Understanding Open Proxies in the Wild

Will Scott, Ravi Bhoraskar, and Arvind Krishnamurthy
{wrs, bhora, arvind}@cs.washington.edu
University of Washington

Abstract

This paper conducts an extensive measurement study of open proxies to characterize how much these systems are used, what they are used for, and who uses them. We scanned the Internet to track proxy prevalence and monitored public statistics interfaces to gain insight into the machines hosting open proxies. We estimate that 220 TB of traffic flows through open proxies each day, making them one of the largest overlay networks in existence. We find that automatic traffic taking advantage of multiple vantage points to the Internet overwhelms the traffic of individual 'end users' on open proxies. We present a characterization of the workload experienced by these systems that can inform the design of future open access systems.

1 Introduction

Web proxies are a widespread Internet phenomenon, but their usage is poorly understood. Many websites promote proxies as mechanisms for privacy, anonymity, and accessing blocked content. In addition, there is a vibrant community of open proxies, which offer access to content without requiring registration or payment. These proxies typically run on well-known ports and offer service through either the HTTP or the SOCKS protocol. While these systems can be seen as providing a valuable service to users, the motivation to run such a system is much less clear.

Open proxies are typically not discovered organically, but are generally found through the use of aggregators. These sites, like xroxy.com, hidemyass.com, and gatherproxy.com, curate lists of active proxies. Beyond simply monitoring uptime, these sites also provide metadata like geographic location, stability, proxy type, and connection quality information to help users choose 'good' proxies. Users can either directly access these aggregator sites or use a variety of browser extensions and client software to configure their proxy settings using data from the aggregators. While aggregators typically do not advertise how they discover their lists, we know that some are volunteer powered[12], some serve as advertisements for commercial services[8], and some accept user submissions of new proxies[18].

There are clearly some things that web proxies do well. They are accessed through an extremely minimal and simple interface and are supported on virtually every operating system. Further, the protocol for an HTTP proxy is simple enough that many different implementations have emerged, with regional communities forming around them. Web proxies also have a wide charter compared to many other protocols. The same software is used to offload popular requests from popular web servers, to reduce costs and validate traffic leaving organizations, and to improve the speed of accessing the web on airplanes.

While open proxies have been around for nearly 20 years [16], and a rich ecosystem thrives around their use, we know little about them. There are few verified statistics on how many open proxies are on the Internet, or how much traffic is served by these systems. We have even less insight into how these systems are used. This paper provides the first comprehensive measurement and characterization of the open proxy ecosystem. Through a combination of Internet scans, scraping of aggregators, and queries to open proxies themselves, we answer the following questions about the stakeholders in the open proxy ecosystem: (1) Who are the users of open proxies, and what do they use open proxies for? (2) Who are the operators of open proxies, and what are their motivations behind running them? (3) What are the characteristics of a typical open proxy server, in terms or load, stability and traffic? We find that many of these proxies are unintentional and short-lived, and that they serve a diverse set of users spanning legitimate organizations, users avoiding their local network, and a variety of automatic and malicious traffic. Further, we describe the observed traffic composition and geographical distribution, and provide case studies to represent the different situations that produce open proxies. We believe that answering these questions will help future overlays understand the traffic they are likely to receive, and spur education on more secure methods of indirection.

The structure of the rest of the paper is as follows. §2 describes the methodology we used when collecting data on open proxies - what data we collected, how, and what we did with it. §3 describes what we learned about open proxy servers in the wild, followed by a discussion in §4 of several specific proxy server instances. We then dive into traffic seen by proxies in §5 and §6 to analyze the behavior of proxy users and the workload seen by open proxies. In §7 and §8, we place these results within the context of related work and conclude.

Category	# Hosts
Port 3128 open	2,133,646
Identify as Squid	28,608
Open proxy	1,880
Open Squid proxy	943
Open Squid proxy with visible traffic	505

Table 1: HTTP proxies on port 3128. For comparison, aggregator sites typically list between 2,000 and 5,000 active proxies across all ports.

2 Methodology

2.1 Discovering Open Proxies

The first challenge in understanding the use of proxies on the Internet is to know where they are. We used two mechanisms to discover and track open proxy servers. First, we crawled aggregator sites once a week over our sample period to learn when open proxies were listed and removed from their indexes. Second, we performed our own probing of the full IPv4 address space using ZMap [4]. While it remains impractical to monitor all services on the Internet, there are a set of well known ports, like 3128 and 8080, which proxy software uses by default. Monitoring ports 3128, 8080, and 8123 allowed us to find what we believe to be the bulk of proxies in an efficient way. We performed individual snapshots on these ports using ZMap to find all open hosts, followed by a full request to see if the server functioned as an open proxy on the 2-3 million hosts which acked our initial TCP request. Scanning each port took us about 1 day as we rate-limited our activity to 50,000 packets per second in our initial probe. Our selection of these ports is backed by aggregator data, which report that the top 5 ports account for about 85% of known servers.

We find that while many hosts are listening on these ports, as shown in Table 1, only a small fraction of those services are open HTTP proxies. To determine if a host provides open proxy services, we attempt to load our department homepage, cs.washington.edu[1], and check to see if the expected title is included in the response. This step allowed us to efficiently filter our list to currently active open HTTP proxies.

Having thus built a pipeline to maintain a list of active open proxies, we can begin to understand the demographics and lifetimes of these services. Many of the commonly used proxy services advertise what software is used in HTTP headers, as shown in Table 2. We discuss more the breakdown of proxies, specifically where they operate, how long, and what software they run in Section 3.

[1]We actually request the IP address of the site, 128.208.3.200, to include servers which are unable to perform DNS resolution.

Port	Squid	Mikrotik	Polipo	Apache	Other
3128	943	232	0	16	689
8080	73	727	0	4	1424
8123	2	0	637	0	779
80	44	0	0	39	416

Table 2: Division of commonly used HTTP proxy servers on standard ports. Many of the unidentified proxies are believed to be instances of general-purpose web servers like Apache and Nginx.

Page	Description
menu	Cache Manager Menu
fqdncache	FQDN Cache Stats and Contents
http_headers	HTTP Header Statistics
info	General Runtime Information
objects	All Cache Objects
counters	Traffic and Resource Counters
client_list	Cache Client List

Table 3: Relevant resources provided by the Squid cache manager. These resources provide insight into both the clients and contents of a significant fraction of Squid open proxies.

2.2 Probing Open Proxies

Of commonly operating HTTP proxies, we notice that two of the most common, Squid and Polipo, include a management interface to their internals that is sometimes accessible from external requests. To understand the clients using these open proxies and the associated workload, we built a measurement infrastructure to monitor these management interfaces and capture information about recorded traffic. Our analysis in Section 5 focuses on understanding the data collected regarding proxy traffic.

The two programs, Squid and Polipo, offer overlapping information about the traffic they serve. Squid provides a cache manager feature, which was publicly accessible in about half of the discovered open Squid proxies[2]. The cache manager feature piggy-backs on the standard Squid HTTP proxy interface, but causes requests for specific URLs to be handled by the proxy and returned information about the proxy itself. For example, we can inspect Squid's DNS cache using this interface. By querying the cache manager, we collected data between April and October of 2014. Table 3 shows the available cache manager keys we queried for analysis and provides a sense of what data was available. In particular, the interface provides information about the cached objects (meaning URLs) in the proxy, the list of connected clients, and the cache of recently resolved domains.

In contrast, Polipo interprets requests to `/Polipo` as requests for information about the proxy. We find that

[2]http://wiki.squid-cache.org/Features/CacheManager

10% of Polipo proxies would respond to these requests, and provide us with information about their state. The Polipo information is more limited than Squid. It provides information on URLs visited and maintains a longer history and cache of these objects than Squid proxies, and it will provide information on connected servers and observed latency and throughput of those connections. Unlike Squid, Polipo does not reveal information about the clients using the proxy.

To be explicit, we believe that both of these interfaces are problematic, because users are generally not aware of their existence or their potential for surveillance. As such, we have informed the abuse contacts for discovered instances of these management interfaces, and requested that they either block access to the proxy or reconfigure their software to keep user information private. Further, as explained in Section 4, we have contacted both software vendors and individuals directly where they could be identified in order to help them fix these issues.

One factor mitigating the privacy and exposure risk presented by proxy servers publicly providing real time traffic information is that secure requests are not included in this information. When a user establishes a secure (HTTPS) connection through an HTTP proxy, they will use the CONNECT verb. In these requests, only the DNS lookup will be recorded, but the proxy will not know either client headers or the destination URL. In over 95% of the proxies we probed, the CONNECT verb was functional, and provided connectivity without revealing specific user intention.

Using data collected from cache management combined with discovery of open proxies, we can make inferences about the workings of the open proxy ecosystem. Open proxies are particularly interesting because they have an extremely low barrier of entry for usage, and process a workload similar to other open access systems, which are categorically difficult to observe. Proxies are not a new phenomenon, and anecdotally we know they are used as a light-weight mechanism to evade filtering in nation states, schools, and businesses. However, despite their ubiquity, the workload we discuss in Section 6 has remained under-characterized.

3 Open Proxy Servers

Once open proxies have been discovered, the next challenge is in understanding why they are operated. Unlike paid services where there is a financial incentive, or open access relays like Tor which may be run to support anonymity, there is no obvious answer to why one might run an open HTTP proxy. It is also unclear how expensive these services are to run, or even if the operators are actually aware that they are operating a service.

Country	ASN	Service Provider	#
UK	35662	Redstation Limited	102
DE	24940	Hetzner Online AG	74
US	16509	Amazon.com, Inc.	50
FR	16276	OVH SAS	48
CN	4134	Chinanet	46
CA	54718	Synaptica	46
BR	28573	Servios de Comunicao S.A.	42
PT	24768	Almouroltec, Portugal	40
BR	4230	Brasileira de Telecomunicaes	31
ID	17974	PT Telekomunikasi Indonesia	28

Table 4: Autonomous systems running open web proxies. Proxies are most dense in commercial data center subnets.

3.1 Open Proxy Diversity

One clue we can use to begin to break apart the open proxy system lies in the different software used to run open relays. The different proxy systems default to different ports (as seen in Table 2), and also appear to cater to specific localities - our snapshot showed 42% of Mikrotik proxies are located in Indonesia, Brazil and Russia while Polipo servers were almost entirely (90%) located in China. This is tempered somewhat by the many proxies marked as "Other", due to their lack of finger-printable headers.

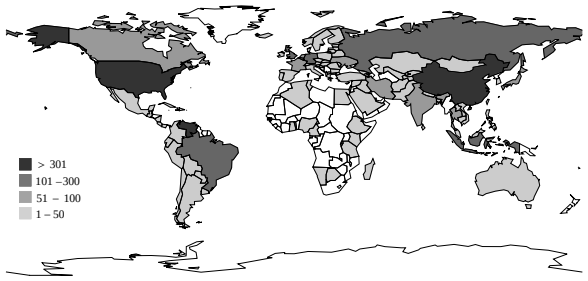

Figure 1: The geographical distribution of 4250 observed open proxies.

To look deeper at the locations of discovered proxies, we geolocated discovered IP addresses, shown in Figure 1. The most concentrated AS hosts are listed individually in Table 4. In this process we find that that the US has the highest concentration of open proxy servers, closely followed by Brazil and Venezuela. Our case study in Section 4 helps to explain the prevalence of proxies observed in South America. More generally, we observe that proxies appear to operate in locations with relatively cheap broadband access and relatively low liability associated with forwarding traffic for others. However, China and Russia also run large numbers of proxy servers, indicating a more complex picture. We also note that the top 4 ASNs where open proxy servers are located are those of large scale Infrastructure as a Service providers.

3.2 Open Proxy Lifetime

We next consider whether open proxies are primarily operated on purpose, or if they are largely the result of misconfiguration, we look at the observed uptimes and traffic loads of discovered proxies. Uptimes, shown in Figure 2 show a median proxy life of only 7 days. It would be preemptive to say that a transient lifetime implies that operation is unintentional. An alternative explanation could be that proxies are moved purposefully to avoid being blocked, or that they operate on hosts without fixed IP addresses. To rule out these possibilities, we look for a pattern where traffic to a proxy increases until it reaches a threshold where the operator notices it and fixes their configuration.

To help understand the distribution of load across proxy lifetime, we plot in Figure 3 how the total observed requests were distributed. Data from this plot was gathered from the 500 monitored squid proxies with open cache interfaces. The jaggedness of this plot represents an uneven distribution of traffic load across proxies. We do notice in this chart that that there is a relatively long tail of more stable proxy servers. These machines are likely run on purpose, since they remain stable over long periods of time. Secondly, we notice a large amount of requests handled by the small set of proxy servers which have been online for a significant amount of time. The inflection point at 60 days is a manifestation of several related IPs which all began operating at the same time.

We further characterized a sample of 100 open proxies in our dataset that have an uptime of over 50 days using NMap. 73 of the sampled proxies run Linux, two run Mac OS X, while 14 are routers or embedded devices. (NMap was unable to fingerprint the remaining 11.) 57 of the proxies run an SSH server, 15 run SMTP, and 45 run a non-proxy HTTP server on port 80. Additionally, we find that four of these machines have reverse DNS entries explicitly listing them as 'web caches', at least two of them are being run by commercial anonymity services, and at least two others are run from within universities.

Using collected load data, we can extrapolate the breadth of the open proxy ecosystem. We already see that our sampled servers transited a cumulative 10,000 requests per second. Figure 4 shows the distribution in request and response sizes observed by these proxies, notably featuring a mean HTTP response size of 31KB per request. This distribution indicates that our sample of measured proxies transit 300MB of data per second, equivalent to 27 TB per day. Extrapolating this sample to the full population of observed open proxies, we estimate that 220 TB of traffic is transited by open proxies on a given day. These numbers indicate that open proxies continue to service a significant amount of traffic. This magnitude of traffic is comparable to the roughly 168 TB of anonymous traffic exited daily by the Tor anonymity network in October, 2014 [14].

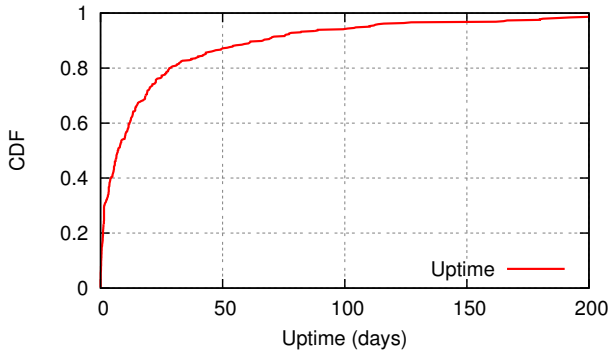

Figure 2: Observed Uptime. Half of all open proxies at a given time will have been up for less than a week.

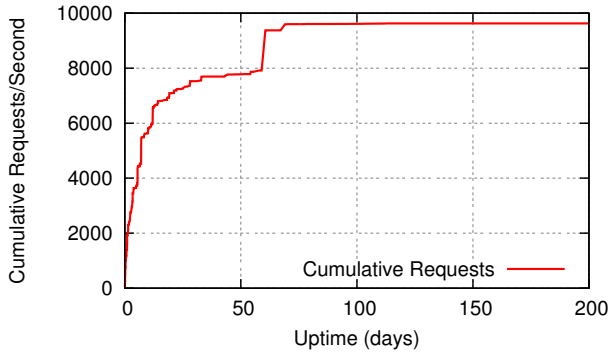

Figure 3: Requests handled as a function of uptime for monitored Squid proxies. The bulk of proxy traffic is transferred by transient proxies, but there are also more stable 'intentional' open proxies.

3.3 Open Proxies as Caches

Partially due to their transience, we found that the open proxies observed in practice showed significantly different caching characteristics from how Squid may have originally been imagined. Of the proxies reporting data, the average cache hit rate was only 2.5%, Significantly lower than the 10% hit rates achieved by previous caches [9].

The data provides two partial explanations for this low hit rate. First, we notice that the cache storage on the observed caches was smaller than previously estimated, with an average allocation of 70MB. Under 1% of proxies had swapped cache objects to disk. Secondly, we noticed that many requests include explicit instructions not to be cached. Table 5 shows the distribution of different cache control mechanisms as reported by the Squid servers. Note that this distribution is skewed by the number of requests which do not include any cache control header. Overall we find 48% of responses to be cacheable, which

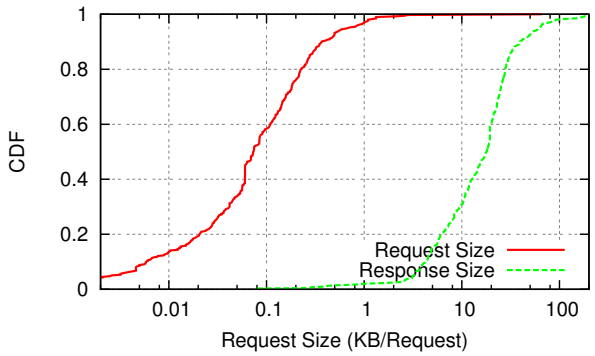

Figure 4: Resource size distribution across 5 billion observed requests and responses. Calculated as total traffic transited in a direction divided by total requests. Requests, as expected, are smaller than responses.

Cache Control	Count (millions)	Percentage
public	216	4%
private	1372	25%
no-cache	1493	27%
no-store	905	16%
no-transform	22	0%
must-revalidate	581	10%
proxy-revalidate	49	1%
max-age	924	17%
s-maxage	10	0%
max-stale	0.1	0%
stale-if-error	0.7	0%

Table 5: Observed cache control policies in server responses. 79% of responses with cache control headers were explicitly uncacheable, making up 48% of all HTTP responses.

is only somewhat depressed from the 60% number found by [5].

4 Proxy Operators

In this section, we present case studies of three long-running open Squid proxies in an attempt to highlight some of the motivations and ownership behind these devices. Given the nature of traffic we present next in §6, it seems likely that the operators of long running proxies are aware that they are running a relay. What they might not be aware of is that they're running a relay for the wider internet. These examples help to explain why they are okay with running a relay in the first place.

4.1 Spanish Small Business Server Linux Distribution

One common signature of open proxies we observed, which we believe partially explains the prevalence of Squid servers with exposed cache managers found in South America, came from a small business Linux distribution with Spanish localization. While this distribution does not have a Squid proxy enabled by default, when it is enabled its configuration has the cache manager interface enabled without firewall restrictions. This software is intended to be used by largely non-technical organizations looking to set up file sharing and related LAN services for their internal network. This makes it plausible that an administrator may check the 'HTTP Proxy' button without performing subsequent configuration.

The customers who enable the Squid proxy on this distribution may even use the proxy for their internal LAN, and will likely expect the machine to relay traffic for their organizations. This means that the traffic it proxies will not appear as unexpected behavior, meaning the fact it acts as an open proxy may (and in fact does, as we've shown) escape detection for quite some time. However, not all instances of the software were configured as open proxies; there were a significant number of open proxies which identified themselves as having been customized for this distribution.

The software vendor was responsive to our report, and is working to fix the configuration in the future. However, given the nature of their distribution, it does not sound like they will be able to directly contact or push a hotfix to their existing customer base.

4.2 Chinese Statistics Department

Another class of open proxies are run by individuals and appear to be primarily meant for personal services. One example of this class is a proxy server run on a server used for a course web page within a Chinese university. The value of such a proxy is two-fold - The Chinese educational network has a more relaxed policy than the consumer network, so a proxy would allow those associated with the proxy to more easily access foreign content from off campus. A second hypothesis is that when traveling abroad, a proxy back to campus allows access to local content and media that is inaccessible abroad. We found many personal proxies in many countries which appear to have been set up initially for personal access to content.

Many of these proxies are hosted at universities and on consumer ISP networks. While we might expect many of these proxies to be short lived, this particular proxy was running continuously for multiple months. Anecdotally, we can expect lifetimes of proxies to correspond with the network management policy of the organization where it is hosted. In many school settings, especially in countries where older, more vulnerable versions, of software are predominant, the network administrators give up on attempting to identify and stop individual 'malicious' hosts, instead preferring to mitigate malicious activity through DPI and routing policy that they can directly control.

After contacting the professor involved in this case, we were put in touch with the graduate student who adminis-

tered the computer. He had been unaware of the proxy's existence[3], but was proficient enough to ssh into the machine and with our guidance kill the proxy process. This interaction followed a pattern we saw several times in our investigation, where computers are set up with good intentions which due to subsequent configuration changes (e.g., giving the computer a public IP address in order to host a blog) result in security issues.

4.3 Russian Wi-Fi Hotspot

A final class of proxy we observed was a set of devices meant to provide public network access. This class includes proxies operated by anonymity services, along with more common place devices. One proxy we wanted to focus on is operated on a Russian Wi-Fi Hotspot. This is a gateway meant to provide public wireless Internet access at a hotel or restaurant, which has been configured such that it can be accessed from the whole Internet. The main page of the device allows you to register the MAC address it receives, and provides a set of default navigation links indicating it is operated by wifiroute.ru.

Here we find an instance where an organization is running a relay they expect to be public, and have filtering in place, along, presumably, with a robust abuse pipeline. The only unexpected aspect of this device is the scope of users receiving access. Looking at the device configuration of the traffic filter in question, we believe this is due to installation issues. The 'captive portal web server' is meant to be connected to a LAN as the gateway for connected wifi access points. In this case, the physical LAN interface of that device had instead been given a world-routable IP address.

From these examples, we can conclude that most of the open proxies running publicly online are not intentional. Rather, there are many legitimate reasons to be running a proxy, and it will not be immediately apparent if your infrastructure is also being used for public traffic relay. These examples also help to color the traffic distribution presented in the subsequent section. We reveal a mix of automated, malicious, and legitimate traffic, and these examples help to show users who are using open proxies for their legitimate traffic, and may not even know the proxy is open.

5 Characterizing Proxy Clients

With our newfound understanding of the reasons open proxies exist, we next consider the clients using these proxies. Anecdotally, we hear proxies discussed as ways to circumvent content blocking at schools or workplaces, to gain anonymity while browsing, and to circumvent country-wide censorship. From manual inspection of proxy content, we also notice that a significant amount of traffic appears to be automated, and we weren't too surprised to find evidence of use by spammers and botnets. Finally, amongst the client base we also expect to see legitimate traffic originating from the organizations or WiFi clients that don't realize the extent of their shared connection.

5.1 Client Geographic Diversity

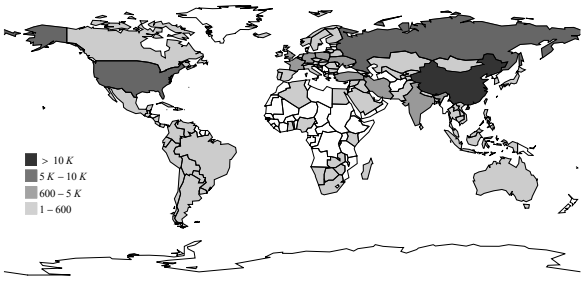

Figure 5: The geographical distribution of 51365 observed proxy clients.

Figure 5 shows the distribution of countries where clients are located. An extended view of this data is provided on our project website[4]. We observed a total of 51365 clients across the 505 monitored squid servers over a one week span. The two countries with highest number of clients were China and Russia, with 11922 (23%) and 6309 (12%) clients respectively. These figures are comparable to the 9000 daily Chinese IP addresses reported by VPN-gate[12], even without comparable mechanisms to avoid being blocked by the national firewall. It is unclear to what extent filtering is taking place on proxies. In total, the number of clients observed from countries employing "Pervasive Filtering" as defined by the OpenNet Initiative [13] was 13548 (26%), and this number goes up to 25886 (50%) when we broaden the countries to include ones employing "Selective Filtering". We also observed a significant number of users from western countries. USA, Canada and Western Europe accounted for 8408 (16%) of the clients observed. Western users can also gain a degree of anonymity while browsing using proxies, or this number may be biased by automated requests originating from data-centers in western countries. One specific attraction to western users may be that proxies can provide access to otherwise country-restricted content.

We also notice that some clients are observed connecting to many different proxies. A relatively small subset of 52 clients appear have connected to over 100 of the proxies we queried. 41 out of these 52 clients are from the same Chinese AS, identified as "CNCGROUP China169

[3]His initial response translates as: Hi, I am the system administrator for the site, but didn't know about that proxy. However, I don't study computer science, and don't really understand it. What should I do to solve the problem? Thanks!

[4]http://netlab.cs.washington.edu/squid

Backbone", and lie on the same 110.249.208.0/24 subnet. We discuss the types of activity that we believe are affiliated with these frequent clients in the Section 6.

5.2 Client Overheads

One common complaint made against anonymity and indirection tools is the overhead they impose on traffic. In order to understand the tradeoff of using an open proxy, it is interesting to consider the increased latency imposed by the proxy. To find this latency, we take two approaches. First, we find the locations of clients for each proxy to understand how much latency must be imposed on connections purely as a function of distance. Similarly, we geolocate destinations to understand how far the proxy will be from the destination content. Finally, we compare these numbers to estimates of geographic locality for the Internet in general to find the additional latency imposed by the proxy system.

Figure 6 shows the CDF of distances observed between proxies and clients. Distances are calculated as the geodesic distance between geolocated IP addresses. One interesting insight from this data is that under 20% of proxied connections involve a trans-Pacific link, countering the assumption that proxies are widely used to provide US Internet access from countries with filtered Internet. In fact, we find that only 28% of clients (3831 out of 13548 clients) in countries with "Pervasive Filtering" are seen on proxies in the US. Figure 6 also plots the distances observed between proxies and destination servers. In both cases we see that connections to proxies are on median closer to source and destination than if sources were to randomly connect to observed destinations on the same proxy (marked "Direct Distance"), showing that there is some degree of geographic locality, although it is relatively small.

Figure 7 provides an alternative view into the latency between the proxy and destination through use of reported statistics. The figure shows the CDF of HTTP request fulfillment times as reported by the statistics interface. The median latency is 0.52 seconds to load 31KB (Response size distribution is shown in Figure 4). The presence of proxies reporting zero page load time latency is an artifact of the latency running as a counter over the previous hour. Those proxies are ones which had not served requests recently. This request latency is consistent with the calculated latency overhead associated with indirection through a proxy. These numbers represent the times for individual resources to be retrieved, page load times will be much higher.

Our results also show that many of the assumptions underlying proxy software fail to hold today. We note that caching is largely ineffective for these systems, with an average 3% hit rate for the workload we observed. Indeed, performance is the least of the reasons why clients use

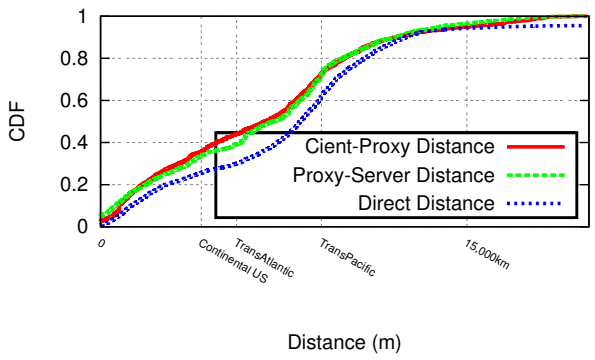

Figure 6: Observed distances incurred by proxying traffic. There was no preference observed for nearby open proxies.

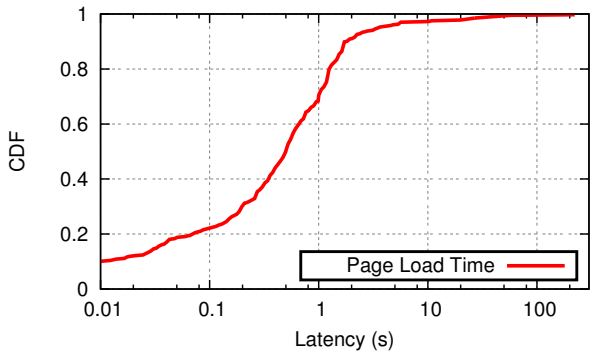

Figure 7: Observed page load times for proxies. Proxies reporting zero or minimal load time have not served requests within the previous hour.

open proxies, and current users have demonstrated willingness to trade performance for access to an alternative vantage point.

6 Open Proxy Workload

Having looked at the players involved in the open proxy ecosystem, the final and perhaps most interesting aspect of this phenomenon is to understand the workload and data accessed through open proxies. We have already alluded to some of the types of traffic observed in this study: forum spam and vulnerability scanning, legitimate user traffic with a biased geographical origin, and aggregators monitoring the health of the proxies themselves. In this section we try to make these traffic patterns more concrete by first characterizing the content overall, then focusing on specific details like observed search patterns, and finally offering specific examples of representative traffic patterns.

6.1 Traffic Distribution

For a high-level understanding of the open proxy workload, we first compare the observed traffic with overall

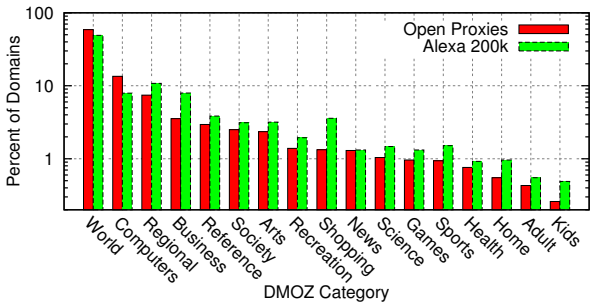

Figure 8: Distribution of top-level DMOZ categories of domains in our dataset vs Alexa Top 200,000. Open proxies appear skewed towards the long tail, with a stronger bias towards 'World' and 'Computer' related sites and away from 'Business' and 'Shopping' related sites.

Internet traffic rankings. In Figure 8, we cluster domains into the DMOZ[5] defined categories, and compare the distribution observed through open proxies to the distribution of the Alexa World top 200,000 domains. Open proxies showed a skew towards 'World' or international domains, which represents an over-representation of non-English language content. Likewise, these proxy users were more likely to look at non-commercial content, but were less likely to perform shopping or business through the proxies. Only one fifth of domains in the open proxy workload were indexed by DMOZ, and we caution that part of this distribution skew is due to the relative concentration of different categories. From looking at smaller Alexa lists we note that as more domains are included some categories, like 'World', trend higher, which indicates that this workload may not be as divergent from a typical web workload as it might initially appear.

Looking at how often domains were resolved by proxies, we also can generate a list of the most frequent domains. This top 10 list is provided in Table 6, along with a comparison to the Alexa Top 10. We notice several interesting things from this listing. First, the prominence of Bing is surprising. We find in practice that there were many cached results for Google queries, but many of them were made to a pre-resolved IP addresses. We also notice that several Chinese sites obtain more popularity in the open proxy list than their Alexa ranks would suggest, which can be partially attributed to the large number of Chinese clients using these systems.

Remember that this top 10 list is not based on number of accesses, but on how many proxies a domain has been visited through. What this highlights are automated scrapers, and not domains visited by normal users, since most individuals will make requests only using a single server at a time. An example of this is the inclusion of zennolab.com. This domain is a Russian product for blackhat SEO which obfuscates itself through the use of open proxies.

[5]www.dmoz.org is an open directory for URL categorization started by Netscape and maintained by AOL.

Alexa Top 10

Domain	Rank
google.com	1
facebook.com	2
youtube.com	3
yahoo.com	4
baidu.com	5
wikipedia.org	6
qq.com	7
twitter.com	8
taobao.com	9
amazon.com	10

Open Proxy Top 10

Domain	Rank
bing.com	1
baidu.com	2
scorecardresearch.com	3
google.com	4
blog.sina.com.cn	5
toolbarqueries.google.com	6
tieba.baidu.com	7
soso.com	8
ib.adnxs.com	9
chekfast.zennolab.com	10

Table 6: Most common domains found in the DNS caches of open proxy servers. The distribution shows a more international focus than is shown in Alexa rankings. Baidu and Bing each appear in 60% of the proxy server caches.

From this initial characterization of traffic combined with manual inspection of cached URLs, we focus on three specific slices of the proxy workload which appear to deviate from what would be expected of a general Internet workload. First, we notice that a significant fraction of traffic appears to involve search. We extract searches from the observed cache objects, and consider the topics of search and the fraction of 'human' searches next. We also notice unexpected regularity in other accesses, which we discuss in 6.3. Finally, we consider 'long sessions' to see if we can find evidence of end-user proxy use in 6.4.

6.2 Search Trends

To understand searches made, we first select all observed cached urls with a query string and a path involving 'search', or a parameter 'q'. We manually confirm that these selectors include the HTTP searches made to Google, Bing, Baidu, Yahoo, and Yandex. From monitoring the caches of Squid and Polipo for a one month period in the Summer of 2014, we collected 120 thousand searches across 291 hosts with exposed cache interfaces.

It was difficult to find a meaningful aggregation of these searches, given the limitations of what is revealed

Open Proxy Widest Searches

Search	Event	Servers
东海县桃林镇脱衣舞	Funeral Striptease (December 2013)	17
广州可悠信息有限公司		17
岳阳	City with Cake-centric corruption scandal (December 2012)	15
莒南		15
偃师		14
焦点访谈山西大同公路乱收费	Traffic corruption Scandal (October 2013)	13
google		13
宜宾		12
故宫游客群殴	Forbidden City Brawl (September 2013)	12
晋城宝马女打人	BMW Hit-and-run (November 2013)	11
yandex		10

Table 7: Common searches with context provided when known, and the number of proxy servers they appeared on.

by the presence of a URL in an open proxy cache. We can't directly claim popularity of a search from these metrics, and instead choose to use the metric of how many proxies we see a search cached on as a measure for how many distinct visitors are requesting a topic. This form of query aggregation results in Table 7. There are a couple immediate things that are indicated by these results. The first is that we have clearly captured a narrow sample of searches both temporally and contextually. Many of these queries which have been repeated on many different proxies are Chinese language, and many can be linked with scandals that have occurred over the last year.

This bias can be explained in a couple ways. One is that some entity is interested in understanding the popularity of these terms and is using proxies to actively monitor them. Looking specifically at the top term, a query related to a revelation of lewdness last December at a large Funeral party held in the town of DongHai, we find 430 instances of that search term cached by open proxies. This is because on many of the proxies we find the search has been repeated for the first 20 pages of results. Given our data we cannot however distinguish between queries made by unique individuals and those repeated by a single entity across many proxies.

Beyond the common set of Chinese terms which bubbled to the top, we can also see some queries that may be more suggestive of legitimate, widely made queries: searches for Google, Yandex, and other popular sites also ranked highly.

6.3 Flight Price Monitoring

We also notice another fraction of cached URLs which help to corroborate our supposition that the previous search results for sensitive Chinese terms were likely not due to many unique users but rather a result of an entity interested in monitoring coverage of the event from many vantage points. In the fall of 2014 there was a large increase in the listings of airline flights - requests to list flights on specific days between pairs of cities. These requests appeared across many proxies, and covered both US aggregators like expedia.com and orbitz.com and Chinese aggregators like ctrip.com and elong.com. However, these requests were also extremely regular in terms of both their starting time and url structure - more so than we would expect from legitimate user requests. In fact, the rate jumped from .03% of cached URLs over the summer to 8.1% of cached URLs in October, 2014.

Upon investigation, we found reference to a partially active instance of a Chinese PHP software interface titled "Plane ticket data crawling management system"[6]. This system provided monitoring of the status of open proxies loaded into the system, activity of slave machines making data requests, and configuration of the workload of aggregator sites, cities, and dates on which to collect information. The existence of this software has made us suspicious that much of the observed traffic is due to automated rather than actual user traffic.

6.4 Automated & 'Manual' Traffic

Visits	Servers	Domain
21494	1	oviprox:3128
12385	1	ftp.se.debian.org
2548	1	cdn.wikiimgs.com
6681	2	157.7.147.146 (Game Asset Server)
742	2	www.boswp.com
553	2	video.yandex.ru
882	3	www.google.kg

Table 8: Selected domains with many resources accessed through a low numbers of proxies. The top entry of each discussed type of access is reported.

While the previous understanding of URLs has considered popularity to be based on URLs accessed by many

[6]In Chinese: 机票数据抓取管理系统 .

proxies, we realize that this biases our analysis towards many of the illegitimate uses for proxies. Beyond URLs accessed from many vantage points, we also looked at domains accessed the most times from individual proxies. These 'long sessions' are another way of data slicing that we believe better reveals legitimate use cases. The top domains from this analysis are shown in Table 8. This view of the data shows a very different set of accesses from overall top domains. The first entry, `oviprox`, was an internal name of the proxy itself, and represents legitimate traffic to the site which was operating as an open proxy. We see that software updates and CDNs (entries 2-4) also rank highly in this ordering, since a single page will include many cacheable files. `bowswp.com` is a mobile proxy for Windows phones. A significant amount of legitimate browsing activity is also visible when content was sorted in this way, including examples of searching, news, video watching, video games, and pornography. We find that overall 50% of domains were visited on less than 40 of the 500 observed proxies, while the most frequent domains were cached by 300 of the proxies.

7 Related Work

Web proxies have existed for almost as long as the Internet itself. The CERN HTTP server in 1994 was the first to feature caching, and subsequent work like Harvest [1] optimistically predicted 50% hit rates in such caches. Squid itself is a derivation of Harvest and IRCache [10]. In 2004 at the conclusion of the IRCache project, they reported a 25%-50% cache hit rate.

Recent work on understanding web traffic, such as CoDeeN [9], have been more pessimistic about the effectiveness of caching a the web workload. CoDeeN find an ideal HTTP hit rate of 15%-30% of traffic with an infinite cache size and handle 5%-10% of requests with a 500M cache.

CoDeeN provides important insight into what the view of the web through an open proxy looks like. In particular, CoDeeN has provided a valuable view into how content on the web has changed as web pages have evolved into full fledged applications. Our analysis of content usage corroborates these observations of the growing sizes of individual resources in recent years. However, CoDeeN does not ask the question of who is using open web proxies. Neither do they analyze what types of content are being requested in their workload, and how that compares to the broader Internet. In these regards, our work sheds additional insight into why CoDeeN saw the distribution of traffic that it did.

A recent measurement study of proxies [15] found that 14% percent of traffic to the Netalyzer traffic analysis tool was proxied in some way, which confirms the prevalence of proxies. This analysis attempts to understand the purpose of these proxies based on how they modify traffic, and find a combination of caching, network security, captive portals, and malware. However, they do not measure traffic through these portals, nor do we expect the proxies which passively capture machine traffic from an internal LAN, the majority of what was measured, to overlap with the open proxies we measure.

Many insights into Internet traffic have come from the differences found in specialized workloads. In addition to studies of proxies and CDNs[9, 6], researchers have looked at the traffic workloads originating from universities, businesses, and developing countries [17, 7]. This work however has primarily focused either on the caching techniques in use and how to optimize performance of middle boxes, or on characterizing the workloads observed in these specific instances. [2] measured the distribution of web proxy traces to compare whether they follow Zipf's Law. Other work has used analysis of proxy servers to measure criminal activity [3].

There have also been attempts at characterizing the traffic of anonymity networks like Tor[11]. We note that open proxies provide more geographically diverse vantage points than exiting through Tor. (The Tor network currently comprises just over 1000 exit nodes operating in 66 countries, while open proxies are found in over 100 countries.) We also understand the motivations behind many of the operators of Tor nodes, an outcome of the network's clear focus on anonymity.

8 Conclusion

In this paper we have tried to clarify the nature of open proxies on the Internet today. We have offered evidence that these proxies continue to traffic a large amount of data, and help to explain who operates them, use them, and why. We expect this usage characterization will prove useful for developers of anonymity and privacy preserving systems. We believe this is the first characterization of usage patterns for systems offering anonymous Internet access, and provides some insight into the motivations of end users of these systems, and the forms of abuse such systems can expect. Perhaps even more importantly, we demonstrate what a passive observer can learn about visitors of open proxies, and hopefully motivate a shift towards more secure forms of anonymous traffic indirection.

9 Acknowledgment

We would like to thank Raymond Cheng and Antariksh Bothale for their help, and the uniformly responsive proxy operators we contacted while conducting this study.

10 References

[1] C. M. Bowman, P. B. Danzig, D. R. Hardy, U. Manber, and M. F. Schwartz. The harvest information discovery and access system. In *WWW*, 1994.

[2] L. Breslau, P. Cue, P. Cao, L. Fan, G. Phillips, and S. Shenker. Web caching and Zipf-like distributions: Evidence and implications. In *INFOCOM*, pages 126–134, 1999.

[3] C. Castelluccia, M. A. Kaafar, P. Manils, and D. Perito. Geolocalization of proxied services and its application to fast-flux hidden servers. In *Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference*, pages 184–189. ACM, 2009.

[4] Z. Durumeric, E. Wustrow, and J. A. Halderman. ZMap: Fast Internet-wide scanning and its security applications. In *USENIX Security*, 2013.

[5] J. Erman, A. Gerber, M. T. Hajiaghayi, D. Pei, and O. Spatscheck. Network-aware forward caching. In *Proceedings of the 18th international conference on World wide web*, pages 291–300. ACM, 2009.

[6] M. J. Freedman. Experiences with coralcdn: A five-year operational view. In *NSDI*, 2010.

[7] S. D. Gribble and E. A. Brewer. System design issues for internet middleware services: Deductions from a large client trace. In *USENIX Symposium on Internet Technologies and Systems*, 1997.

[8] Hide My Ass — Free Proxy and Privacy Tools. http://hidemyass.com/.

[9] S. Ihm and V. Pai. Towards understanding modern web traffic. In *IMC*, 2011.

[10] The ircache project, 1995. http://www.ircache.net.

[11] D. McCoy, K. Bauer, D. Grunwald, T. Kohno, and D. Sicker. Shining Light in Dark Places: Understanding the Tor Network. In *PETS*, 2008.

[12] D. Nobori and Y. Shinjo. VPN Gate: A Volunteer-Organized Public VPN Relay System with Blocking Resistance for Bypassing Government Censorship Firewalls. In *NSDI*, 2014.

[13] OpenNet Initiative. https://opennet.net/.

[14] Tor metrics: Bandwidth. https://metrics.torproject.org/bandwidth.html.

[15] N. Weaver, C. Kreibich, M. Dam, and V. Paxson. Here be web proxies. In *Passive and Active Measurement*, pages 183–192. Springer, 2014.

[16] Wingate proxy server. http://www.wingate.com.

[17] A. Wolman, G. Voelker, N. Sharma, N. Cardwell, M. Brown, T. Landray, D. Pinnel, A. Karlin, and H. Levy. Organization-based analysis of web-object sharing and caching. In *USITS*, 1999.

[18] XROXY.COM — More than just a proxy. http://xroxy.com/.