APT Group Attribution

Paul-David Jarvis
Faculty of Science & Technology
Bournemouth University
Bournemouth, Dorset
s5115232@bournemouth.ac.uk

***Abstract*—A paper detailing why the collection of data and analysing the attributes of APT groups and their attacks.**

Keywords— Advanced Persistent Threat (APT), Indicators of Compromise (IoC), Digital Threats: Research and Practice (DTRAP), Computer Security Incident Response Team (CSIRT), National Electronics and Computer Technology Center (NECTEC), Electronic Transactions Development Agency (ETDA), Internet Service Provider (ISP), Forum of Incident Response and Security Teams (FIRST), Asia Pacific Computer Emergency Response Team (APCERT), Command and Control (C2), Distributed Denial of Service (DDoS)

I. Introduction

APT groups have sought to create the deadliest cyber weapon to achieve their cause, whether it be to destroy or to steal. The focus of this paper is to study APT groups and the malicious software that they used in their campaigns against their targets such as nation-states or financial institutions. We will be studying and collecting an enormous amount of data relating to the groups and providing a technical view on the methods they employ using reputable engines and data manipulation software to analyse the hashes associated to their respective IOCs. We will be looking at many attributions of an APT group such as their origin, sectors they target and the countries that they attack. Anyone in the field of information technology and more especially cyber security and cyber threat intelligence would be interested in reading and knowing what we will discuss further down.

II. APT Groups

APT groups are considered the most dangerous and most sophisticated attackers by most people in the cybersecurity field ranging from companies and institutions to individual researchers. Their main goals are usually theft and espionage including intellectual property, classified data, and personally identifiable information. They also partake in sabotage including deleting important databases and completely taking over their targets website.

The reasons that they're so dangerous is because these threat actors are usually backed by the resources of nation states and corporations; some including APT 28 also known as Fancy Bear and APT 34 also known as OilRig. The most dangerous thing about the attacks that these APT groups carry out is that their main objective is to get a foothold access into their target and maintain their secretiveness so they're able to collect intelligence over a long period of time. They're usually knowledgeable so their tactics and techniques are usually very sophisticated including dropping rootkits, DNS tunnelling and social engineering. They are also very meticulous with mapping out all potential weaknesses in a system to attack the efficient weakness possible.

A. Stages of Attack

There are five stages to an attack orchestrated by an APT group. Firstly, they need to gain initial access to their targets which is done by attacking either a web-based system, networks, or employees. Secondly, once they gain access, the APT group but maintain their access so they deploy malware with the goal of creating a backdoor into their targets system. They now begin to map the network to expand their access and begin to find ways to escalate their privileges. They also may attempt to move laterally across networks by creating tunnels across networks. When they have the necessary access and privileges, they now need to get ready, so they begin to stage the account by readying the target data they extracted. Now that an APT group is ready, they will launch their attack by quietly extracting the stolen data outside the network to a C2 server, they might cause a distraction while they do this to draw the attention of the I.T department away usually by launching a DDoS attack. Finally, the APT group will follow up their attack depending on their main goal, they might cause as must destruction and damage they can by either destroying or taking over critical assets such as data centres and websites or they might choose to stay silent in attempt to not get caught and maintain access to their target as well as dropping other back doors in case they are caught and need to regain access. [7]

III. hypotheses

These reasons are why it is vital to know and understand the behavior of APT groups and the methods, tactics, and techniques that they employ to properly mitigate and counter their attacks. As the project moves forward so will our hypotheses and ideas that we formulate will also evolve, they will all be surrounding the topic of how dangerous an APT group can be.

IV. AIMS AND OBJECTIVES

This project aims to conduct an empirical study on APT groups and the malware they used as well as to attempt to provide a look into the attributions of our APT groups. The project objectives are:

- Collect enough data to allow us to provide a thorough investigation into APT groups and the malware they used by scraping and storing the data we collected.

- Develop our Jupyter notebook files to include our collection and gathering methods from APIs that we use as well as the data analysis code to help investigate our hypotheses.

A. *Success Critera*

- The code that we write in Jupyter notebooks must compile correctly and return the correct and accurate data up to date from Virus Total. They also must read in the correct data sets and create the necessary data frames that are later joined together.
- The relationship data we collect on our samples of hashes associated to our APT group's malicious files must contain enough data to analyse. For example, some of the samples have no parent files, IP addresses or domains that they contacted associated with them. This limitation means that we are unable to provide a deeper look into this specific hash.
- Another problem that we ran into was the number of requests I was able to make using the Virus Total API I was given. Due to the sheer amount of data we will collect, we will also be making a similar amount of requests and to counter this issue I had to contact Virus Total and request that my API key was made into an academic key allowing me to make requests with a daily quota of 20 thousand and a monthly quota of 1 billion requests.

V. DIGITAL THREATS: RESEARCH AND PRACTICE

This idea came originally from a special call for papers from a reputable online computer science website called Association for Computing Machinery and one of their journals Digital Threats: Research and Practice that targets the prevention, identification, mitigation, and elimination of digital threats. DTRAP aims to bridge the gap between academic research and industry practice. Accordingly, the journal welcomes manuscripts that address extant digital threats, rather than laboratory models of potential threats, and presents reproducible results pertaining to real-world threats [5]. They were looking for researchers in the field to submit papers based on malware and one of their suggested ideas were APT groups and malware lineage. So, I considered a paper where I combined both ideas and looked at malware that APT groups have used.

VI. LITERATURE REVIEW

A. *VX-Underground*

VX-Underground is an online website that hosts the largest collection of malware source code, samples, and papers on the internet. They currently have over 3 million samples on their website with a goal of 26 million. They have a range of sections but the most important one for our project is their APT collection that has their collection of samples separated by years and split up into groups of their respective published papers [2]. VX-Underground will be the main source of our data using only the recent years of APTs to provide as much of an up-to-date view. VX-Underground is an amazing source of APTs and malware hashes, but the main issue comes down to the integrity and liability of the website. They are an individual website with no connections to any government or academic institute and have caught some backlash from other companies about providing malware openly to anyone.

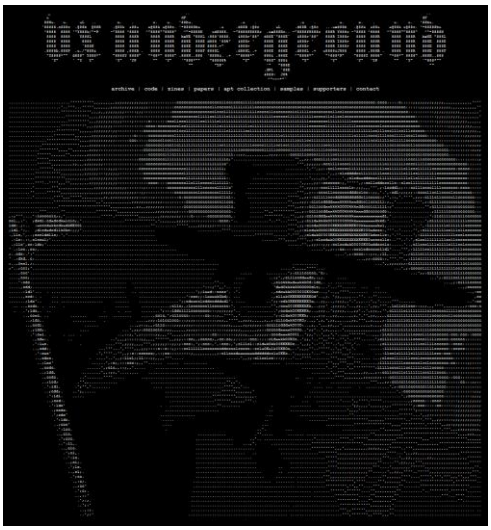


Fig. 1. VX-Underground [2].

B. *VirusTotal*

VirusTotal is a free to use analysis engine that security researchers can analyse files and URLs for viruses, worms, trojans and other kinds of malicious content. Their goal is to make the internet a safer place through collaboration between members of the antivirus industry, researchers, and end users of all kinds. Fortune 500 companies, governments and leading security companies are all part of the VirusTotal community, which has grown to over 500,000 registered users [1]. We will be using VirusTotal's API to automate the analysis of our collection of malware sample hashes from VX-Underground. The type of data that VirusTotal will provide is useful chronological data such as when the malware was first seen on the web, first and lastly submitted as well as other extremely useful data relating to a large range of security vendors labelling the malware hash as either malicious, undetected, or unable to process the file type.

The type of information about our hashes will first be directly relating to the file such as its SHA256 hash, what the file is called, the type of file that it is, how many times the file has been submitted, how many vendors have detected it as malicious, the first and last time it was submitted, the domains that that it connect to, the IP addresses that it connect to and lastly the files that drop these this file. We can then further enrich this data by analysing the IP addresses to try and pinpoint a geolocation such as the country it originates from. We can also analyse who the owner of the server is to help us narrow down any common patterns.

Fig. 2. VirusTotal [1]

C. *ThaiCERT*

ThaiCERT is the CSIRT for Thailand and provides an official point of contact for dealing with computer security incidents in the Thai Internet community. Founded by NECTEC under the Ministry of Science and Technology, ThaiCERT has been the first and only non-profit CSIRT in Thailand. In February 2011, by the resolution of the Thai cabinet, ThaiCERT operations were transferred to a new administrative team in a new public organization named ETDA under the supervision of the Ministry of Information and Communication Technology. ThaiCERT collaborates with the Thai government sector, organizations, universities, ISPs, and other relevant entities to handle computer security incidents in Thailand. Additionally, as a full and active member of FIRST and APCERT, ThaiCERT coordinates with both globally and regionally trusted CSIRTs in responding to computer security incidents [3]

They are an incredible source for technical information for our APT groups as they provide an open-source library where we can find our respective APT groups as well as other names from them, names, and other sources for their attacks as well as the tools that they used in their campaigns. Considering the reliability of the source, as previously mentioned ThaiCERT collaborates with government, educational institutions, and tech companies so I believe that there are a credible source and the data we gathered from them could be use accurately without bias.

Fig. 3. ThaiCERT [3].

D. *Project Jupyter*

Project Jupyter are the developers of the jupyter notebook. An open-source web application that allows you to create and share documents that contain live code, equations, visualisations, and narrative text. Uses include data cleaning and transformation, numerical simulation, statistical modelling, data visualization, machine learning, and much more [4]. We will be using jupyter notebook to import all our data we collect and to analyse and visualise our data to achieve our goals in this project. With the use of the API key that VirusTotal gave us access to, we can write code python that uses the key and calls commands within that gives us the data we are looking for, For Example: the parent file that created our file, the IP addresses, and domains that our file connected too. With all this information we can enrich our data to see the lineage of our hash.

Fig. 4. Our Project Jupyter Work [4].

E. *Analysing the fall 2020 Emotet campaign*

Throughout writing this paper and this project, we noticed that there is another paper that has a small overlap with our project. "Analysing the fall 2020 Emotet campaign" written by Constantinos Patakis and Anargyros Chrysanthou at the University of Piraeus in Athens, Greece. Their work is considerably in the same area of research whereas their work is mainly analysing Emotet; the banking trojan. Hopefully while this project advances forward, we can reach out to Constantinos and Anargyros to discuss ideas and our research. They have created a data set that includes email headers, documents, executables, and domains; This data set will include similar pieces of information that we will collect on APT groups [6].

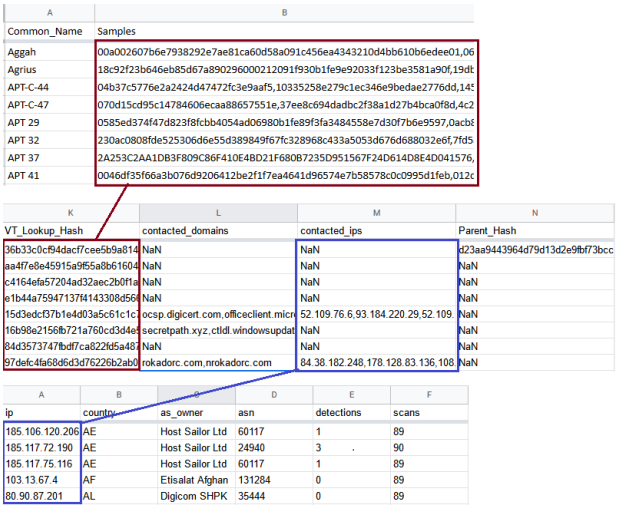
They have also used Project Jupyter and their notebooks to perform data analysis on their data set and results. One of most interesting pieces of manipulated they did were using the seaborn library to plot all their data about C2 servers on heat maps allowing readers to get a quick and easy view on where these servers are located the most. We will also be considering doing something using Seaborn library and heat maps to help portray the locations where servers are that are being contacted by our malicious files. Inferences could be made that the servers located in these countries are those that are being targeted by our APT groups or the locations of where the information is being sent to by our APT groups.

Fig. 5. A Heatmap from the Emotet Paper [6].

VII. Methodology

A. Data Set

The data set that we will create will be the main artefact for this project in Excel. I am hoping to have two main spreadsheets: one for samples and one for APT groups. These two will be linked through the samples and they will both hold additional pieces of information that will be linked to other spreadsheets. For Example, the samples sheet has a column for IP addresses and domains that the samples contact, and these will be used to join to other spreadsheets dedicated to information on those IP addresses and domains. Figure 6 shows the above example, you can see clearly how my data sets can be linked together to provide a more detailed analysis.



Common_Name	Samples
Aggah	00a002607b6e7938292e7ae81ca60d58a091c456ea4343210d4bb610b6edee01,06
Agrius	18c92f23b646eb85d67a8902960002120 91f930b1fe9e92033f123be3581a90f,19dt
APT-C-44	04b37c5776e2a2424d47472fc3e9aaf5,10335258e279c1ec346e9bedae2776dd,145
APT-C-47	070d15cd95c14784606ecaa88657551e,37ee8c694dadbc2f38a1d27b4bca0f8d,4c2
APT 29	0585ed374f47d823f8fcbb4054ad06980b1fe89f3fa3484558e7d30f7b6e9597,0acb8
APT 32	230ac0808fde525306d6e55d389849f67fc328968c433a5053d676d688032e6f,7fd5
APT 37	2A253C2AA1DB3F809C86F410E4BD21F680B7235D951567F24D614D8E4D041576,
APT 41	0046df35f66a3b076d9206412be2f1f7ea4641d96574e7b58578c0c0995d1feb,012c

VT_Lookup_Hash	contacted_domains	contacted_ips	Parent_Hash
36b33c0cf94dacf7cee5b9a814	NaN	NaN	d23aa9443964d79d13d2e9fbf73bcc
aa4f7e8e45915a9f55a8b61604	NaN	NaN	NaN
c4164efa57204ad32aec2b0f1a	NaN	NaN	NaN
e1b44a7594713 7f4143308d56	NaN	NaN	NaN
15d3edcf37b1e4d03a5c61c1c7	ocsp.digicert.com,officeclient.micr	52.109.76.6,93.184.220.29,52.109.	NaN
16b98e2156fb721a760cd3d4e5	secretpath.xyz, ctldl.windowsupdat	NaN	NaN
84d3573747fbdf7ca822fd5a48	NaN	NaN	NaN
97defc4fa68d6d3d76226b2ab0	rokadorc.com,nrokadorc.com	84.38.182.248,178.128.83.136,108.	NaN

ip	country	as_owner	asn	detections	scans
185.106.120.206	AE	Host Sailor Ltd	60117	1	89
185.117.72.190	AE	Host Sailor Ltd	24940	3	90
185.117.75.116	AE	Host Sailor Ltd	60117	1	89
103.13.67.4	AF	Etisalat Afghan	131284	0	89
80.90.87.201	AL	Digicom SHPK	35444	0	89

Fig. 6. Some of our data sets showing connections.

References

[1] Virustotal.com. 2004. VirusTotal. [online] Available at: <https://www.virustotal.com/> [Accessed 4 November 2021].

[2] Vx-underground.org. n.d. vx-underground - home. [online] Available at: <https://vx-underground.org/> [Accessed 4 November 2021].

[3] Thaicert.or.th. 2000. ThaiCERT. [online] Available at: <https://www.thaicert.or.th/about-en.html> [Accessed 4 November 2021].

[4] Jupyter.org. 2014. Project Jupyter. [online] Available at: <https://jupyter.org/> [Accessed 9 November 2021].

[5] Dl.acm.org. n.d. DTRAP Home. [online] Available at: <https://dl.acm.org/journal/dtrap> [Accessed 26 November 2021].

[6] Patsakis, C. and Chrysanthou, A., 2020. Analysing the fall 2020 Emotet campaign. [online] arXiv.org. Available at: <https://arxiv.org/abs/2011.06479> [Accessed 27 November 2021].

[7] Cynet. 2020. Advanced Persistent Threat (APT) Attacks. [online] Available at: <https://www.cynet.com/network-attacks/advanced-persistent-threat-apt-attacks/> [Accessed 2 December 2021].