基于聚类分析的僵尸网络识别系统①

Botnet Recognition System Based on the Clustering Technology

李晓桢[1] 程 佳[1] 胡 军[2]

(1.江南计算技术研究所 江苏 无锡 214083; 2.电子部 14 所 上海 210013)

摘 要： 僵尸网络是一种恶意的攻击平台，为攻击者提供了灵活、高效并且隐蔽的控制和攻击方式，对网络安全造成了严重威胁，尤其是我国大陆地区。本文从僵尸网络的基本要素着手，把握其本质特征，介绍了一种基于聚类分析技术的僵尸网络识别系统。它独立于僵尸网络的协议和结构，具有较好的适应性和较高的识别率。

关键词： 僵尸网络 聚类技术 x-均值

1 引言

僵尸网络是随着智能程序的应用而逐渐发展起来的，它起源于 1993 年在 IRC 聊天网络中出现的 Eggdrop，这是一种智能程序，能够自动地执行如防止频道被滥用、管理权限、记录频道事件等功能。但这种设计思路被黑客所利用，他们编写出恶意的 bot 程序，对大量的受害主机进行控制，并利用其资源达到恶意的目的。

在之后的近 15 年时间里，随着网络的普及和发展，各种攻击技术也不断出现和成熟，这在无形中带动并促进了僵尸网络的发展和演变。它从传统的恶意代码形态开始，在计算机病毒、网络蠕虫、特洛伊木马和后门工具的基础上进化，并通过相互的融合发展成为一种复杂的攻击方式，并逐渐成为近年来危害互联网安全的重大威胁之一。

僵尸网络就是攻击者手中的一个攻击平台，通过该平台，攻击者不仅可以发起强有效的 DDoS 攻击、发送海量垃圾邮件和传播恶意代码，还可以通过 bot 收集受感染主机中的敏感信息和进一步组建规模更大的 botnet。据有关安全报告指出[1]，目前我国已成为感染僵尸程序计算机数量最多的国家，且这些计算机多数是被其他国家或地区所控制，给我国的互联网安全造成了极大的威胁和隐患。我们需要有一种行之有效的方法来对其进行识别和检测。

2 僵尸网络简介

2.1 僵尸网络的定义

所谓僵尸网络，是指攻击者利用僵尸程序在互联网用户的计算机中秘密建立起来的可以统一控制的计算机群。攻击者控制计算机的方式就是采用一种或多种传播手段，在可控的计算机中植入一种称为僵尸程序(bot)的恶意程序，该程序秘密运行在被控计算机中，可以接收预定义的命令和执行预定义的功能，其本质就是一个网络客户端，可以主动连接控制服务器读取控制指令，按指令执行相应的代码。

如图 1 所示，僵尸网络主要是由控制者、僵尸主机和命令与控制服务器(Command & Control Server,C&CS)构成，控制者通过命令与控制服务器对受控的僵尸主机实行一对多的控制。

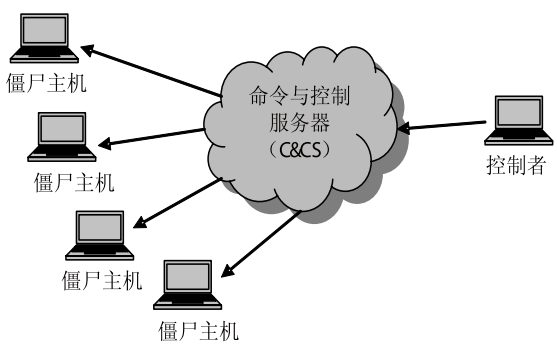

图 1 僵尸网络的基本结构

① 收稿时间:2008-12-08

2.2 僵尸网络的命令与控制机制

僵尸网络是在其他恶意软件的基础上发展起来的，但与其他的恶意软件不同，僵尸网络是一种有组织的计算机群，这种组织关系主要是通过其命令与控制机制来实现。总的来说，僵尸网络的命令与控制机制主要有以下三种：

(1) 集中式的命令与控制机制

该类型僵尸网络的代表主要是 IRC 僵尸网络和 HTTP 僵尸网络。在这种控制机制下，攻击者通过中心式的命令与控制服务器与受控的僵尸主机进行交互，向其传达攻击指令并接受其状态报告。这种机制的优点是效率较高，但缺点是容易存在单点故障。其典型案例有 GTBot、AgoBot、Bobax 等。

(2) 基于 P2P 的命令与控制机制

该类型的 P2P 僵尸网络不存在充当命令与控制服务器的中心节点，网络中的节点同时充当服务器和客户机，相互之间是对等的关系，所以不存在单点故障，很难发现和摧毁。但其实时性和可控性不如集中式的命令与控制机制。其典型案例主要有 Sinit、Phatbot、Storm 等。

(3) 随机命令与控制机制

这是 Evan Cooke[2]在 2005 年提出的一种模型。这种模型在现有的僵尸网络中还没有被发现，但它具有很高的存活性。在这种模型中，bot 不主动与其他 bot 或攻击者通信。攻击时，攻击者通过扫描 Internet 来发现 bot。随机命令与控制模型易于实现且很难被发现和摧毁，但它不具备可测量性，可控性差，很难实现大规模协同攻击。

2.3 僵尸网络的三要素

僵尸网络之所以称之为僵尸网络，主要是因为它包含了三个方面的要素：恶意的、可控的、主机群。这三个方面缺一不可，下面分别进行介绍：

(1) 恶意的

僵尸网络之所以成为互联网面临的重大威胁之一，其中一个主要的原因就是它为攻击者提供了一个大规模的攻击平台，使其可以方便地进行各种有效的网络攻击，而这种攻击通常都是出于恶意的目的。

根据[3]的介绍，当前监测到的活跃的 IRC 僵尸网络中，大约 53%的命令都是用来进行扫描的，进行扫描的目的主要是为了进一步的传播或者进行 DDos 攻击，大约 14.4%的命令是为了进行二进制的文件下载(出于僵尸程序更新的目的)。而 HTTP 和 P2P 僵尸网络则大部分是用来发送垃圾邮件。

(2) 可控的

僵尸网络之所以能形成规模并且方便地让攻击者使用，主要是因为它的形成是建立在可控的基础上。攻击者通过特定的命令与控制信道，向网络内的僵尸主机发布各种命令，僵尸主机也通过此信道向攻击者汇报执行命令的情况或者自身的状态等等。这样，在攻击者和命令控制信道之间，尤其是僵尸主机与命令控制信道之间，不可避免地就会出现相应的通信流量(无论是集中式的控制方式还是分布式的控制方式)。

(3) 主机群

僵尸网络之所以存在，其基本的条件就是有大量的僵尸主机供其差遣。如今的僵尸网络较之以前相比，其灵活性、隐蔽性更强，规模也更小，但正是因为如此，同一僵尸网络内的主机往往都会被指示去做同一种的攻击，这样，各台主机便会具有类似的通信模式和相同的攻击行为。否则，整个网络内的主机若是各司一职，就与单个的肉鸡无异，失去了僵尸网络本身的意义。

3 基于聚类技术的主机分类算法

目前已有的僵尸网络识别技术大多数是针对 IRC 类型的，针对 HTTP 僵尸网络和 P2P 僵尸网络的识别技术则相对欠缺。且当前已有的识别技术多数是针对某种特定的僵尸网络，局限性较大，有效性和灵活性也不是很好。

本文主要是从僵尸网络的三要素出发，抓住其本质的特征，设计出这样一种识别系统：它独立于僵尸网络所采用的协议及控制方式，不受僵尸程序结构的变化，并且可以适应不断变化的僵尸网络命令与控制服务器地址的变迁，且不需要预先知道僵尸网络的任何特征。

该系统主要采用一种改进的 k-均值聚类算法来对网络中的流量进行分析，下面首先对其进行介绍：

3.1 聚类技术简介

聚类技术采用一种无监督学习的方法，将一批数据依照它们的相似特征分成不同的类簇(cluster)，同一簇中的对象具有较高的相似度，不同簇中的对象彼此差别较大。总的来说，聚类技术可以分为以下几大类：分裂法，层次法，基于密度的方法，基于网格的

方法和基于模型的方法。本课题采用的其中的分裂法(Partitioning Methods)，其原理如下：给定一个有 N 个元组或记录的数据集，分裂法将构造 k 个分组，每一个分组代表一个聚类。对于给定的 k，算法首先确定一个初始的分组方法，并通过不断的迭代改变分组，使每一次改进后的分组都较前一次好，即使同一分组中的距离越来越近，不同分组的距离越来越远。

3.2 k-均值聚类算法

k-均值聚类算法是分裂法中较为典型的一种算法，它的工作过程说明如下：①从 n 个数据对象中任意选择 k 个对象作为初始聚类中心；②对于所剩的其他对象，则根据它们与这些聚类中心的相似度(距离)，分别将它们分配给与其最相似的聚类；③计算每个聚类的新聚类中心(该聚类中所有对象的均值)；④不断重复②③直至标准测度函数(通常都采用均方差作为标准测度函数)开始收敛。

k-均值聚类算法思想简单，具有很好的局部收敛性和广泛应用性，但它也存在着计算复杂度高、运行时间长以及可能收敛到局部最小等缺点，尤其是该算法中的 k 值必须事先给定，对多变的网络数据来说其效果就大打折扣。

3.3 x-均值聚类算法

针对上述 k-均值算法的各种不足，本文采用了一种改进的 x-均值算法[4]。x-均值首先加快每一次 k-均值，在 k-均值的每一次迭代中，用每一个聚类中心的信息代表它所在的子集，并且利用 kd-树的层次结构大幅度提高其求解速度。有了快速的 k-均值算法就可以求解最优的 k 值，x-均值算法在指定的上、下届范围内寻找最优的 k 值，其最优性的判断标准是基于贝叶斯信息标准(BIC：Bayesian Information Criterion)：$BIC(C|X)=L(X|C)-(p\log n)/2$，其中 $L(x|c)$ 是数据集 X 关于模型 C 的对数似然，$p=k(d+1)$ 是 d 维的 k 个聚类中心的模型 C 中的参数个数，n 是样本数。

总的来说，x-均值主要由以下两步组成：(1)改进聚类参数，即搜索最优的 k 个中心；(2)改进聚类结构，即搜索最优的 k 值。

4　基于聚类分析的僵尸网络识别系统的设计与实现

本系统主要是从网络流量出发，对其中的通信过程和恶意行为进行检测，并引用 x-均值聚类算法对检测的结果进行聚类分析，再经过关联分析模块进行综合考评，从中识别出可疑的僵尸主机。

4.1 系统的结构和框架

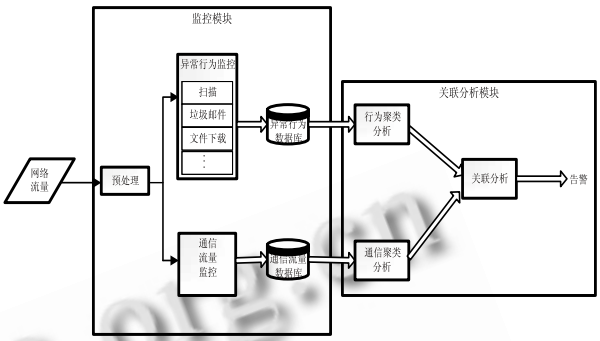

图 2　识别系统的结构和框架

如图 2 所示：该系统主要由两个模块构成：监控模块和关联分析模块。其中监控模块主要部署在要监控网络的出入口处，对进出口的流量进行监控和采集，关注并记录当前网络中的通信流量数据，同时对异常行为(如网络扫描、发送垃圾邮件、恶意程序下载等等)进行检测，并将结果存至相应的数据库，以便关联分析模块进行分析；关联分析模块分别读取异常行为数据库和通信流量数据库中的内容，对其进行聚类分析，并将聚类的结果送至关联分析模块综合进行评估，从中发现可疑的僵尸主机。

在理想的情况下，监控模块应该分布在多个网络出入口处，对整个网络情况进行监控，并将采集的日志信息上报给统一的关联模块进行分析。

4.2 监控模块

监控模块主要由预处理、异常行为监控和通信流量监控三个部分组成。其中预处理主要是通过协议匹配、黑白名单等手段实现对网络中不必要的数据流进行过滤，以提高整个系统的运行效率；异常行为监控主要是从僵尸网络的恶意性着手，针对其工作过程中可能出现的一些恶意行为进行检测，如对外扫描、发送垃圾邮件、恶意文件下载等；通信流量监控则关注僵尸主机与命令控制服务器的通信过程，对进出的流量实施监控，并以特定的格式将其存入数据库，以方便关联分析模块从中提取合适的特征进行聚类分析，进而发现可疑的僵尸主机与服务器的通信过程。

4.3 关联分析模块

关联分析模块主要是由三个部分组成：通信聚类

分析、行为聚类分析和关联分析。

通信聚类分析关注的是 who is talking to whom，它首先按照协议、源地址、目的地址和目的端口将具有相同元素的数据流划分到一个集合 C_i，针对每个集合计算出其 fph、ppf、bpp、bps(即每小时的数据流数、每个数据流的收发包数、每个数据包的平均字节数、每秒钟传输的平均字节数)，将其整合后作为该集合的向量表示用 x-means 算法进行聚类分析，将原本杂乱无章的网络流量划分为一个个小的类簇，其中每个类簇中的集合都具有类似的通信模式。

行为聚类分析关注的是 who is doing what，在这里由于监控的恶意行为有限，我们仅对其进行粗略的聚类，将具有相同异常行为的主机划分到一个行为类簇中。

单独通过通信聚类分析和行为聚类分析都不足以准确地定位网络中的僵尸主机，我们必须将二者结合起来进行关联分析。其思路如下：首先，我们从行为聚类分析的结果中选出有恶意行为的主机 h，然后计算其可疑度 s(h)，并与给定的检测阈值 θ 进行比较，若大于 θ，则认为是可疑的僵尸主机，否则就认为是正常的。其中可疑度 s(h)的定义如下：

假设有恶意行为的主机集合为 H(由行为聚类分析的结果给出)，对其中的每台主机 h(h∈H)，

$$s(h)=\sum_{\substack{i,j\\ j>i\\ t(A_i)\neq t(A_j)}} w(A_i)w(A_j)\frac{|A_i\cap A_j|}{|A_i\cup A_j|}+\sum_{i,k} w(A_i)\frac{|A_i\cap C_k|}{|A_i\cup C_k|}$$

其中，A_i, $A_j\in A_{(h)}$,$C_k\in C_{(h)}$,$A_{(h)}=\{A_i\}i=1,\cdots,m_h$ 为包含主机 h 的 m_h 个行为类簇的集合，$C^{(h)}=\{C_i\}$ $i=1,\cdots,n_h$为包含主机 h 的 n_h 个通信类簇的集合，$t(A_i)$ 代表行为类簇 A_i 的类型(扫描、发送垃圾邮件等)，$w(A_i)$ 代表不同 A_i 的权重，$w(A_i)\geq 1$ 且不同 A_i 的权重各不相同，取决于该类簇代表的恶意行为出现的几率，恶意行为出现的可能性越大，其 $w(A_i)$越大。

从 s(h)的定义中我们可以看出，主机 h 所执行的恶意行为的种类越多，其可疑程度越大；其所在的类簇与其他类簇的重叠部分越多，h 的可疑程度也越大。举个例子来说，假设 h 同时进行了对外扫描和漏洞破解两种行为，A_1、A_2 分别代表了 h 所在的扫描类簇和漏洞破解类簇中所有主机的集合，若 A_1 和 A_2 的重叠部分越多，则 $A_1\cap A_2$ 的值越大，相应 s(h)的值也越大。同理，若 h 所在的恶意行为类簇与通信类簇重叠部分越多，则说明相应的主机不仅执行相同的恶意行为，而且具有相同的通信模式，它们是同一僵尸网络中的僵尸主机的可能性也就越大。

在计算出每台主机的可疑度 s(h)后，我们将其与预先定义好的检测阈值 θ 相比较，若大于该阈值，则认为该主机是可疑的，否则便认为是正常的并予以过滤。在本系统中，为了方便起见，我们设所有的 $w(A_i)$ 均为 1，且 θ =0。也就是说所有出现在多个行为类簇中的主机以及所有与其他主机拥有相同通信模式并且执行了恶意行为的主机都被认为是可疑的。

5 结束语

本系统主要是从僵尸网络的本质特征着手，抓住其不可或缺的三个要素，关注并识别僵尸主机与命令控制服务器的通信过程，结合僵尸网络可能实施的各种恶意行为，通过聚类和关联分析实现对可疑僵尸主机的识别。该系统在实验环境下取得了很好的效果，具有较高的识别率和较低的误报率。但该系统也存在着很多的不足之处，如只能对已感染的活跃的僵尸主机进行识别，对感染过程中或暂时休眠的僵尸主机则不能准确识别等。

随着网络技术的不断发展，僵尸网络也在不断进化，从 IRC 协议到 HTTP 协议，再从 HTTP 协议到 P2P 协议，僵尸网络的控制模式由集中转为分布，其行为越来越隐蔽，扩展性也越来越好。我们在识别和检测僵尸网络的道路上还有很长的路，需要更多的关注和努力。

参考文献

1 CNCERT/CC2007 年网络安全工作报告 http://www.cert.org.cn/UserFiles/File/CNCERT2007AnnualReport_Chinese.pdf.

2 Cooke E, Jahanian F, McPherson D. The Zombie Roundup: Understanding Detecting, and Disrupting Botnets. SRUTI Workshop July7, 2005.

3 Zhuge J, Holz T, Han X, Guo J, Zou W. Characterizing the irc-based botnet phenomenon. Peking University & University of Mannheim Technical Report, 2007.

4 Pelleg D, Moore AW. X-means: Extending k-means with efficient estimation of the number of clusters. Processings of the Seventeenth International Conference on Machine Learning.