Multi-Agent Collaboration in Incident Response with Large Language Models

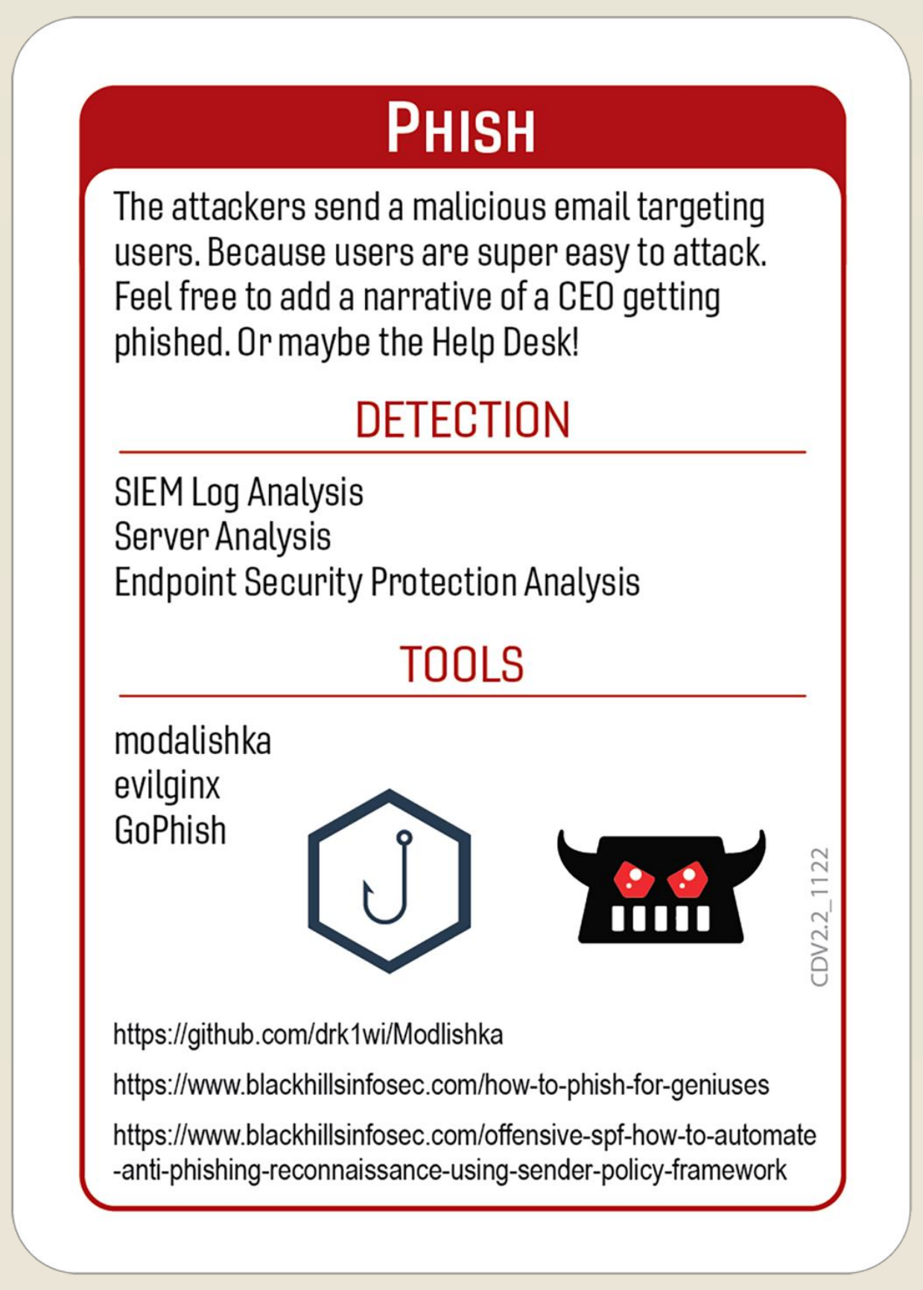
Zefang Liu (Georgia Institute of Technology)
Contact: liuzefang@gatech.edu

Overview

- **Incident response (IR)** is essential for cybersecurity, requiring **quick decision-making and coordination**.
- **Large Language Models (LLMs)** can serve as **intelligent agents** to enhance collaboration and efficiency.
- **LLM-based multi-agent collaboration** in **cybersecurity** is explored using **Backdoors & Breaches**, a tabletop game designed for IR training.
- Different **team structures (centralized, decentralized, hybrid)** are analyzed to evaluate their impact on IR effectiveness.

Backdoors & Breaches

- A **tabletop game** that simulates **real-world cybersecurity incidents**.
- Players take on roles as **incident captain** and **defenders**, working together to **uncover and mitigate attack vectors**.
- The game includes different card types:
  - **Attack Cards:** Represent stages of a cyberattack (e.g., initial compromise, pivot and escalate, command and control (C2) and exfiltration, and persistence).
  - **Procedure Cards:** Defensive strategies used to detect and counter threats.
  - **Inject Cards:** Unexpected events that introduce new challenges.
- The goal is to **reveal all hidden attack cards** within limited turns through strategic decision-making.



PHISH

The attackers send a malicious email targeting users. Because users are super easy to attack. Feel free to add a narrative of a CEO getting phished. Or maybe the Help Desk!

DETECTION

SIEM Log Analysis
Server Analysis
Endpoint Security Protection Analysis

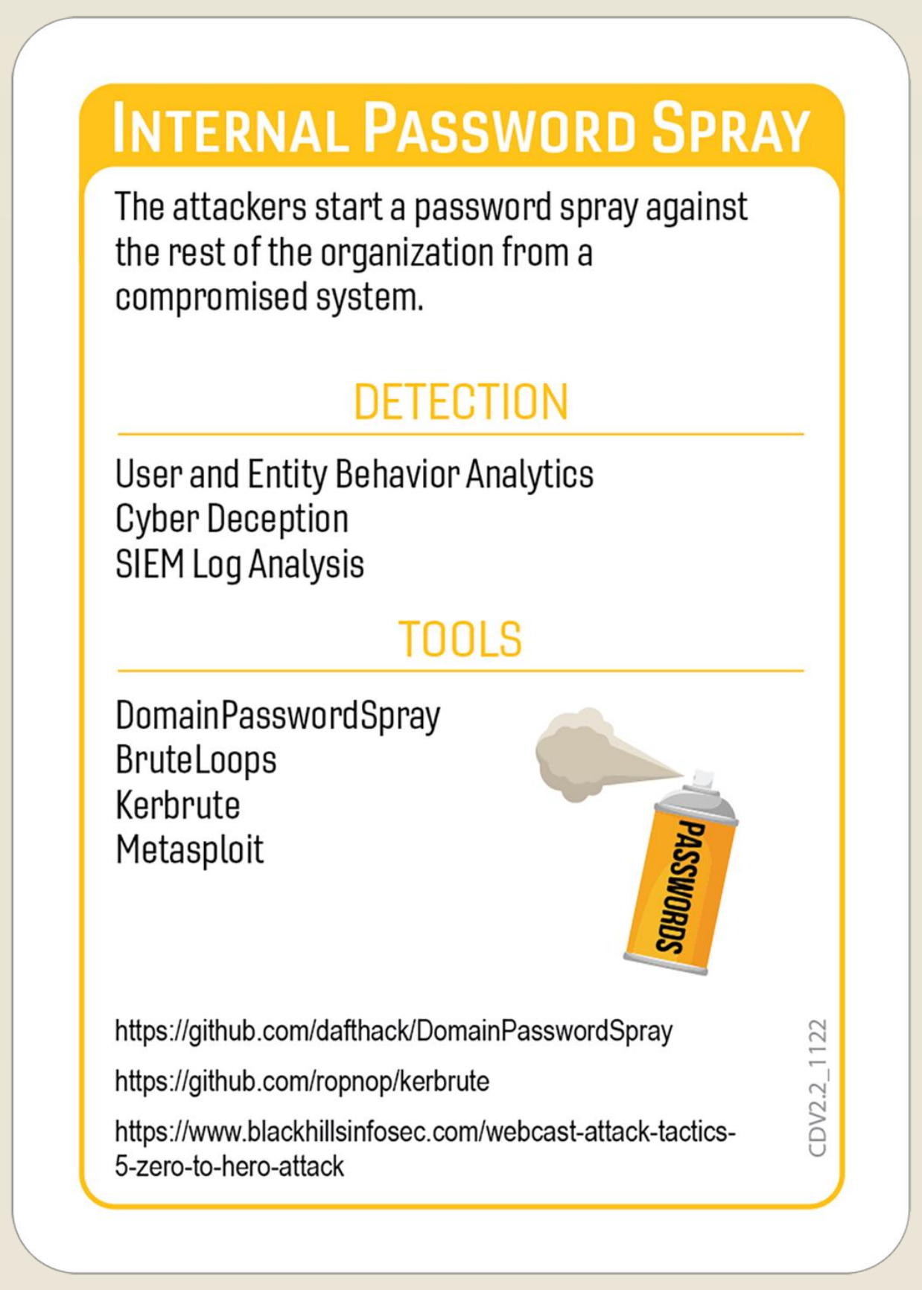
TOOLS

modalishka
evilginx
GoPhish

https://github.com/drk1wi/Modlishka
https://www.blackhillsinfosec.com/how-to-phish-for-geniuses
https://www.blackhillsinfosec.com/offensive-spf-how-to-automate-anti-phishing-reconnaissance-using-sender-policy-framework

CDV2.2_1122

Initial Compromise



INTERNAL PASSWORD SPRAY

The attackers start a password spray against the rest of the organization from a compromised system.

DETECTION

User and Entity Behavior Analytics
Cyber Deception
SIEM Log Analysis

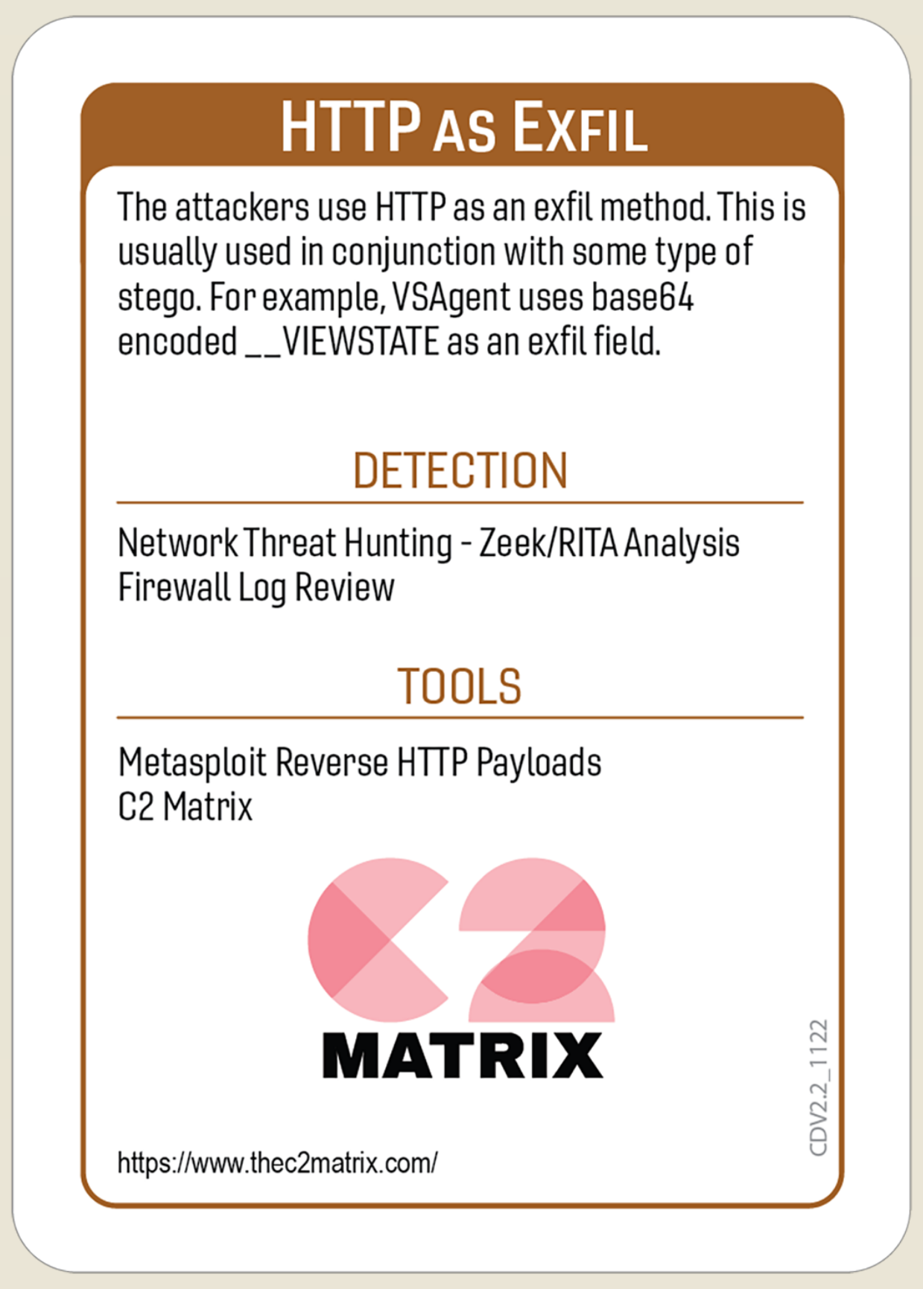
TOOLS

DomainPasswordSpray
BruteLoops
Kerbrute
Metasploit

https://github.com/dafthack/DomainPasswordSpray
https://github.com/ropnop/kerbrute
https://www.blackhillsinfosec.com/webcast-attack-tactics-5-zero-to-hero-attack

CDV2.2_1122

Pivot and Escalate



HTTP AS EXFIL

The attackers use HTTP as an exfil method. This is usually used in conjunction with some type of stego. For example, VSAgent uses base64 encoded __VIEWSTATE as an exfil field.

DETECTION

Network Threat Hunting - Zeek/RITA Analysis
Firewall Log Review

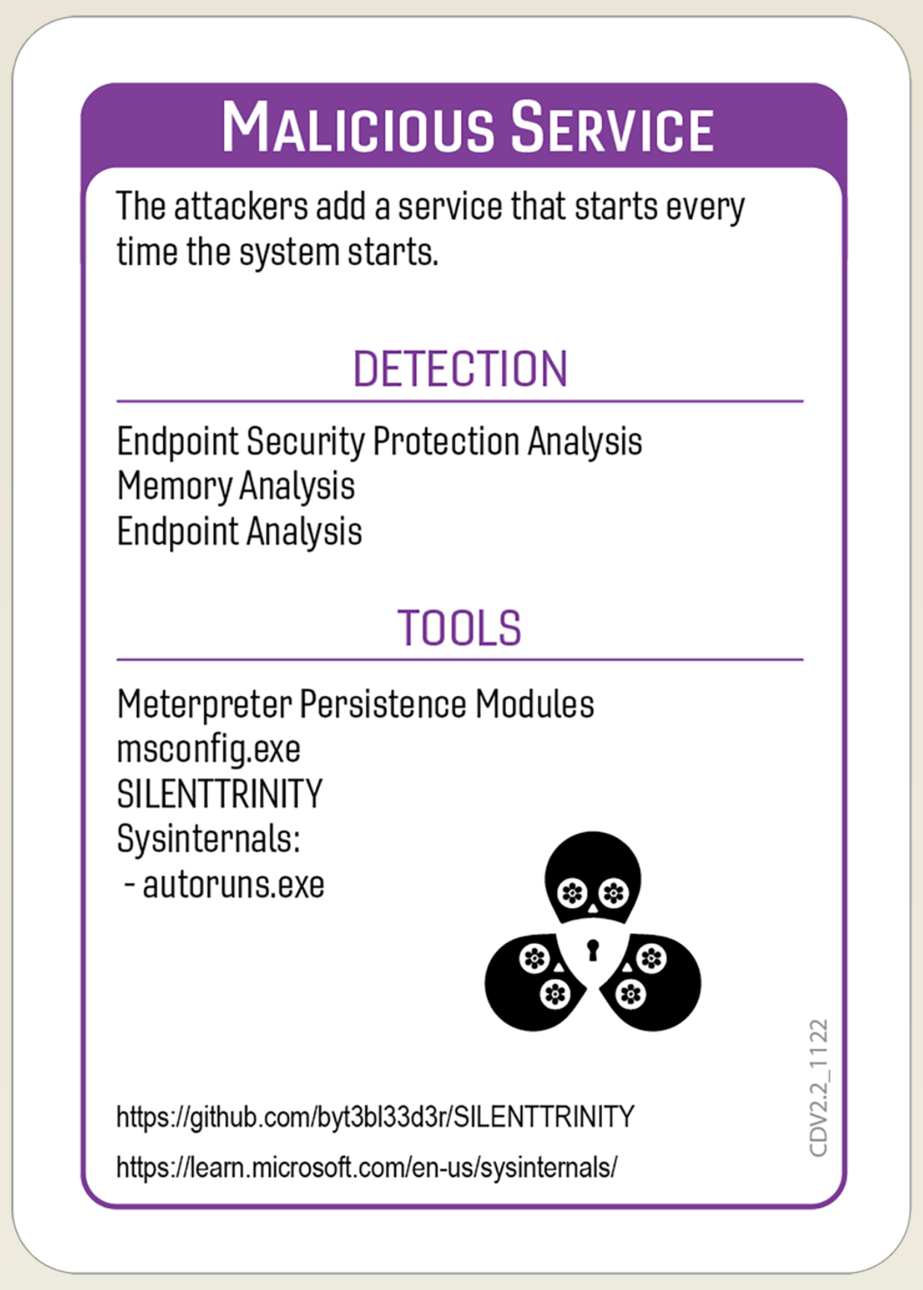
TOOLS

Metasploit Reverse HTTP Payloads
C2 Matrix

https://www.thec2matrix.com/

CDV2.2_1122

C2 and Exfil



MALICIOUS SERVICE

The attackers add a service that starts every time the system starts.

DETECTION

Endpoint Security Protection Analysis
Memory Analysis
Endpoint Analysis

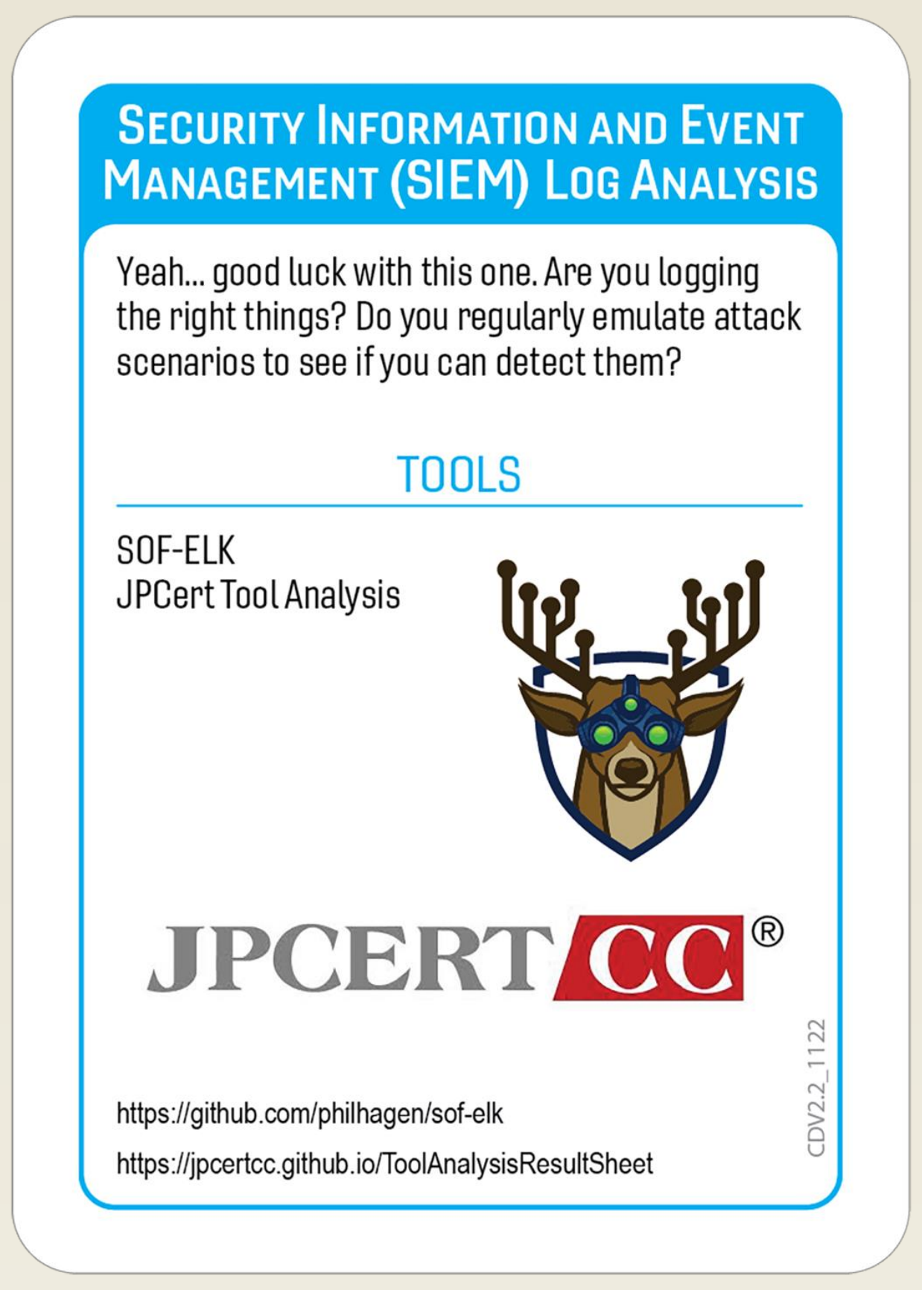
TOOLS

Meterpreter Persistence Modules
msconfig.exe
SILENTTRINITY
Sysinternals:
 - autoruns.exe

https://github.com/byt3bl33d3r/SILENTTRINITY
https://learn.microsoft.com/en-us/sysinternals/

CDV2.2_1122

Persistence



SECURITY INFORMATION AND EVENT MANAGEMENT (SIEM) LOG ANALYSIS

Yeah... good luck with this one. Are you logging the right things? Do you regularly emulate attack scenarios to see if you can detect them?

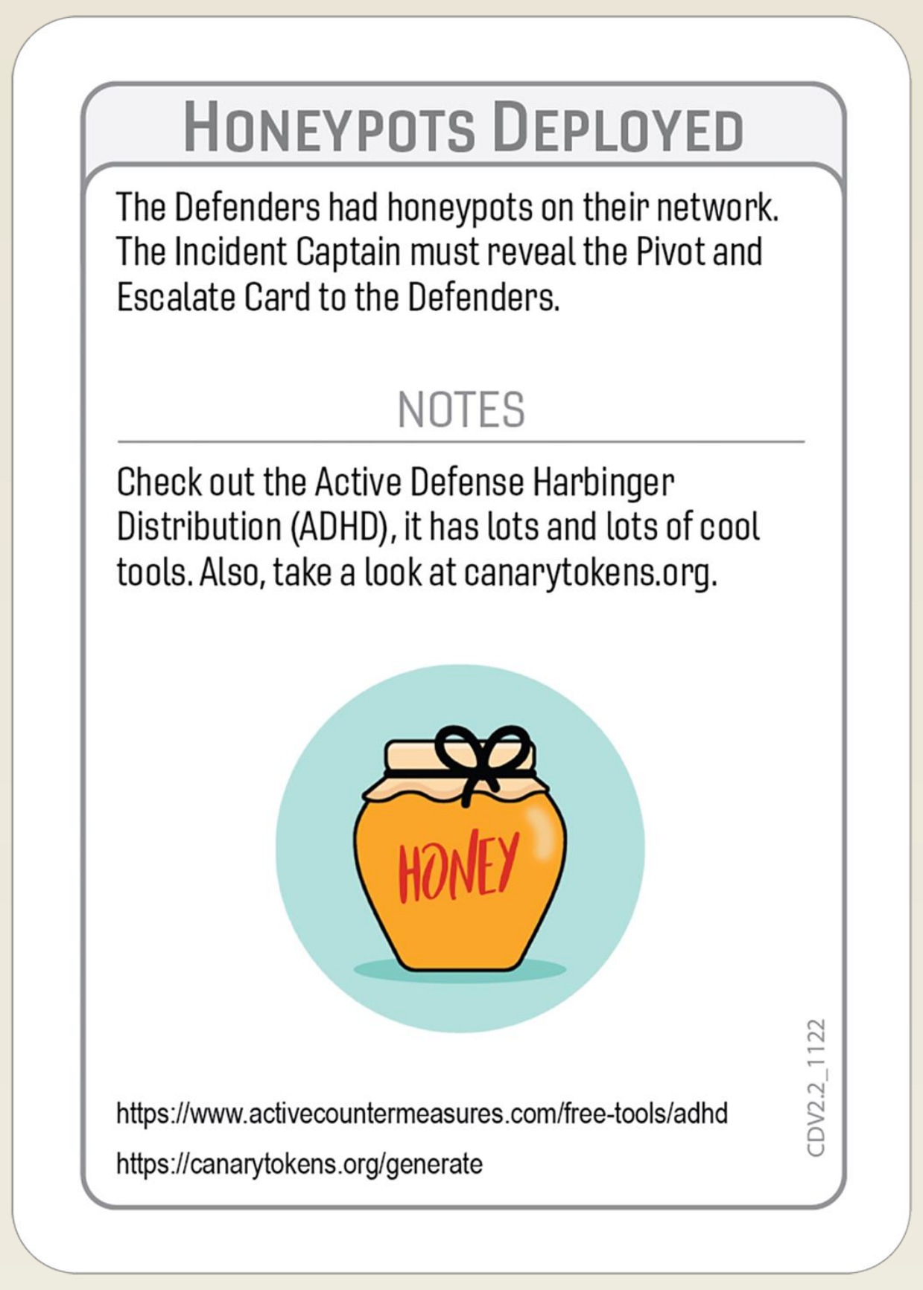
TOOLS

SOF-ELK
JPCert Tool Analysis

https://github.com/philhagen/sof-elk
https://jpcertcc.github.io/ToolAnalysisResultSheet

CDV2.2_1122

Procedure



HONEYPOTS DEPLOYED

The Defenders had honeypots on their network. The Incident Captain must reveal the Pivot and Escalate Card to the Defenders.

NOTES

Check out the Active Defense Harbinger Distribution (ADHD), it has lots and lots of cool tools. Also, take a look at canarytokens.org.

https://www.activecountermeasures.com/free-tools/adhd
https://canarytokens.org/generate

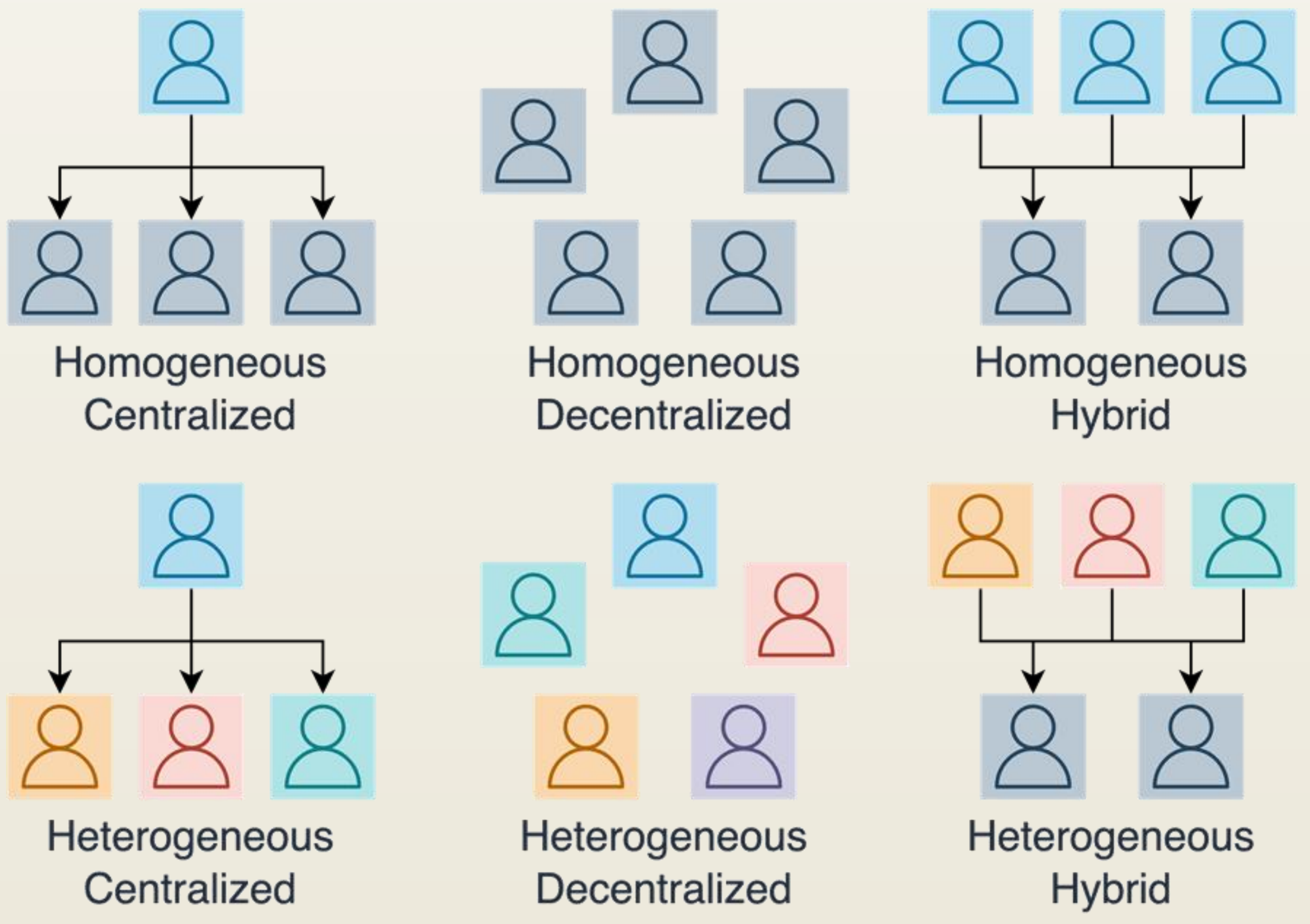
CDV2.2_1122

Inject

Experimental Setup

- **LLM-based agents:** Implemented using **AutoGen**, with structured roles and interactions.
- **Team structures:**
  - **Centralized:** Leadership-driven decision-making.
  - **Decentralized:** Equal decision-making across all agents.
  - **Hybrid:** Mix of leadership and collaboration.
- **Evaluation metrics:**
  - **Success rate** in uncovering all attack stages.
  - **Failure patterns** across different team structures.

Team Structures



Homogeneous Centralized
Homogeneous Decentralized
Homogeneous Hybrid
Heterogeneous Centralized
Heterogeneous Decentralized
Heterogeneous Hybrid

Experimental Results

- **Centralized teams** perform well due to clear leadership but may struggle with adaptability.
- **Decentralized teams** leverage diverse expertise but can have coordination challenges.
- **Hybrid teams** balance structured leadership with flexibility, leading to strong performance.
- **LLM-based agents** facilitate IR processes by assisting in decision-making and coordination.

Team	Success	Failure	Pentest	Invalid
Homo-Cen	14	1	2	3
Heter-Cen	13	3	3	1
Homo-Dec	13	1	1	5
Hetero-Dec	12	3	3	2
Homo-Hyb	14	3	2	1
Hetero-Hyb	13	1	2	4

Case Studies

- **Homogeneous Centralized:** Over-reliance on high-modifier procedures led to poor adaptability.
- **Heterogeneous Centralized:** Struggled with prioritization and aligning expert inputs.
- **Homogeneous Decentralized:** Slow decision-making and overuse of standard procedures.
- **Heterogeneous Decentralized:** Lack of coordination led to missed attack indicators.
- **Homogeneous Hybrid:** Misprioritized investigations delayed threat detection.
- **Heterogeneous Hybrid:** Expertise misalignment caused early attack stages to be overlooked.

Conclusion & Future Work

- **LLMs** demonstrate strong potential in **multi-agent collaboration** in **incident response** and **cybersecurity**.
- **Future directions** includes:
  - Improving adaptability of LLMs for unpredictable cyber threats.
  - Extending simulations to real-world cybersecurity environments.
  - Exploring human-LLM hybrid teams for incident response.

References

Young, Jacob, and Farshadkhah, Sahar. "Backdoors & Breaches: Using a Tabletop Exercise Game to Teach Cybersecurity Incident Response." *Proceedings of the EDSIG Conference ISSN*. Vol. 2473. 2021.
Wu, Qingyun, et al. "AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation." *ICLR 2024 Workshop on Large Language Model Agents*.
Liu, Zefang. "Multi-Agent Collaboration in Incident Response with Large Language Models." *AAAI 2025 Workshop on Multi-Agent AI in the Real World*.

Acknowledgment: Backdoors & Breaches, developed by *Black Hills Information Security* and *Active Countermeasures*, provided the structured framework for simulating incident response scenarios in this study.

SCAN ME