

The Case of the Missing Cases: Inferring Sealed U.S. Federal Cases

Ye Shu
UC San Diego
La Jolla, CA, USA

Elisa Luo
UC San Diego
La Jolla, CA, USA

Paul Chung
UC San Diego
La Jolla, CA, USA

Geoffrey M. Voelker
UC San Diego
La Jolla, CA, USA

Stefan Savage
UC San Diego
La Jolla, CA, USA

Abstract

In the U.S., the public right of access to court proceedings arises from the common law and the First and Sixth Amendments. However, courts have long recognized limited exceptions to this right, and certain court cases may be “sealed” (inaccessible to all but authorized parties) and thus, even the existence of such proceedings may be shielded from public scrutiny. This paper develops empirical measurement techniques to infer both the prevalence of Federal case sealing and the duration of such sealing orders in practice. Using a combination of comprehensive crawling, undocumented Web APIs, and heuristics for inferring per-court case-numbering policies, we are able to use existing online legal databases (the U.S. Courts’ CM/ECF and IDB systems and the Free Law Project’s CourtListener RECAP service) to tease out the presence of case sealing via “absences” in the online record. Using this approach, we provide an initial analysis of sealing practices across twenty Federal court districts, how these relate to different “kinds” of proceedings, and how our contemporary findings compare to the U.S. Courts’ internal study conducted seventeen years prior.

CCS Concepts

• **General and reference** → **Measurement**; *Empirical studies*; • **Applied computing** → **Law**.

Keywords

Sealed Court Cases, Federal Courts, Judicial Transparency, Public Court Records, Side-channel Inference

ACM Reference Format:

Ye Shu, Elisa Luo, Paul Chung, Geoffrey M. Voelker, and Stefan Savage. 2026. The Case of the Missing Cases: Inferring Sealed U.S. Federal Cases. In *Proceedings of the 2026 ACM Internet Measurement Conference (IMC '26)*, October 12–16, 2026, Karlsruhe, Germany. ACM, New York, NY, USA, 21 pages. <https://doi.org/10.1145/3777912.3839786>

1 Introduction

In the U.S., the default presumption is that judicial proceedings and their contents are open to the public. However, in a variety of circumstances the courts will abridge this right for good reason:

e.g., to protect the privacy of victims, the integrity of ongoing investigations, or the value of trade secrets. Some such protections can be afforded simply by redacting key information from a document, but if this is impractical (or if the court must review the document in its entirety) then individual documents can be “sealed” — that is, closed to the public and only made available to the court and authorized parties in the dispute. However, in other situations, an entire *case* must be sealed. This last outcome is one that causes particular concern for civil libertarians, as the nature of such a proceeding, the claims made in motions on its docket, and even the names of the parties and their representatives are no longer available for public scrutiny.

Unfortunately, characterizing the nature of *case sealing* is challenging since, by definition, the proceedings are not reviewable and thus, the answer to questions such as: “What fraction of cases is sealed? What kinds of cases are they? Do they ever get unsealed? How does sealing differ across the country and how is it changing over time?” are generally not available. The gold standard for such an investigation is the Federal Judicial Center’s 2009 study characterizing sealing in the Federal courts for cases filed in the calendar year 2006 [24]. However, seventeen years later, we are unaware of any group that has repeated such an effort. We surmise that it is not for lack of interest, but a combination of access (the FJC, as part of the Federal judicial branch, had unfettered access to court records, sealed or not, and was well positioned to elicit data from the clerks of each Federal court), monetary cost, and the labor involved in examining and validating tens of thousands of cases.

Adapting classic network crawling and inference techniques to this space, we make four contributions in this paper. First, we show how the basic approach of the prior FJC study can be replicated *entirely using inference*. Using a combination of crawled online data from the U.S. Courts’ Case Management/Electronic Case Filing (CM/ECF) system, combined with undocumented, but consistent, behaviors of their case-number API endpoint, we are able to precisely identify which cases have sealed dockets. In particular, for 20 Federal districts, we provide data about *currently sealed* civil, criminal and magistrate cases filed over the last six years — notably showing significant increases in sealing rates for magistrate cases.¹ Second, we develop and validate an alternative approach for inferring such case sealing *without* incurring CM/ECF fees, using data obtained by crawling the online API for the Free Law Project’s CourtListener RECAP service. Third, using data from both CourtListener and the U.S. Courts Integrated Database (IDB), we develop new techniques for inferring the existence of cases that have been



This work is licensed under a Creative Commons Attribution 4.0 International License. *IMC '26, Karlsruhe, Germany*

© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2327-8/26/10
<https://doi.org/10.1145/3777912.3839786>

¹For most of these districts we have obtained research exemptions, forgiving our costs, and for the remainder we have paid to scrape their CM/ECF Case Reports.

sealed *in the past* as well as the *duration* of the sealing periods — an analysis that we believe is the first of its kind. This analysis allows us to understand sealing not just as a static phenomenon, but as a dynamic one, where cases are sealed and unsealed over time, and to distinguish sealing as a temporary procedural tool versus a long-term barrier to public access. Finally, using this approach we show that civil cases are rarely sealed. However, when they are sealed they are likely to stay sealed for years, thus raising the suspicion that they may not reflect truly compelling interests.

We make our dataset of currently and historically sealed cases available to the public.²

2 Background

The public’s right of access to judicial proceedings in the U.S. is rooted in the common law (“the courts of this country recognize a general right to inspect and copy public records and documents, including judicial records and documents” [44]), in the First Amendment [45], and, for criminal cases, via the Sixth Amendment’s Speedy Trial clause [46]. Moreover, the exercise of this right continues to be litigated (and generally upheld), creating a body of case law supporting public access as well (e.g., most recently in Jason Leopold’s ultimately successful efforts to gain press access to search and surveillance orders [33, 35]). These rights are not just abstract legal ideals, but also serve an important societal function. Openness allows journalists, researchers, and the general public to monitor the use of the courts, to identify patterns of abuse, and ultimately to establish a record to hold institutions accountable and support trust in the rule of law.

However, the right of public access is a qualified one and there are an array of situations in which judges may seal whole cases. Among the more common exceptions found in the Federal court system: 1) criminal indictments are frequently sealed if there is concern that a suspect may flee before arrest, 2) court orders for criminal legal process (e.g., search and tracking warrants, pen registers, wiretaps, etc.) are frequently sealed to protect the integrity of an ongoing investigation (and may stay sealed even after an investigation is completed for reasons of privacy or policy), 3) Qui Tam actions under the False Claims Act are automatically filed under seal to give the government time to investigate allegations and decide whether to take on a case (the so-called “tolling” period), and 4) juvenile cases may be sealed to protect their identities and to support future expungement. Moreover, any party to a matter can make a motion to seal individual documents or the entire docket, and it is within judicial discretion to accommodate such requests. However, aside from such generalities, a quantitative notion of case sealing prevalence, how it is being applied, for how long and for what purpose — let alone how these factors change over time and geography — is not well understood.

The last comprehensive study on case sealing was conducted by the Federal Judicial Center (FJC) in 2008 and released in 2009 [24]. Using a combination of mining electronic court records (via CM/ECF), extensive interaction with the clerks of each district court, querying the presiding judges where necessary, and a range of sampling and extrapolation, the study concluded that nationwide case sealing rates in 2006 were only 0.2% for civil cases, 1.6% for criminal cases

and 16% for magistrate cases. While this study was, without dispute, the gold standard for characterizing case sealing it also left much unclear — including questions of sealing duration, variation by court, and changes in sealing behavior over time. The FJC study has also been criticized for underrepresenting sealing by focusing on *case sealing* without documenting the large amount of *per-document sealing* [20, 30]. This limitation equally applies to our work and establishing cost-effective ways to characterize per-document sealing remains an open research problem.

In the seventeen years since the release of the FJC report, we are unaware of any effort to repeat such a study. However, this is not for lack of interest, and case sealing has remained a topic of controversy in academia, civil society, and within the U.S. Courts system as well. One such vein of work concerns Qui Tam False Claims Act suits, which allow private individuals to sue on behalf of the government over alleged fraud against government programs or contracts. By statute, the complaint “shall remain under seal for at least 60 days, and shall not be served on the defendant until the court so orders” [32, (b)(2)]. In reality, these cases are often sealed for much longer [8, 22]. In particular, Engstrom used a combination of a FOIA’ed list of Qui Tam cases from the Department of Justice, docket sheets for each case, and interviews with attorneys, to establish that many such cases remained sealed long after closure, potentially due to neglect by the courts or DoJ [8]. This finding was quantitatively supported by Li, who built a dataset of such cases using Bloomberg Law’s commercial database and Nature of Suit (NOS) codes (filed in civil cases) to show that such cases are frequently sealed for over five years [22]. More broadly, Smith highlighted methodological deficiencies in the FJC report with respect to the sealing of criminal process (e.g., search warrants) and used contemporaneous data from docket sheets at three district courts to estimate the magnitude of the error [30]. Abraham, Abdo, and Manes, in advocating for new sealing rules to the U.S. Courts Committee on Rules of Practice and Procedure, document how the divergent sealing rules between district courts result “in inconsistent and improper sealing of court records, often in violation of the public, constitutional and common law rights of access” [1]. Finally, Rowe argues that motions to seal are among the most common court filings (using searches for “motion to seal” in LexMachina) and laments that the “FJC has not produced a similar study since this one even though the legislature has continued to discuss the issue of sealing” [28].

We believe important reasons for the lack of contemporaneous analysis include access, cost, and effort. By their nature, sealed records are not visible to the public and thus there is no direct way to measure them absent the privileged access available to the FJC; they can only be inferred. However, any such inference technique is further complicated by the significant access cost, as well as the considerable manual effort employed by the FJC in their analysis; we describe these constraints and how they shaped our selected district sample in Section 3. Thus, in addition to demonstrating how we can use mechanical inference to replicate the FJC’s methodology (for those districts where we can afford to do so), we have also focused on how highly similar results can be derived using entirely free public data sources (i.e., CourtListener) which did not exist during the FJC study. Furthermore, since such archives only report cases once they become public, we also show how longitudinal sealing behavior can be inferred as a side effect of their data collection.

²<https://github.com/ucsdysnet/us-sealed-cases>

Dataset	Measurement/inference	Signal and methodology	Evidence status and principal limitation
CM/ECF	Current sealing (Section 4.1)	Parse case metadata and sealing indicators from Case Reports; validate reconstructed case numbers when redacted using the case-number endpoint.	Direct observation of current case status. Ground truth, but expensive to query at scale.
IDB	Historical criminal sealing (Section 5.1)	Interpret sealing- and unsealing-related proceeding codes as indicators that a case was sealed previously.	Proceeding-code usage dependent. Different districts use and update the codes inconsistently, and IDB omits magistrate cases.
	Historical civil sealing (Section 5.1)	Identify cases whose plaintiff and defendant names are both recorded as “SEALED.”	Lower bound. The method detects a case only if an IDB snapshot captured it while sealed.
CourtListener	Current sealing from case-number gaps (Section 4.2)	Use case metadata to infer sealed cases from gaps in otherwise predictable case-number sequences.	Validated estimate against CM/ECF. Reliability varies depending on CourtListener coverage and district numbering.
	Historical sealing from ingest delay (Section 5.2)	Classify filing-to-ingest delays longer than four days as evidence that a case became public after filing.	Validated estimate for criminal cases in districts with full RSS coverage; exploratory for civil cases with substantial overestimation.
	Docket entries validation (Section 5.2 and 6)	Use an LLM to parse docket entry descriptions for (un)sealing-related entries and extract dates.	Secondary validation only. Many districts do not record (un)sealing events in docket text, and the LLM may misinterpret partial unsealing.
Combined	Duration under seal (Section 6)	Combine current sealed status from CM/ECF, historically sealed cases from IDB, and historical sealing durations from CourtListener.	Approximate duration for historically sealed cases identified by IDB. Inherits limitations of each source, with the civil cases subject to small sample sizes.

Table 1: Quick reference table from each dataset to its derived measurements and the validity of the resulting evidence.

3 Datasets

Our study makes use of three distinct data sources to investigate sealing in Federal district courts. Table 1 provides a quick reference from these sources to each derived measurement and its validity, with references to the section in which each measurement is presented; the remainder of this section describes the sources in detail.

Federal case numbers provide the common identifier across these sources, relative to a particular district. They generally have the form “0:YY-TT-NNNN”: 0 identifies a divisional office, YY is the filing year, TT is the case type (such as cr for criminal, cv for civil, or mj for magistrate), and NNNN is the sequence number. For example, 1:25-cv-11472 is a civil case filed in 2025.

CM/ECF. The Case Management/Electronic Case Filing system (CM/ECF) [7] is a federated system operated by each of the Federal districts used for managing court dockets and documents, and serves as the ground truth for the *current* sealing status of cases. We crawl and parse the CM/ECF’s “Case Report”s to obtain lists of cases and case metadata, including case numbers and titles, for specified filing periods and case types. While CM/ECF has no particular feature designed to identify sealed cases, we have determined that such cases are implicitly revealed within Case Report results by the lack of a link to the case’s docket sheet (they are typically also explicitly labeled as such, e.g., “SEALED v SEALED”, and have partially-redacted case numbers). In addition, when a Case Report result partially redacts a case number, we use an undocumented case-number API endpoint that returns a special message for currently sealed cases. Together, the Case Reports and the case-number API endpoint behavior provide direct observations of current sealing, and the endpoint helps reconstruct partially-redacted case numbers. We access each district court’s CM/ECF instance individually, using the Public Access to Court Electronic Records (PACER) [2]

federated login system, with a custom-built scraper designed to accommodate the idiosyncrasies of each district’s CM/ECF instance.

Of the 94 Federal district courts, we selected 20 for which we obtained comprehensive data for criminal, civil, and magistrate dockets³ for the period between January 2020 and June 2025 (Table 2). These twenty districts balance a set of practical constraints: in part prioritizing the subset of districts for which we were able to obtain PACER research exemptions to defray data costs, and in part prioritizing districts that provided granular “proceeding code” data to the U.S. Courts IDB (used in later stages of our analysis). PACER charges \$0.10 per report page, and printing *just the case names* for the District of Arizona’s civil and criminal cases in 2024 ran thousands of pages and cost over \$200 in fees. Without the research exemptions, our analysis in Section 4.1 would have cost more than \$12 000. Replicating such queries across 94 districts, across time, and then including the contents of their dockets as well becomes quite expensive.

We therefore cannot claim that this set of districts is “representative” in any statistical sense (especially given the financial bias in our selection). Nevertheless, the selected districts are diverse, spanning 10 of the 12 Federal circuits, and account for substantial shares of filings: for the 12-month period ending in June 2025, these twenty districts filed 66 915 new civil cases (23% of the national total), 11 108 new criminal cases (20%), and their magistrates issued 43 639 search warrants (21%) [6].

To access the data at scale, we built a CM/ECF crawler by extending CourtListener’s open-source Juriscraper [11], which handles session management via PACER, to fetch comprehensive Case Reports of district filings and to query the undocumented case-number

³Magistrate judges handle a variety of matters in lieu of district judges, including criminal legal process, Habeas petitions, various civil settlement actions and criminal complaints, among others.

District	Circuit	CL	Numbering
D. Puerto Rico	1	✓	Unified
S.D. New York	2	○	Unified
W.D. Pennsylvania	3	▲	Separate
D. South Carolina	4	▲	Unified
W.D. North Carolina	4	✓	Separate
W.D. Louisiana	5	✓	Unified*
E.D. Louisiana	5	✓	Unified‡
M.D. Louisiana	5	▲	Unified
N.D. Texas	5	○ [†]	Separate
W.D. Michigan	6	○	Separate
N.D. Ohio	6	○	Unified
S.D. Illinois	7	▲	Unified‡
S.D. Indiana	7	○	Separate
S.D. California	9	✓	Unified
D. Nevada	9	✓ [†]	Separate
D. Oregon	9	✓	Unified*
E.D. Washington	9	✓	Unified
N.D. Oklahoma	10	✗	Unified
W.D. Oklahoma	10	✓	Unified
N.D. Georgia	11	✗	Separate

Table 2: The twenty U.S. district courts we study and their characteristics. CL: CourtListener uses ✓ full or ○ partial RSS coverage and satisfies empirical timeliness; CL uses ▲ some or ✗ no RSS coverage and does not satisfy empirical timeliness. [†] the court introduced RSS coverage during our study period, leading to changes in CL coverage. Numbering: whether the district uses a Unified case numbering system or Separate across offices, * indicates that the district has limited overlap across offices. ‡: the district mixes numbering systems across case types.

API endpoint when a report partially redacts a case number. We then parse the HTML report pages using BeautifulSoup 4 [26] to extract case information, while attending to the slightly different formats and conventions used by each district. Since CM/ECF is a federated system, each district can make arbitrary modifications to its instance. Thus, we had to implement custom logic for each district to accommodate the idiosyncrasies of its CM/ECF instance. Specifically, when a report omits the office and sometimes the case type, we enumerate candidate full case numbers and submit only those candidates to the district’s undocumented case-number endpoint. We interpret its response as evidence of current sealing. The response format varies by district, so we empirically determined each district’s behavior and implemented custom logic accordingly.⁴

Integrated Database (IDB). The Integrated Database (IDB) [9] is a public (and free) dataset provided by the Federal Judicial Center (FJC) and the Administrative Office of the U.S. Courts (AOUSC) containing longitudinal case metadata (not documents) for criminal, civil, bankruptcy, and appellate cases (but not magistrate cases). While we do not know if it is intended, the IDB includes some very

⁴The endpoint is https://ecf.<district>.uscourts.gov/cgi-bin/possible_case_numbers.pl?<case_number>. For example, the District of Puerto Rico returns the generic error “Cannot find case,” whereas the District of South Carolina returns “Sealed v. Sealed. This case is SEALED.”

limited metadata on currently sealed cases and some metadata indicators on cases that were historically sealed but are now unsealed. These indicators give us the rare opportunity to revisit part of the FJC’s sealing-basis analysis, which originally relied on queries to court clerks [24], a privilege not available to us. However, the IDB is known to have gaps in coverage (compared with CM/ECF) [29], and information can occasionally be lost during quarterly updates [10].⁵ Furthermore, IDB’s more detailed metadata, combined with its historical snapshots, allow us to trace the history of a case’s sealing status over time, thereby enabling our analysis of historically sealed cases in Section 5.1. This analysis is not possible with CM/ECF.

CourtListener RECAP. Finally, we use the CourtListener RECAP service [14], a free and open-source project from the Free Law Project that crawls and archives CM/ECF data. CourtListener gets its information from a variety of sources, including crowdsourced contributions from users [17], monitoring individual district court RSS feeds for updates (for the districts that provide them) [16], and active scraping via PACER [11]. Our primary measurements use CourtListener case metadata and do not rely on its crowdsourced access to docket *content*. The LLM validation in Section 6 additionally uses available docket-entry descriptions, but not the underlying documents. In our experience, the quality of CourtListener’s data for a district depends directly on the configuration of the RSS feed the district provides (i.e., how comprehensive it is in the kinds of data it shares). In districts with comprehensive feeds, the coverage for unsealed cases is very good, so we can use the absence of a case number as a strong signal for case sealing (Section 4.2).

We validate our CourtListener current-sealing estimates by comparing them with CM/ECF. Moreover, because CourtListener is continuously updated, case unsealing can be implicitly revealed as a side effect: when a docket’s creation date in CourtListener is significantly later than its filing date, this delay is indicative of an unsealing action that made the case available to the public after some time under seal. This approach allows us to infer whether a case was sealed in the past (Section 5.2) and, if so, the length of the sealing period (Section 6). For historical criminal cases, we evaluate this timestamp signal against IDB.

We access the CourtListener data with a custom API client, using Python Requests [25] to query their REST API [18], while respecting their data rate limits and smoothing out queries over time to avoid overloading their servers.

Linking data sources. We link records using the district-relative Federal case number described above. IDB records typically embed an analogous representation of the case number⁶ and CourtListener provides a “Case Lookup” API that maps case numbers (plus court identifiers) to CourtListener’s native case ID representation.⁷

⁵For example, in some cases a previously-sealed indictment becomes unsealed yet remains marked with a sealed proceeding code even several years later, and some cases have an indictment unsealed code but no history of having the sealed code.

⁶The docket number field in IDB contains only the year and sequential number, omitting the office and case type (e.g., 2501234). While a separate field provides the office number, IDB does not provide the case type for civil cases. This omission created some challenges matching civil cases that had non-cv case types. An example is Multi-District Litigation (MDL) cases, which may have a different case type than the original case [41, 42].

⁷Note that CourtListener creates individual cases for criminal cases with multiple defendants, resulting in multiple CourtListener IDs for the same case number. Thus, we need to re-aggregate these records to avoid double counting.

4 Currently Sealed Cases

We start our analysis by characterizing cases that remained under seal in the selected district courts as of September 2025. CM/ECF’s Case Reports and its undocumented case-number endpoint directly expose current sealing status. When a Case Report redacts a case number, we construct candidate full case numbers and query the endpoint to resolve it and confirm current sealing status (Section 4.1). Because CM/ECF Case Reports are expensive for the public to generate at scale, we also infer sealed cases from freely-available CourtListener data. Comparing the two sources shows where our gap-based technique is reliable: it closely matches CM/ECF for districts and case types with adequate RSS coverage and consistent case-number allocation patterns.

4.1 Establishing Ground Truth by Crawling and Parsing CM/ECF Case Reports

Methodology. For each of the twenty districts, we use our custom crawler to request comprehensive “criminal reports” for criminal (cr) and magistrate judge (mj) cases and “civil reports” for civil (cv) cases, all filed between January 2020 and June 2025. In general, such reports include sealed cases,⁸ which are usually indicated by the case title “SEALED v. SEALED”, a partially redacted case number (i.e., “24-1234” instead of “1:24-cv-01234”), and/or the lack of a URL to the case’s docket sheet. We parse the resulting Case Reports to extract the case numbers, titles, and report-level indications of sealing. Moreover, many districts do not provide full case numbers for sealed cases, eliding the office and sometimes the case type. We therefore use the sequence of unsealed case numbers to infer the missing office and case type⁹ for each sealed case, allowing us to reconstruct candidate full case numbers for each sealed case. We then submit these candidates to CM/ECF’s undocumented per-court case-number endpoint. For a public case, the endpoint returns the full case number and name; for a sealed case, it instead returns a district-specific sealed-case response, which also lets us distinguish it from a non-existent case.

Results and Comparison with FJC Study. We use the original FJC study as a historical benchmark. It examined cases filed in 2006 that were still sealed two years later [24]; we therefore restrict our comparison to cases filed in 2023 that remained sealed in CM/ECF Case Reports two years later in 2025.¹⁰ We emphasize, however, that our contemporary sample contains 20 districts selected through research exemptions and validation needs, whereas the FJC study covers all 94 districts. Our comparison is therefore descriptive since it is not nationally comprehensive. From these reports, we identified 98 664 cases filed during this period, of which 14 488 (14.68%) are sealed, which we compare to the FJC data on a per-case-type basis.

For *civil* cases, the selected 2023 sample has a sealing rate of 0.27% (median of 0.19% across 20 districts), compared with 0.20%

(median of 0.14% across 94 districts) in the 2006 national statistics [24, p. 4]. For *criminal* cases, the selected 2023 sample has a sealing rate of 3.18% (median of 1.22% across 19 districts), compared with 1.6% (median of 0.96% across 94 districts) in the 2006 national statistics [24, p. 17]. This 2023 aggregate is highly sensitive to the D. of South Carolina, where requests for investigative criminal process are filed on the criminal rather than magistrate docket. Excluding that district reduces the 2023 selected-sample rate to 1.17%.

For *magistrate* cases, 58.6% of cases in the selected 2023 sample remained sealed two years after filing (median of 60.2% across 19 districts), compared with 16% (median of 20% across 94 districts) in the 2006 national statistics [24, p. 21], constituting a significant increase. One likely contributor to this difference is the composition of magistrate caseloads. Search-warrant applications comprised less than 10% of the federal magistrate caseload nationwide in 2006 and over 34% in 2023 [3, 4].

Per-district sealing rates. While the FJC study did not report sealing rates for each district, we were able to identify significant inter-district variations which signal institutional differences in sealing practices. Using the previous analysis of cases filed in 2023, Table 3 documents the per-district sealing rates (shown in the “ECF” column), further broken down by the three types of cases (cr, cv, and mj). *Civil* case sealing is largely consistent across districts, with every district except Puerto Rico and Nevada seeing a sealing rate below 1%. *Criminal* sealing is also typically close to its median (1.4%), and nearly always under 4%. The significant exception is the D. of South Carolina where, as mentioned above, search warrants and related criminal process are handled on the criminal docket, thus inflating its criminal sealing rate and proportionally shrinking its magistrate sealing rate (which at 8% is dramatically lower than other districts).

Finally, *magistrate* cases have the highest inter-district variation of all. For instance, while the W.D. of Louisiana and S.D. of California seal around 7% and 20% of their mj cases, respectively, the S.D. of New York seals more than 60% and the W.D. of Pennsylvania seals over 90%. We believe aspects of this inter-district difference may relate to differences in the volume of criminal legal process workload. For example, in the 12-month period ending in December 2023, search warrant issuance comprised a small fraction of the matters before W.D. Louisiana (16.75%) and S.D. California (17.73%) magistrates, but were substantially larger fractions of the caseload in S.D. New York (80.71%) and W.D. Pennsylvania (47.09%) [5]. However, this factor alone does not fully explain the observed variations, and it is known that local rules and practices play a significant role in sealing decisions [1, 19].

Underlying Sealing Causes. Since cases currently under seal are not publicly visible and their dockets cannot be interrogated directly, it is inherently challenging to validate speculations about the underlying reasons. However, a modest exception to this rule can be found by correlating with per-case metadata in IDB. In particular, of the 1 874 currently sealed *criminal* cases identified from CM/ECF, 547 (29.19%) also exist in IDB. Within this subset, 81.0% had a “Sealed Indictment” proceeding code, consistent with our expectation that indictment was a major cause of criminal sealing. Of the remainder, 7.9% had the “Juvenile/FJDA case” proceeding code (which are statutorily sealed [34]), 5.5% were “Felony with

⁸A key exception is the N.D. of Georgia, which only provides the ability to generate “External Reports”. The external criminal reports do not include sealed cases and do not allow us to list magistrate judge cases. Thus, we are only able to study its currently sealed civil cases.

⁹For example, we discovered that the criminal report may sometimes contain sealed mj (magistrate judge) cases beyond cr.

¹⁰In Appendix D we use the full set of CM/ECF data as a window into longitudinal unsealing behavior.

District	Civil (cv)			Criminal (cr)			Magistrate (mj)		
	ECF (%)	CL (%)	Diff (pp)	ECF (%)	CL (%)	Diff (pp)	ECF (%)	CL (%)	Diff (pp)
D. Puerto Rico	3.97	4.42	0.45	2.32	2.12	-0.20	73.81	72.95	-0.86
S.D. New York	0.19	0.19	0.01	2.71	4.98	2.27	60.42	73.83	13.41
W.D. Pennsylvania	0.21	0.21	0.00	0.00	0.29	0.29	92.30	90.20	-2.10
D. South Carolina	0.11	1.56	1.45	23.10	25.67	2.57	9.09	10.87	1.78
W.D. North Carolina	0.35	3.23	2.88	0.71	2.34	1.62	73.54	75.72	2.18
W.D. Louisiana	0.11	1.15	1.04	0.69	2.71	2.02	7.41	9.09	1.68
E.D. Louisiana [‡]	0.06	9.16	9.10	1.05	1.74	0.69	34.85	35.07	0.23
M.D. Louisiana	0.06	1.53	1.47	1.82	4.46	2.65	60.17	60.50	0.33
N.D. Texas	0.32	2.04	1.72	1.03	1.52	0.49	71.81	72.13	0.32
W.D. Michigan	0.00	1.90	1.90	2.84	3.93	1.09	75.75	77.81	2.07
N.D. Ohio	0.20	0.53	0.32	1.50	1.95	0.44	76.75	75.45	-1.31
S.D. Illinois [‡]	0.00	4.14	4.14	1.99	3.90	1.91	86.31	82.33	-3.98
S.D. Indiana	0.15	1.23	1.09	0.37	0.73	0.36	89.20	88.66	-0.54
S.D. California	0.34	0.34	0.00	0.75	0.61	-0.14	20.11	22.99	2.87
D. Nevada*	1.04	1.06	0.02	3.56	4.24	0.68	55.97	58.01	2.04
D. Oregon	0.36	1.81	1.45	2.42	3.36	0.94	12.73	12.33	-0.40
E.D. Washington	0.92	2.47	1.54	1.64	4.00	2.36	89.96	87.15	-2.82
N.D. Oklahoma*	0.18	0.53	0.35	1.22	1.22	0.00	51.83	51.83	-0.01
W.D. Oklahoma	0.17	0.25	0.08	0.38	0.75	0.37	46.71	46.95	0.24
N.D. Georgia*	0.25	0.35	0.10	–	2.30	–	–	69.26	–
Total	0.27	2.00	1.73	3.18	4.07	0.89	58.62	61.25	2.63

Table 3: Rates of currently sealed cases filed in 2023, by district and case type. ECF is the rate observed in CM/ECF Case Reports; CL is the rate estimated from gaps in CourtListener; with their differences in percentage points (pp). An asterisk * indicates suboptimal CourtListener coverage due to RSS feeds in 2023. [‡] indicates a mixed civil case-number space: the reported CL civil rates exclude known unsealed miscellaneous mc cases in E.D. Louisiana and Paraquat MDL pq cases in S.D. Illinois from both inferred sealed counts and civil denominators. Dashes denote unavailable CM/ECF values.

waiver of indictment”, 4.0% only had the generic “Indictment filed” proceeding code without any indication of being sealed, and 1.1% had the “misdemeanor information / rule 42(b) contempt notice” proceeding code. While the data provides us rare insight into the causes of sealing, we also caution that the inclusion of a sealed case in IDB may not be a representative sample of all currently sealed criminal cases since the detailed causes cannot be verified without access to the sealed cases themselves.

On the *civil* side, of the 1 036 currently sealed civil cases identified from CM/ECF, 902 (87.07%) also exist in IDB. Similar to criminal cases, civil cases have a Nature of Suit (NOS) code reflecting the underlying matter. Among these cases, 63.9% had their Nature of Suit listed as “(Qui Tam) False Claims Act” (again consistent with prior expectations due to the statutory sealing requirement for such cases [32]), 9.9% were coded as “Other Statutory Actions”, 7.6% had Prisoner-related Petitions (Mandamus, Civil Rights, Habeas Corpus, Vacate Sentence, etc.) and 3.3% had an NOS of “Other Forfeiture and Penalty Suits”. Sadly, IDB does not contain data about *magistrate* judge cases.

4.2 Inferring Currently Sealed Cases with Freely Available CourtListener Data

Having established current sealing observations with CM/ECF, we next develop a gap-based estimate using freely-available CourtListener data. We compare the two sources to evaluate our inference

methodology, and to understand the conditions under which the lower-cost estimate is reliable.

The premise behind our approach is to use *missing* case numbers in CourtListener data as an indication that a case is sealed. This heuristic assumes that case numbers are allocated predictably and that a sealed case will consume such a number, but be absent on CourtListener. This inference also requires sufficiently complete case-level CourtListener coverage. Because coverage varies with each district’s RSS configuration, we evaluate it empirically against CM/ECF and validate the separate timeliness requirement used by the historical analysis in Appendix C. Therefore, we have turned our problem of identifying sealed cases into a problem of identifying gaps in the sequences of case numbers that are available on CourtListener.

Identifying District Numbering Systems. Given the federated nature of the district courts, there is no guarantee that case numbers are allocated in the same way across all districts. Hence, our first step is to understand the case-numbering systems used by each district, and to identify their valid case-number ranges. For each district we study, we query CourtListener’s search interface for all case numbers with the pattern <year>-<type>, which matches all cases with the given year and case type, regardless of the office number or the sequence number. For example, querying 24-cv in txnd returns all civil cases filed in the N.D. of Texas in 2024, such as 1 : 24-cv-00001,

2: 24-cv-00002, etc. Using this approach, we queried for all cases filed from January 2020 to June 2025 across three case types (cr, cv, and mj), and used cases filed in 2023 to directly compare against CM/ECF results.

Using this data, we classify each district into one of two case-numbering systems. In districts that use a Unified numbering system, all offices share the same sequence of case numbers and, hence, case numbers should not overlap across offices (absent filing errors). In districts that use a Separate numbering system, each office maintains its own distinct sequence of case numbers and different offices may have case numbers that overlap (but are distinguished using different office numbers). As a result, we can categorize the numbering system used based upon the degree of case-number overlap across offices within the district. If there are negligible (less than 0.005%) overlapping case numbers across offices, we classify the district as using a Unified system, and otherwise we classify it as using a Separate system.¹¹ While some districts also have limited office overlap or share a sequence across case types, we treat these as district-specific caveats rather than a third numbering system and account for their effects on the gap estimates below.

In addition, we also identify allocated segments within numbering sequences. Districts commonly reserve certain ranges of case numbers for different purposes or offices. For example, the S.D. of California allocates case numbers 0xxxx for most criminal cases, but reserves 1xxxx for transferred supervised release cases. The E.D. of Washington allocates 00xxx for its Spokane office, 02xxx for Yakima, and 06xxx for Richland. Offices in these districts allocate case numbers sequentially within each segment, but the case-number gaps between segments correspond to non-existent cases. To properly consider such gaps as non-existent case numbers, we use a heuristic to detect segments by flagging gaps larger than a fixed threshold. We use an empirical threshold of 50 case numbers, treating a sequential run greater than that size as a boundary between allocated segments rather than as 50 or more sealed cases.

CourtListener Gap Methodology. Finally, having identified district case-numbering systems and the segments they use, we enumerate the possible case-number ranges spanning the identified segments and look for gaps. For districts with a Unified system, we construct a single sequence across the entire district and verify whether any office has filed a case with the corresponding number. For districts with a Separate system, we repeat the same process for each office separately. We infer that the gaps within established segments are likely sealed cases, and ignore the gaps between segments.

Results. Using this methodology, we identified 1 097 segments of continuous sequences across twenty districts, three case types, and five and a half years. These segments cover a total of 567 491 possible case numbers, among which CourtListener reported 455 360 as currently unsealed cases (it only has access to cases that are publicly available). CourtListener reports 2.35% fewer cases than our ground truth from the CM/ECF systems (461 181 currently unsealed cases).

¹¹As with all real systems, there can be complexity in how districts implement such case-number allocation. Notably, while the seven Separate districts have a significant percentage of overlapping numbers across offices, in some districts as high as 25%, the W.D. Louisiana and the D. Oregon have small percentages of overlapping numbers in their civil cases (3.09% and 1.32%, respectively), but not for criminal or magistrate judge cases. Hence, we also treat them as having a Unified numbering pattern, but indicate the anomalies with an asterisk in Table 2.

Most of this difference comes from the N.D. of Georgia and D. Nevada (before 2024), which do/did not provide any RSS feeds.

Validation of the CourtListener Estimate. We validate the CourtListener estimate against current sealing status ground truth observed in CM/ECF. We report both per-case agreement and differences in aggregate sealing rates because good agreement under one measure does not imply good agreement under the other. Both comparisons are conditional on CourtListener coverage and on the district's numbering practices. For the 17 districts that had RSS coverage in 2023, 89.04% of the *individual* currently-sealed cases identified from CM/ECF were *also* identified as sealed by our gap methodology applied to CourtListener data. If we only consider the 14 districts with more comprehensive RSS coverage (✓ and ○ in Table 2), the per-case agreement increases to 94.60%. For most district and case-type combinations, the estimated *aggregate* sealing rates from CourtListener (the “CL” column of Table 3) are close to the ground truth rates from CM/ECF. Here, all but two of the districts (which we discuss below) are within 3% of CM/ECF for civil cases. For criminal cases, all districts for which we have CM/ECF data are within 5%, and if we exclude the M.D. of Louisiana this tightens to 3%. Finally, for magistrate cases, all except the S.D. of New York and N.D. of Texas are within 5 percentage points,¹² with the vast majority within 3 percentage points.

The largest errors in our estimates coincide with identifiable institutional numbering practices that violate the assumptions of our gap-based methodology. Specifically, in the E.D. of Louisiana and S.D. of Illinois, civil cases share their number space with other case types. In the S.D. of Illinois, multidistrict litigation (MDL) cases, such as the Paraquat MDL case [41, 42] with case type pq, occupy a significant portion of the case numbers, sharing the same number space as normal civil cases despite having separate case types.¹³ On the other hand, the E.D. of Louisiana shares its civil case-number space with miscellaneous (mc) cases [36], which is also an uncommon practice among federal districts. These other case types create gaps in the sequence of civil cases despite not having the usual cv case type, leading our methodology to overestimate the number of currently sealed civil cases.

We therefore queried CourtListener for the known unsealed pq and mc cases and removed them from both the inferred number of sealed civil cases and the civil-case denominator. For the S.D. Illinois, removing 2 984 pq cases reduces the inferred rate from 74.01% (3 030 of 4 094) to 4.14% (46 of 1 110). For the E.D. Louisiana, removing 97 mc cases reduces the rate from 10.35% (769 of 7 432) to 9.16% (672 of 7 335). However, because mc cases are commonly sealed and thus not identifiable, we cannot completely remove them from our analysis. Thus, the corrected rate for the E.D. Louisiana still an overestimate. We report these corrected, civil-only rates in Table 3 and mark these rows with ‡.

¹²We do not yet have an explanation for these two outliers.

¹³In 2023 alone, 2 985 cases were filed under the pq case type. However, S.D. of Illinois only had 1 064 normal civil cases filed in the same year, causing numerous gaps in the sequence of civil case numbers that would be mistakenly inferred as sealed cases without special handling. Notably, other districts with a lot of MDL cases, such as N.D. of Ohio, do not have this issue as they typically use a separate number space for MDL cases. For example, N.D. of Ohio assigns op-4xxxx for cases under the Opioid MDL case [37] and sf-6xxxx for cases under the Suboxone MDL [38], which do not overlap with normal civil cases.

Overall, the comparison validates CourtListener and our gap-based methodology as a reliable and free estimate of currently sealed cases for districts with adequate RSS coverage. However, we caution that the method is not a full replacement for its sensitivity to institutional practices, such as the availability of comprehensive RSS feeds (known a priori [15]) and shared case-number spaces (which requires empirical investigation).

5 Historically Sealed Cases

One novel feature of our analysis, and a side effect of our established CourtListener methodology, is that we can go back in time beyond the current state of sealed cases and infer whether cases were historically sealed, even if they are now unsealed. In addition to CourtListener, we use quarterly snapshots of the Integrated Database (IDB) to identify historically sealed cases. IDB proceeding codes provide a validated subset of criminal cases, while the civil name heuristic provides a lower bound on cases detectable in the quarterly snapshots. We evaluate the CourtListener timestamp-based estimates against the IDB criminal labels and then apply it to criminal, civil, and magistrate cases.

5.1 Inferring Historically Sealed Cases with Special IDB Codes and Case Names

Methodology for Criminal Cases. To identify historically sealed criminal cases within IDB, we used the Proceeding Code (PROCCD) flag for each case. The field can take one of 27 possible values [9], and represents the status of the case at the time of the quarterly snapshot of IDB submitted by each court to the FJC [10]. We used seven proceeding codes that correspond to criminal cases with sealing-related actions. Proceeding codes 19 (Indictment unsealed), 23 (Misdemeanor information unsealed), and 25 (Felony information unsealed) correspond to unsealing of case information. Codes 22 (Sealed misdemeanor information), 24 (Sealed felony information with waiver of indictment), and 26 (Sealed indictment) correspond to the sealing of case information. Finally, code 9 (Juvenile proceeding) corresponds to juvenile cases, typically filed under seal [34].

However, we note that not all districts use these proceeding codes with the same regularity, and some not at all. For broad context, Appendix B shows the percentage of cases containing at least one sealing-related proceeding code for 94 federal districts over the past ten years. For the twenty districts in our study, all used sealing-related proceeding codes in at least 5% of their criminal cases over the past ten years. We take these results as a strong indication that the districts use the proceeding codes with some regularity, further supporting our use of IDB to identify their sealed criminal cases.

Methodology for Civil Cases. To identify historically sealed civil cases, we searched in IDB for civil cases where the plaintiff and defendant were both named “SEALED”. A limitation of this heuristic is that it only identifies civil cases that were still sealed at the time of the quarterly IDB snapshot submitted by each court to the FJC [10]. Consequently, the count and rate produced by this heuristic are lower bounds on historically sealed civil cases. For the results returned from IDB, both criminal and civil, we then identify and remove the currently sealed cases that we previously identified in Section 4 to focus exclusively on currently unsealed cases that were historically sealed.

Results of Historical Sealing Rates. Historically, as expected, sealing is far more common in criminal cases than civil cases. Employing the IDB proceeding code methodology on criminal cases in the twenty districts, we identified 10 477 (18.37%) out of 57 008 unsealed criminal cases filed between 2020 and 2025 as having been historically sealed. Similarly, employing the IDB sealed-name methodology on civil cases, we identified 278 (0.08%) out of 328 567 unsealed civil cases that were historically sealed over the same time period. We plot the historical sealing rates per district per year in Figure 1. The median criminal sealing rate among the twenty districts was 22.9%, and civil 0.05%. Unsurprisingly, districts with higher historical sealing rates among currently unsealed cases also have more currently sealed cases (e.g., W.D. Pennsylvania and D. South Carolina).

However, note that, unlike CM/ECF, neither IDB result is a complete census. The criminal rate describes the proceeding-code-labeled cases and may omit sealed cases lacking those codes. The civil rate is explicitly a lower bound because it captures only cases still sealed at a quarterly snapshot submitted by each court to the FJC [10]. In addition, while criminal cases have unsealing-related proceeding codes that continue to identify unsealed cases that were previously sealed, civil cases do not have such codes. Thus, the sealed-name methodology only identifies cases that were sealed at the time of the IDB snapshot.

Underlying Sealing Causes. Of the 10 477 historically sealed criminal cases identified in IDB, most of them (75.2%) used the “Indictment unsealed” proceeding code. Only 35.2% of them ever used the “Sealed indictment” code, an indication that IDB’s data is not complete with respect to all proceeding codes used in a case’s lifecycle. Overall, these causes are similar to the cohort of currently sealed criminal cases. However, civil cases show more differences.

On the civil side, of the 278 IDB-identifiable historically sealed cases, “(Qui Tam) False Claims Act” is the most common (32.3%) Nature of Suits (NOS), as expected. Interestingly, “Trademark” (31.7%) becomes the second most common, and represents a much higher percentage of historically sealed cases compared to those currently sealed. Some of this may be due to a practice, recently highlighted by the Electronic Frontier Foundation, in which intellectual property cases are directed to certain jurisdictions (e.g., the Eastern District of Texas) where judges are highly amenable to granting litigant motions to seal [19]. The other common (>2%) NOS codes include “Other Statutory Actions” (5.8%), “Patent” (4.7%), “Copyright” (3.2%), and various “Prisoner Petitions” (4.6%). These results are qualitatively consistent with the FJC report analyzing cases filed in 2006 [24], which also found (Qui Tam) False Claims Act cases to be the most common sealed civil category and discussed the same set of other case types as prominent reasons.

5.2 Inferring Historically Sealed Cases with CourtListener Timestamp Differences

CourtListener Timestamp Methodology. As mentioned in Section 3, since CourtListener is continuously updated for publicly available cases, we can use the timestamp differences between when a case was filed and when it was first recorded on CourtListener to infer whether a case was historically sealed but is now unsealed.

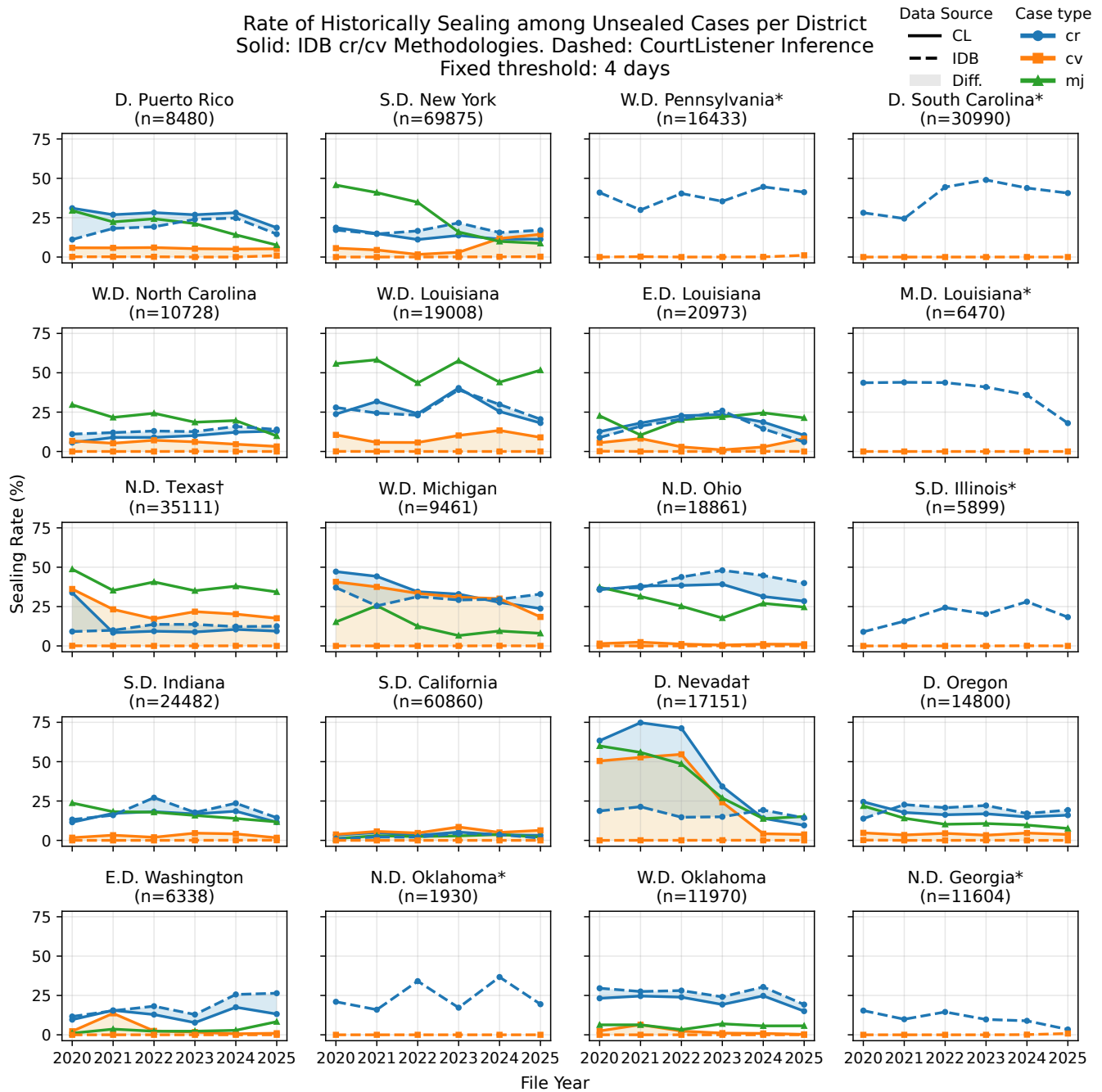


Figure 1: Percentage of currently unsealed cases that were historically sealed, as identified by our CourtListener and IDB methods. * indicates districts where CourtListener has suboptimal coverage due to lack of RSS feeds or empirical timeliness. We leave out their results as unreliable. † indicates districts that introduced RSS feeds during our study period, improving CourtListener’s timeliness.

This approach relies on two assumptions that our empirical results support. First, it assumes that if a case was historically sealed, then there is a significant delay between when the case was filed and when it was first made publicly available. Second, it assumes that CourtListener records cases within a short time after they are made publicly available. We evaluate these assumptions in Appendix C and categorize the CourtListener data reliability and timeliness for each district in the ‘CL’ column in Table 2.

Thus, if a case was filed on date D_f and was first recorded on CourtListener on date D_c , then we can infer that the case was likely sealed if $D_c - D_f > T$, where T is a threshold that we need to determine empirically. First, we query CourtListener for all cases as in Section 4.2 to get a list of all case numbers filed in our twenty districts from 2020 to 2025. We then compute the difference between the `date_filed` and `date_created` fields for each case. To determine a reasonable threshold T , we used the interquartile range (IQR) outlier detection method (commonly used in exploratory data analysis and the de facto standard for identifying outliers in box plots [31]), and chose $T = 4$ days as the threshold.¹⁴

Finally, we apply the chosen threshold to all cases in our dataset and identify those with a timestamp difference greater than the threshold as likely historically sealed cases.

Comparing Results from CourtListener and IDB. The overall quality of the CourtListener results depends heavily on the RSS coverage and timeliness of each district. As an aggregate sensitivity analysis, we first applied our CourtListener timestamp method to all three case types across the twenty districts, including the six districts that do not meet our coverage and timeliness requirements. This analysis identified 19 548 out of 67 709 (28.87%) currently unsealed criminal cases that were historically sealed, 90 305 out of 332 975 (27.12%) civil cases, and 15 529 out of 64 193 (24.19%) magistrate judge cases. These diagnostic values are significantly higher than the rates inferred from IDB and are heavily influenced by a few districts that violate our timeliness assumption (e.g., N.D. of Georgia and D. of South Carolina), which we accordingly exclude from our subsequent analysis.

If we only consider the fourteen districts that meet the timeliness requirement (marked by ✓ and ○ in Table 2, discussed in more detail in Appendix C), we identify 9 923 out of 53 867 (18.42%) criminal cases, 29 421 out of 232 789 (12.64%) civil cases, and 10 416 out of 57 904 (17.99%) magistrate judge cases. We plot these coverage-conditioned estimates for the fourteen districts in Figure 1 and omit the other six districts due to their lack of RSS feeds and suboptimal timeliness. We also plot the IDB criminal labels and civil lower bounds, and shade the area between them and the CourtListener timestamp estimates to show the difference. Across the fourteen districts, the median district historically sealed rate is 31.74% for criminal cases, 6.13% for civil cases, and 37.18% for magistrate cases, according to CourtListener timestamp estimates.

For *criminal* cases, the CourtListener timestamp method produces accurate estimates where coverage assumptions hold. We calculated the precision/recall scores for each of these districts and

each year from 2020 to 2025. These fourteen districts all have at least an 80% average precision in 2024, with ten districts maintaining this level for the past five years. Thus, the IDB-labeled subset validates the CourtListener timestamp method as a reliable estimator of historically sealed criminal cases, given sufficient RSS coverage. The two districts providing no RSS feeds (N.D. of Georgia and N.D. of Oklahoma) have performance no better than random guessing. We leave the full precision/recall curves for Figure 6 in Appendix E.

The importance of RSS coverage is further highlighted by correlations between longitudinal improvements in RSS coverage and improvements in the accuracy of our CourtListener results, as shown by the decreasing gap between the two methodologies in Figure 1. For example, D. Nevada shows a noticeably narrower gap between the estimate and IDB-labeled subset from 2023 (Average Precision of 0.37 in Figure 6) to 2024 (AP=0.89). This increase corresponds with the district transitioning from providing only a partial RSS feed for court orders in 2023 to providing a full RSS feed in 2024 [13]. Similarly, the N.D. of Texas also shows a significant improvement from 2020 (AP of 0.20) to 2021 (AP of 0.91), coinciding with when its RSS feed started including complaints and criminal charges [12].

In contrast, the *civil* historical sealing estimates are much less reliable because filing and data collection delays can resemble time under seal. When compared to IDB’s 0.08% lower bound, the CourtListener estimate of 12.64% is likely a significant overestimation. To diagnose this discrepancy, we used an LLM as a secondary aid to identify sealing-related text in available docket entries and manually checked a random sample of its outputs (Appendix F). This analysis is not ground truth: many districts do not document sealing and unsealing in docket entries, and CourtListener’s docket entries are often incomplete (see Section 6). For the 3 216 civil cases filed in 2023 that were identified as historically sealed by CourtListener but not by IDB, the LLM found no unsealing-related entry for 3 096 (96.27%) of them.

Although this result cannot establish that those cases were never sealed, the pattern is consistent with two known sources of delay unrelated to sealing. First, 1 883 (58.55%) of the 3 216 cases came from two districts whose partial RSS feeds omitted “petition,” “motion,” and “complaint” entries [15]. Spot checks confirmed that their docket entries were missing some earlier entries related to civil complaints, which may explain why they were picked up by CourtListener days later than their filed date. Second, 1 332 (41.43%) had a Nature of Suit associated with prisoner petitions (including Habeas Corpus, Prison Condition, Vacate Sentence, Mandamus, etc.). These cases are typically filed by inmates through prison mail, and are considered filed when they are handed to the prison authorities per the “mailbox rule” established by the Supreme Court [47]. Moreover, since these cases are often mailed in, they can take days to reach the court, before being processed by the clerk. We accordingly observed that many of these cases were delayed in being added to CM/ECF and then backdated to their filed date. These delays can therefore cause the CourtListener timestamp method to misclassify civil cases as historically sealed.

Despite these limitations, CourtListener provides two additional benefits when compared to IDB. First, IDB contains only criminal and civil cases, but CourtListener includes magistrate cases as well, which are more commonly sealed. Second, IDB provides only a binary label of whether a case was historically sealed, whereas

¹⁴ Given the subset of historically sealed criminal cases derived from the proceeding codes in IDB, we can also choose the threshold automatically by maximizing the F1 score [27]. We experimented with this methodology and the results were very similar. Thus, we used a fixed threshold for simplicity.

CourtListener allows us to infer when the case was unsealed. This feature makes CourtListener particularly useful for our next stage of analysis to approximate duration under seal for the historically and currently sealed criminal and civil cases.

6 Duration Under Seal for Currently and Historically Sealed Cases

In our final analysis, we combine our data sources and analysis methodologies to answer a different, but perhaps more important question about sealing practice: if sealed, how long do such cases remain unavailable to the public? To the best of our knowledge, ours is the first large-scale analysis of federal sealing duration.

We analyze criminal and civil cases filed between January 2020 and June 2025 that are either labeled as historically sealed in IDB or observed as currently sealed in CM/ECF as of September 2025. Current sealing status is directly observed in CM/ECF. The historical criminal cases are from IDB proceeding code labels, the historical civil cases from IDB name-based lower-bound, and all unsealing times approximated from CourtListener timestamps. For historically sealed cases, we approximate duration under seal as the timestamp difference between the date filed and the date it appears in CourtListener. For cases that remain currently sealed, we calculate the duration under seal as the difference between the date filed and September 2025, which is the date of our analysis. We treat currently sealed cases as right-censored data since we only know their current duration under seal but not when they will be unsealed (if ever). Then, using the Kaplan-Meier estimator [21], we calculate survival curves for these currently or historically sealed cases and plot them in Figure 2. The survival curves show the probability (y -axis) that a case in each district remains sealed after a certain amount of time since it was filed (x -axis), with right-censored cases marked by '+' on the curves for their current durations.

Durations Under Seal Differ for Criminal and Civil Cases. The results show that unsealing practices differ significantly between criminal and civil cases. While there are many more sealed *criminal* cases, they are unsealed much faster. The median duration under seal for criminal cases is less than 30 days for all districts with good CourtListener coverage (i.e., districts with no '*' in Figure 2), with some districts having a median duration under seal of less than 10 days (D. Puerto Rico, S.D. New York, and W.D. Michigan).

On the other hand, *civil* cases stay sealed for much longer durations. In the majority of districts, more than half of the civil cases still remain under seal. Of the 1 172 historically sealed civil cases we identified, 842 (71.84%) are still sealed as of September 2025. Among the remaining civil cases that were eventually unsealed, in nearly all districts, a majority of them were sealed for more than three years. However, we note that our IDB name-based methodology is only a lower bound for historically sealed civil cases, and thus may bias our results towards civil cases sealed for longer durations. While these results are based on a small population size of sealed civil cases, they suggest that, while civil cases are rarely sealed, when they are, they are likely to remain sealed for a long time.¹⁵

¹⁵ A few civil cases appear to be unsealed very quickly on the same day they were filed. Upon manual inspection, many of these cases were actually sealed for a much longer time (and many are still sealed today). However, they were usually filed first in CM/ECF as normal cases with public access but then became sealed through motions

This analysis changes how we should think about sealing practices. A high sealing rate does not necessarily imply long-term secrecy if most cases are unsealed quickly. Conversely, a low sealing rate with long durations can still pose a significant barrier to transparency, as they may prevent public access to information perhaps for years or indefinitely.

Validation. Because no ground truth for duration under seal exists, we use manual docket review as a validation subset and an LLM as a secondary, larger-scale extraction aid. We first randomly sampled 5 criminal and up to 5 civil cases from each district (177 cases total) and read their docket entries on CM/ECF to identify the actual unseal date, if any. Of the 177, we found that 22 cases were still sealed, all of them civil cases. Moreover, among the remaining 155 unsealed cases, we were only able to find the exact unsealing date for 75% of criminal and 49% of civil cases. Among the manually reviewed cases for which we found an exact unsealing date, we compared the reviewed duration to our CourtListener timestamp estimate. In this validation subset, the median error was 1.64 days for criminal cases and 6.85 days for civil cases.

For a larger secondary comparison, we used an LLM to parse all currently unsealed cases. Specifically, we provided GPT-4.1 (gpt-4.1-2025-04-14) with all CourtListener docket entries available for each evaluated case. The prompt instructed the model to extract the unsealing date, if present, from currently unsealed criminal and civil cases (Appendix F describes the full prompts and indicator rankings in more detail). The LLM identified an unseal date for 1 961 (18.5%) of the criminal cases and 178 (53.9%) civil cases. For cases where the LLM found an unsealing date, the extracted and CourtListener-estimated durations have a median difference of 0.68 days for criminal cases, and only 1.51 days for civil cases.¹⁶

Limitations of LLM Parsing. The LLM was useful for extracting sealing and unsealing events when the relevant docket entries were available, but it could not recover omitted events. This limitation reflects the underlying data practices: the usage and inclusion of sealing and unsealing information in docket entries varies significantly across districts. Notably, S.D. Illinois and E.D. Washington never include unsealing events for criminal cases, and many districts do not include explicit unsealing orders on civil cases. Even when they do, the language varies case by case. Moreover, partial-unsealing events can make the unseal date ambiguous and require additional context for the LLM. Retrieving and processing docket entries also requires substantially more pages of data transfer and thus cost from CM/ECF, whereas CourtListener's docket entries are freely available but often incomplete. We therefore use the LLM as a validation where district practices support it, rather than a reliable methodology. CourtListener's first seen timestamps provide a distinct, more scalable signal, but the resulting duration measurements remain approximations subject to the coverage and timestamp limitations discussed above.

to seal the case. Due to these districts providing up-to-date RSS feeds, CourtListener picks up these cases very quickly after they are filed but before they are sealed. As a result, our timestamp method underestimates the time these cases were under seal, and slightly (< 0.95%) underestimates the number of currently sealed cases as well.

¹⁶ Most differences between the LLM and manual review are due to the existence of multiple unsealing events for the same case (e.g., partial unsealing to the defendant), which makes it hard to identify the exact unseal date for the entire case.

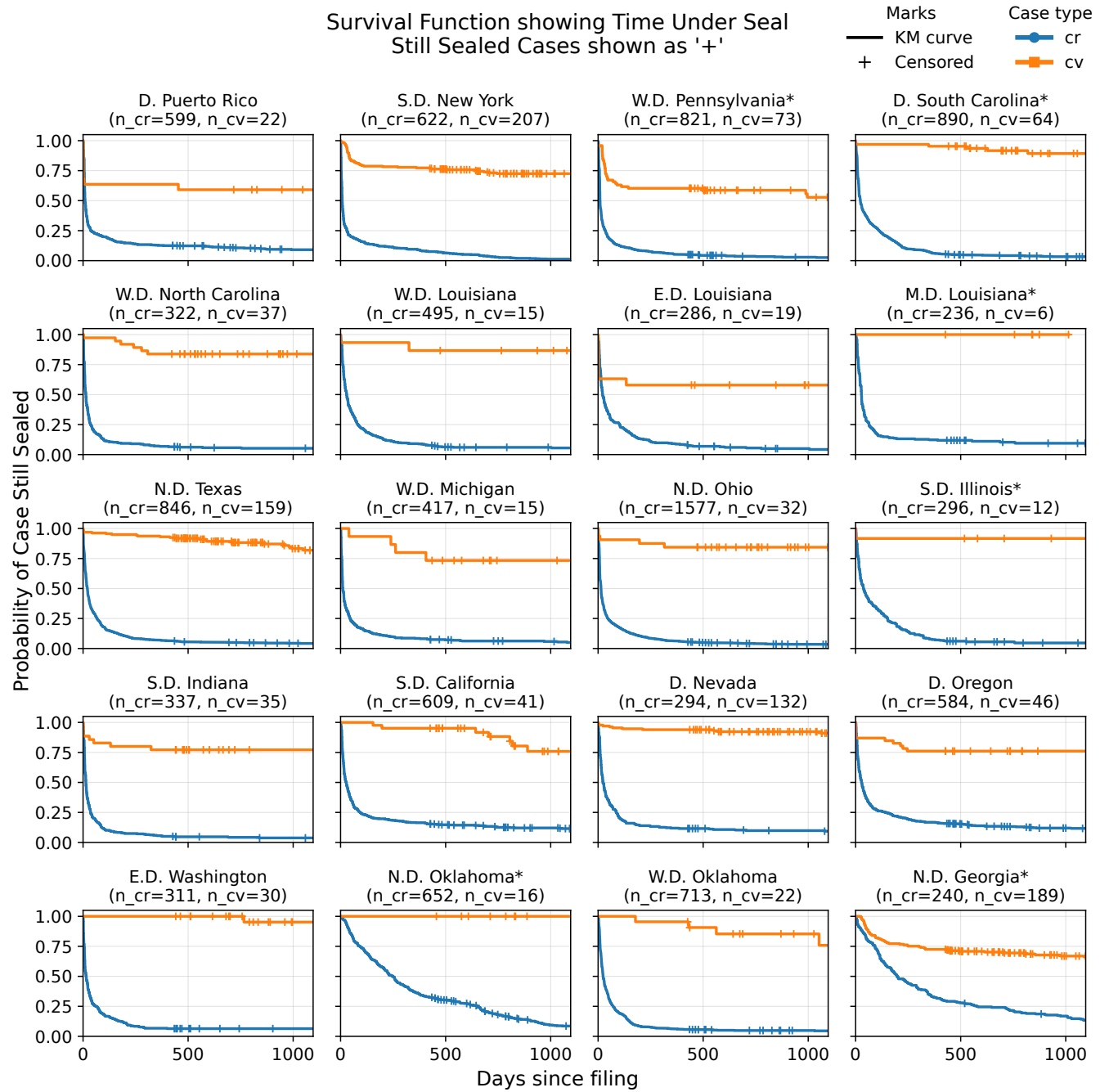


Figure 2: Survival curves of historically sealed cases identified using IDB, with current sealing status from CM/ECF and historical duration approximated using CourtListener timestamps. A * indicates a district where CourtListener has suboptimal coverage due to lack of RSS feeds or failing empirical timeliness tests. The results for those districts should be interpreted with caution.

7 Discussion

Our paper provides four classes of contributions. The first is to show how unprivileged researchers can use inference *alone*, along with the existing CM/ECF Web interface, to identify and enumerate currently sealed Federal district court cases. Using this approach we provide the first contemporary analysis of U.S. Federal case sealing in seventeen years. We show that while the sealing rates of civil and criminal dockets are not significantly changed since that time, there have been substantive increases in magistrate docket sealing rates that bear further study. While some of this latter issue may reflect increases in criminal process issuance — between 2006 and 2023, the number of search warrants issued by U.S. magistrates has grown by a factor of *five* (from 32 467 [3] to 162 433 [4]) — this factor alone is not sufficient to explain the increase nor its substantial inter-district variability.

Our second contribution is about cost. While CM/ECF data represents the “gold standard” for Federal district court data, at ten cents per “page” of data it remains costly to use at scale. While there have been multiple bills in the U.S. Congress to make PACER free to the public, all have so far failed to pass. However, in this work we have shown that for most districts, our inference techniques combined with the freely available (and easier to use) CourtListener data are sufficient to explore many questions of case sealing with only minor error compared to using the more costly CM/ECF data. We believe this finding should help democratize the study of case sealing, and we expect that the existing error will only be reduced as CourtListener improves and more districts adopt full RSS feeds.

Third, using the data available to us, we establish new longitudinal inference techniques that allow us to identify cases that were sealed in the past and, for the first time, the *duration* of their sealing period. These analysis techniques fill a blank in the literature, as indications of sealing and unsealing events are not systematically recorded in the public record, and thus, to the best of our knowledge, no prior work has attempted to estimate the scale of historical sealing or the duration of sealing orders. Our analysis is only made possible by the combination of continuous and comprehensive crawling of CM/ECF data by CourtListener, and our development of side-channel inference techniques. These new data sources and analysis techniques allow us to understand the practices of judicial secrecy and dynamics of sealing in a more nuanced way, identifying the extent to which sealing is used as a temporary procedural tool versus a long-term barrier to public access.

Finally, our last analysis highlights a concern about sealing in general — a measure only meant to supersede the public right of access when there are “compelling interests” at stake. Our data shows that civil cases, once sealed, are likely to stay sealed — frequently for years — raising the suspicion that these may not reflect truly compelling interests. This seems to be particularly likely since these appear to predominantly be False Claims Act cases, which have no broad remit for long-term open-ended sealing. To quote one of the authors of the original FJC study, “Closed False Claims Act cases ordinarily should not be sealed” [23]. Indeed, the problem appears to be that Federal sealing is a system that “fails open”. If a failure to unseal occurs, it is extremely hard to correct, as few have the standing or means to do so. This structural challenge argues for general sealing rules that require affirmative resealing to maintain

secrecy over time, and to ensure that there remain truly compelling interests at stake.

Acknowledgments

We thank our anonymous reviewers for their insightful suggestions and feedback.

As well, this work was only possible due to datasets, access and guidance provided by others who work with Federal case data. Notably, we are grateful to the various district courts who granted us PACER research exemptions, without which our research would have been cost prohibitive. In the same vein, we are deeply appreciative of the Free Law Project’s choice to make CourtListener available for free to all educational researchers. In particular, we would like to thank Michael Lissner and Eri O’Diah of the Free Law Project for facilitating access to the CourtListener REST API and for answering our questions about CourtListener’s data-collection methods, RSS coverage, and historical changes in crawler behavior. Finally, Seamus Hughes of Court Watch provided us with free training and advice on various aspects of the PACER system and highlighted to us the variable nature of how different district courts file matters.

Outside data collection, we offer our thanks to Cindy Moore and Jennifer Folkestad for operational and administrative support of our research. Funding for this work was provided by the Irwin Mark and Joan Klein Jacobs Chair in Information and Computer Science, the CSE Professorship in Internet Privacy and/or Internet Data Security, and operational support from the UCSD Center for Networked Systems.

References

- [1] Heather Abraham, Alex Abdo, and Jonathan Manes. 2021. Suggestion 21-CV-T. Administrative Office of the U.S. Courts. <https://www.uscourts.gov/forms-rules/records-rules-committees/suggestions/heather-abraham-alex-abdo-and-jonathan-manes-21-cv-t>.
- [2] Administrative Office of the U.S. Courts. 2001. Public Access to Court Electronic Records | PACER: Federal Court Records. <https://pacer.uscourts.gov/>.
- [3] Administrative Office of the U.S. Courts. 2006. Table M-3—U.S. District Courts—U.S. Magistrate Judges Judicial Business (September 30, 2006). <https://www.uscourts.gov/data-news/data-tables/2006/09/30/judicial-business/m-3>.
- [4] Administrative Office of the U.S. Courts. 2023. Table M-3—U.S. District Courts—U.S. Magistrate Judges Judicial Business (September 30, 2023). <https://www.uscourts.gov/data-news/data-tables/2023/09/30/judicial-business/m-3>.
- [5] Administrative Office of the U.S. Courts. 2023. Table M-3—U.S. District Courts—U.S. Magistrate Judges Statistical Tables For The Federal Judiciary (December 31, 2023). <https://www.uscourts.gov/data-news/data-tables/2023/12/31/statistical-tables-federal-judiciary/m-3>.
- [6] Administrative Office of the U.S. Courts. 2025. Statistical Tables for the Federal Judiciary - June 2025. <https://www.uscourts.gov/data-news/reports/statistical-reports/statistical-tables-federal-judiciary/statistical-tables-federal-judiciary-june-2025>.
- [7] Administrative Office of the U.S. Courts. [n. d.]. Electronic Filing (CM/ECF). <https://www.uscourts.gov/court-records/electronic-filing-cm-ecf>.
- [8] David Freeman Engstrom. 2012. Harnessing the Private Attorney General: Evidence from Qui Tam Litigation. *Columbia Law Review* 112, 2012 (2012), 1244–1325. <https://papers.ssrn.com/abstract=2290847>.
- [9] Federal Judicial Center. 2025. Integrated Database (IDB). <https://www.fjc.gov/research/idb>.
- [10] Federal Judicial Center. [n. d.]. The Integrated Database: A Research Guide. <https://www.fjc.gov/sites/default/files/IDB-Research-Guide.pdf>.
- [11] Free Law Project. 2009. Freelawproject/Juriscraper: An API to Scrape American Court Websites for Metadata. <https://github.com/freelawproject/juriscraper>.
- [12] Free Law Project. 2020. Help with Alerts – CourtListener.Com. <https://web.archive.org/web/20200305003115/https://www.courtlistener.com/help/alerts/> (archived 2020-03-05).
- [13] Free Law Project. 2023. Help with Alerts – CourtListener.Com. <https://web.archive.org/web/20230323193536/https://www.courtlistener.com/help/alerts/>.

- (archived 2023-03-23).
- [14] Free Law Project. 2025. RECAP Archive. <https://www.courtlistener.com/recap/>.
 - [15] Free Law Project. [n. d.]. Help with Search and Docket Alerts – CourtListener.Com. <https://www.courtlistener.com/help/alerts/>.
 - [16] Free Law Project. [n. d.]. RECAP Archive Coverage – CourtListener.Com. <https://www.courtlistener.com/help/coverage/recap/>.
 - [17] Free Law Project. [n. d.]. RECAP Suite – Turning PACER Around Since 2009. <https://free.law/recap/>.
 - [18] Free Law Project. [n. d.]. REST API – v4.3. <https://www.courtlistener.com/help/api/rest/>.
 - [19] Betty Gedlu. 2026. EFF Challenges Secrecy In Eastern District of Texas Patent Case. <https://www.eff.org/deeplinks/2026/04/eff-challenges-secrecy-eastern-district-texas-patent-case>.
 - [20] Elizabeth Goitein. 2016. *The New Era of Secret Law*. Technical Report. Brennan Center for Justice at New York University School of Law. https://www.brennancenter.org/media/226/download/Report_The_New_Era_of_Secret_Law_0.pdf.
 - [21] E. L. Kaplan and Paul Meier. 1958. Nonparametric Estimation from Incomplete Observations. *J. Amer. Statist. Assoc.* 53, 282 (June 1958), 457–481.
 - [22] Tiffany Li. 2025. Government Control over Qui Tam Suits and Separation of Powers. *Yale Journal on Regulation* 42, 1 (2025), 383–448.
 - [23] Robert Timothy Reagan. 2010. *Sealing Court Records and Proceedings: A Pocket Guide*. Technical Report. Federal Judicial Center. https://www.fjc.gov/sites/default/files/2012/Sealing_Guide.pdf.
 - [24] Tim Reagan and George Cort. 2009. *Sealed Cases in Federal Courts*. Technical Report. Federal Judicial Center. <https://www.fjc.gov/content/sealed-cases-federal-courts-0>.
 - [25] Kenneth Reitz. 2026. Requests: A simple, yet elegant, HTTP library. <https://github.com/psf/requests>.
 - [26] Leonard Richardson. 2025. BeautifulSoup4: Screen-scraping Library. <https://www.crummy.com/software/BeautifulSoup/bs4/>.
 - [27] C. J. Van Rijksbergen. 1979. *Information Retrieval*. Butterworths, London.
 - [28] Elizabeth A. Rowe. 2024. Judicial Secrecy. Virginia Public Law and Legal Theory Research Paper No. 2025-36 (SSRN). <https://ssrn.com/abstract=5154667>.
 - [29] Scales. 2021. Crosswalking PACER to the IDB. <http://livingreports.scales-okn.org/#/idbCrosswalkReport>.
 - [30] Stephen W. Smith. 2012. Gagged, Sealed & Delivered: Reforming ECPA’s Secret Docket. *Harvard Law & Policy Review* 6, 2012 (May 2012), 313. <https://papers.ssrn.com/abstract=2071399>.
 - [31] John Wilder Tukey. 1977. *Exploratory Data Analysis*. Addison-Wesley Publishing Company, Reading, MA, USA.
 - [32] United States Code. 2023. 31 U.S. Code § 3730 - Civil Actions for False Claims. <https://www.govinfo.gov/app/details/USCODE-2023-title31/USCODE-2023-title31-subtitleIII-chap37-subchapIII-sec3730>.
 - [33] U.S. Court of Appeals for the D.C. Circuit. 2020. In the Matter of the Application of Jason Leopold to Unseal Certain Electronic Surveillance Applications and Orders. 964 F.3d 1121 (D.C. Cir. 2020). <https://www.courtlistener.com/opinion/4766181/in-the-matter-of-the-application-of-jason-leopold-to-unseal-certain/>.
 - [34] U.S. Department of Justice. 2018. Justice Manual: 9-8.000 - Principles of Federal Juvenile Prosecution. <https://www.justice.gov/jm/jm-9-8000-juveniles>.
 - [35] U.S. District Court for the District of Columbia. 2018. In the Matter of the Application of Jason Leopold to Unseal Certain Electronic Surveillance Applications and Orders. 300 F. Supp. 3d 61 (D.D.C. 2018). <https://www.dcd.uscourts.gov/sites/dcd/files/13mc712.pdf>.
 - [36] U.S. District Court for the Eastern District of Louisiana. 2025. Sealed Case - Unavailable for Public Access. No. 2:25-mc-00010, (E.D. La.).
 - [37] U.S. District Court for the Northern District of Ohio. 2017. In Re: National Prescription Opiate Litigation. No. 1:17-md-02804, (N.D. Ohio 2017). <https://www.courtlistener.com/docket/62647370/in-re-national-prescription-opiate-litigation/>.
 - [38] U.S. District Court for the Northern District of Ohio. 2024. In Re: Suboxone (Buprenorphine/Naloxone) Film Products Liability Litigation. No. 1:24-md-03092, (N.D. Ohio. 2024). <https://www.courtlistener.com/docket/68224010/in-re-suboxone-buprenorphinenaloxone-film-products-liability-litigation/>.
 - [39] U.S. District Court for the Southern District of California. 2024. United States v. 7413 Canyon Breeze Drive, San Diego, CA (“Target Location”). No. 3:24-mj-03851 (S.D. Cal. 2024). <https://www.courtlistener.com/docket/69248609/united-states-v-7413-canyon-breeze-drive-san-diego-ca-target/>.
 - [40] U.S. District Court for the Southern District of California. 2024. United States v. 7413 Canyon Breeze Drive, San Diego, California. No. 3:24-mj-03648 (S.D. Cal. 2024). <https://www.courtlistener.com/docket/69203105/united-states-v-7413-canyon-breeze-drive-san-diego-california/>.
 - [41] U.S. District Court for the Southern District of Illinois. 2021. In Re: Paraquat Products Liability Litigation v. Syngenta Crop Protection, LLC. No. 3:21-md-03004 (S.D. Ill. 2021). <https://www.courtlistener.com/docket/59988561/in-re-paraquat-products-liability-litigation-v-syngenta-crop-protection/>.
 - [42] U.S. District Court for the Southern District of Illinois. [n. d.]. Attorney Instructions for Opening Paraquat Cases in Re: MDL 3004. <https://www.ilsd.uscourts.gov/sites/ilsd/files/AttorneyInstructionsForMDL3004.pdf>.
 - [43] U.S. District Court for the Southern District of New York. 2023. Southern District Court Announces Electronic Filing of Criminal Applications And Pro Se Filing by Email. <https://nysd.uscourts.gov/sites/default/files/2023-07/Notice%20to%20Bar%20-%20ECF%20Rules%20Revision%20-%2020072423.pdf>.
 - [44] U.S. Supreme Court. 1978. Nixon v. Warner Communications, Inc. 435 U.S. 589, 596-97 (1978). <https://www.courtlistener.com/opinion/109830/nixon-v-warner-communications-inc>.
 - [45] U.S. Supreme Court. 1980. Richmond Newspapers, Inc. v. Virginia. 448 U.S. 555 (1980). <https://www.courtlistener.com/opinion/110339/richmond-newspapers-inc-v-virginia>.
 - [46] U.S. Supreme Court. 1984. Waller v. Georgia. 467 U.S. 39 (1984) <https://www.courtlistener.com/opinion/111186/waller-v-georgia/>.
 - [47] U.S. Supreme Court. 1988. Houston v. Lack. 487 U.S. 266 (1988). <https://www.courtlistener.com/opinion/112126/houston-v-lack>.

A Ethics

Our study analyzes public U.S. federal court records drawn from multiple sources. While querying the public facing CM/ECF system and CourtListener RECAP archive, we have respected their API rate limits and scheduled our data collection to minimize load on their systems and avoid disruption to other users.

Although these records are publicly accessible, linking and publishing them may potentially increase privacy harms by improving discoverability and enabling re-identification for individuals involved in these cases and their personally identifiable information, such as names, addresses, financial information, immigration status, or other sensitive information. We therefore do not collect or disseminate party names and court documents, and restricted our analysis to case-level metadata, to respect the practical obscurity and avoid revealing sensitive information about individuals involved in these cases. We also consider downstream misuse (e.g., harassment or discriminatory screening) and design our dataset release to limit such risks while preserving accountability and transparency.

Furthermore, we provide only aggregate statistics and respect court orders on matters that remain sealed or restricted to avoid harming the interests of involved parties. We report results primarily in aggregate and avoid publishing or citing examples for currently or historically sealed matters to prevent potential harms from re-identification and avoid singling out individuals or entities. We also do not attempt to identify or profile parties involved in sealed cases. On encountering cases or case documents that are later sealed during our data collection, we ceased further data collection to respect the court’s sealing order and suppressed naming such cases in our analysis.

B Reliability of IDB Flags

To establish context, we calculated the percentage of criminal cases in the IDB that use sealing-related proceeding codes over the past ten years for the 94 U.S. lower district courts. Figure 3 shows the detailed results as a heat map. Some district courts, such as the District of Rhode Island and the Northern District of Virginia, never use any sealing-related proceeding codes. For such districts, our IDB-based methodology for identifying historically sealed cases (Section 5.1) will not apply. Also immediately apparent is that the District of North Dakota routinely seals nearly all of its criminal cases. A small random sample of criminal cases from this district suggests that these cases indeed contain sealing and unsealing events in their docket entries, indicating that the proceeding codes are being used correctly in this district.

In our analysis, we specifically focus on the 20 districts that used sealing-related proceeding codes in at least 5% of their criminal cases over the past five years. We believe this is a reasonable threshold to ensure that they use the proceeding codes with some regularity, while still including a diverse set of districts.

C Timeliness of CourtListener’s Data and District Categorization

The accuracy of our CourtListener inference methods assumes that CourtListener’s case-level coverage is sufficiently complete and timely. CourtListener draws its data from a wide range of sources, such as RSS feeds provided by courts, crowdsourcing through the

RECAP browser extension, and their own Web scraping efforts [11, 16]. Our data suggest that it performs well at case-level coverage relative to the CM/ECF Case Reports used in this comparison. This does not imply complete docket entries for those cases; some feeds omit early entries, as discussed further in Appendix F.

Our raw data from CourtListener contains 464 877 cases filed in the twenty districts from January 2020 to June 2025.¹⁷ When comparing against CM/ECF data over the past 5.5 years, CourtListener found all except 143 of them (99.97%).

However, the timeliness requirement is a higher bar to meet, as it requires that CourtListener’s data is updated shortly after a case is filed on CM/ECF. We hypothesize that the timeliness of CourtListener’s data is highly dependent on the availability and reliability of RSS feeds provided by courts, as they are among the earliest methods for CourtListener to obtain new cases and are deeply integrated into their alert pipeline [15]. Therefore, we categorize the districts we studied based on their use of RSS feeds. As of September 2025, ten out of the twenty districts provide full RSS feeds for all types of events and filings, eight provide partial RSS feeds for a subset of events, and only two do not provide any RSS feeds at all [15].

Furthermore, we also empirically evaluated the timeliness of CourtListener’s data by comparing the filing dates of cases with the dates they were first available in CourtListener, which we term the delay time. We plot the distribution of delay times for criminal, civil, and magistrate cases per district per year in Figure 4. Across all twenty districts and all years, the median time between a case being filed and being added to CourtListener is 36 hours for criminal cases, 20 hours for civil cases, and 18 hours for magistrate cases. We therefore use a conservative threshold of four days to rate the timeliness of districts.

Across all years, fourteen districts have median delay times of less than four days for all three case types, of which nine have full RSS feeds and five have partial RSS feeds. We mark them respectively with ✓ and ○ in Table 2. Examining the districts with partial feeds in more detail, we discover that these RSS feeds typically include complaint and criminal charge events, which are among the earliest filings within new cases. One district with full RSS coverage and three with partial RSS coverage fail the four-day test, as do the two districts without an RSS feed, N.D. of Georgia and N.D. of Oklahoma. We mark these districts with an ✗ in Table 2.

D Visualization of Currently Sealed Cases across Districts and Years

While Section 4 presented the rate of currently sealed cases for those filed in 2023, we have also generated CM/ECF Case Reports and queried CourtListener for all cases filed from 2020 to 2025 for the twenty districts we studied. Employing the same methodology as in Section 4, we plot the sealing rates per district for cases filed in 2020 to 2025 in Figure 5.

There are many interesting longitudinal changes and lack thereof. While we had previously expected the rate of sealing to increase for cases filed in more recent years, as older cases become unsealed, the effect was much less pronounced than we had anticipated. For

¹⁷Our analysis in Section 4 uses data after preprocessing, which removes cases with same sequence numbers in districts deemed to have Unified or Unified* numbering systems according to Table 2.

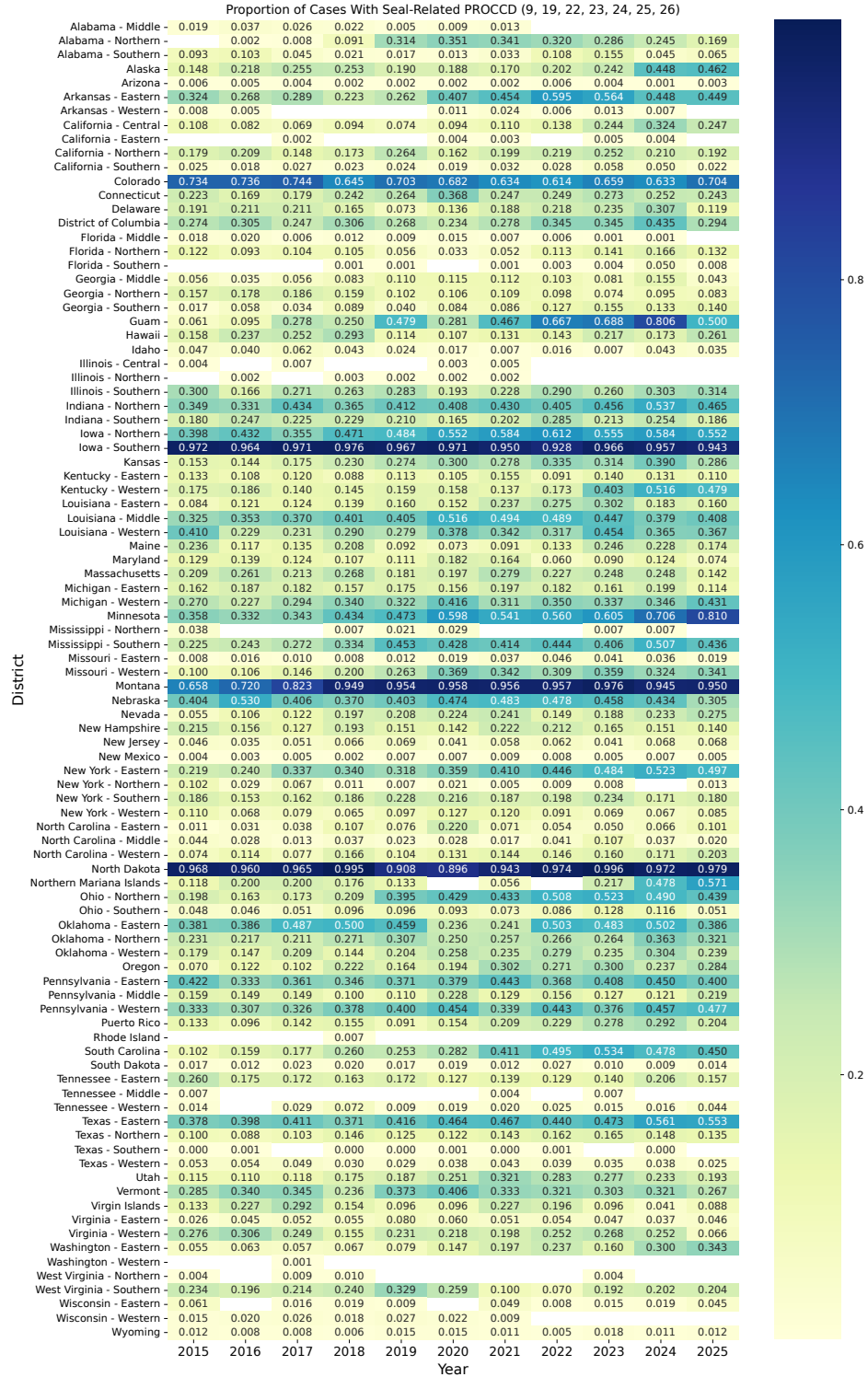


Figure 3: Percentage of criminal cases in the IDB that use sealing-related proceeding codes in the past ten years. Empty cells indicate that there were no cases marked with a sealing-related proceeding code for that district and year.

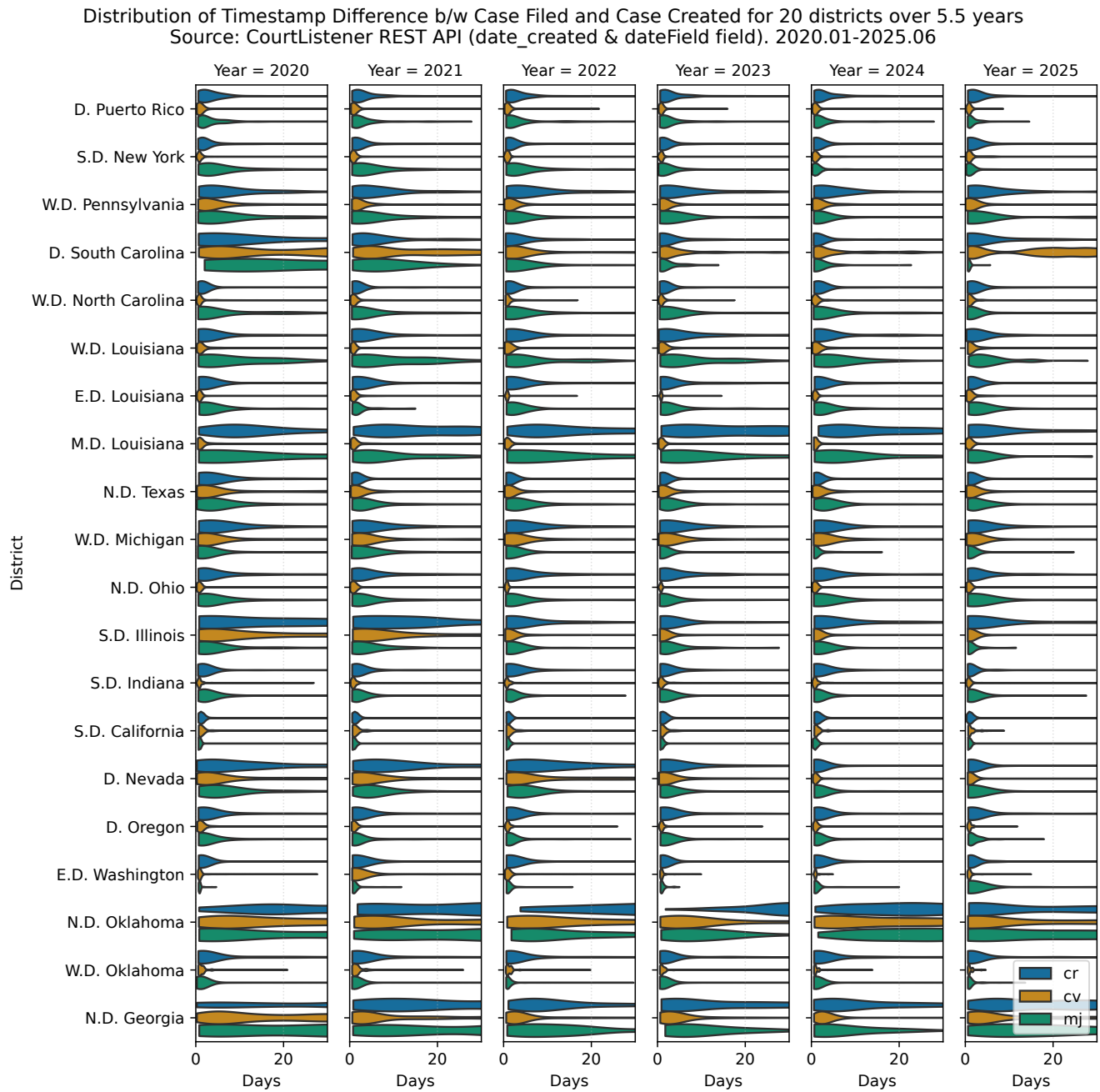


Figure 4: Time delay between case filing and case availability on CourtListener for cases in the twenty districts from 2020 to 2024.

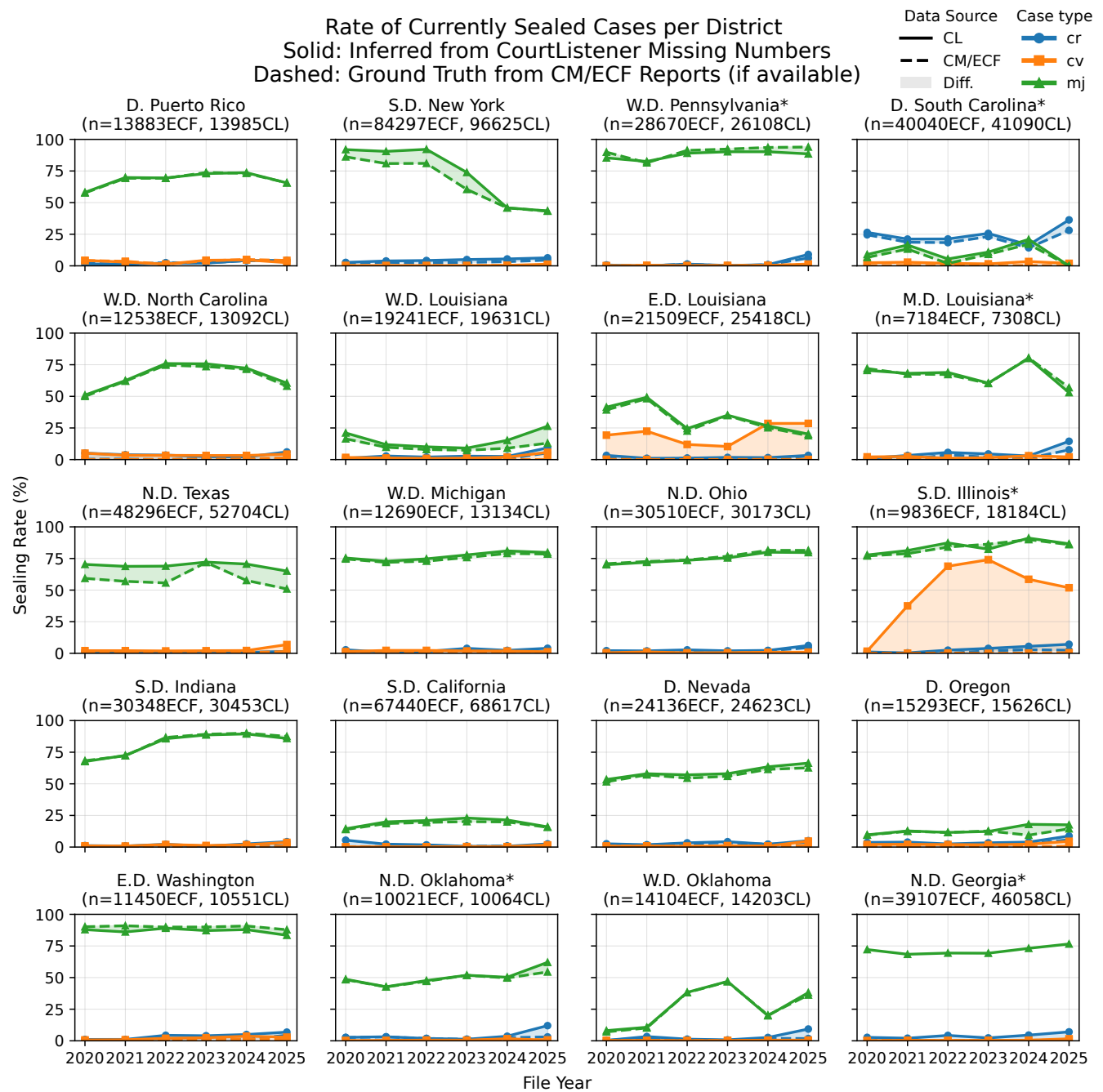


Figure 5: Rate of currently sealed cases per district from 2020 to 2025. Dashed lines are rates obtained from CM/ECF Case Reports and considered ground truth. Solid lines are rates inferred from CourtListener and approximations. * indicates districts where CourtListener has suboptimal coverage due to lack of RSS feeds or failing empirical timeliness tests. Their results should be interpreted with caution.

example, quite a few districts experienced an increased sealing rate of criminal (cr) cases but only for those filed in 2025. These criminal cases likely contain pending indictments where the defendant is not yet arrested, and thus are sealed until their arraignment. Otherwise, older criminal cases are typically sealed for a very short duration and most do not remain sealed for years (Section 6).

However, we did not observe similar increases in magistrate (mj) cases filed over the years. The most likely explanation is that many of the warrants are never unsealed even after their execution. For example, the search warrant filed in 3:24-mj-03122 was executed in August 2024 and even included as attachments in further unsealed applications for search warrants against the same target location [39, 40]. However, the case containing the original executed search warrant is still sealed and has not been unsealed as of September 2025. Furthermore, the Northern District of New York (nysd) has a sharp drop in the sealing rate of mj cases starting in 2023. Investigating further, the change coincides with the court’s issuance of Standing Order 1:23-mc-213, which mandates the electronic filing of all electronic criminal applications and restricts the sealing of magistrate cases to limited scenarios [43], and this order would explain the drop.

E Evaluation of Inferring Historically Sealed Cases with CourtListener

We evaluated our CourtListener timestamp methodology (Section 5.2) end-to-end against the IDB proceeding-code labels for currently unsealed criminal cases filed from 2020 to 2025 in the twenty districts. Then, we used the timestamp difference between when a case was filed and when it was added to CourtListener to infer whether the case was historically sealed. We plot the Precision/Recall curve for each district and each year in Figure 6 when varying the time delay threshold. Each curve represents the performance of using different thresholds to identify historically sealed cases in a specific district and year. The x -axis represents the Recall (i.e., the percentage of historically sealed cases identified), and the y -axis represents the Precision (i.e., the percentage of identified cases that were actually historically sealed). The larger the area under the curve, the better the performance. To better illustrate the performance, we also present the average precision (AP) score for each curve in the legend. In addition, for each curve, we mark the point where the threshold is four days, the threshold we use in the paper. We also automatically calculate the threshold that maximizes the F1 score and mark them as crosses along the curves.

Overall, fifteen out of the twenty districts have at least 80% average precision in 2024, indicating that our CourtListener timestamp method performs well in identifying historically sealed criminal cases. Moreover, three out of the five districts with lower precision are also the ones that failed to meet the timeliness requirement, reiterating that the accuracy of our CourtListener methodology is closely tied to the assumption of CourtListener’s timeliness. In addition, accuracy changed slightly in the most recent nine months measured. For example, ncwd, lamd, and ilsd all had small improvements, but laed and insd had degradations. Interviews with CourtListener staff members suggest that these improvements are

(1) Civil Seal-indicator Extraction Prompt

You will be given a list of docket entry descriptions from a civil case. Your task is to find any indicator if the case has ever been sealed.

—

If any of the docket entry descriptions indicate the case has/had been sealed, return “true” for “ever-sealed” but an empty list for “missing-docket-entries”. If not, return “false” for “ever-sealed” and list all earlier docket entries (that might contain indication of sealing) that seem missing for a typical civil case in a list format.

—

Return the output in a following json format.

```
{
  "ever_sealed": "false",
  "missing_docket_entries": ["Complaint", "Motion to Seal Case", "Case Sealed"],
}
```

{RANKING PROMPT CONTINUED in Table 6}

Table 4: Prompt given to GPT-4.1 for Task 1 in Appendix F.

(2) Unseal-date Extraction Prompt

You are an expert at analyzing clerical data from U.S. courts. You will be given a list of docket entry descriptions from a case.

—

Your task is to find the first date that the **case** has been unsealed (can be partial).

—

Return the output in a following json format. Provide the date and the complete description as is provided:

```
{
  "date": "2024-05-01",
  "description": "Case unsealed",
}
```

{RANKING PROMPT CONTINUED in Table 6}

Table 5: Prompt given to GPT-4.1 for Task 2 in Appendix F.

likely due to the introduction of new methodologies in their scraping efforts to automatically guess case IDs on CM/ECF, which they have gradually rolled out to more districts in the past nine months.

F Parsing Docket Entries with an LLM

Supplementing the metadata signals of IDB proceeding codes and CourtListener timestamps, we used GPT-4.1 (gpt-4.1-2025-04-14) as a secondary validation and extraction aid. The LLM is neither ground truth nor the primary classifier: it can identify only events represented in the docket entries available from CourtListener. For each evaluated case, we retrieved and provided all available docket entries to the model with a structured prompt for two tasks. In the first task, we evaluated a subset of civil cases to determine whether they were historically sealed (Section 5.2). In the second task, we evaluated all currently unsealed civil and criminal cases to determine when they were first unsealed (Section 6).

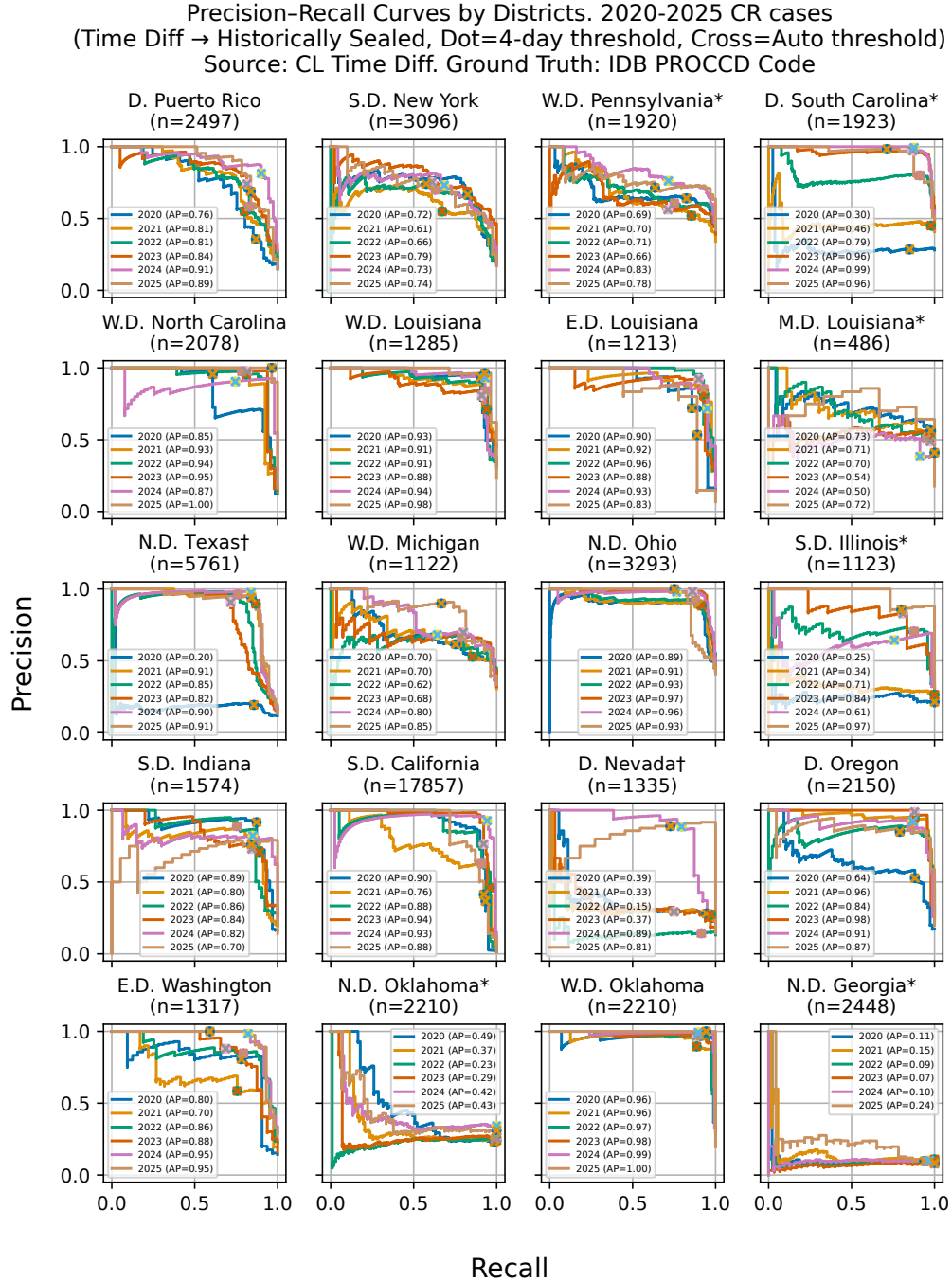


Figure 6: Precision/Recall curves for identifying historically sealed criminal cases using CourtListener, evaluated against the IDB proceeding-code labels.

* indicates districts where CourtListener has suboptimal coverage due to lack of RSS feeds or failing empirical timeliness tests.

† indicates districts that introduced RSS feeds during our study period, leading to changes in CourtListener’s timeliness.

Refer to the following rankings for what you'll be looking for. If you cannot find any docket entry that even remotely indicates the case being sealed or unsealed, return a null value for the corresponding result.

—

Civil Cases – Ranked Sealing Indicators

1. 'Case sealed' / 'Entire case sealed' / 'Order sealing case'
→ Most explicit – the docket entry itself confirms the whole case is sealed.
2. 'Case filed under seal' / 'Complaint filed under seal'
→ Common at case initiation; indicates the full docket is not public.
3. 'Order sealing complaint'
→ Direct judicial order applying to the entire case, not just a filing.
4. 'Motion to seal case' **granted**
→ If no explicit "case sealed" entry exists, the granting of a case-level sealing motion is the effective seal date.

—

Criminal Cases – Ranked Unsealing Indicators

1. 'Order unsealing' / 'Case unsealed' / 'Indictment unsealed' / 'Unsealed in open court'
→ Most direct and explicit indicator – the docket date here is usually the **true unseal date**.
2. 'Motion to unseal' **granted**
→ If no direct "order unsealing" entry exists, the granting of the motion signals the effective unseal date.
3. 'Arraignment'
→ Arraignment cannot occur under seal – the case must be unsealed by this date.
4. 'Initial appearance'
→ Like arraignment, this public hearing forces unsealing of the case.
5. 'Indictment returned and unsealed' / 'Superseding indictment unsealed'
→ Explicit for indictments; sometimes appears instead of a general order.
6. 'Plea hearing' / 'Change of plea'
→ Requires public docket – fallback indicator if earlier entries missing.
7. 'Sentencing'
→ By sentencing, case must already be unsealed. Only useful as a *latest possible* date if others are absent.

—

Civil Cases – Ranked Unsealing Indicators

1. 'Order unsealing' / 'Case unsealed' / 'Complaint unsealed'
→ Explicit, best marker of the unseal date.
2. 'Motion to unseal' **granted**
→ If no "order" entry, the grant date is usually the effective unseal.
3. 'Qui Tam' / 'False Claims Act' specific: 'decided/declined to intervene' / 'intervened' / 'unsealed'
→ These cases have statutory sealing; unsealing is often tied to U.S. deciding to intervene and serve as plaintiff.
4. 'Summary judgment hearing' / 'Trial'
→ Public proceedings – case must be unsealed by this point. Serves as a fallback latest date.

Table 6: Ranking Prompt of Docket Entries referenced in Table 4 and Table 5.

We designed two separate prompts for the two tasks, as shown in Table 4 and Table 5. For the first task, the prompt instructed the model to extract whether a docket entry indicates sealing for a civil case and, if not, which expected docket entries are missing (as an indication that CourtListener may have incomplete docket entries for the case). For the second task, the prompt instructed the model to extract the first unsealing date for a civil/criminal case (if any).

To provide the LLM with more legal context, we developed a ranked list of textual indicators tailored separately for criminal and civil cases. For sealing, we instructed the LLM to look explicitly for orders such as "Indictment sealed" or "Order to seal case". For unsealing, in criminal cases we ranked explicit orders such as "Indictment unsealed" or "Case unsealed" highest, while public proceedings (e.g., arraignment, initial appearance) served as fallback proxies. In civil cases, we prioritized unsealing orders, granted

motions to unseal, and qui tam unseals/interventions, with public hearings used as last-resort indicators (Table 6).

We then manually checked samples of the LLM outputs. For the first task, we randomly sampled five cases per district and manually checked whether a seal-indicating docket entry exists based on the same ranking table provided to the LLM. For the second task, we randomly sampled 100 criminal and civil cases each, for both cases that the LLM labeled as having an unsealing docket entry and those that it labeled as not having one. We then checked whether an unseal-indicating docket entry indeed exists or not for these cases. Across these reviewed samples, we observed zero disagreements between the LLM outputs and the available docket text. This checks extraction from observed text only; it cannot detect events absent from CourtListener, and it does not resolve ambiguity between partial- and whole-case unsealing dates.