesses (including, but not limited to, Instagram and/or WhatsApp), and such other relief sufficient to restore the competition that would exist absent the conduct alleged in the Complaint").

[167] For an illustration, see Harvard L. Sch., *Historical Trends*, CASELAW ACCESS PROJECT, https://case.law/trends/ [https://perma.cc/RTH4-467Q] (search for "us: market power").

[168] *See, e.g.*, United States v. Conn. Nat'l Bank, 362 F. Supp. 240 (D. Conn. 1973), *vacated*, 418 U.S. 656 (1974).

[169] Here, the top cases are *Conn. Nat'l Bank*, 362 F. Supp. 240; United States v. Phillipsburg Nat'l Bank & Tr. Co., 306 F. Supp. 645 (D.N.J. 1969), *vacated*, 399 U.S. 350 (1970).

[170] 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 cases have held steady. *See* Harvard L. Sch., *Historical Trends*, CASELAW ACCESS PROJECT, https://case.law/trends/ [https://perma.cc/RTH4-467Q] (search for "us: antitrust, market power").

The only other topic to undergo such a drastic decline is Topic 11, 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).[171] 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.[172] 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.[173] 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.[174] Even though the Supreme Court continued to treat tying as per se illegal,[175] 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.[176]

In place of tying and bank merger cases, litigation-related topics have assumed greater prominence. These include a general litigation topic (Topic 9),[177] two litigation topics dominated by conspiracy claims (Topics 10 and 13),[178] and

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

[172] *See, e.g.*, Einer Elhauge, *Tying, Bundled Discounts, and the Death of the Single Monopoly Profit Theory*, 123 HARV. L. REV. 397 (2009).

[173] *See* HOEVENKAMP, *supra* note 89, at 459.

[174] *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).

[175] *See Hyde*, 466 U.S. at 2; *Eastman Kodak Co.*, 504 U.S. at 451.

[176] *See, e.g.*, Patrick Rey & Jean Tirole, *A Primer on Foreclosure*, *in* 3 HANDBOOK OF INDUSTRIAL ORGANIZATION 2145 (Mark Armstrong & Rob Porter eds., 2006); Elhauge, *supra* note 172, at 397; *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).

[177] In Topic 9, the top terms are "summary," "allegations," "dismiss," "injury," "allege," "alleges," and "count." The top cases are *United Tactical Sys., LLC v. Real Action Paintball, Inc.*, 143 F. Supp. 3d 982 (N.D. Cal. 2015), *Meridian Project Sys., Inc. v. Hardin Const. Co., LLC*, 404 F. Supp. 2d 1214 (E.D. Cal. 2005), and *Levine v. McLeskey*, 881 F. Supp. 1030 (E.D. Va. 1995).

[178] In Topic 10, the top terms are "conspiracy," "allegations," "allege," "agreements," "compl[aint]," and "dismiss." The top cases are *In re Flash Memory Antitrust Litig.*, 643 F. Supp. 2d 1133 (N.D. Cal. 2009), *Williamson Oil Co., Inc. v. Philip Morris USA*, 346 F.3d 1287 (11th Cir. 2003), and *Simpson v. Sanderson Farms, Inc.*, 744 F.3d 702 (11th Cir. 2014). In Topic 13, the top terms are "american," "ex," "doc [document]," "conspiracy," "dfw," "dr [doctor]," and "information." The top cases are *United States v. AMR Corp.*, 140 F. Supp. 2d 1141 (D. Kan. 2001), *Med. Diagnostic Imaging, PLLC v. CareCore Nat., LLC*, 542 F. Supp. 2d 296 (S.D.N.Y. 2008), and *Ross v. Am. Express Co.*, 35 F. Supp. 3d 407 (S.D.N.Y. 2014).

a class actions topic that took off in 1995 (Topic 2).[179] 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.[180] This coincided with U.S. regulators easing up on consolidation on the assumption that localized markets would be open to greater competition, with smaller and less efficient banks being absorbed.[181] 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.[182] 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.

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 11[183] and, to a lesser extent, Topic 3.[184]) Coinciding with their decline, general antitrust litigation topics rose sharply.[185]

[179] In Topic 2, the top terms are "class," "dr [doctor]," "certification," "common," "damages," and "members." The top cases are *In re Hydrogen Peroxide Antitrust Litig.*, 240 F.R.D. 163 (E.D. Pa. 2007), *In re Titanium Dioxide Antitrust Litig.*, 959 284 F.R.D. 328 (D. Md. 2012), and *In re Chocolate Confectionary Antitrust Litig.*, 289 F.R.D. 200 (M.D. Pa. 2012).

[180] *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 (2002).

[181] *Id.* at 250–51.

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

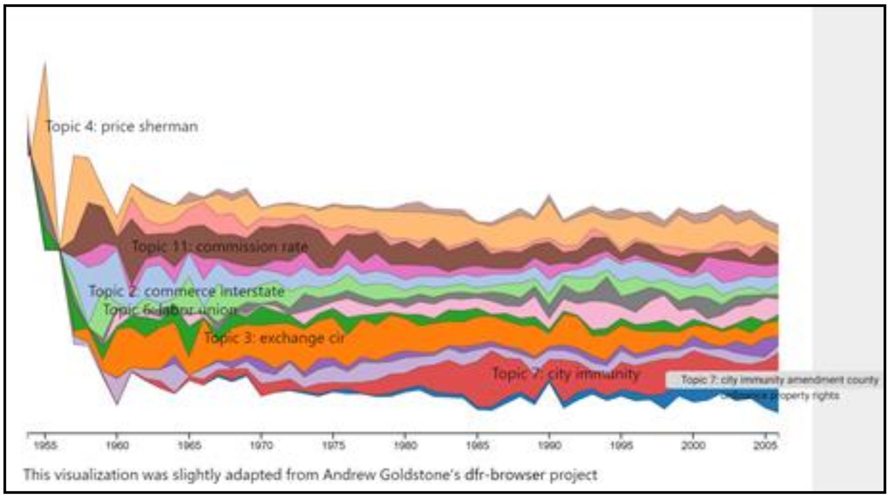
[183] In Topic 11, the top terms are "commiss[ion]," "rate," "gas," rates," "price," "cost," "costs," "ferc [F.E.R.C.]," "natural [gas]," "prices," "transportation," "carrier," "carriers," "customers," "electric," "energy," and "icc [I.C.C.]." *See generally* Cal. Gas Producers Ass'n v. Fed. Power Comm'n, 421 F.2d 422 (9th Cir. 1970); Brizendine v. Cotter & Co., 4 F.3d 457 (7th Cir. 1993), *vacated*, 511 U.S. 1103 (1994); S. La. Area Rate Cases v. Fed. Power Comm'n, 428 F.2d 407 (5th Cir. 1970). Interestingly, these cases span both energy and transportation. Both types of cases engage with similar vocabularies. Note these visualizations were run with fifteen topics over a corpus of 2,720 cases, from filtering down to cases where "antitrust" and "regulation" recur approximately twenty times or more.

[184] In Topic 3, the top terms are "exchange," "cir [circuit]," "conspiracy," "sherman [Sherman Act]," "summary," "american," "arbitration," "damages," "foreign," "price," and "securities." *See generally* Hunt v. Mobile Oil Corp., 444 F. Supp. 68 (S.D.N.Y. 1977); Zuckerman v. Yount, 362 F. Supp. 858 (N.D. Ill. 1973); Apex Oil Co. v. DiMauro, 641 F. Supp. 1246 (S.D.N.Y. 1986).

[185] This includes Topic 1, whose top terms are "class," "allegations," "dismiss," "fsupp2d [F. Supp. 2d]," "wl," allege," "california," "consumer," and "damages." *See generally* Counts v.

These swings cohere with a broader pattern that scholars have previously noticed, where cases pertaining to the Interstate Commerce Commission 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.[186] 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 the Telecommunications Act of 1996, a trend commonly (though not altogether accurately) called *deregulation.*[187]

FIGURE 9: ANTITRUST–REGULATION CASES THROUGH THE DECADES



This visualization was slightly adapted from Andrew Goldstone's dfr-browser project

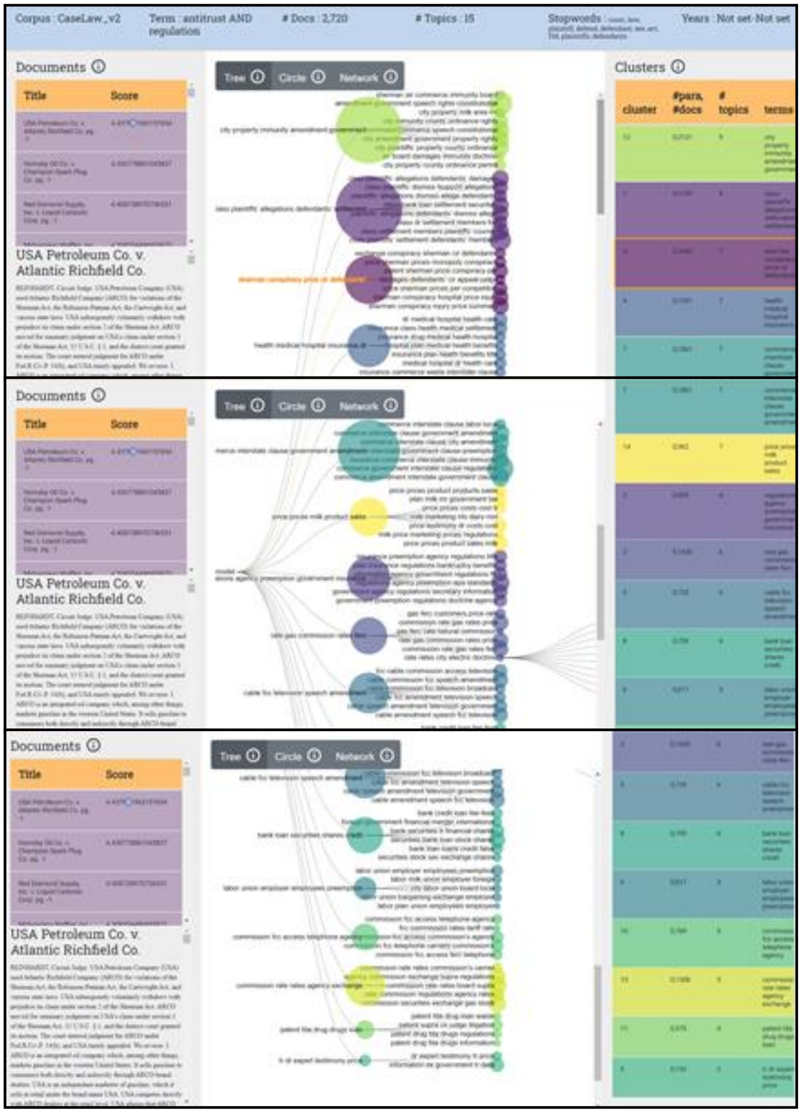
Topic browser histograms are a good starting point for historical trends. However, because topic browser view lists dozens of terms for each topic, the details can quickly overwhelm. As a supplement, then, we use the multilevel visualizations of aggregated modeling to eliminate the "noise" and scale 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).

Gen. Motors, LLC, 237 F. Supp. 3d 572 (E.D. Mich. 2017); *In re* Packaged Seafood Prods. Antitrust Litig., 242 F. Supp. 3d 1033 (S.D. Cal. 2017); Chavez v. Blue Sky Nat. Beverage Co., 268 F.R.D. 365 (N.D. Cal. 2010).

[186] Kearney & Merrill, *supra* note 28, at 1330–33.

[187] *Id*.

FIGURE 10: MULTILEVEL VISUALIZATION OF ANTITRUST–REGULATION CORPUS



Corpus : CaseLaw_v2 Term : antitrust AND regulation # Docs : 2,720 # Topics : 15 Stopwords : Years : Not set-Not set

In Figure 10, the Antitrust–Regulation Corpus is broken down into fifteen clusters that correspond to the regulatory frameworks where antitrust litigation frequently arises. These include patent,[188] health care,[189] telecommunications,[190]

[188] *See supra* Figure 10 (Cluster 11).

[189] *Id*. (Cluster 4).

[190] *Id*. (Clusters 5 and 10).

banking and securities,[191] labor,[192] energy,[193] banking,[194] and dairy industries,[195] as well as general regulatory[196] and litigation[197] clusters. Significantly, Cluster 0, the largest topic cluster, covers seven topics that share the terms "conspiracy" and/or "price." These topics cover almost every imaginable industry, including oil, diamonds, restaurants, retail space, health care, telecommunications, and beverages.

The prevalence of "conspiracy" and "price" in Cluster 0 suggests that a plaintiff's framing of the defendants' actions as a conspiracy, contract, or other agreement over price 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.[198] Indeed, collusive acts such as price-fixing and market division are often viewed as the core of antitrust prohibitions.[199] 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 conscious parallelism to coordination.[200] Against this doctrinal backdrop, many of the cases in Cluster 0 feature agreements permitted by regulatory frameworks but nonetheless charged by plaintiffs as anticompetitive.[201]

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

[191] *Id.* (Cluster 8).

[192] *Id.* (Cluster 6).

[193] *Id.* (Cluster 3).

[194] *Id.* (Cluster 7).

[195] *Id.* (Cluster 14).

[196] *Id.* (Clusters 12, 7, 2, and 13).

[197] *Id.* (Clusters 1 (class actions), 0 (general antitrust litigation), and 9 (expert witnesses)).

[198] 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 Elec. Indus. Co. v. Zenith Radio Corp., 475 U.S. 574 (1986).

[199] *See* Jonathan B. Baker, *Exclusion as a Core Competition Concern*, 78 ANTITRUST L.J. 527, 545 (2013).

[200] 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 Litig.*, 630 F.3d 622 (7th Cir. 2010), and *In re Text Messaging Antitrust Litig.*, 782 F.3d 867 (7th Cir. 2015).

[201] *See, e.g.*, Kartell v. Blue Shield of Mass., 542 F. Supp. 782 (D. Mass. 1982); Bd. of Com'rs v. Fed. Maritime Comm'n, 440 F.2d 1312 (5th Cir. 1971); Metro. Intercollegiate Basketball Ass'n v. Nat'l Collegiate Athletics Ass'n, 337 F. Supp. 2d 563 (S.D.N.Y. 2004).

manufacturing was overrepresented in private antitrust suits.[202] In their seminal article analyzing data from the Georgetown Private Antitrust Litigation Study (the "Georgetown Study"), Professors Salop and White found that 44.3% of defendants and 24.1% of plaintiffs hailed from the manufacturing sector.[203] 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.[204]

Data from the Georgetown Study ends 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.[205] 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.[206] 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.[207] We can partially preempt such results by screening for cases where query words (e.g., "market power") appear more than a desired number of times (e.g., ten times). In this way, the visualizations will be compiled only out of those cases.

[202] 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*, 54 ANTITRUST L.J. 59 (1985).

[203] *Id.*at 1004.

[204] *Id.* at 1005.

[205] 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.

[206] *See, e.g.*, Verizon Commc'ns Inc. v. L. Offs. of Curtis V. Trinko, 540 U.S. 398 (2004).

[207] *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.[208]

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.[209] 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% of the corpus, a fairly sizeable proportion.[210] Relatedly, 2.4% of the Georgetown corpus featured cases where an antitrust action was appended as a counterclaim.[211] 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,[212] antitrust counterclaims are almost induced by their quasi-per se treatment.[213] 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*

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

[209] *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 Int'l, 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).

[210] Salop & White, *supra* note 202, at 1048–49.

[211] *Id.* at 1048.

[212] 12 U.S.C. § 1972.

[213] For more on the bank tying provisions, see generally Felix B. Chang, *Death to Credit as Leverage: Using the Bank Anti-Tying Provision to Curb Financial Risk*, 9 N.Y.U J.L. & BUS. 851 (2013). Tying counterclaims are often found in cases where a lender moves against a defaulting borrower.

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 7 (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.[214] 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.[215] Parker immunity, or the antitrust state action doctrine, covers certain state and local regulations that affect competition, exempting them from federal antitrust laws.[216] It is a variation on federalism questions more typically encountered in constitutional law. By contrast, antitrust immunity in the regulatory setting usually implicates the question of whether regulation displaces antitrust—and the extent to which an antitrust savings clause resuscitates private antitrust litigation from regulatory pre-emption.[217]

We can confirm that *Trinko*[218] and the old cases on repugnancy such as *Silver*[219] and *Gordon*[220] do appear in Topic 9—they are just not among the top results.[221] In fact, *Trinko* has a closer association with other topics (e.g., pertaining to telecommunications, federal legislation, and antitrust procedure) than with antitrust immunity (see Figure 11).[222] 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

[214] On the general balance between antitrust and regulation, see generally Shelanski, *supra* note 157.

[215] *See* Parker v. Brown, 317 U.S. 341 (1943). The top terms in Topic 7 are "city," "immunity," "amendment," "county," "ordinance," "property," "rights," "constitutional," "doctrine," "government," and "local." Further down the list, terms also include "noerr" (after the Noerr-Pennington Doctrine) and "parker." The top cases are *Traweek v. City and Cnty. of San Francisco*, 920 F.2d 589 (9th Cir. 1990), *Falls Chase Special Taxing Dist. v. City of Tallahassee*, 788 F.2d 711 (11th Cir. 1986), and *Jacobs, Visconsi & Jacobs Co. v. City of Lawrence*, 715 F. Supp. 1000 (D. Kan. 1989). Further down the list, two other top cases are *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), a set of classic cases on state action exemption.

[216] *See Parker*, 317 U.S. at 341.

[217] *See* Shelanski, *supra* note 157, at 690–94.

[218] Verizon Commc'ns Inc. v. L. Offs. of Curtis V. Trinko, 540 U.S. 398 (2004).

[219] Silver v. N.Y. Stock Exch., 373 U.S. 341, 357 (1963).

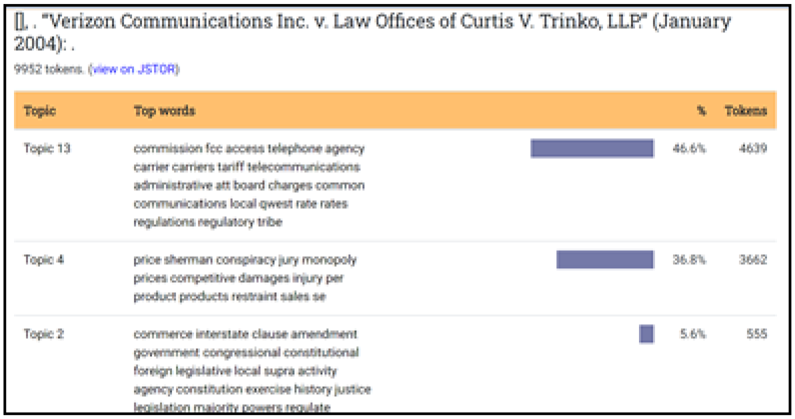
[220] Gordon v. N.Y. Stock Exch., 422 U.S. 659 (1975).

[221] The platform has a "bibliography" feature that lists all cases.

[222] *Trinko* has a 46.6% association with Topic 13 (top words "commission," "fcc," "access," and "telephone"), a 36.8% association with Topic 4 (top words "price," "sherman," "conspiracy," "jury," and "monopoly"), and a 5.6% association with Topic 2 (top words "commerce," "interstate," "clause," "amendment," and "government"). It has only a 2.2% association with a topic that has "immunity" as a top term, Topic 3 (top words "exchange," "cir [circuit]," "conspiracy," "sherman," "summary," "american," and "arbitration").

facilities doctrine, and antitrust immunity.[223] Put differently, a case can constitute precedent in a number of areas.

FIGURE 11: CROSS-REFERENCE OF *TRINKO* AND ASSOCIATED TOPICS



[], . "Verizon Communications Inc. v. Law Offices of Curtis V. Trinko, LLP." (January 2004): .

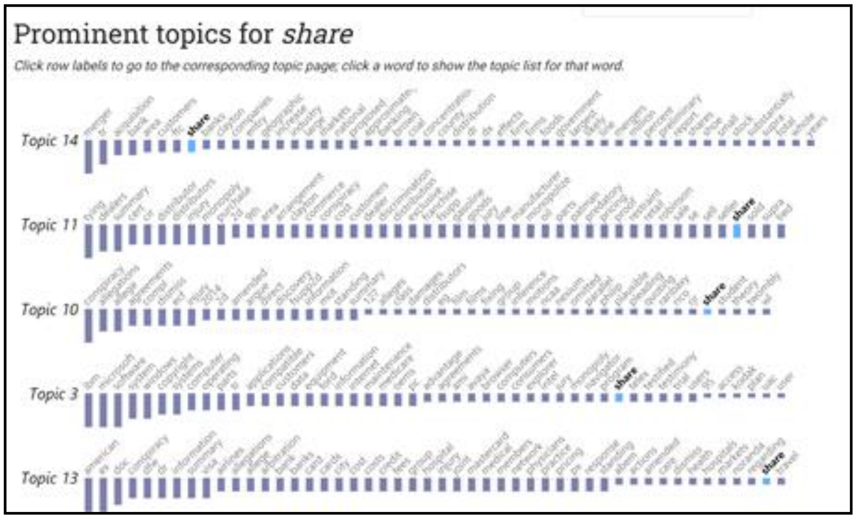
9952 tokens. (view on JSTOR)

Topic	Top words	%	Tokens
Topic 13	commission fcc access telephone agency carrier carriers tariff telecommunications administrative att board charges common communications local qwest rate rates regulations regulatory tribe	46.6%	4639
Topic 4	price sherman conspiracy jury monopoly prices competitive damages injury per product products restraint sales se	36.8%	3662
Topic 2	commerce interstate clause amendment government congressional constitutional foreign legislative local supra activity agency constitution exercise history justice legislation majority powers regulate	5.6%	555

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 "share" does not appear across even half of the topics in the Market Power Corpus (see Figure 12).

[223] Interestingly, *Trinko* is not among the top twenty 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 fourth result.

FIGURE 12: TOPIC BROWSER VIEW OF TOPICS CONTAINING "SHARE"

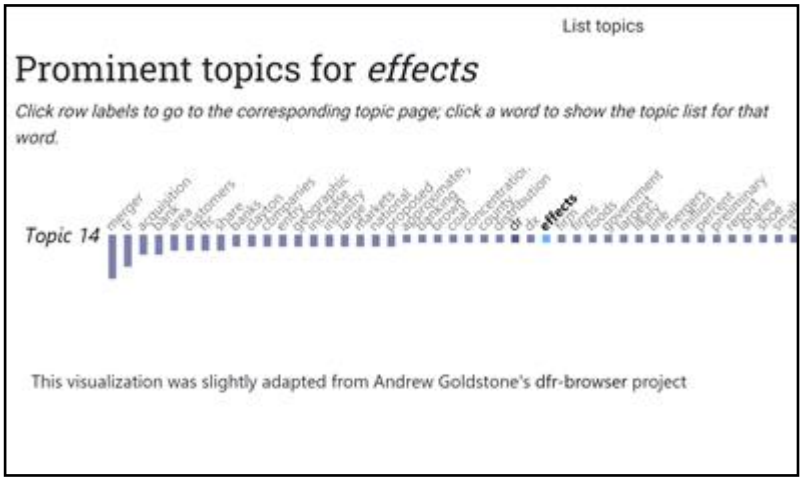


Prominent topics for *share*

Click row labels to go to the corresponding topic page; click a word to show the topic list for that word.

We might reasonably attribute this to two possibilities: either a court has accepted one party's market definition and market share calculation, 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 13), which is hardly surprising, since anticompetitive effects are difficult enough for economists to measure and even harder for courts to articulate. Significantly, the terms "share" and "effect" do not overlap in topics, so we might also postulate that courts are using them as alternative proxies for market power.

FIGURE 13: TOPIC BROWSER VIEW OF TOPICS CONTAINING "EFFECTS"



Prominent topics for *effects*

Click row labels to go to the corresponding topic page; click a word to show the topic list for that word.

This visualization was slightly adapted from Andrew Goldstone's dfr-browser project

As we read the cases in the topics, however, we see that these inferences must be cautiously drawn. For instance, even within topics where "share" is not highlighted as among the words (each topic lists approximately fifty top words), we find that courts often do take up the relevant product market share, even if in cursory form. There simply may have been fifty other words that show much more frequently in the topic than "share."[224] 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.

IV. 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.[225] 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."[226] 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 Part 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 disciplinary silos and cut through citation biases.[227] Some of the first computer scientists to experiment with collaborative

[224] *See* Chuang et al., *supra* note 49, at 444 ("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.").

[225] *See id.*

[226] *Id.*

[227] *See* Wang & Blei, *supra* note 50, at 448–50; *see also supra* notes 50–54 and accompanying discussion.

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.[228] Consequently, a scholar will tend to cite others within their discipline, at the expense of finding relevant literature in another field.[229] 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.[230]

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,[231] 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 Securities (USA) LLC v. Billings*,[232] might otherwise miss this connection, directing the researcher simply to the antitrust–regulation balance. A few scholars writing on antitrust immunity have already observed the connections between regulatory pre-emption and state action,[233] as has at least one court.[234] 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 fifty times or more, the algorithms return decisions where "regulation" may appear forty-nine times in a document but "antitrust" appears only once. Several of the top decisions in Topic 14 of the Antitrust–Regulation Corpus, for instance, feature the term "antitrust" only in the context of quoting antitrust cases as precedent on

[228] Wang & Blei, *supra* note 50, at 450.

[229] *Id.*

[230] To quote a critic of topic modeling, its utility may well be a "content-based recommendation [system] (such as Facebook advertising products to its users)." Da, *supra* note 10, at 625.

[231] *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.

[232] Verizon Commc'ns Inc. v. L. Offs. of Curtis V. Trinko, 540 U.S. 398 (2004); Credit Suisse Sec. (USA) LLC v. Billing, 551 U.S. 264 (2007). These are top two cases for savings clauses.

[233] *See* Darren Bush, *Mission Creep: Antitrust Exemptions and Immunities as Applied to Deregulated Industries*, 2006 UTAH L. REV. 613, 661 (2006); Daniel F. Spulber & Christopher S. Yoo, *Mandating Access to Telecom and the Internet: The Hidden Side of* Trinko, 107 COLUM. L. REV. 1822, 1822 (2007); *see also* HOVENKAMP, *supra* note 89, at § 19.3c.

[234] *See* Am. Agric. Movement v. Bd. of Trade, 977 F.2d 1147, 1155 (7th Cir. 1992).

injunctions.[235] 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.[236] An easy improvement, however, is to extend the numerical filters to both "antitrust" and "regulation."

The results from Topic 14 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 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.[237] 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

[235] *See, e.g.*, Kimberlin v. Quinlan, 6 F.3d 789, 807 (D.C. Cir. 1993); Real Truth About Obama, Inc. v. FEC, 575 F.3d 342, 345 (4th Cir. 2009); Kiser v. Kamdar, 831 F.3d 784, 791 (6th Cir. 2016); O Centro Espirita Beneficiente Uniao Do Vegetal v. Ashcroft, 389 F.3d 973, 978–79 (10th Cir. 2004).

[236] *See, e.g.*, Bell Atl. Corp. v. Twombly, 550, U.S. 544 (2007); Matsushita v. Zenith Radio Corp., 475 U.S. 574, 585–86 (1986).

[237] *See, e.g.*, Cheryl I. Harris, *Whiteness as Property*, 106 HARV. L. REV. 1709, 1707–09 (1993); Samuel Issacharoff & Richard H. Pildes, *Politics as Markets: Partisan Lockups of the Democratic Process*, 50 STAN. L. REV. 643, 646 (1998); Jonathan R. Macey & James P. Holdcroft, Jr., *Failure Is an Option: An Ersatz-Antitrust Approach to Financial Regulation*, 120 YALE L.J. 1368, 1370 (2011); Darrell A.H. Miller, *Text, History, and Tradition: What the Seventh Amendment Can Teach Us About the Second*, 122 YALE L.J. 852, 855–56 (2013).

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.

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 Co. v. United States*,[238] *United States v. Alcoa*,[239] *Grinnell*,[240] and *du Pont*,[241] all of them classic market power cases, do not appear as terms in the Market Power Corpus.[242] In a narrowly focused topic—say, on tying—landmark cases such as *Eastman Kodak*[243] and *Jefferson Parish*[244] do not materialize as terms either. (The notable exception is *Microsoft*,[245] which shows up more frequently, even being picked up as a term in multilevel view.)[246] 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.[247] Another discrepancy from commercial databases is that topic modeling occasionally returns decisions that have been overturned or vacated.[248] 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.

[238] Lorain J. Co. v. United States, 342 U.S. 143, 144 (1951).

[239] United States v. Aluminum Co. of Am. (*Alcoa*), 148 F.2d 416, 421 (2d Cir. 1945).

[240] United States v. Grinnell Corp., 384 U.S. 563, 566 (1966).

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

[242] *See supra* text accompanying note 193. (We know these cases are part of the corpus because they appear in the bibliography.)

[243] Eastman Kodak Co. v. Image Tech. Servs., Inc., 504 U.S. 451 (1992).

[244] Jefferson Parish Hosp. Dist. No. 2 v. Hyde, 466 U.S. 2 (1984).

[245] United States v. Microsoft Corp., 253 F.3d 34 (D.C. Cir. 2001).

[246] 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 by checking against the bibliography. This may simply be because *Trinko* is still relatively recent and has not been cited by other cases incorporated into the modeling.

[247] *See supra* text accompanying note 145 (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* text accompanying note 169 (search in "citing references" for Town of Hallie v. City of Eau Claire, 471 U.S. 34 (1985) (seminal Parker immunity case)).

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

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 astonishingly little overlap in the top cases when a query is run.[249] As Professor Mart notes, these inexplicable results are frustrating because the platform operators reveal virtually nothing about their algorithms.[250] On a different level, algorithms compound human biases, and society is urgently re-evaluating the use of artificial intelligence for predictive purposes.[251] The lack of "algorithmic accountability" on the part of commercial databases is a detriment to research and the legal profession.[252]

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.[253] They promise to harness innovations in information technology to deliver "faster and smarter" legal research.[254] 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

[249] Mart, *supra* note 38, 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]").

[250] *Id.* at 389.

[251] *See* Cade Metz & Adam Satariano, *The Algorithm That Grants Freedom, or Takes It Away*, N.Y. TIMES (Feb. 6, 2020), https://www.nytimes.com/2020/02/06/technology/predictive-algorithms-crime.html [https://perma.cc/3LUJ-MY5R].

[252] *See* Mart, *supra* note 38, 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.").

[253] *See id.* (empirical comparison of legal research providers).

[254] *What Is Fastcase?*, FASTCASE (Feb. 3, 2020), https://www.fastcase.com/about/ [https://perma.cc/Y628-K4ME].

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,[255] we anticipate that users will press providers for an explanation.

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 compromise the function of uncovering patterns 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.

[255] *See* Mart, *supra* note 38, at 389.

APPENDIX

This Appendix lists the fifteen topic clusters in the Market Power Corpus (2,591 total decisions) and the fifteen topic clusters 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, the Antitrust–Regulation Corpus was filtered down to decisions where the key terms ("antitrust" and "regulation") occur more than approximately twenty times in each decision, resulting in a total of 2,720 documents.

For both corpora, we excluded the stop words "court," "law," "plaintiff," "defend," "defendant," "see," "act," "f3d," "plaintiffs," and "defendants" from the visualizations.

Note the tables are sorted in descending order of cluster size, as measured in the number of topics in each cluster (as opposed to the number of decisions in each cluster). Note, too, that because aggregation is over six different models, some of the clusters will have more cases than are in each corpus. (Each model can have a maximum of all cases in the corpus; over six cases, that maximum could grow six-fold.)

MARKET POWER CORPUS

Cluster	# Cases	# Topics	Top Terms in Cluster	Top Decisions in Cluster
2	915	9	patent, infringement, patents, said, generic	Engel Indus., Inc. v. Lockformer Co., 96 F.3d 1398 (Fed. Cir. 1996); Sunrise Med. HHG, Inc. v. AirSep Corp., 95 F. Supp. 2d 348 (W.D. Pa. 2000); Undersea Breathing Sys., Inc. v. Nitrox Techs., Inc., 985 F. Supp. 752 (N.D. Ill. 1997); Beckman Instruments, Inc. v. Chemtronics, Inc., 439 F.2d 1369 (5th Cir. 1970)
7	3,853	9	cir [circuit], tying, dealers, jury, damages	Rea v. Ford Motor Co., 497 F.2d 577 (3d Cir. 1974); Fox Motors, Inc. v. Mazda Distribs. (Gulf), Inc., 806 F.2d 953 (10th Cir. 1986); Petrol. for Contractors, Inc. v. Mobil Oil Corp., 493 F. Supp. 320 (S.D.N.Y. 1980); Siegel v. Chicken Delight, Inc., 311 F. Supp. 847 (N.D. Cal. 1970)
13	3,999	8	injury, allegations, defendants', dismiss, summary	Levine v. McLeskey, 881 F. Supp. 1030 (E.D. Va. 1995); Bushnell Corp. v. ITT, 973 F. Supp. 1276 (D. Kan. 1997); Wolf Concept S.A.R.L. v. Eber Bros. Wine & Liquor Corp., 736 F. Supp.2d 661 (W.D.N.Y. 2010); Yong Ki Hong v. KBS Am., Inc., 951 F. Supp.2d 402 (E.D.N.Y. 2013)
1	1,155	7	commission, fcc [F.C.C.], icc [I.C.C.], carriers, gas	Cent. & S. Motor Freight Tariff Ass'n v. United States, 757 F.2d 301 (D.C. Cir. 1985); Sea-Land Serv., Inc. v. Fed. Maritime Comm'n, 653 F.2d 544 (D.C. Cir. 1981); Am. Trucking Ass'ns v. United States, 755 F.2d 1292 (7th Cir. 1985); Am. Trucking Ass'ns v. Interstate Com. Comm'n, 656 F.2d 1115 (5th Cir. 1981)

11	1,012	7	merger, tr [trial], acquisition, bank, ftc	United States v. Black & Decker Mfg. Co., 430 F. Supp. 729 (D. Md. 1976); United States v. Falstaff Brewing Corp., 410 U.S. 526 (1973); United States v. Mfrs. Hanover Tr. Co., 240 F. Supp. 867 (S.D.N.Y. 1965); United States v. Ivaco, Inc., 704 F. Supp. 1409 (W.D. Mich. 1989)
0	925	6	nfl [NFL], ncaa [NCAA], league, se, dr	Metro. Intercollegiate Basketball Ass'n v. Nat'l Collegiate Athletic Ass'n, 339 F. Supp.2d 545 (S.D.N.Y. 2004); Smith v. Pro Football, Inc., 593 F.2d 1173 (D.C. Cir. 1978); Ass'n for Intercollegiate Athletics for Women v. Nat'l Collegiate Athletic Ass'n, 558 F. Supp. 487 (D.D.C. 1983); Chi. Pro. Sports Ltd. v. Nat'l Basketball Ass'n, 874 F. Supp. 844 (N.D. Ill. 1995)
3	1,117	6	trial, jury, cost, cir [circuit], pricing	Weber v. Wynne, 431 F. Supp. 1048 (D.N.J. 1977); Sunkist Growers, Inc. v. Winckler & Smith Citrus Prod. Co., 284 F.2d 1 (9th Cir. 1960); United States v. Aluminum Co. of Am. (*Alcoa*), 91 F. Supp. 333 (S.D.N.Y. 1950); Barry Wright Corp. v. Pac. Sci. Corp., 555 F. Supp. 1264 (D. Mass. 1983)
4	606	6	cable, license, television, programming, fcc [F.C.C.]	Turner Broad. Sys., Inc. v. FCC, 520 U.S. 180 (1997); Turner Broad. Sys., Inc. v. FCC, 512 U.S. 622 (1994); Time Warner Ent. Co. v. United States, 211 F.3d 1313 (D.C. Cir. 2000); Turner Broad. Sys., Inc. v. FCC, 819 F. Supp. 32 (D.D.C. 1993)
5	872	6	class, dr, members, certification, plaintiffs'	*In re* Processed Egg Prods. Antitrust Litig., 312 F.R.D. 171 (E.D. Pa. 2015); *In re* Chocolate Confectionary Antitrust Litig., 289 F.R.D. 200 (M.D. Pa. 2012); *In re* Titanium Dioxide Antitrust Litig., 284 F.R.D. 328 (D. Md. 2012); *In re* Blood Reagents Antitrust Litig., 283 F.R.D. 222 (E.D. Pa. 2012)

8	603	6	gas, ferc [F.E.R.C], rate, energy, commission	Pub. Utils. Comm’n v. FERC, 456 F.3d 1025 (9th Cir. 2006); Pub. Utils. Comm’n v. FERC, 462 F.3d 1027 (9th Cir. 2006); New Jersey Bd. of Pub. Utils. v. FERC, 744 F.3d 74 (3d Cir. 2014); PPL Energyplus, LLC v. Nazarian, 974 F. Supp. 2d 790 (D. Md. 2013); Pub. Util. Dist. No. 1 v. Dynegy Power Mktg., Inc., 384 F.3d 756 (9th Cir. 2004)
10	1,357	6	plaintiffs’, allegations, allege, dismiss, conspiracy	Myun-Uk Choi v. Tower Rsch. Cap. LLC, 165 F. Supp. 3d 42 (S.D.N.Y. 2016); *In re* Commodity Exch., Inc. Silver Futures and Options Trading Litig., 560 Fed. App’x 84 (2d Cir. 2014); Burtch v. Milberg Factors., Inc., 662 F.3d 212 (3d Cir. 2011); *In re* Interest Rate Swaps Antitrust Litig., 261 F. Supp. 3d 430 (S.D.N.Y. 2017)
9	1,105	5	hospital, dr [doctor], medical, hospitals, care	Koefoot v. Am. Coll. of Surgeons, 610 F. Supp. 1298 (N.D. Ill. 1985); Westchester Radiological Assocs. P.C. v. Empire Blue Cross and Blue Shield, Inc., 707 F. Supp. 708 (S.D.N.Y. 1989); Johnson v. Blue Cross/Blue Shield of N.M., 677 F. Supp. 1112 (D.N.M. 1987); Ball Memorial Hosp., Inc. v. Mutual Hosp. Ins., Inc., 603 F. Supp. 1077 (S.D. Ind. 1985)
6	220	4	microsoft, software, windows, operating, microsoft’s	United States v. Microsoft Corp., 84 F. Supp. 2d 9 (D.D.C. 1999); United States v. Microsoft Corp., 65 F. Supp. 2d 1 (D.D.C. 1999); Sun Microsystems, Inc. v. Microsoft Corp. 87 F. Supp. 2d 992 (N.D. Cal. 2000); New York v. Microsoft Corp., 224 F. Supp.2d 76 (D.D.C. 2002); Massachusetts v. Microsoft Corp., 373 F.3d 1199 (D.D.C. 2004)

14	281	3	milk, insurance, plaintiffs', rico, labor	*In re* Midwest Milk Monopolization Litig., 510 F. Supp. 381 (W.D. Mo. 1981); Alexander v. Nat'l Farmers Org., 687 F.2d 1173 (8th Cir. 1982); Case-Swayne Co. v. Sunkist Growers, Inc., 389 U.S. 384 (1967); Cochran v. Veneman, 252 F. Supp.2d 126 (M.D. Pa. 2003)
12	122	2	license, ascap [ASCAP], music, generic, patent	United States v. Am. Soc'y of Composers, Authors and Publishers, 157 F.R.D. 173 (S.D.N.Y. 1994); Broad. Music, Inc. v. DMX Inc., 683 F.3d 32 (2d Cir. 2012); *In re* THP Capstar Acquisition Corp., 756 F. Supp.2d 516 (S.D.N.Y. 2010); United States v. Am. Soc'y of Composers, Authors and Publishers, 981 F. Supp. 199 (S.D.N.Y. 1997)

ANTITRUST–REGULATION CORPUS

Cluster	# Cases	# Topics	Top Terms in Cluster	Top Decisions in Cluster
12	2,121	9	city, property, immunity, amendment, government	World Wide Rush LLC v. City of Los Angeles, 605 F. Supp.2d 1088 (C.D. Cal. 2009); Anselmo v. Cty. of Shasta, Cal., 873 F. Supp.2d 1247 (E.D. Cal. 2012); Congregation Rabbinical Coll. of Tartikov, Inc. v. Vill. of Pomona, 915 F. Supp. 2d 574 (S.D.N.Y. 2013); Dolls, Inc. v. City of Coralville, Iowa, 425 F. Supp. 2d 958 (S.D. Iowa 2006)
1	1,799	8	price, product, prices, sales, products	Shady Grove Ortho. Assocs., P.A. v. Allstate Ins. Co., 293 F.R.D. 287 (E.D.N.Y. 2013); *In re* Relafen Antitrust Litig., 221 F.R.D. 260 (D. Mass. 2004); Jacob v. Duane Reade, Inc., 293 F.R.D. 578 (S.D.N.Y. 2013); Belfiore v. Procter & Gamble Co., 311 F.R.D. 29 (E.D.N.Y. 2015)

0	3,342	7	Sherman [Sherman Act], conspiracy, price, cir [circuit], defendants'	USA Petrol. Co. v. Atl. Richfield Co., 859 F.2d 687 (9th Cir. 1988); Hornsby Oil Co. v. Champ. Spark Plug Co., 714 F.2d 1384 (5th Cir. 1983); Red Diamond Supply, Inc. v. Liquid Carbonic Corp., 637 F.2d 1001 (1st Cir. 1981); Midw. Waffles, Inc. v. Waffle House, Inc., 734 F.2d 705 (11th Cir. 1984)
4	1,291	7	health, medical, hospital, insurance, dr [doctor]	U.S. *ex rel.* Villafane v. Solinger, 543 F. Supp. 2d 678 (W.D. Ky. 2008); Catanzano by Catanzano v. Dowling, 847 F. Supp. 1070 (W.D.N.Y 1994); Cospito v. Heckler, 742 F.2d 72 (3d Cir. 1984); Wash. Hosp. v. White, 889 F.2d 1294 (3d Cir. 1989)
7	1,861	7	commerce, interstate, clause, government, amendment	Garcia v. San Antonio Metro. Transit Auth., 469 U.S. 528 (1985); Gonzales v. Raich, 545 U.S. 1 (2005); United States v. Morrison, 529 U.S. 598 (2000); Camps Newfound/Owatonna, Inc. v. Town of Harrison, Me., 520 U.S. 564 (1997)
14	962	7	price, prices, milk, product, sales	Longview Refining Co. v. Shore, 554 F.2d 1006 (Temp. Emerg. Ct. App. 1997); *In re* Midw. Milk Monop. Litig., 510 F. Supp. 381 (W.D. Mo. 1981); Reynolds Indus. v. Mobil Oil Corp., 569 F. Supp. 716 (D. Mass. 1983); Pac. Supply Co-op v. Shell Oil Co., 697 F.2d 1084 (Temp. Emerg. Ct. App. 1982)
2	905	6	regulations, agency, preemption, government, insurance	Long v. U.S. Dept. of Just., 450 F. Supp. 2d 42 (D.D.C. 2006); United States v. Sunny Cove Citrus Ass'n, 854 F. Supp. 669 (E.D. Ca. 1994); Garrelts v. SmithKline Beecham Corp., 943 F. Supp. 1023 (N.D. Iowa 1996); Stainback v. Mabus, 671 F. Supp. 2d 126 (D.D.C. 2009)

3	1,430	6	rate, gas, commission, rates, ferc [F.E.R.C.]	Pac. Gas & Elec. Co. v. Lynch, 216 F. Supp. 2d 1016 (N.D. Ca. 2002); Town of Norwood v. FERC, 202 F.3d 392 (1st Cir. 2000); New York v. FERC, 535 U.S. 1 (2002); E. & J. Gallo Winery v. EnCana Corp., 503 F.3d 1027 (9th Cir. 2007)
5	735	6	cable, fcc [F.C.C.], television, speech, amendment	Satellite Broad. & Commc'ns Ass'n v. FCC, 275 F.3d 337 (4th Cir. 2001); Time Warner Ent. Co. v. United States, 211 F.3d 1313 (D.C. Cir. 2000); Turner Broad. Sys., Inc. v. FCC, 819 F. Supp. 32 (D.D.C. 1993); Time Warner Ent. Co. v. FCC, 93 F.3d 957 (D.C. Cir. 1996)
8	709	6	bank, loan, securities, shares, credit	CSX Corp. v. Children's Inv. Fund Mgmt. (UK) LLP, 562 F. Supp. 2d 511 (S.D.N.Y. 2008); CSX Corp. v. Children's Inv. Fund Mgmt. (UK) LLP, 654 F.3d 276 (2d Cir. 2011); *In re* WorldCom, Inc. Secs. Litig., 346 F. Supp. 2d 628 (S.D.N.Y. 2004)
6	817	5	labor, union, employer, employees, preemption	Sears, Roebuck & Co. v. San Diego Cnty. Dist. Council of Carpenters, 436 U.S. 180 (1978); Signatory Negot. Comm. v. Local 9, Int'l Union of Operating Eng'rs, 447 F. Supp. 1384 (D. Colo. 1978); Connell Const. Co. v. Plumbers and Steamfitters Local Union No. 100, 421 U.S. 616 (1975); Local 210, Laborers' Int'l Union v. Lab. Rels. Div. Assocd. Gen. Contractors of Am., 844 F.2d 69 (2d Cir. 1988)
10	769	5	commission, fcc [F.C.C.], access, telephone, agency	Cap. Network Sys., Inc. v. FCC, 3 F.3d 1526 (D.C. Cir. 1993); Pub. Serv. Comm'n v. FCC, 909 F.2d 1510 (D.C. Cir. 1990); Fones4All Corp. v. FCC, 550 F.3d 811 (9th Cir. 2008); AT&T Corp. v. Adventure Commc'n Tech., LLC, 207 F. Supp. 3d 962 (S.D. Iowa 2016)

13	1,308	5	commission, rate, rates, agency, exchange	New England Power Co. v. United States, 693 F.2d 239 (1st Cir. 1982); App. of FTC Line of Bus. Rep. Litig., 595 F.2d 685 (D.C. Cir. 1978); Am. Trucking Ass'ns v. Interstate Com. Comm'n, 656 F.2d 1115 (5th Cir. 1981); W. Coal Traffic League v. United States, 694 F.2d 378 (5th Cir. 1982)
11	575	4	patent, fda [F.D.A.], drug, drugs, loan	Jack Winter, Inc. v. Koratron Co., 375 F. Supp. 1 (N.D. Ca. 1974); *In re* Terazosin Hydrochloride Antitrust Litig., 335 F. Supp. 2d 1336 (S.D. Fla. 2004); Lexmark Int'l, Inc. v. Impression Prods., Inc., 816 F.3d 721 (Fed. Cir. 2016); Ark. Carpenters Health & Welfare Fund v. Bayer AG, 604 F.3d 98 (2d Cir. 2010)
9	193	2	tr [trial], dr [doctor], expert, testimony, price	United States v. Visa U.S.A., Inc., 163 F. Supp. 2d 322 (S.D.N.Y. 2001); FTC v. Staples, Inc., 190 F. Supp. 3d 100 (D.C. Cir. 2016); United States v. Am. Exp. Co., 88 F. Supp. 3d 143 (E.D.N.Y. 2015); FTC v. Sysco Corp., 113 F. Supp. 3d 1 (D.D.C. 2015)