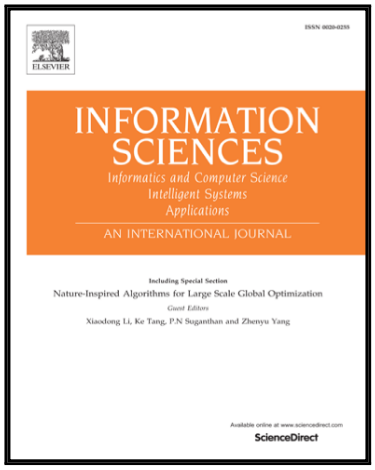
Journal Pre-proof



A Fully Scalable Big Data Framework for Botnet Detection Based on Network Traffic Analysis

S.H. Mousavi, M. Khansari, R. Rahmani

PII: S0020-0255(19)30970-3
DOI: https://doi.org/10.1016/j.ins.2019.10.018
Reference: INS 14934

To appear in: *Information Sciences*

Received date: 27 May 2018
Revised date: 3 October 2019
Accepted date: 12 October 2019

Please cite this article as: S.H. Mousavi, M. Khansari, R. Rahmani, A Fully Scalable Big Data Framework for Botnet Detection Based on Network Traffic Analysis, *Information Sciences* (2019), doi: https://doi.org/10.1016/j.ins.2019.10.018

A Fully Scalable Big Data Framework for Botnet Detection Based on Network Traffic Analysis

S.H. Mousavi[a], M. Khansari[a,*], R. Rahmani[b]

[a] *Faculty of New Sciences and Technologies, University of Tehran, North Kargar Street, Tehran, Iran*
[b] *DigikalaNEXT Research, Tehran, Iran*

Abstract

Many traditional Botnet detection methods have trouble scaling up to meet the needs of multi-Gbps networks. This scalability challenge is not just limited to bottlenecks in the detection process, but across all individual components of the Botnet detection system including data gathering, storage, feature extraction, and analysis. In this paper, we propose a fully scalable big data framework that enables scaling for each individual component of Botnet detection. Our framework can be used with any Botnet detection method - including statistical methods, machine learning methods, and graph-based methods. Our experimental results show that the proposed framework successfully scales in live tests on a real network with 5Gbps of traffic throughput and 50 millions IP addresses visits. In addition, our run time scales logarithmically with respect to the volume of the input for example, when the scale of the input data multiplies by 4x, the total run time increases by only 31 percent. This is significant improvement compared to schemes such as Botcluster in which run time increases by 86 percent under similar scale condition.

Keywords: Botnet Detection, Big Data, Hadoop, Spark, Machine Learning, Scalability

1. Introduction

The rapid growth and scale of current internet applications and the rise in network security threats have rendered many traditional network security methods ineffective[23]. Today, Botnets are among the most serious network security threats[27]. Each Botnet consists of a network of infected hosts called Bots that are usually distributed throughout the internet. Each Bot places a software agent in its infected host that is controlled by a unique person or business called the Botmaster[20]. The communication between Bots and Botmaster is known as Command & Control channels (C&C channels). Hosts are commonly infected without any awareness of their owners and so Botmasters use them to perform malicious activities throughout the Botnet network. These malicious activities may include Distributed Denial of Service (DDoS) attacks, spam emails, malicious adware, identity theft, sensitive information collection, malware distribution, and cyber terrorism[21, 15]. The rapid growth and scale of current internet applications and the rise in network security threats have rendered many traditional network security methods ineffective[23]. Today, Botnets are among the most serious network security threats[27]. Each Botnet consists of a network of infected hosts called Bots that are usually distributed throughout the internet. Each Bot places a software agent in its infected host that is controlled by a unique person or business called the Botmaster[20]. The communication between Bots and Botmaster are known as Command & Control channels (C&C channels). Hosts are commonly infected without any awareness of their owners and so Botmasters use them to perform malicious activities throughout the Botnet network. These malicious activities may include Distributed Denial of Service (DDoS) attacks, spam emails, malicious adware, identity theft, sensitive information collection, malware distribution, and cyber terrorism[21, 15].

*Corresponding author
Email addresses: s.hadi.mousavi@ut.ac.ir (S.H. Mousavi), m.khansari@ut.ac.ir (M. Khansari), r.rahmani@digikala.com (R. Rahmani)