NAT 识别调研报告

指导老师：杨嵘

郑延钦

2020.08

目录

1. 研究背景

随着 Internet 的迅猛发展，访问 Internet 已经成为普通人日常生活中不可或缺的重要组成部分，人们的需求不断地增长，网络互联规模不断地扩大，接入 Internet 的计算机数量持续猛增。然而，与此相对应的，全球 IP 地址资源匮乏的问题也日益突出，可用 IP 地址的数目现在已经明显不足，近乎枯竭。特别是在我国，问题尤其突出。由于我国人口众多，需求巨大，IP 地址早已成为十分紧缺的资源，显得非常捉襟见肘。

在如此环境下，作为用于暂时解决 IP 地址资源匮乏问题的过渡技术，NAT 技术应运而生。NAT 技术是由 IETF 制定的一个标准，属于接入广域网技术，它是一种将 IP 数据包中的内部私有 IP 地址与一个合法的公网 IP 地址进行相互转换的技术，实现了私有网络访问公共网络的功能。NAT 技术大大地减缓了 IP 地址空间枯竭的问题，它在当今的网络世界中正发挥着无可取代的重要作用，已成为众多公司、学校、政府部门，以至于一般用户的必然选择。

NAT 技术不仅有效地解决了 IP 地址资源匮乏的问题，而且由于建立了私有网络 IP 地址与公共网络 IP 地址间的映射关系，使得内外网隔离开来，从而将局域网中的所有计算机都隐藏并保护了起来，公共网络将无法直接对其进行访问，有效地防范了来自公共网络的各种非法攻击。因此，NAT 技术也为防火墙技术的发展提供了新的思路。使用 NAT 技术，未经授权的用户可以十分轻易地私自接入 Internet，并在网络上拥有一个属于自己的私有网络地址空间。对于网络的管理者或 ISP，也就是运营商来说，像这样的非法的私有网络地址空间的存在，给网络的正常运营带来诸多的困难，出现了大量严重的问题，比如网络整体的服务品质的下降、运营商网络运营总成本的增加、运营商网络服务费用难以正常地回收、合法用户的权益无法保证、合法用户的上网账号被非法盗用、正常授权合法用户的流失等等。而且，一旦有黑客利用 NAT 设备对外部网络中的计算机发动非法的攻击，对于运营商来说，其追查的难度将被远远加大，这给网络安全带来了巨大威胁。

因此，如何对这种非法的共享接入上网行为进行管理与监控，就成了放在运营商面前的一个迫切需要解决的问题。而要解决这个问题，首先就要对网络中所有使用 NAT 设备进行共享上网的用户进行识别。然而，由于 NAT 设备把一整个局域网隐藏到一个 IP 地址后面，将一个私有网络伪装成了一台单一的普通主机，私有网络中的所有主机相对于其他外网的网络设备来说都是透明的，私有网络的所有信息（包括网络的主机数量，各台主机的真实 IP 等）都是外部网络所不可见的，因此，如何有效区分一个公网 IP 地址对应的是单个的普通主机还是 NAT 设备，变得十分困难。基于净化网络环境、保障网络安全、有效地管理与监控网络的需要，必须要找到一种行之有效的 NAT 流量识别方法。

2. NAT 技术简介

NAT 网络拓扑如图 1 所示：

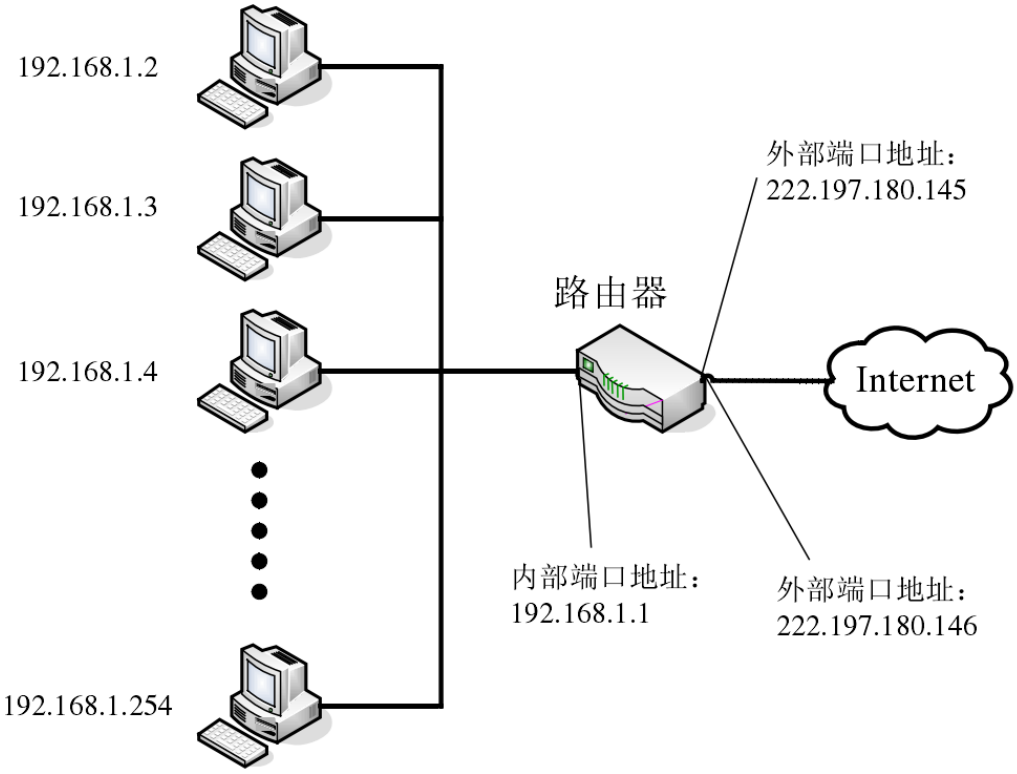

图 1 NAT 网络结构示意图

2.1. 静态地址转换

这是最为简单的一种方式。在此种方式下，私有网络内部的 IP 地址是与合法的公网 IP 地址一一对应的，属于一对一的转换方式。一个私有 IP 地址只固定地与其相对应的唯一的一个公网 IP 地址进行相互转换，这两个 IP 地址之间的映射关系一旦建立之后，就不会改变。

2.2. 动态地址转换

私有网络内部的 IP 地址与合法的公网 IP 地址间的映射关系是动态分配的，随机的，属于多对多的转换方式。NAT 设备维护了一个由多个合法的外网 IP 地址所组成的 IP 地址池，这些地址池中的 IP 地址将根据一定的策略动态地分配给内部网络的主机。当管理员授权允许内部网络的某台主机可以访问外网时，这台主机将向 IP 地址池申请一个合法的外网 IP 地址，申请成功后，NAT 设备将根据事先确定的策略从 IP 地址池中随机地选择一个外网 IP 地址，并在这个申请到的外网 IP 地址与这台主机的私有 IP 地址间建立起临时的映射关系，直到租用时间过期或用户主动断开连接之前，申请到的这个外网 IP 地址都将被这台主机所占用。

2.3. 端口地址转换

这是目前最为流行的一种转换方式，被广泛地应用于接入设备中。在此种方式下，私有网络内部的多个 IP 地址将映射到同一个合法的公网 IP 地址的不同端口上，属于多对一的转换方式，即端口多路复用的方式。当一台内部主机需要与外网进行通信时，NAT 设备会为这台主机分配一个唯一的端口号，然后在设备内部维持的一张端口转换表中将建立起内部 IP 地址、端口号与分配

到的一个外部端口号之间的映射关系，从而用端口转换的方式实现了内外网之间地址的转换。

3. NAT 识别方法概述

3.1. 基于特征字段的识别方法

3.1.1. IPID 识别法

IPID 是指 IP 数据包首部中的标识域字段，长度为两个字节，被用于唯一地标识该主机发出的每一个 IP 数据包，设置该字段的最初目的其实是为了便于分片重组，但现在 IPID 早已被操作系统当做一个计数器来使用了。在实际的使用中，不同操作系统有不同的 IPID 更新算法。对于使用 Windows 操作系统的主机来说，采用的是计数器累加和算法。使用这种算法，开机之后的 IPID 初始值一般默认为 0，随后其每发出一个 IP 数据包，不管这个包是属于哪一条流、哪一个连接，也不管是使用的什么协议，数据包中的 IPID 值都会递增 1。对于同一个网络中的不同主机来说，当它们同时访问网络时，其 IPID 值都是按照各自的序列独立地递增的，相互之间没有影响，而且由于每台主机启动的时间不可能完全一致，其各自操作的频率、活跃的程度、发包的时间都不一样，这些不同的主机产生的 IPID 序列理论上将各不相同。因此，通过统计分析同一个 IP 地址后发出的所有 IP 数据包的 IPID 值，绘制出 IPID 序列图，如果是 NAT 设备，那图中将会出现多条离散的 IPID 序列线，否则将只有一条孤立的 IPID 序列线，这样便可以直观地查看这个 IP 地址对应的是普通主机还是 NAT 设备，如果是 NAT 设备，通过查看这些离散序列线的条数便可以知道在设备后有多少台主机了。

该方法的缺点：IPID 识别法完全依赖于 IPID 字段的数值，因此不可避免会有许多的缺陷。首先，该方法的适用范围有限，该方法只适用于使用 Windows 操作系统的主机（Windows 操作系统采用计数器累加和算法，其 IPID 序列是线性的），而有些操作系统并不采用计数器累加和算法（而是采用随机、逆字节序、阶跃式等算法），这样将不会产生线性的 IPID 序列，而呈现乱序，IPID 识别法也就失效了；其次，当 NAT 设备后的不同主机在内网有大量数据的相互间通信，或者其使用多线程进行下载时，其 IPID 值都会增长，使得从外网获取的 IPID 序列图不再具有特定的规律，从而导致出现误判或漏判；再次，现在有许多 NAT 设备已经具有了修改 IPID 值的功能，它会将来自内网不同主机的数据包的 IPID 值进行修改，使其看起来像一条连续的 IPID 序列，从而将自己伪装成一台单独的主机，逃避 IPID 识别法的检测；另外，由于网络状况等因素，致使 IP 数据包发送与接收的顺序不一致，出现乱序时，同样也会导致误判；最后，在计数器累加和算法中，当 IPID 值增长到 65535 后，将归 0，并从头开始计数，这样一条连续的 IPID 序列分成了两段，将对 IPID 识别法的准确性带来影响。

3.1.2. TTL 识别法

TTL 值是 IP 数据包首部中一个 8 位字段，它表征了 IP 数据包的生存时间，其避免了数据包在网络中永不停止地传递下去。TTL 的初始值通常都是固定的，为系统缺省值，Windows 操作系统中一般为 128，其他操作系统通常设置为 64。IP 数据包在网络中的传递过程中，每经过一跳，路由器在转发数据包

时，TTL 值就会减 1。TTL 值实际上表示了 IP 数据包在传递过程中所经过的路由器的个数。因此，对于相同网络条件下的不同主机来说，NAT 设备后的私有网络主机发出的数据包的 TTL 值，将会比没有通过 NAT 设备的普通主机发送的小 1。通过检测相同网络条件下的不同主机发送的数据包的 TTL 值，便可判定其是普通主机还是 NAT 设备了，NAT 设备的 TTL 值会小 1。

该方法的缺点：TTL 识别法完全依赖于 TTL 字段的数值，同 IPID 识别法相类似，该方法同样受到来自操作系统的制约，不同的操作系统有不同的 TTL 初始值，将会严重影响到该方法的准确性；而且，现在有许多 NAT 设备已经具有了修改 TTL 值的功能，当 TTL 字段值被 NAT 设备修改后，该方法便会完全失效。

3.1.3. TCP 时间戳识别法

时间戳是 TCP 数据包中的一个选项字段，其表征了发送主机的开机时间。一般情况下，时间戳在刚开机时默认设置为 0，此后随着时间单调递增，其单位是毫秒，在系统重启后又重新置 0。在 TCP 协议中，发送主机可将时间戳放入 TCP 数据包中，接收主机在收到该数据包后，也在回复的确认包中放入这个时间戳数值，当发送主机收到该确认包后，再通过当前系统时间与时间戳来计算 RTT 值。因此，通过时间戳字段便可以获取到所有主机的开机时间，而不同主机的开机时间一般都有不同程度的差异，由此通过对来自不同 IP 地址的数据包时间戳与标准时间的差值的分析，便可以识别出哪些 IP 地址是普通主机而哪些是 NAT 设备了，同时也可确定 NAT 设备后的主机数量。

该方法的缺点：TCP 时间戳识别法完全依赖于时间戳字段的数值。因此，该方法同样受到操作系统的制约，一旦操作系统没有启用时间戳功能，那么该方法将会完全失效。

3.1.4. TCP 初始序列号识别法

TCP 协议为了实现数据的可靠传输，采用了三次握手来建立一个连接。其中，第一次握手时，客户端将发送一个 SYN 包到服务器，以建立连接。TCP 协议为了确认数据包发出的先后顺序，在 TCP 协议头中使用了一个 32 位的 TCP 序列号字段。在 SYN 包中，这个序列号字段实际上是 ISN。在 Windows 操作系统中，每隔一段默认的时间，ISN 值便会自动加上一个较小的数值。对于同一个网络中的不同主机来说，当它们同时访问网络时，其 ISN 值都是按照各自的序列独立地递增的，相互之间没有影响，而且由于每台主机启动的时间和 ISN 初始值不可能完全一致，这些不同的主机产生的 ISN 序列理论上将各不相同。因此，通过统计分析同一个 IP 地址后发出的所有 IP 数据包的 ISN 值，绘制出 ISN 序列图，如果是 NAT 设备，那图中将会出现多条离散的 ISN 序列线，否则将只有一条孤立的 ISN 序列线，这样便可以直观地查看这个 IP 地址对应的是普通主机还是 NAT 设备，如果是 NAT 设备，通过查看这些离散序列线的条数便可以知道在设备后有多少台主机了。

该方法的缺点：该方法与 IPID 识别法相似，只不过其依赖的特殊字段换成了 ISN，故缺点是类似的，比如不同的操作系统会采用不同的生成 ISN 值的算法，因而不同操作系统的主机会对 TCP 初始序列号识别法的准确性产生干扰。

3.1.5. OS 指纹信息识别法

对于不同的操作系统来说，其发出的数据包中都将有自身独特的指纹信息，这些可用于识别的指纹信息包括了 TTL 值、数据包首部未定义字段、初始 TCP 窗口的大小等。与 IPID 识别法和 TCP 时间戳识别法相类似，这种方法本质上也是要找出同一个 IP 地址的数据包相互之间的差异，并通过这些差异来判断这个 IP 地址是普通主机还是 NAT 设备，并且可通过统计这些数据包的指纹信息的差异来判断 NAT 设备后的主机数量。

该方法的缺点：OS 指纹信息识别法同样完全依赖于特征字段的数值，前述的那些问题同样难以避免。而且，假如 NAT 设备后的所有私有网络主机都使用的同一个版本的操作系统，该方法将无效。

3.1.6. CookieID 识别法

在 HTTP 协议中，为了在 UserAgent（一般为浏览器）与 Web 服务器之间有效地传输状态信息，便于网站辨别用户的身份，定义了一个 Cookie 数据值。当用户浏览某个网站时，Web 服务器将生成一个包含有用户 ID、时间日期等信息的 Cookie 值，并将该 Cookie 值连同用户访问的相应内容一并返回给请求访问的浏览器，浏览器则将其存储于用户本地的终端中。当下次该用户再次浏览同一个网站时，用户会将上次保存在本地的 Cookie 值一并发送，网站通过该 Cookie 值便可以得到用户信息了。一般情况下，对于首次访问该网站的用户，Web 服务器会在 Cookie 值中设置一个有效期（假如没有有效期，则默认在浏览器关闭之前有效）。在有效期内，同一个网站下不同的用户的 Cookie 值中的用户 ID 是不同的。因此，通过统计分析同一个 IP 地址后发出的所有 HTTP 请求数据包的 Cookie 信息，便可以直观地查看这个 IP 地址对应的是普通主机还是 NAT 设备了，如果是 NAT 设备，那么同一个 IP 地址访问同一个网站的 HTTP 数据包中将会有多个 CookieID 值，其具体的数量就是 NAT 设备后的主机数了。

该方法的缺点：CookieID 识别法很大程度上受到了用户的上网习惯的制约，很难保证在同一段检测的时间内不同的用户都访问了相同的网站，而且，有些用户会随时清理本机上的 Cookie 信息，这些因素都将使得 CookieID 识别法的检测误差相对较大。

3.1.7. User-Agent 识别法

User-Agent 是用户浏览器使用的一个特殊的只读字符串头，每当用户浏览某个网站时，浏览器发送的 HTTP 请求数据包的用户代理头中就包含该 User-Agent 值。Web 服务器通过该 User-Agent 值，便可以知道用户使用的操作系统及版本是什么、浏览器及版本是什么、CPU 类型是什么了。又由于同一个网络中的不同主机的操作系统及版本、浏览器及版本、甚至于打的补丁都不尽相同，因此通过统计分析同一个 IP 地址后发出的所有 HTTP 请求数据包中的 User-Agent 字段，便可以确定是普通主机还是 NAT 设备，如果是 NAT 设备，NAT 设备后有多少台主机了。

该方法的缺点：User-Agent 识别法很大程度上也受到了用户的上网习惯的制约，一些操作系统、浏览器的使用情况，如同一台主机打开两个不同的浏览器都会使其产生误判，而且由于 User-Agent 字段可以轻易地被用户修改，该方

法的检测效果没有保障。

3.2. 基于网络流量特征的识别方法

NAT 流量识别的问题，本质上是将网络中的 IP 地址划分为 NAT 设备与普通主机两大类的一个分类问题。该问题其实可以看作是数据挖掘技术中最为典型的二类分类问题，是两类数据挖掘技术——分类和聚类方法重点解决的问题。该方法的总体思想是获取网络中所有 IP 地址的网络流量特征，然后以各个 IP 地址作为数据挖掘中的实例，以其特征参数作为数据挖掘中的属性，通过数据挖掘将 IP 地址划分为 NAT 设备与普通主机两大类，进而完成 NAT 的识别。

分类和聚类的机器学习算法有很多，比如决策树、朴素贝叶斯算法、K-means 算法和深度学习算法。具体使用那种算法可以依据情况选择。该方法最重要的部分是在特征的选取上，这些特征参数在识别过程中起到的作用肯定并不完全相同，考察哪些特征参数对本方法的结果影响更大，无疑也具有很重要的意义。

3.2.1. 流量大小特征

总体来看，由于 NAT 设备后拥有多台主机，相比于一台普通主机，其网络流量应该会比较大。这可以看做是一个最为基本的 NAT 流量特征。因此，对于 NAT 流量的这一特征，流量特征参数将采用不同 IP 地址的网络流量的总报文数和总字节数来进行标示。但是，在实际网络环境下，这个特征不一定总是有效，因为一台普通的主机，其网络流量也可能很大，许多用户常常都会进行大数据量文件的下载，或者在线观看高清视频等，这都会导致一台普通主机的网络流量同样会比较大。因此，这一流量特征有时无法获得理想的结果。但通常情况下，由于一台普通主机的用户不可能长时间进行大数据量文件的下载，或者在线观看高清视频，因此大部分时间内一台普通主机的流量相对于 NAT 出口的总流量来说，还是要小很多。

3.2.2. 流的数目特征

由于 NAT 设备后拥有多台主机，相比于一台普通主机，NAT 网络的流的数目总体比较多。相对于 NAT 网络总的流的数目，一台普通主机同时传输的流相对较少。因此，流量特征参数可采用不同 IP 地址的网络流量的总的流的数量来进行标示。需要特别注意的是，在本文的数据准备工作中，由于 TCP 协议与 UDP 协议的不同连接特点，TCP 流与 UDP 流采用了完全不同的标准进行提取，二者之间具有很大的差异，不适合放在一起进行统一计数。因此，考虑到 TCP 流与 UDP 流的差异，将分别对其各自的流的数量进行统计。

3.2.3. 端口的数目特征

大多数 NAT 设备的公有 IP 地址都是十分有限的，通常只有一到两个，因此目前大多数 NAT 设备都是采用了端口地址转换（PAT）的方式。在这种方式下，内部网络的多个 IP 地址将映射到同一个合法的公网 IP 地址的不同端口上，因此，在内部网络的多台主机与 Internet 的通信过程中，NAT 设备通常总是使用了多个端口，而与此相反，一台普通主机一般情况下都只使用少数几个端口。通常情况下，拥有多台主机的 NAT 设备使用的端口的数量要远多于一台普通主机。因此，流量特征参数可采用不同 IP 地址下用于通信的总端口数来进行标

示。

3.2.4. TCP 连接数特征

由于 NAT 设备后拥有多台主机，不同的用户都在同时进行不同的 TCP 连接，相比于一台普通主机，NAT 网络的 TCP 连接应该更为频繁，其总的并发 TCP 连接数较多，而一台普通主机同时打开的 TCP 连接相对较少。因此，流量特征参数可采用建立和拆除 TCP 连接时通信的特定报文的数量来进行标示，如 FIN 报文的数量、RST 报文的数量、SYN 报文的数量等。

3.2.5. 跳变报文数量特征

由于 NAT 设备后拥有多台主机，在同一时刻，NAT 设备上可能拥有多个流在同时进行传输，因此 NAT 网络的网络流量会显得比较跳跃，随时不停地在不同的流之间进行切换，一条流的报文紧接着另一条流的报文，报文所属的流随时都在变化。而一台普通主机同时传输的流的数量相对较少，网络流量中的流不会发生频繁的跳跃式变化。因此，流量特征参数可采用一段时间内发生跳变的总的报文的数量与总报文数的百分比来进行标示。

3.2.6. IP 地址数特征

由于 NAT 设备后拥有多台主机，与其进行通信的网络设备就相对比较多，因此 NAT 网络流量的报文中的 IP 地址（上行流量里的目的地址、下行流量里的源地址）相对较多，而一台普通主机一般情况下同时只与一个或少数几个 IP 地址进行通信。因此，流量特征参数可采用一段时间内同一 IP 地址通信的总 IP 地址数来进行标示。但是，应当注意到，通常情况下 P2P 协议访问的 IP 地址也比较多，因此 IP 地址数这一特征可能会受到 P2P 流的干扰。当然，大多数时候，一台普通主机通信的 IP 地址数还是要远少于拥有多台主机的 NAT 设备。

3.2.7. 上下行流量差异特征

由于 NAT 设备后具有多台主机，这些主机各自的网络行为不可能完全一致，必定存在各种上传和下载的差异，当多台主机的流量合并到一起时，其总的上传和下载的差异将会在一定程度上被缩小，因此通常情况下，NAT 网络的上下行流量差异较小。而一台普通主机的网络行为则比较单一，短时间内上传和下载的差异通常较大，甚至可能只单独进行其中一种操作，因此其上下行流量差异较大。因此，流量特征参数可采用网络的上下行流量差异值与网络总流量的百分比来进行标示。

3.2.8. 网络空闲时间特征

由于 NAT 设备后具有多台主机，不同的用户访问网络的时间、访问网络的率、访问网络的操作等都不相同，因而 NAT 网络总体上会表现为持续稳定的流量，而一台普通主机的网络流量则具有突发性，与该用户的网络操作密切相关，大部分的时间几乎没有网络流量。因此，流量特征参数可采用一段时间内该 IP 地址没有网络流量的空闲时间长度与总时间长度的百分比，以及该 IP 地址没有网络流量的最大空闲时间这两种参数来进行标示。但是，应当注意到，如果一台普通主机的用户长时间进行大数据量文件的下载，或者在线观看高清视频时，总体上同样也会表现为持续稳定的流量，因此这一特征同样可能会受

到 P2P 流的干扰，有时可能无法获得理想的结果。

3.2.9. DNS 报文的数量特征

由于 NAT 设备后具有多台主机，不同的用户访问网站的数量、种类、频率都各不相同，因而 NAT 网络访问不同域名时发出的 DNS 请求应该会比较频繁，其 DNS 请求的总的数量会比较多，而一台普通主机短时间内不会产生较多 DNS 请求。因此，流量特征参数可采用 DNS 请求报文的数量来进行标示。

3.2.10. 具体特征总结

序号	特征参数	参数描述
1	total packets num	总的报文的数量
2	total packets sent	总发送报文的数量
3	total packets recv	总接收报文的数量
4	total packets bytes	总的报文的字节数
5	total bytes sent	总发送报文的字节数
6	total bytes recv	总接收报文的字节数
7	total TCP flow num	总的 TCP 流的数量
8	total UDP flow num	总的 UDP 流的数量
9	total port num	总的使用的端口的数量
10	FIN pkts num	总 FIN 报文的数量
11	RST pkts num	总 RST 报文的数量
12	SYN pkts num	总 SYN 报文的数量
13	jump pkts ratio	发生跳变的报文的比例
14	total IP address num	通信的总 IP 地址数
15	up-down diff num	上下行流量的报文数差值比例
16	up-down diff bytes	上下行流量的字节数差值比例
17	idle time ratio	无网络流量的空闲时间比例
18	maximum idle time	无网络流量的最大空闲时间
19	DNS REQ pkts sent	总发送 DNS 请求报文的数量
20	total http pkts num	总 HTTP 报文和 HTTPS 报文的数量
21	total http website num	总的访问网站的数量
22	TCP pkts num	总的 TCP 报文的数量
23	UDP pkts num	总的 UDP 报文的数量
24	TCP pkts bytes	总的 TCP 报文的字节数
25	UDP pkts bytes	总的 UDP 报文的字节数
26	TCP UDP diff num	TCP 与 UDP 报文的报文数差值比例
27	TCP UDP diff bytes	TCP 与 UDP 报文的字节数差值比例
28	TCP UDP jump pkts ratio	TCP 与 UDP 跳变报文的比例

4. NAT 识别总结

本次调研的 NAT 方法主要分为两大类，一类是基于特征字段的方法，有 IPID、TTL、User-Agent 等特征字段可以用于识别，另一类是基于网络流量特征的方法，通过分类或聚类算法对网络流量特征进行数据挖掘。

基于特征字段的方法比较传统，其本质是对 IP 数据报内的几个特征字段进行判断，但是大多数字段的特征不具有唯一解，即单个用户也有小概率可能可以操作产生类似 NAT 运行的结果，故具体实用性较差。较为可行的是基于 TTL 的识别，但目前存在可以修改 TTL 的方法，即可以人为躲避 NAT 识别。总体来说，该方法易于实现，但是也易于破解。

基于网络流量特征的方法比较新兴，是针对 NAT 上网时与个人上网时在网络流量上存在的偏差进行判定，在数据量足够的情况下，会有很好的识别效果，且其特征较难人为隐藏。总体来说，该方法实现会有较大开销，需要一定数据集来训练，但是其不容易被破解。目前还没有相应的开源数据集，自己造数据集其泛化能力可能较低。