OSTrICa – Open Source Threat Intelligence Collector
An Open Source plugin-oriented framework to collect and visualize Threat Intelligence Information

Roberto Sponchioni
Independent Security Researcher
Contact: rsponchioni@yahoo.it

Abstract— **Current approaches to protect sensitive data, such as Intrusion Detection Systems, Anti-Virus programs, traditional Incident Response methodologies by themselves are no longer enough to face today's relentless threats. Cybercrime used to be a hobby, now is highly organized and more financially driven. Organizations need a holistic view of the threat landscape to proactively fight a multitude of new threats that companies can face every day.**

SOC analysts, incident responders, attack investigators or cyber-security analysts need to correlate IoCs (Indicator of Compromise), network traffic patterns and any other collected data in order to get a real advantage against cyber-enemies. This is where threat intelligence comes into play, but unfortunately, not all the companies have enough budget to spend on Threat Intelligence Platform and Programs (TIPP); this is the main motivation behind OSTrICa's development. OSTrICa is a free and open source framework that allows everyone to automatically collect and visualize any sort of threat intelligence data harvested, from open, internal and commercial sources using a plugin based architecture. The collected intelligence can be analysed by analysts but it can also be visualized in a graph format, suitable for link analysis. The visualized information can be filtered dynamically and can show, for example, connections between multiple malware based on remote connections, file names, mutex and so on so forth.

I. INTRODUCTION

Cybercrime is highly organized, more financially driven and in many cases operating much like legitimate businesses, complete with organizational charts, C-level executives and even human resources departments. Cyber-threat actors are constantly improving their tools, techniques and procedures (TTP) to gain access to valuable companies' data. According to the "2015 Verizon Data Breach Investigations Report", in 60% of cases attackers are able to compromise organizations within minutes and 75% of the attacks spread from victim 0 to victim 1 within a day.

Cyber-attacks have changed and is extremely important to implement additional levels of protection to identify incidents, malicious events and link together attacks. Organizations need a holistic view of the threat landscape to proactively fight a multitude of new and dangerous threats that companies can face every day. Any security professional working in the field has to protect company's assets; and threat intelligence can help them in this important task. Correlating and linking together IoCs (Indicator of Compromise), network traffic patterns and any other harvested information can give analysts a real advantage against cyber-enemies, but unfortunately, not all the companies and researchers have enough budget to spend on Threat Intelligence Platform and Programs (TIPP). There are few options available for users:

- Commercial Option
- Free Option

Regarding commercial options, we can cite Maltego [1] or ThreatConnect [2] or Palatir [3], etc which all of them are powerful and well known in the industry. Unfortunately, all of them are more or less costly and not everyone can afford it (especially small enterprises or independent researchers).

Concerning instead free tools, which can easily be found online, most of them show blocked URLs or malware behaviour or botnet domains (botnet trackers), etc. Only few of them are capable of linking the information together: Maltego and ThreatCrowd [4]. Free version of Maltego has many drawbacks like:

- it is not for commercial use
- it has a maximum of 12 results for transform
- API keys expire every couple of days
- it runs on a server that is shared with all the community users affecting its performance
- it has no updates until a major version is released and it does not have any end user support
- it can only gather information from online Paterva servers

ThreatCrowd is a very nice and interesting online service and totally free, but it has its drawbacks too, such as:

- it can only link together some type of information (eg.: domains, URLs, MD5s, emails)
- it is an online service, which means it cannot be extended by the community with additional plugins
- it cannot be updated dynamically with new information collected during one investigation
- it does not have any filter to hide unnecessary information
- it does not clearly identify how the intelligence is linked together (eg.: it is not clear how an IP is associated to an MD5. Is the file downloaded from one IP or is the MD5 connecting to it?)

Instead, as said before, other free online tools/services cannot link all the information together, consequently they do not provide additional valuable information to the investigators.

In this paper, will be described how and why a new open source tool, named OSTrICa, has been developed. As discussed throughout the paper, the developed framework allows anyone to create a relevant and accurate threat profile, for free, based on all the collected information. This tool can help to identify and proactively protect company information by using multiple data sources (open, commercial and internal) as it is developed in a way that it allows analysts to add new modules in the form of plugins. Moreover, it can help during incident response and attack investigations since it can draw dynamically and automatically a graph containing all collected intelligence and links them together.

Analysts can play around with the visualized data by removing or filtering out unnecessary nodes and by identifying IoCs (Indicator of Compromise) source and destination nodes. The end goal of OSTrICa is to identify malicious or suspicious data inside the organization in order to proactively protect the company's assets, link the information together and potentially identify specific targeted attacks.

II. OSTrICa

OSTrICa stands for Open Source Threat Intelligence Collector and is a modular framework, plugin oriented and completely written in Python. The tool is divided in four modules:

- Command Line Interface
- Plugin Interface
- Report Interface
- Visual Report Interface

A. *Command Line Interface*

Analysts are provided with a command line interface which give them the possibility to perform different actions:

- Request information about domains
- Request information about IPs
- Request information about MD5/SHA256 files
- Request information about ASNs
- Request information about emails
- Show available plugins
- Generate a graph of all the collected intelligence. The graph can be updated dynamically every time the analyst needs it, with new intelligence data.

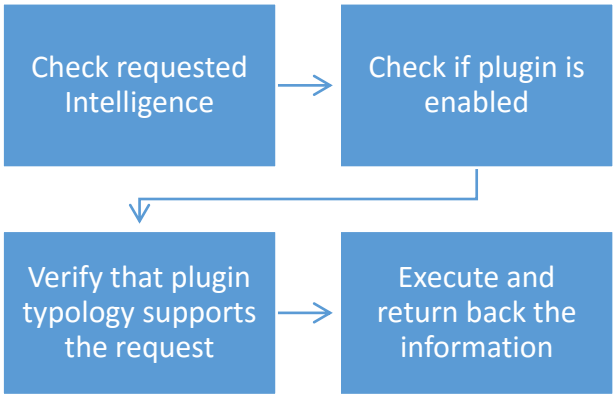
```
OSTrICa v.0.1 - Open Source Threat Intellicence Collector
Developed by: Roberto Sponchioni <rsponchioni@yahoo.it>
write "help" for help
> help
Following options are available

        domain - used to collect domains information
        Example: domain=google.com or domain=google.com,yahoo.com
        ip - used to collect IP information
        Example: ip=8.8.8.8 or ip=8.8.8.8,173.194.68.99
        md5 - used to collect MD5 information
        sha26 - used to collect SHA256 information
        asn - used to collect ASN information
        email - used to collect email information
        graph - generate a graph based on all the information collected
        gclean - clear graph information
        show - show all information that will be collected
        run - extract intelligece information
        help - this help
        plugins - show available plugins
```

B. *Plugin Interface*

OSTrICa comes already with a different set of plugins; but it has been developed with the aim to let analysts write their own components. Plugins are written in Python and each new module should be added in the "Plugin" directory within the tool directory. This will allow the framework to automatically load it in memory if needed.

Currently the configuration file, "cfg.py", contains the type of intelligence that can be collected: MD5, SHA256, email, domain, IP and ASN. Each plugin will be executed only if the typology (or typologies) associated with it matches the original request provided by the analyst via the command line interface.

For example, the "DeepViz Plugin" is only associated with the MD5 typology, so if the analyst will request information about one MD5, if enabled, that plugin will be executed and will return the collected data (if available). Instead, if a domain request is made by the analyst, "DeepViz Plugin" will never be executed.



Once the analyst requests information about an IoC (Indicator of Compromise), it goes inside a queue and one by one each requested information is popped out and delivered to the appropriate plugin that is able to handle it. The module that executes the plugins is called via 2 different directives: "run" and "graph" through the command line.

Once the analysts executes "run", if the typology is correct, OSTrICa will call the function named "run()" within the plugins source code and the plugin starts collecting the information and eventually it returns the collected data back.

Instead, "graph", will call the function named "data_visualization()", still within the plugin source code, and all the collected information previously harvested, will be parsed and two dictionaries of nodes and edges are filled with the information so that the intelligence data can be linked together and displayed through a graph.

A plugin just needs 2 functions "run()" and "data_visualization()" to be available, so that it can be called from the framework.

Current available plugins use different ways of collecting information such as via REST API or by scraping online resources. For example, DeepViz[5] plugin uses REST API to collect information, while instead VirusTotal [6] plugin uses web scraping [7] techniques to extract information (eg.: detections, behavioural information, IP and domain information, etc). But using web scraping, does not mean that

VirusTotal plugin cannot benefit of the power of VirusTotal Intelligence APIs. Since these APIs are not free and have a limited number of requests (based on the purchased option), I opted to use web scraping techniques (although in some cases cannot be used due to legal issues) so that the framework does not need any API key and there might be no limitations on the number of requests.

C. Report Interface

The report interface is another important part of OSTrICa, and it contains all the collected intelligence.

Every time a new information is collected by the plugins, whether they come from open source, internal or commercial sources, they are saved in a file (the name is randomly generated) under the "report" directory in JSON format so that the analyst can go back and read through it, in case any specific information is needed.

Each plugin returns a JSON in the following form.

TABLE I
REPORT FORMAT

```
{
   "plugin_name": "VT"
   "requested_intel": "fullmooncalendar.net"
   "requested_intel_type": "domain_infomation"
    "intelligence": { PLUGIN_DEPENDANT }
}
```

As per table above, "intelligence" is the dynamic part and can change from plugin to plugin and contains the actual intelligence returned by the plugin. For example, in the "VT" plugin, "intelligence" dictionary can contain detected URLs, based on the requested intelligence obviously, detection names, behavioural information and so on. It basically contains all the information that can be extracted by the plugin. Of course not all of them can be visualized in an easy way but, if needed, an analyst can look them up and decide to analyse the IoCs further and possibly block specific IPs or take additional steps. Table below explain a bit more into details what the four keys in the dictionary mean.

TABLE III
REPORT FIELD EXPLAINATION

Field name	**Description**
plugin_name	Name of the plugin that was called to collect some specific intelligence data
requested_intel	Original request made by the analyst
requested_intel_type	Typology of request made by the analyst. It can be anything from MD5, SHA256, email, domain, IP or ASN information
intelligence	Returned dictionary from the plugin. This field can be very different from plugin to plugin and it depends on the code written by the analyst.

As an example, following table shows a partial output of the VirusTotal plugin regarding a request made by the analyst about a specific MD5.

TABLE IIIII
PARTIAL VT REPORT

```
{
"plugin_name": "VT",
"requested_intel": "3a0d3a4cbed00926ad8c6d9a7f93e9d9",
"requested_intel_type": "md5"
"intelligence": {
        "extraction_type": "md5",
        "intelligence_information": {
          "av_results": {
            "ALYac": [
              "Trojan.Inject.BAY",
              "20160315"
            ],
            "AVG": [
              "FileCryptor.HAX",
              "20160315"
            ],
            "AVware": [
              "BehavesLike.Win32.Malware.wsc (mx-v)",
              "20160315"
            ],
[ ...REMOVED FOR SEMPLICITY ... ]
          "filenames": [
            "dump.bin",
            "locky_unpacked.ex$",
            "locky.unpacked"
          ],
          "first_submission_date": "2016-02-17 08:47:56",
          "last_submission_date": "2016-03-09 11:41:50",
          "threat_behaviour": {
            "copied_files": [
              "SRC:
C:\\2d6120701bd48c6395aa199211ebe5db01229ac48d98ea
da89da962769d05122\nDST:
C:\\DOCUME~1\\<USER>~1\\LOCALS~1\\Temp\\svchos
t.exe (successful)"
            ],
            "created_mutexes": [
              "RasPbFile (failed)"
            ],
[ ...REMOVED FOR SEMPLICITY ... ]
            "opened_mutexes": [
              "ShimCacheMutex (successful)",
              "RasPbFile (successful)"
            ],
[ ...REMOVED FOR SEMPLICITY ... ]
            "tcp_connections": [
              "195.154.241.208:80"
            ],
[ ...REMOVED FOR SEMPLICITY ... ]
```

D. Visual Report Interface

Visual Report Interface is another important, if not the most important, part of the framework that could be of help during an investigation. This component makes incredibly easy to visualize the relationships among domains, registrants, IP addresses, malware samples and all the collected data. This module can accelerate investigations by allowing analysts to visually explore and uncover connections in the harvested intelligence. The technology behind this capability is very powerful. The data are presented in an HTML page that can be automatically updated by the analyst with new IoCs. HTML pages are really powerful because without much hassle it is possible to generate powerful graph using Cytoscape [8], show visual appealing data with CSS [9] and JQuery [10]. The developed interface allows investigators to identify interesting nodes, move them around, search or remove and/or filter out unnecessary information.

A. Nodes connections

Thanks to this technology, for example, if an analyst clicks on a node, that specific intelligence will be highlighted with all the neighbours associate to it as per image below.

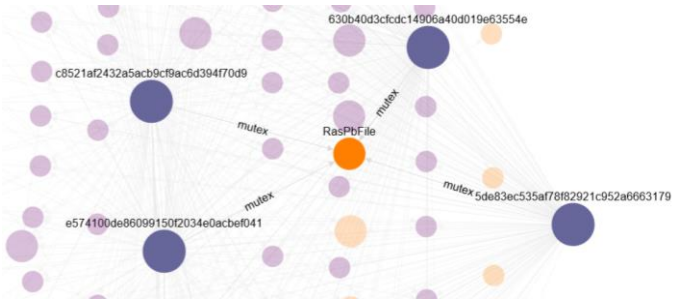
1 - Domain and IP connections

By clicking on the malicious domain "bpocpxyxwggt.me" it is possible to see what is associated to it, namely 4 different MD5s and one IP address. With this information the analyst can start writing a report where for example he can identify the command and control server, its IP address and possibly block them via firewall rules. Moreover, the figure clearly shows the link between the 4 MD5s, which confirm the fact that all of them are sharing similar characteristics such as the remote command and control server or maybe a mutex or a file name.

Another important characteristic of the visual module is that it is possible to identify how these nodes are all associated to each other by looking at the edges label. For example, in the picture above all MD5s are associated directly to the IP address "194.58.121.186", since all the samples are connected to it, and the IP address is associated to the domain "bpocpxyxwggt.me" as that domain resolved to that IP. Furthermore, the MD5s are also associated to that specific malicious domain since the threats directly connected to it.

B. Node sizes and colours

An additional interesting feature that has been implemented is the size and the colour of the nodes. For example, the original requested intelligence, which is the starting point, the data originally requested by the investigator, is specifically coloured (#666699) and its size is statically set to 60 pixels to easily identify those nodes. Instead, other nodes have a dynamic size which depends on how many time a specific data has been seen in the collection process. So, for example if a mutex name has been seen multiple times, because several malware used it, its node size will be bigger which means that more than one collected data is associated to it. For example, the image below shows that the mutex "RasFbFile" has been used by four different MD5s and hence its size will be bigger because it could mean that all four files are in relation to each other.



2 - Mutex size and connection

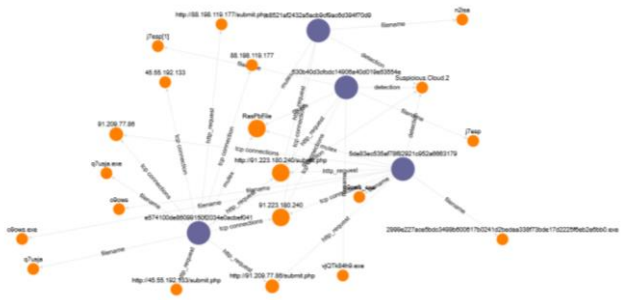
Regarding the colour instead, each new plugins has its own colour so that it becomes easier to identify specific information in the graph and possibly filter out unnecessary intelligence. For example, in the image above, the orange colour identify the VirusTotal plugin.

C. Filters and Search

Other important features that have been implemented are the filters and the search capability. Filters can hide unnecessary information that does not give any value to the graph and to the investigation. For example, if multiple plugins are executed the graph might contain lots of information that for a specific investigation might not be needed. Thanks to a multiple sets of filters, this superfluous information can be removed or hidden from the graph. The system implements three level of filters:

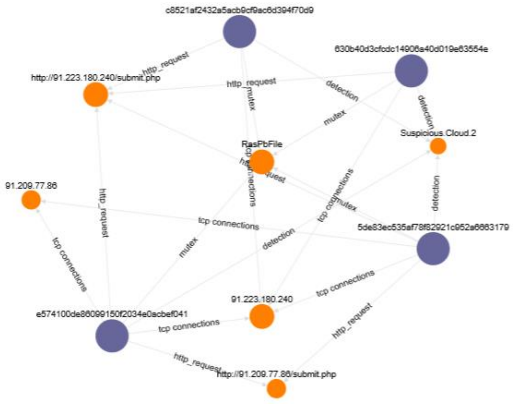
- *Plugin level filter*. The analyst has the ability to hide/show specific plugin related data so that only particular information can be seen.
- *Node level filter*. Nodes can be connected to each other with different edges and sometimes not all of them are needed, although all plugin data is still relevant. In this case investigators are capable of removing for example specific nodes associated to domains, detection names or file names and so on so forth.
- *Nodes connection filter*. Sometimes all plugin information is still relevant but all the collected data can be noisy and investigators want to see only nodes that are actually connected with the original intelligence or with nodes that are in turn connected with two or more other nodes. This is possible by "double clicking" on one node (like for example the original intelligence nodes) and automatically all the nodes that are not relevant will be hidden.

For example, figure 3 below, shows all the nodes in the graph. But it might be still a little bit noisy.



3 - All nodes in the graph

By "double clicking" on all the original intelligence nodes, the graph becomes more readable and all the associations result more clear as per image below.



4 - Interesting nodes

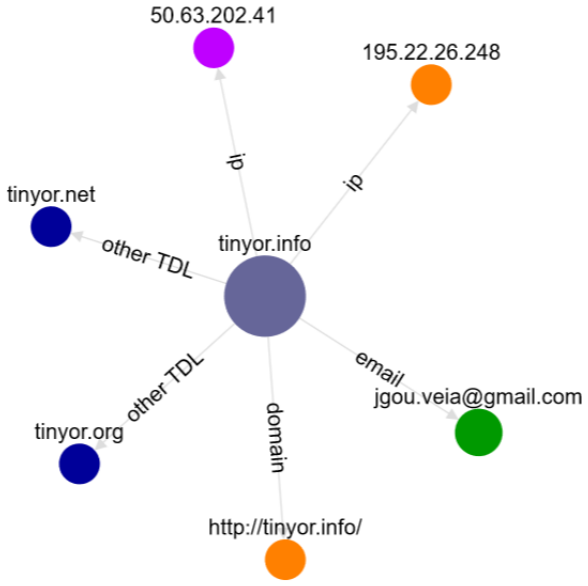
As evident, it is possible to see that all the original intelligence is associated with specific information such as same mutex name, same remote IP address or remote host. This picture clearly shows that the 4 malware are actually related and can be a good starting point to track the attack and find other possible malware in the infected network by looking for example at specific IPs and possibly find out the malicious campaign run behind it.

Furthermore, a *search option* has been implemented in order to help the investigators to find the information he needs inside the graph. By using the search box, analysts can identify the needed node since it gets automatically highlighted in the graph. This may help to track down links with specific intelligence that have been researched by analysts.

III. Scenario

In this last section a scenario* is described as part of an example investigation with the aim to find potential malicious domains that are worth to look into more details to possibly block attacks via the network firewall or IPS (Intrusion Prevention Systems). Assuming that during an internal investigation the investigation team came across a suspicious connection to a domain "tinyor.info". By running OSTrICa on that specific domain they were able to collect the following information:

- Current associated IP address
- IP address that was resolving to that domain back in March 2015
- Email address used to register that specific domain
- Additional details



5 - Details about tinyor.info

As part of the ongoing investigation, the team decided to collect additional details about the IP address "195.22.26.248" from both their firewalls and by using OSTrICa. The logs showed that a specific file tried to connect to that IP. That specific file was identified by the following MD5 "747b3fd525de1af0a56985aa29779b86". After the analysis they identified the sample to be malicious and started collecting intelligence about it as well. The produced report is very interesting since the team was able to:

- Identify additional domains linked to the IP address
- Identify additional domains and IP linked to the malware "747b3fd525de1af0a56985aa29779b86"
- Identify the malware type, Trojan.Bayrob!gen6, which is known to be used to steal confidential data
- Identify Portugal as the country associated to the malicious IP
- Identify newly created processes and files via the textual report

All the above information can be helpful to the analysts to block suspicious domains, IPs and scan the company's machines to find artefacts related to the identified malware to block the attack in a timely manner.

Another interesting piece of intelligence that can be useful to the investigators is the email address "jgou.veia@gmail.com" which was originally associated with "tinyor.info". By running another OSTRiCa query, investigators were able to collect further domains registered with this email. Example of these domains are: "earjinnmqsk.info", "b7d3f7f191b16f0cd7a1997cd90f1986.info", "badybayxdlnmzhofdymnbmup.info" and so on. The team was then able to identify those domains as malicious as part of another attacks and, consequently, by promptly blocking these domains they were able to protect the company's assets.

* Note: The scenario is not associated to any real investigation. It was taken as an example to show how OSTRiCa could be used.

IV. Conclusion

The developed tool is in its beta stage but I think that it is very powerful and already able to collect relevant information that can be used during Incident Response or attacks' investigation by companies or independent security researchers. Since it is in its early stage of development, this tool can be improved, for example, in the following areas:

- *Visualization*: improving the filtering options
- *Visualization*: importing timeline details. Some of the collected information, like for example domain expiry date, malware first seen date
- *Visualization*: improve the graph update capability in a way that all the removed nodes will no longer be available once the page gets refreshed or a new intelligence is added in the graph
- *Visualization*: improve the way the graph is generated and make it more appealing

References

[1] Maltego, https://paterva.com/web6/products/download2.php
[2] Threat connect, https://www.threatconnect.com
[3] Palantir, https://www.palantir.com
[4] ThreatCrowd, https://www.threatcrowd.org
[5] DeepViz, https://www.deepviz.com
[6] VirusTotal, https://www.virustotal.com
[7] Web Scraping, https://en.wikipedia.org/wiki/Web_scraping
[8] Cytoscape, http://js.cytoscape.org
[9] CSS, http://www.w3schools.com/css
[10] JQuery, https://jquery.com