ALGORITHMIC ANTITRUST

FELIX B. CHANG [*]
ERIN MCCABE [†]
ZHAOWEI REN [‡]
JAMES LEE [§]

Abstract

This paper illustrates how machine learning can disrupt legal scholarship through the algorithmic extraction and analysis of enormous volumes of data. Specifically, we utilize data from Harvard Law School's Caselaw Access Project (the "Caselaw Project") to model how courts tackle two thorny question in antitrust: the measure of market power and the balance between antitrust and regulation.

Unveiled in October 2018, the Caselaw Project has digitized all book-published case law in the U.S. It greatly expands legal access while also simplifying big data research. We have written computer code (Python script) to pull two groups of cases from the Caselaw Project: (i) roughly 36,000 federal cases containing the term "antitrust" and (ii) roughly 305,000 federal cases containing the term "regulation." Each corpus can be further refined—for instance, by extracting antirust cases that consider "market power" or regulation cases bearing the term "antitrust."

[*] Professor of Law and Co-Director, Corporate Law Center, University of Cincinnati College of Law. E-mail: felix.chang@uc.edu. We thank the Andrew W. Mellon Foundation for providing financial support for this project.

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

[‡] Software Development Engineer, Amazon Web Services (previously, Software Developer, Digital Scholarship Center, University of Cincinnati).

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

We then utilize a machine learning platform to perform topic modeling on each refined dataset through statistical algorithms. In this way, we create visualizations that (i) group terms within cases into topics, (ii) map the relationships among terms, and (iii) map the relationships among topics. We hope these visualizations can uncover pattens in antitrust cases while also introducing to legal scholars how machine learning can be applied to large corpuses of freely available data.

I. Introduction

Machine learning abounds in finance, policing, employment, politics, and health services,[1] but as a research technique, it is just gaining traction in legal academia.[2] This aversion to machine learning, despite its potential for processing large amounts of data, can be attributed to at least three factors. First, only a few repositories hold a corpus of easily extractable legal data—or, for common law jurisdictions, case law.[3] Second, even if data could be easily extracted, its interpretation is limited by modeling that can translate machine analysis into

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

[2] One exception is the emerging application of corpus linguistics to statutory interpretation to discern the ordinary meaning of language. *See* Stefan Th. Gries & Brian G. Slocum, *Ordinary Meaning and Corpus Linguistics*, 2017 B.Y.U. L. REV. 1417. Recently, BigML also started to provide machine learning services to academics. *See* BigML, https://bigml.com/ (last accessed Jan. 15, 2020).

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

intuitive visualizations.[4] Finally, legal scholars have been reluctant to employ algorithms that, despite utopian promises, amplify rather than eliminate human biases.[5]

Making use of recent technical advances, we have overcome two of those hurdles—data extraction and data interpretation—to reveal how algorithmic processing can transform legal research. In October 2018, Harvard Law School unveiled its Caselaw Access Project (the "Caselaw Project"), which had digitized all book-published U.S. case law between 1658 and 2018, some 40 million pages.[6] The Caselaw Project will disrupt legal research. By making freely available all published decisions in every U.S. jurisdiction, it threatens the Westlaw and Lexis paywalls, greatly expanding legal access for anyone with an Internet connection. Further, the Caselaw Project provides cases in a clean, digestible form, so users need not write application programming interfaces ("APIs") to extract data, which greatly simplifies big data research.[7]

We have built a machine learning platform that analyzes large volumes of data, primarily through the algorithmic construction of visualizations through topic modeling. Topic modeling is a way of representing the probability distribution of terms and their co-occurrence within a

[4] *See* Jason Chuang et al., *Interpretation and Trust: Designing Model-Driven Visualizations for Text Analysis, in* CHI '12: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (2012).

[5] After all, law scholars are among the most vocal critics of algorithmic processing. *See, e.g.*, Amanda Levendowski, *How Copyright Law Can Fix Artificial Intelligence's Implicit Bias Problem*, 93 Wash. L. Rev. 579 (2018); Jack Balkin, *The Three Laws of Robotics in the Age of Big Data*, 78 OHIO ST. L.J. 1217 (2017); Dan L. Burk, *Algorithmic Fair Use*, 86 U. CHI. L. REV. 283 (2019); Kiel Brennan-Marquez & Stephen E. Henderson, *Artificial Intelligence and Role Reversible Judgment*, 2 J. CRIM. L. & CRIMINOL. __ (2019).

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

[7] In fact, the Caselaw Project provides stock APIs for software developers to tinker with. *See id*. https://case.law/api/ (last accessed Sept. 15, 2019).

dataset.[8] Utilizing this platform, we analyze large numbers of federal antitrust cases up to 2018, which we extracted from the Caselaw Project, to see how federal courts tackle various concepts in antitrust.

This paper presents visualizations and preliminary results in two vexing areas of antitrust law: market power and the balance between antitrust and regulation. In antitrust, the measure of market power is fraught with controversy.[9] The prevailing paradigm—using market share as a proxy for market power—has been the target of fierce criticism.[10] Yet examinations of collusion and exclusion are seldom complete without market power analysis of the constituent markets. Another contested issue is how courts approach competition in regulated industries such as finance, telecommunications, and health care. Since the Supreme Court recalibrated the balance between antitrust and regulation in 2004 in *Trinko*,[11] academics have offered a flurry of

[8] *See* Chuang, *supra* note 2; Chuang et al., *Termite: Visualization Techniques for Assessing Textual Topic Models*, *in* AVI '12: Proceedings of the International Working Conference on Advanced Visual Interfaces (2012).

[9] *See, e.g.*, Louis Kaplow, *Why (Ever) Define Markets?*, 124 HARV. L. REV. 437, 440 (2010); Gregory J. Werden, *Why (Ever) Define Markets? An Answer to Professor Kaplow*, 78 ANTITRUST L.J. 729, 740 (2013).

[10] While market definition/market share presents only circumstantial evidence of market power, it has become the prevailing way of gauging market power. Direct evidence, such as of anticompetitive effects, if often too hard to come by.

[11] 540 U.S. 398 (2004). *See also Credit Suisse Sec. (USA) L.L.C. v. Billing*, 551 U.S. 264 (2007). Silver v. N.Y. Stock Exch., 373 U.S. 341, 357 (1963); Gordon v. New York Stock Exchange, Inc., 422 U.S. 659 (1975).

proposals to overhaul the role of antitrust.[12] Particularly in an era of regulatory abdication, scholars have looked to antitrust to step into the voids.[13]

We have chosen to start with market power and the antitrust-regulation balance in part because of these doctrinal ambiguities. More importantly, our method of research can help legal scholars become comfortable with algorithmic processing, by revealing how we are employing our own algorithms. 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,[14] our machine-generated visualizations can provide a fresh take on thousands of cases.

Our work sits at the intersection of several major questions in legal scholarship. One question is whether alternatives to Westlaw and Lexis can emerge to streamline big data research projects. Currently, data extraction through Westlaw and Lexis is cumbersome—researchers must negotiate limited licenses and then write elaborate APIs to pull data in batches. We also test

[12] *See* Maurer & Scotchmer, *supra* note 22; Adam Candeub, *Trinko and Re-Grounding the Refusal to Deal Doctrine*, 66 U. PITT. L. REV. 821 (2005); Brett Frischmann & Spencer Weber Waller, *Revitalizing Essential Facilities*, 75 ANTITRUST L.J. 1 (2008); Howard A. Shelanski, *The Case for Rebalancing Antitrust and Regulation*, 109 MICH. L. REV. 683 (2011).

[13] *See, eg.*, Samuel N. Weinstein, *Financial Regulation in the (Receding) Shadow of Antitrust*, 91 TEMP. L. REV. 447 (2019); Tim Wu, *Antitrust via Rulemaking: Competition Catalysis*, SSRN.

[14] *See* SAFIYA UMOJA NOBLE, ALGORITHMS OF OPPRESSION: HOW SEARCH ENGINES REINFORCE RACISM (2018); Nathan Newman, *How Big Data Enables Economic Harm to Consumers, Especially to Low-Income and Other Vulnerable Sectors of the Population*, 18 J. INTERNET L. 11 (2014); Frank Pasquale, *Typecastes: Big Data's Social Stratifications*, JOTWELL (March 18, 2015).

the limits of the capacity and ethics of machine learning. Even within digital humanities, machine learning techniques remain novel; within legal scholarship, they are nascent altogether.[15] To be sure, machine analysis has its restrictions—our models are only as good as our algorithms and the Caselaw Project's data. Additionally, big data projects can gloss over the nuance and context that legal analysis favors. Yet at a time when scholars are obsessed with the effects of algorithms upon law and society, our project turns machine learning upon legal scholarship itself, an inversion that produces fascinating results.

Section II of this paper introduces the Caselaw Project. Section III summarizes our methodology. Section IV presents our preliminary findings and hazards some inferences. While it is still premature to draw any firm conclusions from our visualizations, we end with these general inferences that both affirm and complicate previous antitrust research.

II. The Harvard Caselaw Access Project

It took over three years for the Caselaw Project to simply digitize all court decisions published in the 40,000 bound volumes in the Harvard Law School Library.[16] The 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

[15] For a debate on the promises and pitfalls of corpus linguistics for law, compare Thomas R. Lee & Stephen C. Mouritsen, *Judging Ordinary Meaning*, 127 YALE L.J. 788 (2018), with Carissa Byrne Hessick, *Corpus Linguistics and the Criminal Law*, 2017 BYU L. REV. 1503 (2017).

[16] *See About*, *supra* note 6.

Islands.[17] This text is presented in a machine readable format, which permits easy extraction for research projects. In fact, the Caselaw Project has shared APIs for public use.[18]

Notably, the Caselaw Project excludes cases published after June 2018 and cases not designated as officially published, such as some lower court decisions. The scope limitations also leave out unpublished trial documents, such as filings and exhibits. Nonetheless, these is enough data to compile rich models and graphs. For instance, the Caselaw Project enables searches for historical trends.[19]

III. Methodology

A. *Data and Access*

Data for this project was made available through the Caselaw Project, which contains 6.7 million unique cases (and over 1.7 million federal cases). Having applied for and obtained researcher access from the Caselaw Project, we gathered data by writing python-based calls to its API. The Caselaw Project's APIs feature tools that permit searching through all text in selected cases (as opposed to searches using tags or other metadata). We created two pools of cases: all federal cases with the word "antitrust," a total of approximately 36,000 cases; and all federal cases with the word "regulation," a total of approximately 305,000 cases.[20] Antitrust cases

[17] *Id.*

[18] *See supra* note 6.

[19] A simple search reveals that antitrust cases rose to a high of 4% of all federal cases in the 1980s. *See Historical Trends*, CASELAW ACCESS PROJECT, https://case.law/trends/ (search for "us: antitrust").

[20] At first glance, these numbers seemed very small to us, particularly the count of 36,000 for all federal antitrust cases. However, we attribute this to the Caselaw Project's corpus, which excludes unreported decisions. The Caselaw Project has a little over 1.7 million unique federal cases in its corpus, and a search in historical trends

bearing the term “market power” total 2,591, and regulation cases bearing the term “antitrust” total 7,308.

Manual assessment quickly becomes impracticable when examining a corpus as extensive as the Caselaw Project. Thus, the application of machine learning provides a more manageable approach. We use algorithms to sort through each case’s natural language and produce models of topics based on the clustering of frequently recurring words. These are statistical processes that have been developed and refined in other fields that have experimented with text visualizations.[21] This computational approach to language allows us to see certain trends through topics generated from the case law documents’ own semantic and syntactic structures, themselves rather than applying human data and metadata structures to a dataset. Put differently, machine learning has the potential to provide a neutral way of ordering this volume of case law, devoid of human—and doctrinal—preconceptions.

B. *Modeling and Visualizations*

reveals that antitrust cases have comprised a low of about 0.1% to a high of almost 4% of all federal cases, with a median roughly short of 2% (or about 34,000 cases). Comparisons to Lexis are not far off. For instance, we have extracted approximately 16,664 federal cases with “antitrust” and “regulation” from the Caselaw Project. Excluding unreported decisions, the total for cases with “antitrust” and “regulat!” (root expander) is 20,770 in Lexis. Curiously, however, Westlaw results using these search terms are about double the size of Lexis results. We have been communicating with both Westlaw and Lexis to reconcile the difference. Nonetheless, we have more than a robust sampling for federal antitrust cases.

[21] Chuang et al., 2012; Sievert Shirley;

Using Elasticsearch (a full-text search and analytics engine)[22] and the python Gensim package,[23] we developed a web-based platform. The platform performs topic modeling by using unsupervised machine learning clustering algorithms, specifically latent Dirichlet allocation ("LDA"), to sift through cases. LDA models are generated based on the distribution of latent topics in a document and the distribution of words in those topics.[24] Each topic is constructed based on a probability distribution of words.[25] For instance, one topic might feature the term "market" with high probability, whereas its appearance in another topic will not be as strong. Similarly, one document might have a high presence of topic 1, pertaining to procedural and evidentiary matters, whereas that same topic only features faintly in another document. In this project, we have defined a "document" as an individual case from our dataset.

As with any empirical project based on copious amounts of data, the term relevance and topic modeling are subject to margins of error, which we affectionately call the "wobble." We

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

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

[24] D.M. Blei et al., *Latent Dirichlet allocation*,3 J. Machine Learning Research 993 (2003).

[25] *See* Chuang et al., *supra* note 4:

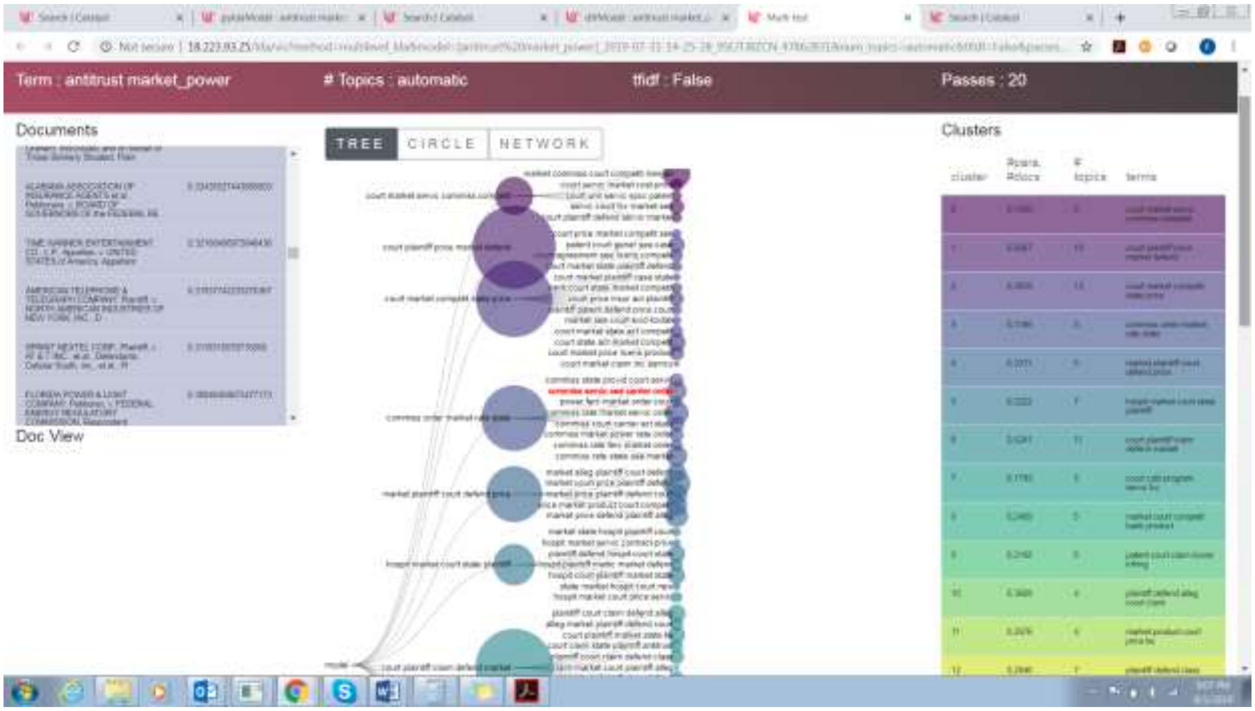
Given as input a desired number of topics K and a set of documents containing words from a vocabulary V, LDA derives K topics β_k, each a multinomial distribution over words V. For example, a "physics" topic may contain

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

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 (multilevel).

The models provide visualizations of cases grouped by recurring terms, depicting both the relationships among terms and the relationships among groups of cases. We rely on three types of visualizations, all built around topic modeling. The remainder of this Subsection explores all three types, using antitrust market power cases as the dataset.

First, multilevel (or model-of-model) visualizations provide a hierarchical view of topics and topic clusters in three different formats—tree, circle, and network (see Figure 1a–d).



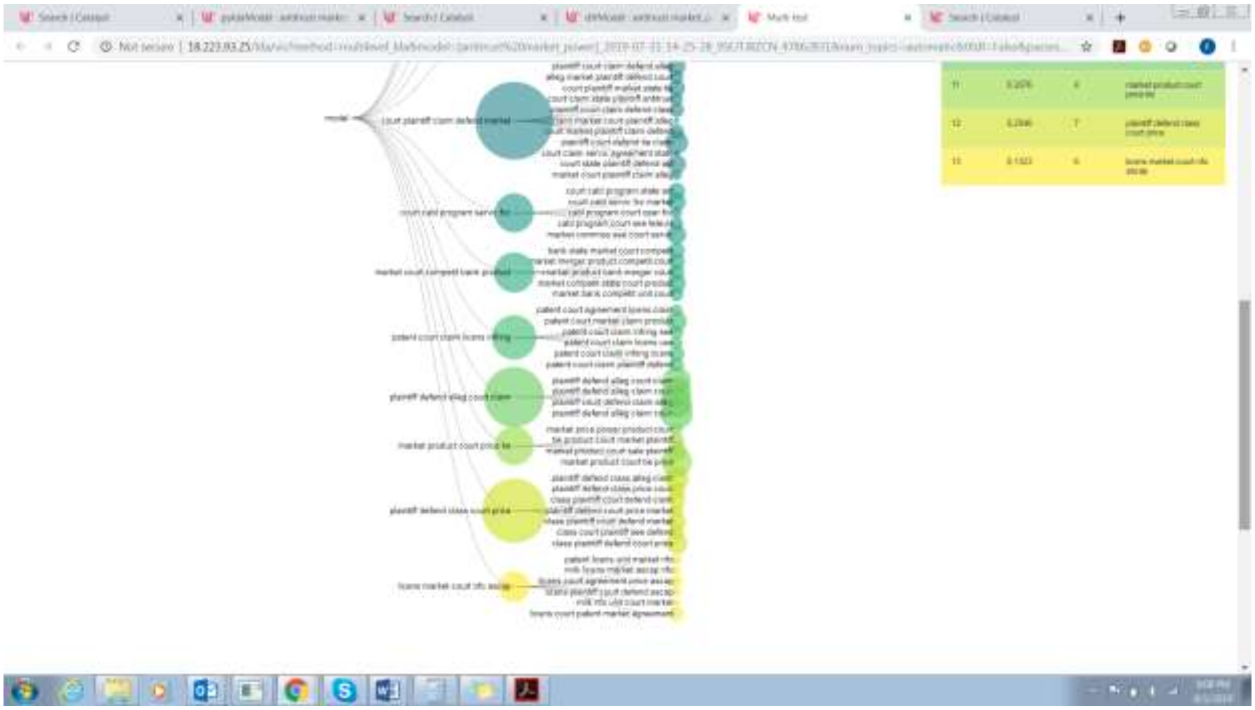
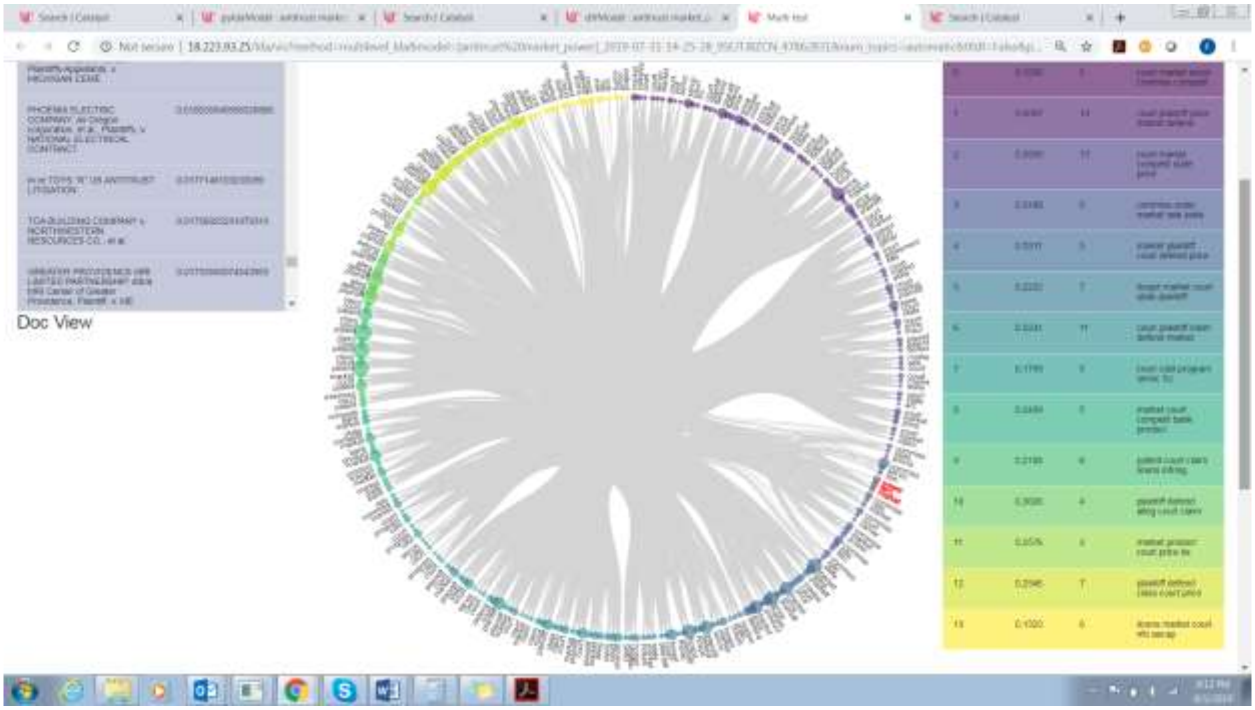

model
court plaintiff claim defend market
court calif program servic be
market court compet bank product
patent court claim licens infring
plaintiff defend alleg court claim
market product court price be
plaintiff defend class court price
licens market court nfo assign

Figure 1a: Multilevel Visualization of Market Power Cases in Tree Format

In the tree format above, the smaller nodes on the right represent topics (i.e., machine-grouped terms “commiss[ion],” “servic[e,]” “see,” “carrier,” and “order”), while the larger nodes represent clusters of topics (e.g., a cluster with “court,” “market,” “servic[e,]” “commiss[ion],” and “compet[ition]”). The size of each cluster node or topic node represents the significance of the cluster or topic to the overall corpus. The right-hand bar shows the number of topics within each cluster (thereby functioning as a proxy for the cluster’s diversity), and the left-hand bar lists the top cases in each topic.



Plaintiffs-Appellants, v. MICHIGAN CEME

PHOENIX ELECTRIC COMPANY, an Oregon corporation, et al., Plaintiffs, v. NATIONAL ELECTRICAL CONTRACT [illegible]

In re TOYS "R" US ANTITRUST LITIGATION [illegible]

TCA BUILDING COMPANY v. NORTHWESTERN RESOURCES CO., et al. [illegible]

GREATER PROVIDENCE MRI LIMITED PARTNERSHIP d/b/a MRI Center of Greater Providence, Plaintiff, v. ME [illegible]

Figure 1b: Multilevel Visualization of Market Power Cases in Circle Format

Circle view presents the same information, but in a format that more clearly conveys the topics where each word appears. For example, Figure 1c shows the recurrence of the term "market" within all topics.

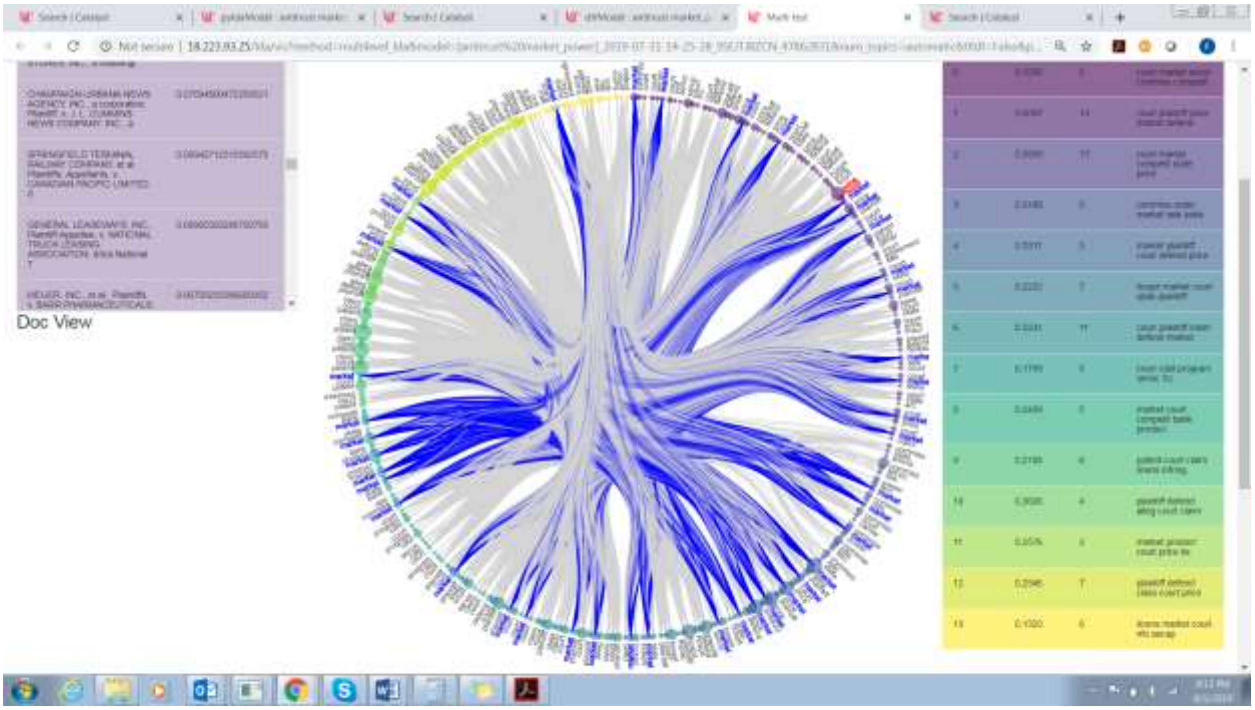


Figure 1c: Multilevel Visualization Showing the Recurrence of the Term “Market”

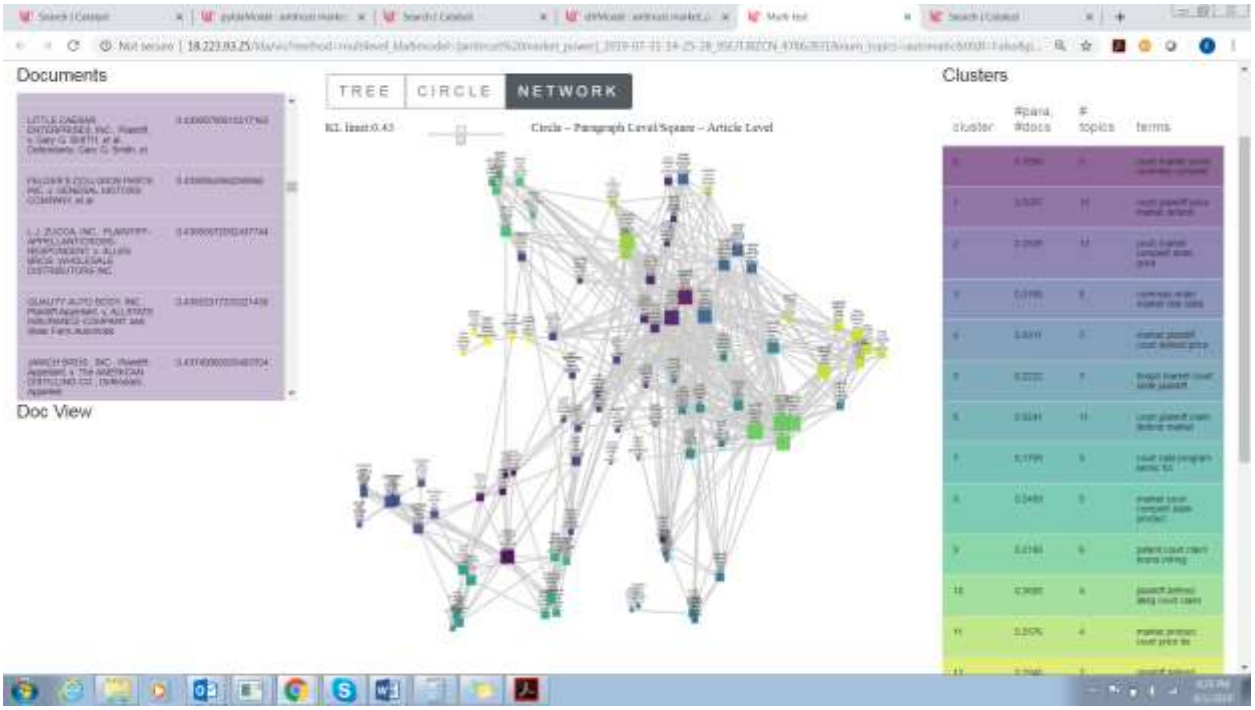
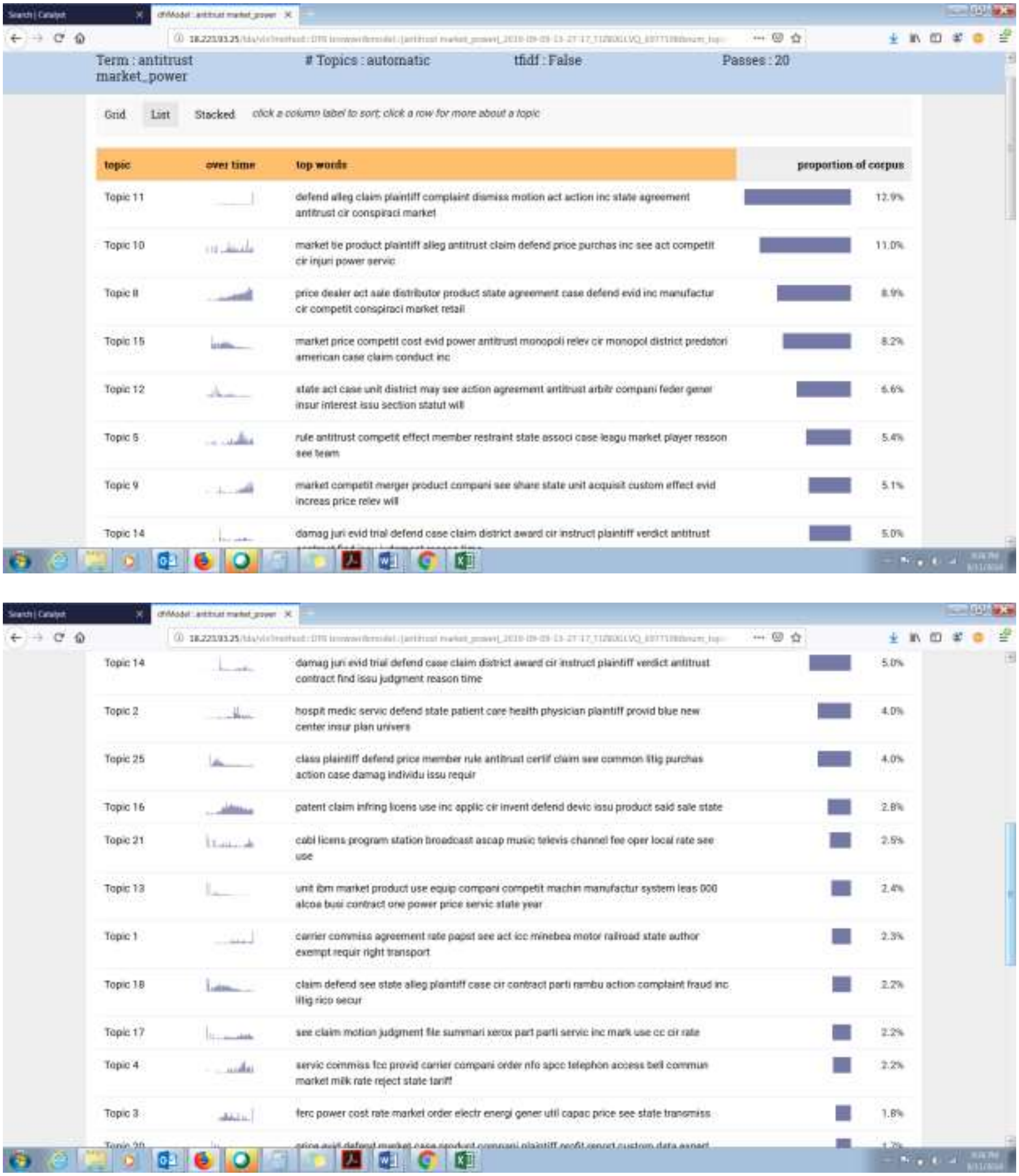


Figure 1d: Multilevel Visualization of Market Power Cases in Network Format

Second, topic browser visualizations organize cases into topics, enabling detailed analyses of where (i.e., in what topics) certain terms recur (see Figure 2a–c).

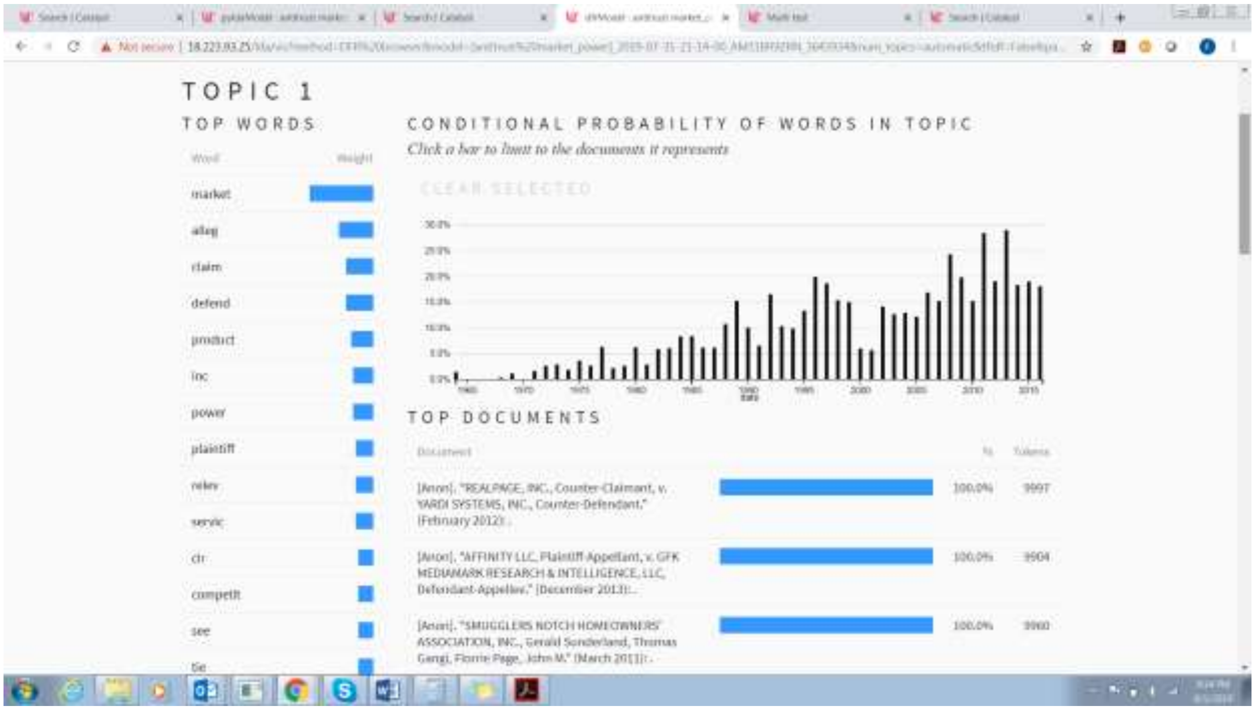


Term : antitrust market_power # Topics : automatic tfidf : False Passes : 20

Grid List Stacked *click a column label to sort; click a row for more about a topic*

topic	over time	top words	proportion of corpus
Topic 11		defend alleg claim plaintiff complaint dismiss motion act action inc state agreement antitrust cir conspiraci market	12.9%
Topic 10		market tie product plaintiff alleg antitrust claim defend price purchas inc see act competit cir injuri power servic	11.0%
Topic 8		price dealer act sale distributor product state agreement case defend evid inc manufactur cir competit conspiraci market retail	8.9%
Topic 15		market price competit cost evid power antitrust monopoli relev cir monopol district predatori american case claim conduct inc	8.2%
Topic 12		state act case unit district may see action agreement antitrust arbitr compani feder gener insur interest issu section statut will	6.6%
Topic 5		rule antitrust competit effect member restraint state associ case leagu market player reason see team	5.4%
Topic 9		market competit merger product compani see share state unit acquisit custom effect evid increas price relev will	5.1%
Topic 14		damag juri evid trial defend case claim district award cir instruct plaintiff verdict antitrust contract find issu judgment reason time	5.0%
Topic 2		hospit medic servic defend state patient care health physician plaintiff provid blue new center insur plan univers	4.0%
Topic 25		class plaintiff defend price member rule antitrust certif claim see common litig purchas action case damag individu issu requir	4.0%
Topic 16		patent claim infring licens use inc applic cir invent defend devic issu product said sale state	2.8%
Topic 21		cabl licens program station broadcast ascap music televis channel fee oper local rate see use	2.5%
Topic 13		unit ibm market product use equip compani competit machin manufactur system leas 000 alcoa busi contract one power price servic state year	2.4%
Topic 1		carrier commiss agreement rate papst see act icc minebea motor railroad state author exempt requir right transport	2.3%
Topic 18		claim defend see state alleg plaintiff case cir contract parti rambu action complaint fraud inc litig rico secur	2.2%
Topic 17		see claim motion judgment file summari xerox part parti servic inc mark use cc cir rate	2.2%
Topic 4		servic commiss fcc provid carrier compani order nfo spcc telephon access bell commun market milk rate reject state tariff	2.2%
Topic 3		ferc power cost rate market order electr energi gener util capac price see state transmiss	1.8%
Topic 20		price evid defend market case product compani plaintiff profit report custom data expert	1.7%

Figure 2a: Topic Browser Visualization of Market Power Cases in List Format



TOPIC 1

TOP WORDS

Word	Weight
market	
alleg	
claim	
defend	
product	
inc	
power	
plaintiff	
relev	
servic	
ctr	
competit	
see	
tie	

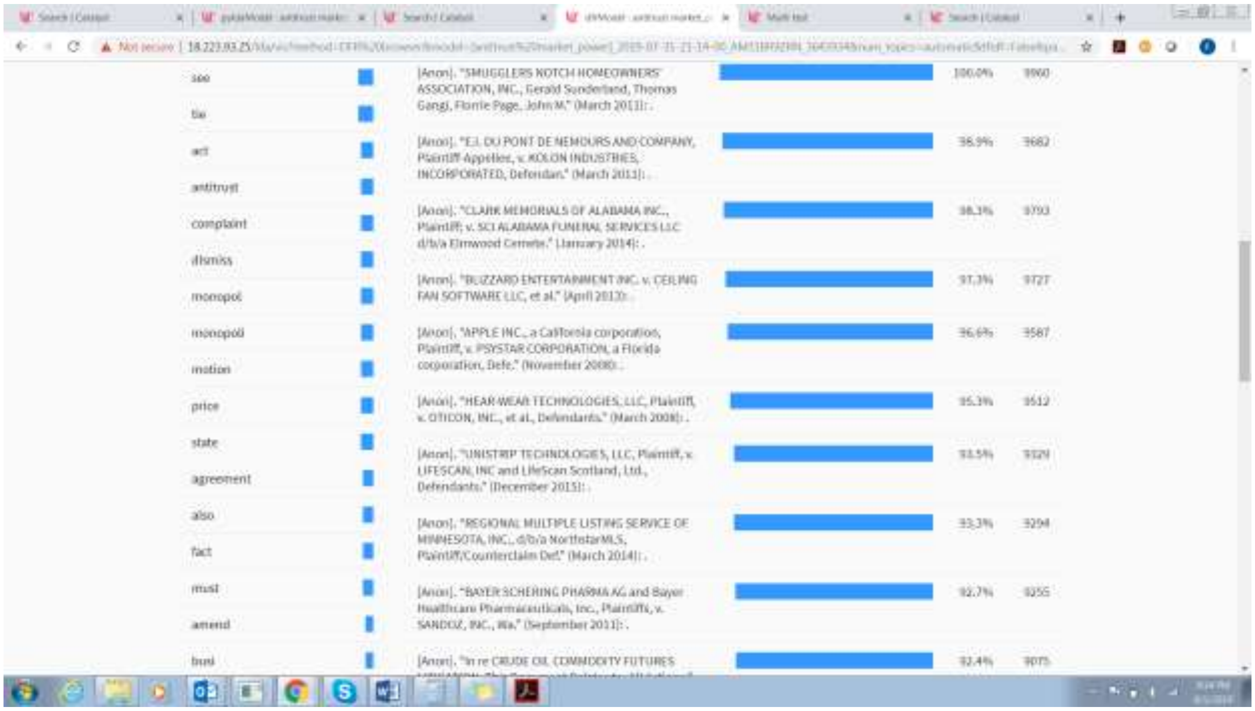
CONDITIONAL PROBABILITY OF WORDS IN TOPIC

Click a bar to limit to the documents it represents

CLEAR SELECTED

TOP DOCUMENTS

Document	%	Tokens
[Anon]. "REALPAGE, INC., Counter-Claimant, v. YARDI SYSTEMS, INC., Counter-Defendant." (February 2012): .	100.0%	9997
[Anon]. "AFFINITY LLC, Plaintiff-Appellant, v. GFK MEDIAMARK RESEARCH & INTELLIGENCE, LLC, Defendant-Appellee." (December 2013): .	100.0%	9904
[Anon]. "SMUGGLERS NOTCH HOMEOWNERS' ASSOCIATION, INC., Gerald Sunderland, Thomas Gangi, Florrie Page, John M." (March 2011): .	100.0%	9960

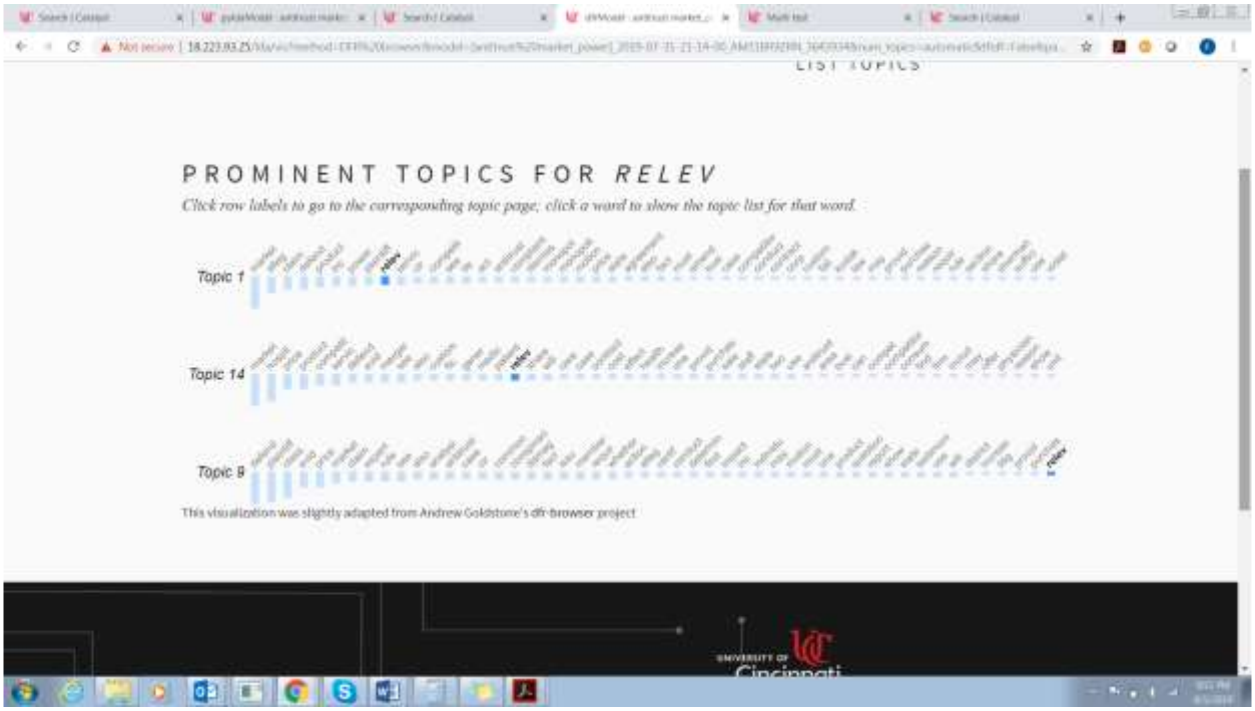


Word	Weight
see	
tie	
act	
antitrust	
complaint	
dismiss	
monopol	
monopoli	
motion	
price	
state	
agreement	
also	
fact	
must	
amend	
busi	

Document	%	Tokens
[Anon]. "SMUGGLERS NOTCH HOMEOWNERS' ASSOCIATION, INC., Gerald Sunderland, Thomas Gangi, Florrie Page, John M." (March 2011): .	100.0%	9960
[Anon]. "E.I. DU PONT DE NEMOURS AND COMPANY, Plaintiff-Appellee, v. KOLON INDUSTRIES, INCORPORATED, Defendan." (March 2011): .	98.9%	9682
[Anon]. "CLARK MEMORIALS OF ALABAMA INC., Plaintiff; v. SCI ALABAMA FUNERAL SERVICES LLC d/b/a Elmwood Cemete." (January 2014): .	98.3%	9793
[Anon]. "BLIZZARD ENTERTAINMENT INC. v. CEILING FAN SOFTWARE LLC, et al." (April 2012): .	97.3%	9727
[Anon]. "APPLE INC., a California corporation, Plaintiff, v. PSYSTAR CORPORATION, a Florida corporation, Defe." (November 2008): .	96.6%	9587
[Anon]. "HEAR-WEAR TECHNOLOGIES, LLC, Plaintiff, v. OTICON, INC., et al., Defendants." (March 2008): .	95.3%	9512
[Anon]. "UNISTRIP TECHNOLOGIES, LLC, Plaintiff, v. LIFESCAN, INC and LifeScan Scotland, Ltd., Defendants." (December 2015): .	93.5%	9329
[Anon]. "REGIONAL MULTIPLE LISTING SERVICE OF MINNESOTA, INC., d/b/a NorthstarMLS, Plaintiff/Counterclaim Def." (March 2014): .	93.3%	9294
[Anon]. "BAYER SCHERING PHARMA AG and Bayer Healthcare Pharmaceuticals, Inc., Plaintiffs, v. SANDOZ, INC., Wa." (September 2011): .	92.7%	9255
[Anon]. "In re CRUDE OIL COMMODITY FUTURES	92.4%	9075

Figure 2b: Breakdown of Terms and Cases within a Topic in Topic Browser View

Both the overview in Figure 2a and the topic in Figure 2b provide histograms showing the time periods when certain topics were more prevalent. Clicking on each topic also brings out the topic's top terms and cases (Figure 2b).



PROMINENT TOPICS FOR *RELEV*

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

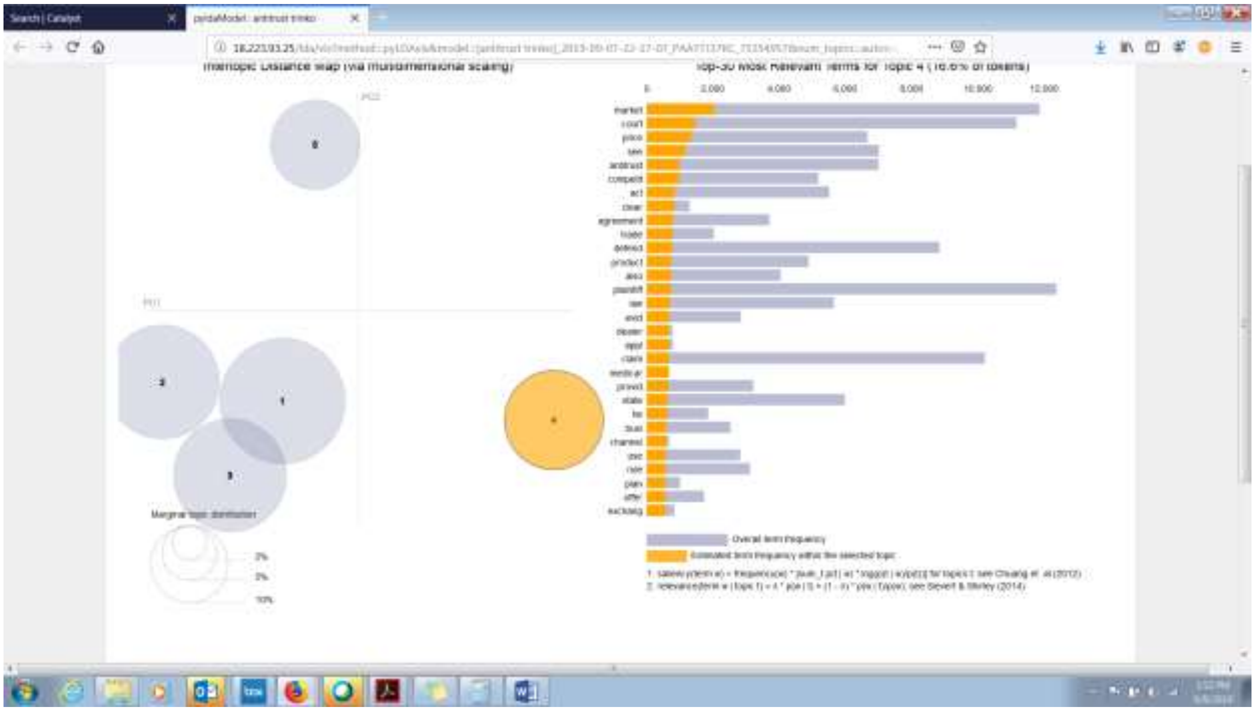
This visualization was slightly adapted from Andrew Goldstone's dfr-browser project

Figure 2c: Recurrence of the Term "Relev" in Topic Browser View

Clicking on each term pulls up the topics where the term appears.[26] For instance, Figure 2c shows that "relev" (as in relevant market, which would come up in market definition) appears in only three topics—a slightly surprising result in a corpus of cases dealing with market power.

[26] The topic browser visualization is adapted from Andrew Goldstone's dfr-browser project. *See* https://agoldst.github.io/dfr-browser/.

Third, python-based LDA visualizations ("pyLDAvis") depict the distance between topics (see Figure 3).[27] Additionally, the size of each topic bubble represents the weight of that topic. When a topic is highlighted, the platform pulls up the top probable words contained in that topic.[28]



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

[27] pyLDAvis is adapted from package led by Ben Mabey. *See* https://pyldavis.readthedocs.io/en/latest/index.html.

[28] For a mathematical expression of one of the key concepts in this statistical analysis, see Carson Sievert & Kenneth E. Shirley, *LDA vis: A Method for Visualizing and Interpreting Topics*, *in* PROCEEDINGS OF THE WORKSHOP ON INTERACTIVE LANGUAGE LEARNING, VISUALIZATION, AND INTERFACES 63, 66 (Jason Chuang et al. eds. 2014). Here The probability of any term within a topic is its *relevance* within that topic. Relevance can be expressed as

$$r(w,k) \mid = \lambda \log(\varphi_{kw}) + (1 - \lambda) \log (\varphi_{kw} / p_w),$$

where λ is the weight of the probability of term *w* under topic k relative to its lift.

Figure 3: pyLDAvis View of Antitrust Cases Containing "Trinko"

Figure 3 shows that our algorithms have sorted antitrust cases with the word "Trinko" into four topics.[29] In the screen shot, topic 4 is highlighted, bringing up its top terms. With pyLDAvis and the other visualizations, the platform user can set the number of topics manually. Here, the model was sorted into five topic bubbles.

Two additional points are notable. First, generic words such as "court," "see," "claim," and "plaintiff" are prevalent in the initial results. Although their presence renders the topics more generic, their appearance validates our machine learning because antitrust cases are replete with these words—words that algorithms are not trained to filter out.[30] We can refine the results by excluding generic words from the visualizations.[31] Second, these three types of visualizations are different than Word2Vec, which has dominated corpus linguistics and is the visualization of choice for legal scholars so far.[32] From a methodological perspective, our project therefore pushes machine learning in legal scholarship beyond word-level analysis, by building topic and even meta-topic models.

Finally, we have begun to read the top cases within each topic to see how courts think through market power. For example, we are reviewing cases within the three topics highlighted in Figure 2c above where "relev[ant]" was a top word; then we review cases in other topics,

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

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

[31] At this point, the platform can only filter out up to nine terms.

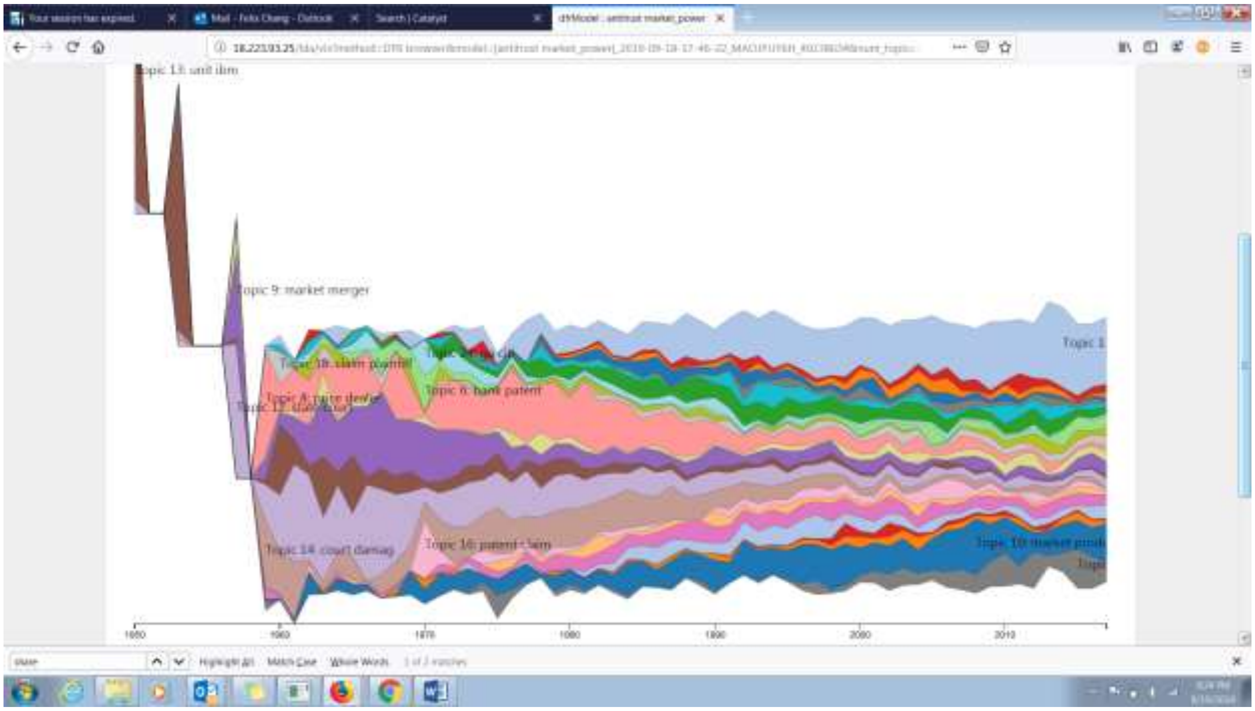
[32] For a description of Word2Vec, see Levendowski, *supra* note 5.

where “relev[ant]” had a lower probability distribution, presumably because the relevant product and geographic markets were not defined. (In each topic, cases are ordered by the probability score of that topic’s appearance in the case.)

IV. Inferences and Preliminary Observations

This section discusses some preliminary observations.

First, through various histograms, we can immediately spot several macrotrends, such as how the nature of market power cases has changed through the decades (see Figure 4).



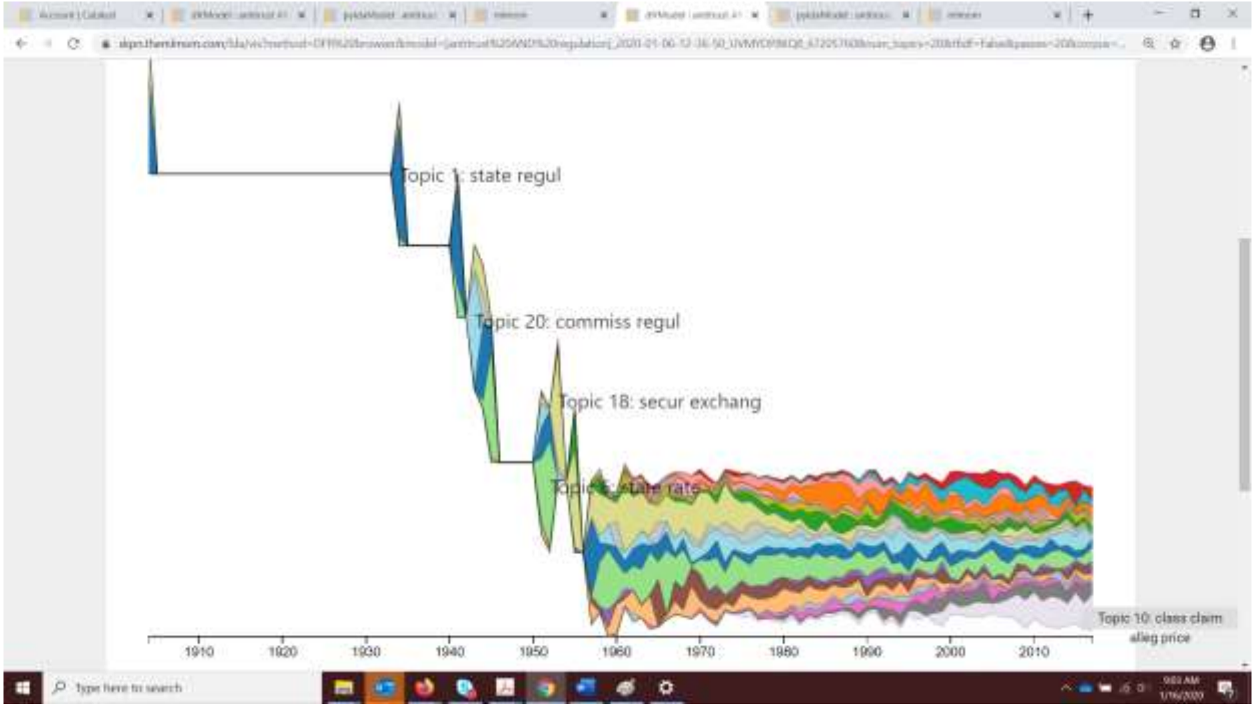
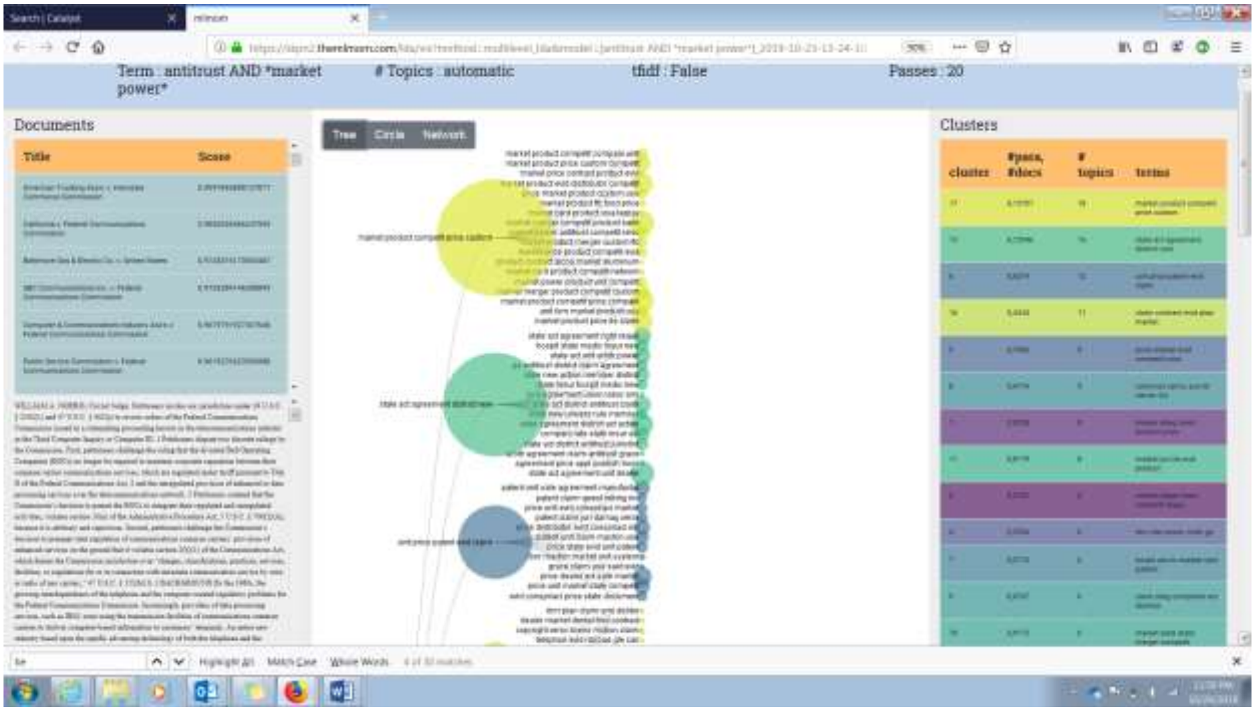


Figure 4: Topic Browser Stacked View with Histogram of Market Power (top) and Regulation (bottom) Cases

Starting in the late 1950s, there was an explosion of market power cases. Initially these cases were concentrated in manufacturing sectors and pertained to mergers, with *Alcoa* and market definition/market share featuring prominently in the analyses. Over time, the types of cases diversified, straddling various industries and types of horizontal and vertical conduct. Most notably, cases pertaining to the Interstate Commerce Commission ("ICC") were supplanted by telecommunications cases. 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. The dwindling of ICC cases is consistent with the 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.* [33]

Other macrotrends include an explosion of antitrust class actions starting around 1995 as well as the decline of certain topics over time, such as bank merger cases.[34] In addition, market power cases fall into a few large buckets: patent cases, health care cases, telecommunications cases, tying cases, banking and financial cases, and cases delving deeply into civil and evidentiary procedure (see Figure 5).



Term : antitrust AND *market power*

Topics : automatic

tfidf : False

Passes : 20

Documents

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

[34] Further research would have to be done, but the decline of bank merger cases might be attributable to the globalization of finance, which would push toward larger relevant geographic markets, thereby diminishing findings of market power.

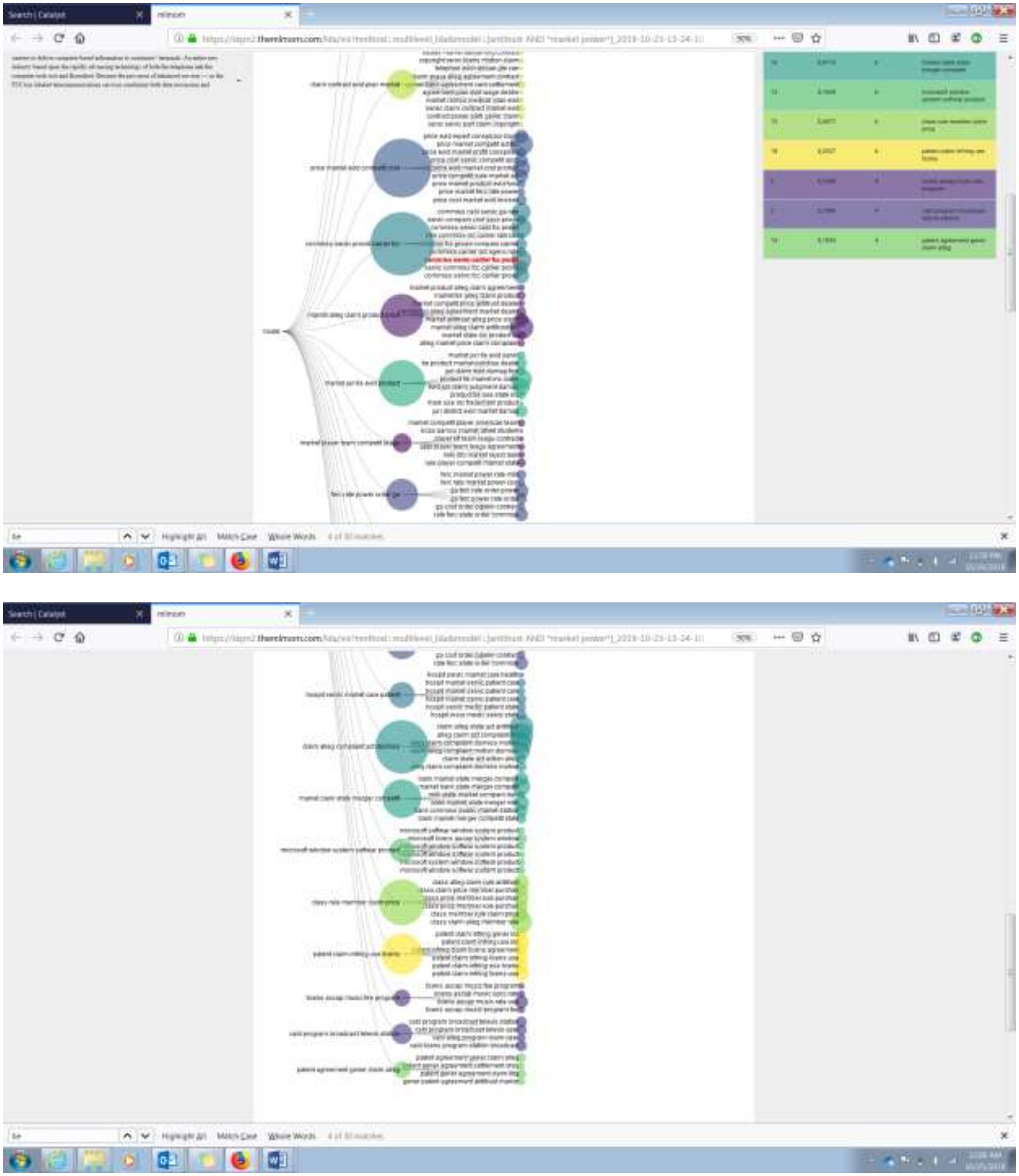


Figure 5: Multilevel Visualization of Market Power Cases

Here, the prevalence of the terms such as “agreement” (see Figure 6), “conspirac[y/ies],” and “contract” confirms the emphasis that U.S. antitrust laws places upon horizontal conduct. Because of per se treatment of conspiracies, plaintiffs strive to plead some sort of agreement among competitors, even if such evidence can be difficult to come by.

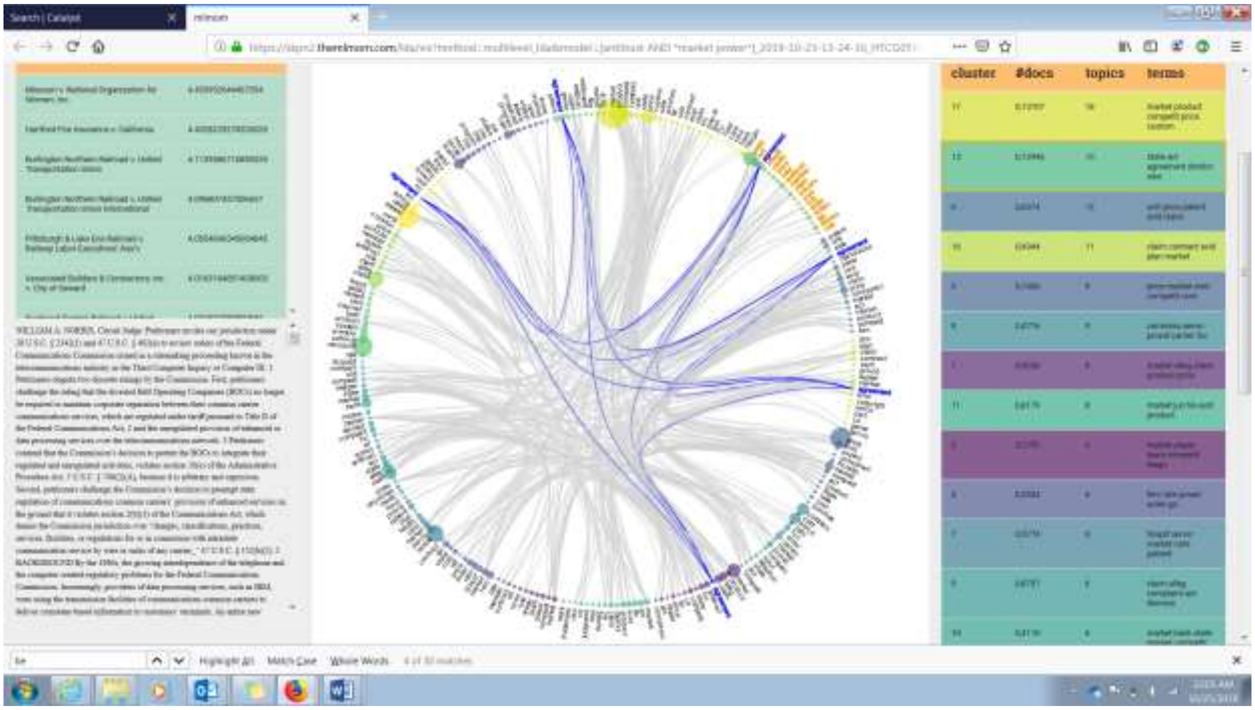
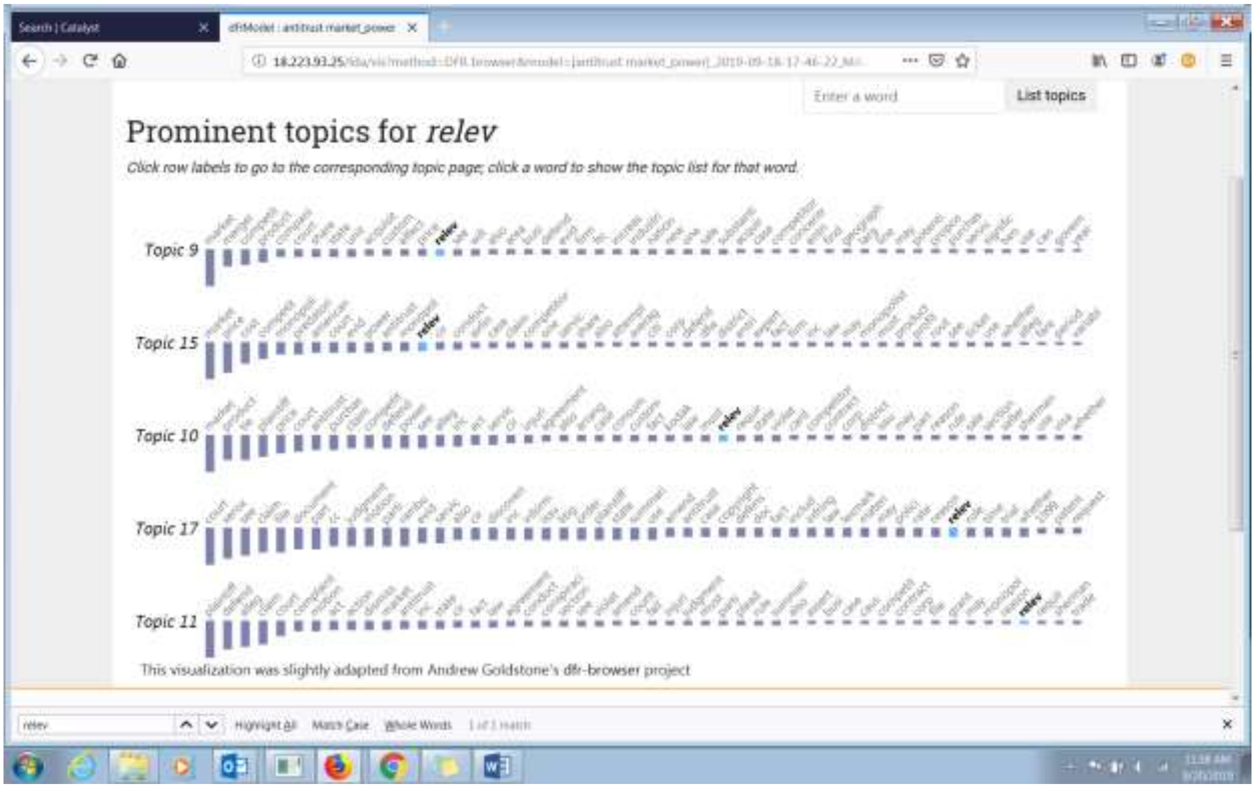


Figure 6: Multilevel Circle View with the Term “Agreement” Highlighted

The *second observation*, however, is that our visualizations present some challenges for drawing inferences. For instance, topic browser visualizations suggest that courts do not frequently define the relevant market, since the term “relev[ant]” does not appear across even half of the topics in market power cases (see Figure 7).



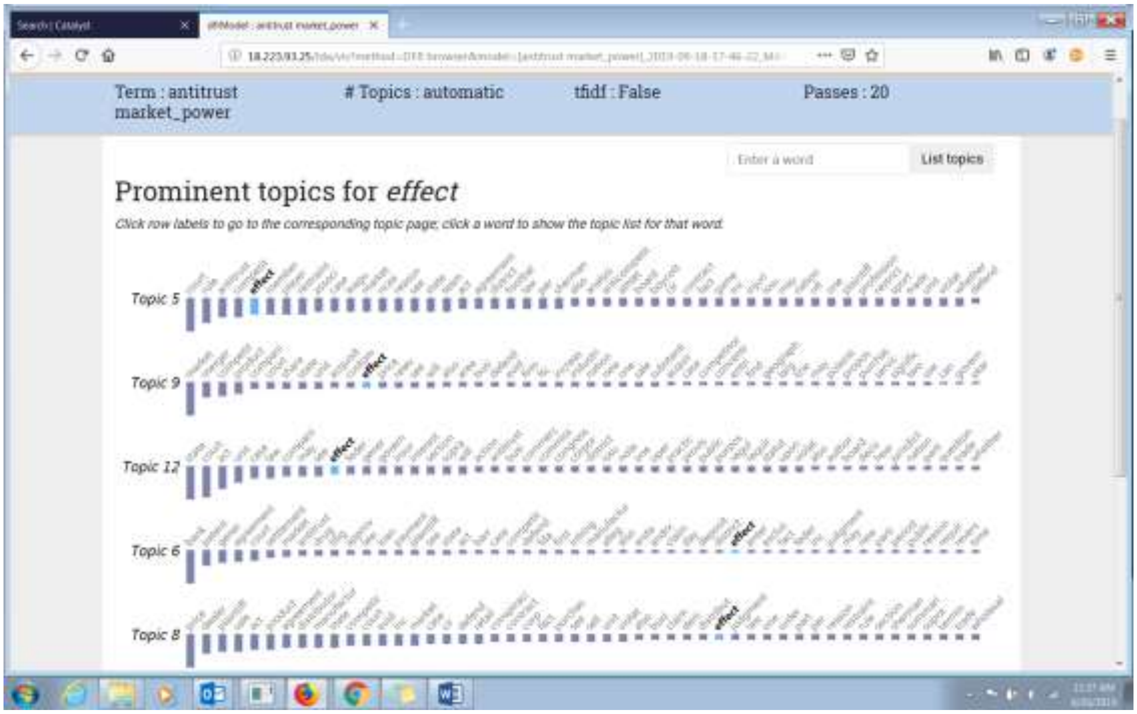
Prominent topics for *relev*

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

Figure 7: Topic Browser View of Topics Containing "Relev[ant]"

We might reasonably attribute this to two possibilities: either a court has accepted one party's market definition, or a court directly finds market power because there is evidence of anticompetitive effects. Yet "effect" also does not appear across many topics (see Figure 8), which is hardly surprising, since anticompetitive effects are difficult enough for economists to measure and even harder for courts to articulate. Significantly, the terms "relev[ant]" and "effect" do not overlap in topics, which suggest that courts may be using them as alternative proxies for market power.



Term : antitrust market_power

Topics : automatic

tfidf : False

Passes : 20

Prominent topics for *effect*

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

Figure 8: Topic Browser View of Topics Containing "Effect"

As we read the cases in the topics, however, we see that these inferences must be cautiously drawn. For instance, even within topics where "relev" is not highlighted as among the words (each topic lists approximately 50 top words), we find that courts often do take up the relevant product market, even if in cursory form. There simply may have been 50 other words that show much more frequently in the topic than "relev."[35] Yet we should not automatically

[35] *See* Chuang et al., *supra* note 5 ("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.").

filter out too many generic words. Algorithmically removing too many such terms strips away the benefit of the machine discerning relationships among terms that scholars might gloss over.[36]

Third, these visualizations nonetheless raise interesting questions about the way we read cases and understand precedent. In each of the datasets, case names hardly ever show up as top terms. *Lorain Journal*,[37] *Alcoa*,[38] and *DuPont*[39] do not appear as terms in the market power cases (even though *Microsoft*[40] shows up more frequently).[41] This suggests that courts are haphazard in their approach to market power. Surprisingly, the cases that appear as top results within each topic tend to be infrequently cited in legal scholarship—they are not precedent-setting, though they can be heavily cited by practitioners' manuals or by other courts within a particular circuit.

[36] However, we should note that the machine visualizations may highlight terms whose meanings have strong associations for readers in certain contexts. As a more precise example, in the regulation cases, there are two topics where "immunity" figures prominently as a recurring term. In reviewing the topic cases within those topics, we discover that these are actually Parker immunity cases concerning antitrust immunity for state action, as opposed to antitrust immunity by way of the presence or absence of an antitrust savings clause.

[37] Lorain Journal Co. v. U.S., 342 U.S. 143 (1951).

[38] U.S. v. Alcoa, 148 F.2d 416 (2d Cir. 1945).

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

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

[41] In the antitrust-regulation cases, *Trinko* does not appear as a top term. Yet we can confirm that this case is picked up in the topic modeling, because there is a "bibliography" function on the platform that lists all the cases. This may simply be because *Trinko* is still relatively recent and has not been cited by other cases incorporated into the modeling.

V. Conclusion

There is still much to be done with our platform and visualizations. Looking ahead, we plan to improve the platform's capability to eliminate more generic words. As this happens, the visualizations will be more informative, and the cases will be grouped more accurately. Of course, we must exclude terms with care, lest we comprise the function of uncovering patters that the machine's algorithms illuminate. Ultimately, we see our project as an initial step in the use of algorithmic processing in legal research, especially as a complement to commercial databases.