Detecting P2P Botnets through Network Behavior Analysis and Machine Learning

Wu You

2018 年 5 月 3 日

目录

1. Botnet 简介

1.1. Botnet 定义

Botnet 也就是我们所说的 Botnet 网络，是指采用一种或多种传播手段，将大量主机感染 bot 程序（Botnet 程序），从而在控制者和被感染主机之间所形成的一个可一对多控制的网络。

往往被黑客用来发起大规模的网络攻击，如分布式拒绝服务攻击 (DDoS)、海量垃圾邮件等，同时黑客控制的这些计算机所保存的信息，譬如银行帐户的密码与社会安全号码等也都可被黑客随意“取用”。因此，不论是对网络安全运行还是用户数据安全的保护来说，Botnet 网络都是极具威胁的隐患。Botnet 网络的威胁也因此成为目前一个国际上十分关注的问题。然而，发现一个 Botnet 网络是非常困难的，因为黑客通常远程、隐蔽地控制分散在网络上的“Botnet 主机”，这些主机的用户往往并不知情。因此，Botnet 网络是目前互联网上黑客最青睐的作案工具。

Botnet 的主要特征表现是其网络架构，根据这一特征可将其分为两类：一种是基于互联网中继聊天或 HTTP 协议的中心化 Botnet，另一种则是基于端到端（P2P）网络结构的非中心化 Botnet。其中中心化 Botnet 是比较容易检测和摧毁的，而端到端 Botnet 的检测工作则难度较大。

有人将 Botnet 攻击过程分为四个阶段，分别为：形成阶段 (Formation)，命令及控制 (Command and Control, C and C)，攻击阶段以及攻击后 [1]。多束研究工作都致力于在形成或攻击阶段对 Botnet 进行检测，本文主要研究在命令及控制阶段的探测问题。

2. 研究方法

对网络信息流通的分析方法分为基于端口、基于协议以及基于网络行为的分析。其中，基于端口的分析方法错误率较高，因为有大量的网络应用并不基于 TCP/UDP 协议来工作的。基于协议的分析方法主要通过对数据包的有效载荷（Payload）来实现，这是错误率最低的一种方法，但它同时存在两个问题，首先这种方法对计算量要求较大，会对网络性能造成影响，其次，对数据包内容的分析将涉及隐私问题。基于网络行为的分析也就是通过检测网络中的特殊行为来判断异常，由于网络日志可以从设备中轻松提取，所以这一方法对于网络性能和服务可用性的影响很小。

本文采用了基于网络行为的分析方法，这一方法在实际应用中所面临的问题是，Botnet 的网络行为会经常发生变化，因此会产生查新问题，从而影响准确率。本文共使用了 17 数据特征分为基于流量和基于主机两类特征。如下图所示。

本文选用了 NNC（Nearest Neighbor Classifier），SVM, ANN, NBC 以及 GBC（Gaussian based classifer）算法，对其性能进行对比。

Feature	Description	Type
SrcIP	Flow source IP address	flow
SrcPort	Flow source port	flow
DstIP	Flow destination IP address	flow
DstPort	Flow destination port	flow
Protocol	Transport layer protocol	flow
Pack length	Payload size in bytes	flow
APL	Average packet length per flow	flow
FPL	The length of the first packet in the flow	flow
TPC	The total number of packets per flow	flow
TBT	Total number of bytes per flow	flow
IOP	The ratio between the number of incoming packets over the number of outgoing packets	flow
DPL	The total number of subsets of packets of the same length over the total number of packets in the same flow	flow
PL	The total number of bytes of all the packets over the total number of packets in the same flow	flow
SPDP	The ratio between the number of source ports to the number of destination ports	host
CDA	The number of connections over the number of destination IP addresses	host
TPDA	The sum of the numbers of different transmission protocols (usually 1 or 2) used per destination IP over the total number of destination IPs	host
DASP	The number of destination IPs connected to the same open port in the monitored host over the total number of open ports in the monitored host	host

3. 实验结果及总结

以四个指标来衡量算法的性能，分别为：训练速度、分类速度、正确识别率、整体错误率。最终，NNC、ANN 和 SVM 算法是整体性能最佳的三种算法，SVM 拥有最高的正确率和最低的错误率，但训练时间和分类时间都最长。存在的问题：所有的算法都无法满足查新和适应性的要求，需要一种新的算法或对算法进行混合。

参考文献

[1] Justin Leonard, Shouhuai Xu, and Ravi Sandhu. A framework for understanding botnets. In International Conference on Availability, Reliability and Security, pages 917–922, 2009.