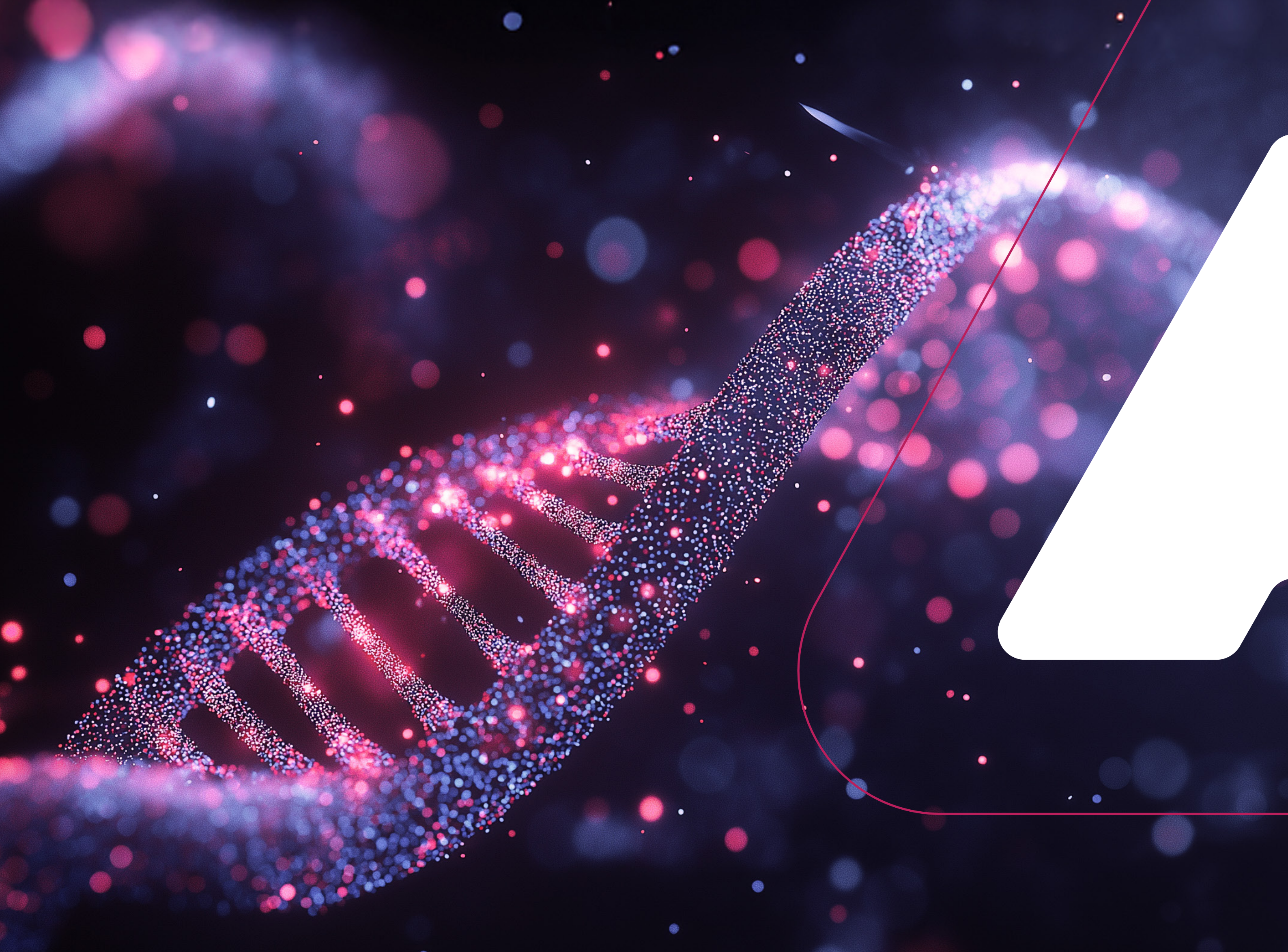
Lab 1

1ST EDITION / 2025

The Anatomy of a Breach



Contents

Foreword

In today's digital economy, data has become both our greatest asset and our greatest liability.

Cyberattacks and fraud are growing in sophistication and intent, targeting the very structure of our organizations. To achieve this, threat actors are studying us, probing every corner of our business, from third-party vendors to forgotten cloud and IT assets, to find our weakest points to exploit.

We've entered an era where cybercriminals behave like data scientists. They scrape the dark web for breach datasets, not just to find high-value data assets, like customer or corporate Personally Identifiable Information (PII), but to uncover the overlooked fragments. Internal documents, credentials, and supplier communications that, when pieced together, paint a detailed picture of how we operate.

These fragments become tools for social engineering, deepfake generation, and lateral movement. A single leaked access key, project document or HR file can now trigger a cascade of downstream attacks and fraud.

This is why **we can no longer think of a breach as a leak of data, it's a weapon** to be used over and over again at an alarming rate.

This report, ***The Anatomy of a Breach***, arrives at a critical moment. It challenges a dangerous assumption that still lingers in boardrooms and dashboards alike: that the size of a breach is the best indicator of its impact. It's not.

The real damage lies in the content of what's exposed. A breach of 500 files might seem minor until you realize those files contain cryptographic keys, customer account data, wealth statements, or sensitive commercial contracts.

Accurately assessing the impact of a breach now means uncovering not just if data was leaked, but what was leaked, how it can be used, and who might be affected. And faster than it can be used against us.

I encourage you to read this report not just as a benchmark, but as a call to action to move beyond incident counts to content-aware breach response.

Robin Brattel
Co-Founder & CEO, Lab1

Key findings

The key findings are derived from an analysis of **141 million unstructured files** - such as emails, PDFs, spreadsheets and code files - across **1,297 data breach incidents** analyzed by Lab 1's AI Data Intelligence Application.

The average breach contains

22,647
Files

13.44GB
Data

14
File Types

22
File Classifications

482
Organizations Impacted

Customer and Corporate PII exposed in most data breaches

67%
of incidents contain customer data

82%
of incidents contain HR data

51%
of incidents contain U.S. Social Security Numbers

54
email addresses exposed per file

The significant proportion of breaches containing customer and corporate PII introduces not only a high compliance risk, but supports hyper-targeted phishing and social engineering campaigns.

The emergence of sophisticated generative AI fraud makes this particularly urgent, warranting Finra to warn investors[1] early in 2025, especially when bad actors have access to materials that help them impersonate employees or customers with high fidelity.

Financial data exposure increases the risk of fraud for both the organization and its customers

93%

of incidents contain financial data

36%

of incidents contain IBANs

49%

of incidents contain bank statements

14%

of incidents contain wealth statements

Financial data was the most exposed classification, making up 41% of all files. Corporate documents, such as bank records, balance sheets, and billing data are a prime target for fraud, extortion, and monetization on darknet markets as they often contain banking details, tax identifiers, or invoicing information.

IBANs, and banking and wealth account details can enable fraud through techniques such as mandate scams, payment redirection, or synthetic identity construction for customers and organizations alike.

Exposed code, credentials, and keys create high-impact breach pathways

87%

of incidents contain code

17%

of all incident data is code

18%

of incidents contain SSH & RSA keys

79%

of incidents contain System Logs

Code files - particularly .xml and .json, which were the most exposed code file types - frequently expose sensitive configurations, credentials, or customer data. System Logs often contain technical information alluding to the configuration of critical applications.

However, the exposure of cryptographic private keys, such as SSH and RSA Keys, especially in unencrypted form, poses the most severe threat, enabling attackers to bypass authentication, access secure systems, and elevate the risk of lateral movement and credential forgery.

Executive Summary

Data breaches are escalating in both scale and complexity. As data creation accelerates at an exponential rate - projected to grow 23% between 2024 to 2025[2] - leaks are stemming not only from attackers, but also from employees, suppliers, 4th parties, contractors, cloud & IT misconfigurations, and vendors. This creates a risk landscape so fragmented and unpredictable that anticipating and defending against every breach vector is virtually impossible.

≠

Not all breached data is equal

It has also led to an overemphasis on data volume. Sensational media headlines expose the huge volumes of simple, structured data, like email addresses and user names, and overlook the leak of unstructured files, including emails, PDFs, spreadsheets, code files, and invoices. Not all breached data is equal, and these files often hold high-value information that can severely undermine an organization's security, people, and its safety, compliance obligations, and commercial integrity, such as customer data, PII, IBANs, and HR records.

This report addresses the critical gap in understanding around the Anatomy of a Breach by conducting the largest known content-level analysis of over 141 million files across 1,297 data breach incidents. Using a custom pipeline that integrates machine learning, semantic labeling, and validation-driven regex extraction, we have classified the structure, sensitivity, and function of exposed files, ranging from financial records and personal identifiers to source code, credentials, and healthcare documentation.

The result is a first-of-its-kind annual benchmark report that business, risk and security leaders can use to inform themselves and their executives about the increasing and ongoing risk of breached data.

The report exposes that, contrary to conventional assumptions, when it comes to data breaches, size doesn't matter. Many incidents analyzed for this report with relatively low file volumes involved high-impact files that introduced significant risks.

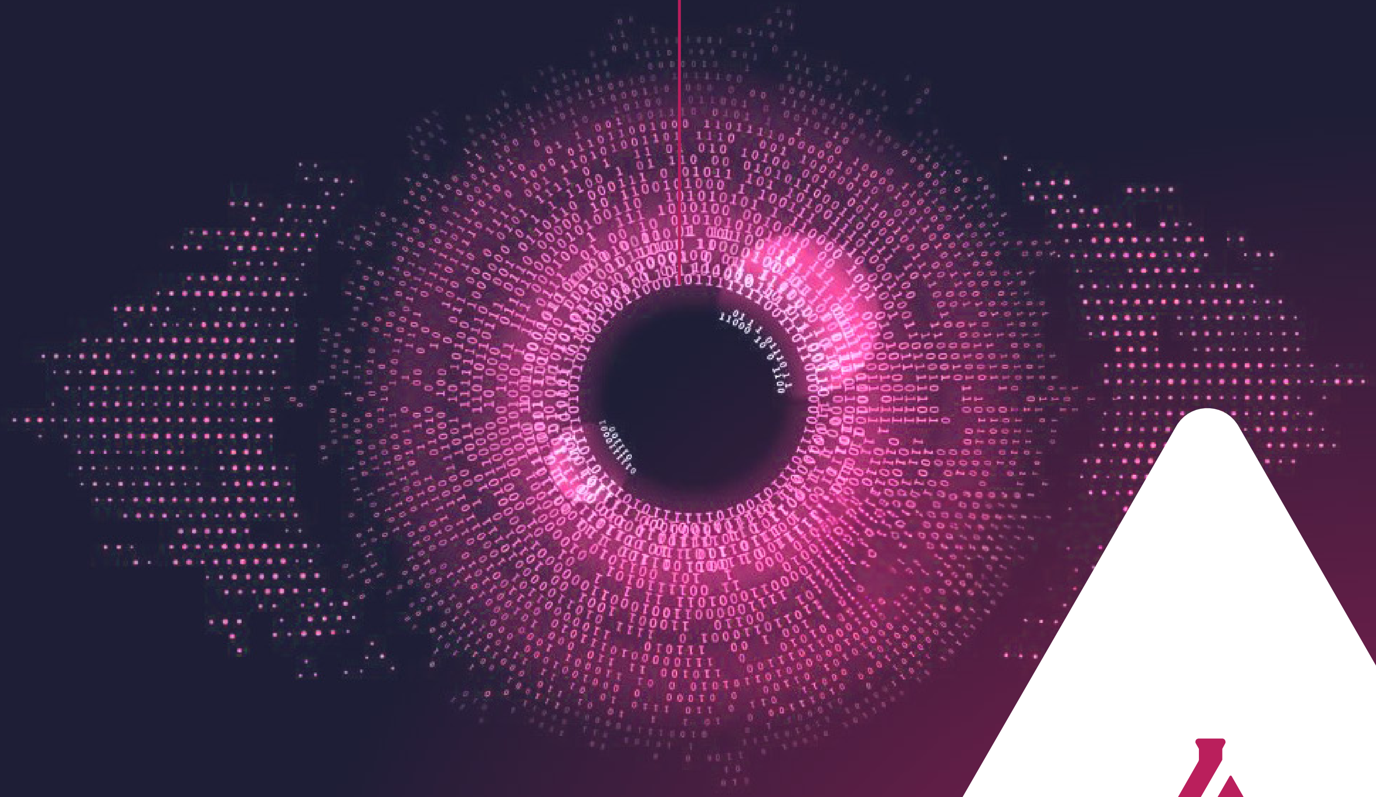
With adversaries now using artificial intelligence (AI) and machine learning (ML) tools to rapidly analyze stolen content and extract structured data, the weaponization of breached data is accelerating. From supplier relationships and email exchanges for social engineering, open credentials and keys for network access and lateral movements, and code to introduce Software Bill of Materials (SBOM) vulnerabilities, breached high-impact files are powering downstream fraud, deepfake generation, and targeted attacks.

The Anatomy of a Breach report, for the first time, reveals the contents and blast radius of breaches, and the composition of the 'average' breach. This report demonstrates the need for organizations to review and assess every file in a breach to take a content-aware breach response, in which the nature of what is leaked, not just how much, is the key determinant of damage and regulatory exposure.

Contents of a Breach

In this section, we explore the contents of a breach through three lenses:

- File Classification →
- File Type →
- Sensitive Information Types →

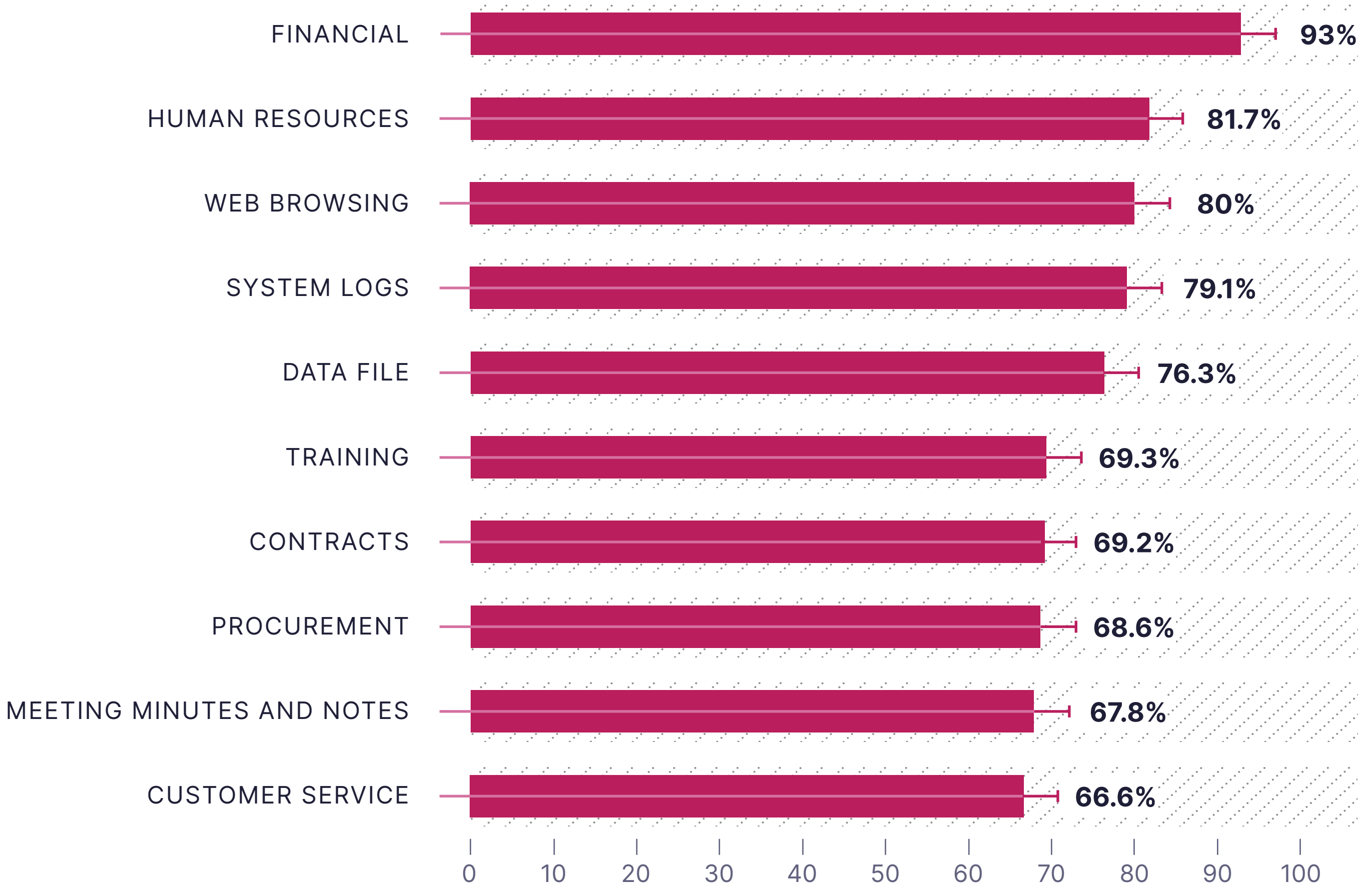


File Classification

The classification of the file reveals critical implications for data exposure risk and post-breach weaponization potential. Classification across the corpus of 141 million exposed files revealed the significant prevalence of high-risk content types.

Figure 1:
Top 10 File Classifications Most Commonly Identified in Data Breaches

Percentage of the 1,297 analyzed data breaches that includes each file classification



82%

of incidents include HR data

Rich HR data at risk of AI-enabled weaponization and advanced social engineering

Human Resources data appeared in 81.7% of breaches (Figure 1), often containing PII, payroll, and resumes. Further to the compliance risk, breaches rich in HR content are particularly suited for AI-enabled weaponization as these narrative-rich datasets can be used to generate synthetic identities, deepfake content, or voice-clone phishing attacks with high fidelity. It also significantly elevates the risk of advanced social engineering and psychological operations, especially in highly regulated industries.

Figure 2:

Top 10 Most Common File Classifications of Total Files Analyzed

Percentage of the 141 million breached files analyzed that belong to each file classification

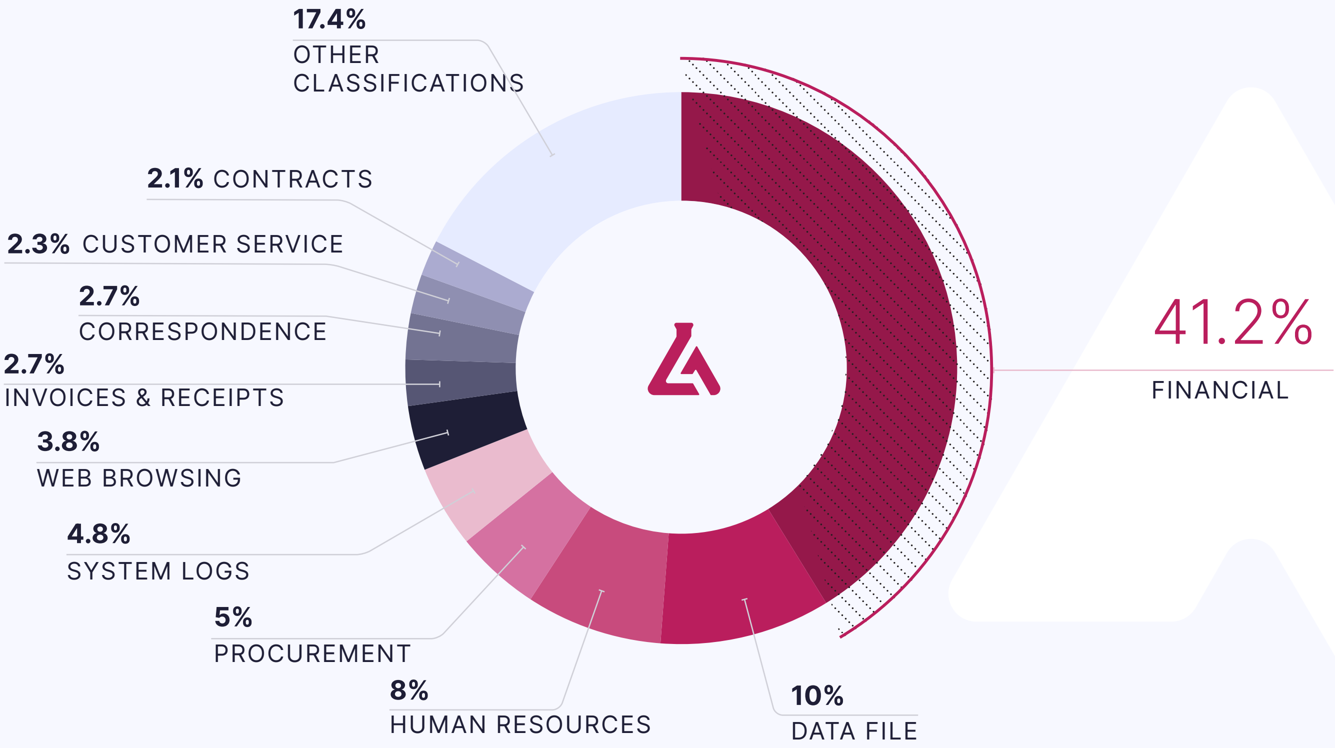
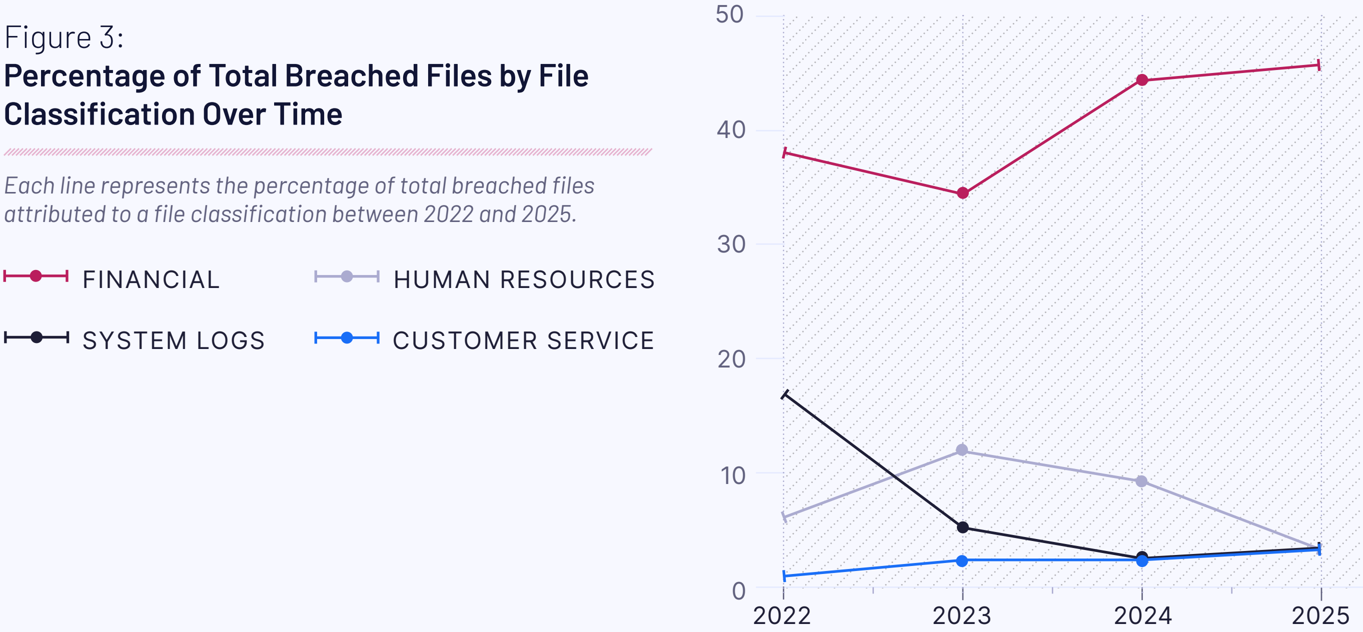


Figure 3:

Percentage of Total Breached Files by File Classification Over Time

Each line represents the percentage of total breached files attributed to a file classification between 2022 and 2025.



93%
of incidents contain financial data

Financial documents are the most breached file classification

Financial documents were the most dominant classification, appearing in 93% of incidents (Figure 1) and accounting for over 56 million of the 141 million total exposed files (Figure 2). This file type ranked as the most breached file classification every year from 2022 to 2025 (Figure 3). Often containing banking details, tax identifiers, or invoicing information, financial documents are a prime target for fraud, extortion, and resale on darknet markets.

Two-thirds of incidents involve customer service data

Two-thirds of incidents involved communications and records concerning customer service interactions and support, including customer details (e.g. contact details, addresses, etc.) (Figure 1). Exposure of customer service records containing personal data can lead to targeted phishing, identity theft, and regulatory violations under laws like the General Data Protection Regulation (GDPR) or the Federal Trade Commission Act (FTC Act). Such breaches risk substantial fines, legal action, and erosion of customer trust, especially if complaints or sensitive interactions are made public.

The 2023 23andme data breach, for example, resulted in a $30 million settlement for a U.S. lawsuit accusing the genetics testing company of failing to protect its customers' privacy[3] and a £2.31 million fine from the UK's Information Commissioner's Office for failing to protect UK users' genetic data[4], which contributed to the company filing for bankruptcy protection[5].

79%
of incidents contain System Logs

High prevalence of System Logs enables attackers to exploit and navigate systems

System Logs, present in 79.1% of incidents (Figure 1) and making up 4.8% of all breached files (Figure 2), are instrumental in understanding system behavior, user activity, and environmental configurations. This metadata can help attackers map out the system and detect vulnerable endpoints or misconfigured services, which attackers can use to craft targeted exploits, identify privilege escalation, and hide their activity once inside an organization's system.

File type

The types of documents exposed offer a clear view into how the organization operates and reveal multiple points of vulnerability that malicious actors can exploit after a breach.

Most common file type overlooked by DLP

Text files, dominated by the .txt extension, appeared most frequently in data breaches and were identified in 93% of the 1,297 incidents analyzed in this report (Figure 4). Typically humanreadable and unstructured, their simplicity and ubiquity often result in them being overlooked by traditional data loss prevention (DLP) mechanisms, despite their frequent sensitivity and serving as a common vector for leaked credentials and configuration notes.

Figure 4:
Top 10 File Types Most Commonly Identified in Data Breaches

Percentage of the 1,297 analyzed data breaches that included each file type

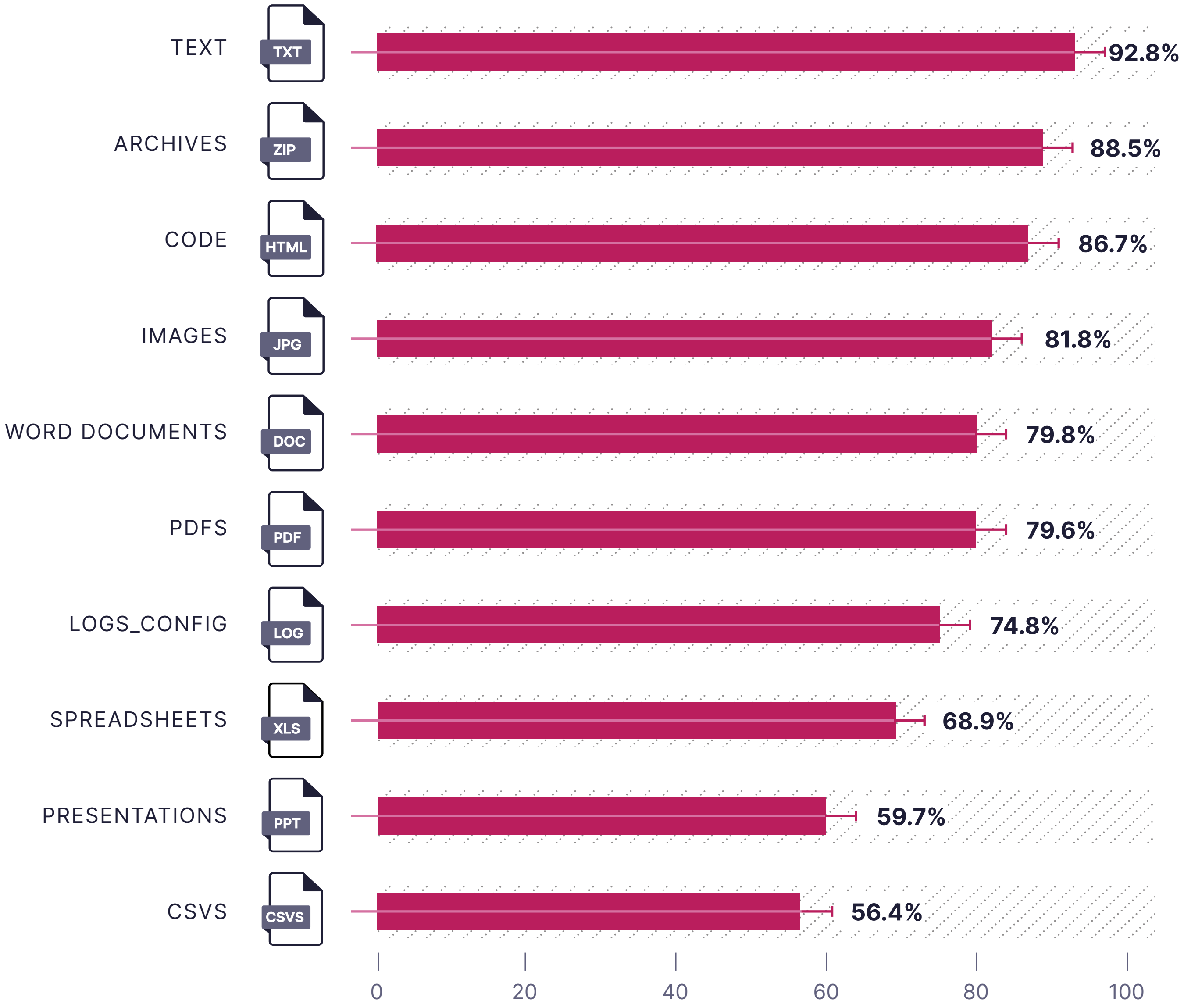


Figure 5:
Top 10 Most Common File Type of Total Files Analyzed

Percentage of the 141 million breached files analyzed that belong to each file type

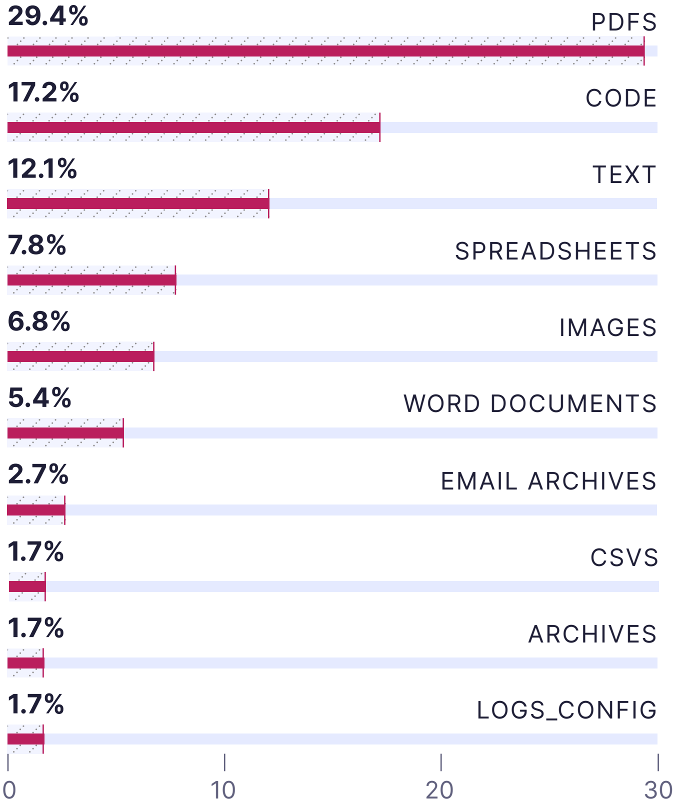
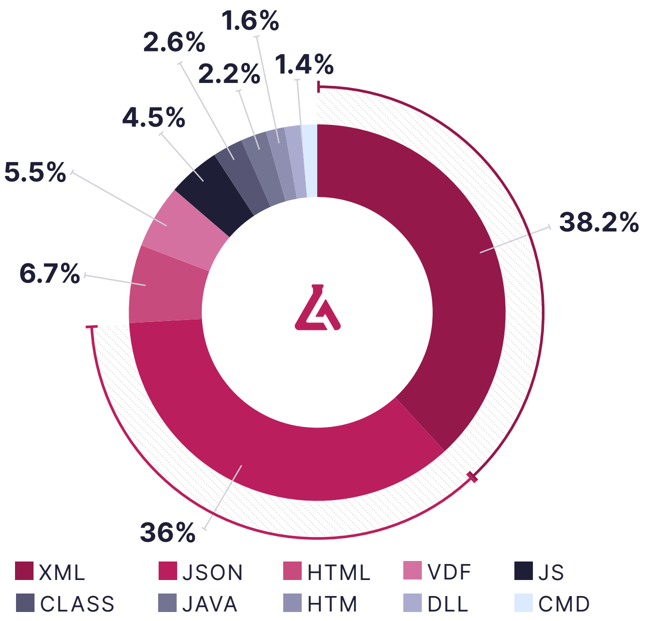


Figure 6:
Distribution of File Extensions Within the "CODE" File Type

Percentage of the 24,272,193 breached code files analyzed that have each file extension



Significant code leakage introduces major SBOM vulnerabilities

Code appeared in 86.7% of incidents (Figure 4), accounting for 17.2% of all 141 million files analyzed (Figure 5). Code leakage introduces vulnerabilities to the Software Bill of Materials (SBOM) by undermining the integrity, visibility, and trustworthiness of the software supply chain, highlighting a deeper structural risk in today's digital ecosystems. XML (34.3%) and JSON (32.4%) file formats constitute the majority of leaked code-related files (Figure 6). Frequently serving as configuration payloads or API artifacts. If exposed, they can reveal hardcoded secrets, environment variables, or backend schemas.

Leaked Server/Application Credentials and Access Keys pose a critical security risk

While not exposed at the same frequency, Certificates (e.g., .gpg, .pgp) and Remote Access configurations (e.g., .rdp, .ovpn) pose an especially acute threat. Present in 29% of incidents, leaked certificates may allow impersonation, enable man-in-the-middle (MITM) attacks, or be used to sign malicious code. Remote access configurations were exposed in 18% of incidents and can lead directly to lateral movement in the post-exploitation phase of an attack, particularly if passwords are cached or embedded. These file types serve to show how even breaches with relatively low file volumes can introduce significant risks when high-impact files are exposed.

Sensitive Information Types

Analysis of several sensitive information types reveals the scale of the security and compliance risks associated with the 1,297 breach incidents that informed this report.

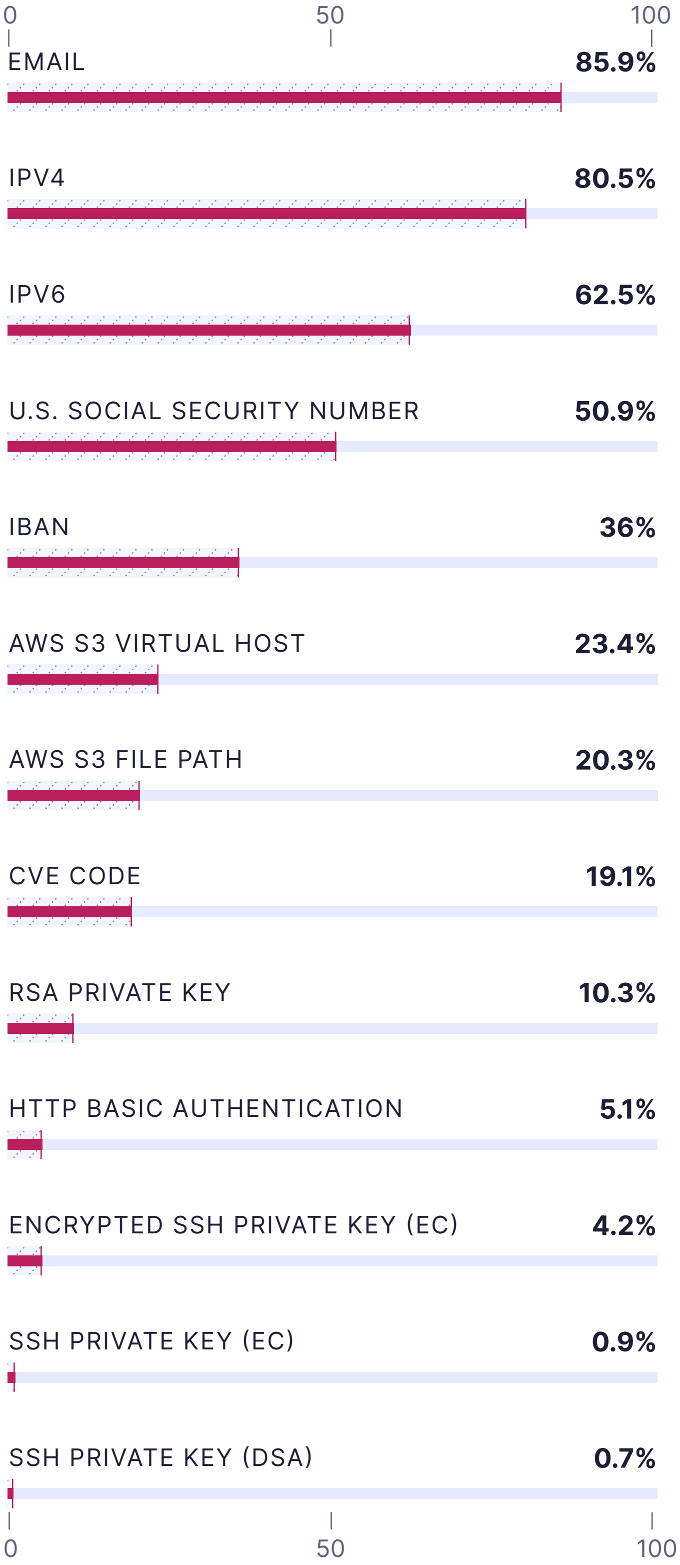
Sensitive information types are files or data types that carry a heightened risk if exposed, for example, by increasing an organization's vulnerability to a cyberattack, fraud, or regulatory or compliance risk. Their exposure can lead to legal, financial, or reputational harm.

51%

of incidents exposed U.S. Social Security Numbers

Figure 7:
Top 10 Sensitive Information Types Most Commonly Indentified in Data Breaches

Percentage of the 1,297 analyzed data breaches that included the sensitive information type



- **Email addresses** were by far the most frequently extracted sensitive information type, appearing in 85.9% of incidents (Figure 7) and at a rate equivalent to 54 email addresses per incident. At such a high volume, email addresses not only present a significant risk of phishing and impersonation, but adversaries could train social engineering models or conduct highly targeted campaigns at scale.

- **U.S. Social Security Numbers** were identified in half of all incidents analyzed (Figure 7) and were the second most prevalent personal identifier. Due to this PII being commonly exploited in identity theft and benefits fraud, they are highly regulated under U.S. law.

- **Cryptographic private keys** (RSA, SSH, DSA) are arguably the most severe from a security posture perspective. While exposed in a smaller number of incidents - 18% of incidents exposed SSH and RSA Keys (Figure 7)- their exposure enables attackers to bypass authentication, access secure systems, and elevate the risk of lateral movement and credential forgery.

- **IBANs** were identified in 36% of incidents (Figure 7). Leaked IBANs, although not sufficient alone for account takeover, are widely used in automated financial transactions and can enable fraud through techniques such as mandate scams, payment redirection, or synthetic identity construction.

- **Cloud and Infrastructure indicators**, such as AWS S3 paths and virtual hosts, showed more modest frequencies, featuring in 20-23% of incidents (Figure 7). However, they present a significant risk as they can hint at possible misconfigurations or hard-coded secrets, which may facilitate data exfiltration or the discovery of unsecured cloud storage endpoints. It's long been known that IP addresses offer an easy route for potential secondary attacks, which is concerning given that IPv4 addresses are found in over 80% of incidents, closely followed by IPv6 addresses at 62.5% (Figure 7).

- **Bank Statements** were present in 49% of incidents, while wealth statements were present in 14%. Bank and wealth statements, which typically contain sensitive customer PII, can be misused to commit fraud by helping bad actors impersonate individuals, falsify financial standing, or gain unauthorized access to services like loans or credit.

14%

of incidents involved wealth statements

> *"Breaches do not respect organizational boundaries. A single leak can ripple across the supply chain, exposing weaknesses hidden in plain sight. When fragmented data from partners and providers is pieced together, it reveals far more than intended. Protecting our enterprises means understanding the full ecosystem of exposure and acting before others connect the dots."*
>
> **Yonesy Núñez,** Chief Cybersecurity Risk Officer, DTCC

Blast Radius of a Breach

The exposure of multiple organizations within a single data breach - often referred to as the “blast radius” - has significant implications for systemic risk, regulatory obligations, and reputational damage.

Across the 1,297 breach incidents analyzed, the median number of distinct organizations exposed was 482 organizations. This sits amid a huge range of blast radii. The incident with the highest number of impacted organizations had a blast radius of over 1.73 million affected organizations. The lowest number of impacted organizations from an incident involved a single entity.

Figure 8:

Percentage of incidents with up to 500 distinct organizations exposed

Percentage of the 1,297 analyzed data breaches where the distinct number of organizations impacted was either 500 or below

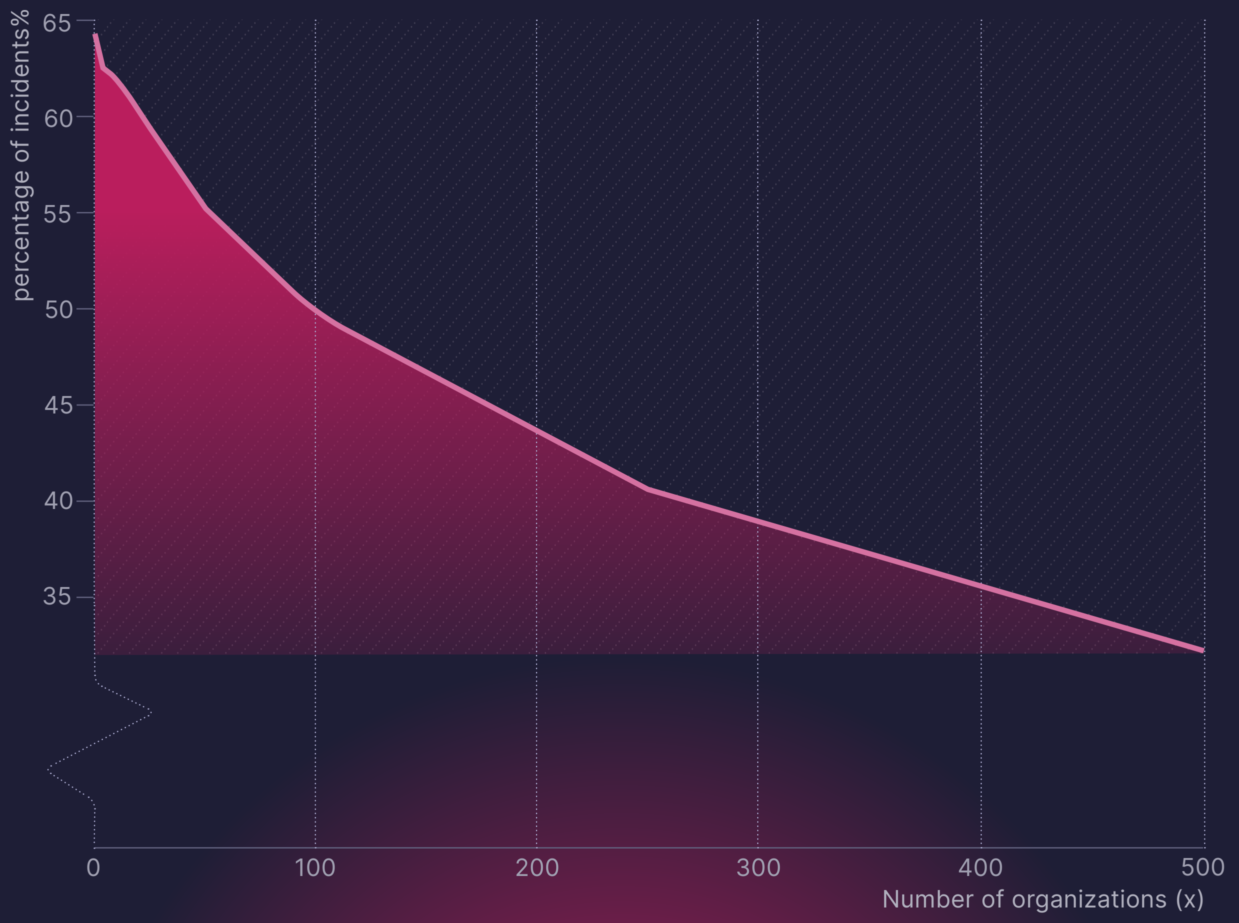
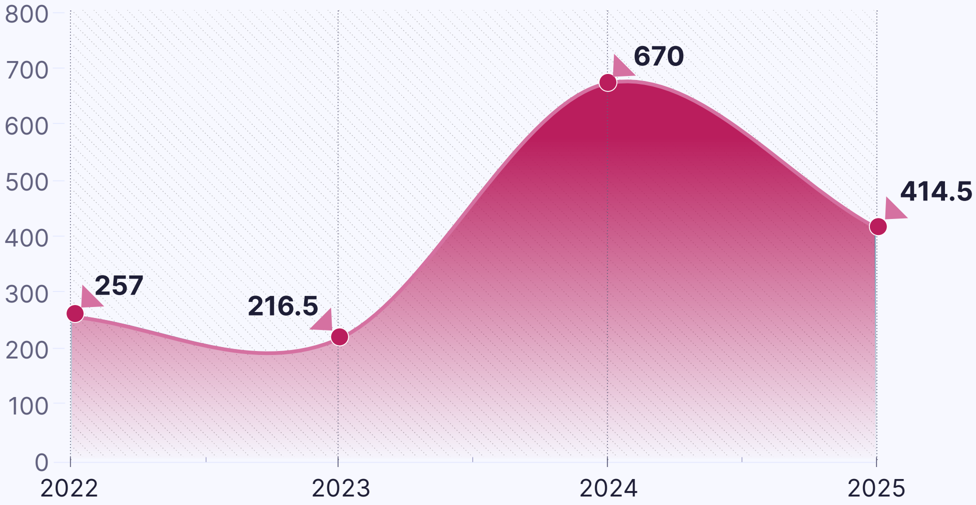


Figure 9:
Average blast radius over time

The average number of organizations impacted by a data breach between 2022 and 2025

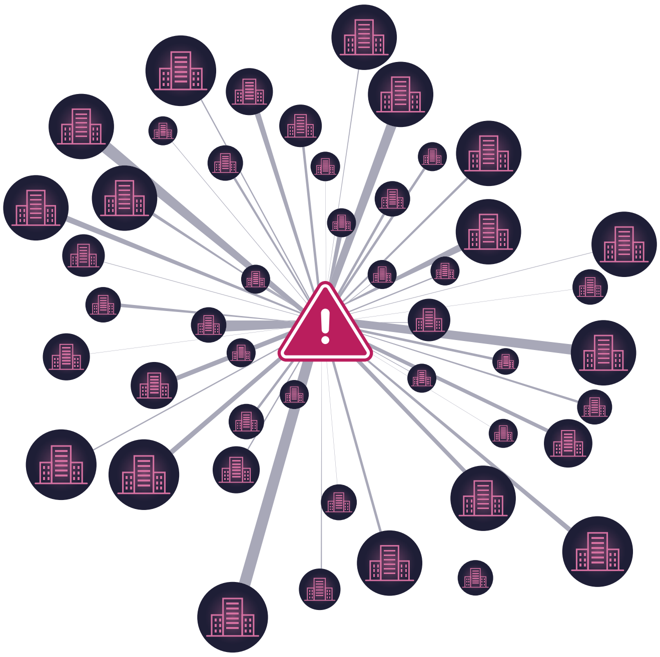


4,468

average blast radius for financial services organizations

> *"We support innovation in assessing the impacts of potential third-party data breaches – innovations that can help the financial services industry protect our customers and maintain our vigilance on their behalf."*
> **Stephen Sparkes**, Chief Information Security Officer (CISO), TD

How Blast Radius impacts Concentration Risk



Organizations assess third-party and vendor concentration risk to ensure that they are not over-reliant on a single or limited number of third parties. However, a limited number of third parties may be connected to many organizations within industry sectors and/or in a geographical region.

To demonstrate the interconnected nature of our supplier networks, Cloudflare, a popular service used to increase the security and performance of business websites and services, experienced an outage in June 2025 caused by one of their suppliers, which had itself experienced an outage. The supplier outage in Google Cloud Platform (GCP) not only impacted Cloudflare (its customer), it also impacted other customers of Google, those hosting in GCP, Enterprise SaaS, and consumer applications such as Discord and Spotify[6].

Therefore, while a high blast radius in a single incident may have a wide-reaching impact across an industry, when you examine the connections between incidents, suppliers and customers, it’s clear multiple incidents have the potential to significantly increase concentration-related risks.

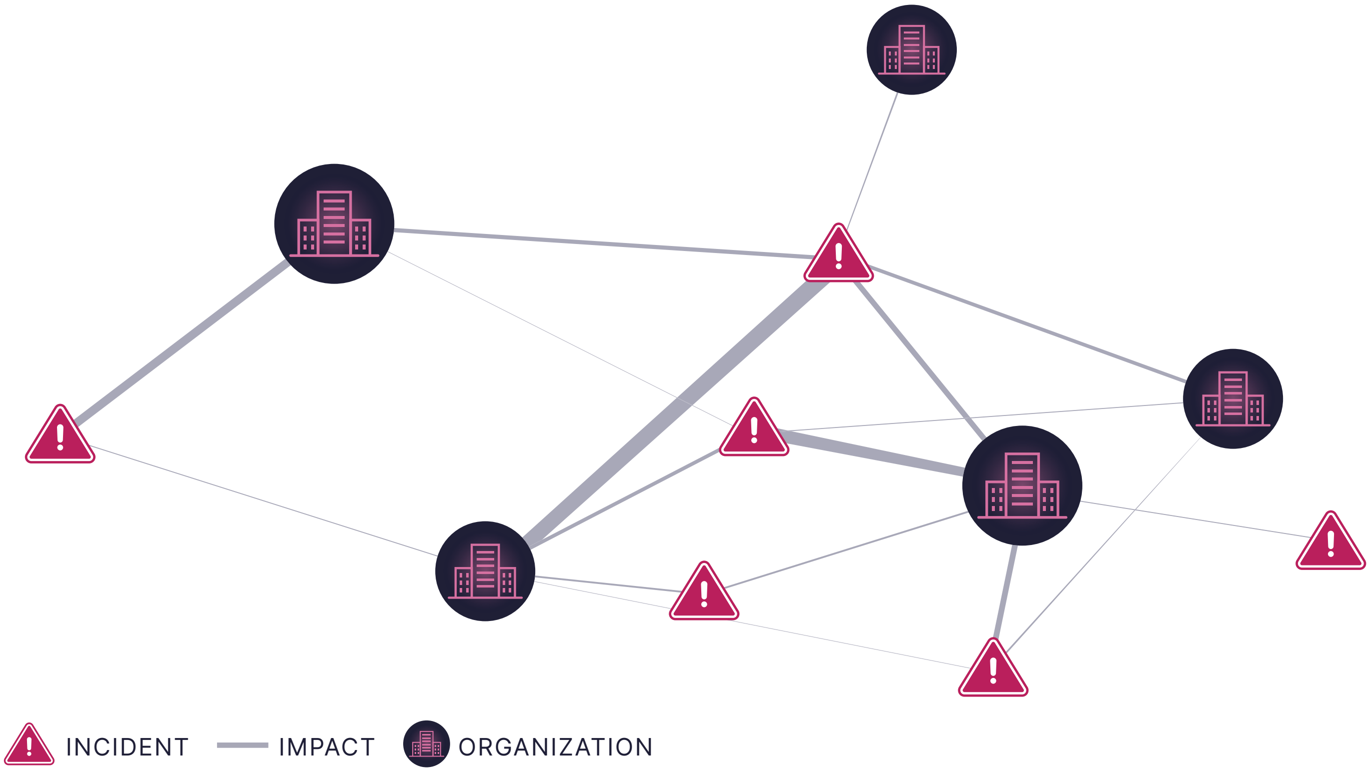
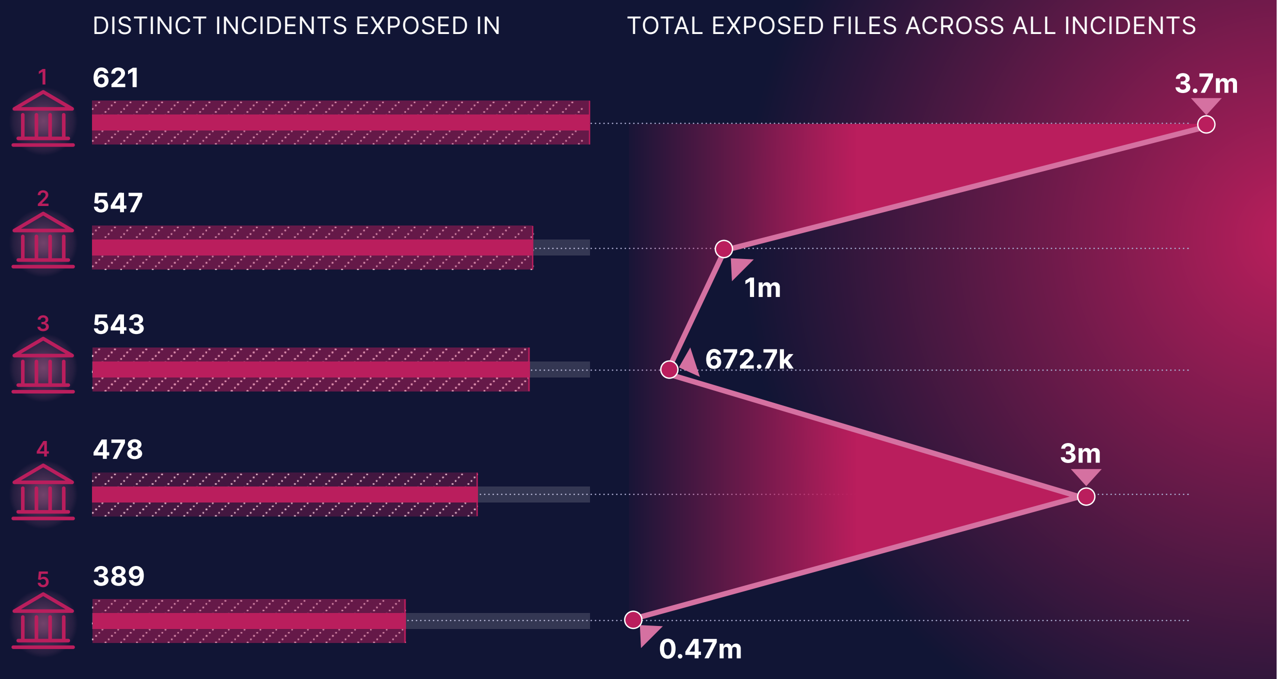


Figure 10:

Blast radius for the top 5 US and European banks

The number of distinct breaches involving the top five global banks across the 1,297 data breaches analyzed, and the total number of their files exposed across those breaches.



> *"We need to stop thinking of breaches as isolated incidents. Each one adds to a growing mosaic of exposure that, over time, paints a detailed picture of how our organizations operate. The real risk lies in the concentration of this intelligence. And not just that held within our systems, but information held across our entire supply chain. To manage that, we need visibility not only into our direct suppliers, but into the full ecosystem of nth-party relationships where we may be unknowingly exposed."*
>
> **Damian Sutcliffe**, Former EMEA CIO, Goldman Sachs

Composition of the Breach

Having examined the full dataset of files from 1,297 breaches, we now turn to the median breach, offering a clearer view of what the mean average breach looks like in terms of content and composition.

The average breach contains

22,647 Files

13.44GB Data

14 File Types

22 File Classifications

482 Organisations impacted

Figure 11:

Typical composition of a breach by file classification

Typical percentage of exposed files per file classification, based on analysis of 1,297 data breaches

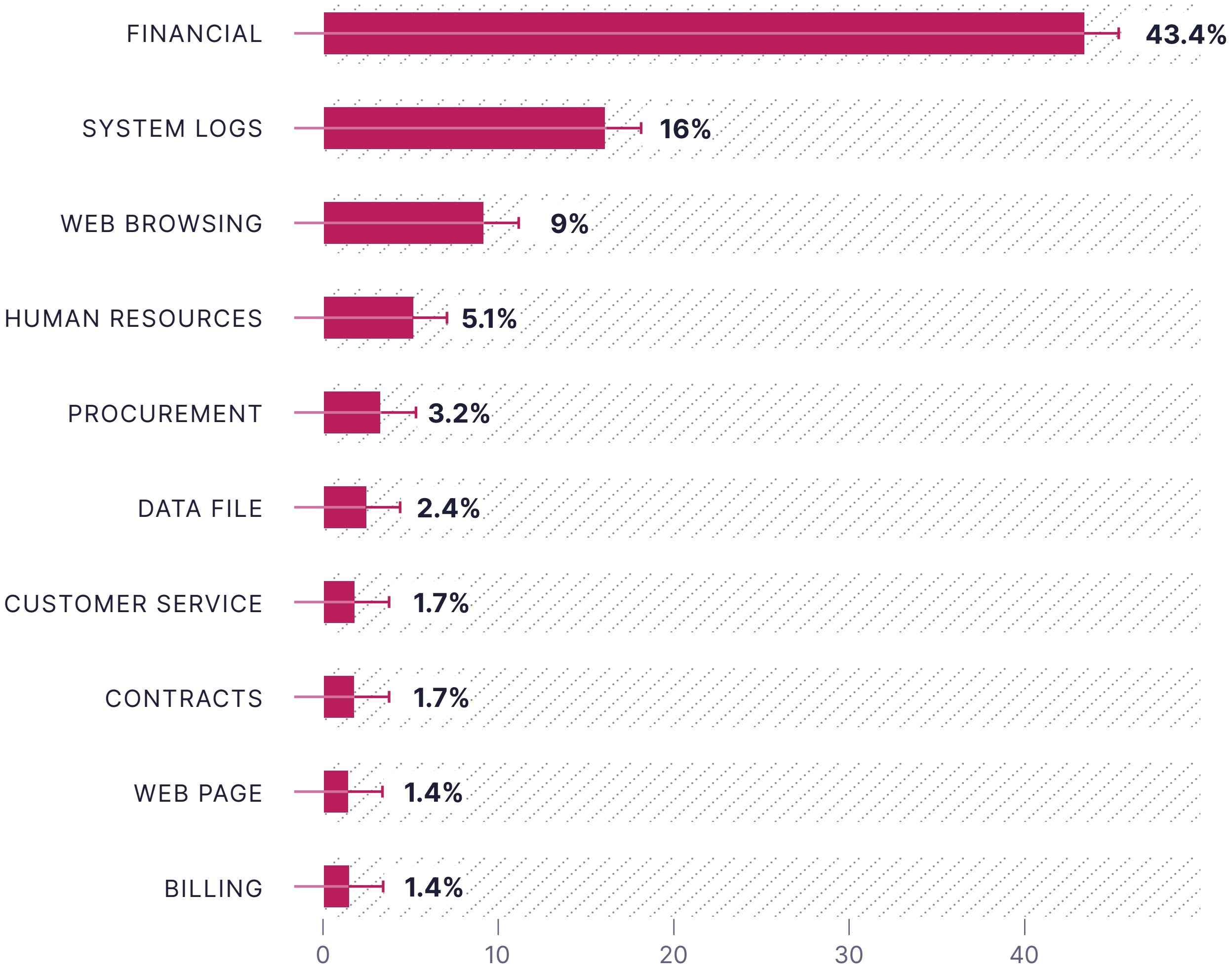
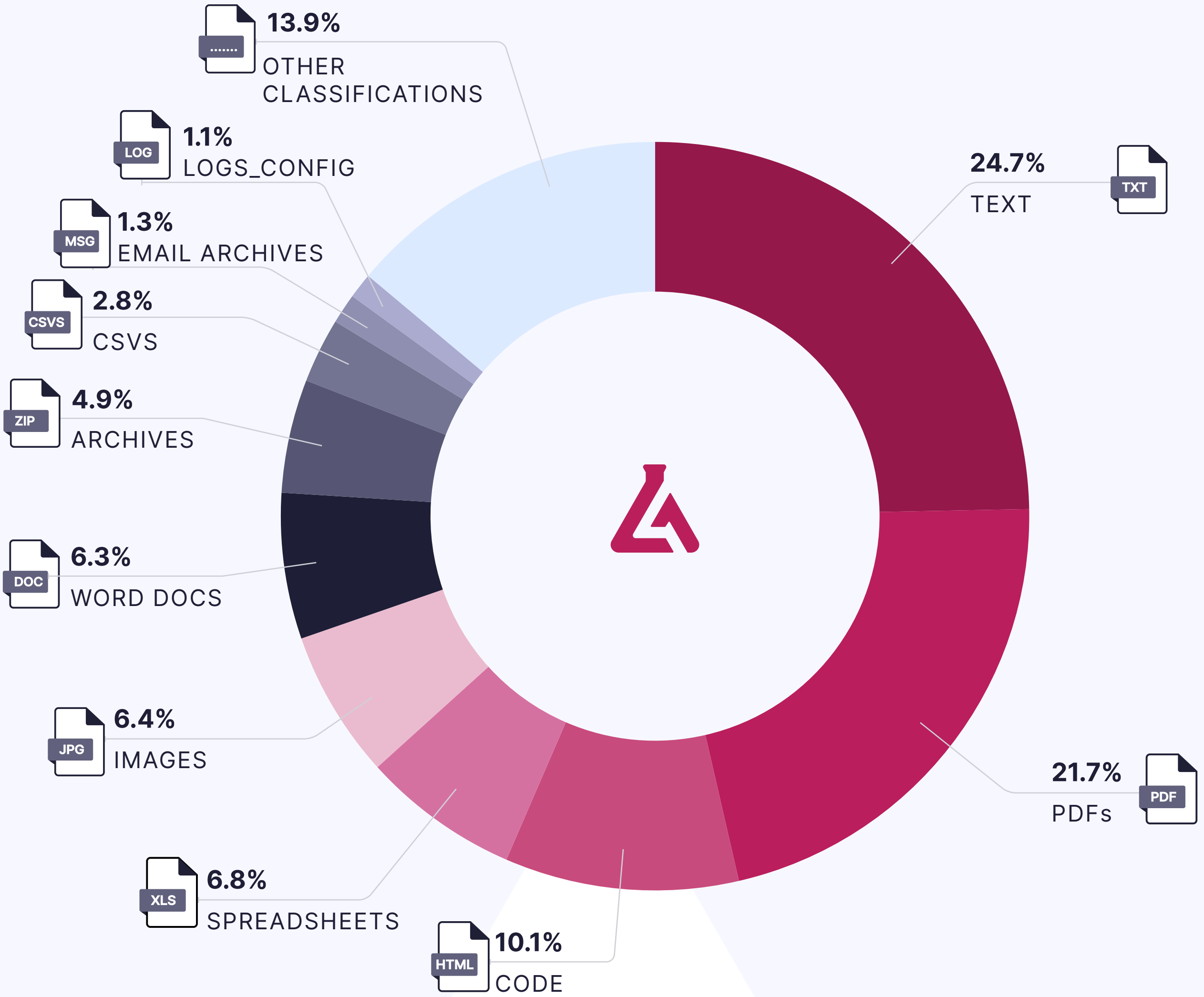


Figure 12:
Typical composition of a breach by file type

Average percentage of exposed file types per breach based on analysis of 1297 data breaches



14
file types are present in the average breach

But breaches materially differ and, in turn, the risks that they pose to an organization. To demonstrate this, we have compiled the incident profiles of two data breaches that occurred.

Incident Profile 1:

With over 70,758 unique organizations implicated, against a mean dataset average of 16,476, this breach is a high "blast radius" event. It suggests a failure at the platform or supply chain level, raising implications for cross-jurisdictional legal liability, reputational damage, and contagion effects within interconnected digital ecosystems.

The breach exposed a significant volume of sensitive information types classed as PII, including 16.7 million instances of U.S. Social Security Numbers (SSN) and 718 million instances of email addresses (Figure 13). As the number exceeds the number of exposed documents, it indicates repeated mentions or multiple identifications per document.

Figure 13:
Composition of the breach of financial institution 1

To ensure our case studies accurately reflect the impact of high-volume breaches, we calculate percentage differences using the mean—which incorporates the skew from these large file events—instead of the median.

Data points	Median	Mean	Financial Institution 1	Percentage difference (mean)
Number of files	21,749	108,842	1,800,000	↑ 1,554%
Number of organizations exposed	482	16,476	70,758	↑ 330%
File Type: Text	784	12,741	40,636	↑ 219%
File Type: PDF	426	32,859	91,8078	↑ 2,695%
File Type: Code	828	18,859	64,876	↑ 244%
File Classification: Customer Service Data	7	2,613	751	↓ -71%
Classification: Financial	7,050	4,6511	1,185,380	↑ 2,449%
Classification: System Logs	210	4,949	71,671	↑ 1,349%
Classification: Human Resources	70	10,471	1,652	↓ -84%
Content Analysis: SSNs	59	98,856	16,725,734	↑ 16,819%
Content Analysis: Email addresses	34,500	6,970,006	718,182,107	↑ 10,196%
Content Analysis: Remote access keys	14	285	0	↓ -100%
Content Analysis: Certificates	12	498	0	↓ -100%
Content Analysis: IBAN	57	4,550	227146	↑ 140,772%

Incident Profile 2:

With 3.9 million exposed files (Figure 14), this breach is in the extreme upper decile of breach magnitude. The coexistence of financial records, personal identifiers like Social Security Numbers, IBANs, and Email Addresses, infrastructure metadata, and source code creates a multifaceted attack surface.

This would enable adversaries to engage in identity theft, financial fraud, infrastructure reconnaissance, and insider impersonation. Notably, this breach reported 82,005 code files (Figure 14) which indicates the exposure of system logic, configuration scripts, or potentially sensitive development assets.

Figure 14:
Composition of the breach of financial institution 2

To ensure our case studies accurately reflect the impact of high-volume breaches, we calculate percentage differences using the mean—which incorporates the skew from these large file events—instead of the median.

Data points	Median	Mean	Financial Institution 2	Percentage difference (mean)
Number of files		108,842	3,900,000	↑ 3,483%
Number of organizations exposed	482	16,476	3,690	↓ -78%
File Type: Text	784	12,741	163,502	↑ 1,183%
File Type: PDF	426	32,859	2,012,440	↑ 6,025%
File Type: Code	828	18,859	82,005	↑ 335%
File Classification: Customer Service Data	7	2,613	6,752	↑ 158%
Classification: Financial	7,050	46,511	2,324,295	↑ 4,894%
Classification: System Logs	210	4,949	8,273	↑ 67%
Classification: Human Resources	70	10,471	109,755	↑ 948%
Content Analysis: SSNs	59	98,856	685,954	↑ 594%
Content Analysis: Email addresses	34,500	6,970,006	5,847,506	↓ -16%
Content Analysis: Remote access keys	14	285	8	↓ -97%
Content Analysis: Certificates	12	498	23	↓ -95%
Content Analysis: IBAN	57	4,550	8,992	↑ 98%

Understanding the Anatomy of a Breach

Understanding the anatomy of a breach - the specific content and structure of exfiltrated data - is essential to effective risk management.

Breach severity cannot be assessed solely by the number of files or records compromised. It must account for the type, sensitivity, and exploitability of leaked content and, in turn, its implications and how it could be weaponized. To assess materiality, you must look at every file.

Risk Area	Implication / Weaponization
Credential Exposure & Infrastructure	The exposure of .rdp, .ovpn, .pem, and SSH/RSA keys facilitates immediate system access, lateral movement, and persistent backdoors. In multi-tenant setups, this can escalate to supply chain compromise.
Synthetic Identity & AI Abuse	HR files, internal emails, and training materials containing PII and behavioral data are leveraged to train deepfake and voice clone models, enabling impersonation, phishing, and synthetic identity fraud.
Cloud and Infrastructure Attack Surface	Leaked AWS S3 paths and virtual host data allow adversaries to reconstruct cloud architecture and automate the discovery of unsecured buckets or dev/test environments, ideal for staging malware or exfiltration. Leaked IP addresses provide information used to create further attacks, such as denial of service or targeted endpoint attacks (e.g. VPN hijacks and targeted attacks).
Financial Exploitation	IBANs, invoices, and payroll documents are exploited in wire fraud, invoice redirection, and BEC schemes. Even minimal exposure (e.g., 0.01%) poses significant financial risk at scale.
Nation-State or Espionage Use Cases	Leaks involving Military Material, Contracts, Legal files, and System Logs—though low in volume—may contain classified or strategic data, enabling geopolitical leverage, insider targeting, or espionage.
Mass Phishing & Social Engineering	Exposure rates equivalent to 54 email addresses per incident enable hyper-targeted phishing campaigns, especially when cross-referenced with leaked internal org charts, HR files, or client data.
Reverse Engineering of Software/Operations	Leaked source code, configuration files, and logs allow reverse engineering of proprietary systems, discovery of hardcoded secrets, and preemptive exploitation of zero-days before defenders can patch.
Reputational and Legal Blackmail	Leaks containing contracts, legal documents, internal correspondence, and audit files enable “double extortion” or blackmail scenarios, where attackers threaten to leak damaging data to the media or regulators.

If you want to understand what content from your organization has been exposed.

Request a demo of the Lab 1 Exposed Data Intelligence Platform

References

1. FINRA. “Protecting Your Investment Accounts From GenAI Fraud.” Investor Insights, 15 January 2025, https://www.finra.org/investors/insights/gen-ai-fraud-new-accounts-and-takeovers. Accessed 18 July 2025.

2. Duarte, Fabio. “Amount of Data Created Daily (2025).” Exploding Topics, 24 April 2025, https://explodingtopics.com/blog/data-generated-per-day&sa=D&source=docs&ust=1751797023828215&usg=AOvVaw3zrqzL2pp8-2KSzWdVRvip. Accessed 6 March 2025.

3. Guo, X., Zheng, X., Yang, Y. and Yu, Y. (2021). ExSense: Automatically Extracting Sensitive Information from Massive Text Leaks. In: Proceedings of the 30th USENIX Security Symposium. USENIX Association.

4. Stempel, Jonathan, and Jonathan Oatis. “23andMe settles data breach lawsuit for $30 million.” Reuters, 16 September 2024, https://www.reuters.com/technology/cybersecurity/23andme-settles-data-breach-lawsuit-30-million-2024-09-13/. Accessed 8 July 2025

5. “23andMe fined £2.31 million for failing to protect UK users’ genetic data.” Information Commissioner’s Office, 17 June 2025, https://ico.org.uk/about-the-ico/media-centre/news-and-blogs/2025/06/23andme-fined-for-failing-to-protect-uk-users-genetic-data/. Accessed 8 July 2025.

6. McMahon, Liv. “23andMe files for bankruptcy protection.” BBC, 24 March 2025, https://www.bbc.co.uk/news/articles/c9q4r9xy9wro. Accessed 8 July 2025.

7. Thomson, Iain, and Simon Sharwood. Google Cloud goes down, takes Cloudflare and its customers with it, 12 June 2025, https://www.theregister.com/2025/06/12/google_cloudflare_outage/. Accessed 10 July 2025.

Appendix

Lab 1

About Lab 1

Lab 1 is the first platform to apply AI and data science at scale to identify and analyze exposure to data breaches. Its AI Intel Agent continuously scans breached datasets across the surface, deep, and dark web, extracting and classifying exposed files. These are safely previewed within the Lab 1 Platform, eliminating the need to download potentially dangerous files for lengthy manual analysis. Organizations receive AI-generated alerts of exposure and summaries of the information revealed, enabling them to understand and act on their exposure quickly and securely. Backed by information security leaders from Goldman Sachs, Credit Suisse, UBS, and Revolut, Lab 1 has now analyzed over 160 million exposed files.

For more information, visit https://lab-1.com/

Methodology

The dataset used in this study comprises 141,168,340 individual file records sourced from 1,297 ransomware and data breach incidents, all of which are in the public domain and were reconstructed from forensic acquisitions of compromised systems. A machine learning-based pipeline was developed to classify file paths according to their likely semantic content (e.g., Finance, Legal, HR). File extensions were mapped to canonical document classifications, and each classification's presence was aggregated by incident count and total file volume. To detect potentially sensitive information, we used a set of curated tools and techniques executed across the extracted file content. Extracted values were further validated using Python packages and appropriate libraries to enhance measurement accuracy.