

Data Breaches in the News, 2019-2026: Attention, Geography and Tone of Global Breach Coverage, and How It Compares with Research and Invention

Fliar BI · Prism Research Layer · Working paper, October 2026 · Dataset and code: github.com/fliarbi/prism-news-api · Licence CC BY-NC 4.0

Abstract

We measure how data breaches were covered in a global news index from January 2019 to July 2026, using 38,972 de-duplicated articles about data breaches and data leaks, and set that coverage beside 1,646 research papers and 617 patent families on the topic. Six results stand out:

1. **Attention, not occurrence.** Data breaches took a 42% smaller share of news in 2020 than in 2019. Their share recovered to 2.98 articles per 10,000 in 2024 but stayed below the 2019 level of 3.43.
2. **Event peaks.** The largest monthly share was July 2019 (7.04 per 10,000), the month Capital One disclosed its breach. The largest monthly count was July 2024 (1,027 articles), the month AT&T disclosed the theft of call records.
3. **Geography.** The United States holds 49.23% of located data breach news, close to its weight in all news (lift 1.02). South Korea (4.77), Australia (2.35) and Singapore (2.1) are the most over-represented countries; India is under-represented (0.52). Among US states, Maine (18.91) leads.
4. **Shifting attention.** Data breach coverage rose 8.94-fold as a share of South Korea's news between 2023 and 2025. Australia's share peaked in 2022 at 16.0 per 10,000.
5. **Tone and urgency.** 78.5% of data breach articles are negative in tone (all news: 33.6%), and 86% are rated critical priority (all news: 42.6%).
6. **Research and invention.** Data breach papers grew from 0.15 to 0.35 per 10,000 papers between 2015 and 2025. Patent families grew 3.38-fold from 2015 to 2024, and China's office held 54.4% of their office filings.

Findings stated for citation

- **F1.** Data breaches took a 42% smaller share of news in 2020 than in 2019 (1.99 vs 3.43 per 10,000 articles).
- **F2.** July 2019 was the peak month by share, at 7.04 articles per 10,000.
- **F3.** July 2024 had the second-largest monthly share (5.25 per 10,000) and the most articles of any month (1,027).
- **F4.** The United States' share of located data breach news (49.23%) matched its share of all news (lift 1.02). The United Kingdom's lift was 1.8 and India's 0.52.
- **F5.** South Korea (4.77), Australia (2.35) and Singapore (2.1) were the most over-represented countries.
- **F6.** 78.5% of data breach articles were negative in tone, 2.34 times the all-news share.

- **F7.** 86% of data breach articles were rated critical priority, 2.02 times the all-news share.
- **F8.** Breach notification appeared in 13.88% of themed articles, and dark-web monitoring in 5.61%.
- **F9.** Data breach papers rose 2.33-fold as a share of all papers between 2015 and 2025.
- **F10.** Data breach patent families grew 3.38-fold between 2015 and 2024; China's office held 54.4% of office filings.
- **F11.** Among US states, Maine (18.91) and Oklahoma (2.97) had the most over-represented coverage; New Jersey (0.11) the least.
- **F12.** Data breach coverage as a share of South Korea's news rose 8.94-fold from 2023 to 2025.
- **F13.** Australia's share peaked in 2022 at 16.0 per 10,000 located Australian articles.
- **F14.** India's share of located data breach news was 0.52 times its share of all news, and fell from 2023 to 2025.
- **F15.** The gap between data breach tone and each country's own news was largest in South Korea (63.1 points) and smallest in India (35.7).

1. Introduction

A data breach is the unauthorised exposure, theft or leak of data held by an organisation. Breaches reach the public mostly through news: a company's disclosure, a regulator's notice, a lawsuit or a hacker's claim. News coverage shapes how customers, investors and regulators judge an incident. Yet coverage is rarely measured as data in its own right. Breach registers record incidents. Bibliometric studies record research. Few sources record attention, meaning how much of the news breaches take, and when, where and in what tone.

This paper and its dataset do that for one topic. They combine three records: a global news index, research papers and patent families. Each comparison uses a denominator, so that growth in the index or in publishing is not mistaken for growth in interest. The companion dataset publishes an article-level sample (9,845 rows) and full coverage tables, so every figure here can be re-derived. The [Prism News API](#) serves the full history.

2. Data and methods

News. The Prism news index holds articles from outlets worldwide; non-English articles are indexed in English translation. Each article carries its opening text (the first few sentences), a primary location, a graded tone label, an editorial priority rating (1–9), themes, keywords and named organisations. We count unique article identifiers; repeated rows for the same article are removed. When this release was built, 92.88% of the index was readable, so all counts are lower bounds.

Matching. An article is about data breaches when its opening text contains one of these phrases as whole words: *data breach(es)*, *data leak(s)*, *personal data breach*, *breach of personal data*, *exposed (the) personal data*, *personal information was exposed*, *stolen data* or *leaked data*. Two broader phrases were tested and dropped. *Security breach* often describes physical security, such as an intrusion at a parliament or an airport. *Hackers stole* most often describes cryptocurrency theft. Two families of matches are excluded:

- video-game and product leaks ("leaked data points to a release date");
- market-research listings for data-leak prevention software.

These removed 113 of 39,085 raw matches (0.29%). The phrase *data breach* alone matched 20,322 articles (52.15% of the total; one article can match several phrases). One reviewer read 40 randomly drawn matched openings. 36 were about the exposure, theft or leak of data, or about preventing, reporting or litigating it (precision 90%). In 28 (70%) a breach or leak was the main subject. The false positives were a software-support review and a marketing piece that named a data breach only as an example, and a product press release that used breaches as a backdrop.

Measures. Coverage over time is expressed per 10,000 articles in the index that month or year, because the index grew from 11,807,177 articles in 2019 to 26,990,006 in 2025. Geography, tone, priority and themes use January 2023 – July 2026. Geography is expressed as lift: a country's share of data breach articles divided by its share of all located articles. State lift uses the same ratio within one country. Countries need at least 100 located articles, and states are discussed only with 50 or more. Place tags derived from ordinary words (714 words such as "Congress", "Earth", "Cloud" or "Mobile") were treated as unlocated in every location field, which removed 15.3% of city-tagged articles.

Research and patents. Research papers are de-duplicated articles and reviews (2000–2025) with *data breach*, *data leak*, *data leakage*, *personal data breach* or *privacy breach* in the title. Repository deposits, the machine-learning sense of "data leakage" (training and test contamination, model inversion) and physical leaks (pipelines, membranes, memory leaks) are excluded. Patent families (2005–2024) use a similar title rule, are counted once per filing office, and exclude memory, fluid and battery leaks.

3. Results

3.1 How much and when

The index holds 38,972 data breach articles from January 2019 to July 2026. Raw counts rose from 4,048 in 2019 to 7,360 in 2025, but the share of all news did not: 3.43 per 10,000 in 2019, 2.98 in 2024 and 2.73 in 2025 (Table 1). The low point was 2020, at 1.99 per 10,000. That is 42% below 2019. Whether pandemic coverage crowded breaches out is not tested here. The share then rose steadily to 2024, when it stood at 1.5 times the 2020 level.

Year	Data breach articles	All articles in index	Per 10,000
2019	4,048	11,807,177	3.43
2020	3,228	16,230,014	1.99
2021	3,905	17,281,220	2.26
2022	4,448	19,115,515	2.33
2023	5,911	20,484,506	2.89
2024	6,954	23,311,896	2.98
2025	7,360	26,990,006	2.73

Table 1. Data breach articles per year and as a share of the index. 2026 (January–July) is partial and omitted.

Coverage moves in spikes rather than seasons (Figure 1). The highest monthly share was July 2019, at 7.04 per 10,000, the month Capital One disclosed its breach. The second highest was July 2024, at 5.25, the month AT&T disclosed the theft of customer call records; with 1,027 articles it was also the largest month by count. October 2022 (4.25), the month of the Medibank breach in Australia, ranks third.

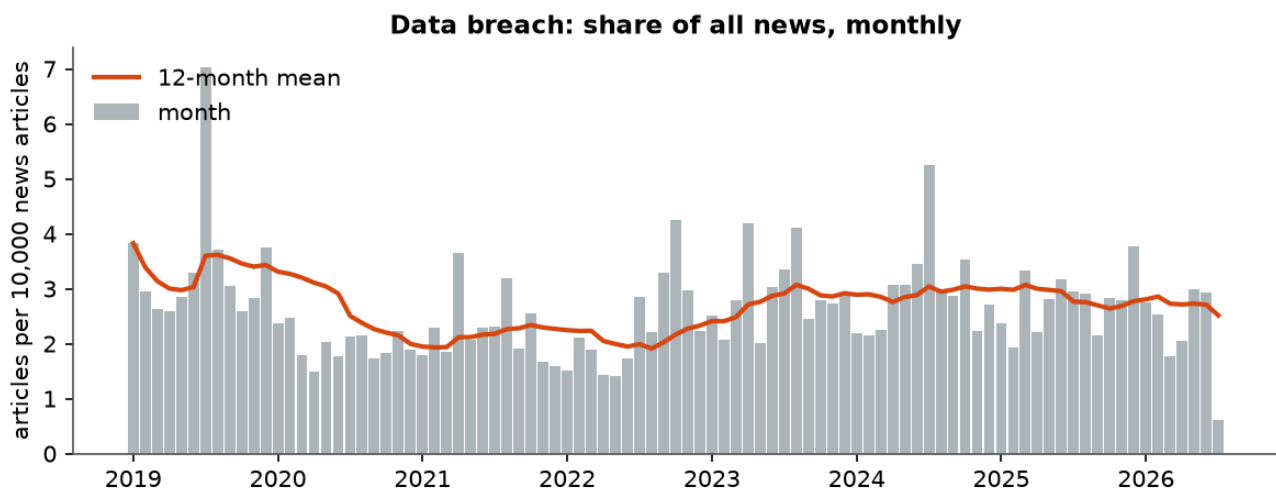


Figure 1. Data breach coverage as articles per 10,000 news articles, monthly, January 2019 – July 2026. Bars are months; the line is the 12-month mean.

Averaged over 2019–2025, July holds 11.05% of a year's articles and February the least (6.95%), against 8.3% for an even spread (Figure 2). The July excess comes from the two large disclosures above, not from a recurring season.

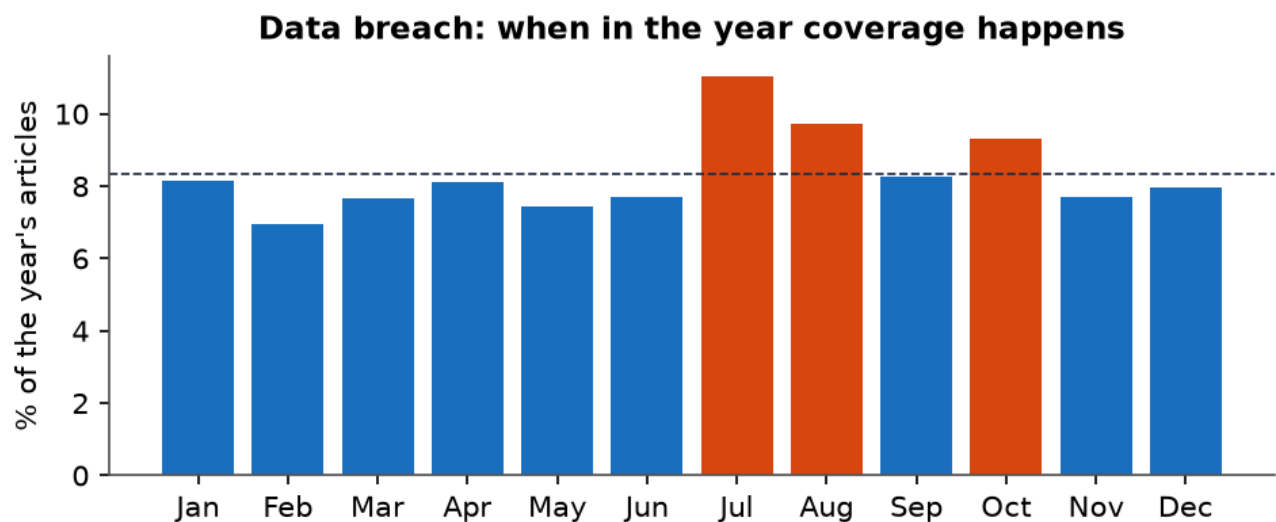


Figure 2. Share of each year's data breach articles published in each month, averaged over 2019–2025. Dashed line = an even spread (8.3% per month).

3.2 Where

Between January 2023 and July 2026, 9,713 data breach articles carried a primary country. The United States holds almost half of them (49.23%), but its lift is 1.02: breaches take about the same share of American news as of news overall. The United Kingdom (1.8) and Australia (2.35) are over-represented. India, the third-largest source of located articles, is under-represented at 0.52 (Table 2, Figure 3).

Country	Located data breach articles	Share %	Lift vs all news
United States	4,782	49.23	1.02
United Kingdom	1,021	10.51	1.8
India	663	6.83	0.52
Australia	490	5.04	2.35
South Korea	278	2.86	4.77
China	268	2.76	1.05
Singapore	182	1.87	2.1
Canada	171	1.76	0.82
France	113	1.16	0.59
Japan	112	1.15	1.09
Brazil	104	1.07	1.24
Nigeria	103	1.06	0.77
New Zealand	103	1.06	1.64

Table 2. Countries with at least 100 located data breach articles, January 2023 – July 2026, with lift against all located news.

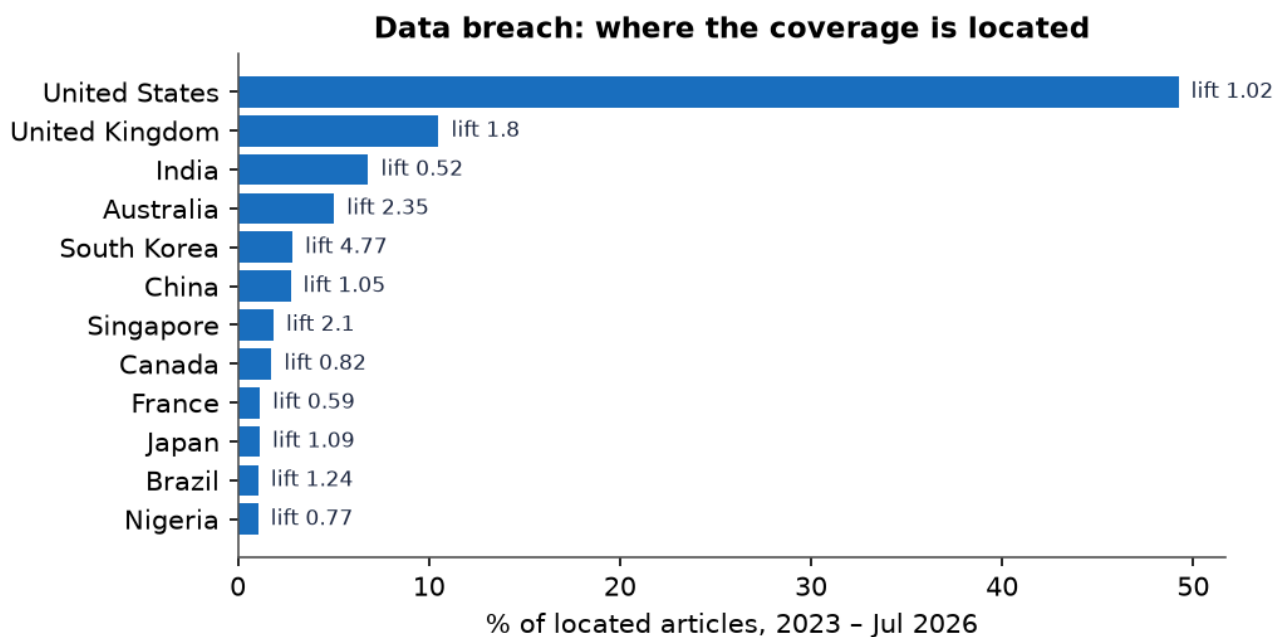


Figure 3. Leading countries by share of located data breach articles, 2023 – Jul 2026, labelled with lift against all located news.

3.3 Geo intelligence: continents, states and shifting attention

The world map (Figure 4) shows the pattern at a glance. Data breaches are over-represented in the news of South Korea, Australia, Singapore, the United Kingdom and New Zealand, and under-represented in India, France, Nigeria and Canada. Most countries are grey because they have fewer than 100 located articles. By continent, Oceania (2.23) and South America (1.42) are over-represented, Europe (1.05) and North America (1.03) sit near parity, and Asia (0.8) and Africa (0.66) are under-represented (Figure 6, left). Asia's low value hides a split: South Korea and Singapore are high, India is low.

Data breach: which countries over- and under-cover the topic

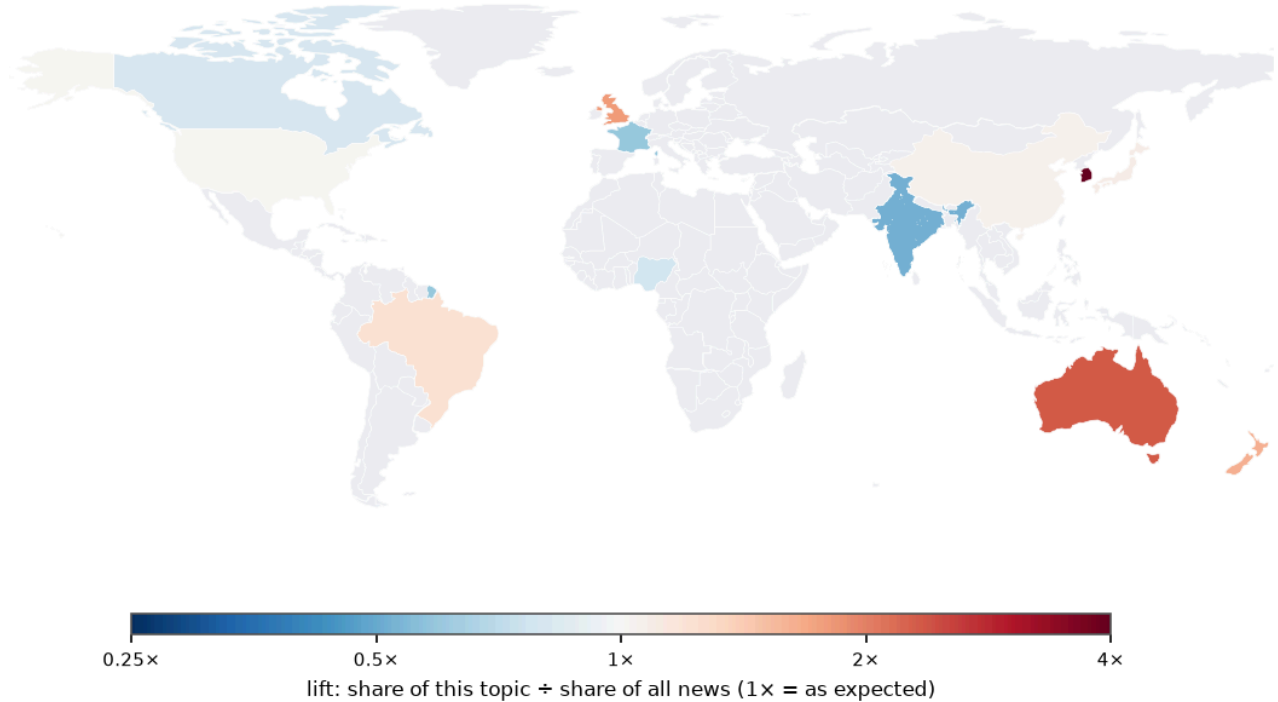


Figure 4. Lift of each country's share of located data breach articles, January 2023 – July 2026 (red = over-represented, blue = under-represented, grey = fewer than 100 articles). India is drawn with its official outline.

Within the United States the picture is sharper (Table 3, Figure 5). Of 3,359 located US articles, measured against each state's share of American news, Maine (18.91) is far ahead, followed by Oklahoma (2.97). California (1.51) is above parity with the most articles of any state (507). New Jersey (0.11), Pennsylvania (0.66) and Florida (0.79) are below their weight in American news. Maine's value may reflect reporting of breach notices filed with state authorities rather than breaches in the state, and most of Oklahoma's sampled rows are wire press releases datelined Oklahoma City; the data do not show more. Massachusetts (1.87) and Alabama (1.85) also rank high, but in the sample many of their rows come from company and regulation names read as towns, so they are not treated as findings.

Highest lift: state	Articles	Lift	Lowest lift: state	Articles	Lift
Maine	273	18.91	New Jersey	68	0.11
Oklahoma	122	2.97	Pennsylvania	133	0.66
Massachusetts	149	1.87	Florida	143	0.79
Alabama	71	1.85	Michigan	63	0.8
Oregon	53	1.58	New York	392	1.0
California	507	1.51	North Carolina	54	1.01

Table 3. US states with the highest and lowest lift in data breach coverage, January 2023 – July 2026 (states with at least 50 located articles). Lift = the state's share of the country's located data breach articles ÷ its share of all located news from that country.

Data breach: states of United States by lift

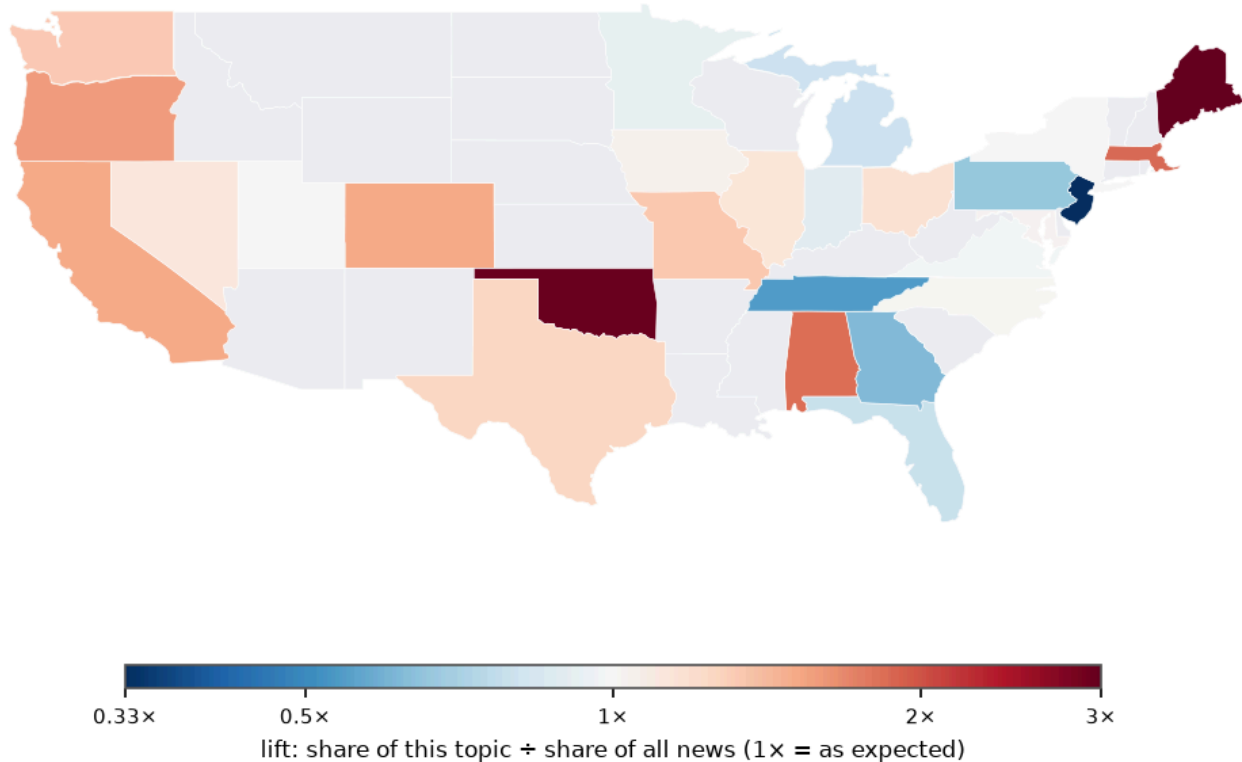


Figure 5. Lift of each US state's share of located data breach articles against its share of all located US news, January 2023 – July 2026.

India has no state map. Only 219 Indian data breach articles carry a state, and no state reaches 50 articles, so state lifts there are too unstable to report.

Attention also moves between countries from year to year (Figure 6, right). Measured as data breach articles per 10,000 of each country's located news, South Korea rose from 2.83 in 2023 to 25.3 in 2025 (8.94 times), the year of the Coupang customer-data breach. Australia rose slightly (6.04 to 6.73, 1.11 times). The United States was unchanged (2.39 to 2.4), while the United Kingdom fell to 0.68 times its 2023 level (5.91 to 4.01) and India to 0.58 times (1.85 to 1.07). Australia's largest year was 2022, at 16.0 per 10,000, the year of the Optus and Medibank breaches.

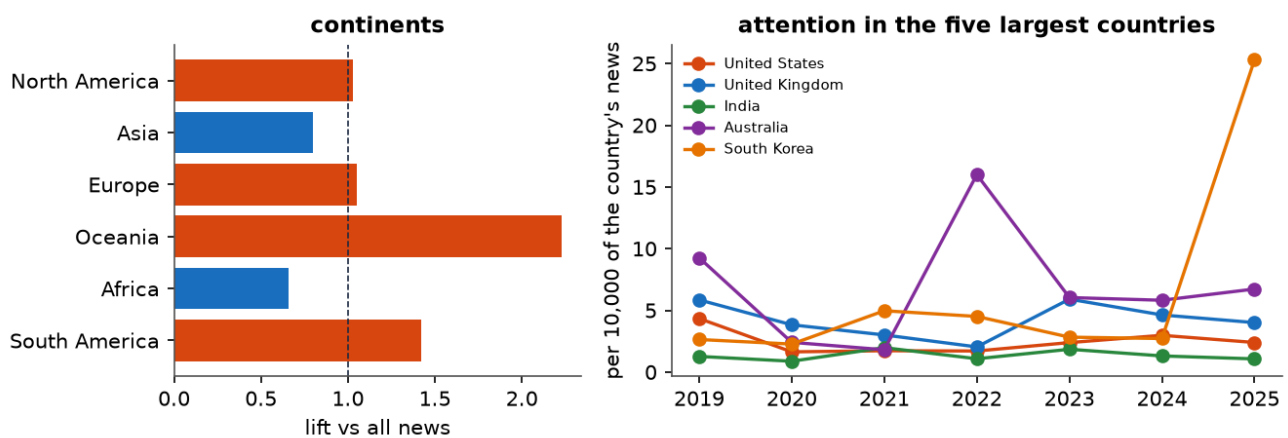


Figure 6. Left: lift by continent, January 2023 – July 2026. Right: data breach articles per 10,000 of each country's located news, 2019–2025, for the five countries with the most articles.

Data breach coverage is negative in tone everywhere, but the gap from each country's own baseline varies. In South Korea 88.1% of data breach articles were negative, against 25.0%

of all South Korean news, a gap of 63.1 points. China (89.6% against 27.2%, 62.4 points) and the United States (87.1% against 33.1%, a gap of 54.0 points) follow. India has the smallest gap: 76.9% against 41.2%, or 35.7 points.

3.4 Tone, urgency and themes

Data breach coverage is far more negative and more urgent than news in general (Figure 7). Among articles with a tone label, 78.5% are negative, against 33.6% across the index (2.34 times). A machine-assigned editorial priority rating puts 86% of data breach articles at critical level (7 or higher), against 42.6% of all news (2.02 times).

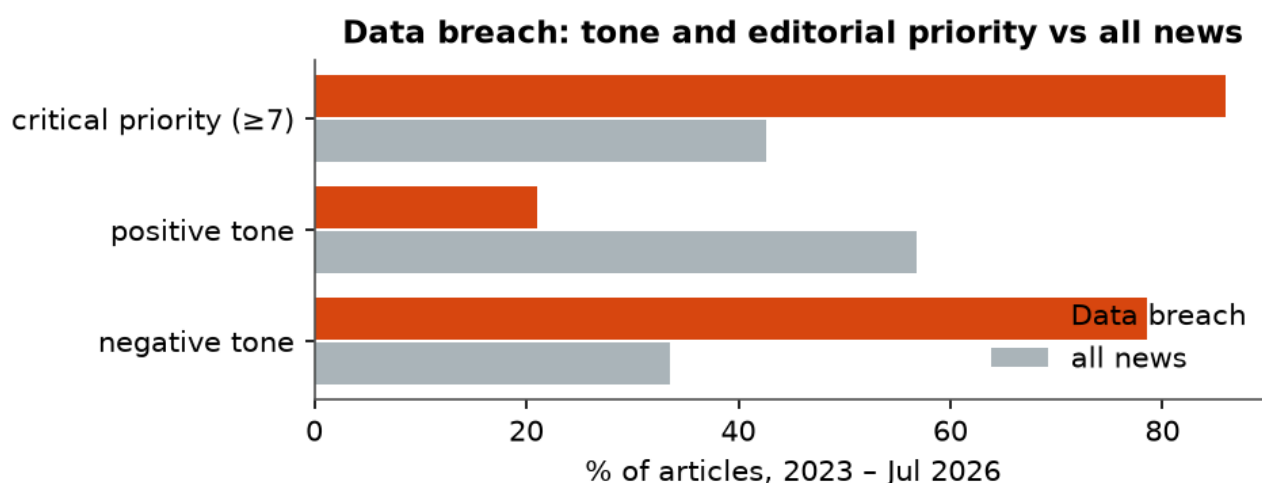


Figure 7. Tone and the share rated critical priority, data breach articles vs all news, 2023 – Jul 2026.

Themes show what data breach stories are about (Table 4). The most specific are about what happens after a breach: notification systems (13.88% of themed articles), monitoring personal data on the dark web (5.61%) and social security number monitoring (5.6%). Each appears more than a thousand times as often as in all themed news. Data loss prevention is the most common theme by count (14.9%).

Theme	Articles	Share %	Lift
monitoring personal data on the dark web	607	5.61	1672.64
data breach notification systems	1,501	13.88	1668.31
medical identity theft prevention tips	277	2.56	1475.71
social security number monitoring tools	606	5.6	1279.88
dark web scanning for stolen information	250	2.31	1239.14
monitoring public records for identity misuse	116	1.07	777.77
data breach execution strategies	285	2.64	740.03
data loss prevention (dlp) in networks	1,611	14.9	636.35

Table 4. Themes most specific to data breach coverage (lift against all themed news), January 2023 – July 2026.

The organisations named most often are technology firms that are both breach subjects and commentators: Google (1,042 articles), Microsoft (875) and IBM (789). Companies with large breaches follow, including AT&T (493), Equifax (441), Coupang (292) and Optus (230). Wire services and technical tokens also rank high, so the list reflects sourcing habits as much as incidents.

3.5 Research attention and invention

Research on data breaches grew faster than the scientific literature. Papers with the topic in the title rose from 0.15 per 10,000 papers in 2015 to 0.35 in 2025, a 2.33-fold increase in

share, while the total literature in the corpus was about the same size in both years (factor 1.0). Between 2015 and 2025, 1,646 such papers were published. The most-cited studies cover healthcare breaches [1], causes and prevention in enterprises [2], the statistical distribution of breach sizes [3], customer responses to breach announcements [4], breach litigation [5], breaches of protected health information in the United States [6], the contextual risk of a breach [7] and corporate crisis response [8].

News attention does not track research attention (Figure 8). Research share rose in most years, while news share fell sharply in 2020 and has moved with large disclosures since.

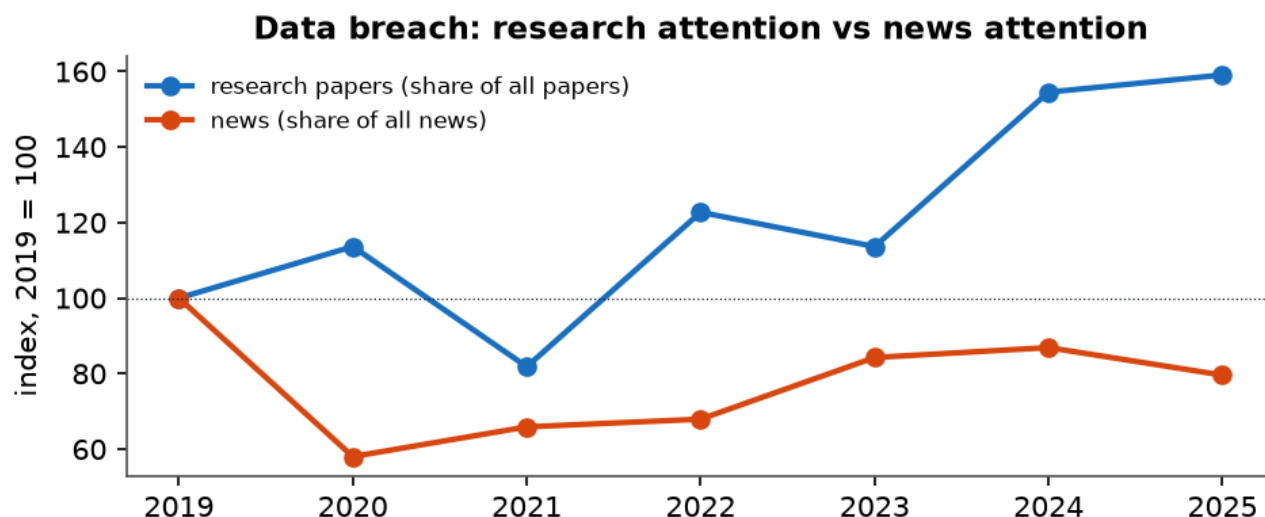


Figure 8. Research papers with the topic in the title (share of all papers) and data breach news (share of all news), each indexed to 2019 = 100.

Patenting is small but growing: 26 patent families with data breach or data leakage in the title were published in 2015 and 88 in 2024 (3.38 times). China's office holds 54.4% of office filings of these families from 2015 to 2024, followed by the United States (25.3%) and international applications (7.4%) (Figure 9). The leading technology classes are electric digital data processing (G06F, 228 families) and transmission of digital information (H04L, 119).

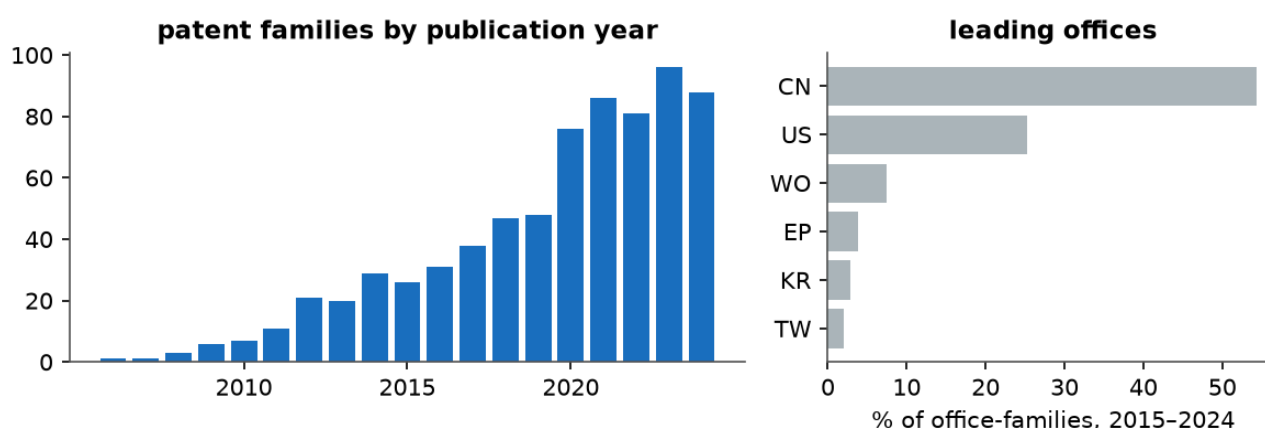


Figure 9. Data breach patent families by publication year (left) and the leading filing offices, 2015-2024 (right).

4. Discussion

Three points matter for anyone who uses news as a record of data breaches. First, coverage measures attention. The 2020 trough and the July peaks show when breaches took more or less of the news. They do not show the number of breaches. Second, raw shares mislead.

The United States leads in raw terms but is at parity relative to its weight in the index, while South Korea, Australia and Singapore give breaches far more of their news. The sharp national rises (Australia in 2022, South Korea in 2025) each coincide with a year of large consumer breaches. The data show the timing, not the cause. Third, coverage is about consequences. The most specific themes are notification and the monitoring of exposed data, which suggests that much of breach news addresses affected customers.

Research, invention and news run on different clocks. Research share grows steadily from a small base. Patent family counts tripled from 2015 to 2024, and filings are concentrated at the offices of China and the United States. News follows disclosures. The dataset lets researchers link the three, for example to test whether heavily covered breaches are followed by regulatory action or research in the affected country.

5. Limitations

- **Index availability.** 92.88% of the index was readable at build time, so counts are lower bounds. Shares and lifts are less affected, because the missing part is spread across topics.
- **Coverage after May 2026.** The index thins after May 2026, so 2026 (January–July) is partial and is excluded from year comparisons.
- **Matching.** Matching uses the article opening. Articles that mention a breach only later are missed, and 10% of matches are not about breaches. Vendor commentary and promotional content are included when they discuss breaches.
- **Machine labels.** Location, tone, priority and themes are machine-labelled. The city field is noisy and is not analysed. Only 41.0% of sampled rows carry a country.
- **Small baselines.** Several country and state values rest on a few hundred articles. India's states cannot be analysed.
- **Place words.** Place tags derived from ordinary words were treated as unlocated (15.3% of city-tagged articles). Real places that share a name with a common word are therefore also unlocated, and some company or regulation names read as towns remain (for example Norton and Dora).
- **Outlet mix.** The index over-represents English-language outlets. This shapes geography and tone.
- **Research and patent counts.** These match titles only, and the topic is a small share of both records.

Data availability

The dataset (article sample, monthly and country tables, research and patent tables, all metrics) is released under CC BY-NC 4.0 with this paper. The sample holds 9,845 of 38,970 matched articles to 31 July 2026. Full article text, headlines, publishers, later articles and other topics are served by the Prism News API: <https://github.com/fliarbi/prism-news-api> (base URL <https://radar-api.fliarbi.com/news>). A worked query: `prism-news trends "data breach" --since 2019-01-01 --until 2026-07-31 --granularity monthly`.

References

1. Adil Hussain Seh et al. (2020). Healthcare Data Breaches: Insights and Implications. *Healthcare*. <https://doi.org/10.3390/healthcare8020133>

2. Long Cheng et al. (2017). Enterprise data breach: causes, challenges, prevention, and future directions. *Wiley Interdisciplinary Reviews Data Mining and Knowledge Discovery*. <https://doi.org/10.1002/widm.1211>
3. Ben Edwards et al. (2016). Hype and heavy tails: A closer look at data breaches. *Journal of Cybersecurity*. <https://doi.org/10.1093/cybsec/tyw003>
4. Ramkumar Janakiraman et al. (2017). The Effect of a Data Breach Announcement on Customer Behavior: Evidence from a Multichannel Retailer. *Journal of Marketing*. <https://doi.org/10.1509/jm.16.0124>
5. Sasha Romanosky et al. (2014). Empirical Analysis of Data Breach Litigation. *Journal of Empirical Legal Studies*. <https://doi.org/10.1111/jels.12035>
6. Vincent Liu et al. (2015). Data Breaches of Protected Health Information in the United States. *JAMA*. <https://doi.org/10.1001/jama.2015.2252>
7. Ravi Sen et al. (2015). Estimating the Contextual Risk of Data Breach: An Empirical Approach. *Journal of Management Information Systems*. <https://doi.org/10.1080/07421222.2015.1063315>
8. Kholekile L. Gwebu et al. (2018). The Role of Corporate Reputation and Crisis Response Strategies in Data Breach Management. *Journal of Management Information Systems*. <https://doi.org/10.1080/07421222.2018.1451962>
9. Fliar BI (2026). *Data Breaches in the News 2019-2026: article-level dataset and coverage tables*. Prism News API. <https://github.com/fliarbi/prism-news-api>