Towards a Methodology for Evaluating Threat Intelligence Feeds

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

Piotr Kijewski (CERT Polska / NASK)
piotr.kijewski@cert.pl

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

28th Annual FIRST Conference
Seoul, South Korea, 15th July 2016

CERT.PL

CERT

Agenda

Motivation

Focus

Prior work

Methodology

Results

Discussion

Agenda

Motivation

Focus

Prior work

Methodology

Results

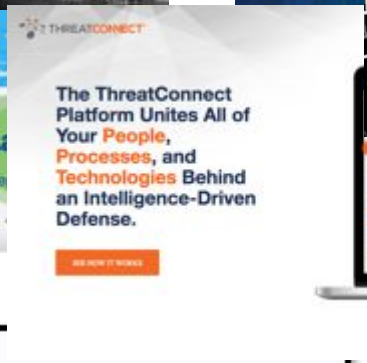
Discussion

A little psychology, some economics and a little about intel providers as "middlemen".


200+ EXPERTS. 24 LANGUAGES.
16 COUNTRIES. 1 MISSION.


KPMG


SOLUTIONARY
Know the Threat La
SERT Q2 2014
Quarterly Threat Intelligence Re
Learn More


THREATCONNECT
The ThreatConnect Platform Unites All of Your People, Processes, and Technologies Behind an Intelligence-Driven Defense.


COMPANY
BLOG
NEWS & EVENTS
SUPPORT
LOGIN
ANOMALI
PRODUCT
WHY ANOMALI
PARTNERS
RESOURCES
FREE TRIAL


Press Releases
Mandiant® Launches New Threat Intelligence Offering
Mandiant Intelligence Center™ provides security teams with the knowledge and context to effectively handle security incidents and combat the most advanced threat actors.
February 26, 2013
San Francisco, CA


Intelligence
Advisories
CTU Research Team
Cyber Security Index
Global Threat Intelligence
Targeted Threat Intelligence
Global Threat Intelligence
Overview
What We Deliver
Service Options
Other Resources

Our Claim

Until evaluation is a more integrated part of the commercial “threat intelligence” ecosystem, progress will be slow...

One small step

Assign value (a “price”) to a stream of information

threatbutt.com

Threatbutt About Careers Contact

Threatbutt

Defense in derpth™

Maximum protection from threatening threaty threats like

cyber hacking

So tell me a little bit about your process…

“We know **our network**, **our users** and **our needs** best. We’re going to do it ourselves.”

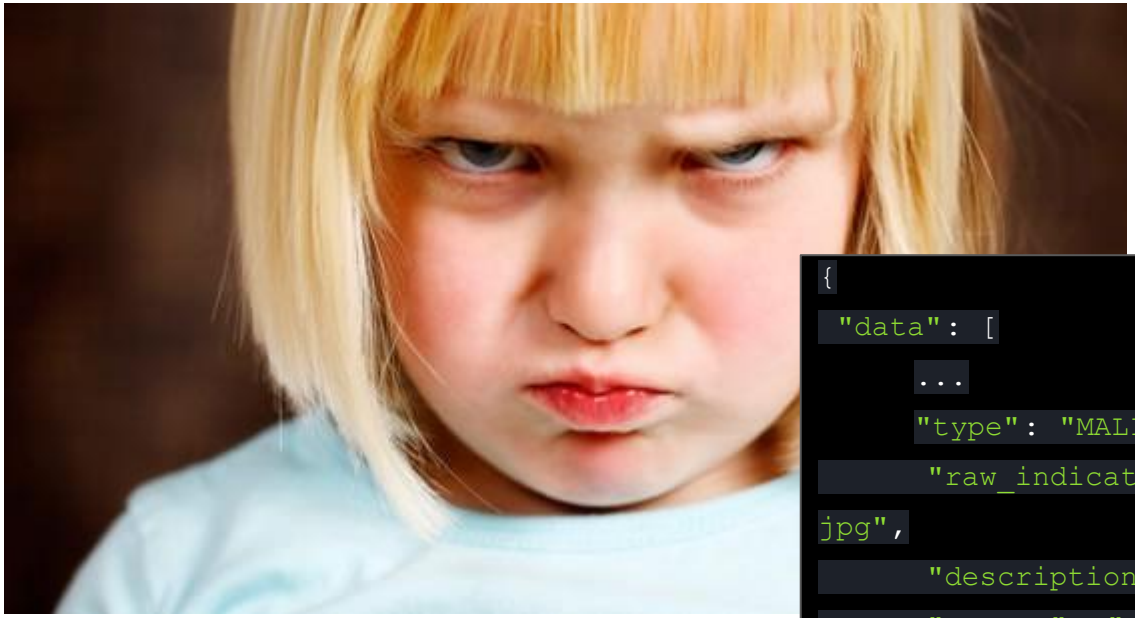
Anonymous

Respect my authoritah!

Clean up your netblock…

or I’m *going home.*

“Suzie doesn’t like the puppy”



facebook for developers | Products | Docs | Tools & Su

ThreatExchange

Learn about threats. Share threat information back. Everyone becomes more secure.

```
{
 "data": [
      ...
      "type": "MALICIOUS_URL",
      "raw_indicator": "http://dawgs.com/puppy.
jpg",
      "description": "Meen looking dawg",
      "status": "UNKNOWN"
]
```

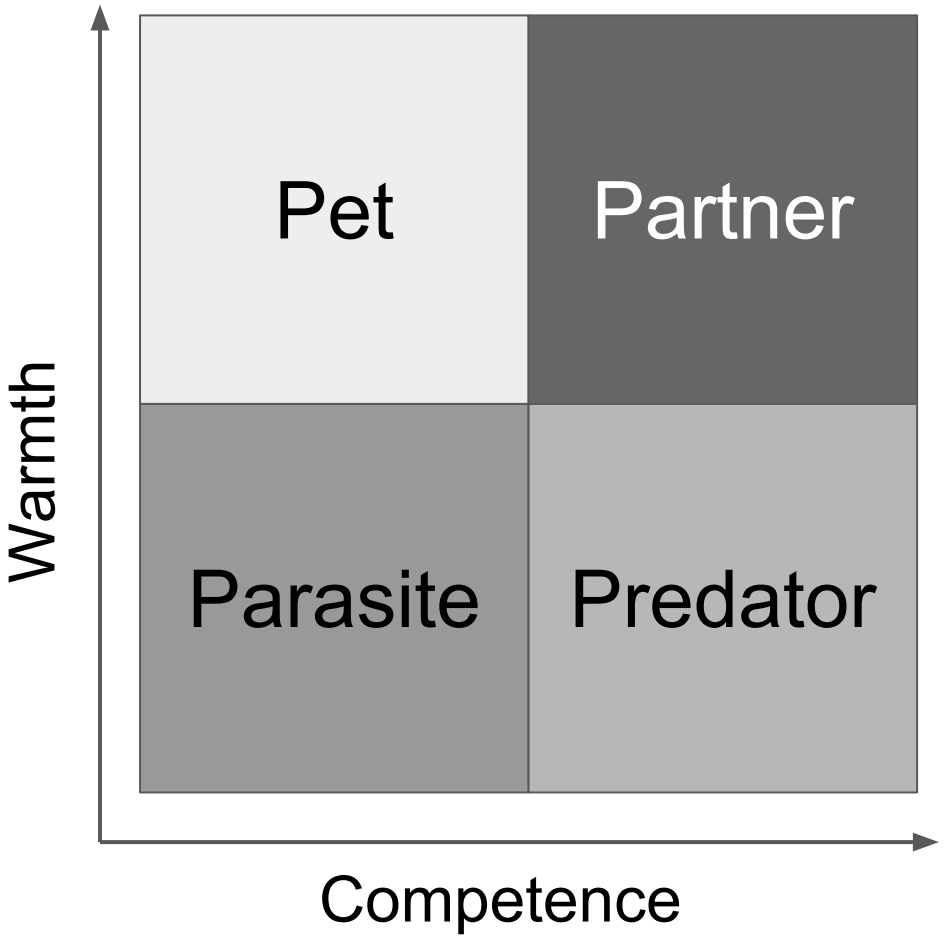
and neither should you

Is there an
echo in here?

“[The Internet will be] the ultimate go-between, the universal middleman [such that] **the only humans involved in a transaction will be the actual buyer and seller**,”

Bill Gates in *The Road Ahead* (1995)

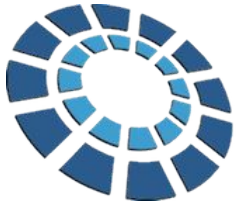
So what use do I have for this guy?



Based on “Universal Dimensions of Social Cognition”, Fiske, Cuddy and Glick.


APWG


INTELMQ




MISP
Threat Sharing


ThreatExchange

Competence: opportunities to add value

Technical value close to the source:

- Collection footprint
- Innovative detection technology

Value added in processing:

- Filtering and quality control
- Distribution

Analytical value added, the hard problem: **Synthesis and interpretation**

Warmth: building networks

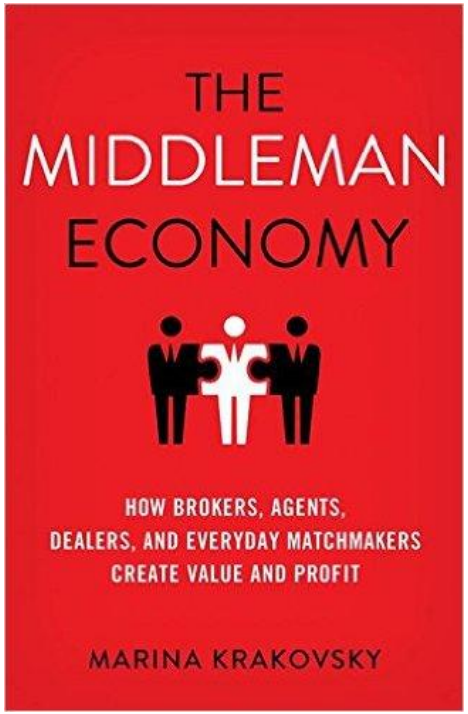
Reduction the cost of connecting providers and consumers:.

- Sensitivity to both consumer need & environment
- Knows space of consumers & producers
- Impedance matching and filtering of data
- Equities management, information protection
- Trust building and maintenance

And of course:

Equipped with tools for evaluation and matching

The Middleman, explained and rehabilitated



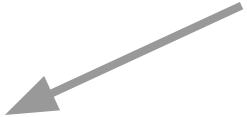
Interesting model and anecdotes:

- A look at the biases against “middlemen” in the economy
- A framework for thinking about their value


Lemon Mart
레몬마트
제한없이 배달해 드립니다
8285~6 종로점
레몬마트
레몬마트
레몬마트

Agenda

So what constitutes a threat intelligence feed anyway?



Motivation

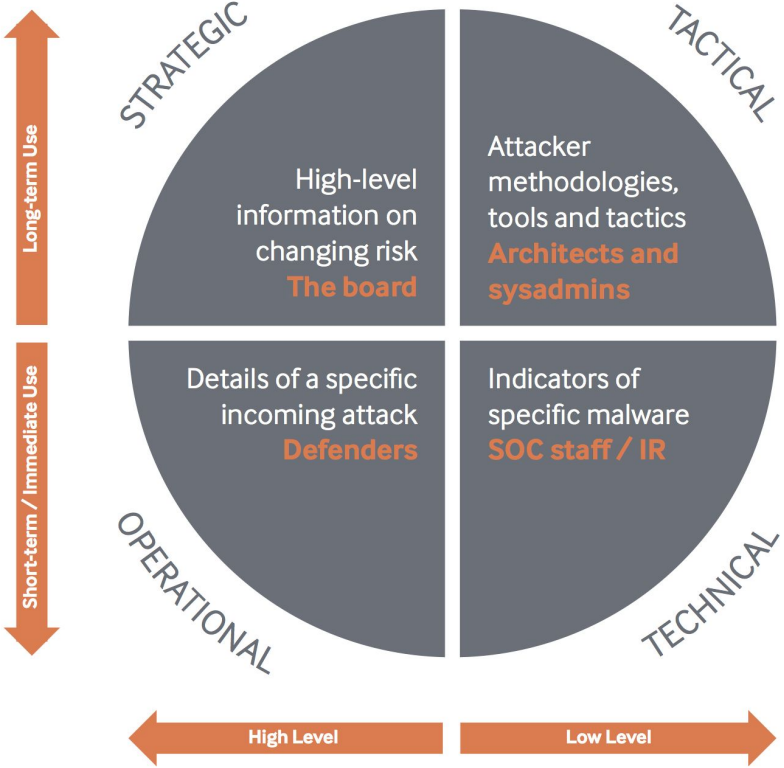
Focus

Prior work

Methodology

Results

Discussion

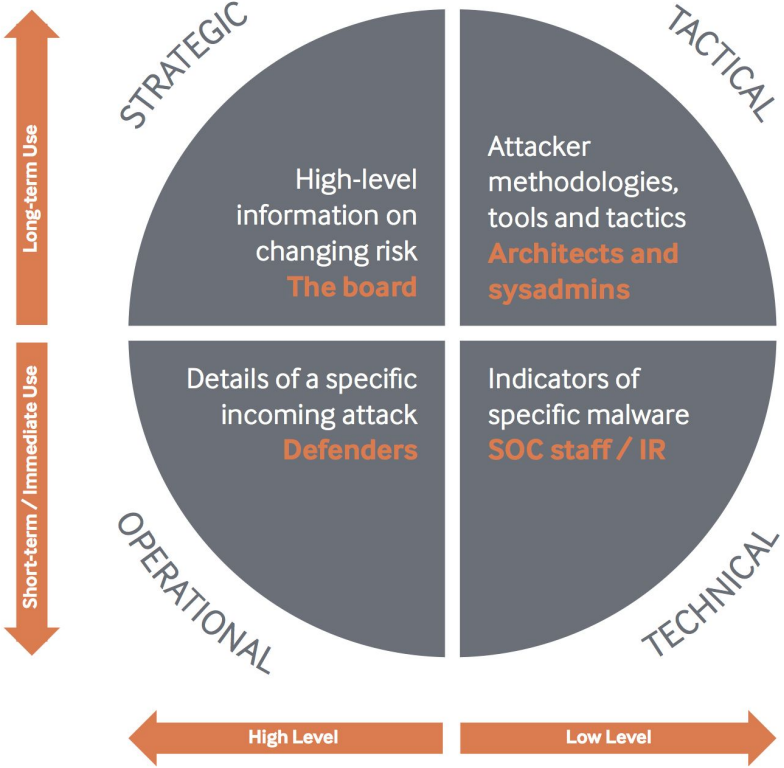


Dimensions

- Scope of use
- Abstraction level of data

Assessment of *value* very different for each case

Threat Intelligence: Collecting, Analysing, Evaluating. Accessed April 28,2015
https://www.mwrinfosecurity.com/system/assets/909/original/Threat_Intelligence_Whitepaper.pdf



Dimensions

- Scope of use
- Abstraction level of data

Assessment of *value* very different for each case

Threat Intelligence: Collecting, Analysing, Evaluating. Accessed April 28,2015
https://www.mwrinfosecurity.com/system/assets/909/original/Threat_Intelligence_Whitepaper.pdf

- Risks in real world domain
- Business resiliency the driver
- In report form

Measurement is hard

- Technical domain
- Decisions about security controls, sensing design
- Still mostly reporting

Value measurement in terms of what's blocked

- Real world & technical
- Reactive mode of use

Easier to assess. Was our response effective?

- Technical domain
- Proactive use (block, monitor)
- Automated measurement feasible

What we're focusing on right now:

- Technical indicators to drive remediation actions

Measurement rubric

Measures of quality:

Relevance	Do I care?
Accuracy	Is it true?
Completeness	Enough context?
Timeliness	Still valid?
Ingest-ability	Effort to process it?

Measures of scope:

Volume	How much data?
Vantage	Where are the sensors?
Detection	How was it detected?

Agenda

Motivation

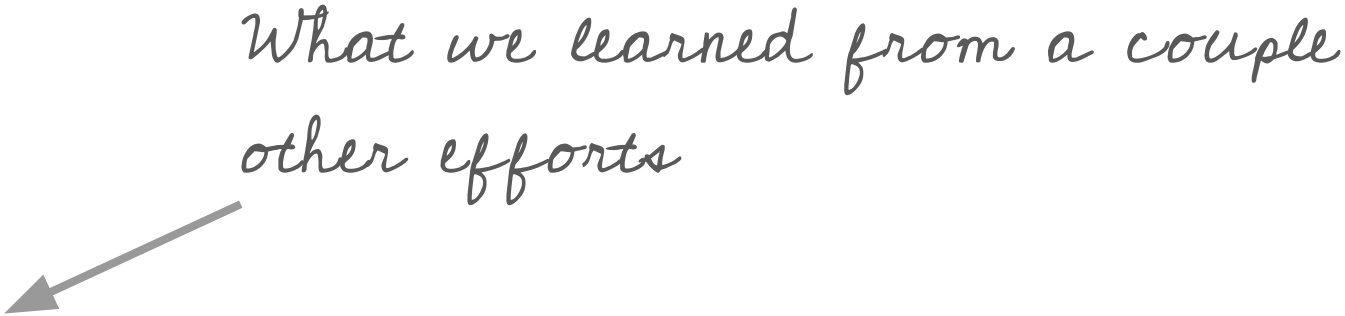
Focus

Prior work

Methodology

Results

Discussion



Prior work

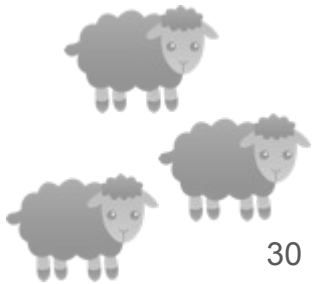
Related evaluations of sources of technical indicators

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

“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

Updates in 2014 and 2016, more coming.



Dataset (2012-2014)

Types of data: **“blacklists”**

Anonymized, origin not disclosed

67 domain-based lists, **18** IP-based lists

30 months of observations

122M IPs, **31M** domains (2nd year)

Measurements and results (2014)

Studied overlap as a characterization of **scope**:

- Number of lists on which an indicator appears
- Pairwise **intersection** between lists

Key results:

- More than **96%** of domain names are unique to one list
- IP addresses are unique to one list **82%-95%** of the time

Insights

Less overlap than expected:

- Blacklists paint fragmented picture of malicious infrastructure
- Providers have very different **scope** of collection

“Measuring the IQ…”

Measuring the IQ of your Threat Intelligence
Alexandre Pinto, Kyle Maxwell, DEFCON 22, August 2014

Data-Driven Threat Intelligence
Alexandre Pinto, Alexandre Sieira, FIRST Conference 2015, June 2015

Verizon DBIR 2015, *Indicators of Compromise* chapter, May 2015

https://github.com/mlsecproject/tiq-test

Dataset

Similar types of data

54 unnamed blacklists

Inbound & outbound indicators

6 months of observations

Measurements and results

Descriptive statistics for **scope**:

- Rate of change
- Overlap
- AS / CC distribution

And **accuracy:**

- Indicator aging

Results confirm the previous study (97% uniqueness).

Volume Vantage

Insights

DIY approach is feasible, some tools 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

Types of data: **C&C** & **“malicious” domains**

Sources: **15 public** blacklists + **4 AV databases**

2 years of observations, 500k domains

“Paint it Black…”

Measurements and results

- **Domain classification**: unregistered, parked, sinkholed, active
  - Worst public sources over half of the domains not active
- **Coverage**: are actual C&C listed?
  - All public sources: **26%** average across families
  - AV sources combined: **90%** average across families
- Compute **reaction time** of blacklists relative to sandbox data
  - Over a month for “slow” sources

Paint it Black: Insights

- “Ground truth” allows the estimation of effectiveness
- AV sources do better than expected
- Some families are not covered enough
- Reaction time - worth checking

Agenda

Motivation

Focus

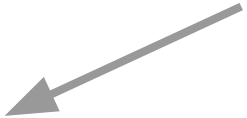
Previous work

Methodology

Results

Discussion

How we approached on the analysis of our CERT.pl data



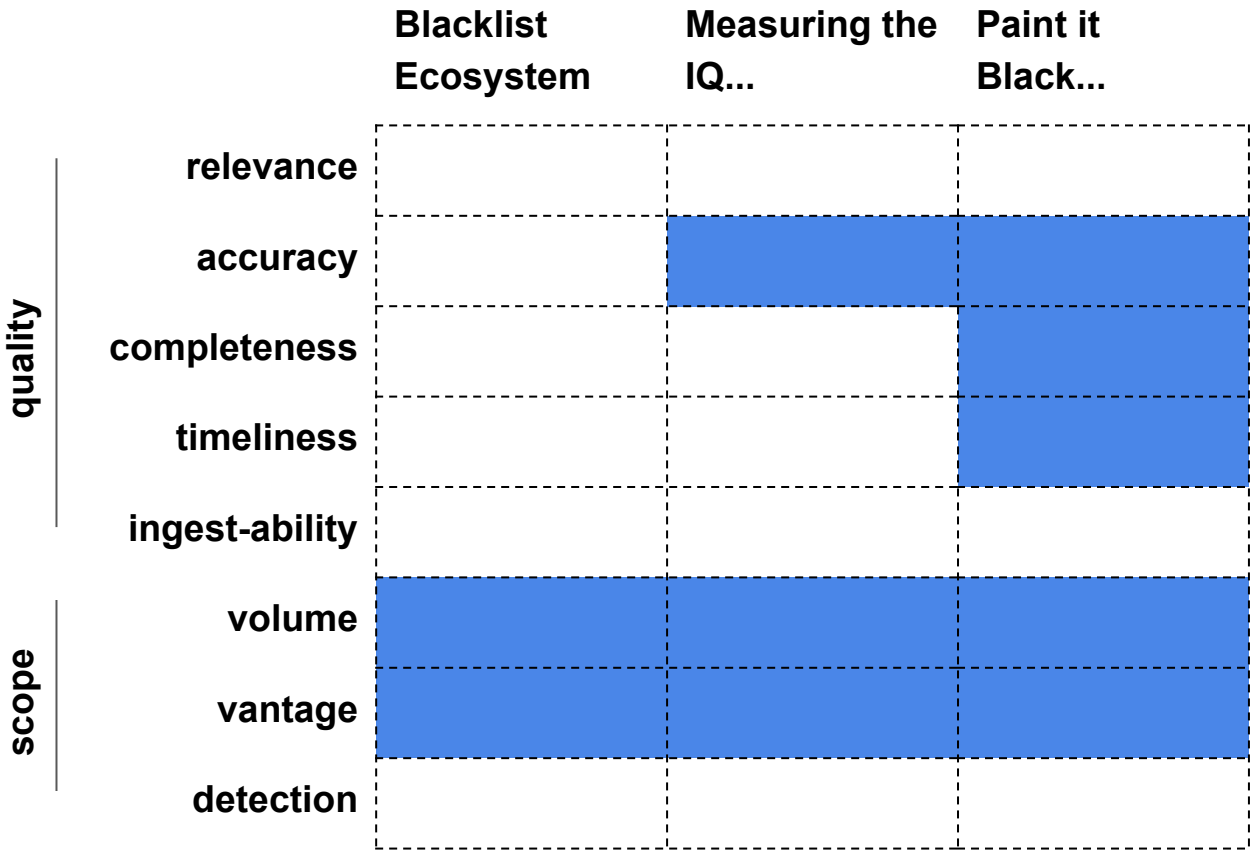
Measurement rubric

Measures of quality:

Relevance	Do I care?
Accuracy	Is it true?
Completeness	Enough context?
Timeliness	Still valid?
Ingest-ability	Can I process it?

Measures of scope:

Volume	How much data?
Vantage	Where are the sensors?
Detection	How was it detected?


Blacklist Ecosystem
Measuring the IQ...
Paint it Black...
quality
relevance
accuracy
completeness
timeliness
ingest-ability
scope
volume
vantage
detection

Dataset

Typical data collected by a national CERT:

- Data from 3rd parties: C&C, phishing, EKs
- Information on victims
- Attacks originating in the constituency
- Own sources
  - Sinkhole and honeypots
  - Malware tracking
  - Operational activities

1B security events in 2015, sharing with 300+ organizations

Mostly automated feeds

www.necoma-project.eu

Deliverable 2.2: Threat Analysis Platform, Dataset rating

Methodology

Measurements

- Rate
- Delivery delay
- False positive rate
- Cross-dataset linkage
- Representativeness
- Utility

Methodology

Measurements

- Rate
- Delivery delay
- False positive rate
- Cross-dataset linkage
- Representativeness
- Utility

Agenda

Motivation

Focus

Previous work

Methodology

Results

Discussion

What running our analysis on the data we've got told us...

Dataset details

Total of **45** sources:

- **7** of our own, **38** anonymized
- public & private

IPs & domains separately

3 weeks of observations in July 2015

55M unique records (record = indicator + source + day)

Delivery delay

Measurement

- Delay = **t(**report**) - t(**detect**)**
- Introduced by: source, intermediaries, exchange mechanism

Results

- Insufficient precision to determine: **27%** (mostly URL sources)
- (Too) Many feeds with delay over 24h: 25% of botnet victim feeds

Timeliness **Completeness**

False positives

Measurement

- Simple white lists created - upper bound of FP rate

Results

- Unfiltered sandbox: **5.1%,** 2nd *worst*: **3.1%**
- Potential problems: **0.5%+**
- Most IP sources were close to 0%

Accuracy

User / utility rating

Measurement

- Count analyst queries

Results

- **2k+ analysts' queries, top** dataset **35.9% (URLs)**, also the 2nd noisiest
- Most “useful”: **phishing, bots, scans**
- Not “useful”: vulnerable servers, amplifiers
- Own sources are above average
- *Observation*: Some correlation with volume (within categories)

Scope Relevance

Case study: closed intelligence sharing groups

- **3 groups**
- Manually verified indicators (in theory)
- Compared against all n6 sources
- **1 year** of data: July 2015 - June 2016

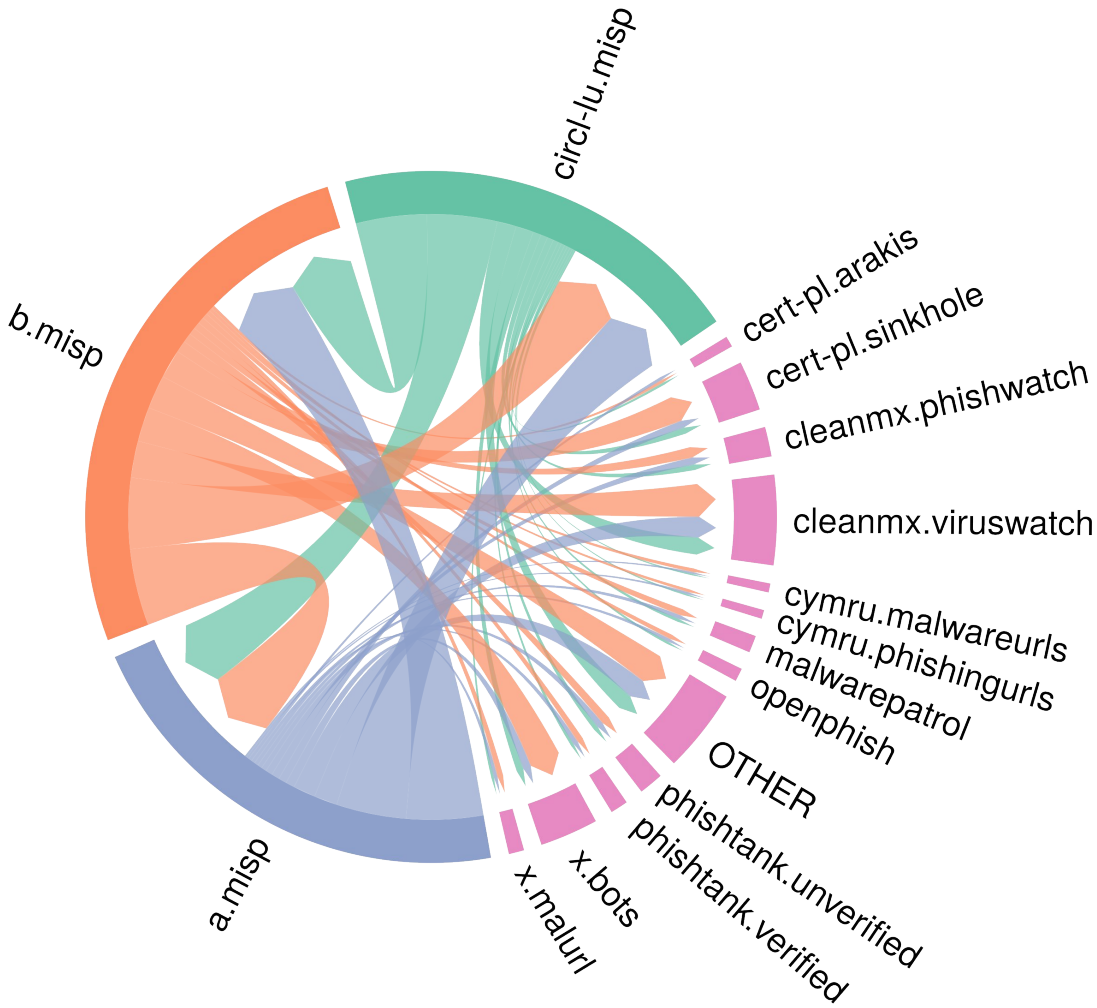
Linkage / Overlap

Analyze relationships between sources

Check overlap for IPs - including data expanded via DNS

Volume | Detection | Vantage

Instance	IPs	Overlap
a	12k	95%
b	26k	68%
CIRCL	10k	99%


circl-lu.misp
b.misp
a.misp
x.malurl
x.bots
phishtank.verified
phishtank.unverified
OTHER
openphish
malwarepatrol
cymru.phishingurls
cymru.malwareurls
cleanmx.viruswatch
cleanmx.phishwatch
cert-pl.sinkhole
cert-pl.arakis

		Overlap	
Instance	**IPs**	**MISPs**	**Other**
a	12k	89%	44%
b	26k	41%	45%
CIRCL	10k	97%	43%

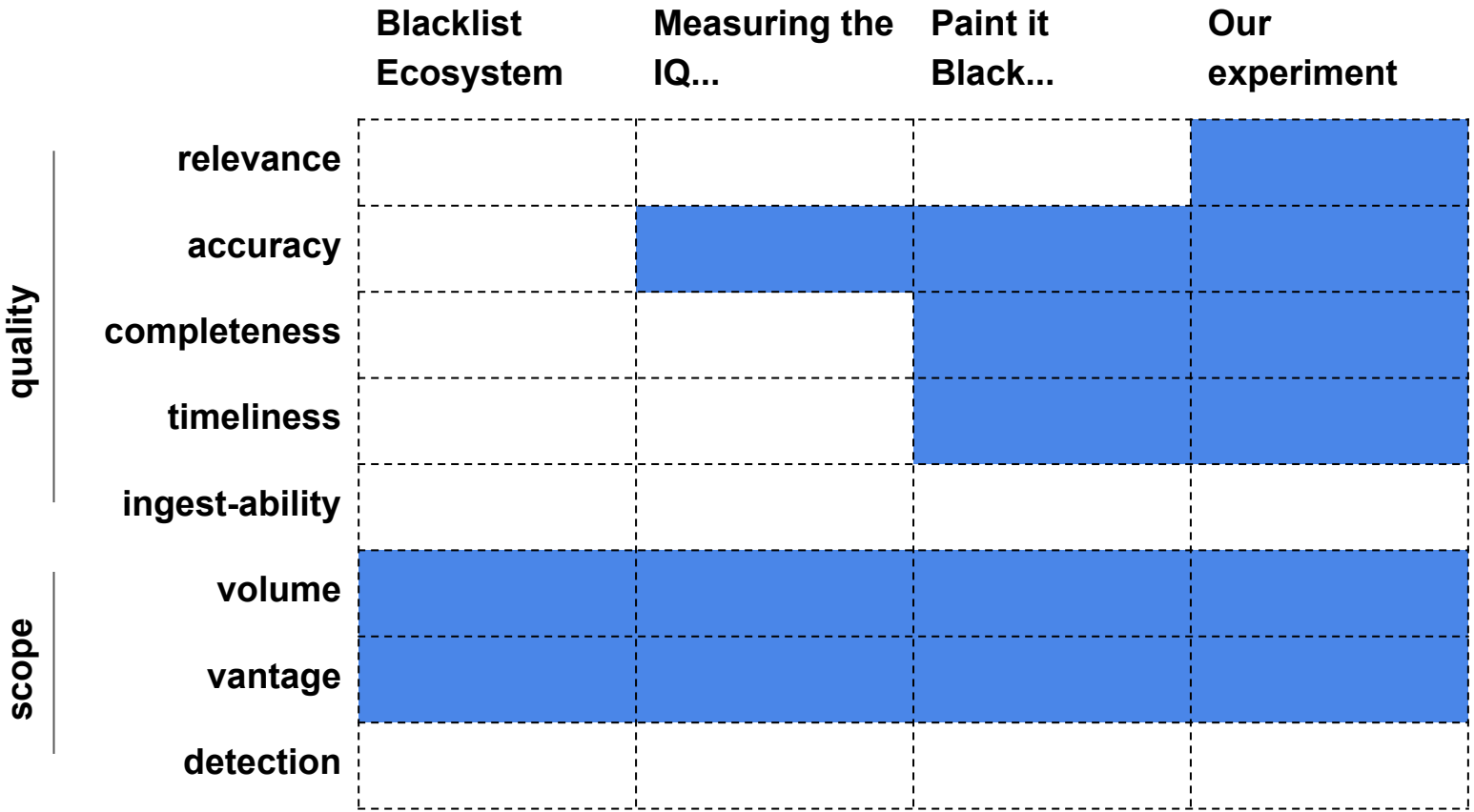
Insights

- Overlap for threats relevant to the community is higher than for blacklists
- Sharing between MISP instances - high (as suspected)
- Many indicators confirmed by public / commercial sources

Code and **data** to reproduce results:

https://github.com/pp-/feed-evaluation

(only overlap for now)


Blacklist Ecosystem
Measuring the IQ...
Paint it Black...
Our experiment
quality
relevance
accuracy
completeness
timeliness
ingest-ability
scope
volume
vantage
detection

Agenda

Motivation

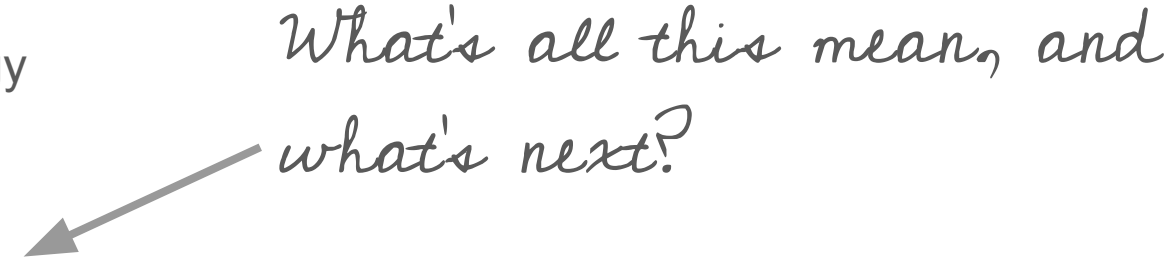
Focus

Previous work

Methodology

Results

Discussion



Agenda

Motivation

Focus

Previous work

Methodology

Results

Discussion

Conclusions

Much work remains

- Best practice guidance for measurement (this is a start)
- Integration of evaluation measurement into tools
- Decision-making framework for acquisition decisions

Are there any motivated entrepreneurs out there?

Next Steps: Interest in community efforts?

Best practice guide (methodology?) for measurement

Catalog of feeds and measurements

Plug-ins for sharing infrastructures

Ideas? Interest?

Acknowledgements (1)

Part of this research has been supported by the Strategic International Collaborative RD Promotion Project of the Ministry of Internal Affairs and Communication, Japan, and by the European Union Seventh Framework Programme (FP72007--2013) under grant agreement No. 608533 (NECOMA) and co-financed by the Ministry of Science and Higher Education, Poland (grant 3011/7.PR/13/2014/2).

The opinions expressed in this presentation are those of the authors and do not necessarily reflect the views of the Ministry of Internal Affairs and Communications, Japan, of the European Commission, or of the Ministry of Science and Higher Education, Poland.

Acknowledgements (2)

Part of this material is based upon work funded and supported by Department of Homeland Security under Contract No. FA8721-05-C-0003 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center sponsored by the United States Department of Defense.