基于通信特征提取和 IP 聚集的僵尸网络相似性度量模型

李润恒[1)] 王明华[2)] 贾 焰[1)]

[1)](国防科学技术大学计算机学院 长沙 410073)

[2)](国家计算机网络应急技术处理协调中心 北京 100029)

摘 要 IRC 僵尸网络(botnet)是攻击者通过 IRC 服务器构建命令与控制信道方式控制大量主机(bot)组成的网络. IRC 僵尸网络中 IRC 服务器与 bot 连接具有很强的动态特性,为识别使用不同 IRC 服务器的同一僵尸网络,文中提取并比对僵尸网络的通信量特征、通信频率特征,建模估算 bot 重叠率,通过融合以上度量指标,提出了僵尸网络相似性度量模型. 实验验证了模型的有效性,计算了其准确率,并分析了僵尸网络的迁移.

关键词 僵尸网络;通信;聚集;相似性度量;迁移

中图法分类号 TP393 **DOI 号**: 10.3724/SP.J.1016.2010.00045

Modeling Botnets' Similarity Based on Communication Feature Extraction and IP Assembly

LI Run-Heng[1)] WANG Ming-Hua[2)] JIA Yan[1)]

[1)](*School of Computer, National University of Defense Technology, Hunan* 410073)

[2)](*National Computer Network Emergency Response Technical Team/Coordination Center of China, Beijing* 100029)

Abstract IRC botnet can be regarded as a collection of compromised computers (called Zombie computers) running software under the command-and-control infrastructure constructed by the IRC servers. The connection between the botnet server and the bots are usually very dynamic. In order to describe a botnet at a finer granularity, the paper proposes a method that measures the similarity of botnets by extracting and comparing the metrics such as communication volumes, frequency, and the overlap rate of bots. A novel model for botnet similarity measuring is proposed by combining those metrics mentioned. Experiments are carried out for validation purposes, the confidence of the accuracy is evaluated and shown, and the migration situation of botnet are also discussed.

Keywords botnet; communication; assemble; similarity measure; migration

1 引 言

僵尸网络是攻击者出于恶意目的,传播僵尸程序控制大量主机,并通过一对多的命令与控制信道(Command and Control, C&C)所组成的网络. 僵尸网络为攻击者提供了隐匿、灵活且高效的一对多命令与控制机制,可以控制大量僵尸主机实现信息

收稿日期:2009-07-15;最终修改稿收到日期:2009-09-07. 本课题得到国家"八六三"高技术研究发展计划项目基金(2007AA010502, 2007AA01Z474, 2006AA01Z451)资助.**李润恒**,男,1982 年生,博士研究生,研究方向为僵尸网络、数据挖掘. E-mail: lirunheng1982@gmail.com. **王明华**,男,1978 年生,博士,工程师,研究方向为互联网安全监测、应急响应处理. **贾 焰**,女,1960 年生,教授,博士生导师,研究领域为网络安全、数据库.

窃取、分布式拒绝服务攻击和垃圾邮件发送等攻击目的.僵尸网络正步入快速发展期,对因特网安全已造成严重威胁.

僵尸网络主要分为 IRC 僵尸网络、HTTP 僵尸网络和 P2P 僵尸网络.IRC 僵尸网络是最早产生而目前仍然大量存在的一类僵尸网络,基于标准 IRC 协议在 IRC 聊天服务器上构建其命令与控制信道,控制者通过命令与控制信道实现对大量受控主机的僵尸程序版本更新、恶意攻击等行为的控制,其控制者、命令与控制服务器(IRC 服务器)、受控主机(bot)、被攻击对象的关系如图1所示;HTTP僵尸网络与

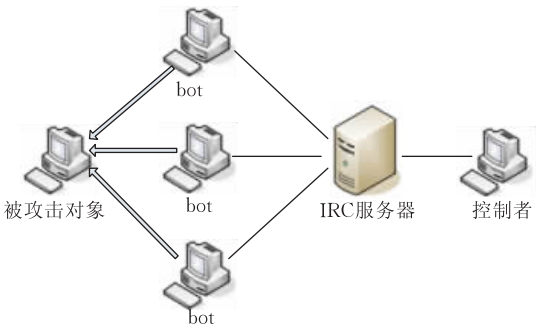

图 1 IRC 僵尸网络关系示意图

IRC 僵尸网络的功能结构相似,所不同的是 HTTP 僵尸网络控制器是以 WEB 网站方式构建;P2P 僵尸网络是一种较新型的僵尸网络,在 P2P 僵尸网络中僵尸程序同时承担客户端和服务器的双重角色.

图 1 所示的 IRC 僵尸网络健壮性差,存在单点失效问题,可通过摧毁单个 IRC 服务器来切断僵尸网络控制者与 bot 的联系,导致整个僵尸网络瘫痪.针对这一问题,bot 的僵尸程序使用域名而非固定的 IP 地址连接 IRC 服务器,僵尸网络控制者使用动态域名服务将僵尸程序连接的域名映射到其控制的多台 IRC 服务器上,一旦正在工作的 IRC 服务器失效,僵尸网络的受控主机会连接到其他的 IRC 服务器,整个僵尸网络继续运转,如图 2(a)所示.此外,将僵尸网络的控制权出租出售谋取经济利益是目前僵尸网络产业链的重要组成部分.僵尸网络主动或者被动改变其 IRC 服务器的行为称为僵尸网络的迁移.此外,出于安全的考虑,某些大型僵尸网络采用分层管理模式,如图 2(b)所示,由多个 IRC 服务器控制各自不同的 bot 群体,而所有的 IRC 服务器同时由僵尸网络控制者统一控制.

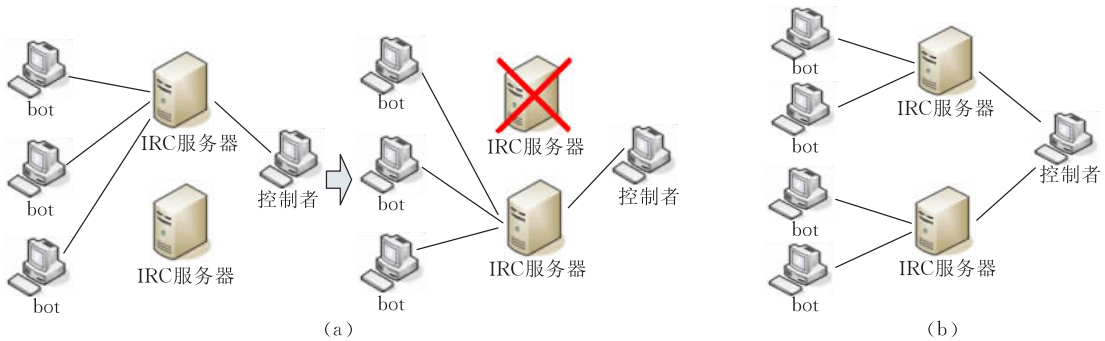

图 2 IRC 僵尸网络衍变示意图

因此 IRC 服务器与僵尸网络(控制者)并不一定是一一对应关系,并且 IRC 服务器与僵尸网络(控制者)的对应关系可能随时间发生转变.利用 IRC 服务器与 bot 的一对多映射关系,使用聚类等数据分析方法可以有效地检测 IRC 服务器与 bot 的 C&C 通信,以此获得 IRC 服务器与 bot 的对应关系.但是僵尸网络控制者与 IRC 服务器是一对一映射关系,很难使用数据分析方法获得僵尸网络控制者与 IRC 服务器的对应关系.

IRC 僵尸网络中,bot 与控制者是实体,IRC 服务器只是其中间桥梁.要准确地掌握僵尸网络,必须掌握僵尸网络(控制者)与 bot 的对应关系.由于僵尸网络 IRC 服务器与 bot 连接的复杂衍变特性(如图 2 所示)以及 IRC 服务器与控制者通信检测的困难,目前相关研究主要集中在 IRC 服务器与 bot 的 C&C 通信检测,存在局限.此外,由于很难获取大规模僵尸网络通信数据,实验数据由少量已知僵尸网络通信数据集仿真产生,无法对大量僵尸网络通信数据进行特征比对等关联分析.本文首先将 IRC 服务器与所关联的 bot 看作一个僵尸网络,在此基础上建立僵尸网络相似性度量模型,根据僵尸网络相似性距离值,分类识别相同的僵尸网络,以此准确地掌握僵尸网络.准确地掌握僵尸网络有利于度量僵尸网络的大小,评估僵尸网络的危害;研究僵尸网络的生命周期,掌握其衍变特性等.

本文基于国家网络安全监测平台监测到的僵尸网络 IRC 服务器与 bot 的 C&C 通信数据,从不同角度对僵尸网络的相似性进行度量:提取并比对僵

尸网络的通信特征；bot 重叠率的建模估算. 由于僵尸网络间 bot 群体的差异、僵尸程序版本的差异等因素，通信特征是僵尸网络区别其它僵尸网络的显著特征，包括通信量特征和通信频率特征. 由于大多数 bot 夜间关机下线，僵尸网络通信量有明显的以一天为周期的周期规律，提取僵尸网络通信量日周期曲线和通信频率日周期曲线. 通信量日周期曲线反映了 bot 群体的普遍上线时间习惯，而通信频率日周期曲线反映了僵尸网络控制者的使用习惯以及僵尸程序版本等特征. 度量僵尸网络相似性的另一个方法是建模估算 bot 的重叠率. 考虑到互联网上众多 ADSL 上网的主机使用动态 IP 地址，直接计算 bot IP 的重叠率会导致很大的误差. 本文通过 bot IP 地址的聚集操作，将 bot IP 地址集合，映射为 bot 集合，估算僵尸网络间 bot 的重叠率，以此来度量僵尸网络的相似性. 两类方法各有优缺点，适合不同的情况，融合其相似性度量的结果，本文提出僵尸网络相似性度量模型. 通过蜜网蜜罐跟踪、域名监测系统日志分析等手段确认相同僵尸网络，对模型进行有效性验证，计算其准确率，并分析导致错误的各类原因. 最后分析僵尸网络的迁移.

本文第 2 节介绍相关研究；第 3 节介绍基于通信特征提取和 IP 聚集的相似性度量模型，3.1 节介绍国家网络安全监测平台，3.2 节介绍通信量特征提取，3.3 节介绍通信频率特征提取，3.4 节介绍 IP 聚集，3.5 节介绍相似性度量模型；第 4 节为实验和验证；第 5 节为结语及未来工作的展望.

2　相关研究

僵尸网络是在网络蠕虫、特洛伊木马、后门工具等传统恶意代码形态的基础上发展、融合而产生的一种新型攻击方式. 采用灵活且高效的一对多控制机制，利用僵尸网络，攻击者可以轻易地控制成千上万台主机对因特网任意站点发起分布式拒绝服务攻击，并发送大量垃圾邮件. 因此，僵尸网络得到了攻击者的关注并进一步发展成为因特网最为严重的威胁之一. 近年来，僵尸网络的活跃已经引起国内外安全业界的充分重视，僵尸网络已成为安全领域的学术研究和讨论的热点问题.

目前主流的僵尸网络是 IRC 僵尸网络，基于标准 IRC 协议构建其命令与控制信道，其控制服务器可构建在公用 IRC 聊天服务器上，但攻击者为保证对僵尸网络控制服务器的绝对控制权，一般会利用其完全控制的主机架设专门的僵尸网络命令与控制服务器. IRC 僵尸网络的工作机制：攻击者通过各种传播方式使得目标主机感染僵尸程序；僵尸程序加入到攻击者私有的 IRC 命令与控制信道中；攻击者登陆并加入到 IRC 命令与控制信道中，通过认证后向僵尸网络发出各种指令；僵尸程序接受指令，执行指令，必要的情况下返回执行指令的结果.

IRC 僵尸网络的跟踪与检测方法可以分为 3 大类：蜜网蜜罐跟踪僵尸网络[1-4]、协议与结构相关检测方法[5-9]、协议与结构无关检测方法[10-11]. 蜜网蜜罐通过捕获并分析恶意代码获取僵尸网络命令与控制信道的相关信息，然后模拟受控的僵尸主机加入僵尸网络，对僵尸网络的内部活动进行观察和跟踪，但是这类方法依赖于蜜网蜜罐布控点的分布，无法有效地检测出全部活跃的僵尸网络. 协议有关的检测方法利用跟踪方法了解僵尸网络内部工作机制，抽象出僵尸网络行为特征，通过异常检测等方法检测僵尸网络. 协议无关的检测方法采用聚类算法将网络流量分类，从而识别僵尸网络流量和正常流量.

关于僵尸网络的动态性、相似性度量方面的研究，文献[12]从评估僵尸网络规模的角度提出了僵尸网络相似性度量问题，文章指出评估僵尸网络规模的难点之一是僵尸网络的动态性，通过蜜网蜜罐跟踪僵尸网络获取其僵尸程序版本、IRC 服务器 IP、IRC 服务器域名、IRC 频道名、控制者 ID 等信息，提出了僵尸网络相似性度量模型，最后分析了僵尸网络的迁移情况；文献[2]对僵尸网络的迁移及复制现象进行了分析，但是只针对僵尸网络在同一个 IRC 服务器上不同频道的迁移与复制；文献[13]从研究僵尸网络传播模型的角度，考虑到大多数计算机在夜间关机下线，从而僵尸网络的通信量呈现周期现象，提取了僵尸网络在全球不同时区的通信量日周期曲线.

3　基于通信特征和 IP 聚集的相似性度量模型

3.1　国家网络安全监测平台

863-917 网络安全监测平台[14]是国家“八六三”计划设立的网络安全应急项目(917 工程)建设的网络安全监控平台. 该平台是保障国家网络安全和网上重要信息系统安全的重要监测平台，由 CNCERT/CC 负责建设并运行.

863-917 网络安全监测平台底层为网络型 IDS

系统,实时监测我国互联网中特定安全事件,诸如僵尸网络、木马通信事件等.采用协议与结构相关的僵尸网络检测方法,利用蜜网蜜罐获取僵尸网络信息、提取僵尸网络报文级通信特征,在国家重要路由器节点部署网络型 IDS,对路由报文使用特征匹配检测僵尸网络 C&C 通信.检测到的僵尸网络 C&C 通信包括 IRC 服务器与 bot 间的控制命令、定期存活检测通信等.863-917 平台记录了僵尸网络通信事件的 bot IP 地址、IRC 服务器 IP 地址、通信时间等属性.根据 863-917 平台的检测结果,能够获取 IRC 服务器与 bot 的映射关系.

3.2 通信量日周期曲线

由于僵尸网络间 bot 群体的差异、僵尸程序版本的差异等因素,僵尸网络的通信特征是僵尸网络区别其他僵尸网络的显著特征.通信特征包括通信量特征和通信频率特征,通信量特征反映了 bot 群体的普遍上线时间习惯,由于僵尸网络可能是针对特定的漏洞(比如 Windows 2000 SP2 漏洞)而发展形成的,其 bot 群体的上线时间习惯具有一定的相似性[13],而通信频率特征反映了僵尸网络控制者的使用习惯以及僵尸程序版本等特征.首先给出两个定义:

通信量(Communicate Count) $CC_i(t)$,僵尸网络 i 的通信量随时间变化的函数,它是一个统计值函数,需要给定统计时间间隔大小 w.其中 i 为僵尸网络标号,在不引起歧义的情况下,本文省掉 i.

在线(online)bot 数量 $Obot_i(t)$,僵尸网络 i 在线 bot 数量的统计函数.

大多数计算机在夜间关机下线,僵尸网络的通信量在夜间有明显的下降,具有明显的周期性,如图 3 所示.实验显示僵尸网络的通信量每一天的变化曲线相似,统计 n 天的数据计算僵尸网络的通信量日周期函数 $C(t)(0\leq t\leq 24h)$ 如下:

(1)计算每一天的通信量 $CC(t)$;

(2)对每一天的数据进行归一化;

(3)平均 n 天的数据;

(4)对(3)的结果进行归一化,得到 $C(t)$;

为了度量僵尸网络间通信量特征的相似性,计算其通信量日周期曲线的距离,曲线的距离有欧氏距离、DTW、LB_Keogh、LB_PAA 距离等[15],此外通信量日周期曲线还有其显著的特点:由于 bot 群体的相似性,上线时段集中,有上线高峰和低谷,通信量日周期曲线有明显的曲线峰、谷.本文采用欧氏距离计算两曲线的距离.

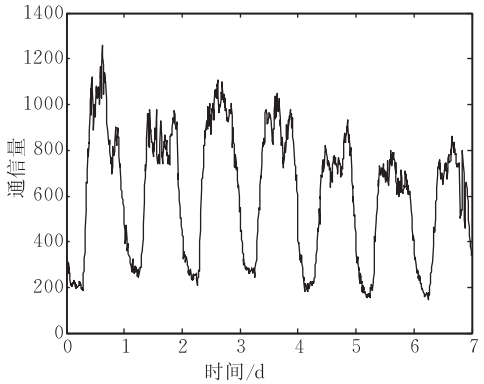

图 3 僵尸网络通信量变化曲线图

3.3 通信频率日周期曲线

通信量一定程度上反映了僵尸网络在线 bot 数量,而通信频率即单位 bot 主机的通信量,反映的是僵尸网络 IRC 服务器与 bot 间通信的频繁程度.实验显示,僵尸网络的通信频率在一天内的平均值趋于常量,如图 4 所示.

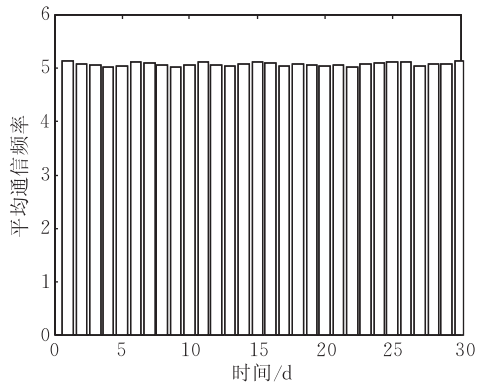

图 4 僵尸网络日平均通信频率变化图

但是僵尸网络的通信频率在一天内并不恒定,而是同样呈现明显的周期性.统计 n 天的数据,计算僵尸网络通信频率日周期函数 $CF'(t)(0\leq t\leq 24\text{h})$ 如下:

(1)把每天的通信数据分成 $24\text{h}/w$ 份(w 为统计时间间隔大小,它的含义是:认为在 w 间隔内有通信的 IP 数为该时间跨度内在线肉机数 $Obot(t)$,根据僵尸网络 IRC 服务器与 bot 通信数据的特点,本文 w 取 10min),每一份时间跨度为 w,计算每一份数据中不同 IP 个数,得到在线肉机函数 $Obot(t)$ 的统计值;

(2)计算通信量 $CC(t)$,通信频率函数 $CF(t)=CC(t)/Obot(t)$,即单位 bot 的通信量.若 $Obot(t)=0$,使用线性插值的方法计算 $CF(t)$.

(3)平均 n 天的数据,得到 $CF'(t)(0\leq t\leq 24\text{h})$;

(4)为了去掉噪声的影响,使用多项式拟合

$CF'(t)$，得到 $CFS(t)$.

由于互联网 IP 地址紧缺，一些局域网内部网络采用 NAT(Network Address Translation)技术，使多台计算机使用一个 IP 共享 Internet 连接，在局域网内部网络中使用内部地址，而当内部节点要与外部网络进行通信时，就在网关将内部地址替换成公用地址. bot 中这类 IP 的通信频率明显大于所属僵尸网络的通信频率，如图 5 所示，图中两曲线分别为僵尸网络通信频率和该僵尸网络某 bot IP 的通信频率. 因此计算僵尸网络通信频率时，应该剔除掉这些 IP. 由于共用 bot IP 为静态 IP，通信时间跨度较长，在计算僵尸网络通信频率时，剔除通信时间跨度超过阈值 m 的 bot IP，本文实验 m 取 10d.

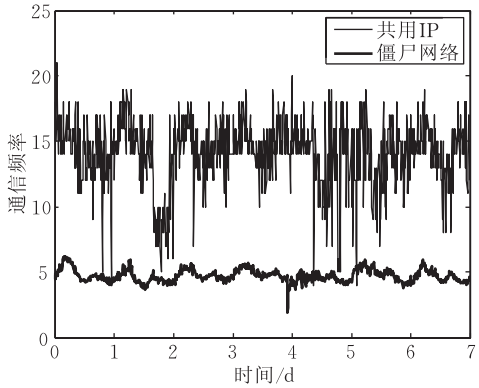

图 5　共用 IP 与其所属僵尸网络通信频率曲线对比图

3.4　IP 聚集

度量僵尸网络相似性的另一个方法是计算 bot 的重叠率. 考虑到互联网上众多 ADSL 上网的主机没有一个固定的 IP，当主机联网，互联网服务提供商(ISP)从一个 IP 库中对其随意分配一个未经使用的 IP 地址. 这一 IP 地址只会在该主机上网的时间段中保留，下一次上线可能分配不同的 IP 地址. 因此 bot 的 IP 地址存在大量动态 IP，直接计算 bot IP 的重叠率会导致很大的误差. 本小节通过 bot IP 地址的聚集操作，将 bot 的 IP 地址集合，映射为 bot 集合，再计算僵尸网络间 bot 的重叠率，以此来度量僵尸网络的相似性.

bot IP 聚集理想的结果是每一个 bot 使用过的 IP 聚集到同一个集合，不同 bot 对应聚集后的集合不同，即聚集后的集合与 bot 集合一一对应. 对于给定的僵尸网络，设其 bot 集合为 B，$B=\{b_1,\cdots,b_n\}$，bot 数量为 n，即 $|B|=n$. 这些 bot 使用过的 IP 地址集合为 I，$|I|=m$，$m\geq n$，$f(B)=I$，f 为 B 到 I 的1 对多映射.

IP 地址是 4 个小数点隔开的十进制整数，考虑到 ISP 给 bot 主机动态分配的 IP 地址集合具有局部性，对 bot IP 地址进行聚集操作，去掉 IP 地址的小数点间隔的第 4 部分，这样的操作记作映射 g.

容易证明以下定理.

定理 1.　若 $\forall IP_i, IP_j\in f(b_k)(k=1,2,\cdots,n)$，$g(IP_i)=g(IP_j)$，则 $|g(I)|\leq|B|$；

若 $\forall IP_i\in f(b_k)$，$\forall IP_j\in f(b_l)(k,l=1,2,\cdots,n,k\neq l)$，$g(IP_i)\neq g(IP_j)$，则 $|g(I)|\geq|B|$.

由定理 1 得到定理 2.

定理 2.　若 $\forall IP_i, IP_j\in f(b_k)(k=1,2,\cdots,n)$，$g(IP_i)=g(IP_j)$，$\forall IP_i\in f(b_k)$，$\forall IP_j\in f(b_l)(k,l=1,2,\cdots,n,k\neq l)$，$g(IP_i)\neq g(IP_j)$，则 $|g(I)|=|B|$.

根据定理 2 的假设，对僵尸网络的足迹(footprint)(给定监测时间内所监测到的 bot IP)即集合 I 进行聚集操作，得到 $g(I)$，它与 bot 集合一一对应，计算僵尸网络间 bot 的重叠率以此来度量僵尸网络的相似性.

3.5　相似性度量模型

僵尸网络的通信量日周期曲线距离、通信频率日周期曲线距离、bot 重叠率均可以度量僵尸网络的相似性. 但是这几种方法各有优缺点，适合不同的情况，根据单独的一个特征不能准确地判断僵尸网络的相似性. 譬如僵尸网络间没有 bot 的重叠，也可能是同一个僵尸网络，它们是同一个僵尸网络的不同 bot 群体，如图 2(b)所示. 本小节融合以上方法的度量结果，建立僵尸网络相似性度量模型，第 4 节将验证模型的有效性.

僵尸网络相似性度量指标：bot 重叠率、通信量日周期曲线距离、通信频率日周期曲线距离.

相似性度量函数应该满足下面的性质：

单调性. 函数值随某个指标的值的增加而增加或者随某个指标的值的增加而减小.

敏感性. 函数值随各指标值变化的变化速度不同，对更重要指标，函数值对其变化更敏感.

鲁棒性. 若某个指标误差较大，函数值能够一定程度地屏蔽其对结果的影响.

bot 重叠率、通信量日周期曲线距离、通信频率日周期曲线距离的值进行归一化处理后分别计为 S_1，S_2，S_3，其权值系数记为 w_1，w_2，w_3.

相似性度量函数：

$$S=w_1(1-S_1)+w_2S_2+w_3S_3.$$

僵尸网络对的相似性度量函数值越小，表示僵尸网络对的相似性越大. 容易验证，函数满足单调性、敏感性、鲁棒性. 权值系数的确定，最优分类判别值的计算以及模型有效性验证见第 4 节.

4 实验及验证

对 863-917 网络安全监测平台 60d 内监测到的 723 个僵尸网络，分别计算通信量日周期曲线、通信频率日周期曲线，进行 IP 聚集操作. 4.1 节列出了部分结果；最优分类判别值的计算以及相似性度量模型有效性的验证见 4.2 节.

4.1 实 验

图 6、图 7 为其中两对僵尸网络通信量日周期曲线对比图. 图 6 两条曲线的欧氏距离为 0.0672. 图 7 两条曲线的欧氏距离 0.0745. 两对僵尸网络均为国外 IRC 服务器控制国内受控主机. 僵尸网络 1-僵尸网络 4 控制服务器分别在美国、美国、加拿大、加拿大.

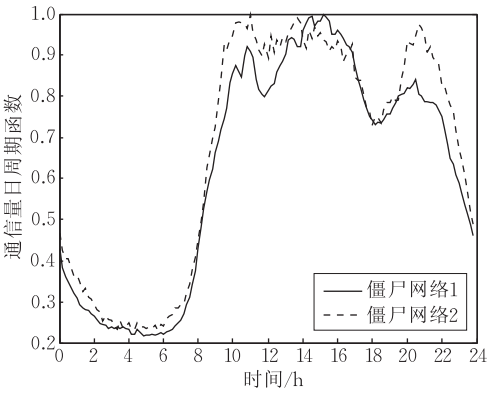

图 6 僵尸网络对通信量日周期曲线对比图 1

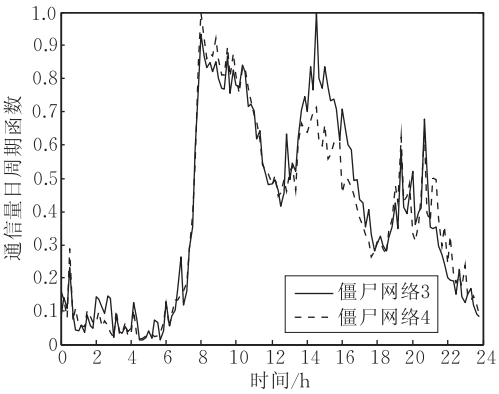

图 7 僵尸网络对通信量日周期曲线对比图 2

图 8～图 11 为僵尸网络 1～4 的通信频率日周期曲线图.

对僵尸网络 1～4 进行 IP 聚集，结果如表 1、表 2 所示.

表 1 僵尸网络 IP 聚集结果 1

	僵尸网络 1	僵尸网络 2	重叠	重叠率/%
聚集前 IP 数	42778	5124	127	2
聚集后 IP 数	8532	1105	503	46

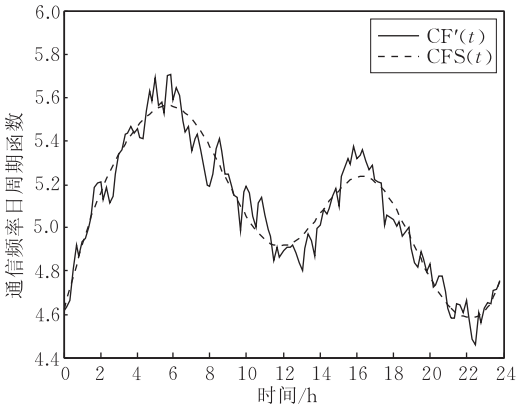

图 8 僵尸网络 1 通信频率日周期曲线图

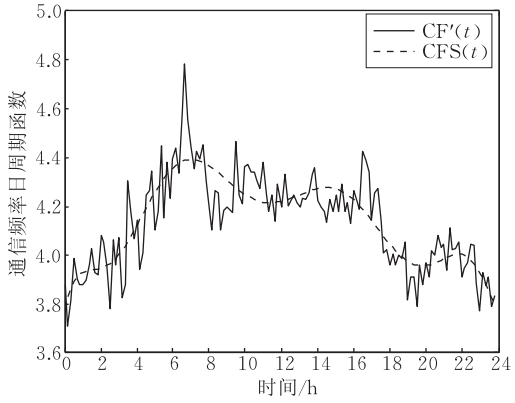

图 9 僵尸网络 2 通信频率日周期曲线图

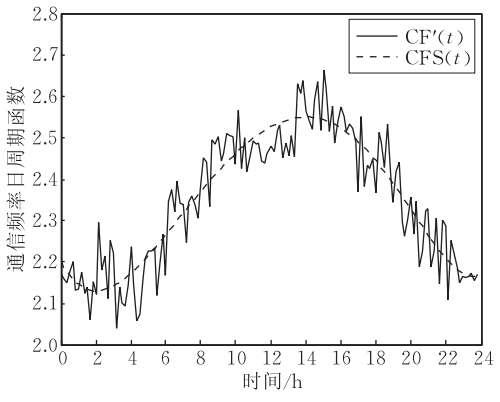

图 10 僵尸网络 3 通信频率日周期曲线图

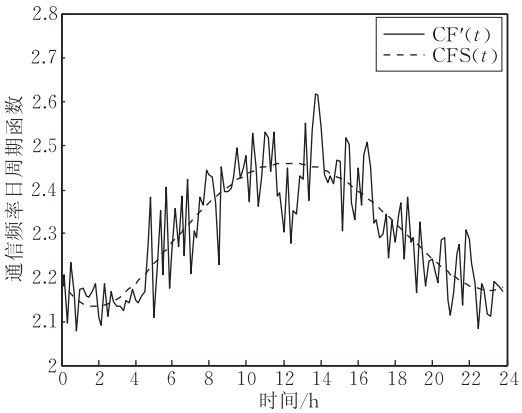

图 11 僵尸网络 4 通信频率日周期曲线图

表 2　僵尸网络 IP 聚集结果 2

	僵尸网络 3	僵尸网络 4	重叠	重叠率/%
聚集前 IP 数	11303	10634	4805	45
聚集后 IP 数	4009	3903	3360	86

这里的 IP 数指僵尸网络 bot 的足迹(footprint)大小，即在给定监测时间内所监测到的 bot IP 总数. 两对僵尸网络的 IP 重叠率聚集后有明显增加.

4.2　有效性验证

对于监测到的 723 个僵尸网络，我们采用蜜网蜜罐跟踪、域名监测系统日志分析等手段确认相同僵尸网络 150 对和不同僵尸网络 150 对. 其中 100 对相同僵尸网络和 100 对不同僵尸网络作为模型中分类方法的训练集，其余作为测试集.

使用僵尸网络通信特征曲线距离、bot 重叠率以及综合各指标的相似性度量模型均可以度量僵尸网络的相似性，本小节通过训练集确定各种方法的最优分类判别值，计算其分类错误率，并对产生错误的各类情况分析了可能的原因. 使用测试集验证了相似性度量模型分类识别相同僵尸网络的有效性.

图 12 为僵尸网络对通信量日周期曲线距离值分布. 横坐标为僵尸网络对标号，标号属于[1,100]和[101,200]分别为相同僵尸网络对和不同僵尸网络对. 记相同僵尸网络对通信量日周期曲线距离为 $D_{\text{true}}(i)$，$1\leq i\leq 100$，不同僵尸网络对通信量日周期曲线距离为 $D_{\text{false}}(i)$，$101\leq i\leq 200$. 对两集合 $\{D_{\text{true}}(i)\mid 1\leq i\leq 100\}$、$\{D_{\text{false}}(i)\mid 101\leq i\leq 200\}$ 分别采用 Shapiro-Wilk 算法检验数据的正态性，取 $\alpha=0.05$，得到两数据集服从正态分布，统计计算其均值 μ 和方差 δ^2.

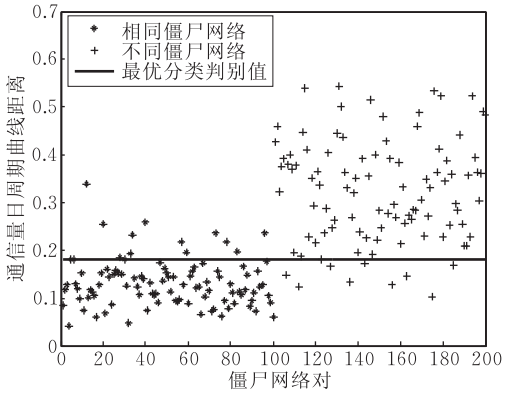

图 12　僵尸网络对通信量日周期曲线距离

根据通信量日周期曲线距离对僵尸网络对进行分类，最优分类判别值 η(分类规则为距离小于或等于 η 认为是同一僵尸网络，距离大于 η 认为是不同僵尸网络)的理论值为

$$\eta=\min_{\xi}(P(F_{\text{true}}(x)>\xi)+P(F_{\text{false}}(x)\leq\xi)).$$

错误率为

$$(P(F_{\text{true}}(x)>\eta)+P(F_{\text{false}}(x)\leq\eta))/2,$$

其中 $P(F_{\text{true}}(x)>\eta)$ 为弃真错误率，弃真错误指相同僵尸网络而作出不同僵尸网络的判断. $P(F_{\text{false}}(x)\leq\eta)$ 为取伪错误率，取伪错误指不同僵尸网络而作出相同僵尸网络的判断. 采用迭代的方法计算最优分类判别值 η 和 3 类错误率如表 3.

图 13 为进行归一化处理的僵尸网络对通信频率日周期曲线距离值分布，同样的方法计算最优分类判别值 η 和 3 类错误率如表 3. 错误率较高的原因是对通信频率日周期曲线只是进行简单的归一化处理，没有考虑曲线形状、均值、方差等与通信频率特征的相关性.

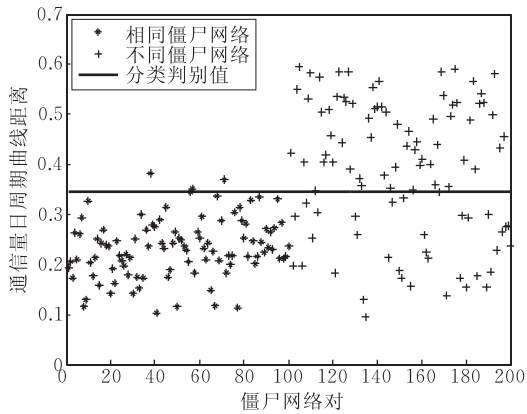

图 13　僵尸网络对通信频率日周期曲线距离

图 14 为僵尸网络对 IP 聚集后 bot 重叠率计算值分布，不考虑样本点的分布，计算最优分类判别值 η 和 3 类错误率如表 3. 弃真错误率为 0，这是因为不同的僵尸网络 bot 重叠率很低，但是取伪错误率较高，因为相同僵尸网络有可能是图 2(b)所示分层管理的情况，其 bot 重叠率低，从图 14 也可以看出，

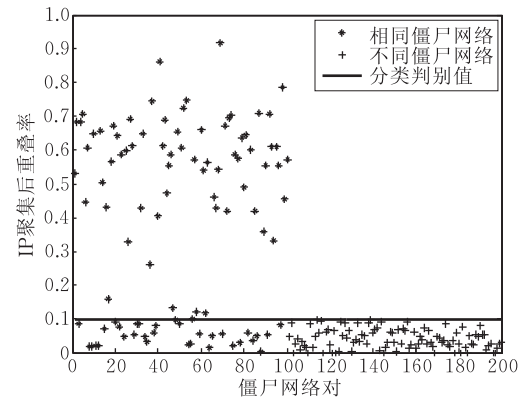

图 14　僵尸网络对 bot 重叠率

相同僵尸网络的 bot 重叠率分布点聚集在两个区域;此外,IP 聚集时定理 2 的假设可能并不严格成立,导致了结果的偏差.

使用相似性度量模型计算僵尸网络对相似性值,相似性度量函数各指标的权值取对应方法的准确率,如表 3 所示,图 15 为僵尸网络对相似性距离值分布,同样不考虑样本点的分布,计算最优分类判别值 η 和 3 类错误率(如表 3 所示).

表 3 僵尸网络相似性度量各方法最优分类判别值及错误率

	最优分类判别值	准确率	错误率	弃真错误率	取伪错误率
通信量特征比对方法	0.1803	0.895	0.105	0.12	0.09
通信频率特征比对方法	0.3453	0.81	0.19	0.03	0.35
计算 bot 重叠率方法	0.0975	0.84	0.16	0	0.32
相似性度量模型(训练集)	1.1339	0.94	0.06	0.07	0.05
相似性度量模型(测试集)	1.1339	0.89	0.11	0.06	0.05

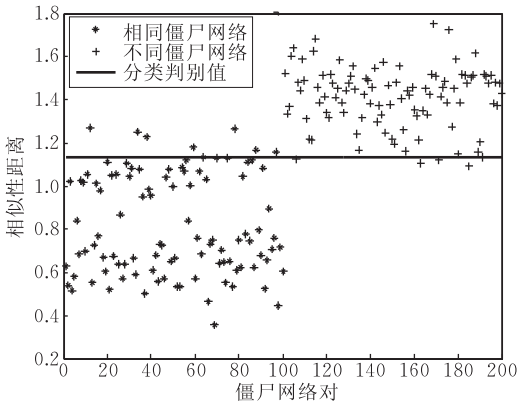

图 15 僵尸网络对(训练集)相似性距离

使用训练集计算所得的最优分类判别值 η,通过测试集检测模型识别相同与不同僵尸网络的准确率结果如图 16、表 3. 从结果可以看出,融合了通信特征提取和 IP 聚集的相似性度量模型识别相同与不同僵尸网络的准确率较高.

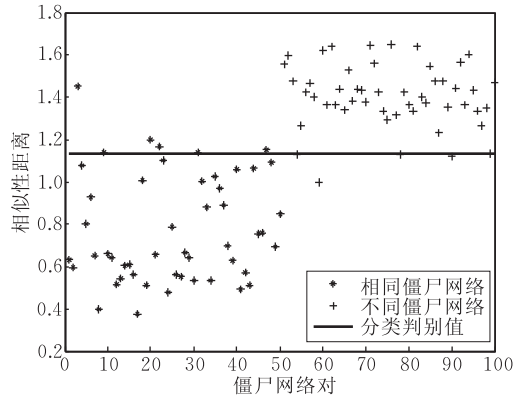

图 16 僵尸网络对(测试集)相似性距离

4.3 迁移分析

利用僵尸网络相似性度量模型计算的结果,对于相同僵尸网络对,本小节定性地分析僵尸网络的迁移,典型的几类情况如下.

图 17 为某对僵尸网络的通信量曲线对比图,横坐标单位为天. 从图 17 可以看出,在某时间段,僵尸网络 1 没有通信,而僵尸网络 2 在该时间段的通信量普遍超过其它时间段的通信量. 图 18 为另一对僵尸网络的通信量曲线对比图,在其中一时间段,僵尸网络 4 没有通信,而僵尸网络 3 在该时间段的通信量超过其"正常"通信量.

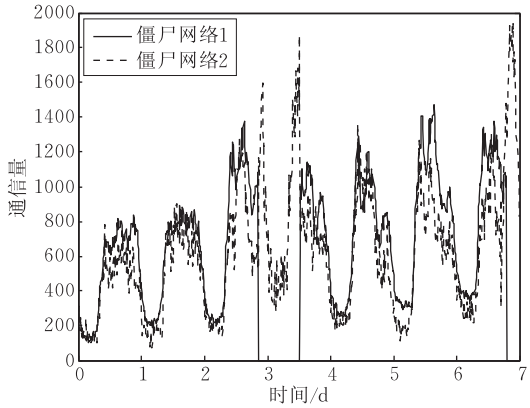

图 17 僵尸网络对通信量曲线对比图 1

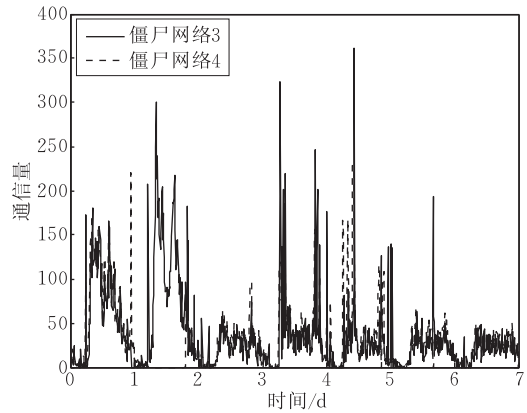

图 18 僵尸网络对通信量曲线对比图 2

图 19 为第 3 对僵尸网络的通信量曲线对比图,从图 19 可以看出,僵尸网络 5 在某时间段通信量为 0,而僵尸网络 6 只在这个时间段有通信,把两僵尸网络通信量叠加,叠加之后的通信量按天递减.

考虑展示时间粒度更细的情况,图 20 为第 2 对

僵尸网络的通信量曲线对比图，横坐标单位为小时. 从图 20 可以看出，在某时间段，僵尸网络 4 没有通信，而僵尸网络 3 在该时间段的通信量超过其他时间段的通信量. 并且在这个时间段的开端，僵尸网络 3 的通信量大小发生了跳跃. 这是因为僵尸网络在迁移初始阶段 bot 与新的 IRC 服务器有额外的认证等通信.

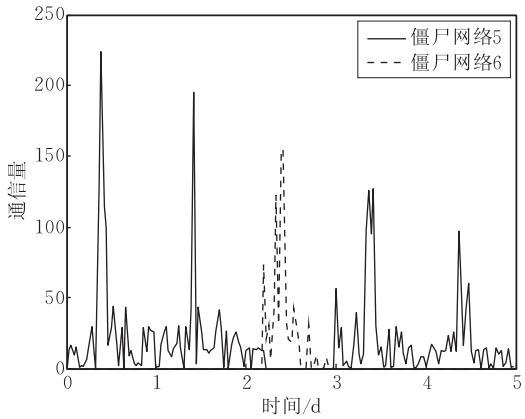

图 19　僵尸网络对通信量曲线对比图 3

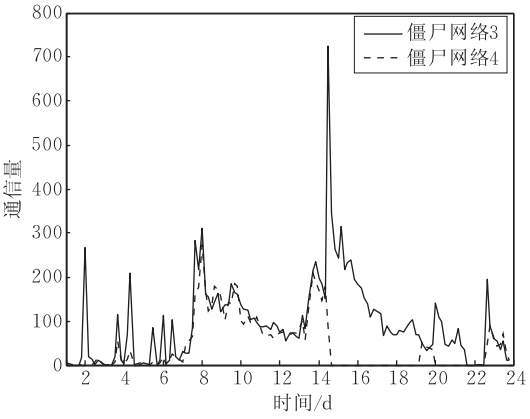

图 20　僵尸网络对通信量曲线对比图 4

从以上几个例子可以看出，僵尸网络的迁移在通信量方面有如下特点：

僵尸网络发生迁移所导致的通信对象的改变会反映到僵尸网络对通信量的变化，即“此消彼长”. 若把两个僵尸网络通信量曲线叠加，叠加后的曲线符合自然的变化规律；迁移初始阶段由于 bot 与新的 IRC 服务器有额外的认证等通信，迁移目的僵尸网络通信量大小在短时间内有较明显的跳跃.

当然，根据僵尸网络通信量的变化并不一定能检测出僵尸网络的迁移；僵尸网络发生迁移也不是相同僵尸网络的必要条件，图 2(b)所示的僵尸子网络没有迁移，但它们是同一个僵尸网络的不同 bot 群体.

5　结束语

本文通过提取并比对僵尸网络通信特征，IP 聚集估算 bot 重叠率，建立僵尸网络相似性度量模型，分类识别使用不同控制服务器的相同僵尸网络. 下一步的工作主要有：

(1) 本文计算僵尸网络通信特征曲线距离采用欧氏距离，考虑到曲线自身的特点：由于 bot 群体的相似性，上线时段集中，有上线高峰和低谷，通信量日周期曲线有明显的曲线峰、谷等. 应该考虑更合理有效的曲线距离计算方法，考虑曲线形状、均值、方差与通信特征的相关性.

(2) 相似性度量模型中的权值系数是训练集所得错误率计算所得，应该再考虑弃真错误率和取伪错误率以及它们的方差等.

(3) 形式化描述并提取僵尸网络迁移时通信量变化的特征，研究僵尸网络迁移检测方法，作为僵尸网络相似性度量模型的指标之一.

(4) 在建立僵尸网络相似性度量模型，准确识别使用不同控制服务器的相同僵尸网络的基础上，评估僵尸网络的规模以及危害；建立僵尸网络生命周期模型，研究僵尸网络的衍变特性.

参　考　文　献

[1] Freiling F, Holz T, Wicherski G. Botnet tracking: Exploring a root-cause methodology to prevent distributed denial-of-service attacks//Proceedings of the 10th European Symposium on Research in Computer Security (ESORICS 2005). LNCS 3679. Milan: Springer-Verlag, 2005: 319-335

[2] Rajab M A, Zarfoss J, Monrose F, Terzis A. A multifaceted approach to understanding the botnet phenomenon//Almeida J M, Almeida V A F, Barford P eds. Proceedings of the 6th ACM Internet Measurement Conference (IMC 2006). Rio de Janeriro: ACM Press, 2006: 41-52

[3] Zou C C, Cunningham R. Honeypot-aware advanced botnet construction and maintenance//Proceedings of the International Conference on Dependable Systems and Networks (DSN 2006), 2006: 199-208

[4] Zhuge J W, Han X H, Zhou Y L, Song C Y, Guo J P, Zou W. HoneyBow: An automated malware collection tool based on the high-interaction honeypot principle. Journal on Communications, 2007, 28(12): 8-13

[5] Binkley J R, Singh S. An algorithm for anomaly-based botnet detection//Proceedings of the USENIX 2nd Workshop on Steps to Reducing Unwanted Traffic on the Internet (SRUTI 2006). 2006: 43-48

[6] Livadas C, Walsh B, Lapsley D, Strayer T. Using machine learning techniques to identify botnet traffic//Proceedings of the 2nd IEEE LCN Workshop on Network Security. 2006: 967-974

[7] Goebel J, Holz T. Rishi: Identify bot contaminated hosts by IRC nickname evaluation//Proceedings of the 1st Workshop on Hot Topics in Understanding Botnets (HotBots 2007). 2007

[8] Gu G, Porras P, Yegneswaran V, Fong M, Lee W. BotHunter: Detecting malware infection through IDS-driven dialog correlation//Proceedings of the 16th USENIX Security Symposium (Security 2007). 2007

[9] Gu G, Zhang J, Lee W. BotSniffer: Detecting botnet command and control channels in network traffic//Proceedings of the 15th Annual Network and Distributed System Security Symposium (NDSS'08). 2008

[10] Gu Guofei, Perdisci Roberto, Zhang Junjie, Lee Wenke. BotMiner: Clustering analysis of network traffic for protocol- and structure-independent botnet detection//Proceedings of the USENIX Security, 2008: 139-154

[11] Karasaridis A, Rexroad B, Hoeflin D. Wide-scale botnet detection and characterization//Proceedings of the 1st Workshop on Hot Topics in Understanding Botnets (HotBots 2007). 2007

[12] Rajab M A, Zarfoss J, Monrose F, Terzis A. My botnet is bigger than yours (maybe, better than yours): Why size estimates remain challenging//Proceedings of the 1st Workshop on Hot Topics in Understanding Botnets (HotBots 2007). 2007

[13] Dagon D, Zou C C, Lee W. Modeling botnet propagation using time zones//Proceedings of the 13th Annual Network and Distributed System Security Symposium (NDSS 2006). 2006

[14] National Computer Network Emergency Response Technical Team/Coordination Center of China (CNCERT/CC), National Certificate of Network and Information Technology-Management Center of China (NTC-MC). Emergency and Practice Guideline of Network Security. Beijing: Publishing House of Electronics Industry, 2008(in Chinese)
(国家计算机网络应急技术处理协调中心(CNCERT/CC), 全国网络与信息技术培训项目管理中心(NTC-MC). 网络安全应急实践指南. 北京:电子工业出版社, 2008)

[15] Han Wook-Shin, Lee Jinsoo, Moon Yang-Sae, Jiang Haifeng. Ranked subsequence matching in time-series databases//Proceedings of the VLDB, 2007: 423-434

LI Run-Heng, born in 1982, Ph. D. candidate. His research interests include botnet and data mining.

WANG Ming-Hua, born in 1978, Ph. D., engineer. His research interests include network security monitor and emergency response technic.

JIA Yan, born in 1960, Ph. D., professor, Ph. D. supervisor. His research interests include network security and database.

Background

The paper aimed at solving problem on modeling IRC botnets' similarity of network security. Some work show the similarity can be measured by multi-dimensional data obtained from the infiltrated botnets, that is, some information, such as server version, IP address of IRC server, DNS name of IRC server, IRC server/network name, and botmaster ID, can be obtained by joining the command and control channel. Because such information doesn't represent the essential characteristic of botnets, and with the upgrade of server version, obtaining the information such as botmaster ID becomes more difficult, the error ratio of the model is hard to be bounded. This paper proposes a method that measures the similarity of botnets by extracting and comparing the metrics such as communication volumes, frequency, and the overlap rate of bots based on command and control traffic data between IRC server and zombie computers monitored by national network security monitoring platform. The experimental results show that the model has a good accuracy, and it dosn't need server version, IRC server name, botmaster ID and other information.

The internet infrastructure is the combination of computer and communications systems that serve as the underlying infrastructure for many areas. However, with the gradual in-depth of information technology, critical operations and assets continue to be highly vulnerable to computer-based attacks, internet security may cause very serious damage to our country at any time. Security of the internet is regarded as the most important security issue.

This paper was supported by the National High Technology Research and Development Program (863 Program) of China (2007AA010502, 2007AA01Z474, 2006AA01Z451). These projects may improve the level of the Internet network security by National Network monitor, threat assessment and crisis management. In this paper, a novel model for botnet similarity measuring is proposed, it may identify the same botnets and improve the capacity of threat assessment of botnets and malicious attacks such as DDoS attack caused by botnets.