Detecting emerging malware in the cloud before VirusTotal can see it

Andreas Pfadler, Anastasia Poliakova
Gan Feng, Thanh Nguyen, Ali Fakeri-Tabrizi, Hongliang Liu, and Yuriy Yuzifovich

Alibaba Cloud and **DAMO Academy of Alibaba**

Botconf 2021/2022, Nantes, France April 2022

達摩院 ALIBABA DAMO ACADEMY

Alibaba Cloud

Who are we?

- Security Innovation Lab of Alibaba Cloud (SIL)

  Anastasia Poliakova
  Gan Feng
  Thanh Nguyen, PhD
  Ali Fakeri-Tabrizi, PhD
  Hongliang Liu, PhD
  Yuriy Yuzifovich

- DAMO Academy of Alibaba

  Andreas Pfadler, PhD

Your speakers today

Anastasia Poliakova

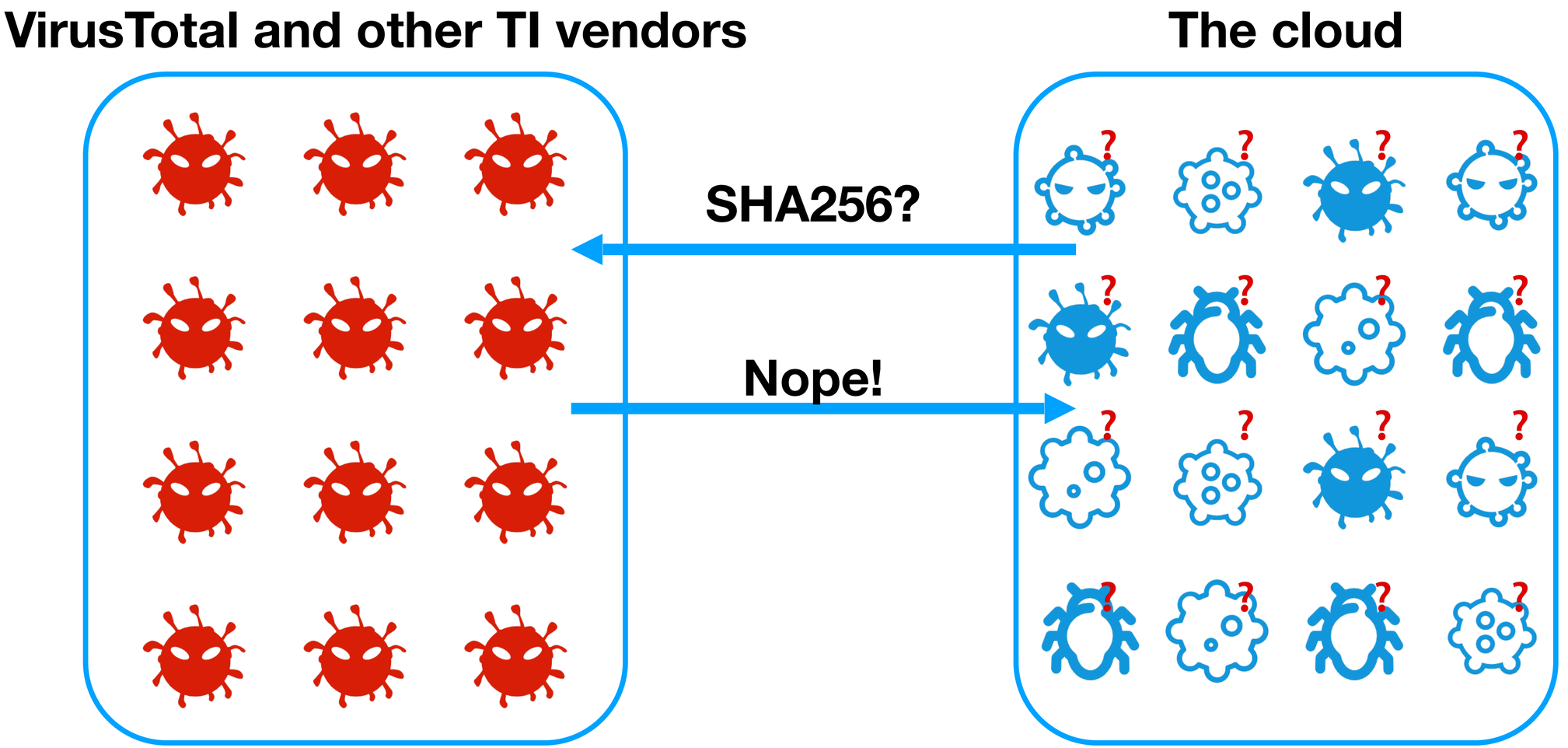
Andreas Pfadler, PhD

Yuriy Yuzifovich

Why: security reasons

1. Third-party validation comes with a price:
   - Latency
   - License coast
   - Dependency
2. Ambiguous VT results interpretation
3. Malware targeting Chinese customers is underreported to VT
4. Cloud-specific threats – is it malware or customer’s binary?

A Tale of Two Cities



Fuzzy hash: ssdeep

Distance(ssdeep1, ssdeep2) ~ portion of shared code

Block size	Hash1	Hash2
98304	Di4jwJ2zVPiuNMuE1y2qldX6i/hIUA+nCvw3c7Mk90oxK0	u4NpPiMMt1y2AdX6i/hIal3ajkoB
98304	xi4jwJ2zVPiuNMuE1y2qldX6i/hIUA+nCvw3c7Mk90oxK	I4NpPiMMt1y2AdX6i/hIal3ajko
98304	Ti4jwJ2zVPiuNMuE1y2qldX6i/hIUA+nCvw3c7Mk90vxK	e4NpPiMMt1y2AdX6i/hIal3ajkv
98304	hi4jwJ2zVPiuNMuE1y2qldX6i/hIUA+nCvw3c7Mk90oxKy	Y4NpPiMMt1y2AdX6i/hIal3ajkoP
98304	xi4jwJ2zVPiuNMuE1y2qldX6i/hIUA+nCvw3c7Mk900xK	I4NpPiMMt1y2AdX6i/hIal3ajk0

ssdeep fuzzy hash for Phoenix Miner family

Ref https://www.virustotal.com/gui/file/599393e258d8ba7b8f8633e20c651868258827d3a43a4d0712125bc487eabf92

Why no one has scaled it up?

Enough samples?

Algorithm and computing power?

Use cases?

Great team?

We are a cloud provider and we have all four, so, why not?

Sample collection

- From our cloud security product, we collect binary file hashes: MD5, SHA256 and ssdeep
- Ssdeep library size on our cloud is 100 million and counting

```
MD5:a4ae6d3308ba52770bf6f8aa429e2a6c

SHA256:e852258786ef31204997cd8f1b7392d2be615b7effe0a8819065cb2c8e65218a

SSDEEP:196608:0lUsEAfUY7F0e6KtTrYP7GILBqMFE2844DYIv8ZFbImQpRTUsliOw8:Js
FUYJUKtTrYSILBqMFE2844DYIv8ZF9
```

Engineering optimization

- Pairwise ssdeep comparison as Levenshtein edit distance

  - O(n*n) or O(m*n) if using dynamic programing

  - When it comes to 100M samples and their pairs, challenges appear

Without any loss of generality, here is how we did it

Pre-filtering: ssdeep normalization

Redundancy cleanup

Pre-filtering: space reduction

Block size ratio filter

7-gram block filter

7-gram heap construction

Optimized edit distance

Modified bitap with Wu, Manber* for Levenshtein

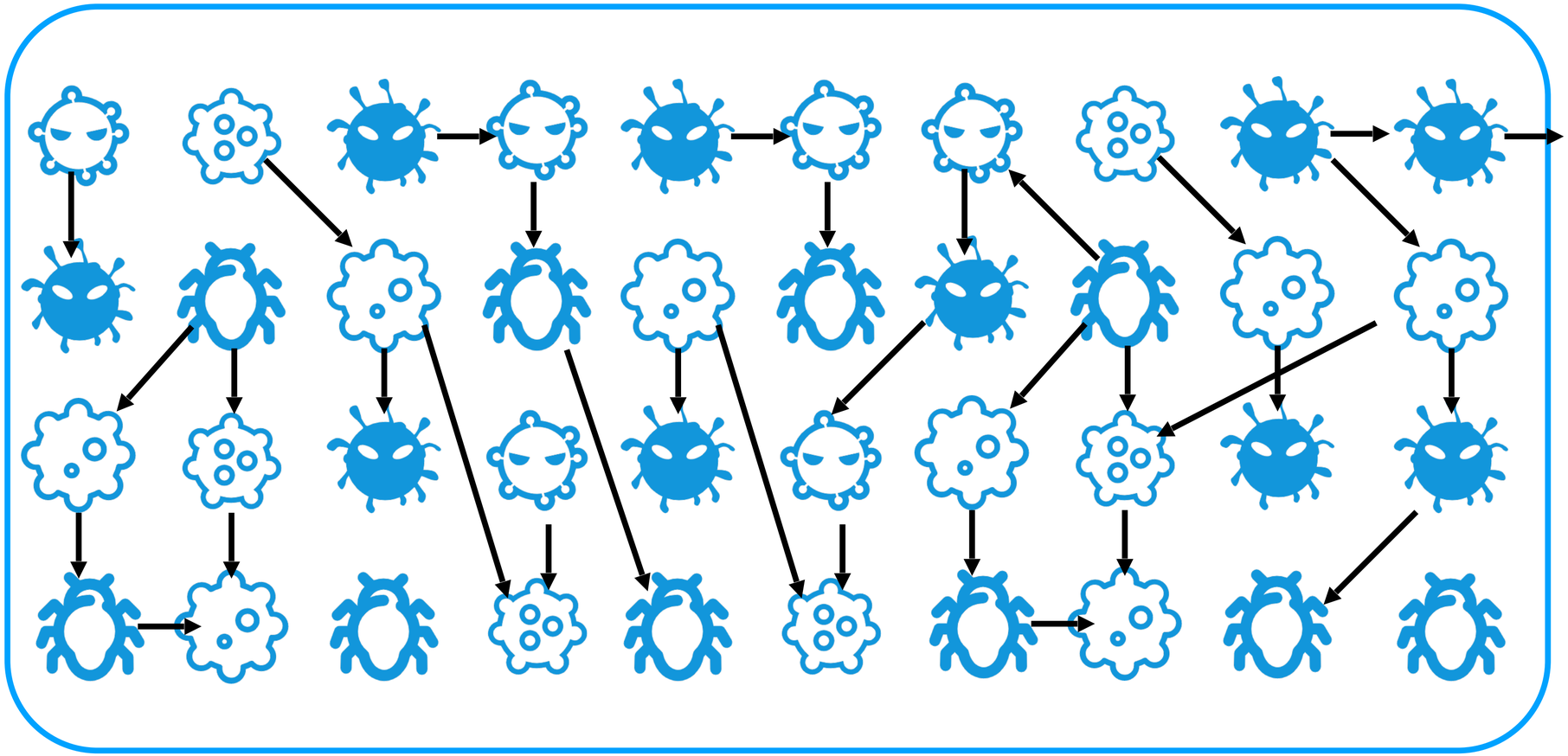
Fast clustering

Parallel ssdeep clustering

Distributed label propagation

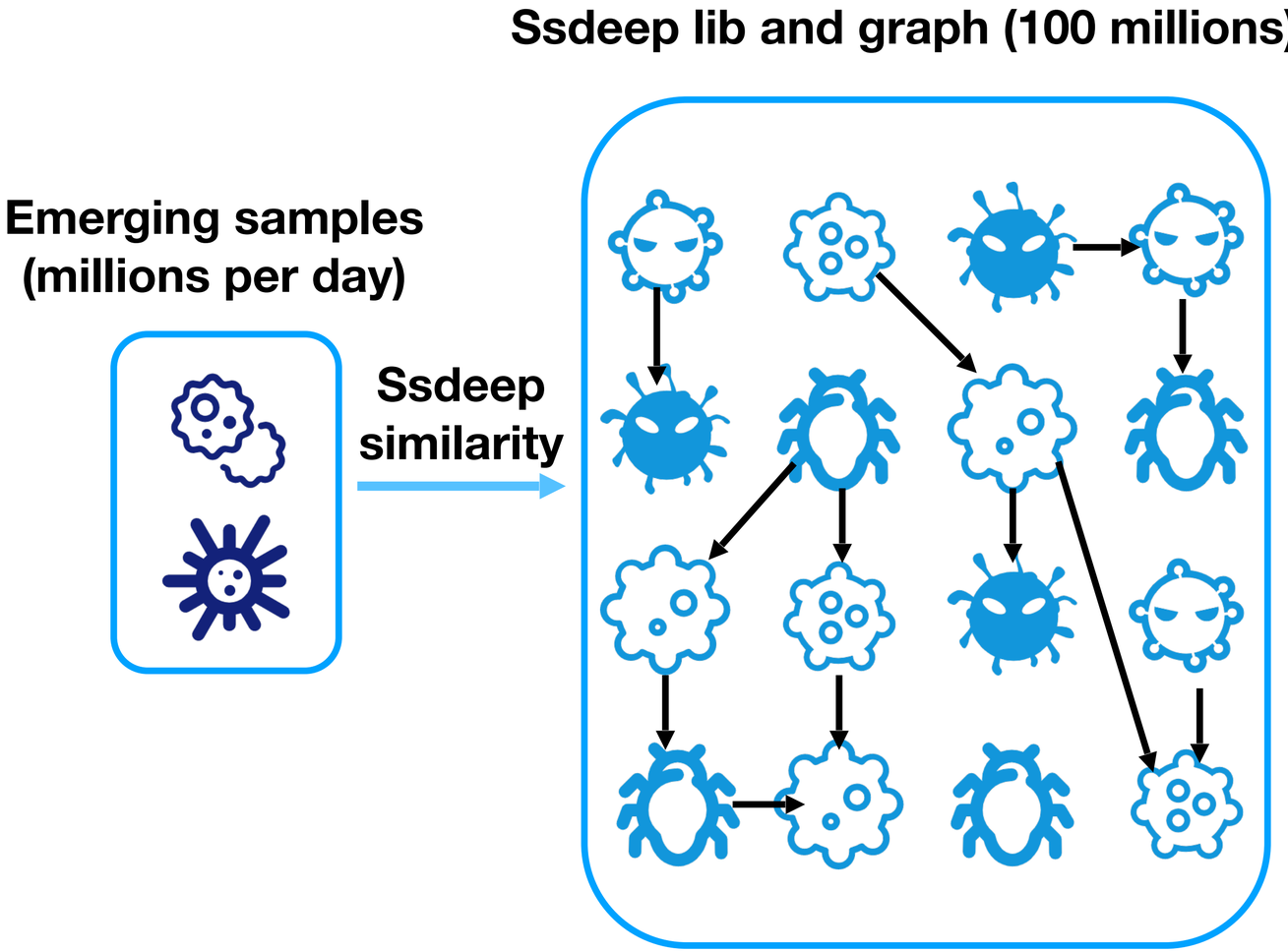
A large graph of known ssdeep

Ssdeep lib and similarity graph of 100 millions and growing



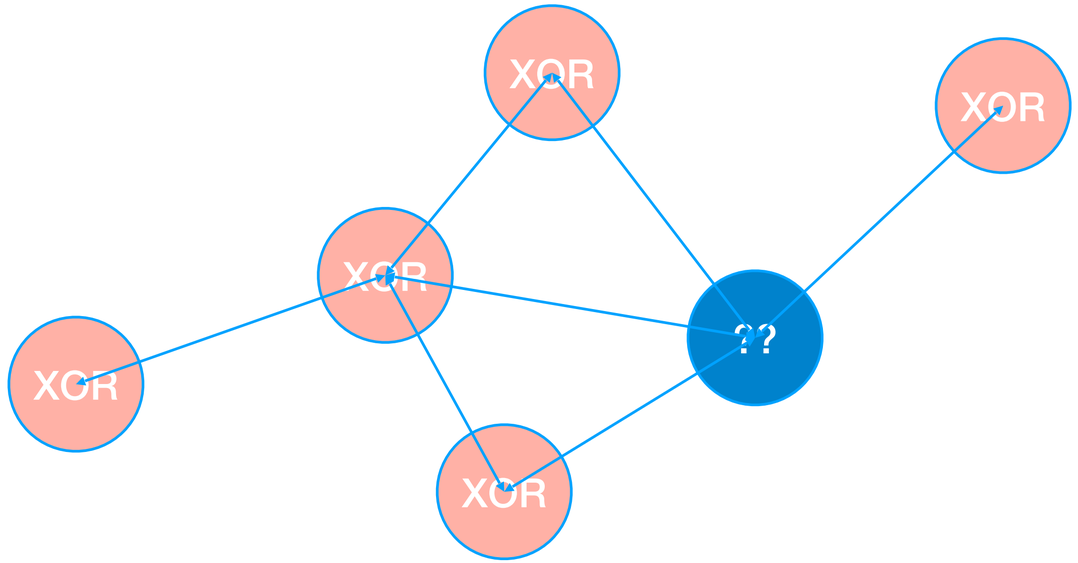
How to keep growing? Anomaly detection!

- 100 million vs 100 million pairwise approach would not work for near real time
- General anomaly detection
  - Find the emerging samples and compare with all known ones
  - Propagate and update the ssdeep library and similarity graph



So we scaled it up in a graph

Zoom-in Sub graph pattern in similarity graph of labeled xorddos nodes with unknown nodes

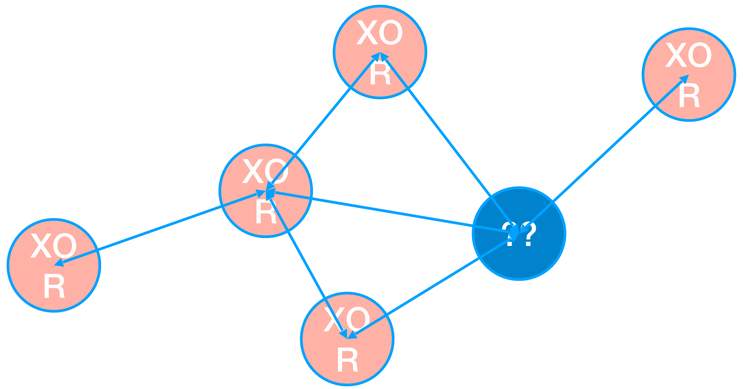


The bird view of the xorddos similarity graph

When scaling up, it starts to rock

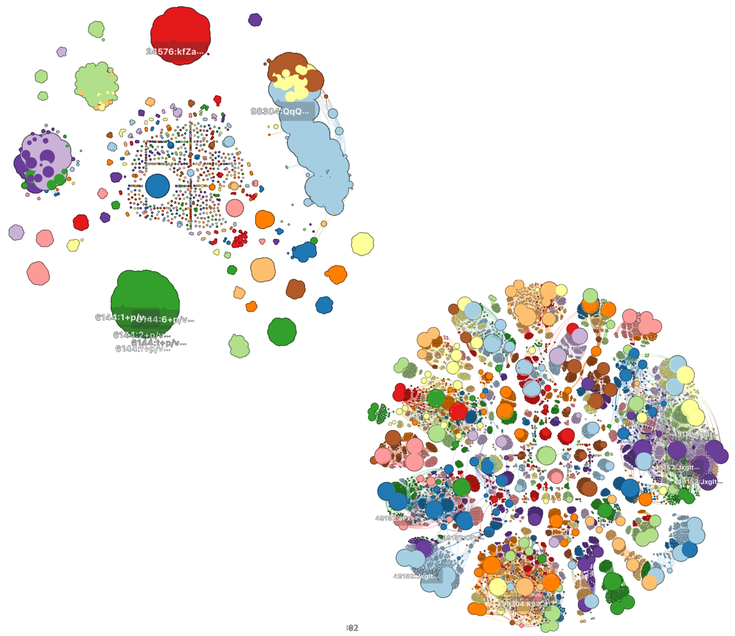
Knowing any seeds?

Look for sub-graphs



Don't have seed?

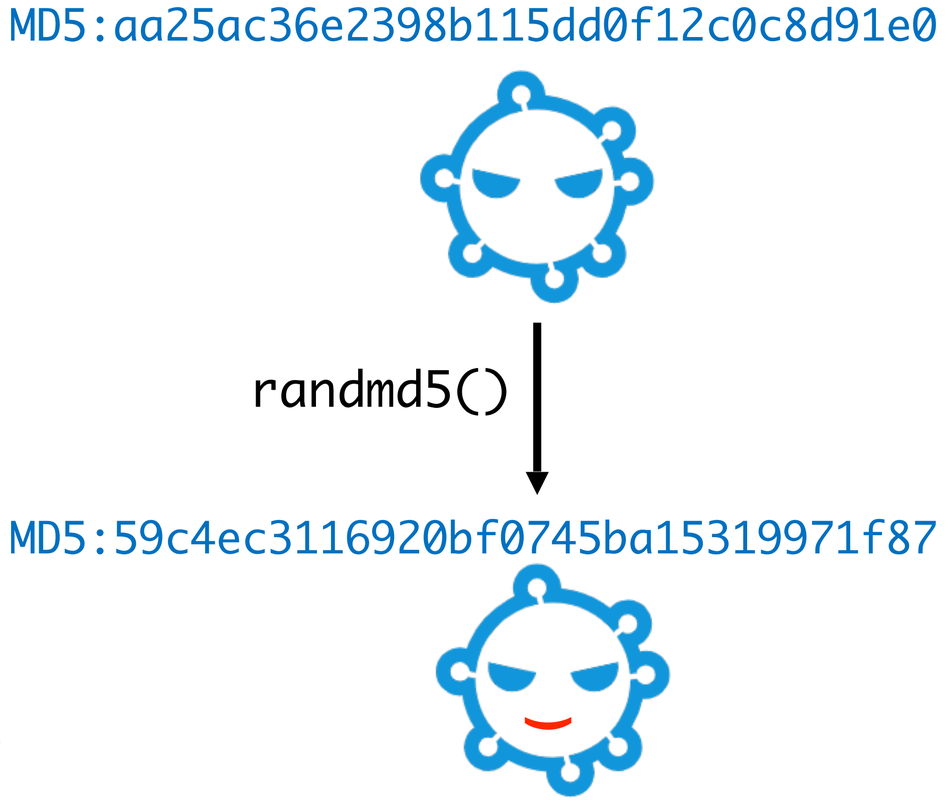
Look for clusters



Polymorphic and code reuse can be detected automatically at scale

XorDDoS as polymorphic malware

- XorDDoS “`randmd5()`” * as a simple polymorphic function to change its file hash like MD5
  - Along with randomized file names like “`/usr/bin/ogjaymscci`”
- A different file hash can effective bypass threat intelligence libs like VirusTotal
  - And xorddos does it much!
- Alibaba cloud has collected xorddos detection rules for years to provide enough seeds in graph



*** For more about Polymorphic XorDDoS, please go to https://blog.malwaremustdie.org/2015/09/mmd-0042-2015-polymorphic-in-elf.html**

Polymorphic means high fuzzy hash similarity

Xorddos variant #1

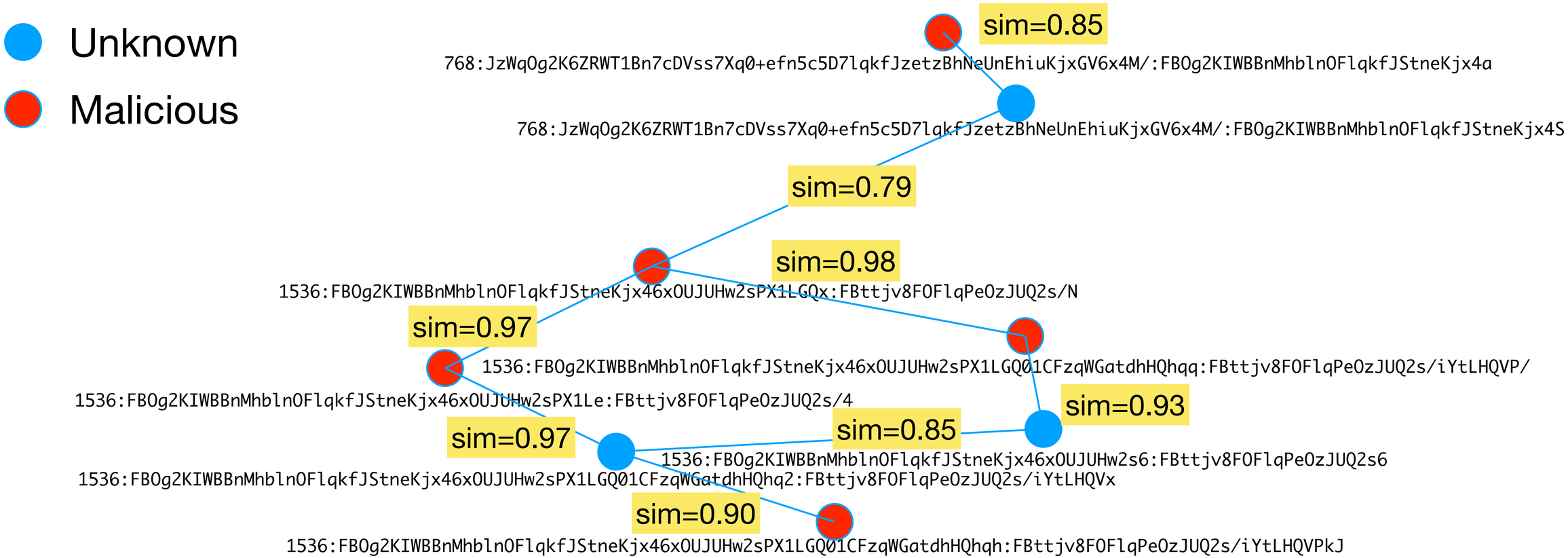
1536:FBOg2KIWBBnMhblnOFlqkfJStneKjx46xOUJUHw2sPX1LGQx:FBttjv8FOFlqPeOzJUQ2s/N

Xorddos variant #2

1536:FBOg2KIWBBnMhblnOFlqkfJStneKjx46xOUJUHw2sPX1Le:FBttjv8FOFlqPeOzJUQ2s/4

Ssdeep similarity (variant #1, variant #2) = 0.97

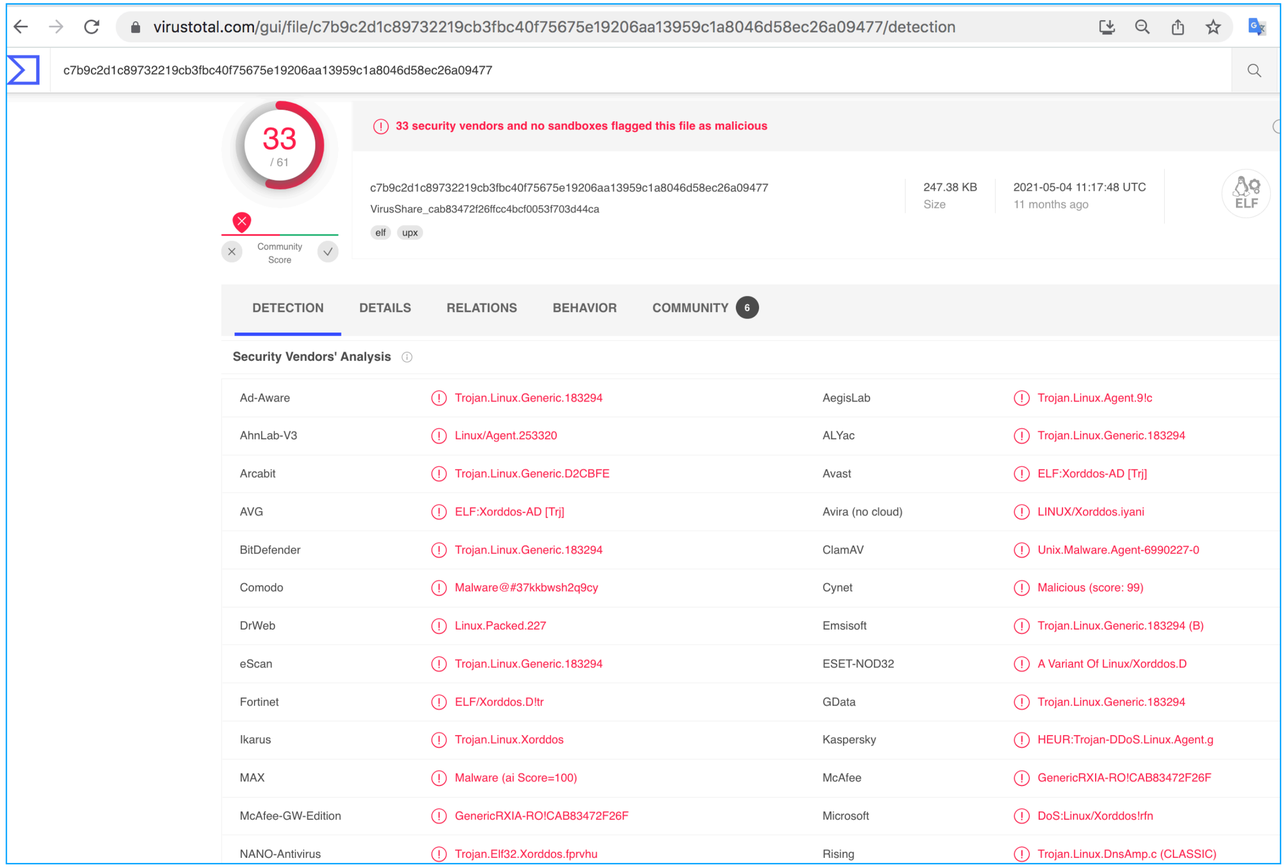
A sub-graph of XorDDoS



XorDDoS on VirusTotal

- Out of 211k XorDDoS samples
  - Avira: 153k detected (73%)
  - VirusTotal: 2.5k listed w/ one or more engines (1.2%)
- Yes, the polymorphic technique can bypassing VirusTotal!

SHA256:c7b9c2d1c89732219cb3fbc40f75675e19206aa13959c1a8046d58ec26a09477
https://www.virustotal.com/gui/file/c7b9c2d1c89732219cb3fbc40f75675e19206aa13959c1a8046d58ec26a09477/community



Code reuse in modern malware dev

Miner #1 MD5:a4ae6d3308ba52770bf6f8aa429e2a6c **VT** 47/70
Miner #2 MD5:35c0e1e371d0de3d069da8824207f4c2 **VT** 48/70

196608:0lUsEAfUY7F0e6KtTrYP7GILBqMFE2844DYIv8ZFbImQpRTUsliOw8:JsFUYJUKtTrYSILBqMFE2844DYIv8ZF9

Large portion of code reuse

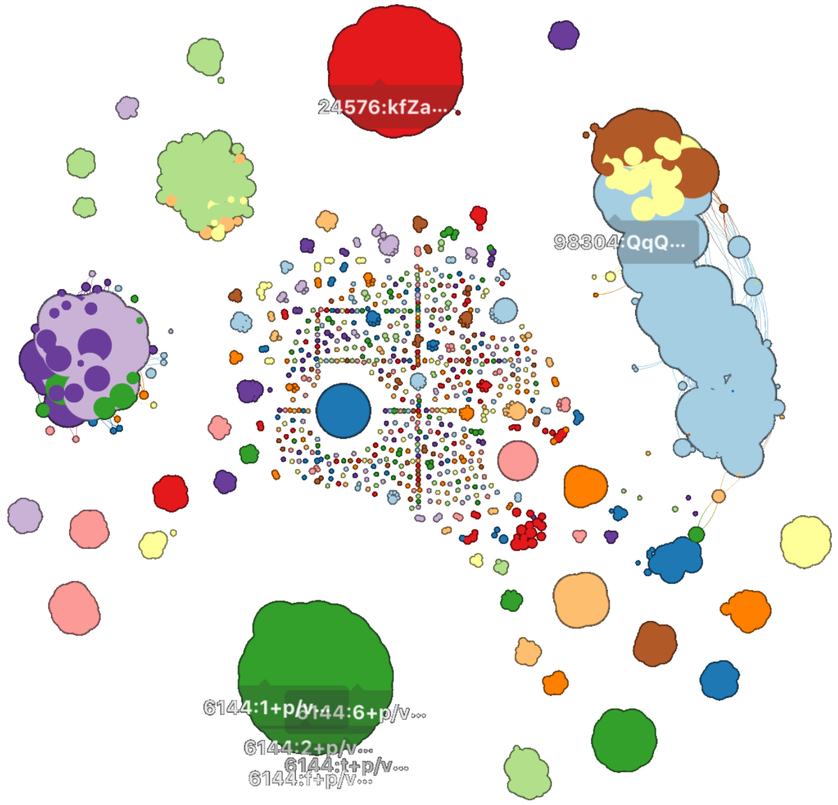
Large portion of code reuse

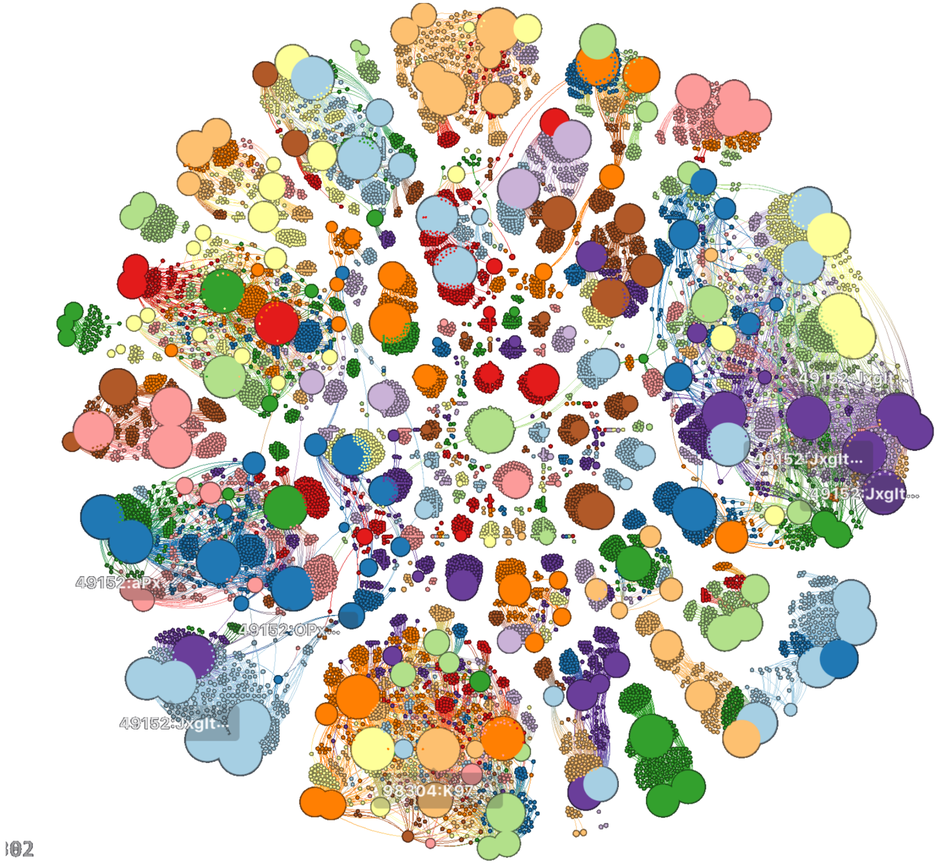
196608:0ey+QJcyXwlxUKtTrYP7GILBqMFE2844DYIv8ZFbImQpRTUsliOw8:YNYxUKtTrYSILBqMFE2844DYIv8ZFbIm

Ssdeep similarity (miner #1, miner #2) = 0.82

Ransomware and miners

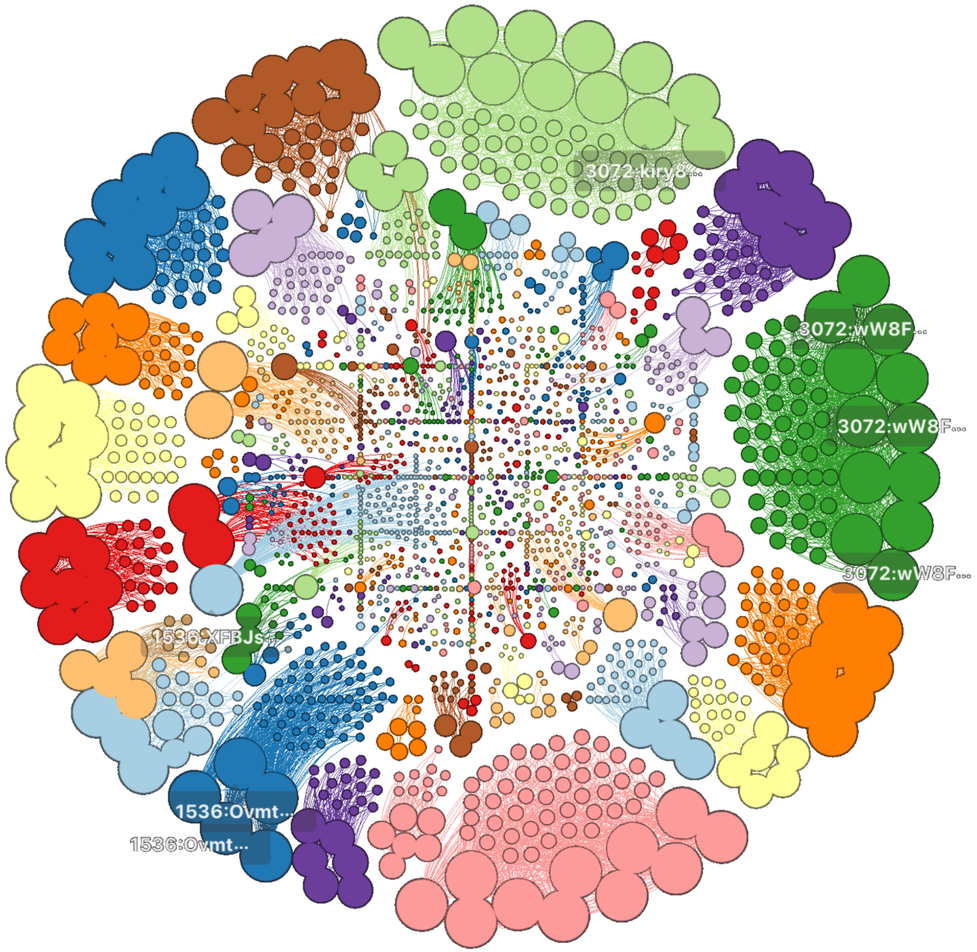
When code reuse becomes a fashion



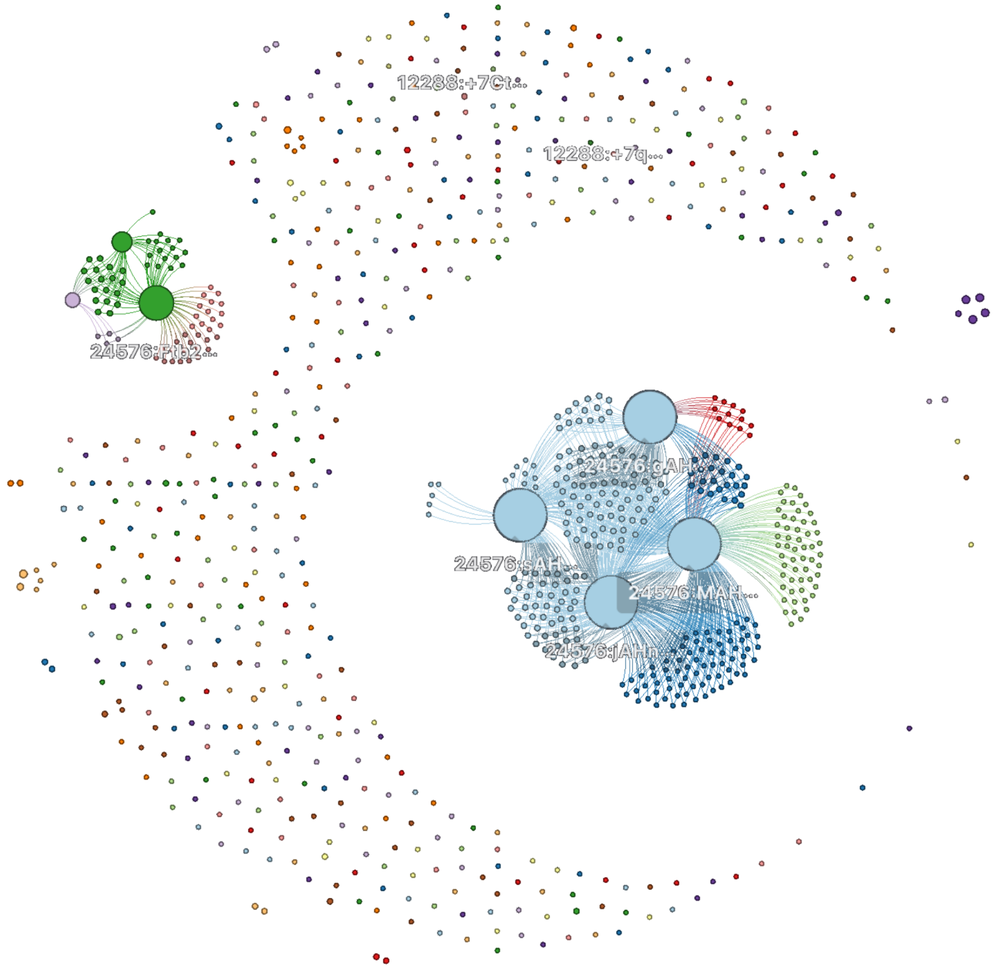


Mirai, Agent Tesla, and so on

Mirai and its siblings



Agent Tesla families (sparse clusters)



Recap the workflow

- Searching the graph for a new binary hash is faster and have better coverage then any third-party validation
- In-home validation approach remove uncertainty of VT result interpretation
- Presented workflow implemented in production and fed to Alibaba Cloud Security products

One question: is ssdeep essential?

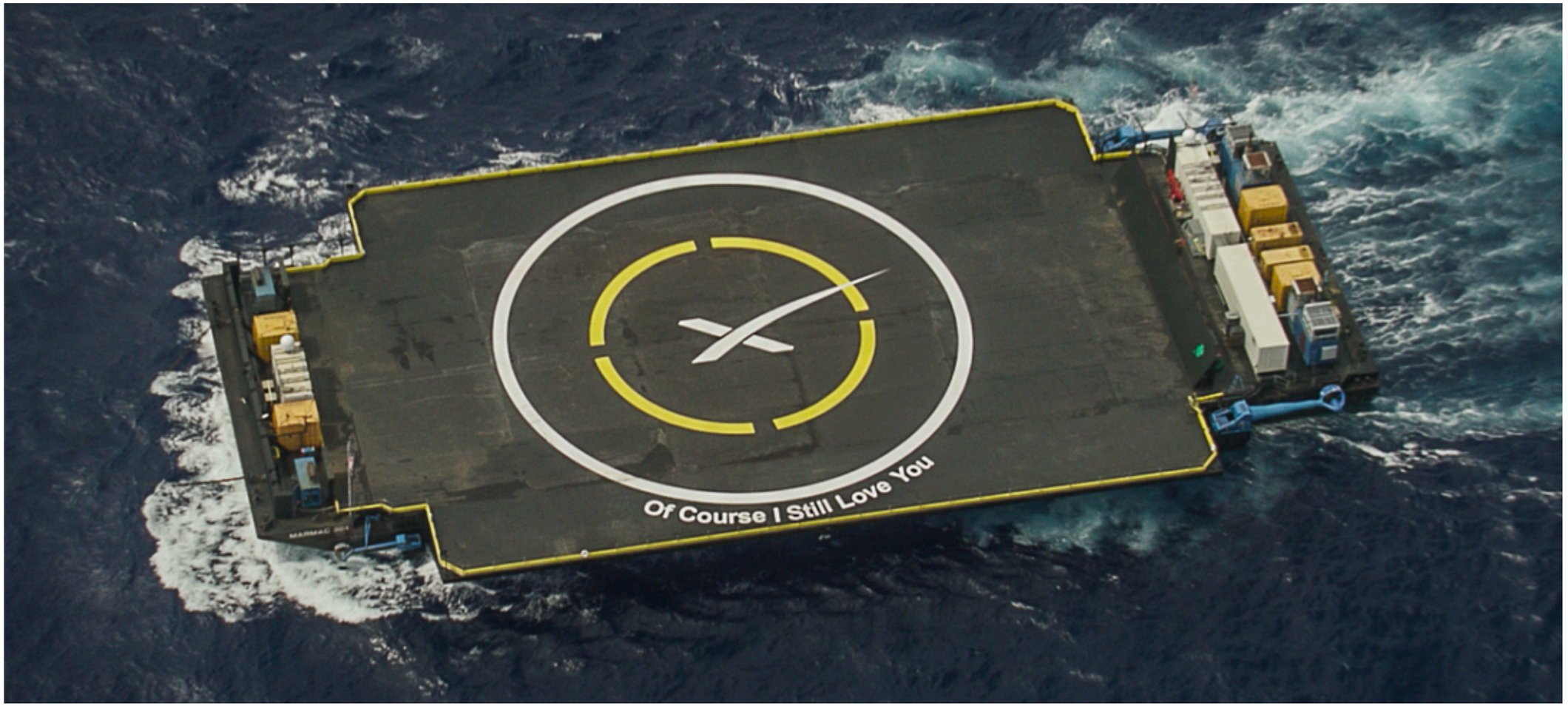
- A short answer: no, any fuzzy hash can work with this approach
- A longer answer:
  - Fuzzy hash and its corresponding similarity function like LZJD can replace ssdeep
  - Machine learning models like learn-to-hash for fuzzy hash would work
  - Non-fuzzy hash but component analysis like Intezer's Genetic Software Mapping* can probably work too
  - Engineering optimization for the hash and similarity at scale is critical

Intezer webpage https://www.intezer.com/technology/

Key takeaways

- Successful malware evolves and comes back again and again as seen today
- Big companies face specific malware threats and have unique security needs. It’s easy to outgrow even best third-party validation
- Label propagation and clustering are the two major techniques in the ssdeep graph framework at massive scale used to amplify discovery with fuzzy hash
  - Ssdeep similarity is only one of many fuzzy hash similarity functions that can be used. Please feel free to replace it with your favorite one, like LZJD
- Beside polymorphic malware, code reuse in malware, fuzzy hash similarity graph can be used to detect supply chain pollution and beyond

To VirusTotal with Love



Source https://commons.wikimedia.org/wiki/File:SpaceX_ASDS_moving_into_position_for_CRS-7_launch_(18610429514).png

Reference

- Ssdeep project: https://ssdeep-project.github.io/ssdeep/index.html
- libffuzzy: https://github.com/a4lg/libffuzzy
- fast-ssdeep-clus: https://github.com/a4lg/fast-ssdeep-clus
- Bitap algorithm: https://en.wikipedia.org/wiki/Bitap_algorithm
- Manber, Wu “Fast text search allowing errors.” doi:10.1145/135239.135244.
- Upcoming conference paper https://journal.cecyf.fr/ojs/index.php/cybin