**文章编号**:1671-8860(2012)02-0247-05　　**文献标志码**:A

# 僵尸网络关系云模型分析算法

臧天宁[1,2,3]　云晓春[1,2,3]　张永铮[1,3]　门朝光[2]

(1　中国科学院信息工程研究所,北京市海淀区闵庄路 27 号,100097)
(2　哈尔滨工程大学计算机科学与技术学院,哈尔滨市南通大街 145 号,150001)
(3　信息内容安全技术国家工程实验室,北京市海淀区闵庄路 27 号,100097)

**摘　要**:通过分析僵尸网络内部的通信行为,提取了相同僵尸网络的通信特征,利用这些特征定义了僵尸网络之间关系的云模型,并设计了基于云模型的僵尸网络关系分析算法。通过典型僵尸程序样本的评测结果表明,即使对采用加密通信和无固定通信时间间隔的僵尸程序,该算法仍然能够有效地识别出这些僵尸网络之间的关系。通过与相关研究工作的对比表明,该算法在分析的准确度、僵尸网络的类型和加密通信等方面均优于相关研究成果。

**关键词**:僵尸网络;云模型;迁移;相似度

**中图法分类号**:P208;TP393

僵尸网络(botnet)是攻击者在互联网上秘密建立的可控计算机群[1]。僵尸程序(bot)与蠕虫等传统网络恶意软件的一个重要区别是其与命令和控制服务器之间建立的通信机制。该机制主要采用 IRC[2]、HTTP 和 P2P 等通信协议,其中基于 IRC 和 HTTP 协议的僵尸网络构建了一对多集中式结构的命令与控制信道[3]。以往的研究报告中,一个典型集中式僵尸网络受控主机的数量在十万台以上[4,5],而实际上,随着僵尸网络的迅速发展,逐渐呈现了小型化、分散化和专业化的趋势,大的僵尸网络被划分成小的僵尸网络群[6,7],从包含十万个以上主机的大型网络演变成含有 1 000～5 000 个主机的小型僵尸网络群[8]。Vogt[9]等提出僵尸网络军团(army of botnets)和超僵尸网络(super-botnet)的概念,几个集中式僵尸网络由控制者统一管理,协同工作。另外,集中式僵尸网络经常更换命令和控制服务器的 IP 或域名发生迁移行为[10],不同时间检测到的两个僵尸网络可能是同一僵尸网络在迁移前后的不同表现。因此,有必要研究僵尸网络之间的关系,以便掌握发生在不同地理位置、不同时间段的安全事件之间的内在关联。Rajab[11]等在评估僵尸网络规模的过程中,根据蜜罐获取的 IRC 通信的相关信息提出了度量僵尸网络关系模型,并分析了僵尸网络的迁移问题。但其前提条件是必须掌握僵尸网络各种详细的通信信息,而且主要是面向应用明文通信的 IRC 僵尸网络。为此,本文针对集中式结构僵尸网络分析了两个已知僵尸主机群之间的关系。

## 1　僵尸网络相似性特征

采用不同协议的集中式僵尸网络,其通信方式存在差别[12]。IRC 僵尸网络应用“推”的方式,控制者以“群聊”或“私聊”的方式主动向连接到同一 IRC 频道的僵尸成员“推”送指令,僵尸程序执行命令后,将结果等信息在该频道反馈回服务器。HTTP 僵尸网络应用“拉”的方式,控制者把命令文件上传到控制服务器,僵尸程序定时发 HTTP 请求拉取命令文件,服务器应答僵尸程序 HTTP 请求,返回命令文件。两种方式都可以看作是僵尸主机与控制服务器间建立起固定虚拟会话信道。僵尸主机在这些虚拟信道上机械地执行相同的通信进程,按固定方式通信,受到相同控制者的操作,在一段时间内,各主机与服务器间的交互信息具有相似性。另外,同一僵尸网络内受控主机

**收稿日期**:2011-12-15。

**项目来源**:国家自然科学基金资助项目(60703021, 61070185, 60873138);国家 863 计划资助项目(2007AA010501)。

利用相同系统漏洞建立,其用户具有相似的因特网使用习惯[10,11],大部分主机夜间关机下线,僵尸网络内部通信有明显以 1 d 为周期的变化规律。

综上所述,隶属于同一僵尸网络的不同僵尸主机群具有几个日周期变化趋势的相似的特征:① 数据流统计量,反映了僵尸主机群体的在线活动情况;② 数据流中数据包统计量,相同僵尸网络内受控主机在一段时间内与服务器交互的信息比较固定,通信流中数据包的数量相近;③ 主机通信量,同一僵尸网络内的受控主机在控制者的统一操作下,其通信量有相似的变化规律。通过综合分析各通信特征变化趋势的相似性,判断两批僵尸主机间的关系。

## 2 僵尸网络关系云模型分析算法

### 2.1 云模型

云模型[13]是李德毅院士提出的一种定性定量转换模型,能实现定性概念与其数值表示之间的不确定性转换。

定义 1 云和云滴[14]。设 $U$ 是一个用数值表示的定量论域,$C$ 是 $U$ 上的定性概念,若定量值 $x\in U$ 是定性概念 $C$ 的一次随机实现,$x$ 对 $C$ 的确定度 $\mu(x)\in[0,1]$ 是有稳定倾向的随机数,

$$\mu: U\rightarrow[0,1],\ x\in U,\ x\rightarrow\mu(x)$$

则 $x$ 在论域 $U$ 上的分布称为云,记为云 $C(x)$,每一个 $x$ 称为一个云滴。

定义中的随机实现是概率意义下的实现,每次实现的随机样本具有一个确定度是模糊集意义的隶属度,同时又具有概率意义的分布,体现了模糊性和随机性关联。云模型的整体特性用云的数字特征即期望 $E_x$、熵 $E_n$、超熵 $H_e$ 表征,$C(E_x,E_n,H_e)$ 称为云的特征向量。正向云算法和逆向云算法是云模型中的两个重要运算。正向云算法实现概念空间到数值空间的转换;逆向云算法将定量值转换为 $\{E_x,E_n,H_e\}$ 表示的定性概念。由两个(多个)定性概念组合的复杂定性概念可用二维(多维)的期望、熵、超熵数字特征构成二维(多维)云模型描述复杂的定性与定量间的转换[14]。

### 2.2 系统架构

首先识别僵尸网络通信协议,按 RFC1495 定义,IRC 通常以 PASS、NICK 和 USER 三个信息开始,而 HTTP 请求的前几个字节是关键字 GET、POST 或 HEAD。因此,通过检测网络链接前几个字节,可快速判断通信协议[15],如协议不同,判为两个僵尸网络;如协议相同或由于加密通信等原因无法识别,再进一步提取僵尸网络通信特征,根据云模型将定量特征转换为定性概念这一特点,建立僵尸网络通信特征变化趋势的相似度正态云模型。将基准场景下(相同僵尸网络的通信数据)的各特征变化趋势的相似度转换为隶属于同一僵尸网络的概念,利用此概念度量不同僵尸主机群隶属于同一个僵尸网络的程度。

### 2.3 分析僵尸网络之间的关系

根据僵尸网络通信行为的日周期性,取观测周期 $P$ 为 24 h,各特征统计时间段 $I$ 为 1 h,建立特征相似度统计函数,每个时间段的分析结果看作一个云滴。

1) 数据流统计量。以每小时数据流的数量(fph)作为统计值,$\mathrm{FPH}_i(t)$ 表示僵尸网络 $i$ 在第 $t$ 小时内的数据流统计函数。两批僵尸主机在两个连续时间段内的数据流统计量变化趋势的相似度为:

$$S_{\mathrm{fph}}(t,t+1)=\frac{|\ \mathrm{FPH}_1(t+1)-\mathrm{FPH}_1(t)\ |}{|\ \mathrm{FPH}_2(t+1)-\mathrm{FPH}_2(t)\ |}$$

$S_{\mathrm{fhp}}(t,t+1)$ 越接近 1,说明两批僵尸主机的数据流统计量在这两个时间段内的变化情况越相近。

2) 数据包统计量。第 $t$ 个时间段,数据流 $f_i$ 中数据包数量(ppf)的统计值为 $\mathrm{ppf}_j(t)$,由于互联网环境的不稳定性,使得常数据包丢失或数量突增,为消除噪声的影响,计算出现频率前 50% 的 ppf 加权算术平均值。第 $t$ 个时间段僵尸网络 $i$ 的 $\mathrm{ppf}_j$ 的统计函数为:

$$\mathrm{PPF}_i(t)=\sum_{j=1}^{m}\left(\frac{n_j}{N}\times|\ \mathrm{ppf}_j(t)\ |\right)$$

其中,第 $j$ 个 $\mathrm{ppf}_j(t)$ 出现 $n_j$ 次;$m$ 为前 50% 频率的数量;$N$ 是出现频率占前 50% $\mathrm{ppf}_j(t)$ 的总数。$\mathrm{PPF}_i(t)$ 在两个时间段变化的相似度为:

$$S_{\mathrm{ppf}}(t,t+1)=\frac{|\ \mathrm{PPF}_1(t+1)-\mathrm{PPF}_1(t)\ |}{|\ \mathrm{PPF}_2(t+1)-\mathrm{PPF}_2(t)\ |}$$

3) 主机通信量。以 IP 表征在线僵尸主机,第 $t$ 个时间段,IP 的数据流数量(fpi)的统计值为 $\mathrm{fpi}_j(t)$。按 fpi 出现的次数加权算术平均值,计算统计函数:

$$\mathrm{FPI}_i(t)=\sum_{j=1}^{m}\left(\frac{n_j}{N}\times|\ \mathrm{fpi}_j(t)\ |\right)$$

其中,第 $j$ 个 $\mathrm{fpi}_j(t)$ 出现 $n_j$ 次;$m$ 为各 $\mathrm{fpi}_j(t)$ 出现频率的数量;$N$ 是 $\mathrm{fpi}_j(t)$ 的总数。$\mathrm{FPI}_i(t)$ 在两个连续时间段内的相似度为:

$$S_{\mathrm{fpi}}(t,t+1)=\frac{|\ \mathrm{FPI}_1(t+1)-\mathrm{FPI}_1(t)\ |}{|\ \mathrm{FPI}_2(t+1)-\mathrm{FPI}_2(t)\ |}$$

根据以上 3 个特征的统计函数,建立以 1 为

中心的半云模型。

4) 云模型的构造。提取基准场景的通信特征,用逆向云算法生成相同僵尸网络的通信特征变化趋势相似度云的模型。

*算法 1*　通信特征变化趋势相似度逆向云算法。

输入:代表相同僵尸网络特征变化趋势的 $N$ 个云滴$\{(x_{\text{fph}1}, x_{\text{ppf}1}, x_{\text{fpi}1}),(x_{\text{fph}2}, x_{\text{ppf}2}, x_{\text{fpi}2}),\cdots,(x_{\text{fph}N}, x_{\text{ppf}N}, x_{\text{fpi}N})\}$。

输出:这 $N$ 个云滴表示隶属于同一个僵尸网络的云模型,其期望值、熵和超熵分别为$(Ex_{\text{fph}}, Ex_{\text{ppf}}, Ex_{\text{fpi}})$,$(En_{\text{fph}}, En_{\text{ppf}}, En_{\text{fpi}})$,$(He_{\text{fph}}, He_{\text{ppf}}, He_{\text{fpi}})$。

1) 分别计算 $x_{\text{fph}i}$、$x_{\text{ppf}i}$ 和 $x_{\text{fpi}i}$ 的样本均值$(\overline{X}_{\text{fph}},\overline{X}_{\text{ppf}},\overline{X}_{\text{fpi}})$、一阶样本中心矩以及样本方差$(S^2_{\text{fph}},S^2_{\text{ppf}},S^2_{\text{fpi}})$;

2) $Ex_{\text{fph}}$、$Ex_{\text{ppf}}$、$Ex_{\text{fpi}}$的估计值分别为 $\hat{E}x_{\text{fph}}=\overline{X}_{\text{fph}}$,$\hat{E}x_{\text{ppf}}=\overline{X}_{\text{ppf}}$,$\hat{E}x_{\text{fpi}}=\overline{X}_{\text{fpi}}$;

3) 计算 $He_{\text{fph}}$、$He_{\text{ppf}}$、$He_{\text{fpi}}$ 和 $En_{\text{fph}}$、$En_{\text{ppf}}$、$En_{\text{fpi}}$的估计值。

由此得到以期望值为$(Ex_{\text{fph}}, Ex_{\text{ppf}}, Ex_{\text{fpi}})$、熵为$(En_{\text{fph}}, En_{\text{ppf}}, En_{\text{fpi}})$、超熵为$(He_{\text{fph}}, He_{\text{ppf}}, He_{\text{fpi}})$的数字特征,表征隶属于同一僵尸网络的各网络特征变化趋势相似度的云模型。

5) 僵尸网络关系分析算法

*算法 2*　通信特征变化趋势相似度正向云算法。

输入:代表两批僵尸主机通信特征变化趋势相似度的云滴$(x_{\text{fph}}, x_{\text{ppf}}, x_{\text{fpi}})$和隶属同一僵尸网络的云模型。

输出:云滴隶属于同一僵尸网络的确定度。

1) 在区间$[En_{\text{fph}}-He_{\text{fph}}, En_{\text{fph}}+He_{\text{fph}}]$、$[En_{\text{ppf}}-He_{\text{ppf}}, En_{\text{ppf}}+He_{\text{ppf}}]$和$[En_{\text{fpi}}-He_{\text{fpi}}, En_{\text{fpi}}+He_{\text{fpi}}]$上分别生成正态随机数 $En'_{\text{fph}}$、$En'_{\text{ppf}}$ 和 $En'_{\text{fpi}}$;

2) 计算$(x_{\text{fph}}, x_{\text{ppf}}, x_{\text{fpi}})$在两个连续时间段内变化趋势的相似度 $q$:

$$q = \mathrm{e}^{-\frac{(x_{\text{fph}}-Ex_{\text{fph}})^2}{2(En'_{\text{fph}})^2}-\frac{(x_{\text{ppf}}-Ex_{\text{ppf}})^2}{2(En'_{\text{ppf}})^2}-\frac{(x_{\text{fpi}}-Ex_{\text{fpi}})^2}{2(En'_{\text{fpi}})^2}}$$

$q$ 在一个观测周期有 23 个结果,有 $n$ 个小于阈值,通过 $n/23$ 判断两批僵尸主机的关系。若判为相同僵尸网络,则该云滴加入云模型更新调整队列,通过先进先出进行动态自学习。

## 3　原型实现及实验分析

本文方法的实现称为僵尸网络关系云模型分析器(botnet relationship cloud model analyzer, BRCMA),应用典型僵尸样本对其进行评测。

### 3.1　实验环境和评测样本数据

基于 VMware,每台服务器安装 6 个 Windows XP 系统,共 20 台服务器实现 120 个僵尸网络客户端,另外两台主机充当 HTTP 和 IRC 命令以及控制服务器。修改 4 个典型 IRC 僵尸程序样本和 4 个 HTTP 僵尸程序样本,使其自动链接到虚拟网络内的服务器。4 个 IRC 僵尸程序样本为 Spybot、Sdbot、rbot、Agobot,依次编号为 $I_1$、$I_2$、$I_3$、$I_4$,其中 $I_4$ 可加密通信。4 个 HTTP 僵尸程序样本为:$H_1$ 修改了 CERT 研究报告[19]中基于 Web 的 bot 源码,每 5 min 通信一次;$H_2$ 是 Bobax 的修改版,通信时间是 0～10 min 之间的随机值;$H_3$ 修改了 Engergy 1.9.2 源码;$H_4$ 对采用 RC4 算法加密通信的 rustock 进行了修改。

### 3.2　实验结果和有效性验证

每个样本在虚拟网络中运行 5 个周期,通过 Wireshark 获取通信特征,并模拟真实的网络状况,随机设置各虚拟主机的开、关机时间,各客户端在不同时刻与服务器建立链接。

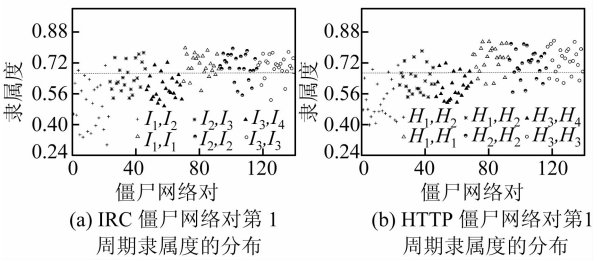
4 个 IRC 僵尸样本两两匹配构成 10 个僵尸网络对,4 个相同僵尸程序对$(I_1,I_1)$、$(I_2,I_2)$、$(I_3,I_3)$、$(I_4,I_4)$的通信数据流为基准场景。5 个周期内共产生 460 个云滴,利用逆向云算法构造三维云模型:$Ex_{\text{fph}}=1$,$En_{\text{fph}}=0.26$,$He_{\text{fph}}=0.1$;$Ex_{\text{ppf}}=1$,$En_{\text{ppf}}=0.24$,$He_{\text{ppf}}=0.12$;$Ex_{\text{fpi}}=1$,$En_{\text{fpi}}=0.27$,$He_{\text{fpi}}=0.11$。同样,通过 4 个相同的 HTTP 僵尸程序对构造 HTTP 僵尸网络通信特征变化趋势相似度云模型:$Ex_{\text{fph}}=1$,$En_{\text{fph}}=0.25$,$He_{\text{fph}}=0.15$;$Ex_{\text{ppf}}=1$,$En_{\text{ppf}}=0.26$,$He_{\text{ppf}}=0.13$;$Ex_{\text{fpi}}=1$,$En_{\text{fpi}}=0.23$,$He_{\text{fpi}}=0.11$。

通过第 1 周期中僵尸网络对产生的云滴计算判别僵尸网络关系的阈值。不同 IRC 僵尸网络对$(I_1,I_2)$、$(I_2,I_3)$、$(I_3,I_4)$和相同 IRC 僵尸网络对$(I_1,I_1)$、$(I_2,I_2)$、$(I_3,I_3)$在第 1 周期中各有 23 个云滴,其隶属度的分布如图 1(a)所示,横坐标为各云滴编号,其中[1,23]、[24,46]、[47,69]、[70,92]、[93,115]和[116,138]分别表示$(I_1,I_2)$、$(I_2,I_3)$、$(I_3,I_4)$、$(I_1,I_1)$、$(I_2,I_2)$和$(I_3,I_3)$产生的云滴。

设 $d$ 代表$(I_1,I_2)$、$(I_2,I_3)$、$(I_3,I_4)$各隶属度的平均值,$s$ 代表$(I_1,I_1)$、$(I_2,I_2)$、$(I_3,I_3)$各隶属度的平均值,则简单计算判别阈值 $T=(s-d)/2+d$。

图 1(b)为 HTTP 僵尸网络对$(H_1,H_2)$、

$(H_2,H_3)$、$(H_3,H_4)$和$(H_1,H_1)$、$(H_2,H_2)$、$(H_3,H_3)$在第1周期中各云滴隶属度的分布图，同样可简单计算HTTP僵尸网络关系的判别阈值。




(a) IRC僵尸网络对第1周期隶属度的分布　(b) HTTP僵尸网络对第1周期隶属度的分布

图1　隶属度的分布

Fig. 1　Distribution of Membership Degree

表1列出了不同僵尸网络对低于判别阈值的云滴在各周期中的数量。根据僵尸网络内部通信的实际情况，在同一周期23个云滴中，如果低于阈值云滴的数量超过12，则认为在该周期识别出了不同的僵尸网络。

**表1　低于阈值的云滴在各周期中的数量**

Tab. 1　Number of Droplets Lower than the Threshold in Every Point

| 僵尸网络对 | 周期 | | | | | 僵尸网络对 | 周期 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | | 1 | 2 | 3 | 4 | 5 |
| $I_1,I_2$ | 14 | 18 | 20 | 18 | 20 | $H_1,H_2$ | 16 | 16 | 18 | 20 | 18 |
| $I_1,I_3$ | 18 | 17 | 20 | 20 | 19 | $H_1,H_3$ | 18 | 18 | 16 | 20 | 19 |
| $I_1,I_4$ | 14 | 17 | 18 | 16 | 18 | $H_1,H_4$ | 16 | 13 | 18 | 19 | 18 |
| $I_2,I_3$ | 15 | 16 | 17 | 18 | 19 | $H_2,H_3$ | 18 | 18 | 15 | 18 | 18 |
| $I_2,I_4$ | 18 | 20 | 21 | 18 | 20 | $H_2,H_4$ | 13 | 19 | 20 | 16 | 18 |
| $I_3,I_4$ | 17 | 20 | 18 | 21 | 19 | $H_3,H_4$ | 16 | 19 | 17 | 20 | 19 |

从检测结果可以看出，不论是IRC僵尸网络，还是HTTP僵尸网络，BRCMA在各周期中低于阈值云滴的数量都高于12，正确区分出所有僵尸网络对之间的关系，没有误报。尽管$H_2$通信时间是随机值，$H_4$和I4应用加密通信，但BRCMA仍很好地识别出了$H_2$、$H_4$与其他僵尸网络之间的关系，进一步验证了本文提出的僵尸网络关系云模型分析方法的有效性。

与本文同样关注僵尸网络之间关系的研究工作有Rajab[11]等提出的度量IRC僵尸网络之间关系的模型。通过比较蜜罐获取的僵尸程序版本、IRC服务器IP、IRC服务器域名、IRC频道名、控制者ID等信息加权和的相似度来确定两个僵尸网络的相近程度，本文称这个方法为$B_{\text{Rajab}}$。其对10个IRC僵尸网络对的分析结果如表2所示，“Y”表示识别出僵尸网络对关系，“N”表示没有识别出。由于$B_{\text{Rajab}}$需要IRC僵尸网络通信的详细信息，而$I_4$应用了加密通信，因此$(I_3,I_4)$和$(I_4,I_4)$都没有区分开。并且由于$B_{\text{Rajab}}$仅通过简单加权和计算各特征的综合相似度，$(I_2,I_3)$关系也没有正确区分出，因此其识别的准确率低于BRCMA方法。

**表2　$B_{\text{Rajab}}$的分析结果**

Tab. 2　Analytical Result of $B_{\text{Rajab}}$

| | $I_1$ | $I_2$ | $I_3$ | $I_4$ |
|---|---|---|---|---|
| $I_1$ | Y | Y | Y | Y |
| $I_2$ | | Y | N | Y |
| $I_3$ | | | Y | N |
| $I_4$ | | | | N |

**表3　两种方法的分析能力比较**

Tab. 3　Comparisons Between Two Algorithms

| | BRCMA | $B_{\text{Rajab}}$ |
|---|---|---|
| 分析类型 | 集中式僵尸网络 | IRC僵尸网络 |
| 加密通信 | 可分析加密通信 | 只分析明文 |
| 分析内容 | 数据流的统计信息 | 需详细信息 |
| 量化方法 | 利用云模型建立隶属于同一僵尸网络概念 | 特征相似度的加权和 |
| 分析准确度 | 没有误报 | 存在误报 |
| 计算量 | 大 | 小 |

## 4　结　语

僵尸网络在迅速发展过程中逐渐呈现出小型化、分散化和专业化的发展趋势，不同僵尸主机群可能隶属于同一僵尸网络。本文设计了基于云模型的僵尸网络关系分析算法。通过典型僵尸程序样本对本文方法进行了评测，结果表明，即使对采用加密通信和无固定通信时间间隔的僵尸程序，该算法仍能有效地计算它们的相似度。通过与相关研究工作的对比表明，该算法在分析的准确度、僵尸网络的类型和加密通信等方面均优于相关研究成果。

## 参　考　文　献

[1] Du Yuejin, Cui Xiang. Malicious Botnet and Its Illumination on Computer Security[J]. China Data Communications, 2005, 7(5):9-13 (Chinese)

[2] Oikarinen J, Reed D. Internet Relay Chat Protocol [S]. Request for Comments (RFC) 1459, IETF, 1993

[3] 于晓聪，董晓梅，于戈. 僵尸网络在线检测技术研究[J]. 武汉大学学报·信息科学版，2010，35(5)：578-581

[4] Lemos A R. Bots Surge Ahead in March[OL]. http://www.securityfocus.com/brief/466, 2007

[5] Wesson R. Botnets and the Global Infection Rate: Anticipating Security Failures[OL]. http://www.

stanford. edu/class/ee380/Abstracts/070606-slides. pdf, 2007

[6] Cullen D. Dutch Smash 100 000-Strong Zombie Army[OL]. http://www. theregister. co. uk/ 2005/ 10/07/dutch_ police _ smash _ zombie _ network/, 2005

[7] Evers J. Bot Herders May have Controlled 1. 5 Million PCs, ZDNet News[OL]. http://news. zdnet. com/ 2100-1009_22-5906896. html, 2005

[8] Vogt R, Aycock J. Attack of the 50 Foot Botnet [R]. Technical Report, Department of Computer Science, University of Calgary, 2006

[9] Vogt R, Aycock J, Jacobson M J. Army of Botnets [C]. The 2007 Network and Distributed System Security Symposium (NDSS 2007), San Diego, California, 2007

[10] 臧天宁,云晓春,张永铮,等. 利用 C-F 模型识别僵尸网络迁移[J]. 武汉大学学报·信息科学版, 2010,35(5):622-625

[11] Rajab M A, Zarfoss J, Monrose F. My Botnet is Bigger than Yours (Maybe, Better than Yours): Why Size Estimates Remain Challenging[C]. The 1st Workshop on Hot Topics in Understanding Botnets, Berkeley, CA, USA, 2007

[12] Gu G, Zhang J, Lee W. Bot Sniffer: Detecting Botnet Command and Control Channel Sin Network Traffic[C]. The 15th Annual Network and Distributed System Security Symposium, San Diego, CA, 2008

[13] 李德毅,杜鹢. 不确定性人工智能[M]. 北京:国防工业出版社,2005

[14] 杨朝晖,李德毅. 二维云模型及其在预测中的应用[J]. 计算机学报,1998,21(11):961- 969

[15] Hi-performance Protocol Identification Engine [OL]. http:// hippie. oofle. com/, 2007

**第一作者简介**:臧天宁,博士,主要研究方向为僵尸网络、协同分析。

E-mail: zhangyongzheng@ict. ac. cn

# A Botnet Relationship Analyzer Based on Cloud Model

*ZANG Tianning*[1,2,3]　*YUN Xiaochun*[1,2,3]　*ZHANG Yongzheng*[1,3]　*MEN Chaoguang*[2]

(1 Institute of Information Engineering, Chinese Academy of Sciences, 27 Minzhuang Road, Haidian District, Beijing 100097, China)

(2 College of Computer Science and Technology, Harbin Engineering University, 145 Nantong Street, Harbin 150001, China)

(3 National Engineering Laboratory for Information Security Technologie, 27 Minzhuang Road, Haidian District, Beijing 100097, China)

**Abstract**: An approach for analyzing the relationship among botnets was presented. Several botnet communication characteristics were extracted, including the amount of data flows within a botnet, the number of packets per data flow, the payload of communication and data packets in the master hosts. Statistical similarity functions of botnet characteristics were defined. Based on the cloud model and the defined statistical similarity functions, the analysis model of botnet relationship was build, and the similarities of botnet characteristics were synthetically evaluated. The analysis experiments were conducted based on a simulation network environment. The experimental results show that the presented method was valid and efficient, even in the case of encrypted botnet communication messages. The result is better than the research production in the report on the interrelated research achievements.

**Key words**: botnet; cloud model; migration; similarity

**About the first author**: ZANG Tianning, Ph. D, majors in botnet and coordinative analysis.

E-mail: zhangyongzheng@ict. ac. cn