A New Step Towards Network Security：

DDoS Protection with Machine Learning and Artificial Intelligence

WHITE PAPER

GenieNetworks

Disruptive Technology Trends：
Artificial Intelligence and Machine Learning

Among the numerous types of cyberattacks existing today, Distributed Denial of Service (or commonly known as DDoS), an attack in which multiple compromised systems attack a target website or online service by flooding its traffic to make its service unavailable, is often used as a tool of extortion, blackmail, politics and business rivalry. DDoS attacks are easy to execute, and along with advancement of new technologies like 5G and IoT, have grew quickly in scope and frequency to become more devastating than ever.

Though DDoS attacks have been known for years, it is still one of the most threatening cyberattacks to date. Due to its sophisticated characteristics, with the help of emerging technologies like Artificial Intelligence (AI) and Machine Learning (ML), DDoS attacks have found ways to get around network security measures and become a favorite cyber-attack tool for hackers worldwide.

Ever since AlphaGo defeated top human Go players in 2016, AI quickly became one of media's hottest tech topics, and successfully proved itself to be more than a subject of theoretical research but practical application. New hackers have begun developing cutting-edge cyber threat tools and custom attack programs based on AI and ML technologies. These attacks are not only based on variants of static algorithms but utilize automation and ML techniques to customize themselves for a specific target. The result is an increase in the scope and scale of the attack, while being less likely to be detected and identified.

One such example is replacing botnets with self-learning intelligent clusters of compromised devices such as Hivenets and Swarmbots. The basic concept is by using swarms of malware or ransomware programs to target vulnerable systems and simultaneously communicating with each other and take actions based on shared local intelligence. A hive can give a swarm of insects a command to implement in ways that suit its circumstances, and then decide on its implementation and timing based on local intelligence as it travels and grows. As it identifies and compromises more devices, a Hivenet is able to grow exponentially and widening its ability to simultaneously attack multiple victims and impede mitigation and response.

One other example is using the Natural Language Processing (NLP) algorithm to analyze public or private content from social media and emails to perform phishing attacks to leverage attack success rate on social engineering.

Another example of ML being used by hackers is Generative Adversarial Network (GAN). Think of it like an art forgery, the forger (Generator) constantly reproduces fake images that constantly undergo evaluations by an art forensics detective (Discriminator). Every time the forger receives feedback from the detective, it is able to make some improvement on its forgery. This learning

process repeats itself until the forgery can no longer be identified by the detective. Attackers today are applying this concept of GAN to cyberattacks to bypass security detections.

The rise of AI and ML has led to dramatic increase in cyber incidents that exploit these disruptive technologies as tools for new attacks, causing devastating threats to network security like never before.

A Battle of AI vs AI

Along with the rapid development of the internet technology, cyberattacks have become automated, smarter, and stealthier. Traditional defenses face a number of challenges against these attacks：First of all, it often relies upon Signature-based detection which only identifies known threats. When an attack takes place, a unique pattern known as a signature, is established about the attack so that it can be identified in the future. To create a signature takes manual analytic effort and eventually takes time. Therefore, the overall response time of the defense would also be higher, resulting in greater impact and damage when an attack takes place. The flaw of signature-based detection is magnified when facing new attacks incorporating emerging technologies like AI and ML. As the attack behavior automatically changes and updates at a rapid pace, the amount of human resources and their ability to analyze and combat these diverse attacks will become very limiting.

Another problem faced by traditional defense method against automated and intelligent attacks is its dependence on single-vector attributes to detect attacks. For virus code detection, it relies on scanning traffic content for virus patterns; for IP blacklisting detection, its detection attribute is the attack source or target IP address; for traffic baselining, it counts on traffic rate (e.g., bit-per-second, packet-per-second), or traffic ratio (e.g., byte-per-packet, packet-per-flow) of specific traffic behavior rules. These examples show that most traditional defense mechanisms identify attacks based on a single attribute or vector. However, components like malware code patterns, IP addresses, and traffic behaviors are all fundamental structures of a cyberattack. Under the aid of AI and ML, these components are constantly and automatically modified and updated so they are less likely to be recognized. In this case, traditional security solutions that rely on manual deterministic approach are thus considered outdated.

Aside from the need of manual intervention and the automation of detection attribute selection, DDoS attacks also possess the characteristic of multiple distributed attack sources. Since the attack could come from various multiple sources, it is essentially difficult to identify and analyze the attack traffic with a single defense system deployed at a single location. Particularly within a large, complex environment like a carrier-grade network, even experts may find it hard to grasp the network-wide, dynamic traffic composition and behavior characteristics. Not to mention

implementing a mechanism to correctly identify attack traffic behavior and determine detection attribute.

Cyber security is an ever-lasting battle. While the costs for hackers to obtain AI technique, algorithms and cloud computing power are getting lower each day, traditional security tools are only proven to be insufficient. Cyber defenders are on a quest to develop new security measures – ones that employ automated, intelligent traffic behavioral analysis and rely less on human intervention. Not only to effectively handle more complex and dynamic traffic behavioral analysis, but also to precisely and quickly detect and react against attacks. Essentially, AI and ML have come to play a major role in the field of cyber security.

Due to the decrease in hardware storage cost, the rise of computing power, and the availability of big data in recent years, today's ML technology realizes artificial intelligence through the capability of learning patterns within data. Common applications of ML involve a number of technical features including：

1. Data Cleaning：To recognize a pattern within data, having a clean and correct dataset is the first key procedure;

2. Feature Extraction：A process that involves excavating informative features such as source address, speed, and time from data, then quantifying and transforming these values into multi-dimensional vectors;

3. Feature Selection：Selecting vital features (variables) by inspecting the results of applying ML models. In attack traffic analysis, these features can be source distribution, speed, time variation, etc., while other vectors may not be as crucial. The best set of features to optimize learning process are selected by gradually testing and implementation of ML algorithms;

4. Model Selection：The process of choosing between different ML approaches depending upon evaluation of the problem, the current data type, and the overfitting level. Generally, ML models are classified into Supervised Learning, Unsupervised Learning, Semi-supervised Learning, and Reinforcement Learning. There are many types and methods of ML algorithms including Artificial Neural Network (ANN), Support Vector Machine (SVM), Clustering, Decision Tree, and Random Forest, just to name a few.

Engaging in this ever-lasting battle against new intelligent cyber threats, many security vendors are beginning to incorporate ML techniques into their solutions to enhance cyber security. Technology could be good or evil, depending on how you adopt it. So as AI and ML. In the world of cyber security, the battle between AI attackers and defenders have just started.

DDoS Protection as an application in Machine Learning

In contrast to traditional signature-based detection, Genie Networks focuses on DDoS protection that leverages network-wide massive data mining to perform Network Behavior Analysis and Automatic Baseline Learning. With the proliferation of Big Data and ML over the past years, Genie has become active in exerting efforts to the research and development of ML-powered detection on DDoS attacks. The main focus of our research includes：

Powerful Traffic Data Mining to Generate Traffic Feature

The behavior of every network traffic flow can be described in a flow record exported by a network device such as a router or a switch. The flow records include basic data information such as IP address, protocol, port, TCP flag, size, time, etc. Based on this information we can perform data aggregation, classification and correlation. For example, correlating each traffic attribute with country, sub-network, BGP routing, etc., or preprocesses like classification, statistics, and sorting to obtain more traffic attribute features. Another way GenieATM obtains attribute features is through the correlation of DDoS anomaly tickets to IP traffic information.

For a carrier-grade detection, the cumulative value of traffic rates is often used as an indicator when observing a traffic burst. Sorting of a traffic feature is commonly used for traffic analysis rather than anomaly detection. However, in the field of MI, we can convert each set of the analyzed sorted values to a single indicator value that represents a particular status in a time series – such as the status of country distribution, application distribution, and service provider source distribution. When comparing a distribution status among different time spans (days, weeks, weekends, etc.), if a difference is observed, i.e., an anomaly is detected, an alert will be triggered immediately.

From aggregated data statistics and correlation of known attack records, we are preparing training data for a ML model that not only inspects a single-vector behavior but a full-scope traffic status.

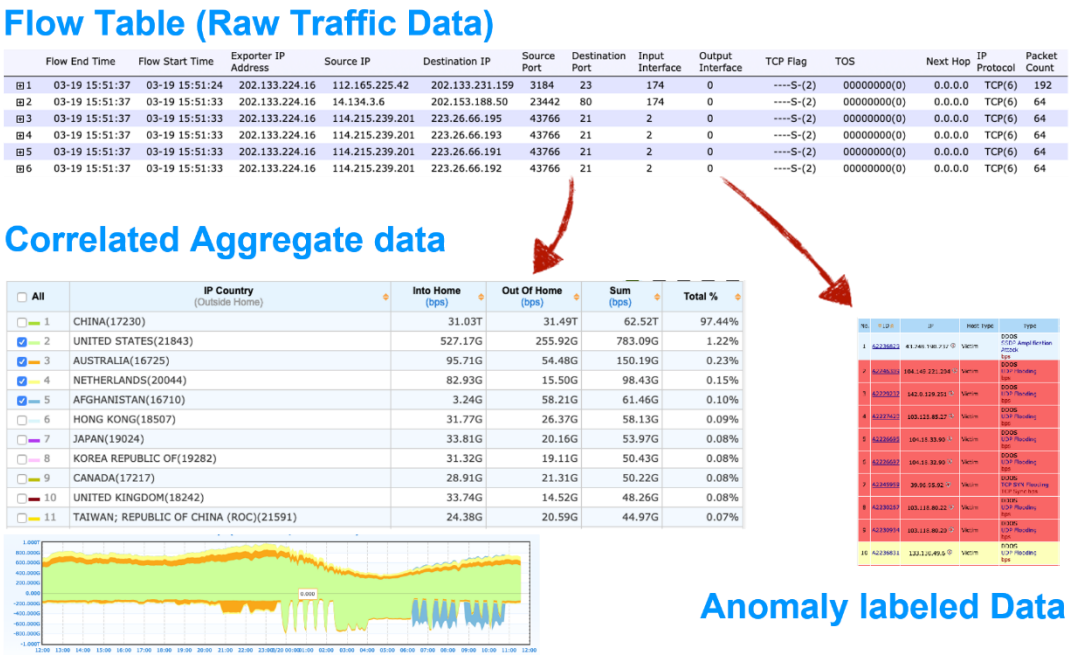

Figure 1: Feature Engineering

Constant Adjustment of the ML Model to Optimize ML Process and Result

Adopting ML technique to anomaly traffic detection requires Feature Engineering – the process of putting suitable traffic attribute features into appropriate ML algorithms to perform classification and analysis. While our training dataset contains many features, it is preferred to extract the principal variables prior to using all the possible random variables for model training. This is commonly known as dimension reduction. We can implement Principal Component Analysis (PCA), an algorithm for feature extraction, to avoid overfitting of features as a result of the curse of dimensionality.

Since network traffic datasets are generated during the data communications over time, the network traffic dataset can be analyzed as time series, and it comes very naturally to use time series analysis and forecasting during anomaly detection. For example, the most basic traffic baselining model compares real-time traffic behavior with that of a specific time range of the past, such as traffic from a week or a month ago. The setting of such comparison is often manual, static, and is not automated or intelligently triggered. Under the aid of ML, we are able to count on more time-series systematic and non-systematic components such as the baseline level, linear trend, seasonality, and noise. Some common ML models that are applied to time-series data analysis include Auto-Regressive Integrated Moving Average (ARIMA) and Long Short Term Memory Network (LSTMN).

For ML based DDoS detection, it is common to use unsupervised learning for unlabeled data and divide traffic into groups on the basis of similarity and dissimilarity in traffic features. This grouping process, also known as automatic clustering, enables us to find the outliers which are extreme values that deviate from other observed data points – i.e. anomaly traffic. Some of these common ML algorithms include K-means Clustering, Isolation Forest, and Local Outlier Factor.

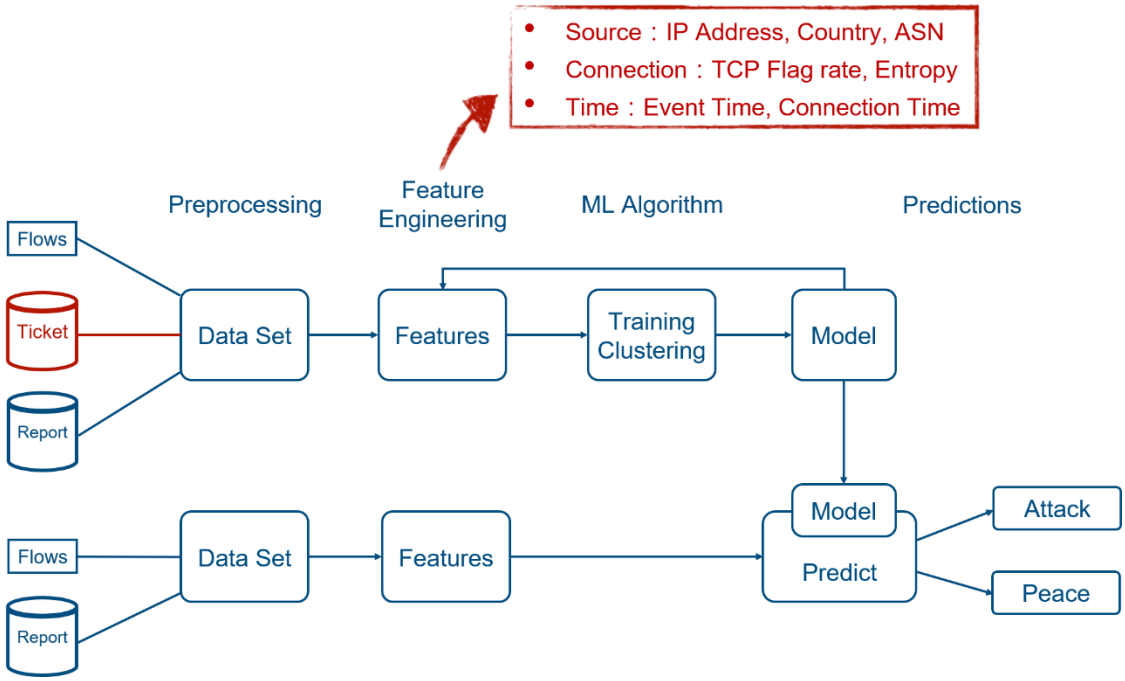

Figure 2: Schematic Diagram of Unsupervised Learning

On the other hand, a real-time continuous DDoS detection system can utilize known anomaly events as feedback to obtain a multi-level Supervised Learning model. One such example involves automatically inputting known anomaly traffic attributes as feedback to ML's labels. This further enhances the precision and efficiency of ML while eliminating the hassle of manual labeling.

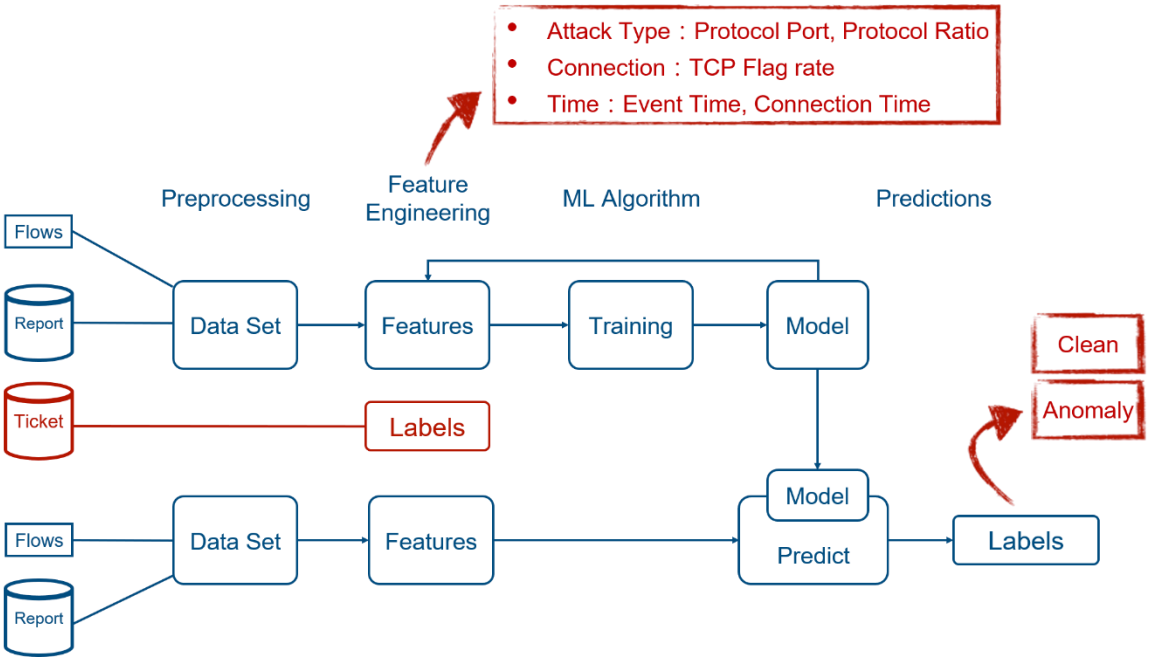

Figure 3: Schematic Diagram of Supervised Learning

Visualized Machine Learning Process

Through visual representation, users can get a glance of the attributes and trends of their network traffic and easily gain insights from the analyzed data. The process and result of ML are displayed as graphs and animations rather than complex algorithms and mathematical formulas, giving users the ability to intuitively and interactively operate and modify parameters of the ML model.

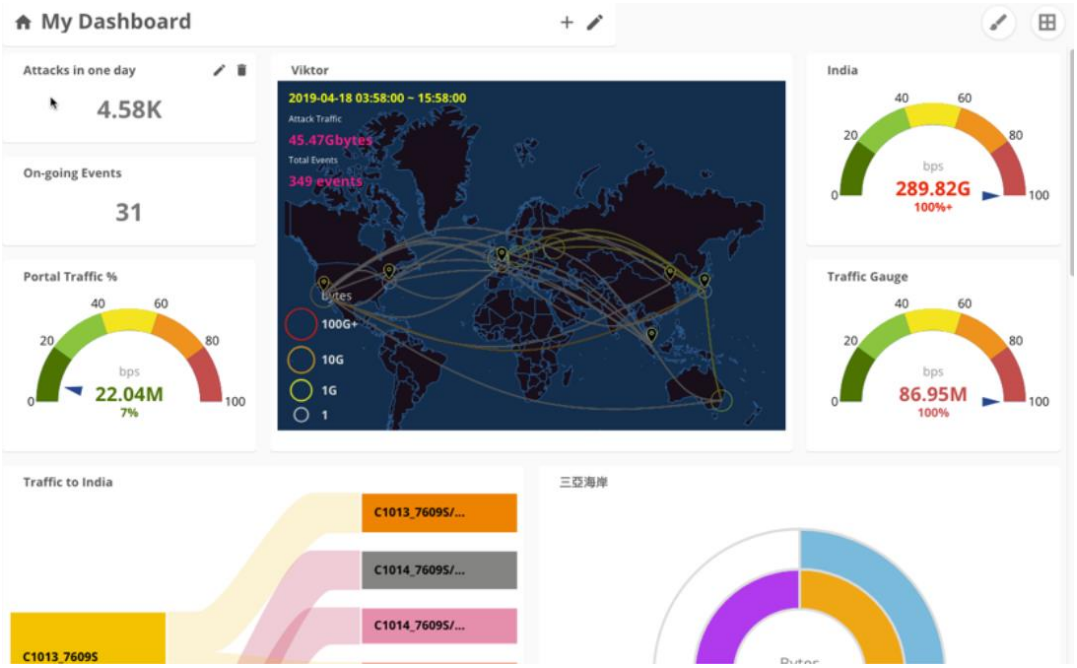

Figure 4: Cyber Threat Dashboard

A New Era of ML-Based DDoS Security

As cyber threats continue to grow and evolve at quick pace, security solution providers are doing all they can to strengthen their defense. Conventional cyber protection relies heavily on manual expertise to generate attack signatures and modify system thresholds and settings. But as threats become intense and complex, human efforts to scale these procedures become less efficient. You may wish to automatically compare the features (content, location, speed, or time series) of a certain traffic behavior with another. Relying on human expertise is no longer an ideal solution when the traffic is huge in scale, complex in structure, or diverse in behavior. This is where ML comes in.

Intelligent detection method based on ML is capable of analyzing complex unstructured data and provides solution to tasks with rules that are hard to define, some of which include IoT-driven vehicles, diversified network applications, exponential heterogeneous data, and most importantly, automated intelligent cyber threats. Thus, ML-powered cyber defense has become a major development focus for network security solution vendors.

For many years, Genie Networks has been playing a major role in the development of network traffic analysis and DDoS security for the communications service provider market. While leading technologies like Big Data and ML continue to evolve and mature, the idea to incorporate these technologies into our solutions has become inevitable. ML allows the collection of massive traffic data, and the extraction and selection of the appropriate attribute features for the construction of a multi-vector anomaly detection model. The goal is to create a human-like system that learns and grows from its experiences and ultimately achieve fast and precise DDoS protection. Through this, we are helping our customers stay ahead of the battle against new intelligent cyber threats.

GenieNetworks

Genie Networks Limited | sales@genie-networks.com
+886 2 2657 1088 | www.genie-networks.com