Gallagher Re

Scanning the Horizon:
How broadening our use of cybersecurity data can help insurers

Building on our previous study from 2023, Gallagher Re explores which cyber datasets can help insurers predict claims and materially reduce loss ratios

BITSIGHT

Executive summary

- Cybersecurity firms' ability to scan and assess companies' defences against cyber attacks has developed quickly in recent years. At Gallagher Re, we have been exploring the potential for this data to help cyber (re)insurers inform their underwriting since 2021.
- To date, (re)insurers' ability to do this has been limited, thanks to uncertainty over which data points are really predictive of claims. Yet with further research, this rich dataset has huge potential and could be transformative for the cyber insurance industry.
- In early 2024, Gallagher Re conducted the largest study of this kind yet published, comparing third-party assessments of companies' security controls to the insurance claims that have arisen from the same firms.
- We performed an independent analysis of cybersecurity performance data provided by Bitsight, whose assessment data formed the primary basis for this study, in combination with our claims data. This study was conducted by Gallagher Re without providing Bitsight access to our data.
- Importantly, this is the first published study in this field to include SPoF (Single Point of Failure) data, which highlights the dependencies a company has on third-party systems and services (from reliance on Amazon Web Services (AWS) products in specific regions, to VPNs and e-mail security tools.)
- This study was conducted before the CrowdStrike incident of July 2024, when a malfunctioning software update led to a widespread outage for many organisations globally, and insurance losses of potentially up to USD1 billion. This incident dramatically highlighted the value of this kind of SPoF assessment.
- This paper represents a snapshot of our ongoing work exploring what is a deep and highly complex dataset. We hope others will be able to build on this research over time, as we will continue to do ourselves.
- Gallagher Re has simultaneously developed a comprehensive suite of tools and services to support (re)insurers utilising technographic (scanning) data. These can help evaluate external scanning products; apply external scanning data in underwriting; and assist in visualising and benchmarking portfolio quality. Please contact us for further information on these.

Key findings of the study

External scanning data can be used to materially reduce loss ratios

Updating the findings of our previous study from 2022, we confirmed that our model excels at identifying the organisations with the weakest cybersecurity controls, with the worst 20% of risks 3.17x more likely to suffer a loss (using scanning data alone). By removing these from the portfolio, we estimate insurers could achieve a reduction in loss ratios of up to 16.4%.

Cyber risk factors are changing in line with the threat landscape

Risk factors linked to hybrid or home-working, such as weak mobile application security, have grown in importance since 2022. Meanwhile, factors associated with traditional on-premises networks (e.g., port security) have decreased. This is intuitive, given workforce trends in the intervening 2 years, and reassures us the model is responsive to shifts in the threat landscape.

IP address count is an important claims predictor

The number of IP addresses a company maintains — its 'cyber footprint' — is a strong predictor of claims. This is significant, as IP count is not a widely used metric even among cyber insurers at present. Despite being a strong indicator for company attack surface size, it also has surprisingly little correlation to company revenue, a metric that is commonly used.

Certain Single Point of Failure (SPoF) data is predictive of claims

This data enables insurers to assess portfolio exposures to particular services and vendors (e.g., AWS) across their portfolios.

This is a comparatively less developed dataset when compared with risk factors like patching cadence. However, this first test of the correlation of this dataset with claims shows strong predictive potential, with SPoF offering an additive view of cyber risk when compared with other scanning data.

Background

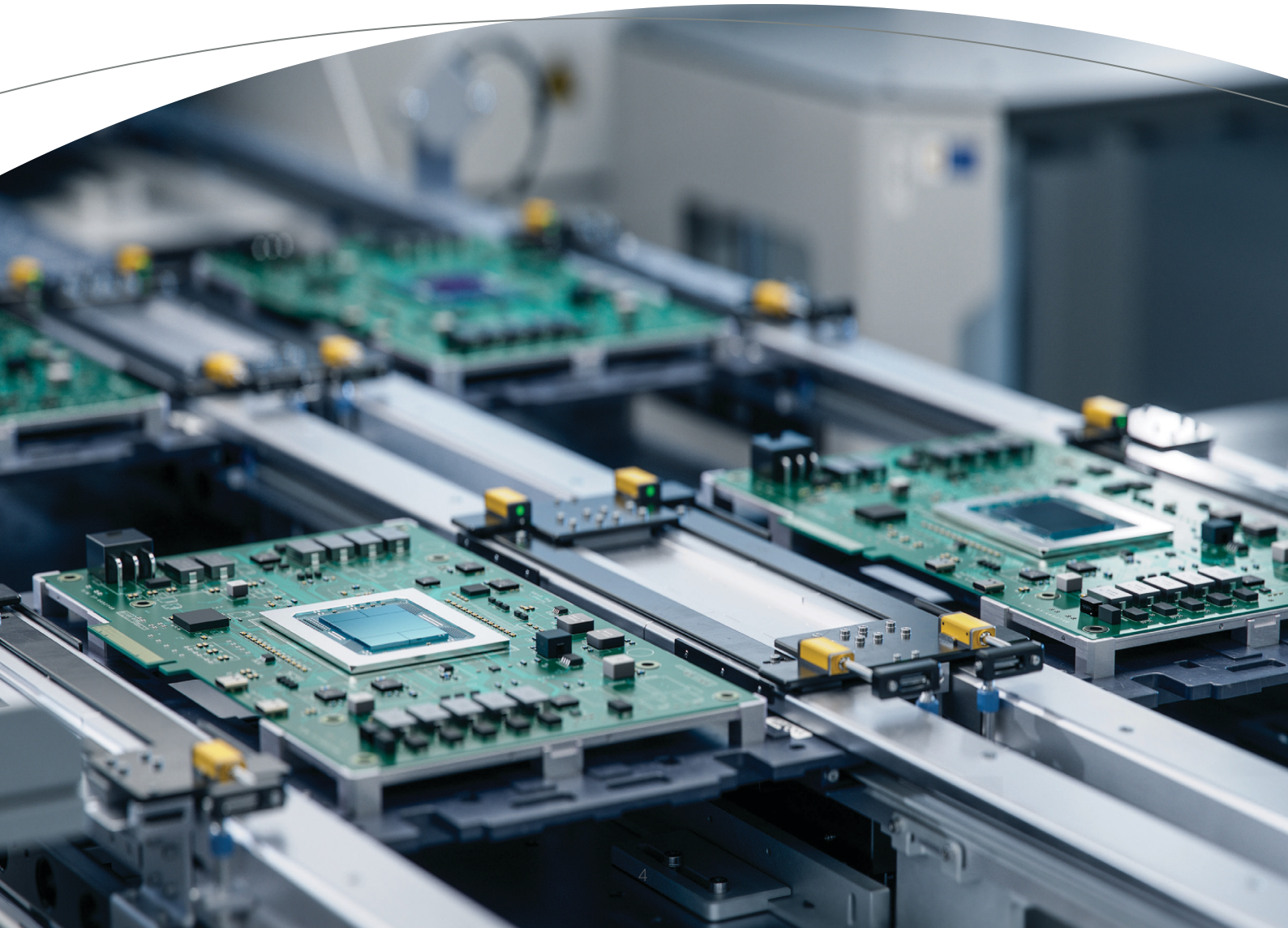
Cybersecurity firms have been able to remotely scan and assess companies' resilience to hacking attacks since at least the early 2010s, and have now built up large databases of this information. And in recent years, cyber insurers have begun to use these to inform their underwriting.

At Gallagher Re, we have been exploring this data's vast potential for several years. In 2022, we published our first paper, Looking from the Outside-In, remarking on the rapid uptake of external scanning data among our insurance clients since the rise of ransomware attacks in 2019–2020. We conducted our first large-scale study into the use of this data in the summer of that year.

The following April, we published the results of that study: Can scanning technologies predict claims? This took an in-depth look into how this data can be useful to insurers, establishing (amongst other things) that while firms' revenue (i.e., size) was the biggest single predictor of cyber claims, other technology-specific factors such as patching cadence (the speed at which companies fix vulnerabilities) were also material.

We also observed how many of these cyber risk factors are highly correlated to one another, and to basic firm characteristics such as revenue or company size. Intuitively, a larger firm has more computer systems, more people vulnerable to phishing, and more revenue for attackers to target. So what additional value can this scanning data bring to an insurer beyond 'large firm = larger risk'? For our 2024 study, we have attempted to answer this question.

This research has informed the development of a suite of proprietary tools and services aimed at supporting the (re)insurance community in realising the potential of cyber data to enhance underwriting and portfolio monitoring. Principal among these is TIDE, our portfolio quality and benchmarking tool.



Gallagher Re 2024 cyber claims correlation study

Rapid improvements in the scanning technology itself are helping to generate more insight. For our 2024 study, we have performed an independent analysis of cybersecurity insights data provided by Bitsight. This has enabled us to make use of IPv4 data, for example — exploring whether a company's count of IP addresses correlates with its size, and whether this is predictive of claims.

We have also been able to incorporate Bitsight's Single Point of Failure (SPoF) data for the first time, examining companies' vulnerabilities to particular providers such as Amazon Web Services (AWS), and whether this, too, is additive to insurers' view of risk.

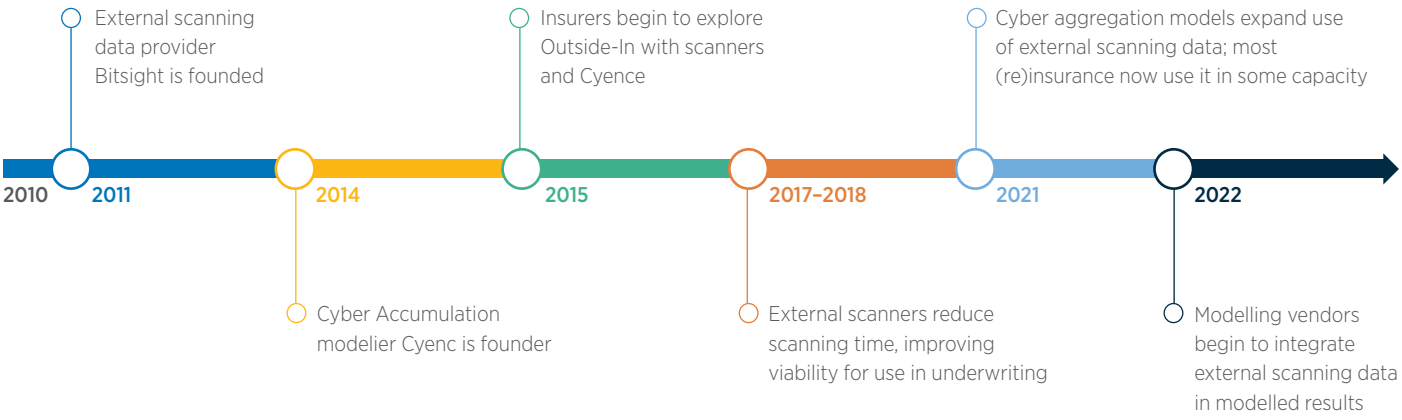
We hope our findings will contribute to an ongoing debate. The cybersecurity industry itself is of course likely to benefit from greater data-driven insights into the effectiveness of its defences, and this in turn can empower cybersecurity leaders within organisations and enable more effective public policymaking in this field.

Yet the industry remains challenged by a lack of consistency of approach. As Daniel Woods and Sezaneh Seymour highlighted in a recent paper in the Journal of Cyber Policy, "there is no authority that collects evidence and ranks cybersecurity controls by efficacy" — something that underlines the importance of empirical studies in this area.

The root cause of this challenge lies with the shortcomings of cyber incident reporting. Companies, of course, can have strong and understandable incentives to minimize their disclosures surrounding successful cyber attacks. Even where regulators have made efforts to standardise reporting — such as the new requirements from the US Securities and Exchange Commission (SEC) — these increasingly appear open to manipulation by companies, and even threat actors.

In the absence of consistent incident reporting, insurance claims data helps fill the gap. That is not to say claims data does not have its own challenges and limitations; some of these are detailed in the Appendix to this paper. Despite these, however, it is our view that insurers, via their claims reports, possess one of the most comprehensive datasets available on cyber incidents. We hope that studies based upon it will prove useful — both to the insurance industry, and beyond.

Figure 1: Timeline of external scanning data developments



[1]Woods, Daniel W and Sezaneh Seymour. "Evidence-based cybersecurity policy? A meta review of security control effectiveness," *Journal of Cyber Policy*, 07 April 2024.

[2]"Cybersecurity Risk Management, Strategy, Governance, and Incident Disclosure," *SEC,* 26 July 2023.

[3]"Hackers Weaponize SEC Disclosure Rules Against Corporate Targets," *DarkReading.com*, 17 November 2023.

Why us? Our methodology and model

Gallagher Re first developed a claims-correlation model in 2022, and we have kept it under continual development since. Our analysis utilises a wide range of machine learning and data science techniques to explore the relationship between data and insurance claim frequency.

Reinsurance brokers are ideally placed to work towards solving the security control efficacy uncertainty the cyber (re)insurance market, public policymakers, and cybersecurity leaders face. Our clients are cyber insurance companies, and in the course of our work for them, we naturally build up a large database of claims across the industry. By using this claims data to assess, at the point of underwriting, which cyber datapoints are predictive of different types of loss and additive to our view of risk, we can facilitate more targeted use of external scanning data and provide a view on security control efficacy. Moreover, we are able to evaluate the strengths and weaknesses of different cybersecurity scanning approaches from an insurance perspective, as well as the modelling firms that aggregate this data for the insurance industry.

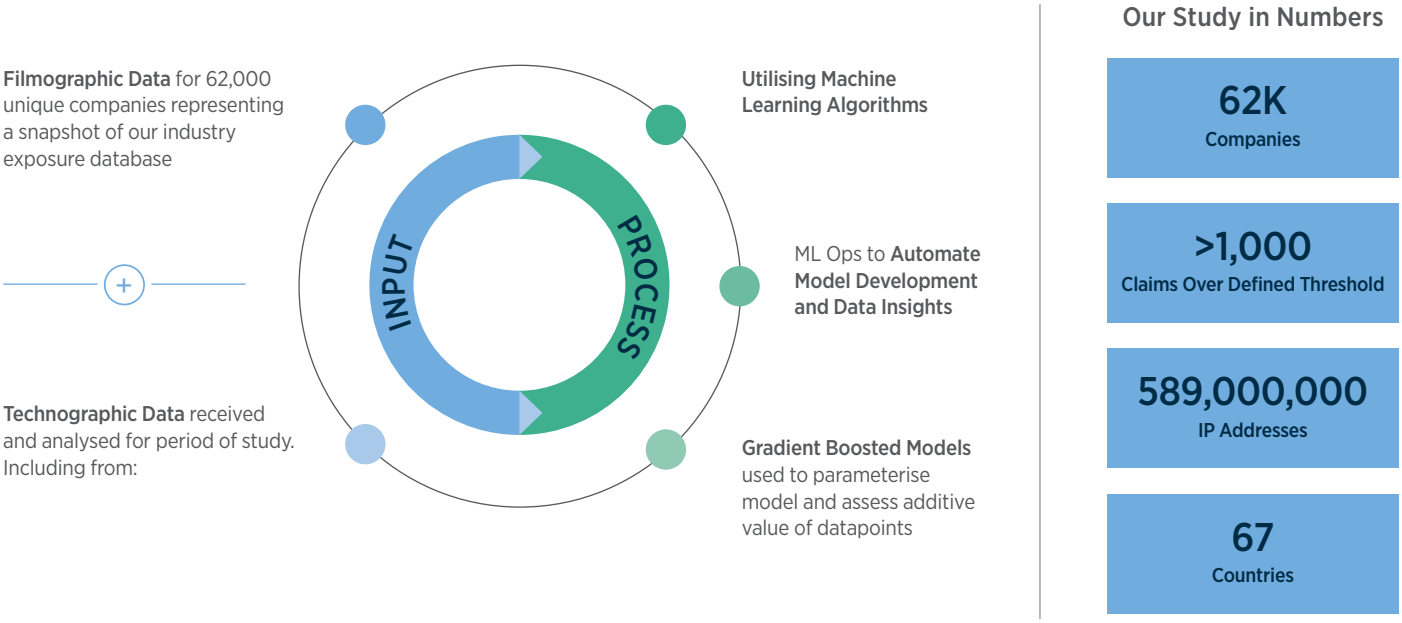
Gallagher Re's in-house Cyber Analytics team compiles huge volumes of claims data together with traditional risk indicators ('firmographic' or traditional company information, such as revenues) in our cyber insurance 'data lake'.

In each iteration of our claims-correlation study, we have then augmented this firmographic and claims data with 'technographic' (cybersecurity control) risk indicators supplied by cybersecurity firms. For our 2024 study, we partnered with Bitsight, a market-leading external scanning provider.

Previously, Gallagher Re has conducted claims correlation studies on multiple other vendors' datasets, and these are also referenced on an anonymised basis, where the insights further enrich our findings.

Our 2024 study looked at a dataset comprising over 62,000 companies in 67 countries and over 589 million separate IP addresses. It involved over one thousand material claims. Methodology highlights are presented below:

Figure 2: Overview of Gallagher Re's 2024 claims correlation study



Our data scientists use machine learning algorithms to create mathematical models of the relationship between various risk factors and specific claim types. Using the patterns captured in these mathematical models, our team can then make predictions about future claim frequency. Our teams also develop more traditional univariate analysis, and benchmark models relying on simple heuristics and business rules, to provide a benchmark to compare model performance against.

Statistical analysis alone is insufficient to gain a complete view of the complex relationships uncovered by the AI models. Nevertheless, by blending this work with actuarial, insurance, and cyber risk expertise, we can better understand the patterns uncovered and navigate the limitations of data and statistical modelling approaches.

For more information on our analysis and its limitations, please see the Appendix. For more insights into how we use ML models, please see the article Gallagher Re's AI initiatives: Cyber team leveraging AI, published in Gallagher Re's Q1 2024 InsurTech Report (page 41).

Findings and conclusions

We believe our 2024 study represents the largest of its kind yet published — comparing third-party assessments of companies' security controls to the insurance claims that have arisen from the same firms.

01 External scanning data can be used to materially reduce loss ratios

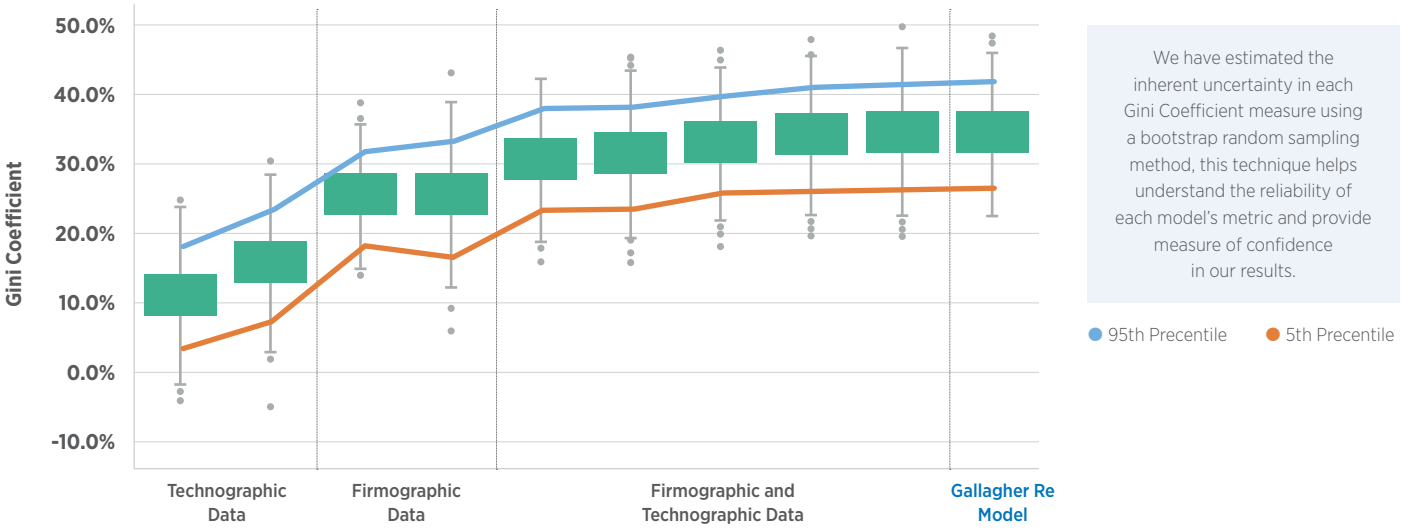
Our previous study found that this use of external scanning data is very good for identifying the worst risks. But our model struggled to distinguish between the strongest 80% of organisations for cyber controls — i.e., a firm with best-in-class external-facing cybersecurity controls did not generate materially fewer claims than a firm with merely good ones.

This trend continued in our 2023/2024 study. Our model ingested firmographic data to complement the vendor technographic data, including industry, revenue, limit, and geography. Based on technographic data alone, the worst 20% of companies identified by our model were 3.17x more likely to suffer a claim than the best 20%. This rose to 6.93x when firmographic data was also included.

Gallagher Re's Actuarial teams are able to assess the real-world performance of our model in predicting insurers' loss ratios. This analysis indicates that external scanning data can be modelled to differentiate profitable and unprofitable business on previously unseen data.

Our conclusion was that this analysis can be highly material for insurers. By removing the worst 20% of risks in our study portfolio, (re)insurers would have enjoyed a 16.35% reduction in loss ratio.

Figure 3: Bootstrapped Gini Coefficients representing model performance



We also cross-checked our conclusion using a bootstrap random sampling method. The results of this are shown in the graphic above, which displays Gini coefficients along the Y axis. These are a statistical method for testing model performance, with a score of 100% representing a model's ability to perfectly rank risks (companies holding cyber insurance policies) in order of their likelihood of suffering a claim. The five vertical plots in the firmographic and technographic data section represent different vendor datasets.

The graph shows that the models perform better when assessing risk through the lens of technographic and firmographic data together. It also offers us reassurance that Gallagher Re's model has achieved strong results in doing so to date (though other vendor models also perform well). This possible further lift in our model has been achieved through feature engineering and manipulation of vendor data to extract additional predictive value.

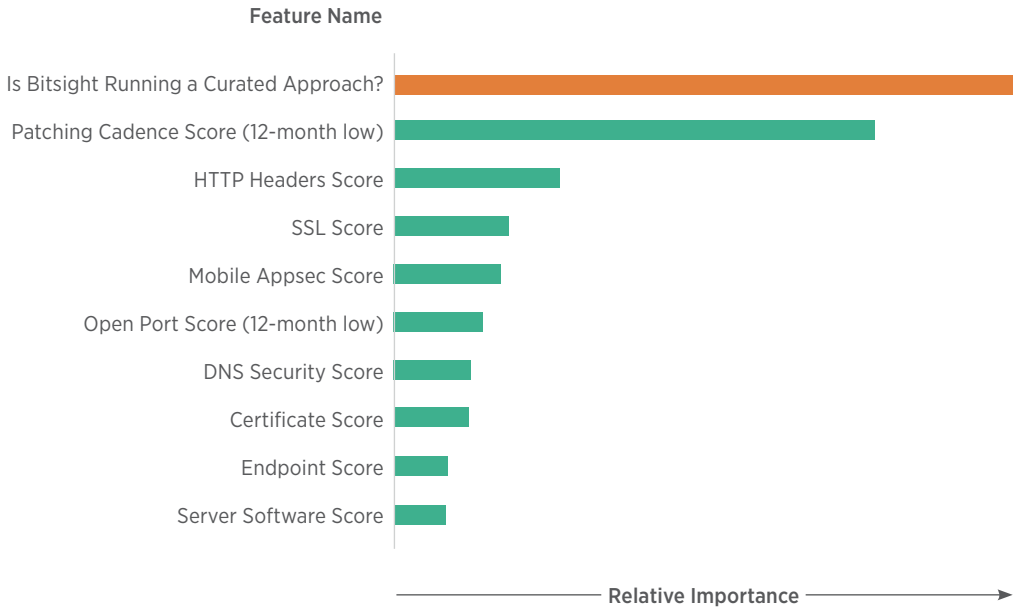
02 The cyber risk features driving claim frequency are shifting

A major weakness of historical correlation studies on external scanning data has been a focus on testing risk factors independently. We find the vast majority of risk factors are able to predict claims to at least some extent. However, significant correlations between the risk factors mean that many of them aren't adding to our view of risk. Therefore, our approach has been to highlight the risk features offering increased value to our view of risk, meaning underwriters can focus on a subset of the most important ones.

The results in Figure 4 below highlight the relative predictive power of different Bitsight risk factors as observed by our model. These continued to evidence patching cadence as the most useful traditional feature, with its utility especially pronounced when looking at the worst score in a 12-month period (a synthetic policy lifecycle).

Compared to our previous study, we identified a shift in the risk factors driving claims frequency, possibly reflecting wider changes in the threat landscape. Factors related to hybrid working and cloud identity management (e.g., mobile application security) have grown in importance, whilst factors associated with traditional on-premises security (e.g., port security) have decreased.

Figure 4: Bitsight risk factors and manual updates



We tested Bitsight data through our model as of September 2023. See connected materials on our modelling approach for background on how charts were developed.

What is a Bitsight-curated approach, and why does it matter?

Bitsight's standard process to attribute an organisation's assets is using a curated approach that combines automated and manual techniques. For some use cases, Bitsight conducts a purely automated approach, commonly used by insurers for rapid underwriting of organisations. Our correlation analysis found that where Bitsight was running a curated approach (this combination of automated/ manual processes), our model flagged a greater likelihood of that company suffering a claim across all revenue bands. This is additive to the view of risk offered by firmographic data and by other risk factor data.

How do the drivers of loss differ across different types of cyber claim?

Whilst broadly consistent across different revenue bands, the Bitsight risk factors (excluding SPoF and footprint data) predictive of different types of cyber events varied significantly. This was particularly pronounced for Business Email Compromise claims, where risk factors important for email security, such as DKIM (DomainKeys Identified Mail) and SPF (Sender Policy Framework), were particularly useful at anticipating loss. For ransomware, Patching Cadence remained the strongest predictor of loss, with residual predictive value split across a range of factors used by threat actors for initial compromise in malware attacks. Due to challenges with claim classification, we have greater confidence in our ransomware findings compared with business email compromise.

What are the drivers of Loss?

Business Email Compromise

1. DKIM Score
2. SPF Score
3. Open Port Score (12 Month Low)
4. Mobile Application Security Score
5. HTTP Headers Score

Ransomware

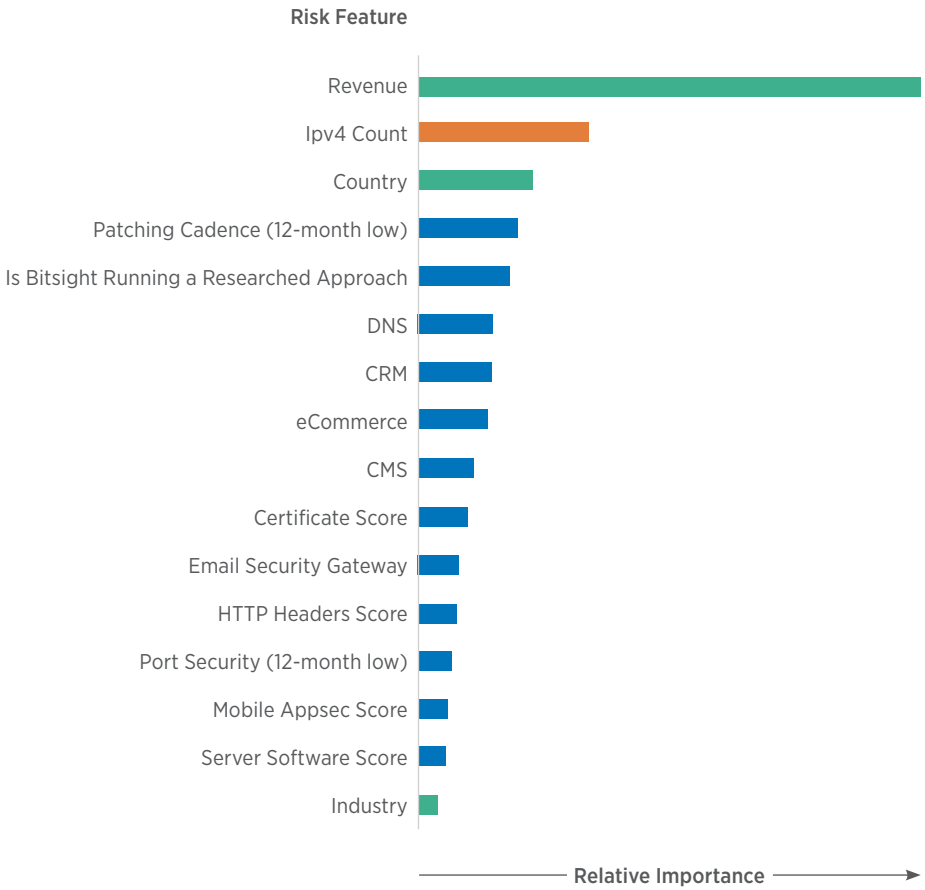
1. Patching Cadence Score (12 Month Low)
2. Open Port Score (12 Month Low)
3. Web Certificate score
4. SSL Score
5. DKIM Score

03 IPv4 count is the strongest technographic predictor of claims

In our previous study in 2022, we found that company revenue was the strongest predictor of claims when considering all factors together.

Our latest model, which includes the number of IPv4 addresses related to each risk for the first time, finds that this ranks as the second-highest predictor (while revenue remains the top predictor). This both validates the importance of company size — since larger companies will tend to have more internet-enabled devices with unique IP addresses and therefore a larger attack surface for threat actors to target — but also goes beyond it. This is because company revenue actually has less correlation to IP address count than you might expect.

Figure 5: Importance of IPv4 Count in Predicting Cyber Claims

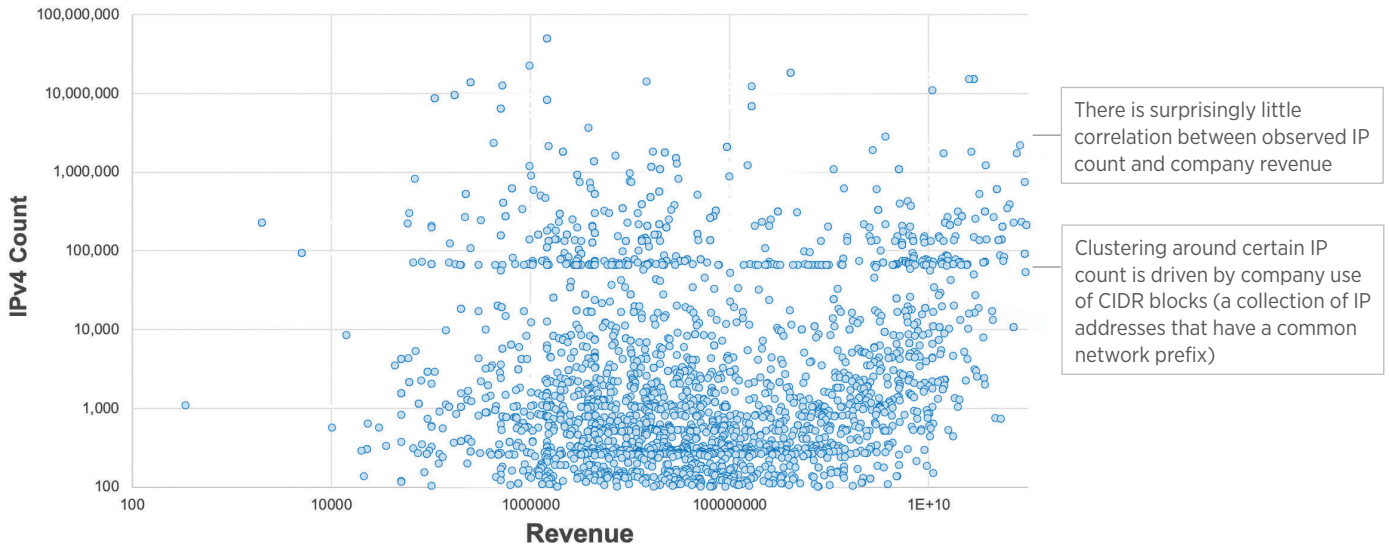


Identifying ‘pockets of value’

It is important to note that a company with high revenues does not necessarily have a large number of cyber assets, and conversely, a large number of cyber assets does not mean high revenues, as shown by Figure 6 below.

The analysis can identify companies with large revenues but comparatively smaller numbers of IP addresses. Under traditional cyber risk analyses, these might generally be assigned a relatively higher level of risk and therefore premium. The inclusion of IPv4 data therefore offers the prospect of identifying ‘pockets of value’ — companies that present less cyber risk than their size might suggest. In any event, both indicators should be considered alongside other technographic data points. We recommend insurers and data providers consider developing technographic measures of attack surface to complement technographic vulnerability scores and firmographic data.

Figure 6: The correlation of IPv4 address count with revenues



The internet has run out of address space — what does that mean for external scanning?

The internet has been expanding way beyond the intentions of its designers, almost since it was founded. In particular, the current Internet Protocol address system, IPv4, dates from the late 1980s and has a maximum of around 4.3Bn IP addresses available. But with the massive explosion in the numbers of internet-connected devices in recent years, that number has become woefully inadequate.

The industry has seen the problem coming for many years. A new system, IPv6, was designed as long ago as the late 1990s and allows for a maximum of 340 undecillion (340 followed by 12 zeros) unique addresses. However, rolling this new protocol out across the global internet has proven to be a slow and arduous process that is still far from complete.

IPv6 will change the way scanning is conducted and likely lead to a shift in the vendor landscape by increasing the barrier to entry for external scanning providers.

However, IPv4 will remain the standard for at least the next decade:

- Challenges to fully migrating to IPv6, such as the replacement of network management assets, mean full adoption of IPv6 might take decades.
- Concerns that threat actors are ahead of security tools and practices for IPv6 will further delay uptake.

The current approach to scanning will need to change:

- It takes scanning vendors roughly 45 minutes to snapshot the entire internet's IPv4 address space.
- This approach is currently impossible for IPv6, since with current technologies it would take over 300,000 years at a minimum (and potentially up to a trillion years).

IPv6 will change companies' attack surface, and external scanning will need to adapt:

- Threat actors will continue to exploit externally facing assets, whether using an IPv4 or IPv6 address space.
- Echoing threat actor behaviour, technology will shift to scanning 'hitlists' of active IPv6 addresses (many companies assign IPv6 addresses sequentially.)
- Additional alternative IP scanning techniques include DNS lookups and reverse DNS lookups.

Bitsight: How we are evolving our approach for IPv6

By Dan Dahlberg, VP of Data and Research at Bitsight

The increasing use of IPv6 requires Bitsight to create robust processes, techniques, and methodologies for identifying IPv6 infrastructure used by organisations around the world.

IPv6 is the next-generation IP standard intended to eventually replace IPv4, the protocol most Internet services still use today. IPv6 was created to address limitations in IPv4 address space and enable the proliferation of millions of new internet-connected devices. IPv6 has many more addresses compared with IPv4 that can be used by different devices connected to the internet. But despite new device proliferation and warnings that the internet is running out of address space, IPv4 is still the dominant IP used around the globe today. Many companies hosting infrastructure will use 'dual stack' IPv4 and IPv6 addresses; few use IPv6 alone.

For an organisations like Bitsight, identifying IPv6 addresses associated with organisations presents an interesting but not insurmountable challenge. One of Bitsight's core capabilities is to associate infrastructure to organisations in order to identify cybersecurity risks and performance issues. While it is possible to scan the entire global IPv4 asset space to enumerate assets, the sheer volume of IPv6 address space makes this approach impossible.

Therefore, Bitsight uses alternative strategies to discover IPv6 assets and attribute assets to organisations. Fortunately, there are many techniques to collect web telemetry and enumerate assets that Bitsight has invented and patented that enable us to identify and associate IPv6 assets to organisations, including passive DNS discovery, DNS resolution, and internet crawling. These techniques allow Bitsight to collect both IPv4 and IPv6 data and perform accurate attributions.

04 Single point of failure data may also help predict claim frequency

Single Point of Failure (SPoF) data goes by many names in the world of cybersecurity. Some external scanning vendors refer to it as 'footprint data; others call it 'fourth-party data' or 'threat intelligence data'. Nonetheless, these terms all refer to the same concept: identifying the external software and services that an organisation is dependent upon. (The term 'SPoF' itself refers to identifying the single one that could potentially fail.)

Whilst the Gallagher Re team has conducted analysis on SPoF data provided by a number of vendors, the SPoF data referenced in this study was provided to Gallagher Re by Bitsight for independent analysis. Our work so far suggests to us that this data holds great promise for the insurance industry. One obvious use case is the proactive identification of companies susceptible to an emerging event, helping them to mitigate losses from either that event or a similar future one. The July 2024 CrowdStrike event provides an excellent illustration of this, as noted in the section below.

SPoF data can also help identify aggregation points for modelling, so it's no surprise that vendors are increasingly making use of it too. CyberCube has incorporated SPoF data directly into its modelled results, looking at actual dependencies of companies in a portfolio where the required input data is provided. Meanwhile, RMS and Guidewire use it to influence model parameterisation.

Nevertheless, the SPoF dataset is at an earlier stage of development than the scanning data that underlies the technographic risk factors referred to in the previous section. There is little standardisation between cybersecurity vendors in how they capture and process SPoF data, and as a result, there is inconsistency in their findings. For example, some weight a company's dependency on an external service differently, according to whether that service is delivered on-premises, or hosted in the cloud.

Due to the limitations of the dataset in its current form across all vendors we have analysed, our study focused on six specific categories of SPoF data. These are set out in Table 1 below. For each SPoF category, we looked at both the change in anticipated claim frequency where an individual company was using a service in the specified category, and the change in anticipated claim frequency for specific vendors within each category.

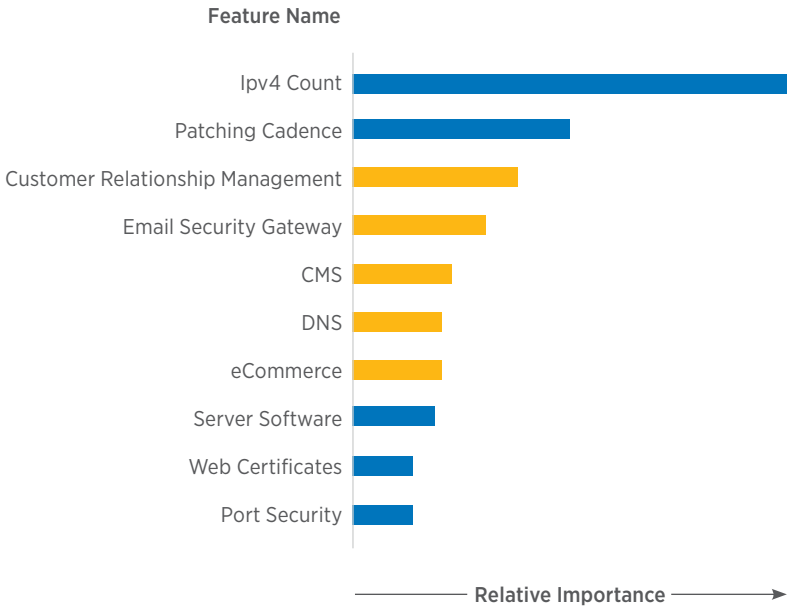
Table 1: Single point of failure categories examined in our study

SPoF Category	Detail
VPN	A **Virtual Private Network** establishes a secure and encrypted connection over the internet. By routing internet traffic through a remote server, it enables private and secure internet access. Whilst the aim of this is to enhance privacy, 'risky VPNs' have been a popular egress point for attackers over the past 2 years.
CMS	**Content Management Systems** are software platforms that facilitate the creation and management of digital content on websites, making the process more user-friendly. They have been widely adopted by internet users.
DNS	The global **Domain Name System** is often described as 'the internet's phonebook': it translates domains (example.com) into IP addresses (192.0.1.1) that computers can understand. This system plays a crucial role in the functioning of the internet by allowing users to access websites and other online services using human-readable domain names.
CRM	A **Customer Relationship Management** system is a software suite or technology package utilised by businesses to handle and evaluate their customer interactions and relationships.
eCommerce	An **eCommerce** software application or platform is one that empowers businesses to conduct online sales of their products or services. Our analysis looked at both popular eCommerce services and those that have been linked to known vulnerabilities in the past or present.
Secure Email Gateway	A **Secure Email Gateway** acts as a barrier between the internet and an organisation's email server, scanning incoming, and outgoing emails. They play a crucial role in safeguarding organisations against cyber threats. See section below.
Payroll	Cybercriminals are increasingly targeting **online payroll accounts** of employees in many different industries. We included both widely used systems and those that have been identified as being systematically targeted by threat actors. The results for this SPoF were inconclusive, but this could be attributed to a lack of available data.

[4] "Data Science Insight: How VPN Vulnerabilities Affect Ransomware Risk," *Corvus Insurance*, 8 September 2022.

Our analysis found that where we had sufficient data, these different SPoF categories could add materially to predictions of claim likelihood. Figure 7 shows a ranking of all technographic factors, with SPoF datapoints highlighted.

Figure 7: Importance of SPoF factors in predicting cyber claims



How external visibility of email security gateways could increase cyber claims

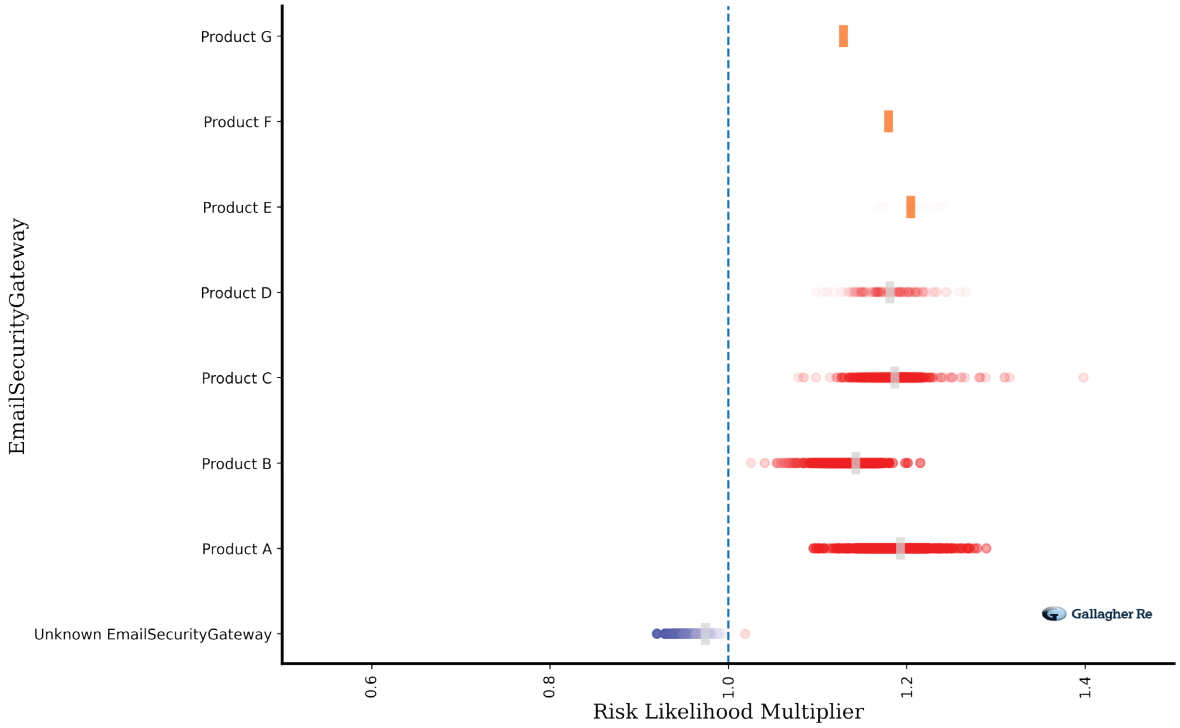
One interesting finding from our study was that the external visibility of a specific email security gateway seemed to increase the anticipated likelihood of claims. This might seem counterintuitive.

As noted above, email secure gateways (ESGs) scan incoming and outgoing emails for cybersecurity threats, playing a crucial role in ensuring the security of an organisation's communication channels. Within an organisation, they help prevent employees from clicking on harmful links or downloading infected files; ensure the confidentiality of information through email encryption; and demonstrate compliance with regulatory requirements by maintaining a secure and compliant email environment.

One might expect, therefore, that a company's use of an email security gateway should decrease its likelihood of making a cyber claim. Our hypothesis here is that this finding is to do with the fact that this gateway is externally visible; this may be indicative of a misconfiguration of an organisation's ESG, or a security control failing.

Figure 8 shows our model's findings on the relative likelihood of a claim dependent on the ESG provider that a company is associated with by external scanning data. The colour shows the spread of organisations using each ESG, and the central bar shows the median of those organisations. Where the bar is orange, there are under 100 organisations identified as using that ESG provider. The bottom line indicates an unknown Email Security Gateway provider.

Figure 8: Relative likelihood of insurance claims, by email security gateway provider



CrowdStrike: Measuring exposure when the TIDE goes out

At Gallagher Re, we have combined SPoF data with our pre-existing Industry Exposure Database (IED) to create a database of cyber insurance policyholders' exposure to external service providers. This database, known as TIDE (our Technographic Insight Detection Engine), can be used by our clients to measure their policyholders' exposures to any given third-party service.

In July 2024, the cyber insurance industry experienced arguably its most significant event since NotPetya in 2017. A faulty configuration update from the cybersecurity firm CrowdStrike triggered a widespread memory error in the firm's Falcon sensor software for Windows PCs, leading to repeated system crashes. The outage impacted approximately 8.5Mn devices across CrowdStrike's 24,000 customers, including nearly 60% of Fortune 500 companies. The aviation industry was particularly affected, resulting in the cancellation of 4.6% of global flights scheduled for that day (July 19). Various industry estimates have put potential global insured losses from this event in the range of USD300Mn to USD1Bn. This is a level unlikely to significantly impact most insurers.

Within 24 hours, we had shared our Gallagher Re TIDE analysis of market exposure with our clients. It showed the use of CrowdStrike services — and hence exposure to the outage — was highest among IT and tech firms, followed by the transportation and logistics sector (including aviation).

[5]"More than 5,000 flights cancelled globally," *BBC News*, 19 July 2024.

Figure 9: Gallagher Re TIDE analysis of industries' relative exposure to CrowdStrike

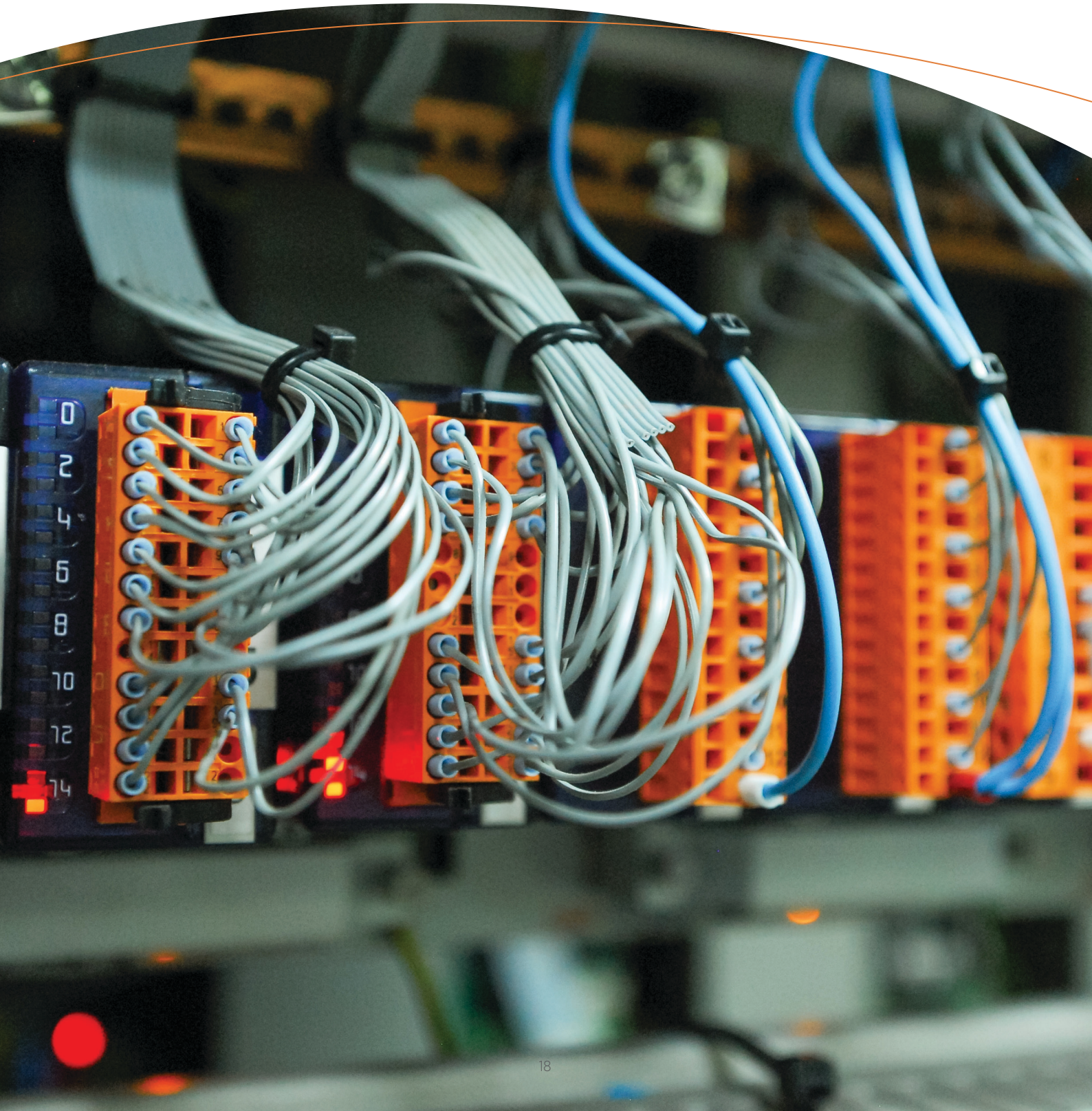
Large Companies	Dependency on CrowdStrike by Industry
Information Technology	HIGH
Transportation and Logistics	
Professional, Technical, and Business Services	
Telecommunications and Media	
Finance	
Real Estate, Property, and Construction	
Manufacturing	
Tourism and Hospitality	MEDIUM
Miscellaneous and Unlisted	
Retail and Wholesale Trade	
Healthcare	
Public Administration and Non-Profit	
Utilities and Energy	
Education	
Entertainment and Recreation	
Agriculture, Forestry, and Fishing	LOW

In the weeks since this event, many of our clients have found value in assessing their own exposures using this dataset, and comparing their portfolios to the wider industry. We anticipate that as the use of online and cloud-based services increases further, this kind of SPoF data analysis will become ever more useful for anticipating claim frequency.

That said, it is worth reiterating that this dataset is markedly more complex and difficult to analyse than the scanning data that underlies the cyber risk factor analysis referred to earlier in this paper. We believe its potential is great, but much more work will be necessary to establish its true value.

Looking Forwards

This report hopefully represents a step in the right direction towards mitigating uncertainty around utilising external scanning data, as well as better understanding security control efficacy. In cyber insurance, there is a common misconception that we 'don't have enough data'. On the contrary, cyber is rich in data, but many of these datasets are complex and will require long-term cross-industry initiatives and dialogue to fully realise their potential. This long-term work is for the benefit of insurers, the cybersecurity community, and policymakers alike. One of our current focusses is developing more sophisticated tooling to measure the financial impact of cyber data. We hope this analysis represents a useful contribution and can be built on by future studies.



Appendix

Limitations of our model

As with any machine learning model, it is important to understand the inherent limitations of the modelling approach.

- **Changing nature of cyber risk —** The cyber threat landscape is constantly changing, and we continue to see an evolution of the threat landscape. Our models learn patterns from historical data, and it is not necessarily true that we will see the same patterns in the future. Additionally, claim development time means that getting a real-time view of insurance risk using claims data alone is very challenging.
- **Correlation vs. causation —** Risk indicators can be correlated with claims activity whilst not being directly causal. Understanding causality vs. correlation is critical, particularly when evaluating new risk indicators. Machine learning models are incredibly effective pattern recognition tools, and can find correlations and patterns where no causal relationship is present. (One famous illustration of this is that ice cream sales correlate with shark attacks, but sharks are not attracted by ice cream. Rather, the underlying cause of both is that hot weather brings large numbers of people to the beach). We have found that understanding this distinction is particularly important when analysing SPoF data using AI and machine learning.
- **Insurance data quality —** Effective management of exposure and claims data remains a challenge across the insurance industry in 2024. The capture of original insured URLs; high-quality claim descriptions and categorisations; and direct links between claims and exposure remain key challenges that we see in our clients' data.
- **Variation in cyber scanning data —** As noted in previous research, we see a wide variation in vendor scores for the same company. Our research continues to indicate that specific sub-scores are more predictive than others. We have also noted that some vendors do not have full company coverage.
- **Claim frequency —** Our research to date has focused on claim frequency. This paper presents some early analysis of financial impact (e.g., loss ratios) but there is still more to do to build a complete picture of cyber risk.
- **Interpreting importance —** The core algorithms that the team has used to develop the models are strong where they are analysing highly-correlated risk factors. Whilst we attempt to remove very highly correlated risk factors, where some degree of correlation remains, our model can 'share out' the importance across these factors.
- **Training data —** Our training data utilises a large number of policy records, making it one of the largest studies of cyber risk. However, in modern machine learning terms, this still constitutes a relatively small dataset. Claim frequency is also low, particularly when developing models on specific claim types. Therefore, there is potential for spurious patterns to emerge or real patterns to be hidden.

Given the above limitations, combining an understanding of statistical learning approaches with domain expertise in insurance and cybersecurity helps to unlock value from machine learning whilst mitigating risks.

Learn more about our client-focused, collaborative approach.
Connect with us today at **GallagherRe.com**.

How can we help?

To find out more, please contact your local associate.

Ed Pocock
Head of Cybersecurity
ed_pocock@GallagherRe.com

James Poynter
Head of Data Science
james_poynter@GallagherRe.com

It's the *way* we do it.

Gallagher Re

GREUK101898