JOURNAL OF
EMPIRICAL LEGAL STUDIES

Journal of Empirical Legal Studies
Volume 10, Issue 2, 253–287, June 2013

Building a Taxonomy of Litigation: Clusters of Causes of Action in Federal Complaints

*Christina L. Boyd, David A. Hoffman, Zoran Obradovic, and Kosta Ristovski**

This project empirically explores civil litigation from its inception by examining the content of civil complaints. We utilize spectral cluster analysis on a newly compiled federal district court data set of causes of action in complaints to illustrate the relationship of legal claims to one another, the broader composition of lawsuits in trial courts, and the breadth of pleading in individual complaints. Our results shed light not only on the networks of legal theories in civil litigation but also on how lawsuits are classified and the strategies that plaintiffs and their attorneys employ when commencing litigation. This approach permits us to lay the foundation for a more precise and useful taxonomy of federal litigation than has been previously available, one that, after the Supreme Court's recent decisions in *Bell Atlantic v. Twombly* (2007) and *Ashcroft v. Iqbal* (2009), has also arguably never been more relevant than it is today.

> The idea of "a plain and short statement of the claim" has not caught on. Few complaints follow the models in the Appendix of Forms. Plaintiffs' lawyers, knowing that some judges read a complaint as soon as it is filed in order to get a sense of the suit, hope by pleading facts to "educate" (that is to say, influence) the judge with regard to the nature and probable merits of the case, and also hope to set the stage for an advantageous settlement by showing the defendant what a powerful case they intend to prove.
>
> Judge Richard Posner
> *American Nurses Ass'n v. Illinois* (1986)

*Address correspondence to Christina Boyd, University at Buffalo, SUNY, Department of Political Science, 520 Park Hall (North Campus), Buffalo, NY 14260; email: cLboyd@buffalo.edu or David Hoffman, Beasley School of Law, Temple University, 1719 N. Broad St., Philadelphia, PA 19122; email: David.Hoffman@temple.edu. Boyd is Assistant Professor of Political Science at University at Buffalo, SUNY; Hoffman is James E. Beasley Professor of Law at Temple University Beasley School of Law; Obradovic is Professor of Computer and Information Sciences and Director of the Center for Data Analytics and Biomedical Informatics at Temple University; Ristovski is a Ph.D. candidate in the Department of Computer and Information Sciences at Temple University.

We appreciate the support for our research from the University at Buffalo's Baldy Center for Law & Social Policy and Temple University's Beasley School of Law. We also gratefully acknowledge the research assistance of Geoffrey Bauer, Keith Blackley, Matthew Canan, Antima Chakraborty, Melissa Jabour, Nicholas Mozal, Jacqueline Sievert, and Elizabeth Young and the preliminary analyses of Daniel Katz and Michael Bommarito. We received helpful comments on this project from colleagues at workshops at Temple Law School, Duke Law School, North Carolina Law School, Rutgers-Camden Law School, and the University of Georgia Department of Political Science, participants at the Midwest Political Science Association and Stanford Conference on Empirical Legal Studies 2012 annual meetings, attendees at a joint research workshop with the Federal Judicial Center and the Administrative Office of the U.S. Courts, and Joe Cecil, Dawn Chutkow, Kevin Clermont, Scott Dodson, Marc Galanter, Lonny Hoffman, William Hubbard, Greg Mandel, Morris Ratner, Paul Stancil, Rick Swedloff, Steve Subrin, and the editors and anonymous reviewers at *JELS*. Authors are listed in alphabetical order.

Judge Posner's opinion in *American Nurses* illustrates the dilemma of the complaint drafter. Attorneys often want to tell a story in their pleading—to frame the litigation favorably for an attentive judge or his or her clerk. But the stories that unlock the courthouse door change. Once, judges appeared to prefer Hemingway's concise prose, as "[t]he draftsmen of the Civil Rules proceeded on the conviction, based on experience at common law and under the codes, that pleadings are not of great importance in a lawsuit" (Wright et al. 2002). But it is now evident that William Gaddis is a better lodestar. After lower court decisions in the 1980s, 1990s, and early 2000s, reams of scholarship, and the Supreme Court's eventual input in *Bell Atlantic v. Twombly* (2007) and *Ashcroft v. Iqbal* (2009), well-counseled plaintiffs will create a detailed and plausible factual narrative, despite what the Rules say. Indeed, plaintiffs' complaints are said to be more important now than at any time since the drafting of the Rules in the 1930s. But, given selection effects, pleading strategy is as difficult as ever to study systematically.

In this article, we focus on a particular aspect of the pleading story: the channeling by plaintiffs of their factual narrative into particularized legal claims, or *causes of action*. These causes of action exemplify the cross-cutting tensions of pleading writ large. The rules of civil procedure nominally permit liberal joinder of claims in one suit, and the failure to plead a particular legal claim will often lead to preclusion in later cases. However, increased judicial skepticism of private plaintiffs, and consequent doctrinal changes in pleading, counsel against bringing causes of actions that the facts do not immediately suggest (Miller 2010). Thus, though the parties may still cast a wide net, it seems likely that the more strategically wise choice is to be attentive to the relationship between causes of action, and to attempt, to the extent possible, to frame a coherent nexus of causes of action in a particular complaint.

To better understand this problem, we collected and culled a set of over 2,000 federal complaints and coded the alleged causes of action in each. We then analyzed the relationship between these complaints based on their underlying causes of action—over 7,400 of them—using spectral clustering. Cluster analysis provides a means to objectively classify large data sets and has been widely used for the sorts of taxonomic exercises that are critical foundational work in many sciences. In this present study, cluster analysis allows us to describe and summarize civil complaints, in isolation and in relationship to one another, in ways that previous work simply could not do. Our analysis demonstrates that there are stable relationships between the causes of action found in this set of complaints—indeed, we find that causes of action cluster into eight typical patterns. These patterns permit us to develop a more precise and therefore useful taxonomy of federal litigation than has been previously available.

I. Complaints and Causes of Action

A. From Writ to Cause of Action

At the heart of doctrine lies the *cause of action*. In every U.S. jurisdiction, parties may join together distinct theories that they believe justify legal relief. That is, they may bring

multiple causes of action; they may even join federal and state legal theories together in federal court if they "form part of the same case or controversy under Article III of the United States Constitution" (28 U.S.C. 1367(a)). But this modern cause of action practice is a *relatively* recent procedural innovation.

In their original incarnation, the ancient system of writs coincided with distinctive theories of legal relief. As Bracton wrote, "there may be as many forms of action as there are causes of action" (Plucknett 1956:37). Each writ was issued in response to fact patterns that reoccurred, and particular writs came to be used for common complaints. Over time, these patterned writs were fixed—fact patterns had to be shaped to fit the available procedural formula. Judges also greatly restricted the joinder—that is, the ability to bring together distinct legal theories in one "case"—of distinct writs. The resulting system was arcane, technical, and extremely expensive to access (Hepburn 1897).

New York's famous Field Code sought to replace this obscure system and start afresh. It employed the term "cause of action" to describe those groupings of facts that would result in judicial intervention. The term originally therefore implied that the plaintiff had identified a set of circumstances for which there was a known remedy (Subrin 1987). Even so, the Field Code limited joinder of these causes of action based on the substantive legal nature of each (Hazard 1988). For example, New York permitted the joinder of just seven general kinds of action in one complaint: contracts; injuries by force to person or property; injuries without force to person or property; injuries to character; claims to recover real property; claims to recover personal property; and claims against a trustee (N.Y. Laws, c.379 (1848)). Arguments over joinder bedeviled theorists, who viewed the intellectual incoherence of the term "cause of action" as a precipitating cause (Gavit 1930).

Reflecting this hostility, Charles Clark, the reporter for and force behind the original Federal Rules, believed that the cause of action was nothing more (or less) than "an aggregate of operative facts, a series of acts of events, which gives rise to one more legal relations of right-duty enforceable in the courts" (Clark 1924). Over time, this realistic conception of the cause of action came to dominate, providing the architecture for the innovative federal rules regime (Bone 1989; Sherwin 2008). The Rules famously avoid the term "cause of action" entirely, instead focusing on a "claim for relief," and the type of factual notice that would apprise the defendant of the nature of the theories arrayed against it. That is, as originally proposed, the Federal Rules do not require plaintiffs to plead causes of action at all, and Rule 18, which governs joinder, enables bringing together theories of relief without regard to the underlying doctrinal categories that had dominated practice. Since most states' procedural codes are modeled on the Federal Rules, one might have imagined that the cause of action, like the writ, was extinct.

B. The Modern Practice and Theory of Multiple Claim Pleading

But nothing could be further from the truth. Most lawyers continue to plead independent causes of action in both federal and state court. They do so for many reasons. Primarily, the conservative nature of local legal culture de-motivates changes to traditional pleading practices (Main 2001), and lawyers are told that increasing the number of causes per case will lead to higher rates of recovery (Berger et al. 2005; Eisenberg 2007). That lesson begins

in law school, where professors teach students to channel fact patterns into discrete causes of action, framed by courses like "Tort," "Contract," "Employment Law," or "Property." Important jurisdictions also continue to model their pleading rules on the Field Code, and lawyers may fairly believe that they are safer complying with that more restrictive set of rules in all complaints. The 1993 Federal Rule Amendments may have encouraged broad pleading by requiring mandatory disclosure of "claim or defense" relevant evidence. Finally, claim splitting may result in preclusion in a later filed case (Restatement (Second) of Judgments § 24 (1982)). Thus, there were traditionally few immediate costs in most cases to pleading as many specific causes of action per complaint as a clever lawyer could possibly imagine.

"Few," but not none. Plaintiffs wishing to preserve their choice of forum must plead carefully: for instance, an explicit (or lurking) federal cause of action may enable removal from state court.[1] And overpleading may irritate the trial judge. Emphasizing how common cause-of-action-centered pleading is, courts often complain that overpleading law obscures the merits, permitting plaintiffs to avoid investing in their cases early on and winnowing their theories of relief. The federal reports are full of such laments, emphasizing Rule 8(a)'s command that a complaint be short and plain. In *Cesnick v. Edgewood Baptist Church* (1996), the exasperated Eleventh Circuit noted that a complaint was "so muddled that it was difficult to discern what the appellants [were] alleging beyond the mere names of certain causes of action." In *Davis v. Coca-Cola Bottling Co. Consolidated* (2008), the Court lamented that "[i]f the framers of the Federal Rules of Civil Procedure could read the record in this case—beginning with the plaintiffs' complaint . . . they would roll over in their graves." Though noting that dismissals for prolixity are supposed to be rare, the Ninth Circuit recently cautioned a plaintiff that it was unfair to "burden her adversary with the onerous task of combing through a 733 page pleading just to prepare an answer that admits or denies such allegations, and to determine what claims and allegations must be defended or otherwise litigated" (*United States ex rel. Cafasso. v. General Dynamics C4 Systems* (2011)).[2] Plainly, district courts do not enjoy the task of "wast[ing] half a day in chambers preparing the 'short and plain statement,' which Rule 8 obligated plaintiffs to submit" (*McHenry v. Renne* (1996)).

All of which is to say that from the passage of the Federal Rules until quite recently, liberal joinder and liberal pleading combined to recommend that attorneys set forth as many causes of action as they felt would pass a very loose (but not nonexistent) judicial

---

[1]Additional strategic complexities abound. Class action plaintiffs are motivated (particularly post-CAFA) to limit the number of causes of action in their complaints and thus decrease the number of potentially class-defeating individual issues. Conversely, "master complaints" in MDL cases will contain numerous causes of action to increase the size of the consolidated case.

[2]Many dismissals resting on Rule 8(a) will result from a motion for a more definite statement, or an initial motion to dismiss under Rule 12, and will be without prejudice. A similar effect will obscure our understanding of the effect of Rule 11 sanctions, which may be imposed when plaintiffs have failed to reasonably investigate the factual basis of their claims. It is highly unlikely that trial judges will write opinions on such nondispositive orders. That is, simply because we observe few opinions relying on Rule 11 or Rule 8, does not mean that those Rules do not meaningful influence attorney practices on the ground.

scrutiny. Whether recent changes in judicial views on pleading will or have changed attorney practices regarding causes of action is a topic we will explore later in this article.

II. Data

To study causes of action empirically, we first developed a large database of civil complaints. As noted above, this project focuses exclusively on the study of federally litigated complaints. Although this ensures that lawyers are playing by the same rules when filing their cases, something that is desirable from an experimental standpoint, it is also a practical requirement for a project with data coming directly from a large number of complaints. Unfortunately, state court complaints remain difficult and expensive to retrieve in large and representative numbers, something that presents certain generalizability limitations to any empirical study, like ours, that focuses exclusively on federal trial courts.

A. RECAP Complaint Data

Truly random selection of federal complaints remains nearly impossible, since, for example, many complaints are not available electronically, paper complaints are archived around the country, and the traditional retrieval of a large sample of them (electronically or not) would be cost prohibitive. To gather our federal district court complaints, we turned to RECAP, a free digital archive of federal district court and bankruptcy case documents developed in 2008 by the Center for Information Technology Policy at Princeton University. RECAP's repository is sourced through Internet users of PACER (Public Access to Court Electronic Records), the federal judiciary's pay service for accessing electronic court records. The RECAP database now contains over 5 million federally filed documents, a number that represents approximately 1 percent of PACER's current library.[3]

Within the RECAP electronic database, we identified approximately 80,000 electronically available civil complaints, from which we could retrieve unique identifying information like a case's district name and docket number.[4] Our goal with these RECAP complaints was to build a data set that somewhat resembles the population of civil complaints filed in (or removed to) federal courts. To do this, we selected a stratified sample of 2,500 complaints from the RECAP database based on an estimation of filed cases' issue

---

[3]RECAP obtains electronic documents from federal courts when individuals install an extension into their Firefox Internet browser, which, after being installed, transfers a copy of any file downloaded from PACER into the RECAP file-sharing directory. RECAP was seeded with several million documents in 2009, when Aaron Swartz entered a library at which the government had begun a free trial of PACER (Schwartz 2009). Swartz managed to download around 20 percent of the entire PACER database at that time, which amounts to 19,856,160 pages of text.

[4]The presence of a federal complaint in the RECAP database and our sample does not guarantee that it is the original, preamendment(s) complaint. However, where multiple complaints from a single case were available, we coded the original filing. Within our data, 99 percent of our coded complaints were the original. In the few instances when we relied on an amended complaint, it was because the original complaint was not available. For purposes of the clustering, these amended complaints were treated identically to original complaints.

areas.[5] Specifically, we used the "nature of suit" (NOS) code, a single-issue code that is designed to serve as a summary of a case identified by the plaintiff's attorney at filing, to develop a sample that roughly reflects the Administrative Office of the U.S. Courts (AO) database's overall distribution of NOS codes, and thus rough issue areas in lawsuits, in federal district courts. According to Eisenberg and Schlanger (2003), "for researchers seeking to identify all federal district court cases in a certain subject matter category, it is clear that the AO database [and its NOS code variable] is the easiest, and perhaps the most reliable, method of doing so . . ."[6] After the selection of our 2,500 complaint sample, we found and removed two duplicate complaints (based on docket-number errors; most duplicates were identified prior to the 2,500 case sample) and 62 complaints with no nonrelief causes of action. We also excluded 427 cases because they were a part of multidistrict litigation (MDL) and, as such, were likely subject to a different pleading process and overall litigation strategy than other cases (Williams & George 2010).[7] After this data partitioning, we are left with a final sample of 2,009 complaints, all of which were filed between 2000 and 2008.

The dark gray bars in Figure 1 depict the NOS code distribution for all cases filed in federal district courts in 2007, as recorded by the AO. Comparing those to the light gray bars, which display the same distribution of NOS codes for our 2,009-complaint database, indicates that our data are overrepresentative across most issue areas. This is due to our exclusion, as noted above, of prisoner petition, Social Security, and MDL cases from our data. When we also exclude the prisoner petition and Social Security cases from the AO's distribution of cases, as we do with Figure 1's black bars, we can see that our data much more closely approximate the overall distribution of cases in the federal district courts for the remaining categories of cases. This comparison between the black bars and light gray bars indicates that we have a lower percentage of personal injury tort, bankruptcy, and real property cases, a noticeably higher percentage of cases with contract, civil rights, property rights, and other statute NOS codes, and relatively similar levels of labor, personal property tort, forfeiture, tax, and immigration cases. Short of being able to draw a random sample of

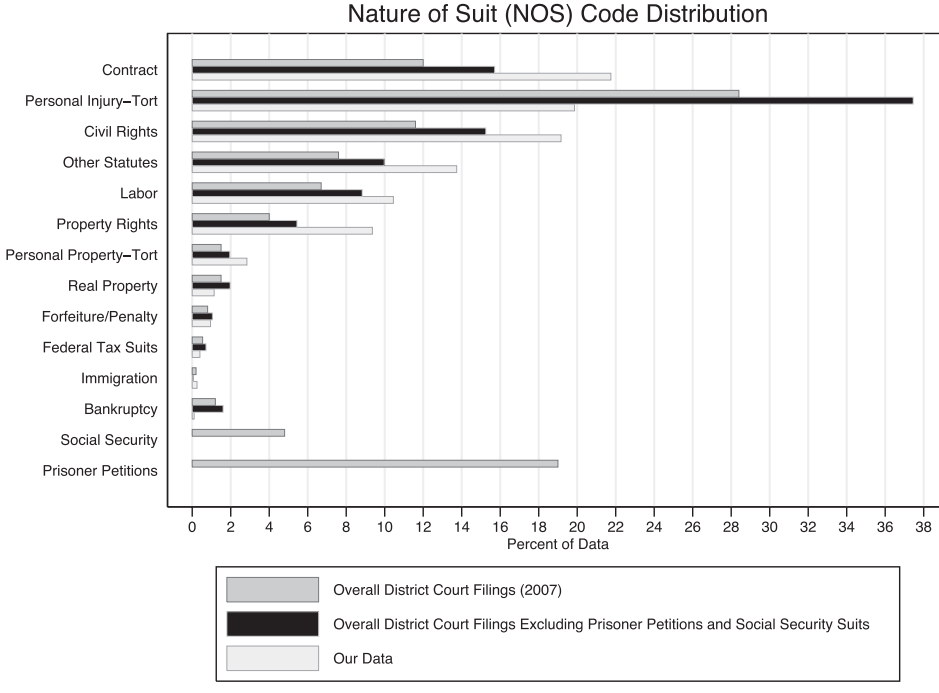
---

[5]Before the selection of our 2,500-case sample, we excluded prisoner habeas petitions and Social Security complaints as well as those complaints filed by a pro se plaintiff. Social Security cases would be difficult to fit into our larger coding scheme, present no opportunity for multiple-claim pleading, and are usually pled as a matter of rote. The exclusion of prisoner petitions, like the exclusion of pro se plaintiffs, represents a judgment call that these cases are unlikely to be subject to the same kinds of pleading strategies as ordinary civil litigation. For one, they are governed by an elaborate set of rules, statutory and otherwise, which police their content and format (see, e.g., The Prisoners Litigation Reform Act, 42 U.S.C. § 1997e). Notwithstanding these regimes, our inspection of these excluded cases suggests that there remains an enormous number of very hard to parse complaints, which would have significantly increased the likelihood of erroneous coding for our purposes. Further study of the content and organization of these excluded complaints and choices that are made are a topic all on their own.

[6]As has been noted elsewhere by, for example, Hadfield (2005) and Schlanger (2003), the NOS codes themselves fall well short of being ideal for summarizing the complex content of a case. In addition, we also note that the nature of RECAP and the way that it is populated nonrandomly by users means that it is likely not perfectly representative of overall federal cases *within* NOS categories. For example, within the personal injury torts NOS category, it is very likely that RECAP's contents contain a higher percentage of large-scale tort cases and fewer individual tort actions than the AO data's distribution.

[7]To identify the MDL cases within our database, we relied on the Administrative Office's "disposition" and "source" variables (e.g., Administrative Office of the U.S. Courts 2007).

Figure 1: The distribution of Nature of Suit (NOS) codes, by broad category.



NOTE: The displayed distributions are for all cases filed in federal district courts in 2007 and for all 2007 filed cases minus those that involve prisoner petitions or Social Security claims.

SOURCES: Administrative Office of the U.S. Courts, Federal Judicial Caseload Statistics, March 31, 2007, Public Access to Court Electronic Records (PACER).

district court complaints, we believe this distribution of data gives us the next best thing in our quest to empirically examine the anatomy of federal complaints.

B. Categorizing Causes of Action

With a data set of usable and relatively representative federal complaints in hand, our next important task was to identify and categorize the causes of action within these complaints. We began by coding each cause of action in every complaint, a task that is greatly eased in federal complaints by relatively standardized pleadings and wide use of labeled counts. The first step of our coding method was simple: we separately listed each cause of action as a distinct item. Where the plaintiff labeled the causes of actions with counts or numbers, this task was anodyne: each count or subsection was coded as its own cause of action. When the plaintiff failed to use divisions, we coded each clearly alleged cause of action from the relevant paragraphs. Our method was intended to be conservative—that is, we did not code a cause of action unless that plaintiff clearly seemed to intend to plead one.

We excluded purported causes of action where the plaintiff simply asserted a claim for relief—for example, for damages, arbitration, an injunction, or attorney fees. In our

Table 1: 19 Cause of Action Categories in Data

Main Category	*Notable Examples*	*% of Causes of Action (Raw* N*)*
Agency	Respondeat superior liability, vicarious liability	1.13% (84)
Bad faith	Bad faith	0.39% (29)
Breach of fiduciary duty	Breach of fiduciary duty, dissipation of trust assets	1.36% (101)
Civil rights/constitutional law	ADA claims, employment discrimination	16.06% (1,191)
Consumer protection	Unfair and deceptive trade practices, antitrust	9.32% (691)
Contract	Breach of contract, warranty	13.89% (1,030)
Enforcement	Civil forfeiture, foreclosure	1.44% (107)
Equitable contract	Account stated, equitable estoppel	3.7% (274)
Fraud	General fraud, fraudulent concealment	6.46% (479)
Intellectual property	Trademark, copyright	6.86% (509)
Labor	ERISA, collective bargaining	5.89% (437)
Process causes	Judicial review, appeal	0.35% (26)
Property	Trespass, eminent domain	1.02% (76)
Racketeering/criminal activities	Common law conspiracy, RICO	0.97% (72)
Regulatory	Administrative Procedure Act, CERCLA	2.54% (188)
Securities	Securities Exchange Act, Investment Advisers Act	1.71% (127)
Tax	Recovery of taxes paid, tax liability	0.19% (14)
Tort	Negligence, defamation, wrongful death	26.15% (1,939)
Obscure, unknown, or unusable	n/a	0.37% (42)

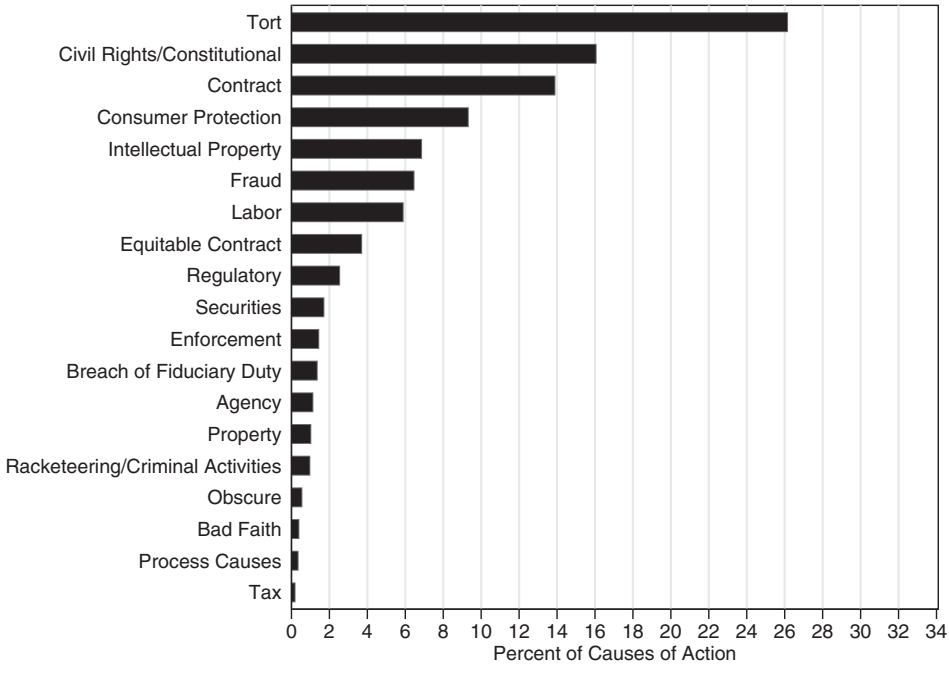
NOTE: The middle column lists notable examples of each category, and the right-hand column provides descriptive statistics (percentages and raw numbers) on the distribution of causes of action within our data. See Appendix A for further details on the coding of the causes of action.

considered view, a claim for a particular remedy is not ordinarily or best understood as a cognizable cause of action. Within our data, there were 480 non-MDL causes of action classified as bare claims for relief.[8] Excluding such claims, our final sample of cases contains 7,415 individual causes of action.

Categorizing these causes of action was not as simple. We first developed a list of general categories of causes of action, which loosely corresponded with the NOS codes, but that also drew on our understanding of the nature of pleading practice and common form-book complaints. The result was 18 general buckets listing types of causes of action (and an eventual 19th "obscure, unknown, or unusable" category). We list these types in the left-hand column of Table 1. Our next step was to assign each of the 7,415 causes of action to a category. That process ranged from easy text normalization (e.g., "Breach of Contract" and "Contract Breach" or "Warranty" and "Warrantee" claims) and the use of similar names to describe a similar concept (e.g., wantonness and recklessness describe a similar legal claim in tort) to more complex coding (ensuring that all causes of action, whether based in

[8]In an earlier version of this article, we included such bare claims for relief in the analysis. As we discuss in footnote 11, the exclusion of these claims for relief from the cause of action data has a modest, but predictable, effect on our clustering analysis. We did not, however, exclude a small number of causes of action we label "process causes"—for example, those seeking judicial review. Those causes of action, unlike the bare claims of relief, are not entirely derivative on substantive actions.

Figure 2: The distribution of coded causes of action, by category, in data.



NOTE: See the text, Table 1, and Appendix A for further details on the coding of causes of action.

common law or statutory in nature, objectively fit within a single category). We list notable examples of these for each category in the middle column of Table 1. The full details of our cause of action classification codebook are reported in Appendix A.

In the left-hand column of Table 1 and in Figure 2, we report the descriptive statistics for the coded cause of action categories in our data. As we can see, tort causes of action dominate, making up over 26 percent of our causes of action. Also composing over 13 percent of the causes of action each are the contract and constitutional law/civil rights categories.

III. METHODS

To better understand the composition of civil complaints, we set out to categorize cases based on the similarity of their individual causes of action. We utilized a quantitative procedure known as *cluster analysis*, which aims to objectively group similar objects based on information found in the data (Everitt et al. 2011). Data classification like this, often referred to as taxonomy, is commonplace in many sciences like biology, zoology, psychiatry, and even medicine. As the work in this area argues, while classification of data through clustering can be informative for summarizing the data, its results can provide a far more

foundational and fundamental understanding of the topic of interest. As Everitt et al. (2011) argue:

> Medicine provides a good example. To understand and treat a disease it has to be classified, and in general the classification will have two main aims. The first will be *prediction*—separating diseases that require different treatments. The second will be to provide a basis for research *aetiology* [etiology]—the causes of different types of diseases. It is these two aims that a clinician has in mind when she makes a diagnosis. (2011:3–4)

A clustering classification of civil complaints can be similarly foundational to the development of a larger, more nuanced understanding of litigation. Just as with basic science and medicine, such classification can serve both prediction and etiological purposes. For prediction, identifying different classes of cases can be informative for legal scholars, practitioners, and educators and can have implications as wide-ranging as how empiricists control for different types of cases, how law schools formulate effective curriculums, when law firms decide to deploy specialist attorneys or pursue particular litigation strategies, and the degree to which we can effectively predict, for example, case outcomes, case lengths, and termination methods. Just one example of the etiological value of this type of work has to do with the dispute formation process prior to and then following civil filings, a topic that previous empirical research has revealed to be a treasure trove of opportunities for understanding what cases eventually make their way through the court system (Miller & Sarat 1980–1981; Boyd & Hoffman forthcoming).

A. Associational Methods in Legal Studies

Several recent papers have employed cluster analysis and other more general data association methods—sometimes referred to as network analysis—to analyze legal data. A number of these authors have examined opinion citations as a network, an effort that allows them to draw inferences about the importance and strategic use of precedent and the overall relationship between courts (Fowler et al. 2007; Lupu & Voeten 2011; Fowler & Jeon 2008). Legal scholars have also used associational data methods to analyze specialized areas of law (Bommarito et al. 2011; Strandburg et al. 2006) and legal actors (Katz & Stafford 2010).

Three recent empirical legal analyses most closely mirror that conducted here. Pleasence et al. (2004) examine the clustering of English and Welsh individuals' justiciable problems and find that problems relating to the family tend to occur together (like divorce, domestic violence, and child-related problems), as do those relating to social exclusion (e.g., homelessness and unfair police treatment) and medical negligence with mental health issues. Cross and Lindquist (2009) and Yung (forthcoming) use cluster analysis to group U.S. circuit court judges based on their decision-making characteristics, the results of which provide novel insight into how to best "judge" judges and classify them based on their varying judicial characteristics.

B. Cluster Analysis of Causes of Action

Turning to our project, we utilize *spectral clustering* to classify and group the cases in our data based on the similarity of their individual causes of action. While there are a variety of

clustering methods available (Tan et al. 2005), many have parameterization issues and are biased based on the particular structure of the data set. Spectral clustering overcomes this by permitting the illustration of complex clusters of arbitrary shapes (Ng et al. 2002). Spectral clustering is based on graph cut theory, which takes into account the similarity function between pairs of data points. The spectral clustering algorithm seeks to cut a weighted undirected graph into k clusters such that the edges within each partition (for us, connections between cases) have a high weight or degree of similarity while the edges between nodes in different partitions have a low weight or dissimilarity among cases.

To determine our clusters via spectral clustering, we (1) defined the proper similarity measure and (2) determined the appropriate number of clusters for our data. The extended Jaccard coefficient (similarity measure) between two case vectors ignores 0-0 matches to prevent a large number of cases being considered similar due to *not* containing many of the same causes of action. The measure also accounts for the presence of causes of action that occur more than once in an individual complaint (Tan et al. 2005).

With the similarity measure in hand, we investigated the appropriate number of stable clusters that captured the inherent structure in our data set. It is reasonable to assume that the method has captured the inherent structure in the data set if clusters obtained on different subsamples of our data set are similar (a similarity measure close to 1). We clustered and compared pairs of subsamples, following Ben-Hur et al. (2002), repeated this 100 times, and also repeated this for different values of k. The process was completed when, for a particular k, the distribution of similarities between pairs of subsamples stop being concentrated close to 1 (for details, see Ben-Hur et al. 2002). After the experiments with our data, we determined that as k moves up to 9, there is a large change in distribution of the similarities away from 1, indicating instability. Therefore, we select $k = 8$ as the number of stable clusters.

To show that this spectral clustering defines reasonable groupings, we plot a gray-scale image of the similarity matrices before and after the clustering in Figure 3. In the figure, brighter pixels signify higher similarities. Before clustering, cases were randomly spread over the data set so there are no interesting patterns, as Figure 3(a) illustrates. By contrast, Figure 3(b) on the right displays eight bright square blocks around the main diagonal, meaning that similarities are high between cases inside our clusters. The size of each square is relative to the number of cases present in the cluster (see Table 2 for descriptive statistics on the clusters).

Appendix B provides the technical details of this spectral clustering, including our clustering algorithm, our similarity measure, and the complete procedure for determining the final number of and assignment to our eight clusters.

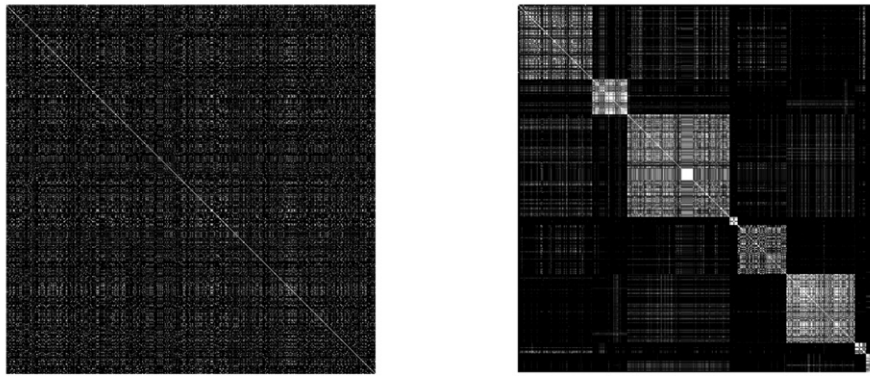
IV. Results

A. Resulting Clusters

The spectral cluster analysis thus results in eight clusters of causes of action. Each of the eight clusters represents a discrete grouping of causes of action—that is, the kinds of causes of action that tend to be brought together in complaints. That there are a limited number

Figure 3: Similarity matrices between cases before and after spectral clustering.

(a) Similarity matrix of the data before clustering (b) Similarity matrix of the clustered data



NOTE: Brighter pixels on gray-scale images represent higher similarity while dark ones indicate low similarity. Figure 3(a) is made with a random arrangement of the cases in the data set while data points in Figure 3(b) are arranged in cluster order, with the eight light boxes on the diagonal indicating the clusters.

of patterns to cause of action pleading makes sense: after all, causes of action must be based on facts that can give rise to a plausible claim for relief. There are only so many general ways that individuals seeking recourse in federal court can be *generally* harmed.[9]

Table 2 details the distribution, both in percentages and raw numbers, of our 19 coded causes of action across the eight clusters yielded from the analysis. Many of the categorized causes of action rest largely within a single cluster. Ninety-two percent of the constitutional law/civil rights claims, for example, are found in Cluster 6. Eighty-eight percent of the labor claims are in Cluster 2, 69 percent of the fraud claims are in Cluster 3, and 81 percent of the regulatory claims are in Cluster 8. The most striking in their consistency are securities and intellectual property claims. Over 98 percent of securities causes of action are located in Cluster 4, and 94 percent of intellectual property claims lie in Cluster 5. As Table 2 indicates, other claims are not nearly as predictable in their ultimate cluster location. Agency claims split nearly evenly between Clusters 3 and 6 and breach of fiduciary duty causes of action are divided largely between Clusters 1, 3, and 4. These resulting cluster locations for different types of legal claims tells us a great deal about the breadth with which certain legal claims are pled as well as detailing, more generally, the underlying content of cases brought in federal trial courts.

---

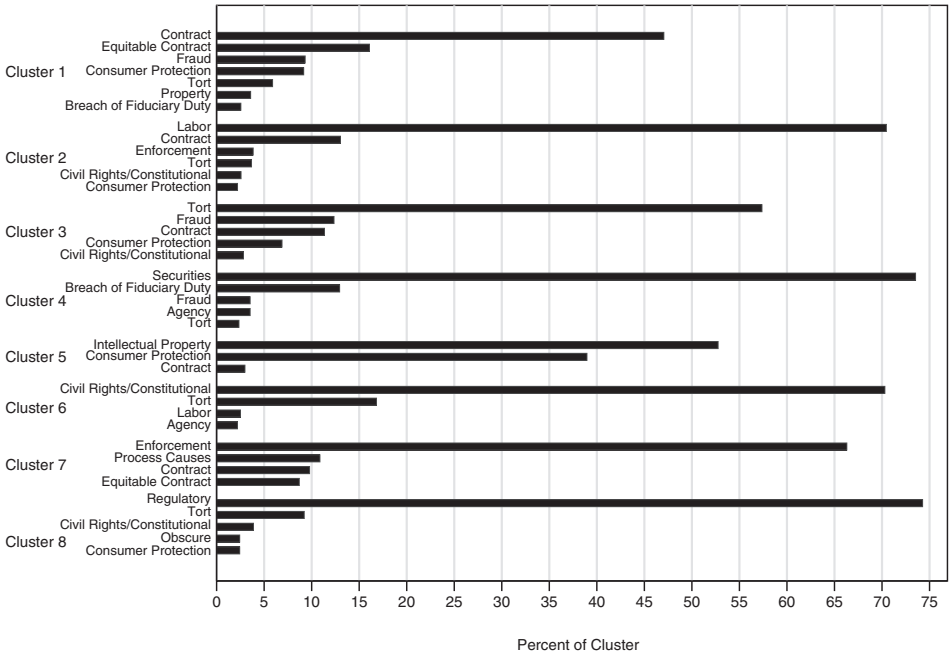
[9]It is worth noting here that we do not claim that there are only eight kinds of cases in federal court. That would reach beyond our data. Rather, we assert that for this sample, we can say that the most replicable cluster pattern finds eight typical groupings of causes of action: any more would divorce causes of action that are more tightly linked together than they are separated from others and any fewer would artificially lump together causes of action that have little to do with one another. As an anonymous reviewer notes, clustering of kinds of causes of action together may reflect the structure, and increasing specialization, of law firm practice.

Table 2: Distribution of 19 Categories of Causes of Action Among the Eight Clusters

	Clusters								
Cause of Action	*#1*	*#2*	*#3*	*#4*	*#5*	*#6*	*#7*	*#8*	*Total*
Agency	3.57%	0	45.24%	7.14%	2.38%	40.48%	0	1.19%	100%
	(3)	(0)	(38)	(6)	(2)	(34)	(0)	(1)	(84)
Bad faith	55.17%	0	44.83%	0	0	0	0	0	100%
	(16)	(0)	(13)	(0)	(0)	(0)	(0)	(0)	(29)
Breach of fiduciary duty	31.68%	4.95%	35.64%	21.78%	1.98%	0.99%	0.99%	1.98%	100%
	(32)	(5)	(36)	(22)	(2)	(1)	(1)	(2)	(101)
Civil rights/ constitutional law	0.08%	1.18%	6.38%	0.08%	0	91.60%	0	0.67%	100%
	(1)	(14)	(76)	(1)	(0)	(1,091)	(0)	(8)	(1,191)
Consumer protection	16.64%	1.74%	26.77%	0	51.09%	3.04%	0	0.72%	100%
	(115)	(12)	(185)	(0)	(353)	(21)	(0)	(5)	(691)
Contract	57.38%	6.89%	29.61%	0.19%	2.62%	2.33%	0.87%	0.10%	100%
	(591)	(71)	(305)	(2)	(27)	(24)	(9)	(1)	(1,030)
Enforcement	14.02%	19.63%	2.80%	0	4.67%	1.87%	57.01%	0	100%
	(15)	(21)	(3)	(0)	(5)	(2)	(61)	(0)	(107)
Equitable contract	73.72%	2.19%	13.87%	0.73%	4.74%	1.09%	2.92%	0.73%	100%
	(202)	(6)	(38)	(2)	(13)	(3)	(8)	(2)	(274)
Fraud	24.43%	1.67%	69.31%	1.25%	0.63%	1.88%	0	0.84%	100%
	(117)	(8)	(332)	(6)	(3)	(9)	(0)	(4)	(479)
Intellectual property	3.14%	0	2.55%	0	93.91%	0	0.20%	0.20%	100%
	(16)	(0)	(13)	(0)	(478)	(0)	(1)	(1)	(509)
Labor	0.46%	87.87%	2.52%	0	0	8.92%	0.23	0	100%
	(2)	(384)	(11)	(0)	(0)	(39)	(1)	(0)	(437)
Obscure	2.44%	2.44%	51.22%	0	4.88%	26.83%	0	12.20%	100%
	(1)	(1)	(21)	(0)	(2)	(11)	(0)	(5)	(41)
Process causes	3.85%	3.85%	3.85%	3.85%	0	30.77%	38.46%	15.38%	100%
	(1)	(1)	(1)	(1)	(0)	(8)	(10)	(4)	(26)
Property	59.21%	0	28.95%	0	0	11.84%	0	0	100%
	(45)	(0)	(22)	(0)	(0)	(9)	(0)	(0)	(76)
Racketeering/ criminal activities	12.50%	1.39%	52.78%	1.39%	2.78%	27.78%	0	1.39%	100%
	(9)	(1)	(38)	(1)	(2)	(20)	(0)	(1)	(72)
Regulatory	1.60%	0.53%	6.38%	0	0.53%	9.57%	0	81.38%	100%
	(3)	(1)	(12)	(0)	(1)	(18)	(0)	(153)	(188)
Securities	0	0	1.57%	98.42%	0	0	0	0	100%
	(0)	(0)	(2)	(125)	(0)	(0)	(0)	(0)	(127)
Tax	92.86%	0	0	0	0	7.14%	0	0	100%
	(13)	(0)	(0)	(0)	(0)	(1)	(0)	(0)	(14)
Tort	3.82%	1.03%	79.53%	0.21%	0.93%	13.46%	0.05%	0.98%	100%
	(74)	(20)	(1,542)	(4)	(18)	(261)	(1)	(19)	(1,939)
Total causes of action falling within cluster	16.94% (1,256)	7.35% (545)	36.25% (2,688)	2.29% (170)	12.22% (906)	20.93% (1,552)	1.24% (92)	2.78% (206)	100% (7,415)

NOTE: Unless otherwise noted, percentages listed are for the row, that is, the percent of a cause of action's occurrence located in a particular cluster. The raw number of causes of action for each cell is located in parentheses.

Figure 4: Causes of action composing each cluster.



NOTE: Cluster numbers are labeled on the far left of the graph. To aid in the graph's readability, causes of action composing 2 percent or less of a cluster are excluded.

We look more closely at the legal composition of each cluster in Figure 4. This figure depicts, for each cluster in our data, the percentage breakdown of causes of action. As we can see from this figure, each cluster has one or two dominant causes of action. From the figure, we can also start to get a strong sense of patterns in the content of civil complaints, as measured through their combinations of causes of action, that we are likely to see in litigation over and over again.[10] These legal patterns can be summarized as follows.

- **Cluster 1:** One of the most heterogeneous clusters, the plurality of the claims in Cluster 1 are of a contract nature (47 percent). Equitable contract claims are about 16 percent, and consumer protection, fraud, and tort claims each are over 5 percent. A common case falling in this cluster is a commercial contract case accompanied with a quasi-contract claim.
- **Cluster 2:** This cluster contains a large number of labor cases (over 70 percent of causes of action), including many claims for enforcement of ERISA plans. Contract causes of action amount to about 13 percent of the cluster. A representative case

[10]That attorneys repeatedly attempt joinder of these kinds of causes of action does not mean that they are properly brought together in federal court under 28 U.S.C. § 1367. Our analysis simply indicates that attorneys wished it to be so.

assigned to this cluster involves a lawsuit brought by a pension fund against an employer.

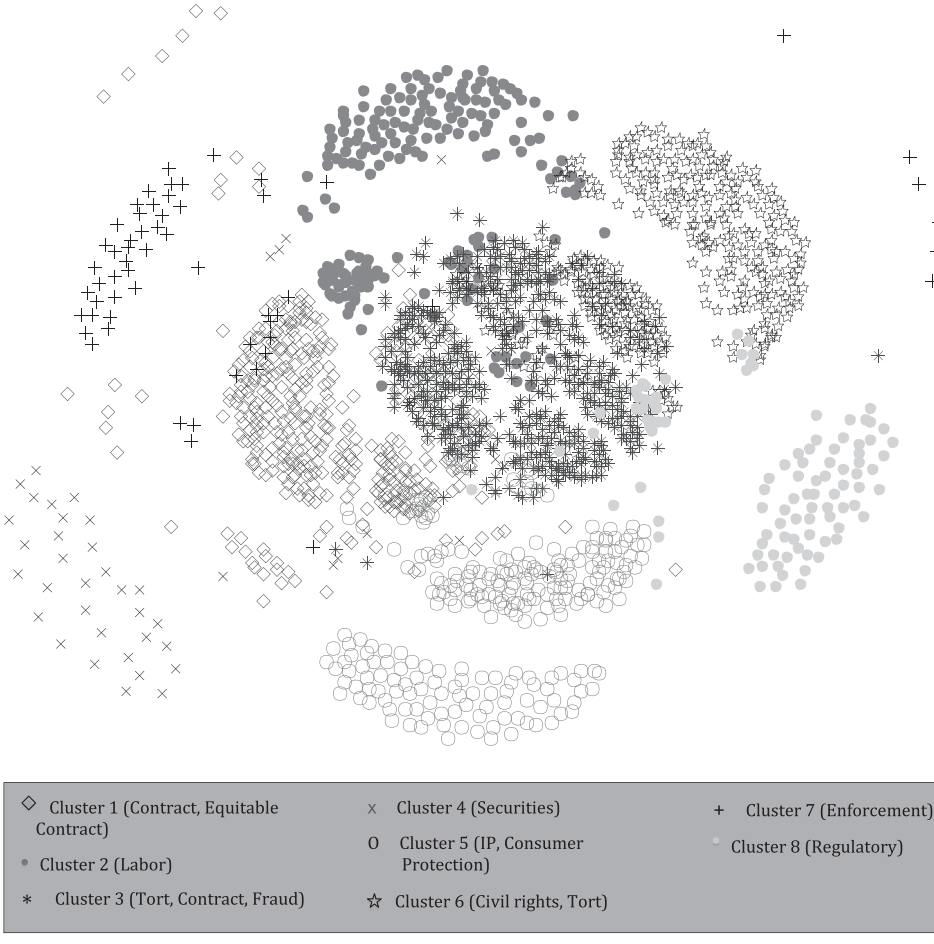
- **Cluster 3:** Tort causes of action make up the majority of this cluster, but do so only at less than 58 percent of the cluster. Contract claims make up about 11 percent, fraud claims 12 percent, and consumer protection 7 percent. The tort claims in this cluster are often products liability disputes (which are often accompanied by contract-warranty claims) and ordinary accident cases. This cluster contains products liability cases as well as more straightforward personal injury torts.
- **Cluster 4:** The cluster is characterized by the presences of securities law claims (nearly 75 percent of causes of action in cluster) plus, to a lesser degree, breach of fiduciary duty claims, with fraud, agency, and tort also accounting for less than 5 percent of the cluster each. This cluster represents, for example, federal securities class action practice.
- **Cluster 5:** Cluster 5 is dominated, more than any other cluster, by two (related) causes of action rather than one: intellectual property (53 percent) claims paired with consumer protection causes of action (39 percent). A representative case assigned to this cluster is a trademark cause of action paired with one for unfair trade practices.
- **Cluster 6:** Seventy percent of causes of action are civil rights/constitutional in nature. The only other notable claim, at 17 percent of the cluster, is tort based. A representative case assigned to this cluster is an alleged Title VII violation paired with a tort-like intentional infliction of emotional distress.
- **Cluster 7:** Enforcement actions dominate this cluster, accounting for 67 percent of the causes of action. Contract, equitable contract, and process-related causes each make up about 10 percent of the cluster. A typical Cluster 7 case involves a civil forfeiture action for money seized for violations (or intended violations) of the Controlled Substances Act.[11]
- **Cluster 8:** Regulatory actions (74 percent of causes of action), in particular claims under the Administrative Procedure Act against the United States, seeking agency action dominate this cluster. Example cases assigned to this cluster seek adjudication of an immigration asylum claim or seek declaratory relief based on an alleged violation of the Endangered Species Act by the U.S. Department of Interior.

To better understand the composition of our eight clusters and how these clusters relate to each other, we graphically and spatially depict them in Figure 5. Each point in Figure 5 represents an individual case and the distance between points represents their

---

[11]As noted above (see footnote 8 and related text), we do not code separately listed damage or relief pleas as causes of action in our data presented here. However, we note that when these relief claims *are* included in the clustering as causes of action (in alternative modeling not reported here), their cases are frequently assigned to Cluster 7. Without relief causes of action, these cases become distributed across our clusters, since the substantive causes of action that the enumerated relief claims are attached to are better able to influence the ultimate cluster assignment of the case. Because of this, Cluster 7 becomes much smaller, more discrete, and more informative on the case's internal structure.

Figure 5: Fruchterman-Reingold force directed graph layout for the clusters of cases.



NOTE: The figure results from a Fruchterman-Reingold force directed graph layout for weighted graphs implemented in R. Distances between vertices (cases) are approximately proportional to the similarity between them. To maximize clarity, we do not display the graph edges.

relationship to one another. Within the figure, points close together typically fall within the same cluster, something we indicate through the use of different symbols.

As Figure 5 illustrates, there is significant overlap between Clusters 6 (civil rights/constitutional) and 3 (tort/contract/fraud) as well as Clusters 3 and 1 (contract/equitable contract), which the reader can observe in the middle of the figure. The remaining clusters are spread further from each other. Indeed, the cluster that is least like the others is one in which federal securities law claims dominate (Cluster 4) and for which the resulting combination of legal theories is very unlike all others in the data. From this, we can conclude that federal securities law cases have less in common—legally—than do cases

based in ordinary commercial torts and contract claims. They rest on a set of facts and doctrine that is consequently more remote.

Some of the other cluster interrelationships are of note. Cluster 5, which is dominated by intellectual property and consumer protection claims and is located in the bottom of the figure, is, spatially speaking, very distant from, for example, Cluster 6 (our civil rights/constitutional law cluster located in the top-center of Figure 5). This sort of separation makes a great deal of legal sense, since it is difficult to imagine many shared causes of action between these two types of cases. A similar division can be seen between Cluster 8 on the bottom right (regulatory actions) and Cluster 2 on the top (labor cases).

B. Causes of Action Relationships

With a more nuanced understanding of our eight clusters in hand, we turn now to a closer examination of the relationship between individual causes of action. As discussed above, the breadth of pleading practice and, more generally, liberal joinder seemingly permit a wide array of legal claims to be regularly pled together. This is generally confirmed in our data based on the legal composition of the cases falling into our eight clusters above. However, since the cluster analysis is case based, it does not provide extensive details on the underlying causes of action, meaning we continue to lack a full understanding of how legal claims of different types interact with each other within lawsuits.

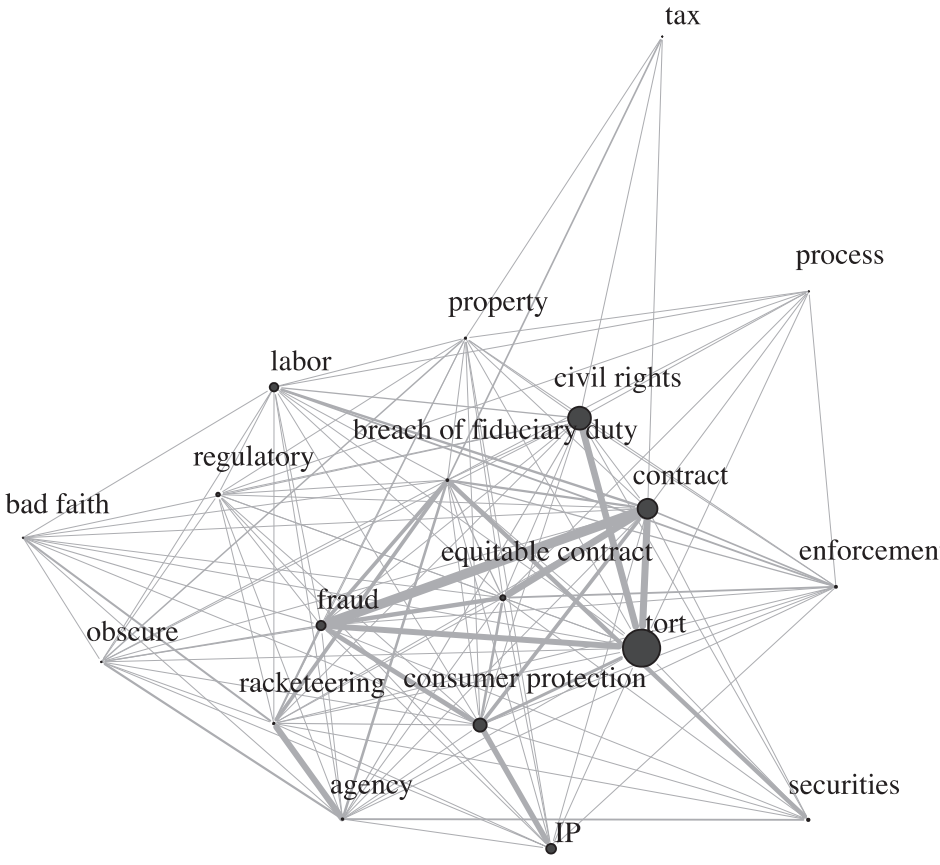
To tackle this next step, we begin with Figure 6's visualization of this cause of action relationship within our data. Within the figure, the nodes (shaded dots) represent the spatial location of causes of action, with the node's relative size indicating the frequency of the cause of action in the data. The edges (gray lines) depict the relationship between these causes of action, with stronger co-occurrences represented with thicker lines.

As Figure 6 shows with its thick edges, contract and fraud claims are often brought together, as are tort and contract and tort and fraud causes of action. Other strong relationships include consumer protection claims to intellectual property claims and agency and securities causes of action to those claims involving breach of fiduciary duty. Figure 6 may be just as interesting for what it tells us about weak relationships between certain types of claims. Some causes of action, like those involving tax, are rather isolated. Other causes of action that make up a sizable proportion of the data and have numerous edges, like constitutional law/civil rights, securities, and labor, simply do not have nearly as consistent patterns in their outward legal relationships as do tort, fraud, and contract claims. Take labor causes of action as an example. As the figure indicates with the numerous edges bursting out from the labor node, there are a number of different cause of action relationships for labor claims. However, none of the edges are darker than others, indicating that no systematically predictable patterns emerge.

To calculate the co-occurrence of two types of causes of action while taking into consideration how common an individual cause of action is in our data, we use the following statistic:

$$\log \frac{f(i, j)}{f(i) f(j)}$$

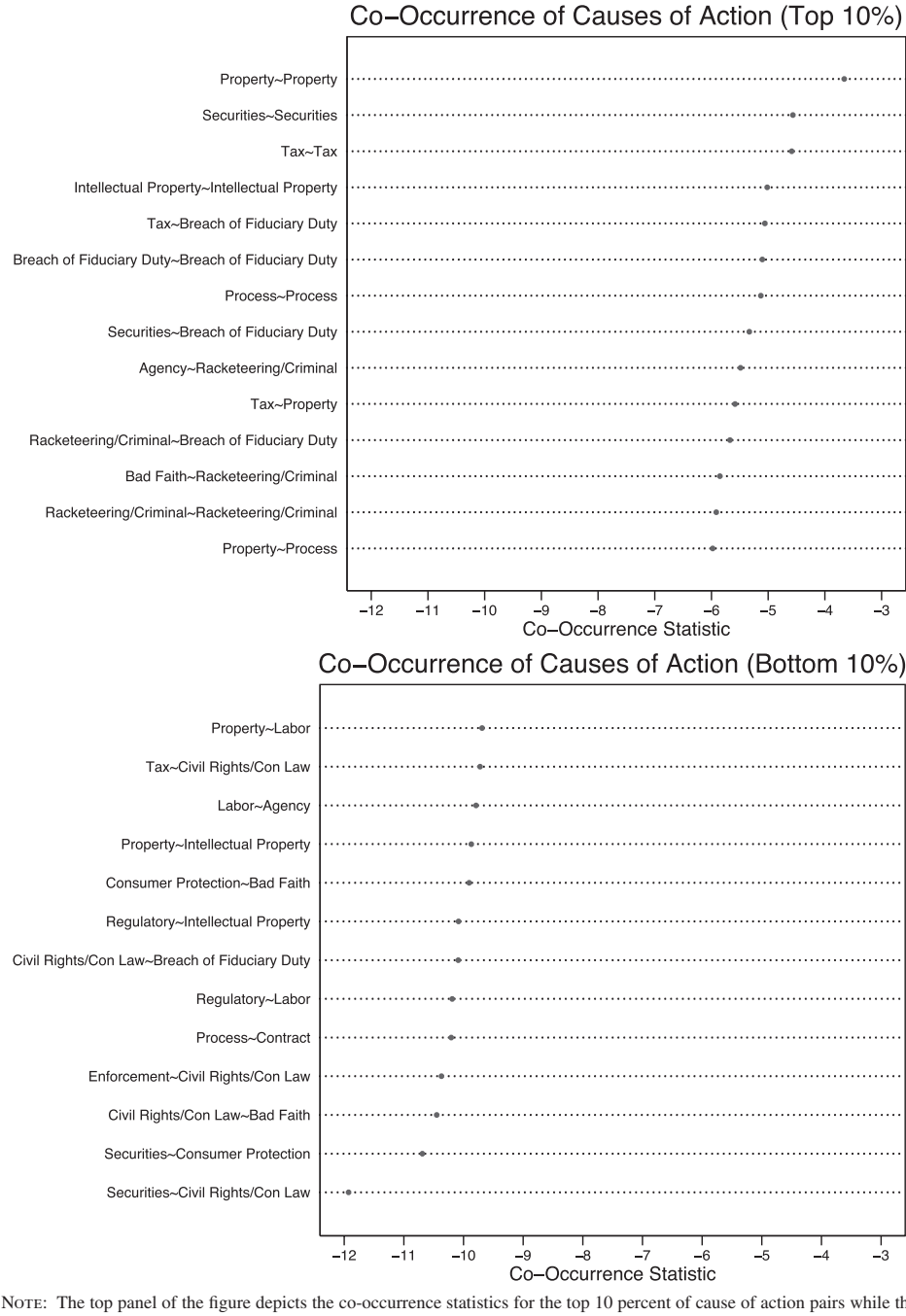
Figure 6: Fruchterman-Reingold force directed graph layout for the clusters of causes of action.



NOTE: The figure results from a Fruchterman-Reingold force directed graph layout for weighted graphs. Distances between nodes (causes of action, the shaded dots) are approximately proportional to the similarity between them. A vector in which each element represents the occurrence of a certain cause of action in a particular case is assigned to the corresponding node. The similarity between causes of actions is measured applying the extended Jaccard coefficient (described in Appendix B) to assigned vectors. The size of each cause of action node is proportional to its incidence in the data.

where *f(i,j)* represents the rate of co-occurrence of causes of action *i* and *j*, and *f(i)* and *f(j)* are the rate of individual occurrence of these causes of action. The higher the calculated co-occurrence statistic is (or, in our case, the closer it is to 0), the stronger a cause of action pairing's relationship. To compute *f(i, j)*, we created and summed each cause of action pair for each case in our data. So, for example, if a case had three causes of action, this meant that there were three pairs: Pair (1) cause of action 1-cause of action 2, Pair (2) cause of action 1-cause of action 3, and Pair (3) cause of action 2-cause of action 3. With this co-occurrence rate, along with rate of occurrence of individual causes of action, this statistic provides insight into how strong cause of action relationships are in a way that should extend beyond our data. Figure 7 depicts the computed results for this statistic for the top 10 percent (top panel) and bottom 10 percent (bottom panel) of the paired causes of action in our data.

Figure 7: Dot plots of the co-occurrence statistics for causes of action pairs within the data.



NOTE: The top panel of the figure depicts the co-occurrence statistics for the top 10 percent of cause of action pairs while the bottom panel does the same for the bottom 10 percent of pairs in our data. Both figures exclude cause of action pairs that include an "obscure" cause of action.

Certain patterns emerge from these paired cause of action statistic depictions. In particular, we can see from Figure 7 that a number of causes of action are frequently paired with causes of action falling in the same legal category, including securities, intellectual property, (real) property, tax, and breach of fiduciary duty. For outward rates of cause of action pairing, the pairing of breach of fiduciary duty with both tax and securities is quite strong. On the low end of the co-occurrence statistic are relatively predictable weak cause of action relationships like, for example, real property with labor, securities with civil rights, and regulatory with intellectual property. In other words, it is just very rare for us to see these types of causes of action being pled together in a complaint. Beyond the cause of action relationships reported in Figure 7, the two most common cause of action pairings in our data, tort with tort ($N = 3{,}034$) and civil rights/constitutional with civil rights/constitutional ($N = 1{,}802$), fare relatively well under the co-occurrence statistic. The former relationship receives a −7.12 score, which places it in the 38th percentile, while the civil rights inward pairing comes in even stronger at −6.67 (21st percentile).

V. Discussion

What follows from this taxonomic exercise? Our cause-of-action-focused data set illustrates how each complaint creates a cloud of possible legal theories: a winnowing litigation follows until only a few, or one, is left. That one, discussed at length in a trial or appellate opinion, suggests that the litigation was a "contract" or "constitutional" or "patent" case (Boyd & Hoffman 2010, forthcoming), but it was originally no such thing. The causes of action that find their way into doctrinal exegesis are the residuum from a cluster of causes of action, *any of which* might have, in another turning of the world, survived. Understanding litigation as a *tournament of selection for causes of action*, beyond being valuable for describing and summarizing the anatomy of civil complaints, provides both predictive and etiological benefits. These, we argue, can readily translate into empirical, theoretical, and practical legal applications based on our taxonomic findings about the clustering of civil cases. In this section, we discuss two concrete applications of studying causes of action, and then describe some broader research paths for future work in this area.

A. Twombly *and Pleading Strategy*

Examining the content of pleadings in federal courts has likely never been so relevant as it is today with recent Supreme Court decisions in *Twombly* (2007) and *Iqbal* (2009) in the forefront of plaintiffs' (and, more generally, Court watchers') minds. Even before *Twombly*, federal courts were moving toward heightened scrutiny of pleading practices (Fairman 2003; Marcus 1986). But the *Twombly* and *Iqbal* cases made the trend more salient and (arguably) signified a change in how seriously trial courts should engage in their gate-keeping tasks. In these decisions, the Supreme Court, explicitly repudiating old case law that discounted the importance of the pleadings, rejected plaintiffs' complaints both because the allegations made were too vague and because they were implausible. This heightened scrutiny, especially in realms with perceived high discovery costs like antitrust

and environmental torts or weak(er) merits like civil rights and conspiracy, ordinarily would thus entail a greater degree of factual specificity by plaintiffs seeking to comply with the Court's demands. *Twombly* and *Iqbal*, whatever they may mean with respect to this perceived trend toward heightened scrutiny of pleading practice, have generated an immense outpouring of scholarly criticism (e.g., Miller 2010; Steinman 2010).

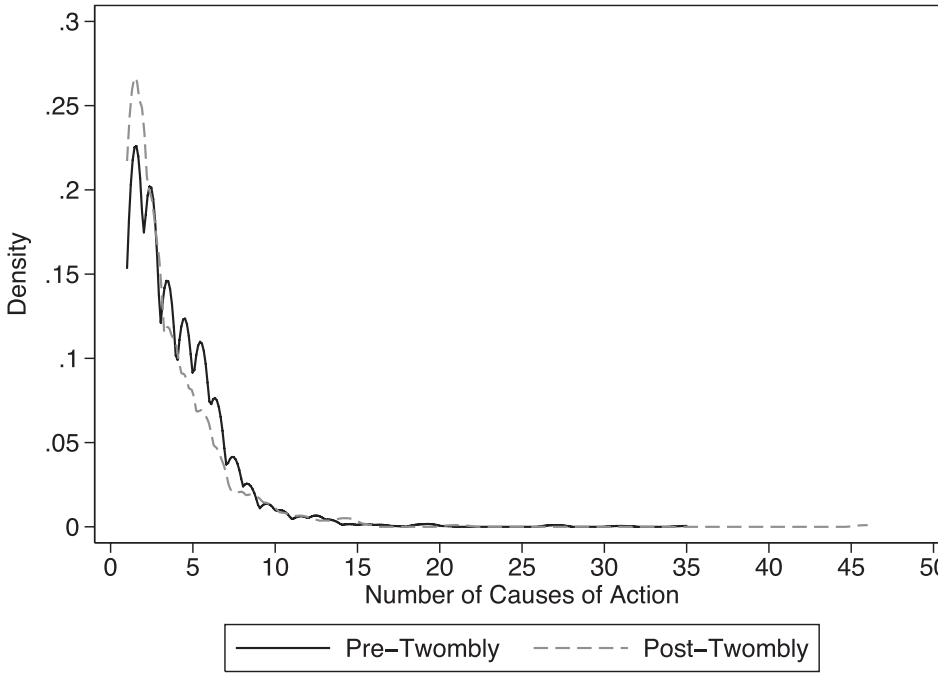
No matter the reaction, empirically assessing the effect of these cases on federal trial court practice has proven difficult. The most comprehensive empirical analysis to date, conducted by the Federal Judicial Center, looked at motions to dismiss filed before and after *Twombly* and *Iqbal* and found few significant changes in courts' grant rates (Cecil et al. 2011). Interpreting this nonfinding may be more difficult (Hoffman 2012). Putting aside concerns about finding appropriate samples, selection bias looms large when we study the operation of motions to dismiss. Perhaps attorneys have changed the content of their complaints—that is, made them "stronger"—after the pleadings revolution. All else equal, this would result in a lower overall grant rate for filed motions to dismiss. However, defense attorneys, who can read opinions and predict district court practice, will evolve to file such motions more rarely, saving their bullets for an especially bad complaint. That is, motion grant rates following *Twombly/Iqbal* will not fully illuminate how those decisions affected the kinds of cases that are prosecuted in federal court (Hubbard 2011).

What may be more informative and less biased moving forward, however, is an examination similar to that which we have conducted here. By focusing first and foremost on the content of the filing documents, we can better understand the strategy of plaintiff's attorneys in anticipation of the litigation to come, including what may be a more searching reading of the 12(b)(6) standard after these two important recent Supreme Court decisions. Indeed, it may be logical to assume that attorneys, growing concerned over the new interpretation of Rule 8(a) to require "plausible" claims of relief, may react by pleading more facts—or more plausible ones (Lee 2012:tab. 28). Causes of action that are difficult to support with facts immediately at hand—like conspiracy claims or ones resting on the defendant's intent—will be more difficult to allege. On net, we would thus expect that the number of causes of action in any given complaint will decrease.

Our data do not equip us to study this systematically—after all, we have no observations occurring after *Iqbal* and only a relatively small number immediately after *Twombly* (30 percent of the cases)—but a preliminary look at the question of the effect of this trend in case law on pleading breadth does show signs of promise. We plot in Figure 8 the kernel density of the number of causes of action per case in cases in our data filed before the Court's decision in *Twombly* on May 21, 2007 (solid line) and after (gray dashed line).

Figure 8 hints at exactly what we would expect. Cases filed after *Twombly* have a distribution in the numbers of causes of action that is centered largely from one to five, with a sharp dropoff thereafter. Cases filed before *Twombly*, however, have a distribution in the numbers of causes of action that is more widely spread from one to eight and includes a more gradual downward slope in density as the number of causes of action increases. Further, descriptive statistics and statistical tests confirm that the number of causes of action per case between pre- and post-*Twombly* cases are different from one another. Cases filed pre-*Twombly* have a median number of causes of action of three, those filed after have a median of two, and the two medians are statistically different from each other ($\chi^2 = 16.045$;

Figure 8: Kernel density plots of the number of causes of action per case.



NOTE: The density plot depicts the number of causes of action per case in the data, pre- and post-*Twombly*'s decision (May 21, 2007).

$p = 0.01$). In addition, the Wilcoxon rank-sum test, which is used for data that are not normally distributed, provides statistically significant evidence that the distributions of the pre- and post *Twombly* populations are not equal ($z = 4.785$; $p < 0.01$).

Thus, while these results are relatively preliminary in nature, they do seem to indicate that cases filed after *Twombly* appear to be, on net, pled with fewer causes of action. If these results are verified in future projects with more complaint-level data that expand into 2009, 2010, and beyond and the addition of systematic regression analysis, they will go a long way toward confirming the evolving nature of our pleading regime and the resulting changing strategies from attorneys in response to the change in the operative rules. Future work might also test if *Twombly*'s effect is issue-area or cluster specific. Though some work on particular issue areas—like employment litigation—has commenced, such studies are often methodologically compromised by relying on courts' reactions to filed motions (Noll 2010). However, because certain types of cases were arguably already subject to heightened pleading standards prior to *Twombly*, discerning the effect, if any, produced by the Supreme Court may remain statistically difficult going forward.

B. Clusters and Reforming NOS Codes?

This project's systematic examination of the contents of federal complaints also presents the opportunity to begin to evaluate the accuracy and value of NOS codes for classifying

their underlying civil lawsuits. As we note above, NOS codes are designated by the plaintiff's attorney at filing and are designed to summarize the case in a single code. Unquestionably, these codes serve important functions for many followers and scholars of the federal trial courts, including aiding in the reporting of subject matter descriptive statistics on filings and terminations and allowing scholars, including ourselves, to develop data samples for further study based on general case issue area. They have, however, been criticized as being "extremely sketchy" (Schlanger 2003:1699) and "not sufficiently reliable" (Eagan 2011:6) and, more generally, are recognized as being an imperfect method for summarizing complex underlying cases.

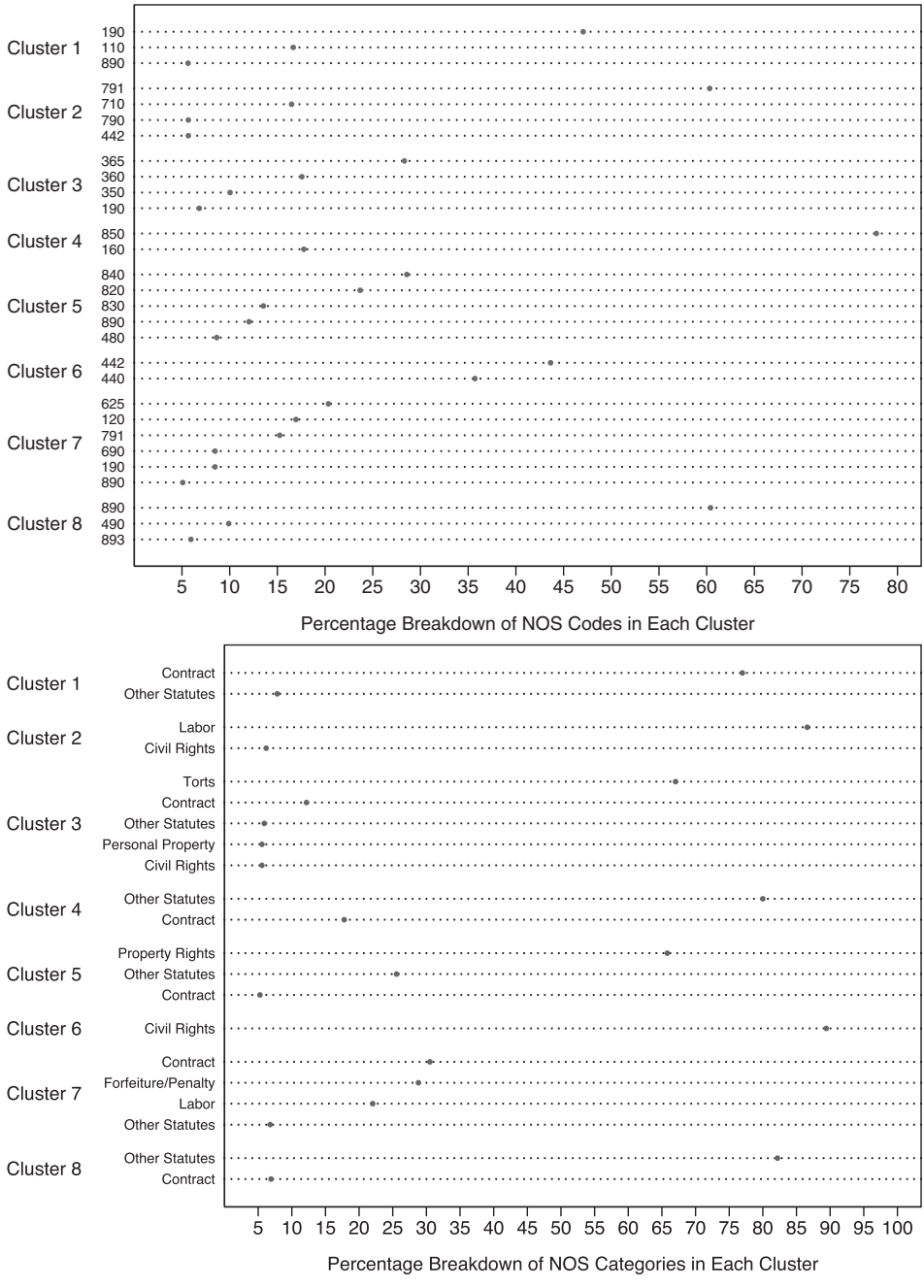
The clustering of cases based on complaints' underlying content presented here creates the potential for evaluating the reliability of NOS codes for serving this summary function. Unlike NOS codes, which are selected by a filing attorney, clustering presents an objective and stable classification of a case based on complaint content. To compare the clusters produced from the cases in our data to the NOS codes selected for the same cases, we turn to Figure 9. There, in the top panel, we depict the most frequently occurring NOS codes for each of our eight clusters. The bottom panel provides similar information but displays the data breakdown for the broader NOS categories rather than the individual NOS codes.

As the bottom panel of the figure indicates, seven of the eight clusters have a relatively homogenous NOS structure. In each of those seven clusters, a single NOS category accounts for at least 65 percent of the case classifications. In some of these clusters, like Cluster 6, that number reaches as high as 90 percent. What is more, the dominant NOS category is probably the category that would be expected given the legal content of the cluster, a conclusion that is aided by comparing Figure 9 with Figure 4. Cluster 6, our cluster containing a large number of civil rights and constitutional-law-based causes of action, is dominated by the "civil rights" NOS category, and specifically, the 442 (Employment) and 440 (Other Civil Rights) codes. The same is true, for example, for Cluster 2, where labor-related causes of action make up 70 percent of the cluster, and "Labor" NOS categories (especially code 791 (ERISA)) compose 85 percent of the classified cases.

This seems to be good news for NOS codes, and it probably is. However, Figure 9 also points to an imperfection in the NOS classification system. The fact that any cluster contains a variety of NOS categories indicates that those cases not classified in the dominant NOS category could well be considered to be NOS classification errors. To put this another way, under the clustering that we have conducted, the underlying cause of action structure of a complaint groups it with other similar cases, but the NOS classification of that case excludes some of these cases from that grouping. If we take this interpretation to the bottom panel of Figure 9 again, we could say that, at best, there is a 10 percent "error" in case grouping with NOS categories (Cluster 6) and, at worst, a 70 percent "error" rate (Cluster 7).

Of course, this is a relatively preliminary examination of this NOS-clustering relationship, but it is one that seems to indicate that revisiting NOS codings and the way that federal trial court cases are classified at filing could well be fruitful. Indeed, to further investigate and implement these possibilities, we would recommend that the Administrative Office of the U.S. Courts consider revising the information that it seeks from filing plaintiffs' attorneys. One potential way that this could be done is to require the plaintiffs' attorneys, who already fill out a civil cover sheet upon filing their complaints, to assign each

Figure 9: Percentage composition of NOS codes and categories for each cluster.



NOTE: Percentage composition of NOS codes (top panel) and NOS categories (bottom panel) for each of the clusters. For visual clarity, only those NOS codes/categories that account for over 5 percent of a cluster are depicted.

of their pled causes of action an NOS-like code. This would replace or augment the single-case-level NOS code assignment that currently takes place on the filed cover sheet. Applied in this way, trial courts would be enabled to easily employ this project's cluster assignment methodology to each case with a simple computerized formula, all the while reducing the problems implicit with attorney error by requiring NOS assignment at the case level.

C. The Future of Clustering and Applications

These two illustrations do not begin to exhaust the possible applications of complaint-level clustering. Future work, with larger data sets, may consider the following kinds of problems, among others.

- **Specialized Courts or Judges:** The ease and usefulness of complaint clustering may revitalize the debate about the normative benefits and practical costs of having more specialized courts (Baum 2010) or utilizing judge assignment based on expertise rather than randomization (e.g., Cheng 2008). With evidence like ours of civil cases combining into a limited number of legal patterns, it becomes far less daunting to think about such a change in the way we structure generalized trial courts.
- **Court Settlement Resources:** This sort of cause of action clustering evidence may provide courts the information that they need to more effectively and efficiently determine, from filing, when and in what cases to use court resources to push settlement.
- **Case Actor Strategy:** Clustering can also help attorneys and courts better plan for discovery and other case events. More generally, the more systematic information that case actors have about their case and how it compares to others, the more able they should be to make strategically wise decisions. For scholars, clustering presents the opportunity to examine a popular topic, like the effect of lawyer experience, specialization, and party resources (e.g., Galanter 1974), in more detail, more systematically, and earlier in a case than ever before.
- **Law School Curriculum:** That legal claims regularly combine in the ways that we have found may cause some to question the way that law schools silo topics into "Contracts," "Torts," "Property," "Intellectual Property," and so forth. We observe significant overlap between these legal areas in practice, and it is likely that the litigation of one cause of action will influence how another comes out.
- **Issue Area Controls in Empirical Work:** In empirical scholarship, we often use NOS codes, opinion text searches, case headers, and keynotes to narrow in on and statistically control for specific types of cases. As is well known, each of these, in its own way, presents a biased method of case summarization. Clustering, like we have done here, overcomes much of this.
- **Case Outcomes:** Finally, we believe that complaint clustering can provide important information on civil case outcomes. As scholarship like Galanter (2004), Eisenberg and Farber (1997), Clermont and Eisenberg (2002), and Clermont (2009) reveals, a case's "issue area" affects its likelihood of going to trial, settling,

or terminating via an adversarial motion, the probability of plaintiff success, and, when successful, the amount that is recovered. We may well be able to gather a more nuanced understanding of these outcome-related concepts, and their probabilities upon filing, by relying on clusters of complaints.

We recognize that this study, while the first of its kind, does have its limitations. One such limitation has to do with representativeness and statistical inference due to our reliance on RECAP to draw our data. Because of this reliance, it is possible that our data set's makeup is somehow different from all filed federal civil lawsuits. Although we cannot be certain that this is not the case, largely for the same data-gathering limitations that forced us to utilize RECAP to begin with, we believe that concerns about our data's representativeness are somewhat ameliorated by the issue-level distribution of our data (but not necessarily intraissue distribution, see footnote 6) discussed above with relation to Figure 1. Our decision to exclude pro se and prisoner petition cases also has implications for our analysis (beyond affecting the representativeness of our data). Perhaps more seriously, we cannot make any systematic claims about the structure of state complaints.[12] It is quite possible that lawyers in Code pleading jurisdictions have different cause-of-action strategies than those in jurisdictions whose procedural system follows the Federal Rules—although differences are likely to be muted by norms and customs.

VI. Conclusion

In this, the first systematic study of federal civil complaints, we illustrated the utility of examining an all but neglected data source on attorney strategy and behavior. As it turns out, most legal theories in the federal court sample we have gathered arise from state law causes of action—particularly, tort and contract claims. When joined together, we found that these underlying legal theories cohered to form predictable clusters of causes of action. Such clusters could easily form a firmer basis for etiological inquiry into litigation than the tools currently at hand. They might also help illuminate the effect of important changes in legal rules on attorney strategy and judicial behavior.

Complaints have long been ignored because pleadings themselves were deemphasized by the Rules. Indeed, we might as well have studied how a lawyer's use of font affected outcomes. But in the new era of revived pleading scrutiny, it seems time to turn our attention to a careful study of the documents that generate litigation. The project provides evidence that such an inquiry will not be fruitless.

References

Administrative Office of the U.S. Courts (2007) *Judicial Business of the United States Courts*. Available at <http://www.uscourts.gov/judbus2007/contents.html>.

[12]Of the cases in our data, some were undoubtedly removed from a state trial court, meaning that a small percentage of the complaints in our data are indeed state complaints. However, since the selection process for these complaints is quite unrandom, these data do not position us well to speak about state complaints more generally.

Baum, Lawrence (2010) *Specializing the Courts.* Chicago, IL: Univ. of Chicago Press.

Ben-Hur, Asa, Andre Elisseeff, & Isabelle Guyon (2002) "A Stability Based Method for Discovering Structure in Clustered Data," 17 *Pacific Symposium on Biocomputing* 6.

Berger, Vivian, Michael O Finkelstein, & Kenneth Cheung (2005) "Summary Judgment Benchmarks for Settling Employment Discrimination Lawsuits," 23(3) *Hofstra Labor & Employment Law J.* 45.

Bommarito, Michael J., Daniel Martin Katz, & Jillian Isaacs-See (2011) "An Empirical Survey of the Population of U.S. Tax Court Written Decisions," 30 *Virginia Tax Rev.* 523.

Bone, Robert G. (1989) "Mapping the Boundaries of a Dispute: Conceptions of Ideal Lawsuit Structure from the Field Code to the Federal Rules," 89(1) *Columbia Law Rev.* 1.

Boyd, Christina L., & David A. Hoffman (2010) "Disputing Limited Liability," 104(3) *Northwestern Law Rev.* 853.

—— (forthcoming) "Litigating Toward Settlement," 29 *J. of Law, Economics, & Organization.*

Cecil, Joe S., George W. Cort, Margaret S. Williams, & Jared J. Batallion (2011) *Motion to Dismiss for Failure to State a Claim After* Iqbal*: Report to the Judicial Conference Advisory Committee on Civil Rules.* Washington, DC: Federal Judicial Center.

Cheng, Edward K. (2008) "The Myth of the Generalist Judge," 61 *Stanford Law Rev.* 519.

Clark, Charles E. (1924) "The Code Cause of Action," 33(8) *Yale Law J.* 817.

Clermont, Kevin M. (2009) "Litigation Realities Redux," 84 *Notre Dame Law Rev.* 1919.

Clermont, Kevin M., & Theodore Eisenberg (2002) "Litigation Realities," 88 *Cornell Law Rev.* 119.

Cross, Frank B., & Stefanie Lindquist (2009) "Judging the Judges," 58 *Duke Law J.* 1383.

Eagan, Jaime A. (2011) "The Americans with Disabilities Act: An Empirical Look at U.S. District Court Litigation Involving Government Services and Public Accommodations Claims," working paper. Available at <http://ssrn.com/abstract=1870601>.

Eisenberg, Jonathan (2007) "Beyond the Basics: Seventy-Five Defenses Litigators Need to Know," 2(3) *Business Lawyer* 1281.

Eisenberg, Theodore, & Henry S. Farber (1997) "The Litigious Plaintiff Hypothesis: Case Selection and Resolution," 28 *Rand J. of Economics* 92.

Eisenberg, Theodore, & Margo Schlanger (2003) "The Reliability of the Administrative Office of the U.S. Courts Database: An Initial Empirical Analysis," 78(5) *Notre Dame Law Rev.* 101.

Everitt, Brian S., Sabine Landau, Morven Leese, & Daniel Stahl (2011) *Cluster Analysis,* 5th ed. Chichester, West Sussex: John Wiley & Sons.

Fairman, Christopher (2003) "The Myth of Notice Pleadings," 45 *Arizona Law Rev.* 987.

Fowler, James H., & Sangick Jeon (2008) "The Authority of Supreme Court Precedent," 30 *Social Networks* 16.

Fowler, James H., Timothy R. Johnson, James F. Spriggs, II, Sangick Jeon, & Paul J. Wahlbeck (2007) "Network Analysis and the Law: Measuring the Legal Importance of Supreme Court Precedents," 15(3) *Political Analysis* 324.

Galanter, Marc (1974) "Why the 'Haves' Come out Ahead: Speculation on the Limits of Legal Changes," 9(1) *Law & Society Rev.* 95.

—— (2004) "The Vanishing Trial: An Examination of Trials and Related Matters in Federal and State Courts," 1(3) *J. of Empirical Legal Studies* 459.

Gavit, Bernard C. (1930) "The Code Cause of Action: Joinder and Counterclaims," 30(6) *Columbia Law Rev.* 802.

Hadfield, Gillian K. (2005) "Exploring Economic and Democratic Theories of Civil Litigation: Differences Between Individual and Organizational Litigants in the Disposition of Federal Civil Cases," 57(5) *Stanford Law Rev.* 1275.

Hazard, Geoffrey C. (1988) "Forms of Action Under the Federal Rules of Civil Procedure," 63(5) *Notre Dame Law Rev.* 628.

Hepburn, Charles M. (1897) *The Historical Development of Code Pleading in America and England: With Special Reference to the Codes of New York, Missouri, California, Kentucky, Iowa, Minnesota, Indiana,*

Ohio, Oregon, Washington, Nebraska, Wisconsin, Kansas, Nevada, North Dakota. Cincinnati, OH: W.H. Anderson & Company.

Hoffman, Lonny (2012) "*Twombly* and *Iqbal*'s Measure: An Assessment of the Federal Judicial Center's Study of Motions to Dismiss," 6(1) *Federal Courts Law Rev.* 1.

Hubbard, William H. J. (2011) "The Problem of Measuring Legal Change, with Application to *Bell Atlantic v. Twombly*," working paper on file with the author.

Katz, Daniel M., & Derek K. Stafford (2010) "Hustle and Flow: A Social Network Analysis of the American Federal Judiciary," 71 *Ohio State Law J.* 3.

Lee, Emery G. III (2012) "Early Stages of Litigation Attorney Survey: Report to the Judicial Conference Advisory Committee on Civil Rules," March *Federal Judicial Center Publication*. Available at <http://www.fjc.gov/public/pdf.nsf/lookup/leeearly.pdf/$?le/leeearly.pdf>.

Lupu, Yonatan, & Erik Voeten (2011) "Precedent on International Courts: A Network Analysis of Case Citations by the European Court of Human Rights," presented at the 2011 Annual Meeting of the American Political Science Association, Seattle, WA.

Main, Thomas O. (2001) "Procedural Uniformity and the Exaggerated Role of Rules: A Survey of Intra-State Uniformity in Three States that Have Not Adopted the Federal Rules of Civil Procedure," 46(1) *Villanova Law Rev.* 311.

Marcus, Richard (1986) "The Revival of Fact Pleading Under the Federal Rules of Civil Procedure," 86 *Columbia Law Rev.* 433.

Meila, Mariana (2001) "The Multicut Lemma," 417 *UW Statistics Technical Report*.

Miller, Arthur R. (2010) "From *Conley* to *Twombly* to *Iqbal*: A Double Play on the Federal Rules of Civil Procedure," 60 *Duke Law J.* 1.

Miller, Richard E., & Austin Sarat (1980–1981) "Grievances, Claims, and Disputes: Assessing the Adversary Culture," 15(3/4) *Law & Society Rev.* 525.

Ng, Andrew Y., Michael I. Jordan, & Yair Weiss (2002) "On Spectral Clustering: Analysis and an Algorithm," in T. Dietterich, S. Becker, & Z. Ghahramani, eds., *Advances in Neural Information Processing Systems 14*, pp. 849–56. Cambridge, MA: MIT Press.

Noll, David L. (2010) "The Indeterminacy of *Iqbal*," 99 *Georgetown Law J.* 117.

Pleasence, Pascoe, Nigel J. Balmer, Alexy Buck, Aoife O'Grady, & Hazel Genn (2004) "Multiple Justiciable Problems: Common Clusters and Their Social and Demographic Indicators," 1 *J. of Empirical Legal Studies* 301.

Plucknett, Theodore F. T. (1956) *A Concise History of the Common Law*, 5th ed. Boston, MA: Little, Brown, & Co.

Schlanger, Margo (2003) "Inmate Litigation," 116(6) *Harvard Law Rev.* 1557.

Schwartz, John (2009) "An Effort to Upgrade a Court Archive System to Free and Easy," *New York Times* A16. Available at <http://www.nytimes.com/2009/02/13/us/13records.html>.

Sherwin, Emily (2008) "The Jurisprudence of Pleadings: Rights, Rules and *Conley v. Gibson*," 52 *Howard Law J.* 73.

Steinman, Adam N. (2010) "The Pleading Problem," 62 *Stanford Law Rev.* 1253.

Strandburg, Katherine J., Gabor Csardi, Jan Tobochnik, Peter Erdi, & Laszlo Zalanyi (2006) "Law and the Science of Networks: An Overview and an Application to the 'Patent Explosion'," 21 *Berkeley Technological Law J.* 1293.

Subrin, Stephen N. (1987) "How Equity Conquered Common Law: The Federal Rules of Civil Procedure in Historical Perspective," 135(4) *Univ. of Pennsylvania Law Rev.* 909.

Tan, Pang-Ning, Michael Steinbach, & Vipin Kumar (2005) *Introduction to Data Mining*. Boston, MA: Addison-Wesley.

Williams, Margaret S., & Tracey E. George (2010) "Who Will Manage Complex Civil Litigation? The Decision to Transfer and Consolidate Multidistrict Litigation," presented at the 2010 Annual Conference on Empirical Legal Studies, New Haven, CT.

Wright, Charles Alan, Arthur R. Miller, & Mary Kay Kane (2002) *Federal Practice and Procedure*, 6th ed. Eagan, MN: Thomson West Group

Yung, Corey Rayburn (forthcoming) "How Judges Judge," 107 *Northwestern Univ. Law Rev.*

Cases & Laws Cited

American Nurses Ass'n v. Illinois, 783 F.2d 716 (7th Cir. 1986).
Ashcroft v. Iqbal, 129 S. Ct. 1937 (2009).
Bell Atlantic v. Twombly, 550 U.S. 544 (2007).
Cesnick v. Edgewood Baptist Church, 88 F.3d 902 (11th Cir. 1996).
Davis v. Coca-Cola Bottling Co. Consol., 516 F.3rd 955 (11th Cir. 2008).
McHenry v. Renne, 84 F.3d 1172 (9th Cir. 1996).
United States ex rel. Cafasso v. General Dynamics C4 Sys., 637 F.3d 1047 (9th Cir. 2011).
Act of Apr. 12, 1848, Ch. 379, 1848 N.Y. Laws 521 (the "Field Code").
Restatement (Second) of Judgments (1982).
28 U.S.C. 1367(a) ("Supplemental Jurisdiction").

Appendix A: Coding Causes of Action

Below, we provide coding content details for the 18 substantive categories of causes of action in our data. Causes of action that could not be coded in one of these 18 categories were classified as "Obscure, Unknown, or Unusable," our 19th cause of action category.

1. Agency
    a. Aiding and Abetting
    b. Premises or Supervisory Liability in Tort
    c. Respondeat Superior
    d. Vicarious Liability
2. Bad Faith
    a. Bad Faith
3. Breach of Fiduciary Duty
    a. Breachf Fiduciary Duty—General
    b. Dissipation of Trust Assets
    c. Failure to Perform Duty as Corporate Officer
    d. Waste
4. Civil Rights/Constitutional Law
    a. 1st Amendment (or state equivalent)
    b. 5th Amendment (or state equivalent)
    c. Age
    d. Conspiracy
    e. Constitution—Non Civil Rights
    f. Disabilities
    g. Employment Federal and State
    h. Employment—Age
    i. Employment—Disabilities

j. Employment—Race
k. Employment—Retaliation
l. Employment—Sex
m. Employment—Termination/Discharge
n. Equal Access to Justice
o. Equal Protection
p. Failure to Intervene
q. False Arrest/Imprisonment
r. Force
s. General Discrimination/Access
t. Housing
u. Municipal/Supervisory
v. Process
w. Race/National Origin
x. Search
y. Sex

5. Consumer Protection

a. Antitrust
b. Debt Collection
c. Deceptive Trade/Business Practices
d. False Advertising
e. False Designation of Origin
f. Federal Miscellaneous
g. Lanham Act
h. State Whistleblower
i. Truth in Lending
j. Unfair and Deceptive Practices

6. Contract

a. Admiralty Contract
b. Contract—General
c. Contributions
d. Creditors Suits for Nonpayment
e. Express Warranty
f. Good Faith and Fair Dealing
g. Implied Warranty
h. Insurance
i. Sales and Secured Transactions
j. Warranty—General

7. Enforcement

a. Accounting
b. Attachment

   c. Audit
   d. Civil Forfeiture
   e. Constructive Trust
   f. Enforcement of Judgment
   g. Foreclosure

8. Equitable Contract

   a. Account Stated
   b. Equitable Estoppel
   c. Equitable Relief
   d. Promissory Estoppel
   e. Quasi-Contract

9. Fraud

   a. Common-Law Fraud
   b. Deceit
   c. Federal FCA
   d. Federal Miscellaneous
   e. Fraud—General
   f. Fraudulent Concealment
   g. Fraudulent Conveyance
   h. Fraudulent Inducement
   i. Misrepresentation

10. Intellectual Property

   a. Copyright
   b. Cyber Piracy/Squatting
   c. Dilution
   d. Patent
   e. Trade Secret
   f. Trademark

11. Labor

   a. Collective Bargaining Agreement
   b. ERISA
   c. FEHA
   d. FELA
   e. FISA
   f. FMLA
   g. Labor—General
   h. LMRA
   i. Overtime/Minimum Wage

12. Process Causes
    a. Appeal
    b. Discovery Related
    c. Judicial Review
    d. Legal Standards
13. Property
    a. Abandonment
    b. Condemnation
    c. Eminent Domain
    d. Eviction
    e. Foreclosure
    f. Liens
    g. Nuisance
    h. Quiet Title
    i. Replevin
    j. RESPA
    k. Restrictive Covenant
    l. Trespass
14. Racketeering—Criminal Activities
    a. Common Law Conspiracy
    b. RICO
15. Regulatory
    a. Administrative Procedure Act
    b. Attorney
    c. Bankruptcy
    d. CERCLA
    e. Communications
    f. Federal
    g. FOIA
    h. General Health
    i. HAZMAT
    j. Immigration
    k. Transportation
    l. Unauthorized Cable Service
16. Securities
    a. Investment Advisers Act
    b. Investment Company Act
    c. Securities Exchange Act
    d. State Securities

17. Tax
    a. Recovery of Taxes Paid
18. Tort
    a. Conversion
    b. Defamation
    c. Detinue
    d. Failure to Warn
    e. Federal Tort
    f. Intentional
    g. Loss of Consortium
    h. Maritime
    i. Medical Malpractice
    j. Negligence
    k. Outrage
    l. Palming Off
    m. Premises Liability
    n. Privacy
    o. Products Liability
    p. Strict Liability
    q. Tortious Breach of Contract
    r. Tortious Interference
    s. Wantonness
    t. Wrongful Death

Appendix B: Technical Clustering Analysis

A. Spectral Clustering Algorithm

Graph $G = (V,E)$ is specified by its vertex set, V, and edge set E. In our problem, vertices represent cases, while edges represent connections among them. Each edge e is undirected and weighted where weight w corresponds to the similarity between cases represented by the nodes connected by e. Weight w ranges between 0 (dissimilar cases) and 1 (identical cases). The adjacency matrix associated with this weighted undirected graph used in the spectral clustering algorithm is defined as:

$$A_{ij} = \begin{cases} w_{ij} & if\ i \neq j\ and\ (i,\ j) \in E \\ 0 & otherwise \end{cases}$$

The objective of the spectral clustering algorithm is to cut a weighted undirected graph into k clusters (k is predefined) such that edges within each partition have large weight while edges between nodes in different partitions have low weight. A solution for this

multicut problem defined in Meila (2001) and proposed in Ng et al. (2002) will be deployed in our experiments. The method is fairly simple and easy to implement through the following steps (Ng et al. 2002):

- Define a set of cases (vertices) $V = v_1, \ldots, v_n$ and specify the number of clusters k
- Define the similarity measure between cases and create affinity matrix A
- Make diagonal matrix D whose (i,i) element is the sum of A's i-th row
- Construct matrix $L = D^{1/2} AD^{1/2}$
- Find $x_1, x_2, \ldots, x_k$ the k largest eigenvectors of L and create matrix $X = [x_1 x_2 \ldots x_k]$ where x_i is an i-th column in X
- Find a matrix Y from X such that $Y_{ij} = \frac{x_{ij}}{\sqrt{\sum_j x_{ij}^2}}$
- Treating each row of Y as a point, cluster all rows into k clusters via simple clustering algorithm K-means (Tan et al. 2005)
- Assign case v_i to cluster j if and only if row i of the matrix Y is assigned to cluster j

B. Similarity Between Cases

Let $V = \{v_1, v_2, \ldots, v_N\}$ be a data set of N cases to be clustered. A case i in the data set can be represented as a 19-dimensional vector $v_i = \{causeofaction_1, \ldots, causeofaction_{19}\}$, where $causeofaction_k$ (k = 1, . . .,19) denotes a count of how many times the cause of action k appears in the case i. This vector contains many zero-valued elements and several elements that are different from zero. The similarity measure between two case vectors ignores 0-0 matches to prevent a large number of cases being considered similar due to *not* containing many of the same causes of action. Since vector elements in our data set can be greater than 1, we will use the extended Jaccard coefficient (*EJ*) (Tan et al. 2005) as the similarity measure between cases. If v_i and v_j are two cases, then similarity between them is calculated as:

$$w_{ij} = EJ(v_i, v_j) = \frac{\sum_{k=1}^{19} v_{ik} v_{jk}}{\sum_{k=1}^{19} v_{ik}^2 + \sum_{k=1}^{19} v_{jk}^2 - \sum_{k=1}^{19} v_{ik} v_{jk}}$$

In the graph representation, the edge between v_i and v_j has weight w_{ij}. Calculated weights are used to construct affinity matrix A.

C. Number of Clusters

Our objective is to find stable clusters that capture the inherent structure in the data set. An effective way of discovering stable clusters based on subsamples is described in Ben-Hur et al. (2002). We determine that cluster partitions are stable when we find similar partitions when we run the clustering algorithm with different subsamples obtained by random sampling without replacement of the original data set. We calculate the similarity between partitions obtained on different subsamples V_1 and V_2 as follows:

- Find $V_{INTERSECT} = intersect(V_1, V_2)$
- Construct squared matrices $C^{(1)}$ and $C^{(2)}$ corresponding to the partitions of V_1 and V_2, respectively, such that for a pair of cases (v_i, v_j) from V

$$C_{ij} = \begin{cases} 1 & \text{if } i \neq j, (v_i, v_j) \text{ are from } V_{INTERSECT} \text{ and belong to the same cluster} \\ 0 & \text{otherwise} \end{cases}$$

If two partitions are similar, then cases that belong to the same cluster obtained on set V_1 would most likely belong to the same cluster obtained on set V_2. In other words, there will be 1s on the same places in both matrices $C^{(1)}$ and $C^{(2)}$ corresponding to the partitions of sets V_1 and V_s.

- Denote

N_{01}—the number of 0-1 matching pairs from $C^{(1)}$ and $C^{(2)}$
N_{10}—the number of 1-0 matching pairs from $C^{(1)}$ and $C^{(2)}$
N_{11}—the number of 1-1 matching pairs from $C^{(1)}$ and $C^{(2)}$

We calculate similarity between partitions made on V_1 and V_s using:

$$Sim(V_1, V_2) = \frac{N_{11}}{N_{01} + N_{10} + N_{11}}$$

To find k and to reduce search space, we will explore stability for $4 \leq k \leq 12$. We use the following algorithm (Ben-Hur et al. 2002):

- Sampling rate $f = 0.9$, $number\_of\_iterations = 100$
- for $k = 4{:}12$
- $i = 1{:}number\_of\_iterations$
- $V_1 = subsample(V,f)$
- $V_2 = subsample(V,f)$
- $L_1 = cluster(V_1)$
- $L_2 = cluster(V_2)$
- $S_{ik} = Sim(L_1,L_2)$
- end for
- end for

where sampling rate f determines the fraction of the original data set used in subsampled sets.