浙 江 大 学 学 报 (工学版)
Journal of Zhejiang University (Engineering Science)

基于 IP 流记录的服务器识别

张状状，龚俭

（东南大学 网络空间安全学院，江苏 南京 211189）

摘要：为了对网络中的服务器进行有效管理和深度分析，合理使用有限的网络监管资源，提出了一种基于 IP 多包流记录的服务器识别方案，使用假设验证的科学研究方法进行算法构造，通过设置一些判定主机的通信端口为服务端口的条件，并结合一个检测周期内的多次通信活动，计算主机的服务器置信水平，最后对该主机进行服务器判定。通过多周期实验得到服务器集合，按照服务器的活跃程度和服务器类型进行分类统计分析，对实验结果进行深度语义挖掘，发现 web 服务器在集合中占比最高，且在大量的网络节点中，服务器数量较少。最后证明了假设的合理性，并对结果的完备性进行分析，结果表明，算法有较好的完备性，且识别准确率达 95%。

关键词：IP 流记录；服务器；服务器识别；网络安全态势感知；服务端口

中图分类号：TP 393　**文献标识码：**A　**文章编号：**

Server identification based on IP flow record

ZHANG Zhuang-zhuang, GONG Jian

(*School of Cyberspace Security, Southeast University, Nanjing, Jiangsu 211189*)

Abstract: In order to effectively manage and deeply analyze the servers in the network and use limited network supervision resources reasonably, a server identification scheme based on IP multi-packet flow records was proposed. The scientific research method of hypothesis verification was used to construct the algorithm. Determined some condition that the communication port of the host is the service port, and combined the multiple communication activities in one detection period to calculate the server confidence level of the host. Finally, performed server determination on the host. The server collection was obtained through multi-cycle experiments, and the statistical analysis was performed according to the activity level of the server and the server type. The deep semantic mining of the experimental results showed that the web server has the highest proportion in the collection, and in a large number of network nodes, the number of servers is small. Finally, the rationality of the hypothesis was proved and the completeness of the result was analyzed. The results show that the algorithm has good completeness and the recognition accuracy is 95%.

Keywords: IP flow record; server; server identification; network security situational awareness; service port

随着网络的飞速发展，网络环境日益复杂，各种层出不穷的安全问题愈演愈烈，这要求网络管理者对网络整体态势要有把控。在这种背景下，网络安全态势感知逐渐发展起来。首先，网络中存在大量功能各异的网络节点，其中服务器作为资源和计算的存储者和提供者，其安全性尤为重要，服务器

投稿日期：2019-07-30　**浙江大学学报（工学版）网址：**www.journals.zju.edu.cn/eng

基金项目：国家重点研发计划课题“互联网基础行为指标体系及测量方法”（课题编号：2018YFB1800202）资助项目

作者简介：张状状（1995-），男，硕士，E-mial：zzzhang@njnet.edu.cn.

通信联系人，龚俭，男，教授。ORCID:0000-0002-2966-3073. E-mial：jgong@njnet.edu.cn.

的安全态势成为网络态势感知的重要组成部分。其次，大规模互联网中包含大量活跃的 IP 地址，对它们进行一视同仁的监测，在有限资源的情况下是不可行也是不合理的，因为使用这些 IP 地址的主机的重要性和对网络的影响程度是不一样的。因此选择使用有限的监测资源优先监测服务器的流量行为将有助于提高这些监测资源的使用效率。而网络中通常拥有大量的服务器，管理者很难逐一记录，因此服务器的发现是对其进行安全性分析管理的前提和基础，也为合理利用有限的检测资源提供技术支撑。

本文基于 IP 流记录，提出了一种服务器识别方法，通过为通信主机和服务端口设置置信水平来进行服务器判定，可有效检测网络中各种类型服务器及其提供的服务。本章首先简单介绍了研究意义和背景，然后详细描述了所提出的算法，接下来对实验情况进行了分析讨论，最后总结全文。

1 问题背景

1.1 问题定义

本文所讨论的服务器有如下定义：在网络环境下，专门用于被其它节点访问，并向其提供某种资源服务的一种主机。常见的服务器有 web 服务器、域名服务器等。本文的识别算法是针对上述定义的服务器而提出的。服务器识别是指通过一些已知的数据得到服务器的 IP 地址，这与一般意义上的服务识别不同。服务识别的目的是发现网络活动中不同的服务及其类型，目前常用的方法是通过主机通信时使用的端口号进行服务发现，通常服务与端口号都是绑定的，因此可以较准确的发现服务。然而，随着 P2P 网络的普及，包含 P2P 应用的客户机也时常向对等节点提供服务，这些主机虽然有服务行为，却不属于上文定义的服务器。

IP 流记录作为一种常见的数据源，能有效进行大规模的流量行为分析，在当今的网络测量和行为观测领域应用广泛。因此，本研究使用流记录作为识别数据源，由一个分布式的 IP 地址综合信息系统（IPCIS）提供，IPCIS(IP Comprehensive Information System)是面向 CERNET 主干网运行管理与安全保障系统下的一个子系统，提供 IP 地址使用信息和 IP 流量信息。

流记录是对遵循某种特定规则的报文集合的摘要，常见的流记录规范为五元组形式{源 IP 地址，源端口，宿 IP 地址，宿端口，传输层协议}和超时约束。除了上述五元组，IPCIS 中流记录还包括出入方向报文数、是否为双向流、时间戳等。IP 流记录可以反映主机通信的对端 IP 数量、通信频率、主机活跃情况、通信端口情况等信息。本算法则是以流记录反映的语义为基础进行服务器识别。

1.2 研究现状

目前，针对服务器识别的研究多集中在单一种类的服务器识别，如 HTTP 服务器、数据库服务器、域名服务器等，上述研究可以为单一类型的服务器识别提供方法，但无法提供一种通用的服务器识别方案，且通常受限于不同服务器的固定服务端口。例如，文献[1]使用 TCP80 端口的双向流记录进行 web 服务器识别，实际上，部分 web 服务器并不使用 80、443 等常用的 http 端口，该方法将很难有效识别。不过作者仅分析双向流，为减轻流记录的复杂程度、高效的提取有效信息提供了可行方案。还有一些研究着重于小型局域网内的服务器检测，这部分研究主要针对少量相互通信的主机进行服务器的判别，并不适用于大规模的网络环境。

本文提出的服务器识别算法可以在大规模网络环境下，准确的识别出不同类型的服务器，为服务器识别研究提供了新的思路，一定程度上弥补了当前研究的不足。

2 服务器识别算法

实际条件下，路由器输出的流记录一般是抽样后的结果，服务器的服务过程很难完整的在流记录中体现。并且，网络中存在大量的诸如扫描等非正常流量，同时，随着 P2P 应用的快速发展，P2P 流量也在飞速增长，这使得网络流量愈发复杂。算法的目标就是在这种复杂的网络环境下，通过分析主机的 IP 流量特征，准确的识别出网络中的服务器。

算法的核心思想是设置一些判定主机的通信端口为服务端口的条件，通过统计观测分析主机的通信活动，判定其是否是服务器，并计算其置信水平。

一般情况下，服务器会将服务与某个端口绑定，其中，端口 0-1023 为固定端口，1024-49151 为注册端口，49152-65535 为动态端口。理论上，服务端口绑定在固定端口或注册端口上，不过也存在一些特殊服务绑定在动态端口上，而客户端使用的动态端口则针对不同系统有所不同，部分系统从 1024 起作为动态端口。因此，服务器的流量特征通常表现为重复使用个某固定端口的多包流，本实验则针对多包流记录进行分析。

根据服务器的特征和在流记录中的体现，本算法作出如下假设，认为满足下述全部条件的待检测主机为服务器：

(1) 目标主机与另一台主机进行交互时,重复使用某个端口。

(2) 与目标主机通信的对端主机使用随机的动态端口。

(3) 与目标主机通信的对端主机有多个。

(4) 目标主机与大部分对端主机的通信过程都满足条件（1）、（2）。

上述假设在流量层面表明了服务器与非服务器的不同行为特征，假设认为服务器通常重复使用端口，而对端主机则使用随机端口，符合实际环境中服务器与客户机的交互特点，由于服务器面向大量主机提供服务，故假设认为服务器的对端主机数量较多。接下来对相关定义和算法进行详细描述。

首先对算法中用到的相关术语进行定义，以更好的介绍本文的服务器识别算法。

定义 1 待检测目标 IP 地址[t]$S_{IP[t]}$：第 t 个检测周期中需要进行服务器识别的 IP 地址的集合。

定义 2 待检测流记录集合S_F：与目标 IP 地址相关的多包流量的集合。

定义 3 目标 IP 地址 AIM_IP：要进行服务器判定的主机的 IP 地址。

定义 4 对端 IP 地址 APPO_IP：与目标 IP 进行通信的主机 IP 地址。

定义 5 服务端口置信水平TLV_{Port}：某个通信端口在此次通信活动中是服务端口的置信水平。

定义 6 阈值 TS_P：在一次通信活动中，作为双方端口最高置信水平接近程度的门槛值。

定义 7 服务端口集合S_P：目标 IP 在通信中使用的服务端口的集合。

定义 8 阈值 TS_IP：目标 IP 是服务器的置信水平的门槛值。

定义 9 服务器置信水平TLV_{IP}：目标 IP 是服务器的置信水平。

定义 10 服务器集合[t]$S_{S[t]}$：第 t 个检测周期中从S_{IP}中检测出的服务器和端口的集合。 接下来，对整体的算法方案进行阐述。

基于上文的假设，设计了一种基于流记录的服务器识别算法。算法总框架如图 1 所示，其中，核心部分为服务器识别模块。

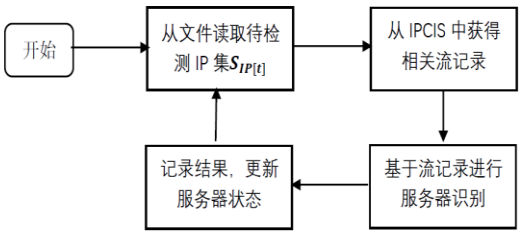

图 1 基于流记录的服务器识别方案流程

Fig.1 Server identification scheme based on stream record

下面详细介绍服务器识别模块的算法流程。

算法思路：以一天作为检测周期，读取在每个周期内，分析待检测 IP 产生的多包流记录，分别考虑每一次通信活动，为此次通信活动中使用的端口设置$TLV_{[Port]}$，随后综合多次与不同主机的交互过程，设置该 IP 的$TLV_{[IP]}$，结合预先定义的阈值，进行服务器判定。算法流程如下：

算法输入：待检测 IP 集$\boldsymbol{S_{IP[t]}}$。

算法输出：服务器集$S_{S[t]}$。

(1) 对于$S_{IP[t]}$中每一个待检测 IP，执行以下步骤

(2) If IP 无多包流记录 then 执行步骤 1

(3) 构造 AIM_IP 的多包流集合S_F

(4) 读取S_F中流记录，对于一条流记录，记录通信双方使用的端口，并用方法 SET_TRUST()来设置双方端口对应的TLV_{Port}，S_F读取完得到多组<AIM_IP,OPP_IP>对应的端口置信水平<aim_port1: $TLV_{Port}1$,… ; appo_port1: $TLV_{\mathrm{P}ort}1$,…>

(5) 对于所有的<AIM_IP,OPP_IP>对，分别找出双方端口中置信水平最高的AIM_TLV_{Port}，$APPO\_TLV_{Port}$，并执行 5，6。

(6) if $\frac{AIM\_TLV_{Port} - APPO\_TLV_{Port}}{APPO\_TLV_{Port}} \geq TS\_IP$ (1)

(7) then 对于所有满足公式（1）的 aim_port，将其加入服务端口集合S_P，

(8) $TLV_{IP} \leftarrow TLV_{IP} + 1$

(9) else $TLV_{IP} \leftarrow TLV_{IP} - 1$

(10) if $TLV_{IP} >$ TS_IP

(11) then 将{$\mathrm{AIM_{IP}}$:S_P}加入集合$S_{S[t]}$

设置服务端口置信水平的 SET_TRUST()算法：

(1) if (一方 IP 使用的端口 p1<1024 and 另一方使用的端口 p2>1023)

(2) then if p1 被使用过

(3) then p1.$\boldsymbol{TLV_{Port} \leftarrow TLV_{Port} + 1}$

(4) else p1.$\boldsymbol{TLV_{Port} \leftarrow 1}$

(5) if p2 被使用过

(6) then p2.$\boldsymbol{TLV_{Port} \leftarrow TLV_{Port} + 1}$

(7) else p2.$\boldsymbol{TLV_{Port} \leftarrow 0}$

(8) else if p1 被使用过

(9) then p1.$\boldsymbol{TLV_{Port} \leftarrow TLV_{Port} + 1}$

(10) else p1.$\boldsymbol{TLV_{Port} \leftarrow 0}$

(11) if p2 被使用过

(12) then p2.$\boldsymbol{TLV_{Port} \leftarrow TLV_{Port} + 1}$

(13) else p2.$\boldsymbol{TLV_{Port} \leftarrow 0}$

3 实验分析

3.1 实验过程

本研究以面向 CERNET 主干网运行管理与安全保障系统下的子系统 IPCIS 作为实验环境。算法在 Linux 系统下运行，基于 Python3 实现。由于算法存在大量的流记录查询过程，并且待检测 IP 集较大，为了提高算法效率，采用多进程并发运行的方式进行实验。

以东南大学所属 IP 作为目标探测集$S_{IP[t]}$，周期 t 为一天，进行 6 个周期的检测。东南大学 IP 共有 112128 个，具体网段如表 1。IPCIS 采集的流量仅是东南大学校园网进出 CERNET 主干网的流量，不包含进出运营商主干网的流量，因此，虽然东南大学 IP 较多，实际上只有少量活跃 IP 地址在 CERNET 边界路由器上产生流量，大部分 IP 采集不到流量或仅产生少量单包流，而服务器只可能存在于活跃的地址空间中，因此本实验将剔除不含多包流的不活跃 IP，针对活跃 IP 地址进行分析。同时，为了减少扫描流量和 p2p 流量对检测结果准确率的影响，将算法中阈值 TS_P 设为 0.5，TS_IP 设为 2。

表 1 东南大学所属 IP 地址

Tab.1 IP address of Southeast University

网段	IP 数量
58.192.*.0/20	4096
58.200.*.0/24	256
58.200.*.0/24	256
121.248.*.0/19	8192
121.248.*.0/20	4096
121.249.*.0/19	8192
202.119.*.0/19	8192
202.119.*.0/20	4096
211.65.*.0/19	8192
211.65.*.0/22	1024
223.*.0.0/16	65535

3.2 结果分析

算法从 2019-06-16 开始运行，经过 6 个周期的检测，共发现活跃 IP 数 5429 个，共识别出服务器 305 台，其中 158 台服务器有域名，147 台没有域名。IPCIS 系统中提供了域名数据库，根据 DNS 服务器的响应报文的 A 记录采集域名和 IP 的对应关

系，且数据库开发者提供了动态更新的功能，经过长期更新，目前域名库中包含 80M 以上的域名信息，因此域名数据库的完备性是可信的。通过对服务端口的分析，将服务器按服务类别进行分类，图 2 展示有域名服务器的分类结果，图 3 展示了无域名服务器的分类情况。

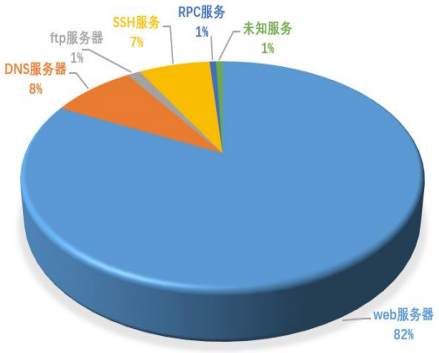

图 2 有域名的服务器分类结果

Fig.2 No domain name server classification results

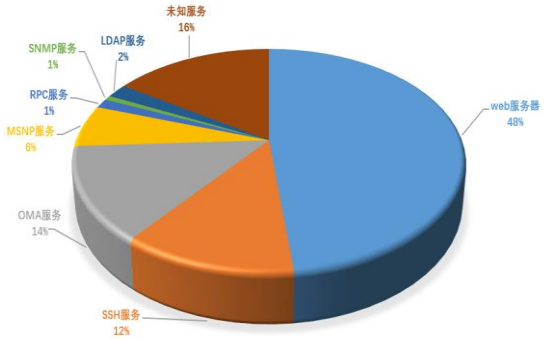

图 3 无域名服务器分类结果

Fig.3 No domain name server classification results

从上图可以看出，有域名服务器中 web 服务占了绝大多数，高达 82%，而在无域名服务器中，服务类型较多，web 服务器依然占比最高，为 48%。

将服务器按照活跃天数进行统计，结果如下：

表 2 服务器活跃天数分类统计结果

Tab.2 Statistics of server active days

活跃天数	4-6 天	2-3 天	1 天
服务器数	145	63	97

从上表可以看出，在检测的 6 个周期中，活跃服务器占比 48%，只在某一个周期内活跃的服务器占比 32%。若进行长期检测，可以更加清晰的反应服务器的活跃情况。

3.3 准确性与完备性验证

下面对实验结果进行分析，使用与假设不同的判定条件以验证假设的合理性。为了证明某台主机的确是服务器，需要分析其是否具有一些只有服务器具备的性质。首先，域名是为了便于记住网络中服务器的地址而出现的，因此，若一台主机有域名，那么可以认为这台主机是服务器；其次，一台服务器通常是不间断的提供服务，若端口是长期开放的，则可认为该主机为服务器；最后，服务端口一般都在 IANA 组织注册某种服务，故服务端口是否注册服务也可以作为服务器判定条件。结合上述分析，定义以下 4 种规则对识别的服务器的准确性进行验证，满足一条规则即认为通过验证：

(1) 待验证主机有域名；

(2) 对服务端口进行开放性探测，端口长期开放；

(3) 服务端口在 IANA 注册某种服务；

(4) 主机 IP 在 t 个周期的服务器集合$S_{S[t]}$中出现 $t/2$ 次以上。

对识别出的结果集$S_{S[t]}$进行验证，表 3 展示了服务器通过验证的条数的统计结果，其中有 287 台主机未通过验证，通过率为 95%以上。

表 3 通过不同验证规则数的 IP 数量

Tab.3 IPs of passing different verification rules

验证规则数	4	3	2	1	0
通过 IP 数	98	97	73	19	18

下面进行结果的完备性验证，通过分析待检测 IP 集中被算法排除的 IP 流记录特征，表明排除原因，以验证该算法较低的漏报率。定义了 3 个类别，为被排除的 IP 进行分类。

类别 1 不活跃 IP 地址：只产生少量的单包流记录，无多包流的 IP。该类别的 IP 绝大部分没有作为源地址的流记录，产生的少量单包流记录一般由端口扫描导致。

类别 2 使用动态端口的 IP 地址：产生的多包流记录使用不同的动态端口，具有客户端性质，不

满足假设中任一点。

类别 3 有服务行为的主机：主机在某次通信中的端口被重复使用，体现出服务器性质，但更多的通信过程使用动态端口，所以达不到设定的阈值。该现象由 p2p 流量导致，也不排除偶然性。

每个测试周期排除的 IP 分类情况和识别出的服务器数量如表 4 所示：

表 4 待检测 IP 集分类情况

Tab.4 Classification of IP sets to be tested

日期	检测到的 IP 总数	类别 1	类别 2	类别 3	服务器数
2019-06-16	111890	109458	2056	232	144
2019-06-17	111891	108580	2898	277	136
2019-06-18	111896	108895	2619	194	188
2019-06-19	111898	108901	2611	264	122
2019-06-20	111893	108825	2737	220	111
2019-06-21	111891	108966	2586	223	116

从上表可以看出，在总的地址空间中，只有极少数活跃地址，且活跃 IP 地址中超过 80%属于类别 2，故不可能为服务器；而类别 3 中的主机在某个不确定的时间段有服务行为，但很可能是 p2p 应用产生的，故不将其判定为服务器。可以看出，服务器数量占活跃地址空间的 5%左右，符合实际场景，算法的完备性得到验证。

4 总结

本文基于 IP 流记录，提出了一种检测网络中各种类型服务器的方案，在 IPCIS 平台实现。该方案可以识别 CERNET 网络中的服务器以及其使用的服务端口，以东南大学为例进行了多周期检测，并按照服务类型和活跃程度对实验结果进行了分析，随后进行准确性和完备性验证。经分析，该方案可以准确识别出大部分服务器。后续工作将结合多周期下的实验结果，以剔除误报的服务器，并尝试使用动态阈值，以得到更为精确的检测结果。

参考文献（References）：

[1] 丁伟, 洪沿, 夏震. 基于流记录的 HTTP 80 端口服务检测和分析[J]. 华中科技大学学报, 2016.6, 44:34-38.

DING Wei, HONG Yan, XIA Zhen. HTTP 80 Port Service Detection and Analysis Based on Stream Recording [J]. **Journal of Huazhong University of Science and Technology.** 2016.6, 44:34-38.

[2] Bo R, Cheney D, Braun H W. Internet flow characterization: adaptive timeout strategy and statistical modeling[C]//**Proc Passive and Active Measurement Workshop**. Amsterdam: Springer, 2001: 45.

[3] CNCERT. 2015 年中国互联网网络安全报告[EB/OL]. [2016-06-02]. http://www.cert.org.cn/pubhsh/main/upload/File/2015annualreport.pdf.

[4] 张凌峰. 丁伟, 龚俭, 等. 基于流记录的主干网活跃 IP 地址空间检测[c]//第四届中国互联网学术会议 (ICoC 2015)论文集. 沈阳: 东北大学出版社, 2015: 347-352.

ZHANG Ling-feng, DING Wei, GONG Jian, et al. Detection of Active IP Address Space in Backbone Network Based on Flow Recording[c].//**Papers of the Fourth China Internet Academic Conference (ICoC 2015)**. Shenyang: Northeast University Press, 2015: 347-352.

[5] 张维维, 龚俭, 丁伟, 等. NBOS: 一个基于流技术的精细化网管系统[J]. 太原理工大学学报, 2012, 43 (S1): 41-45.

ZHANG Wei-wei, GONG Jian, DING Wei, et al. NBOS: A refined network management system based on stream technology

[J]. **Journal of Taiyuan University of Technology,** 2012, 43 (S1): 41-45.

[6] 叶忻. 基于分布式架构的 IP 活动库的设计与实现[D]. 东南大学学报, 2015.

YE Xi. Design and Implementation of IP Activity Library Based on Distributed Architecture [D]. Journal of Southeast University, 2015.

[7] 柳斌, 李之棠, 李佳. 一种基于流特征的 P2P 流量实利识别方法[c]//中国教育与科研计算机, 2007.

LIU Bin, LI Zhi-tang, LI Jia. A Pragmatic Recognition Method for P2P Traffic Based on Flow Characteristics [c]// **China Education and Research Computer,** 2007.

[8] Lee D J, Brownlee N. Host measurement of network traffic[C]//**Telecommunication Networks and Applications Conference**, 2007 ATNAC 2007. Australasian IEEE, 2007 282-287.