No. 37
May 19th, 2020

HardenStance Briefing

Trusted research, analysis & insight in IT & telecom security **PUBLIC/UN-SPONSORED**

MITRE's ATT&CK Evals Are Out: Cheers!

- The new Round 2 of MITRE ATT&CK Evaluations provides very useful comparative data on EDR product effectiveness against the APT29 or Cozy Bear threat group.
- MITRE's Evaluations are primarily of value to the minority of organizations that are at risk from a specific APT and have the skills to filter the data and apply it to unique environments. The Next Round, Round 3, will emulate Carbanak/FIN7.
- The Evaluations are accelerating a market shift from expensive, proprietary cyber security product testing to outcomes that are more transparent and free to users.

Such is the respect in which the MITRE ATT&CK Framework is held nowadays, it's tempting to rush straight into the findings from the new ATT&CK Evaluations and to compare the performance of the 21 different Endpoint Detection and Response (EDR) products. After all, "Come on, who came out best?" just trips off the tongue doesn't it?

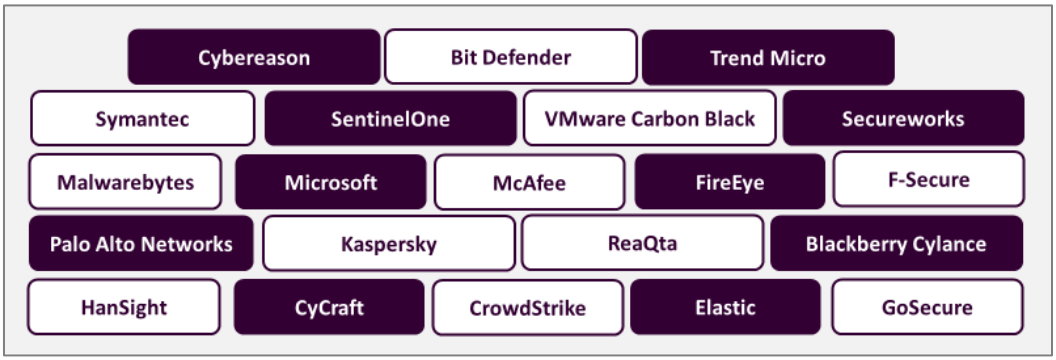
It's tempting but it would be a mistake for a couple of reasons. First, the latest round is part of an inter-linked programme of MITRE Evaluations designed to test EDR product performance against emulations of different Advanced Persistent Threat (APT) groups:

"The latest round is part of a programme of Evaluations designed to test EDR product performance against different APT groups."

- **Round 1**, carried out in early 2019, emulated the tactics and techniques of APT3, a Chinese threat group. Originally APT3 targeted US and UK defence, construction and technology firms before switching to Hong Kong-based companies.
- **Round 2**, released on April 21st, 2020, emulated Russian threat actor, APT29, known as Cozy Bear, which hacked into the Democratic National Committee (DNC).
- **Round 3,** for which MITRE recently opened the call for participation, should begin later this year. This round will subject EDR vendors to the behaviours of the financially-motivated Carbanak/FIN7 APT groups. These use the Carbanak malware to target financial institutions, primarily in the U.S.

These advanced threat groups have different motivations and target different types of organizations. Just as importantly, and as documented in superb detail in the MITRE ATT&CK Framework, they also pursue very different tactics and techniques. Most Hong Kong companies should care more about the Round 1 Evaluations than the latest Round 2. Most banks should care more about the next Round 3 than the latest Round 2.

Figure 1: 21 EDR Vendors Participated in Round 2 of the MITRE Evaluations



Source: MITRE/HardenStance

A second reason for not rushing to conclusions about the results is that the Round 2 Evaluation data sets are huge and interpreting them accurately requires a lot of time and skill. One tremendously positive aspect of the MITRE model is that all the Evaluations data can be viewed by anyone free of charge at https://attackevals.MITRE.org

The Round 2 Evaluations Methodology for APT29/Cosy Bear

- In the latest Round 2, MITRE created a Windows domain architecture in Azure. It then subjected each of the 21 EDR vendors to an emulation of APT29 comprising no less than 134 separate Tactics, Techniques and Procedures (TTPs) over three days.
- MITRE representatives played the role of the red team against the vendor's blue team. A MITRE judge registered a product's detection capability across all the TTPs.
- Whereas testing to detect commodity malware is highly automated, MITRE's APT-focused Evaluations required manual interventions throughout the test cycle.
- Critically, the Evaluations focused exclusively on detection capability. The EDR products had to be configured in passive-only or alert-only mode. Preventative mechanisms were not allowed – no ingestion of logs from other sources such as Next Gen Firewalls (NGFW) or Web Application Firewalls (WAFs); no quarantining of files; no termination of processes. Vendors weren't penalized in any way for generating false positives as they most certainly would be in a live environment.

"The Evaluations focused exclusively on detection capability. The EDR products had to be configured in passive-only or alert-only mode."

- For each of the 134 TTPs that a vendor's product was subjected to, the MITRE team registered whether or not the threat was detected. The quality of each detection was also registered i.e. whether it provided nothing more than raw data, telemetry, or whether it provided processed or (better still) enriched data, serving as an indicator of how much time it would take a SOC operative to be able act on it.

In the four weeks since the Round 2 results were released, participating vendors have highlighted different proof-points for their performance. Among them are the following:

- Breadth of coverage across the many techniques in ATT&CK (Palo Alto Networks)
- Total telemetry detections (Elastic)
- Contexts per alert (CrowdStrike)
- High quality detections - defined as Techniques and Tactics (Sentinel One)
- Products tested without requiring any configuration changes (Secureworks)

For good reason, MITRE declines to apply a universal scoring mechanism on top of its raw data sets to give vendors some kind of overall 'Rating'. This is because it takes no account of unique user requirements in live networks or other overall performance-impacting vendor capabilities that are not taken account of in the Evaluations.

Evaluating the Evaluations

- The Round 2 Evaluations do constitute a valuable, apples-to-apples comparison of EDR detection capabilities against an APT29-like threat.
- MITRE is committed to providing as much transparency as possible regarding its testing methodology. For Round 2, MITRE relied on PupyRAT and Meterpreter, both of which are open source tools. MITRE did use their own PowerShell scripts based on threat intelligence but they then released these scripts openly at the conclusion of the Evaluations. There are some other aspects that could do with improving from a transparency perspective. For example, greater clarity on exactly what evidence meets the threshold for a positive detection would be helpful. On the whole, MITRE is doing a pretty good job with respect to openness and transparency and is continuing to actively solicit participant feedback on how to improve further.

- The methodology was necessarily narrow and simplified and didn't take account of a number of critical performance-impacting factors one would encounter in a real production environment. On their own, a vendor's Evaluation results cannot be used as a wholly reliable proxy for its effectiveness in detection and response against an APT29-like threat in a unique user environment.
- Given the variation in the motives, tactics and techniques of different APT groups, the Round 2 results are even less reliable as a proxy for a vendor's detection and response effectiveness against other unique APTs or against APTs in general.

Key HardenStance Take-Aways

- MITRE's full programme of Evaluations is filling an important gap in the cyber security testing market. In terms of the number of participating vendors, this must rank as the largest ever commercial test of endpoint threat detection against an APT. MITRE appears to be the only organization capable of undertaking such a large scale public test of EDR vendor performance against APTs.
- The cyber security community should dwell on the strong positives of this effort before pointing to its limitations and looking to improve on it in future rounds.
- The Evaluations are only of use to the minority of organizations that face a meaningful risk from APT groups. Each round is of use to an even smaller subset of organisations that are at risk from the specific APT threat actor being emulated.

"This must rank as the largest ever commercial test of endpoint threat detection against an APT."

- The Evaluations are only useable by the minority of cyber security professionals that are employed in advanced threat detection and response roles – i.e. by vendors, MSPs, MSSPs, MDR providers and at-risk organizations. Most cyber security professionals don't have the know-how to apply the data effectively.
- To those that know how to use it, the Evaluations data is useful input to vendor selection. For the reasons discussed, HardenStance declines to single out specific top performers. At the other end, however, buyers should consider that a number of vendors from AV backgrounds do appear to have performed consistently poorly in terms of the basic detection 'table stakes' that were tested. A number of these vendors will have a lot of work to do to become truly competitive in the EDR space.
- For organisations at risk from an APT, obtaining a granular understanding of exactly how different EDR products respond in different phases of an attack isn't just helpful in making buying decisions. It can also help users harden their end to end security posture beyond the confines of their EDR product selection.
- Protecting against APTs is the high end of the EDR market. At this point in time, most users shouldn't make the MITRE Evaluations a significant factor in their selection of an EDR vendor. Instead they should prioritise more generic protection, cost and performance selection criteria across mainstream cyber threats.
- Round 2 of the Evaluations has further strengthened MITRE's role as one of the organizations that is most trusted and valued by cyber security professionals, especially those engaged in advanced threat detection and response. The Evaluations - and the ATT&CK Framework more generally - appeared to HardenStance to be the single most talked about tools at RSA in February. Notably, the ten vendor participants in Round 1 had also grown to 21 in the recent Round 2.
- The MITRE Evaluations are accelerating a market shift from expensive, proprietary cyber security product testing to testing models that are more transparent as well as free of charge to users. This aligns well with the open and transparent cyber security testing being advanced by open standards bodies such as NetSecOPEN and the Anti Malware Testing Standards Organization (AMTSO) ■

More Information

- https://attackevals.MITRE.org
- www.hardenstance.com
- HardenStance White Paper: "Next Steps in Playbook-Driven Cyber Security" sponsored by Cyber Threat Alliance, IBM Security, KPN & Nokia (September 2019).
- HardenStance White Paper: "A New Era in Trusted Network Security Testing" sponsored by Spirent.
- HardenStance Briefing: AMTSO's New Malware Testing Standard: Some Progress in Endpoint Security
- HardenStance received no payment – direct or "in kind" – for publishing this briefing.

HardenStance Ltd Disclaimer of Warranty and Liability

HardenStance Ltd has used its best efforts in collecting and preparing this report. HardenStance Ltd does not warrant the accuracy, completeness, currentness, noninfringement, merchantability or fitness for a particular purpose of any material covered by this report.

HardenStance Ltd shall not be liable for losses or injury caused in whole or part by HardenStance Ltd's negligence or by contingencies beyond HardenStance Ltd's control in compiling, preparing or disseminating this report, or for any decision made or action taken by user of this report in reliance on such information, or for any consequential, special, indirect or similar damages (including lost profits), even if HardenStance Ltd was advised of the possibility of the same.

The user of this report agrees that there is zero liability of HardenStance Ltd and its employees arising out of any kind of legal claim (whether in contract, tort or otherwise) arising in relation to the contents of this report.