U.S. Patent and Trademark Office
OFFICE OF THE CHIEF ECONOMIST
Economic Working Paper Series

Patent Litigation Data from US District Court Electronic Records (1963-2015)

Alan C. Marco, *Chief Economist*
Asrat Tesfayesus, *Economist*
Andrew A. Toole, *Deputy Chief Economist*

USPTO Economic Working Paper No. 2017-06
March 2017

The authors thank Ted Sichelman for valuable comments. All remaining mistakes are our own.

For more information about the USPTO's Office of the Chief Economist, visit www.uspto.gov/economics.

Patent Litigation Data from US District Court Electronic Records (1963-2015)

Alan C. Marco

Asrat Tesfayesus

Andrew A. Toole

U.S. Patent & Trademark Office

March 2017

Abstract

Economists, legal scholars and policy makers are concerned about the impact of patent litigation on the rate and direction of US innovation and on the functioning of the US intellectual property system. At this time, however, there is no reliable, comprehensive, free and publicly accessible source of patent litigation data. As a first step towards overcoming this limitation, the Office of the Chief Economist (OCE) at the United States Patent and Trademark Office, in collaboration with the National Technical Information Service, undertook a patent litigation data pilot. This paper describes two patent litigation databases from which OCE generated comprehensive patent litigation datasets as a result of the pilot and is releasing for public use. First, we obtained the docket reports on the universe of patent litigation cases in PACER and RECAP and created a dataset for the period 1963-2015. Second, we captured the metadata for these cases, which includes information on the case identifier, parties involved, filing date, and district court location. We hope that free access to comprehensive data will help advance research on the evolving patent litigation landscape and its economic impact.

Keywords: Patents, Litigation, Intellectual Property
JEL Classification: O3, K4

Table of Contents

I. Introduction

Economists, legal scholars and policy makers are concerned about the impact of patent litigation on the rate and direction of US innovation and on the functioning of the US intellectual property (IP) system. These concerns are primarily driven by the number of new cases and the costs of litigation. For instance, the number of cases filed in US district courts more than doubled from 2009 through 2012, reaching a peak in 2013 at over 6,200 new filings. And, while the median costs for patent infringement suits have not grown since 2013, these costs remain very high, ranging from $600,000 to $5 million per suit depending on the value at risk (AIPLA 2015).

To explore the potential impacts of patent litigation, academic researchers are studying how firm-level outcomes such as patenting, access to capital, and research and development (R&D) expenditures are responding to patent litigation, especially for startups and small firms. Some evidence suggests that litigation costs are affecting the direction of innovation by shaping the nature of patented technologies among new biotechnology firms (Lerner 1995). Another study found the likelihood of being sued increases as firms perform more R&D (Bessen and Meurer 2013). And still other studies explore the impacts of non-practicing entities (sometimes referred to pejoratively as "patent trolls"), legal procedural and statutory changes, and issues related to patent quality (see Section II).

Among policy makers, concerns about patent litigation fuel the impression that the IP legal system is vulnerable to abuse and exploitation, particularly victimizing small firms and end-users. In 2011, Congress passed the American Invents Act (AIA), the most comprehensive and transformative legal reforms since the Patent Act of 1952. Some of the AIA reforms focused on patent quality while others focused on deterring abusive patent holder behaviors. For example, it established the Patent Trial and Appeal Board to allow for a speedier and less expensive patent validity challenges. Since AIA took effect in 2012, over 5000 *inter partes review* petitions have been filed at the Patent Trial and Appeal Board. Congress has also taken more recent initiatives to curb abusive patent litigation through changes in pleading standards, more specific venue rules, restrictions on end-user targeting, limits on discovery, and fee shifting.[1]

Despite these legislative initiatives, a better understanding of the impacts of patent litigation is sorely needed. Careful and systematic research on pressing questions will help. Questions such as: Do we observe more "abusive" litigation? Are plaintiffs increasingly engaging in strategic forum shopping behaviors that are costly to the public? Can patent litigation filing trends be

[1] These initiatives, mostly recently under consideration in Congress, include the Innovation Act (H.R. 9), the PATENT Act (S. 1137), and the VENUE Act (S. 2733).

explained by the changing procedural and substantive landscapes of patent law? What are the effects of patent litigation on firms' innovative activities, their financial wellbeing, and their contribution to the economy?

Research on these questions requires access to detailed information about plaintiffs and defendants, patents-in-suit, and information on the outcomes of litigated cases. At this time, however, there is no reliable, comprehensive, free and publicly accessible source of patent litigation data. Researchers must rely on proprietary data that are typically costly and often do not provide formats that make it easy to integrate other data elements and sources.

As a first step towards overcoming this limitation, the Office of the Chief Economist (OCE) at the United States Patent and Trademark Office, in collaboration with the National Technical Information Service, undertook a patent litigation data pilot. The project examined the quality, accessibility, and coverage of public information on patent litigation and gauged the potential for collecting district court patent litigation data. The pilot used the Public Access to Court Electronics Records (PACER) database, which provides public access to all cases litigated in US district courts, and RECAP, which is a subset of PACER that serves as a free repository of PACER data. As discussed in Section III, these sources provide comprehensive and detailed information on patent litigation across the 94 district courts of the United States.

The pilot concluded with the production of two sets of patent litigation data that OCE is releasing for public use. First, we obtained the docket reports on the universe of patent litigation cases in PACER and RECAP and created a dataset for the period 1963-2015.[2] Second, we captured the metadata for these cases, which includes information on the case identifier, parties involved, filing date, and district court location.

The docket reports have three sections that contain (1) a header box that provides basic information on the case including the case (or docket) number, the parties involved, the court and judges assigned, the cause of action and jurisdictional basis, and relevant dates in the litigation process; (2) a list of the parties involved as well as information on the representing attorneys; and (3) a list of the individual documents added to the docket along with the dates of filing, the number of documents submitted, number of attachments included, a long and short description of the documents' contents, and the date when the documents were uploaded to PACER.

The docket report data that we provide come in four separate files (i.e. cases, names, attorneys, documents). For each of the 74,623 unique cases, the cases file provides information that is

[2] Note that the PACER patent litigation data coverage pre-1999 is incomplete and decreases substantially going back in time from 1998. Furthermore, as explained in more detail later, the patent litigation data coverage in PACER may be missing some cases even for the later years as they may have been misclassified in other areas of Nature of Suit.

captured from the header box of the docket report. Information reported in this file includes: the case numbers; pacer IDs (when available); the party names; the name of the district court; the assigned judge; the magistrate judge it was referred to (when applicable); the cause of action; the jurisdictional basis of the case; the filing, closing, and last document filing dates; whether either or both of the parties requested a jury trial; any lead or member cases that are associated with the case; related cases that were referenced in the docket report; any record of settlement, and any monetary demand submitted.

The names and attorneys files are both derived from the second part of the docket report where the parties and their representing attorneys are listed. While the former provides the party type information along with their names, the latter provides the names, location, and contact information of the representing attorneys.

The documents file provides records of each document that was added to the docket report. It provides information on when each document was filed, a short and long content description, the document's count in the sequence of recorded documents for a given docket, the number of additional documents attached, and the date of upload to PACER.

The case-level metadata we collected from PACER are contained in the pacer_cases file. While these data are also sourced from PACER and contain a subset of the variables in the docket reports, we provide it as a separate file for two reasons. First, this dataset supplements data elements that may be missing in the docket reports for some cases. For instance, this is helpful for the date of filing, which is frequently missing in the docket reports data. Second, this dataset allows us to assess the reliability of the PACER content by evaluating the internal consistency of its datasets.

The remainder of this paper is organized as follows. In Section II, we provide a brief literature review of recent academic research that uses proprietary data to study a variety of pressing questions about patent litigation. In Section III, we describe the data sources used in the pilot project to obtain the docket reports and the PACER case-level metadata. In Section IV, we describe all of the steps taken to obtain, process, and compile the contents of the final data products we released. We also describe each of the variables contained in the data. In Section V, we provide summary statistics for the categorical variables, merge and compare the content of the two alternative PACER datasets, and give illustrative examples of how the data can be used. We conclude in Section VI.

II. Literature Review

Researchers studying patent litigation use datasets drawn from a variety of sources. Before describing the literature, it is helpful to have an overview of some important data sources used. The six main sources are Lex Machina, RPX Corporation, LexisNexis, Westlaw, Docket Navigator, and Bloomberg Law.[3] For each of these, we provide some information on the origins, related datasets, and content.

1. Lex Machina: This is a legal analytics firm that provides litigation data on patents, trademarks, copyrights, and antitrust cases. It was originally formed in 2006 at Stanford University as a public interest project under the name of IP Litigation Clearinghouse (IPLC). Although it leverages a number of sources to provide value-added and optimized IP data, Lex Machina uses PACER as its primary patent litigation data source.

2. RPX (Rational Patent EXchange) Corporation: Founded in 2008, RPX Corporation is a patent information and consulting firm that provides a variety of services including patent risk solutions, patent intelligence, and insurance services. Similar to Lex Machina, RPX relies on PACER as its primary source of patent litigation data. In 2014, RPX acquired PatentFreedom (also founded in 2008). PatentFreedom is an online platform that provides information to its paying members on the identity and assertion practices of non-practicing entities.

3. LexisNexis: Established in 1970 by Mead Data Central, LexisNexis, among other things, offers access to billions of searchable documents and records. In addition to its extensive content on US statutes and laws, the Lexis database provides published case opinions dating back to the 18th century and unpublished opinions since the 1980s. Furthermore, it covers a number of other jurisdictions, content from academic publications, and news articles. It was acquired by Reed Elsevier in 1994 and has made purchases of its own; including when it acquired Quicklaw, a Canadian legal research database company, in 2002.

4. Westlaw: This is another leading legal database company established in 1992 by West publishing and acquired by Thomson Corporation in 1996. Similar to LexisNexis, it provides online access to cases, statutes, academic writings, and news publications. Westlaw also contains Derwent LitAlert which provides information on patent and trademark infringement lawsuits in all 94 district courts since 1973.[4]

[3] Note that while one of the papers discussed uses PricewaterhouseCoopers database, we did not discuss it separately as it is derived from the records of the Westlaw database.

[4] In addition to PACER data, both LexisNexis and Westlaw source litigation data from paper documents from various courts. Thus, their coverage is more comprehensive than PACER, especially for the pre-1999 period.

5. Docket Navigator: Established with the recognition of the importance of analyzing litigation data to alleviate litigation uncertainty, Docket Navigator began providing its services in 2008 by distributing comprehensive patent litigation information via email. Its research database was then launched at the end of 2010. This database contains patent litigation data from all district courts, the Patent Trial and Appeal Board, and the International Trade Commission.

6. Bloomberg Law: This is another subscription-based online legal research tool that recently started its services and is frequently used by researchers. First introduced as a pilot in 2009, its online platform was launched in 2010. The services it provides include litigation and dockets from state, federal, and select international courts, legal and financial analytics, as well as News and Law Reports.[5]

Turning to the literature, we provide a selective review of recent scholarly work on patent litigation in order to discuss some of the main issues, present selected findings, and highlight the reliance of the literature on proprietary data. As will be clear, most papers use PACER data indirectly through one of the databases described above. The PACER-based datasets accompanying this paper are free and publicly accessible, which should facilitate academic and policy research on important issues related to patent litigation.

We organized the literature into three broad categories. In the first category, the literature examined patent litigation trends, litigation outcomes and their connection to possible determinants of patent quality. The second set of papers looked at changes in assertion practices of non-practicing entities and the influence these changes had on the patent litigation landscape. The third category of studies examined the effect of patent litigation and the broader patent system on the economy. Table 1 provides a summary of the findings and data sources for the literature reviewed.

Determinants of patent quality:

The quality of a patent can be defined from a number of perspectives. One perspective on patent quality has to do with the legal boundaries of the intellectual property rights in a patent. Patents with clear and sharp legal boundaries are considered higher quality while patents with unclear and fuzzy legal boundaries are considered lower quality. A number of factors contribute to the challenge of delineating the legal boundaries of a patent. For instance, patent claims may be interpreted in different ways based on the language and syntax used. Furthermore, changes in the law introduce uncertainty as to what is patentable subject matter. Other perspectives on patent quality focus on the novelty of the invention, the nature of the invention disclosure, the

[5] For more information on patent litigation sources, refer to Schwartz et al. (2016).

actions of the USPTO, or the patent's economic value. Using patent litigation data, several papers in the literature explored potential determinants of patent quality, sometimes with the goal of predicting a patent's value or its probability of future litigation.

For patent cases filed from 2008-09, Allison et al. (2014) took a comprehensive look at litigation outcomes. Using the Lex Machina database, they related validity, infringement, and unenforceability to the characteristics of the parties, the patents, as well as the court location. They found the rate of patent litigation outcomes was stable across the years (e.g. the percentage of cases with a validity finding). However, among successful validity challenges, the statutory grounds and the characteristics of the plaintiffs differed significantly over time. In a similar vein, Miller (2013) considered how repeat litigation of a patent relates to its quality and, indirectly, to its value. To do so, he used data from the Stanford Intellectual Property Litigation Clearinghouse, the predecessor to the Lex Machina database. He found that repeat patent litigation plaintiffs are more successful in obtaining winning judgements. However, this result did not hold for software patents; which is in line with the theory that intellectual property boundaries in the software space are more uncertain.

In a 2013 report, the Government Accountability Office (GAO) used patent litigation data sourced from Lex Machina to learn about the characteristics of recent patent litigation cases and the factors that may drive litigation trends. The report found lawsuits increased by about a third in 2010-2011 compared to the previous decade and that non-practicing entities account for a fifth of the lawsuits. In addition, based on a representative sample of 500 lawsuits in 2007-2011, the report found a 129% increase in the number of overall defendants; 89% of which were attributed to an increase in defendants for software-related patents. Chien (2011) considers how litigated patents differ from unlitigated patents and looks at both pre- and post-issuance characteristics that may predict as to whether a patent will be litigated. Using LexisNexis as the primary source of litigation data, Chien shows that litigated and unlitigated patents have very different post-issuance characteristics. She finds that litigated patents were significantly more likely to be transferred, to experience change in owner size, to undergo an ex parte reexamination, to be renewed, to be collateralized, and to be cited compared to unlitigated patents. Similarly, in 2015, the USPTO conducted a study to explore ways to predict the likelihood of patent infringement litigation based on patent prosecution histories and other patent characteristics (Marco et al. 2015). Along with district court litigation data sourced from Lex Machina, the USPTO included *inter partes reviews* from the Patent Trial and Appeals Board. The study found that entity size of the assignee, foreign origin, and family size of the patent are strongly related to the likelihood of litigation proceedings and that the number and length of independent claims, the GS-level of the examiner, the number of Information Disclosure Statement (IDS), the number of examiner interviews, and first action allowance also seemed to play a role.

Assertion practices of non-practicing entities:

Another strand of the literature focuses on assertion activities of non-practicing entities (NPEs) and their role in changing patent litigation trends. In light of increasing use of litigation by NPEs, many studies explored the factors that might explain NPE litigation behaviors. Also, both academics and policy-makers have focused attention on the costs that litigious NPEs may be imposing on practicing entities and end-users, particularly related to their innovative activities and commercialization efforts.

Jeruss et al. (2012) used Lex Machina data to quantify the increase in the rate of patent litigation filings by NPEs. They found that "patent monetization entities" have increased their lawsuit filing rate from 22% to almost 40% in a span of 5 years ending in 2011. Their study also showed these entities accounted for 40% of cases filed in 2011 alone. Bessen et al. (2014) used RPX data to look at the direct costs that NPEs imposed on the defendants in patent litigation. They conducted a survey of about 250 firms that were involved in a patent litigation as defendants against NPEs. They estimated the direct cost to be $29 billion in 2011, which disproportionately burdened small and medium-sized companies, and imposed more costs on these defendants compared to the money NPEs transferred to inventors.

More recently, Cohen et al. (2015) studied the strategic behavior of NPEs related to the type of firms they target in asserting patents through litigation. They obtained litigation data covering more than 4000 NPEs from PatentFreedom, which was acquired by RPX Corporation in June 2014. They found NPEs' target firms already involved in other lawsuits and possess large cash stocks. The defendant firms were targeted irrespective of whether their patent portfolios were related to that of the NPEs. Furthermore, the authors found that NPE litigation had a negative effect on innovative activity.

The Federal Trade Commission (FTC) recently released a report based on a case study of 22 patent assertion entities (PAEs), 2,500 of their affiliates, and other related entities. Using its subpoena power under Section 6(b) of the Federal Trade Commission Act, the FTC collected confidential data from these PAEs to study how firm characteristics relate to their patent assertion practices. This report defined two types of PAE business models: Portfolio PAEs and Litigation PAEs. They found Portfolio PAEs generated 80% of the total revenue reported to FTC by all study firms, but accounted for 9% of the licenses in the study. Litigation PAEs accounted for 91% of the reported licenses and 96% of the litigated cases, which they tended to settle quickly, generating relatively small licensing royalties. The initial selection of the PAEs into the study used RPX data.

A number of papers have also conducted studies comparing NPEs and practicing entities. In a case study of the ten most litigious NPEs since 2003, Risch (2012) studied litigation practices, patent portfolios, and the original assignees of the patents. He found that NPEs are not a new phenomenon and do not play the role of intermediaries. He also found that their litigated patents were acquired from productive companies and were similar to other litigated patents with respect to the technology category and quality level. In subsequent work, Risch (2015) looked at patents-in-suit that were decided on the merits and were also assigned to the same set of litigious NPEs he studied previously. He analyzed the litigation characteristics of these NPE plaintiffs relative to comparable practicing entity plaintiffs. Among his results, he found NPEs had more complex cases, shorter litigation duration, twice as many invalidations, and more non-infringement findings. However, he also found that case-specific factors predicted likelihood of invalidation better than patent quality indicators, such as citations. In both instances, the author obtained the list of litigious NPEs from PatentFreedom. In addition, the data were supplemented with information from sources such as the Stanford IP Litigation Clearinghouse (later Lex Machina), PACER dockets, and Lexis as well as Westlaw databases.

Mazzeo et al. (2013) analyzed 1,750 patent infringement cases litigated in the period 1993-2011 to evaluate the success rate and award differences between NPEs and practicing entities (PE), and across different types of NPEs. While they observed differences within NPE types on their success rates as well as their assertion practices, they found little difference in success rates between NPE and PE litigants. Their data was sourced from the PricewaterhouseCoopers database, which contained all of the decided patent cases for that period as reported in Westlaw. Love (2012) also compared NPE and practicing entities' litigation behavior using the Westlaw's Derwent LitAlert database. He found that practicing entities enforce their patents early in the life of the patent while NPEs begin litigation close to the expiration of their patent.

Focusing on high-tech patents, Chien (2009) evaluated NPE litigation practices by type of technology, industry, and firm or entity characteristics using the Stanford Intellectual Property Clearinghouse's data on patent litigation filed in the period 2000-2008 in US district courts. Chien also looked at the share of patent suits brought by NPEs. While NPEs brought a minority of patent suits in the eight year period, she found the proportion of cases brought by NPEs went up from 10% in 2000-2001 to 20% in 2006 through March 2008. During that same period, defendants facing NPE suits experienced a larger percentage point increase, from 22% to 36%. In a more recent study, Chien (2014) used RPX's patent litigation database covering the period 2006-2012 to analyze the impact that patent assertion entities have on startups. She found that 55% of the PAE litigation targets (i.e. unique defendants) had annual revenue of $10 million or less. Survey results from this study also indicated that smaller firms suffer more from "significant operational impact" as PAE targets.

Broader economic impacts on the patent system:

A third set of studies examined both the micro and macro-level impacts of patent litigation. Henry (2013) used an event study methodology on stock market reactions to patent litigation outcomes to quantify abnormal returns. These were then used as an input for a patent value measure of publicly traded firms. The study concluded that litigation outcomes for which a patent was found "invalid" cost firms 0.85% of their value while decisions of "Valid & Infringed" only increase firm value by 0.7%. The author also found that an "Invalid" decision subsequent to the establishment of US Court of Appeals for the Federal Circuit (CAFC) added to the average decline of firm value by 0.7%. The data covered the period 1953-2002 and used both the United States Patent Quarterly and Westlaw as sources.[6]

Similarly, Bessen and Meurer (2012) conducted stock market event studies based on patent litigation filings during the period 1984-1999 to estimate the costs of patent litigation for U.S. public firms. Their estimates suggest that by the late 1990s the expected cost to alleged infringers was over $16 billion per year. The primary data source for this study was Westlaw's Derwent LitAlert database. In another paper, Galasso et al. (2015) exploited the random allocation of appeals court judges to patent litigation cases to identify the impact of litigation outcomes on follow-on innovation. They found that patent invalidation led to a 50% increase in citations of the focal patent. However, they also observed significant variations across industries, firm size, and bargaining environment. Their paper used full text data from Federal Circuit decisions on patent litigation from the LexisNexis QuickLaw dataset.

All of the contributions discussed rely mainly on proprietary data. Recognizing that accessing and using proprietary data is costly, Cotropia et al. (2014) took a first step toward providing freely accessible patent litigation data. For 7,705 patent litigation lawsuits for the years 2010 and 2012, the authors compiled and released the corresponding patents-in-suit, the associated technology and a classification of the entity holding the patents. Our data release builds on this by collecting and providing the universe of patent litigation suits as recorded in PACER.

Table 1: Literature Summary

Author	Publication Year	Study & Findings	Data Source
Allison et al.	2014	The authors evaluated patent litigation outcomes for all cases filed in 2008 and 2009 and found consistency across time in terms of patentee success and invalidity holdings at 25% and 45%, respectively. However, they found that courts identified as	Lex Machina

[6] The data were originally collected for a prior article, Henry et al. (2006).

		being the most active changed significantly along with NPE participation and grounds of most successful validity challenges.	
Bessen et al.	2014	In an effort to study whether NPEs deter innovative activity, the authors quantified the direct cost they impose on defendants in their sample, most of which are small and medium-sized firms. They estimated that the direct cost amounted to $29 billion in 2011.	RPX
Bessen et al.	2012	Conducted stock market event studies for U.S. public firms involved in patent litigations filed 1984-1999. They estimated that by the late 1990s, the expected cost to alleged infringer was over $16 billion per year.	Westlaw
Chien	2014	Specifically focused on the impact of PAEs on startups and found that, for patent litigations 2006-2012, 55% of the PAE litigation targets (i.e. unique defendants) had annual revenue of $10 million or less. Survey results also indicated that smaller firms suffer more from "significant operational impact" as PAE targets.	RPX
Chien	2011	Finds that litigated patents were significantly more likely to be transferred, to experience change in owner size, to undergo an ex parte reexamination, to be renewed, to be collateralized, and to be cited compared to unlitigated patents.	LexisNexis
Chien	2009	Provided statistics on NPE activity in patent litigation both as plaintiff and, in declaratory judgment cases, as defendant. Found a 14 percentage point increase in the proportion of defendants sued by NPEs in an eight-year period with significant variations across industries. Noted the prevalence of large to large firm patent litigation, putting the effectiveness of defensive patenting into question.	IPLC
Cohen et al.	2015	Found that, in their patent litigation practice, NPEs targeted firms with large cash stock and those involved in other lawsuits. They sued regardless of whether their patent portfolio was closely related to the defendant firm. The authors also found that NPE litigation had negative effect on innovative activity.	PatentFreedom
Cotropia et al.	2014	Covered a combined 7,705 patent litigation lawsuits in 2010-2012 to provide freely accessible patent litigation data. This carefully processed data contains the patents-in-suit corresponding to the case, the technology that they belong to, and a classification of entity type of the assignee.	Bloomberg Law[7]
Galasso et al	2015	Studied the effect of patents on follow-on innovation by looking at court invalidation through cases that are randomly allocated to judges. They found 50% increase in citations of the focal patent due to patent invalidation. However, they observed	LexisNexis

[7] The Bloomberg Law's Federal Docket Database that the authors use is effectively the same as what is contained in PACER's raw data. The primary data contribution of the authors is in the value-add that includes improving the accuracy of patents-in-suit covered, providing technology grouping, and, most importantly, generating patent holder classification.

		significant variations across industries, firm size, and bargaining environment as to the effect of invalidation.	
FTC	2016	The report distinguished between Portfolio PAEs and Litigation PAEs. The former yielded 80% of the revenue while only accounting for 9% of the reported licenses. Litigation PAEs, on the other hand, filed 96% of the litigated cases which they settled quickly with licenses typically generating relatively small royalties.	RPX
GAO	2013	Compared changes in litigation trends during the periods 2000-2010 and 2010-2011, the GAO found an increase in lawsuits by about a third and that NPEs accounted for a fifth of the lawsuits. In addition, based on a representative sample of 500 cases filed in 2007-2011, GAO found a 129% increase in the number of overall defendants; 89% of which is attributed to defendant increase in software-related patents.	Lex Machina
Henry	2013	Conducted a stock market event study and found that a litigation outcome that found a patent "Invalid" costs firms 0.85% of their value while a decision of "Valid & Infringed" only gained them 0.7%. Furthermore, the author found that an "Invalid" decision subsequent to the establishment of CAFC added to the average decline of firm value by 0.7%.	Westlaw
Jeruss et al.	2012	Found that "patent monetization entities" have increased their lawsuit filing rate by about 18% percentage points in a span of 5 years, accounting for 40% of cases filed in the most recent year observed (i.e. 2011).	Lex Machina
Love	2012	Found that while practicing entities enforce their patents early in the life of the patent, NPEs began litigation close to the expiration of their patent.	Westlaw
Marco et al.	2015	Related patent examination, application, and applicant characteristics to subsequent litigation or IPR proceedings. They found that entity size, foreign origin, and family size are strongly related to litigation proceedings and that independent claims, GS-level of the examiner, the number of IDS, the number of examiner interviews, and first action allowance seem to also play a role.	Lex Machina
Mazzeo et al.	2013	By looking at patent litigation outcomes for the period 1995-2001, the authors found little difference in success rate between NPE and PE litigants. However, they found differences within NPE types in their success rate as well as assertion practices.	PwC, Westlaw
Miller	2013	Found that repeat patent litigating plaintiffs are more successful in obtaining winning judgements. However, this result does not hold for software patents.	IPLC
Risch	2015	Compared characteristics of litigated cases of highly litigious NPEs and PEs. Found that, among other things, the former had more complex cases, shorter litigation duration, twice as many invalidations, and more non-infringement findings. However, case-specific factors better predicted likelihood of invalidation	PatentFreedom, IPLC, Lexis, Weslaw

		as opposed to factors used as patent quality indicators, such as citation.	
Risch	2012	Conducted case studies of the ten most litigious NPEs by looking at their litigation practices, their patent portfolio, and the original assignees of these patents. Found that these NPEs are not new, their litigated patents are similar to other litigated patents and are mostly acquired from productive companies. Did not found evidence that NPEs play the role of intermediaries.	PatentFreedom, IPLC, Lexis, Weslaw

III. Data Sources

A. Data Sources and Content Summary

During a six months period that concluded in March 2016, the USPTO Office of the Chief Economist conducted a data pilot to explore the availability and costs of obtaining patent litigation data. All of the data provided in this release come from the Public Access to Court Electronic Records (PACER) database, which contains the most comprehensive court records of cases litigated in the 94 district courts of the United States. The PACER records were also accessed using RECAP, which serves as a repository for case records that have already been downloaded by users from PACER (RECAP is a computer plug-in available to PACER users). The data collected provide both metadata that summarizes the overall case information and document-level data providing information on the nature of court documents filed. To access the data in PACER, the user must pay a fee, but information contained in RECAP is free. We used the RECAP information when it contained up-to-date information to avoid unnecessary fees. The final datasets contain patent litigation cases filed in PACER during the period 1963-2015. The mechanics of accessing and recording the content from these sources is detailed below.

B. PACER

Initiated through a pilot program in 1989 by the Federal Judicial Center, PACER is a service that provides electronic public access to case and docket information online from federal district, appellate and bankruptcy courts. While all 94 district courts have implemented the PACER system, complete historical data on all cases filed in these courts are not available. PACER underwent a staggered implementation and not all of the older cases have been added.

Nevertheless, PACER appears to be the most comprehensive and authoritative litigation data source. The public can obtain any document available on PACER for a fee of $0.10 per page and up to $3.00 per document.[8]

PACER has records on federal civil litigation cases covering a variety of subject matters that are categorized by *Nature of Suit* codes.[9] In order to identify the subject matter of the case, federal courts require the plaintiff to designate a single *Nature of Suit* per case at the time of filing. PACER then classifies the case based on the plaintiff's designation. The limitations of this system are that the plaintiff can only select one type of suit, some descriptions for the *Nature of Suit* are ambiguous, and the recordation accuracy is subject to the plaintiff's submission error. However, a preliminary view at patent litigation data extracted from PACER using the *Nature of Suit* for patents (i.e. code 830) shows a high degree of accuracy in identifying cases that are related to a patent suit. For example, a check on the cause of action by looking at the entry in the variable case_cause for any random set of cases predominantly indicates that a patent is at issue. Furthermore, Marco et al. (2015), show that Nature of Suit 830 captures about 98-99% of all patent cases in PACER, while also indicating that, starting in 1999, PACER appears to contain a nearly complete set of patent cases when compared to the Administrative Office of the Courts' raw data.

PACER contains a broad spectrum of information with varying degrees of granularity. Limiting the *Nature of Suit* to patents (code 830), we captured: (1) all the case-level patent litigation data recorded in PACER, which provides party names, court codes, case identifiers, and the filing and closing dates; and (2) all docket reports, containing basic information on the cases as well as a list and description of all documents added to the docket. In Section IV, we describe in detail the contents of the docket reports data, which are provided in this data release as four separate files containing a common identifier to allow easy record linking.

C. RECAP

RECAP is an independent project designed to serve as a repository for litigation data sourced from PACER. More specifically, PACER users with Firefox or Chrome web browsers that have the RECAP plug-in will automatically download documents to the RECAP repository as they access them in PACER. Once a document is in the repository, RECAP alerts users of its availability for free access. In other words, RECAP provides free access to a subset of PACER data.

[8] We estimate that a single docket report contains an average of around 12 pages. The PACER system and the resources it provides can be accessed at https://www.pacer.gov/.
[9] A complete listing of the Nature of Suit categories can be found here: https://www.pacer.gov/documents/natsuit.pdf.

Two features of RECAP deserve note. First, if an individual document for a case is on RECAP, the entire docket report for that case will also be available on RECAP. Second, the information and documents listed in the RECAP docket report will be as of the date of download from PACER.[10] Any documents added to PACER after that date will not be in RECAP until a subsequent PACER user with the RECAP plug-in downloads the docket report again. The docket reports obtained from RECAP and included in our data release were obtained after confirming RECAP had the most up to date information, as of our download date of March 6, 2016. Otherwise, the docket reports were obtained from PACER. In Section IV, we discuss the procedure we follow to verify if the docket report data in RECAP is current.

IV. Data structure and files

A. Docket Reports

Docket reports contain all material filed by any party (including amicus curiae) or by the court itself.[11] They have three sections: (1) case details, (2) parties, and (3) documents. Below is an example of a typical docket report selected from the District of Northern Alabama.

The case details section provides basic information on the case, such as the case (or docket) number, the parties involved, the court and judges assigned, the cause of action and jurisdictional basis, and relevant dates in the litigation process. In the example below, the case number is 5:02-cv-00444; the case name, which also identifies the parties, is given as "Summit Specialties v. Hunter's View Ltd., et al" and the case is assigned to Senior Judge Inge P Johnson.

The parties section provides more detailed information on the parties involved as well as the attorneys who are representing them. In addition to separately identifying and providing the names and entity type of the parties, it provides the location and contact information of the representing attorneys. In the example, Summit Specialties is identified as the plaintiff and is represented by two attorneys. One of the representing attorneys is Arthur Gardner at Gardner Groff Greenwald & Villanueva PC, which is located in Atlanta, GA.

The third section, i.e. Documents, provides a list of the individual documents added to the docket along with the date when the document was filed, the number of documents submitted,

[10] We estimate that RECAP contains approximately a third of the docket reports in PACER, but these may not be complete records as described.

[11] Amicus curiae, literally meaning "friend of the court", refers to a disinterested or third party that contributes unsolicited information to a court in regards to a case that is before it.

number of attachments included, a long and short description of the document's nature and content, and the date when the document was uploaded to PACER. Note that entries for **Short Description** and **Upload date** are missing for the case in our example. Data elements in PACER are not always complete. Summaries of percentage of missing values for each of the variables in the documents file is provided in the Appendix.

Case details

Court:	Alnd
Docket #:	5:02-cv-00444
Case Name:	Summit Specialties v. Hunter's View Ltd., et al
PACER case #:	86511
Date filed:	2002-02-20
Date terminated:	2003-01-13
Assigned to:	Senior Judge Inge P Johnson
Case Cause:	15:1125 Trademark Infringement (Lanham Act)
Nature of Suit:	830 Patent
Jury Demand:	None
Jurisdiction:	Federal Question

Parties

Represented Party	**Attorney & Contact Info**
Summit Specialties, Inc. Plaintiff	Arthur A Gardner GARDNER GROFF GREENWALD & VILLANUEVA PC 2018 Powers Ferry Road, Suite 800 Atlanta, GA 30339 770-984-2300 Fax: 770-984-0098 Email: agardner@gardnergroff.com *ATTORNEY TO BE NOTICED* Robert H Harris HARRIS CADDELL & SHANKS PC 214 Johnston Street SE PO Box 2688 Decatur, AL 35602-2688 1-256-340-8000 Fax: 1-256-340-8040 fax Email: rharris@harriscaddell.com *ATTORNEY TO BE NOTICED*
Hunter's View Ltd. Defendant	Michael K Alston HUSCH BLACKWELL LLP 736 Georgia Avenue, Suite 300 Chattanooga, TN 37402

	423-266-5500 Fax: 423-266-5499 Email: michael.alston@huschblackwell.com *ATTORNEY TO BE NOTICED* Robert Eugene Muir HUSCH & EPPENBERGER LLC 401 Main Street, Suite 1400 Peoria, IL 61602 1-309-637-4900 *ATTORNEY TO BE NOTICED*
Doug Smith Defendant	Michael K Alston (See above for address) *ATTORNEY TO BE NOTICED* Robert Eugene Muir (See above for address) *ATTORNEY TO BE NOTICED*
Hunter's View Ltd. Counter Claimant	Michael K Alston (See above for address) *ATTORNEY TO BE NOTICED*
Summit Specialties, Inc. Counter Defendant	Arthur A Gardner (See above for address) *ATTORNEY TO BE NOTICED* Robert H Harris (See above for address) *ATTORNEY TO BE NOTICED*

Documents

Date Filed	Document #	Attachment #	Short Description	Long Description	Upload date
2002-02-20	1	0		COMPLAINT filed, amount paid $ 150.00, receipt # 200 174964 (ASL) (Entered: 02/22/2002)	
2002-02-20	2	0		AFFIDAVIT of Bradley Fitzgerald filed cs (ASL) (Entered: 02/22/2002)	
2002-02-20	3	0		MOTION by plaintiff Summit Specialties for attorney Arthur A. Gardner to appear pro hac vice filed (ASL) (Entered: 02/22/2002)	

2002-02-25	4	0		REQUEST of plaintiff for service by certified mail filed (ASL) (Entered: 02/25/2002)	
⋮	⋮	⋮	⋮	⋮	⋮

B. Data Structure and Process

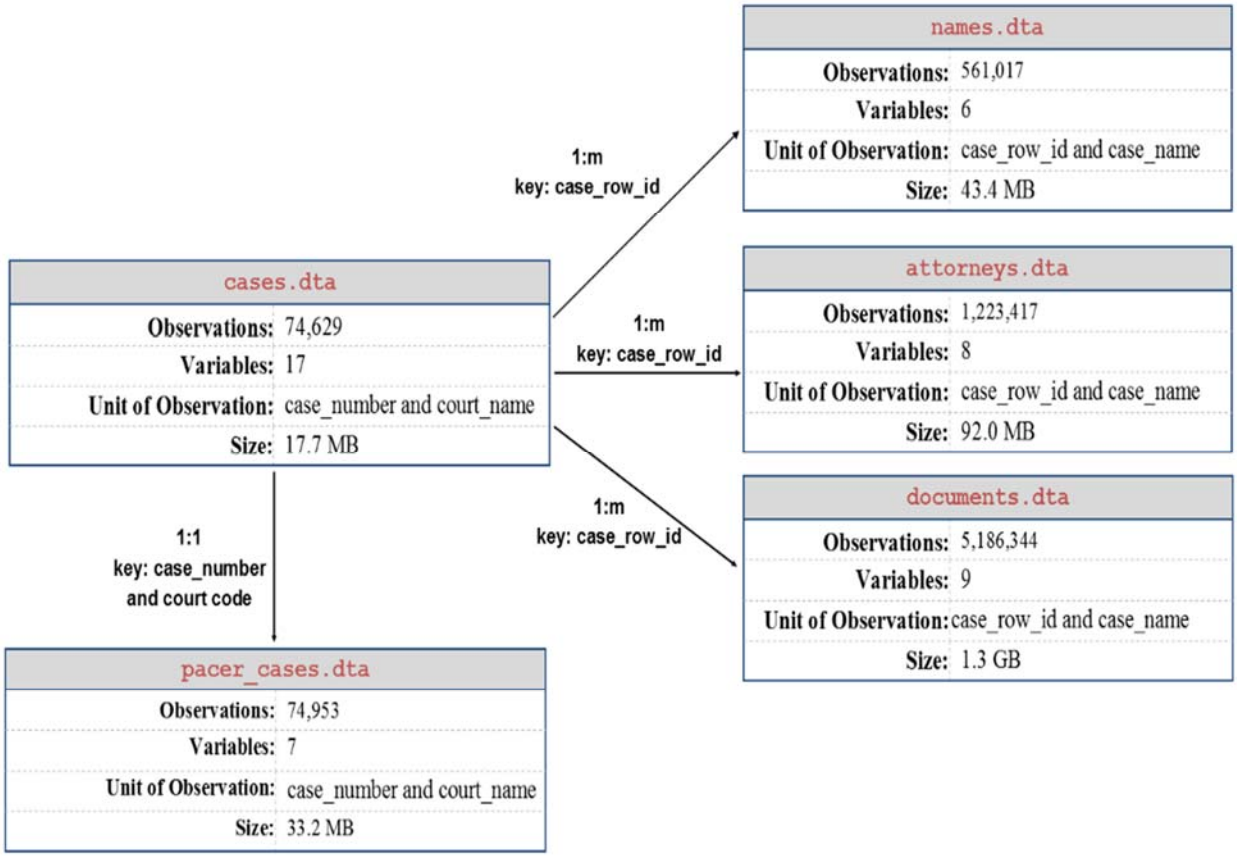
This section describes the data collection process and contents. On March 6, 2016, the docket reports for 74,664 unique cases were downloaded covering all data through that date.[12] In our data release, the type and number of the documents in each report and their metadata covers the full years 1963-2015 and contains a total of 74,623 unique cases. As noted earlier, PACER coverage of pre-1999 patent litigation cases is incomplete and decreases substantially going back in time from 1998. Furthermore, a small number of additional cases are missing due to mislabeling of the Nature of Suit. Thus our data is more likely to have a near complete coverage of the patent litigation cases post-1999.

The downloading and processing was done in multiple stages. First, a computer script was developed to automate the download of all docket reports from PACER, provided that an up to date version was not available in RECAP. It looped through each docket report in PACER and checked if it was available in RECAP. When the record was found in RECAP, the process evaluated if the last date of the RECAP update matched the record in PACER. If so, the docket report is downloaded from RECAP in HTML format. Otherwise, after paying a fee, it was downloaded from PACER in HTML format. Second, docket reports downloaded from PACER were submitted to be processed and posted on RECAP. So, all of the docket reports we collected are publicly available from RECAP free of charge. Third, the docket reports in HTML format were parsed and converted to XML. Fourth, the raw data were converted into both STATA .dta and regular .csv formats.

The final docket reports data are contained in four different files (i.e. cases, names, attorneys, documents). In Figure 1, we provide a data file schema that shows file sizes, the number of variables, number of observations, the unit of observation, and how each of the files are linked to each other. A detailed description of the files and their contents is provided below.

[12] There are 41 unique cases from the partial year (i.e. 2016) included in this count.

Figure 1: Data File Structure



C. Data Files

1. cases File

The cases data file contains metadata information on each of the 74,629 cases (i.e. case numbers) across all district courts for the years 1963-2015, of which 74,623 are unique. A case number is only unique within a district court and not necessarily across district courts. Therefore, a unique observation in the data is defined as a case number plus a district court name (*case_number* + *court_name*). There are six duplicate (*case_number* + *court_name)* observations that are kept in the cases data file.[13] We decided to keep these duplicates because some of the data fields were different across the observations. We defined 17 variables that capture the

[13] The case number and court name of each of these six cases are as follows: (1) 1:05-cv-00590 in the Western District of Michigan; (2) 1:06-cv-00546 in the District of Delaware; (3) 1:13-cv-00009 in the District of Delaware; (4) 2:14-cv-00153 in the District of Utah; (5) 3:05-cv-04374 in the Northern District of California; (6) 5:13-cv-05460 in the Northern District of California.

metadata elements from the header box of the docket reports for this file. A detailed description of each of these variables follows.

When a court receives a newly filed case, it assigns it a case number (or docket number). An example of this number as recorded under the variable case_number is 0:92-cv-00398-MJP. The first digit corresponds to the courthouse (also known as a "division") within the district that received the case. The next two digits indicate the year when the case was filed. In this example, the case was filed in 1992. The letters following the year identify the type of case filed (e.g. civil, criminal, family court, etc.) as designated by the filing attorney. In this example, the letters 'cv' indicate that it is a civil case.[14] The five digits that follow indicate the chronological order in which this case was received by the court in that year. That means that there are 397 other cases that were filed in this court prior to our example case. The last three letters provide the initials of the judge assigned to the case. Looking at the assigned_to variable in the same file, we learned that the case was assigned to the Honorable Matthew J Perry, Jr. Finally, some cases have an additional set of three letters at the end. These letters correspond to the initials of the magistrate judge. The variable referred_to provides the full name of the magistrate judge.

Other identifiers of the case include the pacer ID and the case name. The PACER ID, denoted as pacer_id, is a PACER generated number that tracks each case as it enters the PACER system. The case name, denoted as case_name, is composed of the first listed name from each party in any given case. There is also information on location and date. The location is the name of the district court (court_name) and the dates include the filing, closing, and last document filing date (date_filed, date_closed, date_last_filed).

The cases file also includes variables that provide information on the legal basis of filing and, when applicable, variables that identify other litigation records related to the current case. The case_cause variable indicates statutory basis or the cause of action in the case. An example entry is "15:1126 Patent Infringement". The first part of this entry identifies the code and section of the relevant statute while the second part provides a brief description of the cause of action. The variable jurisdictional_basis provides the stated basis for the court to have subject matter jurisdiction, which is often "Federal Question".[15] Furthermore, the applicant or filing attorney may indicate that there exists a case that is related to the one being filed. In the event that such information exists, the related_case variable provides the relevant case number. The applicant

[14] Note that all patent cases are civil. The case type is a function of the submitted document initiating the suit. For example, a case type of 'cv' should be assigned if the initiating document is a complaint, a notice of removal, or a petition for writ of habeas corpus.
[15] "Federal Question" indicates that the district court has subject matter jurisdiction because the suit raises a federal law question.

may also indicate a lead case (lead_case) with which the current case is associated. When a lead case is specified, many documents are automatically added to the associated cases.[16]

Another set of variables provides information on litigant preferences. These variables include whether either (or both) of the parties requested for a jury trial (jury_demand); the amount of monetary damages demanded (demand); or if there is a record of settlement (settlement).[17] [18]

The common identifier required to link the cases data file to all other *docket report* data files discussed below is called case_row_id. Appendix I provides a definition of each of the variables in the cases data file.

2. names File

The names data file is taken from the docket report section that identifies each of the parties and lists their names. This data file contains: (1) the case number (case_number); (2) a label that describes the named party's affiliation to the case (party_type); (3) the name of the party (name); (4) the case-level identifier used to link across all the data files (case_row_id); (5) a new variable that counts the number of entries for party type, called (party_row_count), which also tracks this count across docket reports. This new variable is also included in the attorneys data file described below and can serve as another variable for the purpose of linking the names and attorneys data files. Another new variable called (name_row_count) counts each new entry of a name and tracks this count across docket reports. Because there are situations where more than one name entry occurs per party type, the count of names is higher than the count of party type. Note that no effort has been made to clean or standardize any of the text name fields. Appendix II provides a variable dictionary for the names data file.

3. attorneys File

Like the names data file, the attorneys data file is taken from the docket report section that identifies each of the parties. That section contains information on the representing attorneys

[16] Not all documents can be transferred to member cases through the lead case. Documents that cannot be automatically added and require separate submission for each case include: complaints, answers, reassigning motions, assigning cases, use of the credit card functionality, and service of process events.
[17] While the demand variable should indicate monetary amounts, the majority of records in this field are "Plaintiff", "Defendant", or "Both". These entries seem to indicate that either or both parties have demanded damages.
[18] The settlement record provides the case number of instances where either party has settled.

and their contact information, although the fields are not always populated. The attorneys data file also contains four of the variables that appear in the names data file. These are: case_row_id, case_number, party_row_count, and party_type. The new variables are: (1) a count variable of every entry identifying an attorney, called attorney_row_count; (2) the attorneys' name (name); (3) their contact information (contactinfo); and (4) their position as representing attorneys in the case (position). Again, no effort has been made to clean or standardize any of the text name fields. Appendix III provides a variable dictionary for the attorneys data file.

4. documents File

The documents data file provides information on each document submitted in the case as recorded in the documents section of the docket reports. The information includes: (1) when the document was filed with the court (date_filed); (2) a short description (short_description) of its contents; (3) a long description (long_description) of its content; (4) its order in the sequence of recorded documents (doc_number); (5) the number of additional documents attached (attachment); and (6) the upload date (upload_date).[19] We generated a variable counting the number of documents recorded in a given docket report (doc_count). There were many instances where our count of documents is smaller than the document's sequence number. This indicates that some docket report documents are missing from the PACER records without explanation. As in the other files, the documents data file contains the case_row_id and case_number variables for linking. Appendix IV provides its variable dictionary.

5. pacer_cases File

In addition to the docket reports, PACER provides case-level metadata information covering the 1963-2015 period. These data are important for supplementing missing data in the docket reports and allows for checks on the internal consistency of PACER. We provide these data in the pacer_cases data file. The variables include: (1) the case number (case_number); (2) the name of the parties (case_name); (3) district court (court_name); (4) district court code (court_code); (5) the pacer ID (pacer_id); (6) the filing date (date_filed); and (7) the closing date (date_closed). Appendix V provides a variable dictionary for the pacer_cases file.

[19] Note that some filed documents may not have a document number.

V. Discussion and stylized facts

In this section we provide descriptive statistics for the datasets and explore some of the ways the data might be used. We first summarize the content of the categorical variables and identify any inconsistencies in the files. Second, we merged the case-level data contained in the cases file and those contained in pacer_cases file to supplement missing data elements and to evaluate PACER's internal consistency. We then provide some simple examples that illustrate how the data can be used to learn about the contents of the docket reports, litigation trends and cases by district court.

Description of categorical variables:

We identified a number of categorical variables contained in the files from the docket reports. The categorical variables in the cases file include *jurisdictional basis*, *cause of action*, *jury demand*, and *damages demand*. Also, *party type* is categorical in the names file and attorneys file while the variable *attorney position* is categorical in the attorneys file.

Some of the categorical variables require further processing before they can be used for analysis. For example, there are 366 different types of cause of action recorded in the variable case_cause. Many of the types observed refer to the same cause of action, but are recorded differently. For example, *15:1136 Patent Infringement* and *28:1338 Patent Infringement* both refer to a patent infringement cause of action although the statutory basis indicated by the numbers are different. Many of the entries found account for a small number of observations in the dataset. For example, there are three cases with *15:2 Antitrust Litigation* as their cause of action. The majority of the cases have some type of patent infringement entry. Also of note, 3,050 of the cases have no record of cause of action (i.e. case_cause has missing values).

A second variable that requires additional processing for analysis is the demand variable. While this variable captures information about the amount of damages demanded, it also includes information about which party submitted a demand (i.e. plaintiff, defendant, or both). Fifteen entries indicating damage ranges and party types are recorded in the demand variable.

Similarly, two of the other data files contain categorical variables that have a number of redundancies. The party_type variable in both the names file and the attorneys file has 131 unique entries, with a large majority of them referring to either the defendant or the plaintiff. The attorneys file has a variable indicating the attorney's role in the case, called position. Most of the records contain "ATTORNEY TO BE NOTICED", "LEAD ATTORNEY", or "TERMINATED" as

entries. However, due to a large number of recording variations for the same type of substantive entry and due to the existence of numerous entries with substantive difference in their information content, there are hundreds of different types of records in this variable.

For the jurisdictional basis variable in the cases file, it has five types of entries. In columns 1 and 2 of Table 2, the names of these entries and the frequency with which they appear are provided. Given that the default appears to be an entry of "Federal Question" and that patent litigation is correctly coded as a federal question, it is unsurprising that "Federal Question" makes up over 98.5% of the records under jurisdictional_basis (71,713 entries). In other words, the court has subject matter jurisdiction primarily on the basis that the suit raises a federal law question. When the U.S. Government is a defendant (U.S. Government Defendant) or a plaintiff (U.S. Government Plaintiff), these are also bases for subject matter jurisdiction. In the data, we observe 529 and 196 records of these jurisdictional bases, respectively. Finally, the parties' diversity of citizenship (i.e. "Diversity") is another basis for subject matter jurisdiction recorded in the data and it covers 29 of the entries.

Again in the cases file, the variable indicating whether the plaintiff, defendant, or both requested a jury trial (i.e. the jury_demand variable) has five possible entries. Unfortunately, most of the observations are missing from the record. However, based on the available information, plaintiffs are the parties that request a jury trial most frequently. There are also two anomalous entries showing values "P" and "y." Although we searched the full text of the complaint documents in PACER, neither case included complaints and we were unable to determine what these two entries mean.

PACER Metadata and PACER Docket Reports comparison:

As explained earlier, we obtained patent litigation data from PACER and RECAP in two different forms. First, we extracted comprehensive metadata elements from the docket reports. Second, we obtained the metadata table that PACER provides for all cases with *Nature of Suit* 830. This allowed us to cross-supplement information that may be missing from each dataset. For example, we found a large portion of the filing date information is missing from the docket report source. However, the PACER metadata has complete records of the filing dates. Using these datasets we can also evaluate the internal consistency of PACER's contents. Table 3 shows the match rate between the two datasets. We observed 822 unique cases in the PACER metadata that were missing from the docket reports, while only 28 cases in the docket reports were missing from the PACER metadata. We did not observe any systematic differences between the datasets to explain these discrepancies. However, it should also be noted that 74,027 (which is ~ 99% of all observations) were found in both datasets.

Table 2: Categorical Variables and their Content

Jurisdictional Basis	
0	306
Diversity	29
Federal Question	71,713
U.S. Government Defendant	529
U.S. Government Plaintiff	196

Jury Demand	
Both	2,989
Defendant	402
P	1
Plaintiff	3,561
Y	1

Table 3: PACER Metadata vs PACER Docket Reports

	Not Contained in Metadata	Contained in Metadata
Not Contained in Docket Reports	-	822
Contained in Docket Reports	28	74,027

Example Text Analytics on the Short and Long Descriptions:

Using the documents dataset, we explored the contents of the short and long description variables. Our objective was to identify the type of document and how often that document type appears in the dataset.[20]

The first step was to perform some basic cleaning. We removed non-alphabetic characters from the *short_desription* and *long_description* variables. We also generated a new variable that started with the information in the *long_description* and added the *short_description* whenever the long description was missing.[21] We call this variable "combined description."

The first type of document we tried to identify is court judgement, which reflects the outcome of patent litigation cases. Many scholars analyze case outcomes (Allison et al. (2014), Bessen et al. (2012), Henry (2013), etc.). One approach is to look for documents recorded as an "order" or a "judgement" in the docket reports. So, using a regular expression tool, we found all entries in the combined description that began with the words "order", "so ordered", "judgement", or "judgment" (notice the alternative spellings of judgement). Our identification process was not case sensitive and the results are likely to underestimate the count as we strongly avoid false positives. We found 631,840 such entries.

Using this same approach, we identified and counted the frequency of four other types of documents. First, we identified all descriptions that began with the word "exhibit" and found 65,364 entries. These documents probably reflect submission of evidentiary information to the court. Second, we identified descriptions that began with the words "complaint" or "amended complaint" and found 83,113 entries. The number of amended complaints was 14,727 and the number of complaints totaled 68,386. Considering that there are 74,623 unique cases in the dataset, it is reassuring we found 68,386 complaints.[22] However, our total does not capture all complaint records in the docket reports. For example, "amended and restated complaint" would not be captured although such entries do exist in the documents file. For our third type, we identified descriptions that began with the words "answer", "counterclaim", "amended answer" or "amended counterclaim" and found 130,231 entries. We believe these are part of the defendants' responses to a plaintiffs' complaints.

The last document type we identified were notices to the USPTO, also called AO120s. Statue 35 U.S.C. § 290 requires clerks of US courts (not limited to district courts) to provide notice in writing to the USPTO of both the filing and disposition of any action involving a patent,

[20] Note that analyses of text fields are quite variable. We tried to implement a "conservative" approach to limit the possibility of over-counting document types.
[21] After combining these two variables only 7,478 of 5,186,344 observations are missing entries.
[22] We found that 67,920 of these complaints have unique case numbers, alleviating the concern that our result was driven by duplicate counts of complaints per case.

including the names of the parties and the number(s) of the patent(s) upon which the action was brought. Identifying these documents in a docket report provides important information as it indicates that an event of substantive significance occurred. After searching combinations of "commissioner" and "patent" or a combination of "patenttrademark" and "AO," we found 33,469 entries. Table 4 summarizes all of our text search findings for the document types and frequencies.

Table 4: Text analytics on documents file

Document Type	Text Searched	Count
Judgments	"order", "so ordered", "judgment" or "judgement"*	631,840
Exhibits	"exhibit"*	65,364
Complaints	"complaint" or "amended complaint"	83,113
Answers	"answer", "counterclaim", "amended answer" or "amended counterclaim"	130,231
Notices to the USPTO	("commissioner" and "patent") or ("patenttrademark" and "AO")	33,469

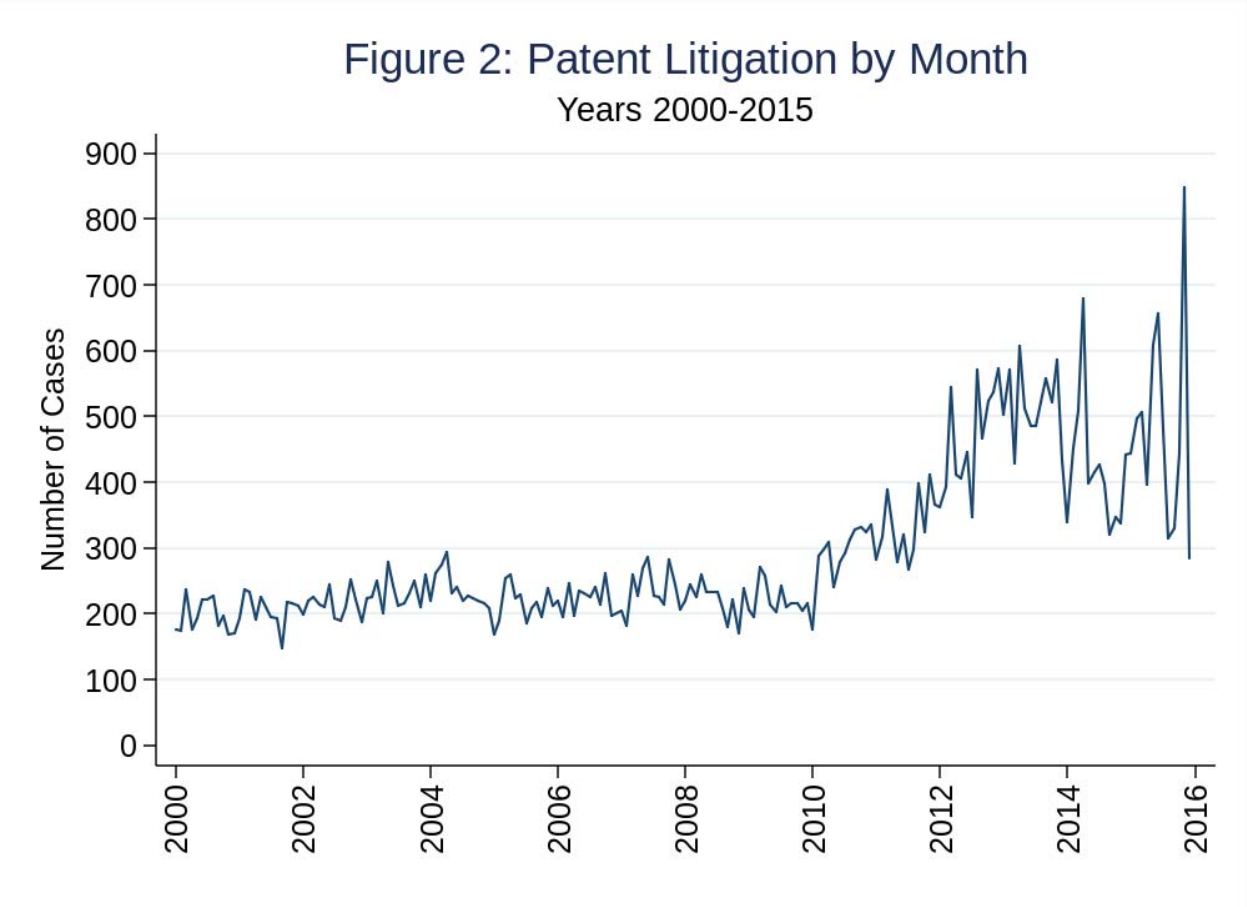
* Indicates that these words were searched at the beginning of the description content.

Other Illustrative Examples:

Given the importance of patent litigation trends, we plotted the rate of new case filings using the case-level data from PACER. Figure 2 shows the monthly time series starting in 2000.[23] Notably, there appears to be a clear break in the trend starting around 2011, after which monthly filings increased substantially and also became more volatile. Both of these changes might be attributed to procedural changes that have affected filing practices. For instance, the AIA Joinder Rule § 299 which took effect in 2011 clearly has had an impact in the patent litigation filing rate.[24] However, this conjecture should be studied more thoroughly and future

[23] Our coverage of the trend starts in 2000 because, as noted earlier, the pre-1999 PACER data on patent litigation cases is significantly less reliable.
[24] Considering that the standards set forth in § 299 are nearly identical to the FRCP Rule 20, it is likely that the former had the effect of stricter adherence and more uniformity in application across courts.

research should consider other factors such as changes in patent quality or litigants' patent assertion strategies.

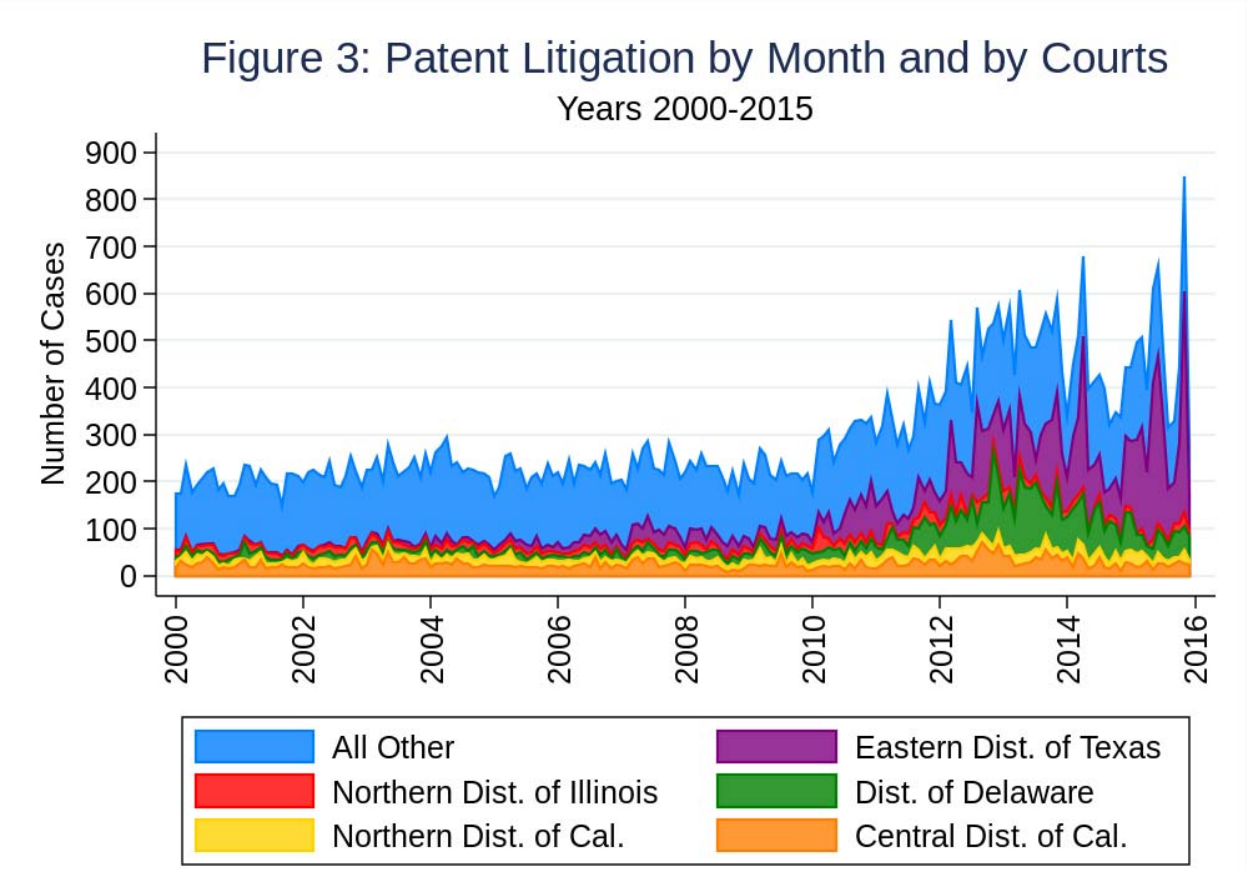


Another pressing issue is the location where patent litigation suits are filed. Policy makers and courts have struggled for over a quarter of a century with the "proper venue" for patent litigation. One case, *TC Heartland,* is currently before the Supreme Court along with a number of amicus curiae, many of which discuss behaviors such as "forum shopping" and "forum selling."[25] [26]

So how have the filing rates across district courts changed? Figure 3 shows the trends in filing across five major district courts over 2000-2015. The courts considered are: Eastern District of

[25] While the former refers to plaintiffs selectively filing their case in a court that they perceive as being more patent-owner friendly, the latter refers to a court's application of both procedural and substantive laws in a manner favorable to a plaintiff.
[26] A number of briefs as Amicus Curiae are submitted to the Supreme Court in this case, including one by 22 law, economics, and business professors in support of respondent. See generally, brief as Amici Curiae, TC Heartland LLC v. Kraft Foods Brands Group, No. 16-341 (submitted March 8, 2017).

Texas, District of Delaware, Northern District of Illinois, Northern District of California, and Central District of California. [27] It is notable that the Eastern District Court of Texas had over half of all of the monthly filings more than once in the past two years. The District of Delaware also had a significant share of the filings, although we observed a slight decline in its share in the last two years. One possible explanation for these changes may be that plaintiffs are increasingly targeting select district courts for their filings.[28] It will be important for future research to determine whether forum shopping behavior explains these changes and to understand the implications for courts and defendants.[29]



Digging even deeper into the data, Figure 4 shows the same five district courts focusing on 2015. The figure clearly illustrates the variation in rates of filing by month in 2015. The Eastern

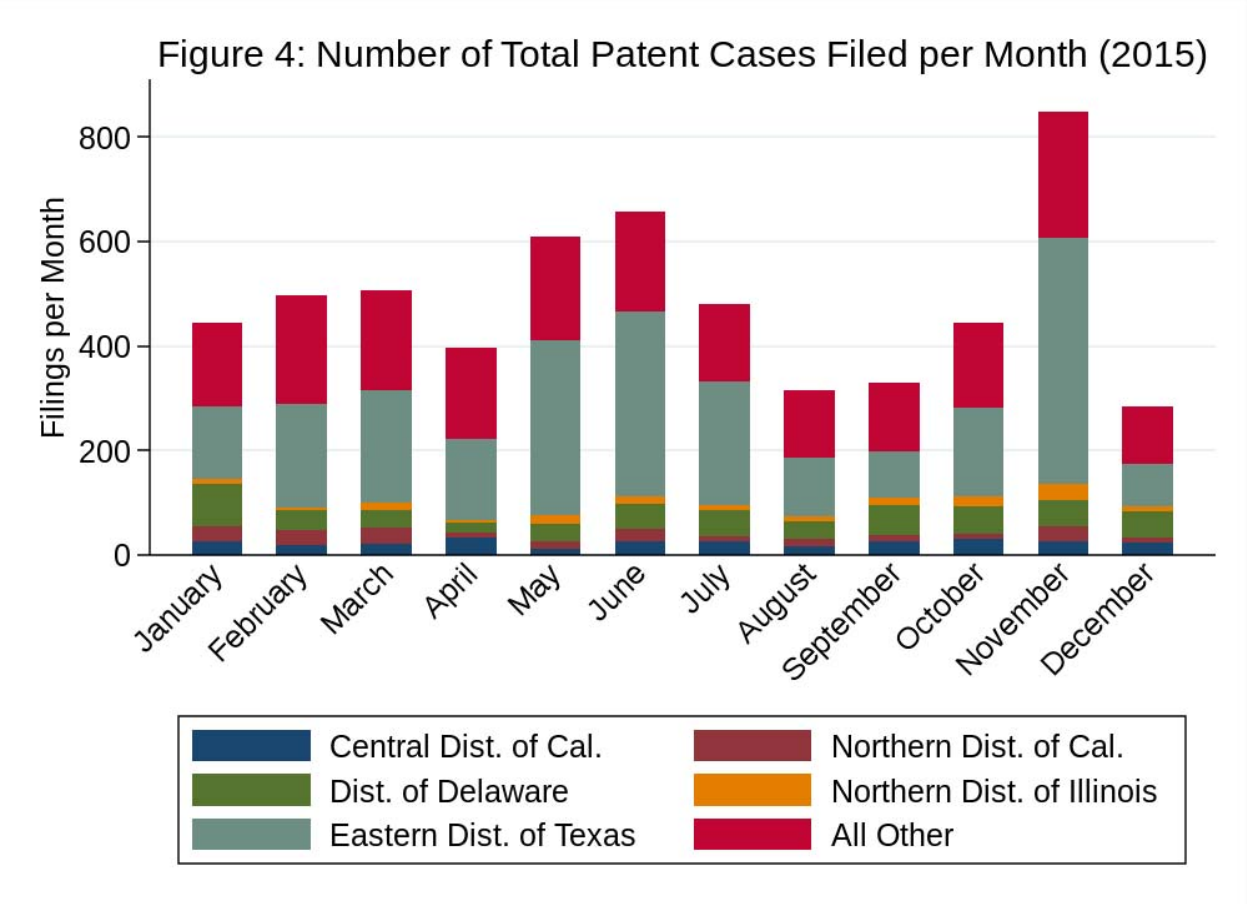
[27] Note that courts adopted usage of the PACER system at different times and have had varying backfilling practices when joining PACER. However, we expect that the proportions in the time series provided give accurate estimates since PACER coverage of patent case filings has been reliably comprehensive for at least the period 2002-2015.

[28] Again our graph is sensitive to the joinder rule passed by the AIA. However, as argued earlier, this rule is more likely to lead to more uniformity in the application of the rule. Thus, the fact that we see greater disparity across courts in recent years is even more noteworthy.

[29] For an empirical study showing forum shopping by validity rates the pre-CAFC era of significant nonuniformity in patent validity rates, see Atkinson et al. (2009). The study also finds strong evidence that the end of validity rates based forum shopping preceded the CAFC by several years and weak evidence that CAFC contributed to uniformity of validity rates.

District of Texas had the highest rate of filings of any court, making up more than half of all filings in some months, and accounting for a lot of the variation in filings per month.[30] The reason for this volatility is not apparent and would benefit from further analysis.

Also, notice that case filings in November of 2015 were particularly high. The increase in filings, while highest in the Eastern District of Texas, was also visible in many other courts. One possible explanation has to do with changes in legal procedures. On December 1st, 2015, the rules for discovery in civil proceedings were due to change. November's surge in filing may reflect a choice by plaintiffs to file before the rule change.



VI. Conclusion

[30] The variance in filing rates per month for the Eastern District of Texas in 2015 is more than three times as high as any of the variances in the other courts.

Strong empirical evidence serves as the foundation supporting good policy formulation. To do this, however, researchers need access to comprehensive data that can be purchased at a price appropriate for limited research budgets. This document and its accompanying data files provide free and public data drawn from the most comprehensive source of patent litigation information in the United States, namely PACER. This document provides details on the data release as well as an overview of the literature and some suggestive uses of the information. The datasets cover both the case-level metadata and the docket reports from PACER for the 1963-2015 period.

While these datasets contain a wealth of information, important elements are still missing and incomplete. First, some variables, such as jury demand information, have a significant number of observations with missing data. Second, variables such as patents-in-suit and coded outcome information are not provided. Enhancing the data in these ways will require extensive manual or automated processing efforts. Third, data from other venues where patent litigation takes place, such as appellate courts and the Patent Trial and Appeal Board are not included. Extracting granular information from these sources and preparing datasets for research would also be a valuable contribution to future research and understanding.

In future work, we hope to enhance the data by identifying patents-in-suit and by better capturing information on the litigation outcomes. We hope that free access to comprehensive data can facilitate and advance research on the evolving patent litigation landscape and its economic impact.

VII. References

Allison, John R., M. A. Lemley, and D. L. Schwartz (2014). *Understanding the Realities of Modern Patent Litigation*. 92 Texas L. Rev. 1769.

American Intellectual Property Law Association (2015). *Report of the Economic Survey 2015*. AIPLA Report.

Atkinson, Scott E., A. C. Marco, and J. L. Turner (2009). *The Economics of a Centralized Judiciary: Uniformity, Forum Shopping, and the Federal Circuit*. 52(3) Journal of Law & Economics 411.

Bessen, James and M. J. Meurer (2014). *The Direct Costs from NPE Disputes*. 99 Cornell L. Rev. 387.

Bessen, James and M. J. Meurer (2012). *The Private Costs of Patent Litigation*. 9 J. L. Econ. & Policy 59.

Chien, Colleen (2014). *Startups and Patent Trolls*. 17 Stan. Tech. L. Rev. 461.

Chien, Colleen (2011). *Predicting Patent Litigation*. 90 Texas L. Rev. 283.

Chien, Colleen (2009). *Of Trolls, Davids, Goliaths, and Kings: Narratives and Evidence in the Litigation of High-Tech Patents*. 87 North Carolina L. Rev. 1571.

Cohen, Lauren H., U. G. Gurun, S. D. Kominers (2015). *Patent Trolls: Evidence from Targeted Firms*. Harvard Business School. Finance Working Paper no. 15-002.

Cotropia, Christopher A., J. P. Kesan, and D. L. Schwartz (2014). *Unpacking Patent Assertion Entities (PAEs)*. 99 Minn. L. Rev. 649.

Galasso, Alberto and M. Schankerman (2015). *Patents and Cumulative Innovation: Causal Evidence from the Courts*. 130(1) Quarterly Journal of Economics 317.

Federal Trade Commission (2016). *Patent Assertion Entity Activity: An FTC Study*. FTC Report.

Government Accountability Office (2013). *Intellectual Property: Assessing Factors That Affect Patent Infringement Litigation Could Help Improve Patent Quality*. GAO Report GAO-13-465.

Henry, Matthew D. (2013). *The Market Effects of Patent Litigation*. 4(1) Technology and Investment 57.

Jeruss, Sara, R. Feldman, and J. Walker (2012). *The America Invents Act 500: Effects of Patent Monetization Entities on US Litigation*. 11 Duke L. & Tech. Rev. 357.

Lerner, Josh (1995). *Patenting in the Shadow of Competitors*. 38(2) Journal of Law & Economics 463.

Love, Brian J. (2012). *An Empirical Study of Patent Litigation Timing: Could a Patent Term Reduction Decimate Trolls Without Harming Innovators?* 161 U. Pa. Law Review 1309.

Marco, A., R. Miller, K. Fonda, P Laufer, P Dzierzynski, and M. Rater (2015). *Patent Litigation and USPTO Trials: Implications for Patent Quality*. USPTO Report.

Marco, A., S. Miller, and T. M. Sichelman (2015). *Do Economic Downturns Dampen Patent Litigation?* 12 Journal of Empirical Legal Studies 481.

Mazzeo, Michael J., J. H. Ashtor, and S. Zyontz (2013). *Do NPEs Matter? Non-practicing entities and patent litigation outcomes.* 9(4) Journal of Competition Law & Economics 879.

Miller, Shawn P. (2013). *What's the Connection Between Repeat Litigation and Patent Quality? A (Partial) Defense of the Most Litigated Patents.* 16 Stanford Technology Law Review 313.

Risch, Michael (2015). *A Generation of Patent Litigation.* 52 San Diego Law Review 67.

Risch, Michael (2012). *Patent Troll Myths.* 42 Seton Hall Law Review 457.

Schwartz, David L., T. M. Sichelman (2016). *Data Sources in Patents, Copyrights, Trademarks, and Other Intellectual Property.* 2 Research Handbook on the Law & Economics of Intellectual Property (Peter S. Menell & David L. Schwartz, eds.)

VIII. Appendix I: Data Dictionary cases File

	Definition	Number of Missing Entries	Type	Formatting
case_number	Case or docket number assigned by the court; identifies unique cases by court location and indicates the filing location, the year of filing, the case type, its filing sequence for that year, and the initials of the judges it was assigned to	0	str40	%-20s
case_row_id	Case-level unique identifier generated during processing and that serves to link the content of the four docket report files	0	long	%12.0f
pacer_id	ID assigned to a case in a chronological order as it is recorded in PACER	65,076	long	%12.0f
case_name	Name given to the case using the first listed entity names from each party	198	strL	%-20s
court_name	Name of one of the 94 district courts in which the case was filed	0	strL	%-20s
assigned_to	Name of the judge to whom the case is assigned	2,202	strL	%-20s
referred_to	Name of the magistrate judge to whom the case was referred	47,700	strL	%-20s
case_cause	Statutory basis or cause of action in the case; the statutory code followed by a brief description is provided	3,050	strL	%-20s
jurisdictional_basis	The basis under which the court has jurisdiction over the case; the default and primary basis is that the case raises a "Federal Question"	1,897	str25	%-20s
date_filed	The court filing date of the case	65,119	int	%td
date_closed	The disposition date or when the case was closed	65,497	int	%td
date_last_filed	The date that the latest document was added to the docket report of the case	71,617	int	%td
jury_demand	Whether "Plaintiff", "Defendant", or "Both" have requested for a jury trial	67,676	str9	%-9s
demand	Amount of damages demanded; however most of the entries in this field are "Plaintiff", "Defendant", or "Both"	22,743	str10	%-10s
lead_case	Case number that the current case is associated with, if any; this information is added by a court staff	71,623	strL	%-20s

related_case	Case number of a case that relates to the current case as identified by the filing agent	59,698	strL	%-20s
settlement	Lists cases in which either of the parties have settled	74,301	strL	%-20s

IX. Appendix II: Data Dictionary names File

	Definition	Number of Missing Entries	Type	Formatting
case_number	Case or docket number assigned by the court; identifies unique cases by court location and indicates the filing location, the year of filing, the case type, its filing sequence for that year, and the initials of the judges it was assigned to	0	str40	%-20s
case_row_id	Case-level unique identifier generated during processing and that serves to link the content of the four different files	0	long	%12.0f
party_type	The named party's affiliation in the case; while there are many variations of the recorded types, the party type predominately refers to either the plaintiff or the defendant	8	strL	%-20s
party_row_count	Count tracking each entry of a given party type across docket reports	0	long	%12.0f
name	Name of the party	10	strL	%-20s
name_row_count	Count tracking each entry of a name across docket reports	0	long	%12.0f

X. Appendix III: Data Dictionary attorneys File

	Definition	Number of Missing Entries	Type	Formatting
case_number	Case or docket number assigned by the court; identifies unique cases by court location and indicates the filing location, the year of filing, the case type, its filing sequence for that year, and the initials of the judges it was assigned to	0	strL	%-20s
case_row_id	Case-level unique identifier generated during processing and that serves to link the content of the four different files	0	long	%12.0f
party_type	The named party's affiliation in the case; while there are many variations of the recorded types, the party type predominately refers to either the plaintiff or the defendant	1	strL	%-20s
party_row_count	Count tracking each entry of a given party type across docket reports	0	long	%12.0f
attorney_row_count	Count tracking each entry of an attorney across docket reports	0	long	%12.0f
name	Name of attorney	1,787	strL	%-20s
contactinfo	Any record on the contact information of the representing attorney; includes physical address, phone number, fax, and email	1,948	strL	%-20s
position	Position of the attorney as it relates to the case; the two main entries are "LEAD ATTORNEY" and "ATTORNEY TO BE NOTICED"	38,129	strL	%-20s

XI. Appendix IV: Data Dictionary documents File

	Definition	Number of Missing Entries	Type	Formatting
case_number	Case or docket number assigned by the court; identifies unique cases by court location and indicates the filing location, the year of filing, the case type, its filing sequence for that year, and the initials of the judges it was assigned to	0	strL	%-20s
case_row_id	Case-level unique identifier generated during processing and that serves to link the content of the four different files	0	long	%12.0f
doc_count	Count tracking each document added to the docket report of a given case	0	int	%8.0f
attachment	Count of attachments included with the added document	4,204,068	int	%8.0f
doc_number	Number, as recorded in the docket report, that represents the order of a given document in the sequence of documents that are added to the docket report	509,726	str14	%-14s
short_description	Brief description of what is contained in the added document	4,834,609	strL	%-20s
long_description	Longer, yet still brief, description of what is contained in the added document	90,541	strL	%-20s
date_filed	Date when the document was filed with the court	47,319	int	%td
upload_date	Date when the document was uploaded into PACER	5,101,894	int	%td

XII. Appendix V: Data Dictionary pacer_cases File

	Definition	Number of Missing Entries	Type	Formatting
case_number	Case or docket number assigned by the court; identifies unique cases by court location and indicates the filing location, the year of filing, the case type, its filing sequence for that year, and the initials of the judges it was assigned to	0	str182	%182s
pacer_id	ID assigned to a case in a chronological order as it is recorded in PACER	0	str7	%9s
case_name	Name given to the case using the first listed entity names from each party	3	str220	%220s
court_name	Name of one of the 94 district courts in which the case was filed	0	str37	%37s
court_code	Code identifying one of the 94 district courts in which the case was filed	0	str4	%9s
date_filed	The court filing date of the case	0	float	%td
date_closed	The disposition date or when the case was closed	6,657	str10	%10s