White Paper

The Need for Threat Risk Levels in Secure Web Gateways

Understanding How Threat Risk Levels Help with Uncategorized Sites, and How It Works with Web Isolation Technologies to Strengthen Security

Introduction

Secure web gateways (SWGs) have long been used by organizations to provide secure web access for their employees. In recent years, with the increased threat level coming from cyber criminals and the increased number of threats coming specifically from the web, SWG technology has become even more important. SWGs have been adapted to the new landscape to detect malicious and suspicious sites in real time, in addition to examining objects downloaded from the web using advanced threat technologies. These advanced threat technologies include anti-malware, sandboxing, behavioral analysis, and others. Symantec™, a division of Broadcom, introduced the Threat Risk Levels feature to our SWG products. This feature helps IT administrators deal with the variety of threats coming from the web.

This document outlines how the Threat Risk Levels feature works, its underlying technology, and how to use it effectively in SWG policies.

Categorization and Real-Time Rating

SWGs started out utilizing on-device category ratings. On-device databases provided the primary means of identifying malicious and undesirable websites. This process allowed for easy policy implementation, such as blocking users from accessing pornography or shopping websites while at work. However, as the number of new websites quickly grew, on-device category databases were no longer able to keep up. The result was unacceptably high percentages of uncategorized sites. Organizations needed a real-time rating technology to bridge the gap between known and unknown websites.

Real-time rating uses cloud-based artificial intelligence (AI) engines to categorize websites as they are requested by users. The use of AI significantly reduces the number of sites the SWG is unable to categorize. Sites categorized from the cloud intelligence are also included as needed in the on-device database for on-premises SWG deployments to maximize performance.

Our dynamic real-time rating system has evolved to become a highly advanced AI engine in the cloud. It comprises over 300 separate modules that identify a web page's content. Categorization of content occurs in real time. In most instances, a category is returned in a fraction of a second once the web page contents are retrieved. Some of these modules can identity over 60 languages.

For more information on the Symantec real-time rating system, see our white paper on WebPulse.

The real-time rating system currently categorizes each website into as many as four categories. There are 85 categories in all. The following 14 categories are generally considered dangerous or potentially risky:

- Malicious sources (malnets)
- Phishing
- Potentially unwanted software
- Malicious outbound data (botnets)
- Spam (questionable legality)
- Suspicious
- Proxy avoidance
- Placeholders
- Dynamic DNS host
- Pornography
- Gore (extreme)
- Gambling
- Spam
- Compromised sites

Every organization should block these 14 *security* categories in their default SWG policy.

Why Uncategorized Sites Exist and Why They Are a Problem

While real-time categorization can classify a majority of new websites, some sites fail to get enough *votes* from the various AI engines to be conclusively classified into a specific category. This issue might be due to a lack of information coming from the website itself. For example, information will be missing if the website is blank, or the website hosts only images. In these instances, the real-time rating system returns an *uncategorized* rating (also referred to as *none*) rather than a category rating.

While the number of uncategorized sites is relatively small, in a given day's traffic they remain a problem for many IT and security administrators. Prior to risk level ratings and web isolation technologies, there were virtually only two ways to handle uncategorized sites. You could allow the sites or deny them. Each option came with its own problems.

Organizations that wanted tighter security typically blocked access to uncategorized sites. This decision resulted in a greater number of help desk calls from users wanting to gain access to blocked sites. On the other hand, organizations that allowed access to uncategorized sites often found higher infection rates and a greater number of compromised systems. Their incident response teams had to spend more time remediating affected systems. Neither option worked well. Security professionals needed a better solution.

Symantec introduced the Threat Risk Levels feature to help with the problems caused by uncategorized sites.

Where Threat Risk Levels Come From

As mentioned previously, sites are uncategorized because the AI engines do not have enough information. The website and its associated metadata do not provide enough information to classify the site definitively into at least one category. Yet even without enough information for categorization, there is still quite a bit of information available to the AI engines (generally referred to as our WebPulse system).

A few years ago, Symantec researchers realized that this information, while insufficient to return a category, could be used to help set a risk level rating for the website. For example, the history of the website's host IP address could be researched, or characteristics of the URL or the server's behavior could be analyzed. While any of these factors considered alone would not be enough to rate a site as a higher risk, they could be combined with other parameters and factors from other AI systems to derive a risk level rating.

To provide a risk level, Symantec researchers developed three systems to work in conjunction with the existing WebPulse system.

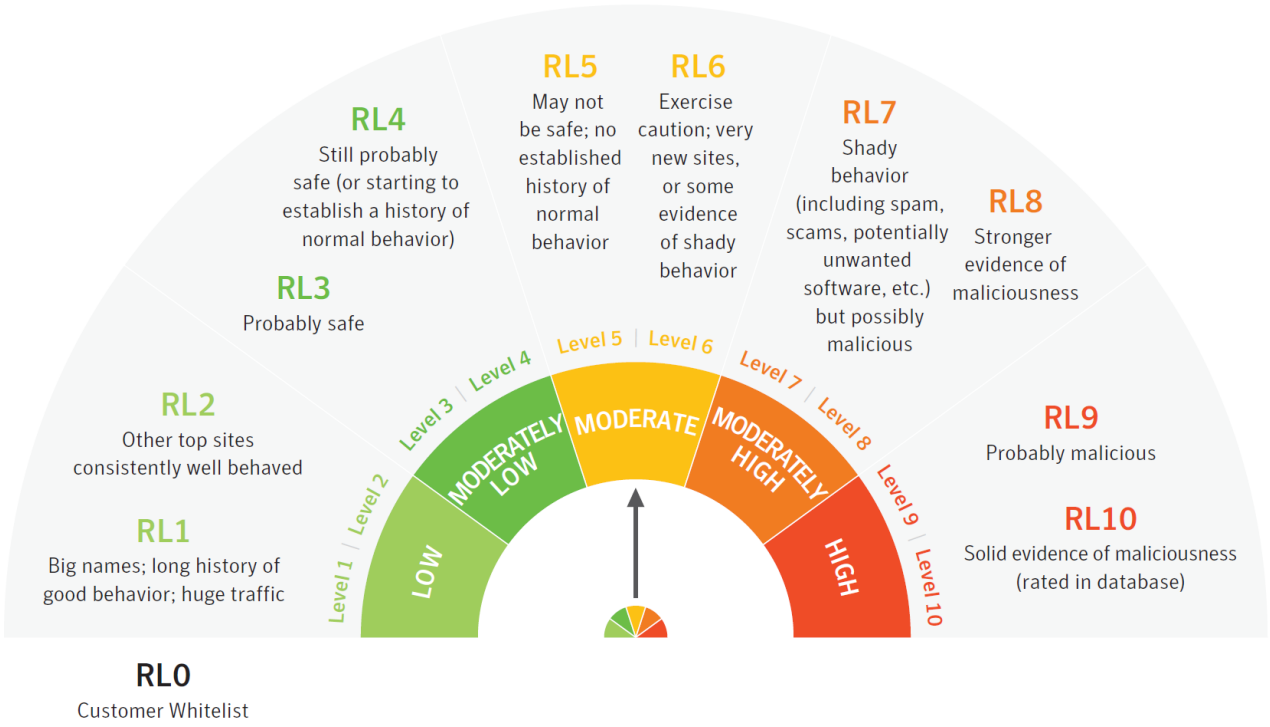
The first system is the Risk Levels Calculator. It is run daily for all entries in our URL database, and it is run whenever new information is added to the URL database. It acts just like its name implies, calculating a risk level for every website. The Risk Levels Calculator uses a snapshot of the current history that might include the reported ratings and recent traffic. Websites are assigned a risk level from 1 to 10 regardless of category (or lack of category). A risk level of 1 is the safest, and a risk level of 10 is the riskiest. Organizations can also separately assign a given website a risk level of 0 (zero), which is the equivalent of an allow list.

The Risk Levels Calculator works on the URL database which contains millions of entries. Our data centers actually see hundreds of millions of domains, subdomains, and IP addresses in daily traffic. But a majority of these items are not typically entered in the database because there are few traffic requests for those sites. These sites are what we call *one-day wonders* (sites that live for only one day or less) or *one-hit wonders* (sites asked for only once).

The second system is the Context Engine. It is a massive-scale AI engine that tracks and renders information about the hundreds of millions of domains, subdomains, and IP addresses that we did not store in our URL database. The Context Engine may add sites to the database if its confidence level is high enough. It also provides summarized context to the WebPulse system for use in its real-time voting system.

The third system is WebPulse real-time voting. It calculates a risk level for all requests not rated by the database. It combines votes from a variety of modules, including the *context* provided by the Context Engine.

Figure 1: Symantec Intelligence Service and Threat Risk Level



A Brief Explanation of WebPulse and the Context Engine

Symantec started with the problem of having to classify the internet for categories and risk levels. Given the massive scale of the internet, accomplishing this goal required that we combine AI with big data—which we call *Big AI*.

We needed to develop risk levels for hundreds of millions of subdomains, domains and IP addresses, but historically we did not have enough information to categorize and place them in our URL database.

Here is what happens when a request comes in for categorization in our SWG solutions (cloud, on-premises, or hybrid). If the risk levels for a website are not included in the URL database and the information was not requested previously by another user, then the request gets sent to the WebPulse service.

WebPulse uses a voting system that is similar to a voting system that uses caucuses and delegates. This type of voting system is used in many primary elections in the United States. If you are not familiar with this voting system, here is a link to a quick primer.

When the request goes to WebPulse, the system checks for existing information on that website. Other organizations may have already asked for the same information, in which case the categorization and risk level rating would be cached in the cloud and available immediately. If no information is cached, then WebPulse checks with the Context Engine for information on the website. Context Engine has more votes to determine categorization and rating in the overall system than any other part of the rating process.

If there is no information for the website on the Context Engine, WebPulse also takes that into account. Next, the dynamic real-time rating system takes over.

The Context Engine ingests all the data it can find, including the billions of classification requests we get daily. In addition to the WebPulse logs, Context Engine uses a number of feeds from third parties, multiple feeds from other security organizations within Symantec, and other data. In addition to contributing votes on the risk level, Context Engine may also provide information about a category for a site. The most relevant information to WebPulse is then pushed out from the Context Engine to WebPulse.

WebPulse uses the metadata (or lack thereof) from Context Engine to help it calculate a dynamic risk level for every request it receives.

WebPulse Dynamic Voting System

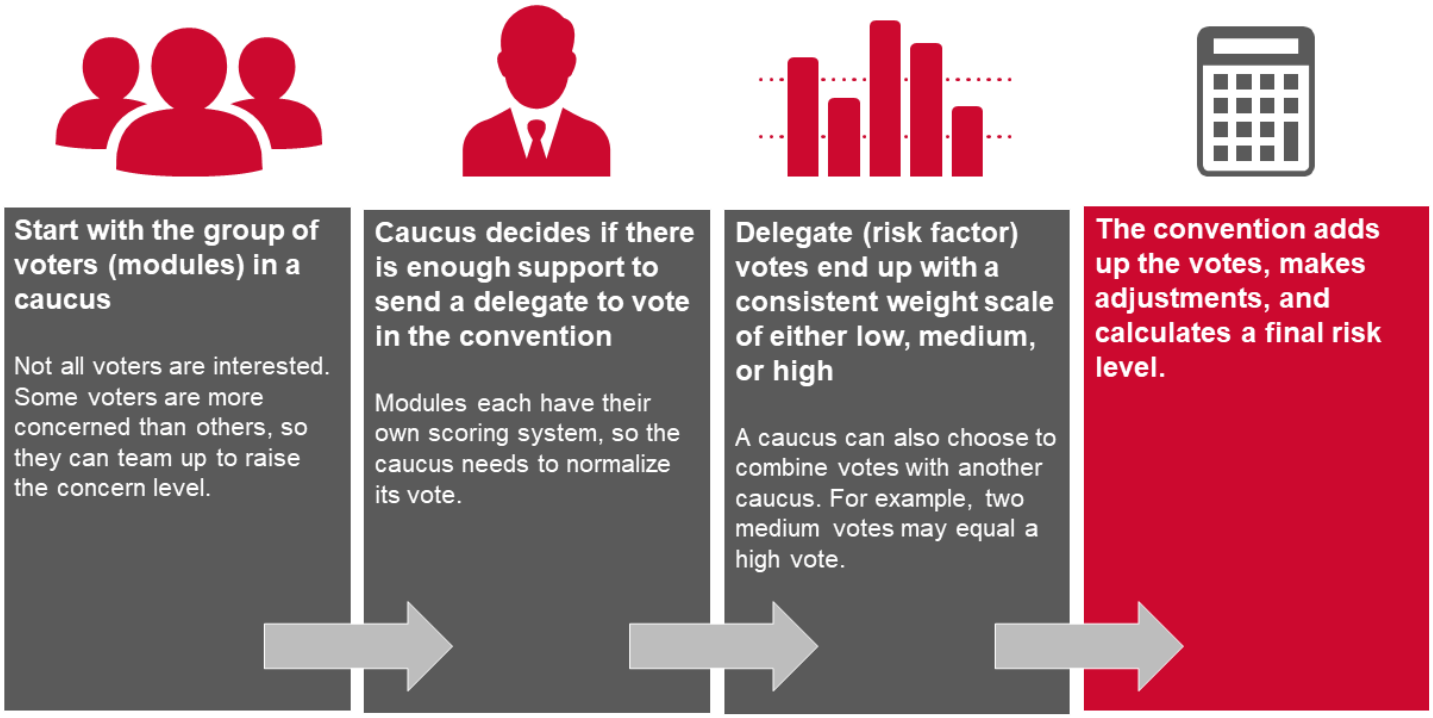
During the WebPulse calculation process, any of the WebPulse voting modules can contribute to the risk level rating. The modules are grouped into related areas (for example, by IP address) to form a caucus. Approximately 20 caucuses are allowed to vote in the current dynamic voting system.

Each caucus can send a vote to help set the final risk level rating assigned to the website. The caucus individually determines if there is enough support to send a delegate (vote) to the convention.

Each of the modules in a caucus may have its own scoring system, so the caucus needs to normalize its vote. Delegate (risk factor) votes end up with a consistent weight scale of low, medium, or high. A caucus can also combine votes with another caucus. For example, two medium votes may equal a high vote.

Each caucus vote goes into the final part of the system, which we call the convention—to maintain the political theme. During the convention, WebPulse analyzes and scores the top four votes. This result takes into account which caucuses voted, and how strong the vote was from the caucuses. The convention adds up the votes, makes adjustments, and then calculates the risk level.

Figure 2: Risk Factor Data Analyzed by Voters Grouped Into Caucuses



These risk levels are calculated with every request, so any significant new data can change the risk level. Finally, WebPulse shares the overall risk level with the requesting SWG system along with one or more content categories—if any categories could be determined. A sample of the caucuses and modules involved in the dynamic voting system is shown in the following figure.

Figure 3: Stages of Analysis

STAGE 0 (Database)	Database has category and risk level. Later stages always produce a risk level (and a category if possible).	**Database is fed by research, background checks (see Stage 5), and feeds (including customers).**
STAGE 1 (Request)	Everything starts with a request (we have the full URL and the HTTP headers to work with).	**Includes:** • Shady Traffic Group (Traffic Cops, and so on) • New/Untrusted Group • Shady Name Group (Weirdo, Ghostbuster, Shady TLD, and so on) • Shady Hosting Group (DynDNS, FreeHost, and so on)
STAGE 2 (DNS)	Now, we have the resolved IP address to work with.	**Includes:** • Server DNA Group (Malnet Tracker, and so on) • Shady Neighborhood Group (IP, Range, and so on)
STAGE 3 (Response)	Look at the response from the destination server (including the content it returned).	**Includes:** • Shady Response Group • Network Error Group • Lie Detector Group • Shady Content Group (EXE, JAR/ZIP, PDF, JScript, ...)
STAGE 4 (Post-Analysis)	Take a second look (or a deeper look) at everything we got to this point.	**Includes:** • Phishing Detectors • Metarule modules • Yara Rule modules **This is also where the caucus and convention voting system runs.**
STAGE 5 (Background)	We want to gather, collate, and analyze as much data about the site and its behavior as possible.	**Includes:** • Malnet Tracker Group • Site Profile Group (through Context Engine) • IP Profile Group (through Context Engine) • Hacked Site Group (through Context Engine) • Chatter Group (through DB) • DeBRa Group (Deep Background Raters)

Web Browser Isolation Primer

The Threat Risk Levels feature is not the only technology available to help with uncategorized sites. Symantec Web Isolation was introduced to our SWG solutions though the acquisition of Fireglass, a leader in web isolation. Web isolation (also known as remote browsing) is a fundamentally new way of handling potentially dangerous websites.

Instead of executing the contents of a requested web page directly in the user's browser, the Web Isolation technology executes the web page using a remote, cloud-based virtual browser, and delivers a safe rendering (visual stream) of the page's information to the user's browser. By executing the web content remotely, the Web Isolation service prevents any malicious code from reaching and executing on the user's system.

For more information about Web Isolation, see: www.broadcom.com/products/cyber-security/network/gateway/web-isolation.

Using Risk Levels and Web Isolation for Categorized and Uncategorized Sites

Used together, risk levels and web isolation are powerful tools for handling both categorized and uncategorized websites. Let us look at the uncategorized problem we described previously in this document.

For sites not in the database, WebPulse will be consulted and a risk level from 1 to 10 will be returned. We generally consider a site with a rating from 1 to 3 to be safe, and we recommend that an organization's policy allow users to freely browse that site. Organizations with extremely sensitive security requirements can set up a policy to send this traffic to isolation, so that the user is further protected from any potential risk.

With a rating from 4 to 6, there is some suspicion regarding the safety of the website. Here the recommended course of actions has traditionally been to put in a coaching page recommending that users do not go to the website. The coaching page would typically offer two options: leave, or proceed with notification that their activity will be monitored.

In this instance, we recommend the use of Web Isolation regardless of the organization's security stance. With Web Isolation, we can offer a smoother, yet safe option for the user and organization. For suspicious sites, web traffic is sent through isolation, the user safely gets to their desired location and content, and the organization is protected. The site can be further analyzed and eventually assigned a category and potentially blocked, or deemed safe.

For the remaining risk levels of 7 to 10, we recommend blocking these sites completely since they have indicators of high risk. For organizations that absolutely need to provide access to these sites, Web Isolation is warranted, as is blocking any downloads.

Organizations that want a higher security profile, and are moving to risk levels for greater security, should consider setting a block level starting at risk level 6. Highly sensitive organizations with skilled security teams familiar with WebPulse should consider blocking sites starting with risk level 5. This recommendation is also true for organizations that have historically blocked uncategorized websites. Now we will look at categorized sites.

Certain categories are inherently risky. For example, it is well documented that the pornography category has a high percentage of malicious sites. The Threat Risk Levels feature gives an added dimension to any website regardless of categorization, and offers an additional tool in deciding how to proceed with data-sensitive web categories such as healthcare and finance.

In the cases of healthcare and finance specifically, many organizations have recently needed to protect users' privacy regarding their records. This need competes directly with the emerging requirement to inspect encrypted web traffic, which cyber criminals are increasingly using to distribute malware and to collect personally identifiable information. Risk levels enable the IT administrator to allow safe sites (rated 1 to 3) to be accessed without SSL inspection—thereby protecting user privacy. Riskier sites can be decrypted for additional inspection, including advanced threat detection and data leakage protection.

Other categories that benefit from Web Isolation technology are webmail and online file storage. For example, Gmail and Dropbox are rated as risk level 1 sites. This risk level means that the site themselves are safe since they are extremely unlikely to be hacked, but these sites could be hosting unsafe files. Because unknown content is stored on these sites, these categories are ideal for use with Web Isolation—which allows for remote viewing of documents without the need to download them to the local machine.

Another category presenting an interesting use case is web ads and analytics. These sites tend to have a wide array of risk levels. It makes sense for organizations concerned about *malvertising* attacks to block these sites more aggressively starting at risk level 4 or 5. It is much more likely that malvertising would be involved with these slightly higher risk level ad sites.

Conclusion

Risk levels help solve problems faced by many organizations using content filtering technologies on their user's web access. Risk levels give organizations more flexibility in handling user's requests to visit new and unknown websites. This is done without the typical problems of completely blocking access to uncategorized sites or the increase in compromised systems that results from completely allowing access to uncategorized sites. Risk levels also provide a way to effectively use advanced technology like Symantec Web Isolation (formerly Fireglass). Risk levels are found in the Symantec Intelligence Services offering, and all SWG products—including the Web Protection Suite. The Web Protection Suite allows customers to deploy a comprehensive web security solution in the cloud, on-premises, or as a hybrid solution.

For more information about the Web Protection Suite, see: www.broadcom.com/products/cyber-security/network/gateway/web-protection-suite

NIS-SWG-WP100 September 8, 2021