Chapter 39
Predictive Network Defense:
Using Machine Learning Algorithms to Protect an Intranet from Cyberattack

Misha Voloshin
Mighty Data, Inc., USA

ABSTRACT

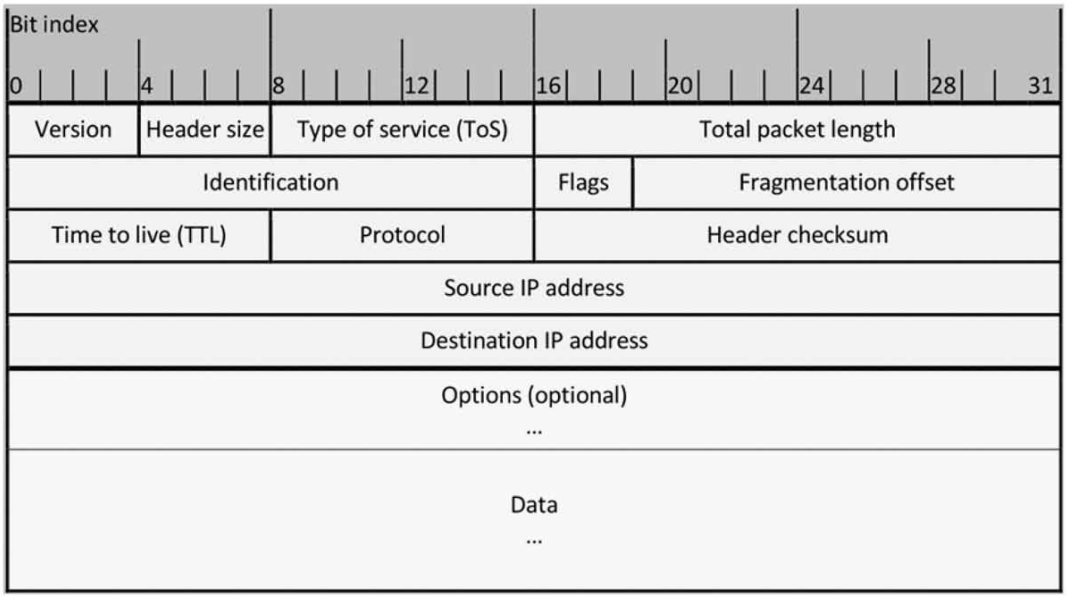
Maintaining electronic devices in today's networked world is not for the faint of heart. The modern network administrator is tasked not only with keeping machines running but also with standing a constant and unerring vigil against cyberattack. A skilled admin learns to identify telltale signs that the network is in trouble, and to quickly evict intruders, repair damage, and reinforce the network's fortifications. No software today can replicate a trained admin's experience and talent. However, just as viruses and rootkits grow progressively more sophisticated with each passing year (Parikka, 2007), so too do the tools to combat them. The information security industry provides admins with alert systems, dashboards, and traffic analysis tools to the tune of $100B per year and growing (Selma Institute of Technology, 2010). This chapter explores ways that algorithms from the fields of machine learning and predictive analytics can be added to this arsenal of the network administrator, helping digital defenders tip the scales of cybersecurity in their favor.

1. TRAINING FIREWALLS TO FILTER PACKETS WITH ID3 DECISION TREES

When it comes to protecting the network against external threats, the standard first line of defense is the network's firewall (Stewart, 2010). This device guards the border between the network and the Internet, monitoring packets as they enter and leave and directing them according to programmable rules. Administrators typically spend countless hours configuring their firewalls, crafting highly specialized rule sets to allow or deny packets of specific types from specific sources. This section explores how to use a type of machine learning algorithm called a decision tree to help the admin automate part of this firewall configuration process.

DOI: 10.4018/978-1-5225-1759-7.ch039

Figure 1. Structure of an IPv4 packet. Each row represents 32 bits, or 4 bytes. The first 20 bytes of the IPv4 packet comprise the packet's header (not counting the optional and rarely-used Options field), which is rich in metadata about the packet. These first 20 bytes can be seen as a 160-bit Boolean vector.



1.1. Background: Parsing IP Packet Structure

When a firewall receives network traffic from the outside world, that traffic comes in the form of packets, i.e. variable-length sequences of bits. These bits encode information such as the traffic's origin, destination, version number, service type, data payload, and so on. A network-capable machine uses the values of bits in specific positions in the packet in order to determine information about that packet and to decide how to handle that packet; in the case of a firewall, the values of bits in specific positions in the packet, among other things, inform the firewall about whether or not the packet should be permitted onto the firewall's subnet.

The order in which specific bits encode these pieces of information is an industry standard called Internet Protocol, or IP. Currently, devices in widespread use predominantly use version 4 of the IP standard protocol (commonly called IPv4), though version 6 (IPv6) is slowly gaining in adoption. The structure of an IPv4 appears below (Fall, 2011).

As the diagram illustrates, the first four bits of a packet carry the IP version number; in the case of IPv4, these will always be set to 0100, representing the quantity 4. The next four bits tell the device that receives the packet how long the packet's header is, measured in 32-bit words; these bits will typically be set to 0101, representing the value 5 (though the header can be longer if the packet contains IP options). The eight bits after that tell the receiver the packet's Type of Service, which is currently used as an aggregate field containing information about the packet's self-declared priority, congestion handling behavior, and so on.

The protocol field and data portion of an IP packet warrant special attention. The values of the eight bits in the protocol field determine how the packet's recipient is intended to interpret the bits in the data