DBIR

2021 Data Breach Investigations Report

Healthcare

verizon

Healthcare NAICS 62

Summary

Basic human error continues to beset this industry as it has for the past several years. The most common Error continues to be Misdelivery (36%), whether electronic or of paper documents. Malicious Internal actions, however, have dropped from the top three for the second year in a row. Financially motivated organized criminal groups continue to target this sector, with the deployment of Ransomware being a favored tactic.

Frequency	655 incidents, 472 with confirmed data disclosure
Top Patterns	Miscellaneous Errors, Basic Web Application Attacks and System Intrusion represent 86% of breaches
Threat Actors	External (61%), Internal (39%) (breaches)
Actor Motives	Financial (91%), Fun (5%), Espionage (4%), Grudge (1%) (breaches)
Data Compromised	Personal (66%), Medical (55%), Credentials (32%), Other (20%), (breaches)
Top IG1 Protective Controls	Security Awareness and Skills Training (14), Secure Configuration of Enterprise Assets and Software (4), Access Control Management (6)

Since 2019, the Healthcare sector has seen a shift from breaches caused by Internal actors to primarily External actors. This brings this vertical in line with the long-term trend seen by the other industries. This is good news actually, as no industry wants their employees to be their primary threat actor. While one of the top patterns for Healthcare continues to be Miscellaneous Errors, with Misdelivery being most common, at least errors are not malicious in nature (Figure 1). The insider breaches that were maliciously motivated have not shown up in the top three patterns in Healthcare for the past several years. But does this mean they are no longer occurring, or are they still around but we just aren't catching them (like Bigfoot)? Only time will tell.

For the second year in a row, we have seen Personal data compromised more often than Medical in this sector. That strikes us as strange, given the fact that this is the one sector where you would expect to see Medical information held most commonly. However, with the increase of External actor breaches, it may simply be that the data taken is more opportunistic in nature. If controls, for instance, are more stringent on Medical data, an attacker may only be able to access Personal data, which is still useful for financial fraud. Simply put, they may take what they can get and run.

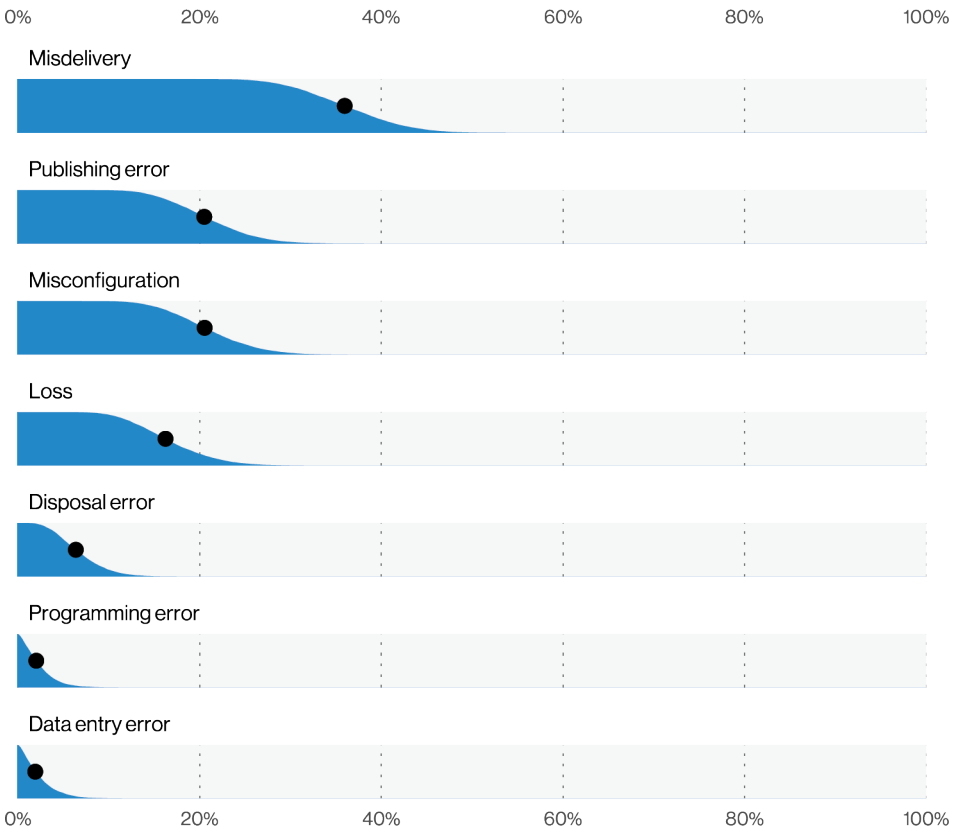


Figure 1. Error varieties in Healthcare breaches (n=70)

Miscellaneous Errors

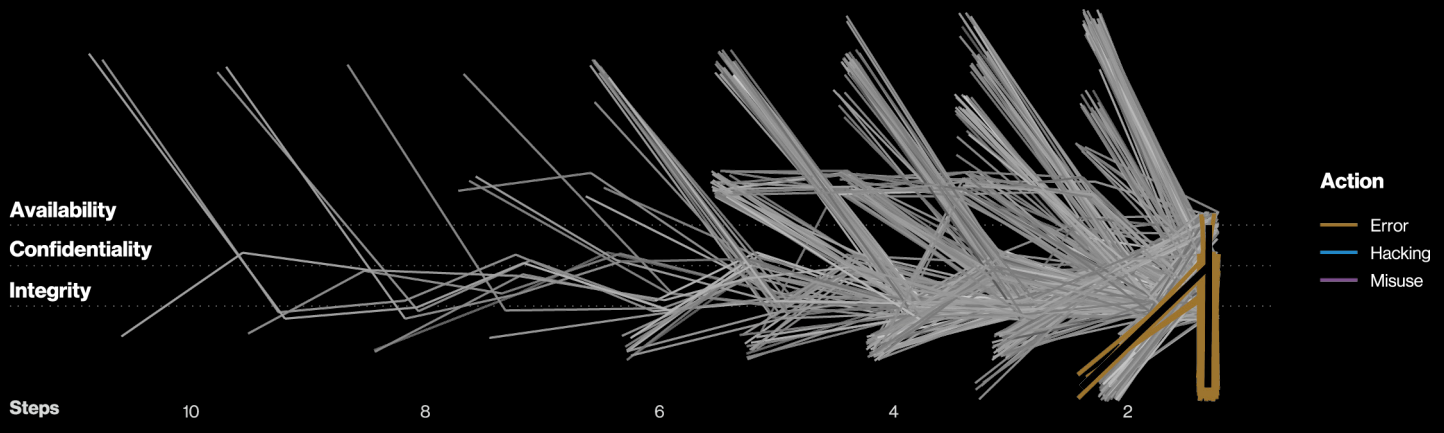


Figure 2. Miscellaneous Errors incident paths (n=126)

Summary

Errors are unintentional actions, typically taken by an Internal actor, but Partner actor errors also occur. Misconfiguration of database assets being found by Security researchers is a growing problem. Employees sending data to the wrong recipients also continues to be a significant issue.

Frequency	919 incidents, 896 with confirmed data disclosure
Threat Actors	Internal (99%), Partner (1%), Multiple (1%) (breaches)
Data Compromised	Personal (79%), Medical (17%), Other (13%), Bank (13%), Credentials (13%) (breaches)

The Miscellaneous Errors pattern should be a familiar frenemy from years gone past. We have included this pattern since the beginning, and the errors have remained constant. What can we really say about this pattern? Humans make mistakes, often at scale. This pattern consists of Internal and/or Partner actors only.

We show the breakdown for Internal actors in Figure 3, and they are relatively intuitive since both system administrators and developers typically have privileged access to data on the systems they maintain. However, the adage of "to whom much is given, much is expected" assuredly applies here. When people in these roles do make mistakes, the scope is often of much greater significance than the foibles of an average end-user.

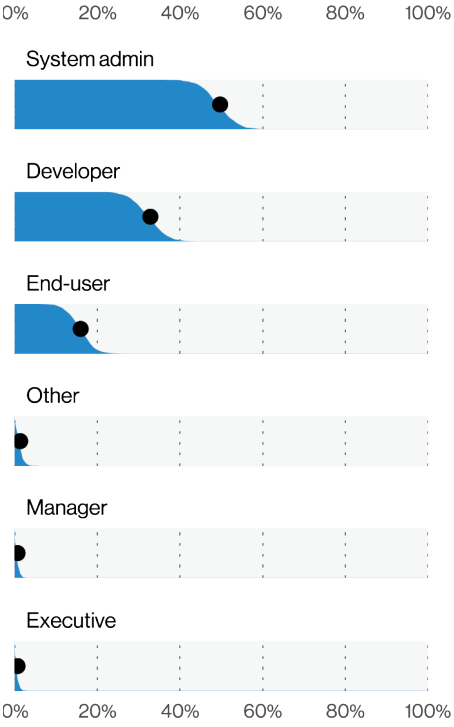


Figure 3. Internal actor varieties in Miscellaneous Errors breaches (n=157)

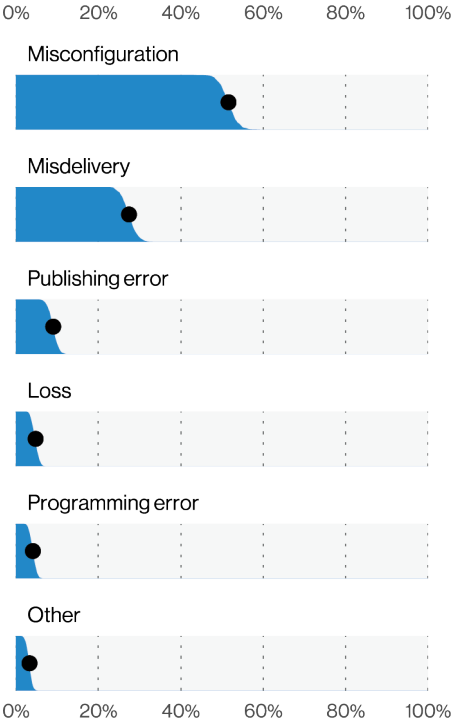


Figure 4. Top Error varieties in Miscellaneous Errors breaches (n=609)

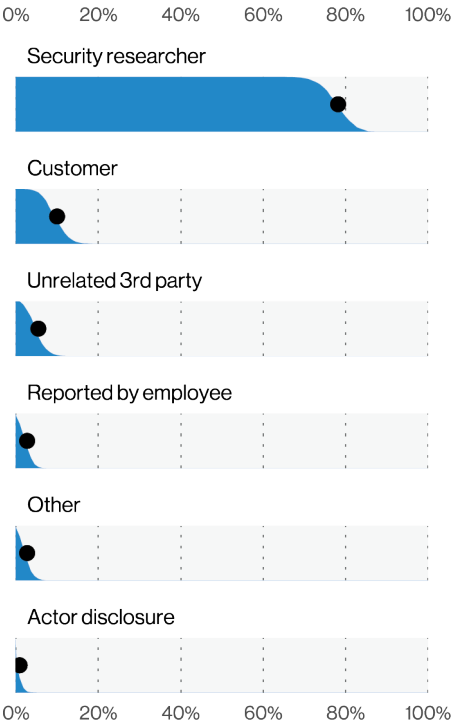


Figure 5. Discovery method varieties in Miscellaneous Errors breaches (n=110)

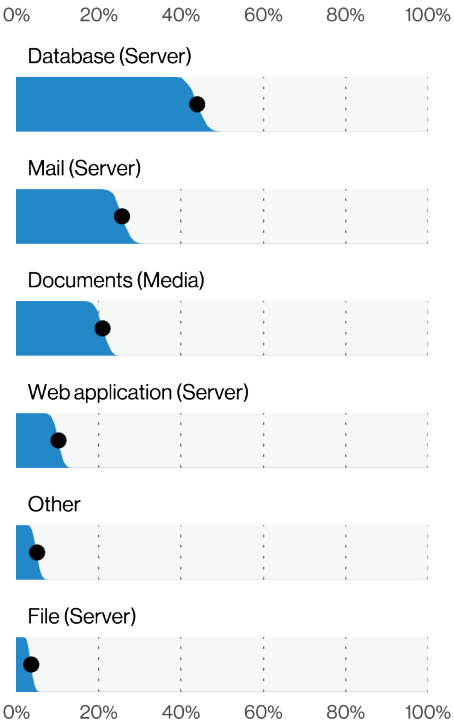


Figure 6. Top Asset varieties in Miscellaneous Errors breaches (n=635)

Sadly, Misdelivery remains alive and well in our dataset, and while a number of these breaches are electronic data only (e.g., email to the wrong distribution list), there remains a significant number that involve paper documents.

Allow us to take you on a tour of pairings—no, not wine and cheese, but Actors and Actions. Given the pairing of sys admins and developers with the Misconfiguration action varieties (Figure 4), you can imagine that this combination can wreak havoc on the confidentiality of an organization's data, or that of their customers' or employees'.

The other pairing we frequently observe is data stores (such as relational or document databases, or cloud-based file storage) being placed onto the internet with no controls, combined with the security researchers who search for them (Figure 5). These rather undesirable combinations have been on the rise for the past few years.

Sadly, Misdelivery remains alive and well in our dataset, and while a number of these breaches are electronic data only (e.g., email to the wrong distribution list), there remains a significant number that involve paper documents (Figure 6). These are particularly common in industries in which large mass mailings are a preferred method of getting information to the customer base. One example being when the envelopes become out of sync with the contents. Many of these events could be avoided by a basic sample check at different points during the mailing process. Nevertheless, we continue to see this occurring regularly, but rarely with any of our bills (those always seem to arrive on time).

Personal data is the most commonly disclosed data type in these cases by a wide margin.

The Assets involved in Error actions run the gamut, from the aforementioned misconfigured databases to physical documents and user devices (Figure 6). A certain portion of this is from Asset loss, although if the device is configured such that unauthorized data access cannot be confirmed, it would be considered an incident rather than a breach.

Personal data is the most commonly disclosed data type in these cases by a wide margin (Figure 7). Medical data is also exposed in this manner, but not nearly as often. The other data varieties represented appear in much smaller quantities.

Just take a gander at that lovely Discovery timeline in Figure 8. See how it flexes all of those breaches discovered within hours and days of the event? Surely this is the story of successful detective controls! Actually, it may be because people usually realize they goofed fairly quickly. But just in case they don't, they have the added safety net of legions of devoted Security researchers out there scouring the internet with their specialized search engines just looking for mistakes.

Figure 7. Top Data varieties in Miscellaneous Errors breaches (n=839)

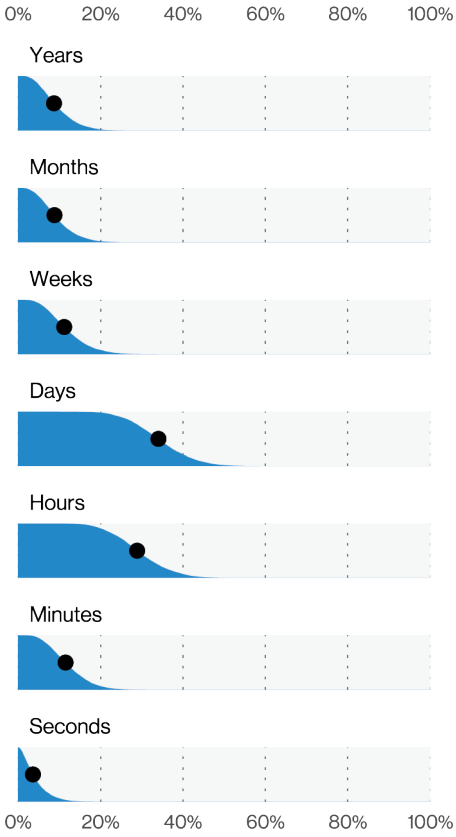


Figure 8. Discovery timeline in Miscellaneous Errors breaches (n=39)

System Intrusion

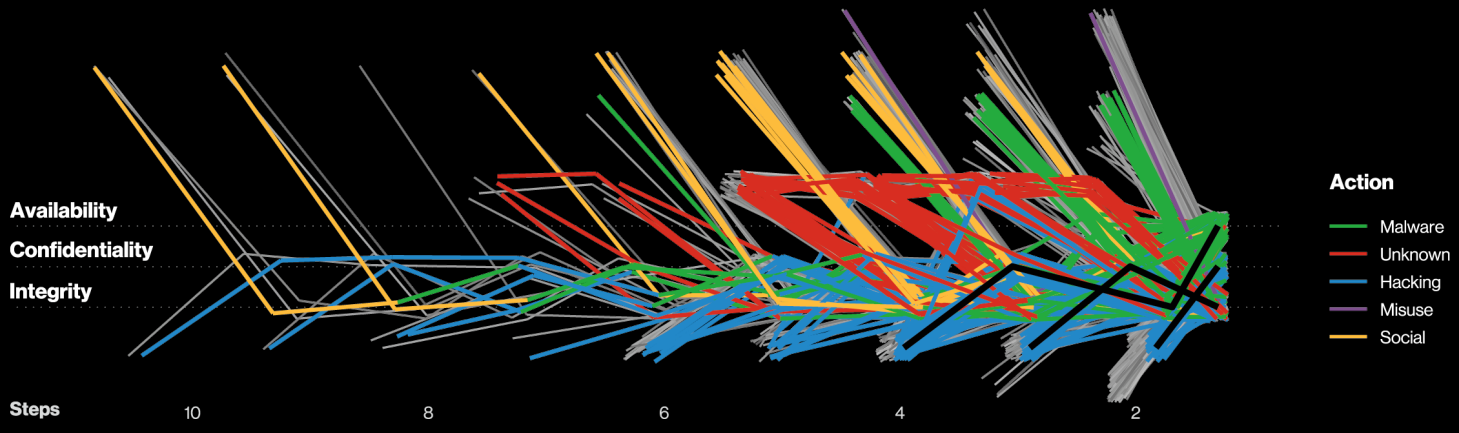


Figure 9. System Intrusion incident paths (n=251)

Summary

This new pattern consists of more complex attacks, typically involving numerous steps. The majority of these attacks involve Malware (70%), usually of the Ransomware variety, but also of the Magecart attack type used to target payment card data in web applications. Hacking (40%) also appears in many attacks and most often consists of the Use of stolen credentials or Brute force attacks.

Frequency	3,710 incidents, 966 with confirmed data disclosure
Threat Actors	External (93%), Internal (8%), Multiple (1%) (breaches)
Actor Motives	Financial (95%), Espionage (6%) (breaches)
Data Compromised	Personal (48%), Other (35%), Credentials (33%), Payment (24%) (breaches)

Not only is this one of the "newer" patterns, it certainly is one of the more interesting ones to talk about, as you'll see in a few. This pattern consists of the more complex attacks, often involving multiple steps as the attackers move through the environment to find the hidden stash of wealth.

In previous years, some of the incidents we discuss in this section would have fallen under the Cyber Espionage pattern, which would have captured most of the hijinks of Nation-states and their affiliated actors looking for Secrets. Still others would have been found in the Crimeware pattern, and lastly, the often-forgotten point-of-sale (POS) server attacks that target servers processing credit cards. Our new System Intrusion pattern is intended to capture those (sometimes only slightly) more elaborate "human-operated" attacks regardless of the motive the actors present. Without further ado, let's get into the details.

Actors in chains

As "trained" data scientists, when we're presented with complex data and detailed charts like Figure 9, representing the event chains associated, we'll go through and quickly triage potential key findings. We pull out gems like "there sure are a lot of colors" and "those lines definitely seem long" to see if they are indeed relevant or statistically significant. In this case, the lines are indeed long, indicating that a lot of the attacks within this pattern involve a variety of different actions done by actors until they finally achieve their goal. Only the Social Engineering pattern has a similar number of steps

involved in both data breaches and incidents. In terms of colors, this pattern has a good combination of mostly Malware events, with some Hacking and a very small smattering of other Action types as a garnish.

Figure 10 describes this differently, and shows Malware being involved in over 70% of the cases and Hacking in over 40%. Lastly, at a very high level, we can tell that the vast majority of the incidents in this pattern are from Financially motivated External actors. The further we dig, the more interesting this pattern becomes.

When we did a deep dive into the data, we found that there are three main "components" that make up this pattern. The first is Ransomware, with 99% of the Ransomware cases falling into this one pattern. The second is Malware in general, and the third is Magecart attacks in which Web applications are compromised with a script to export data as it is processed. Let's go over them.

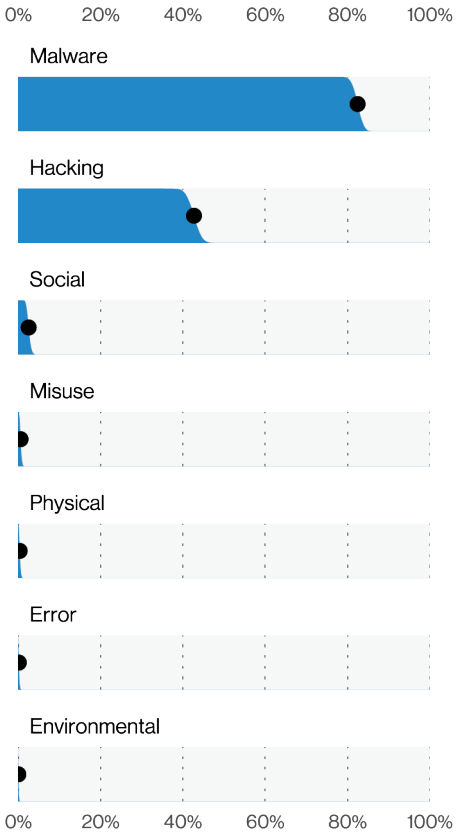


Figure 10. Actions in System Intrusion breaches (n=966)

We're still writing about ransomware?

Unfortunately, this is a section that we've had to write consistently over the last few years and odds are that we'll probably continue to write about this in subsequent reports. This year, we're displeased to report that we've seen yet another increase in Ransomware cases, which has been continuing on an upward trend since 2016 and now accounts for 5% of our total incidents. The novel fact is that 10% of all breaches now involve Ransomware. This is because Actors have adopted the new tactic of stealing the data and publishing it instead of just encrypting it. These attacks have some variety in terms of how the Ransomware gets on the system, with Actors having strong preferences that can be broken into several vectors. The first vector is through the Use of stolen credentials or Brute force. We've seen 60% of the Ransomware cases involving direct install or installation through desktop sharing apps. The rest of the vectors that we saw were split between Email, Network propagation and Downloaded by other malware, which isn't surprising as we found in our web proxy detections dataset that 7.8% of organizations attempted to download at least one piece of known Ransomware last year (Figure 11). For these types of incidents and breaches, we largely see servers being targeted, which makes sense considering that's where the data is located.

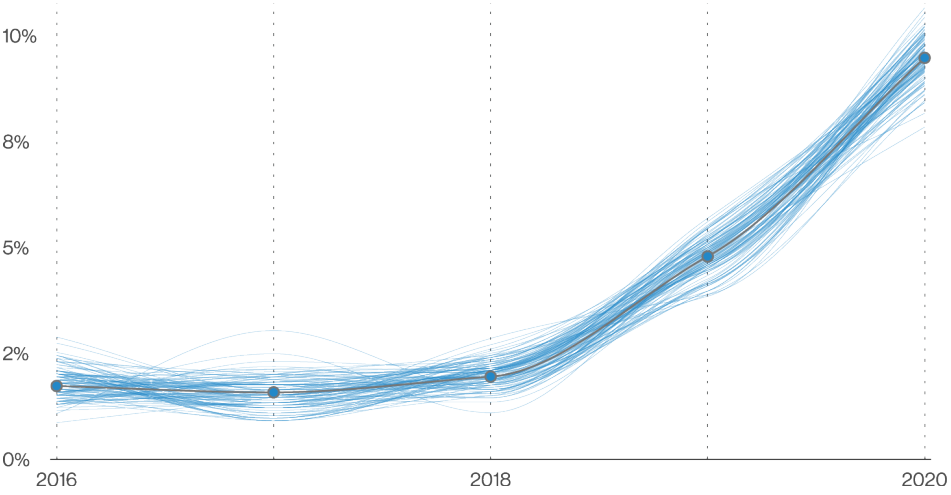


Figure 11. Ransomware in breaches over time

Magecart attacks

The second attack type that we found in this pattern involved the targeting of Web applications processing Payment cards. Now before you interrupt us and ask "but DBIR team, isn't there a whole pattern dedicated to attacks against Web applications?" let us state that the incidents we discuss here are slightly different than those attacks based on a few key components. The biggest differentiator is the subsequent use of Malware to capture Payment card data. In the System Intrusion pattern, we found that of the web servers targeted in this pattern, 60% had malware installed to capture app data and 65% of incidents involved payment cards. These types of attacks follow the trends of attack that we in the biz[1] have been calling Magecart-style attacks based on their original targets. For those who aren't familiar with this attack archetype, attackers will exploit some vulnerability, then use stolen credentials or some other means to access the code of an e-commerce website that processes credit card data. By using that access to the code base or server, they will insert additional code that will ship off the payment data not only to the correct endpoint, but also to their own servers, thereby quietly siphoning off valuable data.

30% of the malware was directly installed by the actor, 23% was sent there by email and 20% was dropped from a web application. While this probably doesn't surprise many people, it does highlight the importance of having a robust defense to cover these three major entry paths for Malware.

General malware

The final breakdown of this pattern involves the general use of Malware that is found on a system. In many of these situations, we may not necessarily know if that Malware would have been used to cause further damage down the road or if it was just there for the sake of being there, doing the kind of things Malware enjoys doing.[2] When we removed the Ransomware cases, we found that 40% of the Malware cases we had left involved the use of C2/Trojans/Downloaders. There was also an interesting split in terms of how the Malware arrived on the system. We found 30% of the malware was directly installed by the actor, 23% was sent there by email and 20% was dropped from a web application. While this probably doesn't surprise many people, it does highlight the importance of having a robust defense to cover these three major entry paths for Malware.

When it comes down to the daily amount of malware incidents, Figure 12 shows that for the majority of organizations, this data has a whole lot of spikiness, which means some days it's probably relatively quiet—until it's not.

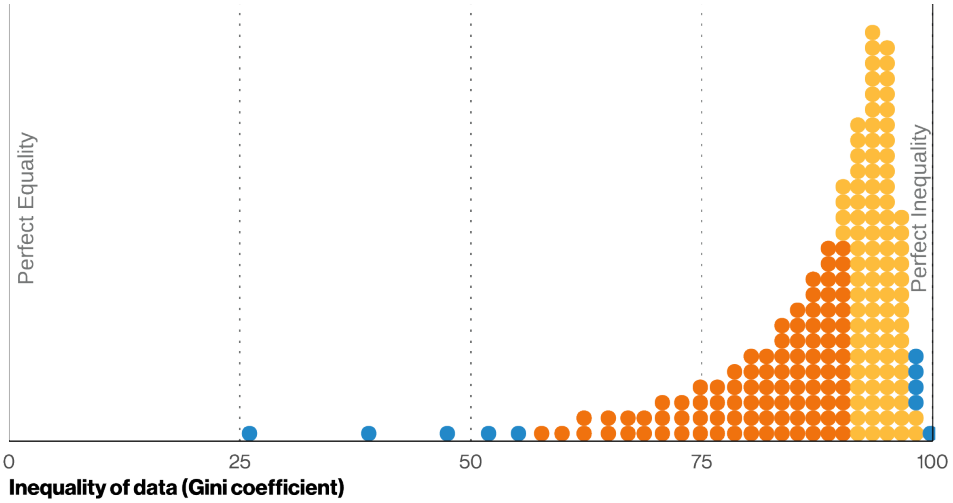


Figure 12. Inequality of Malware per day (n=16,524)
Each dot represents 0.5% of organizations

1 There is no biz like Cyberbiz.
2 Even Malware wants to live its best life.

While we don't necessarily know the severity of these malware events, we do know that data from botnet incidents we reviewed indicates that the majority of botnet infections only compromised three or fewer credentials. So, having malware in your environment, if properly cleaned and handled, probably isn't the end of the world, but it's best to not let it fester.

The big picture shifts.

In the last few iterations of this report, we have mentioned the decrease in the targeting of Payment data. We have continued to see this trend in this pattern. As Figure 13 demonstrates, attackers are less likely to purely target Payment data and are more likely to broadly target any data that will impact the victim organization's operations. This will increase the likelihood that the organization will pay up in a Ransomware incident. As we have often repeated, the monetization through Ransomware seems to have become the preferred method, and the targeting of data will shift to reflect that. The attacks that come out of this pattern impact all of the industries we track at some level, which shows the wide net that these Actors cast to turn a profit.

Attackers are less likely to purely target Payment data and are more likely to broadly target any data that will impact the victim organization's operations. This will increase the likelihood that the organization will pay up in a Ransomware incident.

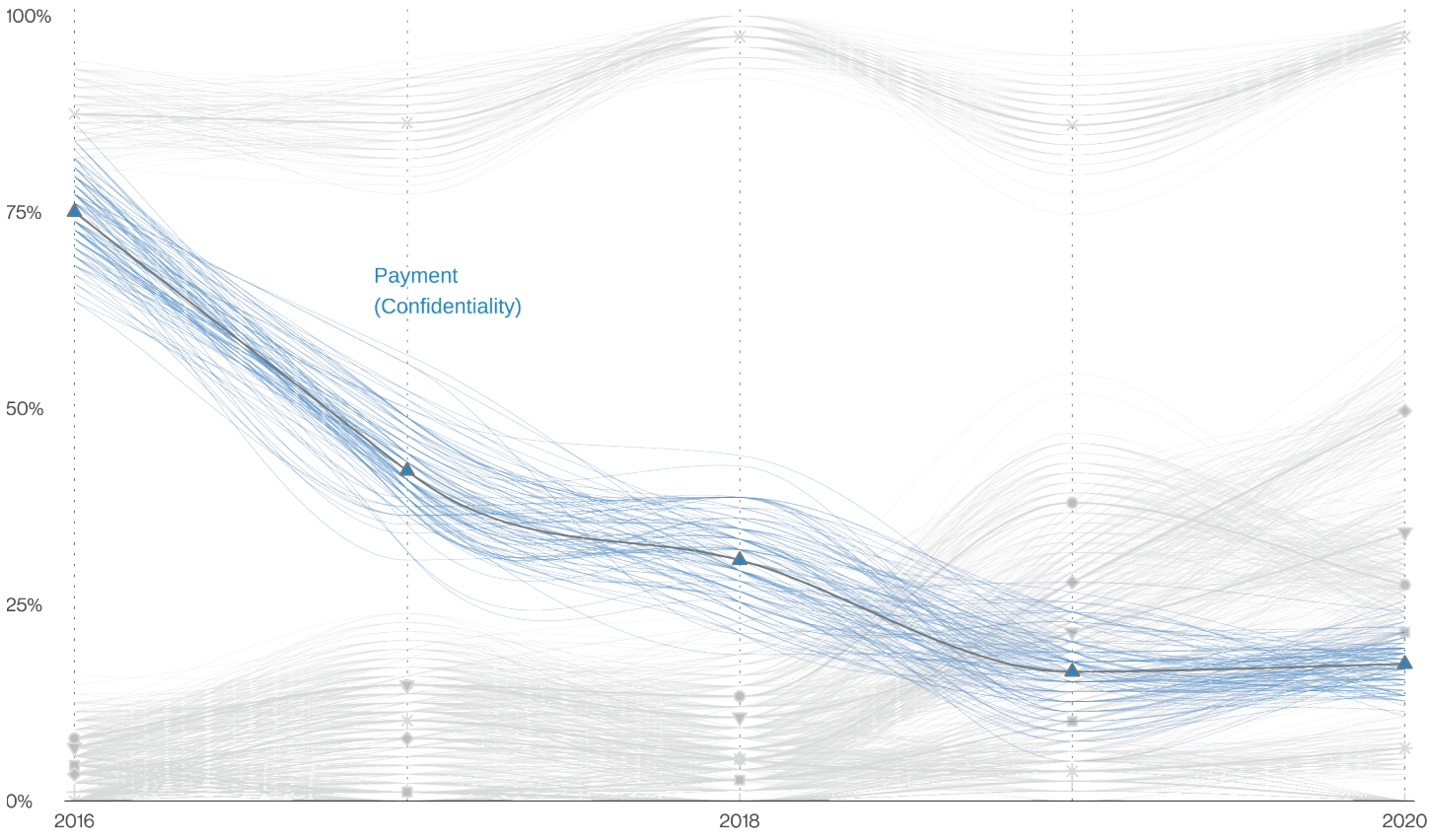


Figure 13. Attribute varieties in breaches over time

Basic Web Application Attacks

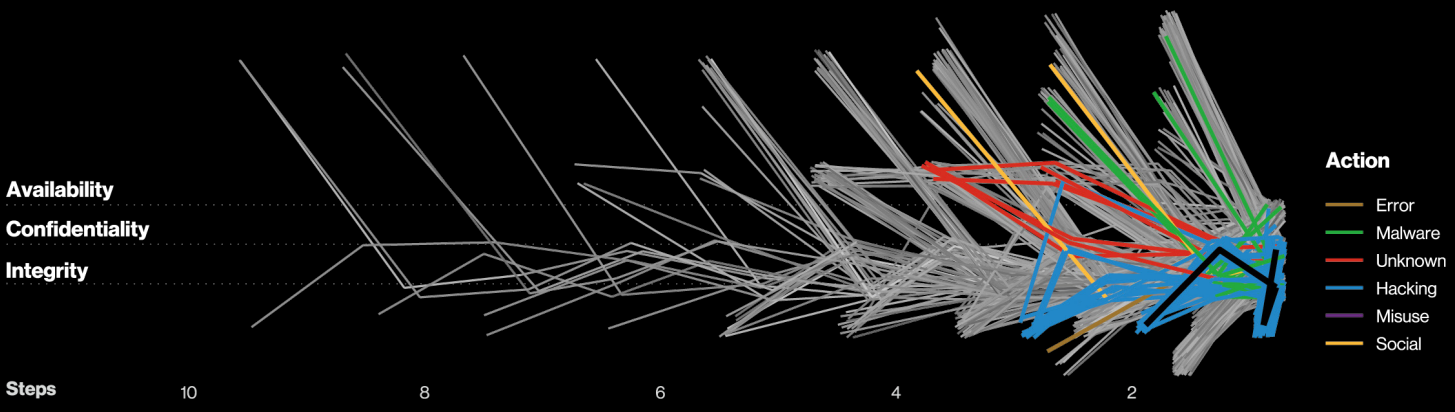


Figure 14. Basic Web Application Attacks incident paths (n=130)

Summary

Basic Web Application Attacks are those with a small number of steps or additional actions after the initial Web application compromise. They are very focused on direct objectives, which range from getting access to email and web application data to repurposing the web app for malware distribution, defacement or future DDoS attacks.

Frequency	4,862 incidents, 1,384 with confirmed data disclosure
Threat Actors	External (100%), Internal (1%), Multiple (1%) (breaches)
Actor Motives	Financial (89%), Espionage (7%), Grudge (2%), Fun (1%) (breaches)
Data Compromised	Credentials (80%), Personal (53%), Other (25%), Internal (12%) (breaches)

Basic Web Application Attacks (or BWAA)—we wanted BWAHA but we couldn't justify the H—is the new and improved version of our trusty Web Applications pattern. We do realize the name is a mouthful, but it better captures the nature of these short and to-the-point attacks that target open web and web-adjacent interfaces (it also freshens breath and whitens teeth). Our other name option was almost as long: Simple Web Attack Group (or SWAG), and perhaps that would have been better, since those attacks are looking for some low-hanging, easily available knickknacks to grab.

While the Assets present in this pattern according to Figure 16 are overwhelmingly represented by the Hacking of Servers, there are a few different sub-patterns encapsulated here, and they are all easy to explain and visualize.

The first sub-pattern covers the Use of stolen credentials and Brute force through a Web application vector to compromise either actual Web apps or Mail servers, as you can see on Figure 14. Almost all (96%) of those Mail servers compromised were cloud-based, resulting in the compromise of Personal, Internal or Medical data.

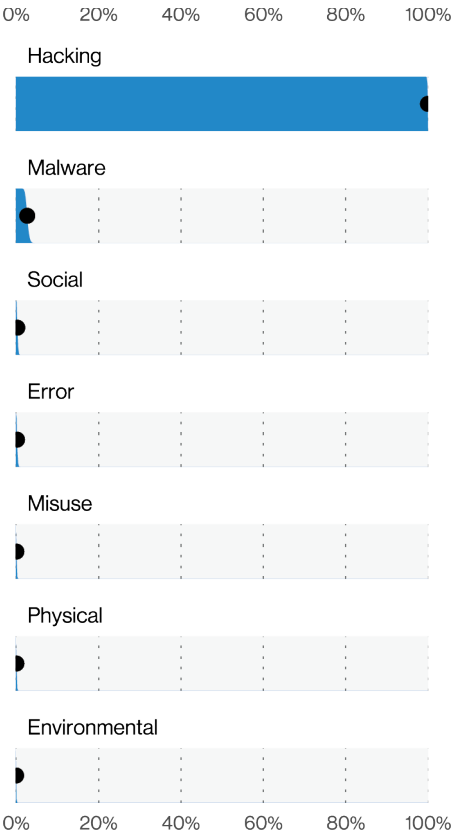


Figure 15. Actions in Basic Web Application Attacks breaches (n=1,384)

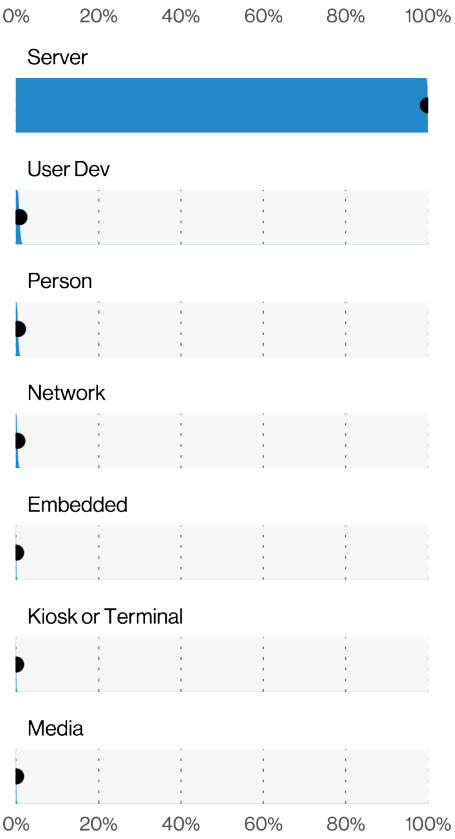


Figure 16. Assets in Basic Web Application Attacks breaches (n=1,369)

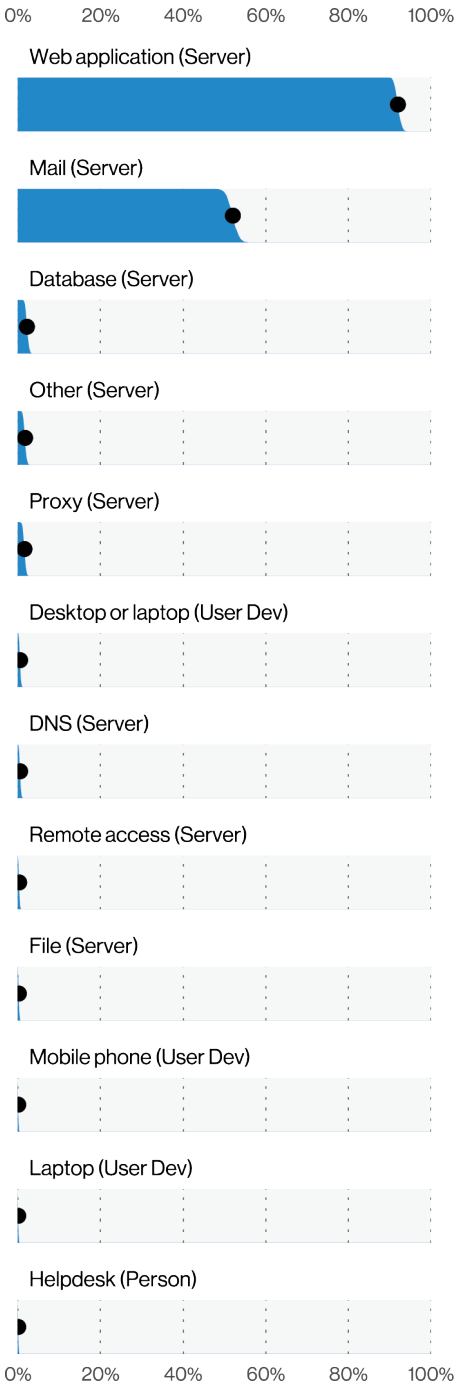


Figure 17. Asset varieties in Basic Web Application Attacks breaches (n=1,324)

All of those Brute force attempts do not happen all at the same time, or even with any predictable regularity.

Astute readers will point out that if using stolen credentials is the leading characteristic of this part of BWAA, how is it differentiated from other threat actor favorites such as Social Engineering and System Intrusion? Glad you asked! It turns out that the credential abuse actions in this pattern were not preceded by any kind of Social attacks as far as the victims were aware. This could mean that either they didn't notice it, or that they were victims of a credential stuffing attack, where the credentials were actually compromised elsewhere and were, sadly, the same on the affected system.

Brute force and credential stuffing attacks are extremely prevalent according to SIEM data analyzed in our dataset. We found that 23% of the organizations monitored had security events related to those types of attacks, with 95% of them getting between 637 and 3.3 billion(!) attempts against them, as Figure 18 demonstrates. This is a very large number at face value, but when you consider the sheer volume of automated bots and worms looking for vulnerable services out there, it feels par for the course.

However, as you may suspect if you have been reading up on the other patterns, all of those Brute force attempts do not happen all at the same time, or even with any predictable regularity. Figure 19 demonstrates that more often than not for the organizations we reviewed, those attacks happened in very uneven intervals. It seems the cost of keeping up with potential credential dumps can't be simplified as something you should do every month or so.

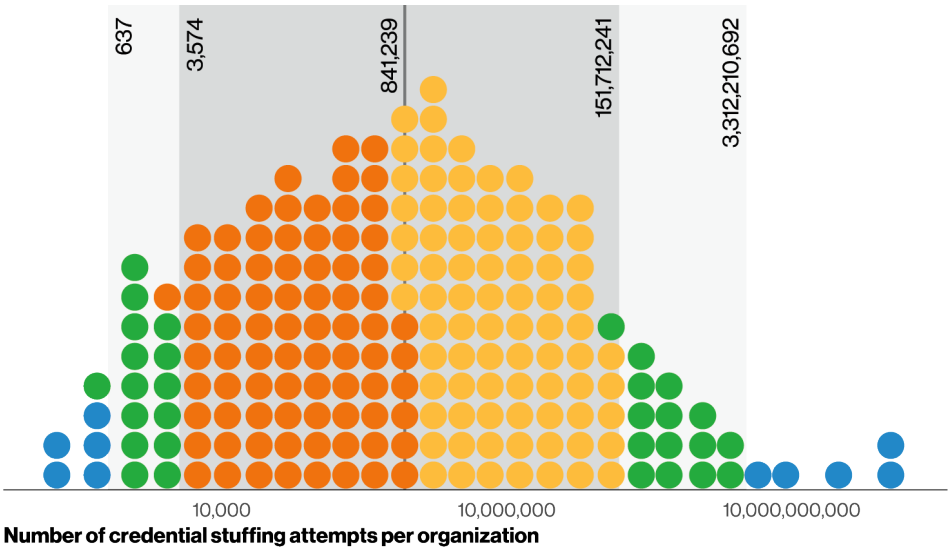


Figure 18. Credential stuffing attempts per organization (n=821)
Each dot represents 0.5% of organizations.

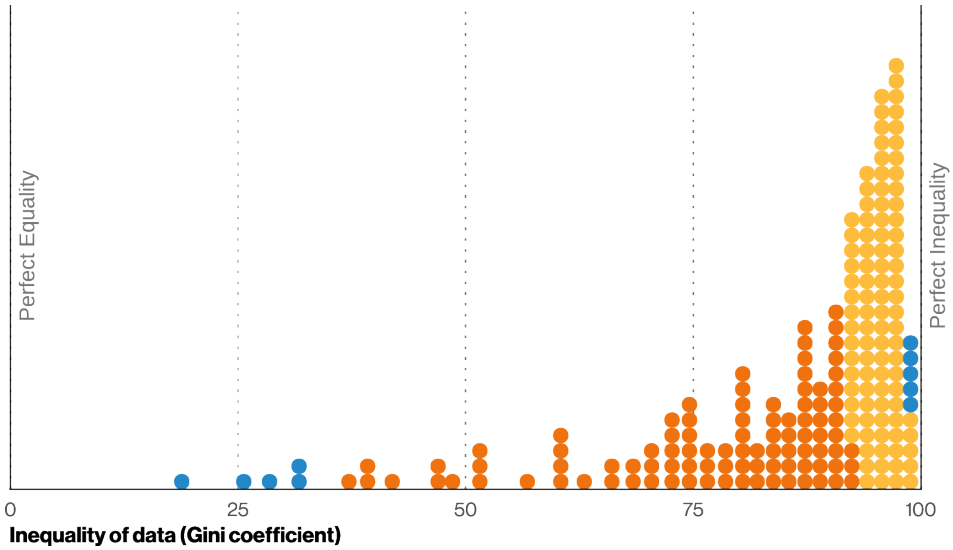


Figure 19. Inequality of login attempts per day (n=328)
Each dot represents 0.5% of organizations.

The other sub-pattern covers the exploitation of vulnerabilities in Web applications. They are not as common as the credential-related ones, as Figure 20 shows, but they are significant. Vulnerability exploitation is also the territory of a sister pattern, System Intrusion, but those present here in BWAA are not only focused on Web applications. They are also attacking with a small number of steps or additional actions after the initial Web application compromise.

In those incidents, the Actor will be focused on repurposing the web app for malware distribution, defacement[3] or installing malware for future DDoS attacks and calling it a day. Needless to say, a lot of the motive here is Secondary, more precisely in 78% of incidents. Threat actors are clearly not wasting the opportunity to shout "It's free real estate!" and expand their nefarious domains. Figure 21 shows this distribution in incidents, as in defacement, cases we often cannot get confirmation of a fully realized breach.

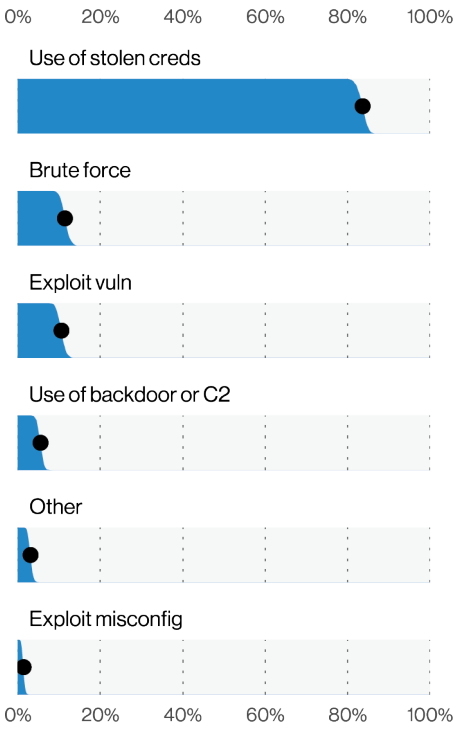


Figure 20. Top Hacking varieties in Basic Web Application Attacks incidents (n=947)

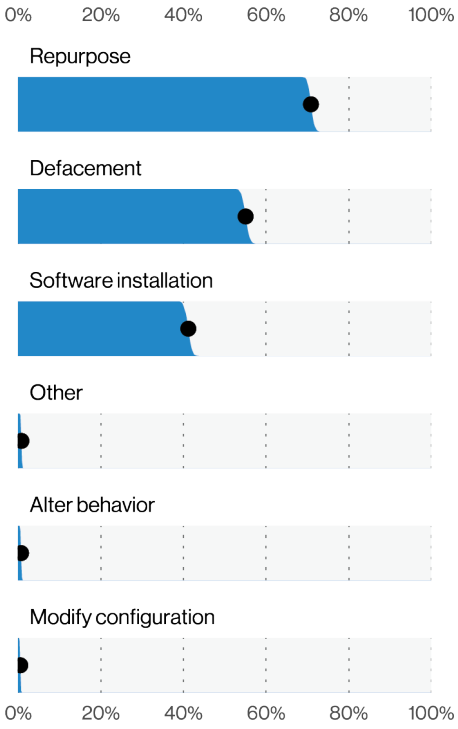


Figure 21. Top Integrity varieties in Basic Web Application Attacks breaches (n=3,653)

3 It's the '90s! Join our DBIR webring in Geocities!

Appendix: Methodology

One of the things readers value most about this report is the level of rigor and integrity we employ when collecting, analyzing and presenting data.

Knowing our readership cares about such things and consumes this information with a keen eye helps keep us honest. Detailing our methods is an important part of that honesty.

First, we make mistakes. A column transposed here; a number not updated there. We're likely to discover a few things to fix. When we do, we'll list them on our corrections page: verizon.com/business/resources/reports/dbir/2021/report-corrections/

Second, we check our work. The same way the data behind the DBIR figures can be found in our GitHub repository,[4] as with last year, we're also publishing our fact check report there as well. It's highly technical, but for those interested, we've attempted to test every fact in the report.[5]

Third, François Jacob described "day science" and "night science."[6] Day science is hypothesis driven while night science is creative exploration. The DBIR is squarely night science. As Yanai et al. demonstrate, focusing too much on day science can cause you to miss the gorilla in the data.[7] While we may not be perfect, we believe we provide the best obtainable version of the truth[8] (to a given level of confidence and under the influence of biases acknowledged below).

However, proving causality is best left to the controlled experiments of day science. The best we can do is correlation. And while correlation is not causation, they are often related to some extent, and often useful.

Non-committal disclaimer

We would like to reiterate that we make no claim that the findings of this report are representative of all data breaches in all organizations at all times. Even though the combined records from all our contributors more closely reflect reality than any of them in isolation, it is still a sample. And although we believe many of the findings presented in this report to be appropriate for generalization (and our confidence in this grows as we gather more data and compare it to that of others), bias undoubtedly exists.

The DBIR process

Our overall process remains intact and largely unchanged from previous years. All incidents included in this report were reviewed and converted (if necessary) into the VERIS framework to create a common, anonymous aggregate data set.

The collection method and conversion techniques differed between contributors. In general, three basic methods (expounded below) were used to accomplish this:

1 Direct recording of paid external forensic investigations and related intelligence operations conducted by Verizon using the VERIS Webapp

2 Direct recording by partners using VERIS

3 Converting partners' existing schema into VERIS

All contributors received instruction to omit any information that might identify organizations or individuals involved.

Some source spreadsheets are converted to our standard spreadsheet formatted through automated mapping to ensure consistent conversion. Reviewed spreadsheets and VERIS Webapp JavaScript Object Notation (JSON) are ingested by an automated workflow that converts the incidents and breaches within into the VERIS JSON format as necessary, adds missing enumerations, and then validates the record against business logic and the VERIS schema. The automated workflow subsets the data and analyzes the results. Based on the results of this exploratory analysis, the validation logs from the workflow, and discussions with the partners providing the data, the data is cleaned and re-analyzed. This process runs nightly for roughly two months as data is collected and analyzed.

4 https://github.com/vz-risk/dbir/tree/gh-pages

5 Interested in how we test them? Check out Chapter 9, Hypothesis Testing, of ModernDive: https://moderndive.com/9-hypothesis-testing.html

6 Jacob F. The Statue Within: An Autobiography. CSHL Press; 1995. By way of Selective attention in hypothesis-driven data analysis, Itai Yanai, Martin Lercher, bioRxiv 2020.07.30.228916;

7 Really. They made printing the data print a gorilla and people trying to test hypotheses completely missed it

8 Eric Black, "Carl Bernstein Makes the Case for 'the Best Obtainable Version of the Truth,'" by way of Alberto Cairo, "How Charts Lie" (a good book you should probably read regardless).

Incident data

Our data is non-exclusively multinomial, meaning a single feature, such as "Action," can have multiple values (i.e., "Social," "Malware" and "Hacking"). This means that percentages do not necessarily add up to 100%. For example, if there are five botnet breaches, the sample size is five. However, since each botnet used phishing, installed keyloggers and used stolen credentials, there would be five Social actions, five Hacking actions and five Malware actions, adding up to 300%. This is normal, expected and handled correctly in our analysis and tooling.

Another important point is that when looking at the findings, "Unknown" is equivalent to "Unmeasured." Which is to say that if a record (or collection of records) contains elements that have been marked as "Unknown" (whether it is something as basic as the number of records involved in the incident, or as complex as what specific capabilities a piece of malware contained), it means that we cannot make statements about that particular element as it stands in the record—we cannot measure where we have too little information. Because they are "unmeasured," they are not counted in sample sizes. The enumeration "Other," however, is counted, as it means the value was known but not part of VERIS. Finally, "Not Applicable" (normally "NA") may be counted or not counted depending on the claim being analyzed.

This year we again made use of confidence intervals to allow us to analyze smaller sample sizes. We adopted a few rules to help minimize bias in reading such data. Here we define "small sample" as less than 30 samples.

1. Sample sizes smaller than five are too small to analyze.
2. We won't discuss count or percentage for small samples. This applies to figures, too, and is why some figures lack the dot for the point estimate.
3. For small samples we may talk about the value being in some range, or values being greater/less than each other. These all follow the confidence interval approaches listed above.

Incident eligibility

For a potential entry to be eligible for the incident/breach corpus, a couple of requirements must be met. The entry must be a confirmed security incident defined as a loss of confidentiality, integrity or availability. In addition to meeting the baseline definition of "security incident" the entry is assessed for quality. We create a subset of incidents (more on subsets later) that pass our quality filter. The details of what is a "quality" incident are:

1. The incident must have at least seven enumerations (e.g., threat actor variety, threat action category, variety of integrity loss, etc.) across 34 fields OR be a DDoS attack. Exceptions are given to confirmed data breaches with less than seven enumerations.
2. The incident must have at least one known VERIS threat action category (Hacking, Malware, etc.)

In addition to having the level of details necessary to pass the quality filter, the incident must be within the timeframe of analysis, (November 1, 2019, to October 31, 2020, for this report). The 2020 caseload is the primary analytical focus of the report, but the entire range of data is referenced throughout, notably in trending graphs. We also exclude incidents and breaches affecting individuals that cannot be tied to an organizational attribute loss. If your friend's laptop was hit with Trickbot it would not be included in this report.

Lastly, for something to be eligible for inclusion in the DBIR, we have to know about it, which brings us to several potential biases we will discuss on the next page.

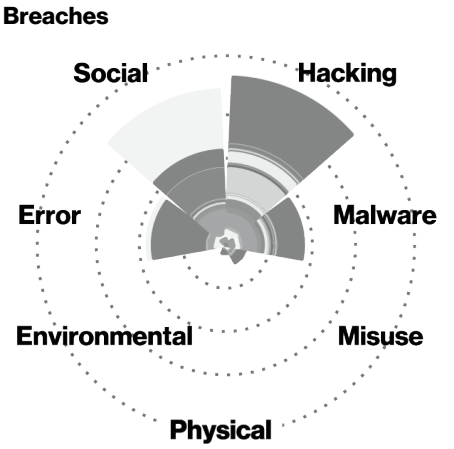


Figure 22. Individual contributions per action

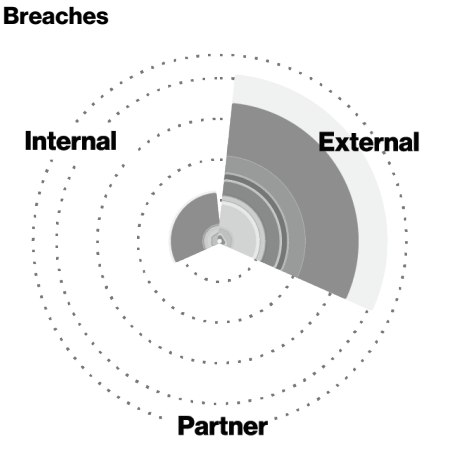


Figure 23. Individual contributions per actor

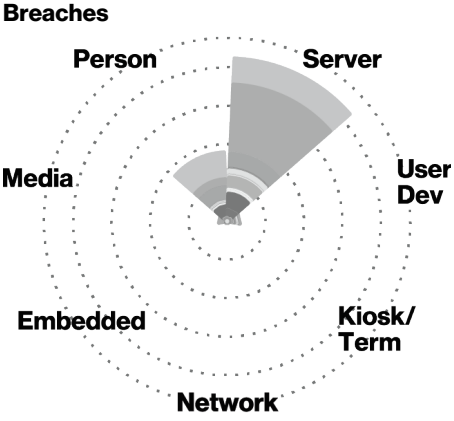


Figure 24. Individual contributions per asset

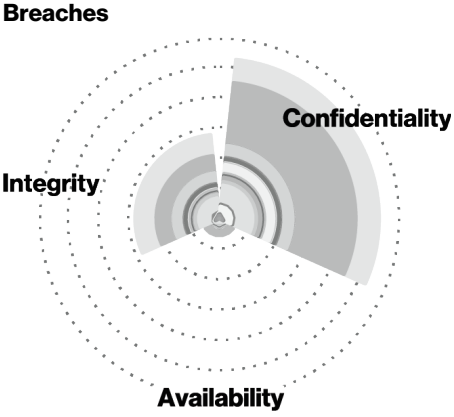


Figure 25. Individual contributions per attribute

Acknowledgment and analysis of bias

Many breaches go unreported (though our sample does contain many of those). Many more are as yet unknown by the victim (and thereby unknown to us). Therefore, until we (or someone) can conduct an exhaustive census of every breach that happens in the entire world each year (our study population), we must use sampling. Unfortunately, this process introduces bias.

The first type of bias is random bias introduced by sampling. This year, our maximum confidence is +/- 0.6% for incidents and +/- 1.5% for breaches, which is related to our sample size. Any subset with a smaller sample size is going to have a wider confidence margin. We've expressed this confidence in the conditional probability bar charts (the "slanted" bar charts) we have been using since the 2019 report.

The second source of bias is sampling bias. Still, it is clear that we conduct biased sampling. For instance, some breaches, such as those publicly disclosed, are more likely to enter our corpus, while others, such as classified breaches, are less likely.

Figures 22, 23, 24 and 25 are an attempt to visualize potential sampling bias. Each radial axis is a VERIS enumeration, and we have ribbon charts representing our data contributors. Ideally, we want the distribution of sources to be roughly equal on the stacked bar charts along all axes. Axes only represented by a single source are more likely to be biased. However, contributions are inherently thick tailed, with a few contributors providing a lot of data

and many contributors providing a few records within a certain area. Still, we mostly see that most axes have multiple large contributors with small contributors adding appreciably to the total incidents along that axis.

You'll notice rather large contributions on many of the axes. While we'd generally be concerned about this, they represent contributions aggregating several other sources, so not actual single contributions. It also occurs along most axes, limiting the bias introduced by that grouping of indirect contributors.

The third source of bias is confirmation bias. Because we use our entire dataset for exploratory analysis (night science), we do not test specific hypotheses (day science). Until we develop a good collection method for data breaches or incidents from Earth-616 or any of the other Earths in the multiverse, this is probably the best that can be done.

As stated, we attempt to mitigate these biases by collecting data from diverse contributors. We follow a consistent multiple-review process and when we hear hooves, we think horse, not zebra.

Data subsets

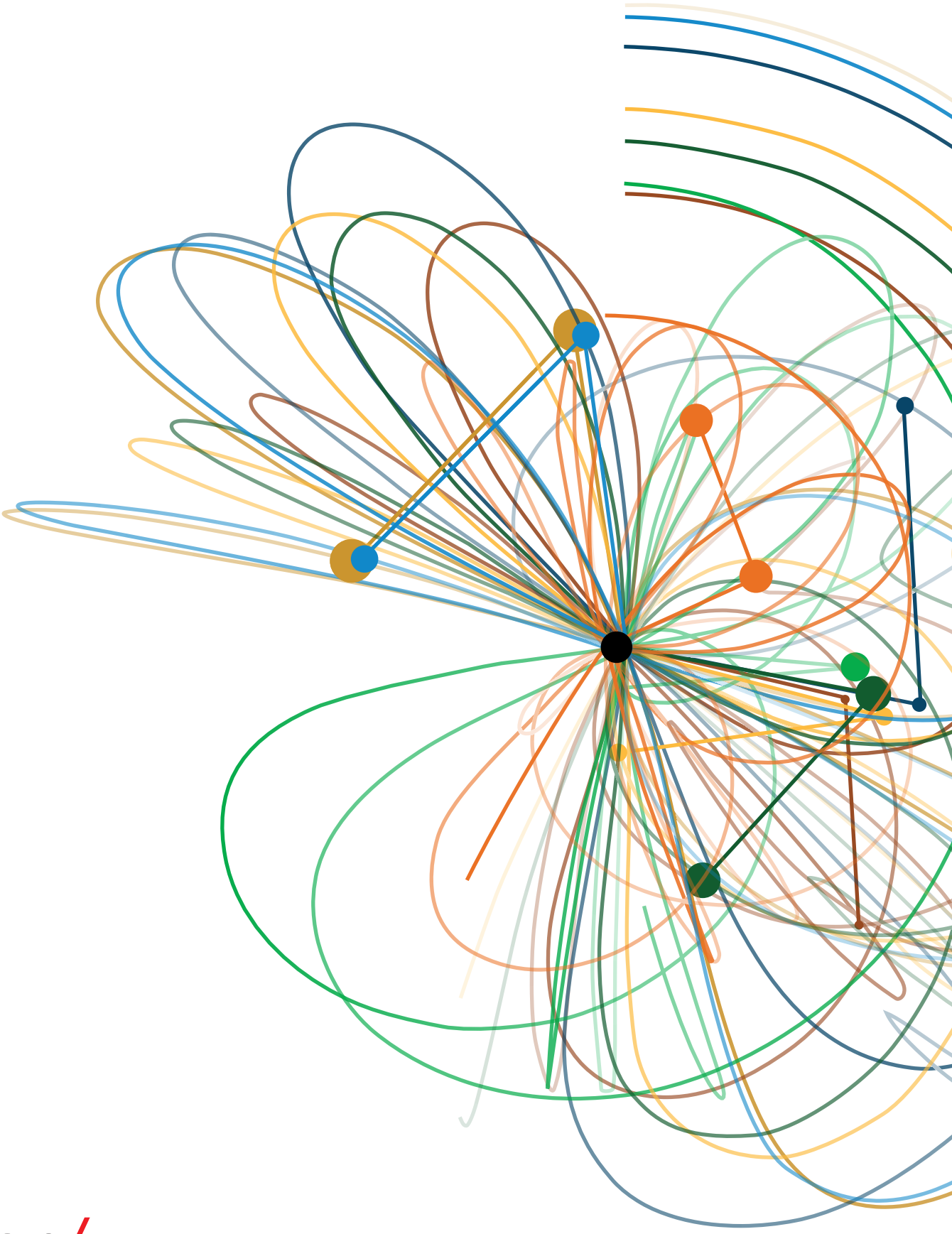
We already mentioned the subset of incidents that passed our quality requirements, but as part of our analysis there are other instances where we define subsets of data. These subsets consist of legitimate incidents that would eclipse smaller trends if left in. These are removed and analyzed separately (as called out in the relevant sections). This year we have two subsets of legitimate incidents that are not analyzed as part of the overall corpus:

1. We separately analyzed a subset of web servers that were identified as secondary targets (such as taking over a website to spread malware).
2. We separately analyzed botnet-related incidents.

Finally, we create some subsets to help further our analysis. In particular, a single subset is used for all analysis within the DBIR unless otherwise stated. It includes only quality incidents as described above and excludes the aforementioned two subsets.

Non-incident data

Since the 2015 issue, the DBIR includes data that requires the analysis that did not fit into our usual categories of "incident" or "breach." Examples of non-incident data include malware, patching, phishing, DDoS and other types of data. The sample sizes for non-incident data tend to be much larger than the incident data, but from fewer sources. We make every effort to normalize the data (for example weighting records by the number contributed from the organization so all organizations are represented equally). We also attempt to combine multiple contributors with similar data to conduct the analysis wherever possible. Once analysis is complete, we try to discuss our findings with the relevant contributor or contributors so as to validate it against their knowledge of the data.



verizon