サイバーセキュリティ研究におけるデータセットの意義

2023年6月29日 MWS2023プレミーティング

NTTコミュニケーションズ株式会社
畑田 充弘

データセットの価値

信頼性（一般化する）

- 調査で明らかにした事象の信頼性を高め、対策を促す
  - 山嵜ら「悪用された脆弱性に関する情報源の比較調査」（MWS2022ベストプラクティカル賞）
    - KEV Catalogue、NVD、Symantec Signature、他
    - 調査の結果，いずれかの情報源にのみ含まれていた悪用された脆弱性は全体の 57.3%に及び，情報を網羅的に収集するためには複数の情報源を活用する必要があることがわかった.

信頼性（一般化）

- 検知、解析技術の評価結果の信頼性を高める
  - Your Router Is My Prober: Measuring IPv6Networks via ICMP Rate Limiting Side Channels (NDSS 2023 Distinguished Paper)
    - Our large-scale ISAV measurements cover ~50% of all IPv6 autonomous systems and find ~79% of them are vulnerable to spoofing, which is the most large-scale measurement study of IPv6 ISAV to date. Our method for reachability measurements achieves over 80% precision and recall in our evaluation.

新規性

- データセットそのものが新規であり、有用なデータを提供することで他の研究にも貢献する
  - EMBER dataset
    - https://github.com/elastic/ember
  - MWS Datasets
- データの収集方法を提案、ツールを公開するものも
  - Zmap - USENIX Security ’13
    - https://github.com/zmap/zmap

データセットの準備≒研究そのもの

パターン

- データセット First、研究ネタ Second
  - 所属組織（企業、研究室、etc.）または個人で保有している
  - せっかくあるので何かできないか？
- 研究ネタ First、データセット Second
  - こういう研究をしたい！
  - そのためにはこういうデータセットが必要
    - 探す
    - 作る

考えないといけないこと

- 何を評価するか？
  - 研究目的そのもの（例：マルウェアを検知したい）
- どう評価するか？
  - ノイズ除去、サンプリング、ラベリング、分割
- データを収集・蓄積・処理するシステムの開発・運用
  - インターネット接続環境
  - 監視
  - ストレージコスト（生データ vs メタデータ）
- 倫理
  - https://www.iwsec.org/csec/ethics/checklist.html