White Paper

Symantec® WebPulse

A Technical Overview of WebPulse Collaborative Defense in the Symantec Global Intelligence Network

Introduction

The web has become an integral part of virtually every business. At the same time, social media has transformed the web into a dynamic and complex environment—ideal for the proliferation of malware in increasingly covert and sophisticated ways.

This evolving environment makes it more difficult to manage web access and bandwidth usage. It also introduces security challenges that web proxies are uniquely suited to address. It is critical that web security solutions provide accurate website categorization and risk assessment, global coverage, and real-time categorization of previously unseen URLs.

Symantec WebPulse, part of the Symantec Global Intelligence Network (GIN), is a cloud-based infrastructure specifically designed to harness the power of user-driven behavior and to translate user input into global web intelligence and web threat intelligence. We believe WebPulse, launched in 2004, is the most advanced and most relevant web security technology in the industry today.

Symantec GIN is a collaborative defense cloud that powers Symantec secure web gateway solutions by integrating input from over 15,000 diverse enterprises, 175 million Symantec Endpoint Protection users, and billions of emails scanned by Symantec email security products.

WebPulse uses multiple technologies to analyze this input to deliver the fastest and most accurate web categorization and ratings of any vendor. Within the WebPulse framework, each incoming URL request is processed by many different threat analysis methods, both automated and manual.

WebPulse uses its cloud infrastructure to deliver web intelligence to Symantec Edge and Cloud Secure Web Gateways. WebPulse seamlessly delivers frequent database updates as well as new defense types, such as analytical methods and additional language support. Users benefit instantly from these new defenses and updates without having to update their appliances or SaaS.

The Symantec threat research team—over 3,500 researchers in nine research and development centers around the world—supports WebPulse and the Global Intelligence Network.

The intent of this document is to provide insight into WebPulse collaborative defense, which is an integral part of Symantec anticipatory security defenses.

Symantec Web Security Architecture

Preemptive, layered web defenses: It is no longer effective to simply detect and block known threats. With the sophistication of malware techniques and the advent of mass-market malware—attacks that require little investment but achieve high penetration—web security solutions need to anticipate malware so they can block it before a business is infected.

An anticipatory, layered web defense requires five key components:

- **A global collaborative cloud intelligence infrastructure**—which requires the following:
  - An integrated, crowd-sourced ecosystem that can take advantage of the experience of real users who, collectively, visit tens of millions of web pages each day.
  - A cloud-based infrastructure that can use multiple threat-detection engines, machine analysis, and human raters to aggregate and analyze data from the community.
- **Real-time content filtering:**
  - Backed by a global collaborative intelligence system, real-time content filtering combines dynamic protection with the granular category control that businesses need to implement their acceptable Internet usage policies.
  - Extensive category coverage and the ability to assign multiple categories to a given URL to provide the multi-dimensional control that it takes to manage today's complex web environment.
- **Inline threat detection:** Inline threat analysis today is itself multilayered as well as an integral component of

a preemptive web security solution. It must inspect SSL-encrypted traffic as more malware is hiding in SSL and using it as a communications channel. It must also inspect user-authenticated software downloads from the web, attachments sent through webmail, and other content. It should not just include basic antimalware scanning, but also include block listing/allow listing, behavioral analysis (including static code analysis), sandboxing, machine learning, and so on.

- **Web and cloud application and content controls:** This layer consists of web content, cloud, and web application controls that prevent downloads from unknown websites, detect masquerading files, and allow or deny web applications or web application operations (for example, post message or upload attachment) based on users, groups, or other policy variables.
- **Protection for remote and mobile users:** The number of remote and mobile users is steadily increasing, and they require the same level of protection as users at the corporate office. Extending the protection of WebPulse through a web-filtering client or SaaS gives these users predictive defenses and reduces the risk they will bring malware into the business network.

Symantec Intelligence Services/ WebPulse: A High-Level Overview

Symantec Intelligence Services, in conjunction with WebPulse collaborative defense, plays several key roles in this multilayered defense. These include preemptively responding against zero-day attacks, preventing 'phone home' attempts from spyware and even botnet-infected systems, and detecting phishing and malvertising threats. By preventing malware from being downloaded from the Internet, combined with in-line advanced threat protection, the defense achieves maximum effectiveness.

Symantec Intelligence Services and WebPulse are designed as a highly responsive, preemptive, front-line defense for advanced threat protection—not as its replacement. WebPulse can simply be described as a basic input-output system. The massive input is generated by more than 15,000 enterprise customers, including those using Symantec Cloud and Edge proxies, Security Analytics, Symantec Endpoint, Email Gateways, and consumer users.

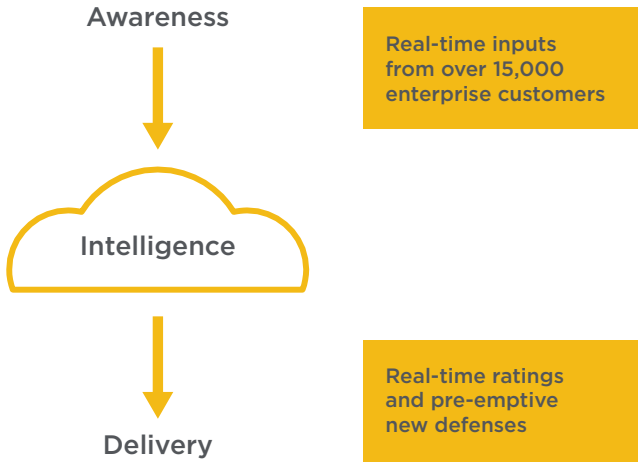
WebPulse, as a black box, performs real-time analysis and outputs a URL categorization and a Risk Level rating. The output is measured in milliseconds. Users get feedback in real time. There is no need for update cycles or patches; WebPulse is always up to date.

This section provided a high-level look at Intelligence Services and WebPulse. Next we will discuss WebPulse technical details.

Note: It is important to understand the principle behind Intelligence Services. In many cases, web-based attacks start by injecting scripts into trusted web pages. These scripts typically generate a dynamic link to a malware host over a dynamic and powerful malware delivery network. The primary goal of Symantec Intelligence Services is to analyze and block links to the malware itself. Users should never be prevented from viewing a trusted web page. The script itself does not harm the endpoint. This is fundamentally different from the approach of many vendors in the web security space, and raises these questions: Do they know where the actual malware is? If yes, why are they blocking the innocent page that hosts the link? If not, why not, since they have found the link?

How It Works

Requests to URLs are first checked against the local Intelligence Services database on the proxy, or the local categorization cache on other Symantec products. If the URL can be categorized locally, the category information can be used to allow or block the request.



Typically, the percentage of locally uncategorized content is about 5%. If the URL is not in the local database or categorizations cache, the URL is dynamically sent to WebPulse. In the cloud, the URL will first be checked against the central master database. This is comparable to the local lookup; if the URL is in the master database, the URL category will be sent back to the requesting WebPulse client and can be used to allow or block the request. The new result is automatically cached locally.

If the URL is also not in the central master database in the cloud, Dynamic Real-Time Rating (DRTR) will be used to analyze and categorize it in real time if possible. Any resulting URL category will be sent back to the requesting WebPulse client and can be used to allow or block the request.

Independent of the real-time categorization result, the URLs will be sent to several background processes in parallel. Some of the background processes are focused on providing new content categorizations for the database. Others are focused on hunting for evidence of malware activity. DRTR is primarily a content categorizer, but it is also used to log a large amount of metadata about each URL it analyzes, and it is this metadata that feeds many WebPulse background processes. Many URLs are not web pages and are not suitable for DRTR categorizing, but WebPulse still gathers as much information about the URL as possible to feed the background processes.

WebPulse uses several methods, including exploit detection techniques, to analyze scripts and detect malicious payloads and referenced domains. When a user accesses a binary file through a URL that WebPulse has not seen before, WebPulse will also download that file and run it through a bank of up to ten different AV scanners with full heuristics, script analyzers (for example, malicious java scripts with heap sprays), exploit detection modules, and other malware- detection mechanisms. New threats are identified within minutes and automatically added to the master URL database to protect other customers. This is one way in which WebPulse cloud users benefit from the network effect of working together to provide broad real-time protection and receive a strong zero-day response to new web threats—when only a few antivirus vendors have even been able to detect them.

In addition to Symantec analysis, several third-party URL feeds covering malware and phishing sites are considered for inclusion in the master database. Further, when used with Symantec Content Analysis, the Edge Gateway can send any URLs that Content Analysis identifies as malware sources to the WebPulse service for verification. It's important to know that malware feeds are quality-checked before being integrated into the Intelligence Services database, preventing false positives.

For security-related categories, incremental database updates are pushed to the proxy every five minutes. This enables the local defense to maintain performance by responding to as many requests as possible from the local databases. If any URLs discovered cannot be matched against an entry in the local database, the proxy will check with WebPulse.

Recommended Features for Malware Protection

Symantec Secure Web Gateway solutions have a broad feature set. The following section provides a brief overview of useful—and recommended—features for malware protection.

URL Filtering

This is the first point at which requests to known malware sources can be blocked. For URLs that are not known or not included in the local database, the Cloud or Edge SWG connects to WebPulse collaborative defense. Uncategorized URLs are then analyzed in real time.

Authentication

The most secure way to authenticate users is to authenticate each new session. If the desktop is infected with malware but is not authenticated, it cannot communicate with systems on the Internet, blocking any potential loss of confidential and private data.

Controlling Data Types

If users have no right to install software on their desktops, why should they be able to download executable files from the Internet? Blocking executable files is another step in protecting against malware. Often malware tries to download software to add malicious content on the infected desktop.

Another reason for blocking executable files is that malicious dynamic links could point to an executable malware file that would be installed on the desktop. Blocking executable files prevents this threat.

File-type blocking can be done based on true file-type detection. Symantec best practice recommends blocking executable files in general for regular Internet users. If this is not acceptable, they should at least be blocked for sites that are uncategorized and/or have a high Threat Risk Level. Threat Risk Level is an extremely useful tool for fine-tuning and customizing web security when file type and categorization alone cannot meet your business needs. For more information, read the Symantec Threat Risk Levels white paper.

Protocol Compliance

Symantec Edge and Cloud Secure Web Gateways use application proxies for several protocols. Because there are two connections—one between client and proxy and one between proxy and server—threats such as buffer overflow attacks on the protocol level can be filtered out. The proxy changes protocol behavior (from server to proxy) to RFC-conforming behavior (from proxy to client).

SSL Interception
SSL-encrypted traffic tunnels require a secure web gateway solution. Terminating SSL at the proxy enables detection of malicious content and tunneled applications. Certificate management can be used to verify X.509 certificates and allow only trusted client or server certificates. Non-SSL traffic attempting to exit via port 443—which may be an indication of a malware infection—can also be blocked by the proxy.

Advanced Threat Protection and Malware Scanning
The last step in malware protection is inline malware scanning. Both inbound and outbound data can be malware-scanned by extending the content inspection capabilities of the Cloud and Edge SWG, powered by Symantec Content Analysis engines. By default, all traffic, including large files, are scanned by Content Analysis.

Content Analysis—included in Symantec Web Protection Suite and Symantec Network Protection—offer advanced threat protection using layers of inspection including, malware signature scanning using two antimalware engines, static-code and behavioral analysis, block/allow listing, machine learning, and the ability to either provide on-box sandboxing or cloud sandboxing, or to broker to an external sandbox solution. The advanced threat protection is a valuable differentiator from most other secure web gateway solutions, many of which use traditional threat protection and/or a selective scanning approach.

Log File Analysis/Reporting
Checking access log files on a regular basis is recommended. This means checking often enough to recognize normal traffic, so that new, unusual, or abnormal traffic can be spotted and investigated. Symantec Reporter is a superb tool for analyzing access log files.

WebPulse Technical Overview

Classification Accuracy
Accuracy refers to the ability of a filtering product to categorize URLs correctly with minimal false positives. To state it another way, accuracy level answers the question, "Of the 100 URLs the filter categorized as X (Pornography, Spyware and Gambling, for example), what percentage were actually X?" The higher the percentage, the greater the filter's accuracy. False negatives provide another accuracy indicator. The question in this case would be, "How many of type X did you miss?" Symantec technology delivers the most accurate categorization of any web security vendor. WebPulse is able to categorize URLs based on multiple levels:

- Domain: all hosts of "symantec.com" could have the same category
- Host: "host1.symantec.com" and "host2.symantec.com" could have different categories
- Directory: "host1.symantec.com/directory1" and "host1.symantec.com/directory2" could have different categories
- File name: "host1.symantec.com/directory1/good_file.jpg" and "host1.symantec.com/directory1/malicious_file.jpg" could have different categories
- Query string: "www.facebook .com/?sk=inbox" and "www.facebook.com/?sk=ff" could have different categories
- IP address: For performance reasons (to prevent reverse DNS lookups) it is possible to add IP address-based categorization to the database
- Protocol and header analysis is an additional categorization option

Note: Usually, content sent to WebPulse by the Edge proxy is typically content that could not be categorized by the local database. However, if necessary, URLs categorized as "web hosting" will also be sent to WebPulse for real-time analysis to apply a more accurate categorization.

Multiple Categories per URL

Web pages do not always fit easily into a single category. One example, "www.facebook.com/?sk=inbox", which is both a social networking site and an email application within Facebook. An accurate web filter recognizes this and classifies the site into both of these categories, giving enterprises the flexibility to control which parts of any site can be accessed by their users. Intelligence Services can provide up to four categories per web page, which reflects web page content much more accurately and makes possible thousands of granular subcategory combinations for flexible and powerful policy enforcement.

Preventing Users from Bypassing the Content Filter Policy
To achieve high accuracy, WebPulse is able to prevent users from bypassing the content filter policy by accurately analyzing and classifying tools such as the following:

- Translation sites that provide online translation of languages
- Archive sites that cache selectable content from the past

- Image searches that are delivered by a search engine
- Proxy anonymizers that relay requests via intermediary sites that are often obscure

Early-generation filtering technology often provides only superficial categorizations (examples: translation site, image search, or archive site) but this is not helpful for implementing a policy. Customers do not want to block all image searches or all translation and archive requests. In contrast, Symantec solutions are able to see the destination web page embedded in the intermediary page to make an accurate and useful categorization. For example, Intelligence Services accurately categorize an archive of "cnn.com" as News.

Note: On policy-enforcing systems such as Symantec Cloud and Edge Secure Web Gateways, a search engine safe-search policy can be enforced. This also helps prevent users from bypassing the content filter policy.

Quality Checks

The WebPulse infrastructure is supported by a set of stringent quality checks designed to reduce false positives and over blocking. All categorization changes and malware identifications must pass Symantec proprietary quality checks before they are released to the customer base.

Performance

When talking about WebPulse, it's important to mention performance. Intelligence Services and WebPulse provide a highly scalable, high-performance solution. Only a small percentage of the overall web traffic has to be analyzed by WebPulse in real-time.

Web filtering is optimized to run on-proxy (on-box). Categorization requests are processed in RAM, usually an order of magnitude faster than when they are run off-box. Intelligence Services typically categorizes around 95% of the web pages requested by a corporate or educational user on-proxy in less than eight milliseconds (ms). For the other 5%, a categorization can be instantly and transparently requested from the WebPulse master database (typically in less than 70 ms) or from the WebPulse Dynamic Real-Time Rating (typically in about 200 ms, although there are some dependencies on the performance of the site in question). Processing categorization requests on-proxy is the fastest possible architecture for high performance and scalability. That's why Symantec provides incremental database updates every five minutes for security-related categories and every six hours for nonsecurity-related categories.

The on-box database also includes IP addresses for the most common websites so that DNS reverse lookups don't slow down the processing of URLs.

Dynamic Real-Time Rating and Dynamic Link Analysis

Over 300 WebPulse libraries are available to rate and categorize new content in real-time. Real-time categorization supports over sixty languages, including "Pornovian," a generic module that detects pornography-related content. This and various threat-detection features are key components of WebPulse. Together they present another unique differentiator.

Real-time threat detection includes dynamic link analysis, which is used when cyber criminals place a script on a trusted web page that forces the browser to download malicious content from a typically uncategorized and quickly changing malware host. The offending URL will be sent to WebPulse in real-time.

Real-time categorization disassembles a web page and analyzes its components. Here is an extract of the kinds of information that is used to assign categories:

- Language (example: English)
- Source code language (example: JavaScript)
- Document type (example: HTML)
- Character set (example: UTF-8)
- External link categories
- Content words
- Scripts
- Iframes

Real-time Malware Detection Modules

Most of the real-time malware detection modules look for characteristics of the content (data or traffic) that may indicate danger. At the same time, they also assess the source for indications of danger—using more than thirteen years of experience in mapping the shady parts of the Internet. If the combination of characteristics is sufficiently suspicious, they trigger. The modules ask, "How does the bad content differ from legitimate content? How are they serving their content? Where are they serving it from?" Access to suspicious content, which triggers a response from the real-time malware detection modules, can be blocked immediately.

URL Background Checker

The background checker system has two modules: a foreground (real-time) module and a background (off-line research) module that checks the background of a URL or site. The research module gathers data on malware delivery networks (MDN) so the real-time module can ask, "Does this URL belong to one of those networks?" Access to URLs pointing to MDNs can be blocked immediately.

Background Analysis Techniques

Not all analysis can be done in real-time. The boundary between real-time and background categorization is crossed when there isn't enough information for a real-time decision, or when the content is not applicable to real-time categorization. For the small volume of content that cannot be categorized in real-time— typically less than 2%—a background analysis service uses sophisticated, proprietary techniques and feeds the analysis back into the master WebPulse categorizations database. As a final step, human raters continuously train the DRTR and the background systems, and investigate and categorize rare sites.

The background systems also process URLs, categorized in real time, to decide if a categorization should be added to the database. This indicates that not all the URLs that have been categorized in real time will be added to the database. One criterion for adding a URL to the database is its number of requests per unit time.

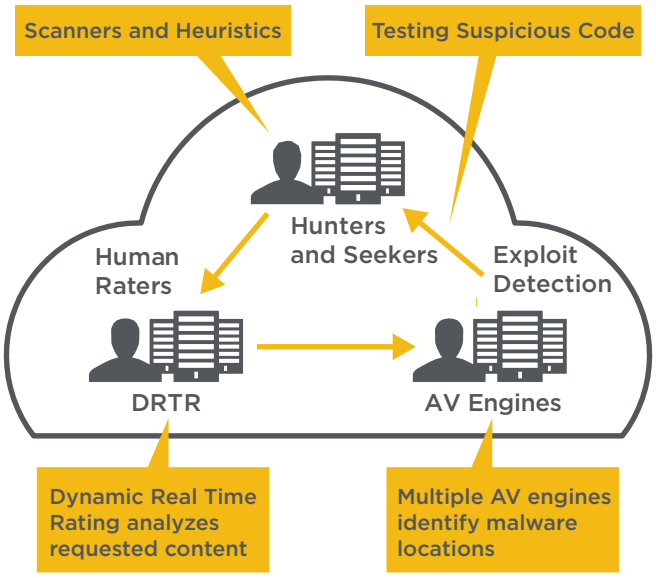
The background systems are looking for additional evidence to supplement what was collected in real-time. Once identified, this information can be used to fine-tune and update real-time categorization modules. As an example, consider an HTTP referrer header: With knowledge about the referrer, WebPulse can analyze the full path of a web-based attack. In many cases, malware hosts present their malicious content only when the requests contain a certain referrer, such as a search engine result.

Malware Detection

Given the rapidly evolving threat landscape, effective malware protection requires a broad set of detection mechanisms. WebPulse technology has a unique approach to protection against malware in Internet traffic, combining the following analysis and identification techniques.

Detection and Analysis of Malicious Traffic

To identify malware distribution mechanisms including intermediaries and malware hosts, WebPulse includes multiple rule engines to flexibly deal with different kinds of traffic and sites in a constantly changing threat landscape. These rules are constantly fine-tuned, expanded, and updated to reflect real-time information identified by Symantec malware experts. They eliminate the need for you to spend time trying to become an expert in web defense.



Malicious Site and Content Identification and Analysis

It's important to constantly evaluate risks associated with all sites that users access. Malware that has been embedded in reputable sites is identified, even if it has been obfuscated. Symantec conducts this analysis in multiple ways:

Malicious Site Fingerprinting

To match the speed with which malicious sites change their domains, WebPulse utilizes advanced fingerprinting modules that quickly recognize similar sites that appear on new servers.

Web Reputation

WebPulse collects many kinds of reputation information about sites. It automatically scores the reputation of websites and categorizes sites with a heightened security risk as Suspicious.

Malicious JavaScript Detection

WebPulse logs information on JavaScript from the millions of web pages that are requested every day. Symantec researchers use this intelligence to identify characteristics that indicate suspicious behavior and create appropriate new defenses for them.

Malware Content Analyzers

WebPulse has proprietary analyzers that identify malicious sites in real time, using statistical analysis techniques to locate suspicious content on web pages.

Phishing Detection

WebPulse includes proprietary real-time algorithms that identify phishing sites posing as financial institutions. These algorithms, coupled with the real-time WebPulse processing of customer requests to uncategorized sites, are able to find new phishing threats almost immediately. The unique Symantec mechanism often detects sites before they appear on any third-party phishing lists.

Integration with Content Analysis and Malware Analysis

Symantec advanced threat detection engines, Content Analysis and Malware Analysis, provide an added layer of protection for the Cloud and Edge Secure Web Gateway solutions with the use of machine learning, behavioral analysis, block/allow listing, dual antimalware scanning, and sandboxing. In addition to providing this added layer of protection, they also integrate and share detection of new threats directly with the Symantec Global Intelligence Network, providing immediate categorization of new risks as they are found to the WebPulse database.

Detection of Illegal or Questionable Sites (Scam Sites)

The background checker (and some of the other modules) can be used to target any large, complex network of websites, and many scam networks fit this description. Several of the large malware delivery networks also contain subnetworks that deal in this sort of material; the background checker blocks this content in real time (generally with Suspicious as the initial categorization).

Third-party Intelligence

Third-party intelligence is used to complement primary Symantec research and analysis. In addition to all the techniques described above, WebPulse gathers information from numerous third-party intelligence sources. These include commercial malware and phishing lists, community-contributed content, and general research and ongoing monitoring of the overall web security threat landscape. Information from third-party sources must meet a very high standard of quality and pass a set of rigorous checks before the source is accepted for inclusion in the WebPulse system.

Active User Community

The worldwide WebPulse community comprises users from more than 15,000 diverse enterprise customers who are extremely active in ensuring the accuracy and effectiveness of the WebPulse service. They send billions of new web requests to WebPulse every week, giving the service a clear, current view of the huge numbers of new and changed web areas and—hiding inside those huge numbers—the locations of the most likely sources of web-borne threats and the hidden paths that lead to them. Symantec is committed to investigating and responding to community requests and feedback usually within 24 hours. The WebPulse community provides Symantec with a significant sample of all traffic traversing the Internet on a second-by-second basis. WebPulse uses its malicious content detection techniques (described above) to analyze traffic, sites, and content, and deliver an extremely high rate of malware identification. The usage patterns of the WebPulse community and its real-time analysis provide invaluable insights to users, and the highest possible level of web-threat protection for Symantec customers.

Immediate Availability of Malicious Content Identification to WebPulse Users

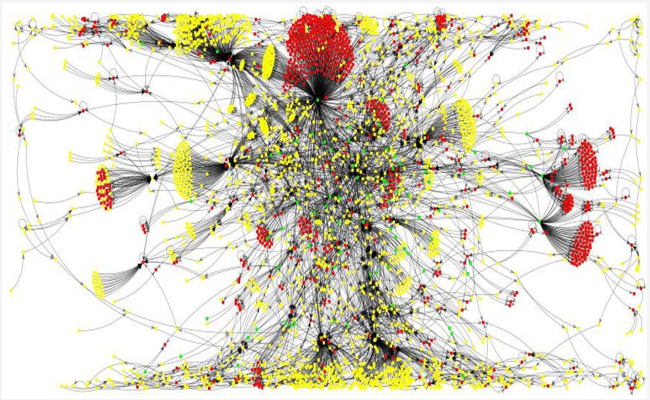
WebPulse ensures that you have what you need to protect yourself as soon as you need it. As soon as a categorization, rating change, or malware identification passes the WebPulse quality checks, all WebPulse users have access to the information. WebPulse clients accessing WebPulse can query the cloud service in real time to receive the new categorization or rating.

Malware Delivery Networks Analysis

Cyber criminals run vast networks of sites and servers to collect victims, relay them to a designated location, infect or entice them, collect payments, serve new or upgraded payloads, or perform other operations. Like any mainstream web architecture, there is provision for redundancy, failover, backup and administration. But these malware delivery networks have an Achilles' heel: their size and complexity presents WebPulse with a large attack surface.

Symantec technology's long experience in the cloud, and its high volume of web traffic, to develop systems that identify and track MDNs. Every day, WebPulse

Graphical mapping software makes it easy to see how the large malware delivery network in the center of the image above pulls unsuspecting users into the attack.



automatically identifies thousands of new sites and servers as members of known MDNs, protecting our users from whatever new exploits and payloads the MDNs may be offering, no matter how well they may be hidden or encrypted.

Web Application and Web Application Operation Controls

In addition to URLs and IP addresses, the Intelligence Services database contains information about web applications:

- Application name (example: Facebook)
- Application operations (example: Post Message)
- Application category (example: Social Networking)

This information can be used to build very granular policies to control web application use. New applications and application operations can be implemented and made available to Cloud and Edge SWG customers without the need for updates. Because this information is part of the database, all changes are available on Edge Proxies for both Content Policy Language (CPL) and Visual Policy Manager (VPM), and for the Cloud Proxy as soon as an automated database update is being installed. If needed, additional application visibility and control is available with Symantec CASB, with the ability to recognize over 35,000 applications.

Both Symantec Edge and Cloud SWG Reporting show new applications and application operations as soon as they're available.

Managing Web Application and Web Application Operation Changes

A critical part of web application control is the early detection of application changes so adjustments can be made promptly. Symantec technology has implemented Q&A processes for all supported applications and application operations. Applications and operations are monitored; when a change is detected, action is taken immediately, rolling out changes using standard database updates.

Conclusion

Symantec products constantly evolve web security solutions to prevent and combat fast-changing web threats. These solutions offer powerful advantages to customers because they take advantage of the following:

- **The cloud:** WebPulse, as a cloud-service component of Intelligence Services, has been in continual development for the past thirteen years—longer than any other cloud security solution. It enables us to constantly enhance and upgrade its capabilities with no impact to customers—no downloads or patches for them to manage. This gives Symantec products the greatest agility of any vendor in dealing with changing threats.
- **The community:** With input from the industry's largest user base — WebPulse has the greatest possible understanding of what users are encountering on the Internet right now. This provides valuable direction and focus for our security research efforts. We have never needed to use web crawlers to search for content.
- **Antimalware detection technology:** Using an easily adjustable bank of cloud-based antimalware technologies allows us to add new advanced threat scanners that have shown recent accuracy improvements or unique capabilities against specific threats. It also allows us to disable scanners that are drifting toward unacceptable levels of false-positive detections. We continuously monitor and modify the configuration of individual systems (and the bank as a whole) for optimum performance and accuracy.
- **Security industry relationships:** As new research organizations appear and others focus on areas of specialization, it's important to adjust our relationships with them to ensure fast information-sharing, and to evolve collaborative processes for addressing new threats.
- **Ongoing expansion plans:** Symantec Security Labs continues to invest in people, equipment, and relationships to build and strengthen our internal expertise.
- **Security expertise:** Symantec products have deep roots in blending machine learning technology with human researchers in joint feedback loops. This is the only solution that meets the challenge of reliably managing the huge volume of web traffic every day.

WebPulse-WP100 November 30, 2021