Data Breach Report:
A Study on Global Data Leaks in H1 2018

Conducted by InfoWatch Analytics Center
Table of contents

Table of contents ................................................................................................................. 2
In figures ................................................................................................................................. 3
Summary ................................................................................................................................. 4
Methodology ........................................................................................................................... 5
Report findings ....................................................................................................................... 7
Leak channels ......................................................................................................................... 13
Industry map ........................................................................................................................... 17
Conclusion and findings ......................................................................................................... 22
In figures

✓ In H1 2018, InfoWatch Analytical Center registered 1,039 data leaks revealed (in the media and other sources) worldwide, which is 12% more than in H1 2017.

✓ The data leaks compromised 2.39 billion personal and payment data records, including social security numbers, bank card details, and other critical information, compared to 7.78 billion records YoY.

✓ External attacks were behind 35.5% of data leaks, while 64.5% of the leaks were caused by insiders.

✓ In H1 2018, there were 21 mega leaks, each resulting in the loss of over 10 million records. The mega leaks harvested 97% of all the records compromised.

✓ Inside companies, employees were responsible for 53.5% of the leaks, while executives and other privileged users caused over 2% of the cases.
Summary

This is a report on confidential data leaks in H1 2018 prepared by InfoWatch Analytical Center.

The reporting period saw a series of notorious leaks: Tesla CEO Elon Musk directly accusing a former employee of stealing trade secrets, a leak of UIDAI (Unique Identification Authority of India) database, and Facebook user details shared with a political consulting firm.

Multi-million data breaches affected 92 million accounts of family genealogy and DNA testing site MyHeritage and 150 million users of the MyFitnessPal app. Researchers occasionally discovered high-capacity data storages with poor protection and databases with citizen details posted on the web. E.g. a massive storage of profile data was found on a public but unlisted Amazon S3 storage bucket without a password, allowing anyone to download the data of 48 million users, scraped from their profiles on Facebook, LinkedIn, Twitter, and real estate site Zillow. Another incident involved someone trying to sell data on 70 million Telegram accounts for eight bitcoins on an online forum.

H1 2018 saw almost daily media reports on large confidential information leaks, mentioning such companies as Adidas, Adobe, Airbus, Amazon, AMD, Apple, BMW, Boeing, Canon, Deutsche Post, Goldman Sachs, FedEx, HBO, Intel, Johnson and Johnson, JPMorgan, Lenovo, Lockheed Martin, Marks & Spencer, Microsoft, PayPal, Schneider Electric, Siemens, Sony, Tesco, T-Mobile, Under Armor, Verizon, Western Union, and Xiaomi. Moreover, data breaches hit the CIA, FBI, U.S. Department of Defense and UK Ministry of Defense, International Olympic Committee, People's Bank of China, as well as BitTorrent, GitHub, Skype, Tinder, WhatsApp, and YouTube.

Employee errors and malicious actions come at an increasing cost to both companies that process large data volumes and society in general. As analyzing large volumes of user data is gaining momentum, the number of reported breaches compromising hundreds of millions of user records is growing too. E.g. Exactis, a U.S. marketing firm, left a database that contained close to 340 million individual records on a publicly accessible server. As a result, governments are increasing legal liability for both leaks and attempts to hide information about them.

theverge.com: The U.S. Securities and Exchange Commission imposed a $35 million fine on Altaba (former Yahoo) to settle charges that it misled investors by not informing them of the major hack until September 2016, despite knowing of it as early as December 2014.

The problem of data leaks has long moved beyond industry level, becoming a country-wide or even international issue. That is why user data protection and responsibility of companies that process large volumes of user data are turning into matters of global concern.
In our Global Data Leakage Report, we tried to find and define the most significant trends that would allow security practitioners to look at the bigger picture of data security. Incident (leakage) statistics vividly show which data channel is the most vulnerable now and why, which industry is the most attractive to intruders, and who is more dangerous: an external hacker or a malicious insider.

The report authors are confident that the research results will be of interest to information and economic security experts, journalists, owners and executives of organizations using confidential information (trade, bank, and tax secrets) or other valuable information assets.

**Methodology**

The report is powered by the InfoWatch Analytical Center’s proprietary database updated and maintained by its experts since 2004. The database aggregates publicly available cases of data leaks, which hit business, non-profit (public, municipal) organizations, and government agencies, and resulted from intentional or negligent actions by employees or other parties.

The InfoWatch leak database consists of several thousand registered incidents.

Each leak being logged into the database is classified according to several criteria, such as organization size, the field of activity (industry), damage size, leak type (by intent), leak channel, types of leaked data, and attack vector.

Incidents are also classified by the nature of intruder’s actions. Along with ordinary leaks, the authors also point out ‘skilled’ leaks when either officers abuse their authorized data access to cash in on payment details, insider information, etc., or employees get data beyond a need-to-know basis (and their access rights).

The authors believe the research covers a maximum of 1% of all assumed leaks due to the extremely concealed nature of incidents involving data compromising. However, InfoWatch selected leak classification criteria in such a way that each category group contained sufficient or

1. Data leaks reported by government agencies, mass media, bloggers, as well as message boards and other open sources.
2. Information (data) leak means losing control over information due to an external intrusion (attack), access abuse, or unauthorized access.
3. Data leaks are divided into intentional (malicious) and unintentional (accidental) depending on whether or not a guilty person intends to cause a data leak. The terms “intentional/malicious” and “unintentional/accidental” are equal and used interchangeably herein.
4. Leaks are classified by person responsible (source), with both malicious insiders and external intruders included.
5. InfoWatch Analytical Center ranks companies by size based on the known or estimated number of personal computers (PCs) installed: small companies with up to 50 PCs; medium-size, with 50 to 500 PCs; and large, with over 500 PCs.
6. Information about damage and the number of records compromised is obtained from mass media publications.
7. Leak channel means a certain scenario (cumulative actions by a corporate information system user in relation to hardware or software services) that results in the loss of control over information or a breach of confidentiality. Leak channels are determined only for those leaks caused by actions of an internal offender.
8. Attack vector means a type of intruder’s actions behind data leakage, including intruders who attack company’s web resources and IT infrastructure from the outside to compromise data, and insiders who obtain unauthorized access, misuse confidential information, etc.
excessive number of items (actual data leaks). This approach to survey fielding allows having a theoretical sample, with the findings and trends identified in the sample being representative for the aggregate total.

When preparing diagrams (breakdowns), we excluded from the sample those leaks, which remained Undefined⁹ according to the key breakdown criteria.

In the ‘Industry Map’ section of the report, we also deliberately excluded both leaks with inadequately large amounts of personal data compromised (over 10 million records) and tiny incidents (below 100 records) from the industry-specific map and diagrams to avoid any misrepresentations. The use of a limited sample for diagrams is expressly specified.

In addition, the sampling does not include confidentiality breaches and other information security incidents (such as DDoS attacks) not followed by any data leaks or leaks from an unclear data source (where a compromised data owner cannot be identified).

The authors did not have a goal of either finding the exact number of data leaks or estimating actual or potential financial damage. The report is aimed at identifying trends and pace in the global, industry-specific, and regional data leakage landscape.

---

⁹ For example, breakdown by attack vector (external threats and insider’s activities) does not contain leaks with an undefined vector. The same goes for breakdowns by person responsible, intent and other criteria.
Report findings

In H1 2018, InfoWatch Analytical Center registered 1,039 confidential data leaks (see Fig. 1), which is 12% more than in H1 2017 (925 leaks).

The total volume of compromised data reached 2.39 billion records, including social security numbers, bank card details, and other critical data (compared to 7.78 billion records over the same period in 2017).

Among the logged data leaks, 651 (64.5%) are caused by internal offenders, while 358 (35.5%) of the cases are triggered by intruders from the outside (see Fig. 2). Compared to H1 2017, the share of leaks caused by external attacks went down by 6.7 percentage points (p.p.).
In H1 2018, for the first time in the history of our reports (since 2006), insiders compromised three times more data than external intruders did. The volume of internally compromised data reduced by 32% YoY to 1.6 billion records, while the number of records compromised from the outside decreased 10 times to 0.56 billion records.

The number of registered mega leaks\textsuperscript{11} remained virtually the same (21 incidents, exposing 10+ million records each, vs. 20 cases YoY).

The mega leaks harvested 2.3 billion records or 97% of all those compromised. In addition to mega leaks, the reporting period also saw 15 registered incidents with over a million records leaked.

Both the total data volume compromised during the period and that exposed by external offenders have reduced most likely due to administrative penalties imposed by governments (primarily, in the U.S. and Europe) on companies that failed to prevent leaks. Evidently, many such large penalties issued in early 2018 compelled executives of organizations that process huge data volumes to consider adopting better confidential information security tools.

Here is an example of one such penalty.

\textsuperscript{10} Attack vector means a type of intruder's actions behind data leakage, including intruders who attack company's web resources and IT infrastructure from the outside to compromise data, and insiders who obtain unauthorized access, misuse confidential information, etc.

\textsuperscript{11} Mega leak is a leak compromising 10+ million records of personal data.
securityprivacyandthelaw.com: The French data protection authority CNIL imposed a 250,000 euros (ca. 300,000 USD) fine on the Optical Center for having insufficiently secured personal data of its customers. The CNIL noted that customers could access more than 300,000 documents (mainly invoices) of other customers on the Optical Center’s website rather easily, by entering several URLs in a browser’s address bar. The invoices contained sensitive personal data: surname, first name, postal address, health data (vision correction) and, in some cases, social security number and date of birth of the data subjects.

Just a few years ago, we talked about a fundamental difference between ‘external’ and ‘internal’ leaks. Back then, external offenders were way ahead in terms of the volume of data compromised per leak (‘leak capacity’). In 2018, this difference no longer exists. An average external leak exposes 1.6 million records, while an internal one affects 2.5 million records.

It should be noted that internally-driven incidents due to their nature more often cause adverse consequences for information owners than those committed by external offenders. A typical external leak compromises data of the same type, such as information on online service users or citizens who obtained any particular service. Leaks caused by a malicious insider, employee errors, or misconfigured databases can reveal any information processed within an organization, even the most sensitive data. Therefore, it is very important to see internal leak breakdown by the person responsible (a person who failed to prevent or triggered a leak).

bleepingcomputer.com: The Coca-Cola company reported a data breach incident after a former employee was found in possession of data of 8,000 colleagues on a personal hard drive.

Over the reporting period, 56% of the leaks were caused by current (54%) and former (2%) employees (Fig. 3).
In more than 2% of the cases in H1 2018, company executives (top managers, heads of departments and divisions) were at fault, while over 3% of the leaks (threelfold growth YoY) were caused by contractor’s personnel who were authorized to access restricted information. The share of external leaks reached 38%.

The share of leaks caused by top managers and system administrators who have virtually unlimited access to corporate information again highlights the problem of privileged users.

_Thehindubusinessline.com_: Indian budget airline GoAir filed a suit against its former CEO who joined GoAir’s rival IndiGo, accusing him of stealing confidential information and poaching its employees.

The aggregate share of personal and payment data leaks was 90%, with personal data accounting for 69% of the cases. The trade and state secret leaks increased slightly (by 1 p.p.) YoY (Fig. 4).

---

12 The breakdown of leaks by person responsible does not include incidents when an attacker was not identified (in previous reports, those fell into ‘Undefined’ category).
13 Leak classification by type of leaked data may be difficult. Personal data (full name, social security number, taxpayer identification number, etc.) may not be easily differentiated from payment details. Therefore, the authors put them in one category ‘Personal data and payment details’.
Over the reporting period, we registered several large personal data leaks that caused obvious adverse effects (see examples below). We no longer need to assume how criminals may use huge volumes of information on social media users. H1 2018 clearly saw some high-profile cases of unlawful data processing that attracted the attention of government bodies and the public, which indicates the relevance of scenarios we predict.

**nytimes.com:** In March 2018, Facebook said that the data of up to 87 million users may have been improperly shared with political consulting firm Cambridge Analytica that paid users small sums to take a personality quiz and download an app, which would scrape private information from their Facebook profiles.

**tribuneindia.com:** The Indian journalists reported that unknown hackers broke into AADHAAR, the largest national identity system keeping biometric data of all Indian citizens. Anyone could pay as little as $8 to hackers and access the data.

However, society and regulatory bodies still are not sure whether user data aggregated by social media and similar services should be considered as a separate protectable asset and whether it may be processed without permission of its owners (service users).

In H1 2018, the share of leaks with further data fraud (bank fraud) went down by 1 p.p. YoY, with the share of leaks associated with unauthorized data access (access abuse and internal espionage) also falling down to 5% vs. 7.6% in 2017 (see Fig. 5).
The appearing reports on user data being used for fraud demonstrate that governments of some countries strive to be tough on such criminals:

{oanow.com}: A Phenix City (Alabama) man was sentenced to 30 years in prison for his role in masterminding the theft of thousands of identities for fraudulent tax refunds. He and his co-conspirators filed more than 8,800 tax returns with the Internal Revenue Service (IRS) that sought more than $22 million in refunds. From those, the IRS paid out approximately $9 million.

In H1 2018, 88% of the incidents associated with losing control over information are the so-called typical leaks where an offender does not commit any additional breaches such as access abuse or fraud. Figure 5 shows those as leaks that resulted in data compromising.¹⁴

Conclusion:

In H1 2018, we registered a dramatic reduction in the volume of information compromised by both internal and external leaks, with the former going down by almost one third and the latter dropping tenfold YoY. For the first time since 2006, internal incidents exposed more data records on average and thus had higher leak capacity than external breaches.

This shift in the leakage landscape calls for the rethinking of the existing data security. Until recently, large companies were considered as main targets of external breaches.

¹⁴ Note that any leak compromises data. However, we established this conventional category in order to differentiate between typical and skilled leaks (using leaked data for fraud, bank fraud, unauthorized access, or access abuse).
offenders, suffering multi-million leaks, as their security often failed to combat attacks from the outside. Evidently, administrative measures introduced by regulatory bodies in different countries forced many companies to improve the protection of large data volumes against external threats. Today, the number one problem is internal threats, i.e. malicious actions/omissions by employees and contractors that are authorized to access corporate information resources.

**Leak channels**

In H1 2018, the network channel and paper documents accounted for slightly more leaks YoY, while the share of leaks via email dropped dramatically (see Fig. 6).

---

**Fig. 6. Leaks by channel, H1 2017 – H1 2018**
When it comes to channels, there is a fundamental difference between intentional and accidental data leaks (see Fig. 7). The majority (87.2%) of intentional leaks went through the network channel, with paper documents, instant messengers, and email together accounting for a little over 10% of all malicious incidents.

**Fig. 7. Leaks by channel, H1 2018**
On the other hand, accidental leak distribution by the channel is much more diverse: around 52% through the network, over 18% via email, 15.5% through paper documents, and 6.6% as a result of equipment theft or loss.

To correctly interpret leak distribution by channel, note that the report is based on information from publicly available sources. Therefore, the breakdown of intentional/unintentional leaks herein is true only for the incidents that really occurred (were not prevented) and were subsequently reported to the public.

The report authors note that small shares of intentional incidents over the majority of channels (except network) should be subject to reasonable doubt. In view of technology maturity nowadays, we can assume that a major share of leaks over these channels remained unknown. As we know, many security systems are not that effective when it comes to non-traditional channels (voice data transmission, messengers, and mobile devices). Moreover, intentional leaks here are harder to prevent. It should also be noted that malicious offenders are becoming more and more skillful every year. Many of them usually know how internal data security systems work and try hard not to fall under suspicion.

Unlike intentional leaks, the distribution of accidental incidents herein correctly shows which channels are a real pain in the neck today. Please also note that despite global digitalization of businesses and government bodies, the share of paper documents is still high.

**abc.net.au**: Australia’s ABC channel obtained access to thousands of top secret and classified files found in two locked filing cabinets purchased at a second-hand government furniture store in Canberra.

Showing data leaks by channel is a matter of practical importance. Indeed, knowing how often data is leaked via any particular channel, we can develop threat models (by industry, region, or certain type of data), implement security tools in an organization or industry, and identify the most challenging channels. Obviously, such models and security tools will be different for external and internal enterprise cyber threats.

As we mentioned before, the distribution of accidental leaks by channel provided in this report gives an accurate picture of the most challenging channels. However, when it comes to malicious leaks, both recorded data breaches and compromised data types should be taken into account.

Evidently, whether data is easy to sell\(^{15}\) means a lot to criminals. That is why we decided to analyze malicious leak distribution by channel for the most easy-to-sell data (payment details).

---

\(^{15}\) ‘Easy-to-sell’ data herein means data which can bring profit to an offender both within a short term and at minimum cost. Traditionally, the most easy-to-sell data is credit card details.
In this case, in addition to the network channel, we can see a higher share of email and paper documents (see Fig. 8).

This distribution shows that a threat model for intentional leaks can only be built for a particular data type and should take human factor (behavioral aspect) into account. This is why a malicious data leak prevention system must combine both things at once, focusing on object security and abnormal behavior of legitimate users.

**Conclusion:**

Leak breakdown by channel clearly shows that malicious and accidental leaks are of different nature. Control over data storage, moving, or use alone is not enough to protect information against intentional leaks. A more effective approach is to combine protected object control with the monitoring of user behavior.
Industry map

In H1 2018, the global share of data leaks from government agencies decreased by 4 p.p. down to 20%, while the share of leaks from businesses grew symmetrically to reach 80% (see Fig. 9).

Data leaks were detected most often in hi-tech (21.3%) and healthcare (19.5%), and most rarely in manufacturing (5.2%) sector. Hi-tech companies (online services) were record holders in terms of compromised data records (25.6%). Data leaks from government and municipal authorities exposed some 13% and 20% of the total volume of compromised data respectively, while retail and HoReCa accounted for 10% (see Fig. Ошибка! Источник ссылки не найден.).
These pie charts show the actual leakage landscape and the volume of compromised data by industry. Let’s look at the most attractive and, therefore, the most vulnerable segments.

First of all, the industry attractiveness directly depends on how easy it is to sell data processed by companies in that sector. An industry is less attractive if hackers believe its data protection is sophisticated enough. The number of intentional leaks in a particular industry indicates the above ratio. To put it simply, here is a formula:

$$\text{Number of intentional leaks} \leftarrow \frac{\text{Ease of data sale}}{\text{Alleged information security maturity}}$$

In other words, any industry with a very large share of malicious leaks is likely to have highly easy-to-sell data and poor security, from the criminals’ perspective.

---

The easier stolen information is to monetize, the more attractive the segment is.
Secondly, for proper analysis, we need to compare malicious leaks of one data type (in our case, personal data) across different industries.

In H1 2018, criminals were mostly interested in banks and manufacturing enterprises, with over 60% of personal data leaks in these sectors being of malicious nature (Fig. 11).

Having defined the most vulnerable industries, let’s look at the picture of personal data leakage across all segments, the so-called ‘industry map’. A bubble size shows the total number of compromised records, while their vertical location indicates the number of leaks across the industry\(^1\) (see Fig. 12).

\(^1\) Industry-specific leakage includes personal data leaks where the exact amount of compromised data is known. However, the compromised data volume calculated for the industry excludes mega leaks, i.e. incidents with over 10 million records compromised.
Hi-tech segment (including online services), government and municipal authorities, HoReCa, and healthcare recorded the largest volume of compromised data.

*theinquirernet*: Hackers breached the systems of Norway’s Health South East RHF, with nearly three million patients’ data potentially compromised as a result. Everything indicates that hackers were ‘advanced’ and ‘professional’, spokesperson of HelseCERT, the Norwegian healthcare sector’s national information security center, told Norwegian publication.

Over H1 2018, small and medium-sized businesses (up to 500 PCs) accounted for 34% of all known data leaks worldwide. At the same time, the share of such companies in the total volume of compromised data was some 24% (see Fig. 13).
Fig. 13. Leaks by organization size, H1 2018

Such small volume of data compromised at SMBs (as compared to the number of leaks) confirms a previously revealed trend: data storages are getting bigger and more centralized—something only rather large businesses can afford.

Conclusion:

In H1 2018, banks and manufacturing enterprises were the most attractive and, therefore, the most vulnerable segments (see Fig. 11). The largest volume of compromised data (excluding mega leaks, as per Industry map) is recorded in online services, government and municipal authorities, retail, and healthcare sectors (see Fig. 12).
Conclusion and findings

In H1 2018, for the first time in the history of our reports (since 2006), the volume of knowingly compromised data went down (yet the number of leaks was still going up), and the volume of data exposed by malicious or negligent insiders exceeded those compromised from the outside. At the same time, the reporting period saw no significant changes in the types of exposed data, a number of mega leaks, or leak breakdown by the person responsible, which is likely due to enormous penalties and other administrative measures that forced confidential information owners to dramatically improve data security.

Evidently, less data was compromised mainly due to lower external leak ‘capacity’ (average number of records compromised per external leak). However, such leaks due to their nature differ radically from those by internal offender (‘internal’ leaks). We believe that the number of external leaks and the volume of compromised data were affected by improved data security and increased liability for companies that fail to prevent leaks.

At the same time, in our opinion, internal (rather than external) leaks cause main business risks today. Internal leaks include both unintentional errors and malicious activities by employees and executives, aimed to compromise secured data, tamper with restricted data (including insider information), and use it for fraud.

With enterprise and government data storages getting larger and big data evolving—a trend noticed back in 2017—it is not surprising that there are more and more reported cases when the use of such technologies does harm to the data owners. This raises a number of questions. How to legally regulate large volumes of user data? Who owns a database and knowledge obtained through its analysis? Who is responsible for a data leak?

The analysis of the leakage landscape in the reporting period allows us to assume that the cybersecurity industry is likely to undergo serious changes. This process may involve the adoption of new data protection approaches and shift from internal/external threat classification to intentional/unintentional paradigm. Finally, it would lay the groundwork for developing a concept of hybrid information protection against malicious compromising, as well as boosting security tools capable of monitoring human activity, such as deviations from employee/privileged user behavior patterns.
Leakage monitoring on the InfoWatch website

InfoWatch Analytics Center regularly posts data leakage reports on its website, as well as the most notorious incidents commented by InfoWatch experts.

In addition, the website contains data leakage statistics for past years, available in the form of dynamic diagrams.

Follow the leakage news, new reports, analytical and popular articles via our channels:

- Email
- Facebook
- Twitter
Glossary

Information security incidents in this research mean cases of compromising confidential information as a result of data leaks and/or destructive actions by employees.

Data leak means losing control over information due to external intrusion (attack), access abuse, or unauthorized access.

Destructive actions by employees mean personnel actions that resulted in the compromising of confidential information, including the use of confidential information for personal needs associated with fraud; illegal access to information (abuse of access rights).

Confidential information in this context means information which can be accessed by a limited number of expressly identified persons subject to its non-disclosure to third parties without the consent of an information owner. In this report, the term “confidential information” also includes personal data.

Intentional/Accidental Leaks. Intentional leaks mean an information leakage when a user, who works with information, could foresee negative implications of his or her actions, knew about their illegal nature, was warned about liability, and acted for personal gain or benefit. This results in a risk of losing control over information and/or committing a confidentiality breach. In this case, it does not matter whether such user’s actions actually led to negative consequences or corporate losses.

Accidental leaks mean information leakages when a user neither foresees negative implications of his/her actions, nor acts for personal benefit. In this case, it does not matter whether such user’s actions actually led to negative consequences or corporate losses. The terms “intentional/malicious” and “unintentional/accidental” are equal and used as synonyms herein.

Attack vector means a classification criterion of intruder’s actions behind data leakage, including intruders who attack company’s web assets and IT infrastructure from the outside to compromise data, and insiders who obtain unauthorized access to classified resources and misuse confidential information, etc.

Data channel means a scenario which results in the loss of control over information and a breach of its confidentiality. Currently, we identify eight separate leak channels:

✓ Theft/loss of equipment (server, data storage, laptop, desktop), with information being compromised during maintenance or due to the loss of such equipment
✓ Mobile devices where data leakage occurs because of unauthorized use or theft of a mobile device (smartphone, tablet) when used as part of BYOD paradigm
✓ Removable media loss/theft (CDs, flash drives)
✓ A network where data is leaked via a browser (sending data to personal email, filling in browser forms); unauthorized use of intranet resources, FTPs, and cloud services; and unauthorized information posting on a website
✓ Email, with data being leaked via corporate email
✓ Paper documents which can cause a data leakage if stored or utilized improperly (with confidential information printed, stolen, or taken out)
✓ Instant messengers (data leakage via voice, chat, and video communications)
✓ ‘Non-defined’ is a category used when incident details appearing in mass media do not allow for the leak channel identification.