Evaluating Threat Intelligence Feeds

Paweł Pawliński (CERT Polska / NASK)
pawel.pawlinski@cert.pl

Andrew Kompanek (CERT/CC)
ajk@cert.org

FIRST Technical Colloquium for Threat Intelligence

Munich, 2016-02-24

Agenda

1 The problem

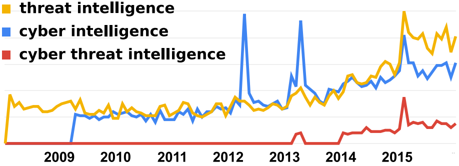
2 Analysis of indicator feeds

3 Our attempt at evaluation

4 Discussion

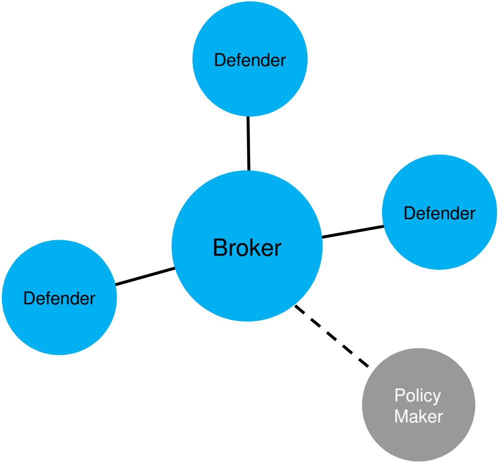
Overview

- Multiple sources of intelligence available
- Ongoing commercialization
- Challenge: assign value to information
- Hypothesis: evaluation needs to be part of consumer ecosystem
- Can we develop an effective approach?

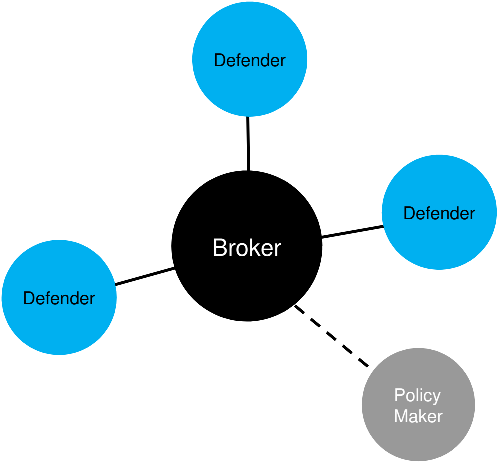


Source: **`www.google.com/trends`**

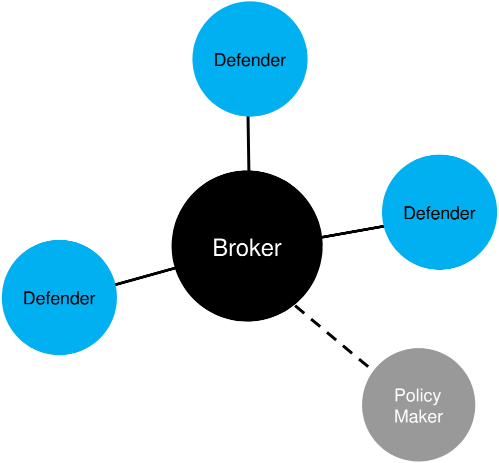
Different points of view



Different points of view

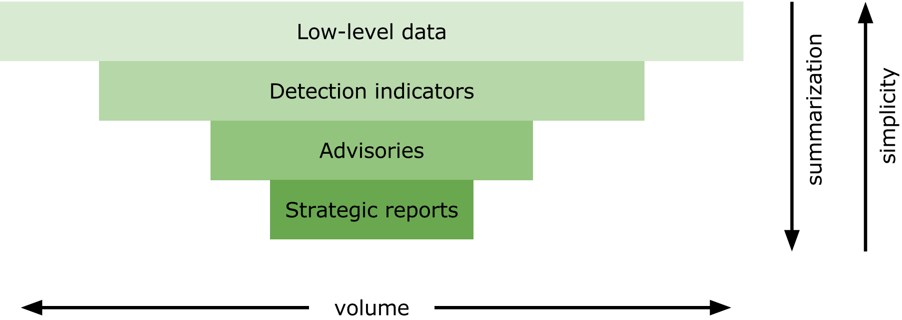


Different points of view



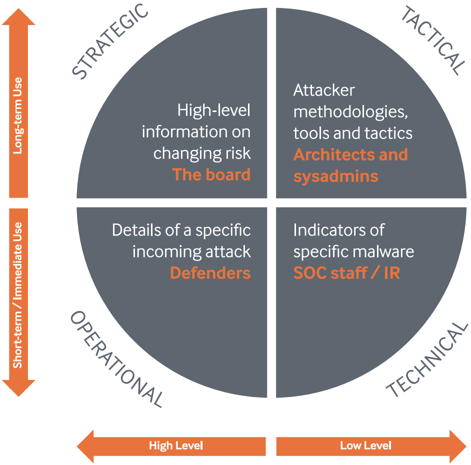
Tip of the day:
Intelligence must be applied at the right spot to provide value

Levels of information



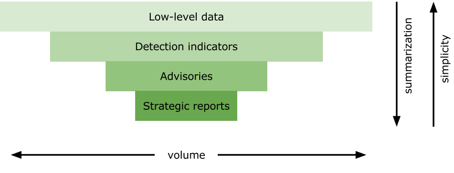
Source: *Actionable Information for Security Incident Response*
`www.cert.pl/news/9684`

Subtypes of intelligence

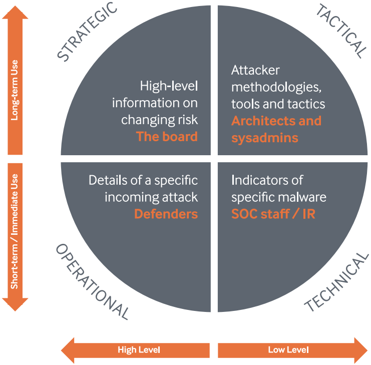

Long-term Use
Short-term / Immediate Use
STRATEGIC
TACTICAL
OPERATIONAL
TECHNICAL
High-level information on changing risk
The board
Attacker methodologies, tools and tactics
Architects and sysadmins
Details of a specific incoming attack
Defenders
Indicators of specific malware
SOC staff / IR
High Level
Low Level

Source: *Threat Intelligence: Collecting, Analysing, Evaluating*
`mwrinfosecurity.com/our-thinking/intelligent-threat-intelligence`

Scope of this talk

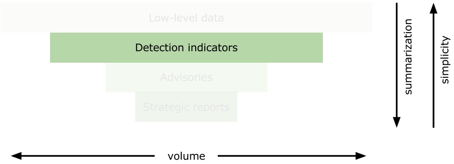


Source: *Actionable Information for Security Incident Response*
`www.cert.pl/news/9684`

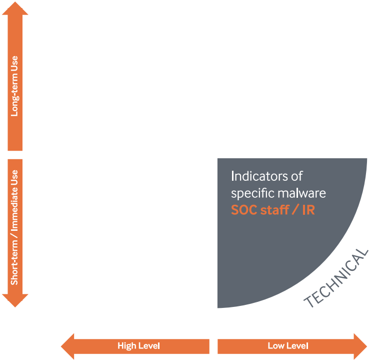

STRATEGIC
TACTICAL
OPERATIONAL
TECHNICAL
Long-term Use
Short-term / Immediate Use
High Level
Low Level
High-level information on changing risk
The board
Attacker methodologies, tools and tactics
Architects and sysadmins
Details of a specific incoming attack
Defenders
Indicators of specific malware
SOC staff / IR

Source: *Threat Intelligence: Collecting, Analysing, Evaluating*
`mwrinfosecurity.com/our-thinking/`

Scope of this talk



Source: *Actionable Information for Security Incident Response*
`www.cert.pl/news/9684`



Source: *Threat Intelligence: Collecting, Analysing, Evaluating*
`mwrinfosecurity.com/our-thinking/`

Properties of (actionable) information

- Quality of information
  - **Relevance** *(Should we care?)*
  - **Accuracy** *(Is it true?)*
  - **Completeness** *(Do we have enough details?)*
  - **Timeliness** *(Is it still valid?)*
  - **Ingestibilty** *(Can we process/interpret it?)*
- Scope of an information source ⇒ coverage
  - **Detection method** *(How the information was obtained?)*
  - **Vantage** *(What is the focus of collection?)*
  - **Volume** *(How much data is provided?)*

Central question:
How do we evaluate available security information?

Central question:
How do we evaluate available security information?

(Ignoring the issue might be a rational approach, too)

Agenda

1 The problem

2 Analysis of indicator feeds

3 Our attempt at evaluation

4 Discussion

Existing work

- Survey of previous data feed evaluation
  1. *Everything You Wanted to Know About Blacklists But Were Afraid to Ask*
  2. *Measuring the IQ of your Threat Intelligence*
  3. *Paint it Black: Evaluating the Effectiveness of Malware Blacklists*

     . . .
  4. Some new ideas applied to CERT.PL data
- Structure of the survey
  - dataset details
  - measurements
  - key conclusions

Everything You Wanted to Know. . .

- *Everything You Wanted to Know About Blacklists But Were Afraid to Ask*
  Leigh Metcalf, Jonathan M. Spring, CERT / SEI, September 2013
- *Blacklist Ecosystem Analysis Update: 2014*
  Leigh Metcalf, Jonathan M. Spring, CERT / SEI, December 2014
- *Blacklist Ecosystem Analysis: Spanning Jan 2012 to Jun 2014*
  Leigh Metcalf, Jonathan M. Spring, CERT / SEI, October 2015

Everything You Wanted to Know. . . Dataset details

Types	“blacklists”, domains & IPs
Sources	anonymized, origin not disclosed domains: 67, IPs: 18
Size	30 months of observations 122M IPs, 31M domains (2nd year)

Everything You Wanted to Know. . . Measurements

→ **SCOPE**

- Descriptive statistics
  - total unique indicators
  - indicators unique to the list
  - intersection
  - **following** relationship

Everything You Wanted to Know. . . Key conclusions

- 96.16% domain indicators unique to 1 list
- 82.46% IP indicators unique to 1 list
- Failed to conclusively determine following relationships

Measuring the IQ…

- *Measuring the IQ of your Threat Intelligence*
  Alexandre Pinto, Kyle Maxwell, DEFCON 22, August 2014
- *Data-Driven Threat Intelligence: Useful Methods and Measurements for Handling Indicators*
  Alexandre Pinto, Alexandre Sieira, FIRST Conference 2015, June 2015
- `http://rpubs.com/alexcpsec/tiq-test-Summer2014-2`
- `http://rpubs.com/alexcpsec/tiq-test-Winter2015`
- `https://github.com/mlsecproject/tiq-test`

Measuring the IQ. . . Dataset details

Types attacking IPs, malicious URLs, C&C, . . .
domains & IPs

Sources 24 public blacklists, 1 private
split into inbound & outbound indicators

Size 2 months of observations, 11k indicators per day
(published example) $\approx$ 0.5M total

Measuring the IQ. . . Measurements

→ **SCOPE**

- Descriptive statistics
  - uniqueness
  - agility
  - overlap
  - AS / CC distribution

→ **ACCURACY**

- Indicator aging

Measuring the IQ. . . Key conclusions

- 97% indicators unique to 1 list (inbound & outbound)
- DIY evaluation (scripts publicly available)

Paint it Black. . .

- *Paint it Black: Evaluating the Effectiveness of Malware Blacklists*
  Marc Kührer, Christian Rossow, Thorsten Holz, Ruhr-Universität Bochum, June 2014

Paint it Black. . . Dataset details

Types	C&C + malicious domains
Sources	15 public blacklists + 4 AV databases
Size	2 years of observations, 0.5M domains

Paint it Black. . . Measurements

→ ACCURACY **→ COMPLETENESS**

- Domain classifications
  - unregistered
  - parked
  - sinkholed
  - active

→ SCOPE

- Blacklist coverage
  - check: C&C in the wild $\in$ blacklist
  - ground truth: 300k sandboxed samples

→ TIMELINESS

- Reaction time
  - **t**(blacklisted) − **t**(appeared)
  - **t**(appeared) based on sandbox data

Paint it Black. . . Key conclusions

- Domain classifications
    - worst public sources: 77% & 57% domains not active
- Blacklist coverage
    - depends on malware family
    - sum of public sources: 0% – 89%, avg 26%
    - sum of AV: 74% – 100%, avg 90%
    - single AV: 26% – 77%, avg 60% (example)
- Reaction time
    - expect > 1 month for “slow” sources

Agenda

1 The problem

2 Analysis of indicator feeds

3 Our attempt at evaluation

4 Discussion

Evaluation experiment

CERT.PL>_

1B security events in 2015, sharing with 200+ organizations
n6: homegrown platform for collection, processing and management

NECOMA

Deliverable 2.2: *Threat Analysis Platform,* Dataset rating
November 2015
`www.necoma-project.eu`

Data collected by a national CERT

- Typical data from 3rd parties: C&C, phishing, . . .
- Information on victims
  - Bots
  - Vulnerable servers
- Attacks originating in the constituency
- Own sources
  - Sinkhole
  - Malware tracking
  - Honeypots
  - Operational activities

Dataset details

- 45 sources:
  - 7 own
  - 38 anonymized
  - public & private
- IPs & domains separately
- 3 weeks of observations
- 55M (indicator – source – day) unique tuples

Variance

→ **SCOPE**

- Quick check of country distribution: deviation from the mode
- Low variance (< 0.1) ⇒ filtered
- Can reveal focus area of a source

Delay

→ **TIMELINESS** → **COMPLETENESS**

- Delay = **t**(report) − **t**(detect)
- Introduced by:
  - source
  - intermediaries
  - exchange mechanism
- Worst case: insufficient precision to determine: 27% (mostly URL sources)
- (Too) Many feeds with delay over 24h

False positive ratio

→ **ACCURACY**

- Simple white lists created – upper bound of FPR
- Unfiltered sandbox: 5.1%
- 2nd *worst*: 3.1%
- Potential problems: 0.5%+
- Most IP sources ≈ 0%

Utility

→ **RELEVANCE** → **SCOPE**

- Idea: see if indicators are useful in operations
- Evaluation dataset: 2k+ analysts' queries
- Top dataset 35.9% (malicious URLs), also the 2nd noisiest
- "Useful" sources:
  1. phishing
  2. bots
  3. scans
- Own sources are above average
- Not "useful": vulnerable servers, amplifiers
- Some correlation with volume (within categories)

Agenda

1 The problem

2 Analysis of indicator feeds

3 Our attempt at evaluation

4 Discussion

Conclusions

- Dataset diversity (not just blacklists of malicious indicators)
- Attempts at analysis of indicator feeds paint interesting picture of the “market”
- Lack of framework for making acquisition decisions
- Missing information:
    - quality
    - scope
    - value vs. cost (in $, effort, false alarms, . . .)
- Even bigger problem for brokers
- *Trust but verify?*

Open questions

- For those of you buying feeds, how did you make those decisions?
- For those of you who do not bother with black lists, your rationale?
- Other studies we should look at that you found useful?
- Other sources of metrics, methodologies, etc.?

Thank you for your attention.

Part of this research has been supported by the Strategic International Collaborative R&D Promotion Project of the Ministry of Internal Affairs and Communication, Japan, and by the European Union Seventh Framework Programme (FP7/2007–2013) under grant agreement No. 608533 (NECOMA). The opinions expressed in this paper are those of the authors and do not necessarily reflect the views of the Ministry of Internal Affairs and Communications, Japan, or of the European Commission.