Distributed Sensor Data Contextualization at Scale for Threat Intelligence Analysis

Jason Trost
January 12, 2016
whoami

Jason Trost

• VP of Threat Research @ ThreatStream
• Previously at Sandia, DoD, Booz Allen, Endgame Inc.
• Background in Big Data Analytics, Security Research, and Machine Learning
• Big advocate and contributor to open source:
  • Modern Honey Network, BinaryPig, Honeynet Project
  • Apache Accumulo, Apache Storm, Elasticsearch
ThreatStream

• Cyber Security company founded in 2013 and venture backed by Google Ventures, Paladin Capital Group, Institutional Venture Partners, and General Catalyst Partners.

• SaaS based enterprise security software that provides actionable threat intelligence to large enterprises and government agencies.

• Our customers hail from the financial services, healthcare, retail, energy, and technology sectors.
Agenda

• Background
• Modern Honey Network
• Sensors
• Enrichment
• Contextualization
• Examples
• Gotchas
• Conclusion
Background

• Huge proliferation of new and old network sensors
  • IDS, Passive Inventory Systems, Malware Sandboxes
  • Honeypots, DNS Sinkholes, Endpoint agents

• Many useful data enrichment sources
  • Passive DNS (PDNS), Whois, IP Geolocation
  • Large Malware Metadata Repositories
  • Network Telescopes / Distributed Sensors / Honeypots
  • Portscan and Web crawl data repositories
  • Internal IT, Security, and IR Systems

• Data overload if not leveraged carefully

• Lots of opportunities for combining these data sets, interpreting them, and contextualizing events for threat researchers

• This research started with Honeypots, expanded to other events...
Honeypots

- Software systems designed to mimic vulnerable servers and desktops
- Used as bait to deceive, slow down, or detect hackers, malware, or misbehaving users
- Designed to capture data for research, forensics, and threat intelligence
- Also useful as sinkhole servers when paired with DNS RPZ
Why Honeypots

• Cheapest way to generate threat intelligence feeds around malicious IP addresses at scale

• Internal deployment
  • Behind the firewall
  • Low noise IDS sensors
  • Can be used in conjunction with DNS RPZ as sinkhole webserver

• Local External deployment
  • Who is attacking me?
  • Outside the firewall and on your IP space

• Global External deployment
  • Rented Servers, Cloud Servers, etc
  • Who is attacking everyone?
  • Global Trends
Modern Honey Network (MHN)

• Open source platform for managing honeypots, collecting and analyzing their data
  • https://github.com/threatstream/mhn
• Makes it very easy to deploy new honeypots and get data flowing
• Leverages some existing open source tools
  • hpfeeds
  • nmemosyne
  • honeymap
  • MongoDB
  • Dionaea, Amun, Conpot, Glastopf
  • Wordpot, Kippo, Elastichoney, Shockpot
  • Snort, Suricata, p0f
MHN Architecture

MHN Server

- Mnemosyne
  - MHN Architecture

MongoDB

- hpfeeds
- honeymap
- Webapp
- REST API

Sensors

- conpot
- dionaea
- Glastopf
- shockpot
- p0f
- Amun
- wordpot
- suricata
- snort
- Kippo
- elastichoney

Integrations

- Splunk
- ArcSight

Malware Sandboxes

- Cuckoo

Users

3rd party apps

© 2015 ThreatStream Confidential
MHN Community

• MHN is also a community of MHN Servers that contribute honeypot events
• MHN Servers and their honeypots are operated by different individuals and organizations
• Sharing data back to the community is optional
• Anyone that does share can get access to aggregated data on attackers
MHN Community

MHN Project

Stats and Indicators on Attackers

MHN Servers

Events

Honeypots/Sensors
Beyond Honeypot Sensors

- Intrusion Detection Systems
- Protocol Analyzers/Decoders
- Passive Device Inventory/Fingerprinting
- Sinkholes
- Malware Sandboxes
- Endpoint Security Products
Enrichment

- Data sets that are useful for joining with events
- Both local and external datasets can be useful
- Examples:
  - Whois
  - Passive DNS
  - Active probing data repositories (portscan, traceroute, web crawl)
  - Malware Metadata Repositories
  - Threat Intelligence Knowledgebase
  - Rollups, Analytics, Facts from your sensors
  - Internal IT, Security, and IR Systems
Contextualization

- Gather details and related information to make an event or an indicator more actionable
- Guide the analyst towards best practices
- Help analysts work faster/better
- Encode expert knowledge in the analytics and presentation
- Building blocks for more automation, decision support, and features for classifiers
- Remove the need for Level 1 SOC analyst?
Honeypot Attacker Profile?

- p0f events?
  - OS?
    - Linux or Windows or other?
  - Uptime?
    - short (less than 1 day)?
    - long (weeks or more)?
  - MTU?
    - Cable?
    - DSL?
    - VPN/tunneled?

- Query PDNS for the IP, filter for recent resolutions
  - Decent number of domains? → could be a web server

- Query Portscan repository
  - recent port 80/443 open?

- Query threat intelligence knowledge database
  - TOR?
  - I2P?
  - Commercial VPN?
  - Open or Commercial proxy?

Infected Windows Workstation?
- home / work

Compromised Webserver?
- shared hosting?
- dedicated?

Ephemeral Exploitation/Scanning server?
Compromised System – How?

- Attacker using a compromised system?
- How did they get in?
  - SSH Brute force?
- Query portscan/webcrawl data repository
Campaign Scope?

• Is this IP attacking just me?
• Are they attacking my vertical?
• Are they attacking everyone?
• Distributed Honeypots or sensors are key here
  • Query external global deployment
  • Query external local deployment
  • Combine Events and summarize
    • first seen / last seen / number of sensors hit / ports involved
    • histogram of activity
    • Summary of exploits used, tools dropped & related C2s
Attacker Toolkit

- Deploying IDS with Honeypots can assist here
- Snort/Suricata are really useful for adding more context
  - CVE Tagging – roughly 1/3 of the Emerging Threat Snort Rules have CVEs
  - Classify traffic
- Honeypots should collect exploit payloads and commands
- Linux Malware Sandbox
  - Execute these commands/scripts (often times wget + execute)
  - Save all payloads
  - Extract host and network IOCs
  - Maintain relationship to original attacker IP
- Query toolsets in VT
Malware Sandbox

- Deploy IDS on Malware Sandbox (Detonate files or URLs)
- Signatures Identify some types of C2 network traffic
- Identify Exploit Kit traffic (CVE tagger)
- Identify sinkhole IPs passively
- Extract indicators, CVEs, Context, make associations
- Any future event regarding these IOCs on your network should be enriched with this context
Sinkholes

- High interaction systems that mimic real services and C2 protocols where possible
- Deploy with IDS sensor
  - tag traffic where possible with C2 protocols
- Local Deployment
  - Use RPZ to sinkhole known malicious / suspicious domains
    - Malware C2
    - Dynamic DNS domains
    - Exploit kit domains
  - Identify internal compromised systems
- External Deployment
  - Register expired malicious domains or seize them
  - Identify infected systems across the globe
Automated Incident Response Collection

- **Starting Point:** Policy Violation, Network IDS Alert, Honeypot Sensor Event, DNS Sinkhole hit, Indicator Match in SIEM, etc.
- Automatically collect host based data
  - Logged in users
  - Running processes
  - DNS cache
  - Open network connections
  - Persistence checks
  - Prefetch files
- Diff the collected data against the previous collection or a “gold image”
- Prepare context for analyst
Enrichments: Whois

- Who registered this domain?
- Was this domain registered with a free email provider?
- Was this domain registered with a disposable email provider?
- Privacy protected?
- Is this domain likely sinkholed?
Enrichments: Internal IT, Security, and IR Systems

- Identity Information
- Asset Data
  - Specific Device
  - Owner
  - Device Characteristics
  - Software Inventory
- Related IR Tickets
Enrichments: Passive DNS (PDNS)

- What other domains resolved to this IP?
- What other IPs did this domain resolve to?

- Is this domain sinkholed?
- Is this a parking IP?
- Is this domain resolving to an IP using DHCP?
- Fast flux domain?
- Often useful to combine with Whois
  - Common registrant across most domains resolving to single IP? -> Sinkholed
  - Diverse registrants, common registrar? -> Parking IP
Enrichments: Active Probing Data

- Portscan, Web crawl, traceroute
- Repositories
  - Build your own or leverage 3rd parties
- Host profile
  - Web server?
  - Embedded Device?
  - Router?
  - Endpoint?
- C2 Panel?
- Vulnerabilities?
  - Many can be determined unobtrusively
- Sinkhole?
  - X-Sinkhole header
Gotchas

• False positives
• Whitelists
• Lots of dead ends, pointing these out to analysts is important
• Rate limiting of enrichments
Conclusion

• Huge proliferation of network sensors and enrichment datasets
• Combining data is useful, let’s do that
• Lots of opportunity to make security analysts better/faster
Contact

Jason Trost

• @jason_trost
• jason [dot] trost [AT] threatstream [dot] com