BEYOND TRANSSUBSTANTIVITY

*Jonah B. Gelbach**

This Article uses a massive collection of data to document, for the first time, the interplay between the substantive subject areas, the intensity, and the procedural complexity of federal civil litigation. The results indicate that like substance, intensity and complexity are pivotal features of litigation. These findings suggest that what might be termed "our uniformity"—the de jure uniform applicability of the Federal Rules of Civil Procedure and other procedural law—should be understood as a broader phenomenon than the traditional focus on transsubstantivity.

The Article documents extensive variation in measures of intensity and complexity among cases sharing the same broad substantive legal subject matter, such that intensity and complexity may vary more importantly within than across substantive areas. For example, it is true that patent cases have comparatively high average intensity as measured by their numbers of docket entries, but there are also plenty of patent cases that terminate without having enormous docket activity. It is also true that patent cases are much more likely than, say, contract cases to be extremely intense, but because there are so many more contract than patent cases, my data have nearly twice as many highly intense contract cases as highly intense patent cases. The Article's twin key conclusions are, thus: (1) even for cases in areas thought to be especially intense or complex, there are cases with both high and low intensity and complexity, and (2) intensity and complexity are transsubstantive—e.g., there are highly intense and highly complex cases in most substantive areas of litigation, not just the usual suspects such as patents, antitrust, or securities.

The Article closes by treading gingerly into normative waters. One unavoidable consequence of our uniformity with respect to formal procedural rules is that judges exercise enormous case-level discretion. Further, in part due to legislative forays in the securities and patent arenas, we have some degree of substance-specificity in our procedure. The Article suggests considering formal procedural tracking as an alternative to each of these

* Herman F. Selvin Professor of Law, University of California at Berkeley. I thank Andrew Baker, Rochelle Dreyfuss, David Freeman Engstrom, Nora Freeman Engstrom, Daniel Hemel, Deborah Hensler, Michael Lissner, Troy McKenzie, Roger Michalski, Austin Peters, Jonathan Petkun, Gwyneth Shaw, Polk Wagner, and workshop participants at NYU, UConn, and the 2022 Civil Procedure Workshop for discussions/comments; Andrew Baker for help constructing maps; Thomson Reuters staff for data help; Yale Law's Oscar M. Ruebhausen Fund, Bill Eskridge, and Robert Post for generous assistance with data acquisition. Over the years I have benefitted enormously from many conversations about American federal civil procedure with Steve Burbank and Tobias Barrington Wolff, without which this Article would be the worse. This paper supersedes previously circulated material appearing in never-published draft form under the title, "The Disuniform Distribution of Litigation: An Amuse Bouche," which I presented at the University of Pennsylvania Law Review symposium and festschrift for the incomparable Stephen B. Burbank in February 2021.

approaches. That approach, which has been suggested in the past and is used in some states and other countries, might allow us to break out of some of our ossified debates about matters such as the rancor set off by the pleading revolution the Supreme Court effected a decade and a half ago in Twombly and Iqbal. If the plausibility standard is here to stay because of its role in limiting intense and complex litigation, perhaps it could be revisited in simpler cases that don't pose the challenges the Supreme Court first flagged in Twombly. Whether this is possible depends on our capacity to identify cases' likely intensity and/or complexity early in the litigation life cycle, which is a topic beyond the scope of this Article. Still, we ought to consider the possibility of a certification process for intense, complex cases, like the one we have for class actions, so that procedure might be adjusted where doing so makes sense.

The Article's contributions include its use of docket-level data on more than 500,000 cases whose dockets could be followed for at least seven years. In addition, the Article offers a novel approach to measuring procedural complexity by showing how links between entries in docket reports may be viewed as mathematical networks.

Introduction

Uniformity is one of the foundational principles of federal civil procedure. Rule 1 declares: "These rules govern the procedure in *all* civil actions and proceedings in the United States district courts."[1] Nowadays, observers tend to regard the interesting aspect of this uniformity as its *transsubstantivity*—the idea that the Rules apply uniformly across substantive areas of the law, whether cases involve tort, contract, federal statutory claims brought under antitrust law or § 1983, and so on.[2] Although the habit is understandable given the earlier Anglo-American history of substance-specific forms of action, this Article demonstrates that viewing uniformity only or even primarily through the lens of transsubstantivity misses empirically important phenomena.

First, there is wild variation across cases in the *intensity* of litigation—i.e., how much litigation activity cases entail. This is true even outside the realm of multidistrict litigation (MDL) and other consolidated actions, which I exclude from my empirical analysis.

Second, there is significant variation across cases in their *complexity*—the degree to which cases depart in their structure from the simple "*P versus D*" model of litigation typically taught in first-year civil procedure courses.[3] In the real world, cases are messier and richer than that, with numerous parties engaged in stipulating, disputing, and hashing out both substantive and procedural issues with conflicting and/or coordinating behavior, across different moments in a litigation.[4]

Both intensity and complexity are familiar to our civil justice system's observers and participants. We are used to discussions of the complexity of class actions, MDLs, public interest litigation, and so on. However much substantive considerations motivate the Supreme Court's majorities behind the scenes, the Court has cited the burdens of intense

1. Fed. R. of Civ. P. 1 (emphasis added). Note, though, the exception for anything "stated in Rule 81," *id.,* affirming certain exceptional substantive areas.

2. As true as that is, it was *geographic* uniformity across federal judicial districts that appears to have motivated the enactment of the Federal Rules' progenitive Rules Enabling Act of 1934. Stephen B. Burbank, *The Rules Enabling Act of 1934*, 130 U. Pa. L. Rev. 1015 (1982). Geographic uniformity was a central feature of the ABA's decades-long campaign in favor of enacting a Rules Enabling Act. And Rule 1 isn't the end of Rules-wise uniformity: As far as a Rule could, Rule 2 ended the distinction in modes of procedure between cases in law and those in equity, by announcing "[t]here is *one* form of action—the civil action." Fed. R. Civ. P. 2 (emphasis added).

3. Alexandra Lahav has noted the mismatch between the casebook model and lived experience in federal court litigation. *See* Alexandra D. Lahav, *Procedural Design*, 71 Vand. L. Rev. 821 (2018).

4. There is also variation in the number of party types, with counterclaimants, third party defendants, and others not of the simple Plaintiff or Defendant sort showing up in numerous cases.

and/or complex litigation when it has retrenched in various procedural areas. A classic move is to couch the retrenching argument in transsubstantive procedural terms, as the Court did a decade and a half ago when it changed the pleading standard in 2007's *Bell Atlantic Corp. v. Twombly*[5] and 2009's *Ashcroft v. Iqbal*.[6] In both cases, the Court justified its decisions in terms of the costs of discovery in intensely litigated, complex cases. *Twombly* was a putative class action antitrust case against multiple major telecom companies and fit that glove well. *Iqbal* involved a different type of litigation burden, namely whether a plaintiff alleging constitutional civil rights violations could hale the Attorney General and FBI Director into deposition rooms. Deciding the answer was no, the Court's *Iqbal* majority pointed to Rule 1's proclamation of uniformity, thereby making plain that *Twombly*'s pleading standard extended to "all civil actions."[7] There are plenty of other examples. Consider *Celotex v. Catrett*, for example, one of the many asbestos cases that filled the federal courts in the 1980s.[8] The Court went out of its way to offer dicta pointing to the liberality of Rule 8's half-century-old pleading standard as a reason to tighten summary judgment's screws.[9]

I define a case's twin traits of intensity and complexity, operating together, as its *intexity*. Intexity implicates uniformity, as well as other foundational principles such as liberality, efficiency, and accuracy.[10]

---

5. Bell Atlantic Corp. v. Twombly, 550 U.S. 544 (2007).

6. Ashcroft v. Iqbal, 556 U.S. 662 (2009).

7. Twombly, 550 U.S. at 555–556.

8. Celotex Corp. v. Catrett, 477 U.S. 317 (1986).

9. *Celotex*, 477 U.S. at 327 (noting Fed. R. Civ. P. 56 should be "construed with due regard not only for the rights of persons asserting claims and defenses that are adequately based in fact to have those claims and defenses tried to a jury, but also for the rights of persons opposing such claims and defenses to demonstrate in the manner provided by the Rule, prior to trial, that the claims and defenses have no factual basis").

10. With respect to efficiency, fairness, and accuracy, consider Fed. R. Civ. P. 1's objective of securing the "just, speedy, and inexpensive determination" of civil actions. Liberality includes liberality of pleading, with complaints originally required to provide fair notice under Fed. R. Civ. P. 8(a)(2) (requiring a pleading that states a claim for relief to provide only "a short and plain statement of the claim showing that the pleader is entitled to relief"; liberality of discovery, so that parties are encouraged to develop a shared understanding of facts well before trial, see Fed. R. Civ. P. 26(b)'s expansive definition of discovery scope; and liberality of joinder, so that parties are encouraged to bring all claims with overlapping facts or legal issues, even when additional parties must be joined. *See, e.g.*, FED. R. CIV. P. 18(a) (allowing a party with even one claim against an adverse person in an action to bring *all* claims the party has against that person in that same action); FED. R. CIV. P. 20(a) (allowing joinder of plaintiffs provided there is overlap in the relief they seek, and of defendants provided there is overlap in the relief sought from them, given that "any question of law or fact common to all" plaintiffs (Rule 20(a)(2)) or defendants (Rule 20(b)(2)) "will arise in the action"); FED. R. CIV. P. 23 (providing for class action litigation). Courts, rule makers, and Congress have modified these three dimensions of liberality in recent decades. *See, e.g.,* recent Supreme Court policy innovations in *Twombly* and

And it does so in ways that can be expected to implicate these principles jointly rather than in isolation. Rule 1's just, speedy, *and* inexpensive determinations are not all possible to a limitless degree—there are unavoidable tradeoffs between the values of justice, speed, and expense.[11] Leaving aside unambiguously wasteful policies, procedure that does well on one of these three dimensions must sacrifice one or both of the others. Such tradeoffs will vary with the intensity and complexity of cases: almost by construction, more intensely litigated cases are more time-consuming than less intensely litigated ones, and it seems likely they are also more expensive. The same goes for more complex cases—where more issues are disputed, more information is subject to discovery, and more parties are involved—by comparison to less complex ones.

Because a more liberal pleading standard allows more cases past the motion to dismiss phase, all things equal, liberal pleading will matter more for more complex and more intensely litigated cases. Similarly, more liberal discovery and joinder will multiply the divide in expense and time-to-adjudication between simple cases and those that are more complex or more intensely litigated.

Thanks to the language of the Enabling Act's[12] Rules, and the legal culture that grew up around them, our system's animating myth of uniformity holds that all cases in federal court face the same basic procedural law.[13] As a result, federal procedure is not just transsubstantive—it's also both transintense and transcomplex.

So federal procedure is *transintex*.

Iqbal, which grafted a textual plausibility and non-conclusoriness pleading requirements onto Rule 8(a)(2); the importation of the discovery proportionality standard into Fed. R. Civ. P. 26(b)'s discovery scope definition (discussed in Jonah B. Gelbach & Bruce Kobayashi, *The Law and Economics of Proportionality in Discovery*, 50 GA. L. REV. 1093 (2016)); the 2005 Class Action Fairness Act of 2005, Pub. L. No. 109-2, 119 Stat. 4 (2005), which channels class litigation into federal courts, as well as Supreme Court decisions, such as Am. Express Co. v. Italian Colors Rest., 570 U.S. 228 (2013), which limit the set of cases that can be litigated in class form.

11. As Professor Bruce Kobayashi has remarked in conversation with me, the phrase "just, speedy, and inexpensive" might better have been written as "just, speedy, *or* inexpensive."

12. The Rules Enabling Act, 28 U.S.C. § 2072 *et seq.* (1988).

13. In using the term "myth," I do not mean to suggest falsity as such. I appeal instead to the sense of "a popular belief or tradition that has grown up around something or someone." *Myth*, MERRIAM WEBSTER (last visited Jun. 9, 2024), https://www.merriam-webster.com/dictionary/myth [https://perma.cc/52QE-AS26] (definition 2a). Of course, as the cited works in note 15 suggest, the open-ended nature of the Rules' text together with the practical reality of varied litigation means there is substantial discretion in the way judges apply the Rules. The Rules themselves recognize this, as is evident from the extensive case management discretion they include. *See* FED. R. CIV. P. 16., FED. R. CIV. P. 26(b)'s highly subjective proportionality provisions, FED. R. CIV. P. 15(a)(2)'s reference to "when justice so requires", and many more.

But intexity puts pressure on the ideal of uniformity. True uniform procedure simply couldn't manage cases with manifestly different features, including complexity and intensity. This is an old theme, brilliantly surfaced by Judith Resnik in her seminal *Managerial Judging* article.[14] For good and possibly for bad, our uniformity is a case-level managerial uniformity, one that firmly reflects equity's discretionary flexibility rather than law's rigid rules.[15] Naturally, case-level discretion allows judges to handle intensely litigated and complex cases differently, applying doctrine and Rules flexibly, so cases, doctrine, and policy discussions all are shot through with discussions of intensity and complexity.[16]

Scholars have lavished attention on the phenomenon of "complex litigation" in recent decades, setting their sights overwhelmingly on class actions and MDL consolidation. But a key thesis of this paper is that intexity is not the sort of trait that cases either have or don't. Rather, intensity and complexity surface along continua. Cases and their litigation can be more or less intense, and more or less complex. So they can be more or less intex. And thus, just as every case has some type of substance—contract, tort, property, violation of the Telephone Consumer Protection Act, the Fair Labor Standards Act, or the Civil Rights Act—every case also has some degree of intexity. It is notable, then, that the categories of intensity and complexity have received limited attention in analytical discussions of the structure of federal civil procedure; I return to this point below.

Perhaps this is because we have generally lacked the ability to measure intexity effectively. That incapacity is a policy choice, being the consequence of the federal judiciary's stinginess with its mountain of litigation data, which pile up in the ordinary course of doing the nation's litigation business. The federal judiciary's case management system,

14. Judith Resnik, *Managerial Judges,* 96 HARV. L. REV. 374 (1982).

15. The procedure literature is replete with discussions of this point and its import. *See e.g.,* Resnik, *supra* note 14 at 432-33 (discussing the importance of "controlling discretion"); Richard L. Marcus, *Slouching Toward Discretion*, 78 NOTRE DAME L. REV. 1561 (2003) (the Federal Rules "draw their essence more from the relaxed and discretionary background of equity than the confining orientation of the common law"); Stephen B. Burbank, *Pleading and the Dilemmas of "General Rules"*, 2009 WISC. L. REV. 535 (2009) (commenting on the discussion by Geoffrey P. Miller, *Pleading After Tellabs*, 2009 WISC. L. REV. 507 (2011) of legislated substance-specific pleading standard in securities cases); and Stephen N. Subrin, *Fudge Points and Thin Ice in Discovery Reform and the Case for Selective Substance-Specific Procedure*, 46 FLA. L. REV. 27 (1994).

16. *See, e.g.*, the deployment of Judge Frank H. Easterbrook's *Discovery as Abuse*, 69 B.U. L. REV. 635 (1989) in *Twombly*, 550 U.S. at 559; *see also* Malibu Media, LLC v. John Does 1, 6, 13, 14, 950 F. Supp. 2d 779, 781 (E.D. Pa. 2013) (discussing arguments for and against Rule 20 joinder in BitTorrent-related copyright cases); for more on the empirics of joinder in such cases, see Shyamkrishna Balganesh & Jonah B. Gelbach, *Debunking the Myth of the Copyright Troll Apocalypse*, 101 IOWA L. REV. 43 (2016).

typically designated "CM/ECF,"[17] contains over a billion documents collected and organized by the federal courts in roughly the last quarter century.[18] But comprehensive access to these data is extremely difficult for researchers to come by, making it prohibitive to study intexity usefully.[19]

This Article uses a massive, bespoke set of docket information to draw the curtain at least partially on the nature of federal litigation's intexity.[20] In broad terms, the data come from a comprehensive collection of docket reports for federal civil cases observed for up to ten years between 2005 and 2014. Using a subset of more than half a million cases, I present an array of empirical findings in Part II. These findings require no knowledge of technical estimation. Most are based on simple summary statistics and tabulations cuts of variables selected to proxy for cases' intensity and complexity. The variables are the number of docket entries in cases observed over a period of at least seven years; the number of parties in these cases; and two novel measures of case complexity constructed by leveraging the inherent structure of docketed information and using simple concepts from the mathematical theory of graphs and networks. The findings provide arguably the most comprehensive assessment to date of interesting aspects of federal civil litigation.[21]

A key contribution of this Article is to quantify the considerable variation in the extent of intexity across cases: some cases are quickly disposed of and simple, whereas others stretch for years and comprise many parties and/or motion practice (which might involve either dispositive or non-dispositive issues). This qualitative result is what anyone acquainted with the contemporary nature of American law would expect. But we have not previously been able to document the extent of these basic facts.

A bigger contribution, though, is to show the degree to which intexity is more importantly *intra*substantive than *trans*substantive. American legal policy and reform controversies often center on, or at least spring partially from, substantive areas of the law, as with securities litigation, antitrust, and patents. This is so even when procedural rather than substantive-standard reforms are involved, as evidenced by the

17. This stands for "Case Management/Electronic Case Files".

18. CHIEF JUSTICE JOHN ROBERTS, 2014 YEAR-END REPORT ON THE FEDERAL JUDICIARY, at 6, https://www.supremecourt.gov/publicinfo/year-end/2014year-endreport.pdf.

19. Along with many others, I have argued in favor of unlocking the door to PACER's data. Jonah B. Gelbach, *Free PACER*, *in* LEGAL TECH AND THE FUTURE OF CIVIL JUSTICE (David Freeman Engstrom, Ed., 2023).

20. I discuss my data below. In brief, the data come from a comprehensive collection of raw docket entries for 10 years of federal civil filings.

21. Even for readers who aren't intrinsically interested in understanding these facts, the results should be of interest because so much substantive law is made in cases that proceed through the federal district courts. On that point, see Florencia Marotta-Wurgler & Samuel Issacharoff, *The Hollowed Out Common Law*, 67 UCLA L. REV. 600 (2020).

statutorily elevated pleading standard in securities litigation,[22] the role of complexity in antitrust cases,[23] and patent law's numerous aspects of procedural exceptionalism.[24] Demonstrating the intrasubstantive nature of intexity—the fact that there is wide variation in intensity and complexity *within* substantive areas of litigation—is important because it helps us see that intexity is not limited to specific subject areas. In turn, that helps us understand that when legislators, rule makers, or courts adopt transsubstantive procedural policies as a way to address concerns about intensity or complexity, the effects of these policies will not be limited to litigation in the "big" substantive areas such as securities, antitrust, or patents. This Article's contribution thus can be thought of as providing an empirical diagnosis of intexity's substantive scope.

To be clear, as an empirical matter, nothing stops substance from having primary importance. Consider an observer interested in predicting whether a case would consume a large amount of court and other social resources, in each of two possible worlds. In World 1, all securities fraud cases are so intex that they eclipse even relatively intex tort cases. In World 2, securities cases are more intex than tort cases on average, but there are highly intex tort cases that are more intex than all but the most intex securities cases. Suppose our observer is confronted with the tort suit of *Smith v. Jones, Inc.*, and with the securities fraud case of *In re ShareCo Securities Litigation*. In World 1, our observer

22. *See, e.g.,* The Private Securities Litigation Reform Act of 1995, Pub. L. No. 104-67, 109 Stat. 737 (1995).

23. Consider this comment from the district court in *Zenith Radio Corp. v. Matsushita Elec. Indus. Co.*, 478 F. Supp. 889, 895 (E.D. Pa. 1979), *vacated sub nom. In re* Japanese Elec. Prod. Antitrust Litig., 631 F.2d 1069 (3d Cir. 1980): "To date over 20 million documents have been produced for inspection. A considerable number of these have had to be translated from Japanese into English. The deposition transcripts completed to date total well over 100,000 pages, and many depositions remain to be taken . . . . We have been inundated with a plethora of discovery motions; in the past few months we have dealt with over 50 Rule 37 motions of various descriptions, and pretrial conferences with counsel for the parties are consuming at least 3 full days per month, mostly to resolve discovery problems . . . . It is anticipated that the trial will consume approximately one year."

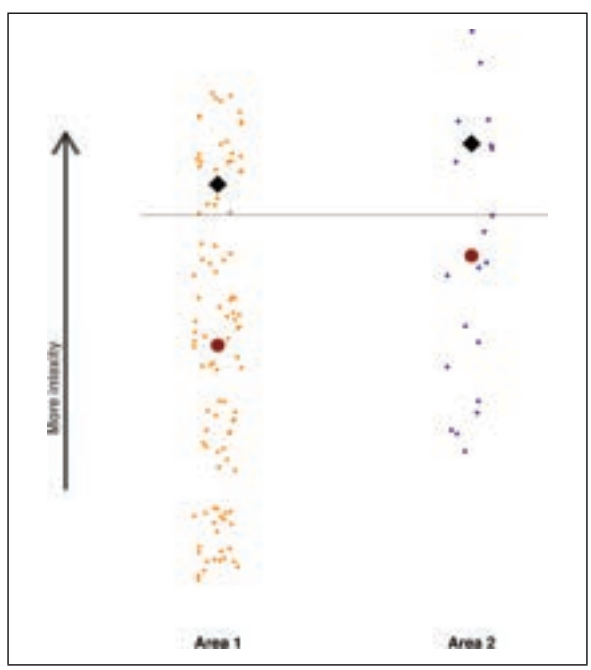
24. These include *Markman*'s allocation of claim construction to judges rather than juries, Markman v. Westview Instruments, Inc., 517 U.S. 370, 390 (1996) (emphasizing "the importance of uniformity in the treatment of a given patent as an independent reason to allocate all issues of construction to the court" rather than allowing juries to decide claim construction); America Invents Acts, 35 U.S.C. § 299 (joinder limitations); restricted venue under 28 U.S.C. § 1400 as interpreted in TC Heartland LLC v. Kraft Foods Grp. Brands LLC, 581 U.S. 258, 262 (2017); and various local patent rules, such as Rules 3-1 and 3-3 of the Local Rules of Practice for Patent Cases before the United States District Court for the Northern District of California, which, the Federal Circuit has explained, "require parties to state early in the litigation and with specificity their contentions with respect to infringement and invalidity." O2 Micro Int'l Ltd. v. Monolithic Power Sys., Inc., 467 F.3d 1355, 1359 (Fed. Cir. 2006).

would immediately know that *ShareCo* would consume more resources, because it's a securities case and *Smith* is a tort case, and in World 1 that's enough to know that *ShareCo* will be more intex. But in World 2, our observer would need to know more—specifically, they'd need to know whether *Smith* was one of the most highly intex tort cases—because if it were, then it might be more intexly litigated than *ShareCo* (so long as *ShareCo* wasn't the most highly intex type of securities case).

If the real world looks like World 1, then designing procedural law in a substance-specific way would solve intexity-related challenges in the process. On the other hand, if the real world is more like World 2, then things aren't necessarily so simple. If (i) there is enough variation in intexity *within* each area, and (ii) there are many more personal injury than securities cases, then it could be more useful to direct intexity-oriented procedural rules at general indicators of intexity rather than at substance.

Figure 1 provides a motivating example to illustrate this discussion. The figure plots the value of an imagined measure of intexity—i.e., some combination of intensity and complexity metrics—with values higher in the figure's space corresponding to more intense and/or more complex cases. In the example, there are two substantive areas of law, labeled Area 1 and Area 2. Cases' intexity levels are plotted using small circles.

Figure 1: A Motivating Example



The large circles indicate the averages level of intexity for the two groups. It is apparent that Area 2 cases are, on average, more intex than those in Area 1. However, the dispersed nature of the dots shows that there is significant variation in intexity within each substantive area, so that comparing only averages could fail to tell the whole story. Think of the horizontal line drawn across the two substantive Areas as representing a threshold, so that cases with intexity above that level are considered highly intex. The following facts are all true about Figure 1:

(1) The share of Area 2 cases that are highly intex exceeds the share of Area 1 cases that are.
(2) But a greater share of all highly intex cases are Area 1 cases rather than Area 2 cases—simply because there are so many more Area 1 cases.
(3) Among the subset of highly intex cases, average intexity—indicated by a diamond for each Area—is greater for Area 2 cases than for Area 1 cases.
(4) Even so, *total* intexity—measured as the product of the number of highly intex cases and their average intexity level—is greater among Area 1 cases; their greater numbers outweigh the greater average intexity among Area 2 cases.
(5) By any reasonable measure, the degree of variation *across* the two substantive areas is small by comparison to the variation *within* each of them.

Although Figure 1 is stylized and not based on actual data, the facts it illustrates aren't just possible in the abstract—they actually hold in my empirical results. This Article's primary empirical take-home points, then, are:

- Intexity is transsubstantive: A given level of intensity or complexity may be found in many substantive areas.
- Substance is transintex: Most substantive areas, broadly considered, have both cases that are highly intense/complex and others that are not.

An additional finding, not described by Figure 1, is that intensity is not simply complexity, or vice-versa. I show this using simple pairwise statistics measuring the association between my various measures of complexity and intensity.

This Article's primary focus is positive—on surfacing and documenting the empirical extent of intexity. But its ideas, results, and analysis have normative bite, too. Thus, Part III treads gingerly into the realm of policy, discussing how Part II's empirical analysis might resonate for procedure policy discussions. One theme involves the case for allowing more access to federal court data so that both federal

court insiders and outside researchers may investigate the empirical landscape. A second normative issue is whether transsubstantive intexity implies that we ought to rethink our default rule of transintex procedure. Specifically, perhaps we ought to ask whether it would make sense to adopt formal rules related to complexity—even outside the MDL context, where such debate already percolates. It seems at least possible that procedural rules and doctrines would be better directed at intexity as such, rather than at particular substantive areas such as securities, antitrust, or patents. Third, and related, is the question of whether fine-tuning discovery according to intexity could relieve some of the pressure that the Supreme Court points to in *Twombly* and *Iqbal* as justifying the plausibility pleading standard.

I. Previous Studies, and this Article's Data

There is a long history of empirical study of federal civil litigation, and it is hardly news that federal civil cases differ in how they play out. Most notable is the famous RAND study of ten districts that implemented Civil Justice Reform Act (CJRA) pilot programs, as well as ten comparison districts that were free to implement required CJRA elements as they saw fit.[25] RAND's authors intensely studied "more than 10,000 cases," with primary focus on cases with "issue joined," generally meaning that defendants answered the complaint.[26]

Other empirical studies of federal civil litigation abound. For example, the Federal Judicial Center has released numerous reports on a variety of matters, including summary judgment[27] and motions to dismiss.[28] Numerous studies have been conducted by legal scholars on the effects of *Twombly* and *Iqbal* on various outcomes;[29]

25. James S. Kakalik et al. An Evaluation of Judicial Case Management Under the Civil Justice Reform Act xvii (1995).

26. *Id.* at 10. The total of 10,000 cases reflects inclusion of roughly 250 cases from each of the pre- and post-CJRA time periods studied for each of the 20 districts.

27. *See*, *e.g.*, Joe S. Cecil et al., *A Quarter-Century of Summary Judgment Practice in Six Federal District Courts*, 4 J. Empirical Legal Stud. 861, 863 (2007); Memorandum from Joe Cecil & George Cort, Fed. Jud. Ctr., to Judge Michael Baylson (Aug. 13, 2008).

28. *See*, *e.g.*, Joe S. Cecil et al., Fed. Judicial Ctr., Motions to Dismiss for Failure to State a Claim After *Iqbal*: Report to the Judicial Conference Advisory Committee on Civil Rules (2011); Joe S. Cecil et al., Fed. Jud. Ctr., Update on Resolution of Rule 12(b)(6) Motions Granted with Leave to Amend: Report to the Judicial Conference Advisory Committee on Civil Rules (2011).

29. *See*, *e.g.*, Jonah B. Gelbach, *Material Facts in the Debate over* Twombly *and* Iqbal, 68 Stanford L. Rev. 369 (2016); Jonah B. Gelbach, Note, *Locking the Doors to Discovery? Assessing the Effects of* Twombly *and* Iqbal *on Access to Discovery*, 121 Yale L.J. 2270 (2012) [hereinafter *Locking the Doors*]; Alexander A. Reinert, *Measuring the Impact of Plausibility Pleading*, 101 Va. L. Rev. 2117 (2015); Patricia

empirical evidence on Rule 11;[30] recent work on transfer[31] and *pro se* representation;[32] securities litigation;[33] and no doubt much more.

These studies, and others,[34] have helped us understand our federal civil litigation system. But through no fault of their authors, they are limited by the extent of their access to data—we can't learn what we can't observe. Even RAND's CJRA studies—which in many ways remain the most comprehensive ones we have before this Article—were limited in two ways. First, they used data from only 20 districts (albeit relatively large ones). Second, within these districts RAND was able to study only about 10,000 cases. Samples of that magnitude are fine—even large—for some purposes, such as computing means or standard deviations of variables such as the time to case termination. But such sample sizes have limited utility, if that much, for making comparisons about extrema, such as statistics involving the top 1% of cases, within substantive areas. As massive an effort as the RAND CJRA study was, its sample size is too small to support the kind of detail I provide here—including discussion of 99th percentiles in various substantive subcategories of the data. To put the difference in data scale in perspective, my analysis set has more than 10,000 cases in just the patent and securities areas. Further, I provide data from many more districts than the 20 represented in the RAND study, and I do not restrict attention to cases in which an answer was filed. For all of these reasons, my collection of data is much better suited to the present study than would be the RAND data.

My data come from a bespoke collection of docket report activity obtained from Westlaw as a result of a grant funded by the Oscar M. Ruebhausen Fund at the Yale Law School.[35] The arrangement with

---

W. Hatamyar, *The Tao of Pleading: Do* Twombly *and* Iqbal *Matter Empirically?*, 59 AM. U. L. REV. 553 (2010); and William H.J. Hubbard, *Testing for Change in Procedural Standards, with Application to* Bell Atlantic v. Twombly, 42 J. LEGAL STUD. 35 (2013).

30. *See*, *e.g.*, STEPHEN B. BURBANK, THE REPORT OF THE THIRD CIRCUIT TASK FORCE ON FEDERAL RULE OF CIVIL PROCEDURE 11: AN UPDATE 511 (1989).

31. *See, e.g.*, Roger Michalski, *Transferred Justice: An Empirical Account of Federal Transfers in the Wake of Atlantic Marine*, 53 HOUSTON L. REV. 1289 (2016).

32. Roger Michalski, *The Pro Se Gender Gap*, 88 BROOKLYN L. REV. 563, 566 (2023).

33. *See, e.g.*, Stanford Law School Securities Class Action Clearinghouse and Cornerstone Research, *Research Reports* [hereinafter Stanford Securities Class Action Clearinghouse] https://securities.stanford.edu/clearinghouse-research.html [https://perma.cc/HG8T-VHK2] (last visited Aug. 6, 2024) (reports based on information from the Securities Class Action Clearing House hosted by Stanford Law School).

34. David A. Hoffman, Alan J. Izenman, and Jeffrey R. Lidicker, *Docketology, District Courts, and Doctrine*, 85 WASH. U. L. REV. 681 (2007).

35. I submitted this grant proposal when I was a law student, jointly with Yale Law Professor William N. Eskridge, Jr., whose assistance was instrumental.

Westlaw provided what is supposed to be (and from what I can tell, appears to be very close to) the universe of federal district court docket activity for cases filed on or after January 1, 2005, with data continuing through December 31, 2014. These data should reflect what one would obtain from PACER searches and downloads, which is how Westlaw obtained the information. One limitation is that this data source does not include any underlying documents such as complaints or briefs.

I am aware of one other study in the law literature that provides information on litigation intensity using analysis of docket reports. That study has useable data on 980 cases filed in four district courts in 2003.[36] To my knowledge, though, no prior study of civil action docket reports has had sample sizes on the same order of magnitude as this Article. Consequently, compared to prior researchers, I am able to study far more cases over an extended period of time, and in all litigation subject matter areas.[37]

That said, one limitation of this data source involves its reliance on the Administrative Office of U.S. Courts (AO) Case Management/Electronic Case Filing (CM/ECF) system. Although U.S. district courts began using this system in May 2002,[38] not all district courts adopted the electronic case filing (ECF) system at that time. As I discuss in Appendix B, this has implications for the set of cases I use here.

Many details about the docket report data I received from Thomson Reuters appear in Appendix A; I offer a brief discussion here. As noted above, I do not have any documents that were filed in any cases, i.e., I do not have complaints, answers, memoranda of law, or attachments

36. Hoffman et al., *supra* note 34, at 708-709.

37. I have used the data in the present study in several previous publications on an array of topics. *See* Gelbach, *Material Facts, supra* note 29, at 393 (evaluating the case-quality effects of pleading standard changes); Jonah B. Gelbach and Deborah R. Hensler, *What We Don't Know About Class Actions but Hope to Know Soon*, 87 FORDHAM L. REV. 65 (2018) (counting class actions); Jonah B. Gelbach, *Rethinking Summary Judgment Empirics: The Life of the Parties*, 162 U. PA. L. R. 1663 (2014) (assessing the role of litigation selection effects in complicating the relationship between judicial characteristics and summary judgment practice); JONAH B. GELBACH & DAVID MARCUS, A STUDY OF SOCIAL SECURITY LITIGATION IN THE FEDERAL COURTS (report to the Administrative Conference of the United States) (July 28, 2016) (investigating correlates of disability appeal remand rates). I have also made some or all of these data available to various other scholars under terms consistent with the data-provision contract between Yale and Thomson Reuters. To date I am unaware of any publications resulting from such access, though happily that will soon change. *See* Jonathan B. Petkun, *Nudges For Judges: The Effects Of The "Six-Month" List On Federal Civil Justice* (working paper), https://jbpetkun.github.io/pages/working_papers/Petkun_SixMoList_May2021.pdf [https://perma.cc/Y2XW-EPYW].

38. *See* FAQs: CASE MANAGEMENT/ELECTRONIC CASE FILES, https://www.uscourts.gov/court-records/electronic-filing-cmecf/faqs-case-management-electronic-case-files-cmecf (last visited Jun. 3, 2024) [https://perma.cc/Y25F-WY27].

filed with docketed events. All I have is case-level information and the docket report text that a PACER search would yield. That turns out to be a lot, though.

The details of my data are lengthy and at times complicated, so I relegate them to Appendix A. I encourage readers to engage with Appendix A because some of the details are important to understanding my data's strengths and limitations. That said, readers uninterested in the gory details can safely skip Appendix A. A final important note, though, is because I exclude cases that have indicia of consolidation, my data should not include any of the mammoth multi-district litigations. Excluding these cases ensures that I not find transsubstantive intexity simply due to MDLs' complexity and intensity.[39]

II. Empirical Facts: Substance, Intensity, and Complexity

A threshold task is to define variables that measure intensity and complexity.

We can think of intensity, roughly speaking, as "lots of litigation activity," for which the number of docket entries in a case is as good a proxy as I can conjure.[40] Section A of this Part thus investigates a number of questions related to this measure of litigation intensity. I find that although intensity varies across substantive areas of litigation, it varies much more importantly *within* these areas. Thus transintexity is distinguishable from transsubstantivity.

What about complexity? It is a commonplace that there is such a thing as "complex litigation" in our civil justice system. The Federal Judicial Center publishes a *Manual* on the subject,[41] casebooks bear the

39. My final data set uses information on 566,315 cases. An additional 64,169 meet all the criteria discussed in Appendix A except indicia-of-consolidation. Including these cases changes some of the particular numbers. One difference is that the extreme right tail of docket-intensity among cases with indicia of consolidation is far above the top of the distribution for the rest of cases; this reflects that the most docket-intense MDLs have ginormous numbers of docket entries—the top value in my database was 80,628, for the case "Asbestos Products Liability Litigation (No. VI)". A second difference is that, as expected, tort MDLs tend to have many more parties, at least at the top of the distribution. My basic qualitative empirical results are unaffected by the exclusion of cases with indicia of consolidation. Although either choice would be defensible, but to avoid any suggestion that my results are simply driven by MDLs, which already have proceduralists' attention, I have chosen to exclude these cases from the analysis.

40. To be sure, docket entries can be generated from non-intense matters, such as motions for *pro hac vice* admission and related orders; conversely, highly intense features of litigation may be reflected by a small number of docket entries or even none at all, as with many aspects of discovery. I make no claim that using the docket-entry count is perfect, or even good for all cases—just that on balance, it is likely a decent measure of intensity.

41. Manual for Complex Litigation (Fourth) 1 (2004).

title,[42] many if not most law schools offer courses in the area, and the term is regularly bandied about in procedure reform discussions.

What is less clear is how to define complex litigation. Most people know it when they see it: *Amchem*,[43] the 9/11 litigation,[44] and *Wal-Mart v. Dukes*[45] were all undoubtedly complex litigations. Certainly, any litigation that involves an MDL docket can be expected to be complex. And many, if not all, putative class actions involve some complexity. So it seems that litigation involving large numbers of parties or absent but represented persons frequently will merit the term. But complexity is not limited to cases with formal markers such as MDL or class litigation. Liberal party joinder by itself creates the possibility of considerable complexity.[46] And even cases with just two parties can engender considerable complexity depending on the nature of the substantive issues involved.

Thus, having many parties, some of whom might be absent, captures some aspect of what I mean by complexity. In addition, procedural complexity can encompass situations in which multiple parties are involved with multiple aspects of contested issues in a litigation. For example, in a securities fraud case covered by the PSLRA, the question of which party will be lead plaintiff[47] may involve jockeying by numerous parties and/or attorneys and law firms. Or consider a tort case with multiple parties, some of whom have been impleaded on indemnification grounds, as in *Asahi Metal Industry Co. v. Superior Court*—where the key question related to personal jurisdiction of a third-party defendant.[48] Issues in which multiple players in the litigation are involved in extended motion-opposition-reply briefing and oral argument can arise across substantive areas. Such issues implicate the kind of procedural complexity I have in mind. Although they are

42. *See, e.g.*, RICHARD MARCUS, EDWARD SHERMAN, HOWARD ERICHSON, AND ANDREW BRADT, COMPLEX LITIGATION: CASES AND MATERIALS ON ADVANCED CIVIL PROCEDURE (7th ed. 2021).

43. Amchem Prods., Inc. v. Windsor, 521 U.S. 591, 641 (1997) (rejecting class certification because of concerns about intra-class heterogeneity and attendant inadequacy of representation).

44. *In re* Terrorist Attacks on Sept. 11, 2001, 392 F. Supp. 2d 539 (S.D.N.Y. 2005).

45. Wal-Mart Stores, Inc. v. Dukes, 564 U.S. 338 (2011).

46. Indeed, this is a good reason why the Rules have safety valves such as Rule 21 (severing claims and dropping parties "on just terms", FED. R. CIV. P. 21) and Rule 42(b) (allowing courts to order separate trials for distinct issues or claims in a given action).

47. On the subject of lead plaintiffs, see 15 U.S.C. § 78u–4(a)(3).

48. Asahi Metal Indus. Co. v. Superior Ct. of Cal., 480 U.S. 102 (1987).

difficult to define with precision, these issues distinguish this kind of complexity from the complexity of a case's substantive questions.[49]

To have a chance at capturing these multiple dimensions of complexity, I use multiple variables. The first is simply the number of distinct formal parties present in a case. Two additional, novel measures will be based on using the mathematical theory of graphs and networks to characterize the relationship between docket entries. These measures, which I explain in detail below, are the number of connected components in a case's docket network and the network's maximum component size. Section B of this Part investigates the same types of questions for these complexity measures that section A considers related to intensity. As with intensity, I find that complexity varies more importantly within substantive areas of litigation than across. It follows that transcomplexity, too, is distinguishable from transsubstantivity.

Section C of this Part then asks whether my measures of intensity and complexity are importantly distinguishable. Using simple pairwise statistical measures, I present evidence suggesting that they are. It follows that "intensity" and "complexity" are not merely synonyms for each other; the variables measuring them represent distinctive phenomena. So, transintensity and transcomplexity are distinct as well.

A. Intensity: Docket Entries and Duration

This section presents an array of information about intensity of litigation, as measured by the number of docket entries observed in each case.[50]

---

49. Here it is useful to consider again *Zenith Radio Corp. v. Matsushita Elec. Indus. Co.*, 478 F. Supp. 889, 895 (E.D. Pa. 1979), *vacated sub nom. In re* Japanese Elec. Prod. Antitrust Litig., 631 F.2d 1069 (3d Cir. 1980), discussed *supra* note 23. In that case, which involved antitrust issues many observers would consider to involve complex economic reasoning, the district court commented that it had been "inundated with a plethora of discovery motions; in the past few months we have dealt with over 50 Rule 37 motions of various descriptions," and that it "anticipated that the trial will consume approximately one year." Even if the case had not involved substantively complex matters of economic reasoning—and instead involved, say, mine-run tortious actions—the "plethora of discovery motions" and Rule 37 activity reasonably would have made the case *procedurally* complex in the sense I have in mind.

50. An alternative would be to use information on case duration, which should be observable effectively even after the dockets data collection ends in 12/31/2014. The Federal Judicial Center's Integrated Database contains termination dates, which would allow duration to be computed for cases closed after 2014. Given this Article's length, I have not conducted the resulting analysis.

1. Pooling data across all substantive areas of law

Table 1 reports basic statistics about the distribution of the case-level number of docket entries. The first column names the percentile or other statistic for each row, and the second shows the value of that statistic for my analysis set.

Table 1: Descriptive Statistics For Number of Docket Entries

Percentile or Other Statistic	Number of Docket Entries
10	4
25	8
50	17
75	33
90	63
95	96
99	216
99.9	533
Mean	30
St. Dev.	48

At the bottom of the distribution, cases have very few docket entries—the 10th percentile is four entries, and the 25th is eight. Even the median is quite low, at just 17 docket entries. The mean of 30 is nearly twice that level, indicating that this distribution is highly skewed; an additional indicum of skewness is that the standard deviation of 48 is nearly twice the value of the interquartile range.[51] The figures for the top of the distribution illustrate just how skewed the distribution is: the 90th percentile of 63 entries is almost four times the median, and the 99th percentile of 216 entries is more than ten times. The 99.9th percentile is a whopping 533 docket entries; to put it differently, the

51. The interquartile range ("IQR") of a distribution is defined as the difference between the 75th and 25th percentile, which is 32–8=24 docket entries here. For a normally distributed variable, the standard deviation is roughly ¾ of the IQR; here the standard deviation is more than twice that.

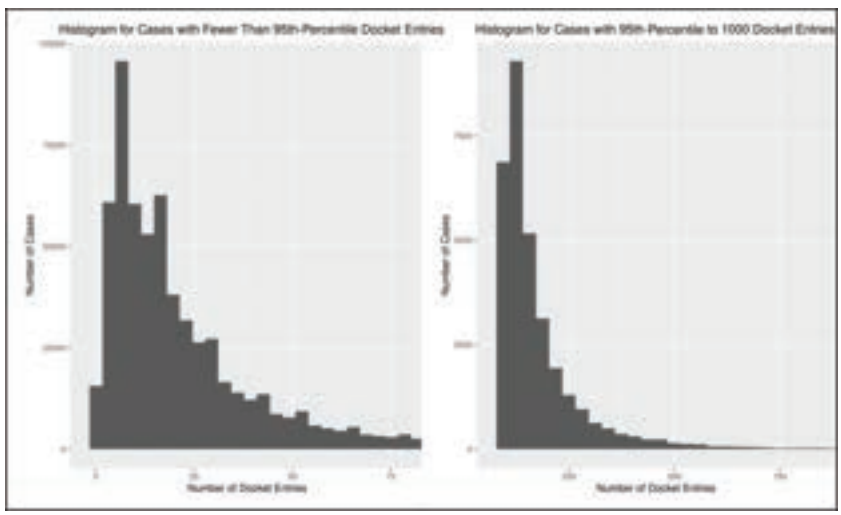
99.9th percentile case accounts for more docket entries than would 30 median cases.[52] The skewness revealed by these statistics underscores the extent to which most cases differ from the relatively small number that have extremely intense docket activity, exemplifying the divide between mine-run and highly intense cases.

Because the skewness is so substantial, a histogram of the full distribution of docket entries is basically uninformative. However, we can get a visual sense for the distribution of docket entries across cases by looking only at those in the bottom 99%. Figure 2(a) shows a histogram for these cases. A key take-home point from this figure is that once we move to the right of the median of 17 docket entries, the share of cases represented by each histogram bin declines rapidly as the number of docket entries rises. Figure 2(b) shows a histogram computed only for the much smaller number of cases with docket entries in the top 5%.[53] (Note the difference in vertical-axis scales for the two figures.) This second histogram shows that top-5% cases cluster toward what could be called "the bottom of the top." By the time we get to the 99th percentile of 216 docket entries, the relative frequency of cases is dropping rapidly as the docket intensity rises.

52. I do not mean to suggest that the number of docket entries is a perfect measure of intensity. In particular, it may miss a lot of discovery-related activity, because such activity need not necessarily be docketed, and in any event not unless they are used in court proceedings. FED. R. CIV. P. 5(d)(1)(A) ("disclosures under Rule 26(a)(1) or (2) and the following discovery requests and responses must not be filed until they are used in the proceeding or the court orders filing: depositions, interrogatories, requests for documents or tangible things or to permit entry onto land, and requests for admission").

53. The relative handful of cases with more than 1,000 entries are excluded to avoid distorting this figure.

Figure 2: Histograms of Number of Docket Entries for Bottom 95% of Cases and Top 5% of Cases, Excluding those with more than 1,000 Entries



To sum up this discussion, when we pool together cases across substantive areas, it is as if there are essentially three sets of cases in the distribution of the number of docket entries:

First is a large number of cases that have very little docket activity—so little that half of all cases have 17 or fewer docket entries.

Second is the next 45% of cases as we move up the docket intensity distribution, which are cases with docketed activity in the low-intensity (17 entries) to the moderate- or even high-intensity level, with the 95th percentile being just under 100 entries.

Third is a set of cases with high to extremely high levels of docket intensity—a set that comprises cases with at least 100 docket entries, whose top echelon includes cases with many times that level.

These statistics demonstrate that, just as the rich are different because they have a lot more money than others, the most intensely litigated cases are very different from other cases in that they have a lot more docket entries. That is, of course, a tautological statement at one level of generality, but the statistics in this section put some fascinating and important meat on the bones of that observation.

2. Cases by PACER Nature of Suit Code Categories

I turn next to an analysis of cases broken down by substantive-law categories (the construction of these categories from PACER nature-of-suit codes is in Appendix C). The first numerical column of Table 2

reports the breakdown of my set of cases according to membership in these 14 categories, which I have listed in rough order of their average number of docket entries.

Many observers likely would guess that patent cases—which often have high stakes and well-resourced litigants—would be at or near the top of litigation intensity. Table 2 bears that out, as it shows that patent litigation (which is the listed nature of suit in only about 1% of the cases I consider) is the substantive area with the greatest average number of docket entries. Patent cases have a mean of 85 docket entries, which is nearly the 95th overall percentile for the overall distribution, as discussed in reference to Table 1, *supra*. Environmental, securities, and antitrust cases, which make up even smaller categories than patents—with averages of 61, 61, and 55 docket entries—have quite high docket intensity relative to the overall distribution. After that the average number of docket entries drops off substantially, with the large categories of civil rights and contract cases having an average of 42 and 36 docket entries each (although both these figures exceed the overall distribution's 75th percentile).

Table 2: Case Counts and Average Number of Docket Entries

Substantive Law Category	Number of Cases in Data	Average Number of Docket Entries
Patent	6,195	85
Environmental	1,723	61
Securities	2,744	61
Antitrust	1,438	55
Civil Rights	80,101	42
Contract	78,334	36
Copyright & Trademark	18,768	30
Labor	40,232	28
Other	60,575	27
Consumer	7,549	26
Tort	104,383	26
Social Security	32,581	21
Prisoner Petitions	131,212	21
Immigration	480	15

Notably, all categories have average (i.e., mean) numbers of docket entries that equal (in the case of immigration) or exceed (all other categories) the median overall number of docket entries. This fact reflects the substantial right skewness in the number of docket entries discussed above. When a distribution is characterized by this much skewness, a simple summary measure such as the average will do a poor job of informing us about what kind of case is typical. As a riff on a criticism of averages I have heard, suppose a billionaire walks into a working class bar. That will cause the average wealth of everyone in the bar to be in the tens or hundreds of millions, maybe even the billions—but it would be a mistake to think that this average told us much about anyone but the richest person there. With a highly skewed distribution, then, it is important to pay attention to multiple parts of the distribution, and I will do more of that momentarily.

Another important fact is that the four substantive law areas with the most intensity as measured by docket entries—patents, antitrust, environmental, and securities—are all quite small. Altogether, the roughly 12,000 cases in these categories account for only about 2% of my analysis set. Accordingly, although many cases of these types have lots of docket entries, there are many more docket entries associated with, say, contract cases than with the top four categories. Even with a relatively small average of 36 docket entries, the 78,334 contract cases accounted for a total of 2.8 million total docket entries, which is more than three times the *combined* number of docket entries for the patent, antitrust, securities, and environmental case categories (hereafter, the "Intense 4"). The key dynamic here is that, although the average number of docket entries for the Intense 4 is about twice that for contract cases, there are more than six times as many contract cases as Intense 4 cases. Six is greater than two, so there is a lot more docketed activity in contract cases.

Table 3 investigates the extent to which substantive law is associated not just with having greater average numbers of docket entries, but also with the share of cases that are at the top of the docket entry distribution. The table's second and third numerical columns collect cases into the subsets that are in the bottom 99% of the docket entry distribution and those that are in the top 1%.[54] The table's rows then report each substantive law category's percentage of all cases in these subsets (thus, the rows in each column sum to 100, up to rounding

54. Recall that the 99th percentile is 216 docket entries. Thus, cases represented in the bottom 99% column are those with fewer than 216 entries, and those in the top 1% column all have 216 or more.

error). To illustrate, the table's top row shows that patent cases make up 1.0% of all bottom-99% cases, but 10.8% of all top-1% cases.

The table's final column reports the "over-representation index," which is the ratio of the top-1% percentage to the bottom-99% percentage. As its name indicates, this ratio measures the degree to which a substantive area of the law is under- or over-represented among the most highly intense cases. The ratio of percentages for patent cases, 10.8, is the highest among all substantive areas, indicating that patent cases are the most over-represented area in top-1% cases.

The substantive areas of law appear in Table 3 in the same order as in Table 2, i.e., in descending order of average number of docket entries. The most notable order reversal between the average number of docket entries and the over-representation index is the fact that the antitrust group ranked fourth in average docket entries but second in over-representation (antitrust cases make up 1.5% of top-1% cases and only 0.2% of bottom-99% cases, for an over-representation ratio of 7.5). Environmental and securities cases are also greatly over-represented in the top 1%, with roughly five times the footprint there as in the bottom 99%.

Table 3: Over-Representation of Cases in the Top-1% in Number of Docket Entries, By Substantive Law Group

	Percentage of Cases with Number of Docket Entries In:		
Substantive Law Category	Bottom 99%	Top 1%	Over-Representation Index
Patent	1.0	10.8	10.8
Environmental	0.3	1.5	5.0
Securities	0.5	2.7	5.4
Antitrust	0.2	1.5	7.5
Civil Rights	14.1	20.8	1.5
Contract	13.8	18.9	1.4
Copyright & Trademark	3.3	3.9	1.2
Labor	7.1	5.4	0.8
Other	10.7	10.1	0.9

Substantive Law Category	Percentage of Cases with Number of Docket Entries In: Bottom 99%	Top 1%	Over-Representation Index
Consumer	1.3	0.6	0.5
Tort	18.5	13.7	0.7
Social Security	5.8	0.0	0.0
Prisoner Petitions	23.3	10.0	0.4
Immigration	0.1	0.0	0.0
Total	*100.0*	*100.0*	*1.0*

Notes:

Over-representation index is the ratio of Top-1% percentage to Bottom-99% percentage.

Only the 566,315 cases with no indicia of consolidation are included.

At the other end of the spectrum, the top 1% of cases by number of docket entries has no immigration cases and only one Social Security case, even though the immigration and Social Security areas together constitute roughly six percent of bottom-99% cases. Civil Rights, contract, and copyright & trademark cases are all over-represented by modest amounts. Labor and other cases are relatively close to par; tort cases are under-represented, making up 18.5 percent of bottom-99% cases and less than 14 percent of top-1% cases (recall that I have excluded consolidated cases, such as MDLs). Finally, although about one in four bottom-99% cases comes from a prisoner petition, only one in ten of top-1% cases does.

This analysis shows that case types vary widely in their relative representation among the most intensely litigated cases. The patterns are pretty much what one would expect, with the intensity of environmental cases being perhaps the only surprise relative to what many generalist observers might have thought ex ante.

We have now seen that (i) there is wide dispersion across case types in the average number of docket entries (Table 2), and (ii) some case types are extremely over-represented in the top 1% group (Table 3). A natural question is whether the dispersion in numbers of docket entries primarily is an artifact of category over-representation among the most intense cases. For example, one might think, "Even though the

Intense 4 have relatively few cases, cases in these groups might be so intensely litigated that they are all at the top of the top 1%."

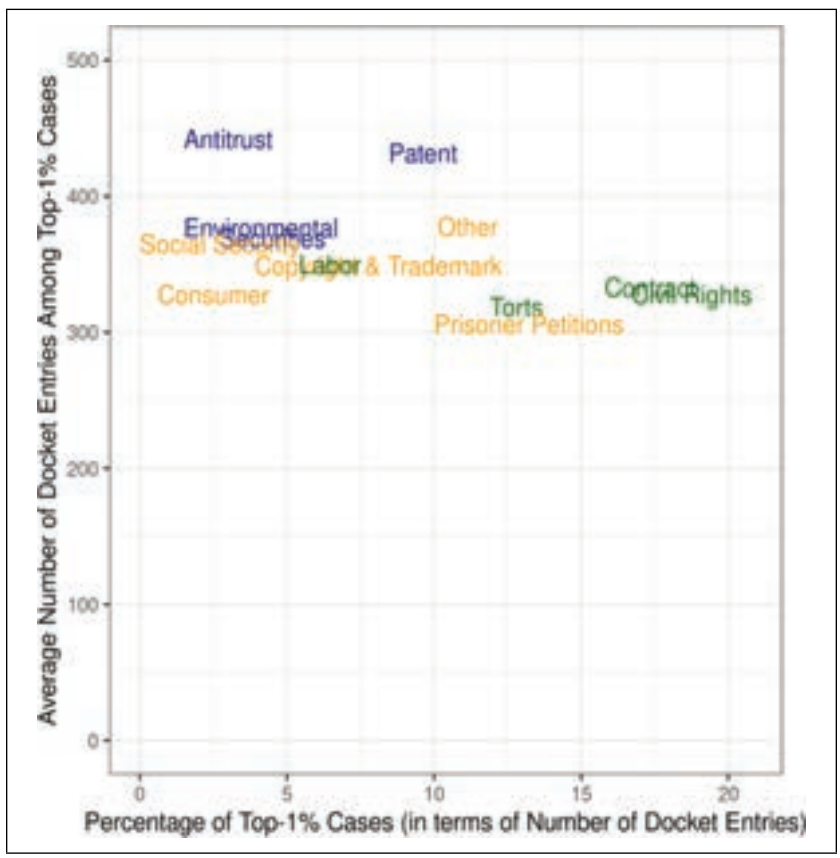
On the other hand, several of the categories that are under-represented in the top 1% of intense cases—what I will call the "Big 4" of civil rights, contract, labor, and tort—comprise very large numbers of cases. Together, these four categories account for a touch more than 300,000 cases—a slight majority of the total analysis set. They also make up the majority of the top-1% cases. In fact, as a group they have an over-representation index value slightly above 1.[55] So unless top-1% cases among the Big 4 substantive areas are lightly litigated *relative* to other top-1% cases, the sheer size of these case categories means that the most intense cases among them will be of great empirical importance.

Figure 3 presents some evidence on this question. The figure presents data by substantive areas. The figure's horizontal axis measures the percentage of cases in the top 1% of docket entry intensity that each substantive area includes; in other words, the horizontal axis measures the "Top 1%" column of Table 3.[56] The figure's vertical axis represents the average number of docket entries among each substantive area's top-1% cases.

55. Together, Intense 4 cases make up 53.5% of bottom-99% cases and 58.8% of top-1% cases. Taken collectively, then, Intense 4 cases have an over-representation index value of 1.1 (58.8 divided by 53.5).

56. Each substantive area's horizontal-axis value is given by the left-most value in the label, e.g., "Social Security" is aligned so that the leading "S" lines up with the 0 value for the horizontal axis.

Figure 3: Average Number of Docket Entries Among Top-1% Cases, and Percentage of Cases in Top 1%, By Substantive Area



The figure's pattern suggests there is a negative association between substantive areas' average number of docket entries among top-1% cases and the areas' percentage of top-1% cases. In other words, substantive areas whose top-1% cases are especially intensely litigated tend also to have relatively few cases overall. For each Intense 4 substantive area, the average number of docket entries among top-1% cases exceeds that for any Big 4 substantive area; this is evident from the fact that the Intense 4 labels are all higher than any of the Big 4 labels on the graph. That said, the average number of docket entries lies between 300–450 for all eight substantive areas, which is not an especially large range given the differences in representation

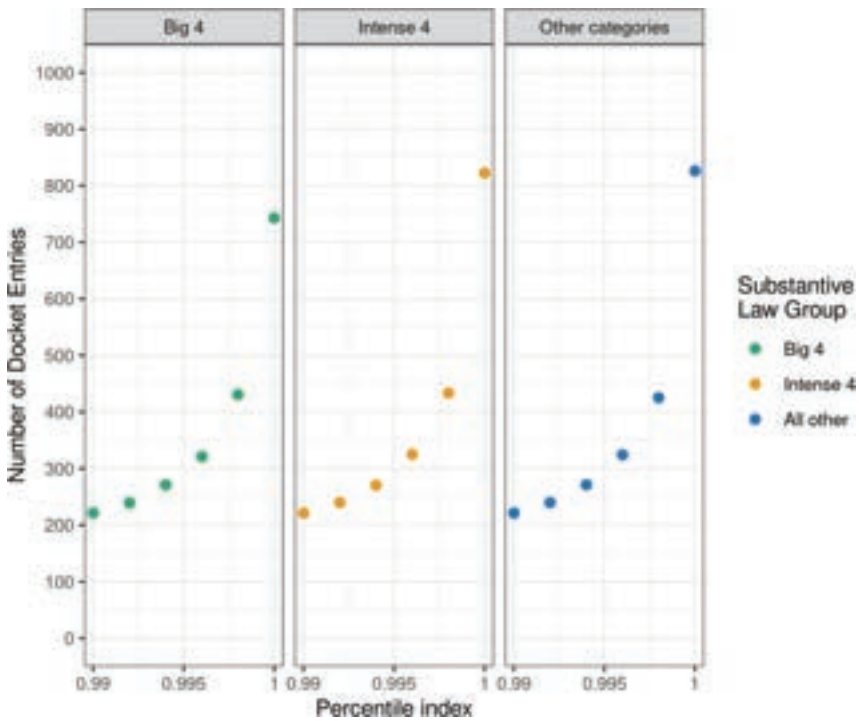
percentages.[57] Further, as noted above, the Big 4 cases have a much larger footprint among top-1% number-of-docket-entries cases: the Big 4 share of top-1% cases is more than three-and-a-half times the corresponding share of Intense 4 cases. So, the greater intensity of highly intense cases in the Intense 4 category is not enough to overcome the much greater number of highly intense Big 4 cases. The end result is that the size difference of the Big 4 category more than offsets the differences in average docket-entry intensity among the most intense Big 4 and Intense 4 cases.

Figure 4 provides more detailed evidence about the distribution of docket intensity within the top-1% cases. The figure contains separate graphs for each of the Intense 4, Big 4, and remaining categories taken as a collection, with each graph showing average numbers of docket entries within the top-1% cases in the group, measured using six bins for each substantive law group.[58] Thus, for cases close to the 99% cutoff of 216 docket entries, each of the three groups shows an average of roughly that value. As we move up the horizontal axis from the 99th percentile, the average docket intensity for each group rises. The increase is similar for the three sets of substantive areas until we get to the very top category. For this category of the most intensely litigated cases within each substantive area group, we do see that Big 4 cases have a noticeably lower intensity (a bit lower than 750 entries on average) than Intense 4 cases (roughly 825).

57. Among categories with non-zero top-1% representation percentages, the top-1% representation percentage ranges from 0.6% for Consumer cases to 20.8% for Civil Rights.

58. To be clear, the data I used to make this figure involve all cases in the top 1% of docket entry intensity as measured across the full set of cases in the analysis set. Thus, cases are included if and only if they have more than 216 docket entries.

Figure 4: Percentiles of the Number of Docket Entries by Substantive Law Group



3. Summary on Intensity

The analysis of docket intensity in this section establishes several broad facts. First, there is tremendous variation in docket intensity across the analysis set of 566,315 cases. Median docket intensity is quite low, at just 17 docket entries. But the most docket-intense cases do involve very large numbers of docket entries. The 99th percentile of 216 seems large, but even so is dwarfed by the 99.9th percentile of 533 entries. And—even though I have excluded cases involving indicia of consolidation—the very most docket-intense cases involve truly enormous amounts of activity.

Second, there is clear evidence that the "usual suspect" substantive areas of patents, antitrust, and securities have relatively more docket entries than other areas of litigation; environmental cases join these three groups to form a discernible Intense 4 when it comes to average numbers of docket entries. The Intense 4 substantive areas unquestionably have both greater average numbers of docket entries

and greater frequencies of representation in cases with top-1% docket intensity.[59]

Third, however, because there are comparatively few of these Intense 4 cases, substantive area is of relatively limited use in explaining total litigation activity as measured by docket intensity. This might seem paradoxical in light of the second fact. But there are many more cases in substantive areas outside the Intense 4 than in it. And although these other case types are less frequently represented among the highly-intense top-1% cases, they are more than highly enough represented to make their contribution to total docket activity multiples of the Intense 4 areas' contribution. The Big 4 areas of civil rights, contract, labor, and tort are slightly *over*-represented in the top-1% cases, so that top-1% cases in the Big 4 areas number more than triple the number of top-1% Intense 4 cases. Although top-1% Intense 4 cases are more intense than top-1% Big 4 cases, this difference turns out to be modest in overall importance.

The fourth fact follows from the third one: although litigation intensity is associated with substantive area, intensity is itself transsubstantive. If what we care about is a case's intensity as measured by the number of docket entries, simply knowing that a case is in the top 1% is much more useful information than knowing its substantive area. One way to see this is to observe that the *least-intense* case among top-1% docket intensity cases, which would have the 99th percentile level of 216 docket entries, has more than 2.5 times the average intensity of patent cases.

Finally, just as intensity is transsubstantive, so, too, is substance importantly transintense. Even substantive areas of law that tend to have relatively low average docket entries have some tremendously intense cases, just as substantive areas with high average docket intensity have many cases that wind up with few docket entries. This observation is surely a tribute both to the variety of disputes within any substantive area, and also to the dynamics of litigation settlement. Some disputes settle early, even though they would be highly intense if litigated to judgment, or even through the summary judgment phase. This topic is beyond the scope of the present Article.[60]

---

59. Of course, the latter fact helps contribute to the former.

60. For work about selection in litigation, see Jonah B. Gelbach, *The Reduced Form of Litigation Selection Models and the Plaintiff's Win Rate*, 61 J. L. & ECON. 125 (2018); Gelbach, *Material Facts*, *supra* note 29; Gelbach, *Locking the Doors*, *supra* note 29; George Priest & Benjamin Klein, *The Selection of Disputes for Litigation*, 13 J. LEGAL STUD. 1 (1984); Daniel Klerman & Yoon-Ho Alex Lee, *Inferences from Litigated Cases*, 43 J. LEGAL STUD. 209 (2014); Hubbard, *supra* note 29, at 35.

B. Complexity

This section addresses case complexity. Although most observers would probably say they know complex litigation when they see it, I am unaware of any widely accepted definition of the term; the most recent edition of the *Manual for Complex Litigation* even proclaims on its first page that "the term 'complex litigation'" is not "susceptible to any bright-line definition."[61]

Complexity is thus less simple to measure than intensity. Whereas intensity seems straightforwardly measured by the number of docket entries, complexity comes in multiple flavors. A case might involve a very complex underlying issue of *substantive* law, as when the scope of a scientifically technical patent claim is at stake. Or a case might involve multiple parties engaging in strategic use of procedural devices—motion practice, discovery, something else, or all of the above. Substantive and procedural complexity often will be associated with each other, because substantively complex cases will have more scope for strategic procedural maneuvering. Nevertheless, these concepts are distinct.

My data provide no particular ability to assess substantive complexity; as noted earlier, I lack the underlying case documents that would shed light on the substance of disputes. But it is difficult to imagine a definition of "complex litigation" that would be widely accepted without at least including party numerosity: with more parties comes more possibility of divergent interests, more persons with the right to take discovery, and potentially more attorneys. Thus, as a first pass, I use the number of parties as a measure of case complexity.[62] I then turn to two complementary measures based on the interrelated structure of docket entries.

61. MANUAL FOR COMPLEX LITIGATION (FOURTH) 1 (2004).

62. I have information on each party's set of attorneys, as well as their firms and locations (and I can tell when parties are unrepresented via the absence of any listed attorneys), but in the interests of brevity I do not use such information here.

1. Number of Parties

Table 4 presents some basic statistics on the distribution across cases of the number of parties. Although many cases have just two parties, even the median exceeds that: *half* of all cases have at least three parties. The table shows that most cases have single-digit numbers of parties. However, at the upper end of the distribution, there are cases with enormous numbers of parties. The 95th percentile is 16 parties, the 99th is 33, and the 99.9th is a whopping 94 parties.

Table 4: Descriptive Statistics for Number of Parties

Percentile or Other Statistic	Cases in Analysis Set
10	2
25	2
50	3
75	5
90	10
95	16
99	33
99.9	94
Mean	5
St. Dev.	15

Consider next Table 5, which describes how the number of parties varies across substantive areas of litigation. Antitrust leads the pack with an average of 13 parties per case, even in cases lacking indicia of consolidation, and the securities category is close behind with ten. Unlike the situation with docket intensity, tort cases are high up the ladder, with a third-ranked average of nine parties per case—again, even though I exclude cases with indicia of consolidation. Environmental cases have an average of eight parties, which is good for the fourth rank. Interestingly, patent cases have just four parties on average—similar to

prisoner, contract, and consumer cases.[63] Thus, three of the Intense 4 from our study of docket intensity are represented among the four most complex cases as measured by average party numerosity, with tort and patents trading places relative to docket intensity.

Table 5: Average Number of Parties Among Cases in Analysis Set

Substantive Law Category	Average Number of Parties
Antitrust	13
Securities	10
Tort	9
Environmental	8
Labor	6
Copyright & Trademark	6
Civil Rights	5
Other	5
Immigration	5
Prisoner Petitions	4
Patent	4
Contract	4
Consumer	4
Social Security	2

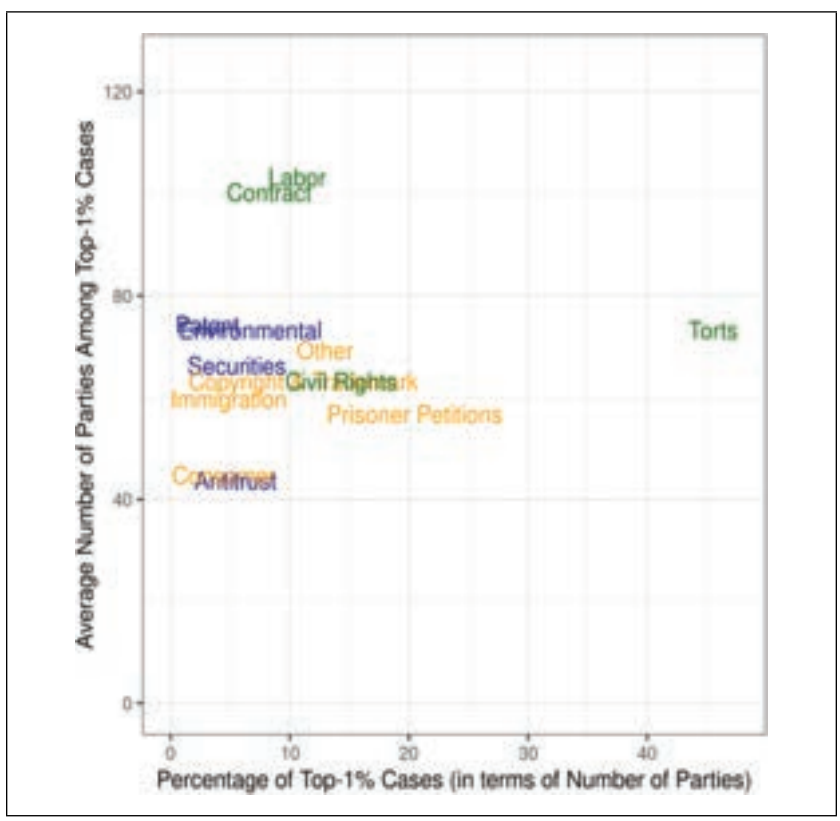
Figure 5 assesses the role of cases with the most parties, repeating Figure 4's analysis for docket intensity. The figure's horizontal axis now measures the percentage of top-1% cases, in terms of the number of parties, for which each substantive area accounts. Tort is the unmistakable outlier, accounting for almost half the cases with

63. Note that this is *not* the result of the joinder restrictions in 2011's America Invents Act, 35 U.S.C. § 299, because all cases in my analysis set were filed between 2005 and 2007.

top-1% numbers of parties—even after excluding cases with indicia of consolidation.

The vertical axis measures the average number of parties in those cases of each substantive area that are top-1% cases in terms of party numerosity. The greatest average is for labor cases (a bit above 100), and the least is for antitrust cases (less than 50). Interestingly, although tort cases are far and away the most frequently observed substantive area in the top 1%, the average number of parties in those many-party tort cases is not an outlier among substantive areas: tort cases are roughly on par with several other substantive areas, and well below both labor and contract cases.

Figure 5: Average Number of Parties in Top-1% Cases and Percentage of Top-1% Cases Accounted for by Cases in Substantive Areas



The discussion above indicates that the Big 4/Intense 4 taxonomy is not as clear-cut as in the intensity discussion, e.g., given that patent

cases have few parties on average and tort cases many. But they are still useful because they allow us to see whether complexity, as measured by the number of parties, follows the same patterns as docket intensity.

For brevity (concededly in short supply here), I will not display an analog to Table 3, which detailed the extent of over- and under-representation of cases in the top 1% of docket-intense cases. I will simply say that the considerable variation in substantive areas' shares of top-1% party-numerosity cases is matched by substantive-area variation in bottom-99% cases. Thus, it is clear that party numerosity varies substantially not only across substantive areas, but also within them.[64]

2. The Network of Docket Entry Links

This section proposes a novel alternative approach to measuring complexity, using the mathematical theory of graphs and networks to characterize docket activity within cases. To understand the idea, it will help first to consider some pictures that represent litigation activity in two cases that I selected to illustrate the ideas presented here.[65]

The first case is captioned *Currie et al. v. Dollar General Corp* and was removed to the Northern District of Florida on June 3, 2005.[66] The docket report indicates that it had three parties—plaintiffs Mary Jo Currie and Adolphus Currie, and defendant Dollar General Corporation. Its PACER nature of suit code is 360, "Torts: Other Personal Injury" (the Key Nature of Suit field additionally lists negligence). The case was terminated on November 14, 2005, about five months after its origination via removal. The docket report indicates that it was terminated via a stipulation of dismissal with prejudice by the defendant, presumably pursuant to Rule 41(a)(1)(A)(ii).[67] Including unnumbered docket entries entered by the clerk, the case included 32 docket entries—roughly the overall median. In other words, at least

64. An exception to this statement is for Social Security cases, which are entirely unrepresented in the top 1% of party numerosity. To the extent that these cases often involve appeals of administrative denials of benefit applications, by nature they involve a single plaintiff suing the Commissioner of Social Security, and thus only two parties.

65. That is, there is no other reason relevant to this Article why these two cases are of particular interest.

66. The case was originally filed in the Circuit Court of the Third Judicial Circuit in and for Dixie County, Florida, as case number 05-0091CA. *See* Notice of Removal at 1, Currie v. Dollar General Corp., No. 1:05-CV-00099 (N.D. Fla. June 3, 2005), ECF No. 1.

67. FED. R. CIV. P. 41(a)(1)(A) ("the plaintiff may dismiss an action without a court order by filing: . . . (ii) a stipulation of dismissal signed by all parties who have appeared").

based on the docket facts I've considered in this Article, *Currie* is a pretty unremarkable case.[68]

Here is docket entry number 3, filed on June 7, 2005: "MOTION to Strike affirmative defenses by MARY JO CURRIE, ADOLPHUS CURRIE. (bkp, Gainesville) (Entered: 06/08/2005)".[69] On June 23, 2005, the defendant responded in docket entry number 8, which reads: "RESPONSE to Motion re 3 MOTION to Strike Defendant's Affirmative Defenses filed by DOLLAR GENERAL CORPORATION. (BERGIN, RUSSELL) (Entered: 06/23/2005)." The "3" in entry number 8 appears as a hyperlink on PACER and is discernible via xml tags in my data. Thus it is possible to see that entry number 8 is linked to entry number 3. More than one docket entry may be linked to any given entry. For example, on July 12, 2005, the following entry was docketed: "ORDER denying 3 Motion to Strike. Signed by Judge STEPHAN P MICKLE. (llt, Gainesville) (Entered: 07/14/2005)."[70] Other docket entries in the case did not have such links. For example, on October 17, 2005, an entry was docketed notifying the parties that a mediation was scheduled,[71] and no other docket entry ever linked to this one.

We can construct a formal graph of the docket activity by drawing a picture in which each docket entry is represented by a *node*, essentially a dot. Node-to-node links—known in graph theory as *edges*—connect docket entries that are linked via references like those to docket entry number 3 in the *Currie* case just discussed. Together, a set of docket entries that can be reached by an unbroken *path* of links is known as a *connected component*. Nodes that are not linked to any other nodes are called *singletons*.

Figure 6 presents a rendering of the graph for the *Currie* case. The activity related to the plaintiffs' Motion to Strike in docket entry number 3 corresponds to the component of three connected nodes highlighted by a circle. The node in the middle is docket entry number 3, and it is connected to each of the two other nodes already discussed. The hollow circles in the graph correspond to singletons; this graph exhibits seven. Including the highlighted one, the graph has a total of seven connected

68. Inspection of the defendant's Memorandum of Law supporting its Notice of Removal, which I downloaded from Bloomberg Law, revealed that the case involved the plaintiffs' allegation that a negligently installed clothes rack at a Dollar General store had fallen on Mary Jo Currie and caused her bodily injuries and loss of consortium damages to her co-plaintiff husband. *See* ECF No. 1 at 6.

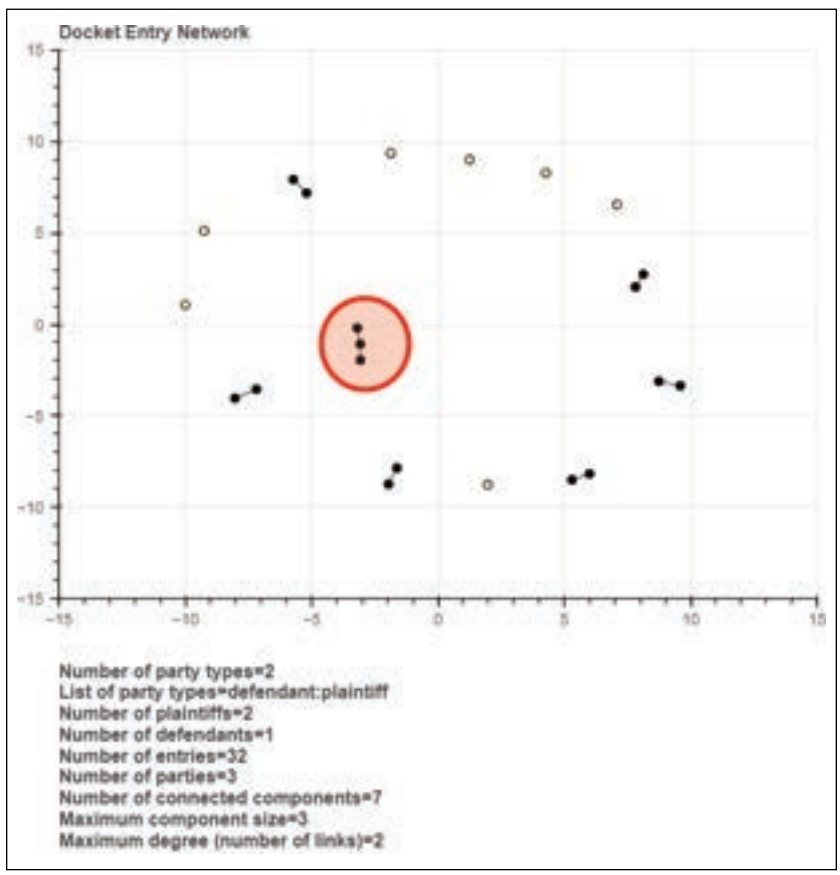
69. ECF No. 3.

70. ECF No. 9.

71. ECF No. 19, Notice of Mediation re Scheduling, 10:00 a.m. 11/4/05, Jacksonville, FL (llt, Gainesville) (entered Oct. 17, 2005).

components, i.e., there are seven subparts of the litigation that involve multiple docket entries.[72]

Figure 6: Network Graph for *Currie et al. v. Dollar General Corporation*



I use the number of connected components for each case as a measure of complexity. This measure of complexity is plausible for at least two reasons. First, the fact that a docket entry links to another

72. In addition to the 22 nodes depicted in Figure 6, the *Currie* case involved an additional 10 docket entries that correspond to unnumbered entries. My discussions to date indicate such entries are typically of limited moment, involving scheduling or other administrative issues coming from the clerk's office (e.g., the need for the plaintiff's counsel to petition for admission to the federal court bar, as evidenced in an unnumbered docket entry from July 14, 2005). Although I count such entries in the docket number intensity analysis above, it is simpler to conduct this section's graphical analysis without accounting for them.

one generally suggests that there is some depth or breadth to litigation activity, as when a party files an opposition to its adversary's motion, or amends its own pleading, and so on. Second, more complex sets of issues—whether procedural or substantive—in a litigation may lead to more components, as procedural areas of contest proliferate. Thus we can expect that more complex cases will tend to have more connected components—each serving as a kind of local region of conflict, or even of coordination, in the case.[73]

The *Currie v. Dollar General* case was a straightforward personal injury case with two plaintiffs and one defendant, which was settled following mediation[74] and before the deadline for discovery to end,[75] with a total time in federal court of under six months.[76] This non-complex case involved 32 total docket entries (counting unnumbered ones[77] and seven connected components).

For comparison's sake, now consider the case of *Teamsters Local 617 Pension & Welfare Funds v. Apollo Group, Inc*. This was a securities class action filed on November 2, 2006, in the District of Arizona. All told, it had 21 distinct parties—14 distinct plaintiffs and seven distinct defendants—some of whom were terminated from the case before it ended.[78] Including appellate activity, the case went on for roughly a decade. It had a total of 165 district court docket entries. By any standard, it is the sort of case that can be considered both intensely litigated and complex, including in the "complex litigation" field sense.

Figure 7 provides a graphical depiction of the Apollo Group securities case's docket network. As with the *Currie* case, the lighter circles correspond to singletons, and the darker ones correspond

73. The number of connected components is reasonably related to intensity because all else equal, a case will need to have more docket entries to have more connected components. To see this, imagine that the *Currie* case is ongoing. Given what has already occurred, as depicted in the figure, the case can add connected components only if (i) some new docket entry is added and links to an existing singleton, or (ii) multiple additional docket entries arrive and involve new links that form additional connected components.

74. *See* docket entry description of ECF No. 20, Mediation Report - Settled (deb, Gainesville) (entered Nov. 10, 2005).

75. *See* docket entry description of ECF No. 17 (stating "Discovery due by 12/6/2005").

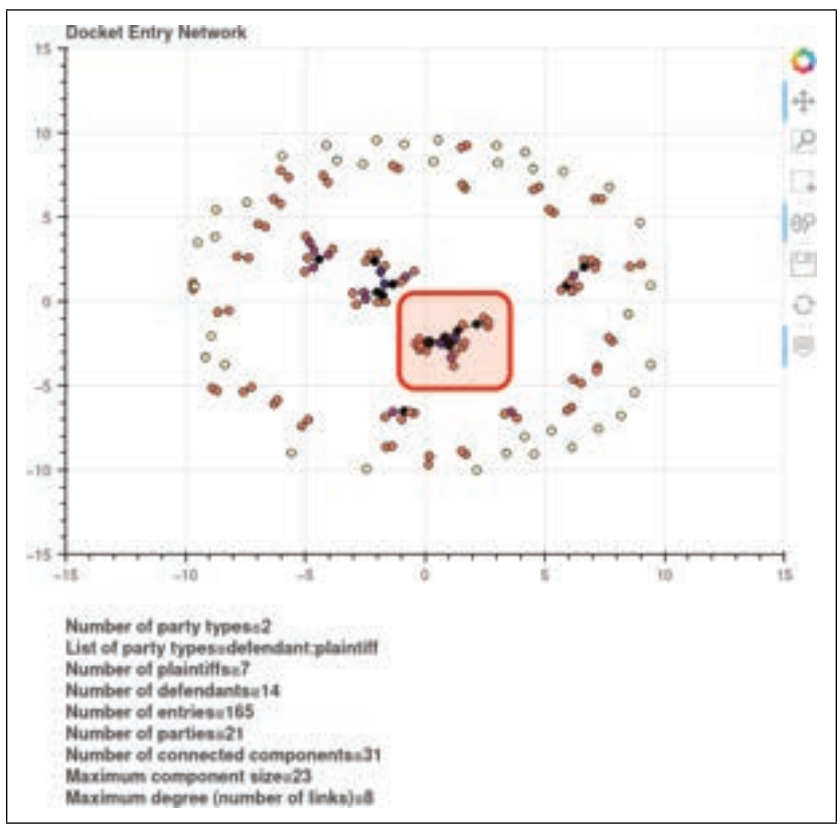
76. The first docket entry was dated June 3, 2005, and the order of dismissal was dated November 30, 2005.

77. *See supra* note 72.

78. There was some overlap in the attorneys representing various parties. The pattern of attorney involvement in cases is another possible future source for characterizing litigation.

to connected components.[79] There are evidently a lot of connected components—distinct clusters of related docket entries. As the figure caption indicates, *Apollo Group* had 31 of them.

Figure 7: Network Rendering for *Teamsters Local 617 Pension & Welfare Funds v. Apollo Group*

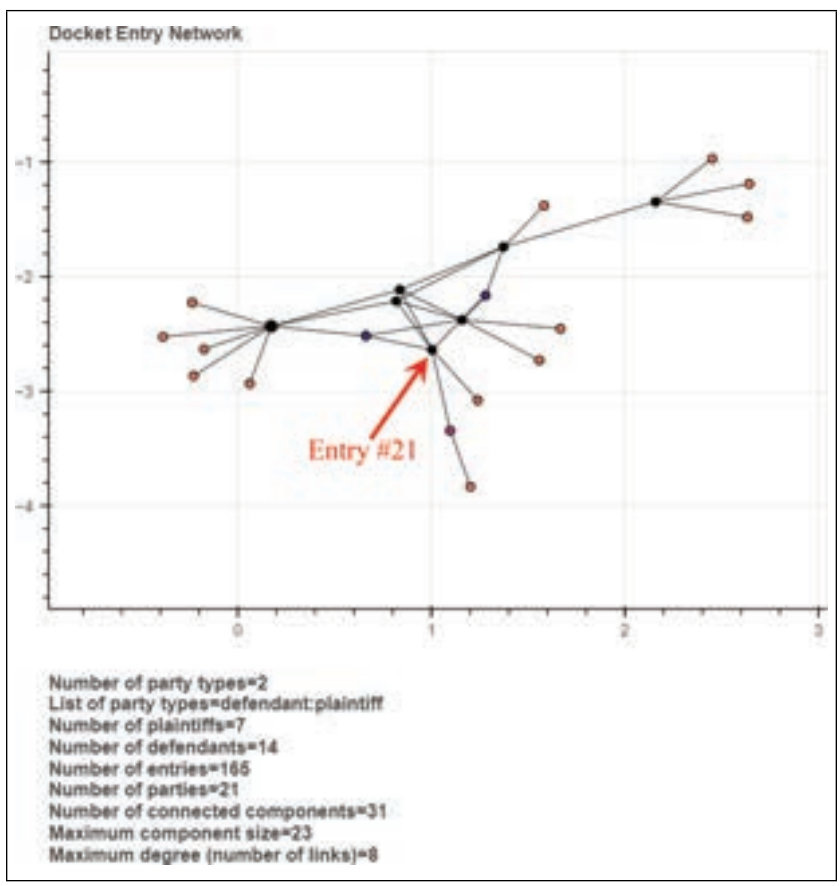


A casual look at Figure 7 suggests that there are some quite large components, rather than just the two- or three-node bunches we saw in *Currie*. It helps to get a close-up of the highlighted part of Figure 7, which is what Figure 8 offers. The highlighted part of the graph corresponds to a single component involving litigation activity over who would be named lead plaintiff in the case; this component was touched

79. I have set up this figure so that docket entries with more links to others—a greater *degree*, as such connectedness is known in the theory of networks—are more darkly colored. Beyond that, I do not use degree as a measure of anything in this Article.

off by docket entry number 21: "MOTION to Appoint Teamsters Local 617 Pension and Welfare Funds as Lead Plaintiff and Approval of Lead Counsel by Teamsters Local 617 Pension and Welfare Funds. (Saltzman, Jay) (Entered: 01/03/2007)".[80]

Figure 8: Zoom-In View of Largest Component in *Teamsters Local 617 Pension & Welfare Funds v. Apollo Group*



The component of entries related to this docket entry is the largest one in the case, with 23 different docket entries connected via links to one or more entries in this component. By way of comparison, recall that the median number of docket entries across all cases in my analysis set was 32.

80. ECF No. 21.

Cases will necessarily have at least as many docket entries as the size of their largest component, of course, so one would expect the size of the largest component to be positively associated with entry-intensity. In addition, though, it is natural to think that a case with a larger largest component may entail either a wider range of issues or more complexity in resolving the most complex ones. Thus, the size of the largest component is a natural measure of complexity. Accordingly, I use this graph feature together with the number of connected components as additional measures of case complexity.

Table 6 reports descriptive statistics for these two variables side by side. Percentiles at the median or below are small for both, with half of all cases having no more than two connected components and having a maximum component of no more than three docket entries. In fact, for one in four cases, there is no connected component, i.e., no docket entries are linked; this means those cases contain nothing but singletons. Because a singleton is not a connected component, this means these cases' maximum component size is zero.

Table 6: Descriptive Statistics for Number of Connected Components and Maximum Component Size

Percentile or Other Statistic	Number of Connected Components	Maximum Component Size
10	0	0
25	0	0
50	2	3
75	4	5
90	8	11
95	12	17
99	26	38
99.9	62	101
Mean	3	5
St. Dev.	6	9

As we saw with the *Apollo* case, both the number of connected components and the maximum component size can become much larger. *Apollo* itself would be quite high up the distribution in both variables:

Apollo's 31 connected components put it between the 99th and 99.9th percentiles (which are 26 and 62), and its maximum component size of 23 lies between the 95th and 99th percentiles (17 and 38).

Next, consider variation across substantive areas. Table 7 reports the average number of connected components and maximum component size separately by substantive area of law. I have listed the areas in descending order of the average number of connected components. Patents, environmental, securities and antitrust cases—our Intense 4 in terms of docket intensity—also have the greatest average numbers of connected components. There is considerable variation between the top category, patent cases, and antitrust cases, which have only half the 10 connected components that patent cases have on average.

Table 7: Average Number of Connected Components and Maximum Component Size

Substantive Law Category	Number of Connected Components	Maximum Component Size
Patent	10	11
Environmental	7	11
Securities	7	9
Antitrust	5	7
Civil Rights	5	7
Contract	4	6
Consumer	3	4
Copyright & Trademark	3	5
Labor	3	4
Other	3	5
Prisoner Petitions	3	4
Social Security	3	4
Immigration	2	4
Tort	2	3

The Intense 4 also are the top group for maximum component size. Patent cases' maximum component averages 11 docket entries, which is matched by environmental cases; securities and antitrust cases come in at nine and seven respectively. For both complexity variables, civil rights cases lead the pack of non-Intense 4 categories. Most other substantive area groups have noticeably lower average values of the network-based case complexity variables.

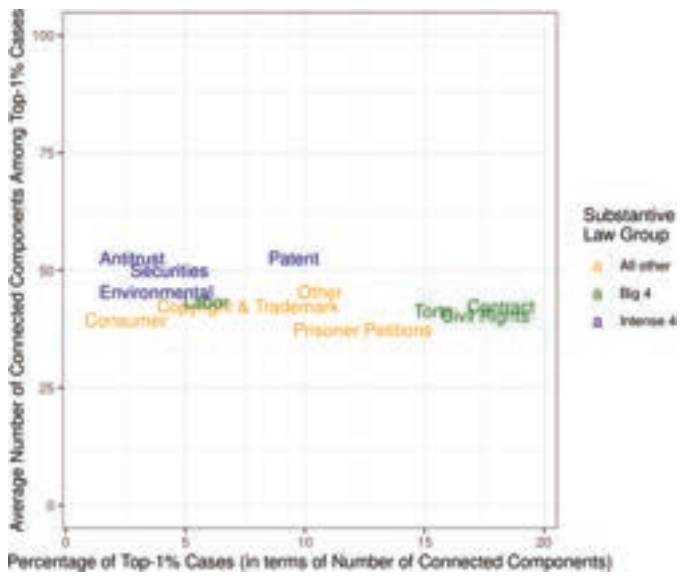
Figure 9 investigates variation across substantive area of law in average complexity measures among the cases at the top. It does so by repeating the comparisons made above between the average value of a variable of interest among top-1% cases (vertical axis) and the percentage of top-1% cases associated with each substantive area of law.

Panel (a) of the figure shows that the average number of connected components among top-1% cases varies within the range of 35 to 55, with substantive area groups' shares of top-1% cases varying across the range from about 5% to about 20%. Except for labor cases, Big 4 cases once again tend to make up large percentages of top-1% cases, and Intense 4 cases tend to make up relatively small percentages.

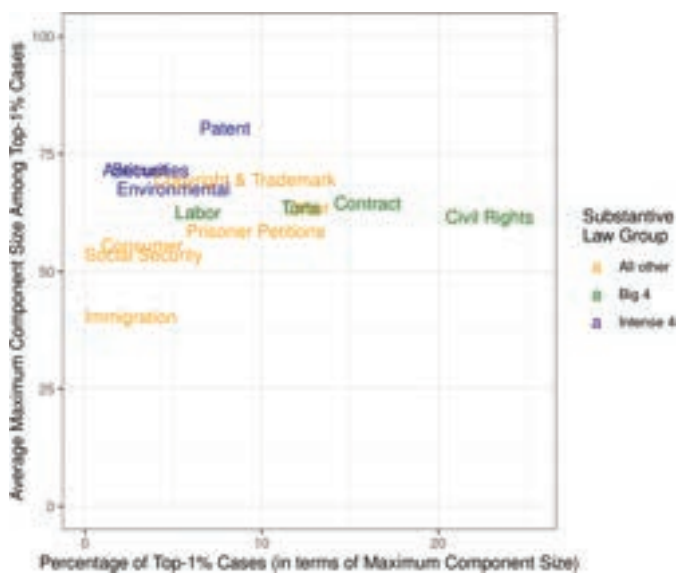
There are a few differences in the patterns for maximum component size in Panel (b) of Figure 9. First, there is a higher upper range across substantive areas of the average value among top-1% cases—with patent cases among the most complex 1% on this measure having an average size of the maximum component equal to roughly 80. To reiterate, a case with a maximum component size of 80 is one that has some aspect of the litigation in which there are 80 distinct docket entries linked together through a set of references to each other. That is a lot!

Second, the Big 4 substantive areas of labor, tort, contract, and civil rights are more spread out in the degree to which they make up the top 1% of cases in terms of maximum component size; whereas tort, contract and civil rights all cluster to the right of the graph for the number of connected components in Panel (a) of Figure 9, they are evenly spread across the horizontal axis of Panel (b). Finally, it appears that there may be more of a negative relationship between a substantive area's share of top-1% cases and the mean of its maximum component size among those cases. This is especially clear if we look only at the Intense 4 cases—patent, securities, antitrust and environmental—and Big 4 ones. Still, as before, simple calculations show that the Intense 4 cases' greater share of top-1% cases more than makes up for their relatively lower values among those cases; although Intense 4 cases among the top-1% are more complex than corresponding top-1% Big 4 cases, that difference is outweighed by the greater representation of Big 4 cases in the top 1%.

Figure 9: Average Numbers of Connected Components and Maximum Component Size Among Cases in Top 1% of Each Variable, Compared to Share of Top 1% Cases by Substantive Area of Law



(a)



(b)

3. Summary on Complexity

Three points stand out from the discussion of complexity as measured by party numerosity and the measures based on docket networks. First, as the patent-tort reversal among the number of parties illustrates, patterns for docket intensity and party numerosity are not universally the same; nor was there a major outlier on the right side of Figure 4, the way there is in Figure 5. Second, just as docket intensity varies significantly across substantive areas, complexity also varies a lot. And third, flipping the comparison, there is significant variation in complexity measures within nearly all substantive areas. Thus, just as we saw in section 0 that substance is transintense, it is evident from this section that substance also is transcomplex.

C. Association Between Intensity and Complexity Measures

This section thus engages the question of whether my intensity measure, the number of docket entries, captures something distinct from the measures of complexity I use. This is useful for two reasons.

First, my measure of intensity and my measures of complexity will naturally be related to each other. As noted above, cases with more connected components (a measure of complexity) will tend to have more docket entries (my measure of intensity) simply because each connected component must have at least two docket entries. Similarly, the maximum component size (a measure of complexity) has more entries than any other component in the case, so a case with a greater maximum component size will tend to have more entries (my measure of intensity). Finally, if distinct parties are represented by different counsel, then a greater number of parties (a measure of complexity) will be associated with more docket entries (my measure of intensity).

Second, if the intensity and complexity variables I use were *too* correlated with each other, that would suggest that one could learn all there is to know from the collection of them using any one of them. In other words, it would suggest that once we know what there is to know about, say, docket-count intensity, we would have little to learn from studying my measures of complexity. It is thus valuable to assess the distinctiveness of the statistical information in these different variables.

A clue that intensity and complexity are distinct may be found in the fact that some litigation areas tend more to extreme values in my intensity variable than in my complexity variables. The poster child for this observation is patent litigation, which is both (i) the most over-represented substantive area in the top 1% of intensity as measured by the number of docket entries and (ii) not notable in terms of

the number-of-parties complexity measure.[81] The remainder of this section investigates more broadly.

A first approach is to ask how correlated my intensity measure and the various measures of complexity are. Table 8 reports the Pearson correlation between each pair of variables.[82] To read this table, one reads down the rows. The right-most entry in each row equals 1, because any variable equals itself, which is a positive and perfect linear relationship.

The second row of the first column reports the correlation between a case's number of parties (the row variable) and its number of docket entries. The value of 0.12 tells us that there is a positive linear relationship between these two variables, but it is small—and perhaps even surprisingly so. Continuing down the first column's rows, we see that the number of connected components and the maximum component size—with correlation coefficients of 0.87 and 0.72, respectively—have much more substantial relationships with the number of docket entries. In light of the discussion above, this is as expected. Given this strong relationship and the weak one between the number of parties and the number of docket entries, it is not surprising that the network-based complexity measures have weak correlations with each of the number of parties (as demonstrated by the second column's entries, which show that both these correlations are less than 0.10). Finally, the Pearson correlation of 0.6 reported in the third column's fourth row shows that the two network measures are themselves relatively highly correlated.

81. Again, this is not the result of the America Invents Act's joinder restrictions, 35 U.S.C. § 299, because all cases in my analysis set were filed years before that Act.

82. The Pearson correlation between two variables is the expected value of the product of the variables after centering and standardizing each variable, which is done by subtracting each variable's mean and dividing the result by the variable's standard deviation. The Pearson correlation always lies between -1 and 1. A value of 0 indicates the absence of any linear relationship between the two variables, and values of 1 and -1 indicate a perfect linear positive or negative relationship, respectively. Values between 0 and 1 indicate that the two variables have some degree of positive linear relationship, i.e., increases in one are associated with increases in the other; values between 0 and -1 indicate the presence of some degree of negative relationship (increases in one variable are associated with decreases in the other).

Table 8: Correlation Between Intensity and Complexity Measures

	Number of:				Connected Components per docket entry:	
	Docket entries	Parties	Connected components	Max component size	Number	Max
Number of docket entries	1					
Number of parties	0.12	1				
Number of connected components	0.87	0.09	1			
Max component size	0.72	0.07	0.6	1		
Connected components per entry	0.08	-0.05	0.32	0.13	1	
Max component size per entry	-0.01	-0.04	0.04	0.32	0.55	1

One natural question is whether the strong correlations between the network-based measures and the number of entries arise purely from the already-cited mechanical reasons why these measures can be expected to move together. To assess this possibility, I created standardized versions of the network-based measures, by dividing each by the total number of docket entries. This will eliminate any simple linear relationship between the network-based measures and the number of docket entries. Looking at the last two rows of Table 8, we see that the per-entry versions of the network-related measures have small correlations, of mixed signs, with the number of entries and the number of parties—indicating that standardization creates complexity measures that have essentially no linear relationship either of these conventional variables. The standardized network-based measures are, nevertheless, reasonably well correlated with their non-standardized versions (correlation coefficient of 0.32 in each case), and they are also relatively highly correlated with each other (correlation coefficient of 0.55).

Another way to investigate the closeness of intensity and complexity measures is to ask how likely cases that are in the extreme part of the distribution for one variable are to be in the extreme part of another variable's distribution. Broadly speaking, we can address this question by asking whether being in the top 1% for one variable indicates a high likelihood of being in the top 1% for the others.

Table 9 addresses this by reporting the share of cases that are in the top 1% of the distribution for variables in the table's rows, given that a case is in the top 1% for the indicated column variable. The second row and first column show that only 9% of cases that are in the top 1% of the number of *entries* are in the top 1% of the number of *parties*. The share leaps to 0.65 (two out of three) and 0.44 (two out of four) for the number of connected components and the maximum component size—not surprising, given how correlated the number of entries is with each of these network-based variables.

Table 9: Share of cases in Top 1% of Row Variable Distribution, Given Presence in Top 1% of Column Variable Distribution

	Number of:				Connected Components per docket entry:	
	Docket entries	Parties	Connected components	Max component size	Number	Max
Number of docket entries	1					
Number of parties	0.09	1				
Number of connected components	0.65	0.08	1			
Max component size	0.44	0.06	0.33	1		
Connected components per entry	0.00	0.01	0.00	0.00	1	
Max component size per entry	0.00	0.02	0.00	0.04	0.32	1

The cases with top 1% values of standardized network-based variables, though, are almost never in the top 1% of the distribution of the number of entries;[83] the same goes for their representation among cases in the top 1% of the number of parties—and even among cases in the top 1% of the unstandardized network-based variables.

Taken together, these findings suggest that my measures of intensity and complexity pick up distinct features of litigation. If we accept both that the number of docket entries is a reasonable measure of intensity and that the other variables are reasonable measures of complexity, then it seems "intensity" and "complexity" are not simply synonyms. Rather, they pick up possibly but not necessarily related aspects of litigation. Thus, although I believe it is helpful to think of *intexity* as the combination of two connected features of a litigation, it is useful to remember that they are distinct as well.

III. Discussion: Data and Procedure Policy

Part II's empirical picture resonates for procedure policy discussions. First, the depth of detail I muster shows the importance of using comprehensive docket data rather than information specific to motion type,[84] district-[85] or substance-specific studies.[86] This kind of evidence is simply impossible to provide without access to massive amounts of data, without which the top of any distribution will consist of few cases, especially when we focus on smaller substantive areas.

That underscores the importance of improving research access to data.[87] Our judicial system hamstrings researchers' capacity to investigate its nature and civil justice performance, because—by choice—the judiciary has limited access to federal court data to those willing and able to pay enormous amounts to acquire data. One estimate that is now more than half a decade old held that it would cost $1 billion to download all the information in

83. The zero shares reported in the table are accurate to two digits, but there are a relatively small share of cases that are in the top 1% of the number of entries, on the one hand, and each of the two standardized network variables, on the other hand.

84. *See*, *e.g.*, Stephen B. Burbank, *Vanishing Trials and Summary Judgment in Federal Civil Cases: Drifting Toward Bethlehem or Gomorrah?*, 1 J. Empirical Legal Stud. 591 (2004) (studying summary judgment motions).

85. *See, e.g.*, David A. Hoffman, Alan J. Izenman & Jeffrey R. Lidicker, *Docketology, District Courts, and Doctrine*, 85 Wash. U. L. Rev. 681, 708 (2007) (studying the District of Maryland, the Northern District of California, the Southern District of New York, and the Eastern District of Pennsylvania).

86. *See, e.g.*, Stanford Securities Class Action Clearinghouse, *supra* note 33.

87. For a discussion of this importance, see *Free PACER*, *supra* note 19.

PACER;[88] the cost can only have grown with PACER's document inventory. This system serves as a revenue center for the judiciary, but it would be easy for Congress to mandate open, or at least much broader, access to court data while replacing the revenues from PACER, which are tiny in any reasonably considered context (for example, I have elsewhere pointed out that Americans annually spend 25 times as much on wild bird seed as the judiciary rakes in from PACER).[89] There has been some legislative momentum in favor of reform, but it seems to have stalled.[90]

Second, my results document the extent to which intexity makes up a distinct aspect of our procedural system's empirical reality. It suggests one can get only so far in addressing challenges with the litigation of perceived substantive problem areas. Usual-suspect areas such as securities, antitrust, and patent actions—three of the Intense 4—may be more intexly litigated on average, but they account for relatively few of the cases that dominate federal district court dockets, and my data indicate that a relatively large number of them do not end up as poster-children for extreme intensity or complexity. Likely, that is to some extent because the threat of intex litigation induces settlements—i.e., I am sure that there is an important element of bargaining in the shadow of intexity.[91]

Nevertheless, it is at least arguable that procedure motivated by intexity considerations should be targeted directly at cases likely to be intexly litigated, regardless of their substantive area. We have certification and other distinctive follow-on procedures for class

88. Michael Lissner, *The Cost of PACER Data? Around One Billion Dollars*, FREE LAW PROJECT (Oct. 10, 2016), https://free.law/2016/10/10/pacer-costs-a-billion-dollars [https://perma.cc/YRY7-G8W9].

89. *Free PACER*, *supra* note 19, at 344.

90. The Open Courts Act of 2021 made progress in the Senate, having been voted out of the Judiciary Committee with bipartisan support on December 9, 2021. *See Judiciary Committee Advances Legislation to Remove PACER Paywall, Increase Accessibility to Court Records*, U.S. SENATE COMMITTEE ON THE JUDICIARY (Feb. 7, 2022), https://www.judiciary.senate.gov/press/dem/releases/judiciary-committee-advances-legislation-to-remove-pacer-paywall-increase-accessibility-to-court-records [https://perma.cc/B2M3-3HPZ]. If that bill, or something like it, became law, then it might be possible for considerably more research to be done, though the bill's text does not specify whether the Judiciary must provide the public with the kind of bulk access necessary to do large-scale research. Open Courts Act of 2021, S. 2614, 117th Cong., § 3(c) (2021). Advocates of this bill were unsuccessful in attempting to attach it to the omnibus spending bill that passed in 2022. *See* Nate Raymond, *No Free PACER as U.S. Lawmakers Exclude Proposal from Spending Bill*, REUTERS (Dec. 20, 2022), https://www.reuters.com/legal/government/no-free-pacer-us-lawmakers-exclude-proposal-spending-bill-2022-12-20/ [https://perma.cc/U75Q-WPAZ].

91. *Cf.* Robert H. Mnookin & Lewis Kornhauser, *Bargaining in the Shadow of the Law: The Case of Divorce*, 88 YALE L. J. 950 (1979).

litigation,[92] and although the motivation there is distinctly related to protecting the interests of absent class members, perhaps rule makers ought to take a page from the class action part of the rulebook and create formal procedure allowing courts to identify and treat distinctively those cases likely to involve great intexity. This discussion overlaps with ongoing suggestions that MDL litigation needs clearer rules to protect the interests of individual plaintiffs.[93] But I emphasize again that I excluded MDL cases from the empirical analysis above; the facts presented above indicate that intexity and substantive area are distinct concepts *outside* as well as inside the MDL context.

Among substance, intensity, and complexity, some dimensions may matter more than others, and this may vary according to the questions at issue. In particular, to the extent that we worry about the costs of uniform procedure,[94] transsubstantivity may matter comparatively little. Perhaps procedure should be transsubstantive, but no longer quite so transintense or transcomplex. So, perhaps transintexity—structuring formal procedural doctrines to cut across cases regardless of the combination of intensity and complexity they involve—warrants policy attention from the Advisory Committee on Civil Rules, or even from Congress.

What would reform look like? One possible avenue would be to design distinct procedural tracks for cases with varying levels of predicted intexity, assessed early in a case's life. This is an idea with a pedigree, both in the context of American procedure and comparatively. Eminent scholars have proposed varying procedural tracks. More than three decades ago, Professor Maurice Rosenberg approvingly discussed an article that proposed "to give the litigants the option of putting the case on a fast track that assures them a trial date of their selection within 12 months," in return for which litigants would "agree to sharply limited pretrial motions and discovery."[95] More recently, Professors Stephen Burbank and Stephen Subrin proposed their own "simple track,"

92. *See* FED. R. CIV. P. 23 and associated case law.

93. *See, e.g.,* Abbe R. Gluck & Elizabeth Chamblee Burch, *MDL Revolution*, 96 N.Y.U. L. REV. 1 (2021).

94. Stephen B. Burbank, *Summary Judgment, Pleading, and the Future of Transsubstantive Procedure*, 43 AKRON L. REV. 1189, 1194 (2010).

95. Maurice Rosenberg, *Federal Rules of Civil Procedure in Action: Assessing Their Impact*, 137 U. PA. L. REV. 2197, 2212 (1989) (citing McMillan & Siegel, *Creating a Fast-Track Alternative Under the Federal Rules of Civil Procedure*, 60 NOTRE DAME L. REV. 431, 431–55 (1985)). Professor Maurice Rosenberg's argument was based on case simplicity: "'Trans-substantive' is a less than helpful concept . . . . Many simple cases, some involving substantive issues drawn from contract law, others from tort law, and still others from the civil rights field, all require the same kind of trial processing despite their diverse substantive sources. On the hand, complex cases often require vastly

concentrating on formal discovery limitations as their primary feature.[96] Although Professors Burbank and Subrin wrote that their proposal would "require that we rethink the transsubstantive assumption of the Federal Rules,"[97] they might have resisted that instinct; as the present Article demonstrates, intexity—and its flipside of simplicity—is not a substantive feature as such.[98]

I mean neither to endorse nor reject such proposals here. But one impact of having a system of formal procedural uniformity filtered through a Supreme Court that doesn't shy away from restrictive procedural rulings is that the Court's doctrinal innovations have a way of remaking the entire ship of procedure, even when they are motivated primarily by the small slice of cases I have characterized as especially intex.[99]

Consider how distinct procedural tracks related to complexity might alter the terms of debate about pleading standards—inarguably one of the most hotly debated areas of procedure in the last decade and a half. A primary argument for the plausibility standard the Court introduced in *Twombly* is about the potentially enormous scope and volume of discovery in a case that featured both procedural and substantive complexity,[100] and one that promised a substantial likelihood

---

different processing from simple ones even though drawn from the same substantive sources. They accordingly belong on different procedural tracks." *Id.*

96. Stephen B. Burbank & Stephen N. Subrin, *Litigation and Democracy: Restoring a Realistic Prospect of Trial*, 46 HARV. C. R.-C. L. LAW REV. 399, 409-411 (2011) (advocating rules for simple track cases that would include "nonnegotiable limits on the number of interrogatories and depositions and on the length of depositions," with exceptions "available only by court order to prevent manifest injustice (or some similarly daunting standard)," as well as "a rule requiring that document requests in simple track cases").

97. *Id.* at 409.

98. Some aspects of their proposal might, though, have involved substance-specificity, such as allowing substance-specific discovery protocols. *See id.* at 412, n. 70 (discussing efforts to create discovery protocols specific to employment discrimination cases).

99. Professor Brooke Coleman has made this point well. *See* Brooke D. Coleman, *One Percent Procedure*, 91 WASH. L. REV. 1005, 1010 (2016). For more on the Supreme Court's use of procedure cases, rather than the rulemaking process, to alter procedure, see STEPHEN B. BURBANK & SEAN FARHANG, RIGHTS AND RETRENCHMENT: THE COUNTERREVOLUTION AGAINST FEDERAL LITIGATION (2017).

100. With respect to procedural complexity, five major corporations were named defendants, there was a sprawling putative class to be represented by two named plaintiffs, and fulsome discovery would surely have involved many depositions of corporate executives as well as an untold amount of electronic discovery, with attendant disputes and jockeying. With respect to substantive complexity, the case involved allegations of collusion of the form that are often associated with the deployment of extensive expert witness testimony related to competition economics, which can be sophisticated and highly complex. *Iqbal* was also quite complex. Even leaving aside the

of highly intense litigation if allowed to continue.[101] By requiring that a complaint's non-conclusory allegations make out a plausible claim for relief (whatever exactly all that means), the Court averted such a litigation. The Court's concerns about *Twombly* may be viewed as importantly connected to practical policy concerns about the case's intensity and complexity.[102]

This Article's empirical evidence adds to an observer's common-sense view that *Twombly* was an outlier in both its intensity and its complexity. There is at most limited empirical reason to think the policy concerns that so obviously motivated the *Twombly* Court apply to the mine run of cases that do not share *Twombly*'s intexity.[103] And although some observers advocated that the Court might read *Twombly* as a substantive antitrust opinion,[104] Justice Kennedy's *Iqbal* majority opinion simply pointed to Rule 1's transsubstantivity, and that was that—*Twombly*'s standard applied to "all civil actions."[105]

But what if there were separate procedural tracks for cases likely to be intense and/or complex? That would allow different pleading standards for simpler cases and more intex ones. Even accepting *arguendo* the wisdom of the Supreme Court's policy innovations in *Twombly*, there would remain a strong case that pre-*Twombly* notice pleading is the more sensible policy for simpler disputes, given the cost and delay associated with litigating motions to dismiss, not to mention the possibility that some meritorious cases will never be

---

involvement of the Attorney General and FBI Director (who were the only petitioners at the Supreme Court), the district court litigation that spawned the Supreme Court case, Elmaghraby et al v. Ashcroft et al, No. 1:04-CV-01809 (E.D.N.Y.), involved dozens of parties and substantial discovery.

101. The order staying district court proceedings due to the Supreme Court's consideration of *Twombly* was the 100th numbered entry on the district court's docket. Twombly v. Bell Atl. Corp., No. 1:02-CV-10220, ECF No. 100 (S.D.N.Y. July 5, 2006).

102. This view is embodied in Justice Souter's opinions in *Twombly* and *Iqbal*. Justice Souter first blessed the plausibility and conclusory requirements in his opinion for a 7-2 *Twombly* majority. Two years later in *Iqbal*, he penned an impassioned dissent against the bare majority opinion, featuring arguments about conclusoriness and plausibility that his own *Twombly* opinion arguably foreclosed.

103. For example, a Federal Judicial Center study of discovery costs conducted around the same time as the Supreme Court's *Twombly* and *Iqbal* decisions found that the median cost of closed cases—including attorney fees—was $15,000 for plaintiffs and $20,000 for defendants. Emery G. Lee III & Thomas E. Willging, FEDERAL JUDICIAL CENTER NATIONAL, CASE-BASED CIVIL RULES SURVEY PRELIMINARY REPORT TO THE JUDICIAL CONFERENCE ADVISORY COMMITTEE ON CIVIL RULES 2 (Fed. Judicial Ctr., Oct. 2009), https://www.fjc.gov/sites/default/files/materials/08/CivilRulesSurvey2009.pdf [https://perma.cc/RHH3-CN98].

104. *See, e.g.*, Burbank, *Pleading and the Dilemmas of "General Rules"*, *supra* note 15.

105. 556 U.S. 662.

brought thanks to the difficulty of pleading without information not available to the plaintiff.[106] Acting on this observation would not require weakening transsubstantivity—just transintexity. If separate pleading tracks could be effectively designed, federal pleading—and, thus, most federal litigation—would be streamlined for the overwhelming majority of cases, even as the value, whatever it is, of the plausibility and conclusoriness aspects of contemporary pleading were preserved in cases likely to be intex.

Pleading aside, the general idea of procedural tracking was discussed to some extent around the topic of evaluating the Civil Justice Reform Act, as some district courts had one form or another of case tracking in effect.[107] At least one federal judge presently has a standing order that allows some version of distinct tracking.[108] Some state court systems have distinct tracks for cases of varying complexity, and so do other countries' systems.[109] Because I use only U.S. federal court data, my findings cannot speak directly to the role of tracking in those systems. But it seems reasonable to think that the broad patterns I find—of overlapping substance, intensity, and complexity—likely occurs in other systems as well.

As I have said, this Article is not the place for a detailed assessment of whether tracking or other intexity-related rule making ideas should be adopted. Any discussion of such issues would require careful consideration of what margins of intensity and complexity ought to be targeted, and what case characteristics—observable to judges and parties, preferably early in a case's life cycle—could be used for this targeting. Obviously the number of docket entries in completed litigation wouldn't be known at a case's outset, and variables such as the number of parties might be too easily manipulated by parties seeking to evade or entrench particular tracks. Thus, the practical

106. For a deeper discussion of these issues, as well as inconclusive empirical evidence, see Jonah B. Gelbach, *Material Facts in the Debate over* Twombly *and* Iqbal, 68 STAN. L. REV. 369 (2016).

107. *See, e.g.*, Donna Stienstra, Molly Johnson, and Patricia Lombard, A STUDY OF THE FIVE DEMONSTRATION PROGRAMS ESTABLISHED UNDER THE CIVIL JUSTICE REFORM ACT OF 1990, Fed. Judicial Ctr. (Jan. 24, 1997), https://www.fjc.gov/sites/default/files/2012/0024.pdf [https://perma.cc/6X7G-RYDW]. And according to Zachary D. Clopton and Marin K. Levy, *Local Rules* (draft on file with author), the Southern District of Florida currently "requires judges to assign each case to one of three defined tracks for case management: expedited, standard, and complex." Clopton & Levy at 13, n. 80.

108. *See Optional Fast Track Procedure for Civil Cases Assigned to Judge DeGiusti*, https://www.okwd.uscourts.gov/wp-content/uploads/Judge-DeGiusti-FAST-TRACK-PROCEDURE.pdf (last visited Feb. 7, 2022) [https://perma.cc/P5R9-FM2V].

109. *See generally* Peter C. H. Chan & C. H. van Rhee eds., *Civil Case Management in the Twenty-First Century: Court Structures Still Matter* vol. 85, IUS GENTIUM: COMPARATIVE PERSPECTIVES ON LAW AND JUSTICE (2021).

aspects of targeting on intexity would be a substantial endeavor, and likely a controversial one with cross-cutting normative considerations. Perhaps that conversation is not worth it, given all the challenges. Or perhaps the common law approach to procedure, working things out one case at a time, is a better road.

But for those who do see value in considering procedural reform pegged to litigation's transsubstantive features, it would be good if procedure-reform discussions were more connected to empirical evidence. This Article has sought to open such a conversation by demonstrating that intensity and complexity can be at least partially measured.

Appendix A: Data Details

A. Raw XML Files[110]

As part of a contract between Yale and Thomson Reuters, I was provided with 360 distinct xml-formatted files.[111] Each xml-file docket contains case-level information including some or all of the following variables:

- Case caption, i.e., title;
- date filed;
- date closed (if that had happened);
- judge assigned to the case as of the time the docket information was pulled;
- court where the case originated in federal court;
- docket number in conventional formats[112];
- PACER nature of suit code information;
- whether a jury was demanded;
- the basis for federal court jurisdiction;
- whether the case is associated with other dockets, as well as whether there is a lead-docket case for cases that are consolidated.

In addition, xml-file dockets contain blocks of information regarding the parties in a case, including their name and role, with separate entries for parties with multiple roles (e.g., a person can be both a plaintiff and a counterclaim defendant). For each party-role, there is information about the attorney(s) representing the party in that role, if any; this information includes attorney name as well as the attorney's firm.[113]

Finally, xml-file dockets contain information about docket entries. For each docket entry, the data include:

---

110. Readers uninterested in the gory data details should feel free to skip this section.

111. To process these, I used the Python software package's *lxml* module, which handles xml parsing automatically. In a small number of cases, some or all of the dockets in a file have corrupted data (whether natively or due to my own early work processing these files) such that I could not parse some or all of the information in those files.

112. Typically, this is of the form O:YY-TT-NNNNN, where 'O' refers to the division with the district, YY is the last two digits of the year when the case was filed, 'TT' is the case type (typically, 'CV', though possibly also 'MD' for multidistrict litigations, 'MC' for miscellaneous, or other strings), and 'NNNNN' is the sequence number within the filing year, so that 'NNNNN' equals 1 for the first case filed in division 'O' in year 'YY', 'NNNNN' equals 100 for the 100th such case, and so on.

113. I do not use attorney data in this Article.

- the date on which it was entered;
- the docket number provided by the CM/ECF system[114];
- the original text describing the docketed event, such as a party's filing of a complaint or a motion, or court's order or memorandum opinion.

Many docket entries include references to other docket entries. For example, if docket entry number 32 represents a plaintiff's memorandum of law responding to defendant's motion to dismiss that was docketed in entry number 16, then entry number 32 will contain text that makes it possible for me to discern this relationship. This is because the CM/ECF system tracks this information, and when Thomson Reuters scraped PACER to obtain docket information, the downloaded docket reports included it. I asked my contacts at Thomson Reuters to wrap xml tags around these fields, which they generously did. Similarly, there are xml tags identifying attachments that parties filed with docket entries.

After reading and parsing the raw xml in the files Thomson Reuters delivered, I uploaded the information into a SQLite relational database, which allows simple queries to retrieve case information.[115]

I received the raw xml files in several feeds from Thomson Reuters, which took place over a period of roughly three years. The first feed occurred in early 2012 and included information on cases filed in the U.S. District Courts on or after January 1, 2005, until roughly the date the data were pulled and then sent to me.[116] Thereafter I received additional feeds approximately every 6 months. The coverage is meant to extend through December 31, 2014, for included cases. Thus I have up to 10 years of docket information for cases filed on January 1, 2005, and a declining window of coverage for cases filed thereafter (e.g., for a case filed on December 30, 2014, I would have one subsequent day of information). This limitation creates a risk of bias in any time-dependent analysis, so I proceed carefully in selecting cases for consideration, as discussed below.

---

114. Some entries have no docket entry numbers. These generally seem to involve scheduling of minor events, corrections of other entries, and the like.

115. *See* SQLite, *SQLite Home Page*, https://www.sqlite.org/index.html (last visited June 6, 2024) [https://perma.cc/8GC5-FSAE]. I created separate tables for case-level metadata such as the title of the case and the court in which it is active; the parties in the case; the attorney(s) for each party; the docket entry information (date filed, docket entry number, and docket entry text); there are also tables that track other dockets associated with cases, whether as lead docket in the event of consolidation or a previous transfer out of another court. This organization makes it possible to quickly query the data base to select cases and features of use for this Article (and future work).

116. There are also some case records for cases filed before January 1, 2005. It is unclear to me why these cases were included, and I do not use them in any of my analysis.

After processing with the code I wrote to parse the xml files and upload to the SQLite database mentioned above, I have the following information:

- Case-level metadata: 6,605,481 records.
- Party-level data: 52,372,588 records.
- Attorney-level data: 63,437,445 records.
- Entry-level data: 172,671,310 records.
- Data on lead dockets associated with cases: 1,567,882 records.
- Data on other dockets associated with cases: 4,584,853 records.[117]

B. Dealing With Duplicate Records[118]

A significant share of the records in the raw xml files are duplicative to some degree. To see why, consider a case filed in 2005. That case will be represented in the initial xml-file data feed, and if the case was still active in the six months that followed the initial feed, there will be docketed activity for the case in the second feed I received. This means I will have a second set of case-level metadata on the case. When I read the xml file for the second feed, there is no way for me to tell that the case was already included in the first feed. Accordingly, I simply add a second row to the caseheader database table, and a third, and fourth, and so on as appropriate, if the case appears in additional feeds.

To avoid the problem of counting cases multiple times, I then construct a database table that includes one record for each already processed case, regardless of the number of times that case appears in the raw xml files. This is possible to do because each xml-file docket includes information identifying the court where the case is docketed and the docket number within that court. The combination of those two pieces of information uniquely identifies a case.[119]

For cases that appear in the raw xml files multiple times, I received multiple copies of the docket entries that were already present in earlier feeds. I thus also create a unique version of the docket entry table, and similarly with parties and attorneys. The unique-record tables thus constructed have the following numbers of records:

---

117. Examples of such cases include the docket information for state-court cases that have been removed, for appellate activity associated with the case itself, and criminal cases that led to a habeas petitioner's incarceration.

118. As with Section A, readers uninterested in the gory data details should feel free to skip this section.

119. It is important to recognize that the docket number generally is *not* enough to uniquely identify cases; for a simple example, the first action filed each year in each district's first division will have the same docket number as the corresponding action for the first division of any other district.

- There are 3,523,880 records in the unique case table.
- There are 79,281,963 records in the unique entry table.
- There are 23,416,322 records in the table that contains information on unique party/party-type records.
- There are 27,876,180 records in the table that contains information on unique attorney records.

C. CM/ECF Go-Live Dates[120]

Not every district court adopted electronic filing at the same time. Many did so around 2005–whether during, shortly before, or shortly after. This matters because it is unclear what level of quality the docket records on PACER have for cases filed in districts that had not yet adopted the CM/ECF electronic system. The term PACER uses for such adoption is "going live". The table in Appendix B provides dates on which all districts but two went live; these two are the Northern District of Ohio and the Western District of Tennessee, which do not provide go-live dates on their websites.[121]

It appears that for at least some cases filed before the go-live date, some if not all docket entries were entered into PACER manually. This means they lack the links to attachments and other entries described above. It also raises questions about coverage (*Were all docket entries entered?*) and quality (*Were docket entries entered accurately?*). Accordingly, I restrict my analysis to cases filed after districts' go-live dates. This requires dropping the Northern District of Ohio and the Western District of Tennessee, as well as many or all cases from a handful of other districts (*see* discussion below).

D. Consolidation and Transfer[122]

Matters are further complicated because of the conceptual difficulty in handling cases that are consolidated with other cases. For unconsolidated cases, I can simply record variables of interest such as the number of docket entries, number of parties, etc. But when cases

120. As with Sections A-B, readers uninterested in the gory data details should feel free to skip this section.

121. For the other district courts, I obtained go-live dates from the Free Law Project, which previously posted by them at Court Version Scraper, https://court-version-scraper.herokuapp.com/courts.json (last visited Feb. 1, 2022). Although the page in question is no longer present at that URL, a copy saved on August 19, 2021, is available via the Internet Archive, at *Court Version Scraper*, https://web.archive.org/web/20210819144426/https:/court-version-scraper.herokuapp.com/courts.json (archived Aug. 19, 2021).

122. As with Sections A-C, readers uninterested in the gory data details should feel free to skip this section.

are consolidated, activity may be docketed on some or all related cases' dockets, which complicates measurement. Further, MDLs involve a single docket for the MDL itself. Some events specific to individual cases may not be docketed in the overall MDL, even as others might appear in the overall MDL docket (e.g., when summary judgment is granted as to a subset of cases). Because it is unclear how best to handle such issues, I have chosen for this paper to focus only on cases that lack certain indicia of having been consolidated with another case as the lead; relatedly, I exclude MDLs.

E. Case-Level Variables Based on Docket Entry Text[123]

Applying standard regular-expression[124] matching routines to all 79 million-plus unique docket entries and 3 million-plus case-level metadata records in the database, I create a number of "flags", i.e., variables that indicate whether one or more conditions is satisfied by the text of each entry. For purposes of this article, the flags of interest are the following:

A flag indicating whether the case title begins with the string "IN RE" (all text provided by Thomson Reuters was capitalized). These cases involve litigation comprising consolidated individual actions.

- A flag indicating that the case has an associated "lead docket", which indicates the case itself is consolidated with one or more other cases. (Note that cases that *are* the lead-docket case will not be flagged this way.)
- A flag indicating whether text related to multidistrict litigation appears in any docket entry for a case.[125]

123. As with previous sections in this Part, readers uninterested in the gory data details should feel free to skip this section.

124. "Regular expressions are specially encoded text strings used as patterns for matching sets of strings." Michael Fitzgerald, *Introducing Regular Expressions*, https://www.oreilly.com/library/view/introducing-regular-expressions/9781449338879/ch01.html (last visited on Feb. 6, 2022) [https://perma.cc/X2KQ-ECDW]. An example is the string "^(\(\d{3}\)|^\d{3}[.-]?)?\d{3}[.-]?\d{4}$", which is "a fairly robust regular expression that matches a 10-digit, North American telephone number, with or without parentheses around the area code, or with or without hyphens or dots (periods) to separate the numbers." *Id.* An application in the civil litigation context is to match docket numbers of the form 1:06-CV-01234, which corresponds to the 1,234th civil case filed in division 1 of a district in 2006. A regular expression that can be used in Python to match strings having this format and also capture the year ("06" here), case type ("CV" here) and sequence number ("1234" here) in variables named "year", "casetype", and "sequencenumber" is "\d\:(?P<year>\d\d)-(?P<casetype>[A-Z]{2})-(?P<sequencenumber>\d{1,5})".

125. To determine this, I used Python's regular expression matching capabilities to determine whether each of the docket entries in my data included any of the strings 'MDL PANEL', 'JPML', or 'JUDICIAL PANEL ON MULTIDISTRICT LITIGATION'.

- A flag indicating whether a case has any flag set that indicates consolidation; this flag is set if any of the preceding three flags is.

I also use the PACER nature of suit code information provided with each docket record to construct a coarser categorization of substantive case types. Based on this code, I assign each case to one of the following 14 substantive law case areas:

- antitrust
- civil rights
- consumer
- contract
- copyright and trademark
- environmental
- immigration
- labor
- other
- patent
- prisoner petitions
- securities
- Social Security
- tort

To construct my grouping, I used a list of codes posted on the PACER website.[126] Appendix C provides a table mapping from each of the 102 PACER codes listed on the PACER website and the 14 groups listed above.

F. Cases With Features Indicating They Should Be Excluded from Some or All of the Analysis

Table 10 shows the total number of unique docket (case) records in my data, by the year cases were filed. There are at least a quarter-million such records in each of the ten years for which I have data, with the total generally rising over time. The total number of cases over the ten filing years covered by the table amounts to a bit shy of 2.9 million. This is roughly 10% fewer than the total number of unique case records described above. The difference is accounted for by the fact that some

If so, I coded the docket entry as referring to MDL litigation, and I then set a flag indicating that the case had such docket entry.

126. For details, see PACER, *Nature of Suit Codes*, https://pacer.uscourts.gov/sites/default/files/files/nature%20of%20suit%20codes.pdf (last visited on February 16, 2023) [https://perma.cc/EU4X-P3X4].

of the cases Thomson Reuters delivered had filing dates that preceded 2005.[127]

Table 10: Docket Counts By Year Case Filed

Filing Year	Number of Cases In My Data
2005	250,712
2006	269,075
2007	253,757
2008	277,932
2009	282,400
2010	329,702
2011	299,182
2012	283,026
2013	309,598
2014	303,321

As discussed above, a number of cases are either impossible for me to use or would raise concerns for some or all of my analysis:

- *Missing or bad court identifier.* In some cases, the reported court was not one of the U.S. District Courts. For example, some cases from the Court of Federal Claims were included, and other cases have reported court codes that are hard to make sense of; a smattering of cases are missing any court identifier. I drop all records associated with these cases.
- *Missing value for nature of suit.* Some dockets have a missing value for the nature of suit code. I drop these as well.

127. An additional issue is that the numbers in this table appear to be greater than the calendar year numbers reported by the judiciary via the Federal Court Management Statistics system. For example, that system reports that 267,989 civil cases were filed in 2012 (*see* first page of table provided at UNITED STATES DISTRICT COURTS, *National Judicial Caseload Profile* 1, https://www.uscourts.gov/sites/default/files/statistics_import_dir/district-fcms-profiles-december-2012.pdf (last visited on June 6, 2024) [https://perma.cc/L8VG-N2YG], whereas my data set has 283,026—about 6 percent more cases. I suspect but have not confirmed that the differences involves inter-district transfers of cases that were filed in earlier calendar years. Such filings cause the origination of a case with a distinct docket number in the transferee district, with the original docket terminated. Such cases are included in my data; at this writing I am not sure whether they are included in the official statistics.

- *Indicia of consolidation.* I drop dockets whose case caption includes "IN RE", which often indicates a large number of cases have been consolidated. I also drop dockets with any docket entry text suggesting they are associated with an MDL. And I drop dockets that have a reference to another case as "lead docket", which indicates the case was consolidated with some other case. (I keep the cases that are themselves lead dockets.)
- *Go-live date*. As discussed above, district courts varied with respect to the date on which they adopted the CM/ECF system for their dockets. I drop cases that were not filed after the go-live date for the relevant district court, as well as those for the two district courts (OHND and TNWD) that do not post their go-live date on PACER.

Table 11 reports statistics related to these issues, again broken down by filing year. Column 1 repeats the total number of records for the year (which also appeared in Table 10). Column 2 reports the number of cases for which the PACER nature of suit code is missing. Column 3 reports the number for which the court identifier either was not for a district court or was not identifiable. In some cases this involves cases apparently filed in the Court of Federal Claims; in others it involves the absence of any identifier, or the presence of one that I could not match with one of the 94 U.S. District Courts. Column 4 reports the number of cases for which the type was neither "CV" (general civil case), "MD" (multidistrict), nor "MC" (miscellaneous).

Before I move forward, I note one data-selection choice: in this Article, except where I explicitly say otherwise, I will consider only cases with no indicia of consolidation. I make this choice partly to avoid having to repeatedly discuss differences. But there is also a more substantive reason. Everyone knows that MDLs and other outlier consolidated litigation[128] can be extremely intense. Disregarding such cases allows me to demonstrate my points about intensity and complexity *even* among cases that are not in that set.

128. For a non-MDL example, consider *In re World Trade Ctr. Lower Manhattan Disaster Site Litig.*, which was managed via the 21 MC 100, 21MC 102, and 21 MC 103 dockets; *see, e.g.*, In re World Trade Ctr. Lower Manhattan Disaster Site Litig., 66 F. Supp. 3d 477 (S.D.N.Y. 2015) (discussing these miscellaneous dockets).

Table 11: Docket Counts By Year Case Filed and Various Bases for Exclusion

Year	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
2005	250,712	6,732	1,529	3,100	10,443	54,786	23,012	169,474
2006	269,075	6,964	958	5,147	12,678	32,436	31,126	196,879
2007	253,757	6,583	906	5,661	12,638	20,079	25,757	199,962
2008	277,932	7,055	934	6,097	13,494	6,815	30,101	230,858
2009	282,400	7,593	906	5,108	13,052	6,932	36,615	228,089
2010	329,702	37,199	896	4,498	42,022	7,988	59,456	252,225
2011	299,182	8,842	917	7,294	16,337	8,822	30,470	248,895
2012	283,026	13,504	934	7,989	21,855	8,862	36,089	222,375
2013	309,598	14,659	1,130	5,581	20,475	6,675	29,323	256,907
2014	303,321	17,026	1,254	4,838	21,952	4,554	29,707	251,240

Column 1: Total number of records for the year.

Column 2: Missing Nature of Suit code.

Column 3: Not in identifiable District Court.

Column 4: Case type not "CV", "MD", or "MC".

Column 5: Case has at least one of the problems in columns 2-4.

Column 6: Case was not filed after a date known to be after the court's PACER go-live date (for TNWD and OHND, there is no such date, so all cases filed in these courts are counted here).

Column 7: Indicia of consolidation: case title begins with 'IN RE' or contains 'MDL', or case has a lead-docket case, or case has a docket entry with text indicating relationship to an MDL.

Column 8: Not in columns 5, 6, or 7.

Column 5 reports the number of cases for which at least one of the problems in columns 2-4 obtains; the number of cases in this column thus can be no greater than the sum of the numbers in columns 2-4 but can be less than that because multiple problems can exist for a single case.

Column 6 reports the number of cases that either were filed before the relevant district court's CM/ECF go-live date (the date when that court adopted the electronic filing and management system) or were filed in one of the districts that has not publicly posted its go-live date. Column 7 reports the number of cases that have indicia of consolidation, as discussed above.

Column 8 reports the number of cases from each year available for use in my analysis. These cases fit in none of columns 2-7: they are cases for which I can identify a PACER nature of suit code and a U.S. district court, for which the case type is "CV", "MD", or "MC", and for which the case is known to have been filed after the filing court's go-live date.

To use cases filed in 2005 as an example, we see that there were 250,712 such cases (the same as reported in Table 10). Of these, 6,732 did not have a PACER nature of suit code (Column 2); 1,529 were not identified as having been filed in an identifiable U.S. district court (Column 3); and 3,100 were not of case type civil, multidistrict, or miscellaneous (Column 4). Altogether, Column 5 reports that there were 10,443 cases with one or more of these detriments. Column 6 indicates that there were 54,786 cases filed in 2005 for which the case wasn't filed after the court's go-live date. Finally, Column 7 reports that there were 23,012 cases with some indication of consolidation. Column 8 thus reports that my analysis set has a total of 169,474 cases from 2005 after eliminating all cases that are represented in at least one of Columns 5-7.

There are two notable departures from the 2005 data for the other filing years. First, in 2010 there were a whopping 37,199 cases for which the nature of suit code was missing and a correspondingly large 59,456 cases with indicia of consolidation.[129] Second, the number of cases in Column 6—those not filed after a court's go-live date—declines rapidly over time. That trend results because 83 of the 94 district courts had gone live by the end of February 2006, 88 had gone live by January 2, 2007, and 91 had gone live by the end of January 2008.[130]

129. As of this writing, I do not know what explains this anomaly, although it is the case that the Deepwater Horizon oil spill occurred in 2010, so it seems likely that event is related.

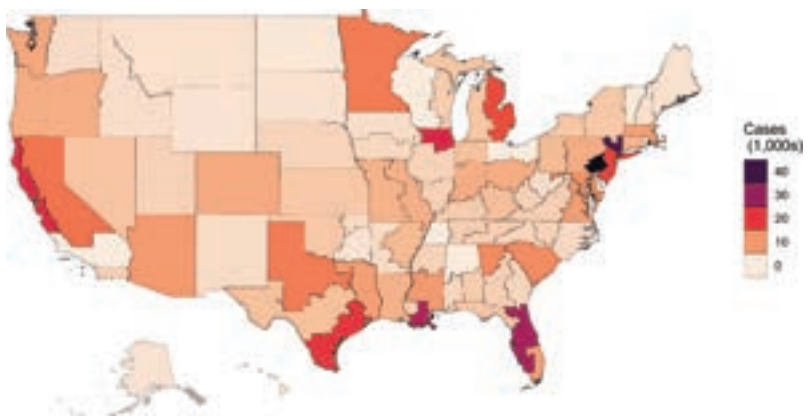
130. The six that hadn't yet gone live by January 2, 2007, were the District Court of the Virgin Islands (go-live date of June 18, 2007), the Central District of California (go-live date of January 1, 2008), the Western District of Wisconsin (go-live date of January 23, 2008), the Northern District of Alabama (go-live date of February 15, 2014), and the Northern District of Ohio and Western District of Tennessee (no publicly posted go-live date).

To provide a sense for the pattern of available cases, Figure 10(a) plots a heat map, with districts represented by more observations in my data set being represented by darker shades.[131] For the most part, the pattern broadly reflects population and economic activity, which explains why, for example, the Northern District of California, the Southern District of New York, the District of New Jersey, and the Eastern District of Pennsylvania have so many cases. On the other hand, the absence of any cases in, e.g., the Central District of California reflects the fact that its go-live date post-dated my window of observation. Figure 10(b) plots a complementary heat map that shows the share of otherwise available observations that I *dropped* because of one or another of the data restrictions discussed above—thus, darker shading indicates that a district had relatively more cases dropped, and so is less represented in my data relative to its share of all filed cases.[132]

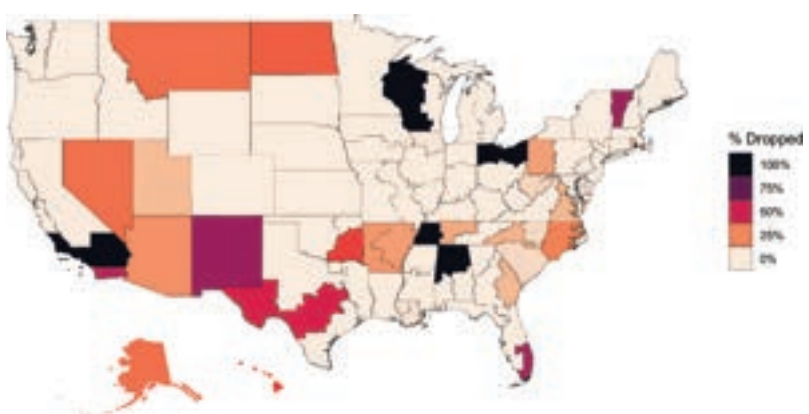
131. The maps do not include non-states other than the District of Columbia; thus the U.S. District Courts for the Districts of Puerto Rico, the Virgin Islands, Guam, and the Northern Mariana Islands do not appear in these maps.

132. I am grateful to Andrew Baker for his help in making these maps.

Figure 10: Maps of Included and Excluded Cases



(a) Included number of cases



(b) Share of available observations excluded

G. Right Censoring

For variables that grow over time as a case continues, such as the number of docket entries, a problem arises related to what statisticians call *right censoring*. Consider the set of cases that will be filed on July 1, 2025, and suppose we will have information on all case activity through the end of 2025.[133] Some cases—*quick terminators*—will terminate soon enough to be closed within the window of observation.

133. I choose these dates in the aspirational expectation that this paper will have been published before they have passed.

Others—*slow terminators*—will not. Now consider two cases that each generate 20 docket entries during the observation window, with one case being a quick terminator and the other a slow terminator. By construction, the quick terminator's total number of docket entries is 20. For the slow terminator, though, all we can say is that its ultimate total number of docket entries is greater than 20; its total number of docket entries is right-censored because it would be impossible to observe activity on dates after—to the right of, on a timeline—December 31, 2025.

This example illustrates how an early closing of the observation window can distort the picture we get of the distribution of docket entries over cases' full lifetimes. Because dockets grow over time for open cases, failing to account for short observation windows will make the longest-lived, most intensely litigated cases seem like they are shorter-lived and more lightly litigated than they ultimately will be.

There are two basic ways to address such a right censoring problem. One is to estimate parameters of a model that relates case termination timing and docket entry generation, thereby allowing predictions of what would happen in right-censored cases after the censoring time. The second approach is to use a subset of cases that can be observed for a long enough period that the censoring problem is either absent or unimportant as a practical matter. Although statistical modelling can be a useful and sometimes indispensable tool, given the initial stage of our knowledge of this Article's subject matter, I choose the latter approach to avoid the possibility of imposing invalid mathematical structure on the underlying litigation process that generates my data.[134]

Because my focus in this Article is importantly related to outliers in litigation intexity, it is important to choose a subset of cases that is only lightly affected by my actual observation window, which closes after December 31, 2014. For cases filed on January 1, 2005, I have a ten-year observation window. For those filed a year later, the window is nine years long, and so on. Choosing the set of cases to consider is complicated by the fact that courts' PACER go-live dates vary, and that many district courts went live at varying dates during calendar year 2005.

134. The issue here is that using statistical modelling to address right censoring requires making mathematical assumptions about the pattern of docket entries that are not observed. In some applied settings there might be good reason to think particular assumptions would be appropriate. I do not think this is such a setting, at least given how little we know to begin with.

As a practical matter, the choice of which cases to include boils down to picking a latest-filed date, such that cases filed before the latest-filed date are included, and those filed afterward are not.[135] To choose a latest-filed date, I created a "days-to-spare" variable that measures the number of days between the last docket entry I observe for a case and the cutoff date of January 1, 2015. I then labeled a case as right-censored if its days-to-spare is 31 or less. This labeling allows for the possibility that a case might be ongoing but just not have any docketed activity in the last month I can observe.[136]

Unsurprisingly, very few cases are right-censored in the early years—fewer than 0.1% (one in a thousand) for each of 2005, 2006, and 2007. Not until we get to 2008 filings does the right-censored share exceed 0.1%. For purposes of this paper, I restrict attention to cases filed in 2005, 2006, or 2007. All told, 99.93% of those cases were unaffected by right censoring using the labeling described above.

135. This is true because my observation window closes after the same date, December 31, 2014, regardless of when cases were filed. Thus, later choices of last-filed date will include more cases, all of which will have a shorter observation window than cases filed earlier do.

136. In principle, PACER's information on case termination dates could be used directly to measure which cases are still active by the latest-filed date. I am not confident enough in the quality of the PACER case termination information I have available, given the way I have set up the data for analysis.

APPENDIX B: GO-LIVE DATES FOR U.S. DISTRICT COURTS

U.S. District Court	Date the Court's CM/ECF System Went Live
Middle District of Alabama	April 05, 2004
Northern District of Alabama	February 15, 2014
Southern District of Alabama	May 01, 2003
District of Alaska	January 03, 2006
District of Arizona	August 01, 2005
Eastern District of Arkansas	July 01, 2005
Western District of Arkansas	August 15, 2005
Central District of California	January 01, 2008
Eastern District of California	January 03, 2005
Northern District of California	April 01, 2001
Southern District of California	September 01, 2006
District of Colorado	November 01, 2004
District of Connecticut	October 14, 2003
District of Delaware	February 14, 2005
District of District of Columbia	January 31, 2001
Middle District of Florida	July 12, 2004
Northern District of Florida	January 01, 2004
Southern District of Florida	October 12, 2006
Middle District of Georgia	October 01, 2004
Northern District of Georgia	July 15, 2004
Southern District of Georgia	May 02, 2005
District of Guam	August 01, 2005
District of Hawaii	December 05, 2005
District of Idaho	January 01, 2005
Central District of Illinois	July 01, 2004
Northern District of Illinois	January 18, 2005
Southern District of Illinois	January 20, 2004

U.S. District Court	Date the Court's CM/ECF System Went Live
Northern District of Indiana	November 03, 2003
Southern District of Indiana	July 01, 2002
Northern District of Iowa	June 16, 2003
Southern District of Iowa	October 01, 2004
District of Kansas	March 03, 2003
Eastern District of Kentucky	March 17, 2003
Western District of Kentucky	August 04, 2003
Eastern District of Louisiana	March 01, 2005
Middle District of Louisiana	January 18, 2005
Western District of Louisiana	June 07, 2004
District of Maine	January 01, 2003
District of Maryland	March 03, 2003
District of Massachusetts	May 10, 2003
Eastern District of Michigan	June 01, 2004
Western District of Michigan	August 01, 2001
District of Minnesota	March 28, 2004
Northern District of Mississippi	January 01, 2005
Southern District of Mississippi	December 20, 2004
Eastern District of Missouri	October 11, 2003
Western District of Missouri	October 01, 1999
District of Montana	November 07, 2005
District of Nebraska	October 11, 2002
District of Nevada	November 07, 2005
District of New Hampshire	October 03, 2003
District of New Jersey	January 01, 2004
District of New Mexico	January 01, 2007
Eastern District of New York	November 01, 1997
Northern District of New York	January 01, 2004
Southern District of New York	December 01, 2003

U.S. District Court	Date the Court's CM/ECF System Went Live
Western District of New York	October 04, 2003
Eastern District of North Carolina	October 01, 2005
Middle District of North Carolina	February 22, 2005
Western District of North Carolina	June 01, 2005
District of North Dakota	November 18, 2005
District of Northern Mariana Islands	November 15, 2004
Southern District of Ohio	September 01, 2003
Eastern District of Oklahoma	February 21, 2006
Northern District of Oklahoma	January 01, 2005
Western District of Oklahoma	October 14, 2003
District of Oregon	April 01, 2000
Eastern District of Pennsylvania	May 01, 2002
Middle District of Pennsylvania	March 03, 2003
Western District of Pennsylvania	July 05, 2005
District of Puerto Rico	January 01, 2004
District of Rhode Island	June 01, 2005
District of South Carolina	February 22, 2005
District of South Dakota	July 07, 2003
Eastern District of Tennessee	May 05, 2004
Middle District of Tennessee	July 05, 2005
Eastern District of Texas	February 01, 2004
Northern District of Texas	February 18, 2003
Southern District of Texas	September 01, 2004
Western District of Texas	May 01, 2006
District of Utah	May 01, 2005
District of Vermont	October 01, 2006
District of Virgin Islands	June 18, 2007
Eastern District of Virginia	May 03, 2005
Western District of Virginia	February 09, 2004

U.S. District Court	Date the Court's CM/ECF System Went Live
Eastern District of Washington	October 12, 2004
Western District of Washington	June 23, 2003
Northern District of West Virginia	March 07, 2005
Southern District of West Virginia	April 19, 2004
Eastern District of Wisconsin	November 16, 2002
Western District of Wisconsin	January 23, 2008
District of Wyoming	July 04, 2002

Source: Free Law Project compilation of information available from PACER website. *See Court Version Scraper*, https://web.archive.org/web/20210819144426/https:/court-version-scraper.herokuapp.com/courts.json (archived Aug. 19, 2021). Note that the Northern District of Ohio and Western District of Tennessee do not post go-live dates and will thus be excluded from analysis that requires that information.

APPENDIX C: NATURE OF SUIT CODE-SUBSTANTIVE LAW GROUPINGS

PACER code	PACER Nature of Suit Description	Substantive Group
410	Antitrust	antitrust
440	Other Civil Rights	civil rights
441	Voting	civil rights
442	Employment	civil rights
443	Housing/Accommodations	civil rights
444	Welfare1	civil rights
445	Amer w/Disabilities-Employment	civil rights
446	Amer w/Disabilities - Other	civil rights
448	Education	civil rights
371	Truth in Lending	consumer
480	Consumer Credit	consumer
110	Insurance	contract
120	Marine	contract
130	Miller Act	contract
140	Negotiable Instrument	contract
150	Recovery of Overpayment & Enforcement of Judgment	contract
151	Medicare Act	contract
152	Recovery of Defaulted Student Loans (Excl. Veterans)	contract
153	Recovery of Overpayment of Veteran's Benefits	contract
160	Stockholders' Suits	contract
190	Other Contract	contract
195	Contract Product Liability	contract
196	Franchise	contract
820	Copyrights	copyright and trademark
840	Trademark	copyright and trademark
893	Environmental Matters	environmental
460	Deportation	immigration
462	Naturalization Application	immigration

PACER code	PACER Nature of Suit Description	Substantive Group
463	Habeas Corpus - Alien Detainee	immigration
465	Other Immigration Actions	immigration
710	Fair Labor Standards Act	labor
720	Labor/Management Relations	labor
730	Labor/Management Reporting & Disclosure Act [c]	labor
740	Railway Labor Act	labor
751	Family and Medical Leave Act	labor
790	Other Labor Litigation	labor
791	Employee Retirement Income Security Act	labor
210	Land Condemnation	other
220	Foreclosure	other
230	Rent Lease & Ejectment	other
240	Torts to Land[a]	other
245	Tort Product Liability[b]	other
290	All Other Real Property	other
400	State Reapportionment	other
422	Appeal 28 USC 158	other
423	Withdrawal 28 USC 157	other
430	Banks and Banking	other
450	Commerce	other
470	Racketeer Influenced and Corrupt Organizations	other
490	Cable/Sat TV	other
610	Agriculture[c]	other
620	Other Food & Drug[c]	other
625	Drug Related Seizure of Property 21 USC 881[c]	other
630	Liquor Laws[c]	other
640	RR & Truck[c]	other
650	Airline Regulations[c]	other
660	Occupational Safety/Health[c]	other
690	Other	other
810	Selective Service	other

PACER code	PACER Nature of Suit Description	Substantive Group
870	Taxes (U.S. Plaintiff or Defendant)	other
871	IRS-Third Party 26 USC 7609	other
875	Customer Challenge 12 USC 34101[c]	other
890	Other Statutory Actions	other
891	Agricultural Acts	other
892	Economic Stabilization Act[c]	other
894	Energy Allocation Act[c]	other
895	Freedom of Information Act	other
896	Arbitration	other
899	Administrative Procedure Act/Review or Appeal of Agency Decision	other
900	Appeal of Fee Determination Under Equal Access to Justice Act[c]	other
950	Constitutionality of State Statutes	other
830	Patent	patent
510	Motions to Vacate Sentence	prisoner petitions
530	General	prisoner petitions
535	Death Penalty	prisoner petitions
540	Mandamus & Other	prisoner petitions
550	Civil Rights	prisoner petitions
555	Prison Condition	prisoner petitions
560	Conditions of Confinement	prisoner petitions
850	Securities/Commodities/Exchange	securities
861	HIA (1395ff)	Social Security
862	Black Lung (923)	Social Security
863	DIWC/DIWW (405(g))	Social Security

PACER code	PACER Nature of Suit Description	Substantive Group
864	SSID Title XVI	Social Security
865	RSI (405(g))	Social Security
310	Airplane	tort
315	Airplane Product Liability	tort
320	Assault, Libel, & Slander	tort
330	Federal Employers' Liability	tort
340	Marine	tort
345	Marine Product Liability	tort
350	Motor Vehicle	tort
355	Motor Vehicle Product Liability	tort
360	Other Personal Injury	tort
362	Personal Injury- Medical Malpractice	tort
365	Personal Injury- Product Liability	tort
367	Personal Injury - Health Care/ Pharmaceutical Personal Injury/Product Liability	tort
368	Asbestos Personal Injury Product Liability	tort
370	Other Fraud	tort
375	False Claims Act	tort
380	Other Personal Property Damage	tort
385	Property Damage Product Liability	tort

Notes:

[a] This case type is listed under the "Real Property" section of the PACER Nature of Suit document.

[b] This case type is listed under the "Real Property" section of the PACER Nature of Suit document.

[c] According to the list provided on the PACER website, codes 610, 620, 630, 640, 650, 660 (all in the PACER Forfeiture/Penalty category) have been eliminated and are listed for reference only; the same is true for codes 730 (Labor/Management Reporting & Disclosure Act), 875 (Customer Challenge 12 USC 34101), 892 (Economic Stabilization Act), 894 (Energy Allocation Act), and 900 (Appeal of Fee Determination Under Equal Access to Justice Act). Because I do not know the date of elimination, I have included these codes in this table and do nothing to exclude them from my analysis.