



APT Sandworm Dataset – Readme File

Fundación Tecnalia Research & Innovation - <https://www.tecnalia.com/>

University of Western Macedonia - ITHACA Lab - <https://ithaca.ece.uowm.gr/>

Public Power Corporation S.A. - <https://www.ppcgroup.com/>

Authors: Eider Iturbe, Christos Dalamagkas, Panagiotis Radoglou-Grammatikis, Erkuden Rios

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101070450 (AI4CYBER). Disclaimer: Funded by the European Union. Views and opinions expressed are, however, those of the authors only and do not necessarily reflect those of the European Union or European Commission. Neither the European Union nor the European Commission can be held responsible for them.



1. Introduction

Due to the rise in Advanced Persistent Threat (APT) attacks, modern digital systems and particularly critical infrastructures, must be equipped to counteract these stealthy, multi-stage intrusions. Unlike simple, isolated attack techniques, APTs are orchestrated by highly skilled adversarial groups that follow strategic, logic-driven plans. These groups employ specific tactics, techniques, and procedures (TTPs) to silently infiltrate systems, perform lateral movements across assets, and ultimately execute impactful actions once their objectives are reached.

To effectively study and defend against such threats, there is a pressing need for datasets that accurately reflect the complexity and sequential nature of real-world APT campaigns.

This dataset introduces the APT Sandworm dataset. Sandworm [1] has been active since at least 2015, notably using the BlackEnergy malware to target Ukraine's energy infrastructure. The group gained further notoriety with the NotPetya attack in 2017, which caused widespread disruption and significant financial losses, particularly in the logistics, shipping, and manufacturing sectors. In 2022, Sandworm was again linked to a cyberattack on Ukraine's electric power infrastructure. Their consistent focus on energy-related targets underscores a persistent and strategic intent to disrupt critical infrastructure systems.

2. Dataset Overview

The dataset is organized as follows:

- README file (this document, APT_Dataset_Readme.pdf): Provides a detailed description of the testbed infrastructure, which emulates a realistic critical infrastructure environment under attack. It also outlines the APT attack scenario and the key features of the dataset.
- PCAP file (SandwormAPT.pcap): Contains the raw network traffic captured during the execution of the APT emulation. This data reflects the observable network-level behaviour of the attack.
- Network flow dataset (SandwormAPT_flow_labelled.csv): Includes labelled network flow records corresponding to the attack procedures that are visible in the captured traffic. Only the steps of the APT campaign that generate detectable network activity are labelled in this file. Other attack stages, such as those occurring at the endpoint or system level, are not represented in the PCAP or flow data due to the lack of observable network evidence.

3. Testbed

Figure 1 depicts the topology of the testbed setup, which aims to replicate a Wide Area Measurement System (WAMS). The testbed was used to produce the dataset and was developed at the premises of the PPC Inspectra Industrial Internet of Things (IIoT) Laboratory and the TECNALIA Research Cloud.

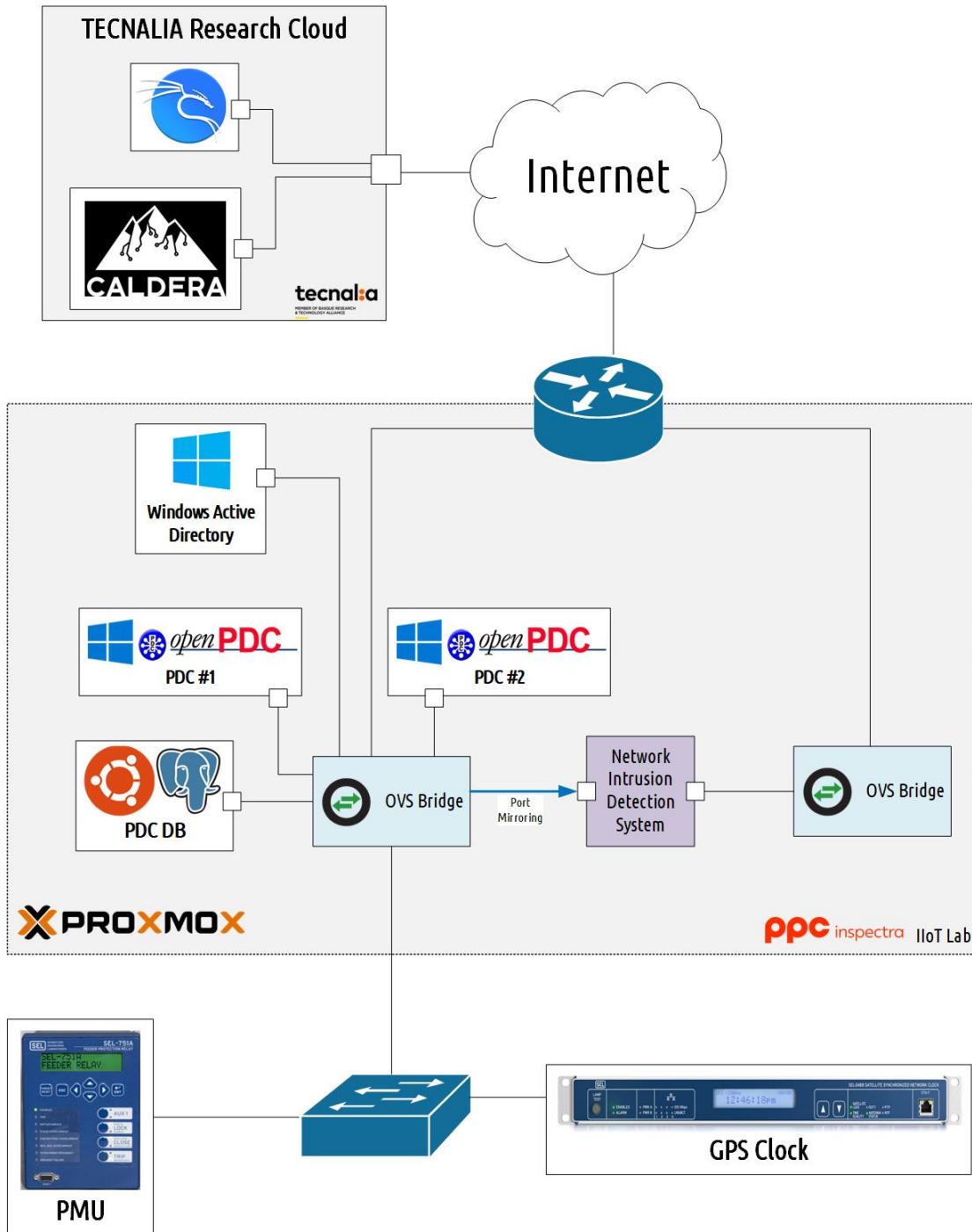


Figure 1: Topology of the testbed utilized to produce the dataset

Table 1 provides further information about each asset of the testbed as well as the IP address of each asset.

Table 1: Testbed Assets

Name	Description	IP Address
PMU	The Phasor Measurement Unit (PMU) is an industrial device that transmits synchrophasor measurements to the PDCs.	192.168.21.206
GPS Clock	This device acts as a Precision Time Protocol (PTP) grandmaster clock, providing high-precision time synchronisation to the PMU. For this purpose, the PTPv2 over Ethernet protocol is used, therefore the asset does not use IP addressing.	N/A
PDC #1	The Phasor Data Concentrator (PDC) is a software that collects the synchrophasor measurements from the PMU. It is based on the openPDC open-source software and is hosted on a Windows 10 Virtual Machine (VM).	192.168.21.230
PDC #2	A second PDC VM is a replica of PDC #1, and it is used for redundancy, in case PDC #1 becomes unavailable.	192.168.21.229
PDC DB	The database (DB) is used by the PDC VMs to store the synchrophasor measurements.	192.168.21.231
Windows Active Directory (WAD)	The Windows Active Directory (AD) provides a Windows domain network for the PDC VMs, allowing the central administration of users, roles and permissions for the Windows VMs.	192.168.21.233
Network Intrusion Detection System	This VM receives mirrored traffic from the Windows AD, PDC #1, PDC #2 and the PDC DB.	N/A
Caldera Mitre	The Caldera Mitre platform is used to perform the multi-step attack against the WAMS components.	192.168.126.176
Kali Linux	The Kali Linux VM is used complementary to perform the attacks against the WAMS components. It works as C&C server used by Caldera.	192.168.126.20

4. APT Sandworm attack

This dataset introduces the APT Sandworm dataset, which emulates the behaviour of the Sandworm threat group using MITRE’s open-source attack emulation framework, Caldera [2], as outlined by the Center for Threat-Informed Defense in their Attack Emulation Library [3].

The targeted system is described in Section 3. The initial compromise occurred on a Linux-based PDC DB system, which served as the entry point for the attack. From there, the adversary employed lateral movement tactics to infiltrate a second device, PDC #1, which was Windows-based. Ultimately, the attacker reached the Active Directory server, the final target, using similar techniques.

A detailed list of the executed Tactics, Techniques, and Procedures (TTPs) is provided in the table below. Each entry includes:

- The start and end time of the attack step
- The host involved
- A description of the TTP, as specified in the emulation plan defined by Center for Threat-Informed Defense’s Emulation Library [4].
- A custom label assigned by the authors to represent specific attack procedures.

The labelling process was based on an analysis of Caldera logs generated during the emulated attack. These logs were correlated with network flow data to identify which attack steps resulted in observable network activity. Labels were assigned only to those attack procedures that were reflected in the network flows. These include:

- ssh_intrusion: Unauthorized access attempts via the Secure Shell (SSH) protocol
- php_insecure_intrusion: Exploitation of insecure PHP-based web applications
- smb_intrusion: Abuse of the Server Message Block (SMB) protocol to access shared resources or perform lateral movement
- rdp_intrusion: Unauthorized access attempts via the Remote Desktop Protocol (RDP)
- remote_system_discovery: Reconnaissance activities aimed at identifying remote systems and services within the network

Table 2. APT Sandworm timeline and attack techniques

Start time	End time	Host	Attack technique
2025-02-18T11:05:59Z	2025-02-18T11:06:05Z	kali	Sandworm Initial Compromise - Transfer Webshell (T1105-Ingress Tool Transfer)
2025-02-18T11:06:09Z	2025-02-18T11:06:20Z	kali	Sandworm Initial Compromise - Deploy Webshell (T1505.003-Server Software Component: Web Shell)
2025-02-18T11:06:24Z	2025-02-18T11:06:29Z	kali	Sandworm Initial Discovery via Webshell - Whoami (T1033-System Owner/User Discovery)
2025-02-18T11:06:34Z	2025-02-18T11:06:41Z	kali	Sandworm Initial Discovery via Webshell - Uname (T1082-System Information Discovery)

2025-02-18T11:06:44Z	2025-02-18T11:07:14Z	kali	Sandworm Initial Discovery via Webshell - ls (T1083-File and Directory Discovery)
2025-02-18T11:07:19Z	2025-02-18T11:07:25Z	kali	Sandworm Initial Discovery via Webshell - cat /etc/passwd (T1003.008-OS Credential Dumping: /etc/passwd and /etc/shadow)
2025-02-18T11:07:29Z	2025-02-18T11:07:47Z	kali	Sandworm Download Agent Executable (T1105-Ingress Tool Transfer)
2025-02-18T11:07:49Z	2025-02-18T11:07:55Z	kali	Sandworm Set Up SUID for Agent Execution (T1548.001-Abuse Elevation Control Mechanism: Setuid and Setgid)
2025-02-18T11:07:59Z	2025-02-18T11:08:05Z	kali	Sandworm Execute Agent via SUID Binary (T1548.001-Abuse Elevation Control Mechanism: Setuid and Setgid)
2025-02-18T11:08:09Z	2025-02-18T11:08:57Z	kali	Sandworm Wait For Priv Esc Agent Beacon (T1548.001-Abuse Elevation Control Mechanism: Setuid and Setgid)
2025-02-18T11:09:00Z	2025-02-18T11:09:05Z	PDC DB	Sandworm Establish Crontab Persistence (T1053.003-Scheduled Task/Job: Cron)
2025-02-18T11:09:10Z	2025-02-18T11:09:15Z	PDC DB	Sandworm Establish Systemd Persistence (T1543.002-Create or Modify System Process: Systemd Service)
2025-02-18T11:09:20Z	2025-02-18T11:09:30Z	PDC DB	Sandworm Read /etc/shadow (T1003.008-OS Credential Dumping: /etc/passwd and /etc/shadow)
2025-02-18T11:09:35Z	2025-02-18T11:09:44Z	PDC DB	Sandworm Get Bash History Modified (T1552.003-Unsecured Credentials: Bash History)
2025-02-18T11:09:45Z	2025-02-18T11:09:56Z	PDC DB	Sandworm Get SSH Keys Modified (T1552.004-Unsecured Credentials: Private Keys)
2025-02-18T11:10:00Z	2025-02-18T11:10:09Z	kali	Sandworm Upload Agent to Windows Host Modified (T1021.002-Remote Services: SMB/Windows Admin Shares)
2025-02-18T11:10:10Z	2025-02-18T11:10:26Z	kali	Sandworm Service Creation/Execution and Registry Persistence Modified (T1569.002-System Services: Service Execution)
2025-02-18T11:10:30Z	2025-02-18T11:10:33Z	kali	Sandworm RDP to Execute Agent Implant Modified (T1547.001-Boot or Logon Autostart Execution: Registry Run Keys / Startup Folder)
2025-02-18T11:10:35Z	2025-02-18T11:11:27Z	kali	Sandworm Wait For Lat Move Target Beacon (T1547.001-Boot or Logon Autostart Execution: Registry Run Keys / Startup Folder)
2025-02-18T11:11:30Z	2025-02-18T11:11:34Z	PDC#1	Sandworm Windows Discovery - Current User (T1033-System Owner/User Discovery)
2025-02-18T11:11:35Z	2025-02-18T11:11:50Z	PDC#1	Sandworm Windows Discovery - Windows Version Info (T1082-System Information Discovery)

2025-02-18T11:11:50Z	2025-02-18T11:15:01Z	PDC#1	Sandworm Windows Discovery - List Entire File System (T1083-File and Directory Discovery)
2025-02-18T11:15:05Z	2025-02-18T11:15:10Z	PDC#1	Sandworm Windows Discovery - List RDP Connections (T1049-System Network Connections Discovery)
2025-02-18T11:15:15Z	2025-02-18T11:15:20Z	PDC#1	Sandworm Download Web Credential Dumper Modified (T1105-Ingress Tool Transfer)
2025-02-18T11:15:26Z	2025-02-18T11:15:35Z	PDC#1	Sandworm Execute Web Credential Dumper (T1555.003-Credentials from Password Stores: Credentials from Web Browsers)
2025-02-18T11:15:36Z	2025-02-18T11:15:50Z	PDC#1	Sandworm Download Keylogger (T1105-Ingress Tool Transfer)
2025-02-18T11:15:51Z	2025-02-18T11:16:07Z	PDC#1	Sandworm Log Keystrokes Modified (T1056.001-Input Capture: Keylogging)
2025-02-18T11:16:11Z	2025-02-18T11:16:42Z	kali	Sandworm Let Keystroke Logger Run (T1056.001-Input Capture: Keylogging)
2025-02-18T11:16:46Z	2025-02-18T11:16:50Z	PDC#1	Sandworm Verify Keystroke File (T1056.001-Input Capture: Keylogging)
2025-02-18T11:16:51Z	2025-02-18T11:17:06Z	PDC#1	Sandworm Upload Keystroke File (T1041-Exfiltration Over C2 Channel)
2025-02-18T11:17:11Z	2025-02-18T11:17:21Z	PDC#1	Sandworm Stop Keystroke Logger (T1056.001-Input Capture: Keylogging)
2025-02-18T11:17:26Z	2025-02-18T11:17:38Z	PDC#1	Sandworm Remote System Discovery (dsquery) (T1018-Remote System Discovery)
2025-02-18T11:17:41Z	2025-02-18T11:17:48Z	PDC#1	Sandworm Cleanup Artifacts (T1070.004-Indicator Removal on Host: File Deletion)
2025-02-18T11:17:51Z	2025-02-18T11:17:55Z	kali	Terminate RDP Session to Second Target (Gammu) (T1021.001-Remote Services: Remote Desktop Protocol)
2025-02-18T11:18:16Z	2025-02-18T11:18:27Z	kali	Upload Agent Executable to Arrakis via SMB (T1105-Ingress Tool Transfer)
2025-02-18T11:18:31Z	2025-02-18T11:18:35Z	kali	Lateral Movement Via RDP (Arrakis) Modified (T1021.001-Remote Services: Remote Desktop Protocol)
2025-02-18T11:18:36Z	2025-02-18T11:19:02Z	kali	Wait for Beacon Arrakis (T1493.003: Virtualization/Sandbox Evasion: Time-based Evasion)
2025-02-18T11:19:06Z	2025-02-18T11:19:15Z	WAD	Sandworm Download NotPetya (T1105-Ingress Tool Transfer)
2025-02-18T11:19:16Z	2025-02-18T11:22:40Z	WAD	Sandworm Executes NotPetya (perf.dat) (T1486-Data Encrypted for Impact)
2025-02-18T11:22:41Z	2025-02-18T11:22:50Z	kali	Terminate RDP Session to Domain Controller (Arrakis) (T1021.001-Remote Services: Remote Desktop Protocol)

2025-02-18T11:22:51Z	2025-02-18T11:23:02Z	kali	Prepare Tools for Ingress Tool Transfer (Arrakis) - Cleanup (T1105-Ingress Tool Transfer)
2025-02-18T11:22:51Z	2025-02-18T11:23:04Z	kali	Sandworm Upload Agent to Windows Host Modified - Cleanup (T1021.002-Remote Services: SMB/Windows Admin Shares)
2025-02-18T11:22:51Z	2025-02-18T11:23:12Z	kali	Sandworm Initial Compromise - Deploy Webshell - Cleanup (T1505.003-Server Software Component: Web Shell)
2025-02-18T11:22:51Z	2025-02-18T11:23:11Z	kali	Sandworm Initial Compromise - Transfer Webshell - Cleanup (T1105-Ingress Tool Transfer)
2025-02-18T11:22:51Z	2025-02-18T11:22:55Z	PDC DB	Sandworm Establish Systemd Persistence - Cleanup (T1543.002-Create or Modify System Process: Systemd Service)
2025-02-18T11:22:51Z	2025-02-18T11:22:57Z	PDC DB	Sandworm Establish Crontab Persistence - Cleanup (T1053.003-Scheduled Task/Job: Cron)
2025-02-18T11:22:51Z	2025-02-18T11:23:00Z	PDC#1	Sandworm Windows Discovery - List Entire File System - Cleanup (T1083-File and Directory Discovery)

5. Dataset Features

Table 3 summarises the TCP/IP network flow statistics that are generated by CICFlowMeter [5].

Table 3. CICFlowMeter TCP/IP Network Flow Statistics - Features

#	Feature	Description
1	Flow ID	The unique ID of the flow; a string composed of the flow's 5-tuple attributes
2	Src IP	The source IP address
3	Src Port	The source TCP/UDP port
4	Dst IP	The destination IP Address
5	Dst Port	The destination TCP/UDP port
6	Protocol	The identifier of the transport layer protocol
7	Timestamp	The datetime indicating the first packet of the flow
8	Flow duration	Duration of the flow in Microsecond
9	total Fwd Packet	Total packets in the forward direction
10	total Bwd packets	Total packets in the backward direction
11	total Length of Fwd Packet	Total size of packet in forward direction
12	total Length of Bwd Packet	Total size of packet in backward direction
13	Fwd Packet Length Max	Maximum size of packet in forward direction
14	Fwd Packet Length Min	Minimum size of packet in forward direction
15	Fwd Packet Length Mean	Mean size of packet in forward direction
16	Fwd Packet Length Std	Standard deviation size of packet in forward direction
17	Bwd Packet Length Max	Maximum size of packet in backward direction
18	Bwd Packet Length Min	Minimum size of packet in backward direction
19	Bwd Packet Length Mean	Mean size of packet in backward direction
20	Bwd Packet Length Std	Standard deviation size of packet in backward direction
21	Flow Bytes/s	Number of flow bytes per second
22	Flow Packets/s	Number of flow packets per second
23	Flow IAT Mean	Mean time between two packets sent in the flow
24	Flow IAT Std	Standard deviation time between two packets sent in the flow
25	Flow IAT Max	Maximum time between two packets sent in the flow
26	Flow IAT Min	Minimum time between two packets sent in the flow
27	Fwd IAT Total	Total time between two packets sent in the forward direction
28	Fwd IAT Mean	Mean time between two packets sent in the forward direction
29	Fwd IAT Std	Standard deviation time between two packets sent in the forward direction
30	Fwd IAT Max	Maximum time between two packets sent in the forward direction
31	Fwd IAT Min	Minimum time between two packets sent in the forward direction
32	Bwd IAT Total	Total time between two packets sent in the backward direction
33	Bwd IAT Mean	Mean time between two packets sent in the backward direction
34	Bwd IAT Std	Standard deviation time between two packets sent in the backward direction
35	Bwd IAT Max	Maximum time between two packets sent in the backward direction
36	Bwd IAT Min	Minimum time between two packets sent in the backward direction

37	Fwd PSH flags	Number of times the PSH flag was set in packets travelling in the forward direction (0 for UDP)
38	Bwd PSH Flags	Number of times the PSH flag was set in packets travelling in the backward direction (0 for UDP)
39	Fwd URG Flags	Number of times the URG flag was set in packets travelling in the forward direction (0 for UDP)
40	Bwd URG Flags	Number of times the URG flag was set in packets travelling in the backward direction (0 for UDP)
41	Fwd RST Flags	Number of times the RST flag was set in packets travelling in the forward direction (0 for UDP)
42	Bwd RST Flags	Number of times the RST flag was set in packets travelling in the backward direction (0 for UDP)
43	Fwd Header Length	Total bytes used for headers in the forward direction
44	Bwd Header Length	Total bytes used for headers in the backward direction
45	Fwd Packets/s	Number of forward packets per second
46	Bwd Packets/s	Number of backward packets per second
47	Packet Length Min	Minimum length of a packet
48	Packet Length Max	Maximum length of a packet
49	Packet Length Mean	Mean length of a packet
50	Packet Length Std	Standard deviation length of a packet
51	Packet Length Variance	Variance length of a packet
52	FIN Flag Count	Number of packets with FIN
53	SYN Flag Count	Number of packets with SYN
54	RST Flag Count	Number of packets with RST
55	PSH Flag Count	Number of packets with PUSH
56	ACK Flag Count	Number of packets with ACK
575	URG Flag Count	Number of packets with URG
58	CWR Flag Count	Number of packets with CWR
59	ECE Flag Count	Number of packets with ECE
60	Down/Up Ratio	Download and upload ratio
61	Average Packet Size	Average size of packet
62	Fwd Segment Size Avg	Average size observed in the forward direction
63	Bwd Segment Size Avg	Average size observed in the backward direction
64	Fwd Bytes/Bulk Avg	Average number of bytes bulk rate in the forward direction
65	Fwd Packet/Bulk Avg	Average number of packets bulk rate in the forward direction
66	Fwd Bulk Rate Avg	Average number of bulk rate in the forward direction
67	Bwd Bytes/Bulk Avg	Average number of bytes bulk rate in the backward direction
68	Bwd Packet/Bulk Avg	Average number of packets bulk rate in the backward direction
69	Bwd Bulk Rate Avg	Average number of bulk rate in the backward direction
70	Subflow Fwd Packets	The average number of packets in a sub flow in the forward direction
71	Subflow Fwd Bytes	The average number of bytes in a sub flow in the forward direction
72	Subflow Bwd Packets	The average number of packets in a sub flow in the backward direction
73	Subflow Bwd Bytes	The average number of bytes in a sub flow in the backward direction
74	Fwd Init Win bytes	The total number of bytes sent in initial window in the forward direction
75	Bwd Init Win bytes	The total number of bytes sent in initial window in the backward direction

76	Fwd Act Data Pkts	Count of packets with at least 1 byte of TCP data payload in the forward direction
77	Bwd Act Data Pkts	Count of packets with at least 1 byte of TCP data payload in the backward direction
78	Fwd Seg Size Min	Minimum segment size observed in the forward direction
79	Bwd Seg Size Min	Minimum segment size observed in the backward direction
80	Active Mean	Mean time a flow was active before becoming idle
81	Active Std	Standard deviation time a flow was active before becoming idle
82	Active Max	Maximum time a flow was active before becoming idle
83	Active Min	Minimum time a flow was active before becoming idle
84	Idle Min	Minimum time a flow was idle before becoming active
85	Idle Mean	Mean time a flow was idle before becoming active
86	Idle Std	Standard deviation time a flow was idle before becoming active
87	Idle Max	Maximum time a flow was idle before becoming active
88	ICMP Code	The ICMP code observed in the flow (-1 for non-ICMP flow)
89	ICMP Type	The ICMP type observed in the flow (-1 for non-ICMP flow)
90	Fwd TCP Retrans. Count	Count of TCP retransmissions in the forward direction
91	Bwd TCP Retrans. Count	Count of TCP retransmissions in the backward direction
92	Total TCP Retrans. Count	The total count of TCP retransmissions
93	Total Connection Flow Time	Duration of the flow in Microsecond
94	Label	The label used for classifying the flow

6. Citation

E. Iturbe, C. Dalamagkas, P. Radoglou-Grammatikis, and E. Rios, “APT Sandworm Dataset,” Zenodo. [Online]. Available: <https://doi.org/10.5281/zenodo.16911636>

7. Acknowledgements

This work has received funding from the European Union’s Horizon Europe research and innovation programme under grant agreement No 101070450 (AI4CYBER).

References

- [1] The MITRE Corporation, “Sandworm Team,” Mitre.org. [Online]. Available: <https://attack.mitre.org/groups/G0034/>. [Accessed: 20-Aug-2025].
- [2] The MITRE Corporation, “Caldera,” Mitre.org. [Online]. Available: <https://caldera.mitre.org/>. [Accessed: 20-Aug-2025].
- [3] The Center for Threat-Informed Defense by MITRE Engenuity, “Sandworm,” adversary_emulation_library (Github), 2022. [Online]. Available: https://github.com/center-for-threat-informed-defense/adversary_emulation_library/tree/master/sandworm. [Accessed: 20-Aug-2025]
- [4] The Center for Threat-Informed Defense by MITRE Engenuity, “Sandworm Emulation_Plan,” Apr-2022. [Online]. Available: https://github.com/center-for-threat-informed-defense/adversary_emulation_library/tree/master/sandworm/Emulation_Plan.yaml. [Accessed: 20-Aug-2025].
- [5] G. Engelen, V. Rimmer, and W. Joosen, “Troubleshooting an intrusion detection dataset: the CICIDS2017 case study,” in 2021 IEEE Security and Privacy Workshops (SPW), 2021, pp. 7–12.