Shining Light in Dark Places: A Study of Anonymous Network Usage

Damon McCoy, Kevin Bauer, Dirk Grunwald, Parisa Tabriz, and Douglas Sicker

CU-CS-1032-07 August 2007

University of Colorado at Boulder

Technical Report CU-CS-1032-07
Department of Computer Science
Campus Box 430
Boulder, CO 80309

Shining Light in Dark Places: A Study of Anonymous Network Usage

Damon McCoy[1] Kevin Bauer[1] Dirk Grunwald[1] Parisa Tabriz[2] Douglas Sicker[1]

[1]University of Colorado at Boulder [2]Google

ABSTRACT

To date, there has yet to be a study that characterizes the usage of a real deployed anonymity service. In this paper, we present observations and analysis obtained by participating in the Tor network. In particular, we are interested in answering the following questions: (1) Who uses Tor? (2) What is the performance of the system? (3) How is the system used? (4) What does the traffic distribution look like? and (5) What are the legal and ethical implications of participating in an anonymous network?

We show that the network is used to fight censorship worldwide. In addition, the system's performance is characterized at the circuit-level and we show that the network traffic can be closely modeled by a Pareto distribution. Finally, the legal and ethical issues that arise from participating in such an anonymous network are discussed.

1. INTRODUCTION

Tor is a popular multi-hop privacy enhancing system that is designed to protect the privacy of Internet users from traffic analysis attacks launched by a non-global adversary [13]. Because Tor provides an anonymity service on top of TCP while maintaining relatively low latency and high throughput, it is ideal for interactive applications such as web browsing, file sharing, and instant messaging. Since its initial development, several researchers have analyzed the system's security in an attempt to understand the degree of privacy that it provides [15, 19, 20, 21]. However, there has yet to be a study aimed at understanding the more practical "real world" aspects of such a privacy enhancing system. In this work, we utilize observations made by our own Tor server to answer the following questions:

Who is using Tor? The design and development of anonymous systems such as Tor are motivated to service people residing in locations where free and uncensored access to the Internet is not guaranteed. Thus far, however, there has been no empirical study describing who the Tor users are and whether or not they match this target audience. For example, many governments around the world actively censor their citizens' Internet access and it has been suggested that services such as Tor can be used to help individuals in such places resist local censorship. To determine if this usage exists, we characterize the geographic distribution of both clients ("Tor proxies") and servers ("Tor routers") in Tor to determine from what parts of the world Tor clients originate and whether this differs from the distibution of Internet users world-wide.

How is Tor being used? It has been suggested that Tor's optimizations for low latency and high throughput traffic have made it ideal for interactive applications [13]. To prove or refute this, we analyzed application layer header data relayed through our router to determine the protocol distribution in the anonymous network. Our results present the types of application use currently sent through Tor, a substantial proportion of which contributes non-interactive traffic through the system. In addition, this study analyzes destination server locations to identify trends in the destinations of exit traffic.

How do the circuits in the network behave and perform? Clients use the Tor network by source routing traffic through multi-hop paths, by default three hops. Since one of the commonly acknowledged costs of anonymity is performance loss, the biggest questions are: How much latency do Tor circuits incur? How much throughput can circuits maintain? Is Tor use heavier at certains times of day? For how long does a client use the same circuit? How much traffic does a circuit transfer over its life time? Our data shows that Tor does, in fact, provide a relatively low-latency transport service. However, throughput is also relatively low and most circuits are short-lived and transport very little data.

What does the distribution of traffic look like? Traditional analytical models of privacy enhancing technologies have often assumed that the distribution of traffic across the network is uniform. Given Tor's optimizations for performance, the distribution is skewed, yet it is unclear how to build an accurate analytical traffic model. We show, based upon our observations, that the network traffic can be best characterized by a Pareto distribution.

What are the challenges involved with participating in Tor? As a final avenue of study, we attempt to expose the challenges that are associated with participating in a privacy enhancing system from legal and policy perspectives. As a consequence of Tor's design, Tor server operators

can be held responsible for illicit traffic that flows through their node when it is used as an exit node. We experienced many such difficulties running our own node and were eventually forced to discontinue our service.

The remainder of this paper is organized as follows: In Section 2, we present a brief overview of the Tor system architecture. Section 3 describes our data collection methodology. In Section 4, we analyze the geopolitical location distribution of Tor proxies, routers, and destination servers. We report observations at the circuit level in Section 5, including circuit latency, throughput, and duration. Section 6 explores the application-level protocol distribution for traffic leaving the Tor network. In Section 7, we measure the internal network traffic distribution. Section 8 explores the legal and ethical questions and arise from participating in an anonymous network. Finally, we conclude in Section 9.

2. BACKGROUND

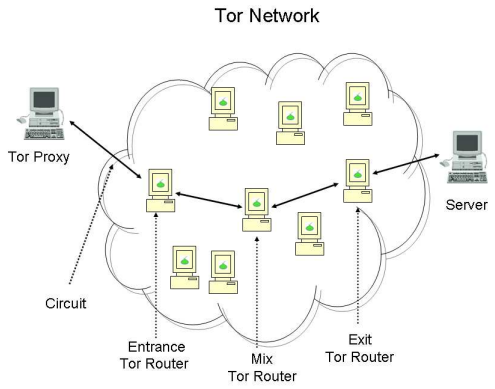


Figure 1: An overview of the Tor system architecture.

Tor's system architecture attempts to provide a high degree of anonymity and strict performance standards simultaneously [13]. At present, Tor provides an anonymity layer for TCP by carefully constructing a multi-hop path, or *circuit*, through the network of Tor servers, or *Tor routers*, using a layered encryption strategy known as *onion routing* [16]. There are precisely three hops in a circuit; the first node in the circuit is known as the *entrance Tor router*, the middle node is called the *mix Tor router*, and the final hop in the circuit is referred to as the *exit Tor router*. It is important to note that only the entrance router can determine the originator of a particular request through the Tor network, and only the exit node can determine the contents and destination of the request. To achieve its low-latency objective, Tor does not explicitly reorder or delay packets within the network. An overview of the Tor system architecture is given in Figure 1.

3. DATA COLLECTION METHODOLOGY

To better understand real world Tor usage, we set up a Tor router that joined the currently deployed network. Our router was configured to use the default exit policy, which allows most exit traffic to leave our router. The server hosting our router was connected to a 1Gb/s network link. This configuration allowed us to record a large amount of Tor traffic in short periods of time. While running, our node was consistently among the top three routers in terms of bandwidth of the roughly 800 routers present.

We understand that there are serious privacy concerns that must be addressed when collecting statistics from an anonymous network. We considered the privacy implications when choosing what information to log and what information was too sensitive to store. In the end, we chose to log information from two sources: First, we altered the Tor router to log information about circuits that were established though our node and cells routed through our node. Second, we logged only protocol header information from exit traffic that was relayed through our node. This logging was handled using `tcpdump` [23], a common protocol analysis tool.

We ran our Tor server for two periods of four days each, in December 2006 and January 2007. During these data collection periods, our Tor server participated in over 2.3 million circuits, and relayed approximately 1 terabyte (TB) of exit traffic alone. This traffic represented a large sample of Tor network traffic, and we believe that our data collection over a total of eight days is an indicative sample of normal Tor usage.

3.1 Tor Router Level Logging

Our router used Tor software version 0.1.1.25 with our own minor modifications to support logging. For every cell routed through our node, we logged the time that it was received, the previous hop's IP address and TCP port number, the next hop's IP and TCP port number, and the circuit identifier associated with the cell. We did not capture the payload of the cell, any of the cryptographic keys, or reassembled TCP packets from exit traffic. While retaining a record of this information could degrade Tor users' anonymity and is not recommended, the system is designed to resist traffic analysis attacks from any individual Tor router. Thus, the information we logged can not be used to link a sender to a receiver.

3.2 Exit Traffic Logging

In order to gather statistics about traffic leaving the network, we ran `tcpdump` locally on the same physical server as our Tor router. `Tcpdump` was configured to capture only the first 150 bytes of a packet using the "snap length" option (`-s`). This limit was selected so that we could capture up to the application-level headers for protocol identification purposes. At most we captured 96 bytes of application header data, since an Ethernet frame is 14 bytes long, an IP header is 20 bytes long, and a TCP header with no options is 20

bytes long. We used `ethereal` [3], another tool for protocol analysis and stateful packet inspection, in order to identify application-layer protocols. As a post-processing step, we filtered out packets with a source or destination IP address and TCP port number of all active routers published during our collection periods. This left only exit traffic, since our server was not running any other network services.

4. GEOPOLITICAL DISTRIBUTIONS

There has been much speculation that privacy enhancing systems such as Tor can be used as a tool to fight Internet censorship, particularly from within countries that actively monitor or filter their citizens' Internet access. As part of this study, we investigate where, *geo-politically*, Tor proxies, Tor routers, and destination servers are located. Recall that a proxy's IP address is visible to a router when that router is used as the entrance node on the client's circuit through the Tor network. Similarly, a destination server's IP address is visible to a router when that router is used as the exit node. Tor router IP addresses are maintained by the Tor directory servers, and we keep track of the router IP addresses by simply polling the directory servers periodically.

In order to map an IP address to its corresponding country of origin, we query the authoritative bodies responsible for assigning IP blocks to individual countries [1, 2, 4, 7]. In order to determine the geopolitical distribution of Tor usage throughout the world, we aggregate IP addresses by country, and present the proxy, router, and destination server location distributions during our two data collection periods.

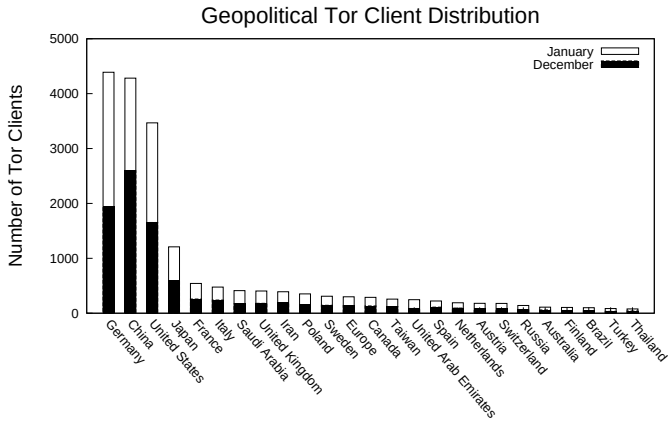


Figure 2: Geopolitical Tor proxy distribution during the December and January data collection periods. IP addresses assigned to the European Union are denoted as *Europe*.

4.1 Geopolitical Tor Proxy Distribution

During both data gathering periods, the client distribution was similar; the United States, Germany, and China composed the top three countries in terms of number of unique clients, having 1,653, 1,940, and 2,597 Tor clients, respectively in December and 1,815, 2,449, and 1,686 clients, respectively, during the January data collection period. The Tor proxy distribution is provided as a histogram in Figure 2.

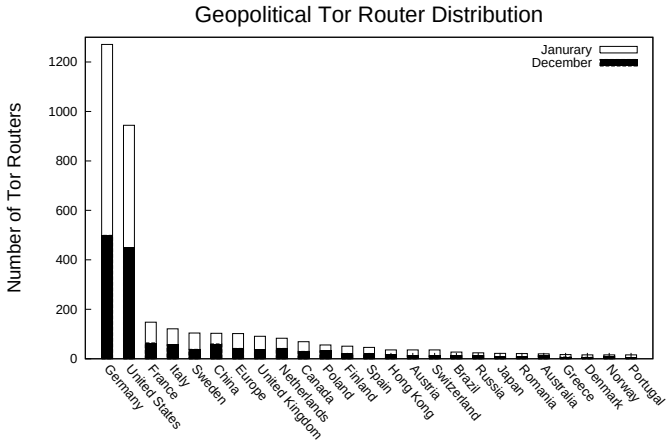


Figure 3: Geopolitical router distribution during the December and January data collection periods. IP addresses assigned to the European Union are denoted as *Europe*.

4.2 Geopolitical Tor Router Distribution

In addition to analyzing client locations, we study the geopolitical distribution of Tor routers. During our data collection periods, we maintained a list of unique Tor server IP addresses, as provided by an authoritative Tor directory server. During the December data collection period, there were 499 routers in Germany and 450 in the United States. During the January data collection period, there were 772 routers in Germany and 494 in the United States. The router distribution is provided in Figure 3.

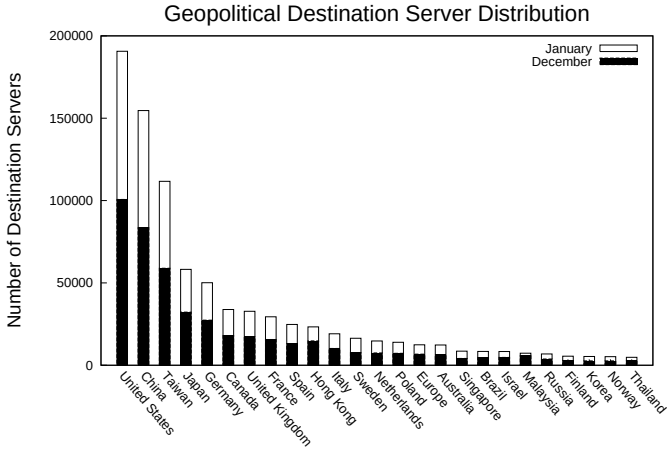


Figure 4: Geopolitical destination server distribution during the December and January data collection periods. IP addresses assigned to the European Union are denoted as *Europe*.

4.3 Geopolitical Destination Server Distribution

In order to obtain insight into *where* Tor users' traffic is destined, we provide the geopolitical distribution of the exit traffic from our router. During both data collection periods, the United States, China and Taiwan were the most popular destination server locations. During the two data collection periods, there were 190,645 and 100,663 connections exiting to the United States, 154,647 and 83,571 connections to China, and 111,668 and 58,873 connections to Taiwan. The destination server distribution is given in Figure 4.

4.4 Tor Used to Fight Censorship

In order to provide evidence that Tor is used to fight censorship, it is necessary to show that Tor use is more likely to occur than real Internet use in a particular country. The difference in the distributions of Tor clients over both data collection periods to the distribution of Internet users where users are more likely to use Tor than the real Internet is shown in Table 1. Data for additional countries is provided in Appendix A. We obtained data on the distribution of Internet users by country from Internet World Stats [17], which is a compliation of current data from a number of trustworthy sources including Nielsen and the International Telecommunication Union.

Country	% Internet	% Tor	%Tor / % Internet
United Arab Emirates	0.12	1.24	10.33
Saudi Arabia	0.23	2.08	9.04
Iran	0.37	1.98	5.35
Germany	4.54	22.12	4.87
Sweden	0.61	1.56	2.56
Switzerland	0.46	0.90	1.96
China	11.85	21.58	1.82
Finland	0.30	0.53	1.77
Poland	1.02	1.77	1.74
Taiwan	1.24	1.30	1.05

Table 1: Percentage of Tor users observed in a country divided by the Percentage of Internet users in that country out of the total number of Internet users in the world.

According to Reporters Without Borders [6], a free press and free Internet advocacy organization, there are thirteen countries, labeled "Internet enemies," where cyber-dissidents are routinely imprisoned or country-wide Internet access is monitored or censored. These countries include Belarus, Burma, China, Cuba, Egypt, Iran, North Korea, Saudi Arabia, Syria, Tunisia, Turkmenistan, Uzbekistan, and Vietnam [5]. In our December data gathering period, clients originating from one of these countries used our Tor router as an entrance node on a circuit with the following frequencies: Belarus - 4, China - 2597, Iran - 194, Saudi Arabia - 172, and Vietnam - 3. During our January data gathering period, clients originated from the following "Internet enemy" countries: Belarus - 1, China - 1686, Iran - 198, Saudi Arabia - 240, Syria - 1, and Vietnam - 3. The geopolitical distribution of Tor clients provides evidence that privacy enhancing systems such as Tor are tools used to combat government sanctioned Internet censorship. It is not possible to directly correlate the entrance and exit traffic; however, there is a large volume of client traffic originating in China and a large amount of router traffic exitting to Taiwan, which suggests that Tor may be used to fight China's policy of censorship.

5. CIRCUIT MEASUREMENTS

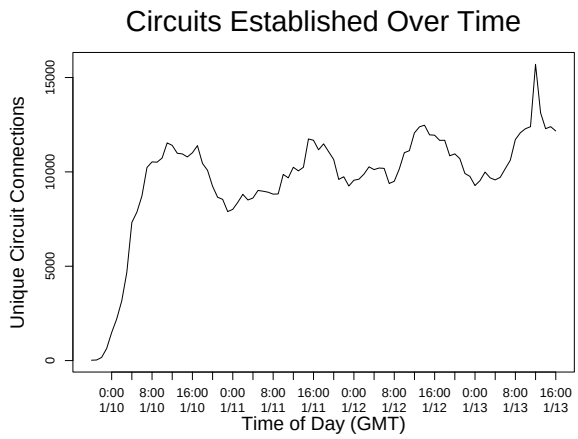
One of Tor's most important design goals is to provide a low latency, high throughput transport service that is suitable for supporting interactive applications. However, a common reason given why most people do not use Tor is that it is "slow." Tor incurs greater latency when compared to direct connections, since Tor routes all packets through a circuit of three hops (by default). Some routers are also highly congested since their limited bandwidth must be shared among several circuits simultaneously. As part of our circuit-level measurements, we examine how Tor use varies with the time of day and we measure the latency across a large number of circuits. Additionally, we study how Tor circuits effect system throughput. We finally look at circuit duration and the amount of data transported by circuits routed through our Tor router.

5.1 Circuit Rhythms

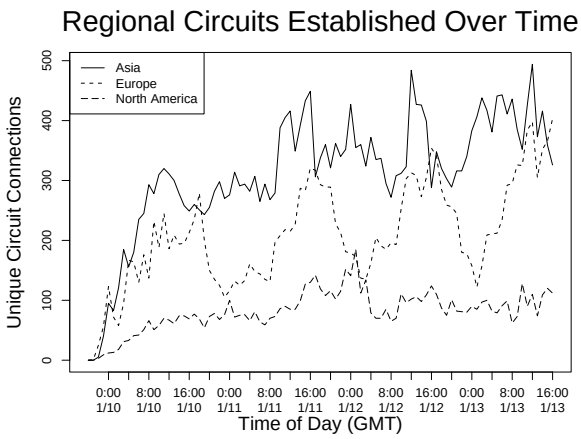
In order to understand Tor use as a function of time of day, we examine the number of circuits observed through our router over the course of the data collection periods. We plot the number of unique connections observed versus time of day in Figure 5(a).

When our Tor router first joined the network, it took several hours to integrate into the network, and the number of connections slowly increases over this warm-up time. Once integrated, the graph indicates that Tor use is cyclical, with a period of approximately one complete day. In addition, peak hours of Tor use occur at 14:00-16:00 GMT, when over 12,000 unique circuits per hour were observed. Tor use is lowest at 0:00 (midnight) GMT, when less than 9,000 circuits per hour were observed. The greatest difference between the high and low times was a 37% decrease from the peak. While the network remains used at all times of day, this shows that a significant correlation between Tor use and time of day exists.

This time correlation is consistent with the client distributions observed in Section 4.1. We examine the nature of Tor usage within Asia, Europe, and North America separately in Figure 5(b). To obtain the locations of Tor users over time, it is only possible to ascertain the client's location when our Tor router is used as the entrance router in a circuit. Using

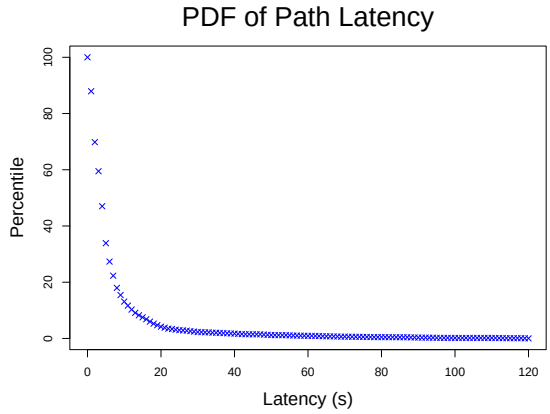


(a) Global circuit connections as a function of the time of day.

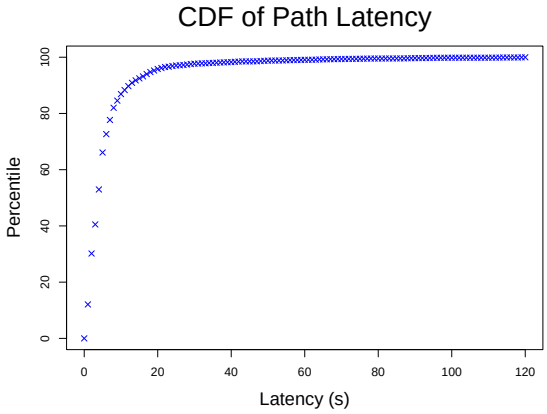


(b) Asian, European, and North American circuit connections as a function of the time of day.

Figure 5: The observed circuit connections are plotted over the course of one data collection period.



(a) PDF of measured circuit latencies.



(b) CDF of measured circuit latencies.

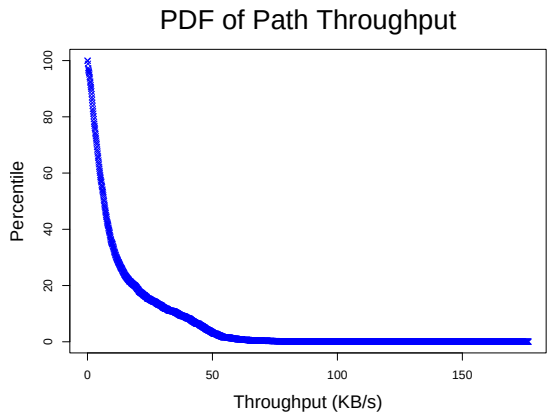
Figure 6: PDF and CDF of measured path latencies through Tor.

this data, users from Asia comprise the most circuit connections, following by Europe and North America. Usage over time in Asia and the United States does not conform to a clear cyclical pattern; however, European users are most frequent during European daytime hours and are least abundant during the night. European users decrease by as much as 62.5% during their off-peak hours. This pattern contributes highly to the global Tor usage pattern shown in Figure 5(a).

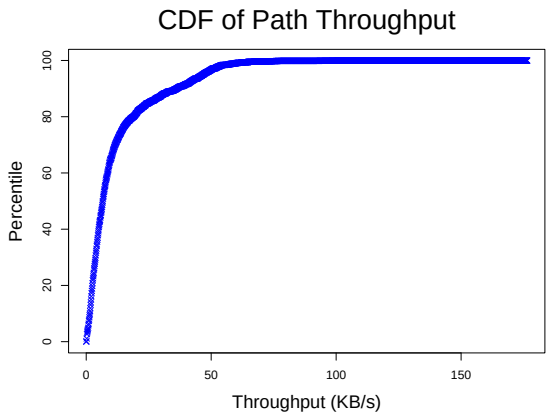
5.2 Circuit Latency

To measure latency of circuits, we used `echoping` [14], a network performance measurement tool, by routing echo requests through the Tor network to an echo server with a high bandwidth link. The server's echo response is also routed through Tor back to the client. This procedure gives an accurate round trip time (RTT) measurement. A base-line measurement was taken by directly sending echo requests to the server, which resulted in an average RTT of 0.47 seconds with negligible variance. Figure 6 shows the probability density function (PDF) and cumulative distribution function (CDF) of the latency measurements through Tor.

The figures reveal that the median latency of a circuit is approximately 4 seconds with a high variance and a maximum observed latency of 120 seconds. At the 25th percentile, a circuit experienced under 2 seconds of latency, at the 75th percentile, a circuit incurred about 7 seconds of latency, and at the 90th percentile, 13 seconds of latency was observed. The mean was 6.09 seconds with a standard deviation of 10.14 seconds. In the case of higher latency circuits (at the 90th percentile and above), it is probable that the cir-

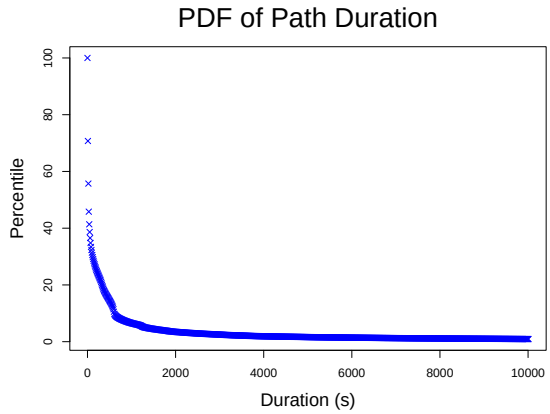


(a) PDF of measured circuit throughput

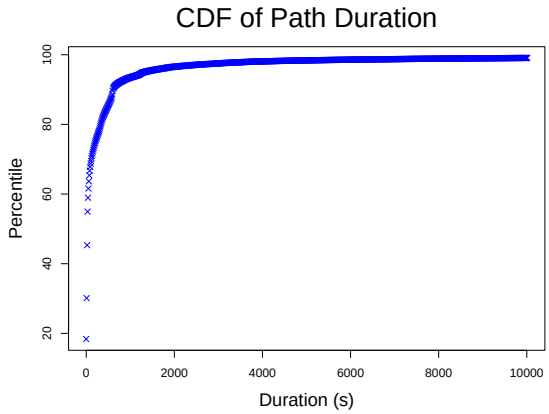


(b) CDF of measured circuit throughput

Figure 7: PDF and CDF of measured throughput through Tor.



(a) PDF of circuit lifespans.



(b) CDF of circuit lifespans.

Figure 8: PDF and CDF of circuit lifespans through Tor.

cuit failed and had to be reconstructed. This shows the cost of privacy in the form of increased latency.

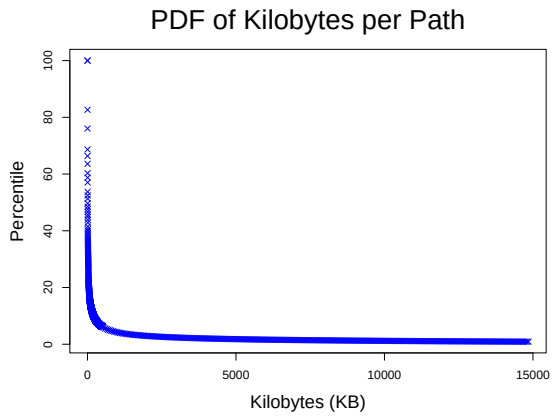
5.3 Circuit Throughput

To collect data about throughput over Tor circuits, we transfered a 128 KB file from a web server through the Tor network and measured the download time. A base-line was produced by measuring the throughput while downloading the file directly. The base-line throughput was 270 KB/s with marginal variance. Figure 7 shows the PDF and CDF of the throughput measurements through Tor. The median throughput was 6.8 KB/s and the mean was 12.6 KB/s with a standard deviation of 15.2 KB/s. At the 25th percentile, circuit throughput was 3.2 KB/s, at the 75th percentile, a circuit provided 14.6 KB/s, and at the 90th percentile, 35.8 KB/s throughput was maintained. The maximum observed throughput during the observation was 180.8 KB/s. This demonstrates that most Tor circuits provide relatively low throughput transport.

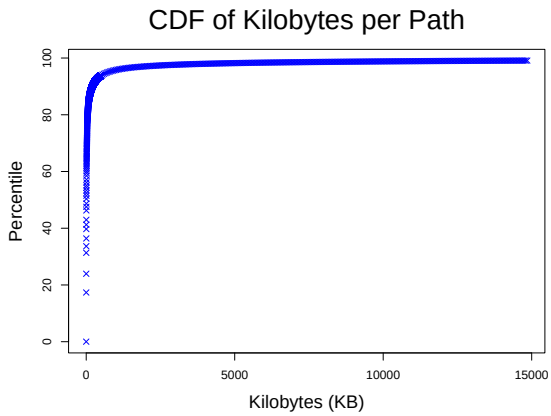
5.4 Circuit Duration

Circuit duration was measured during our data collection while operating a Tor router. As depicted in Figure 8, the median circuit duration was 20-30 seconds. At the 25th percentile, a circuit lasted for under 10 seconds, at the 75th percentile, a circuit is used for 210 seconds, and at the 90th percentile, a circuit is used for 610 seconds. The mean circuit duration is 814 seconds with a standard deviation of 7900.5 seconds. The longest living circuit was observed for 242380 seconds, or 67.3 hours. This shows that the vast majority of circuits are short-lived. The median circuit duration is sufficient for transferring small amounts of data over HTTP, for example.

To demonstrate that the two data collection periods are

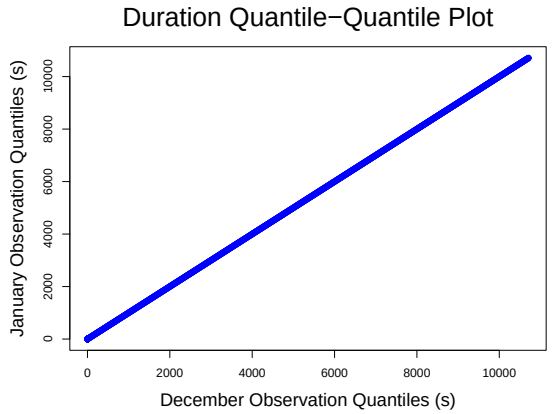


(a) PDF of kilobytes transfered over circuits.

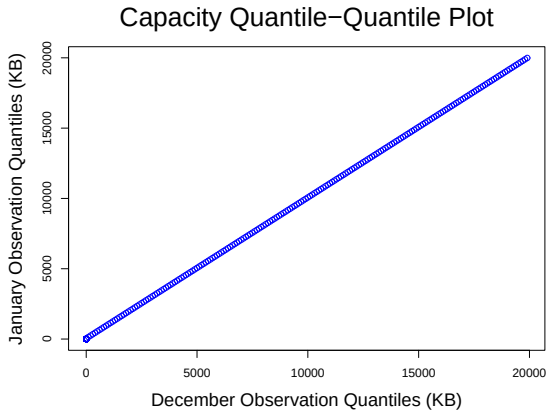


(b) CDF of kilobytes transfered over circuits.

Figure 9: PDF and CDF of kilobytes transfered over Tor circuits.



(a) Quantile-Quantile plot of the duration observations from the January and December data collection periods.



(b) Quantile-Quantile plot of the kilobytes transfered from the Janurary and December data collection periods.

Figure 10: These plots show that the observations from both data collection periods are from the same distribution. This verifies that the data collection was consistent between observation periods.

consistent, we show a quantile-quantile (QQ) plot in Figure 10(a). A linear relationship between the observed quantiles from the two data collection periods indicates that the two data sets are taken from the same distribution [18]. This demonstrates that the data collection was consistent between observations.

5.5 Circuit Capacity

In order to measure capacity of Tor circuits, we observed how many bytes tranversed circuits during our participation in the Tor network. The PDF and CDF of the bytes transfered per circuit are given in Figure 9. At the median, 6.1 KB traversed a circuit. At the 25th percentile, about 1.0 KB flowed through a circuit, at the 75th percentile, 31.2 KB was sent, and at the 90th percentile, 201.7 KB was transferred. However, the mean circuit transported 730.8 KB with a standard deviation of 10312.6 KB.

The maximum amount of data transferred over a circuit was 1.5 gigabytes (GB). This demonstrates that while most circuits transport very little real data, there exist outliers that are able to sustain the circuit for a sufficient amount of time to transfer several orders of magnitude more data - although this is quite rare. The median circuit capacity would be sufficient to transfer a relatively small web page over HTTP. In fact, these circuit-level measurements are highly consistent with the observed protocol distribution.

To demonstrate that the two data collection periods are consistent, we provide a QQ plot in Figure 10(b).

Protocol	December		January		Total	
	Percent	Raw	Percent	Raw	Percent	Raw
Web (HTTP, HTTPS)	89.77	4,678,423	90.74	5,826,413	90.31	10,504,836
Peer-to-peer (BitTorrent)	9.80	510,530	8.78	564,023	9.24	1,074,553
Instant Messaging (AIM,IRC,Jabber,MSNMS,YMSG)	0.26	13,038	0.26	14,696	0.25	27,734
E-Mail (POP,IMAP)	0.14	7,202	0.05	3,127	0.09	10,326
Telnet	0.02	603	0.12	7,798	0.07	8,401
FTP	0.03	1,495	0.04	2,293	0.03	3,788
Total	100	5,211,291	100	6,418,350	100	11,629,641

Table 2: Protocol breakdown of identifiable exit traffic during the December and January data collection periods by number of TCP connections.

Protocol	December		January		Total	
	Percent	Raw	Percent	Raw	Percent	Raw
Peer-to-peer (BitTorrent)	67.58	309GB	64.37	346GB	65.83	655GB
Web (HTTP, HTTPS)	32.00	146GB	35.24	189GB	33.7	336GB
FTP	0.19	882MB	0.19	1GB	0.19	2GB
Instant Messaging (AIM,IRC,Jabber,MSNMS,YMSG)	0.14	654MB	0.14	718MB	0.14	1,372MB
E-Mail (POP,IMAP)	0.08	382MB	0.02	131MB	0.05	513MB
Telnet	0.02	80MB	0.03	156MB	0.02	236MB
Total	100	457GB	100	538GB	100	995GB

Table 3: Protocol breakdown of identifiable exit traffic during the December and January data collection periods by amount of bytes.

5.6 Discussion

In our analysis of Tor performance at the circuit level, we have shown that the volume of Tor traffic is correlated with the time of day. We have also confirmed that Tor has met its goal of providing a low-latency transport service; however, it cannot sustain a high level of throughput. Finally, our data suggests that Tor circuits are short-lived and individually transport a small amount of data. We restrict this performance study to the circuit-level, due to Tor's multi-hop architecture. Given the limited perspective obtained by running a single Tor router, it is only feasible to report statistics at the circuit-level. The only way to study individual Tor routers is to control a significant portion of the network, which is impractical due to the resource requirements and not recommended for security and privacy reasons.

6. PROTOCOL DISTRIBUTION

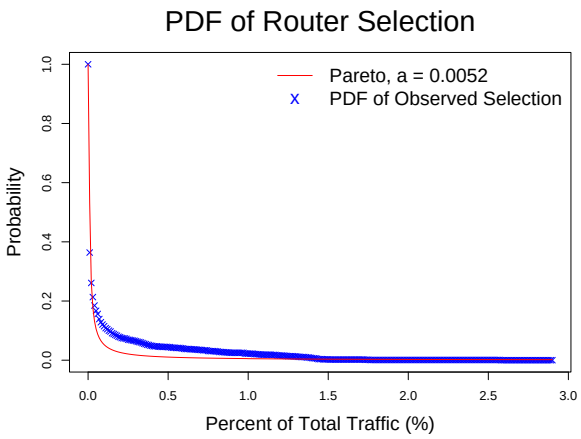
The designers of the Tor network have placed a great deal of emphasis on achieving low latency and reasonable throughput in order to allow interactive applications, such as web browsing, to take place within the network. As part of this study, we were interested in observing which application-level protocols were seen exiting our Tor node. Recall that we used `ethereal`, which can accurately identify most commonly used application-level protocols. For the December data collection period, Ethereal could identify 87% (5,211,575) of the established TCP connections and 83% (457GB) of the total exit traffic. For the Janurary data collection period, Ethereal could identify 85% (6,420,462) of the established TCP connections and 84% (538GB) of the total exit traffic.

The most common protocol, measured by the number of established TCP connections, during both collection periods was HTTP, which is not surprising since there are browser plug-ins [9] that help to tunnel their web traffic through a Tor proxy. This is consistent with the circuit-level observations that the majority of the circuits transported a small amount of data. The next most commonly observed protocol was Bittorrent, a popular peer-to-peer protocol for file sharing. Finally, Secure Socket Layer (SSL) traffic was observed frequently during both collection periods, which could be encrypted HTTP (HTTPS) traffic or other protocols that use SSL for confidentiality. E-mail (POP and IMAP), chat/instant messaging (AOL Instant Messenger (AIM), Internet Relay Chat (IRC), Jabber, Microsoft Messenger Service (MSNMS), Yahoo Messenger Service (YMSG)), FTP, and telnet sessions occurred more sporadically. The complete protocol breakdown by TCP connections is given in Table 2.

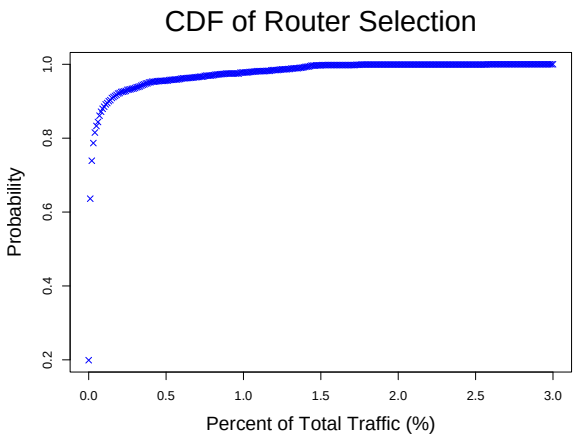
The most glaring difference between viewing the protocol breakdown measured by the number of bytes, shown in Table 3, in contrast to the number of TCP connections is that while HTTP accounted for the majority of TCP connections, it is the BitTorrent protocol that uses the majority of the bandwidth within the Tor network. This is not shocking since BitTorrent is a peer-to-peer (P2P) protocol mostly used to download large files.

6.1 Is Peer-to-Peer Traffic Hurting Performance?

Since the TCP connection statistics show the majority of connections are HTTP requests, one might be led to believe that most users are using the network as a HTTP proxy. However, the few people that do use the network for P2P applications such as BitTorrent consume the majority of the bandwidth. The operators of the network consider P2P traffic harmful, not because of ethical or legal reasons, but simply because it makes the network less useful to those for



(a) The PDF conforms roughly to a Pareto distribution with a shape parameter $a = 0.0052$. Our data fit the distribution with a mean squared error of $MSE = 6.1506 \times 10^{-4}$. This shows what percentage of the network transports what portion of the users' traffic.



(b) The CDF shows that the vast majority of Tor routers (90%) transport little to no traffic (0.13% of the total traffic through our router).

Figure 11: PDF and CDF of Tor Router selection.

whom it was designed. In an attempt to prevent the use of P2P programs within the network the default exit policy was changed to block the standard file sharing TCP ports (1214, 4661-4666, 6346-6429, and 6881-6999). The protocol breakdown by bytes shows that blocking these ports is not effective at impeding P2P applications from using the majority of bandwidth. In fact, it is a losing battle to attempt to block peer-to-peer traffic since applications can run on non-standard ports and can be encrypted to hide application headers.

A potentially better solution to the problem of peer-to-peer applications might be to explore alternative network designs that are tailored to these applications. Such designs may in reality provide slightly less, but sufficient anonymity for file traders while offering better performance over that given by the Tor network. If such a network could be created, most file traders would use it over the Tor network because of the improved performance. Future work in this area could additionally relieve the stress placed on the Tor network by file sharing applications.

7. TRAFFIC DISTRIBUTION

To elucidate Tor's router distribution, we use observations regarding which routers are present as previous or next hops from our router to compute the frequency of occurrence for each router in the network. These frequencies provide insight into the probability distribution associated with the router selection process.

7.1 Modeling Router Selection

Understanding the distribution with which different routers are utilized on circuits can provide valuable insights regarding the system's vulnerability to traffic analysis. In addition, a precise probability distribution can be used to build more realistic analytical models and simulations.

Using the observed router frequencies, we construct a probability density function (PDF) to model how much of the Tor network transports what portion of the users' traffic. The PDF, given in Figure 11(a), shows that during the December data collection period, over 30% of the routers each *individually* transported 0.01% of the total network traffic from the perspective of our router. The PDF curve drops sharply; only 2% of the routers individually transported 1.0% of the traffic. The most traffic that any single router transported was 2.56% of the total traffic. This indicates that the vast majority of Tor traffic is handled by a very small set of routers. This fact may account for a large portion of the variability in the circuit-level measurements of latency, throughput, circuit duration, and circuit capacity.

In addition, we provide a cumulative distribution function (CDF) for the December data collection period in Figure 11(b). The CDF shows that 90% of the routers individually transported only 0.13% of the total traffic, and cumulatively transported a mere 13% of the total network traffic. Inversely, 10% of the routers cumulatively handled the remaining 87% of the traffic.

Probability density functions in which a small subset of the distribution accounts for a large amount of the probability mass can be modeled using a *Pareto* distribution [18]. This model is traditionally used to describe the distribution of wealth in a society, where the top 20% of the population accounts for 80% of the wealth (commonly referred to as the "80-20" rule). The PDF for a Pareto distribution is defined as:

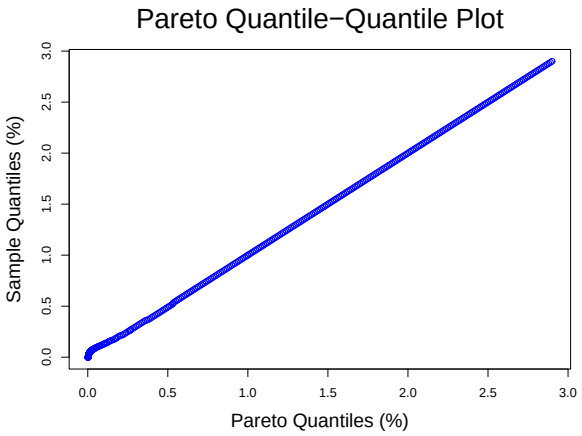


Figure 12: A Pareto quantile-quantile graph is linear. This indicates that the measured traffic distribution fits the theoretical Pareto distribution very closely.

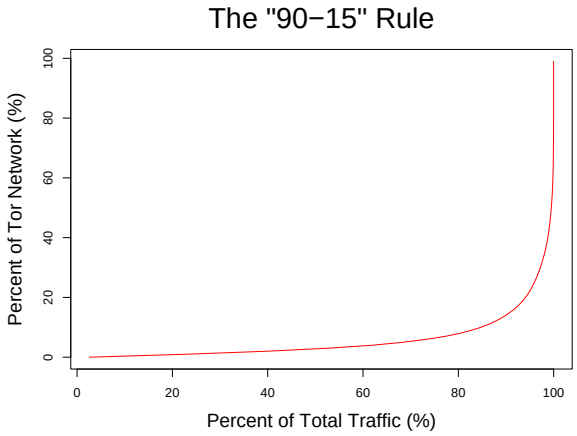


Figure 13: This graph shows that a small portion of the Tor routers- roughly 15% - transport the vast majority of the traffic through the network - almost 90%.

$$p(x) = ax^{-(a+1)},$$

where a is the shape parameter. Using the December data set, a Pareto distribution model is constructed with the shape parameter $a = 0.0052$ that has a mean squared error of $MSE = 6.1506 \times 10^{-4}$. In order to further prove that the measured traffic distribution closely fits a Pareto model, a quantile-quantile plot is generated with the measured traffic distribution and a Pareto distribution. In the graph, the observed quantiles are plotted versus the theoretical quantiles for a Pareto distribution. The resulting curve is linear, which indicates that the observed quantiles are taken from the Pareto distribution, as shown in Figure 12. The Pareto distribution model is plotted with the PDF in Figure 11(a).

7.2 Is the Tor Network as Large as it Seems?

In order to provide a low-latency service, Tor utilizes a *preferential* routing mechanism probabilistically weighted to select high-performing, long-uptime Tor routers more frequently than those that perform poorly and are short-lived. It has been observed that this bias in router selection does, in fact, degrade the anonymity of the system [10].

The total number of routers over the course of each data collection period was approximately 1,500. However, only a few high performing routers are forwarding the vast majority of the traffic. Figure 13 shows the percent of the routers that are involved in forwarding how much of the total traffic through the system. This demonstrates the "90-15" rule, in which 90% of the network traffic is handled by roughly 15% of the network, which are the highest performing routers. Many theoretical analyses of multi-hop privacy enhancing systems make the assumption that all routers in the network participate equally - each node is chosen with probability $\frac{1}{N}$, for a network of N nodes. However, given these observations, such an analysis of Tor is clearly flawed.

Furthermore, since there exists a high probability (approximately 0.9) that only 15% of the network handles a user's traffic, then that user's anonymity set has in reality decreased from N to $0.15N$ with a high probability. This observation was originally theorized in Shmatikov and Wang [22]. If an adversary is able to enter the subset of $0.15N$ nodes that are most likely to forward the most traffic, then their ability to perform attacks such as passive traffic analysis is greatly increased.

8. LEGAL AND ETHICAL IMPLICATIONS

In this section, we will discuss some of our subjective experiences as a Tor operator, including some of the difficulties we encountered while operating as an exit node. Tor provides a flexible set of options to configure a node to filter all or specific exit traffic. For our exit node, we chose to use the default exit policy, which allows most traffic to exit the network. This was done in order to discover which protocols and applications are most commonly used within the network. If we had, for instance, blocked all port 80 traffic, our protocol data would have included much less HTTP traffic.

Unfortunately, since we are forwarding traffic on behalf of Tor users, our Tor node's IP address appears to be the source of sometimes malicious traffic. Our node's liberal exit policy and the large amount of bandwidth that it provided caused us to receive a large number of complaints ranging from DMCA §512 (take-down) notices, reported hacking attempts, IRC bot network controls, and posting of inappropriate and offensive images to message boards. Due to the volume and nature of these complaints, our institution's administration requested that we stop running our node shortly after the data

for this paper was collected. Similar accounts of administrative and law enforcement attempts to prevent Tor use are becoming more common as Tor becomes more popular to the masses [11]. The Electronic Frontier Foundation (EFF), a group that works to protect digital rights, has provided template letters [8] and offered to provide assistance [12] to Tor node operators that have received DMCA take-down notices. The legal questions regarding who is responsible for traffic exiting a Tor node is an ill-defined issue. However, in our case, it was not a matter of legality as much as bad press caused by one especially unfortunate incident involving a child abuse watch dog organization publicly blaming our institution for a message group posting.

A solution to our problems could have been to change our exit policy to reject all exit traffic. A large number of nodes on the currently deployed Tor network have an exit policy that blocks all traffic making the node what is called a "mix-only" node. The majority of circuits built though our node used our router as the exit node. While mix-only nodes do help the Tor network, there must be a subset of at least one-third of the network's bandwidth capacity that allows exit traffic, otherwise the network would have insufficient exit routers to build complete circuits and provide service.

From our experience, the negative activities receive the most attention from legal and administrative authorities. However, the cause of Internet privacy is a noble one since Tor is a successful tool for combating Internet censorship worldwide. We are still in search of a way to run our high-performance Tor router as an exit node on a permanent basis.

9. CONCLUSION

This study is aimed at understanding the usage of Tor. In particular, we provide insight into where, geo-politically, Tor clients, routers, and destination servers are located. In doing so, we uncovered evidence that suggests that Tor is used to fight Internet censorship.

We study the system's performance at the circuit-level by measuring latency, throughput, circuit lifespan, and the amount of data transferred through a circuit. We also examine how Tor use varies with the time of day. We show that the system achieves its goal of providing a low latency service; however, the expected throughput is also quite low. Most of the circuits transported little traffic and were short-lived. We also observe that the circuit-level measurements had a high variability, which is consistent with a highly variable level of router quality in the network.

We also characterize the application-level protocol distribution in an effort to understand the nature of anonymized traffic. The most prevalent protocol observed was HTTP, while the majority of the bandwidth in the system is consumed by the Bittorrent protocol. These protocol observations are consistent with our circuit-level measurements, indicating that the majority of the circuits observed were used to transport HTTP traffic.

Finally, we describe the network traffic distribution in terms of the Pareto distribution. Due to Tor's use of preferential routing to optimize performance, a "90-15" rule emerges, where 90% of the traffic is transported by only 15% of the Tor routers.

It is our hope that this characterization study will reinforce the importance of privacy enhancing systems such as Tor, since we uncovered evidence to suggest that it may be used to fight censorship. In addition, by understanding the application-level protocol distribution and the distribution of traffic throughout the network, more accurate analytical models can be constructed.

10. REFERENCES

[1] American Registry for Internet Numbers. http://www.arin.net/index.shtml.

[2] Asia Pacific Network Information Centre. http://www.apnic.net.

[3] Ethereal: A Network Protocol Analyzer. http://www.ethereal.com/.

[4] Latin American and Caribbean Internet Addresses Registry. http://lacnic.net/en.

[5] List of the 13 Internet enemies in 2006 published.

[6] Reporters Without Borders. http://www.rsf.org.

[7] RIPE Network Coordination Centre. http://www.ripe.net/.

[8] Tor: Response template for Tor node maintainer to ISP. http://tor.eff.org/eff/tor-dmca-response.html.

[9] Torbutton. https://addons.mozilla.org/firefox/2275/.

[10] Bauer, K., McCoy, D., Grunwald, D., Kohno, T., and Sicker, D. Low-Resource Routing Attacks Against Anonymous Systems. *University of Colorado Technical Report CU-CS-1025-07* (2007).

[11] Cesarini, P. Caught in the Network. In *The Chronicle of Higher Education*, vol. 53. Washington, D.C., February 2007, p. B5.

[12] Dingledine, R. EFF is looking for Tor DMCA test case volunteers. http://archives.seul.org/or/talk/Oct-2005/msg00208.html.

[13] Dingledine, R., Mathewson, N., and Syverson, P. Tor: The second-generation onion router. In *Proceedings of the 13th USENIX Security Symposium* (August 2004).

[14] Echoping Performance Measurement. http://echoping.sourceforge.net.

[15] Goldberg, I. On the security of the tor authentication protocol. In *Proceedings of the Sixth Workshop on Privacy Enhancing Technologies (PET 2006)* (Cambridge, UK, June 2006), Springer.

[16] Goldschlag, D. M., Reed, M. G., and Syverson, P. F. Hiding Routing Information. In *Proceedings of Information Hiding: First International Workshop* (May 1996), Springer-Verlag, LNCS 1174, pp. 137–150.

[17] Inernet World Stats. http://www.internetworldstats.com/.

[18] Jain, R. *The Art of Computer Systems Performance Analysis: Techniques for Experimental Design, Measurement, Simulation, and Modeling.* John Wiley & Sons, 1991.

[19] Murdoch, S. J. Hot or not: Revealing hidden services by their clock skew. In *13th ACM Conference on Computer and Communications Security (CCS 2006)* (Alexandria, VA, November 2006).

[20] Murdoch, S. J., and Danezis, G. Low-cost traffic

analysis of Tor. In *Proceedings of the 2005 IEEE Symposium on Security and Privacy* (May 2005), IEEE CS.

[21] Øverlier, L., and Syverson, P. Locating hidden servers. In *Proceedings of the 2006 IEEE Symposium on Security and Privacy* (May 2006), IEEE CS.

[22] Shmatikov, V., and Wang, M.-H. Measuring relationship anonymity in mix networks. In *WPES '06: Proceedings of the 5th ACM workshop on Privacy in electronic society* (New York, NY, USA, 2006), ACM Press, pp. 59–62.

[23] tcpdump. http://www.tcpdump.org/.

APPENDIX

A. GEOPOLITICAL DISTRIBUTIONS

Country	% Internet	% Tor	%Tor / % Internet
United Arab Emirates	0.12	1.24	10.33
Saudi Arabia	0.23	2.08	9.04
Iran	0.37	1.98	5.35
Germany	4.54	22.12	4.87
Sweden	0.61	1.56	2.56
Switzerland	0.46	0.90	1.96
China	11.85	21.58	1.82
Finland	0.30	0.53	1.77
Poland	1.02	1.77	1.74
Taiwan	1.24	1.30	1.05
Netherlands	0.97	0.96	0.99
France	2.77	2.73	0.98
USA	18.95	17.48	0.92
Italy	2.76	2.40	0.87
Japan	7.74	6.10	0.79
Canada	1.97	1.46	0.74
Austria	1.32	0.91	0.69
Spain	1.77	1.12	0.63
UK	3.37	2.04	0.61
Thailand	0.76	0.40	0.53
Russia	2.13	0.72	0.34
Turkey	1.44	0.43	0.30
Brazil	2.88	0.50	0.17

Table 4: Percentage of Tor users observed in a country divided by the Percentage of Internet users in that country out of the total number of Internet users in the world.