# Darktrace – An Engineering Innovation in Cyber Defence

*Zhiheng Zhang, Guildford High School*

Darktrace is the finalist of MacRobert Award in 2017. As a cyber defence company, it created the first unsupervised machine learning software to detect and defend against cyber security threats. Modelled on the human immune system, its Enterprise Immune System can identify and neutralize the cyber threats without requiring human intervention.

What makes Darktrace's cyber security stand out from all conventional IT defence is its success in two fields where the conventional IT defence has failed. One is to fight against insider threats, the other is to prevent new threats which have not been known. These are achieved by its innovative architecture and machine learning.

## I. The innovative architecture

Its architecture is different from the conventional cyber defence which relies on the firewall as a barrier between the protected internal network and the outside world (the Internet or any untrusted network). As Figure 1 illustrates, the firewall, no mater it is a software or hardware, filters all network traffic passing through it, aiming to screen out hackers and viruses, etc. As the firewall is desgined to defend the inside against the attacks from outside, not to police the threats from inside, it is almost helpless to prevent insider threats which are caused by malicious or careless inside users. For example, if a user of a computer unwittingly opens an attachment in a 'weaponized' email which contains malware, or inadvertently clicks a link to a malicious site, or accidentally plugs a USB flash drive loaded with malware into his computer, or logs on a fake website with his online banking password, or even worse, intentionally copy and steal internal information, the firewall is impotent to stop such behaviours or to eliminate their risks or damage. Obviously this disadvantage of the firewall is well known by the attackers. In 2016, 60% of enterprises in USA were victims of social engineering attacks[2]which psychologically manipulate the computer users into installing attacks. According to a 2017 Insider Threat Report, insider threats are the cause of the biggest security breaches, 53% of companies

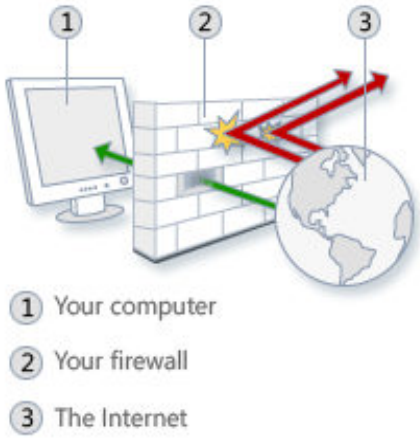


***Figure 1 Firewall*** [1]

1 Source: https://www.microsoft.com/en-us/safety/pc-security/firewalls-whatis.aspx
2 Source of data: https://www.scmagazineuk.com/60-of-enterprises-were-victims-of-social-engineering-attacks-in-2016/article/576060/

estimate remediation costs of $100,000 and more, with 12% estimating a cost of more than $1 million. [3]

The innovative solution of Darktrace is an architecture with Hyper Cylinder model, in which a set of discrete mathematical models called probes are arranged in a hierarchy like a loose pyramid. Each probe acts as a filter, monitoring its underlying network/computer, making decisions based on its own algorithms, and passing its output to another probe higher up the pyramid. At the top of the pyramid is the Hyper Cylinder model which makes the ultimate decision about whether an input is a threat or not.

Figure 2 provides a simplified illustration of the basic level of Darktrace's architecture. For example, 10 is a company's computer system in its London office, which comprise various devices connected via a local network 6. 40 is this company's another computer system in its Tokyo office, which also comprise various devices connected via a local network 43. Both systems communicate with each other via internet 20 and are accessible by other internet users symbolized as 30.

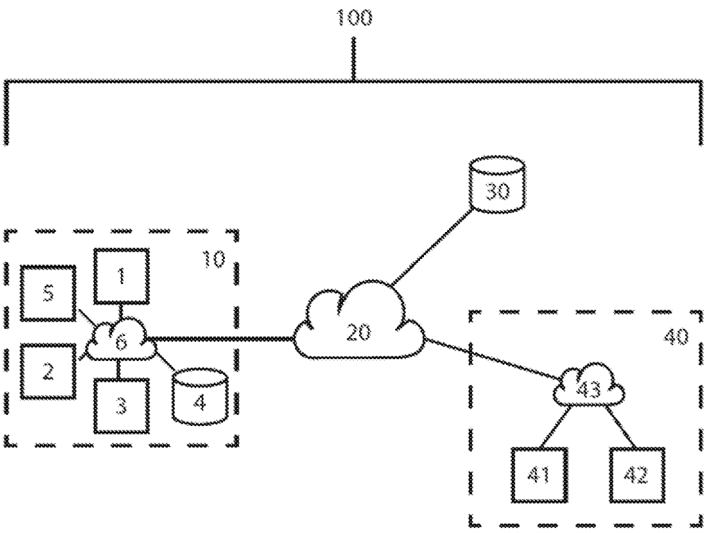


***Figure 2 The basic level structure of ABDS*** [4]

Darktrace installs its anomalous behaviour detection system (ABDS) in computer 1 in system 10. Through the interaction with other devices in 10, ABDS collects the information of every device in system 10 and builds a model of 'normal behaviour' of each one. This 'pattern of life' is dynamically updated when new information is gathered. For example, computer 2 is used by a user who is responsible for sales in UK and usually uses this computer from 8:40am to 5:20pm in working days and occasionally work overtime into late evening, but has no dealings with Tokyo office. So, once a probe at this basic level of ABDS spots that computer 2 is actively approaching the database of Tokyo office in 11:30pm, it will flag this behaviour as anomalous and report to a probe at a higher level for further investigation. [5]

3 Source of data: https://www.tripwire.com/state-of-security/security-data-protection/insider-threats-main-security-threat-2017/

4 Source: US patent No. US2017/0230392 A1, Anomaly Alert System for Cyber Threat Detection by Tom Dean and Jack Stockdale. Sheet 1 of 2

5 Reference: Same as 4. Paragraph [0044] – [0049].

Figure 3 illustrates the structure of the higher levels in ABDS. A network traffic monitor 110 exacts metadata from the probes at the basic level in network 100. These data are proceeded to form behavioural metrics 120 which measure the behavioural properties of every device and user. These metrics are fed into anomaly alert system 200, where historical data 150 are analysed to build models of normal behaviours 170. Then new observation 160 is compared and analysed with 170 to produce a conclusion 180 about if this new observation presents anomaly or not. If it does, the system will take the action 190 to alert the user 210 to the threat. [6]

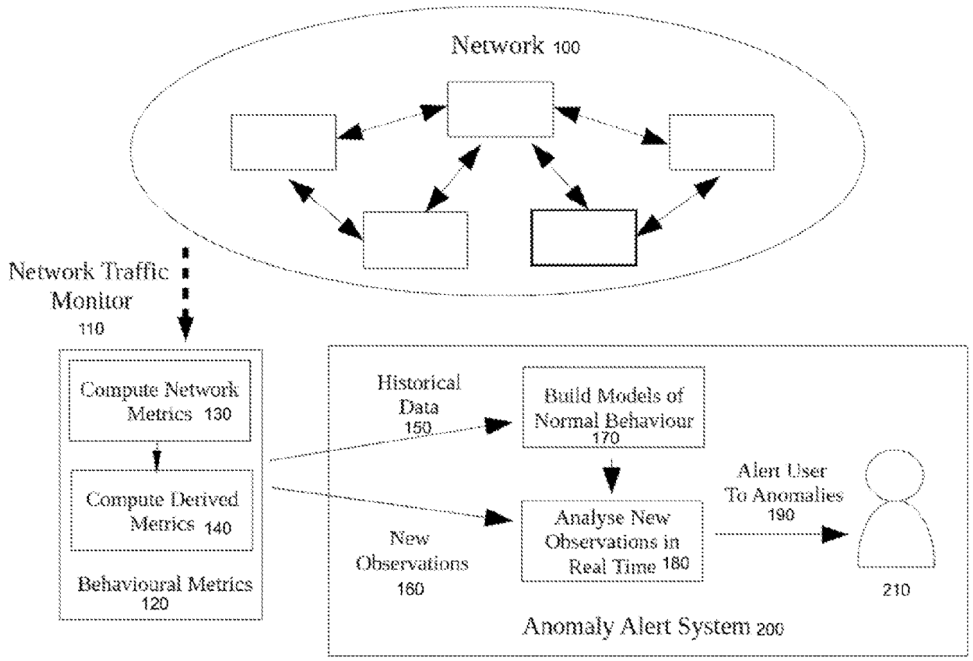


***Figure 3 The high level structure of ABDS*** [7]

In summary, the whole structure can be described as: Receive input date → Derive metrics → Analyse metrics → Calculate probability → Threat determination [8]

Unlike the firewall relying on border/gateway control, Darktrace's system monitors every device and behaviours in the protected network and leave no blind zone. So it can detect and stop the threat no matter where it is from, inside or outside.

6 Reference: Same as 3. Paragraph [0050] – [0052].
7 Source: Same as 4. Sheet 2 of 2
8 Source: US patent No. US2017/0230391 A1, Cyber Security by Matt Ferguson and Maha Kadirkamanathan. Sheet 2 of 3

## II. The innovative machine learning

The breakthrough innovation which makes Darktrace a global leader in cyber security industry is its self-learning. The conventional IT security models relies on pre-determined signature, rules or antiviral updates which are written by human and based on their knowledge about past attacks. For each successful intrusion, a new rule is written, the signature is updated, so the next time that same threat appears, access is denied or challenged. This method is effective only in defending against known threats but is incapable of responding to fresh threats. [9]

To its contrary, Darktrace builds its Hyper Cylinder model on Bayesian probabilistic model which permits automatic detection of cyber threats through probabilistic change in behaviour of normal computers/networks or human behaviours. Its Hyper Cylinder model monitors behaviours of both machine and human to address the threats of malware and wrongdoing of human staff, both from inside and outside of the protected network, and predict and stop the threats before they can do any harm. This is achieved without any supervion from human. Unlike the traditional endpoint defences, a threat does not have to have been known before it can be detected. This enables Darktrace to defend effectively against the fast evolving cyberattack.

In engeneering, Darktrace uses a combination of 12 different machine-learning algorithms that are monitored by a supervisory mathematical model that uses probability theory and Bayesian modelling to learn and adapt the system's output. [10]Firstly, extreme value theory, especially Peak Over Threshhold (POT), are used to assess, from the data gathered by the probes at the basic level of ABDS, the probability of an observed event or value that present extreme deviation from the median of probability distributions. Extreme means either unusually large of unusually small compared to the median. A value can be considered as unusually large if it falls far above a suitable quantile, as unusually small if it falls far below a suitable quantile, either presenting an anomaly, such as the example of Figure 2, the active time of computer 2 as 11:30pm presents an extreme deviation from its usual working time 8:40am to 5:20pm. Mathematically, a value $m$ of some metric $M$ is anomalous if either of the tail possibilities

$$P(M > m), \quad P(M < m)$$

are sufficiently small. [11]

9 Reference: US patent No. US2017/0220801 A1, Cyber Security by Jack Stockdale and Alex Markham. Paragraph [0008].

10 Source: UK's Darktrace aims to lead the way to automatic cyber security http://www.computerweekly.com/news/450299725/UKs-Darktrace-aims-to-lead-the-way-to-automatic-cyber-security

11 Reference: same as 4. Paragraph [0084].

In order to do the assessment, the probes of ABDS need to extract, from a continuous record of the monitored network, the peak values reached in any period during which values exceed a certain threshold or falls below a certain threshold, so called as Peak Over Threshold.

Secondly, Generalized Pareto Distribution (GPD) is used to model the tail of the distribution represented by POT, in order to estimate suitable quantiles (thresholds) and tail probabilities from observed data. When the distribution of has an exponential tail, Gumbel law (also known as Type I) may be used; when the distribution has a heavy tail (including polynomial decay), Fréchet Law (also known as Type II) may be used. [12]

Thirdly, Bayesian model to determine if an observed behaviour of either human or machine is anomalous or normal. To give a simplized example, let us look back to the example of Figure 2. Computer 2 is spotted to be active at 11:30pm. Shall it be determined as anormality and trigger an alert?

As per the Bayesian theorem,

$$P^M(A) = \frac{\pi(A)P(E \mid A)}{\pi(N)P(E \mid N) + \pi(A)P(E \mid A)}.$$

Whereas,

$P^M$ (A) is the posterior probability that the behaviour generating a observation of a metric M is anormalous.

π (A) is the prior probability that the value m is the result of anormalous behaviour.

π (N) is the prior probability that the value m is the result of normal behaviour.

P(E/A) is the conditional probility of the event E given the generating behaviour is anormalous.

P(E/N) is the conditional probility of the event E given the generating behaviour is normal. [13]

For example, suppose the historial data analysis done in the previous stages has found that 98% behaviours in this company's computer system is normal and 2% is anormalous, and possibility of computer 2 in use after 11:00pm as normal behaviour is 4%, then the model can give us the possibility that this behaviour is anomalous, $P^M$(A)as:

$$\frac{0.02 \cdot (1 - 0.04)}{0.98 \cdot 0.04 + 0.02 \cdot (1 - 0.04)} = 0.33$$

The same Bayesian framework is used to compute the anomality possibilities of every network device, based on the GIGO (garage-in-garage-out) principal. That is to say, the anomalious behaviour(s) of a device will result in the anomalious state of the device, no

12 Reference: same as 4. Paragraph [0114].
13 Reference: same as 4. Paragraph [0115] - [0118]

matter how good is the post-processing. So the algorithm to compute the possibility of a device *d* is in anomalous state is

$$P^d(N)=\Pi P_i^M(N),$$

$$P^d(A)=1-\Pi P_i^M(N) \quad {}^{14}$$

For example, suppose a device has only two variables. The possibility that variable 1 is normal is 0.9, and the possibility that it is anomalous is 0.1; the possibility that variable 2 is normal is 0.8, and the possibility that it is anomalous is 0.2. Then the possibility that this device is normal is $0.9 \times 0.8 = 0.72$, and the possibility that this device is anomalous is $1 - 0.72 = 0.28$. So the possibility that the device is normal is always smaller than the possibility that each variable of the device is normal. Obviously this is a prudent approach in cyber security.

With these models and algorithms, Darktrace's Enterprise Immune System allow the software to learn and develop its own tasks and learn from examples, data, and experience over time by itself. This is a self-learning and self-evolving system that doesn't need any human input and is not constrained by pre-conceived human thinking.

Furthermore, Darktrace has developed Antigena modules which are modelled on antibodies that identify and neutralise bacteria and viruses. As Enterprise Immune System detects a threat, Antigena modules are designed to act as an additional defence capability that automatically neutralises those threats without requiring human intervention.

## Summary

By now, Darktrace's cyber security system is the only one in the world that don't need any human intervention. It can be installed within an hour because it does not require any tuning or configuration, and a trial can be completed in four weeks without the need for any Darktrace engineers to be on site. [15] Its self-learning enables it to adopt to any environment, from an organization with a million devices to a network with two devices, from multinational companies to national infrastructure, from banks to heavy industry, with no need of any prior knowledge of what it is looking for.

Can we be optimistic enough to call Darktrace the terminator of cyber attacks? Maybe not, as the attackers have also started using AI. But undoubtedly Darktrace is leading the way to use AI defence against AI attacks.

(word count: 2101 words)

14 Reference: same as 4. Paragraph [0132]
15 Reference: same as 10.

Bibliography:

[1] US patent No. US2017/0230392 A1, Anomaly Alert System for Cyber Threat Detection by Tom Dean and Jack Stockdale.

[2] US patent No. US2017/0230391 A1, Cyber Security by Matt Ferguson and Maha Kadirkamanathan.

[3] US patent No. US2017/0220801 A1, Cyber Security by Jack Stockdale and Alex Markham.

[4] http://www.computerweekly.com/news/450299725/UKs-Darktrace-aims-to-lead-the-way-to-automatic-cyber-security

[5] https://www.microsoft.com/en-us/safety/pc-security/firewalls-whatis.aspx

[6] https://www.scmagazineuk.com/60-of-enterprises-were-victims-of-social-engineering-attacks-in-2016/article/576060/

[7] https://www.tripwire.com/state-of-security/security-data-protection/insider-threats-main-security-threat-2017

[8] https://www.raeng.org.uk/grants-and-prizes/prizes-and-medals/awards/the-macrobert-award/2017-finalist-darktrace

[9] http://www.computerweekly.com/news/450403013/Business-needs-AI-defence-against-AI-attacks-says-Darktrace

[10] https://us.norton.com/internetsecurity-how-to-how-do-firewalls-prevent-computer-viruses.html