基于多测度约束的快速蠕虫传播源定位算法研究*)

顾荣杰[1] 晏蒲柳[1] 黄先锋[2] 杨剑峰[1]
(武汉大学电子信息学院通信工程系 武汉 430072)[1]
(武汉大学测绘遥感信息工程国家重点实验室 武汉 430079)[2]

摘 要 近年来频繁爆发的大规模网络蠕虫对 Internet 的整体安全构成了巨大的威胁,已经造成了巨额的经济损失,新的变种仍在不断出现。目前对于蠕虫的监测与响应都是事后与人工的。本文提出了一种新的基于模式发现的多测度蠕虫快速定位方法,通过源地址活跃度、目标地址离散度和响应度准则等多个测度对监测目标网络已知和未知蠕虫的活动进行快速定位。基于本文的方法在应用中能以较低的资源代价发现未知的蠕虫传播并进行快速源定位。此外为提高算法的效率,本文研究了一种基于双页表结构的攻击树构建方法。

关键词 蠕虫传播,快速源定位,基于双页表结构的信息树构建算法

A New Fast Worm Source Tracing Algorithm of Multi-Constraints

GU Rong-Jie[1] YAN Pu-Liu[1] HUANG Xian-Feng[2] YANG Jian-Feng[1]
(School of Electronic Information, Wuhan University, Wuhan 430072)[1]
(State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079)[2]

Abstract The frequent explosion of massive worm propagation becomes a huge threaten to Internet security and caused countless losses, but endless novel worm species come one after another. Currently, worm monitoring and response are hysteretic and mainly operated artificially. As an improvement, this paper introduces a new fast tracing algorithm for worm infected hosts inside the object network. It uses source IP activity measure, dispersion measure and request response measure as a joint constraint to locate the hosts which have been infected by known and unknown worms. The three advantage of this algorithm is low cost, fast and the ability to detect unknown worm activities without any prior knowledge about the worm. To enhance the algorithm efficiency, a fast attack tree construction algorithm based on 2-page hash structure is proposed in this paper, and the further experiment results prove it can effectively improve the process efficiency.

Keywords Worm spreading, Worm source tracing algorithm, 2-page hash algorithm

1 背景介绍

信息技术的飞速成发展,尤其是"全球信息高速公路"即 Internet 的迅猛发展,使国际间的信息交流日益便利、迅捷,有力地促进了经济全球化趋势的发展。但是,随着互联网应用的不断创新与发展,日益增长的网络安全事件也严重威胁着网络的稳定性和可靠性。蠕虫是一段能够自动运行并且能够通过网络进行自身传播的代码[1]。近几年里,蠕虫对于网络安全的威胁日益严重。1988 年 11 月,Morris 蠕虫[2]造成了世界 6000 多台计算机瘫痪;2001 年 7 月爆发的红色代码[3]在 9 个小时内感染了 25 万台计算机系统;2003 年 1 月 SQL Slammer 蠕虫[4]在爆发后的五天里,全球经济损失超过 12 亿美元,与此同时中国将近有 80%的网络陷于瘫痪。

经历了几次全球性的蠕虫大爆发之后,蠕虫的检测已经成为网络安全研究的热点问题。传统的基于特征检测的方法对于大量未知蠕虫的活动已经失效。Cliff Changchun Zou[5]等人以研究简单病毒传播模型为基础,当监测到非法扫描活动大量增加时,利用 Kalman 滤波算法估计蠕虫的感染率。当估计值在一个正常数附近变化时,则认为发生了蠕虫活动。这种方法只是检测到了异常而无法实现定位。George Bakos 等人[6]通过检测 IP 地址产生的 ICMP-T3(目标地址不可到达)报文数量来发现可能的蠕虫行为,若检测到某个 IP 地址产生的 ICMP-T3 报文数超过设定的阈值并发现该地址与多个系统的同一个端口发生了通信,则判定该地址发生了蠕虫。这种方法在一定程度上能检测到蠕虫的行为,但是它需要修改路由器配置,并将第二份 ICMP-T3 消息拷贝并发送到指定的数据采集点,当网络带宽紧张时,特别是蠕虫爆发时会带来大量的流量,加重网络带宽的压力。

其它也有基于流量阈值基线的检测方法[7,8]、基于流量解析[9,10]等一些基于归并流量的趋势和特征的分析方法,但是基于归并的统计分析方法的最大问题是无法还原独立 IP 地址等底层的细节信息,无法实现蠕虫的基本定位。本文将在蠕虫传播特征分析的基础上,提出一种新的基于多测度约束的快速源定位方法,并提供实验验证。

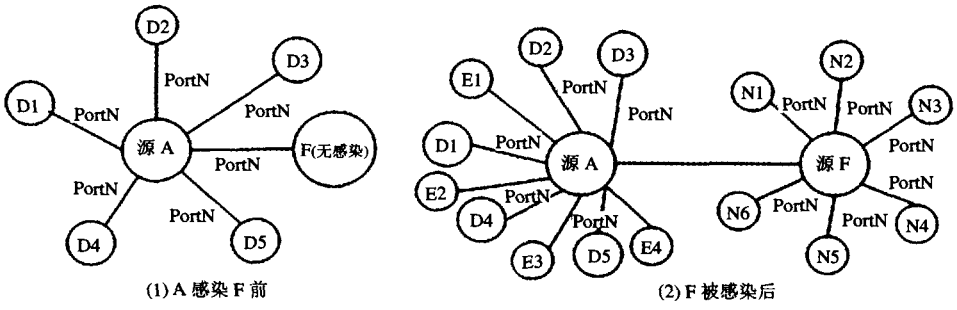
2 蠕虫行为特点与访问兴趣度关系模型

2.1 蠕虫的行为模式

分析蠕虫的典型传播行为,图 1 描述了蠕虫从结点 A 向

*)基金项目:国家自然科学基金(90204008)。顾荣杰 博士研究生,主要研究方向包括智能网络信息处理,网络安全等。晏蒲柳 博士,教授,博士生导师,长期从事计算机网络通信,网络管理等方面的研究。

结点 F 传染前后的模式变化情况。



(1) A 感染 F 前　　(2) F 被感染后

图 1　蠕虫感染前后通信模式特点的变化(结点 F)

从图 1(1)中可以看到,源 A 向 D1～D5 及 F 大量发送指向同一端口的攻击数据包,源 A 的通信模式呈现发散状,且其频繁连接的目标地址是相同的,这样的模式称为花束状(Bloom),而 F 此时尚未被感染,其对外通信无花束状模式存在。图 2(2)是结点 F 被感染以后,其通信行为继承了传染源 A 的特点,也呈现出了明显的花束状。蠕虫的传播具有贪婪的天性,一旦成功占领某台主机以后,它会以最快的速度向尽可能多的目标进行扫描扩散。综上所述,所有感染了同一类型蠕虫的主机对外传播方式具有自相似性的特点,即:

* 在短时间内产生大量对外通信连接,通信行为具有"花束状"(Bloom)

* "花束状"通信连接具有水平扫描的特性

* "花束状"通信的目标地址具有离散分布性

根据蠕虫对外的行为特点,本文将其归纳成图 2,它说明了受随机扫描型蠕虫感染的主机对外行为的一般模式,其中目标端口表示为一个序列是因为某些蠕虫的对外行为存在影响多端口的扫描方式。文[11]讨论了蠕虫行为的双模(Dual Mode)和多模(Multimode)特性,一些蠕虫通过扫描远程主机上的多个服务来发现多个可利用的漏洞,以便更有效地传播自身。

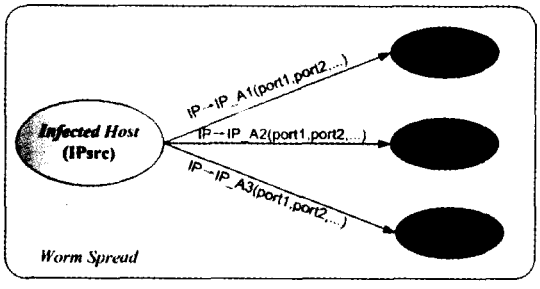


图 2　主动扫描型蠕虫对外传播的一般模式

从传播源方向看,它对远程主机的访问行为可以表达为 $SIP_{infected} \rightarrow DIP_{remote}(dport_1, dport_2, \cdots, dport_n)$,并将其定义为随机扫描型蠕虫的原子行为模式,蠕虫的行为可以看作是这种模式的不断重复。

2.2　访问兴趣度关系(TIR)模型的定义

在实际的网络通信中,会话(Session)的双方之间存在某种访问与被访问的关系,这种关系带有明显的方向性,有时代表了某种攻击的意图,如远程黑客攻击、蠕虫和 DoS 发起的大量攻击性连接。一般来讲,主动发起连接的一方往往包含着某种企图。对源(发起连接者)和目标(响应连接者)之间的连接隐含着的某种关系集合,本文称之为兴趣(Interestedness)。在蠕虫的主动传播中处处体现着这种兴趣关系。图 2 描述的关系模式称为蠕虫访问兴趣度关系(TIR, Threaten Interestedness Relationship)模型。可见该模型中仅包含了 TCP/IP 四元组的基本信息,与负载段的具体内容无关,在 TCP 蠕虫的传播中,每一次会话仅需要记录 SYN 事件,而丢弃所有后续数据包。这在实际建模中可以大大加快速度。

2.3　基于 TIR 模型的攻击树构建

基于上述 TIR 模型的描述,在实际网络监测中,可以按如下方法建立实时攻击树。

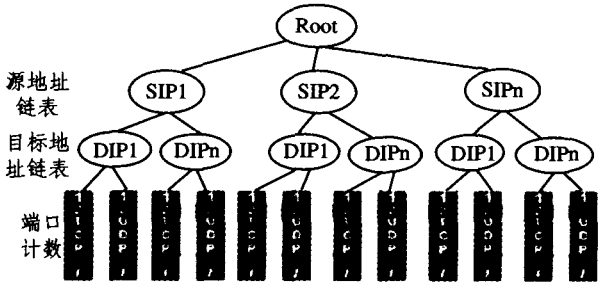


图 3　蠕虫 TIR 攻击关系树结构

最顶端是攻击树的根,监测网络中每到达一个数据包首先判断协议字段,若为 6(TCP),进一步判断 TCP 标志是否为 SYN,若不是则将其丢弃,若是 SYN 则判断该源 SIP 是否在攻击树中,若无则增加节点 SIPx,并将该 DIP 作为该 SIPx 的下级链表结点插入在 DIP 下面的 TCP 链表的插入端口项,并将计数值加 1;若协议字段为 17(UDP),则按上述方法插入相应的 DIP 和 UDP 端口项,并将计数值加 1。这样重复操作,在预定的监测时间片(一般为 5 分钟)结束时,完成攻击信息树的构建,开始后续分析。

3　基于多测度约束的快速蠕虫传播源定位算法

定义 1　Worm 地址频繁度准则:令 C 表示采样周期内所有合法的源 IP 地址集合,$Count_{all}$ 为所有目标 IP 地址 C 被访问次数。对任意 $c_i \in C$,记其某 TCP 和或 UDP 端口被访问次数为 $Count(c_i)^{TDP}_{portid}$ 和 $Count(c_i)^{UDP}_{portid}$,若 $\max\{Count(c_i)^{TCP}_{portid}, Count(c_i)^{UDP}_{portid}\} \geqslant T_{Threshold}$($T_{Threshold}$ 默认取 1000)条件成立,则称 c_i 为活跃地址,$\max\{Count(c_i)^{TCP}_{portid}, Count(c_i)^{UDP}_{portid}\}$ 称为对应的 c_i 频繁度计数。

定义 2　Worm 活动响应度准则:如果蠕虫传输使用 TCP 协议,则令 IP 地址 A 在采样周期内的 Syn 事件累计数为 $Count(Syn)^{tp}_{portod}$,$SynAck$ 的累计数为 $Count(SynAck)^{tp}_{portid}$,定义

$$SymRate=Count(SynAck)^{tcp}_{portid}/Count(Syn)^{tcp}_{portid} \quad (1)$$

为 *SynAck* 对称度，可见 $0<SymRate\leqslant 1$，当 $SynRate=1$ 时称该地址-端口访问是对称平衡的。若有 $SymRate<R_{Threshold}$ 成立，则称地址 A 对外活动响应度低，称 *SymRate* 为其响应度水平。

定义 3 地址离散度准则：令 C' 表示某个 IP 地址在采样周期内对外访问的所有合法的目标 IP 地址的集合，并给定地址离散度衡量函数 D，当 C' 满足 $D(C')\geqslant h_0$ 时，称 C' 中的地址呈离散分布，其中 h_0 为事先设置的离散度阈值。

对于地址离散度的评价，进一步提出下面的基于位图(Bitmap)重量的地址离散度算法。

蠕虫扫描地址离散度算法：令 V 为一面积为 S 的 Bitmap，将 V 初始化为全 0，令 C' 表示某个 IP 地址 *srcHost* 在采样周期里对外访问的所有合法的目标 IP 地址的集合，$\forall c_i\in C'$，取 32 位 IP 地址 c_i 的前 8 位数值内容，令 $Indc_i=c_i\&\&0xFF000000>>24$，将位图数组 V 首部偏移 $Indc_i$ 处的位置为 1。重复上述操作，直至遍历完 C' 集合空间(如图 4)。定义位图 V 的重量为 $Length(V)=\{V$ 中所有非零位的个数$\}$，则 IP 地址空间 C' 的地址离散度定义为 $H(C')=\frac{Length(V)}{S(V)}$，其中 $S(V)$ 为位图 V 的面积。完成对空间中所有的地址的操作后，计算地址离散度 $H(C')$，并与预先设置的阈值 $t(0<t\leqslant 1)$ 进行比较，如果 $H(C')>t$，则将超过阈值的地址 *srcHost* 列入可疑目标清单，一般取阈值 t 为 0.9。实际处理中，为减少计算代价，每次置位后立即判断当前离散度 $H(C')$ 是否已经达到阈值 t，如果已经达到或者超过 t，则结束遍历，跳出循环，进入下一个源 IP 地址结点的判断。

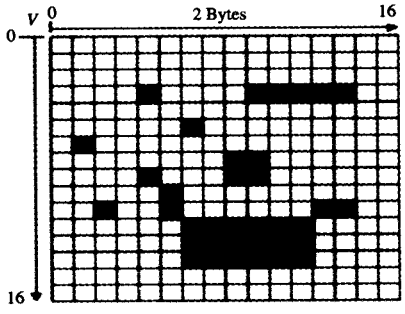


图 4 基于位图重量的地址离散度评价算法

满足上述三个测度约束要求的(srcIP, Prototype, DestPort)可以描述为源地址为 srcIP 的地址感染了基于 Prototype 协议的蠕虫，其影响端口为 DestPort。对于 UDP 协议无响应度约束。

4 数据分析

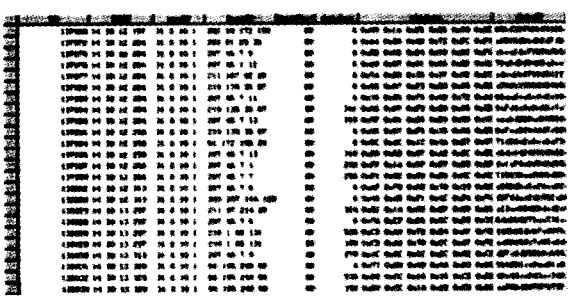


图 5 原始数据

本文的数据取自某政府骨干网，它拥有超过 3,500 台主机。原始数据通过其核心交换机 Cisco Catalyst 6509 的 Spanning 方式采集得到，本文设计的数据包捕获核心模块 Flowcapture 采用 winpcap 库，在完成 TIR 树建模的同时，输出原始流到日志数据库。

表 1 基于多测度约束的快速蠕虫源定位数据分析

IP 地址	目标端口	协议	频繁度水平	响应度水平	离散度水平
31.0.100.1	80	TCP	116,420	11.43%	0.92
31.0.10.1	80	TCP	54,396	9.12%	0.89
31.0.120.1	1434	UDP	261,000	N/A	0.87
31.0.50.1	1434	UDP	245,000	N/A	0.84
31.0.105.2	9136	TCP	65,508	99.4%	0.17
31.0.2.53	80	TCP	71508	96.9%	0.15
…	…	…	…	…	…

(a)频繁度超过阈值 10,000 的候选地址-端口数据

可疑地址	协议	目标端口	包长
31.0.100.1	TCP	80	N/A
31.0.10.1	TCP	80	N/A
31.0.120.1	UDP	1434	300～400
31.0.50.1	UDP	1434	300～400

(b)取离散度水平阈值为 0.6，响应度水平阈值为 50%条件下筛选的最终蠕虫感染列表

表 1(a)是满足频繁度水平条件的所有源 IP 地址和端口协议列表，表 1(b)是经地址离散度水平结合响应度约束进一步筛选后的蠕虫源列表。可见地址 31.0.10.1 和 31.0.100.1 的频繁度水平不如 31.0.105.2 与 31.0.2.53，但通过离散度水平和响应度水平约束加强判断后，筛选算法正确地区分了正常的主机行为和蠕虫行为，将频繁度水平较低的蠕虫行为主机成功筛选出来。从实际情况的数据分析来看，正常的主机访问其目标地址的离散度是比较小的，一般仅为(10～20)%左右。

实际过程中发现传统的链表在高速大容量的建表中效率低下(见图 5 中的 ■)，为提高信息的采集及 TIR 树构建速度，我们提出了基于二级 HASH 页表的高速 TIR 树构建算法[11]，实验表明，传统链表方法消耗时间以指数增长，而改进的二级 HASH 页表算法在 10 万独立地址时最快只要 7.561 秒，完全能满足实时的要求(图 5 中的 ■ 为二级 HASH 页表算法)。

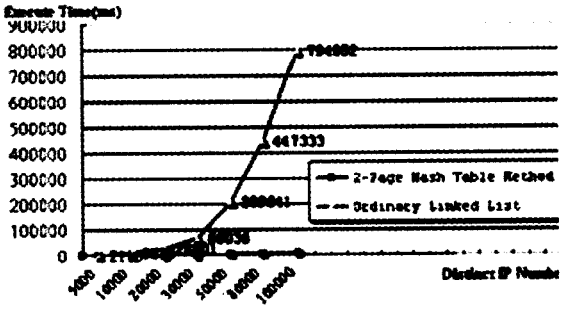


图 5 二级 HASH 页表算法与传统链表方法的效率比较

结论 从当前形势分析，新型蠕虫的检测、定位与防御将是一个长期的过程。本文在现有研究基础上，对蠕虫的一般

传播特点进行了分析，建立了基于 TCP/IP 简单信息的访问兴趣度一般关系模型，在此基础上进一步提出了一种新的基于源地址活跃度、目标地址离散度和响应度准则等多个测度联合约束的快速蠕虫定位算法。其中针对地址离散度评价，又提出基于位图(Bitmap)重量的评价方法。实验表明，基于该约束算法能实现目标网络的蠕虫传播源快速定位。为提高算法在骨干网上的处理速度，研究中提出了基于二级 HASH 页表结构的快速 TIR 树构建算法，实验结果表明该算法极大地提高了信息的处理效率。

参 考 文 献

1 Spafford E H. The Internet Worm: Crisis and Aftermath. Communications of the ACM, 1989, 32(6)678～687
2 The Morris Internet worm. http://snowplow. org/tom/worm/worm. html
3 CERT Advisory CA-2001-19 "Code Red" Worm Exploiting Buffer Overflow In IIS Indexing Service DLL. http://www. cert. org/advisories/CA-2001-19. html
4 Computer worm grounds flights, blocks ATMs - Jan. 26, 2003, CNN. com. http://www. cnn. com/2003/TECH/internet/01/25/internet. attack/
5 Zou C C, Towsley D, Gong W. On the Performance of Internet Worm Scanning Strategies: [Umass ECE Technical Report. TR-03-CSE-07]. Nov. 2003
6 Bakos G, Berk V. Early Detection of Internet Worm Activity by Metering ICMP Destination Unreachable Messages. In: Proc. of the SPIE Aerosense, 2002
7 Maxion R A. Anomaly detection for diagnosis. In: Proc. of the 20th Intl. Symposium Fault-Tolerant Computing(FTCS-20), 1990. 20～27
8 Maxion R A, Feather F E. A case study of Ethernet anomalies in a distributed computing environment. IEEE Transactions on Reliability, 1990, 39(4): 433
9 Huang P, Feldmann A, Willinger W. A non-intrusive, wavelet-based approach to detecting network performance problems. In: Proc. of ACM SIGCOMM Internet Measurement Workshop, San Francisco Bay Area, 2001
10 Lakhina A, Crovella M, Diot C. Diagnosing Network-wide Traffic Anomalies. In: Proc. ACM SIGCOMM, August 2004
11 Gu RongJie, Xia DeLin, Yan PuLiu Yan. An Adaptive Internet Backbone Malicious Activies Detection System Based on Frequent Pattern Mining. [J]GESTS International Transactions on Computer Science and Engineering, 2005, 12: 141～148

(上接第 117 页)

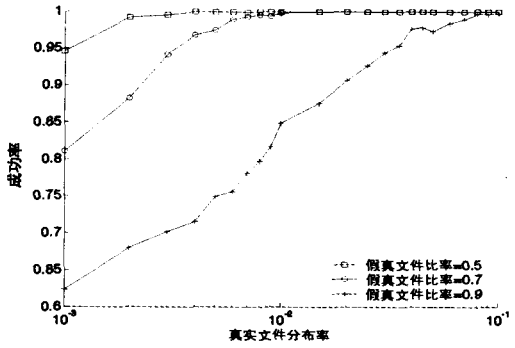


图 4 假真文件比率与认证成功率的关系

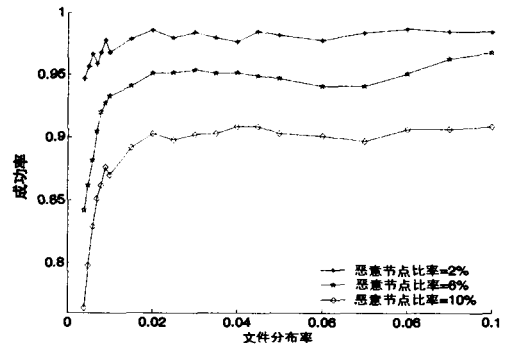


图 5 恶意节点数量与认证成功率的关系

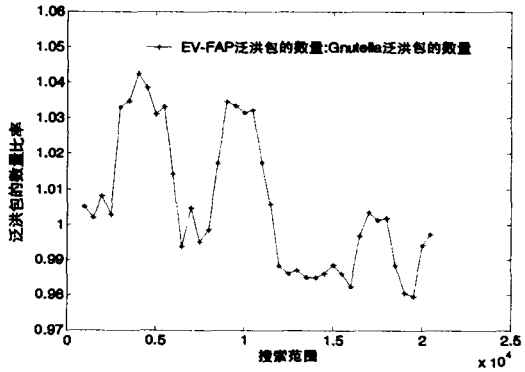


图 6 泛洪包的数量比率

总结 由于匿名性需求，已有的文件真实性认证协议并不能很好地适用于分布式 P2P 环境。本文提出了一种基于电子投票的文件真实性认证协议 EV-FAP，在确定文件真实性的同时实现了协议参与者匿名。我们对协议进行的理论分析表明，EV-FAP 能抵抗网络监听、弃包、恶意投票等攻击。相比已有的文件真实性认证协议，通过实验仿真，我们观察到 EV-FAP 具有协议交互次数少、效率高、认证成功率高等特点。我们相信 EV-FAP 是适合无中心 P2P 网络的文件真实性认证协议。

参 考 文 献

1 Saroiu S, Gummadi K P, Dunn R J, et al. An Analysis of Internet Content Delivery Systems. OSDI '02, 2002
2 Suryanarayana G, Taylor R N. A Survey of Trust Management and Resource Discovery Technologies in Peer-to-Peer Applications: [UCI-ISR Technical Report 04-7]. 2004
3 Daswani N, Garcia-Molina H, Yang B. Open Problems in Data-sharing Peer-to-peer Systems. ICDT2003, 2003
4 Cornelli F, Damiani E, Vimercati S D C D, et al. Choosing Reputable Servents in a P2P Network. In: Proceedings of the 11th World Wide Web Conference, 2002. 441～449
5 Damiani E, Vimercati D C D, Paraboschi S. A Reputation-Based Approach for Choosing Reliable Resources in Peer-to-Peer Networks. In: ACM Conference on Computer and Communications Security 2002, 2002. 207～216
6 Maniatis P, Roussopoulos M, Giuli T, et al. Preserving Peer Replicas By Rate-Limited Sampled Voting. In: Proceedings of the 19th ACM Symposium on Operating Systems Principles, 2003
7 Goldschlag D M, Reed M G F P. Onion Routing for Anonymous and Private Internet Connections. Communications of the ACM, 1999, 42(2): 39～41
8 Xiao L, Xu Z, Zhang X. Low-cost and Reliable Mutual Anonymity Protocols in Peer-to-Peer Networks. IEEE Transactions on Parallel and Distributed Systems, 2003, 14: 829～840
9 冯登国，裴定一. 密码学导引. 北京：科学出版社，1999
10 Gnutella Protocol Specification. www. gnutella. co. uk/library/pdf/gnutella_protocol_0. 4, 2000-07