Scoring model for IoCs by combining open intelligence feeds to reduce false positives

Authors:
Jelle Ermerins
Niek van Noort

Supervisors:
Joao de Novais Marques
Leandro Velasco

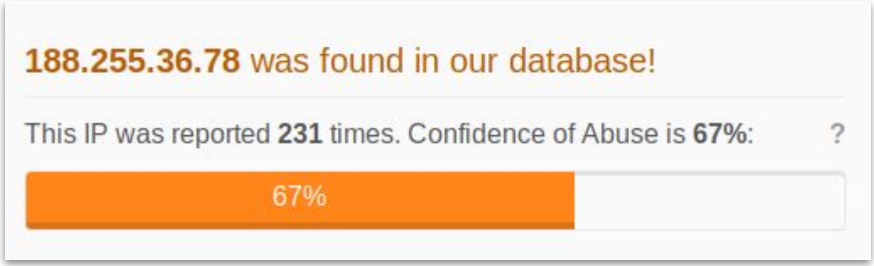
Introduction

Indicators of Compromise (IoCs) identify possible threats

The problem is false positives

Several intelligence feeds available online

Design a scoring model to reduce false positives

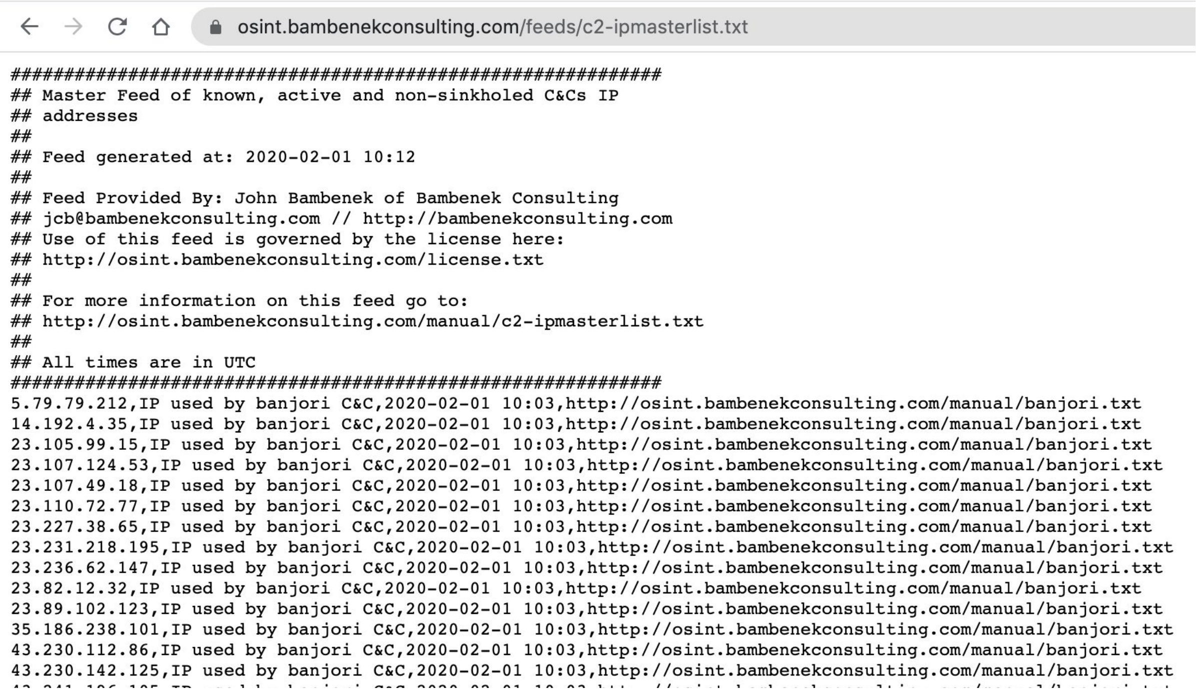


188.255.36.78 was found in our database!

This IP was reported **231** times. Confidence of Abuse is **67%**:

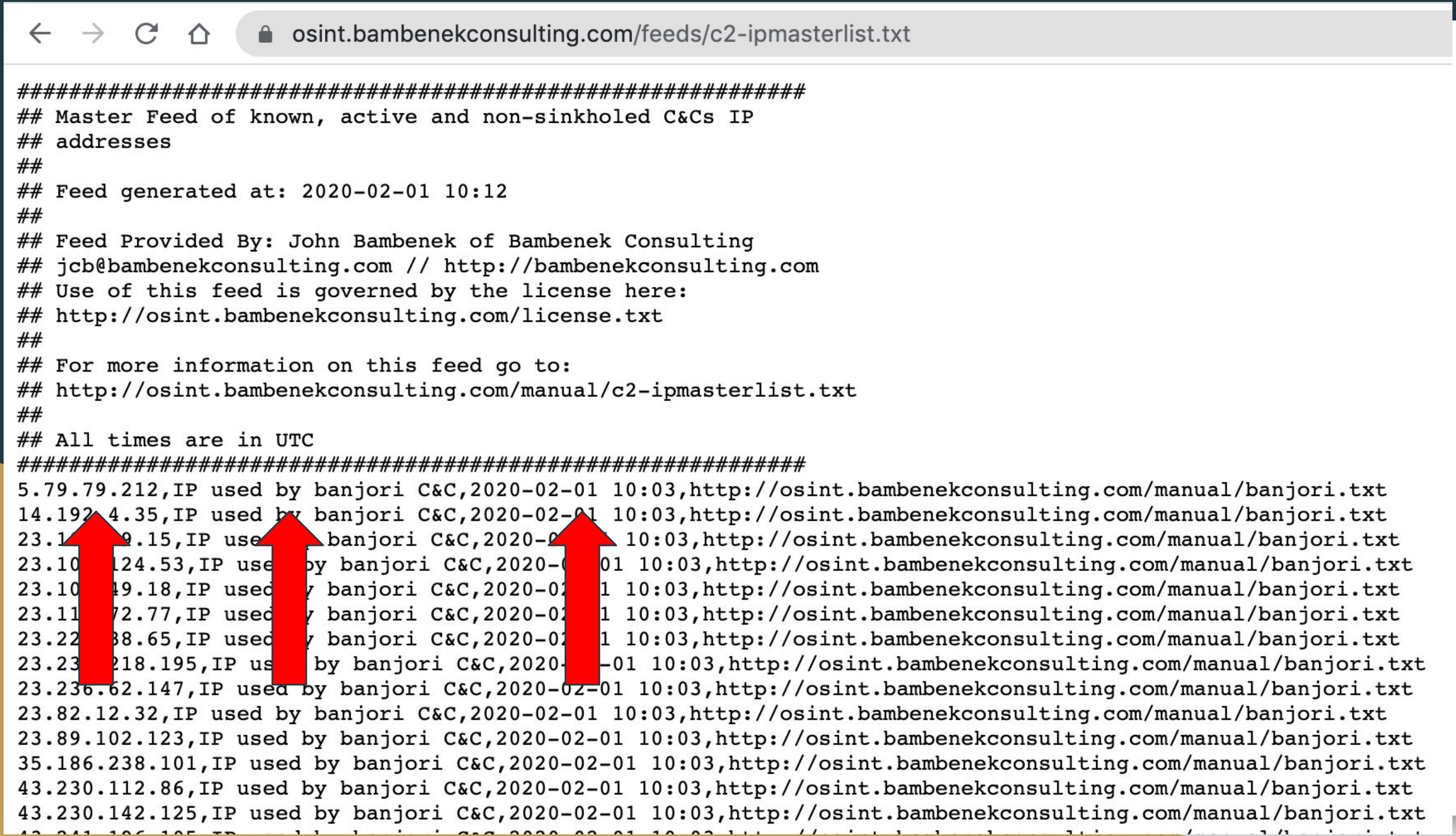
Example of an indicator of compromise (source: AbuseIPDB)

Example of an intelligence feed



osint.bambenekconsulting.com/feeds/c2-ipmasterlist.txt

```
#############################################################
## Master Feed of known, active and non-sinkholed C&Cs IP
## addresses
##
## Feed generated at: 2020-02-01 10:12
##
## Feed Provided By: John Bambenek of Bambenek Consulting
## jcb@bambenekconsulting.com // http://bambenekconsulting.com
## Use of this feed is governed by the license here:
## http://osint.bambenekconsulting.com/license.txt
##
## For more information on this feed go to:
## http://osint.bambenekconsulting.com/manual/c2-ipmasterlist.txt
##
## All times are in UTC
#############################################################
5.79.79.212,IP used by banjori C&C,2020-02-01 10:03,http://osint.bambenekconsulting.com/manual/banjori.txt
14.192.4.35,IP used by banjori C&C,2020-02-01 10:03,http://osint.bambenekconsulting.com/manual/banjori.txt
23.105.99.15,IP used by banjori C&C,2020-02-01 10:03,http://osint.bambenekconsulting.com/manual/banjori.txt
23.107.124.53,IP used by banjori C&C,2020-02-01 10:03,http://osint.bambenekconsulting.com/manual/banjori.txt
23.107.49.18,IP used by banjori C&C,2020-02-01 10:03,http://osint.bambenekconsulting.com/manual/banjori.txt
23.110.72.77,IP used by banjori C&C,2020-02-01 10:03,http://osint.bambenekconsulting.com/manual/banjori.txt
23.227.38.65,IP used by banjori C&C,2020-02-01 10:03,http://osint.bambenekconsulting.com/manual/banjori.txt
23.231.218.195,IP used by banjori C&C,2020-02-01 10:03,http://osint.bambenekconsulting.com/manual/banjori.txt
23.236.62.147,IP used by banjori C&C,2020-02-01 10:03,http://osint.bambenekconsulting.com/manual/banjori.txt
23.82.12.32,IP used by banjori C&C,2020-02-01 10:03,http://osint.bambenekconsulting.com/manual/banjori.txt
23.89.102.123,IP used by banjori C&C,2020-02-01 10:03,http://osint.bambenekconsulting.com/manual/banjori.txt
35.186.238.101,IP used by banjori C&C,2020-02-01 10:03,http://osint.bambenekconsulting.com/manual/banjori.txt
43.230.112.86,IP used by banjori C&C,2020-02-01 10:03,http://osint.bambenekconsulting.com/manual/banjori.txt
43.230.142.125,IP used by banjori C&C,2020-02-01 10:03,http://osint.bambenekconsulting.com/manual/banjori.txt
```



```
#############################################################
## Master Feed of known, active and non-sinkholed C&Cs IP
## addresses
##
## Feed generated at: 2020-02-01 10:12
##
## Feed Provided By: John Bambenek of Bambenek Consulting
## jcb@bambenekconsulting.com // http://bambenekconsulting.com
## Use of this feed is governed by the license here:
## http://osint.bambenekconsulting.com/license.txt
##
## For more information on this feed go to:
## http://osint.bambenekconsulting.com/manual/c2-ipmasterlist.txt
##
## All times are in UTC
#############################################################
5.79.79.212,IP used by banjori C&C,2020-02-01 10:03,http://osint.bambenekconsulting.com/manual/banjori.txt
14.192.4.35,IP used by banjori C&C,2020-02-01 10:03,http://osint.bambenekconsulting.com/manual/banjori.txt
23.1[illegible]9.15,IP used by banjori C&C,2020-02-01 10:03,http://osint.bambenekconsulting.com/manual/banjori.txt
23.10[illegible]124.53,IP used by banjori C&C,2020-02-01 10:03,http://osint.bambenekconsulting.com/manual/banjori.txt
23.10[illegible]9.18,IP used by banjori C&C,2020-02-01 10:03,http://osint.bambenekconsulting.com/manual/banjori.txt
23.11[illegible]72.77,IP used by banjori C&C,2020-02-01 10:03,http://osint.bambenekconsulting.com/manual/banjori.txt
23.22[illegible]8.65,IP used by banjori C&C,2020-02-01 10:03,http://osint.bambenekconsulting.com/manual/banjori.txt
23.23[illegible]218.195,IP used by banjori C&C,2020-02-01 10:03,http://osint.bambenekconsulting.com/manual/banjori.txt
23.236.62.147,IP used by banjori C&C,2020-02-01 10:03,http://osint.bambenekconsulting.com/manual/banjori.txt
23.82.12.32,IP used by banjori C&C,2020-02-01 10:03,http://osint.bambenekconsulting.com/manual/banjori.txt
23.89.102.123,IP used by banjori C&C,2020-02-01 10:03,http://osint.bambenekconsulting.com/manual/banjori.txt
35.186.238.101,IP used by banjori C&C,2020-02-01 10:03,http://osint.bambenekconsulting.com/manual/banjori.txt
43.230.112.86,IP used by banjori C&C,2020-02-01 10:03,http://osint.bambenekconsulting.com/manual/banjori.txt
43.230.142.125,IP used by banjori C&C,2020-02-01 10:03,http://osint.bambenekconsulting.com/manual/banjori.txt
```

Related work

A scoring model was designed by researchers from CIRCL (Luxembourg)

- Using a **decay rate**
- The score of an IoC decays over time

T. Schaberreiter et al. designed another scoring model

- Comparing different sources
- Using features like extensiveness, timeliness, completeness

No research on dependency between intelligence feeds
No practical research

Research questions (Challenges of designing the scoring model)

How can we use multiple open intelligence feeds in a scoring model to determine the quality of IoCs?

How independent are different intelligence feeds from each other?

How do we make the model time dependent?

How do we decide if we can trust an intelligence feed?

How do we calculate one score from multiple feeds with different levels of trust?

How independent are different intelligence feeds from each other?

Independence and overlap of feeds

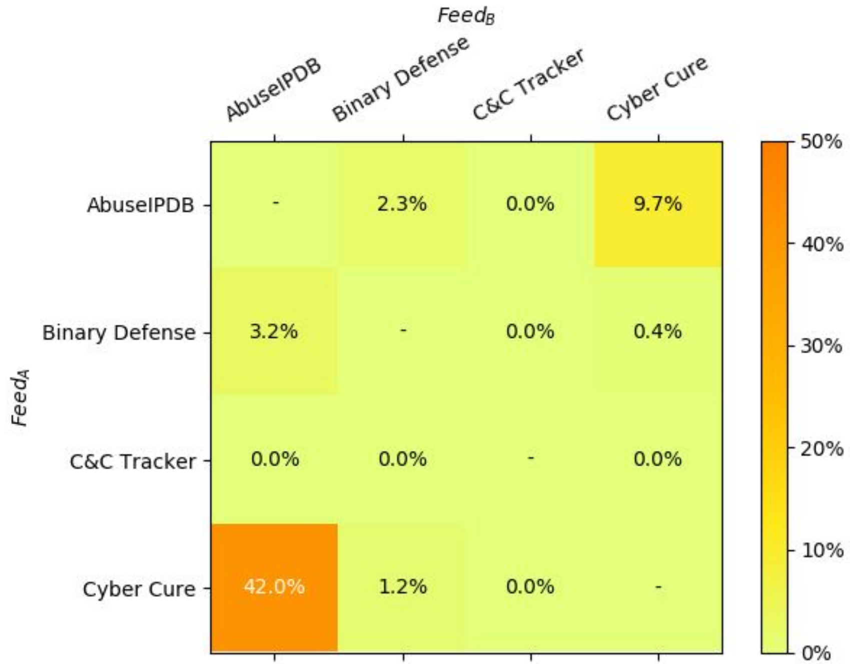
Overlap is important

But intelligence feeds need to be independent

Used intelligence feeds:

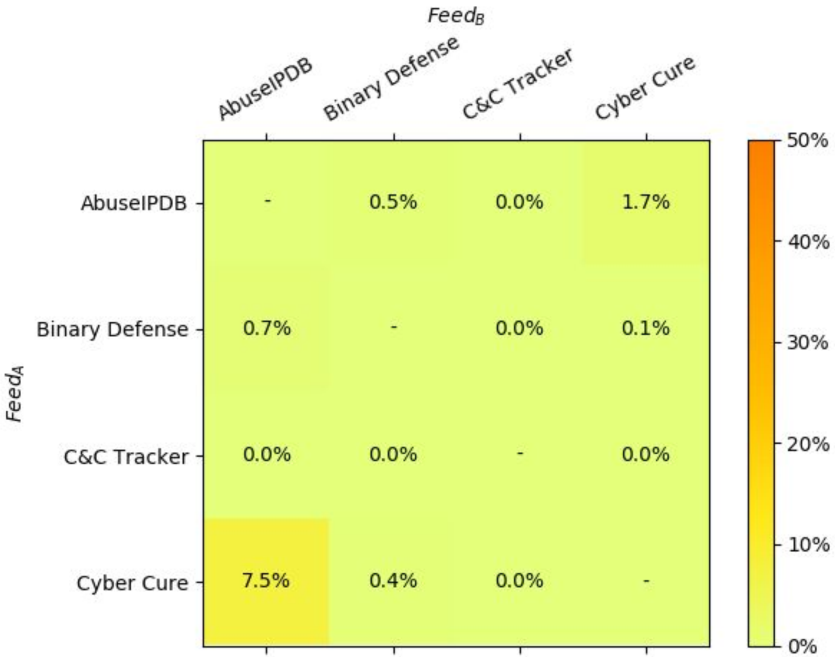
- AbuseIPDB
- Binary Defense Banlist
- C&C Tracker
- Cyber Cure

Overlap matrix of the intelligence feeds



	AbuseIPDB	Binary Defense	C&C Tracker	Cyber Cure
AbuseIPDB	-	2.3%	0.0%	9.7%
Binary Defense	3.2%	-	0.0%	0.4%
C&C Tracker	0.0%	0.0%	-	0.0%
Cyber Cure	42.0%	1.2%	0.0%	-

Overlap matrix, where difference in first sighting is smaller than a day



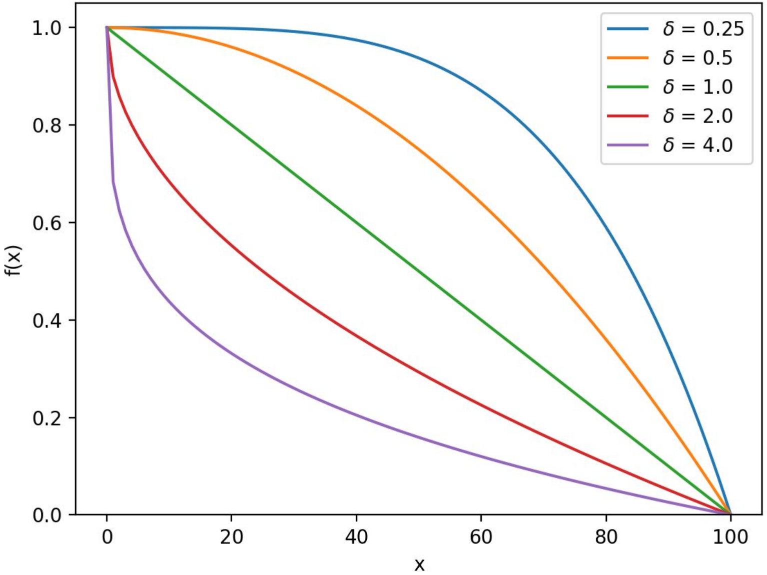
$Feed_A$ \ $Feed_B$	AbuseIPDB	Binary Defense	C&C Tracker	Cyber Cure
AbuseIPDB	-	0.5%	0.0%	1.7%
Binary Defense	0.7%	-	0.0%	0.1%
C&C Tracker	0.0%	0.0%	-	0.0%
Cyber Cure	7.5%	0.4%	0.0%	-

How do we make the model time dependent?

Decay time

IoC will lose value over time when it hasn't been seen

$$f(x) = max(0,\ 1 - (\frac{x}{\tau})^{\frac{1}{\delta}})$$



Decay function with different **δ** parameter values and a fixed **τ** value of 100

How do we decide if we can trust an intelligence feed?

Source confidence

Quality of the source based on some features

Extensiveness

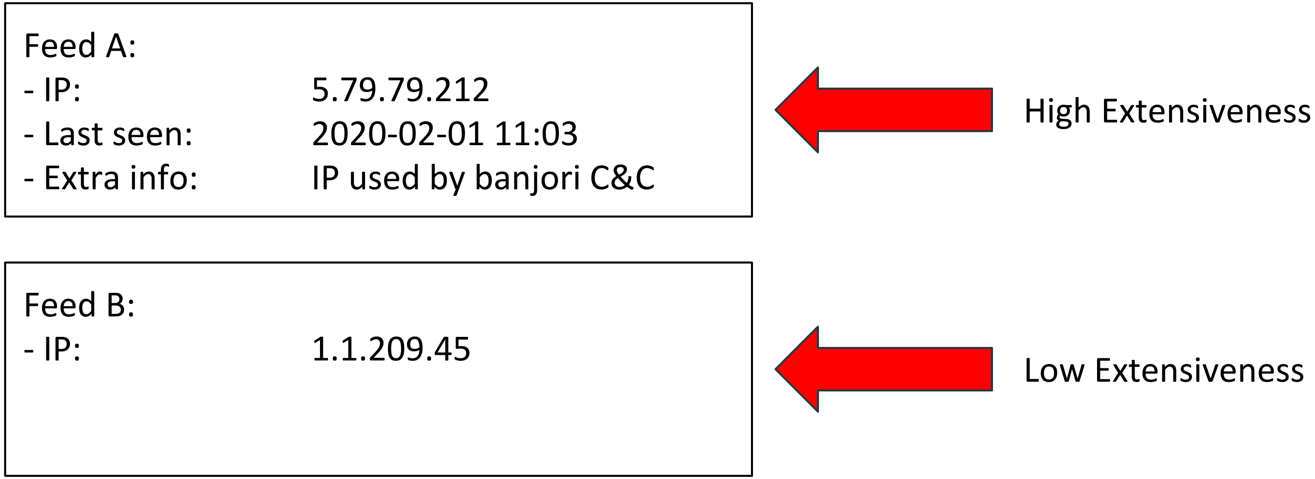
Timeliness

Completeness

Whitelist Overlap Score

Extensiveness

How many properties does the intelligence feed provide?



Feed A:

- IP:	5.79.79.212
- Last seen:	2020-02-01 11:03
- Extra info:	IP used by banjori C&C

High Extensiveness

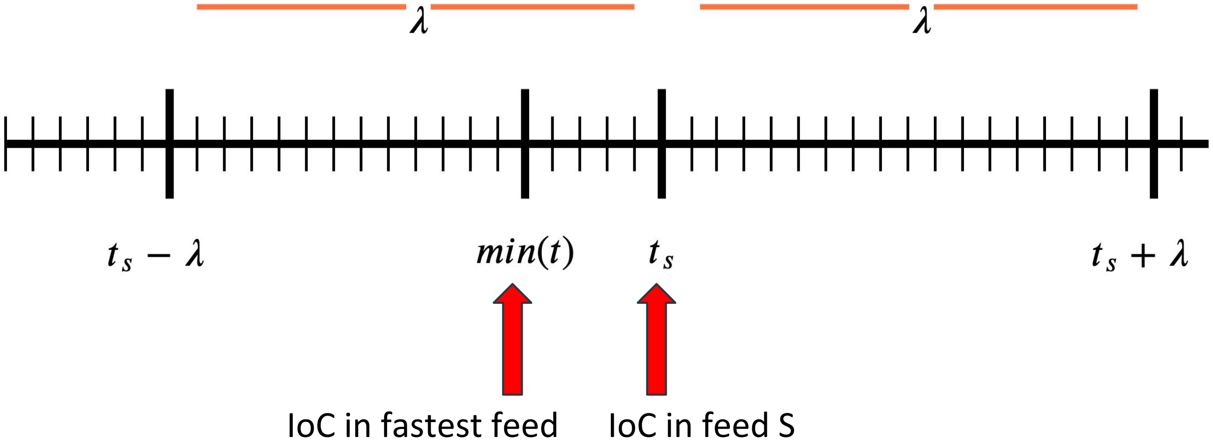
Feed B:

- IP:	1.1.209.45

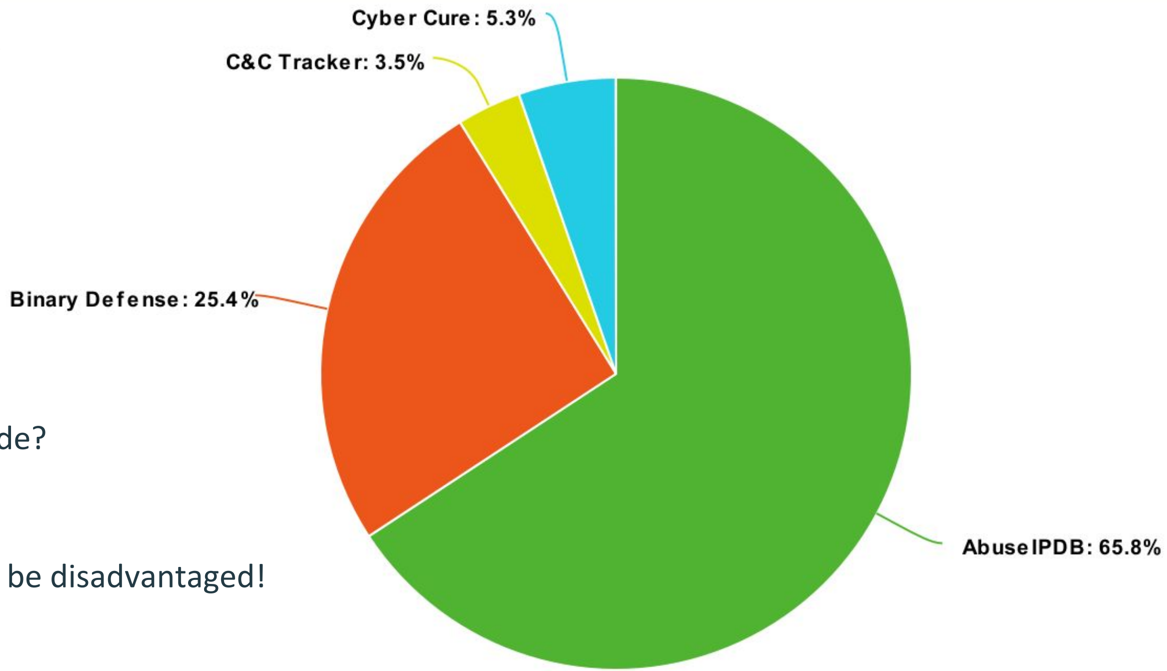
Low Extensiveness

Timeliness

How fast is the intelligence feed?



Completeness



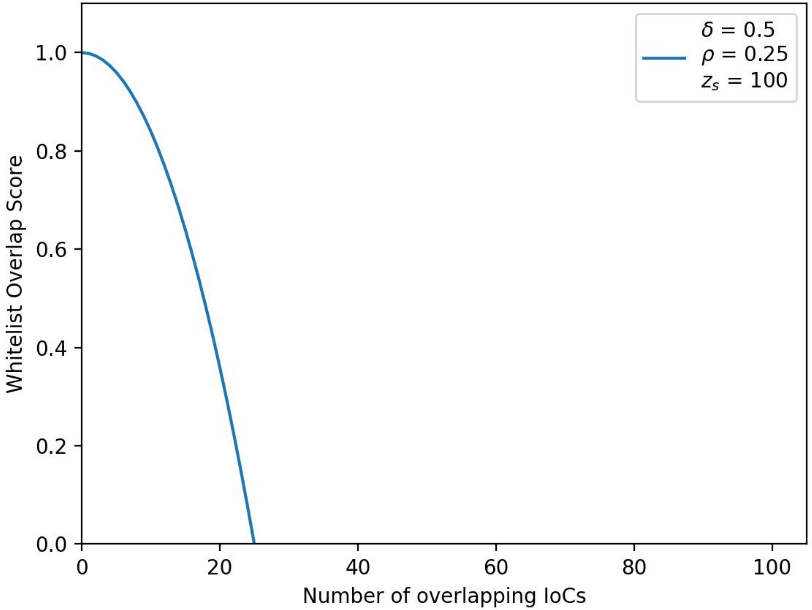
How many IoCs does the feed provide?

Trustworthy small scale feeds could be disadvantaged!

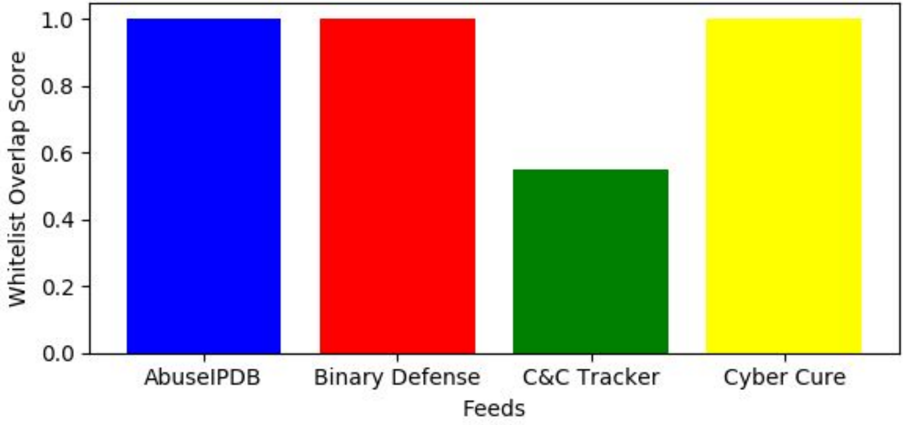
Whitelist Overlap Score

Does the feed have overlap with a whitelist?

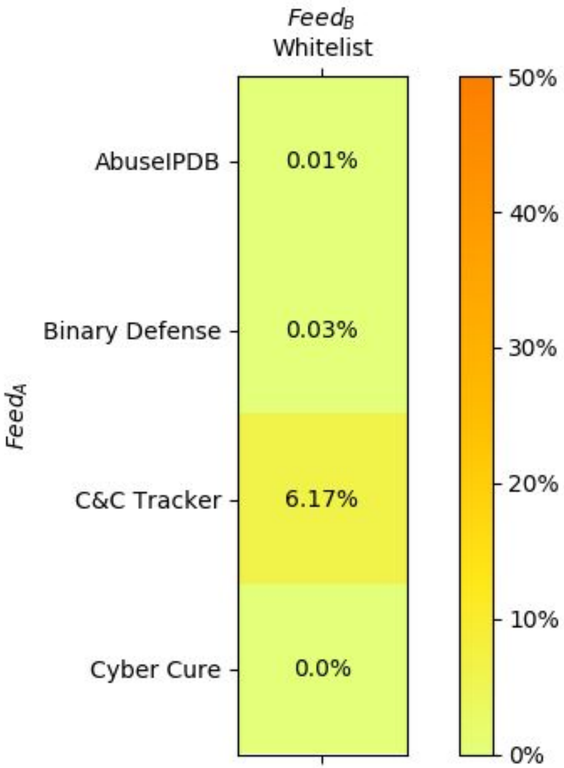
$$max(0,\ 1 - (\frac{u_s}{z_s \cdot \rho})^{\frac{1}{\delta}})$$



Whitelist Overlap Score



Whitelist overlap score of our feeds. (ρ = 0.1)



Whitelist overlap percentage

The Source Confidence

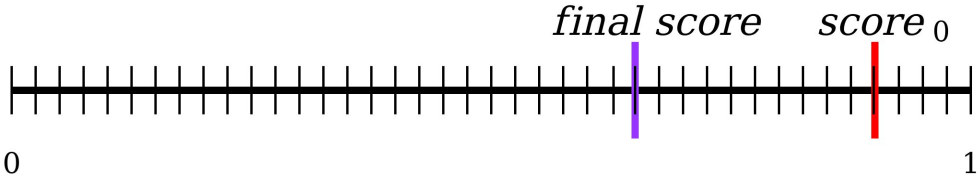
Weighted mean of:	Weight:
- Extensiveness	0.8
- Timeliness	0.6
- Completeness	0.0
- Whitelist Overlap Score	1.0

How do we calculate one score from multiple feeds with different levels of trust?

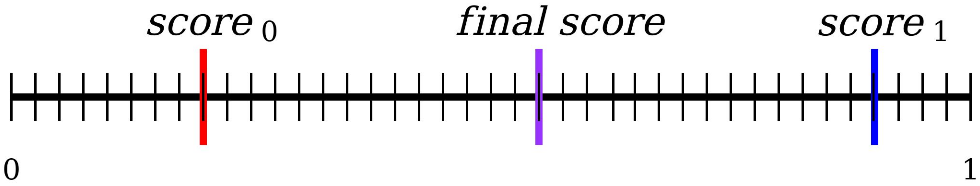
Final Score Calculation

$$final\_score = \frac{1}{N}\sum_{i=0}^{N} source\_confidence_i \cdot score_i$$

Advantage: The source confidence is still useful when an IoC is found in one feed only.



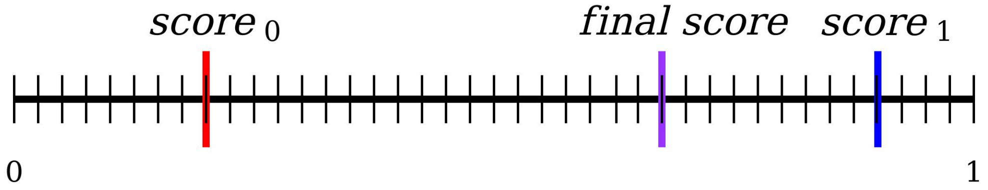
Disadvantage: Each intelligence feed has the same amount of influence on the final score.



Final Score Calculation

$$final\_score = \frac{\sum_{i=0}^{N} source\_confidence_i \cdot score_i}{\sum_{i=0}^{N} source\_confidence_i}$$

Advantage: The source confidence works as a weight on the final score per feed.



Disadvantage: The source confidence is useless when an IoC is found in one feed only:

$$final\_score = \frac{\cancel{\sum_{i=0}^{N} source\_confidence_i} \cdot score_i}{\cancel{\sum_{i=0}^{N} source\_confidence_i}}$$

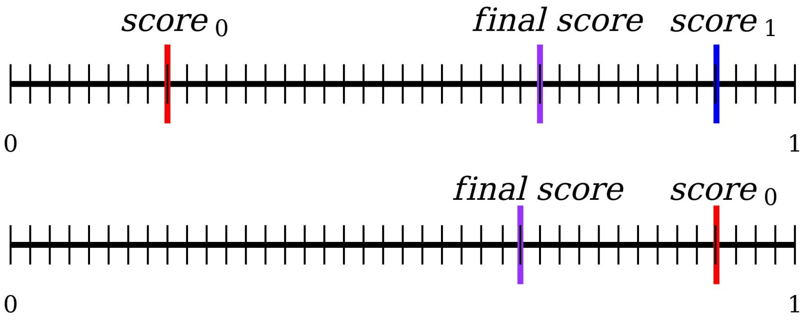
Final Score Calculation

A square has been added

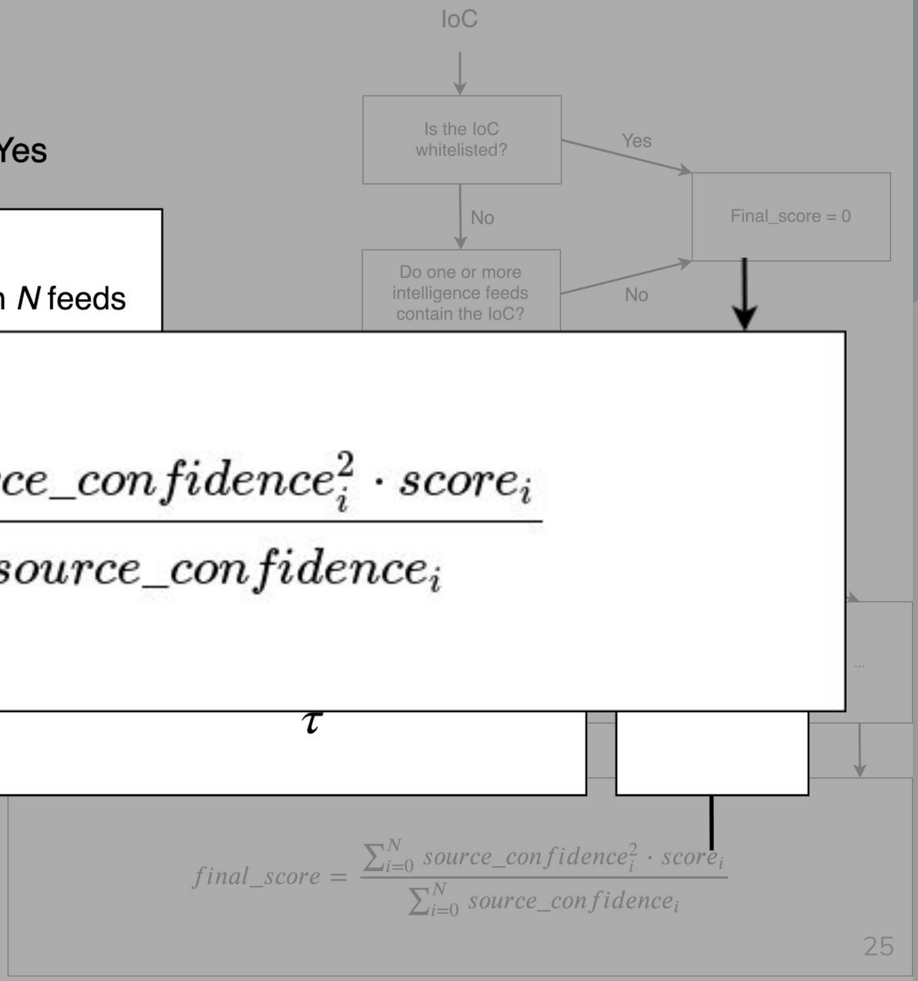
Solution: Combine the two previous functions:

$$final\_score = \frac{\sum_{i=0}^{N} source\_confidence_i^2 \cdot score_i}{\sum_{i=0}^{N} source\_confidence_i}$$

We have both advantages:

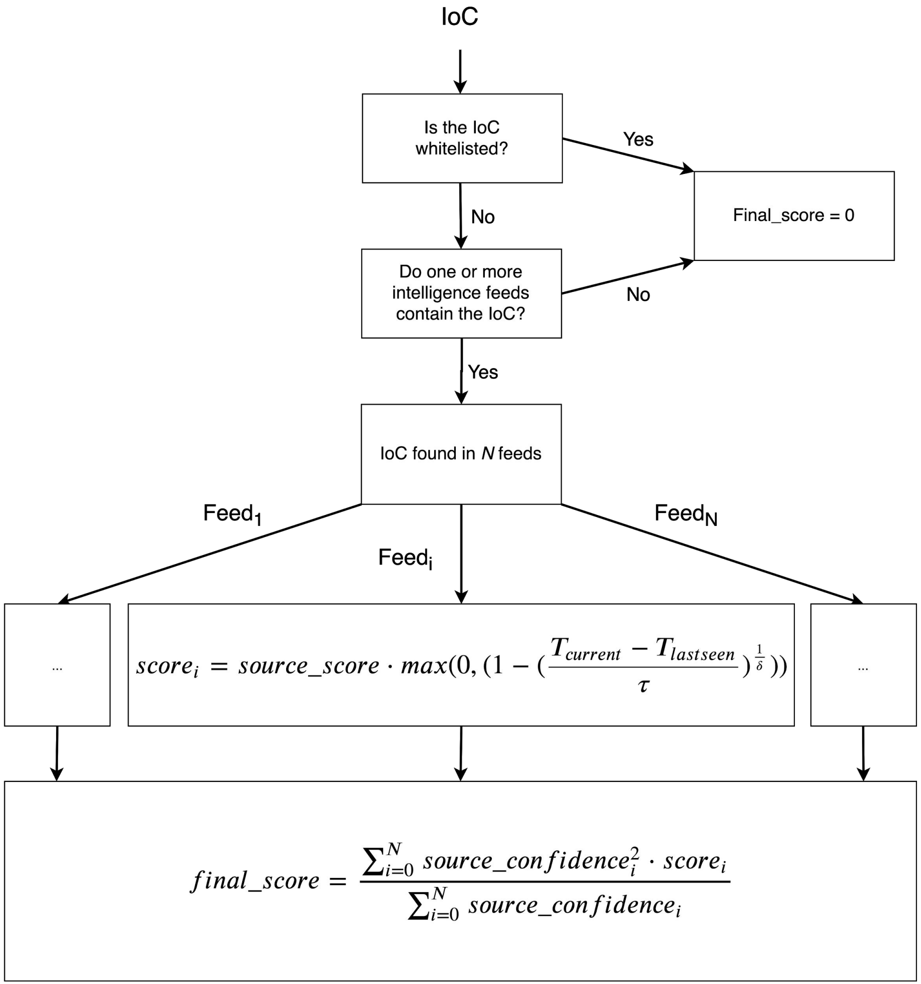


The scoring model



$$final\_score = \frac{\sum_{i=0}^{N} source\_confidence_i^2 \cdot score_i}{\sum_{i=0}^{N} source\_confidence_i}$$

The scoring model



IoC

Is the IoC whitelisted?

Yes

Final_score = 0

No

Do one or more intelligence feeds contain the IoC?

No

Yes

IoC found in *N* feeds

Feed$_1$

Feed$_i$

Feed$_N$

...

$$score_i = source\_score \cdot max(0, (1 - (\frac{T_{current} - T_{lastseen}}{\tau})^{\frac{1}{\delta}}))$$

...

$$final\_score = \frac{\sum_{i=0}^{N} source\_confidence_i^2 \cdot score_i}{\sum_{i=0}^{N} source\_confidence_i}$$

How can we use intelligence feeds in a scoring model to determine the quality of IoCs?

Conclusion

How independent are different intelligence feeds from each other?

The feeds are independent

We want more independent feeds with overlap

How do we make the model time dependent?

Decay rate

How do we decide if we can trust an intelligence feed?

Trust based on extensiveness, timeliness, ~~completeness~~ and whitelist correlation

How do we calculate one score from multiple feeds with different levels of trust?

Source confidence as weight for the feed

And also as part of the IoC score itself

Future work

Parameter optimization

Other characteristics for the source confidence

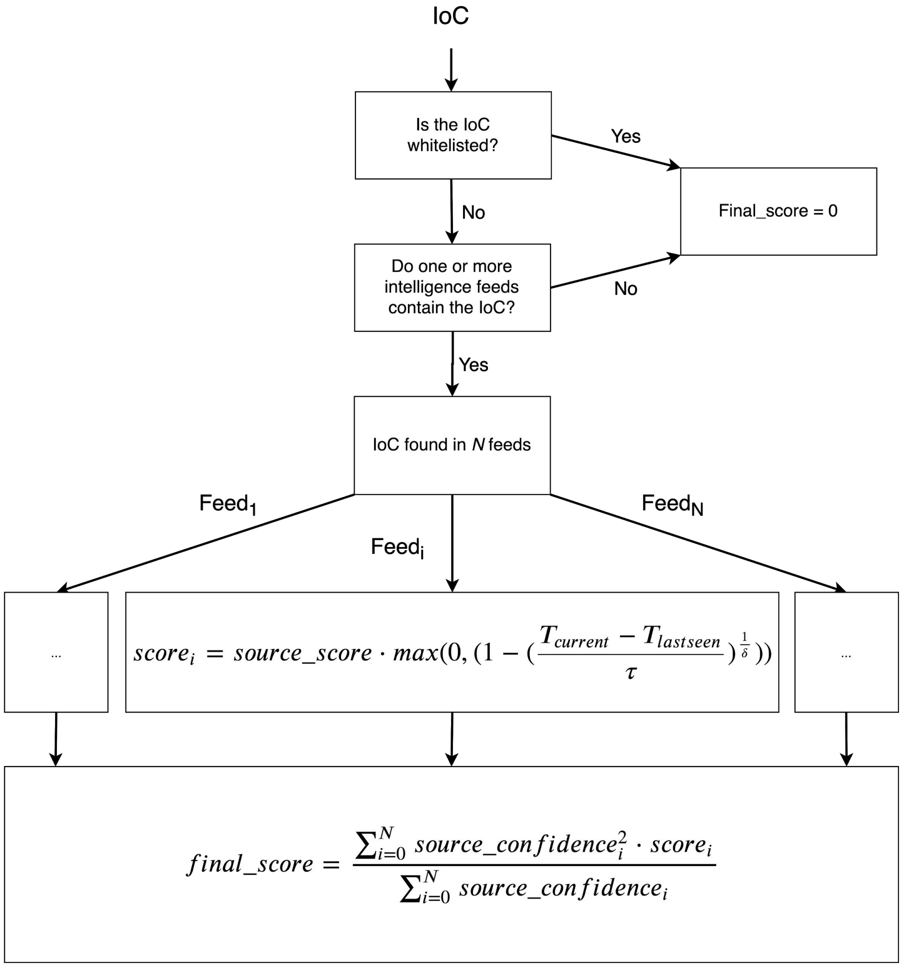
Other intelligence feeds

Scoring whitelists

Thank you!

And special thanks to:

Joao de Novais Marques
Leandro Velasco



IoC

Is the IoC whitelisted?

Yes

Final_score = 0

No

Do one or more intelligence feeds contain the IoC?

No

Yes

IoC found in *N* feeds

Feed$_1$

Feed$_i$

Feed$_N$

...

$$score_i = source\_score \cdot max(0, (1 - (\frac{T_{current} - T_{lastseen}}{\tau})^{\frac{1}{\delta}}))$$

...

$$final\_score = \frac{\sum_{i=0}^{N} source\_confidence_i^2 \cdot score_i}{\sum_{i=0}^{N} source\_confidence_i}$$