Linewize
by Family Zone

TECHNICAL WHITE PAPER

Introduction to VPN-based filter avoidance

A technical guide to using a statistical approach to identify and block unwanted VPN traffic.

Contents

Introduction

For as long as network firewalls have existed, network engineers and vendors have been locked in a cat-and-mouse game with the creators of filtering avoidance applications. The traditional vendor approach is to constantly de-compile and dissect filtering avoidance techniques, looking to find implementation quirks that can identify this traffic. These quirks are then wrapped up in a network signature that uses specific protocol and packet criteria to identify and block unwanted network activity. At the same time, creators of filtering avoidance applications are doing the opposite, constantly evolving their software to get around firewalls and masquerade as innocuous network traffic.

Schools have seen widespread adoption of filtering avoidance software like VPNs, the use of which has now reached almost epidemic levels - to the extent that such behavior has become normalized among students in many schools.

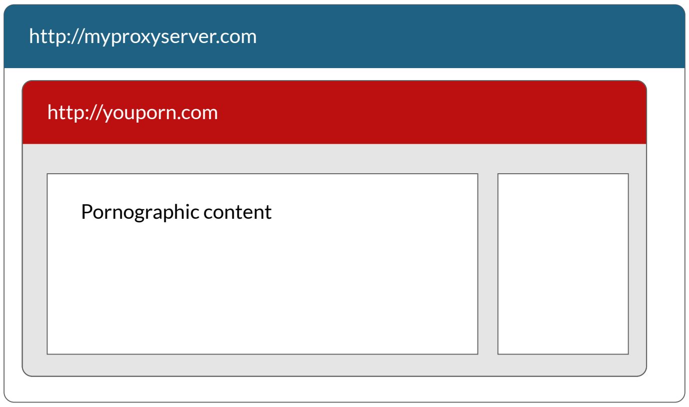
Educational institutions looking to maintain duty of care need to prevent students using filtering avoidance technology to access inappropriate content. This is traditionally performed using an application aware firewall; however, even firewalls that utilise Deep Packet Inspection (DPI) and secure content (SSL) inspection cannot keep up with the rapid pace of change in avoidance techniques, rendering them no longer fit for purpose.

Traditional technologies in use for filtering avoidance

Traditional approaches to filtering avoidance generally take one of two forms, and both exhibit easy patterns that can be identified by a network firewall.

In-browser proxy servers

In-browser proxy servers offer an easy method of viewing websites that would otherwise be filtered. By visiting a website, users are presented with a 'browser within a browser'. Here, at the cost of being subjected to aggressive advertising and substantial malware infection risk, users can enter the URL of another website and their traffic is tunneled through the root website's servers.



In-browser proxy server

By using this type of in-browser proxy, users are vulnerable to many risks - to privacy as well as to infection. As traffic is encapsulated in another website, the authors of that site have complete domain over any content delivered, and HTTPS provides no guarantee of privacy.

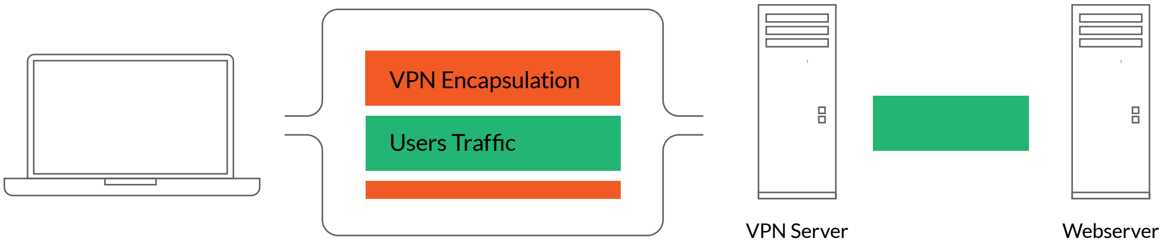
Fortunately for educational institutions, browser-based proxies are the easiest to identify and block. These sites can be identified using standard URL-based filtering and are easily tracked and identified.

Virtual private networks (VPNs) and proxy server tunnels

The second most common type of filter avoidance software is the traditional tunnel. Tunnels have many practical use cases, including granting 'road warriors' remote access to internal network services.

In schools, tunnels can be used to encapsulate internet traffic and route it through a remote server in a fashion that is encrypted and unable to be identified by traditional network content filters and firewalls.

There are several mainstream VPN and proxy servers that provide this functionality, including OpenVPN and Squid Proxy. Users of these services can create their own tunnel servers at home, or on cloud provided hardware like Amazon AWS. Then, by using a client or by configuring their browser, users route their traffic through the tunnel and the traffic content is encrypted.



Tunnel VPN or Proxy

Given that the central use of this technology is for security purposes rather than filter avoidance, this approach is remarkably easy to identify either at the OSI Layer 4 using ports, or with signatures employed with Layer 7 DPI technologies.

A new wave of filter avoidance technologies

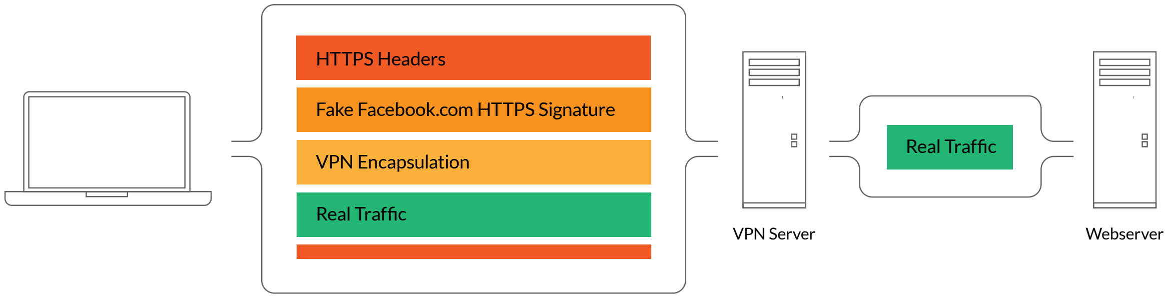
As outlined, both in-browser proxy servers and tunnel-based technologies are easy to identify and filter. To counter this, filter avoidance providers have developed new tunneling technology that is exceptionally difficult to block.

Driving this trend are two factors: oppressive governments that seek to restrict free internet access, and an online population increasingly concerned with privacy. Even customers of media companies such as Netflix are adopting this technology to access content that is not available in their home country.

This VPN-filter avoidance technology is mostly app based. Looking in the App Store you will find hundreds of them available for easy download and purchase. The big difference between these systems and their predecessors is that they masquerade as normal traffic. From a firewall's point of view, the traffic coming from these devices looks like normal, safe internet traffic.

Vendors like Hotspot Shield and Ultrasurf go to extreme lengths to encapsulate users' traffic in a form that masquerades as normal internet-encrypted HTTPS traffic. By doing this, firewalls that are employing DPI techniques will see what looks like normal traffic, and will not be able to distinguish between valid traffic and tunneled VPN traffic.

Some vendors claim to be able to protect against this type of masquerading by deploying client-side certificates to allow for inspecting HTTPS traffic. What they do not mention is that these VPNs often do not utilize the



Tunnel VPN or Proxy

certificate authorities normally installed on client devices, so employing a Man-in-the-Middle (MITM) attack will not identify it. Couple this with the cost of installing client-side certificates on BYOD devices and educational institutions are left in the dark.

This approach of masquerading as normal traffic in itself is brilliant, but it does not end there. With the advent of Software Defined Networking (SDN) and programmatically accessible cloud platforms like Amazon AWS and Microsoft Azure, it is now possible to automatically move and acquire new endpoints on an hourly basis. An entry-level developer can easily deploy VPN endpoints to new servers and release old ones when firewall vendors identify them.

This pushes the cat-and-mouse game to a new level of intensity that leaves traditional vendors and their quarterly signature updates far behind, unable to mount an effective response.

Using machine learning to identify VPNs

The constantly changing landscape of VPNs and filtering avoidance technologies has led Linewize to investigate new techniques of identification in combination with the traditional approach of creating DPI signatures. These education solutions include Linewize School Manager, providing new techniques that involve using machine learning and automated statistical analysis to identify new VPN endpoints on a minute-by-minute basis.

Machine learning and statistical analysis is the art of looking for patterns in large datasets that can identify common groups of behavior and isolate a specific target.

To construct our dataset, Linewize School Manager aggregates network traffic meta-data collected across hundreds of networks summarizing terabytes of network traffic. Within this dataset we look for behavioural patterns of traffic from filtering avoidance systems.

Pattern recognition applied to network traffic

Several statistical and network techniques have been combined to successfully identify target traffic. Specific implementation details are proprietary but the following provides an outline of the high-level approach used to identify VPN traffic.

Abnormal relationships

As unique as humans claim they are, the reality is that their behavior online is fairly predictable. When you start looking at browsing behavior, common trends linked to personas can be identified. Let's look at an example:

Student 1	Student 2	Student 3
facebook.com	youtube.com	facebook.com
youtube.com	facebook.com	emirates.com
nzherald.co.nz	stuff.co.nz	turkiye.gov.tr
Mathletics (app)	Khan Academy (app)	stuff.co.nz
iTunes (app)	Mathletics (app)	paypal.com
Coursera (app)	TedEx (app)	dogs.info

Looking at this dataset, it's remarkably easy for a human to identify which user is different. Student 1 and Student 2 have both been using Youtube, Facebook and some educational and news related websites. Student 3 is different and has been visiting some websites that are less commonly visited - and if we dig a little deeper it gets more interesting.

If we extend our profiling to 10,000 students rather than just 3 and start looking at these patterns we might find that it's very uncommon for a student that is using Facebook, Youtube and Coursera to visit emirates.com. Separately it also seems common among students visiting emirates.com to also visit turkiye.gov.tr and paypal.com. If we pivot the dataset a little, and look at data volumes we can see a closer pattern emerge.

Let's focus our dataset around paypal.com.

Student 1	Data Volume	Student 2	Data Volume	Student 3	Data Volume
paypal.com	10M	paypal.com	500M	paypal.com	1.2G
amazon.com	20M	buzz-zoom.info	1.2G	buzz-zoom.info	500M
stuff.co.nz	10M	whatsapp.com	20M	whatsapp.com	1.4G
trademe.co.nz	5M	stuff.co.nz	1.5G	stuff.co.nz	700M
iTunes (app)	1.2G	edxio.info	5M	paypal.com	12M
pbtech.co.nz	15M	dogs.info	10M	dogs.info	800M

From this dataset, it's easy to see that something is wrong. Student 1 has visited paypal.com and some other shopping related websites has transferred relatively tiny amounts of data to these sites. In comparison student 2 and 3 have no pattern of visiting other shopping sites that could be related to paypal and show huge amounts of data transfer.

For a human, the pattern is clear on a small dataset. For a computer, these sorts of trends can be identified at huge scale. This sort of analysis is great for identifying unusual student behavior, but more importantly it can be combined with other techniques to identify a VPN masquerading as something that it's not.

Common denominators between unlikely partners

Filter-avoidance applications typically tend to share infrastructure that would otherwise be unlikely to be shared. This rule can be used to build on other identifiers and map out infrastructure that VPN providers are using at a rapid pace.

When it is identified that dogs.info is a VPN then it can be inferred that other traffic going to IP addresses related to the same domain will also part of the VPN network.

dogs.info -> 11.2.2.2 <- cats.info

These types of inferred relationships can be used to train identification engines and preemptively block traffic.

High levels of unidentified traffic

Linewize School Manager identifies 98% of traffic, and uses real-time updates of classification information to stay ahead of new applications and websites.

With such a high probability that most traffic will be classified, searching for such traits as users with a high level of unidentified traffic is a rapid way of training the system to identify users that are utilizing filtering avoidance.

If Student 1 is active but traffic is going to only one IP address (perhaps registered to an ISP in Russia) then this is an indicator that something is wrong. If we then search our network flow database for other users with traffic to that IP, we might find other users that also appear to be only transferring data to this IP address. We can be almost certain that this IP address is part of a tunnel VPN network and this can be blocked immediately for all appliances across our schools.

Rapid analysis and constant updates

Linewize Appliances are in constant contact with the cloud platform and continuously receive updates, sometimes on a minute-by-minute basis. A key differentiator between Linewize School Manager and traditional vendors is that we use machine learning to allow our appliances to identify and filter content as it emerges and changes online. At scale, this means that even if your network is not subject to VPN or filter avoidance right now, you will still benefit from Linewize School Manager's ability to identify and block this content should someone emerge on your network with this technology installed.

The Solution

Even with machine learning techniques, identifying and filtering VPNs is still something of a cat-and-mouse game. However, with automatic identification and machine learning, and datasets spanning thousands of networks, it is possible to stay pinned to the mouse's tail.

Linewize School Manager is able to quickly and automatically identify new traffic types and effectively block elusive VPNs like Hotspot Shield and Ultrasurf. By combining real-time updates and machine learning technologies, Linewize School Manager provides visibility and control over unwanted VPN activity on your network.

About Linewize

Linewize is committed to making student Internet management easy, and keeping students safe online on any device, anywhere, any time.

Learn more

Email sales@familyzone.com
Visit us at linewize.com