CaseLinker: 5 Sources, 500 Cases, and Scaling Considerations

Mrinaal Ramachandran
Graduate Student, Department of Computer Science, University of Massachusetts Amherst
Independent Research | github.com/mrinaalr/CaseLinker | Live Demo
Technical Report Series #3 | April 2026

Abstract

CaseLinker is an open-source system for aggregating, extracting, and analyzing public ICAC case reports across jurisdictions. Report #3 scales the dataset to 500 cases across 5 sources — AZICAC, Georgia Bureau of Investigation, NCMEC CyberTipline, Idaho Office of Attorney General, and Michigan State Police — extracting 5,236 features across 9 dimensions via a hybrid deterministic-ML pipeline. This report introduces facet tree navigation for interactive case analysis, documents three primary data use cases (historical trend analysis, offender and crime evolution, and jurisdictional LE response shifts), and formalizes adversarial risk via a utility asymmetry proof showing that defender utility scales with dataset size while adversary utility remains approximately constant. Current coverage represents 5% of an estimated 10,000 public ICAC reports. Report #4 targets 2,000 cases and practitioner validation.

Keywords: ICAC investigations, cross-case analysis, scaling systems, historical trends, adversarial risk assessment, interpretable AI

Contents

1. Introduction

1.1. Monthly Report Series

Report #2 established hybrid deterministic-NER extraction, achieving 87.9% agency coverage across 207 cases (2011–2024), revealing distributed LE networks and early platform signals (2). Report #3 scales to 500 cases across 3 new sources (Georgia Bureau of Investigation (GBI), Idaho Office of Attorney General, and Michigan State Police), introduces facet tree navigation, and paves the path for comprehensive historical and ICAC trend analysis over the past 3 decades.

1.2. ICAC Cases are Not Static, Variations in ICAC Operations

An immediate conclusion from ingesting the GBI cases was the volume of large-scale, multi-perpetrator operations completed by this jurisdiction. A total of 56 known investigations were extracted from the 66 GBI case reports, 19 of which involved multiple perpetrators (34%), 4 of which were online, 1 proactive, and 14 undercover. Compare this against the AZICAC dataset, which shows only 2 cases involving multiple perpetrator operations from the 30 known operations (7%). Further analysis confirms that while each ICAC jurisdiction handles cases across all topics (CSAM, production, family, hands-on, grooming, platform involvement, investigation type), extracted patterns vary in distribution across the 9 dimensions.[1]

2. Improvements in Report #3

2.1. New Functions

Facet Tree Navigation: The facet tree provides interactive, 9-dimensional navigation across the 500-case dataset, converting extracted case features into a precise, investigator-controlled tree. Dimensions include:

- (1) **Topic** (CSAM, possession, production, online-only)
- (2) **Severity** (infant victims, hands-on abuse, grooming, multiple perpetrators)

[1] Preliminary conclusions directly from querying the CaseLinker interface. Their derivation from the facet tree is shown in 5.1. Underlying extraction accuracy and false positives have been checked via the audit system. However, until ablation tests are completed and extraction is rigorously tested, these stats are still an approximation over the dataset, not a hard claim. NHSR determination: HRPO #7668. Public, redacted data analyzed under UMass Amherst HRPO #7668 (Not Human Subjects Research). Refer to sources tab for full disclosure and compliance documentation.

- (3) **Platform** (Discord, Snapchat, Facebook, chat, social media)
- (4) **Inv. Type** (undercover, proactive, reactive, online)
- (5) **Source** (AZICAC, GBI, NCMEC, Michigan, Idaho)
- (6) **Agency** (from 518 unique agencies)
- (7) **Organization** (broader organizations, not necessarily law enforcement)
- (8) **Location** (NER extracted locations)
- (9) **Severity Phrase** (perpetrator admissions and case severity phrases)

New Sources: 66 cases from GBI, 244 cases from NCMEC CyberTipline annual report, 134 from Idaho Office of Attorney General and 11 from Michigan State Police.

2.2. Data Utility Analysis

This report introduces a comprehensive analysis of CaseLinker's data utility across key use cases

- Historical trend analysis tracking platform/offender/crime evolution from 2010–2026
- ICAC responses and evolution (operation outcomes/case load) across sources/jurisdictions
- Mathematical proof formalizing offender-defender utility disparity (investigators gain comprehensive success patterns; adversaries receive only shallow, non-actionable warnings from public successes)
- Bounded risk assessment (mosaic re-identification capped by source documents; evasion limited by absence of failure modes and operational details)

Full elaboration follows in Sections 4, 5, and 6 below.

Table 1: Source coverage (500 cases). "Features" = non-empty values among extracted dimensions.

Source	Cases	≥3 feat.	Inv. type	Severity	Platform
AZICAC	47	100.0%	63.8%	80.9%	27.7%
GBI	66	100.0%	100.0%	74.2%	98.5%
NCMEC	244	100.0%	59.2%	38.5%	33.9%
Idaho ICAC	134	100.0%	31.3%	35.8%	17.2%
Michigan ICAC	11	100.0%	100.0%	18.18%	100.0%

3. Scaling Considerations

3.1. From 500 to Comprehensive Coverage

There are 61 ICAC task forces investigating and arresting offenders since 1998 (4). Although it is challenging to estimate the total number of public ICAC cases, as there is no aggregate database or source that collects outcomes of ICAC or CSAM related cases, it can be estimated that there are at least 10,000 public reports that can be found and analyzed. This number is derived from taking the current ICAC sources, averaging their case published volumes, and scaling from 2010-2026 back to 1998.

Table 2: Sample Sources for Public ICAC Case Volume Estimation

Source	Date Range	Cases	Years	Cases/Year
AZICAC	2011–2014	47	4	11.8
GBI CEACC	2010–2026	66	16	4.1
Idaho AG ICAC	2021–2026	134	6	22.3
Michigan State Police	2025–2026	11	2	5.5
Average				**10.9**

Average annual publication rate: 10.9 cases/source/year across 4 task-force sources. NCMEC is excluded from this estimate as a national aggregator rather than a single jurisdiction. Scaling to 61 task forces dating back to 1998:

$$\text{Total Public Reports} = 61 \times 10.9 \times 28 = 18{,}608 \quad (1998\text{–}2026)$$

. Note this value represents linear scaling of caseload, which is not true. ICAC case volumes have exploded, especially in 2010 and beyond. As a result, public coverage may show few cases in 1998-2010, a substantial increase in 2010-2020, and average to the 10 cases per source per year from 2020-2026 onward. As a scaled estimate, I bound this number at 10,000, not 18,000 to estimate for the total number of public ICAC reports that should be publicly available.

Current dataset represents $500/10,000 = 5\%$ coverage. At scale, CaseLinker would analyze the complete census of successful public ICAC operations across all jurisdictions and agencies—enabling comprehensive trend analysis on online harms and their enforcement. Note, this depends on if public archives exist for each jurisidiction. A preliminary scan shows that most jurisdictions including Texas, California, and Florida include public reports.

Ingestion and Processing: Cases analysis is currently pre-computed; PDFs are ingested, processed, and analyzed with a one time cost. On a local machine, the current end-to-end pipeline takes roughly 5 minutes on 500 cases to extract 5236 features from 5 PDF sources, complete clustering, and store within the local SQLite / deployed PostgreSQL server. The projected time for 10k cases is 1.5-2 hours.

3.2. Challenges and Solutions

One emerging challenge is not computational resources and scale but rather issues relating to data bias, normalization, reporting details, and downstream impacts from differences in the length and quality of reporting. For example, the Michigan ICAC cases tend to be brief and succinct. This does *not* mean that there was not a substantial case load, investigative burden, or offender patterns worth studying, simply that the system was unable to extract it from the reported data as it was presented.

A quick visual example is comparing the cases reported by the AZICAC and Michigan ICAC respectively. AZICAC includes detailed accounts of the perpetrator behaviors, violations, and even severity phrases indicating offender attributes *ex: azicac_2011_005*. Michigan ICAC cases appear to be higher level, summarizing charges, investigative methods and agencies but lack explicit details on the offenders. This disparity continues within

reports, cases within the NCMEC 2024 annual report vary, with some having brief success outcomes and prosecutions and others having detailed insights into the specific crime.

These are limitations at the available data level, not analytical level. These are also important limitations to have, as public reports have mosaic reidentification and adversarial risks, which are discussed in section 6. While there are different methods for normalization, at this time *no effort will be made to normalize this data.* To be clear, all cases will be extracted, processed, and analyzed in their current format, regardless of reporting detail. This is to serve CaseLinker's primary functionality: visualizing and analyzing public reports. Downstream statistical analysis and broader inferences would require normalization, especially if attempting to come to conclusions on platform involvement for example.

As a final footnote, the analytical layer can be adapted. The current system supports broad feature extraction across the 9 dimensions to collect as much information as possible for investigators, journalists, and case studies. However, the processing layer is not set in stone and can be adapted to be focused on statistics, though that is not the primary research direction CaseLinker's public release is evolving towards.

4. Data Use Cases

4.1. Broad Historical Comparisons and Evolution: 1998–2026

- Cross-era platform involvement: early internet → P2P → social media → modern platforms, traceable across the full date range of the dataset
- Case volume and charge type shifts across three decades, filterable by source, jurisdiction, and year range
- Correlation of caseload surges with legislative changes, major platform policy shifts, and new technology.

4.2. Offender and Crime Evolution

- Distribution of offense type (possession, production, distribution, hands-on) over time and across jurisdictions
- Shifts in grooming behavior, online-only vs. hands-on operations, and multi-

perpetrator coordination across the dataset

- Perpetrator admission patterns and severity phrases extracted across 500 cases, enabling longitudinal behavioral comparison

4.3. LE Response Shifts Across Jurisdictions

- Investigation type distribution (undercover, proactive, reactive) by source and agency, revealing how different task forces allocate investigative effort
- Multi-perpetrator operation rates vary significantly by jurisdiction: GBI at 34% vs. AZICAC at 7%, invisible without cross-source aggregation
- Recurring agency networks and inter-agency coordination patterns across the 518 unique agencies in the dataset

5. Facet Tree: 9-Dimensional Navigation

5.1. From Chaos to Precision

The initial tree represents 500 distinct cases and contains 2459 nodes. This immediately demonstrates that there are a variety of branches that can be taken to group, analyze, partition, and otherwise engage with the extracted features from these cases.

The method presented is a facet tree with interactive pruning and drill-down to dynamically focus on topics of interest.

Analysts can select one or more topics from each dimension to analyze, both individually and in relation to other dimensions. To concretely explain navigation and how conclusions on GBI responses were made for this paper, you can first

1. Select Split on: *Source* and narrow to GBI, pruning the (4) other sources from the tree, reducing total cases from 500 to 66 and nodes from 2459 to 298.
2. Identify topics of interest. To support analysis of investigation types, select all known investigation types, pruning the tree further to include 56 cases and 256 nodes.
3. To case study GBI larger operations involving multiple perpetrators (large scale operations likely involving weeks–months of planning and resource allocation), select

the multiple perpetrators tag from severity indicators. This narrows the search to 19 cases and 93 nodes.

4. Remove all splits that do not improve knowledge for studying GBI operations involving multiple perpetrators (severity phrases, location, organization, agency, topic, platform). This reduces nodes from 93 to 11.

5. The resulting tree (referenced below) shows GBI cases involving multiple perpetrators, split by investigation type. Analysts could instead split by platform, victim age, date range, location, or any other combination and permutation of interest.

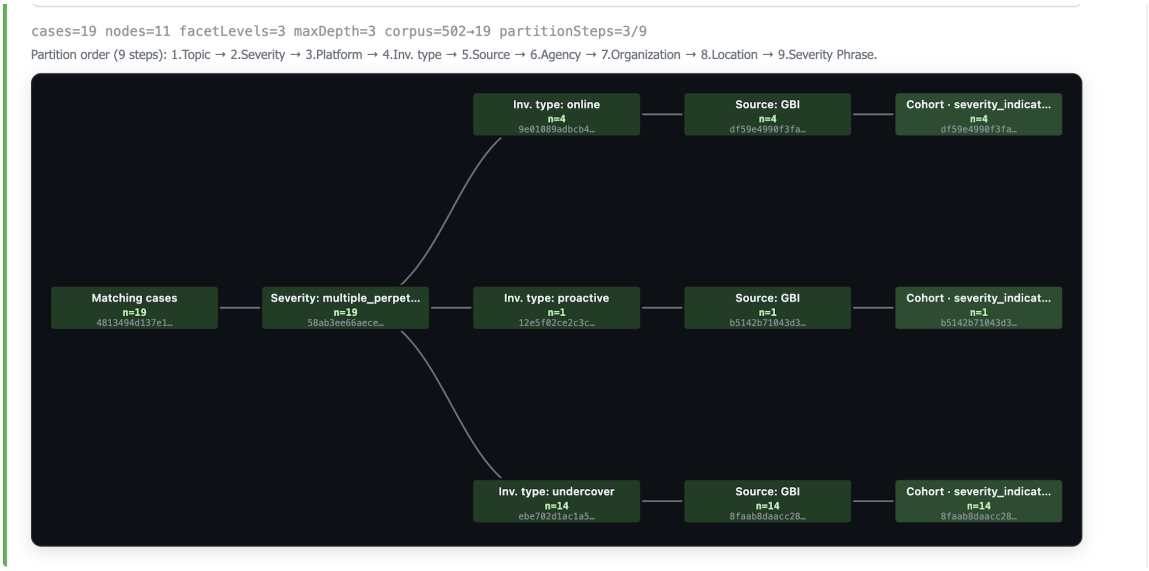

Figure 1: Resulting Facet tree from running the pruning steps above.

6. Adversarial Usage Risk Assessment

Below is a mathematical approach to analyzing adversarial risk of the CaseLinker system. As was argued in my position piece, Why Internet Crimes Against Children And Retrospective Case Analysis Matters (3), all technology can be misused by bad actors and CaseLinker is no exception. The question is not does CaseLinker present *no risk* but rather does the defender utility outweigh offender insights and are appropriate steps are taken to prevent misuse.

6.1. Definitions

Let $\mathcal{D} = \{c_1, c_2, \ldots, c_n\}$ denote the CaseLinker dataset of n public ICAC case reports, where each c_i is drawn exclusively from published law enforcement press releases and annual reports representing *successful* prosecutions. Define:

- $U_D(\mathcal{D})$: aggregate utility to defenders (investigators, researchers, journalists)
- $U_A(\mathcal{D})$: aggregate utility to adversaries extracted from $\mathcal{D}$
- $\mathcal{R}(c_i)$: re-identification risk associated with case c_i
- $\mathcal{S}(c_i)$: operational signal extractable from case c_i (tactics, tools, failure modes)

By construction, $\mathcal{D}$ contains only success outcomes. Failure modes, operational tradecraft, undercover methodologies, and tip sources are absent — agencies do not publish these.

6.2. Proposition 1: Re-identification Risk is Source-Bounded

Proposition 1. $\mathcal{R}(c_i) \leq \mathcal{R}(\text{source}(c_i))$ for all $c_i \in \mathcal{D}$.

Proof. CaseLinker ingests public documents without processing victim names, perpetrator personal identifiers, addresses, or case numbers beyond what appears in structural metadata. Feature extraction operates across 9 dimensions (platform, investigation type, severity indicators, agency, location, etc.) and introduces no new identifying information. The re-identification ceiling is set by what the originating agency chose to publish and redact — not by CaseLinker's processing. Aggregating redacted summaries does not reconstruct the underlying case details; it surfaces distributional patterns across cases. Therefore the mosaic risk is bounded above by the source document itself. □

6.3. Proposition 2: Operational Signal is Absent by Construction

Proposition 2. $\mathcal{S}(c_i) \approx 0$ for all $c_i \in \mathcal{D}$, and this bound holds independent of n.

Proof. Each c_i is a redacted success narrative. The information available in any c_i is limited to: charge categories, investigation type label (undercover, proactive, reactive), platform mentions, agency names, and outcome (conviction, sentencing). Crucially, *how* law enforcement operated — communication interception methods, undercover persona construction, tip source routing, digital forensics tooling — is absent not because CaseLinker omits it, but because the source document does not contain it. An adversary reading all $n = 500$ cases manually derives the same operational signal as reading 10: approximately

zero. Aggregation does not recover what was never published.

Furthermore, adversarial statistical inference over $\mathcal{D}$ does not track: since $\mathcal{D}$ contains only successful operations, it cannot surface the conditions under which law enforcement *fails*. An adversary seeking evasion strategy requires failure mode data; no such data exists in $\mathcal{D}$ or its source documents. □

6.4. Bounded Misuse Cases

Despite the above, three adversarial misuse vectors are worth naming explicitly, along with their practical ceiling.

Platform Avoidance. An adversary observing that Facebook appears in a repeated number of successful prosecutions may attempt to migrate to less-documented platforms. This signal is available from repeated manual reads of any public ICAC source and provides no marginal lift from CaseLinker's aggregation. Platform avoidance is further negated by CyberTipline IP tracing and cross-platform device correlation, which bind the offender to their device independent of which platform was used.

Geographic and Agency Avoidance. An adversary may attempt to infer which jurisdictions or agencies produce high prosecution rates and avoid them. However, internet-based ICAC offenses are not geographically bounded in the same way as physical crimes — CyberTipline reports and device forensics route investigations to the offender's location regardless of which jurisdiction initiates the case. Geographic avoidance provides minimal protective value.

Victim Demographic Targeting. An adversary may attempt to use severity and victim age distributions to identify demographic groups that appear less prosecuted. This risk is low in practice: the dataset represents *successful* prosecutions across all age ranges and offense types, meaning there is no demographic "gap" visible in $\mathcal{D}$ that reflects a true enforcement blind spot rather than a reporting artifact. CaseLinker surfaces no information on how to obtain CSAM, how offender communities form, or how platforms are regulated beyond what is publicly known.

6.5. Proposition 3: Utility Asymmetry

Proposition 3. $U_D(\mathcal{D}) \gg U_A(\mathcal{D})$ for $n \geq 50$.

Proof. Defender utility U_D is superlinear in n. Investigators and researchers gain cross-jurisdictional success patterns, platform evolution trends (early chatlogs → P2P → Discord),

multi-perpetrator operation typologies, recurring agency networks, victim and perpetrator age distributions, common severity indicators, perpetrator admission patterns, and offense type distributions — none of which are visible from any single source or from manual review of a small case set. These insights require $n \gg 10$ to emerge and are not available from any existing aggregate public resource.

Adversary utility U_A is approximately constant in n. The marginal gain from case 11 through case 500 is near zero: no failure modes, evasion tactics, or operational details accumulate with scale. Formally:

$$\frac{\partial U_A}{\partial n} \approx 0 \quad \text{while} \quad \frac{\partial U_D}{\partial n} > 0$$

Therefore $U_D(\mathcal{D}) \gg U_A(\mathcal{D})$ for any practically relevant n. □

6.6. Access Controls and Risk Boundary Conditions

The above propositions hold for the current system and data scope. Two boundary conditions warrant explicit statement.

Access Control. For cases involving elevated sensitivity indicators (infant victims, hands-on severity phrases), CaseLinker restricts direct case retrieval to access-key holders. Facet navigation and aggregate statistics remain public. This prevents the system from functioning as a targeted lookup tool for the most sensitive case details while preserving full analytical utility for investigators, researchers, and the general public.

Scope Boundary. This risk assessment is bounded to the current data scope: public, redacted case summaries. If the system were extended to incorporate non-public case data, active investigation records, or unpublished operational details, the threat model would require fundamental reassessment. Similarly, if future extraction surfaces qualitatively new sensitive features not present in the current 9-dimensional schema, risk must be re-evaluated. The propositions above do not generalize beyond public redacted sources.

6.7. Summary of Proofs

Public ICAC reports = successful ops only. Adversary gains: "avoid x platform" (obvious from manual reads and not a strong insight for evasion). Defender gains: comprehensive success patterns across 61 TFs, 28 years.

Re-identification: Capped by source docs (no names processed). Risk = publicity of

origin.

Misuse Ceiling: No tactics/tools/failures. Platform and Geographical avoidance negated by CyberTipline/IP tracing. Victim demographics reflect prosecuted CSAM, not sourcing gaps.

7. Conclusion

Report #3 scales CaseLinker to 500 cases across 5 sources, introduces facet tree navigation across 9 dimensions, and formalizes the adversarial risk argument that has been implicit since the project began. The numbers are useful — 5,236 extracted features, 518 unique agencies, 2,459 facet tree nodes — but they are not the point.

The point is that ICAC enforcement is a 30-year-old ecosystem that has never had a tool for looking at itself in aggregate. Jurisdictions operate in parallel, catching offenders and publishing successful outcomes, but lack mechanisms for researchers, journalists, or task force commanders to ask: what does the full picture look like? Which platforms recur? How do multi-perpetrator operations distribute across jurisdictions? What do perpetrator admissions look like at scale? Which funding, policy, and enforcement actions should be taken based on emerging trends. These are not easy questions to answer. However, they are the baseline questions any serious analysis of a law enforcement ecosystem should start with, and until now they have been unanswerable without months of manual work and reviewing hundreds of graphic cases across thousands of pages.

Five percent coverage of estimated public ICAC reports is the beginning. The analytical value compounds with scale — the asymmetry proof in Section 6 is not just a risk argument, it is a description of how pattern recognition works. At 500 cases, you have a window into this ecosystem. At 10,000 you can see the structure. That is the direction this project is going.

CaseLinker is not a replacement for investigative judgment, institutional knowledge, or the expertise of the people who have been doing this work for decades. Often, investigators, commanders, and experts who have been protecting children, catching offenders, and

shaping policy recognize patterns and trends extremely well. CaseLinker is a way of making what is known visible at a scale, uncovering deeper patterns that manual review cannot reach, supporting these claims in an auditable, concrete, evidence tied manner, and providing material insights into the nature of these crimes for those unaware, without requiring the consumption of graphic case material. The goal has not changed since the initial release: give investigators, researchers, and the general public the ability to understand this landscape without requiring repeated direct exposure to its worst material (1; 5).

Report #4 targets 2,000 cases, semantic similarity across cases, and the first round of practitioner validation.

CaseLinker is available at github.com/mrinaalr/CaseLinker under the MIT License.

Live demo: web-production-13a2.up.railway.app.

References

[1] Perez, L. M., Jones, J., Englert, D. R., & Sachau, D. (2010). Secondary traumatic stress and burnout among law enforcement investigators exposed to disturbing media images. *Journal of Police and Criminal Psychology*, 25(2), 113–124.

[2] Ramachandran, M. (2026, March). *CaseLinker: Interpretable ML Approaches for Analyzing Internet Crimes Against Children Reports. Technical Report Series #2.* University of Massachusetts Amherst. https://mrinaalr.github.io/website/CaseLinker-%20Interpretable%20ML%20Approaches%20for%20Analyzing%20Internet%20Crimes%20Against%20Children%20Reports.pdf

[3] Ramachandran, M. (2026, March). *Why Internet Crimes Against Children And Retrospective Case Analysis Matters* University of Massachusetts Amherst. https://mrinaalr.github.io/website/Why%20Internet%20Crimes%20Against%20Children%20And%20Retrospective%20Case%20Analysis%20Matters.pdf

[4] Office of Juvenile Justice and Delinquency Prevention. (2024). *Internet Crimes Against Children Task Force Program.* https://ojjdp.ojp.gov/programs/internet-crimes-against-children-task-force-program

[5] Burns, C. M., Morley, J., Bradshaw, R., & Domene, J. (2008). The emotional impact on

and coping strategies employed by police teams investigating internet child exploitation. *Traumatology*, 14(2), 20–31.