网络追踪对象的活跃状态预测

吴祎[1]，龚俭[1]

（1.东南大学网络空间安全学院，南京，210000）

摘　要：针对当下 SDN 架构的系统普遍存在的流表项资源数目有限，应用响应效率低的问题，本文面向 CERNET 安全保障系统对应用需求进行了分析，明确问题的研究方向，设计并实现了一种网络追踪对象的活跃状态预测方法，该方法基于 Markov 链，经实验表明在时间跨度合适的情况下该方法具有较好的预测准确率，同时由于算法复杂度低，具有良好的实现效率。然后为了增加预测结果对调度过程的参考价值，分析了动态窗口相比固定窗口的优越性，提出了一种窗口动态调整思路。

关键词：网络追踪对象；OpenFlow 流表项；Markov 链；状态预测

Active Status Prediction of Network Tracking Objects

Wu Yi[1], Gong Jian[1]

(1. School of Cyber Science and Engineering, Southeast University, Nanjing, 210000)

Abstract: Aiming at the problems of limited number of flow table items resources and low efficiency of application response that commonly exist in the systems of SDN architecture nowadays, this paper analyzes the application requirements for CERNET security assurance system, clarifies the research direction of the problem. Designs and implements active state prediction method, which is based on Markov chain. And experiments show that the method has good prediction accuracy in the case of suitable time span. The method is based on Markov chain, and the experiments show that the method has good prediction accuracy in the right time span, and good implementation efficiency due to the low complexity of the algorithm. Then, in order to increase the reference value of the prediction results to the scheduling process, the superiority of dynamic windows compared with fixed windows is analyzed, and a window dynamic adjustment idea is proposed.

Key words: Network Tracking Objects; OpenFlow flow entries; Markov chain; state prediction

随着计算机网络的普及和应用的迅猛发展，大型网络所面临的网络安全问题种类也越来越多。为了提高网络安全系统的可扩展性和自动化响应能力，许多大型网络的安全保障系统都应用了软件定义网络（SDN，software-defined networking）的思想和技术。CERNET 主干网运行管理与安全保障系统（BIG-CHAIRS）[1]就是这样的一类系统，该系统部署在江苏省教育网网络边界，为 CERNET 提供安全事件管理功能，对安全事件指示的网络追踪对象（即与安全事件相关联的追踪 IP 集合）进行流量采集及回溯分析，以提供给管理员可信的建议结果及有效信息。负责系统逻辑中追踪事件响应的子系统 HYDRA（hybrid detection response agent）[2]，其功能实现依托于 SDN 架构的分层思想，实际的系统追踪响应逻辑如图 1 所示。HYDRA 系统未对底层设备有特殊要求，同时拥有松耦合的处理逻辑，具有较高的可扩展性和自动化响应能力[3]。从图上可知，追踪事件的底层表达是 SDN 交换机中的流表项。

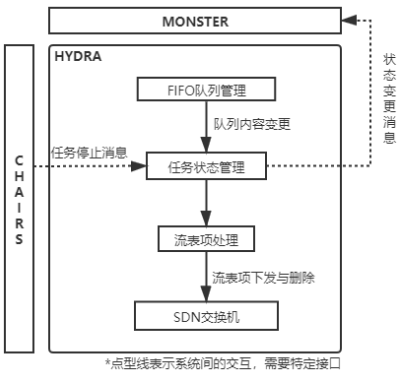

图 1　系统追踪响应处理逻辑

作者简介：吴祎，（1997-），女，硕士研究生，E-mail: wuyi@njnet.edu.cn；龚俭，（1957-），男，教授，E-mail: jgong@njnet.edu.cn

根据 CERNET 安全保障系统的应用特点和实际情况，安全事件通常需要一个较长周期的追踪回溯，因此随着系统中安全事件数量的不断增长，现有的 HYDRA 系统追踪机制面临一个亟需解决的资源问题。现有的 SDN 交换机因存储技术限制，流表项存储空间有限，如果在案件追踪过程中不执行一定的策略让不活跃的追踪 IP 让出流表项资源，HYDRA 系统的追踪响应效率将会受到相当大的影响，进而影响上层系统的分析效率。因此，研究追踪 IP 的活跃状态变化过程，是提高 CERNET 安全保障系统整体性能的重要部分。

基于上述研究背景和研究现状，本文研究了一种网络追踪对象活跃状态预测方法，先分析追踪 IP 的流记录在长间隔粒度上的特点，设计了一种基于 Markov 链的状态预测方法，实验结果表明该方法在状态序列时间跨度不过大的情况下，具有较好的预测准确率和执行效率。后提出了一种观测窗口的优化思路，该方法可以使预测方法对调度过程的有更高的适应性，为后续深入研究流表项的资源调度问题打下基础。

1 应用背景及问题分析

在 CERNET 安全保障系统中，一个安全事件到达后，需要经过追踪响应、IDS 检测、通信活动关联分析、行为画像[4]多个步骤，如图 2 所示。其复杂性以及过往的经验表明安全事件的回溯分析是长期且不断迭代的过程。时间上，通常需要采集较多天数的数据才便于完善分析安全事件。

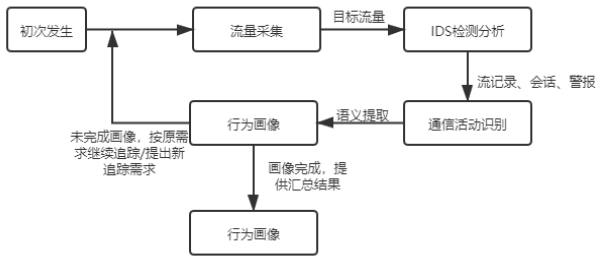

图 2 安全事件生命周期

过往研究对于流表项资源不足的解决方案有两种角度，一种是基于流表项的底层存储方式[5]，优化 TCAM（三态内容寻址存储器，ternary content addressable memory）；一种是对流表项本身的利用率进行优化。随着安全事件的累积，前者显然难以解决本文涉及的资源问题，而后者多是面向负载均衡场景或是流表项溢出攻击[6]，主要的研究方法可以分为三类：

1）从流表项的报文匹配情况出发。此类方法利用流表项的统计特性，按照有无数据包匹配来判断资源是否浪费的原则，计算出单一流表项的利用率，排序后淘汰利用率低的流表项；

2）从流表项对应的单条流出发。分析该条流的到达过程，利用流表项的超时机制实现流表项的淘汰，通过对超时时间的统计分析，得到合适的 hard_timeout（硬超时）值或 idle_timeout（空闲超时），也有研究[7]对流到达间隔分布和流持续时间进行预测，通过强化学习和随机森林算法去选择淘汰的流表项；

3）从交换机整体出发。该种方法通过对新流产生数或交换机吞吐量情况分析[8]，设置合适的超时值，达到提高流表项资源利用率的问题。

上述方案具有最小调度单位为流表项，只关注匹配流量大小、不考虑流量分析价值的特点。在本文所述安全保障系统场景下，对同一追踪 IP 的流量采集基于匹配规则分别为{src_ip = 追踪 IP}、{dst_ip = 追踪 IP }的两条流表项，考虑对后续分析的影响，对单一流表项进行分析并做淘汰选择是不合适的。在该场景下，资源调度涉及的最小操作单位是追踪 IP 对应的流表项集合。所以，在 CERNET 安全保障系统中，我们可以将流表项资源不足的问题转化为研究追踪 IP 是否在活跃，进而通过对流表项资源的调度来提高系统效率。

2 基于 Markov 链的网络追踪对象活跃状态预测

本文所研究的数据集来源于 CERNET 安全保障系统中历史安全事件涉及的追踪 IP 的流记录信息，流按照五元组的形式{源 IP 地址，源端口，宿 IP 地址，宿端口，传输层协议}进行定义。流记录的关键属性如表 1 所示。涉及追踪 IP 个数 1357 个，流记录条数 56179562 条。

2.1 活跃状态定义

流记录数据量的多少并不能反映流量的分析价值，即使当前时间段网络流量较小，也不会对该时间短的流量进行舍弃。

一般情况下，网络追踪对象的活跃状态可以定义为：按一个观测窗口统计，对象发生网络通信活动的时间段。具体来说，当该观测窗口有流记录，认为该追踪 IP 在此期间发生了通信活动，反之认为没有在活跃。得追踪 IP 的可选状态集为{活跃，不活跃}，可以将观测窗口内的追踪 IP 活跃情况表现为含两种状态的序列。

表 1 流记录关键属性列表

字段	字段描述
task_ip	追踪 IP
duration	流持续时间
src_ip	源地址
src_port	源端口
dst_ip	目的地址
dst_port	目的端口
protocol	传输层协议
ts	该流的起始时间

2.2 观测窗口研究

由于网络流量的自相似性特征[9]以及不同流之间持续时间的差异性，观测窗口大小的选取，会影响最终得到的观测序列。为了选取合适的观测窗口，我们先对整体流记录进行统计分析，再考虑寻找合适的时间粒度及大小。

2.2.1 流持续时间

对流记录的持续时间进行统计，得到流持续时间的基本分布如表 2 所示。

表 2 流持续时间的基本分布

区间（单位：秒）	流记录占比
[0,1]	56.3%
(1,10)	37.3%
(10,300)	6.3%
[300,max)	0.4%

流持续时间的分布有很强的重尾特征[10]，单包流占比很高，按照帕累托法则，将流记录中 80%的流视为短持续时间流，那么得到短持续时间流和长持续时间流的划分界在 6 秒左右，远远小于上层系统的最小分析周期 300 秒。

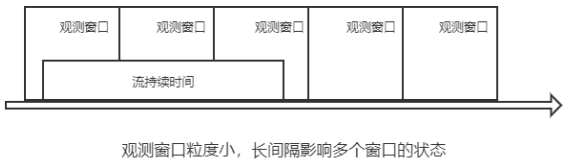

图 3 长持续时间流对窗口状态的影响

结合长持续时间流在网络中的分布以及图 3 可以推断出，当观测窗口粒度相对网络中绝大多数流的流持续时长明显偏大的情况下，可以认为将来的活跃状态只与当前活跃状态相关。

2.2.2 观测窗口设置

本文中追踪 IP 的活跃状态按一个足够长的观测窗口统计，默认在该情况下，要预测的状态只与前一个状态相关。

考虑系统愿意接受的观测周期以及追踪 IP 流间间隔的平均情况，以 30 分钟的观测窗口作为对比对象，再按 30 分钟的步长设置 1 到 5 小时的值为观测窗口 A 的值，计算各追踪 IP 在不同观测窗口下的状态序列，得到追踪 IP 在观测窗口 A 下的序列与观测窗口 30 分钟下状态序列的差异率。差异率的计算方式比较简单，以 30 分钟步长得到的序列和以 1 小时步长得到的序列为例，将后者的一个状态按顺序对应到前者的两个状态上，若这两个状态不相同，那么该状态就认为是一个相异状态，这类状态数量占后者状态总数的百分比即为相异率。

当差异率小于阈值时，该窗口认为是一个合适的观测窗口，如果有多个可用观测窗口，那么取最大的观测窗口作为最合适的观测窗口。

在不考虑极端不活跃 IP 和极端活跃 IP 的情况下，由差异率在可接受范围内的 IP 个数占总追踪 IP 数的百分比可以得出，占较多数的观测窗口最佳大小在 1 小时到 2 小时之间。我们选择 2 小时作为观测窗口，默认在 2 小时大小的观测窗口下得出的状态序列，追踪 IP 当前的活跃状态都只与前一个状态相关。

2.3 基于 Markov 链的网络追踪对象活跃状态预测

2.3.1 Markov 链

Markov 链是一种将来状态只受当前状态影响，要求过程具有无记忆特征的性质，下一状态的概率分布只与上一状态相关联，通过条件概率和状态转移概率矩阵即可进将来状态的概率预测。

根据前述分析，本文所定义的活跃状态符合使用 Markov 链的前提条件。

2.3.2 网络追踪对象活跃状态预测

本文设计了基于 Markov 链的活跃状态预测方法，其流程如图 4 所示。

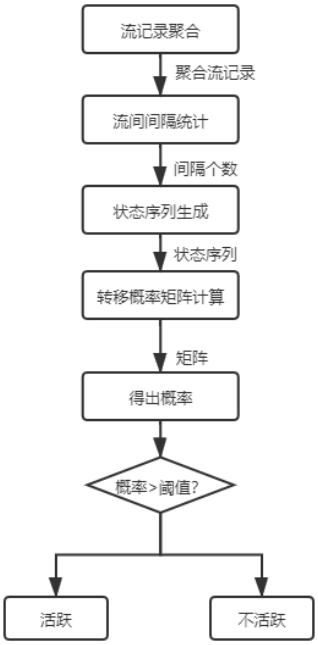

图 4 网络追踪对象的活跃状态预测方法流程

现以 IP 地址为 103.*.151.50 的追踪对象为例，具体说明本文中使用的 Markov 链预测方法。

按照 2 小时的观测窗口对该 IP 的流记录进行处理，观测窗口内如果存在流记录，则认为 103.*.151.50 在该两小时为活跃状态，记为 A，不存在则记为 B。得到 A 和 B 构成的活跃状态序列 S 后，对序列 S 中{A→A}，{A→B}，{B→A}，{B→B}计数，四种转移概率的计算方式如下：

$$P_{A\to A}=\frac{\mathrm{Count(A\to A)}}{\mathrm{Count(A\to B)+Count(A\to A)}} \quad (1)$$

$$P_{A\to B}=\frac{\mathrm{Count(A\to B)}}{\mathrm{Count(A\to B)+Count(A\to A)}} \quad (2)$$

$$P_{B\to A}=\frac{\mathrm{Count(B\to A)}}{\mathrm{Count(B\to A)+Count(B\to B)}} \quad (3)$$

$$P_{B\to B}=\frac{\mathrm{Count(B\to B)}}{\mathrm{Count(B\to A)+Count(B\to B)}} \quad (4)$$

得到转移概率矩阵：

$$P=\begin{bmatrix}P_{A\to A} & P_{A\to B}\\ P_{B\to A} & P_{B\to B}\end{bmatrix} \quad (5)$$

根据状态序列的最后一个状态是 A 还是 B，查找转移概率矩阵得到要预测状态是 A（活跃状态）的概率，当结果大于设定的概率阈值时，认为 103.*.151.50 接下来是活跃的。

2.4 实验及效果评估

本文设计了两种实验，分别研究概率阈值，以及状态序列长度对预测准确率的影响。实验通过随机数生成器，每次从数据集中随机选出追踪 IP 及对应子序列 500 个，将预测结果按真正率（TPR，true positive rate）进行记录，实验 100 次取平均值，得到最终实验结果。

2.4.1 概率阈值对预测准确率的影响

在前述实验条件基础上，设置概率阈值步长为 0.05，范围在 0.5-0.8 之间，状态序列长度设为 24，观测窗口为 2 小时（两天长度的数据）。实验得到的结果如表 3。

表 3 概率阈值对预测效果的影响

概率阈值	TPR(%)
0.5	82.644%
0.55	83.4%
0.6	86.232%
0.65	87.596%
0.7	85.323%
0.75	82.2%
0.8	81.423%

可以看出，概率阈值在 0.6-0.7 时，在本数据集上表现最为良好，且准确率在 85%以上。

2.4.2 状态序列长度对预测准确率的影响

将概率阈值定为 0.65，状态序列长度设置为步

长 12，范围 12-84（即取 1 天到 7 天）的值，得到实验结果如表 4。

表 4 状态序列长度对预测效果的影响

序列长度	TPR(%)
12	86.8%
24	87.632%
36	87.2%
48	86.921%
60	84.323%
72	76.34%
84	69.255%

2.4.3 实验评价

1）基于实际的流记录数据集，本文提出的预测方法，能够有效的对追踪 IP 的活跃状态进行预测，且预测准确率良好；

2）状态序列的建立过程未对网络流量的长相关特征进行分析，当状态序列的时间跨度足够大时，预测准确率呈较快速的一种下降趋势；因此将活跃状态预测结果用作流表项调度依据时，应当考虑历史数据的时间跨度，选取合适的序列长度或者优化预测方法，来达到准确预测的效果。

3 观测窗口的动态调整思路

预测活跃状态的最终目的是进行流表项的调度，而不同追踪 IP 之间的流分布存在差异性，而且定长的观测窗口可能会出现某个预测周期出现正在追踪的 IP 其状态预测结果均为活跃，很难进行流表项的调度决策。为提高预测准确度和对调度的适应性，提出一种窗口的动态调整思路。

结合调度问题的性质，为了不同追踪 IP 预测结果的可比较性，对于同一时刻交换机中所有在执行的追踪 IP，其观测窗口应当是相同的。但如果观测窗口在设置了基准值（前述的 2 小时）后，一直不变，那么在触发调度时，可能出现下列两种情况：

1）当前交换机内所有追踪 IP 的状态预测结果全部为{活跃}：由于系统遵循最小牺牲原则，在不考虑其他测度的情况下，被判断为活跃的追踪 IP 无需让出交换机资源，因此很可能出现经过多个调度时机，但交换机中无任何追踪 IP 让出资源，系统资源在该情况下效率降低；

2）交换机中多数追踪 IP 的预测结果经过多个调度时机，连续被预测为{不活跃}：这是 IP 本身活跃程度较低的原因导致的，为了对这些 IP 进行提供一定的适应能力，也应当调整观测窗口。

据此，本文设计了一个观测窗口的动态调整思路，其遵循的主要规则如下：

1）触发窗口缩小：当交换机中多数追踪 IP 连续 n 个调度时机被预测为不活跃；

2）触发窗口扩大：当交换机内所有追踪 IP 的状态预测结果全部为{活跃}；

3）停止缩小/扩大：窗口值具有上下阈值，当窗口缩小到下阈值时，不再进行缩小，若追踪 IP 仍被预测为{不活跃}，那么可以确定该 IP 需要有限让出资源，或者通过优先级调整让其未来得到资源的概率降低；当窗口扩大到上阈值时，不再进行扩大；

4）窗口恢复：若在扩大缩小过程中，出现一个预测周期结束，但这两种情况均未发生，那么观测窗口恢复基准值。

该思路可以通过步长和阈值来实现对提高对调度过程的作用，且方便追踪 IP 间预测结果的比较，减少出现极端调度情况的出现。

4 总结

本文从 CERNET 安全保障系统的背景和应用需求入手，确定了解决系统中流表项资源不足问题的研究思路，并基于 Markov 链提出了一种网络追踪对象活跃状态预测的方法，经分析，当用于计算概率矩阵的序列的时间跨度在一定范围内，该预测方法具有较好的预测准确率。最后提出了一种观测窗口的动态调整思路，有助于提高预测逻辑应用在调度问题中的适应性，为后续工作的研究提供了指向性作用。未来将结合流量的长相关性、节律性等特征对追踪 IP 流间长间隔的规律进行更进一步的分析，以提高流表项资源调度的合理性，进而提高安全保障系统的整体工作效率。

参考文献

[1] 朱礼智. 分布式网络应急响应管理系统 CHAIRS 的设计与实现[D].东南大学,2015.

[2] 金磊，龚俭.基于 SDN 技术的网络入侵阻断系统 HYDRA

的设计与实现 [D].南京：东南大学计算机科学与工程学院，2016.

[3] 程俊. 面向 SDN 网络安全协同系统[D].东南大学,2019.

[4] 郑飞飞. 基于多源数据关联分析的攻击意图推断[D].东南大学,2019.

[5] Zhou Yadong, Chen Kaiyue, Zhang Junjie, et al. Exploiting the vulne rability of flow table overflow in software-defined network: Attack model, evaluation, and defense [J/OL]. Security and Communication Networks, 2018 [2020-05-07].http://downloads.hindawi.com/journals/scn/2018/4760632.pdf

[6] 周亚东,陈凯悦,冷俊园,胡成臣.软件定义网络流表溢出脆弱性分析及防御方法[J].西安交通大学学报,2017,51(10):53-58.

[7] Yang Hemin, Riley G F. Machine learning based flow entry eviction for OpenFlow switches [C] //Proc of the 27th Int Conf on Computer Communication and Networks (ICCCN). Piscataway, NJ: IEEE, 2018: 1-8.

[8] 付应辉. 基于 SDN 的多路径负载均衡算法及流表分配优化算法研究[D].安徽大学,2017.

[9] 徐艳,周明中,王加俊.IP 流到达分布研究[J].河北科技大学学报,2009,30(04):333-339.

[10] 吴桦,周明中,龚俭.大规模网络中 IP 流长分布统计模型[J].计算机工程,2008(06):112-114+117.