BASS Automated Signature Synthesizer

Jonas Zaddach
@jzaddach

Mariano Graziano
@emd3l

INTRODUCTION

Mariano Graziano

Jonas Zaddach

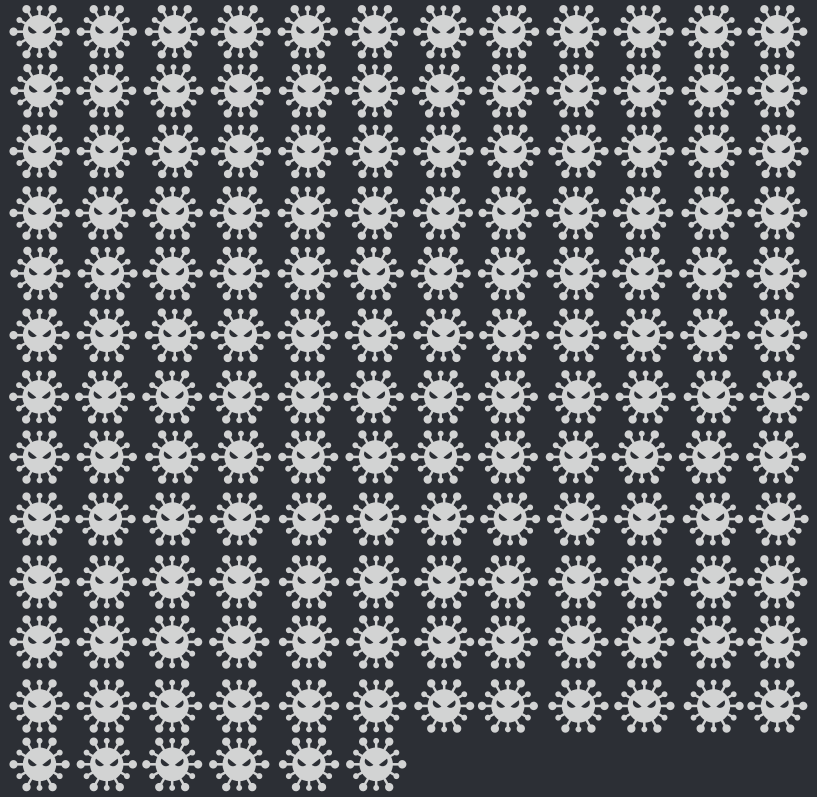
Security Researchers in

cisco > TALOS > ClamAV®

TALOS

LET'S TALK ABOUT THE THREAT LANDSCAPE

THREAT LANDSCAPE



= 10,000

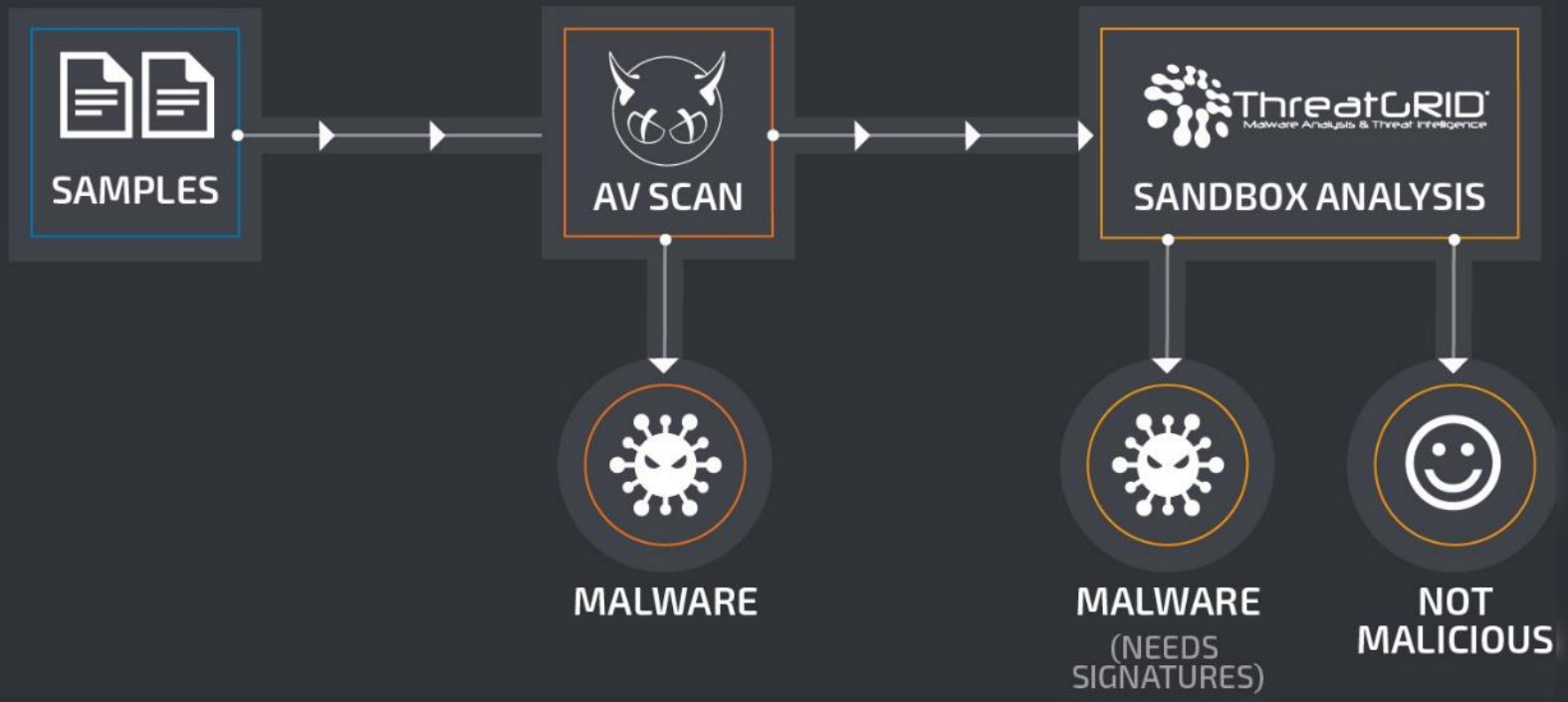
1.5 MILLION

Malware Samples

DAILY

TALOS

AV PIPELINE OVERVIEW



TALOS

MALWARE DETECTION CHALLENGE

≈ 560,000 signatures
over a 3-month period

≈ 9,500 Signatures

DAILY

→ Huge number of signatures
→ Pattern-based signatures can reduce resource footprint compared to hash-based signatures

TALOS


bass
AUTOMATED SIGNATURE SYNTHESIZER

BASS OVERVIEW

TALOS

CLUSTERING

- Clustering is NOT a part of BASS!
- Several cluster sources feed BASS
  - Sandbox Indicator of Compromise (IoC) clustering
  - Structural hashing
  - Spam campaign dataset

UNPACKING & INSPECTION

- Extract all content ClamAV can extract
  - ZIP archives
  - Email attachments
  - Packed executables
  - Nested documents: e.g., PE file inside a Word document
  - …
- Gather information about file content
  - File size
  - Mime type/Magic string
  - …

TALOS

FILTERING

- Reject clusters with wrong file types
  - In the near future BASS will handle any executable file type handled by the disassembler (IDA Pro)
  - Currently limited to PE executables
- Clean outliers with wrong file types from clusters

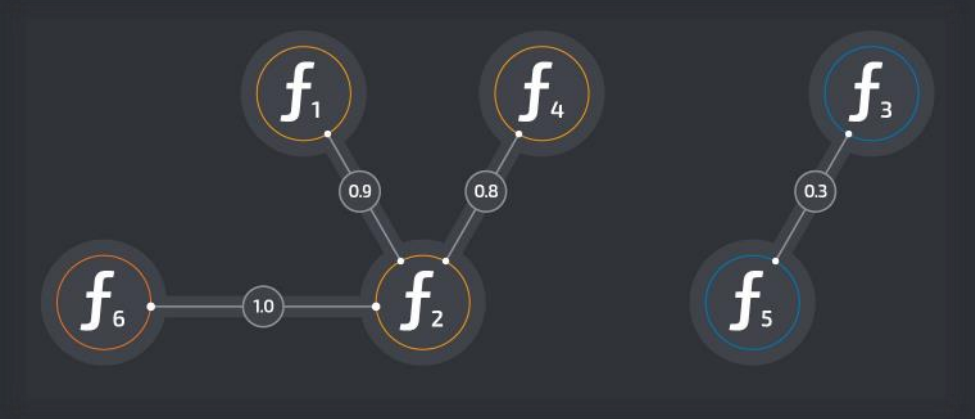
SIGNATURE GENERATION

DISASSEMBLING

- Export disassembly database
- Currently uses IDA Pro as a disassembler
  - Others are possible in the future

FINDING COMMON CODE

- Use binary diffing to identify similar functions across binaries
- Build similarity graph between functions and extract largest connected subgraph



TALOS

FINDING COMMON CODE

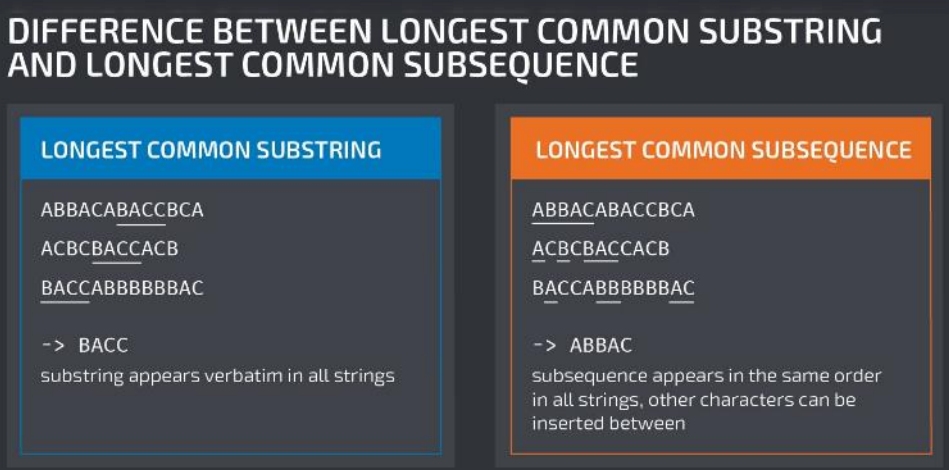
- Test found function against a database of whitelisted functions
  - Kam1n0, a database for binary code clone search, contains functions of whitelisted samples
  - If a found function is whitelisted, take the next-best subgraph

Kam1n0

TALOS

FINDING AN LCS

- Use k-LCS algorithm to find a longest common subsequence



LCS

- Implemented Hamming-kLCS described by C. Blichmann [1]

TALOS

FINDING AN LCS

- Hamming distance between all strings is computed
- 2-LCS algorithm (Hirschberg algorithm) is applied to strings with lowest distance
- Resulting LCS is kept → Rinse and repeat

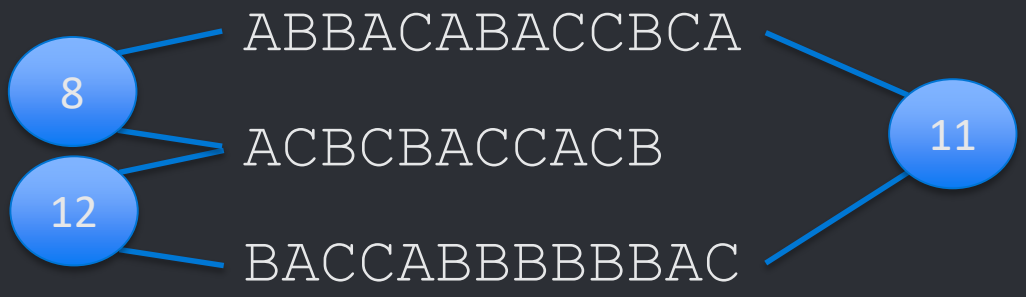
```
ABBACABACCBCA

ACBCBACCACB

BACCABBBBBBAC
```

FINDING AN LCS

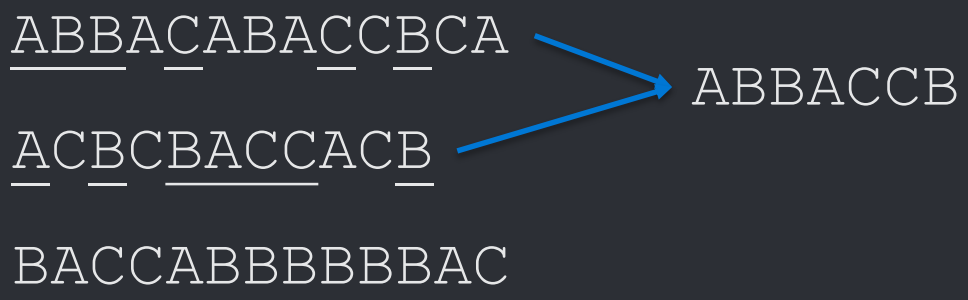
- Hamming distance between all strings is computed
- 2-LCS algorithm (Hirschberg algorithm) is applied to strings with lowest distance
- Resulting LCS is kept → Rinse and repeat



TALOS

FINDING AN LCS

- Hamming distance between all strings is computed
- 2-LCS algorithm (Hirschberg algorithm) is applied to strings with lowest distance
- Resulting LCS is kept → Rinse and repeat



FINDING AN LCS

- Hamming distance between all strings is computed
- 2-LCS algorithm (Hirschberg algorithm) is applied to strings with lowest distance
- Resulting LCS is kept → Rinse and repeat

```
ABBACCB

BACCABBBBBBAC
```

TALOS

FINDING AN LCS

- Hamming distance between all strings is computed
- 2-LCS algorithm (Hirschberg algorithm) is applied to strings with lowest distance
- Resulting LCS is kept → Rinse and repeat

TALOS

GENERATING A SIGNATURE

- Create ClamAV signature
  - Find possible “gaps” in result sequence
  - Delete single characters
- Find a common name
  - Use AvClass to label cluster

SIGNATURE: Win.Trojan.Example:0:*cafebabe*dead*beef

ORIGINAL FILES:

```
1. ca fe ba be 31 de ad 35 37 be ef 38 |....1..57..8|
2. 31 ca fe ba be de ad be ef 35 38 37 |1........587|
3. 30 ca fe ba be 31 de ad 37 be ef 38 |0....1..7..8|
```

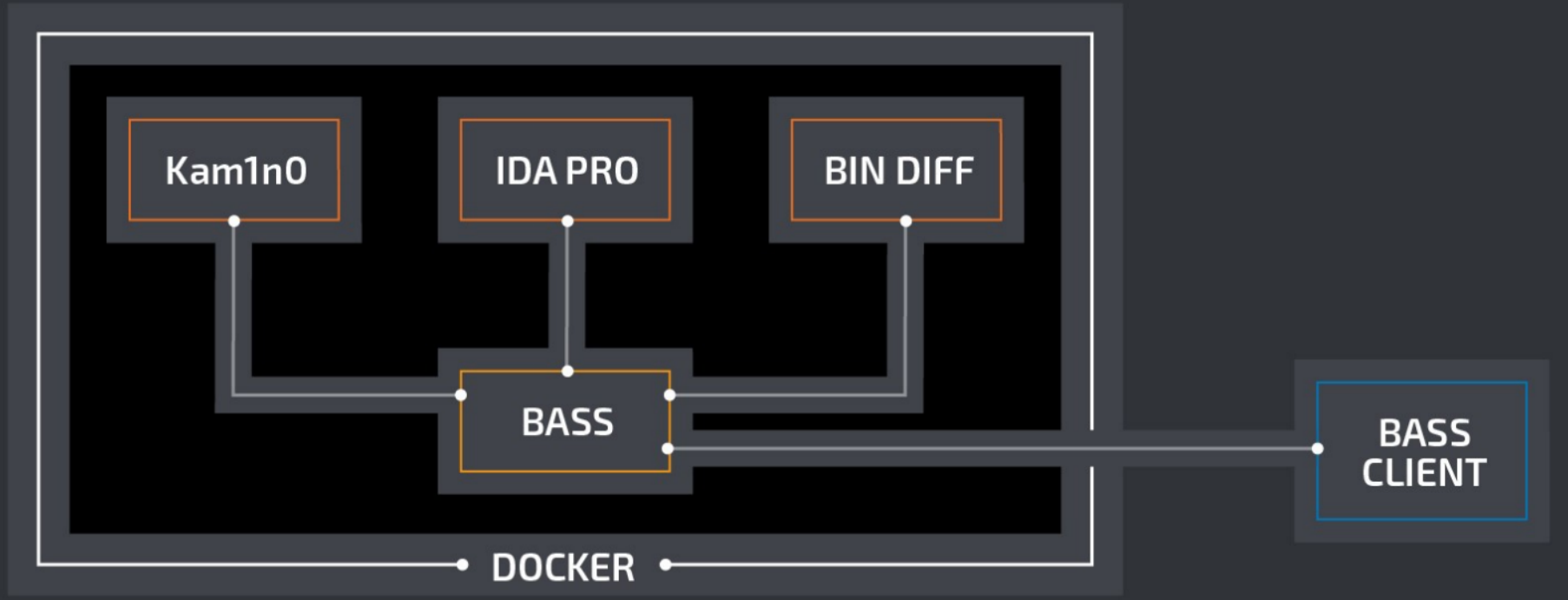
TALOS

VALIDATION

- False Positive testing
  - Against a set of known clean binaries
- Manual validation by Analyst
  - Assisted by CASC plugin [4]
  - Matched binary parts are highlighted in IDA Pro

TALOS

TECHNICAL IMPLEMENTATION



TALOS

DEMO

CONCLUSION

LIMITATIONS

- Only works for executables
- Does not work well for
  - File infectors (Small, varying snippets of malicious code)
  - Backdoors (Clean functions mixed with malicious ones)
- Alpha stage

CONCLUSION

- Presented automated signature generation system for executables
- Implemented research ideas not available as code
  - VxClass from Zynamics
- Code will be available open-source
  - For others to try, improve and comment on

https://github.com/CISCO-TALOS/bass

TALOS


Talos
talosintel.com
blogs.cisco.com/talos
@talossecurity
CISCO

RESOURCES

1. "*Automatisierte Signaturgenerierung für Malware-Stämme*", Christian Blichmann
   https://static.googleusercontent.com/media/www.zynamics.com/en//downloads/blichmann-christian--diplomarbeit--final.pdf
2. "*AVClass: A Tool for Massive Malware Labeling*", Sebastian et al., https://software.imdea.org/~juanca/papers/avclass_raid16.pdf
3. "*Kam1n0: MapReduce-based Assembly Clone Search for Reverse Engineering*", Ding et al., http://www.kdd.org/kdd2016/papers/files/adp0461-dingAdoi.pdf
4. CASC IDA Pro plugin, https://github.com/Cisco-Talos/CASC
5. VxClass – Automated classification of malware and trojans into families https://www.zynamics.com/vxclass.html

TALOS