Principled Data-Driven Decision Support for Cyber-Forensic Investigations

Soodeh Atefi, Sakshyam Panda, Manos Panaousis, Aron Laszka

UNIVERSITY OF HOUSTON

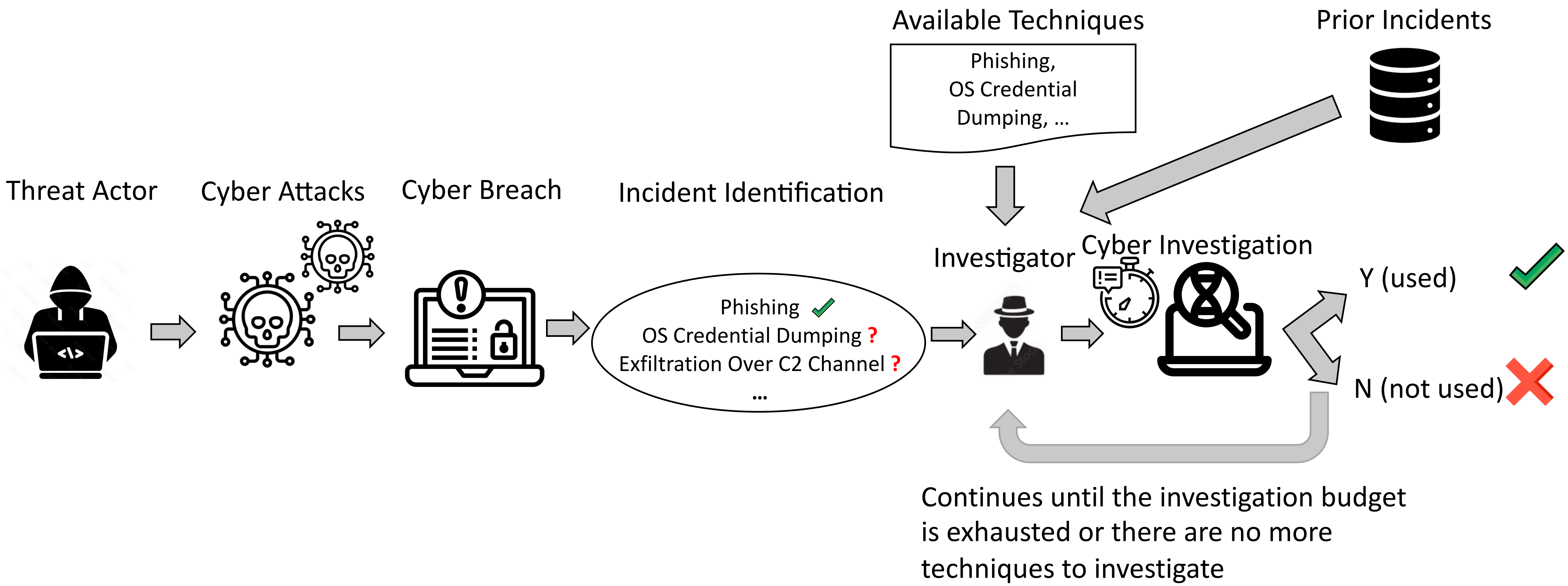
UNIVERSITY of GREENWICH

PennState

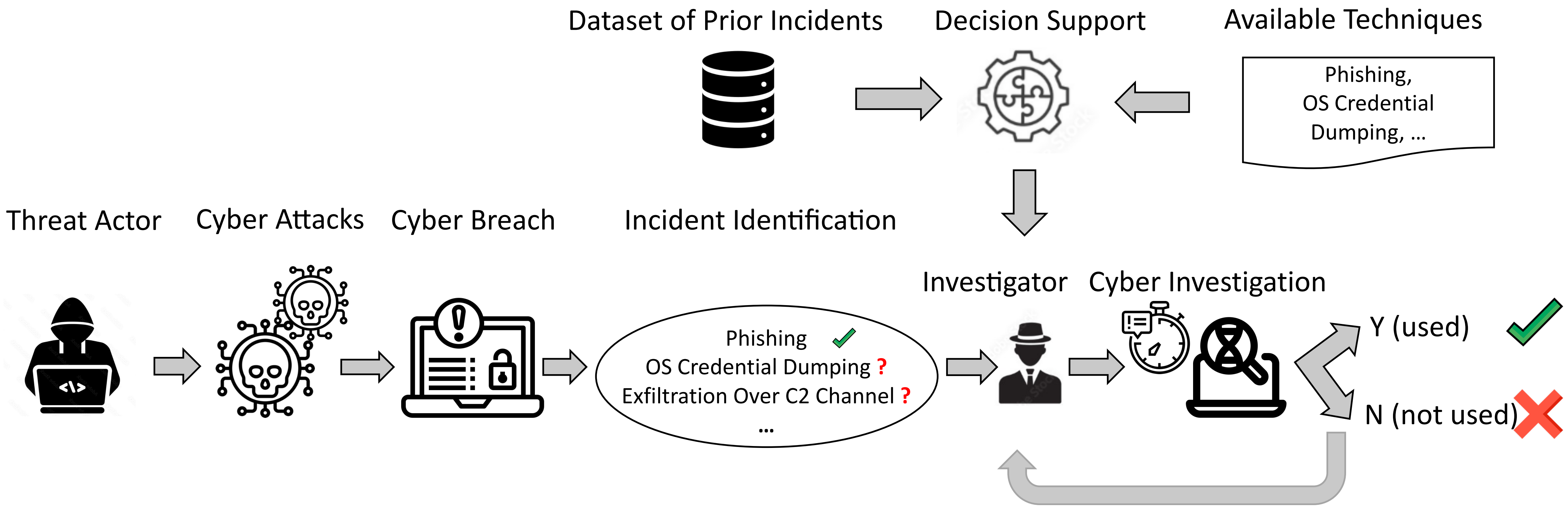
AICS 2023

This material is based upon work sponsored by the National Science Foundation under Grant No. CNS-1850510.

Cyber Forensics Investigation



Goal of Decision Support



Continues until the investigation budget is exhausted or there are no more techniques to investigate

State of the Art: DISCLOSE Framework

- **DISCLOSE** (Nisioti et al. 2021) is a data-driven decision-support framework
- The objective of the framework is to **maximize the benefit** obtained during the investigation without exceeding a **given investigation budget**
  - Investigation of each technique has a **benefit** and a **cost** (denoted by $\boldsymbol{B}$ and $\boldsymbol{C}$)
  - **Budget** is the total cost that the investigator can spend during the investigation
- DISCLOSE **outperforms prior approaches**, such as CBR-FT (Horsman et al. 2014)
- Approach:
  - Computes conditional probabilistic relations between techniques
  - Computes proximity values between techniques (based on the life cycles of an attack)
  - Recommends techniques based on these relations

Nisioti A, Loukas G, Laszka A, Panaousis E. Data-driven decision support for optimizing cyber forensic investigations. *IEEE Transactions on Information Forensics and Security*. 2021 January;16:2397-412.

Limitation of DISCLOSE

- Decisions are based on **heuristic likelihood values**
- Decisions are **myopic**, considering only immediate benefit (but not subsequent steps of the investigation)
- DISCLOSE is a **heuristic approach** that does not approximate optimal decisions under some reasonable objective

Our Approach: Investigation as a Markov Decision Process

- Model the cyber-forensic investigation of an incident as a **Markov decision process** (MDP)
- **State space**: state corresponds to the set of techniques investigated by step t, which were either employed (Y_t) or not employed by the attacker (N_t)
- **Action space**: set of actions is the set of techniques $A \setminus (Y_t \cup N_t)$ that have not been investigated by step t
  - A is a set of all adversarial techniques
- **Transition probability**:
  - probability that the chosen technique was employed by the attacker in the incident
  - estimated based prior incidents (details later)
- **Rewards**:
  - B_a if technique a was used (state $\langle Y_t, N_t \rangle$ to state $\langle Y_{t+1}, N_{t+1} \rangle = \langle Y_t \cup \{a\}, N_t \rangle$)
  - 0 if technique a was not used (state $\langle Y_t, N_t \rangle$ to state $\langle Y_{t+1}, N_{t+1} \rangle = \langle Y_t, N_t \cup \{a\} \rangle$)

Cyber-Forensic Decision Support Problem

- **Policy π**:
  maps a state $\langle Y_t, N_t \rangle$ to a recommended action $a \in A \setminus (Y_t \cup N_t)$
- Objective is to find a policy that maximizes the expected rewards obtained during the forensic investigation:

$$\max_{\pi} \mathbb{E}_{I_Y} \left[\sum_{t=0}^{T_{limit}} 1_{\{a_t \in I_Y\}} \cdot B_{a_t} \,\middle|\, a_t = \pi\left(Y_t, N_t\right) \right]$$

where T_{limit} is the last step before the investigation budget G is exhausted:

$$T_{limit} = \max_T \sum_{t=0}^{T} C_{a_t} \leq G$$

Computational Approach

- To solve the decision-support problem, we propose a **k-nearest neighbor** (k-NN) based **Monte Carlo tree search** (MCTS)
- Monte Carlo tree search
  - in each step of an investigation, run a search from the current state $\langle Y_t, N_t \rangle$
  - **action selection:** apply **Upper Confidence Bound 1 rule** to balance exploration and exploitation
  - **expansion:** sample transitions with uniform probability
  - **backpropagation:** use the transition probabilities (estimated by k-NN, *discussed later*) to update expected rewards
- Computational tricks *(see paper for details)*
  - **myopic pruning**: focus on actions that are optimal w.r.t. myopic objective
  - **values estimation**: estimate the value of unexplored states by assuming that probabilities would be frozen when expanding that state

Probability Estimation

- Our goal is to estimate state-transition probability $\Pr[a \mid Y_t, N_t]$ based on prior incidents
  - **computational challenge**: there are a limited number of prior incidents, so empirical conditional probabilities may be inaccurate or inexistent
- Approach: use ***k*-nearest neighbor** regression to estimate probability
  - non-parametric model estimates directly based on dataset
  - distance metric: similarity between current and prior incident

$$d(\langle Y_t, N_t \rangle, \hat{I}) = \left| Y_t \cap \hat{I}_N \right| + \left| N_t \cap \hat{I}_Y \right|$$

  - number of neighbors *k* is dynamically adjusted during the investigation

Numerical Evaluation

- Baselines:
  DISCLOSE and a **static policy** (i.e., fixed order of investigation)
- Three versions of **MITRE ATT&CK Enterprise dataset** (v6.3, v10.1, and v11.3 latest)
  - our approach can be applied to newer versions without any changes
  - leave-on-out cross validation (i.e., all other incidents are prior)
- For fair comparison, we consider the same 31 techniques as DISCLOSE
- Benefit and cost of each technique (same as DISCLOSE):
  - benefit: based on **Common Vulnerability Scoring System**
  - cost: based on **interviews with cyber forensic experts**

Numerical Results

- Our approach outperforms both baselines on all datasets
  - we considered two scenarios: investigation up to budget 45 and up to 65
- Running times are negligible compared to the investigation time

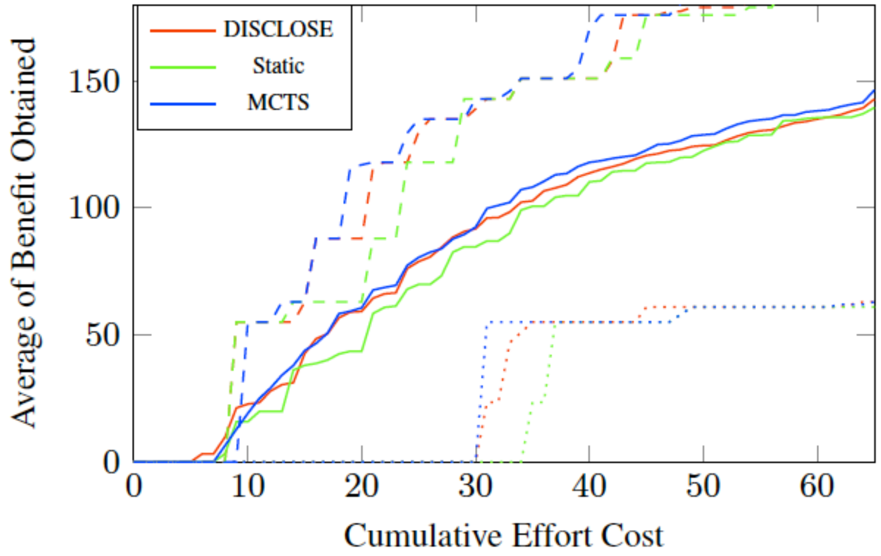


Figure 2: Average benefit obtained as a function of cumulative effort cost (**up to budget 65**) on **v6.3**.

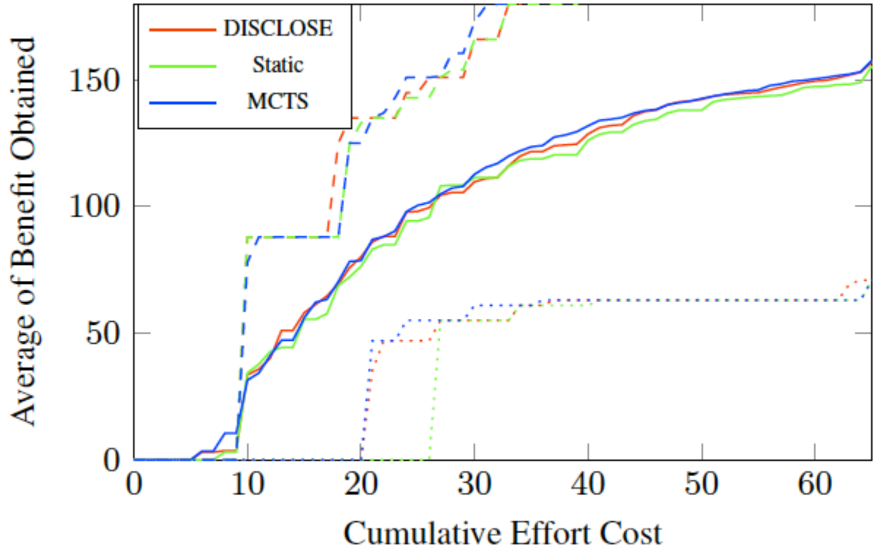


Figure 4: Average benefit obtained as a function of cumulative effort cost (**up to budget 65**) on **v11.3**.

Conclusion

- To address the limitations of DISCLOSE, we introduce a **principled approach for cyber-forensic decision support**
- Key challenge: **limited prior data** vs. large action space
- Proposed approach:
  - model cyber-forensic investigation as **Markov decision process**
  - **k-NN for estimating transition probabilities** (non-parametric model makes best use of limited data)
  - **Monte Carlo tree search** with computational tricks
- Our approach is computationally efficient and outperforms SOTA

Thank you for your attention!