计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2021,30(11):63−70 [doi: 10.15888/j.cnki.csa.008232]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于 VPN 日志的远程学习和科研态势研究①

赖清楠, 郭　强

(北京大学 计算中心, 北京 100871)
通讯作者: 赖清楠, E-mail: laiqn@pku.edu.cn

摘　要: 新冠疫情的暴发, 学生教工无法返校的情况下, 绝大多数高校采用 VPN 的方式保证远程学习和科研. 为了解具体情况, 采集了 2020 年 2 月至 2020 年 9 月新冠疫情期间北京大学的 VPN 日志, 从使用人数、登录登出时间、使用时长、聚类分析、用户类别 5 个方面进行讨论. VPN 每日使用人数最高约 1.5 万, 同时在线人数最高约 0.5 万, 每日的用户平均使用时长最高 325 min, 几项数据表明, 学生教工高度依赖 VPN 进行远程学习和科研, 根据用户的每日平均使用时长和使用天数对用户进行聚类分析, 可以大致将用户分为 4 类; 对用户类别进行分析, 理工科用户 VPN 使用时间比文科用户略长, 但变化趋势基本一致. 这些数据对于了解 VPN 使用情况, VPN 资源调整具有参考价值. 虽然新冠疫情期间 VPN 为校内资源的获取提供了便利, 但其带来的安全隐患也不容忽视, 用户终端如果保护措施不足就容易受到黑客的攻击, 从而作为跳板窃取校内资源或者攻击校内其他机器. 在远程学习和科研将成为新常态的趋势下, 是“甘饴”还是“毒药”? 高校都应该做好充足的准备来应对.

关键词: VPN; 日志分析; 新冠疫情; 远程学习和科研; 网络安全

引用格式: 赖清楠,郭强.基于 VPN 日志的远程学习和科研态势研究.计算机系统应用,2021,30(11):63–70. http://www.c-s-a.org.cn/1003-3254/8232.html

Situation Research on Remote Learning and Scientific Research Based on VPN Logs

LAI Qing-Nan, GUO Qiang

(Computer Center, Peking University, Beijing 100871, China)

Abstract: With the outbreak of the COVID-19 pandemic, most colleges and universities have adopted VPN to ensure remote learning and scientific research when students and teachers cannot return to school. To understand the specific situation, we collect the VPN logs of Peking University during the COVID-19 pandemic from February 2020 to September 2020 and discuss the number of users, login and logout time, usage time, cluster analysis, and user categories. The maximum number of daily users of VPN is about 15 000, and the maximum number of concurrent users is about 5000. Moreover, the average daily usage time of VPN is up to 325 minutes. These data show that students and teachers rely highly on VPN for remote learning and scientific research. According to the daily average usage time and the number of days of use, the users can be roughly divided into 4 categories. The VPN usage time of science and engineering users is slightly longer than that of liberal arts users, but the trend of change is the same. These data are of reference value for understanding VPN usage and adjusting VPN resources. Although VPN has facilitated the acquisition of school resources during the COVID-19 pandemic, the security risks it brings cannot be ignored. If protections are not enough, the user terminal will be vulnerable to attacks by hackers, who can steal resources or attack other machines in school using the user terminal as a springboard. As remote learning and scientific research will become the new normal, is it “sweet” or “poisonous”? Colleges and universities should be well prepared to deal with it.

Key words: VPN; log analysis; the COVID-19 pandemic; remote learning and scientific research; network security

① 收稿时间: 2021-01-15; 修改时间: 2021-02-07, 2021-04-09, 2021-04-16; 采用时间: 2021-04-20; csa 在线出版时间: 2021-10-22

新型冠状病毒肺炎是近百年来人类遭遇的影响范围最广的全球性大流行病, 对全世界是一次严重危机和严峻考验. 人类生命安全和健康面临重大威胁[1]. 2020 年 1 月下旬, 新冠疫情在国内暴发, 为了保障人民的生命安全, 人员流动受到限制. 虽然目前国内新冠疫情趋于缓和, 各地逐步复工复产, 但国外依然紧张, 仍然不能放松警惕.

在这期间, 高校正常的教学科研受到影响, 教师学生在无法返回学校的情况下, 只能居家进行远程学习和科研. VPN 能够极大地方便高校师生校内资源、电子期刊、学术资源的获取, 满足远程学习和科研的需求. VPN 的使用分为客户端和服务器, 用户使用学校统一身份认证登录 VPN 客户端后, 服务器会随机分配 VPN 地址池的某一 IP 给用户, 再去访问资源的时候就和校内用户没有任何差别. 本文通过采集 VPN 系统的日志, 分析新冠疫情期间师生的远程学习和科研的情况.

1 研究现状

新冠疫情暴发后, 为了防止大范围的扩散, 一线二线城市提倡高校远程学习和科研, 这使得远程学习和科研在国内普及开来. 从全世界来看, 远程学习和科研在欧美国家普及率较高[2], 在中国, 更多的只是作为日常学习和科研的辅助手段. 虽然远程学习和科研保障了正常的学习科研工作, 但也会带来网络安全问题. 与校内场景相比, 远程学习和科研的硬件设备、网络环境均不同于内网, 防护做得不足[3]. 使用 VPN 技术, 身份安全以及访问权限可能会成为其弱点. 黑客通过盗用身份登录 VPN, 完全有可能利用 VPN 进入网络内部并进行大肆搜掠[4]. 但是如果能够给 VPN 增加双因素认证, 如移动电话或者软令牌, 将会提高 VPN 使用上的安全性[5]. 我们在使用 VPN 提供的便利的同时, 也不能忽略 VPN 带来的安全问题.

为了了解师生在线学习和远程办公情况, 分析了新冠疫情期间北京大学 VPN 日志. 在网络日志分析及 VPN 日志分析方面很多学者也做了一些相关的研究. 余慧佳等[6] 对搜狗引擎在一个月内的查询日志进行了分析, 从独立查询词分布、同一 session 内的用户查询习惯及用户是否使用高级检索功能等方面对用户行为进行了分析. Mat-Hassan 等[7] 基于 AutoDoc 搜索和导航文档系统的日志数据, 提出了一个用户搜索会话模型, 并对用户的链接或点击选择行为和搜索策略模式进行了分析. Lu 等[8] 为了从日志中统计出真实的 VPN 用户数, 使用特征提取和日志分析方法, 提出了一种新的 VPN 用户识别算法, 论文提取了用户的基本信息、源 IP 地址、账号名称等特征来区分不同的用户, 最后用 2 个月的 VPN 日志验证了算法的有效性和准确性. 武凌等[9] 设计了一个基于 Hadoop 的 VPN 访问日志分析平台, 将 VPN 日志与流量关联, 产生用户的轨迹追踪报表, 找出资源滥用者和潜在的安全威胁. 本文采用了基于日志特征的分析方法, 根据 VPN 日志的类别对日志进行特征提取, 将源 VPN 日志分解成数据量较小的特征日志, 最后对特征进行汇总, 得到所需要的分析结果.

2 日志采集

本文研究对象为新冠疫情期间北京大学 VPN 服务器产生的日志, 为了缓解压力, 学校部署了多台 VPN 服务器, 从这些服务器上采集了 2020 年 2 月 20 日到 2020 年 9 月 20 日, 共 7 个月的 VPN 日志, 大小约为 21 GB, 日志条数约 103 989 852 条.

初步分析后, 日志大概可以分为 Login、System、VPN Tunneling、WebRequest 4 大类, 如表 1 所示. Login 记录用户登录过程中产生的日志; System 记录用户认证成功后, 系统对用户权限的分配日志; VPN Tunneling 记录的是 VPN 通道的建立日志; WebRequest 记录的是用户使用 Web 方式连接 VPN 访问资源的请求, 目前只有极少用户采用此方式, 因此绝大部分用户资源获取的流量日志并未记录在 WebRequest 中.

3 分析方法

本文提出了基于特征的 VPN 日志分析方法, 从 VPN 服务器上采集的源日志是以日期命名, 以天为单位生成的, 每天日志大约为 25 MB. VPN 日志具有很明确的日志类别以及格式规则, 日志里带有源 IP 地址、用户登录账号、客户端类型等信息. 截取了一段真实的 VPN 日志, 如下所示, 其中部分敏感信息采用 xxxx 代替.

2020-03-10 00:01:06 - vpn_D - [xxxx] xxxx(pku's users)[标准用户角色] - VPN Tunneling: Session started for user with IPv4 address xxxx, IPv6 address xxxx, hostname xxxx

2020-03-10 00:01:07 - vpn_D - [xxxx] xxxx (pku's users)[标准用户角色] - Closed connection to xxxx after 1874 seconds, with 4402960 bytes read and 24096373

bytes written

日志里各个字段的含义如下:

时间-VPN 服务器标识-[源 IP 地址] 用户账号 (pku's users)[角色分组]-日志内容

不同类别的日志, 日志内容格式是固定的, 例如 Session started 日志格式固定为:

VPN Tunneling: Session started for user with IPv4 address xxxx, IPv6 address xxxx, hostname xxxx

里面包含了登录后分配的 IPv4 地址和 IPv6 地址, 用户的 hostname 信息. 根据 VPN 日志的类别以及格式规则, 把时间、账号、源 IP 地址、连接时长等信息作为日志的特征, 以这些特征对日志进行分类, 分析方法如图 1 所示.

表 1 VPN 日志分类

序号	类别	子类
1	Login	Login succeeded/failed
2		Testing password
3		Primary authentication succeeded/failed
4		Agent login
5		Users logged out (logout)
6	System	Source IP realm restrictions
7		User limit realm restrictions
8		User chose
9	VPN Tunneling	Session started/ended
10		Transport mode
11		Connected to TUN-VPN
12		Users changed from
13		Session resumed/extended
14		Session timed out
15		Max session timeout
16		Closed connection
17	WebRequest	WebRequest ok/completed

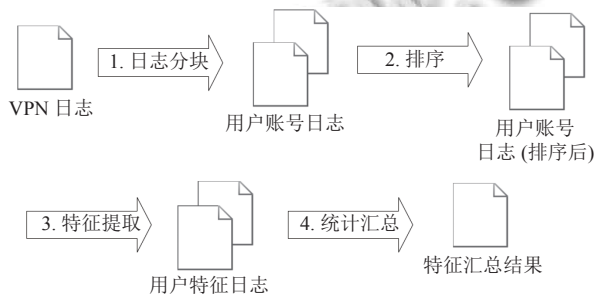

图 1 基于特征的日志分析方法

(1) 日志分块. 源日志按照登录用户账号进行分类, 为每个用户账号建立一个日志文件存储这个账号相关的所有日志.

(2) 排序. 同一账号不同时间段的登录可能会分配到不同 VPN 服务器上, 分块后的日志时间上有可能是错乱的, 为了不影响后续处理, 需要对日志按时间进行排序.

(3) 特征提取. 以统计用户 VPN 使用时长为例, 逐个读取分块后的日志文件, 提取出 Closed connection 日志, 按照 Closed connection 日志规则匹配出使用时长, 为每个用户生成一个只包含时间、账号、使用时长的用户特征日志.

(4) 统计汇总. 将用户特征日志进行汇总.

4 VPN 使用情况

本文分析了新冠疫情期间学生和教工远程学习和科研时 VPN 的使用情况, 从使用人数、登录登出时间、使用时长、聚类分析、用户类别 5 个方面进行讨论.

4.1 使用人数

VPN 日志里记录了账号信息, 在校园网的场景下可以用账号来代表用户, 虽然存在一些公共账号, 但仍可以用账号数量近似的表示用户数量. VPN 日志 Closed connection 类别, 如下所示, 包含账号信息、VPN 使用时长等, 加上日志的时间戳可以推算出用户登录 VPN、登出 VPN、以及在线时间段.

2020-03-15 00:01:47 - vpn_C - [xxxx] xxxx(pku's users)[标准用户角色] - Closed connection to xxxx after 27485 seconds, with 13847842 bytes read and 45740435 bytes written.

图 2 为从 2 月 20 日到 9 月 20 日, VPN 使用人数以及同时在线人数变化情况, 整体上看使用人数和同时在线人数处于下降趋势. 图 3 为 2 月 20 日到 9 月 20 日全国新增确诊人数变化情况, 数据来自中华人民共和国国家卫生健康委员会 (http://www.nhc.gov.cn). 4 月份开始疫情趋于缓和, 6 月份和 7 月份小规模的复发, 8 月份开始逐渐又呈现缓和趋势.

两者对比来看, 2 月份到 6 月份春节学期期间疫情较为严重, 师生远程学习和科研, VPN 使用人数有所波动但保持在较高的水平, 周末和假期休息时间段人数下降明显. 7 月、8 月暑假开始后, 使用人数下降到之前的一半左右, 9 月份秋季学期开学后, 疫情逐渐平稳, 学生和教工返校, 不再依赖 VPN, 使用人数再次下降.

表 2 统计了 2 月到 9 月平均在线人数的情况, 2 月到 4 月每天的平均使用人数突破 1 万, 最高使用人数

接近 1.5 万一天, 同时在线人数在 3800 左右, 最高同时在线人数达到 0.5 万. 5 月份开始使用人数和同时在线人数开始下降, 9 月份下降到 2 月到 4 月的 1/4 水平.

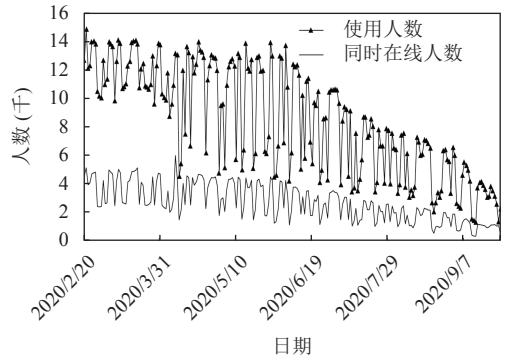

图 2 使用人数和同时在线人数

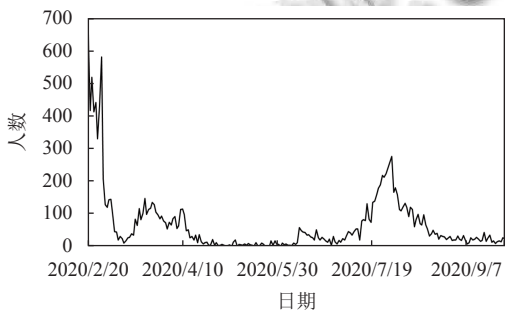

图 3 全国新增确诊人数

表 2 使用人数和同时在线人数统计

月份	使用人数			同时在线人数		
	平均值	最高	最低	平均值	最高	最低
2	12 442	14 830	5142	3949	5142	2408
3	12 286	14 095	5147	3834	5147	2495
4	11 050	13 942	6004	3629	6004	1515
5	9681	13 922	4513	3219	4513	1286
6	9033	13 700	4260	2844	4232	1162
7	6870	10 636	3419	2217	3419	1069
8	5100	7579	2502	1607	2502	573
9	3534	6590	1925	1101	1925	364

4.2 登录登出时间

图 4 是 2 月 20 日到 9 月 20 日按照时间段统计的用户登录和登出 VPN 人次, 横轴为时间段, 0 表示 0 点到 1 点之间, 以此类推, 纵轴表示人次. 从图中可以看出登录人次和登出人次曲线趋势是一致的, 夜晚少白天多, 7 点后学生进入学习状态, 教工进入科研状态, 因此登录 VPN 人次逐步上升, 9 点后登录人次趋于稳定. 10 点、15 点、21 点出现峰值, 10 点、15 点正好是开始上课的时间, 21 点是学生最活跃的查资料以及做实验时间, 因此这几个时间点 VPN 使用人数较多. 22 点登录人次逐渐下降, 12 点、17 点、22 点登出人次出现峰值, 12 点、17 点为下课时间, 22 点为用户休息时间. 登录和登出时间, 基本上与学生教工的学习科研规律符合.

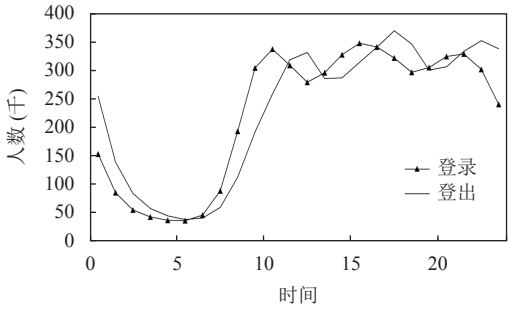

图 4 登录登出人次统计

4.3 使用时长

与往常不同, 新冠疫情期间, 学生需要进行长时间的远程学习, 教工需要进行长时间的科研活动, 因此大量 VPN 用户的使用时长要比往常高. 图 5 表示的是 2 月 20 日到 9 月 20 日 VPN 用户的平均使用时长. 平均使用时长计算方法如下:

$$\text{平均使用时长} = \frac{\text{当天所有用户的使用时长总和}}{\text{当天的用户总数}}$$

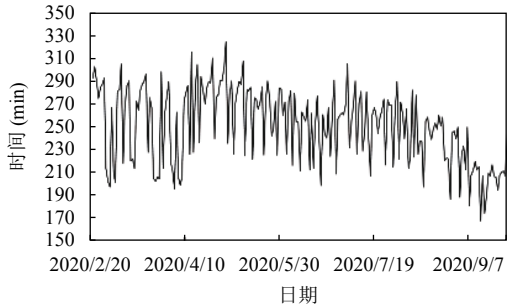

图 5 平均使用时长

平均使用时长从 2 月份到 8 月份并没有明显的变化, 约为 250 min, 9 月份平均使用时长有所下降, 约为 200 min, 期间周末和假期平均使用时长下降明显. 2 月份到 6 月份的使用人数上是 7 月份到 9 月份的 3–4 倍, 虽然平均使用时长相差不大, 但是总的使用时长要高很多, 说明了新冠疫情期间 VPN 为学生教工远程学习和科研提供了很好的支持. 9 月份之后, 有一些常驻校

外的教工需要连接 VPN 进行科研活动, 因此平均时长趋于稳定.

再对每小时的平均使用时长进行分析, 每小时的平均使用时长计算方式如下:

$$\text{每小时平均使用时长} = \frac{\text{该小时所有用户使用时长总和}}{\text{该小时的用户总数}}$$

将 2 月 20 日到 9 月 20 日每天每小时的平均使用时长汇总再平均后结果如图 6 所示, 0 表示 0 点到 1 点之间, 以此类推. 白天时间段平均使用时长大约在 40 min 左右, 8 点达到最低值, 晚上平均使用时长较长, 凌晨 5 点达到峰值, 约为 50 min. 白天 6 点钟开始登录 VPN 人数逐渐增加, 使得平均使用时长逐渐降低, 当人数增加到一定数量后, 8 点钟开始, 平均使用时长开始回升, 但白天使用人数多, 频繁有人登录登出 VPN, 因此白天平均使用时长整体要比晚上低. 晚上因为 VPN 使用人少, 加上部分用户夜晚期间使用 VPN 进行长时间的数据传输和计算等, 因此平均使用时长较白天要长.

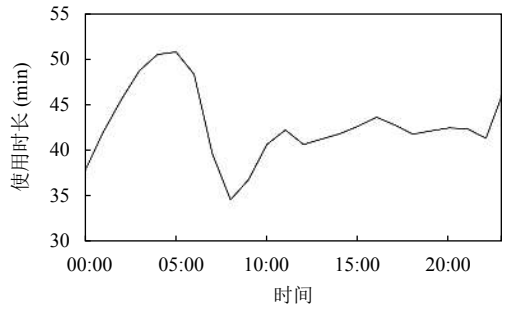

图 6 每小时平均使用时长

从另一个角度来看, 白天时间段使用人数多, 登录登出 VPN 人数也多, 也可能是人数太多 VPN 服务器压力较大, 稳定性不够好, VPN 会有自动断开的现象, 而夜晚使用人数少, VPN 稳定性较好, 使得平均使用时长夜晚比白天要高.

4.4 聚类分析

根据每个用户的每日平均使用时长 (使用总时长除以使用天数) 和使用天数对用户进行聚类分析, 了解 VPN 用户的分布情况, 聚类算法采用的是 K-means 方法. 将每日平均使用时长和使用天数作为 K-means 的输入, 对于 K 值的选择, 采用手肘法进行确定, 如图 7 所示为不同 K 值取值, 聚类误差 (各个点到其中心点的距离的平方和) 的变化情况, K<4 时聚类误差下降较快, K>4 时聚类误差下降缓慢, 因此可以取 K=4.

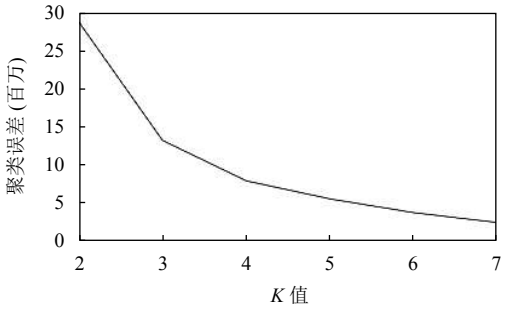

图 7 不同 K 值聚类误差的变化情况

聚类后的结果如表 3 所示, 第 1 类中接近 45% 的 VPN 用户在统计期间每日平均使用时长和使用天数均不高, 这部分用户 (例如学校本科生用户) 不需要长时间使用 VPN 进行校内资源的访问和获取. 第 2 类、第 3 类、第 4 类每日平均使用时长和使用天数逐步增加, 不同程度的依赖 VPN 进行远程学习和科研.

表 3 聚类结果

类别	每日平均使用时长 (h)	使用天数 (天)	账号数量	所占比例 (%)
1	1.48	7.17	24 223	44.30
2	2.50	35.65	14 321	26.19
3	3.35	77.1	10 401	19.02
4	4.46	139	5733	10.49

聚类结果可以让学校了解 VPN 用户的分布情况, 并以此为依据进行 VPN 资源的划分及调整. 例如, 对于每日平均使用时长和使用天数较长的用户, 需要划分的 VPN 服务器资源较多, 而第 1 类用户, 虽然账号数量多, 但是由于每日平均使用时长和使用天数都较少, 所以资源并不需要太多. 如果仅仅是按照用户数量来划分资源的话, 就可能会造成服务器资源的浪费.

4.5 用户类别

根据用户账号查询出用户所属院系, 如表 4 所示, 大致将用户分为文科院系、理工科院系、教工、其他、未知 5 类, 由于 VPN 用户数量较多, 情况复杂, 无法做到精确划分.

从平均使用时长和每小时平均使用时长对比理工科院系和文理科院系的 VPN 使用情况, 分别如图 8 和图 9 所示, 两者变化趋势基本一致, 但理工科要比文科略高, 平均使用时长整体高约 40 min, 每小时平均使用时长整体高约 3 min.

表 4 用户类别统计

序号	类别	所占比例 (%)
1	理工科院系	30.29
2	文科院系	23.47
3	教工	1.67
4	其他	3.34
5	未知	41.22

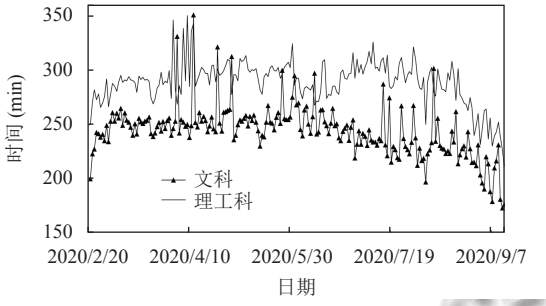

图 8 文科和理工科用户平均使用时长

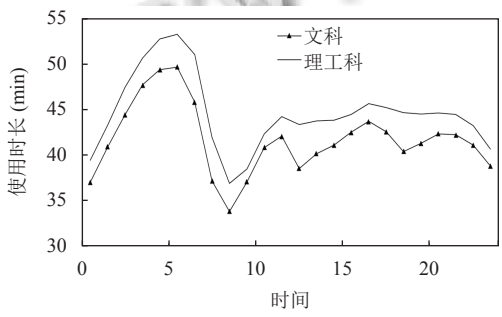

图 9 文科和理工科每小时平均使用时长

5 VPN 安全问题

校园网是一个相对封闭的环境, VPN 提供了一个进入校园网的通道, 在为用户访问校内资源提供便利的同时, 也带来了安全隐患, 比如永恒之蓝、挖矿病毒等, 都有可能通过 VPN 而进入校园内部. 弱密码和撞库攻击会导致用户账号被不法分子利用, 进入到校园内部, 对校内资源进行窃取或者攻击.

在对 VPN 日志分析的过程中, 发现存在同一个源 IP 地址对应多个账号, 同一个账号对应多个地理位置的情况. IP 对应的地理位置信息来自 IP2Location™ LITE IP-COUNTRY-REGION-CITY Database (https://lite.ip2location.com/database/ip-country-region-city). 经分析, 11.13% 的源 IP 地址对应了两个及以上的账号, 原因在于运行商为用户提供网络服务的时候使用的是动态 IP 地址, 不同的时间不同的用户可能拿到相同的 IP 地址, 再登录 VPN 时, 就出现了同一个源 IP 地址对应了不同的账号的情况. 56.63% 的账号对应了 2 个及以上的地理位置, 考虑到大部分师生会往返家乡及北京, 因此 2 个地理位置也是正常现象.

换个角度来看, 同一个源 IP 地址对应多个账号, 同一个账号对应多个地理位置还有可能是账号被盗用了. 盗用者在同一个网络环境下使用不同账号来登录 VPN, 以及盗用者与正常用户在不同地点登录 VPN 也会出现上述现象. 如表 5 所示, 为同一非公共账号地理位置变化情况. 3 月 28 日当天出现在了浙江和山东, 3 月 31 日和 4 月 1 日频繁出现在山东和内蒙古, 因此极有可能该账号已被盗用.

表 5 某账号地理位置变化情况

日期	源IP地址	地点
2020/3/28	xxx.xxx.70.6	山东
2020/3/28	xxx.xxx.5.114	浙江
2020/3/28	xxx.xxx.70.6	山东
2020/3/30	xxx.xxx.20.11	内蒙古
2020/3/31	xxx.xxx.146.88	山东
2020/3/31	xxx.xxx.20.11	内蒙古
2020/3/31	xxx.xxx.146.88	山东
2020/4/1	xxx.xxx.146.88	山东
2020/4/1	xxx.xxx.5.57	内蒙古
2020/4/1	xxx.xxx.146.88	山东
2020/4/1	xxx.xxx.5.57	内蒙古

以时间间隔为 1 天, 地理位置跨越省份为原则, 找出疑似盗用的账号, 将这些账号提交 VPN 管理员, 并且结合账号的身份、账号在其他系统中的使用情况等信息, 考虑是否对账号进行封禁处理.

以一周时间为例, 从学校部署的安全态势感知设备里面统计了 2020 年 6 月 15 日至 2020 年 6 月 21 日的告警类型来源地址分布情况, 如图 10 所示, 共有 12 339 条告警信息. 来源 IP 中 47.14% 的地址是 VPN 地址池中的 IP, VPN IP 告警数量占总告警的 54.46%.

VPN IP 告警数量前 10 的 IP 中, 排行第一的 IP 告警次数达到 3930 次, 告警内容包括频繁访问 445 端口 (每分钟超过 100 次)、MS17-010 永恒之蓝漏洞探测等. 通过 VPN IP 以及 VPN 日志找到告警时间段该 VPN IP 对应的账号信息, 如表 6 所示, 这些账号使用过的终端极有可能感染了病毒或者木马.

对于学校校园网来说, 校园网用户设备终端类型复杂, 操作系统繁多, 大多数缺少专人维护, 并且安全防护措施缺乏, 因此一些简单的病毒或者木马极容易

通过常见的系统漏洞进入到校园网内部进行扩散. 虽然能确定用户的账号, 但也无法对用户进行封禁处理, 因此 VPN 在保障正常的远程学习和科研的同时, 带来的安全隐患也不容忽视.

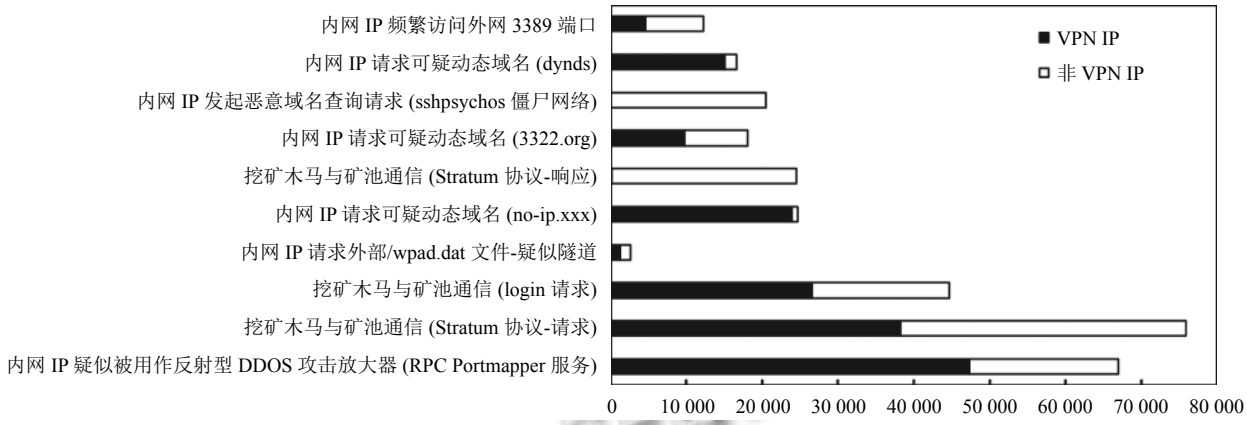

图 10 安全态势感知告警类型来源地址分布

表 6 告警数量前 10 的 VPN IP 及其账号

序号	VPN IP	告警	使用账号
1	xxx.xxx.xxx.81	3930	17xxxx41
2	xxx.xxx.xxx.143	3280	stxxxxp2和19xxxx89
3	xxx.xxx.xxx.2	1507	gkxxxx26
4	xxx.xxx.xxx.111	1267	stxxxxp3
5	xxx.xxx.xxx.183	289	18xxxx82
6	xxx.xxx.xxx.20	209	gsxxxx43和14xxxx26
7	xxx.xxx.xxx.60	198	16xxxx42
8	xxx.xxx.xxx.131	170	18xxxx22
9	xxx.xxx.xxx.227	134	17xxxx05
10	xxx.xxx.xxx.176	130	18xxxx23和14xxxx26

6 结论及展望

新冠疫情的暴发, 学生教工无法返校的情况下, 绝大多数高校采用 VPN 的方式保证远程学习和科研. 为了解具体情况, 采集了 2020 年 2 月至 2020 年 9 月疫情期间的 VPN 日志, 从使用人数、登录登出时间、使用时长、聚类分析、用户类别 5 个方面进行讨论. 新冠疫情期间, VPN 在线人数达到一个较高的水平, 最高使用人数接近 1.5 万一天, 最高同时在线人数达到 0.5 万. 从登录登出时间来看基本符合学生教工的学习和科研规律, 10 点、15 点、21 点出现登录峰值, 12 点、17 点、22 点出现登出峰值. 平均使用时长从 2 月份到 8 月份并没有明显的变化, 约为 250 min, 9 月份有所下降, 约为 200 min. 根据用户的每日平均使用时长和使用天数对用户进行聚类分析, 大致将用户分为 4 类. 对用户类别进行分析, 理工科用户 VPN 使用时间比文科用户略长, 但变化趋势基本一致. 以上这些数据, 可为 VPN 设备的负载优化、链路调整、资源分配提供指导, 对 VPN 设备选型也具有参考意义.

新冠疫情的暴发使得远程学习和科研普及开来, 但也伴随着一些问题的产生. 由于家庭环境和个人的电子设备安全防护措施做得不够, 同时用户对于电子邮件、视频会议等“虚拟”通信的依赖, 使得用户的终端更容易受到黑客的攻击, 如果攻击成功, 黑客就可以利用用户终端作为跳板窃取校内资源或者进行下一步的攻击. 弱密码和撞库攻击会导致用户账号被不法分子利用, 对校内资源造成威胁. 通过分析同一个源 IP 地址对应多个账号, 同一个账号对应多个地理位置的情况, 可以找到一些疑似被盗用的账号, 再结合学校部署的安全态势感知设备的数据来看, VPN 带来的安全隐患不容忽视.

总之, 疫情下的远程学习和科研非常考验高校的信息化水平, 而在远程学习和科研将成为新常态的趋势下, 是“甘饴”还是“毒药”? 高校都应该做好充足的准备来应对.

本文的不足之处在于, 仅分析了北京大学的 VPN 日志数据, 得出的结论有限; 分析过程中没有建立完善的数据分析模型, 分析数据之间的关联性; 提出了 VPN 的安全问题, 但并未做更深入的分析, 这些将是本文需要进一步研究的地方.

参考文献

1 白静. 抗疫斗争充分彰显中国治理能力和科技创新支撑——

《抗击新冠肺炎疫情的中国行动》白皮书发布. 中国科技产业, 2020, (6): 35–36.

2 Brynjolfsson E, Horton JJ, Ozimek A, *et al*. COVID-19 and remote work: An early look at US data. NBER Working Paper No. w27344, 2020. https://ideas.repec.org/p/nbr/nberwo/27344.html. [2020-10-16].

3 李汇. 远程办公需要注意的网络安全问题. 计算机与网络, 2020, 46(4): 52–53. [doi: 10.3969/j.issn.1008-1739.2020.04.049]

4 冯君贺, 汪晨. 基于 XDR+零信任架构的远程办公安全方案研究. 信息安全研究, 2020, 6(4): 296–300. [doi: 10.3969/j.issn.2096-1057.2020.04.003]

5 Yeboah-Boateng EO, Kwabena-Adade GD. Remote access communications security: Analysis of user authentication roles in organizations. Journal of Information Security, 2020, 11(3): 161–175. [doi: 10.4236/jis.2020.113011]

6 余慧佳, 刘奕群, 张敏, 等. 基于大规模日志分析的搜索引擎用户行为分析. 中文信息学报, 2007, 21(1): 109–114. [doi: 10.3969/j.issn.1003-0077.2007.01.018]

7 Mat-Hassan M, Levene M. Associating search and navigation behavior through log analysis. Journal of the American Society for Information Science and Technology, 2005, 56(9): 913–934. [doi: 10.1002/asi.20185]

8 Lu BB, Zhang HP, Liu B, *et al*. Research on user identification algorithm based on massive multi-site VPN log. Proceedings of the 17th IEEE International Conference on Communication Technology (ICCT). Chengdu: IEEE, 2017. 1372–1381.

9 武凌, 杨家桂, 陈劲松, 等. 基于 Hadoop 的 VPN 访问日志分析平台的研究与实现. 沈阳大学学报 (自然科学版), 2016, 28(6): 488–496.