HTTP Bot Detector – DETECTIVE

Minor Project

IBM Career Education

Disclaimer

This Software Requirements Specification document is a guideline. The document details all the high level requirements. The document also describes the broad scope of the project. While developing the solution if the developer has a valid point to add more details being within the scope specified then it can be accommodated after consultation with IBM designated Mentor.

INTRODUCTION

The purpose of this document is to define scope and requirements of a HTTP Bot Detector – DETECTIVE for a small forward looking cooperative bank who wanted to provide enhanced protection to its internal machines. Currently the bank's internal machines are behind a DMZ/Firewall but it does not have mechanism to meaningfully detect any botnets i.e. one or more "compromised" computers that are remotely controlled by a Bot Master under a Command & Control infrastructure.

The proposed system will provide an effective way to detecting potential HTTP command & control activity based on repeated HTTP connections to a Bot Master website.

This document is the primary input to the development team to architect a solution for this project.

System Users

The IT Security Staff of the bank will primarily use the HTTP Bot Detector, DECTECTIVE.

Assumptions

1. It will use firewall logs to track any repetitive HTTP activity(s).
2. It will not support other mechanisms such as IRC.
3. A white list of regularly accessed site will be maintained separately.

REQUIREMENTS

A host infected with HTTP driven malware does not receive instructions directly from a Bot Master website because a TCP connection cannot be initiated and maintained by it. Instead, the infected host must initiate a TCP session and send an HTTP GET request to the Bot Master website to request a command. The website will then send commands to the infected host as a response to the HTTP GET request. This means the botnets have to regularly "poll" a Bot Master.

DETECTOR tries to spot such polling activities and report them. It also takes in to account the fact that a system may not be in use continuously and may be intermittently switched off (e.g. laptop may be taken out or user may put it in sleep mode).

Basic System Operation

All the essential operation steps are outlined below:

1. DETECTOR removes all the access to white listed websites from the log. Such a list is separately maintained and will typically include common business related site along with commonly accessed websites such as Google, twitter, etc.

2. DETECTOR will first build a table of each machine wise http access activities. Activities, in terms of time intervals, will be listed for each uniquely accessed site. For example, if a machine accessed "acme.com" at 10:10, 10:20 and 11:30, 11:36 hours then for that machine, access time intervals will be 10, 70, and 06 minutes. Only high volume activities (# of accesses > 20) are shortlisted for further analysis.
3. It will then find out the outliers time intervals for each shortlisted machine & website combination. Outlier time interval will be detected using a robust statistical technique, where all the entries that are 2.5 times MAD (median absolute deviation) away from the median are removed.
4. From the balance entries, mean and standard deviation is computed. If the standard deviation is very small in context of the mean then such machine/website combination is reported as suspect for further analysis. A smaller standard deviation indicates "polling" activity by the botnet.

DEVELOPMENT ENVIRONMENT

DETECTIVE will be developed as a web application using Java/JSP and DB2 database. Eclipse will be used as the IDE for the same. Please refer to statistics textbooks on median, mean, standard deviation and MAD. Details about botnet may be found at http://en.wikipedia.org/wiki/Botnet URL.