Cybersecurity Datasets: A Mirage

Jelena Mirkovic (USC/ISI)

Stephen Hayne (CSU), Michalis Kallitsis (Merit Network), Wes Hardaker (USC/ISI), John Heidemann (USC/ISI), Christos Papadopoulos (U Memphis), Devkishen Sisodia (U Oregon)

NSF Workshop on Overcoming Measurment Barriers to Internet Research

2021-01-12

Cybersecurity Event Detection: Needs

- normal and abnormal events (e.g, attack vs leg. traffic, clean vs infected hosts)
- revolving (capture new events) and curated (benchmark) data
- accurately labeled data – very hard! no ground truth
- levels of event sophistication or multiple datasets – avoid overfitting
- possible to cross-correlate (join) with other datasets – challenge: privacy

Privacy vs Utility

- Problem: Often at odds, no good technical way to meet both needs
- Spectrum of access: collaborator to public
- Solution: Fall back to social regulations:
  - Vet researchers, sign MOAs
  - Slowly increase access privileges
    - Have ways to grant fine-grained access to data, trace leaks, revoke access
  - Provider/researcher partnerships
    - Providers benefit from research findings

Data Labels

- Problem: No ground truth
  - Can use commercial systems but they are making best guesses too
- Solution: Crowdsourced labeling, multiple labels
  - Different algorithms can be used to label events
  - E.g., “this approach has 90% true positives and 0.01% false positives on Mao-Smith labels”
  - Enable research in spite of uncertainty