USC Viterbi School of Engineering *Information Sciences Institute*

CICI: RSSD: DISCERN: Datasets to Illuminate Suspicious Computations on Engineering Research Networks

Jelena Mirkovic, Brian Kocoloski[1], Sam Liang, Spencer Stingley[2], Rishit Saiya[3], USC Information Sciences Institute

URL: https://steelisi.github.io/DISCERN/

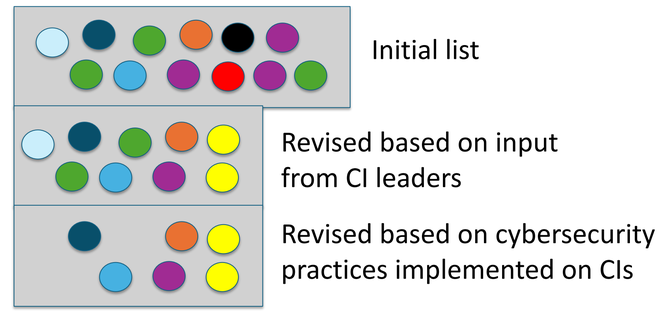
Motivation

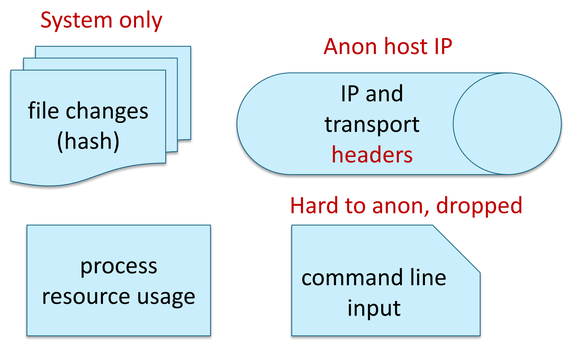
Scientific cyberinfrastructure (CI) is very complex. It often includes large amount of third-party software, external code and data and it serves external users. There is also usually high turnover in CI team members, as students graduate and leave:

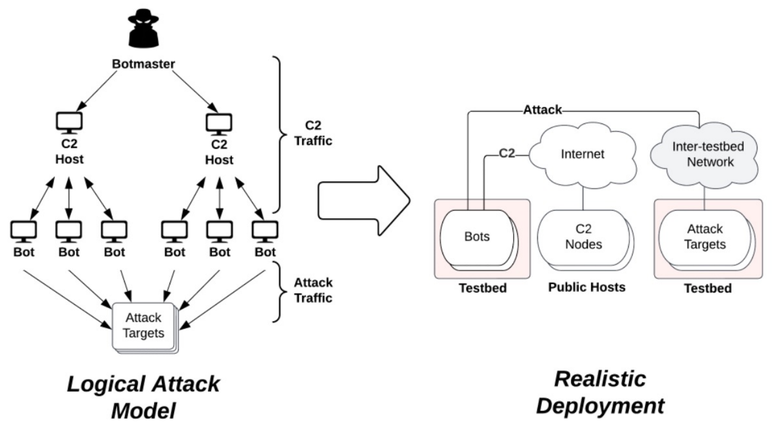
- Scientific CI must remain mostly open to meet research needs
- Attackers can target scientific CI to steal data, misuse resources or damage CI
- We need better datasets about scientific CI use patterns and how attacks may manifest on it

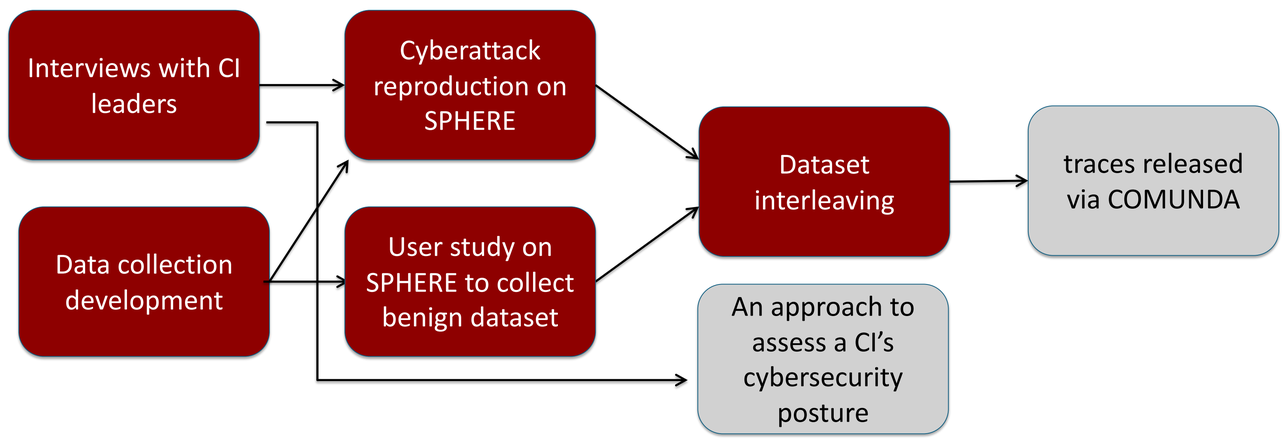
Technical Approach

- Cyberattack scenarios:
  - Assume bot nodes within a testbed, targeting the testbed or misusing resources to attack others
  - Create a preliminary list of cyberattacks
  - Modify the list based on interviews with PIs from Cloudlab, POWDER, FABRIC and Chameleon about cyberattacks they observed or are concerned about
  - Narrow down the list based on interviews with PIs about cybersecurity practices they implement
- Data collection:
  - Instrument default OS images on SPHERE to collect data that may show evidence of cyberattacks
  - Optimize for performance (CPU, memory, disk) to ensure non-interference with experiments
  - Remove or anonymize private or identifying data to ensure adoption
- Reproduce relevant cyberattacks within SPHERE using BYOB, gather attack data
- Collect data from benign experiments with user consent (no attacks) and with IRB approval
- Interleave benign and attack data, generating datasets for attack detection on CIs









Cyberattacks: cryptomining, Internet scanner, port scanner within CI, ransomware, email spread, data exfiltration

Benefits to Scientific Cyberinfrastructure

- A labeled dataset for ML for cyberattack detection on Cis
- A new approach to evaluate a cyberinfrastructure's security posture and prioritize improvements
- Portable tools for CI activity monitoring
- More secure future CI

Result Dissemination

- Results disseminated via our project Web page
- One publication under submission, preparing a second one
- Data collection tools planned for permanent integration with SPHERE
  - Discussing adoption by other CIs
- Datasets released via COMUNDA portal (https://comunda.isi.edu)
  - Easy request and approval process

Risk/Benefit Analysis

- We minimized the privacy risk of our data collection by using anonymization and excluding potentially sensitive data
- We minimized the security risk of attack scenarios by ensuring all code is open source and vetted by us
- Benefits to CI community outweigh any remaining risks
  - Improved security of cyberinfrastructure