基于会话流聚合的隐蔽性通信行为检测方法

陈兴蜀[1,2]，陈敬涵[3]，邵国林[3]，曾雪梅[2*]

(1. 四川大学网络空间安全学院 成都 610065；2. 四川大学网络空间安全研究院 成都 610065；3. 四川大学计算机学院 成都 610065)

【摘要】采用隐蔽技术对抗安全检测并实现长期潜伏与信息窃取的网络攻击已成为当前网络的重大安全问题。目前该领域面临3个难题: 1) 攻击本身的强隐蔽性使其难以检测; 2) 高速网络环境中的海量通信数据使检测模型难以细粒度构建; 3) 隐蔽通信的持续性和复杂性使标签数据缺乏进而加大了模型的构建难度。针对上述3个问题，该文在对长时间的校园网流量进行大数据统计分析的基础上，对基于隐蔽会话的隐蔽性通信行为进行了描述和研究，提出了一种隐蔽性通信行为检测方法。该方法首先通过并行化会话流聚合算法聚合原始会话流，然后从集中趋势和离散程度的角度刻画隐蔽通信行为，并引入标签传播算法扩展标签数据，最后构建多分类检测模型。通过仿真和真实网络环境下的实验，验证了方法对隐蔽性通信行为的检测效果。

关 键 词 大数据分析； 隐蔽性通信； 网络行为； 会话流聚合

中图分类号 TP393 **文献标志码** A doi:10.3969/j.issn.1001-0548.2019.03.013

A Covert Communication Behavior Detection Method Based on Session Flow Aggregation

CHEN Xing-shu[1,2], CHEN Jing-han[3], SHAO Guo-lin[3], and ZENG Xue-mei[2*]

(1. College of Cyber Security, Sichuan University Chengdu 610065; 2. Cyber Security Research Institute, Sichuan University Chengdu 610065; 3. College of Computer Science, Sichuan University Chengdu 610065)

Abstract Network attacks that employ covert techniques to against security detections and achieve long-term latency and information theft have become major security issues in the current network. There are currently three challenges in this field. The strong concealment of the attack makes it difficult to detect, massive communication data in a high-speed network environment makes it difficult to build a detection model in a fine-grained manner, and the persistence and complexity of covert communication make the lack of tag data and increase the difficulty of model construction. Aiming at the above problems, based on the statistical analysis of campus network traffic, this paper describes and studies the hidden communication behavior based on covert conversation, and proposes a hidden communication behavior detection method. The original session flow is aggregated by parallelized session flow aggregation algorithm, and the covert communication behavior is characterized from the perspective of concentration trend and dispersion degree. The tag propagation algorithm is introduced to extend the tag data, and finally the multi-class detection model is constructed. The simulation results and the experiments in real network environment verify the detection effect of the method on the hidden communication behavior.

Key words big data analysis; covert communication; network behavior; session flow aggregation

采用高级隐蔽技术以对抗不断增强的安全威胁检测并实现长期潜伏与信息窃取的网络攻击方式已成为网络信息安全领域的最大威胁之一[1]。APT、远控木马、僵尸网络等均大量采用隐蔽通信手段，以达到网络保持、敏感信息窃取以及关键数据破坏等目的[2-4]。因此准确有效地检测隐蔽通信行为从而防范此类网络攻击一直是信息安全领域的重要问题。

与此同时，基于行为分析的网络安全防护一直是学术界和产业界研究的热点。文献[5]基于流量结构稳定性对网络行为进行建模，检测模型对常见网络攻击及未知网络流量异常均具有有效性。文献[6]分析了不同应用所表现出的网络行为，可有效监测

收稿日期：2018 – 02 – 11；修回日期：2018 – 06 – 19

基金项目：国家自然科学基金(61272447)；国家“双创”示范基地之变革性技术国际研发转化平台(C700011)；四川省重点研发项目(2018G20100)；四川省科技支撑计划(2016GZ0038)；中央高校基本科研业务费专项(2017SCU11059,2017SCU11065,SCU2016D009)

作者简介：陈兴蜀(1968 –)，女，博士，教授，主要从事云计算/大数据安全与网络危胁检测等方面的研究.

通信作者：曾雪梅，E-mail: zengxm@scu.edu.cn

云计算环境下已知和未知的网络异常行为。文献[7]通过时间、空间和网络协议3个层面数据关联，描述事件网络行为，通过行为特征构建的分类器可有效识别恶意网络事件。文献[8]提出一种基于TDRI的DNS流量分析模型，可帮助发现DNS流量中表现出的恶意行为。但上述研究所提出的方法对基于隐蔽会话交互行为的隐蔽通信缺乏检测能力，无法应对具有隐蔽性的网络威胁。然而，当前对于隐蔽性通信行为的检测存在极大挑战：

1) 基于隐蔽会话交互行为的安全威胁力求在不引起察觉的情况下进行隐蔽性通信[9]。相较而言，对具有隐蔽性和伪装性的通信行为检测，传统安全防护技术基本无能为力。如何从隐蔽会话交互行为和持续通信的角度描述和发现隐蔽通信行为成为了亟待解决的问题。

2) 网络环境日趋复杂，基于隐蔽会话的隐蔽性通信行为对内网安全有着极大威胁[10]，现有基于数据包的隐蔽性恶意程序检测方法[11-12]在通信量巨大的环境下缺乏应对能力，文献[13]提出的包级检测方法的有效性随数据量增大而降低。大数据分析技术给网络安全分析带来了新的机遇[14]，但海量的通信数据给也给隐蔽通信模型的细粒度构建带来了巨大挑战。

3) 隐蔽通信相较于其他在短时间内引起明显流量特征变化的网络攻击[15]，是一个隐蔽、持续且长期的过程。其通信情况复杂多样，通信样本难以获取，现有数据缺乏准确标注，都加大了隐蔽通信检测模型构建的难度。因此需要一种行之有效的方法扩展隐蔽通信样本中的标签数据。

针对上述3方面的问题，本文对四川大学校园网网络流量进行基于大数据的统计分析，并以木马隐蔽会话交互过程为例对隐蔽性通信行为进行了特征分析。以此为研究对象，提出一种并行化会话流聚合方法。在此基础上多角度提取隐蔽通信行为特征，对其准确刻画，引入标签传播算法扩展标签数据。最后通过支持向量机(support vector machine, SVM)构建检测模型，实现对真实网络环境中隐蔽性通信行为的检测。本文主要工作及贡献有：

1) 对真实环境中会话流和隐蔽会话交互行为进行分析，明确了海量数据下区分难度极大的问题，发现了隐蔽性通信与正常通信之间的细小差异，及隐蔽通信会话流本身的相似性。在此基础上提出一种并行化流量聚合算法，通过聚合会话流，使海量数据下隐蔽通信模型的细粒度构建成为可能；

2) 借助描述集中趋势和离散程度的统计量，多角度刻画聚合后会话流特征，并提出会话流互异度的计算方法，综合评估聚合后流量间相异程度。实现大数据环境下细粒度模型构建，从而在海量数据环境下检测隐蔽通信行为；

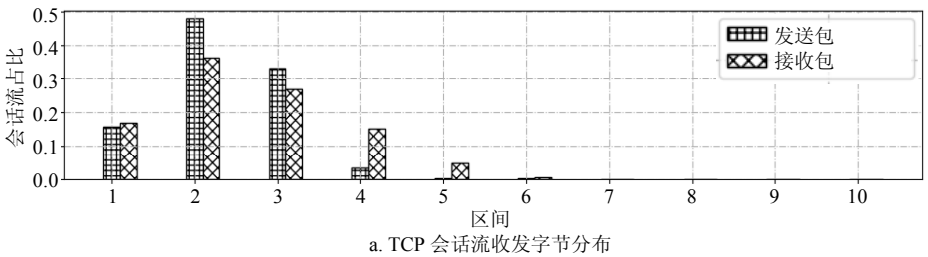
3) 将在音频、视频及图像等领域中有较强实用性的标签传播算法[16-17]引入到标签数据的扩展，实现对隐蔽性通信行为不同阶段数据的自动化标注。

1　流量研究及特征分析

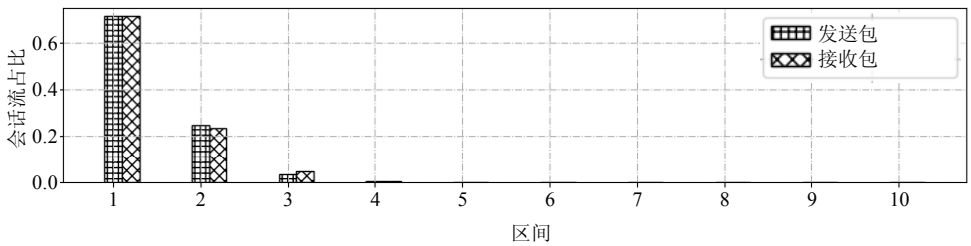
本文分别针对四川大学校园网实际流量和不同隐蔽性恶意程序的通信行为进行了长时间的观测和分析。对隐蔽通信行为进行了刻画，研究了基于隐蔽会话交互的隐蔽通信流量与正常网流量的异同。

1.1　基于统计的真实环境流量分布分析

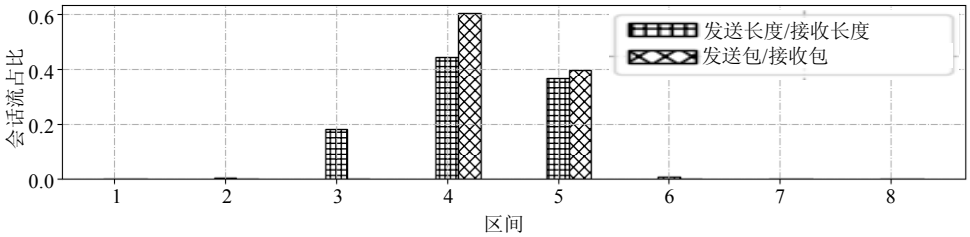
本文首先通过基于统计的大数据分析方法对四



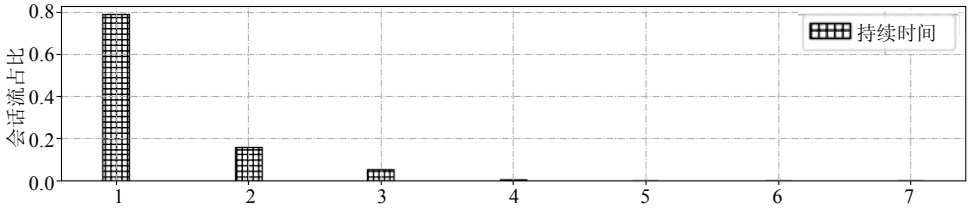
a. TCP 会话流收发字节分布



b. TCP 会话流收发包分布



c. TCP 会话流发收字节比和发收包数比分布



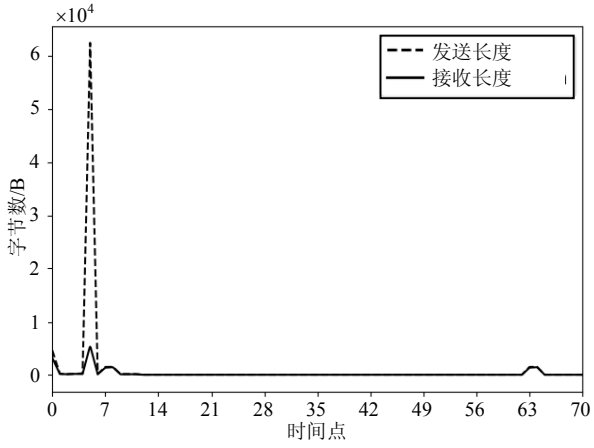
d. TCP 会话流持续时间分布

图1 真实环境会话流分析

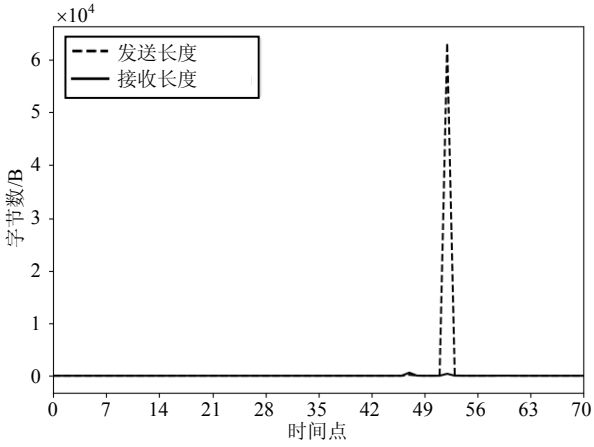
川大学校园网2017年10月共计约10亿条TCP会话流进行了观测。图1a～1d分别从收发字节数、收发包数、发收字节比、发收包数比和会话流持续时间(s)5个维度，按区间进行统计，其中横坐标代表区间，纵坐标分别代表上述5个量在相应区间内的会话流占总会话流的比例。其中，收发字节数分为$(0,10^2],(10^2,10^3],\cdots,(10^{10},+\infty)$共10个统计区间，收发包数分为$(0,10^1],(10^1,10^2],\cdots,(10^9,+\infty)$共10个统计区间，发收字节比和发收包数比分为$(0,10^{-3}],(10^{-3},10^{-2}],(10^{-2},10^{-1}],(10^{-1},10^0],\cdots,(10^3,+\infty)$共8个统计区间，会话流持续时间分为$(0,10],(10^1,10^2],\cdots,(10^6,+\infty)$共7个区间。如图1所示，正常情况下大多数会话流收发字小于10^4 B；收发包小于100；发收字节比以及发收包数比小于1；会话流的持续时间小于等于10 s。

1.2 隐蔽通信流量分析

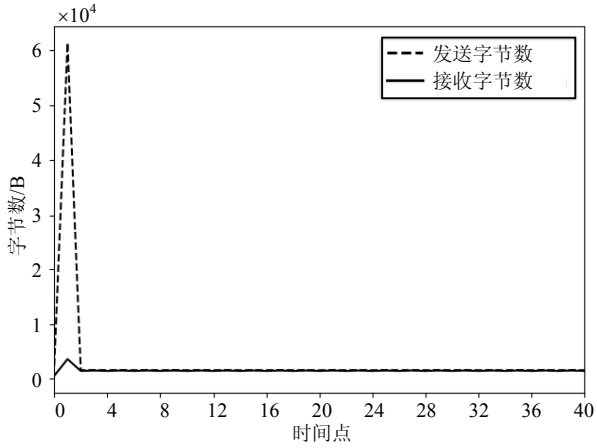
木马作为当前互联网中数量最多，影响最大的一类典型隐蔽性恶意程序，对网络信息安全的危害尤为突出[18]，其通信包括活跃期连接互操作、为保持连接状态而进行的隐蔽性连接保持、为通过不引起察觉的方式建立连接而进行的隐蔽性尝试连接3个典型状态[19]。本文以木马这一典型隐蔽性恶意程序为例，分析了包括Xtreme、Poison、灰鼠等在内的9个恶意程序产生的隐蔽通信流量的共性。图2以njRat、PcShare、Jaws及Bifrost为例，通过收发字节数对比，对隐蔽通信双方的交互行为的共性进行了描述，横坐标表示会话流序列编号(起始时间点为0)，纵坐标表示会话流收发字节数。由图2可知，除在连接互操作阶段因控制端主动向被控端发送控制命令，被控端执行相应操作并返回执行结果，因而发送字节数远大于接收字节数。在更一般的状态下，通信双方希望在不引起察觉的情况下进行交互，其会话流收发字节均无陡增的情况，且表现出平稳性，从而保持强隐蔽性。



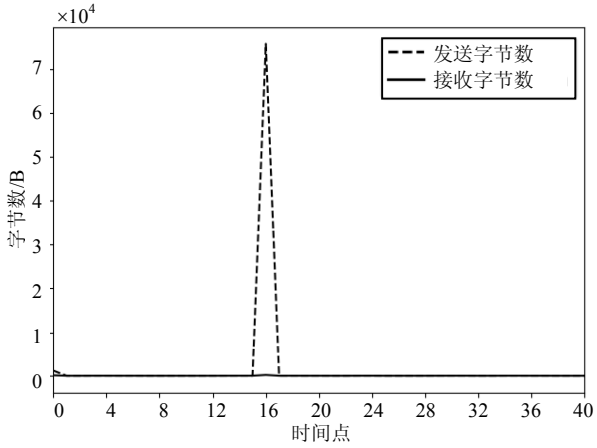
a. njRat



b. PcShare



c. Jaws



d. Bifrost

图2　隐蔽性恶意程序交互行为分析

在上述分析的基础上，本文做出如下定义：

定义 1　隐蔽性连接保持行为：通信双方连接建立后，表明存活性且不引起察觉而进行的隐蔽会话交互行为。

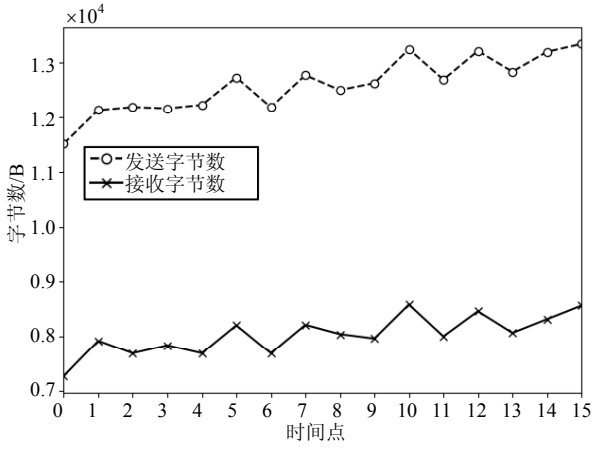
定义 2　隐蔽性尝试连接行为：通信双方中一方试图在不引起察觉的情况下与另一方建立连接而尝试性发起连接的行为。

1.2.1　隐蔽通信隐蔽性与发收比特征分析

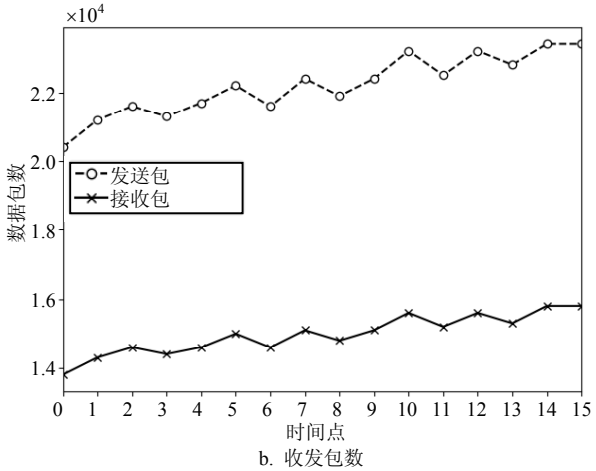
本文进一步利用隐蔽性恶意程序，针对性地分析了隐蔽性连接保持行为和隐蔽性尝试连接行为。图3a和3b以一个隐蔽性恶意程序为例，分别分析了隐蔽性连接保持通信中收发字节数、收发包数。与90%的真实环境中会话流一样，隐蔽会话收发字节数均约为10^4，收发包数小于100。因此，在日均千万条会话流的真实环境下，无论从收发长度还是收发包数的角度，隐蔽通信会话流均具有隐蔽性高、伪装性强、区分度低的特点。显然，对于隐蔽通信的检测，基于会话流开展常规的检测方式缺乏有效特征。

从图3c可知，和正常情况下的会话流不同，隐蔽通信中发送字节数大于接收字节数且发送包数大于接收包数。结合图3d与图1d可知，隐蔽性连接保持阶段会话持续时间相对较长。对于隐蔽性尝试连接通信，接收方无回应，因此其持续时间相对固定且很短。

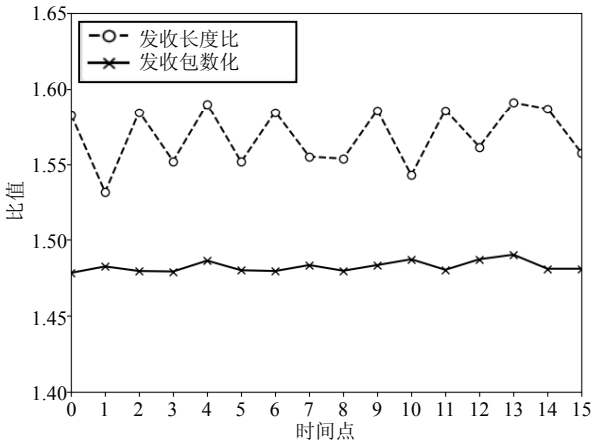
从图3还可以看出，隐蔽通信的通信双方在不同时间点产生的会话流之间表现出了较强相似性，收发字节数、收发包数、收发字节数比及收发包数比、持续时间等各属性值曲线均表现出较高的平稳性。



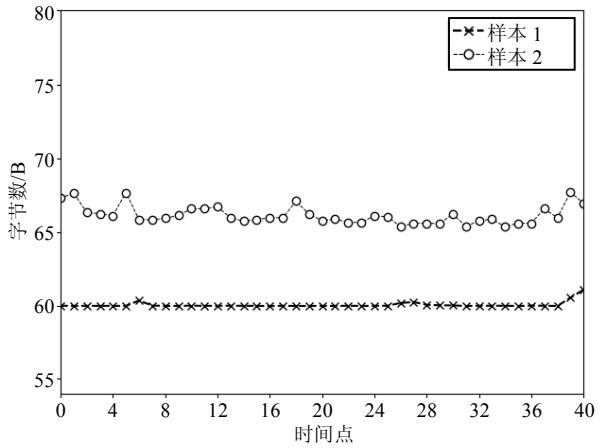
a. 收发字节数



b. 收发包数



c. 收发长度比和收发包数比

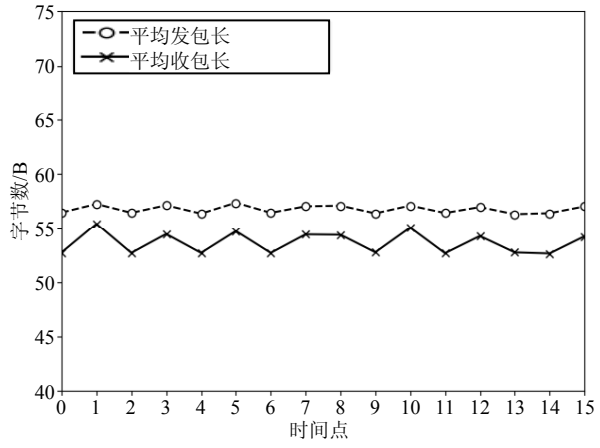


d. 持续时间

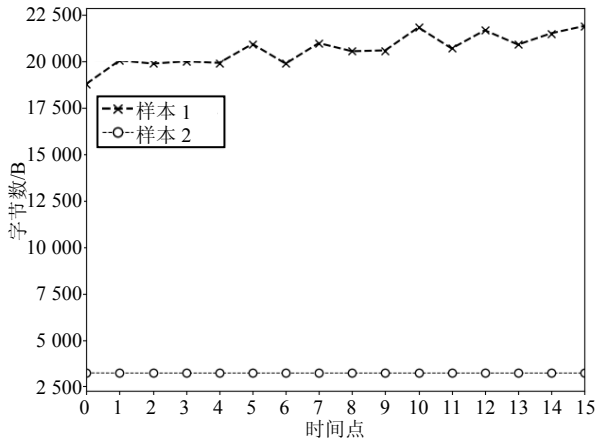
图3 隐蔽性连接保持行为分析

1.2.2 隐蔽通信会话流相似性分析

进一步地，本文对隐蔽通信会话流平均收发包长及会话流总字节数进行分析，如图4所示，验证了隐蔽通信中会话流间表现出的相似性。



a. 平均收发包长



b. 会话流总字节数

图4 隐蔽性连接保持行为进一步分析

1.2.3 隐蔽通信端口变化分析

本文对正常情况下某一服务进行访问时，通信双方产生的连续会话流的源端口变化情况和隐蔽通信双方产生的连续会话流中的源端口的变化情况进行了对比。隐蔽通信所产生连续会话流的源端口号具有连续性，而正常情况下访问服务时所产生的连续会话流的源端口不具有这种连续性。

2 基于会话流聚合的隐蔽性通信行为检测方法

本文提出一种运行在Spark框架下的并行化流量聚合算法。在此基础上，结合对隐蔽通信行为的分析，提取了隐蔽通信行为特征，并对其进行准确刻画。同时，为对隐蔽通信数据进行扩展，采用了标签传播算法对数据集中隐蔽通信不同阶段进行了标注。最终利用SVM算法，实现了海量数据下的隐蔽通信模型构建，对隐蔽通信进行检测。检测过程如图5所示。

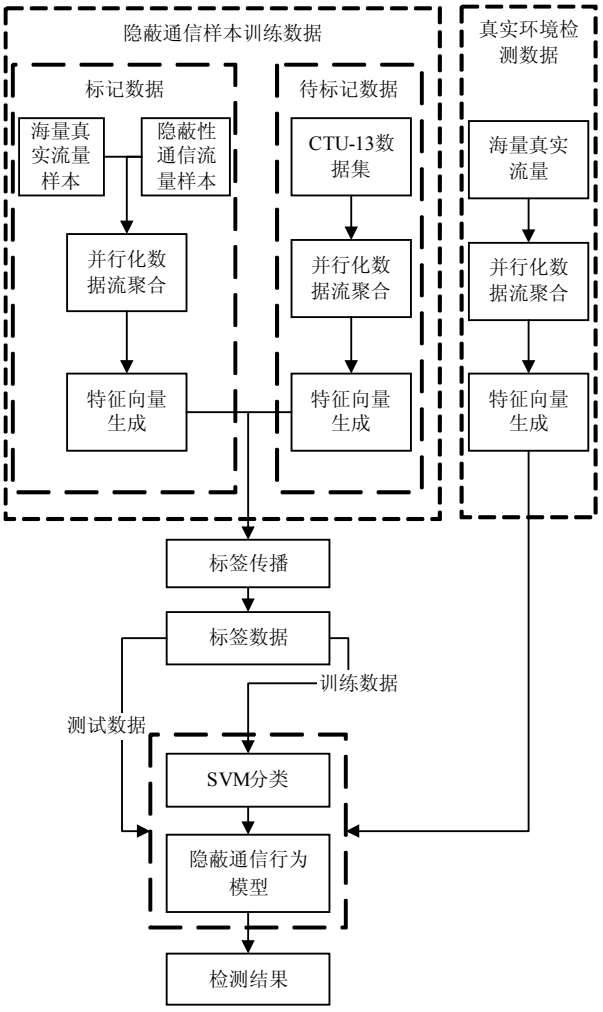


图5 基于会话流聚合的隐蔽通信行为检测流程

2.1 并行化会话流聚合算法

算法1描述了并行化流量聚合算法实现流程，输

入为HDFS中的session数据，输出为按聚合后的数据流。首先，将单条session数据转化成形如((srcIp, dstIp, dstPort), (startTime, session))键值对，即以源IP、目的IP、目的端口为键，会话流开始时间和会话流数据为对应的值。接着，将键值对按(srcIp, dstIp, dstPort)分组，然后对每一个分组的流量数据根据开始时间排序，对在时间阈值aggregation_time(单位：秒)之内的流量数据进行聚合，本文取aggregation_time=180。算法具体实现如下。

算法1　Parallel Flow Aggregation Algorithm

输入: RDD[session]: session data fetched from HDFS and stored in RDD

aggregation_time: flow aggregation time threshold

输出: combined session

keyValueRDD ← RDD[session].map(_.getKV)

groupedKVRDD←keyValueRDD.GroupByKey()

for each $S_i \in$ groupedKVRDD do

valueSeq ← S_i.getValue.toSeq

startTime ← S_i.getStartTime

sortedSeq ← valueSeq.sortBy(startTime)

index ← 0

for each $I_j \in$ sortedSeq do

if I_j.getStartTime − I_{index}.getStartTime ≤aggregation_time then

sessionList ← I_j

else

index ← j

sessionArray ← sessionList

sessionList ← ϕ

sessionList ← I_j

end if

end for

if sessionList≠ϕ then

sessionArray ← sessionList

end if

aggregatedSessionArray ← (S_i.getKey, sessionArray)

end for

return aggregatedSessionArray

2.2　特征提取

本文借助集中趋势和离散程度的统计量，提取聚合后会话流对应特征如下。

1) 会话流数量：聚合在一起的会话流数量N；

2) 收发长度比均值：聚合在一起的会话流收发字节数比平均值，计算方式为：

$$\text{MeanSRLenRatio} = N^{-1}\sum_{i=1}^{N}(\text{SendLen}_i / \text{RecvLen}_i) \tag{1}$$

3) 收发包数比均值：聚合在一起的会话流收发包数比平均值，计算方式为：

$$\text{MeanSRPktRatio} = N^{-1}\sum_{i=1}^{N}(\text{SendPkt}_i / \text{RecvPkt}_i) \tag{2}$$

4) 持续时间均值：聚合在一起的会话流持续时间均值，计算方式为：

$$\text{MeanDur} = N^{-1}\sum_{i=1}^{N}\text{Dur}_i \tag{3}$$

5) 会话流互异度：聚合在一起的会话流之间的差异程度，计算方式为：

$$\text{conv\_htrg} = N * C.V_{\text{sum}} \tag{4}$$

式中，

$$C.V_{\text{sum}} = \sum_{i=1}^{6} C.V_i \tag{5}$$

式中，$C.V_1, C.V_2, C.V_3, C.V_4, C.V_5, C.V_6$分别表示收发字节比、收发包数比、持续时间、发送长度、总长度、平均发包长的变异系数。以持续时间为例，其变异系数为：

$$C.V_{\text{Dur}} = \begin{cases} \text{SD}_{\text{Dur}} / \text{MN}_{\text{Dur}} & \text{MN}_{\text{Dur}} \neq 0 \\ 0 & \text{MN}_{\text{Dur}} = 0 \end{cases} \tag{6}$$

式中，SD_{Dur}和MN_{Dur}分别为：

$$\text{SD}_{\text{Dur}} = \left(N^{-1}\sum_{i=1}^{N}(\text{Dur}_i - \overline{\text{Dur}})^2 \right)^{\frac{1}{2}} \tag{7}$$

$$\text{MN}_{\text{Dur}} = N^{-1}\sum_{i=1}^{N}\text{Dur}_i \tag{8}$$

6) 端口有序度：聚合在一起的会话流端口有序程度，其计算公式为：

$$\text{PtOrderDegree} = \begin{cases} \left((N-1)^{-1}\sum_{i=1}^{N-1}(\text{port}_{i+1} - \text{port}_i)^2 \right)^{\frac{1}{2}} & N > 1 \\ 0 & N = 1 \end{cases} \tag{9}$$

2.3　基于标签传播的隐蔽通信数据集标注

为对样本进行扩展，本文借助标签传播的思想，对CTU-13数据集[20]中隐蔽性通信行为的不同阶段进行了分类标注。

2.3.1　标签传播算法描述

定义 3　$L=\{s_1, s_2, \cdots, s_l\}$为有标签聚合会话流集合。$Y=\{y_1, y_2, \cdots, y_l\}$为$L$对应的标签集合，

$C=\{c_1,c_2,c_3\}$ 为类标集合，分别表示隐蔽性连接保持行为、隐蔽性尝试连接行为及非隐蔽通信行为，且 $y_i \in C$， $i=1,2,\cdots,l$ 。 $U=\{s_{l+1},s_{l+2},\cdots,s_{l+u}\}$ 为无标签聚合会话流集合。集合 $S=L\bigcup U=\{s_1,s_2,\cdots,s_{l+u}\}$，其中， $s_i \in R^m$， $m=6$ 。

具体算法实现如下：

1) 聚合后会话流样本点间距离计算。每个聚合后会话流样本均由特征向量 $\boldsymbol{F}=(f_1,f_2,f_3,f_4,f_5,f_6)$ 描述，其中 $f_1,f_2,\cdots,f_6$ 分别表示2.2节中6个特征，计算每一个样本点与其他点之间的欧式距离。任意两个样本 $s_i,s_j \in L\bigcup U$ 间的距离为：

$$D(s_i,s_j)=\left(\sum_{k=1}^{6}\left|s_i^{(k)}-s_j^{(k)}\right|^2\right)^{\frac{1}{2}} \tag{10}$$

2) 根据计算得到的距离，构建 $N\times N$ 的相似矩阵 $\boldsymbol{P}=\{p_{ij}\}$ 。对于样本点 s_i， 由1)中计算出的距离，得到 k 临近样本集合为 $\mathrm{SK}_{(s_i)}=\{\mathrm{sk}_l \mid \mathrm{sk}\in L\bigcup U, 1\leqslant l\leqslant k\}$，本文 k 取 $3^{-1}u$ 。对该集合元素按距离由小到大排序后，若 s_j 为 $\mathrm{SK}_{(s_i)}$ 中第 l 个元素，则对相似矩阵 $\boldsymbol{P}$ 有 $p_{ij}=(k+l-1)^{-1}$，否则 p_{ij}=0，其中， p_{ij} 为点 s_i 到 s_j 的传播概率。

3) 构建 $(l+u)\times 3$ 的标签矩阵 $\boldsymbol{F}$ 。对于有标签聚合后会话流，构建 $l\times 3$ 的标签矩阵 $\boldsymbol{Y}_L=\{y_{ij}\}$，若第 i 个样本类别是 j，则 $y_{ij}=1$，且第 i 行其他元素为 0 。对于无标签聚合后会话流，构建 $n\times 3$ 的标签矩阵 $\boldsymbol{Y}_U=\{u_{ij}\}$ 。最后得到 $\boldsymbol{F}_{(l+u)3}=[Y_L;Y_U]$ 。

4) 通过 $\boldsymbol{F}=\boldsymbol{PF}$ 进行标签传播。

5) 已标注标签重置，并跳转至步骤4)。直至收敛，本文取收敛阈值为10^{-5}。

2.3.2 数据集标注结果

表1 数据集标注结果

恶意程序名称	聚合后会话流数	隐蔽通信		非隐蔽通信
		隐蔽性连接保持行为	隐蔽性尝试连接行为	
LuminosityLink	25	1	23	1
Gh0st	5	5	0	0
WisdomEyes	19	19	0	0
Regin	46	17	20	9
共计	95	42	43	10

为对CTU-13数据集中的木马通信隐蔽性通信行为数据的不同阶段数据进行标注，本文首先采集了Jaws、PcShare及njRat的隐蔽性连接保持行为和隐蔽性尝试连接行为会话流，以及校园网数据中心真实流量数据。通过本文提出的会话流聚合算法，将聚合后的会话流作为已标记样本。将CTU-13数据集中数据作为未标记数据，利用标签传播算法的标记结果如表1所示。

2.3.3 隐蔽通信模型构建

本文通过支持向量机(SVM)这一有较好分类性能且支持多分类问题的算法，构建隐蔽通信检测模型。为同时对隐蔽性连接保持行为和隐蔽性尝试连接行为进行识别，本文通过OvO(One vs. One)的方式，将非隐蔽通信会话、隐蔽性连接保持行为会话和隐蔽性尝试连接行为会话3个类别两两配对，从而产生 $3\times(3-1)/2$ 即3个分类器。检测阶段，将检测样本同时提交给所有分类器，得到 $3\times(3-1)/2$ 即3个分类结果，最终结果通过投票产生，把被预测得最多的类作为检测样本最终分类。

3 实验结果及分析

针对本文提出的基于会话流聚合的隐蔽通信行为检测方法，参考文献[12]的实验类型，分别设计了仿真及对比实验和真实环境中的检测实验。仿真及对比实验中，首先通过实验对SVM的核函数进行了选择，再通过召回率(true positive rate, TPR)和误报率(false positive rate, FPR)对检测方法进行评价，同时，将本文和已有研究进行了对比分析。为进一步说明本文方法的有效性，将检测方法应用于真实网络环境，并通过威胁情报对检测结果进行验证。

3.1 仿真及对比实验

为验证本文提出的检测方法，本文进行了仿真及对比实验。

首先通过交换机端口镜像的方式获得四川大学校园网原始流量。通过会话流聚合算法将2017年11月某日会话流进行聚合，并从中随机抽取1 000条聚合后会话流作为正常背景流量。为确定本文将使用的SVM算法的核函数，在背景流量中插入已标注的隐蔽性连接保持行为标签数据。引入ROC曲线(receiver operating characteristic curve)[21]作为标准，当核函数为RBF(radial basis function)时分类效果最佳，如图6所示，因此后续实验本文选取RBF作为核函数。

接着，在背景流量中插入已标注的隐蔽性连接保持行为标签数据和隐蔽性尝试连接行为标签数据共181条，构成仿真训练流量数据集合 X 。将 X 按 7∶3 划分为训练样本和测试样本两部分。

图7用归一化后的混淆矩阵表示检测结果。行表

示实际的类的实例，列表示类的实例预测。其中，type_1至type_3分布表示非隐蔽通信会话，隐蔽性连接保持行为会话及隐蔽性尝试连接行为会话。

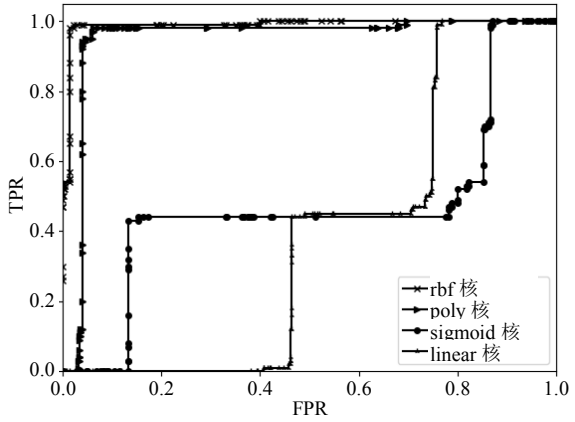


图6　不同核函数下的ROC曲线

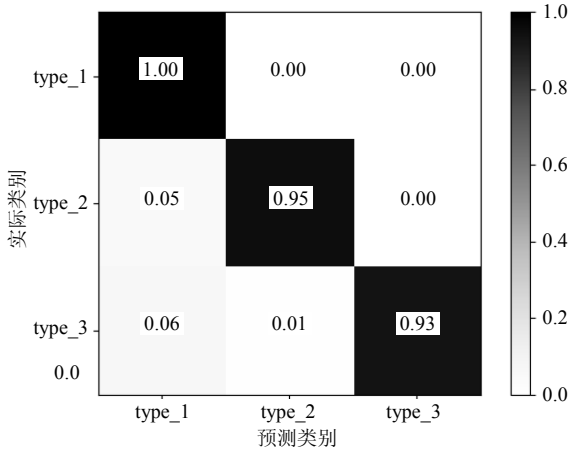


图7　检测结果混淆矩阵

此外，本文计算了提出的检测方法对隐蔽性连接保持行为和隐蔽性尝试连接行为的TPR和FPR，其结果如表2所示。

表2　仿真实验检测结果

隐蔽通信行为类别	TPR/%	FPR/%
隐蔽性连接保持行为	96.34	0.43
隐蔽性尝试连接行为	92.78	0

文献[13]共进行了数据规模不等的3组基于数据包[22]的实验，而本文实验基于聚合后会话流，选取了该文献最大规模的一组数据共769个数据包进行实验。文献[13]的该组实验可有效检测出80%的隐蔽通信数据包，而本文方法可检测出所有隐蔽通信数据，可见，对于隐蔽通信行为的检测，尤其在大规模数据下，本文方法优于文献[13]，如表3所示。

表3　功能性对比实验

隐蔽通信行为类别	本文方法	文献[13]
隐蔽性连接保持行为	全部检出	部分检出
隐蔽性尝试连接行为	有效	——

3.2　真实环境下检测实验

为进一步验证本文检测方法的有效性，将该方法应用于四川大学校园网真实环境。选取2017年11月15日共1 400万条会话流数据进行会话流聚合，得到聚合后会话流共460万条，检测结果数量中如表4所示。

表4　真实环境检测结果

检测类型	数量/条	占总量比例/%
隐蔽性连接保持	10 127	0.21
隐蔽性尝试连接	24 184	0.51

表5　真实环境检测结果样例分析

编号	检测类型	SrcIP-DstIP	DstIP威胁情报
1	隐蔽性连接保持	202.*.*.109-182.*.*.7	成都-僵尸网络
2	隐蔽性连接保持	202.*.*.84-45.*.*.166	荷兰-僵尸网络
3	隐蔽性连接保持	202.*.*.159-220.*.*.178	福州-僵尸网络
4	隐蔽性连接保持	202.*.*.159-113.*.*.59	菏泽-僵尸网络
5	隐蔽性连接保持	202.*.*.159-180.*.*.48	南通-僵尸网络
6	隐蔽性连接保持	202.*.*.159-60.*.*.100	玉溪-僵尸网络
7	隐蔽性连接保持	202.*.*.140-220.*.*.145	成都-僵尸网络
8	隐蔽性连接保持	202.*.*.159-223.*.*.119	芜湖-僵尸网络
9	隐蔽性尝试连接	202.*.*.113-51.*.*.211	沙特-僵尸网络
10	隐蔽性尝试连接	202.*.*.206-223.*.*.136	成都-僵尸网络

检测结果包括隐蔽性连接保持记录10 127条，隐蔽性尝试连接24 184条。其中包含大量校园服务器发起的通信记录。利用威胁情报[23]对这些通信的目的IP进行查询可知，多被标注为僵尸网络。进一步对这些记录进行人工分析发现，其通信行为满足本文所分析的隐蔽通信特征。部分检测结果如表5所示，验证了该方法能在真实网络中有效检测隐蔽通信行为。

4　结 束 语

本文对真实网络通信流量进行统计分析，重点对基于隐蔽会话的隐蔽性通信行为进行了研究。提出一种并行化会话流聚合方法，解决了海量数据下隐蔽性通信与正常通信之间区分难度极大的问题。借助集中趋势和离散程度的描述方式多角度刻画隐蔽通信行为，实现大数据环境下细粒度模型构建。针对隐蔽通信的长期性和复杂性带来的现有数据集

缺乏准确标注的问题，引入标签传播算法对标签数据进行扩展。最后通过支持向量机构建了多分类检测模型。本文将基于会话流聚合的隐蔽性通信行为检测方法应用于真实环境后，检测出了校园网中存在的隐蔽性通信行为。在下一步研究工作中，将持续关注相关领域最新文献，并与之对比，以在现有实验的基础上进一步说明本文方法的有效性；同时，将从隐蔽通信的持续性的角度进一步开展研究。

参 考 文 献

[1] 曹自刚. 隐蔽式网络攻击检测关键问题研究[D]. 北京: 北京邮电大学, 2015.
CAO Zi-gang. Research on key problems of hidden network attack detection[D]. Beijing: Beijing University of Posts and Telecommunications, 2015.

[2] LI M C, HUANG W, WANG Y B, et al. The optimized attribute attack graph based on APT attack stage model[C]// IEEE International Conference on Computer and Communications. Chengdu, China: IEEE, 2017: 2781-2785.

[3] LU J Z, ZHANG X S, WANG J F, et al. APT traffic detection based on time transform[C]//International Conference on Intelligent Transportation, Big Data & Smart City. Changsha, China: IEEE, 2017: 9-13.

[4] WANG X, ZHENG K, NIU X, et al. Detection of command and control in advanced persistent threat based on independent access[C]//IEEE International Conference on Communications. Kuala Lumpur, Malaysia: IEEE, 2016: 1-6.

[5] 邵国林, 陈兴蜀, 尹学渊, 等. 基于流量结构稳定性的服务器网络行为描述:建模与系统[J]. 电子科技大学学报, 2017, 46(1): 102-108.
SHAO Guo-lin, CHEN Xing-shu, YIN Xue-yuan, et al. Profiling structure-stability-based server traffic: Behavior models and applications[J]. Journal of University of Electronic Science and Technology of China, 2017, 46(1): 102-108.

[6] YE X, CHEN X, WANG H, et al. An anomalous behavior detection model in cloud computing[J]. Tsinghua Science and Technology, 2016, 21(3): 322-332.

[7] BOCCHI E, GRIMAUDO L, MELLIA M, et al. MAGMA network behavior classifier for malware traffic[J]. Computer Networks, 2016, 109(P2): 142-156.

[8] 陈兴蜀, 陈敬涵, 曾雪梅, 等. 基于TDRI的多视图关联DNS流量可视分析[J]. 工程科学与技术, 2018, 50(4): 123-129.
CHEN Xing-shu, CHEN Jing-han, ZENG Xue-mei, et al. Correlative visual analytics for DNS traffic with multiple views based on TDRI[J]. Advanced Engineering Sciences, 2018, 50(4): 123-129.

[9] SHAO G, CHEN X, ZENG X, et al. The analysis of malicious group based on suspicious communication behavior aggregation[C]//Chinese Conference on Trusted Computing and Information Security. Singapore: Springer, 2017: 143-164.

[10] YAMADA M, MORINAGA M, UNNO Y, et al. RAT-based malicious activities detection on enterprise internal networks[C]//International Conference for Internet Technology and Secured Transactions. London UK: IEEE, 2015: 321-325.

[11] JIANG D, OMOTE K. An approach to detect remote access trojan in the early stage of communication[C]// International Conference on Advanced Information Networking and Applications. Guwangiu, South Korea: IEEE, 2015: 706-713.

[12] WU S, LIU S, LIN W, et al. Detecting remote access trojans through external control at area network borders[C]//Symposium on Architectures for Networking & Communications Systems. Beijing, China: IEEE, 2017: 131-141.

[13] PALLAPROLU S C, NAMAYANJA J M, JANEJA V P, et al. Label propagation in big data to detect remote access Trojans[C]//IEEE International Conference on Big Data. Washington DC, USA: IEEE, 2017: 3539-3547.

[14] 陈兴蜀, 曾雪梅, 王文贤, 等. 基于大数据的网络安全与情报分析[J]. 工程科学与技术, 2017, 49(3): 1-12.
CHEN Xing-shu, ZENG Xue-mei, WANG Wen-xian, et al. Big data analytics for network security and intelligence[J]. Advanced Engineering Sciences, 2017, 49(3): 1-12.

[15] SELVARAJ R, KUTHADI V M, MARWALA T. Ant-based distributed denial of service detection technique using roaming virtual honeypots[J]. IET Communications, 2016, 10(8): 929-935.

[16] 刘建伟, 刘媛, 罗雄麟. 半监督学习方法[J]. 计算机学报, 2015(8): 1592-1617.
LIU Jian-wei, LIU Yuan, LUO Xiong-lin, et al. Semi-supervised learning method[J]. Chinese Journal of Computers, 2015(8): 1592-1617.

[17] GONG C, TAO D, LIU W, et al. Label propagation via teaching-to-learn and learning-to-teach[J]. IEEE Transactions on Neural Networks and Learning Systems, 2017, 28(6): 1452-1465.

[18] 国家计算机网络应急技术处理协调中心. 2017年我国互联网网络安全态势综述[EB/OL]. [2018-05-30]. http://www.cac.gov.cn/wxb_pdf/2018year/situation.pdf.
National Internet Emergency Center. Summary of internet security situation in China in 2017[EB/OL]. [2018-05-30]. http://www.cac.gov.cn/wxb_pdf/2018year/situation.pdf.

[19] LIANG Y, PENG G, ZHANG H, et al. An unknown trojan detection method based on software network behavior[J]. Wuhan University Journal of Natural Sciences, 2013, 18(5): 369-376.

[20] GRILL M, STIBOREK J, ZUNINO A. An empirical comparison of botnet detection methods[J]. Computers & Security, 2014 (45): 100-123.

[21] ZHANG Y, MERATNIA N, HAVINGA P. Outlier detection techniques for wireless sensor networks: A survey[J]. IEEE Communications Surveys & Tutorials, 2010, 12(2): 159-170.

[22] Chris Sanders. Packet capture files[EB/OL]. [2017-09-19]. http://chrissanders.org/packet-captures/.

[23] 微步在线. 威胁情报分析平台[EB/OL]. [2017-09-19]. https://x.threatbook.cn/.
ThreatBook. Threat intelligence analysis platform[EB/OL]. [2017-09-19]. https://x.threatbook.cn/.

编 辑 叶 芳