7
Threat Intelligence Data Sources

Intelligence is produced based on analyzing the information and data that's been collected from different sources and feeds. Data collection is an arbitrary operation that directly links to the objectives and the requirements set for the CTI project. For such, it is essential to acquire the correct data. Therefore, the more reliable and appropriate the data feed (or data sources), the better the understanding of cyber threats, which supports the organization in adapting the defense system to the threat landscape. The primary objective of this chapter is to understand *what* data needs to be collected for intelligence and *where* we can get it from. Three threat data sources will be studied in this chapter: **Open Source Threat Intelligence** (**OTI or OSINT**), **Shared Threat Intelligence** (**STI**), and **Paid Threat Intelligence** (**PTI**). PTI is also referred to as closed threat intelligence.

This chapter focuses on identifying different threat intelligence sources and linking them to the overall CTI program. We will equip you with the knowledge necessary to evaluate the need for one or more threat feed types and perform intelligence data collection using various sources.

By the end of this chapter, you should be able to do the following:

- Understand the different intelligence data sources and how to define the appropriate one for your program or organization.
- Understand the role of OSINT and its application in CTI programs.
- Understand malware data parsing and analysis in building intelligence.
- Understand the benefits of shared and paid threat feeds.
- Understand how to structure and store intelligence data for future applications.

In this chapter, we are going to cover the following main topics:

- Defining the right sources for threat intelligence
- Open Source Intelligence Feeds (OSINT)
- Malware data for threat intelligence
- Other non-open source intelligence sources
- Intelligence data structuring and storing

Technical requirements

For this chapter, no special technical requirements have been highlighted. Most of the use cases will make use of web applications if necessary.

Defining the right sources for threat intelligence

Selecting the data source is part of the data collection phase of CTI. Hence, it is a crucial step in using intelligence for security enhancement. Organizations that possess a basic security defense system manage to collect network traffic, logs, and any other activities that happen in the system. This data is a good source of intelligence. However, most companies look at external sources to enrich the **Threat Intelligence Platform** (**TIP**) or SIEM to produce reliable threat intelligence results. There are two main categories of threat data sources: *internal* and *external*. Let's discuss the difference between the two.

Internal threat intelligence sources

Internal sources include all data coming from within internal systems. These sources include *network logs* (network element logs such as firewalls, IDSes, IPSes, proxy servers, application servers, and more), *user logs*, *application logs*, *internal malware analysis*, *historical cyber incident reports*, and *security reports*. However, for the internal data to constitute a good TI data source, the CTI team or analyst must transform it into valuable and meaningful content. This is where the multi-skills capability of a CTI analyst comes into play. Although there are tools for parsing data into a specific format, a CTI analyst must process system logs, traffic, and reports into a user-friendly format that's easy to use by the rest of the organization. Gathered internal data is sent to the TIP or SIEM for autocorrelation and indexing. Once the internal data has been formatted, it can be queried to produce actionable intelligence.

Applying Internal Intelligence – Simple Case

We assume that malware infects the system through the **Cross-Site Scripting** (**XSS**) vulnerability and leaves traces in the application logs. It tries to communicate with a remote server (C2 server). It is detected and blocked when trying to access a sensitive directory. (1) TIP/SIEM automatically creates three alerts on XSS execution (medium), attempts to communicate with an external domain (major), and provides sensitive directory access (critical) – TIP/SIEM is connected to the system logs, traffic, flows, and events feed. (2) The analyst uses scripting to parse system logs, network traffic, and events and stores it in a database.

The analyst can identify the possible *entry point* (and *vulnerability*), the *path taken*, and the *files* and *directories* accessed by the malware. As such, they can extract the spread of the attack and the malware file. Every time threats are detected, the CTI team catalogue all the details (path, entry point, IP address, domain information, and so on) and builds a similarity metric between them. This can help deduce threat groups, their TTPs, and their points of interest in the organization to enhance the defense system.

The TIP or SIEM acts as a single central aggregation point and accommodates various data source formats.

Internal data sources can be used for threat intelligence to a certain extent. They are rich and represent the complete insight of activities happening in the system. However, using the internal TI data only provides two possible disadvantages: *reactive intelligence* and a *partial view of the threat landscape*, as follows:

- **Reactive intelligence**: Using internal sources, intelligence is built based on observed events and flows from the internal network. The CTI team builds intelligence as more observations are stored and processed by the TIP or SIEM. Therefore, the system is not provisioned to uncover *unknown* or *global* threats that it has not seen yet (attackers can leverage this loophole to orchestrate new attacks). It reacts to threats that it knows.
- **Partial view of the threat landscape**: By limiting the intelligence to system data, the CTI team only gets a glimpse of threats that have been used against the system. However, it could be that there are more threats that other organizations of the same profile have registered. Attackers can leverage this disadvantage to compromise your organization using known TTPs that haven't been covered by your intelligence yet.

It is essential to mix internal data with external data sources to generate a more reliable, reactive, and preventive intelligence. In the next section, we will look at external data sources and how they enrich the organization's threat intelligence program.

External threat intelligence sources

External TI sources involve data collected outside the organization. This data comes from different providers and origins such as communities, forums, governments, other law enforcement, search engines, the dark web, and security magazines. All this data can be acquired freely (OSINT and STI) or through paid subscriptions (PIT). The most popular external threat data includes the following:

- **TI data feeds**: These are streams of data coming from one or more sources. Those sources can include feeds related to malware data, customer telemetry, human intelligence, SSL/TLS certificates, internet scans and crawls, **Indicators of Compromise** (**IoC**), and counterintelligence (honeypots, DNS monitoring, and more). There are two types of TI feeds: subscribed TI feeds (which require monetary or simple user subscriptions) and open source TI feeds. Communities and companies provide this data.

- **Government sources and reports**: This is data that's provided by governments and law enforcement agencies. The US Department of Homeland Security's **Automated Indicator Sharing** (**IAS**) is one of the government sources that shares threat indicators such as IP addresses, domains, and the phishing details of malicious groups. The **Federal Bureau of Investigation** (**FBI**) is another law enforcement agency that partners with organizations in intelligence matters. The US-CERT **Cybersecurity and Infrastructure Security Agency** (**CISA**), the UK **National Cyber Security Center** (**NCSC**), and the US **National Security Agency** (**NSA**) are other examples of government intelligence sources and reports. Law enforcement constitutes a reliable external source as government services are one of the most targeted by cybercriminals. However, it depends on the breach case status as well because law enforcement data relies on breaches that have occurred.
- **Crowdsourced data**: Crowdsourced intelligence data allows the CTI team or an organization to collect public information to extract local context on security issues. Also known as collective intelligence, it allows you to tap through the collective knowledge of security experts, analysts, blogs, and more to extract information that can be used to build intelligence. An example of crowdsourced data is information coming from a blog of security analysts and managers sharing news, issues, events, and activities that they have experienced. In the case of a breach, involved security analysts share the incident details with the community. Crowdsourced data has the advantage of covering foreign languages and groups, allowing the organization to have a broader view of the threat landscape. FeedReader is an example of a tool (application) that can be used to aggregate and manage RSS feeds (security magazines, reports, blogs, and so on).
- **Business commonality**: When adversaries attack an organization (banking, health, retail, and others), they are likely to expand the attack to more organizations of the same industry. Therefore, organizations with similar business profiles have started creating groups and communities to facilitate intelligence sharing. These groups share intelligence indicators, reports, and TTPs for a common cause. An example of such a group is the **Information Sharing and Analysis Center** (**ISAC**).

External sources are not organization-specific. Thus, the need to select an appropriate one arises. Threat data sources can provide multiple streams of indicators, actors, tactics, techniques, procedures, and others, which come in a random and non-contextualized form. These sources can overwhelm the CTI team or analyst, making them difficult to consume or prove their worth. Hence, the analyst needs to select the correct feeds from the data sources and use them appropriately. We will look at three components to ease the pain of choosing the right source and the suitable feeds for intelligence: the organization profile, the feed's evaluation, and the data quality.

Organization intelligence profile

The organization's threat profile provides the baseline of the necessary data for intelligence. It is a combination of the results of several processes in the CTI program's execution. The organization profile must consider the business objectives, the CTI requirements, the threat intelligence frameworks and platforms, the tradecraft and standards, and the output of the threat modeling operation (to evaluate the assets, the attack surfaces, and the threat vectors). Building the organization threat profile to determine the data feed links all the previous six chapters of this book together. The following points detail the intelligence profile considerations:

- **Business objectives**: The CTI team or analyst must understand the business scope of the organization. This includes knowing the system infrastructure, line of business, available resources, budget for threat intelligence, and so on. Please refer to *Chapter 1*, *Cyber Threat Intelligence Life Cycle*, for more details on business objectives, planning, and the direction for CTI. Business objectives are used to limit the scope of data collection as we can sideline threat feeds that do not fall into the organization's line of business. We can use them to isolate feeds that fall out of the allocated budget.
- **CTI requirements**: The CTI team or analyst must understand the organization's security history (passed threats, attacks, and breaches) and use the CTI short-, medium-, and long-term goals to select the correct data feed. For example, by relying on CTI requirements, the analyst can sideline threat data feeds that are not linked to threat groups that target the organization. The current security stance also plays a vital role in selecting the correct data feeds. Please refer to *Chapter 2*, *Requirements and Intelligence Team Implementation*, for more details on CTI requirements.

- **Threat intelligence frameworks and platforms**: Threat intelligence platforms and frameworks are vital in selecting threat data feeds. Each TIP has characteristics and capabilities, including, but not limited to, the techniques used to aggregate data coming from different sources, the data formats supported, and storage management. Each framework has an architecture and model to analyze threat groups and TTPs. Therefore, the framework, TIP, or SIEM that's used for the CTI program drives the type of feeds that can be used. For example, an analyst can sideline feeds that cannot be used with the MITRE ATT&CK framework. They can skip feeds whose formats and APIs are not supported by the used TIP or SIEM. Please refer to *Chapter 3*, *Threat Intelligence Frameworks*, and *Chapter 5*, *Goals Setting, Procedures for the CTI Strategy, and Practical Use Cases*, for more details on frameworks and platforms for intelligence.
- **Tradecrafts and standards**: CTI analysts use APIs to connect to threat data sources. Those APIs might support one or more threat data formats (STIX, TAXII, CSV, TXT, and more). While this is also linked to the framework and the platform, the analyst needs to select data feeds that can easily be integrated with the existing security infrastructure. More details on tradecrafts and standards were provided in *Chapter 4*, *Cyber Threat Intelligence Tradecraft and Standards*.
- **Threat modeling**: Threat modeling provides excellent insight into adversaries, their potential interest in your organization, the assets that they can target, the possible attack surfaces, and the methods that they can use to compromise your organization. Data feeds provide a great stream of actors, TTPs, scope of interest, and so on. Hence, by matching the threat modeling output with the threat source data, the analyst can select the correct feed for the program. Please refer to *Chapter 6*, *Cyber Threat Modeling and Adversary Analysis*, for more details on threat modeling.

The intelligence profile is a perfect assessment of the current system's capabilities and what needs to be done in terms of threat data investment. Parameters such as industry-related risks and attacks are matched to the overall organizational goals and used to select the data feeds. Paid data sources can be expensive; thus, knowing the budget and the requirements can help protect expenses. The second component will be discussed in the next section.

Threat feed evaluation

The organization intelligence profile may be the baseline of the selection, but the data feed itself is the leading player when you're looking at the suitable data sources to ingest into the organization's security system for threat intelligence. The data feed needs to be evaluated using the criteria highlighted in the following list. Note that several criteria can be used, but we have only provided the most globally used ones here:

- **Data feed's source**: There are several sources of intelligence data feeds. The CTI team must determine the source type that is needed for the program. A CTI team might be interested in external IoCs only. In this case, counterintelligence or human intelligence data sources would be out of scope.
- **Data period**: A good data source or feed must be up to date. The information provided must be relevant, and the source provider must indicate the relevant period of the information contained in the source. The CTI must evaluate whether the data source can be used for short-, medium-, or long-term CTI goals.
- **Source authentication**: The CTI team must validate a data source regarding its transparency to know whether it is relevant and valuable to the program – hence the need to understand where the data has been taken from. For example, data feeds from government agencies and other law enforcement organizations can be considered authentic and transparent.
- **Percentage of unique data**: No one wants to pay for the same data twice. The CTI team needs to highlight the overlap between feeds or sources to minimize the possibility of redundant information. When using paid threat data sources, it is vital to look at the data's uniqueness.
- **Potential Return on Investment** (**ROI**): The CTI team must analyze the feed's content, assess the integration effort, highlight the benefits that can be taken from the source or feed, and calculate the potential ROI of the operation. The objective is to understand whether the information brings value to the organization.

Analyzing the feed or the source does not guarantee quality in the information's content. The analyst must track how each feed is used and what security actions are taken from them. For level 1 organizations (refer to *Chapter 5*, *Goals Setting, Procedures for the CTI Strategy, and Practical Use Cases*), it is crucial to review the community rating of the selected feeds or sources. Apart from evaluating the feed itself, CTI analysts must also look at the quality of the data provided by the feed or source. This means using any necessary method to assess the content of the feed. Transparent sources provide a glimpse of the data's content to help threat analysts and security professionals make informed decisions on threats feeds and sources. In the next section, we will look at assessing threat data quality.

Threat data quality assessment

Quality assessment is an essential contributor to the threat data source's value proposition. That value is defined by many parameters, such as the type of indicators and information in the data. However, some additional parameters must be considered when assessing the quality of a threat data source. Data source maintenance, updates, research for enhancements, and unique vantage points are examples of additional threat data quality assessment considerations. TI data quality is optimal input for source retention. The following list provides some parameters that can be used to assess the quality of the collected data:

- **Coverage**: The CTI team must ensure that the data source (or feed) lives up to its expectations. It has to cover everything it is supposed to. Indicators or information expected must be observed at a higher or maximum proportion. Hence, a coverage metric must be added to the quality assessment. *True positives*, *false negatives*, *true negatives*, and *false positives* are indicators that can be used to calculate the coverage of a TI data source.
- **Accuracy**: After acquiring the data, the CTI team or analyst must assess its accuracy by analyzing the rate of true positives compared to false positives. A data feed that generates a higher number of false positives might not be ideal for a security use case. Let's imagine a priority security alert that calls for a war room, and after you know that it was a false positive.
- **Latency**: One of the vital components for TI quality assessment is latency. It reflects the time it takes for the feed or source to provide an alert from a detected potential threat vector. The CTI team must ensure that the data source and the tool used for integration allow you to be notified of potential threats quickly.
- **Ease of automation**: Raw threat data can be significant in volume. Thus, it takes a lot of human effort to process the data. However, through API calls, scripts, and platforms, it should be possible to automate the entire process of data feed ingestion – hence the need to ensure that the ease of automation requirement is met.

> **Important Note**
>
> Data quality, in most cases, is effectively measured after the data source's acquisition. This means that the CTI team can integrate the data for open sources and evaluate the quality. However, for paid sources, it might not be easy to integrate the data first. Therefore, it is recommended that organizations go through a trial or proof of concept first before purchasing the data.

The CTI team must ensure that the data is relevant to the business or the operating industry and valuable (investment and revenue protection). When the organization's intelligence profile is built correctly, data quality assessment is simplified to a certain extent.

Defining the right source of TI data is a crucial step for the CTI program's success. You should be able to use the three components (organization intelligence profile, threat feed evaluation, and threat data quality assessment) to select the correct data sources (or feeds) for intelligence. In the next section, we will look at one of the most popular intelligence sources: **Open Source Intelligence** (**OSINT**).

Open Source Intelligence Feeds (OSINT)

Threat intelligence sources can be expensive to acquire from private sources. For small and medium enterprises, spending thousands of dollars on TI data subscriptions could be unrealistic (financially disadvantageous). However, organizations can leverage public data from open sources to build intelligence. OSINT sources and feeds are a result of collective intelligence in a public fashion. Organizations, analysts, and researchers aggregate and structure their security output results and publish them as feeds for free. OSINT sources include overt feeds, search engines, usernames, email addresses, domains, social networks, IP and DNS lookups, and URLs. The list of OSINT sources is long, and the CTI team must be able to select the correct OSINT data for a specific intelligence program. Let's have a look at some benefits of OSINT.

Benefits of open source intelligence

As the name implies, open source data sources are publicly available – even though some skills and extra work might be needed to access them. For that, they have become an attractive field of threat information for all kinds of security organizations. The following points highlight the main reasons for the explosion of OSINT:

- **Low or no cost**: Data sources and feeds from security and CTI vendors are expensive and can be out of reach for small and medium enterprises or organizations on a tight budget. Open source intelligence feeds save on the budget and ensure a potentially higher **Return on Investment** (**ROI**) while providing great technical value.
- **Public maintenance**: Most open source threat intelligence data is maintained by the information security community around the globe. Researchers and experts keep on updating and enriching OSINT platforms for general purposes: fighting cybercrimes and protecting assets.

- **Data accessibility**: OSINT is accessible to all. This means that security analysts can always use them when necessary. Most of them are accessible through **application programming interfaces** (**APIs**) or simple registration and download.
- **Data availability**: OSINT such as social media, search engines, and human interactions are always there. The people forums on Twitter or articles on LinkedIn are always available and open to the public, and topics are frequently updated.
- **Information sharing**: The goal of OSINT is to allow the infoSec community to share information about threats, vulnerabilities, exploits, breaches, best practices, and more. This benefit allows organizations to set up preventive measures. For example, an exploit discovered in Asia can help organizations in South Africa take precautions (such as patching or suspending the attack surface, which could lead to the exploit being used against them).

Many organizations, including national security agencies, law enforcement firms, and organization security leaders, are relying more and more on public sources to develop intelligence strategies. For example, by monitoring public forums, social media, news outlets, and internet traffic, cyber threats can be anticipated. Black hat hackers sell exploits and sensitive data on the dark web (or darknet). The dark web as OSINT, can be used to identify breaches that have occurred or possible information leaks. In the next section, we will look at some popular public data sources.

Open source intelligence portals

There are several open source feeds that organizations can use to start collecting data for the CTI project. We can't cover all of them here. Instead, we will discuss some of the most used open source threat intelligence platforms that help an organization initiate a CTI program. These sources are free and could require registration or membership.

Department of Homeland Security – Automated Indicator Sharing (AIS)

Automated Indicator Sharing (**AIS**) permits real-time sharing of threat indicators and defensive procedures to provide organizations with the minimum components to minimize cyberattacks and manage damages they can cause (`https://www.cisa.gov/ais`). Many communities, US federal departments, agencies, and foreign companies are part of AIS. Intelligence is shared at no cost. Using AIS, participants share indicators, defense tactics, and mitigation procedures in near-real time.

AIS uses **Structured Threat Information Expression** (**STIX**) and **Trusted Automated Exchange of Indicator Information** (**TAXII**) for threat indicators and machine-to-machine communication, respectively. It follows a client-server architecture where participants use a STIX/TAXII client to communicate with the **Cybersecurity and Infrastructure Security Agency** (**CISA**) server. The following steps are required to get access to the AIS indicator feeds (more details are given on the CISA website):

1. **Register with CISA**: Organizations need to register with CISA and sign the terms and conditions to initiate membership processes.
2. **Deploy a STIX/TAXII infrastructure**: Organizations must install a STIX/TAXII client to communicate with the CISA server and exchange threat indicators and defense measures. STIX/TAXII is an open source standard; hence, its infrastructure can be built internally or acquired through vendors. In *Chapter 5*, *Goals Setting, Procedures for CTI Strategy, and Practical Use Cases*, we deployed a basic STIX/TAXII client to connect to public STIX/TAXII servers and get indicator feeds.
3. **Acquire a Public Key Infrastructure** (**PKI**) **certificate**: Each participant must have a PKI certificate provided by a **Federal Bridge Certificate Authority**.
4. **Sign an interconnection agreement and provide the IP address to CISA**: The potential participant must sign an agreement to connect to the CISA server and access shared intelligence. They must provide the public IP address to CISA to be added to the trusted list of participant addresses. Now, the organization is ready to exchange intelligence with all AIS members.

CISA protects participants' information by anonymizing intelligence submissions. This protection is done through the US CISA Act of 2015, granting liability, privacy, and all the necessary protections to the AIS members. Note that any organization can become a member of AIS (question to follow the required steps).

AlienVault – Open Threat Exchange (OTX)

Open Threat Exchange (**OTX**) (`https://cybersecurity.att.com/open-threat-exchange`) is an open TI community that facilitates collaborating and sharing the latest threat data, trends, and techniques. Security professionals and researchers around the globe participate in growing the platform by submitting threat indicators frequently. To use OTX, you need to subscribe to **OTX Pulses** as it provides IoCs related to threats. OTX can be integrated into the TIP or SIEM by using the DirectConnect API. The API synchronizes OTX intelligence with the internal system tools.

OTX can work as a STIX/TAXII server, making it usable with any STIX/TAXII client. It can also be integrated into the MISP platform. Some of the IoCs that are generated include IP addresses, domains, hostnames, emails, URLs and URIs, file hashes, CIDR rules, CVE numbers, file paths, and more. Using OTX data can help you address the following points:

- **Organization exposure level**: How badly the organization is exposed to threats.
- **Threat relevance**: By analyzing OTX intelligence, the CTI analyst can determine how relevant a threat is to the organization.
- **Threat actors and groups**: OTX IoCs allow you to identify the adversaries and motives behind threats.
- **Targeted assets**: Based on the data collected, OTX can help you identify the most targeted assets.

> **Important Note**
>
> AlienVault OTX is one of the intelligence sources and services that organizations can use to start a CTI program. Not only does it provide integration with SIEM and TIP, but it also acts as STIX/TAXI server, which saves and protects costs for commercial STIX servers.

Intelligence sources such as OTX and AIS can be considered standalone intelligence platforms. They extract information from multiple data sources such as search engines, forums, emails, logs, and more to make them available to the public in a format such as STIX/TAXII or database files.

InfraGard Portal – FBI

InfraGard defines partnerships between the FBI and other sectors (public and private). It is a collaboration where the FBI and organizations share cybersecurity incidents and threat intelligence to fight against cybercrimes in the US. The requirements and nature of information sensitivity make it a US-based membership source – only US companies, academies, and individuals can subscribe to the data source.

InfraGard divides critical infrastructures into 16 sectors that represent the major business areas. Organizations (members) have access to recent cyber threats and crimes being tracked by the FBI. More information can be obtained from the official website (`https://www.infragard.org/`).

> **Important Note**
>
> It is essential to vet **Indicators of Compromise** (**IOCs**) coming from third parties such as law enforcement and government sources. Some of them (such as hashes, IP addresses, and domains) might be legitimate programs or application indicators. Not vetting IOCs can result in business disruption. Refer to the intrusion analysis use case in *Chapter 10*, *Threat Modeling and Analysis – Practical Use Case*, for more information.

Internet Storm Center (ISC) and DShield

Internet Storm Center is an open source and accessible platform for cybersecurity events sharing. The SANS Institute powers ISC by bringing security experts and organizations together to analyze internet traffic (on a large scale) for malicious activities. It gathers intrusion detection log entries daily using sensors. It covers over 500,000 IP addresses in over 50 countries (`https://www.dshield.org/about.html`). ISC commits to identifying sites used by adversaries to commit cybercrimes and availing data on the types of attacks orchestrated in various industries and regions.

Sensors collect information about suspicious traffic on the internet. These sensors that are used by the ISC work with most network security elements (firewalls, IDSes, broadband devices, operating system data) and send the data to the DShield database to be analyzed by volunteers and machines to identify behavior. The result is posted on the ISC site (`https://isc.sans.edu/dashboard.html`) to be viewed or queried through scripts. The SANS Institute also sponsors the DShield service. CTI analysts can register with DShield to get the benefits of *accessing the logs* collected and submitted to the database. The registration can allow the CTI team to download the logs and integrate them with internal security systems (TIPs and SIEMs).

Many organizations detect cyberattacks long after the system has been infected. ISC encourages organizations to submit their network security logs to Dshield for analysis. Hence, another advantage of registering with DShield is to activate the *fightback capability*, which allows the ISC to notify an organization of an occurring or occurred attack.

> **Important Note**
>
> Any organization can leverage the SANS Institute DShield service to monitor intrusion attempts and protect the system. While the registration provides some benefits, it is not required to submit network element logs.

For more details about the ISC and DShield, please visit the official website provided in the second paragraph of this section.

Information Sharing and Analysis Centers (ISACs)

Information Sharing and Analysis Centers (**ISACs**) (`https://www.nationalisacs.org/`) provide central points for collecting cyber threat information affecting different critical infrastructures. An organization can join an ISAC depending on the sector. ISAC membership may come with a cost, but the intelligence provided is affordable or free to the public. Each sector's ISAC (such as financial services, communication, health, and others) collects and analyzes threat information. The generated intelligence is then shared with the members, along with the tools to mitigate risks and improve security. Some benefits of joining an ISAC include the following:

- 24/7 sector-specific monitoring and alerting of cyber threats through the portals, platforms, and applications
- Hands-on training, exercises, summits, and security events for registered members to ensure that they get the full benefits of the platform
- Reliable support for all the steps of a cyber incident
- Supports the STIX/TAXII standard

Organizations can leverage the benefits of ISAC to integrate intelligence into business operations. Membership fees are linked to the organization's size and infrastructure (critical assets, revenue, customer base, and company size).

Abuse.ch

Abuse.ch (`https://abuse.ch/`) is an open source platform owned by a Swiss professional for malware tracking and monitoring. The platform tracks indicators such as IP addresses, domains, URLs, distributed sites, payment sites, and C2 servers associated with different malware types. Several security professionals, vendors, researchers, and law enforcement agencies use abuse.ch to understand mechanisms used by malware and how threat actors use those mechanisms to launch cyberattacks. Abuse.ch has main projects including MalwareBazaar (used to share malware information), *Feodo tracker* (used to share botnet C2 servers associated with the Feodo malware), *I got phished* (for acquiring information on phishing victims – domains, email addresses, IP addresses, and others), and *SSL blacklist* (used to identify malicious SSL certificates that can help organizations detect fraudulent SSL connections). SSL blacklist also helps detect and block malware botnet C2 channels on the TCP layer. Additionally, there is *URLhaus*, used for sharing information on URLs mainly used for malware distributions, and *threat fox*, which is used for sharing **Indicators of Compromise** (**IoCs**) with the cybersecurity arena.

Abuse.ch is a non-profit organization; hence, it allows CTI analysts, security vendors, and IT security professionals to collect information freely and build better cyber defense systems. There are two standard ways to collect data from abuse.ch: use the provided APIs or download the data as a CSV file and load it into the TIP or SIEM.

> **Important Note**
>
> Abuse.ch's MalwareBazaar provides a Python API (`https://bazaar.abuse.ch/api/`) that the CTI team can use to integrate malware samples (data) into the TIP or SIEM system. The API can also be used to upload malware samples to the MalwareBazaar database. The latter can also be browsed for malware information. Malware samples must be dealt with carefully to avoid infection during analysis.

It is essential for security analysts and researchers to contribute to the platform by uploading malware samples, phishing data, and botnet C2 channel information. Not only does it enrich the platform, but it also helps the security community in fighting against cybercrimes.

The dark web, DarkReading.com, BleepingComputer.com, and other news portals

The dark web is considered the underworld for malicious activity. The dark web hosts some of the most valuable hacking forums and marketplaces for malicious hacking tools and resources in the security scope. According to a Forbes report in July 2020, over 15 billion logins from 100,000 data breaches were stolen and sold on the dark web (`https://bit.ly/3uc4qpO`). The CTI team and analysts must develop mechanisms to collect information on the dark web to identify potential threats. Most of the paid intelligence sources crawl the dark web for additional intelligence and threat warnings. However, small enterprises with a limited budget can leverage the dark web to collect threat data by developing in-house web mining operational systems (subject to having the necessary skills). The dark web can reveal important threat information such as *newly developed malware*, *new exploits*, and *zero-day vulnerabilities* that aren't known or available on the surface web yet. The CTI analyst needs to align the data they've collected from the dark web with the CTI objectives and requirements to answer basic questions such as the following:

- **The trending topics**: What are the trending security topics in the dark web, and how do they impact us?
- **Vendor's presence**: What vendors are mostly discussed in dark web forums and marketplaces?
- **Sector's presence**: What sectors have a higher presence in forums and marketplaces? The CTI can establish a link with the CTI consumer business area.
- **Exploits and malware**: What exploits are being developed, used, or sold on the dark web? The CTI team or analyst needs to understand whether the organization could be a target of such exploits and malware.
- **Vulnerabilities**: What are the newly found vulnerabilities and their exploits? The CTI team can then evaluate the organization's stance against those vulnerabilities.
- **New threats and groups**: Are there any new threats or groups marking their presence in the dark web?

The data that's collected from the dark web can help the CTI team make security decisions to protect the organization from known and unknown threats. However, there are surface websites that provide news obtained from the dark web. One of them is Darkreading.com.

Darkreading.com (`https://bit.ly/3omPMuh`) is an online community for security news. The site provides the latest news about threats, vulnerabilities, and industry trends. It also allows the information security environment to discuss and suggest reliable defenses against those threats. Darkreading.com crawls the dark web and the standard internet to produce one of the most reliable reports and news on cybersecurity matters. It englobes 14 communities addressing enterprise security challenges: Analytics, Attacks and Breaches, Application Security, Careers and People, Cloud Security, Endpoint, IoT, Mobile, Operations, Perimeter, Physical Security, Risk, Threat Intelligence, and Vulnerabilities and Threats. DarkReading.com provides important information (CVEs, attacks, threats, IoCs, and so on) that CTI analysts can use to build intelligence.

BleedingComputer.com (`https://www.bleepingcomputer.com/`) is a popular security and technology news source that provides the latest articles, reports, and news on threats and other computer topics through its vast forums. Another vital news portal for threat intelligence feeds is Hacker News (`https://news.ycombinator.com/`). As a CTI analyst, you must be aware of essential news portals for threat data feeds.

Search engines and social media

Search engines contain databases of information that can be useful to CTI analysts and the public. Internet users go to standard search engines (such as Google, Bing, and Safari) for simple searches, but particular search engines might be required for security information searches and retrieval. Adversaries use search engines to plan and build the attack profile (reconnaissance, vulnerability searches, zero-day exploits, devices connected to the internet, system archives, and so on). There are several security search engines, and it is the CTI team or analyst's responsibility to identify them. Some of the most widely used engines for threat intelligence data collection include the following:

- **Shodan**: Shodan (`https://www.shodan.io/`) is a search engine for internet-connected devices. The CTI team can use Shodan data to monitor and track your organization's exposed devices (network elements or endpoints). By using Shodan data, the analyst can identify vulnerabilities and open ports that adversaries may exploit. It provides an API to integrate data into SIEM or TIP for correlation with other data for reliable intelligence.
- **ZoomEye**: The number of connected devices is still escalating. ZoomEye (`https://www.zoomeye.org/`) is a search engine that retrieves internet-connected devices and fingerprints them for open ports and services. The CTI team can use the ZoomEye-python API to collect data through various queries. The API documentation is available on the official website.

- **Censys**: Censys (`https://search.censys.io/`) is more than just a search engine. However, as a search engine, it helps monitor and discover devices, domains, and site configurations. It reports open ports, vulnerable protocols, services, SSL certificates, and others. CTI analysts can use Censys data to monitor the organization's attack surfaces.
- **Hunter**: During reconnaissance and attack planning, adversaries gather information such as email addresses that could be used for spearphishing. Hunter (`https://hunter.io/`) is a search engine used by professionals to collect corporate email addresses. By just entering the company's name, the system responds with verified emails tied to the company. Hunter provides an API that analysts can use to analyze publicly available information linked to the organization.
- **GreyNoise**: GreyNoise (`https://greynoise.io/`) is an internet scanning finder engine. It provides data of all the devices (IP addresses) that are scanning the internet. By using GreyNoise data (through their APIs), the CTI team can identify benign and malicious scans targeted at the organization and create alerts. GreyNoise data can be integrated with TIP, SIEM, or **Security Operation Center** (**SOC**). You have to register and get the API key.

  A simple search is shown here. First, it returns all the devices scanning the internet for port `554` and their tags (malicious or unknown). Then, it returns the devices searching the internet for the matching TLS/SSI fingerprint:

```
raw_data.scan.port:554
raw_data.ja3.fingerprint:795bc7ce13f60d61e9ac03611dd36d90
```

> **Important Note**
>
> Greynoise can easily be integrated with MISP using the community API. Hence, for level 1 organizations, this can be a great start.

- **WiGLE**: WiGLE (`https://wigle.net/index`) can be used to explore and map surrounding wireless networks. It does so by using the wireless network signal strength and the address specified (or location for smartphones). It allows the security analyst to pinpoint the location (building, apartment, room) and view the information of nearby networks. WiGLE provides various security mechanisms (open, WPA, WEP, WPA2, and WPA3) for highlighting the vulnerability level of all scanned wireless networks. Security or CTI analysts can use WiGLE information to locate and assess the organization's Wi-Fi. They can also use it to evaluate whether the same vendor provides Wi-Fi services in a particular area. Another valuable feature for CTI analysts is to analyze whether the Wi-Fi has been spotted or marked by anyone using WiGLE before – which could be a sign of possible reconnaissance. WiGLE provides a Python API that can be used to access and collect data from their database.
- **Pipl**: While mostly used by hackers and pentesters for spearphishing targets, Pipl (`https://pipl.com/`) allows CTI analysts to assess the amount of individuals' personal information exposed publicly. The site is used by security agents, government agencies, and many other organizations to search for information about individuals. Pipl crawls most public sources and the deep web to gather information. The CTI team or analyst can use the Pipl API to collect and integrate data into the organization's security tools. Although it is commercial, its price is pretty much affordable for small businesses and individuals.
- **Google Hacking**: Google hacking or Google dorking is a search engine that security researchers use to find public information and vulnerabilities in code and configurations. It is a good source of OSINT and is also integrated into security databases such as exploit-DB. The search engine relies on advanced operators to refine searches. CTI analysts need to understand how to integrate Google dork queries into their internal security systems to assess an organization's exposed resources (documents, codes, configurations, devices, cameras, and more).

There are many open source or low-cost search engines that the CTI team can use to enrich the TIP or SIEM system. The list of sites mentioned here is not exhaustive. Hence, the CTI analyst's responsibility is also to identify the search engines that could be valuable to reach the CTI objectives.

Nowadays, social media provides a lot of information than just connecting people. With APIs, developers and security analysts can access social networks such as *Twitter and Tweetdeck (https://tweetdeck.twitter.com/)*, *Facebook*, *LinkedIn*, and *YouTube* to perform *searches*, *retrieve analytics*, and harvest *archives* and *documents*. These functionalities help collect social media data for intelligence – following security researchers, outlets, and agencies to find out about exploits, malware samples, and vulnerabilities news. We can retrieve hot security topics about threats, vulnerabilities, exploits, malware, breaches, and more. In the following subsection, we will look at some OSINT or low-cost malware data collection sites.

Publicly accessible malware portals

Most data breach reports and investigations have shown that a high percentage of cyberattacks involve malware, directly or indirectly. Hence, the CTI team or analyst needs to collect malware-related data for intelligence. Many sites or portals facilitate malware analysis openly. We will not cite all of them, but we will provide some malware sites and links that can freely be used for malware analysis and data collection:

- **VirusTotal**: VirusTotal (`https://www.virustotal.com/gui/`) is a free online malware analysis platform that allows security professionals to submit malware files and get analysis results. The result is shared with the requester and the organizations that are partners with VirusTotal. Through the VirusTotal API (public or premium), the CTI team can enrich the internal security tools with telemetry data to facilitate alert reports and integrate the data with TIPs and SIEMs.
- **VirusShare**: VirusShare (`https://virusshare.com/about`) is a malware samples repository that provides organizations and individuals with access to malicious code samples. VirusShare is free, but access is subject to an invitation. Thus, an organization or individual needs to request whether they can be added to the list to access the data. However, it has commercial APIs for data feeds and specialized searches of the data.

- **Any.run community**: Any.run (`https://any.run/`) is a malware analysis and sandbox platform that allows information security analysts to access live malware data processing, analysis, and IoC data. It can also be used to enrich TIPs and SIEMs. You can register for the community edition of the tool and download samples of malware data.
- **Intezer community edition**: Intezer (`https://www.intezer.com/`) is a malware analysis, threat hunting, and incident response platform that allows analysts to classify and reverse engineer malware files. Intezer supports multiple data formats, including STIX. Hence, it can be integrated into modern TIP and SIEM.

Many malware sites are offering free analysis and access to data. The CTI team or analyst must search for those sites and understand how they can benefit the program.

In this section, we have looked at some of the publicly available open source and low-cost data sources. OSINT is a broad area of research and application. Hence, it demands a thorough understanding and selection. Most of the sources provide mechanisms to connect and collect data (APIs or database downloads). In the next section, we will look at the basic structure of open source intelligence resources.

OSINT platform data insights (OSINT framework)

Collecting OSINT data can be challenging (where to fetch the data from, how to collect it, and what to look for in that data). It requires a good knowledge of the underlying data sources. The sources listed in the previous section, *Open source intelligence portals*, are ideal for when you wish to start collecting data. Much more information can be collected as part of the OSINT framework (`https://osintframework.com/`). This framework summarizes the resources and tools that can be used to collect and gather security information.

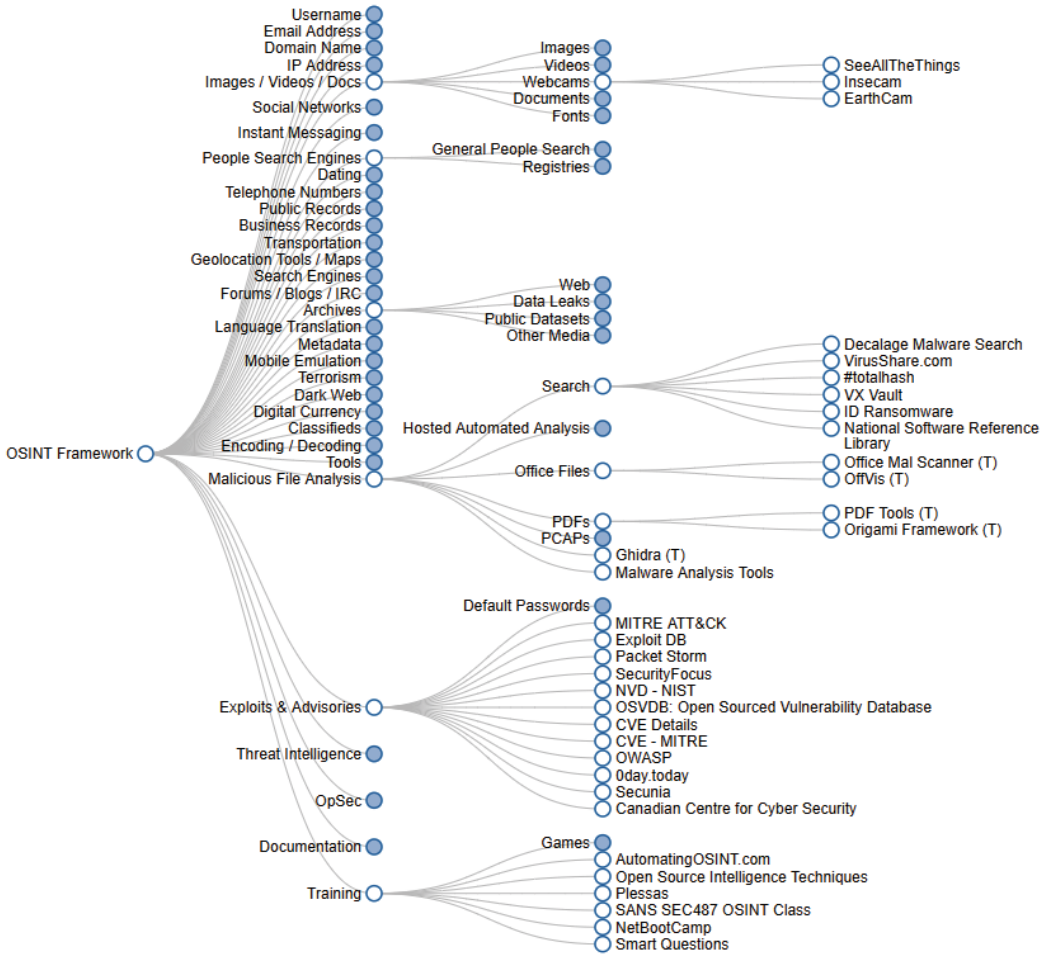

Figure 7.1 – OSINT framework tools and resources

You can browse each OSINT category and access the tool or resource for an in-depth understanding of them. Tools and resources marked with *R* require registration, while those marked with *T* need to be installed locally. Although the tool is still being updated and improved, it can be used to create a list of OSINT data sources. If we click on **Exploits & Advisories > MITRE ATT&CK**, the site will redirect us to the MITRE official website for more details. In the following subsection, we will discuss the probable limitations of OSINT data.

OSINT limitations and drawbacks

OSINT sources provide a large amount of data – data that could be structured or unstructured, controlled or uncontrolled (in terms of overhead). Some of the challenges in working with OSINT include the following:

- **Demanding data cleaning**: Most OSINT data or feeds are raw and not ready for consumption in their initial state. Many post-processing operations need to be performed on them to get them into consumable formats (ready to query, correlate, or integrate with other tools).
- **Source validation**: Source validation and evaluation are essential in working with threat intelligence data and feeds. The voluminous nature of OSINT makes it challenging to evaluate, especially manually. Imagine trying to validate 10 sources of data. It is possible but could require time and resources. How do you validate OSINT coming from a source that conducted corporate or government espionage? This ties OSINT to the data accuracy problem.
- **Information filtering**: Finding the required information in data with a lot of noise is a challenge. And because of that, there is a chance of getting misleading information, resulting in many false positive alarms.

Two components can be invoked to address some of the limitations of OSINT: *data collection automation* and *field expertise*. Data collection automation reduces the probability of human errors and improves the data collection cycle's speed. And by having expertise in the collection process, the CTI team can find optimal ways to separate noise from data insight and the generated false positives. Thus, it is recommended to automate the collection process as much as possible.

OSINT is a good source for intelligence project initiation. Therefore, level 1 and 2 organizations can leverage that to start building intelligence. However, it is essential to understand that the objectives and the CTI requirements must be the drivers of the OSINT data collection process. When suitable sources are selected, they can be correlated with internal data to provide reliable intelligence. In the next section, we will look at malware data and its use for threat intelligence.

Malware data for threat intelligence

Malware is one of the most commonly used words in cybersecurity. Its invocation brings turmoil and torment to organizations and bliss to attackers. Its concept is complex to comprehend as it requires a different set of skills and expertise (architecture, analysis, and design). However, as a CTI analyst, understanding malware data and its collection must become second nature. In this section, we will look at malware data that is fed to the TIP or SIEM for intelligence.

Sites (or sources) such as VirusTotal, VirusShare, and other malware sandboxes are powered by malware analysis engines that allow them to analyze the behavior of files or links that have been uploaded to classify them as malicious or not. Although open source sandboxes can help CTI teams and analysts perform malware analysis automatically, it is essential to understand the basic information about malware data.

> **Important Note**
>
> Malware analysis is a topic on its own and is not part of this book. We have only provided basic information on malware and the fundamental indicators that CTI analysts can extract from malware data.

Modern malware design techniques can bypass security systems and avoid detection. Advanced analysis is required to detect malware threats reliably. The CTI team or analyst aims to extract IOCs from the data and feed that to the TIP or SIEM by connecting to malware-related data sources. They can then correlate that with internal data to create alerts for future threat detection.

Benefits of malware data collection

The key benefits of using malware data for threat intelligence include the following:

- **Malware detection**: Identify malware IOCs that need to be blocked. The CTI can implement notification services to report on possible malware threat detection.
- **Threat prioritization**: The CTI team can create high-priority alarms on malware detection where malware threats have high priority over other IOC alerts. Automated threat prioritization makes the triage task simpler.

- **Threat hunting**: Malware data IOCs and artifacts can be used by threat hunters to identify identical malicious operations across the entire network. For example, if malware modifies directory files in a certain way, threat hunters can use the same behavior to check similar modifications across the network.
- **Reference for future research**: Security researchers can leverage malware data to understand patterns, techniques, technologies, codes, and infrastructures used by adversaries (or groups). It helps to anticipate changes in adversaries' designs.

Malware data is essential when implementing the TI program. It gives the organization the necessary means to make informed decisions about malware-related threats. The data needs to be correlated with the system traffic.

Malware components

Malware is designed to carry out specific tasks in the victim's system. Although there are a lot of functions malware can carry out, the most probable include stealing information, destroying the system, and modifying the system. The CTI analyst needs to understand the different components of malware, which define its purpose. They are illustrated in the following diagram:

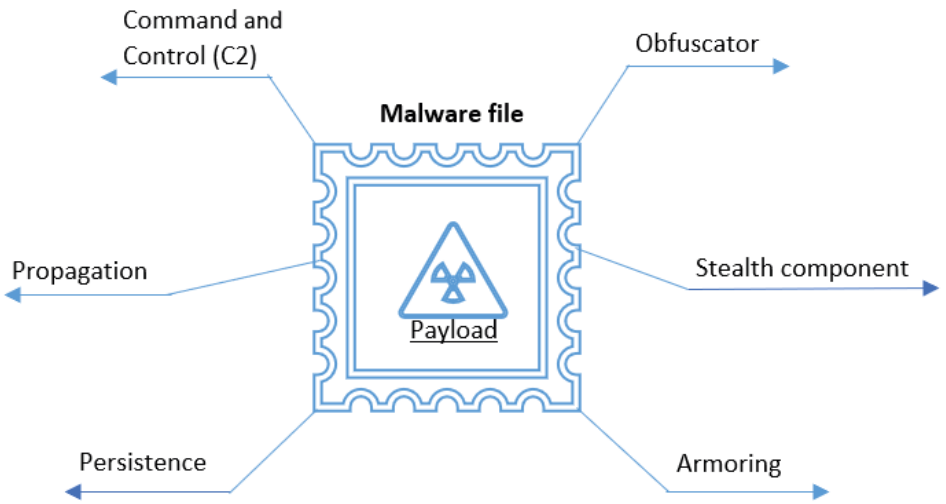

Figure 7.2 – Malware components

Every well-designed malware will contain the following:

- **Payload**: The payload is the central part of the malware and performs the intended malicious activity. It is written using programming languages, instructed to perform one or more tasks. A payload can be programmed to hide its presence.

- **Obfuscator**: It masks the payload so that it is undetectable by antiviruses and other security devices and software. Obfuscation is achieved using wrappers, protectors, and encoders to modify the look of the payload. One of the tasks of malware analysts is to *reverse engineer* the obfuscator to access the payload's code. A good malware designer ensures that the obfuscator is solid and irreversible. Compression and encryption are two methods used to obfuscate malware.
- **Stealth component**: Malware will hide from security defenses by using legitimate processes and ports, changing its properties, or using rootkits. Using a trojan with spoofed extensions such as `new_mercedes.jpg.bat` or using the `WriteProcessMemory()` API to highjack a legitimate process is an example of how malware can escape the defense system, hence the need to efficiently monitor system processes and highlight changes in process behavior.
- **Armoring**: Attackers design malware in such a way that they can detect tools that can threaten them. They tend to change their behavior in the presence of tools such as monitoring tools, virtual environments, and troubleshooting tools. Sophisticated malware can disarm some of the tools to protect itself or modify its execution logic. Wireshark, VMWare, VirtualBox, and debugging tools are examples of tools that malware can armor itself against. Hence, malware analysts need to consider malware behavior in different circumstances and environments.
- **Persistence**: Unless specified by the designer, malware is likely to persist to the victim's system, allowing the attacker to access the victim's system when necessary – even after a system reboot. In most operating systems, attackers rely on registries, services, persistent programs, and memories to make malware persistent. For example, any malware stored in `C:\...\Programs\Startup` (Windows) or `/etc/systemd` (Unix) is likely to start with the operating system at each reboot.
- **Propagation**: Malware will find ways to spread inside the victim system. Depending on its type (virus, trojan, worm, botnet, and others), its propagation can be automated (standalone) or initiated by an external trigger. Worms, for example, can replicate and spread without external intervention. When assessing malware behavior, understanding the propagation mode is critical.
- **Command and control**: C&C or C2 is *the way* the implanted malware communicates with the external world, taking commands from the attacker to perform more actions on the victim. C2 relies on IP addresses and domains to communicate with the outside world. The IPs and domains can either be embedded in the payload or generated dynamically using modern algorithms to avoid being blocked.

Each component assumes a specific role in the malicious task that the malware needs to accomplish. Those components are also the fundamental elements in understanding the behavior of malware. For organizations planning to have in-house malware analysis labs, understanding these details is crucial in understanding the malware itself.

Malware data core parameters

The result of malware analysis, either on-premises or using data collected from malware sources, is a set of *indicators* that will aid in fighting and protecting the system against malware attacks. Those indicators are extracted from malware components, and they help group malware by functionalities and families. The following diagram shows various malware data indicators:

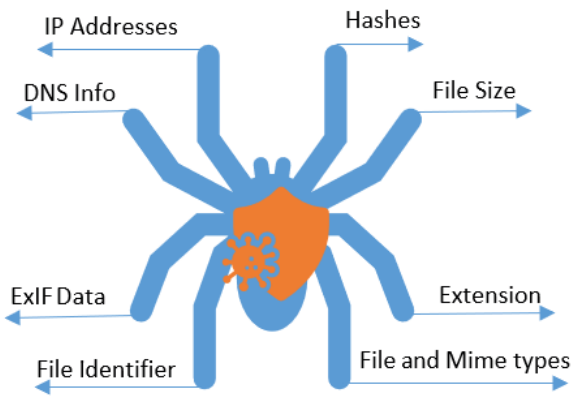

Figure 7.3 – Malware data main indicators

Security vendors are employing advanced techniques to detect, isolate, and remove malware. The *signature-based* method is becoming less and less efficient due to sophisticated obfuscation techniques used by attackers. But combined with *behavior-based* analysis (modernly empowered by AI), vendors are developing sophisticated ways to detect malware. The following points dig a little into the malware parameters that help the CTI in mining the data:

- **Hashes**: Malware uses many cryptographic hashes for identification and as a protection mechanism (evade security by modifying the hash value, for example). Attackers use hash algorithms such as MD5 and the SHA family (SHA1, SHA256, SHA224) to protect the payload. The CTI analyst must extract all the hashes that are used in the malware.
- **File size**: The file size is another parameter that needs attention. The malware codes (or content) determine the file's size.

- **Extension**: Malware files come with extensions that also determine the environment where they can run. Windows malware can have `.exe` or `.bat` extensions, depending on the designer, or `.sh` for Unix environments. However, it is essential to know that malware extensions can be obfuscated (*fileless malware*, for example, running exclusively in memory) or spoofed (to mimic popular unsuspicious files). However, some tools and methods can be used to analyze fileless (*memory monitoring*) and spoofed malware.
- **File and mime types**: The file and mime types determine the extension and the content, respectively. These are the two ways to identify a file type correctly. An attacker can spoof the extension, but the mime will show the correct identity. `PE32 executable (GUI)` and `application/x-dosexec` are examples of file and mime types, respectively.
- **File identifier**: The *TrID*, also known as the file identifier, identifies the file type using binary signatures. The TriD can be used to recover files as well. The TrID tool can also be used standalone for investigation and forensics cases.
- **ExifTool data**: Exchangeable Image File Format is a standard for most files, such as pictures. The Exif tool can be used to manipulate images, videos, and document files such as PDFs. ExIF data can have fields such as the operating system, the code language in the malware or document, character set, and many more.
- **DNS info**: If you're lucky or skillful enough, the malware data can report the domain information of the remote or the C&C server.
- **IP addresses**: This is the same as the DNS information; IP addresses can be extracted from the malware data.

Depending on the data, the analyst can extract the necessary parameters to classify the malware.

The CTI team is not the only beneficiary of integrated malware data indicators. The incident response team and the security operation analysts leverage the data to uncover malicious behavior, create alarms, and monitor the system's efficacy against malware threats.

Malware data is essential for a reliable intelligence program, and there are several OSINTs available to collect malware data (some of them were mentioned in the previous section, *Open source intelligence portals*). In the next section, we will look at paid intelligence sources.

Other non-open source intelligence sources

Paid intelligence sources are becoming more and more the preferred way of conducting intelligence – for organizations that can afford the service cost. The concept of paid intelligence can be divided into *curated paid intelligence* and *original paid intelligence* sources. Certain providers leverage OSINT by processing and aggregating data from multiple open and low-cost sources to provide curated and rich intelligence data. This data is then sold to organizations (curated PTI). Other providers leverage industry-advanced technologies, research, proprietary lead features, and customized information to provide expanded intelligence services (original PTI). Curated PTI services are less expensive compared to original PIT services and can be affordable. However, the value proposition from the original PTI could be the game-changer between that and OSINT and curated PTIs.

Benefits of paid intelligence

Logically, any security analyst would consider paid intelligence to be higher quality and more reliable than open source intelligence. The objective is not to affirm or deny this hypothesis but to provide the CTI analyst and security stakeholder with the necessary knowledge to make an informed decision on paid or open intelligence sources. We can use specific threats combined with feed evaluation metrics to highlight some of the benefits of paid intelligence, such as the following:

- **Information accuracy**: Accuracy is a vital metric in evaluating an intelligence feed, either open, privately shared (from communities), or paid. Inaccurate information can be detrimental to an organization's security stance – misleading information and false positives. PTIs rely on dedicated experts who analyze the feed's accuracy before delivering it to the customers. Information *noise* is dealt with expertly – in the case of mistakes, there is a certain level of recourse (commercial, legal, or operational, depending on the agreement).
- **Data period** (**time**): How often data is updated and contextualized is also essential when evaluating the feed. Open source intelligence sources mostly rely on attack cases' results. First, breaches and threats are analyzed to identify critical IOCs and malicious components and then shared with the public or the community. This process sometimes results in a longer delay and detection cycle – by the time an attack is thoroughly analyzed, several organizations might be victims, or the adversaries might have changed TTPs. PTIs rely on continuous expert analysis to provide preventive strategies to customers by constantly updating the sources with a forward projection look.

- **Processing and integration**: Paid TIs provide an advanced integration and mediation layer to ingest information into other security platforms. They support most of the industry intelligence standards and tradecrafts. This characteristic optimizes the time for the CTI team in trying to get the data in a user-friendly format.
- **Requirement alignment**: Intelligence must be actionable. Therefore, the data sources must align with the requirements. Paid TIP vendors tend to sit with the CTI team and security stakeholders to discuss the requirements and ensure that the feeds or provided TI data align with the organization's goals. Data relevance is of the utmost importance.
- **Information protection**: While OSINT is publicly exposed, PTI protects its value and exclusivity. The attackers do not know when their actions are detected and dealt with because PTI vendors tend to cover their technology. This point gives an advantage to the PTI consumers.
- **Vendor extra services**: Certain PTI vendors such as IBM, Microsoft, Cisco, Spirent, and others are also equipment and product vendors. An organization infrastructure built on top of those vendors' technologies can benefit from additional security services at a lower or free cost. Other PTI vendors also offer additional services on top of these feeds – such as spam filtering libraries, malware detection boxes, and phishing identification services.
- **Wide coverage**: Some PTIs include OSINT. It means that PTI vendors can access open and shared sources and process them (validate, clean, and reformat) before sending them to the consumers. This alleviates the burden of the in-house evaluation and analysis of open source feeds.
- **Customer support**: Instead of just focusing on threat indicators, paid intelligence vendors provide platforms and portals to download reports, ask questions, submit issues, or get the latest news on threats and data breaches. Packaged service support can be an attractive component to organizations as they facilitate vendor-customer interaction.

PTI provides many advantages and some unique benefits that can help organizations collect relevant data that match the intelligence requirements so that you have enough accurate information to build an effective cyber defense system.

Paid threat intelligence challenges

PIT offers several benefits but is not spared from challenges. While it provides quality in data, there are parameters that a CTI analyst or team or security stakeholders must consider when selecting PTIs:

- **Cost**: The cost of the original PTI sources can be out of reach for small and medium organizations. And even for large enterprises, it is essential to ensure that the cost is not outside the budget's boundaries. Feeds can range from $1,500 to $100k based on a monthly subscription, with curated PTI feeds being the less expensive option. It is easy for an organization to find itself paying over $1M for threat data feeds. However, it also depends on the number of feeds required.
- **Feeds overlap**: The CTI team might need more than one paid intelligence source to address the security requirements effectively and have more coverage. However, combining several sources might result in information overlap. Many paid intelligence feeds cover the same scope or even use the same private sources. The team must identify the areas where threat sources overlap to avoid paying for the same thing more than once.
- **Individual feed context and scope**: Yes, certain PTI sources integrate OSINT and other private data, but a source can be made up of one or more feeds. In most cases, organizations pay per feed and not per source. And there could be a need for more feeds to cover the organization's security goals. Each feed category can look at a different aspect of the organization as different security departments might need various feeds. The higher the number of feeds, the higher the subscription fee as well.

Different vendors address PTI challenges differently, and the CTI team must assess each challenge (understand its impact on the CTI project) before investing in the PTI.

Some paid intelligence portals

There are many PTI feeds from top vendors that CTI can use to build intelligence. We are not providing an exhaustive list or giving the best sources here; instead, we cite some of the feeds that can help you move in the right direction. We are also not focusing on the vendor but the intelligence feed (even though the feed or service is linked to the vendor). Some paid threat intelligence feeds include AT&T *AlientVault*, *FireEye* CTI, *CrowdStrike Intelligence Exchange*, *RecordedFuture*, *HackSurfer*, Symantec *DeepSight*, *ThreatConnect*, IBM *X-Force*, *SecureWorks*, *Vipre*, *Kaspersky* TI, Microsoft *Graph Security*, Cisco *openVuln API*, *WildFire*, *Anomali*, *PhishLabs*, *DeCYFIR*, *Flashpoint Collab*, and many more.

Gartner (`https://gtnr.it/3fkEkMZ`) provides a good list of TI feeds, products, and services reviews that analysts and stakeholders can use as the primary source of vendor reconnaissance. Most of the TI feeds' vendors provide *APIs* to facilitate integration with SIEM and TIPs. They also provide intelligence products and services (TIP, SIEM, SOC, network monitoring, and more).

> **Important Tip**
>
> Selecting the feed is also trusting the vendor or provider, which might not be a simple task. However, when planning to purchase intelligence feeds, on top of the benefits, you should consider the vendor's reputation, their experience in the domain, use cases that have already been addressed, references, the roadmap, and user reviews. As a CTI analyst or security stakeholder, you should evaluate these six points.

The CTI team needs to research the source or vendor before purchasing or subscribing to data feeds. The CTI's responsibility is also to understand a vendor's security profile and ensure that the organization maximizes the solution. Regarding open and paid sources, the CTI analyst or team needs to know how to structure and store the raw and processed information. In the next section, we will look at intelligence data structuring and storing.

Intelligence data structuring and storing

This section assumes that you have selected the relevant data sources and feeds and have connected to them through APIs to get the data. At this point, you would like to organize and store this data efficiently and reliably. Data structuring and storing are related to how the data is presented and kept, respectively. Good *intelligence exploitation* heavily depends on how data is structured and stored.

CTI data structuring

Intelligence data must be structured so that it is easy to manipulate. Publicly accessed data is at everyone's mercy – including attackers. So, if there is any chance that they can compromise the data, they will take it. Hence, when collecting data, it is essential to have it presented securely in a trustworthy way. Structuring also involves some of the best practices for handling intelligence data. We will look at three points that must be considered when structuring intelligence data:

- **Maintain the CIA of the data**: Ensure that the CIA triad is kept during the entire cycle of data collection and manipulation. APIs must be secure because source feeds connect to the organization's internal network. Any breach in *confidentiality* can give attackers a window to compromise the APIs and launch malicious attacks. The CTI team must ensure that the collected or shared intelligence is not *altered* in transit or at rest. Reliable API keys must be used to ensure the integrity is maintained. The CTI team must also ensure that the data is available.
- **Maintain a standard format**: Ensure that collected intelligence and data (including reports and others) are using standard formats. This feature makes it easy to use the data by everyone familiar with standard formats. The team and the data must speak one language.
- **Share the data**: You might want to share your CTI reports (results) or processed data with the infosec community. Ensure that the shareable intelligence is secured and can easily be accessed by internal and external consumers (always consider access-level security when sharing information with external users).

When data is structured correctly, it becomes easy to store and share. The CTI team must strive to present intelligence in a language adapted to the target audience (*tactical* and *technical,* or *operational*). Intelligence data contains valuable information, but it needs to be structured and appropriately stored for you to benefit from it. In the next section, we will look at how intelligence can be stored.

CTI data storing requirements

Collected data and intelligence must be stored in the organization's environment to be accessible when needed. Intelligence is for everyone; hence, any security function should quickly use the data or information. When storing intelligence data, three metrics need to be considered:

- **Accessibility**: Security personnel, CTI analysts, and relevant stakeholders must be able to access the data or information at any time. The incident response team can use the data to detect, analyze, and remediate cyberattacks. The security operation team can access the data to perform system monitoring. The CTI team can still access the data for assessments and make informed decisions.
- **Availability**: Collected data must be stored so that it is available when needed. The organization must ensure that the storage infrastructure (hardware and software) is properly maintained and that all applications are functioning well and up to date. In worst-case scenarios, the storing infrastructure must be recovered promptly to avoid data loss.
- **Retrieval speed**: Intelligence data and information volume can be high. It falls under the big data category. Hence, the storage infrastructure must be designed to be queried quickly. Scanning through billions of indicator records can be annoying if the backend is not designed correctly.

These storage requirements are essential for reproducing and consuming intelligence within and outside the organization. When selecting feeds, the CTI team must evaluate the complexity of storing the data.

Intelligence data storing strategies

Companies store intelligence data and information in two different ways: *organization-specific storage infrastructure* and *threat intelligence platforms*. Large companies can combine both methods to ensure data *replication* and *archiving*. (This is one of the best ways to maintain high availability. In the case of a failure, one system is used as a backup for the other.)

Organization-specific storage infrastructure

An organization can build its storage infrastructure or use a cloud provider to host the data. Data that's been collected from feeds is parsed and stored internally in relational or non-relational databases. However, when building the infrastructure, the requirements (metrics) must match. Some of the approaches for storing the infrastructure used by most organizations include the following:

- **Relational databases**: Relational databases can be used to store intelligence data. However, relational systems such as native SQL can become bottlenecks in large data querying, hence the need to leverage non-relational storage and big data.
- **Non-relational databases**: An organization can use non-relational databases to store data. Document and graph-based are two examples of non-relational databases that the organization can use to store data.
- **Big data platforms**: An organization can leverage big data platforms such as Hadoop and Spark to deploy solid intelligence storage and facilitate data accessibility and querying through the use of common popular languages such as Python, SQL, and Java.

The organization must not ignore the challenges of building a data storage infrastructure from scratch or adopting a cloud solution for threat intelligence storage. Some of these challenges are as follows:

- **Infrastructure maintenance**: Although this does not apply to public cloud solutions, maintaining in-house databases comes at a cost. Acquiring cloud processing services (databases, **Extraction Transform Load – ETL** services, parsers, and others) is not free either.
- **Effort and efficiency**: The organization might need to acquire the skills required to build resilient mediation layers for parsing, correlating, and aggregating data from different sources and feeds. Mediation layers must be optimized to support fast processing and streaming capabilities.
- **Data formatting**: The intelligence output must be stored while following specific standards (such as STIX/TAXII, CSV, and JSON) to be shared with the infosec community. Therefore, the system must allow conversion from whatever internal format is used into standard formats.

The advantage of building an in-house storage infrastructure could relate to cost protection. It is likely to cost less to build a big data platform based on open source applications such as Hadoop and Spark than buying an expensive TIP. In terms of efficiency and effort, in-house data storage systems might not be ideal for small and medium businesses.

Threat intelligence platform for storing data

Probably the most popular choice (especially for level 1 and some level 2 organizations), TIPs can be used as intelligence storing infrastructure. This is justified by the fact that TIPs are built while following CTI standards (for example, they automatically support STIX/TAXII and YARA formats) and are ready for use. They also facilitate information sharing with the rest of the infosec community. Most paid TIPs are likely to provide storage mechanisms. However, open source TIPs such as MISP and **Collaborative Research Into Threats** (**CRITs**) also provide storage capabilities.

> **Important Note**
>
> For organizations that are new to CTI (level 1), I recommend saving your budget by leveraging open source TIPs such as MISP before diving into creating internal infrastructure or expensive TIPs. MISP is widely used because it provides several functionalities and is properly maintained. MISP is used throughout this book for practical use cases.

The CTI team and relevant stakeholders must select the correct strategies for storing intelligence data. When choosing one or both storage infrastructures, you must consider the requirements and how intelligence will be shared – CTI is a field that's growing due to the common cause of the community: fighting against cybercrimes.

Threat intelligence data feeds are passive in their original state. The CTI team or analyst must make them active by integrating them into existing security systems, TIPs, or **Security Information and Event Management** (**SIEM**). By doing so, TIPs or SIEMs can perform one of their principal tasks – correlating the external data feeds with the internal system data to provide more insights into threats.

Summary

Collecting the correct data for the CTI program is one of the most significant indicators of the success or failure of the program. The CTI team must evaluate all the selected sources and ensure that they match the requirements that were set during the planning phase. Whether you choose OSINT or PTI, it is essential to consider some parameters (budget, accuracy, data update frequency, data relevance, available resources, and business needs), as described in this chapter. Certain parameters might be negotiable for an organization or CTI team when selecting data sources, but others might not, depending on the objectives and requirements. The appropriate solution depends on the organization's needs, requirements, and resources. However, for small- and medium-sized organizations, OSINT can be sufficient to build intelligence. You should know how to select the correct sources, where to get the data, how to store it, and the prerequisites for integrating the data with other security platforms. In the next chapter, we will look at intelligence from the organization's perspective to effectively defend against threats and protect data.