Security Ratings Are a Dangerous Fantasy

Inaccurate Results, Lousy Data, No Predictive Power, False Confidence, False Security. How Did We Get Here, and How Can We Do Better?

Security professionals don't like security ratings, also known as cybersecurity risk scores.[1] Partly this is because people don't like being criticized. But mostly it's because security ratings don't work, and cannot work as presently conceived and sold. The industry is a marketing façade. Security ratings do not predict breaches, do not help people make valuable business decisions, and do not make anyone safer.

In this white paper, we explain why the above statements are true. In disclosure, Cortex® Xpanse™ is also a cybersecurity company. We decided in 2016 that we would not launch a security rating product because, while believing there to be a substantial market for such scores, we refused to ship a product we could not fully stand behind.

We are not competing with security rating companies, and we do not directly benefit from criticizing them. But our customers have grown tired of spending valuable time explaining the results of these products to their leadership and boards. We hope that summarizing the arguments here will help save some higher-leverage labor. It's also time for the cybersecurity industry to find a better way to achieve our common goal of measurably improving external network postures.[2]

What's a Security Rating?

In theory, security ratings, sometimes alternatively branded as "cybersecurity risk scores," are analytical products that attempt to quantify the cyberrisk associated with an organization by correlating various external data about that organization. A security rating also allows for comparison between organizations, assigning the significance of various factors based on company- or sectorspecific knowledge, and measuring changes in the external data as a proxy for the organization's cybersecurity posture. In reality, security ratings are the output of a subjectively weighted function at the sophistication level of a spreadsheet cell. Security ratings arrived on the scene in the early 2010s. They aimed to solve a growing problem in third-party risk management. Before security ratings, organizations had to rely on questionnaires to assess the security of third parties, if any assessment took place at all. This process had many problems.

Self-reporting was:

- **Highly manual:** Filling out forms and reviewing them was not a scalable way to assess hundreds or thousands of vendors.
- **Imprecise:** Many questions were overly ambiguous. For example, they would ask "Do you have a patching program?" as a yes/no question without specifying the cadence or coverage of the program.
- **Inaccurate:** Surveys were self-reported. The lack of independent data often resulted in blatantly incorrect answers.
- **Point-in-time:** Surveys were typically self-reported annually, meaning that changes in IT systems and configurations that occurred after an assessment was complete were not reflected until the next point-in-time assessment.

Security ratings aimed to solve this by providing a risk score derived from data that could be gathered independently, usually outside of the organization. The basic idea was that if an organization had really poor external security (e.g., lots of expired certificates), then the inside of their enterprise network was also probably a mess.

The introduction of security ratings was an improvement over self-attested surveys, since independent data could be used to rate companies. At a minimum, security ratings can confirm that an organization's publicly attributable Internet Assets are secured and operated to a baseline of hygiene. In other words, a mature organization should have no observable risks to ratings organizations, even if having no observable risks does not mean that an organization is mature (more on this below). In this way, security ratings have the potential to complement more comprehensive security assessments. It should be relatively easy for a mature IT organization to have a good security score, even if a good score doesn't mean the organization is secure.

But these tools have introduced a host of other issues that can lead to poor decision making, and they have failed to solve the core problem of assessing and mitigating third-party cybersecurity risk.

Factual Pathologies of Security Ratings

The quality of security ratings is contingent on the quality of the underlying data, and the science with which those data are interpreted. In other ratings industries, brands like Nielsen and Gallup pride themselves on having the highest quality data, and that is a key differentiator relative to, say, a website poll. Unfortunately, the cybersecurity ratings industry has nowhere close to the depth and breadth of data of other ratings sectors, and there is no hope that the condition will improve any time soon. As we like to say in the tech industry: garbage-in, garbage-out.

High Rate of False Positives / Misattribution

One of the most frustrating aspects of security ratings for organizations is the prevalence of false positives. These often take the form of stale registration, where an IP range or domain name is technically assigned to an entity but has not been used by them in years. Internet registration is often notoriously difficult to update, meaning that naive, registration-based attribution can take months to resolve, and weeks of effort.

Even when an asset is technically owned by an organization, an outside-in rating often lacks the context to connect it with actual business risk. For example, many organizations have a portion of their business where they act as a service provider that hosts services on behalf of their customers. Exposures on these assets are the responsibility of their customers, and should not negatively impact their security ratings.

Similarly, security ratings treat parts of an enterprise network in an undifferentiated manner. A common complaint we have heard from our customers pertains to guest or research and development networks. Security ratings frequently misattribute signals of compromise or

1 If you are the first exception to the rule that we have encountered, and you are not presently an employee of or investor in a risk scoring company, please email our CEO to tell him about it (tim@Xpanseinc.com).

2 We applaud initial industry efforts to hold security ratings companies accountable to best practices, particularly the "Principles for Fair and Accurate Security Ratings" project led by Phil Venables of Goldman Sachs and others in the financial industry. US Chamber of Commerce, June 20, 2017. (https://www.uschamber.com/issue-brief/principles-fair-and-accurate-security-ratings).

security risk to an organization's network, when in fact the risks are on a sub-network that is deliberately segmented and not accredited to the same standards as the rest of the organization.

The process whereby security ratings vendors attempt to overcome these basic problems is cumbersome for companies, point-intime, and yields inaccurate results. Specifically, ratings vendors use ombudsman processes to manually review customer network maps, a process which detracts valuable customer labor to engage in the protests, and may take months to complete. Updated network ranges require manual review should they change again, and may never have been accurately documented — particularly in the case of large, complex organizations. Major events like mergers and acquisitions can wreak havoc on this process.

Security ratings companies do not have accurate network maps, and ratings are regularly deflated due to misattribution or improper understanding of network configurations (e.g., what a segmented guest network is).

Incomplete Data

Another fundamental limitation of security ratings is incomplete data, meaning that rating vendors can miss important assets that cause the most risk. For example, security ratings are known to have poor visibility into cloud environments, where dynamic hosting makes assets difficult to find, and multi-tenancy makes assets hard to attribute. Security ratings vendors use quick-and-dirty methods to map out a network footprint, and they'll often miss subsidiary assets.

Domain Name System (DNS) mapping is one of the primary wa in which security ratings vendors identify corporate assets. DNS is an appealing data source because it is a quick and easy way to find Internet Protocol (IP) addresses and domains associated with an organization. WHOIS, one of several public registration databases for IP addresses and autonomous systems (ASs), is another key data source.

DNS and WHOIS miss assets in commercial IP space, particularly those that are not domain routable. For example, if a large company is leasing IP space from a major Internet Service Provider (ISP), attribution may not be possible from any DNS or public registration data, or the relationship may be nested within regional registration information rather than returned in the top-level WHOIS lookup.

The quality of these public IP, AS, and domain registration data is, in our best estimation, declining, meaning that attribution by security ratings vendors is likely to decrease in quality below today's alreadyinsufficient bar. A reason for this is the General Data Protection Regulation (GDPR), an artifact of EU privacy efforts that requires the redaction of some kinds of public Internet records. Another important reason is the changing nature of enterprise IT. As large organizations increasingly rely on multi-cloud, hybrid-cloud, and zero-trust architectures, the traditional methods of Internet Asset attribution are becoming obsolete.

Ultimately, these deficiencies mean that security ratings are based on observable parts of an organization's network, and miss significant portions of that network. From experience with Xpanse customers, the difference between a security ratings report and reality range between tens to hundreds of percent, and our customer-validated network maps are nearly never congruent with scoring company maps.

Security ratings vendors do not effectively disclaim this limitation in their capabilities, and, from what we know, do not attempt to quantify the representativeness of their data. This means that seemingly executive-ready reporting

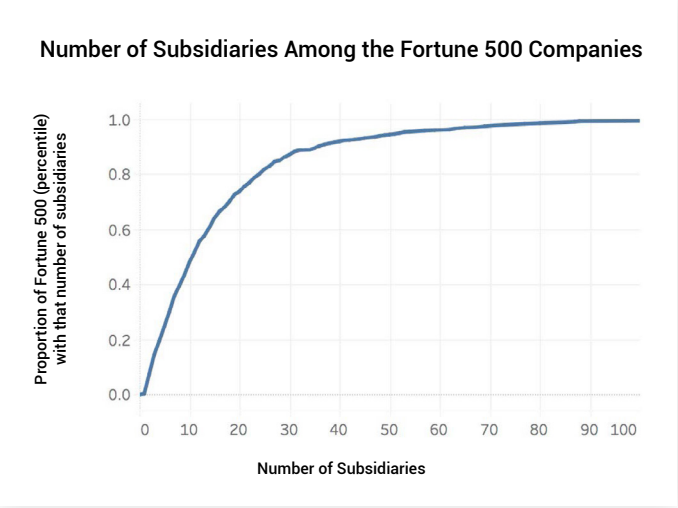

Figure 1: Number of Subsidiaries

3. Title: Number of subsidiaries among the Fortune 500 companies.
X-axis: Number of subsidiaries.
Y-axis: Proportion of Fortune 500 companies (percentile) with that number of subsidiaries.
Sources: Xpanse-generated from public company SEC filings and private industry business intelligence databases.

Typical Security Ratings

Corporate Network Map

✓ True

- Major CIDR Blocks/Assets
- Top-Level Domains

✗ False

- Expired/Transferred Registrations
- Guest Networks
- Divested Businesses or Assets

○ Potentially Missing

- Cloud Resources
- Multi-Tenant Assets
- Subsidiaries
- Commercially Leased Non-DNS
- Managed IoT and SaaS

Security ratings companies typically use incomplete third-party data, and do not communicate caveats or error estimates to their customers.

Low Data Refresh Rates

Security ratings companies advertise that their data are regularly updated, and that ratings change from day to day. This is an improvement over annual self-reporting of network status. However, what is not clear from ratings companies is which information is updated and when. Our customers report that some of the most important network change events are not reflected in their scores for as long as hundreds of days. Such update cadences are insufficient to make security ratings relevant to the required pace of defense against today's Internet-based cyberattacks, and seriously detract from their value for board-level reporting.

An important example is what Xpanse calls "active sensing," a broad category for port and web application scanning and other network reconnaissance methods that often reveal the most severe security risks to an organization. Some ratings companies conduct active sensing on cadences measured in weeks, or rely on thirdparty data updated monthly. Attackers, on the other hand, can index the entire Internet in under an hour for a given protocol. Waiting two weeks to learn about an exposure leaves too much time for an attacker to find and exploit the device. Once an exposure is identified, it can also then take weeks for the scorers to verify that the issue has been fixed by the organization.

Given that exposures and remediations change an organization's reality in an instant—such as when services are opened or closed to the Internet, or software versions change—a given organization's security rating is almost certainly inaccurate at any moment someone looks at it.

By the time you read them, security ratings are already out of date.

Conceptual Pathologies of Security Ratings

Even if the data were high quality, which they are not, security ratings cannot work as presently conceived and sold.

Lack of Predictive Power

Security rating vendor marketing targets a primary use case of assessing risk to organizations, and helping organizations evaluate the risk to their businesses from third parties. Common frameworks, like Certified Information Systems Security Professional (CISSP), require that risk managers attempt to quantify these risks.

Security ratings, however, cannot offer reliable predictive power regarding the possibility of a cybersecurity breach, and therefore fail to meet the basic product criteria laid out in their own messaging. Examples:

- Which are the top ten companies most at risk of a cybersecurity breach next year? Ratings companies cannot tell you with confidence. If they could, we would see annual lists published, and be able to evaluate the merits of the assessments in the following year.
- What is the probability that your company will be breached next year? A security rating does not tell you, and cannot inform internal frameworks required for risk professionals.
- If you have a perfect risk score, does that mean you are safe? No.

Several problems contribute to the inability of security ratings companies to reliably predict cybersecurity incidents. One of the most important is the lack of outcome-variable data to build good generalizable risk models. Major breaches are rare events. Regular, day-to-day breaches are not publicly reported. Companies have almost no incentives for self-reporting on minor security breaches. Major security breaches happen only a couple of times per year. Public datasets on breaches, privately compiled data, and even insurance claim data are incomplete, noisy, and use inconsistent cost measures.

We studied this problem with a private dataset of breaches provided to Xpanse. We found a correlation between breaches and a handful of very obvious exposure types, like Telnet open in core corporate IP space. About half a dozen metrics produced more than 95% of the predictive power of the model.

When we added hundreds of additional exposure types as observable over the public Internet, the predictive power of the model didn't meaningfully improve. This was the case even though we had direct control over quickly refreshed, time-series, complete Internet data and laboriously vetted network maps.

The implication is simple: if cyberhygiene is so abysmally bad for a given organization that anyone could find a few examples of the worst exposure types, the odds of getting breached are a little higher than for everybody else. But beyond that, predicting data breaches isn't precise.

This is one of the reasons why risk scorers present a score, and not a probability. Presenting probabilities would make it clear that the differences between large groups of companies are overshadowed by the inherent randomness of events, and even the error bars.

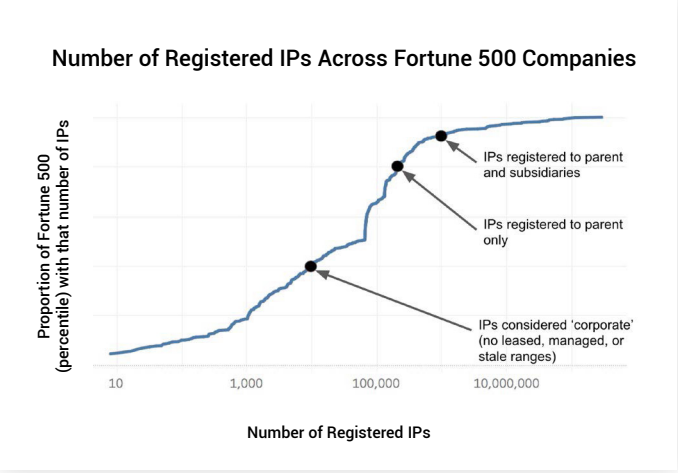

Figure 2: One large technology company has a network size in the 92nd percentile of the Fortune 500. However, if subsidiaries aren't included, it drops to the 84th percentile. And if only ranges used for corporate purposes are included, it drops to the 40th percentile. Almost all companies have similar variability depending on how network assessments are done, which leads to huge implicit error bars.

some kind of breach is more likely, security ratings cannot predict how bad those breaches will be. This diminishes the predictive value of ratings products to almost useless, because they cannot identify where the most costly and consequential risks are.

Security ratings cannot tell you if you or a third party are at risk of a costly cyberattack.

Unknown Local Controls and Layering / Defense in Depth

Security ratings data are almost exclusively based on observations external to the organization. This is by necessity: ratings products attempt to non-intrusively identify and assess risks. An underlying principle motivating this aspect of ratings is not unreasonable: exposures that are directly Internet-accessible are the worst, because anyone in the world can find them, automated tools are getting better at finding them, and, empirically, most breaches start with an insecure public Internet Asset.

However, an unsolvable, fundamental problem is that it is impossible to tell from the public Internet what local mitigating controls may be in place, and what devices on the public Internet are connected to what other things inside the corporate network. Layering cybersecurity is a basic principle in industry certifications and best practices. Ratings companies also cannot identify which parts of a network are properly segmented versus those that are bridged to more. Unknown Local Controls and Layering / Defense in Depth.

Internet Exposure	Mitigation not externally visible
Exposed RDP	Multi-factor authorization
Unpatched server	IP-based whitelisting does not allow connections
Unpatched WiFi router	Not connected to corporate network (guest WiFi)
Unencrypted FTP	One-way uploads of non-sensitive docs not connected to corporate network
Operational technology admin interface	Device in read-only mode
Malware user agent hits honeypot	Guest laptop on guest network, not connected to corporate network

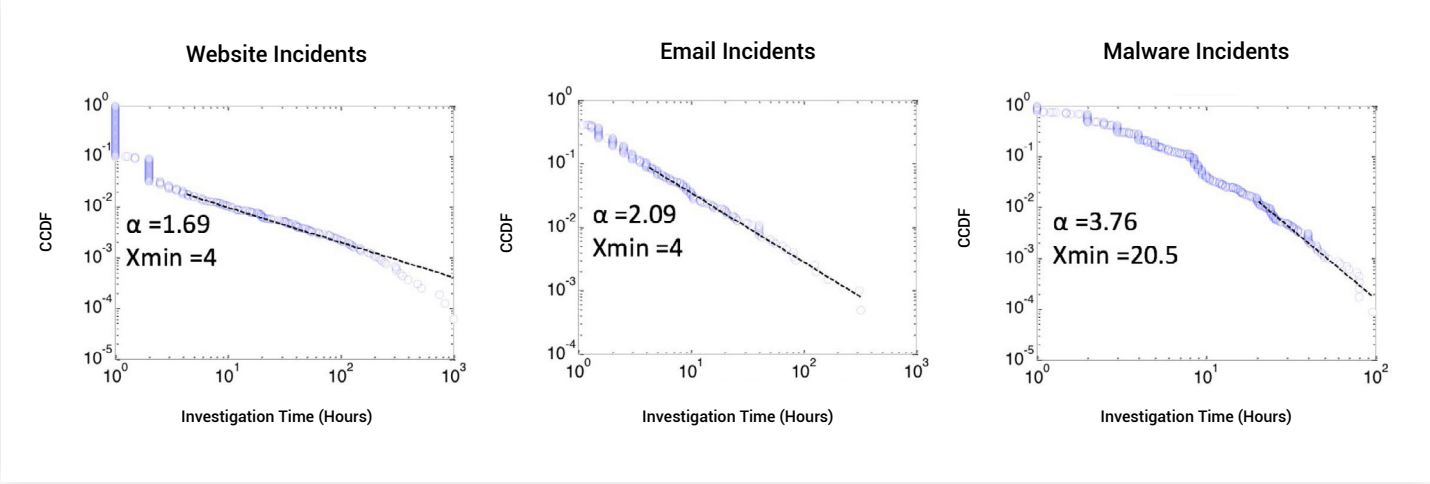

Figure 3: The Complementary Cumulative Distribution Function (CCDF) for several types of cyber incidents, recorded at a large organization. The impact (hours of investigation time to resolve the incident) is shown on the X axis. Note that the linear relationship on a log-log plot indicates a heavy-tailed distribution.
Taken from Kuypers, Marshall. "Risk in Cyber Systems." 2017. Stanford University Dissertation.

For most breaches that began with an Internet-facing exposure, the exposure was only the first step in an attack sequence that involved downstream failures.

One organization could have 1,000 exposures that we would consider very bad in the industry, and another organization could have only one, but if the 1,000 are not on a core business network and the one is, then the latter is a serious security problem and the former possibly are not. No one can tell from the outside.

Security ratings companies typically use incomplete third-party data, and do not communicate caveats or error estimates to their customers.

Like Many Catastrophes, Tail Events Dominate

Data breaches are not like car insurance claims, where most of the losses are about the same size. They are like wildfires, where the most damaging fire is more destructive than nearly all the other fires combined. These events form a heavy-tailed distribution, which is a special type of relationship characterized by many small events and rare massive events that are orders of magnitude larger.

Data breaches are well known to follow heavy-tailed distributions, and this probably agrees with your experience. The most costly incident at your organization last year was probably more impactful than every other incident you had that year. Dealing with heavy-tailed distributions when you have a portfolio of business partners or vendors is difficult, because your priority is avoiding the one big incident, not dropping the top 10% of risky vendors.

Security ratings do not lend themselves to these types of events because the scales don't match up. Losses from cyber range from $10,000 for small incident cleanups to hundreds of millions in losses for the largest events. How does a score of 720 or A, B, C, D, or F grade compare to these losses? Representing extreme events with simple scores makes it difficult to compare companies and trends accurately.

Security ratings cannot tell us what to care the most (or least) about.

Methodological Pathologies of Security Ratings

Subjective Weighting of Key Variables Based on Trailing Indicators

Given the endemic data and conceptual challenges of security ratings, vendors committed to a ratings product have no choice but to hack their way to a partial solution. The partial solution manifests in a subjective weighting of multiple factors that will almost never perfectly align with the security priorities of a given organization.

Every organization has a set of "crown jewels" with subjectively assigned business value that is unknown to outside parties like risk scoring companies. This is part of the normal information security risk assessment process codified in standards like CISSP and the NIST Framework.

For most businesses, the two most important categories of risk are intellectual property theft and business interruption. But which of these is a priority, to what degree, and for which assets? All of this is highly dependent on the business, the internal network structure, and the other security controls that are in place.

For instance, security ratings can generally indicate that an Internet exposure is a cybersecurity deficiency. Ratings vendors cannot meaningfully approximate how much to reduce a given business's score based on that exposure. The determination of how much a given observation impacts a given business's score is a subjective judgment made by a person as applied to a given cohort of companies — like an industry/sector — and is not informed by that organization's years of work in assigning business value to its assets.

A couple of examples help to demonstrate this point. An exposed A couple of examples help to demonstrate this point. An exposed database server could be a sign of shadow IT or a misconfiguration. It could also be a sanctioned DevOps experiment. Or an irrelevant business risk, because the database contains no sensitive data. In the case of subjective industry weightings, a ratings company could choose to weight database exposures more heavily for healthcare. But a clinical practice might use an electronic health records (EHR) system for patient data and consider that system a crown jewel. An exposure of an internal web page for that system would be considered far more serious than a separate test database. Ratings companies do not have the sensitivity or context to discriminate between them, and are far more likely to overweight our hypothetical database exposure and underweight a web login. Another clinical practice performing the same services but running a different IT architecture, on the other hand, may indeed keep radiology images in a local database for research or archival purposes, and so a breach for that practice could be devastating in terms of insurance and regulatory punitive costs.

So, by how much should a security ratings company decrease a clinical practice's score when it observes a database server?

Another example is that almost every organization has expired certificates, and certificate hygiene is generally a low-rated category. For retailers, browser warnings for consumers could entail significant business interruption costs, especially during a holiday season. Security ratings' subjective weightings cannot capture the true business-dependent risk associated with even these routinely observable asset types.

These two hypotheticals demonstrate a baseline problem of all security ratings, and the complexity explodes across many thousands of Internet-observable data types. Scientific methods do not inform whether a given observation decreases a score by 5% or 20%; the people working at the security rating company decide. Even in a best-faith effort for ratings to adhere to our best known reality and industry standards as conditions change, this problem can never be overcome in the absence of organization-specific context.

Ratings are whatever product managers want them to be, and are not based on standards or risk science.

Network Size

Security ratings often overestimate the health of small organizations, especially those that are highly deployed in the cloud. Smaller organizations have a much smaller public attack surface and may receive high scores simply because there is little to find, not because their security practices are more mature than larger organizations. Indeed, the reverse is probably true; a large organization with professional IT, security, and audit teams almost certainly has better cybersecurity programs and employee training than a small business with largely outsourced IT. But when ratings are based on score reductions from the discovery of Internet-routable assets, a large organization could appear to have far worse security than a single-IP small business.

Ratings do not make sense for the vast majority of businesses, which are small, third-party-managed networks with a tiny Internet attack surface.

Ratings Are Not Actionable, and Therefore Waste Valuable Time

Given the above problems, security ratings cannot be used to make business decisions, either about one's own organization or in critical assessment of third parties. At best, security ratings can be considered informative and as part of a conversation regarding Internet attack surfaces and what they should look like. When major cybersecurity events happen, ratings data can help the industry understand where major categories of risk reside and in what quantities. This directional guidance is useful because it can help inform the creation of new standards, training, and, potentially, products. Ratings data are not useful, however, in effecting specific business outcomes.

We see a tacit acknowledgement in the industry that ratings are not actionable. Have you ever chosen not to do business with a company because they have a poor security rating? In the absence of representative survey data, we evaluated this question in light of customer feedback (including among customers with many thousands of suppliers and vendors), and found that other variables, like cost and quality of product, consistently win out over cybersecurity posture as evaluated by security ratings companies.

Risk scoring has become a box-checking exercise where organizations need to meet a minimum threshold. While meeting a minimum threshold does add value over legacy self-reporting, that value diminishes rapidly in the context of driving meaningful business decisions.

Because security ratings are unreliablefor all of the above reasons, companies cannot use them to make important business decisions or to drive security outcomes.

So, If Not Security Ratings, What Then?

A major bank customer recently remarked to us that "as an industry, we have to blow up the ratings companies and completely start over." Naturally, that begs the question of what should be done instead, as the pre-ratings world of relying on self-reporting and written checklists was also deeply problematic.

We propose three paths forward based on what we've learned from experience and discussed with customers. Naturally, each of these is ambitious, and answers won't come overnight. We've sequenced them here in order of achievability with current products and technologies.

Subjective Weighting of Key Variables Based on Trailing Indicators

A promising trend we've seen is for large companies to provide various kinds of coverage for key suppliers that cannot afford their own cybersecurity programs. Consider that small businesses may have a single-digit-sized IT staff, and that IT staff's job is usually to manage contractors and software subscriptions. Such a company likely has a managed security service provider (MSSP), and does not have its own security program. Most MSSP-dependent businesses will not have access to or budget for the most cutting-edge data and technology; security already comes out of operating expenses and thin margins.

Large companies can help by providing notifications of security exposures to those suppliers as detected by their in-house threat intelligence functions, third-party services that are able to continuously monitor networks for risks, or subscriptions (software or data) that can provide both the parent company and its key supplier with alerting. The most mature sector where this trend is playing out seems to be defense.

Promote Within-Sector Information Exchanges

Many large organizations have a high degree of overlap in suppliers and a shared interest in securing their supply base. Financial services corporations, defense contractors, and other large players can team up to monitor public-facing exposures on their supplier networks, and then notify the supplier to require remediation. Information-sharing collectives have been effective in some industries, like the FS-ISAC for financial services and the H-ISAC for healthcare companies. Large organizations can push this expertise downstream by encouraging smaller suppliers to take part in sector-specific exchanges.

Build Risk-Assessment Partnerships Across Levels of the Security Stack

Another direction is for major vendors to team up to develop and measure progress against a cybersecurity maturity model, combining the best externally available data with internal technology, network, and process maturity observations. We've heard variations on this theme for years, usually in the context of something like an "Internet weather report." The problem, as noted above, extends well beyond what is on the Internet, though, into getting a holistic view of an organization's cybersecurity profile. The best we have seen to date are local solutions, like Microsoft building the Azure Security Center for its customers with a security score associated with Windows and O365 configurations. A partnership dream would be for Microsoft's internal telemetry to be joined with ISP flow data, vulnerability scanning, and other active sensing data, firewall logs, APT signatures, and so on.

Our idea may sound fantastic or cliché, depending on where in the world you sit. It can sound fantastic because it requires a lot of cooperation between companies that sometimes compete with each other. But it is also what the world's most sophisticated organizations are already doing for themselves. For a small number of very skilled software companies, like Google and Facebook, the combined IT risk view results from the company simultaneously being its own ISP, CDN, intelligence shop, custom hardware designer and manufacturer, and system integrator.

There is no product that other companies can sign up for; the only solution to date has been massive investments in labor for data integration and engineering, and therefore implementation has been limited to only a handful of very large organizations (think banks and defense contractors).

This vision is consistent with broader industry trends in what we know customers want: the consolidation of various products, tools, and data feeds; the simplification of information that is presented; and the centralization of visibility and control over modern enterprise IT.

Conclusion: Why Dangerous?

Invention is hard. Selling and marketing are hard. We know.

But integrity is more important than living the startup dream. We'll give the founders and early employees of security ratings companies the benefit of the doubt, and hope that they invented their first products with good intentions to try to help fix cybersecurity problems. We assume that their plan was to iterate, learn, and improve on their first products.

Unfortunately, the marketing hype has vastly exceeded the technology and product grasp of these businesses. After nearly a decade in the market, the accuracy of security ratings products still isn't very good, and the data aren't very useful. But the marketing has become so effective that many companies are even contractually required to use risk scorers by their business partners. In other words, the idea of risk scoring has become more successful than the actual technology. The security ratings industry has not just failed its customers; it has also created a dangerous condition that has to be fixed.

The dangers are that companies feel a false sense of confidence regarding their security posture based on a high risk score (which, as we've explained, does not mean that the organization has good cybersecurity). Security ratings have created false insecurity and internal turmoil at organizations where executives wonder why their rating is not best-in-class, but the on-the-ground reality known to security and IT professionals is that the rating is not reflective of reality, because the networks were not mapped correctly… or the data weren't refreshed recently… or assets were deliberately open to the Internet and segmented according to best practices… and so on.

In other words, ratings companies have distorted reality for the sake of a cheap, nearsighted market advantage. These distortions have the potential to misallocate valuable and scarce resources, like expert labor-hours and dollars for technology.

If we really want to make cybersecurity and Internet safety better, then we have to start with a common understanding of the problems, and proceed to then build technology and process solutions. Reducing the complexity and nuance of a highly technical practice to a round number or letter grade takes us farther away from reality. And is an unwelcome distraction for those of us still living in it.

paloalto NETWORKS

3000 Tannery Way
Santa Clara, CA 95054

Main: +1.408.753.4000
Sales: +1.866.320.4788
Support: +1.866.898.9087

www.paloaltonetworks.com

cortex-xpanse-Security-Ratings-are-a-Dangerous-Fantasy-wp-05042021