LOOKINGGLASS

MACHINE READABLE THREAT INTELLIGENCE

COLLECTING INTERNET INTELLIGENCE FOR CYBER THREAT ANALYSIS

CONTENT

INTRODUCTION: ON MARGINLESS CIRCLES

In 1879, an obscure school teacher and bank clerk named James Murray accepted the editorship of a revolutionary new book then entitled, A New English Dictionary on Historical Principles, Founded Mainly on Materials Collected by the Philological Society. The book's purpose was to provide the first, comprehensive description of every word in the English language. As the book's third principal editor in its 22 years under development, Murray was charged with managing a complex editorial process that involved collecting a massive quantity of submitted words from outside contributors and then devising a system to convert every word into a formal definition. Murray would spend the next 36 years working on the project, until his death in 1915. The book – which took 64 years, six editors, and 47 editions to complete – was finally published in 1921 as The Oxford English Dictionary (OED).

- The editorial challenges to developing the OED will sound immediately familiar to today's security analysts. One need only to replace "word" with "threat" and "language" with "Internet" in these challenges that Murray and his team faced:
- How to ensure every word [threat] in the language [Internet] is identified.
- How to standardize the format and presentation of words [threats] so that readers can easily understand what each word [threat] in the language [Internet] means.
- How to ensure every word [threat] is documented, without duplication, in its complete and proper context, including different senses (e.g., "free" is a noun, verb, and an adjective) and the evolution of each word's [threat's] usage over time (e.g., contemporary versus obsolete).

In 1880, one year into his editorship, Murray described the task by writing,

"The English language is not a square with definite sides containing its area; it is a circle, but a circle such as Euclid never contemplated […] clear and definite and unmistakable at the centre, but whose circumference melts marginless into the surrounding nothingness, and seems to have a margin, only till we try to trace it."

Just as Murray was trying to provide a complete, accurate, and updated description of every word in the English language, so are today's security analysts trying to provide a complete, accurate, and updated description of every threat on the Internet. As with the OED's editorial process, today's security analysts face major challenges in efficiently gathering, organizing, and analyzing data so that new and evolving threats can be identified, classified, prioritized, and acted on.

Murray later characterized the strenuous nature of his career as, "The work of a machine & not of human beings struggling with some of the most difficult problems of human history." [1] Murray never had the option of employing machines in his monumental task, but today's security analysts do.

METRIC	VALUE	NOTES	SOURCE
Worldwide Internet Users	3.3 billion		Internet Live Stats
Websites	1.035 billion		Internet Live Stats
Emails Sent so far in 2016	34.4 trillion	As of June 2, 2016	Internet Live Stats
Internet Traffic so far in 2016	470 trillion gigabytes	As of June 2, 2016	Internet Live Stats
Skype Calls so far in 2016	28.4 billion	As of June 2, 2016	Internet Live Stats
Websites Hacked so far in 2016	9.869 billion	As of June 2, 2016	Internet Live Stats
2015 Worldwide Social Network Users	2.04 billion		Statista
Projected Worldwide Social Network Users in 2016	2.22 billion		Statista

The Digital Library of Babel

MACHINE READABLE THREAT INTELLIGENCE

Machine Readable Threat Intelligence (MRTI) provides organizations with data that are collected by automated means from the Internet. Raw data form the basis for threat intelligence, which reveals active and potential cyber threats on internal enterprise networks and the external Internet. MRTI helps to solve the following threat intel challenges:

- Gathering enough data to ensure all major threats are identified
- Properly classifying threats
- Automatically sifting out useless and/or duplicate data, without human effort
- Providing data on all current and historical variants of threats
- Formatting the data so that machines can act, and humans can easily contextualize, correlate, and prioritize

MRTI's raw data are gathered from a network of deployed Internet sensors and surface/deep web and darknet link crawlers. The sensors and crawlers identify threats active in the wild, such as:

Malicious Internet Protocol (IP) addresses and domain names
Links to legitimate websites that are compromised (e.g., distributing malware)
Command and control (C2) infrastructure (e.g., botnets)
Active malware campaigns and variants

Almost any threat intelligence data set can be turned into an MRTI feed given the appropriate resources. Intel from the data feeds is usually converted into a standardized format, such as Structured Threat Information Expression (STIX) and then

distributed via a means such as Trusted Automated Exchange of Indicator Information (TAXII). Such methods make it easier to share information between security products and among teams.

MRTI is intended to fit within an organization's enterprise security ecosystem. MRTI solutions offer varying degrees of automation and integration with other security products. Baseline MRTI solutions enable intel to be incorporated into Threat Intelligence Platforms (TIP) for human analysis and monitoring or into a Security Information and Event Management (SIEM) solution to help correlate external threat activity with an organization's internal systems. Some MRTI solutions allow the intel to be used to automatically update firewall rules, domain name servers (DNS), web content filter whitelists/blacklists, and other defenses. The degree of automation and integration available in MRTI technology differs, but there are three primary approaches:

- Application programming interface (API)-based integration (e.g., JSON, XML, CSV, etc.)
- API + libraries (e.g., SDK, bundled scripts, etc.)
- Platform-based integration (e.g., Splunk, ArcSight, Maltego, etc.)

By efficiently collecting the data to connect dots between internal operations and outside activity, organizations gain the context to prioritize and act on threat intel.

Sidebar: Rings of the Digital Marginless Circle

The raw data that can become actionable threat intelligence are scattered across the Internet and reside in one of several "rings," to borrow Murray's circle metaphor for the English language. The Internet's rings are the surface web, the deep web, and the darknet.

Much like the "center" of the human language, which for Murray was "clear and definite," the surface web is what most people think of as the World Wide Web. The surface web contains information that is easily findable by common search engines (e.g., Google, Bing, etc.). Surface web information is characterized as being highly visible, freely accessible, and reliably available over time.

The next ring of the Internet, like the English language, is less visible and accessible. This ring is the so-called "deep web," a name popularized in the 2001 white paper entitled, "The Deep Web: Surfacing Hidden Value." [2] The paper's author, Michael K. Bergman, used the term "deep web" to refer to parts of the World Wide Web that are not searchable with standard search engines and that can be less accessible than the surface web. Deep web examples include information in databases, behind paywalls, or otherwise restricted (e.g., by account log-in). In his paper, Bergman estimated that the deep web was "400 to 550 times larger than the commonly defined World Wide Web." No one knows how large the deep web is today, but many assume it has grown exponentially alongside the surface web and that the deep web is still proportionally larger.

In the wake of recent news stories, many people have become familiar with a third ring, the so-called "darknet." The darknet is where the Internet, much like the English language, begins to appear marginless, opaque, and untraceable. The darknet is characterized by its difficult navigability and the anonymity it affords users. Information on the darknet is often hosted on sites with obscure domain names, which do not appear in search engine results and which cannot be accessed without special technologies (e.g., the Tor browser). Many darknet sites are private, invitation-only communities. The most notorious example was the Silk Road, an online marketplace for drugs, stolen identities, malware, and other illegal services and controversial products. The darknet is a popular hangout for cyber criminals and is therefore chock full of valuable threat intelligence.

THE LOOKINGGLASS APPROACH: MRTI PORTFOLIO

LookingGlass's MRTI data can be fed directly into security analytics systems and mitigation products to enable timely response to ongoing threats.

LookingGlass's portfolio of MRTI includes the following:

Cyveillance Infection Records Data Feed (Sourced from Virus Tracker)

What it does: Based on the global botnet monitoring system Virus Tracker (virustracker.net), creates a list of newly identified and historical infections originating on any network in the world without requiring a presence on those networks. Virus Tracker is the world's largest botnet monitoring system. A botnet is a network of devices (i.e., robots or bots) that have been compromised and can be controlled remotely (zombies). Virus Tracker identifies three million newly infected zombies daily and maintains a searchable library of two billion historical infections dating back to 2012. Virus Tracker provides a global, real-time map of botnets.

Why it matters: To succeed, threat actors need space in which to operate. This space may be networks, servers, and domains they create and control, or the compromised networks, servers, and domains of unknowing individuals and organizations. Botnets are an effective means to this end and enable many of the Internet's most infamous threats and inconveniences, including:

- Spreading malware
- Sending spam emails
- Carrying out distributed denial of service (DDos) attacks, which use hundreds, thousands, or even millions of zombies to flood a server with requests that overwhelm the server, thereby making the resource(s) it hosts (e.g., website) unavailable to users

Threat actors also must avoid detection – by people and security products. The success of their campaigns is directly tied to their ability to remain anonymous and covert. That's why many threats are designed to be polymorphous, meaning they take different forms that evolve over time to become undetectable as known attack signatures. The most sophisticated threat actors today operate similarly to traditional IT shops, always enhancing and improving their malicious products in a remarkably professional way.

Every tool, technique, and procedure (TTP) used by threat actors leaves a trace – even if only brief and ephemeral. Detecting these traces can reveal how a threat is growing, spreading, and evolving. Such information is invaluable to contextualizing threats, calculating the risk to organizations, and determining prudent actions for protection.

As will be detailed below, threat actors have become savvy at developing attacks that bypass traditional network security measures. In some cases, the only effective method to prevent initial compromise is to block users from viewing, accessing, or downloading certain things.

The Infection Records Data Feed delivers positive confirmation that a host on an organization's network is infected because it has contacted a sink-holed LookingGlass Virus Tracker Command and Control (C2) domain or IP address requesting further tasking information. Over 65,000 new malicious domains are identified every week.

Cyveillance Malicious Command and Control (C2) Data Feed (Sourced from Virus Tracker)

What it does: Reverse engineers malware and extracts the long-lived and ephemeral domains of C2 servers that control malware. Delivers the ability to block embedded malware from reaching back to the C2 servers that are exfiltrating data or coordinating the attacks.

Why it matters: Sophisticated threat actors are always evolving their TTPs for attack. C2 servers play various important roles in today's exploits.

Some attacks employ a multiphase method of compromise. The first step in these attacks often use common exploit kits, such as Angler or Neutrino, or custom-designed "droppers." These tools do not include the full malicious payload, but they allow the attacker to gain information about a newly compromised system or device (zombie) before delivering a full version of the malware. For instance, some droppers detect whether the infected operating system is running in a native or virtualized environment. The latter could indicate that the dropper has been downloaded by a security researcher or threat analyst. If a virtualized environment is discovered, some droppers will instruct the C2 server to not deliver the full payload at the risk of being detected. Other droppers look for specific versions of operating systems or applications, which contain the vulnerability necessary to compromise the device. If patched or otherwise immune systems are detected, then the full malicious payload hosted on the C2 server may never be delivered to that particular zombie.

The use of exploit kits and droppers enables attackers to widely distribute only a small portion of the malicious code, while retaining control of the full payload on their own C2 servers. This decreases the chance of detection. For instance, in-house IT teams who find a dropper on their web servers may not fully understand what the program is unless they reverse engineer the code, which is not a task generalist IT staff usually have the time or skill to do.

C2 servers are also used to remotely issue commands to zombies. Modern malware has become modularized, allowing attackers to be selective in the functionality they use to achieve their goals. C2 servers also act as temporary storage for stolen information (e.g., intellectual property, passwords, credit card information, etc.).

The C2 servers hosting malware and issuing commands to zombies are often associated with multiple domain names, even while the IP addresses remain the same. Threat actors are able to change domain names regularly via the use the domain generation algorithms, which automatically change domain names randomly. This helps them to avoid detection by domain name. Identifying all domain names associated with an IP address helps to track and block their evolving attacks.

Identifying and blocking malicious C2 servers by domain and IP address renders ineffective exploit kits, droppers, and the full versions of malware hosted on those servers.

Cyveillance Malicious URL Data Feed

What it does: Provides a continuous feed of infected web pages on the Internet. Delivers the ability to block access to pages by internal assets, significantly reducing the chance of infection from drive-by downloads.

Why it matters: Drive-by downloads have proven popular with attackers because they are stealth and are not blocked by many traditional network security tools.

The most common type of drive-by download is enabled by malicious code that is covertly embedded in a web page's legitimate code. The malicious code loads with the web page and compromises web browsers without any action by (or, usually, knowledge of) the user. The most common method uses malicious JavaScript inside tiny HTML IFrames, but drive-by downloads can be delivered using any client/server pair (e.g., email, instant messaging, etc.).

Notably, many legitimate websites contain third-party content. For instance, a major news site contains its content, along with the content of affiliate sites, wire services, and advertisers. This third-party content may or may not be malicious. Only by understanding the full path of malicious URLs can access to the content that is bad be blocked, while still enabling access to the pages that are good and not impacting the user experience.

Drive-by downloads are the most common method of distributing exploit kits and droppers in the first step of multiphase attacks because drive-by downloads can bypass traditional security in various ways. For instance, many traditional network security tools do not inspect the content of hypertext transfer protocol (HTTP) traffic (i.e., Transmission Control Protocol/IP [TCP/IP] Layer 7) and are therefore incapable of blocking client-side attacks, in which the vulnerability exists in the code of the client operating system or a client-based application (e.g., web browser). In these scenarios, the compromise does not occur until the client processes all of the code – benign and malicious. Usually such attacks can be blocked only if the entire site is

blacklisted by IP address (e.g., TCP/IP Layer 3) and never allowed to load on a device.

The most effective ways to prevent web-based drive-by downloads are to disable dynamic elements in web pages (e.g., JavaScript), which can seriously impair functionality, and to block the IP address of websites known to host the malicious code. Additional protections against client-side attacks include keeping operating systems and applications patched and hardened (e.g., antivirus).

Cyveillance Phishing URL Data Feed

What it does: Provides a continuous feed of locations used in thousands of phishing attacks across hundreds of companies, websites, and industries each day. Delivers the ability to block access to phishing sites by internal assets, significantly reducing the chance of stolen credentials and other phishing-related activities.

Why it matters: Phishing is a form of social engineering that tricks users into thinking someone they trust has requested them to take an action or provide information. Email is the most common phishing attack vector, and threat actors have enjoyed widespread success with this method of attack. Phishing campaigns have become so sophisticated as to be virtually undetectable on the surface, even by trained security professionals.

The first category of phishing attack asks users to do something, such as download an attachment or follow a link embedded in the email. Usually the attachment is malicious (e.g., contains malware), and the link leads to an infected website (e.g., where a drive-by download could deliver a dropper).

The second category of phishing attack asks users to provide sensitive, confidential, or proprietary information. Sometimes phishers ask users to reply directly to the email with the information (e.g., posing as a colleague), and other times the attackers direct users to a fake website that is practically indistinguishable (visually) from the legitimate site (e.g., password reset page for a bank, social media, etc.).

Attackers have devised phishing schemes to successfully compromise a range of victims, from high-level government officials and business executives to IT professionals and consumers. As with websites hosting drive-by downloads, an effective way to deal with many phishing campaigns is to block user access to sites at the IP level.

The Phishing URL Data Feed can be used by email security appliances to block emails that contain links to known malicious phishing sites or web security gateways. In fact, both the Phishing and the Malicious URL feeds are currently used by the most widely deployed commercial email and web security appliances.

Cyveillance Newly Registered Domain Data Feed

What it does: Gathers the authoritative "master list" of registered domain names for the more than 100,000 new domain names that are registered every day. Provides the ability to identify and block access to new websites, which may be used as part of new attack campaigns, until the reputability of the domain can be established.

Why it matters: As detailed in the previous descriptions, attackers have numerous uses for malicious domains. Many domains used in active campaigns "go live" days, hours, or minutes before a campaign actively begins in order to avoid being blocked by existing security rule sets. By identifying newly registered domains, security controls can block access to any domain that, for instance, is less than seven days old, which is typically a strong indicator of a malicious domain that has recently gone live.

In addition, domains created via domain generation algorithms are automatically registered and will show up as a newly registered domain even if there is no other available threat intelligence about the domain (due to the absence of a history). For example, the domain may not be classified as a malicious URL or a phishing URL, but it will be flagged as new. Many organizations block access to new domains until the domains have had some time to become reputable.

Many traditional security safeguards are ineffective against attacks delivered via malicious and compromised websites. The

most effective method to protect users, assets, and data is to block these sites at the domain level, never allowing users to view or load the sites on their devices.

Cyveillance Compromised Information Monitoring Service

What it does: Allows clients in the identity monitoring and protection industries to enroll end users to be notified of leaked personally identifiable information (PII) discovered on the web (i.e., surface, deep, and darknet).

The automated monitoring solution covers the following types of PII:

- Credit card numbers
- Social security numbers
- Compromised account credentials (e.g., username/email + password combinations)
- The solution supports the following uses cases:
- Early Warnings: Alert end users to PII leaks
- Security Alerts: Provide information specific to a user's compromised account
- Password Resets: Prompt resets for detected compromised account(s)
- Augment ID Theft Solutions: Add another dataset to identity theft monitoring
- For example, a client can enroll 100,000 end users for monitoring of any (or all) of the three types of PII. Once enrolled, the client is notified if PII for enrollees is discovered.

Why it matters: Most of the business and financial damage from data breaches does not result from the breach itself. The most serious damage occurs when the attackers either destroy data (e.g., Sony Pictures, Saudi Aramco, etc.) or when the pilfered data are posted and/or sold online.

If the data stolen contain PII, then consumers face the brunt of the consequences, such as:

- Identity theft
- Unauthorized purchases in consumers' names
- Unauthorized lines of credit opened in consumers' names
- Exposure to blackmail or extortion schemes

In many cases, the impact is not only the initial out-of-pocket financial cost, but additional financial costs associated with restoring victims' reputations and reversing fraudulent transactions. The sooner such data breaches are discovered, the quicker consumers can act to proactively minimize or eliminate the risks.

Cyveillance Malware Total Life Cycle Protection Data Bundle

What it does: Provides pre-infection protection from malicious attacks and prevents malware from communicating and coordinating with C2 servers.

Why it matters: Security professionals often speak of "defense in depth." The idea is to provide multiple layers of defense in the event one of those layers fails or is compromised. This increases the time and effort required by threat actors to succeed, as well as the likelihood that attacks are detected.

The same principle can be applied to threat intelligence. Relying on only one source or type of information may lead to missing other valuable information. The more useful, relevant, actionable intel that is available, the better the decisions.

While none of the data feeds replace the need for traditional enterprise security, they supplement these protections by providing valuable information about how to configure security products to protect against new threats and emerging risks. The Malware Total Life Cycle Protection bundle detects and prevents attacks in a single bundle of domain/URL-based data feeds.

CONCLUSION: TRACING THE DIGITAL MARGINLESS CIRCLE

Today's IT professionals charged with security face countless challenges. One of the biggest is gaining a complete picture and clear understanding of the threat landscape. The Internet is so large, the threats are so numerous, and the volume of data is so high, yet the key information needed to make informed, timely decisions is out there. The challenge is to efficiently acquire, aggregate, and act on it.

MRTI automates the intensive collection and classification of threat data across the Internet – from surface and deep web to the darknet. It provides data in standardized formats that can be acted upon by machines and easily shared with human decision makers.

MRTI is one key piece in the threat intelligence life cycle of:

1. Acquiring
2. Aggregating
3. Actioning

LookingGlass is the only company that incorporates MRTI into an end-to-end threat intelligence solution. Learn more about LookingGlass's MRTI and other solutions by visiting LookingGlass.

ABOUT LOOKINGGLASS CYBER SOLUTIONS

LookingGlass delivers the most comprehensive threat intelligence-driven solutions in the market, enabling security teams to efficiently and effectively address threats throughout the cyber threat life cycle.

With a scalable solutions portfolio of threat data feeds, a threat intelligence management platform, threat mitigation solutions, and threat intelligence services, LookingGlass enables security teams to prevent, detect, understand, and respond to analyzed, prioritized, relevant threats.

Additionally, with a deep knowledge of the global Internet topology and near real-time activity, LookingGlass helps organizations understand threats inside and outside their perimeter – including threats that may be impacting third party trusted partners, other organizations in their industry, and the latest threat trends impacting the global Internet at large. Know More. Risk Less.

REFERENCES

1. Mugglestone, Lynda. "'Pioneers of the Untrodden Forest': The New English Dictionary." Lexicography and the OED: Pioneers of the Untrodden Forest. Ed. Lynda Mugglestone. Oxford: Oxford University Press, 2000.
2. Berman, Michael K. "The Deep Web: Surfacing Hidden Value." Volume 7, Issue 1: Taking License, August, 2001. <http://quod.lib.umich.edu/j/jep/3336451.0007.104?view=text;rgn=main>