入侵检测系统数据集评测研究*)

史美林 钱 俊 许 超
(清华大学计算机科学与技术系 北京 100084)

摘 要 入侵检测技术已经成为信息安全保障体系的重要组成部分。但是到目前为止,还没有广泛认同的入侵检测系统(IDS)评测标准,用户和研究人员对IDS和新的检测算法的有效性抱有疑问。解决这些问题的关键在于对IDS进行完善的评测。研究者对此提出了多种不同的IDS评测方案,如MIT Lincoln Lab提出的数据集评测和Neohapsis提出的OSEC(Open Security Evaluation Criteria)等。通过对评测结果的分析,能发现现有技术的不足,从而为IDS技术今后的研究提供指导。本文对MIT LL提出的数据集评测方法进行了详细分析,阐述了数据集评测方法中的关键问题,并在MIT LL研究的基础上,提出了相关改进方案,作为进一步的研究。

关键词 入侵检测,数据集,IDS评测,IDS测试

Research of Intrusion Detection System Dataset Evaluation

SHI Mei-Lin QIAN Jun XU Chao
(Department of Computer Science and Technology, Tsinghua University, Beijing 100084)

Abstract Intrusion detection technology has been an indispensable part of information security metrics, but till now no standard is widely accepted to evaluate the effectiveness and accuracy of intrusion detection systems (IDS). Thus, both the end users and researchers have doubt on the effectiveness of IDS and algorithms. The key point of the solution is to construct a precise and comprehensive methodology for IDS evaluation. Researchers have proposed several IDS evaluation methods, such as dataset evaluation proposed by MIT Lincoln Lab and Open Security Evaluation Criteria proposed by Neohapsis, etc. According to the evaluation results researchers may look through the weak point of current intrusion detection technologies and thus improve on them. In this paper we present a detail analysis of dataset evaluation methodology proposed by MIT Lincoln Lab and point out the key problems of dataset evaluation methodology. We also propose an improving research plan as our furthering work in order to update the evaluation dataset.

Keywords Intrusion detection, Dataset, IDS evaluation, IDS testing

1 概述

1.1 入侵检测技术的困境

经过20余年的发展,入侵检测系统(IDS)在信息安全防御体系中越来越受到重视。与加密等传统安全技术相比,IDS仍然很不完善,无论是理论模型还是实际应用的检测效果,都没有达到最初设计的期望。尽管业界不断宣扬IDS技术的进展有多么神速,然而事实无疑是被夸大了。新的技术在进入实用之前总是很谨慎的,方法学也并没有如此快的步伐更新换代。很多成熟的技术仍然在使用而且将会继续使用。多算法检测、基于状态的特征分析、后门流量异常、协议解析、启发式检测方法,层出不穷的名词更多地是一种市场策略行为。然而由此带来的负面影响却很大,因为它往往会向用户传达一种误导,这也是入侵检测遭到用户质疑的一方面原因。

随着IDS的普及和广泛使用,如何衡量和检验IDS的检测效果是很现实的问题。对于IDS有效性的怀疑不仅来自最终的用户,同时来自研究人员自身。如今入侵检测领域的研究已经向多学科交叉方向发展,很多不同领域的方法都被借鉴并尝试性地应用于入侵检测。如何比较不同的检测算法所获得的检测效果,是一个很实际的问题。目前入侵检测领域的研究存在一个恶性循环:找到一种算法,应用在入侵检测领域,发现没有得到满意的检测效果;然后换另外一种算法,重复上面的过程。这样做虽然能够验证很多算法在入侵检测领域的可用性,但并不能带来认知和检测的根本性突破。出现这种现象的部分原因在于没有有效的评测方法学。不能明确IDS技术需要改进的方向。到目前为止,还没有一个能被广泛接受作为标准的评测方法学,但是无论是研究者还是厂商都已经意识到了制定评测标准的重要性,而且相关的研究也一直在不断进行中。

1.2 入侵检测系统评测的意义

从科学研究的角度来看,任何一种理论都需要通过实践的检验;从软件工程的角度而言,软件系统和算法也必须通过相应的测试。对于IDS而言,这两个角度都具有实际意义。任何一种IDS都使用了一种或多种检测算法。无论是最简单的字符串匹配还是复杂的神经网络,这些算法本身都有其理论背景,但是将这些算法应用于入侵检测是否有效则需要进行验证。同时,每一个入侵检测系统同时也是一个软件系统,一个软件系统必然有其设计目标,因此对这些设计目标进行测试也是必需的。

*)国家863项目(编号:2001AA142020)。史美林 教授,博导,主要研究方向为计算机支持的协同工作、网络安全、网格技术;钱 俊 博士研究生;许 超 硕士研究生。

入侵检测系统评测的基本目标，一方面能够证实 IDS 的有效性；另一个方面，评测也能反映 IDS 的缺陷和不足，明确改进的方向，所以 IDS 评测对于研究人员而言非常重要。

IDS 评测更深远的意义在于，能够为入侵、入侵检测甚至于网络行为的模型建立提供支持。如果能够获得入侵行为、用户行为和网络行为的数学模型，那么所有的问题都会迎刃而解。但是无论是用户行为还是互联网本身都具有难以处理的复杂性，因此为它们建立精确的数学模型十分困难[1]。提出一种 IDS 评测方案，必须对入侵和入侵检测有深入的认识。同时，在实现评测方案的过程中，也会涉及到网络行为模式、用户行为模式的研究和模拟。因此，对于 IDS 评测的研究会为建立上述的几种模型提供经验和数据，从而帮助研究者逐步地发现入侵和入侵检测的本质。

2 相关工作

2.1 测试技术分类

目前所使用的测试技术通常可以分为以下 3 类：有效性测试、极限测试和综合测试。

有效性测试主要的测试目标是检验入侵检测系统是否达到了设计目标。比如对于基于特征的入侵检测系统，它应该能够准确无误地检测到其特征集所对应的所有攻击。在进行有效性测试的过程中，需要区分真实的攻击和伪造的攻击。有些工具专门针对基于特征的入侵检测系统来伪造攻击数据包。在获取了基于特征的入侵检测系统的特征集之后，这类工具可以根据这些特征自动生成伪造的攻击数据包。如果入侵检测系统不能够识别这些伪造的攻击数据包，那就会造成对入侵检测系统的拒绝服务攻击，使其消耗大量的运算资源，出现丢包的现象，从而有可能漏过了真正的攻击数据包。

极限测试通常是利用发包机来模拟高带宽网络中的背景流量，在流量中插入攻击，然后测试入侵检测系统的检测效果；并逐渐提高流量的带宽，直到入侵检测系统出现丢包现象或者流量达到了网络所能承受的极限。极限测试对于基于特征的入侵检测系统非常重要，因为测试表明，基于特征的入侵检测系统在模式匹配上会消耗大概 60%～70%的运算资源[2]。因此当网络带宽提高时，入侵检测系统是否能够保持有效性，就需要通过测试来确定，这也是极限测试的目标所在。进行极限测试一般需要使用专门的硬件设备，比如 Smartbit 或者 IXIA 等。

综合测试的目标是希望检验入侵检测系统在真实环境下的检测效果。相对于前两种测试方法，综合测试要复杂得多。首先，综合测试采用的测度(metric)比有效性测试和极限测试采用的评价标准复杂得多。对于有效性测试而言，可以用一个百分数来描述 IDS 能够实际检测出的攻击数量和特征集对应的攻击数量的比率；对于极限测试，可以给出一个最大的无丢包工作带宽；而对于综合性测试而言，需要制定一套可以综合衡量入侵检测系统的评估标准。综合测试的复杂性还反映在测试方案的实现上，由于综合测试的目的是检验入侵检测系统在真实网络环境中的检测效果，因此综合测试必须考虑如何仿真一个“真实的”网络环境。

2.2 研究工作分类

自从 IDS 步入成熟的商业化运作之后，入侵检测研究也逐渐分化成学术界和工业界两大阵营，围绕 IDS 评测的研究也分别来自学术界和工业界。

2.2.1 学术界的研究工作

学术界的研究工作主要包括加州大学 Davis 分校计算机安全实验室、IBM Zurich Research Lab 和 MIT Lincoln Lab (以下简称 MIT LL)的研究工作。

加州大学 Davis 分校计算机安全实验室在 1997 年提出了入侵检测系统软件测试平台[3]的研究工作。该研究工作的目标是希望能够为 IDS 测试提供一个通用的软件平台。在这样的平台上，使用者可以设计自己的测试方案。更进一步，这一平台甚至可以被扩展，用于其他软件系统的相关测试。该研究工作的重点是对于真实用户行为的模拟，然而在文[3]发表之后没有进一步的研究进展发布。

IBM Zurich Research Lab 在 1997 年进行了入侵检测评测的相关研究[4]。当时进行该研究的初衷是希望能够对比和评价实验室开发的不同的异常检测算法的检测效果。该研究首次提出了使用模拟背景流量进行测试的方法并做了初步的尝试，研究人员认为模拟异构网络中真实的用户行为是一件需要大量时间的工作，这一点在 MIT LL 的研究工作中得到了验证。该研究也没有进一步的工作进展公布。

1998 年 MIT LL 集前人研究之大成，提出了入侵检测系统数据集评测方法[5]，不仅是 IDS 综合测试的典范，也是目前为止学术界最有影响力的入侵检测评测研究。详细分析将在下一节展开。

2.2.2 工业界的研究工作

工业界的研究主要包括 Anzen Computing 公司的 NIDSbench、Neohapsis 公司的开放安全评测标准(Open Security Evaluation Criteria，OSEC)和 NSS 的安全产品评测。

严格说，NIDSbench[6] 只是一套用于测试网络入侵检测系统(Network Intrusion Detection System，NIDS)性能的工具集，由 Tcpreplay、Fragrouter 和 IDSTest3 部分组成。NIDSbench 并不包含用于测试的数据，也不提供对评测方法和评测过程的指导，它仅仅为研究者提供了一套可用于 NIDS 测试的工具。这个项目最终没有完成，其中最核心的 IDSTest 并没有实现，之后也没有后续的研究公布。

Neohapsis 是一个提供网络安全咨询服务的公司，它在 2002 年 8 月发起了以建立安全系统评测标准为目标的项目——开放安全评测标准(Open Security Evaluation Criteria，OSEC)。该项目的观点是：安全产品评测的标准应该是公开的，应该能接受他人批评而不断改进，这些批评可以来自厂商也可以来自于最终用户，评测的细节应该随评测结果一起公布。OSEC 对所有的网络安全产品定义了一套核心的测试，然后对于不同的系统会分别加入针对系统性能和安全性的测试。虽然 OSEC 是 Neohapsis 的商标，但是它所对应的标准是公开的[7]。

NSS 是欧洲的一家专业的商业评测机构[8]，该组织的评测工作开始于 1991 年。其评测的产品范围非常广泛，囊括了 IDS、IPS、防火墙、安全网关、PKI 等，基本上与网络安全相关的产品都提供评测。NSS 对于 IDS 的评测比较全面，甚至包括 IDS 的软件构架、检测引擎、安装和配置等各个方面都会进行评估。然而，由于是商业评测，因此无论是参加评测，还是希望得到评测报告和评测结果都需要付费，评测细节也不公开。NSS 最近发布的一份评测报告中自称该机构的评测方法和评测结果已经成为实际的标准。但是如果要成为普遍接受的标准，必须符合开放和自由原则。

3 入侵检测数据集评测研究

1998 年，MIT LL 集前人研究之大成，提出了入侵检测系

统数据集评测方法[5]，不仅是IDS综合测试的典范，也是目前为止学术界最有影响力的入侵检测评测研究。目前MIT LL公开发布的评测数据集一共有3年，分别是MIT'1998、MIT'1999和MIT'2000。下面按时间顺序进行分析。

3.1 MIT'1998

1998年提出的评测分为离线评测和在线评测。由于在线评测需要维护专门的网络环境，代价昂贵而且不具有移动性，因此离线评测体现出优势。离线评测所使用的就是评测数据集。评测数据集一般包括两部分：训练数据集和评测数据集。两者之间的区别在于，训练数据集中的攻击已经被标定出来，而评测数据集中的攻击并没有被标定出来。参评者可以利用训练数据集对入侵检测系统进行工作状态调整(tuning)，使其工作在最佳状态；然后使用评测数据集进行测试，并将结果提交MIT LL，由MIT LL给出最终评测分析报告。整个数据集由几个部分组成，包括网络流量的tcpdump文件、被攻击主机的日志文件以及相关的文件系统的备份，其中训练数据集还包括相应的攻击标记说明文件。

3.1.1 设计方案

综合测试面对的核心问题就是如何在真实的环境下测试入侵检测系统。任何一个IDS最终都是在实际的环境中工作，因此最简单直接的方法就是让IDS在实际的应用环境中测试。但是作为评测，实验的可重复性、真实网络流量中的攻击标定和一些私人敏感信息都必须考虑并加以解决。同时，使用真实流量测试无法提供一个基线作为参考，比较不同IDS的检测效果，这就失去了评测的重要意义。因此，在MIT LL看来，IDS评测面对的核心问题就是如何提供一个"真实的环境"用于测试IDS。

MIT LL从入侵检测系统要考察的两个关键指标——检测率和误报率入手分析问题。从网络数据流的观点来看，检测率是相对攻击流量而言，即入侵检测系统正确地检测出了多少攻击流量中的攻击；而误报率是相对于不包含攻击和恶意流量的干净的数据流量而言，即入侵检测系统对于正常流量发出了多少错误的警报。考察一个实际的网络环境，如果该网络环境中存在攻击或恶意行为的话，那么在逻辑上可以把网络流量划分成两部分：正常流量和恶意流量。如图1左边所示，通常一个实际的网络中正常流量远远大于恶意流量。其中，正常流量又可以按照不同的协议进行划分，如图1右边所示。正常流量中的协议和统计特征会随不同的网络环境、用户群和应用程序而表现出很大的差异。所以，一个实际网络的特征几乎是由正常流量体现的。

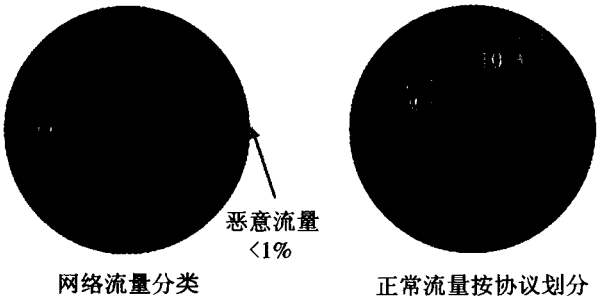

图1 实际网络环境中流量的划分

在这个思路的指引下，MIT LL提出如果可以人为地模拟出一个网络的正常流量和恶意流量，那么就达到了在"真实的环境下"测试IDS的目的，同时解决了实验可重复性、保护用户隐私、评测数据集公开发布等一系列问题。实验中对应地把需要模拟的正常流量称为背景流量，把需要模拟的恶意流量称为攻击流量。

3.1.2 背景流量

对于一个实际的网络而言，网络的基本特征是由背景流量而不是由攻击流量体现的。因此，如何生成背景流量十分关键，可选择的方案有3种：

- 将真实的网络流量作为背景流量；
- 将真实的网络流量除去敏感信息后作为背景流量；
- 使用合成的背景流量。

采用真实网络流量直接作为背景流量的优点是很明显的，因为它体现了真实网络环境的特征。但是它的缺点也很突出，首先是采用真实网络流量不可避免地会带来隐私问题。无论是用户的密码还是包含在电子邮件中的敏感信息，都会侵害用户的隐私。如果仅仅是内部测试使用，这个问题可以接受。但如果是提供给其他测试者使用的数据集，那么就可能带来很严重的问题：不但用户的隐私有可能受到侵害，甚至实验网络本身的一些信息，包括网络拓扑分布等，都会泄露，这些都是需要避免的。另外一个缺点就是无法保证真实的网络流量中不包含恶意流量。对于定量的测试而言，这可能会带来测试结果的误差，导致评测得到的结果不准确，进而影响公正性和公平性。

第二种方案其实是第一种方案的技术性改进方案。主要是通过一些后续的处理，将真实网络流量中的敏感信息过滤掉，包括更换数据包的IP地址、替换用户名、将敏感信息过滤掉等等。但实际上，这样的改进既因为信息量庞大变得不可行，也不会得到显著的效果，甚至情况变得更差。首先是受到技术的限制，现有的过滤技术往往只是将IP替换掉，对于数据包中的敏感信息，则采取全部删除的方法，这对于基于特征的网络入侵检测系统而言是致命的，因为它们的主要检测目标恰好是数据包中的负载部分。同时，仅仅通过净化技术还是无法很好地解决另一个问题，即背景流量中的攻击或恶意流量。

因此，考虑到数据集的公开性和定量的精确评测，使用合成的背景流量成为MIT LL最终的选择。这种方案既能较好地解决隐私问题，又能保留真实网络的特征。紧接着面临的关键问题就是采用什么样的模型生成背景流量，在这一点上是MIT LL的薄弱点，其背景流量建模方法主要是基于专家经验和粗略的统计模型，包括在网络层对网络流量特征的统计和对用户访问行为的特征统计，这一点也往往成为其他研究者所诟病的地方[9]。

3.1.3 攻击流量

对于攻击流量的生成，MIT LL实现了部分自动化调度，一部分攻击用脚本自动实现，一部分比较复杂的攻击还是手工实现。由于参加评测的入侵检测系统可能是基于主机的入侵检测系统，攻击都是实际发生的，因此不能简单地回放攻击流量。

首先需要考虑的问题就是选取什么样的攻击加入到测试中。入侵检测评测的目标之一就是为入侵检测技术的发展提供指导，而反映在实际的评测数据集中，就是数据集中的攻击覆盖。在数据集中加入所有已知的攻击会使工作量过于庞大而变得不现实，因此必须在完备性和覆盖性之间做折衷。目前解决这个问题的比较好的一种方法是采用某种分类方法对攻击进行分类，在每一类中选取一种或几种最具有代表性的

攻击，作为实验中使用的攻击。MIT 采用的攻击分类方法[10,11]并没有遵循一定的理论依据，只是根据专家的经验制定了 4 大类攻击类型。表 1 给出了攻击总表，黑体字的攻击表示只在评测数据集中出现，没有在训练数据集中出现。

表 1　MIT'1998 数据集攻击总表

	Solaris	SunOS	Linux	Cisco Router
Denial Of Service	**apache2** back **mailbomb** neptune ping of death **process table** smurf syslogd **udp-storm**	**apache2** back land **mailbomb** neptune ping of death **process table** smurf **upd-storm**	**apache2** back **mailbomb** neptune ping of death **process table** smurf teardrop **udp-storm**	
Remote to Local	dictionary ftp-write guest **http-tunnel** phf **xlock** **xsnoop**	dictionary ftp-write guest phf **xlock** **xsnoop**	dictionary ftp-write guest imap **named** phf sendmail **xlock** **xsnoop**	**Snmp-get**
User to Root	**at** eject ffbconfig fdformat **ps**	loadmodule	Perl **xterm**	
Surveillance/ Probing	ip sweep **mscan** nmap **saint** satan	ip sweep **mscan** nmap **saint** satan	ip sweep **mscan** nmap **saint** satan	ip sweep **mscan** nmap **saint** satan

3.1.4　网络环境

上面谈到的所有设计方案的实体化就是 MIT LL 用于生成入侵检测评测数据集的实验网络。如图 2 所示[5]，这个实验网络的原型是一个美国空军专用网络，模拟网络的各种特征数据，都是通过对原型网络统计分析得到的。从图中我们可以看出实验网络的拓扑结构以及配置和布署。整个实验网络分为内网和外网两个部分，内网主要是受害主机，而外网主要是攻击主机和提供应用层服务的模拟服务器。

整个实验网络和外界隔离，因此外网的模拟服务器需要模拟整个互联网络能够提供的资源和服务，比如电子邮件和 WEB 服务。MIT LL 使用一台机器来模拟所有提供 WWW 服务的 WEB 服务器，所有在统计工作中被保存下来的网页都存储在模拟的 WEB 服务器上。这台模拟的 WEB 服务器相当于成百上千台互联网络中的 WEB 服务器，所有访问外部网络中 WWW 网页的请求都会被定位到这台模拟的 WEB 服务器上。MIT LL 使用了特殊的技术来实现这一设计，修改了 Linux 的内核，使得一台机器可以同时有成百上千个 IP，每一个 IP 对应于一台虚拟的 WEB 服务器，每一台虚拟的 WEB 服务器会被映射为一个进程。同时，对于应用层的服务器程序，比如 Apache 等也要进行相应的修改。对于内网的用户而言，这一切都是透明的。

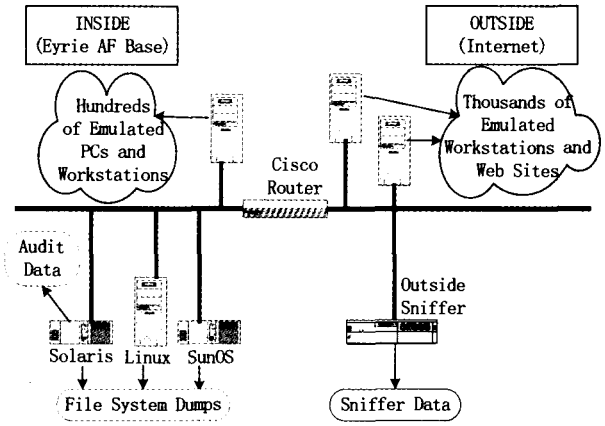

图 2　MIT' 1998 网络环境示意图

内网受害主机的软件配置需要和攻击选择同时考虑。这样做的原因是因为每种攻击通常都只针对某一特定版本的操作系统和应用程序。从图 2 可以看出，在 1998 年的实验中，服务器所使用的操作系统还主要是 UNIX 或者 Linux，因此所有受害主机使用的操作系统都是 UNIX 和 Linux。

3.1.5　评分系统

评分系统是评测方法的具体实现，分为两个方面：一方面是如何标定真实的攻击，另一方面是如何根据这些真实攻击的信息来评判被测试的入侵检测系统的测试效果。

这两个方面问题的一个交叉点就是如何选择分析粒度。这里的分析粒度是指用来对标记攻击并且对入侵检测系统进行评测的基本操作单元。以基于网络的入侵检测系统为例，从网络协议模型的角度来看，任何一层的代表性分析单位都可以选作分析粒度，比如 IP 数据包、TCP 会话、应用层的用户行为。每一种粒度都有其特有的抽象程度和数据细节。MIT LL 选择 TCP 会话作为分析的粒度，给出的最终评分标准是一张会话列表。每一个会话都会标明是否含有攻击，如果含有攻击，还会给出攻击的相关信息，如攻击名称，参加评测的入侵检测系统的最终检测结果也被要求以同样的形式给出。表 2 列出了从某天的训练数据集中抽取出来一些网络连接和攻击标注。

检测率和误报率是描述入侵检测系统性能的两个关键指标。MIT LL 首次提出使用 ROC(Receiver Operating Characteristic)曲线来描述误报率和检测率之间的关系。ROC 最初是用于信号检测领域的一种分析方法[12,13]，如今也广泛地用于语音识别[14]和医疗风险预测[15]。如图 3 所示，X 轴表示误报率，Y 轴表示检测率，ROC 曲线能在同一个二维坐标系中表示出检测率和误报率之间相互影响的关系。入侵检测系统可能的工作状态可以是 ROC 曲线上的任何一点。我们也可以以此来调整入侵检测系统的工作状态，使其处于我们所希望的最佳工作状态。在实际的使用中，由于误报率并不是一个具有直观意义的参数，所以 X 轴的坐标通常会被换成更直观的每天的误报数。如果数值变化范围比较大，X 轴也可以相应地采用对数坐标。图 3 中系统 1 为异常检测系统，系统 2 为基于特征的检测系统。

表 2 训练数据集中的部分网络连接和攻击标注

#	Start Date	Start Time	Duration	Service	Src Port	Dest Port	Src IP Address	Dest IP Address	Attack Score/Name
1	07/03/1998	08:00:01	00:00:01	eco/i	-	-	192.168.1.5	192.168.1.1	0 -
4	07/03/1998	08:00:02	00:00:01	domain/u	53	53	172.16.112.20	192.168.1.10	0 -
8	07/03/1998	08:01:03	00:00:01	smtp	1026	25	172.16.113.84	194.7.248.153	0 -
9	07/03/1998	08:01:06	00:00:02	smtp	1027	25	172.16.113.84	135.13.216.191	0 -
42	07/03/1998	08:01:50	00:00:29	ftp	1106	21	172.16.112.49	97.218.177.69	0 -
43	07/03/1998	08:01:51	00:00:01	http	1107	80	172.16.116.44	167.8.29.15	0 -
44	07/03/1998	08:01:51	00:00:01	http	1104	80	172.16.116.44	167.8.29.15	0 -
53	07/03/1998	08:01:52	00:00:01	http	1297	80	172.16.116.44	167.8.29.15	0 -
73	07/03/1998	08:01:52	00:00:02	ftp-data	20	1685	197.218.177.69	172.16.112.149	0 -
76	07/03/1998	08:01:53	00:00:01	snmp/u	161	1523	192.168.1.1	194.27.251.21	0 -
8383	07/03/1998	11:12:16	00:00:26	telnet	20504	23	197.218.177.69	172.16.113.50	1 Loadmodule
9966	07/03/1998	11:46:39	00:00:01	tcpmux	1234	1	205.160.208.190	172.16.113.50	1 Portsweep
10096	07/03/1998	11:49:39	00:00:01	2	1234	2	205.160.208.190	172.16.113.50	1 Portsweep

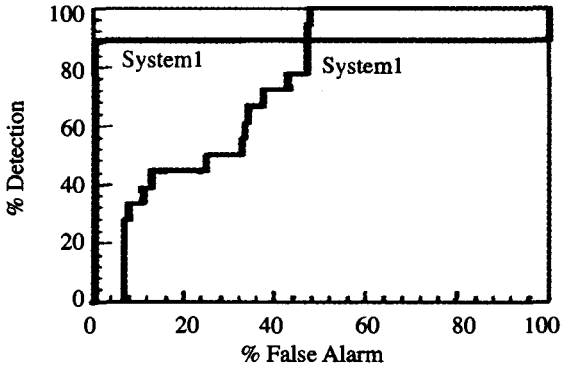

图 3 ROC 曲线图

3.2 MIT'1999

1998 年的评测取得了比较好的效果，同时得到很多研究者的认同。MIT LL 在 1998 年工作的基础之上，同时参考了参评者的反馈意见，发布了 1999 年的评测数据集。1999 年的工作和 1998 年的工作十分相似，但也有一些比较明显的改进。鉴于上面已经详细介绍了 1998 年的工作，下面就只介绍改进部分的工作。

3.2.1 数据集构成

首先，1999 年的数据集中，用于训练的数据集增加了完全不含有攻击流量的训练数据。增加这一部分训练数据集是为了满足一些基于异常的入侵检测系统的训练需要。基于异常的入侵检测系统需要使用不包含攻击流量的数据集进行训练，建立正常行为模型，确定检测参数，然后再进行实际的检测。

其次，在 1999 年的数据集中，增加了对内网网络流量的记录文件以及 Windows NT 的主机日志。来自于内部网络的攻击可能具有更大的破坏力和隐蔽性，而这一类攻击在以往的测试中往往被忽视。MIT LL 在 1999 年的实验中加入了源自内部网络的攻击实例，对于基于网络的 IDS，需要内部网络流量的记录文件来检测这些攻击。在 1999 年，Windows NT 作为服务器操作系统已经被广泛地使用，针对它的攻击也越来越多。考虑到参加评测的可能有使用 Windows NT 主机日志的 IDS，因此 Windows NT 的主机日志也成为了 1999 年数据集的一部分。

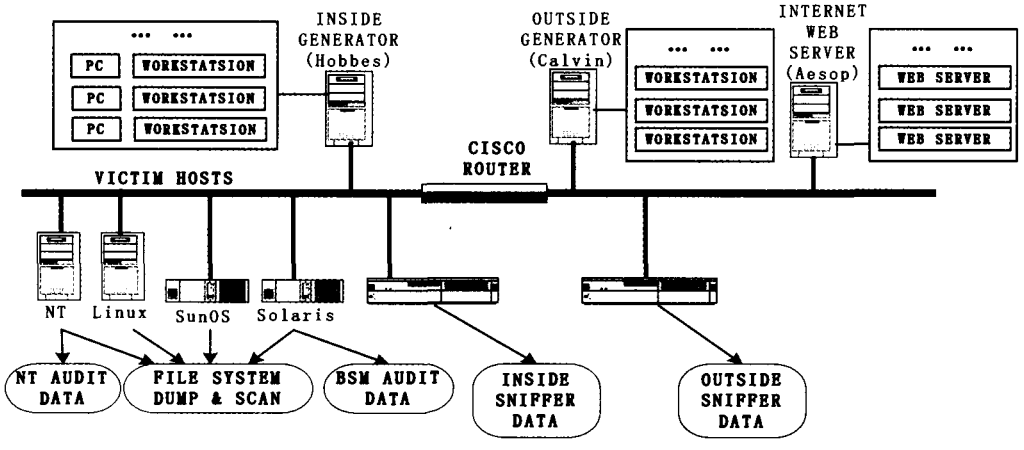

图 4 MIT'1999 网络环境示意图

3.2.2 攻击流量

与 1998 年相比，1999 年的攻击无论是种类还是数量都有所增加，攻击流量部分的改进有两个方面。

首先，由于 Windows NT 作为服务器操作系统的加入，MIT LL 在这一年的攻击中加入了针对 Windows NT 平台的攻击。其次，根据 1998 年的评测结果，绝大部分参加评测的入侵检测系统对于隐蔽性强的或者新的攻击的检测效果都很差，因此在 1999 年的攻击流量中，MIT LL 增加了这类攻击所占的比例。与 1998 年有所不同的是并非所有隐蔽性强或者是新的攻击都曾经在训练数据集中出现，其中一部分攻击只在评测数据集中出现，这样可以更深入地测试入侵检测系统对于未知攻击的检测能力。

3.2.3 实验网络

整个实验网络的基本构架和1998年相比保持不变，但是在内网要增加Windows NT平台的受害主机和一台用于记录内网流量的主机，如图4所示[17]。

3.2.4 评分系统

1999年的评分系统中一个比较大的改变是分析粒度的变化。1998年采用的分析粒度是TCP会话，但是采用TCP会话作为分析粒度计算出来的检测率和误报率会不够精确。比如，某一次攻击行为可能分布在多个TCP会话中。系统A针对这些会话中的两个发出了警报，但是并没有准确地描述出攻击的全貌；而系统B的警报综合程度比较高，仅仅发出了一次警报，但是准确地给出了复杂攻击的相应信息。那么依照1998年的评分系统，系统A会得到比系统B更好的评测结果，但这与实际情况不符。因此，在1999年的评测中，参加评测的IDS可以在检测结果中给出相应的附加信息，比如攻击名称和其他一些细节，这些可以作为额外加分的参考依据。

另一方面，有些参加评测的入侵检测系统给出的警报是针对抽象层次比较高的事件，也有一些具有多检测单元的系统对于抽象层次比较低的警报还要进行协同和综合分析，生成更高态势的警报。对于上述两类入侵检测系统而言，如果采用TCP会话作为分析的基本粒度，并在此基础上给出检测结果，会导致实际操作的困难。因此在MIT'99中改为使用更高抽象层次的事件作为分析的基本粒度。每一次攻击都作为一个单独的事件进行标定，给出相关的信息，比如攻击源、攻击目标、发起时间、结束时间等等。采用这样的分析粒度会有助于提高分析的精度，特别对于多阶段攻击或者是比较复杂的协同式攻击，采用TCP会话作为分析粒度可能无法给出攻击的全貌。

3.3 MIT'2000

MIT'2000的工作无论是设计方案还是最终实现的数据集都与1998年和1999年的工作有很大的不同。2000年的评测没有使用多种攻击进行测试，而是挑选了特定的一类攻击DDoS作为测试对象。相应地，测试数据集的规模也变得较小，仅有两个数据集，每个大概记录了4小时左右的网络流量。这种变化反映了MIT LL评测思想的变化。1998年和1999年的评测中有很多种类的攻击实例，可以认为是一种广度测试；2000年所采用的这种方案，可以认为是一种深度测试。广度测试可以测试出入侵检测系统对攻击检测的覆盖度，也就是能够准确检测出多少种攻击；而深度测试则可以集中地测试入侵检测系统对于某一种攻击的检测效果，对检测算法和检测机制可以进行深入的分析。如果要设计综合全面的入侵检测系统的测试，那么这两种测试都是必不可少的。

由于是针对某一种特定攻击的测试，因此对于攻击场景的设计就有更高的要求。每一个数据集中都包含了一个典型的、完整的DDoS攻击，包括试探、侵入、安装攻击程序、发动DDoS等多个步骤。这样的攻击场景设计，要求被测试的入侵检测系统能够通过检测、分析和综合等步骤，最终识别出攻击者的实际意图——发动DDoS攻击。

MIT LL在2000年的工作并没有详细的相关文档和论文进行介绍，所以很多的细节问题都没有答案。但是沿着1998—1999—2000年的研究轨迹来分析，2000年数据集的这种变化并非是完全没有征兆的。这种演变的理由就是希望能够研究入侵检测系统的工作原理，找出检测失败的更本质的原因，从而促进入侵检测技术的发展。

3.4 MIT'2000以后的工作

2000年以后，MIT LL的研究工作从公开转为政府内部专用，就没有再发布评测数据集，也没有足够的论文和技术报告来描述他们的工作，只能从其他的文献中[18,19]看到一些工作扩展计划，但是都没有进行详细的描述。2000年以后，MIT LL的工作重点是LARIAT(Lincoln Adaptable Real-time Intrusion Assurance Testbed)。LARIAT是用于入侵检测评测的专用评测网络，并不对外界公开。LARIAT的前身就是进行1998年和1999年实验的testbed。MIT LL对其进行了扩展，增加了用于配置实验网络的GUI和相关软件，使得实时的IDS评测可以很迅速地完成配置和准备工作。

3.5 小结

评测对于技术发展和研究导向有着积极的推动作用，例如由DARPA支持的每年一度的语音识别领域的评测[20]就是促使技术快速发展的一个范例。使用系统的方法学对入侵检测系统进行综合的评测是入侵检测技术发展的趋势。数据集评测作为开放的研究是一个很好的解决方案，其开放共享的原则对于研究者很有益处，既避免大家斥巨资去做试验，又可以实现综合全面的测试。其次，数据集不仅仅只用于入侵检测评测，在入侵检测研究领域还有其他方面的很多应用，如专门用于入侵检测数据挖掘的数据集KDD'1999(Knowledge Discovery in Databases)。KDD'1999数据集来源于1998年的MIT LL数据集，只是针对数据挖掘的需要进行了专门的格式化处理[21]。因此，数据集研究对于入侵检测技术的发展来说具有很积极的意义。

MIT的数据集评测研究工作获得了广泛的认可，很多研究者都采用MIT LL的数据集作为实验测试数据[22~27]。但是另一方面，MIT LL的数据集也并非完美，最突出的问题是数据集内容已经较为陈旧。一方面，MIT LL数据集的网络应用环境带宽是10M以太网，鉴于目前的网络发展和实际应用情况，100M以太网环境的数据集才能符合目前的应用需要。另一方面，数据集中的攻击大部分现在看来都很陈旧了，无法适应现阶段的应用环境。同时，MIT LL在数据集研究中使用的一些方法和模型，在设计和具体实现时仍然有值得改进的地方。下一节将提出数据集评测研究中的关键问题并分析MIT LL研究中的不足，从而使评测数据集更加合理、更加完善。

4 数据集评测中的关键问题与改进

数据集生成包括以下几个关键环节：背景流量的生成和验证、攻击方案设计和评分系统。每一个环节在具体实施的时候较为复杂，因此在这节只作概要叙述，更细节的内容将在另外的论文中述及。

4.1 背景流量生成方案

其他研究领域的学者也在研究网络流量的生成问题[28,29]。但是由于研究的目标不同，所关注的重点也不同。这些研究主要关注协议行为，比如数据包的到达时间、连接请求的响应延迟等。入侵检测评测的流量生成则更关注用户行为或者是应用程序的行为对于网络的影响。因此，用户的网络行为模拟是背景流量生成中重要的一环，需要提取相关的网络和用户行为特征建立模型，同时提供可行的验证方法。MIT LL的背景流量生成主要是依据专家经验和统计模型，即研究者根据经验，抽象出对入侵检测系统的运行有影响的

网络参数，比如流量的时间分布、协议分布以及简单的用户行为特征等，然后对样本网络进行采样统计，得到参数的统计模型，再依据模型生成背景流量。MIT LL 所采用的用户行为模型非常简单而且固定，所以模拟时缺乏灵活性，而灵活的配置是新一代数据集应该具备的特点和功能。以 WEB 流量模拟为例，MIT LL 所采用的 WEB 流量模拟方法存在的不足表现在：

(1)用户数据采集。在网络层进行数据采集，仅仅获得所有用户访问的 URL 地址集合，没有对不同用户的行为特征进行区分。

(2)流量模拟。在流量模拟过程中，所有用户共享同一个资源列表，并且完全随机地抽取访问地址，每个用户的会话长度也在 1～15 之间随机抽取。

这些与真实的网络环境和用户行为是不符合的。对于基于网络特征和用户行为学习的异常检测系统，会造成误报信息，从而导致评测精结果的偏差。针对 MIT LL 数据集背景流量的设计和模拟方法的缺陷，我们也提出了一种基于用户轮廓的流量模拟方法，采用该方法生成的数据集可以更真实地模拟用户行为和背景流量，并具有良好的扩展性。图 5 给出了 MIT'99 数据集 WEB 流量分析、我们的数据集 WEB 流量分析和实际网络 WEB 流量分析的对比。从图 5(a)中可以看到曲线频域很窄，主要集中在 10 至 100 之间。这是由于 MIT LL 在建模时采用了随机模型，因此整个曲线表现出比较明显的统计特征。在 x 小于 10 时，图 5(a)中的曲线特征消失，并且在 $10<x<100$ 区间内低频至高频的过渡也不明显。对比而言，图 5(b)中的曲线频域很宽，低频特征很明显，低频至高频的过渡也非常清晰。

另外，MIT 也没有对模拟结果进行验证和评估，这也成为有些研究者质疑的地方。图 5(c)是我们样本网络真实 WEB 流量的特征分析。对比图 5(b)和(c)可以看出，模拟流量保留了实际网络的特征，低频特征和实际流量吻合得非常好，高频部分与实际流量吻合效果稍差。一方面，这种偏差是由模拟算法本身决定的；另一方面，高频部分的偏差对实验的影响是可以接受的。因为实验结果表明，基于特征学习的 IDS 的误报率对低频特征的偏差非常敏感，对高频特征的偏差不敏感，误报率基本都是由短会话引起的，即对应曲线的低频部分。

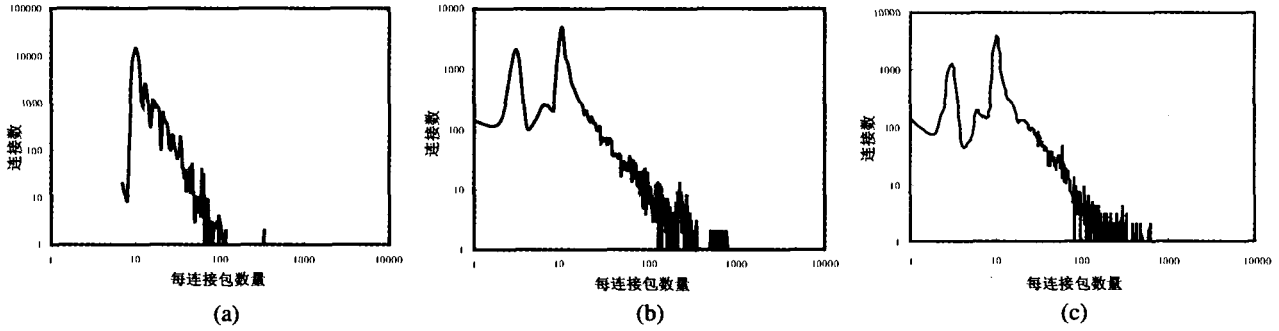

图 5 (a)MIT'99 数据集 WEB 流量分析，(b)我们的数据集 WEB 流量分析，(c)实际网络的 WEB 流量分析

4.2 攻击方案设计

攻击方案的设计包括攻击样本库和攻击场景设计。为建立符合入侵检测评测需要的攻击样本库，首先需要建立攻击样本库所依据的分类标准。目前尚没有被广泛采用的分类标准，因为随着新的攻击方式不断地出现，攻击的分类方法会不断地演化，甚至当某些新的攻击方式出现后，整个攻击分类方法都会面临挑战。目前信息安全组织 SANS 公布的分类标准[30]和很多入侵检测系统采用的 CVE[31]或 Bugtraq[32]分类标准都是以软件的漏洞作为分类的目标，并不完全符合入侵检测对于攻击分类的要求，因为一个良好的分类方法必须同时考虑攻击者和检测者。MIT LL 采用了自行定义的分类标准，以攻击者的角度进行分类。我们认为可以采用一个二维的攻击分类标准，从攻击者的角度和检测者的角度同时考虑攻击分类，这样 CVE 中的每一项都可以映射到二维分类矩阵的每一个矩阵单元中。

在具体实施上，MIT'98 和 MIT'99 数据集涵盖了很多类型的攻击，这些攻击中的大部分现在看来都很陈旧了，无法适应现阶段的应用环境。攻击样本库缺乏更新和升级也是 MIT LL 数据集最大的问题所在。

攻击场景设计和背景流量设计不同，攻击场景没有可以依据的理论模型，所以设计什么样的攻击场景很大程度上取决于专家经验。一个值得注意的问题是，作为入侵检测系统的研究者，通常看待攻击的角度都是从检测的角度来看的。反映到实际的数据上，就是数据包中的特征或者是主机日志中的异常。然而，攻击场景要实现的是真正的攻击过程，具有完整的攻击流程，同样也应该经过精心的策划和准备，这就要求设计者必须从攻击者的角度来考虑问题。不同的网络环境对于不同的攻击关注度也是不一样的。比如商业站点对于扫描不是非常关心，但是希望检测和避免大规模的 DDoS；而政府或信息敏感部门却对攻击的前兆，如扫描等有着较高的警惕性。因此，攻击设计和网络环境有很大关系，MIT 的数据集设计中并没有考虑到这部分因素。

另外一个值得注意的问题是，随着攻击技术的发展，新的攻击方式不断产生，特别是有一些极端的攻击。在实验中我们发现，有些极端攻击会导致主机失去响应，而背景流量仍然在产生，这是一个技术上必须解决的问题。比如分布式拒绝服务攻击，冲击波病毒等攻击就是很典型的一类攻击，占用带宽很大，造成数据集也非常大。因此我们认为，对于一些特殊类别的攻击应当分开测试，单独生成不同的数据集进行测试，这在 MIT'2000 里面已经有所体现。

4.3 评分系统

近年来，信息深度关联分析在工业界被认为有较强的实用性，它可以大大减少无关警报信息并提高入侵检测精度。很多入侵检测系统能结合网络及检测的上下文环境，给出抽象层次比较高的事件警报。分布式和协作式 IDS 也为广度上的关联分析提供了数据支持。它们通常具有多个检测单元，这些检测单元给出抽象层次比较低的警报，需要进行协同分析和综合，从而生成更高层次的警报。随着越来越多的 IDS

都支持各种聚合和关联分析，这项功能会越来越普遍。因此，在进行入侵检测评测时，应考虑到对关联分析的支持。有些实施关联分析的入侵检测系统的输入不是原始的数据流，而是其他设备的警报。因此，如果评测数据集对这些系统支持的话，需要直接产生真实警报或日志，文件作为输入。MIT'99数据集曾经给出了基于不同操作系统平台的日志供参评系统使用，虽然当时很少有系统利用这些信息，但这无疑是支持关联分析的一个初步尝试。MIT'99数据集的不足在于没有为关联分析日志提供统一的描述格式和管理接口，我们认为采用XML可以有效解决这个问题。为了获得更客观公正的评价结果，更好地支持关联分析将是数据集必需具备的一个功能。

如何进行标记攻击是评测数据集的一个关键问题，因为攻击标记不仅直接影响评测精确性，同时是评测数据集有别于其它数据集的一个显著差别。如Honeypot[33]和DEFCON组织的黑客竞赛[34]记录的数据集，这些数据集具有很好的研究价值，可以用于黑客行为和攻击模式研究，但是不能作为评测数据集使用，因为它没有经过攻击标记。虽然MIT的数据集较好地解决了这个问题，但是由于有些新的攻击在行为和方式上比以前的攻击复杂得多，甚至在实验中我们发现，即使相同的攻击导致的场景还存在不可预测的现象，这些都给攻击标记带来了很大难度和新的挑战。因此，更好的攻击标记和场景取证是维护评测数据集精确性的重要保证。

结论与进一步的工作 与简单的测试不同，IDS评测应该遵循一定的标准，并逐步形成科学的方法学。IDS最终需要部署在真实的网络环境中进行实际使用，因此评测的实验环境应该尽可能模拟真实的工作环境。如果评测包含的内容仅仅是一些现实情况中根本不可能出现的古怪情形，那就会失去评测的意义。

目前IDS评测研究可以分为学术界和工业界两大派。学术界的IDS评测研究主要针对原型系统的有效性测试，注重算法和系统性能的评测，研究遵行开放原则。工业界的IDS评测面向商业IDS产品，除了有效性之外，还注重易用性、维护性等非技术指标的评测。我们应提倡安全产品的评测标准遵循开放原则。除了学术界之外，Neohapsis公司也提出了安全产品评测标准的公开化，并提出相应的开放安全评测标准。就商业模式而言，公司可以从出售服务和技术支持来获得利润，而不应该以标准本身和垄断标准为盈利手段，这将不利于技术的发展。

到目前为止，国内尚没有全面而细致剖析MIT LL数据集研究的论文。本文简要介绍了IDS评测研究，并重点分析了MIT LL的数据集评测研究工作，希望能为国内的相关研究提供一些有价值的参考信息。数据集评测研究对于入侵检测技术发展来说具有很积极的意义。MIT的工作获得了研究者们的普遍认可，很多研究者都采用MIT LL数据集作为实验测试数据。但是另一方面，MIT的数据集也并非完美，在理论方法和具体实现上仍有值得改进的地方。文中提出了数据集评测中的几个关键问题和有待改进的方面，限于篇幅，将在另外的文章里阐述进一步的研究工作进展和实施细节。

参考文献

1 Floyd S, Paxson V. Difficulties in simulating the Internet. IEEE/ACM Transactions on Networking, 2001, 9(4): 392～403
2 Spyros A, Kostas G A, Evangelos P M. Generating Realistic Workloads for Network Intrusion Detection System. In: Proc. of the 4th International Workshop on Software and Performance (WOSP), 2004
3 Nicholas P, Mandy C, Ronald A O, et al. A Software Platform for Testing Intrusion Detection Systems. IEEE Software, 1997, 14(5): 43～51
4 Debar H, Dacier M, Wespi A, et al. An Experimentation Workbench for Intrusion Detection Systems: [Tech-Report]. RZ1998. IBM Zurich Research Lab, 1998
5 Lippmann R P, Fried D J, Graf I, et al. Evaluating Intrusion Detection Systems: The 1998 DARPA Off-Line Intrusion Detection Evaluation. In: Proc. of the 2000 DARPA Information Survivability Conference and Exposition (DISCEX), Los Alamitos, CA, 2000, 2: 12～26
6 http://packetstorm.widexs.nl/UNIX/IDS/nidsbench/
7 http://osec.neohapsis.com/about.html
8 http://www.nss.co.uk/default.htm
9 McHugh J. Testing intrusion detection systems: A critique of the 1998 and 1999DARPA intrusion detection system evaluations as performed by Lincoln Laboratory. ACM Transactions on Information and System Security, 2000, 3(4): 262～294
10 Kristopher K. A Database of Computer Attacks for the Evaluation of Intrusion Detection System: [Master thesis]. Massachusetts Institute of Technology, 1999
11 Kumar J D. Attack Development for Intrusion Detection Evaluation: [Master thesis]. Massachusetts Institute of Technology, 1999
12 John A S. The Relative Operating Characteristic in Psychology. Sceince, 1973, 182: 990～1000
13 James P E. Siganal detection theory and ROC-analysis. Academic Press, 1975
14 Martin A, Doddington G, Kamm T, et al. The DET Curve in Assessment of Detection Task Performance. In: Proc. of EuroSpeech, 1998, 4. 1895～1898
15 Lippmann R P, David M S. Coronary Artery Bypass Risk Prediction Using Neural Networkds. Annals of Thoracic Surgery, 1997, 63: 1635～1643
16 Joshua W H, Lippmann R P, David J F, et al. 1999 DARPA Intrusion Detection System Evaluation: Design and Procedures: [Tech-Report]. TR1062. MIT Lincoln Laoratory, 2000
17 Lippmann R P, Haines J W, Fried D J, et al. The 1999 DARPA Offline Intrusion Detection Evaluation. Computer Networks, 2000, 34(2): 579～595
18 Haines J A, Rossey L M, Lippmann R P, et al. Extending the DARPA Off-Line Intrusion Detection Evaluations. In: Darpa Information Survivability Conference and Exposition (DISCEX) II, 2001, 1: 77～88
19 Nicholas A, Randal A, John L, et al. Intrusion Detection Testing and Benchmarking Methodologies. In: Proc. of the 1st IEEE International Workshop on Information Assurance (IWIA'03), Washington DC, USA, 2003. 63～73
20 Cole R A, Mariani J, Uszkoreit H, et al. Survey of the State of the Art in Human Language Technology. Center for Spoken Language Understanding CSLU, Carnegie Mellon University, Pittsburgh, PA, 1995
21 http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html
22 Ning P, Cui Y. An intrusion alert correlator based on prerequisites of intrusions: [Tech-Report]. TR2002-01. Department of Computer Science, North Carolina State University, 2002
23 Matthew V M, Philip K C. Learning nonstationary models of normal network traffic for detecting novel attacks. In: Proc. of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. Edmonton, Alberta, Canada, 2002. 23～26
24 Sekar R, Gupta A, Frullo J, et al. Specification Based Anomaly Detection: A New Approach for Detecting Network Intrusions. In: Proc. of the 9th ACM conference on Computer and communications security, Washington DC, USA, 2002. 265～274
25 Lee W, Stolfo S. A Framework for Constructing Features and Models for Intrusion Detection Systems. ACM Transactions on Information and System Security, 2000, 3(4): 227～261
26 Ke W, Salvatore J S. Anomalous Payload-based Network Intrusion Detection. In: Proc. of the 7th International Symposium on Recent Advanced in Intrusion Detection (RAID), 2004. 201～222
27 杨德刚. 基于模糊C均值聚类的网络入侵检测算法. 计算机科学, 2005, 32(1): 86～91
28 Kamath P, Lan K, Heidemann J, et al. Generation of High Bandwidth Network Traffic Traces. In: Proc. of the 10th IEEE International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunications Systems, 2002. 401～412
29 Lan K, Heidemann J. A Tool for RApid Model Parameterization and its Applications. In: Proc. of the ACM SIGCOMM Workshop on Models, Methods and Tools for Reproducible Network Research, Karlsruhe, Germany, 2003. 76～86
30 The Experts Consensus: The Twenty Most Critical Internet Security Vulnerabilities. SANS Institute: [Version 5.0], 2004. http://www.sans.org/top20/
31 Mell P, Grance T. ICAT Metabase CVE Vulnerability Search Engine. National Institute of Standards and Technology, 2002. http://icat.nist.gov/
32 Security Focus Bugtraq Vulnerability Database. http://securityfocus.com
33 Spitzner L. Honeypots: Tracking Hackers. Addison-Wesley, 2003
34 Cowan C, Arnold S, Beattie S M, et al. Defcon Capture the Flag: Defending Vulnerable Code from Intense Attack. DARPA Information Survivability Conference and Expo (DISCEX III), Washington DC, 2003