基于流相似性的两阶段P2P僵尸网络检测方法

牛伟纳[1,2]，张小松[1,2]，孙恩博[2]，杨国武[2]，赵凌园[2]

(1. 电子科技大学网络空间安全研究中心 成都 611731; 2. 电子科技大学计算机科学与工程学院 成都 611731)

【摘要】僵尸网络利用诸如蠕虫、木马以及rootkit等传统恶意程序，进行分布式拒绝服务攻击、发送钓鱼链接、提供恶意服务，已经成为网络安全的主要威胁之一。由于P2P僵尸网络的典型特征是去中心化和分布式，相对于IRC、HTTP等类型的僵尸网络具有更大的检测难度。为了解决这一问题，该文提出了一个具有两阶段的流量分类方法来检测P2P僵尸网络。首先，根据知名端口、DNS查询、流计数和端口判断来过滤网络流量中的非P2P流量；其次基于数据流特征和流相似性来提取会话特征；最后使用基于决策树模型的随机森林算法来检测P2P僵尸网络。使用UNB ISCX僵尸网络数据集对该方法进行验证，实验结果表明，该两阶段检测方法比传统P2P僵尸网络检测方法具有更高的准确率。

关 键 词 僵尸网络检测; 会话特征; 流相似性; P2P流量识别

中图分类号 TP311 **文献标志码** A doi:10.3969/j.issn.1001-0548.2017.06.019

Two Stage P2P Botnet Detection Method Based on Flow Similarity

NIU Wei-na[1,2], ZHANG Xiao-song[1,2], SUN En-bo[2], YANG Guo-wu[2], and ZHAO Ling-yuan[2]

(1. Center for Cyber Security, University of Electronic Science and Technology of China Chengdu 611731;

2. School of Computer Science and Engineering, University of Electronic Science and Technology of China Chengdu 611731)

Abstract The botnet has been one of the most common threats to the network security since it exploits multiple malicious codes like worm, Trojans, Rootkit, etc. toperform thedenial-of-service attack, send phishing links, and provide malicious services. Peer-to-peer (P2P) botnet is more difficult to be detected compared with IRC, HTTP and other types of botnets because it has typical features of the centralization and distribution. To solve these problems, we propose an effective two-stage traffic classification method to detect P2P botnet traffic based on both non-P2P traffic filtering mechanism and machine learning techniques on conversation features. At the first stage, the non-P2P packages are filtered to reduce the amount of network traffic, according to well-known ports, DNS query, and flow counting. At the second stage, the conversation features based on data flow features and flow similarity are extracted. Finally, the P2P botnet is detected by using Random Forest based on the decision tree model. Experimental evaluations on UNB ISCX botnet dataset shows that our two-stage detection method has a higher accuracy than traditional P2P botnet detection methods.

Key words botnet detection; conversation feature; flow similarity; P2P traffic identification

当今时代，网络环境错综复杂，安全问题日益突出。由于僵尸网络的C&C服务器具有更高的隐蔽性，僵尸程序经常被实施大规模网络攻击的黑客所采用，几乎所有的DDoS攻击和80%～90%的垃圾邮件攻击都是由僵尸网络发起的[1]。因此，僵尸网络已成为网络安全中不容忽视的问题。

早期的僵尸网络主要采用IRC[2]和HTTP[3]作为通信协议，具有单点失效问题，很容易被检测和摧毁。如今，大多数僵尸网络使用P2P技术来创建C&C(命令和控制)机制以增强网络通信隐蔽性[4]。相比采用IRC和HTTP协议的僵尸网络，不具有中心节点的P2P僵尸网络具有更大的威胁性和隐蔽性。所以，P2P僵尸网络越来越受到攻击者的青睐，P2P僵尸网络检测[5]也成为安全领域的研究热点。

目前，P2P应用已经引起了互联网流量爆炸式的增长，这对数据存储以及实时分析来讲都是一个巨大的挑战。因此，在检测P2P僵尸网络的早期，对网络中的非P2P流量进行过滤就显得尤为重要。

本文针对P2P僵尸网络提出一种两阶段的检测方法：第一阶段基于端口判断、DNS查询以及会话中数据流计数来过滤非P2P流量；第二阶段基于会话特征来识别P2P僵尸网络，其中本文使用基于会话特

收稿日期：2016－06－28；修回日期：2017－03－09
基金项目：国家自然科学基金(61572115，61502086，61402080)；四川省重大基础研究课题(2016JY0007)
作者简介：牛伟纳(1990－)，女，博士，主要从事网络攻击检测与软件脆弱性方面的研究.

征的检测方法有效降低了需要分析的数据条数。然后采用基于决策树模型的随机森林算法对流量进行分类识别，从而检测出僵尸网络。同时，在UNB数据集上将本文算法与另外两种已有算法做了实验对比和分析，实验结果表明随机森林算法对P2P僵尸网络的检测准确率更高。

1 相关工作

根据检测策略的不同，P2P僵尸网络检测方法包括以下4种类型：基于特征码[6-7]、基于主机行为[8-9]、基于流行为特征[10]和基于流相似性[11]。

1.1 基于特征码的检测

基于特征码的检测[6-7]是通过分析僵尸网络应用程序或者通信流量提取其特征(如MD5、PE头格式等)来设计检测规则。但是最初的检测规则将会在僵尸网络应用程序改变它们的通信方式和数据包格式之后失效。与此同时，如果当前使用的特征码不能有效表示僵尸程序的特征，该检测策略就会有较高的误报率。

1.2 基于主机行为的检测

基于主机行为的检测[8-9]是通过在一个可控环境中监测主机中进程、文件、网络连接、注册表内容的更改来检测僵尸程序。该方法不能检测新型和变种的僵尸网络程序，如攻击者可以使用诸如rootkit、反调试等新的攻击和隐藏技术躲避此种检测策略。

1.3 基于流行为特征的检测

基于流行为特征[10]的检测主要是在僵尸网络C&C控制阶段使用[12]，因为C&C控制阶段的流量与正常的网络流量在流特征与通信规律上存在差异，这些差异包括平均数据包大小、周期性连接等。因此，可以结合机器学习[13]、神经网络[14]对僵尸网络实时监控。

基于流行为特征的僵尸网络检测方法主要分析如下两个特征：连接失败率和流特征。其中，流特征又包括上下行数据包的数量，上下行传输字节的大小，上下行数据包的平均长度、最大长度、平均方差，数据流的持续时间以及在一个流中已加载的数据包的总长度。

这种方法具有较高的检测率，因为它不依赖于僵尸网络的类别来提取流的共同特征向量。所以，该检测策略广受流量分析领域专家学者的关注。在高速、复杂、多变的网络环境中，决定检测效率和准确率的主要因素是提取的特征和使用的分类策略。

1.4 基于流相似性的检测

研究表明[11]，加入同一个僵尸网络的僵尸主机之间的通信行为具有相似性。所以，P2P僵尸网络流量识别可以采用如下方案：首先对获取到的网络流量进行分析处理，并提取特征；然后结合聚类算法对上一阶段提取的流数据进行聚簇；最后分析判断P2P僵尸网络流量位于哪一个簇中。

该方案是通过设置阈值的方式来提高检测准确率，无需使用现有的僵尸网络数据流进行训练。但是，如果当前网络中只有一台僵尸主机，或者在已捕获的数据包中未发现不同僵尸主机的通信流量，此方法也不会有太大效果。

2 两阶段检测方法

本节描述了本文提出的分两个阶段检测僵尸网络流量的方法。该方法的重点在于对非P2P流量的过滤以及对会话的特征的提取，其结构如图1所示。本文将从数据包过滤规则、会话特征和分类算法3方面依次展开介绍。

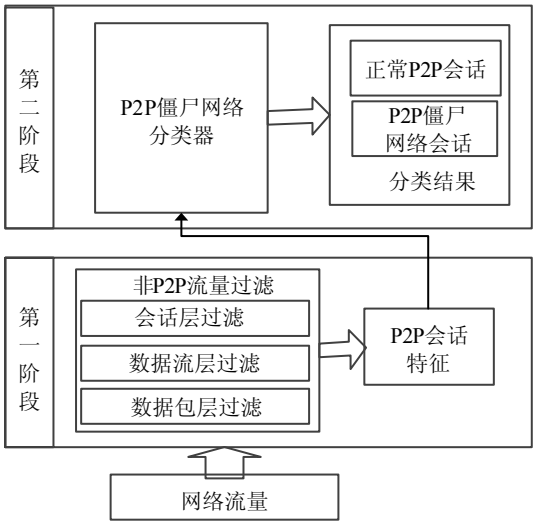

图1　所提出方法的架构

2.1 非P2P流量过滤

目前，常用的P2P流量识别方法包括端口识别[15]、特征码识别[16]以及基于流特征的识别[17]。然而，端口识别方法不能识别出采用随机端口或自定义端口的P2P应用程序，DPI(深度包检测技术)[18]不能识别加密的P2P流量，基于流特征的识别方法只能判断出P2P应用程序的部分流量，且具有很高的误报率。所以，结合快速启发式的P2P协议流量识别方法，本文采用非P2P协议知名端口过滤机制、DNS查询、流计数规则来过滤非P2P流量，如图2所示。

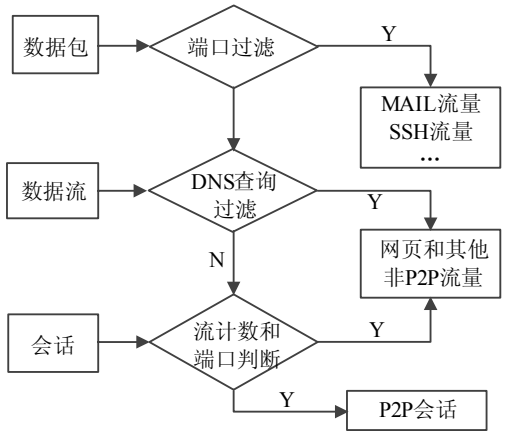

图2 非P2P流量过滤

端口过滤是数据包级别的过滤方法，主要过滤掉常用的非P2P应用程序流量; DNS查询是流级别的过滤方法，流计数和端口判断是会话级的过滤方法，这两种规则主要是过滤掉网页等非P2P流量。其中，基于端口的过滤方法可以识别出一些普通的非P2P应用程序流量，比如SSH一般使用22号端口，Telnet(远程登录)使用23号端口。常用应用程序及其对应端口号如表1所示。

表1 常用应用程序及其对应端口

应用程序	端口号
SSH	22
Telnet	23
MAIL	25, 110, 143, 465, 220, 993, 995
NetBios	125, 137, 139, 445
Remote	3 389
FTP	20, 21
NTP	123

通常，P2P节点通信不需要进行域名解析，而是直接读取存储在本地配置文件中的IPs列表来获取IP。然而，对于非P2P的应用程序必须使用DNS域名解析来获取IP。因此，判断诸如Web、Mail等非P2P网络数据流的依据之一可以是：网络流中的目的IP地址是否是通过域名解析得到的。

用户正常发送Web应用服务请求时，在一个页面中，该Web应用程序会对一个IP地址采用多端口、并行请求的连接方式。因此，会有多条数据流出现在同一个会话中。P2P网络节点每次使用一对随机的源端口和目的端口进行通信。因此，本文可以使用流计数和端口判断来过滤非P2P流量，如果某个会话使用了TCP协议和80、8080或443端口，且会话中的流数量超过了阈值，那么该会话便可以认为是网页流量会话。判断准则为：

$$\begin{cases}\text{sum}(f_i) > \delta_c \\ \text{port} = 808\ 080\ 443\end{cases} \tag{1}$$

式中，$\text{sum}(f_i)$ 表示一个会话中有效的流数量；δ_c 表示阈值，其取值是根据正常网页访问会话中出现的流数量选择的。利用抓包工具分别采集简单和相对复杂的网页请求数据包并对其进行分析，分析结果表明，简单网页一般为3～4个连接请求，而复杂化网页的连接请求是5～8个。所以，本文将阈值设定为3。

尽管这一阶段的方法不能准确地检测出确定的P2P应用程序，但是它可以在真实的网络环境中过滤掉绝大部分非P2P流量和少量安全的P2P流量。

2.2 会话特征提取

通过对P2P僵尸网络的数据流特征分析，加入同一个僵尸网络的僵尸主机之间的流量特征具有相似性。所以，本文使用基于会话的策略进行特征提取，也就是把具有相同目的地址的数据流聚合在相同的会话中，降低了流特征个数和数据条数，进而提高了检测效率。

2.2.1 会话中流持续时间

P2P僵尸主机与其他僵尸主机的通信过程是由僵尸程序自动完成的，流的持续时间一般较短且很固定。因此，可以提取会话中流持续时间的平均值、最大值、最小值和标准差以及会话中上行(下行)流数据包平均间隔时间作为特征。

2.2.2 会话中流的分布

P2P僵尸网络中两个节点进行通信过程中，传输数据包的大小以及传输数量均比较小，且同一个僵尸网络中，僵尸主机产生的C&C通信流具有极大的相似性。因此，可以通过使用会话中流的分布来区分正常的P2P网络通信流与P2P僵尸网络流。提取会话中上行(下行)流最大数据包长度的平均值、平均数据包长度的平均值、最小数据包长度的平均值、平均数据包长度的标准差、有效数据包个数的平均值、有效数据包个数的标准差、传输字节数的平均值、传输字节数的标准差作为特征。

综上所述，将从会话中提取到的这22个特征如表2所示。

表2 会话特征

特征值	特征值的说明
avg_dura	相同会话中不同网络流的持续总时间的均值
std_dura	相同会话中不同网络流的持续总时间的标准差
min_dura	相同会话中不同网络流的持续总时间的最小值
max_dura	相同会话中不同网络流的持续总时间的最大值

(续表)

特征值	特征值的说明
avg_f(b)int	相同会话中不同网络流的上行(下行)数据包传输的平均间隔时间
max_f(b)pl	相同会话中不同网络流的上行(下行)传输数据包长度的最大值的均值
avg_f(b)pl	相同会话中不同网络流的上行(下行)传输数据包长度的均值的均值
min_f(b)pl	相同会话中不同网络流上行(下行)传输数据包长度的最小值的均值
std_avg_f(b)pl	相同会话中不同网络流上行(下行)传输数据包长度的平均值的标准差
avg_f(b)pen	相同会话中不同网络流上行(下行)传输的有效数据包个数的平均值
std_avg_f(b)pen	相同会话中不同网络流上行(下行)传输的有效数据包个数的标准差
avg_f(b)pb	相同会话中不同网络流上行(下行)传输的总字节数的平均值
std_f(b)pb	相同会话中不同网络流上行(下行)传输的总字节数的标准差

2.3　分类器选择

到目前为止，很多有监督的机器学习算法都可以被用来对数据进行分类。本文使用以下3种分类算法来验证所提方法的检测率：贝叶斯网络分类算法、REPTree分类算法以及随机森林分类算法。虽然，这些算法都是基于会话特征来检测P2P僵尸网络流量的，但随机森林算法展现出了较高的准确率。

随机森林算法使用二叉树作为分类树，每个分类树的构建原则是从上到下递归划分的，且其训练集是通过对原始训练数据集有放回采样得到的。为了尽可能降低发生过拟合现象的情况，随机森林使用Bagging随机采样方式构建法分类树。所以，本文使用随机森林分类算法进行高速网络环境中P2P僵尸网络流量的检测。

3　实验结果和分析

3.1　实验设置

本文用于检测僵尸网络的数据集是由ISOT组织，Stratosphere和UNB的信息安全中心公开的僵尸网络流量数据。其中，Stratosphere数据集包括具有端口扫描、C&C通信、攻击等行为的僵尸网络流量；ISOT数据集包括Waledac、Storm、Zuse 3种不同类型的僵尸网络流量和一些背景流量。从UNB信息安全中心所获取的流量包含了许多类型的P2P僵尸网络流量。

根据对上述3家公司所提供的数据集及其流量进行分析，本文选择采用UNB信息安全中心的数据集[19]来做对P2P僵尸网络检测的实验。其中，训练集中只有3种类型的P2P僵尸网络，测试集包含8种不同类型的P2P僵尸网络。

3.2　过滤非P2P流量

流计数和端口判断识别Web流量的识别率与设定的阈值有关，本次试验使用不同的阈值(1,2,3,4,5,6,8,10)来识别Web流量，结果如图3所示。当阈值大于3的时候，随着阈值的增大，检测率急速下降，而误报率基本没有发生变化。但是，当阈值小于3的时候，误报率会随着阈值的减小而增加，而检测率变化不大。因此，根据测试结果，在整个P2P流量的检测过程中将阈值设定为3比较合适。

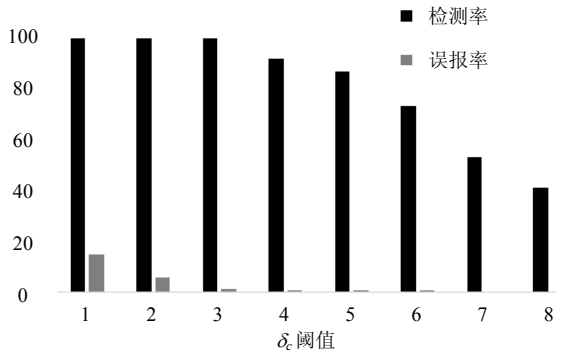

图3　不同阈值下的Web流量识别

3.3　基于会话特征的识别

基于流特征的检测一般使用单一类型的僵尸网络数据集作为实验数据，并采用十指交叉验证策略对分类准确率进行分析。所以，本文采用十指交叉验证的方法来对比分析无过滤的会话特征与过滤后的会话特征的检测效果，僵尸网络的检测率分别为：91%、92.4%，正常流量的误报率分别为：9%、7.6%。结果说明，经过知名端口、DNS、流计数过滤后的会话特征的分类结果不仅比没有过滤的 P2P僵尸网络检测效果好，正常流量的误报率也进一步降低。

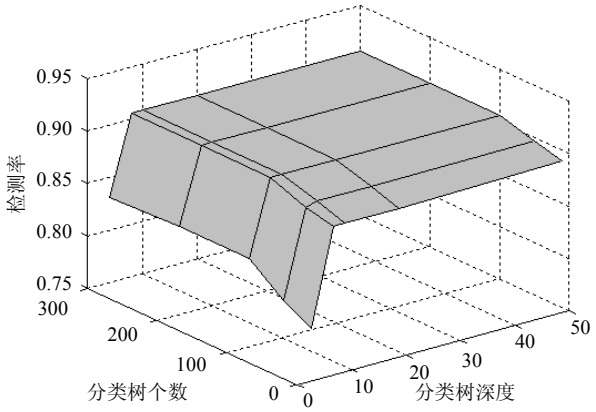

图4　分类树的个数与深度对检测率的影响

一般来说，随机森林算法的分类准确率与分类树个数和分类树深度正相关。然而，随机森林算法的检测速率却与它们呈负相关关系。为了确定用于识别高速网络环境中P2P僵尸网络流量的随机森林模型，本文对不同的分类树个数和分类树深度对检测率的影响做了对比分析，实验结果如图4所示。其中，分类树个数为0～300，分类树深度为0～50。

本文在基于会话特征的基础上，分别采用了贝叶斯网络分类算法、REPTree分类算法和随机森林分类算法来检测P2P僵尸网络流量。REPTree和随机森林算法的深度设定为8，分类树设定为100。P2P僵尸网络流量的检测结果如图5所示。

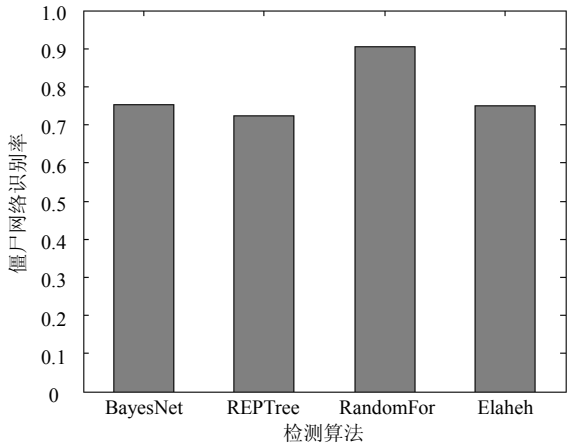

图5 P2P僵尸网络检测结果

使用贝叶斯网络分类算法和REPTree分类算法来检测P2P僵尸网络流量的检测率分别是75.3%和72.4%，但是，使用随机森林算法的检测率高达90.6%。所以，使用随机森林算法对各种类型的P2P僵尸网络流量的检测精度要比使用贝叶斯网络分类算法和REPTree分类算法的检测精度要高。而文献[19]统计的最大检测率只有75%，因此基于会话特征的随机森林检测算法对P2P僵尸网络的检测率有较大提升。

4 结 束 语

本文提出了一种基于会话特征的P2P僵尸网络检测方法，首先分别从包、流和会话层面过滤非P2P流量，然后使用基于会话特征的有监督的机器学习算法检测P2P僵尸网络，该方法同时结合基于流特征的检测方法与基于流相似性的检测方法的优点。最后通过使用公开的数据集验证所提方法的有效性，实验结果表明，该方法能高效地检测P2P僵尸网络流量。

未来将致力于非P2P流量过滤算法的优化，进一步提升其性能。此外，希望将基于会话特征的检测方法推广到其他类型僵尸网络的检测与分类中。

参 考 文 献

[1] ZHU Z, LU G, CHEN Y, et al. Botnet research survey[C]//IEEE International Computer Software and Applications Conference. Turku: IEEE, 2008: 967-972.

[2] LIVADAS C, WALSH R, LAPSLEY D, et al. Using machine learning techniques to identify botnet traffic[C]// 31st IEEE Conference on Local Computer Networks. Tampa: IEEE, 2006: 967-974.

[3] CAI T, ZOU F. Detecting HTTP botnet with clustering network traffic[C]//2012 8th International Conference on Wireless Communications, Networking and Mobile Computing (WiCOM). Shanghai: IEEE, 2012: 1-7.

[4] ZEIDANLOO H R, MANAF A B A, AHMAD R B, et al. A proposed framework for P2P botnet detection[J]. International Journal of Engineering and Technology, 2010, 2(2): 161-168.

[5] HADDADI F, CONG D L, PORTER L, et al. On the effectiveness of different botnet detection approaches[C]// International Conf on Information Security Practice and Experience. Beijing: ACM, 2015:121-135.

[6] WANG J S, LIU F, ZHANG J. Botnet detecting method based on group-signature filter[J]. Journal on Communications, 2010, 31(2): 29-35.

[7] ZHANG J, PERDISCI R, LEE W, et al. Detecting stealthy P2P botnets using statistical traffic fingerprints[J]. Journal of Child Psychology & Psychiatry, 2011, 14(14): 271-282.

[8] ABDULLAH R S, ABDOLLAH M F, NOH Z A M, et al. Preliminary study of host and network-based analysis on P2P botnet detection[C]//TIME-E': International Conference on Technology, Informatics, Management, Engineering & Environment. Bandung: IEEE, 2013: 105-109.

[9] ZHAO Y. The novel approach of P2P botnet node-based detection and applications[J]. Journal of Chemical and Pharmaceutical Research, 2014, 6(7): 1055-1063.

[10] ZHAO D, TRAORE I, SAYED B, et al. Botnet detection based on traffic behavior analysis and flow intervals[J]. Computers & Security, 2013, 39(4): 2-16.

[11] ZHANG J, PERDISCI R, LEE W, et al. Building a scalable system for stealthy P2P-Botnet detection[J]. IEEE Transactions on Information Forensics & Security, 2014, 9(1): 27-38.

[12] SHARIFNYA R, ABADI M. Dfbotkiller: Domain-flux botnet detection based on the history of group activities and failures in DNS traffic[J]. Digital Investigation, 2015, 12(12): 15-26.

[13] BUCZAK A L, GUVEN E. A survey of data mining and machine learning methods for cyber security intrusion detection[J]. IEEE Communications Surveys & Tutorials, 2015, 18(2): 1153-1176.

[14] YIN C, AWLLA A H, YIN Z, et al. Botnet detection based on genetic neural network[J]. International Journal of Security and Its Applications, 2015, 9(11): 97-104.

(下转第948页)

[6] KIM W R, THOMAS B. Evaluation of APRI and FIB-4 scoring systems for noninvasive assessment of hepatic fibrosis in chronic hepatitis B patients[J]. Journal of Hepatology, 2016, 64(4): 773-780.
[7] PINZANI M, VIZZUTTI F, ARENA U, et al. Technology insight: Noninvasive assessment of liver fibrosis by biochemical scores and elastography[J]. Nat Clin Pract Gastroenterol Hepatol, 2008, 5(2): 95-106.
[8] BOOZARI B, POTTHOFF A, MEDERACKE I, et al. Evaluation of sound speed for detection of liver fibrosis[J]. J Ultrasound Med, 2010, 29(11): 1581-1588.
[9]NIGHTINGALE K. Acoustic radiation force impulse (ARFI) imaging: a review[J]. Current Medical Imaging Reviews, 2011, 7(4): 328.
[10] 杨挺青. 粘弹性力学[M]. 第1版. 武汉: 华中理工大学出版社, 1990.
YANG Ting-qing. Viscoelastic mechanics[M]. 1st ed. Wuhan: Huazhong University of Science and Technology Press, 1990.
[11] 何川, 林江莉, 陈科. 超声辐射力弹性成像微小应变检测[J]. 四川大学学报(工程科学版), 2014, 46(s2): 93-98.
HE Chuan, LIN Jiang-li, CHEN Ke. Study on microstrain of ultrasonic radiation elasticity imaging[J]. Journal of Sichuan University (Engineering Science Edition), 2014, 46(s2): 93-98.
[12] HSIEH Y Y, TUNG S Y, LEE I L, et al. FibroQ: an easy and useful noninvasive test for predicting liver fibrosis in patients with chronic viral hepatitis[J]. Chang Gung Med J, 2009, 32(6): 614-622.
[13] CHEN S, FATEMI M, GREENLEAF J F. Quantifying elasticity and viscosity from measurement of shear wave speed dispersion[J]. J Acoust Soc Am, 2004, 115(6): 2781-2785.
[14] CHEN Xin. Quantification of liver viscoelasticity with acoustic radiation force a study of hepatic fibrosis in a rat model[J]. World Federation for Ultrasound in Medicine & Biology, 2013, 39(11): 2091-2102.

编 辑 蒋 晓

(上接第906页)

[15] CONSTANTINOU F, MAVROMMATIS P. Identifying known and unknown peer-to-peer traffic[C]//IEEE International Symposium on Network Computing & Applications. Cambridge: IEEE, 2006:93-102.
[16] MADHUKAR A, WILLIAMSON C. A longitudinal study of P2P traffic classification[C]//MASM' 06: 14th IEEE International Symposium on Modeling, Analysis, and Simulation. Monterey: IEEE, 2006: 179-188.
[17] KARAGIANNIS T, BROIDO A, FALOUTSOS M, et al. Transport layer identification of P2P traffic[C]//ACM SIGCOMM Conference on Internet Measurement. Taormina: ACM, 2004: 121-134.
[18] WANG C, ZHOU X, YOU F, et al. Design of P2P traffic identification based on DPI and DFI[C]//CNMT' 09: Computer Network and Multimedia Technology. Wuhan: IEEE, 2009: 1-4.
[19] BEIGI E B, JAZI H H, STAKHANOVA N, et al. Towards effective feature selection in machine learning-based botnet detection approaches[C]// CNS' 14: 18th IEEE Conference on Communications and Network Security. San Francisco: IEEE, 2014: 247-255.

编 辑 蒋 晓