基于 Markov Chain 的协议异常检测模型*)

李 娜[1] 秦 拯[1] 张大方[1] 陈蜀宇[2]

(湖南大学计算机与通讯学院 长沙410082)[1] (重庆大学计算机学院 重庆400044)[2]

摘 要 本文介绍了基于 Markov 链的协议异常检测模型,此外,通过对 MIT Lincoln 实验室1999评估数据的分析,证明此模型的正确性和有效性。

关键词 入侵检测系统,Markov 链,协议异常

Protocol Anomaly Detection Model Based on Markov Chain

LI Na[1] QIN Zheng[1] ZHANG Da-Fang[1] CHEN Shu-Yu[2]

(College of Computer and Communication, Hunan University, Changsha 410082)[1]

(College of Computer, Chongqing University, Chongqing 400044)[2]

Abstract The model for protocol anomaly detection based on Markov Chain is introduced in this paper. We also prove our model's correctness and effectiveness by analyzing MIT Lincoln laboratory 1999 evaluation data set.

Keywords Intrusion detection system, Markov chain, Protocol anomaly

1 引言

对入侵检测(Intrusion Detection)的研究工作开始于20世纪80年代,目前已经成为计算机安全领域的研究热点。入侵检测系统在加强系统安全性方面有着不可忽视的作用,已经成为系统安全的第二道防线。

入侵检测系统(IDS)分为误用检测(Misuse Detection)和异常检测(Anomaly Detection)。误用检测是目前很多 IDS 使用的方法,如著名的 Snort[7]和 Paxson 的 Bro[8]。它们都是将已知攻击的特征提取出来,建立一个攻击特征数据库,然后对那些符合攻击特征的行为进行报警,但缺点是高漏警率(false negative rate),即只能对已知攻击进行检测,对那些变种的和新的攻击(unknown attack)却无能为力,并且检测率的提高依赖于攻击特征数据库的不断更新。异常检测是对正常状态下的系统行为建立模型,它依赖于用哪种方法对哪些正常行为进行建模,通过观测一组测量值偏离度进行决策。采用这种方法能检测到一些未知的攻击,做到防患于未然,但这势必会引起高误警率(false positive rate)的问题,过多的报警会给 IDS 和管理员带来沉重的负担,其误警率的降低通常依赖于统计分析方法的改进。

协议异常检测是 IDS 异常检测的一个新技术[1,5,6],具有一定的研究价值。Juan M. Estevez-Tapiador 在文[1]中提出了一种基于 Markov Chain 的随机协议异常检测模型,他们首先对 TCP 首部的六位标志字段进行量化,即对每个正常的 TCP 数据包用一个唯一确定的值加以标识,因此对一系列 TCP 数据包,就可以得到一系列离散值的序列,然后通过 Markov chain 对这些序列进行建模,从而得到正常状态下的协议模型,任何偏离此模型的序列都被认为是入侵行为。本文对文[1]所建的协议模型进行了改进,改进后的模型具有更好的完备性和精确性,能够有效地检测到协议异常,并且提出了一个基于马尔可夫链的协议异常检测系统框架。

本文首先介绍马尔可夫链(Markov Chain)的基本知识,然后在第3部分说明如何用马尔可夫链对应用层各协议进行建模,并且提出了一个基于马尔可夫链的协议异常检测系统框架。第4部分给出实验结果。第5部分对实验进行扩展,在更高一层即传输层对 TCP 协议进行建模。最后给出本文的结论。

2 马尔可夫链

2.1 马尔可夫过程假设

马尔可夫过程是一个具有无后效性的随机过程,所谓无后效性是指:当随机过程在时刻 t_0 的状态已知的条件下,随机过程在时刻 $t(t>t_0)$ 所处的状态仅与时刻 t_0 的状态有关,而与过程在时刻 t_0 以前的状态无关。那些时间离散、状态离散的马尔可夫过程称为马尔可夫链,简称马氏链[13]。

2.2 马尔可夫链

定义为:设随机变量序列 $\{X(n),n=1,2,\cdots\}$ 的状态空间 I 是离散的,过程 $X_n=X(t_n)$ 所处的状态为 $a_1,a_2,\cdots,a_N$ 之一,若在任意一个时刻 n,以及任意状态 $a_1,a_2,\cdots,a_i,a_j$,下面的条件概率公式成立

$$P\{X_n=a_j|X_{n-1}=a_i,\cdots,X_2=a_2,X_1=a_1\}=P\{X_n=a_j|X_n-1=a_i\} \quad (1)$$

则称 $\{X(n),n=1,2,\cdots\}$ 为一个马氏链。现在进一步假设 $X_n=a_j$ 的状态概率为:

$$P_j(n)=P\{X_n=a_j\} \quad (2)$$

在 $X_i=a_i$ 的条件下,$X_n=a_j$ 的条件概率或转移概率为:

$$P_{ij}(s,n)=P\{X_n=a_j|X_i=a_i\} \quad (3)$$

显然,根据全概率公式,有:

*)本课题得到国家863项目(No. 2003AA118201)和国家自然科学(No. 60273070)资助。李 娜 硕士生,研究方向为网络安全。秦 拯 博士,研究方向为计算机网络安全与通信。张大方 教授,博士生导师,研究方向为可信系统与网络。陈蜀宇 教授,博士生导生,研究方向为计算机网络安全与体系结构。

$$P_j(n)=\sum_{i=1}^{N}P\{X_n=a_j,X_s=a_i\}$$
$$=\sum_{i=1}^{N}P\{X_n=a_j|X_s=a_i\}P\{X_s=a_i\}$$
$$=\sum_{i=1}^{N}p_{ij}(s,n)P_i(s) \quad (4)$$

由转移概率构成的矩阵为：

$$P(s,n)=\begin{bmatrix}P_{11}(s,n) & P_{12}(s,n) & \cdots & P_{1N}(s,n)\\ P_{21}(s,n) & P_{22}(s,n) & \cdots & P_{2N}(s,n)\\ \cdots & \cdots & \cdots & \cdots\\ P_{N1}(s,n) & P_{N2}(s,n) & \cdots & P_{NN}(s,n)\end{bmatrix}$$

此矩阵称为马尔可夫转移矩阵。

3 协议建模

对异常检测而言，需要对正常状态下的系统行为(host-based)或网络数据(network-based)进行建模，很多文章已经对正常状态下的系统调用建立了模型，通过分析用户的系统调用序列来判断该调用序列是否异常[2,4,12]。

本文对文[1]进行了改进，文[1]中所采用的正常状态下的TCP流量只包含有限的服务(SSH,FTP,HTTP)，并且样本空间太小(不足一天的网络流量)，统计出来的结果必然和实际情况有偏差。基于此问题，本文采用MIT Lincoln实验室DARPA 1999入侵检测系统评估数据(以下简称为DARPA 1999)，此数据含有两个星期(第一周和第三周)正常状态下的网络流量，一个星期(第二周)含有攻击的网络流量[9]。本实验首先分析正常状态下的网络流量，通过Markov Chain建立模型，发现建立的模型和文[1]不同，但我们相信我们所建立的模型更具说服力，并且相信改进后的协议模型更能反映正常状态下实际网络中的协议使用情况。第5部分我们进一步分析了对TCP建立模型的可行性。

3.1 标识化过程

为了对每一个TCP数据包用一个唯一的值进行标识，我们通过TCP首部的标志比特加以标识，见图1，比如说一个TCP数据包的标志比特为000010(SYN位置1)，我们就将这六个比特所组成的二进制数转化为所对应的十进制数2，然后用2对这个数据包进行标识，又如下一个数据包的标志比特位为010010(SYN和ACK均置为1)，对应的十进制数为18，我们就用18来标识这个数据包，它表示对上一个SYN进行确认。因此每个TCP数据包都可以用一个唯一的十进制数表示，这个十进制数的范围为$0\sim2^6-1$，但并不是在此范围的数都合法[10]。文[1]中将TCP首部的标识字段进行了调整，见图2，因此计算出的十进制数会和我们的不同，由于该十进制数只是对TCP数据包进行标识，它并不影响协议模型的建立。

URG	ACK	PSH	RST	SYN	FIN

图1 TCP首部的标志字段

FIN	URG	RST	PSH	ACK	SYN

图2 文[1]中调整过的TCP首部的标志字段

3.2 训练建模过程

该过程仍然采用文[1]中的训练建模过程，见图3。

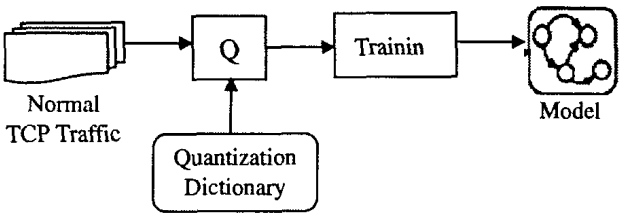

图3 马尔可夫链的训练建模过程

在异常检测中，关键问题是如何对正常状态下的行为建立模型。基于此问题，本文采用正常状态下3.1节所提到的标识序列进行建模。首先用tcpdump[11]对DARPA 1999的评估数据(第一周和第三周)进行协议分离(通过端口号)，然后分别对应用层协议(FTP,SSH,HTTP,TELNET等)建立模型，本文只列出FTP(端口21)的模型，见图4，文[1]中所建立的FTP模型见图5，图中每个圆圈代表状态，圆圈中的数字代表3.1节中六位二进制数所对应的十进制数，圆圈外的数字代表状态空间I中各个状态所占的比例，箭头上的数字代表状态之间的转换概率。注意，图5中状态的标识和图4中状态的标识有所不同，原因见3.1节。可以看到，文[1]中对FTP协议所建立的模型缺少必要的状态以及状态之间的转换，并且各个状态之间的转换概率和我们的结果有所偏差，由于我们是根据DARPA 1999两个星期的数据进行统计的，样本空间很大，因而更具有说服力。

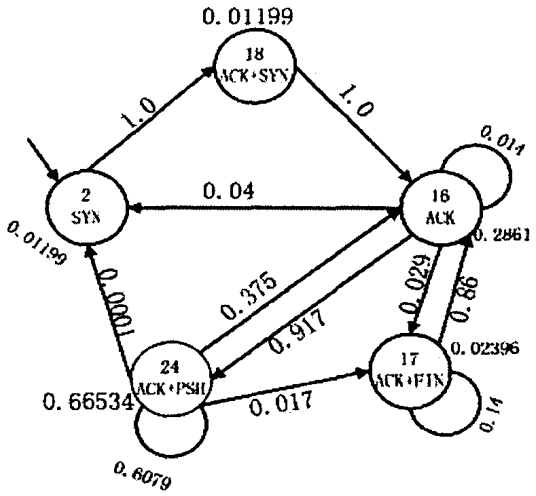

图4 FTP的状态转换图

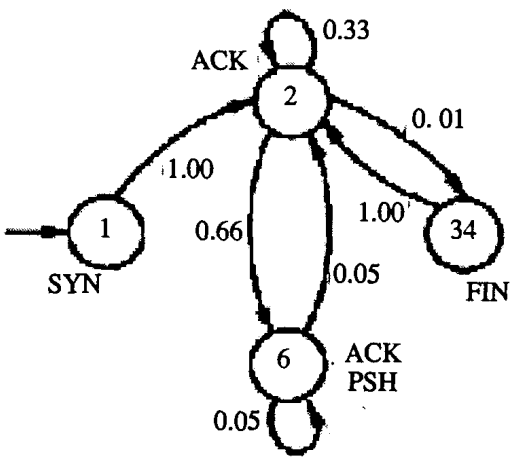

图5 文[1]中所建立的FTP状态转换图

3.3 检测过程

这里，我们提出了一个基于马尔可夫链的协议异常检测

系统框架，见图6。首先，我们对流进主机的 TCP 流量根据不同应用层协议所对应的端口号进行分离，然后将这些数据分别送入3.2节所建立的应用层协议模型中进行检测，得到的结果由一个称为分类器(Classifier)[3]的模块进行分类，将攻击和非攻击的流量进行区分，对含有攻击的流量产生相应的报警。

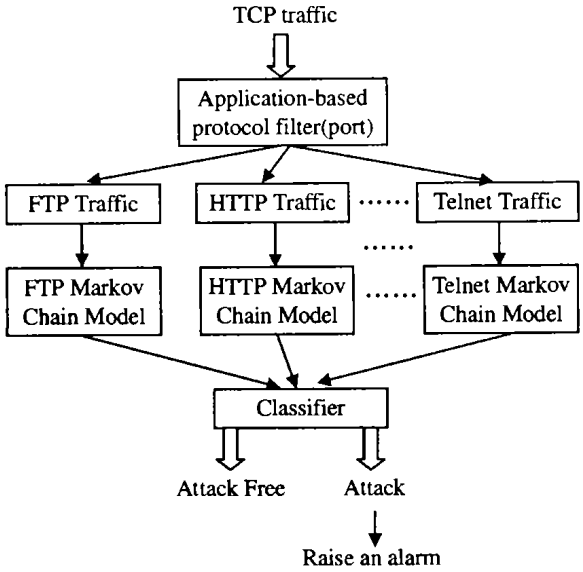

图6　检测过程

4　实验结果

为了验证实验结果的正确性，我们采用了文[1]中的马尔可夫链评估公式。

$$\log MAP(t)=\log(\pi_{o_t})+\sum_{t}^{T-1}\log(a_{o_t o_{t+1}}) \tag{5}$$

$$MAP(t)=\pi_{o_t}\cdot\prod_{t}^{T-1}\log(a_{o_t o_{t+1}}) \tag{6}$$

$$\log MAP(t)=\log(\pi_{o_t})+\sum_{t}^{T-1}\log(a_{o_t o_{t+1}}) \tag{7}$$

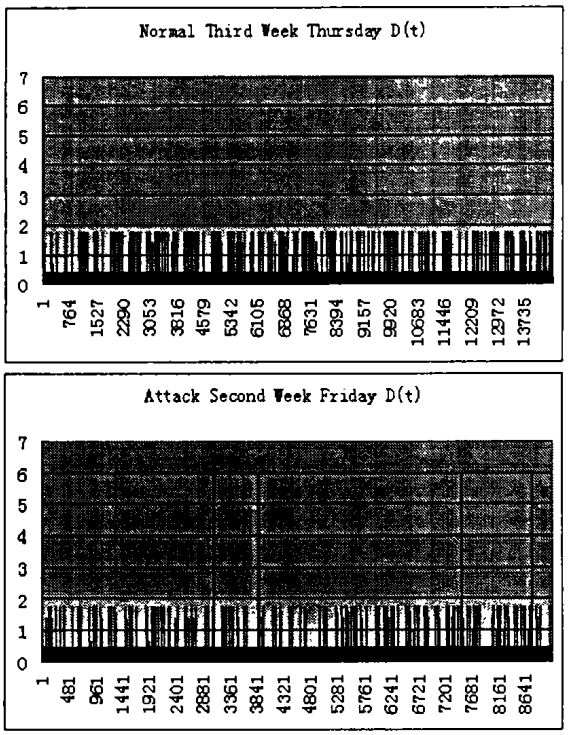

图7　实验结果(采用图4的 FTP 模型)

这里，$A=[a_{ij}]$表示状态转移矩阵，$\pi=(p_i)$表示图4中每个状态的初始概率，$O=\{o_1,o_2,\cdots,o_T\}$表示 FTP 流量所对应的标识序列。本实验是在 Linux c 环境下实现的。图7仅提供了正常状态下第三周星期四和含有攻击的第二周星期五 FTP 流量的实验结果(其它天的结果与此类似)，图中横轴代表公式(7)中的 t，纵轴表示 $D(t)$的值，采用的模型见图4，从图7中可以清楚地看到正常状态下和异常状态下的区别，正常状态下所有的 $D(t)$的值都低于2，而发生协议异常时有的 $D(t)$值达到了6，因而可以有效地检测协议异常。对于其它应用层协议同理可以建立类似图4的模型。

5　TCP 的马尔可夫链

在第3节中本文对应用层诸协议建立了模型，那么是不是可以对 TCP 协议建立模型呢?答案是肯定的。为了验证假设的正确性，我们同理为 TCP 建立类似图4的模型。实验结果见图8(仍采用4节所提到的 FTP 流量)，可以清楚地看出其 $D(t)$的值较之图7有所变化，这是因为所建立的 TCP 模型是针对基于不同服务的协议的综合流量建立的，因此所得的状态转移概率和单个应用层协议所对应的转移概率不同，这就引起 $D(t)$值的不同。我们认为对单个应用层协议建立的模型更能清楚地反映异常和正常之间的差别。如果要用所建立的 TCP 模型进行检测，这也是可行的，但它依赖于阈值的选取，大于这个阈值的就被认为是攻击，本实验选取2.6作为阈值，就可以检测到 FTP 的协议异常，见图8中的箭头所示。同理，使用这个 TCP 模型还可以检测到其它应用层协议的异常，但是需要为不同的协议采用不同的阈值。

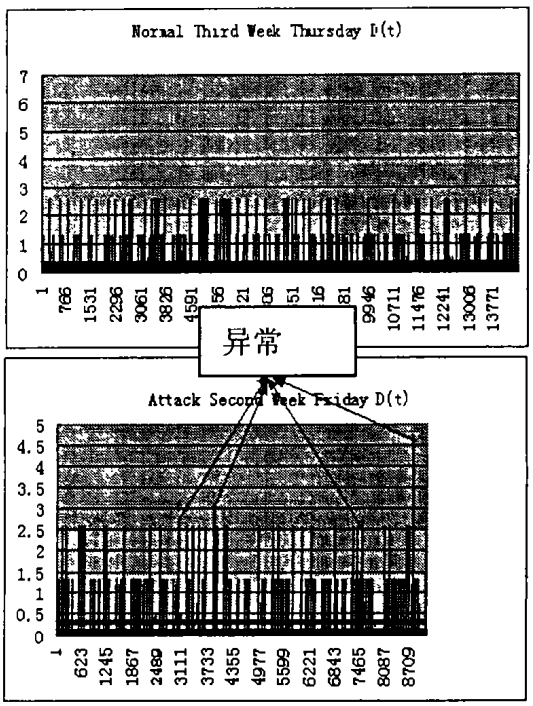

图8　采用 TCP 协议模型的实验结果

结论　根据所得的实验结果，本文中所建立的基于马尔可夫链的网络协议模型能够有效地检测到协议异常。此外，本文还提出了一个基于马尔可夫链的协议异常检测系统框架。今后的工作是如何确定 $D(t)$的阈值来降低误警率。

参 考 文 献

1　Estevez-Tapiador J M, et al. Stochastic Protocol Modeling for

(下转第95页)

计算结果存放在文本文件中。根据文[4,5]的思想，尽我们最大的能力实现了算法 MinCube[4]和 DFS[5]。

weather 数据集被多个算法所采用[4~7,9]，它有1,015,367个元组(大约 27.1MB)，9个维：*station-id*(7037)、*longitude*(352)、*solar-altitude*(179)、*latitude*(152)、*present-weather*(101)、*day*(30)、*weather-change-code*(10)、*hour*(8)和 *brightness*(2)，括号中的数字是基数。我们通过投影的方法从 *weather* 数据集中产生了8个数据集，它们的维数是2、3、…、9，即前2个、3个、…、9个维的投影。去掉相同的元组后，8个数据集中的元组个数分别是7037、298215、298215、540769、1004515、1005318、1015367和1015367。

实验比较了算法 TCUBE 和 MinCube，TCUBE 和 DFS 在生成 condensed cube 和 quotient cube 所需要的时间(秒)，结果如表1和表2所示。实验结果表明，TCUBE 计算速度要快于 MinCube 和 DFS。主要原因在于 MinCube 需要输出大量的 Bitmap 索引，并且随着维数和基本关系中元组个数的增多，索引的个数和索引占用的空间也增多；而 DFS 会产生大量相同的 upper bound，需要通过排序来去除重复。

表1 TCUBE 与 MinCube 的执行时间

Dims	TCUBE	MinCube
2	9	0
3	7	17
4	4	36
5	15	197
6	63	917
7	141	1570
8	256	4271
9	600	15428

表2 TCUBE 与 DFS 的执行时间

Dims	TCUBE	DFS
2	6	8
3	9	14
4	5	16
5	9	44
6	33	199
7	52	595
8	94	1014
9	161	1420

结束语 研究工作者提出了许多类型的数据立方体，需要不同的存储空间，查询响应时间也各不相同，给用户提供了丰富的选择，不同的数据立方体需要不同的生成算法。本文分析了可以使用关系系统作为存储结构的一般数据立方体、部分数据立方体和浓缩立方体的特点，发现可以用合作伙伴的概念来统一描述这些立方体，并设计了算法 TCUBE 用于生成这三类数据立方体。我们还通过实验验证了 TCUBE 的性能，结果表明 TCUBE 生成浓缩数据立方体的速度要快于原有的算法。

参 考 文 献

1 Gray J, Bosworth A, Layman A, et al. Data Cube: A Relational Aggregation Operator Generalizing Group-By, Cross-Tab, and Sub-Totals. In: S. Y. W. Su, ed. Proc. of the Twelfth Intl. Conf. on Data Engineering. New Orleans: IEEE Computer Society, 1996. 152～159

2 Harinarayan V, Rajaraman A, Ullman J D. Implementing Data Cubes Efficiently. In: H. V. Jagadish, Inderpal Singh Mumick, eds. Proc. ACM SIGMOD Intl. Conf. on Management of Data. Montreal: ACM Press, 1996. 205～216

3 Shukla A, Deshpande P M, Naughton J F. Materialized View Selection for Multidimensional Datasets. In: A. Gupta, O. Shmueli, J. Widom, eds. Proc. of 24rd Intl. Conf. on Very Large Data Base. New York City: Morgan Kaufmann, 1998. 488～499

4 Wang W, Lu H, Feng J, Yu J X. Condensed Cube: An Effective Approach to Reducing Data Cube Size. In: Proc. of the 18th Intl. Conf. on Data Engineering. San Jose: IEEE Computer Society, 2002. 155～165

5 Lakshmanan L V S, Pei Jian, Han Jiawei. Quotient Cube: How to Summarize the Semantics of a Data Cube. In: Proc. of 23rd Intl. Conf. on Very Large Data Bases. Hong Kong: Morgan Kaufmann, 2002. 778～789

6 Lakshmanan L V S, Pei Jian, Zhao Yan. QC-Trees: An Efficient Summary Structure for Semantic OLAP. In: Alon Y. Halevy, Zachary G. Ives, AnHai Doan, eds. Proc. ACM SIGMOD Intl. Conf. on Management of Data, San Diego, California, USA: ACM Press, 2003. 64～75

7 Sismanis Y, Deligiannakis A, Roussopoulos N, Kotidis Y. Dwarf: Shrinking the PetaCube. In: M. J. Franklin, B. Moon, A. Ailamaki, eds. Proc. ACM SIGMOD Intl. Conf. on Management of Data. Madison: ACM Press, 2002. 464～475

8 Barbara D, Sullivan M. Quasi-cubes: Exploiting approximation in Multidimensional Databases. SIGMOD Record, 1997, 26: 12～17

9 Beyer K, Ramakrishnan R. Bottom-Up Computation of Sparse and Iceberg CUBEs. In: A. Delis, C. Faloutsos, S. Ghandeharizadeh, eds. Proc. ACM SIGMOD Intl. Conf. on Management of Data. Philadelphia: ACM Press, 1999. 359～370

10 Zhao Y, Deshpande P M, Naughton J F. An Array-Based Algorithm for Simultaneous Multidimensional. In: Joan Peckham, ed. Proc. ACM SIGMOD Intl. Conf. on Management of Data. Tucson: ACM Press, 1997. 159～170

11 Hahn C, Warren S, London J. Edited synoptic cloud reports from ships and land stations over the globe. http://cdiac.esd.ornl.gov/cdiac/ndps/ndp026b.html

(上接第68页)

Anomaly Based Network Intrusion Detection. In: Proc. of the First IEEE Intl. Workshop on Information Assurance(IWIA'2003)

2 Warrender C, Forrest S, Pearlmutter B. Detecting Intrusions Using System Calls: Alternative Data Models. In: IEEE Symposium on Security and Privacy, 1999

3 Jha S, et al. Classifiers, and Intrusion Detection. In: 14th IEEE Computer Security Foundations Workshop(CSFW'01), June 2001

4 Gao B, et al (HMMS (HIDDEN MARKOV CHAIN MODELS) BASED ON ANOMALY INTRUSION DETECTION METHOD. In: Proc. of the First Conf. on Machine Learning and Cybernetics, Beijing, Nov. 2002. 381～385

5 Das K. Protocol Anomaly Detection for Network-based Intrusion Detection. http://www.sans.org/rr/papers/30/349.pdf

6 Bykova M, Ostermann S, Tjaden B. Detecting Network Intrusions via Statistical Analysis of Network Packet Characteristics. In: Proc. of the 33rd southeastern Symposium on System Theory, 2001

7 http://www.snort.org/

8 Paxson V. Bro: A System for Detecting Network Intruders in Real-Time. In: Proc. of the 7th USENIX Security Symposium San Antonio, Texas, Jan. 1998

9 http://www.ll.mit.edu/IST/ideval/data/data_index.html

10 Postel J. Transmission Control Protocol. RFC 793, Sep. 1981

11 http://www.tcpdump.org/

12 谭小彬，王卫平，奚宏生，殷保. 系统调用序列的 Markov 模型及其在异常检测中的应用. 计算机工程，2002，12：189～191

13 罗鹏飞，张文明，刘福声. 随机信号分析. 国防科技大学出版社，2000. 176～183