Traffic Statistics of a High-Bandwidth Tor Exit Node

Michael Sonntag and René Mayrhofer
Insitute of Networks and Security, Johannes Kepler University Linz, Altenbergerstr. 69, 4040 Linz, Austria
{michael.sonntag, rene.mayrhofer}@ins.jku.at

Keywords: Tor, Anonymization, Traffic Analysis.

Abstract: The Tor anonymization network supports (and is widely used for) circumventing censorship, evading intrusive mass-surveillance, and generally protecting privacy of Internet users. However, it also carries traffic that is illegal in various jurisdictions. It is still an open question how to deal with such illegal traffic in the Tor network, balancing the fundamental human right for privacy with the need for assisting executive forces. By operating and monitoring a high-bandwidth Tor exit node as both a technical and legal experiment, we statistically analyse where popular servers are located and how they are used based on connection metadata of actual exit node usage. Through this we identify inter alia that cooperation only in comparatively few countries would be needed – or any illegal use would be very small. In this paper, we provide more in-depth statistical insight into Tor exit node traffic than previously publicly available.

1 INTRODUCTION

The Tor network is a widely used system to anonymize source IP addresses of Internet users, especially regarding the server they contact, as well as observers on the network in-between, including the user's Internet Service Provider (ISP). Normally when contacting a server, it receives the client's IP address, which (with added information from e.g. the user's ISP) allows tracing the request to a specific computer, and often an individual. Observing the traffic on the routing path discloses this information too. Through inserting intermediate stations (and encrypting transfer, randomly selecting/changing intermediaries, etc), the Tor network replaces and so hides this source IP address, while still allowing arbitrary TCP connections.

While this is obviously beneficial for many use cases, e.g. private communication in the age of digital mass surveillance, anonymously looking up information that may be legal but somehow embarrassing, or circumventing censorship, it can also be used for nefarious purposes: attacking other IT systems or accessing/downloading illegal content. Especially such undesirable use is one reason why the Tor network is often seen with suspicion and presented by some as a tool that is used mostly (or even "solely") for illegal use. However, actual and statistically relevant data on how people use the Tor network is curiously lacking.

The institute of Networks and Security operates a Tor exit node with a bandwidth of approximately 200 Mbit/s, making it the fastest exit node in Austria at the time of this writing. On average a daily exit traffic of 1.2 TB is transmitted to/from the Internet. Passively monitoring this node, we gather usage data on actual Tor network use. As we need to guarantee continued anonymity, several precautions are taken, both when gathering the data and during analysis.

As in Austria a court sentenced the operator of a Tor node to jail on probation (LG Strafsachen Graz 30.6.2014, 7 Hv 39/14p), significant efforts were taken to ensure legal operation. A letter was submitted to the Austrian telecommunication authority (Rundfunk und Telekom Regulierungs-GmbH), registering the exit node as a telecommunication network or service – however with the intent of this application getting rejected, as being such an operator would entail additional work and responsibilities (plus potentially fees). In the application we insisted on a binding answer, which we finally received (file number RSON 64/2015-2). Through this it is now "official" that running a Tor exit or relay node does not need official permission in Austria (Sonntag, 2015). As the applicable laws are based on EU law, the result likely applies to other member states too.

Note that while Tor also supports anonymizing (hiding) IP addresses of servers ("hidden services"), this feature is out of scope of this paper. We focus

Sonntag, M. and Mayrhofer, R.
Traffic Statistics of a High-Bandwidth Tor Exit Node.
DOI: 10.5220/0006124202700277
In *Proceedings of the 3rd International Conference on Information Systems Security and Privacy (ICISSP 2017)*, pages 270-277
ISBN: 978-989-758-209-7

on anonymized use of standard Internet servers via IPv4 (IPv6 traffic is subject to future analysis and currently not supported by our exit node).

Specifically we investigate these questions:

- How many countries must be "handled" to cover most data? I.e., should illegal traffic be discovered, in how many and which countries would police cooperation be necessary? As an exit node we can only provide data on attack targets (or servers with illegal content), but not their sources (or users of illegal content).
- How "concentrated" is Tor traffic to (perhaps only a few) enterprises (Autonomous Systems, AS)? I.e., would we need only a few contacts to notify about attacks, or too many to realistically handle? Note that large providers typically operate data centres in several countries.
- Would in-depth investigation of the traffic content be useful at all? If traffic is encrypted, then e.g. intrusion detection or prevention systems (IDS/IPS) for automatically handling "bad" (based on various definitions of good vs. bad) traffic tend to be useless.
- What kinds of services (based on port numbers) are people using? Are any of these known as being used by malware or are there any other hints towards illegal use?

We contribute a statistical analysis of network traffic data from our high-bandwidth Tor exit node and draw conclusions towards options for handling illegal traffic. However, this is a snapshot of one month of a single exit node, and the statistical distribution of future traffic may change at any time, potentially rendering some of this analysis obsolete. More and current data is available online (INS).

2 RELATED WORK

A lot of research focuses hidden services, e.g. (asn), (Biryukov et al., 2014) or (Loesing et al., 2008). As we are only investigating exit traffic, i.e. communications to the public Internet, these are of less relevance here.

Akamai investigated HTTP (not HTTPS) attacks based on whether they were originating from known Tor exit nodes or other computers, finding that 1 in 380 Tor requests were malicious, while requests from other sources only had a probability of 1 in 11.500 (Akamai, 2015). However, DDoS attacks, consisting by definition of many connections, were excluded (limited Tor network bandwidth).

(Chaabane et al., 2010) focus more on the use of the BitTorrent protocol via Tor and employ deep packet inspection to identify specific kinds of traffic. As an additional limitation, for that study six exit nodes were used with only 100 Kbit/s bandwidth. In comparison we investigate a significantly faster node, obtaining a larger and more representative share of users (our exit node is located solely in one country, as in that study too).

And while (Loesing et al., 2010) investigated both incoming and outgoing traffic, they only provide statistics on outgoing traffic per port, ignoring the final destination, i.e. the target IP address. This makes perfect sense as they also investigated incoming traffic – gathering both together is a significant risk identified by them. We safeguard privacy by only analysing outgoing traffic, but include the target address.

(McCoy et al., 2008) also investigated both ends of Tor connections. Additionally they also used deep packet inspection on the first 20 bytes of content data. Regarding geographical locations, only those of clients and entry nodes were investigated.

(Ling et al., 2015) explored how malicious traffic can be detected through adding an IDS after a Tor exit node. They also identified that at least some illegal (or at least undesirable) traffic takes place. E.g. with 86 million flows they received 3.6 million alerts, however most of them seemed to concern P2P traffic (which is not necessarily illegal, and if it is, then mostly on the comparatively low level of copyright violations), namely 77%, leaving 820,000 alerts (a bit less than 1% of all flows, assuming each flow produces at most one alert).

(Jansen and Johnson, 2016) describe a distributed system of several nodes for collecting, aggregating, and tallying statistics from multiple Tor (both entry&exit) nodes simultaneously. They add "noise" to the collection, which is later removed in tallying to blind each node's contribution against potentially malicious other nodes. While useful for a large and open collection system (=malicious nodes must be expected), for a small/closed group it is less suitable.

In contrast, the present study assumes a generally benevolent operator of the Tor node(s) and aims to identify how illegal content transmitted via Tor could be combated: Which/how many countries or providers are involved? Would it be impossible to detect illegal traffic because of encryption? On which protocols would detection have to focus? These aspects seem not to have been investigated up to now, but are highly relevant in practice for legal operation of Tor exit nodes in European

jursidictions. Additionally, in most of the previous studies it remains unclear how relay traffic has been excluded – Tor exit nodes function simultaneously as relays (=middle nodes) as well. This traffic must be excluded in the investigation to avoid systematic bias. As a secondary contribution, we propose a method to do this without unduly interfering with the Tor protocol, significantly modifying the software, or e.g. relying on lists of Tor nodes.

3 DATA COLLECTION

Data collection when investigating Tor nodes is problematic from several points of view: legal, ethical (Ailanthus, 2015), and technical (Soghoian, 2011). We decided to currently not analyse any kind of communication content, not even if extracted, classified, and anonymized immediately and automatically, but solely (header) metadata.

To prevent aiding deanonymization, only exit traffic is monitored. This was implemented by the Tor node relaying between two different interfaces and monitoring only the "outside" via duplicating this traffic on the switch. On a dedicated monitoring server flow data for this traffic is collected via (Pmacct project). As this server cannot ever see any input traffic (i.e. traffic from any middle to our exit node), strict in- and outside separation is enforced. So even should this system be breached, correlating in- to output remains impossible. Similarly, no content data is stored at all even briefly. As a design principle enforced by the implemented passive monitoring network architecture, no traffic (neither data nor metadata) is modified or manipulated in any way through the monitoring itself.

We collect only the following data: source IP and port, destination IP and port, TOS (see below), protocol, and date/time. These are aggregated to count number of flows, packets, and bytes for these tuples. One of the IP addresses is always the exit IP address of our Tor node, depending on whether it is an outgoing or incoming flow. As the other IP address might still be sensitive data, it is further anonymized by identifying the Autonomous System (AS) it belongs to as well as the country. To avoid passing any information to third parties, these lookups are performed locally through the free country and ASN databases by (MaxMind).

As our exit node provides high bandwidth (200 Mbit/s maximum, typ. 15 MByte/s average), the data collected is significant, even though only metadata. To ease handling, it is collected and stored in one-hour chunks. Compressed this produces on average 45 MB per hour (uncompressed: 380 MB). Because of this approach we may lose some information on connections spanning exactly the brief export period, but we argue that this does not signif. change statistical results. Recording data in brief chunks aids privacy too, as after calculating the statistics the raw data can be deleted immediately.

While all such (anonymous) data is stored, we further restrict analysis during evaluation: Only the top 50 entries are considered (countries and AS) with the further restriction of a lower limit of 10 flows per entry, applying to each one-hour chunk. While this removes some details, we can still estimate the error introduced by comparing the sum of all traffic matching these criteria (e.g. the sum of all traffic bytes to/from the top 50 countries) with the total aggregated traffic (e.g. all bytes sent/ retrieved). Apart from the AS statistics (see below) this is a negligible quantity for our investigations. Therefore, e.g. not all countries appear in every one hour period (too little traffic and therefore not in the top 50/below 10 flows). This results in a slight underestimation of traffic per AS/country and an overestimate of the "other" category.

The data gathering process works as follows:

1. Collect data for one hour in a database
2. Dump the database to disk, then purge it
3. Aggregate data: Per country, per AS, per port, total sum of all traffic
4. Remove all entries from countries and AS with less than 10 flows
5. Sort countries/AS according to total bytes descending; remove all entries after place 50
6. Output list of top countries and AS, ports, and total values

Data collection is limited by our exit policy, as traffic that is not allowed cannot be encountered. E.g. like many exit nodes we prohibit connections to port 25 (SMTP) to discourage sending Spam, but we do allow a wide range of 77 ports.

Data was collected for this investigation during the whole April 2016. This took place on a separate IP range (to avoid repercussions on the "official" university IP range, but still using ACOnet, the Austrian university network, as upstream provider.

As we see all outgoing traffic, we still need to differentiate between actual exit and relay traffic. As relay traffic is always encrypted and directed to another Tor server, it must be excluded from our investigation. For this we modified the Tor source code to mark all exit (but not relay) packets by setting the TOS field to 0x28. In this way we can filter outgoing traffic to remove everything unmarked. But replies always arrive unmarked, so

incoming relay traffic would be included. Consequently, all traffic is stored and during evaluation unmarked traffic, as well as all its mirror traffic (source and destination reversed) is deleted. This process is the reason for recording the TOS field. Combined with the hourly recording period this produces a few artefacts for long connections starting in the previous hour (=flow with lacking TOS mark), but ending in the next (incoming reply flows no longer have a matching outgoing flow).

4 DATA ANALYSIS

The first questions regarding traffic destination countries can be answered easily: almost all traffic is directed towards few countries. We further refined the results to group all EU member states, as e.g. mutual enforcement of judgements is typically simple and many legal rules are unified in the EU. The result is that 99% of all traffic (regarding bytes transferred) is directed towards 14 countries (or 28 EU plus 13 others, i.e. 41 countries); 99.74% are covered by 60 countries. The next (61st) country then amounts to a traffic volume of 123 MB per day only, which is very little on both an absolute and relative scale. The traffic distribution is shown in Table 1.

Mutual legal assistance exists with various non-EU countries as well, based on bi- or multilateral treaties. Whether these apply also to IT information and in what form data is disclosed upon request must be determined individually. The specific problem of the so-called international silver-platter doctrine (not exactly, but an analogon: voluntarily sending data to a country which would not be entitled to request it; used e.g. by intelligence agencies) should be considered too: if data is gathered in one country, it might be disclosed to other countries, depending on national (esp. privacy), laws. Note that at the time of collection data is "anonymous" and only through combination with other data (e.g. collected independently by the recipient) for example traffic correlation might enable identification.

Therefore any illegal activities are either using only very little traffic (could still be many connections), or could be prosecuted in just a few countries. Note however, that these countries are not necessarily the countries of the illegal activity. If this is about hosting criminal data like child pornography, it matches exactly. But attackers attempting to break into computers would be on the other side of the Tor network.

This approach has one notable shortcoming: On rank 8 (0.31% traffic share) lies country "A1", the special code of the GeoIP database for "Anonymous Proxy". So this is not really a single country but could be many countries lumped together.

As another result, the rest of the traffic is extremely widely distributed: the top 70 countries together achieve a traffic share of only 99.81%. So the next 39 countries after the table above account for 0.79%, with 0.19% still missing. These are countries that never made it into the top 50 list, as well as traffic towards countries that made it onto the list only occasionally (numbers on the lower end of the list are underestimated by our approach).

Table 1: Traffic per destination country.

Country	*Traffic per day [GB]*	*Share of traffic*	*Flows per day*	*Share of flows*	*Average traffic per flow [KB]*
EU (European Union)	568.82	48.86%	8,437,102	39.80%	69.0
US (USA)	449.56	38.62%	8,385,531	39.56%	54.9
KR (Korea)	50.56	4.34%	1,776,953	8.38%	29.1
RU (Russia)	46.77	4.02%	1,141,645	5.39%	41.9
JP (Japan)	9.40	0.81%	210,553	0.99%	45.7
CA (Canada)	7.08	0.61%	234,800	1.11%	30.8
CN (China)	4.27	0.37%	163,820	0.77%	26.7
A1	3.59	0.31%	1,087	0.01%	3,380.0
VN (Vietnam)	3.17	0.27%	54,459	0.26%	59.6
UA (Ukraine)	3.02	0.26%	64,791	0.31%	47.7
HK (Hong Kong)	2.10	0.18%	32,935	0.16%	65.3
CH (Switzerland)	1.76	0.15%	34,485	0.16%	52.3
SG (Singapore)	1.35	0.12%	44,457	0.21%	31.0
TW (Taiwan)	1.23	0.11%	23,264	0.11%	54.1
Aggregated rest	11.45	0.98%	590,500	2.79%	19.8

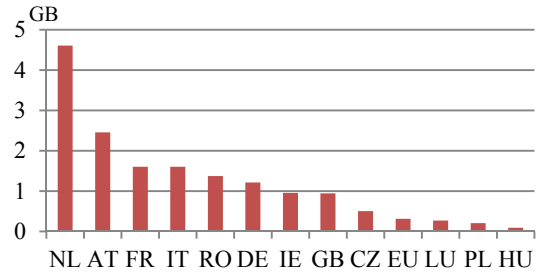

Figure 1: Average daily traffic per EU country in GB.

A similar but less extreme distribution occurs within the EU (see Figure 1): Almost 28% of all traffic is going to Netherlands, Austria, France, Italy, Germany, Romania, Germany, Ireland, and Great Britain. The Netherlands can be explained by several large hosting centers, and the Austrian share by the proximity to the exit node (e.g. there are Google servers within the Austrian university network, so appropriate traffic exiting our node is directed to them and not the "main" servers in the US).

When taking into account the number of flows per country, the EU has a slightly smaller share of the connections, but a larger amount of data transferred per flow. Exactly the reverse is true for Korea. However, really outstanding are the anonymous connections (country A1), which are comparatively huge (3.3 MB, so approx. 60 times the size of the average flow, which is roughly 56 kB large). Also interesting is the small size of the remaining traffic: While the table covers 99% of all bytes, it only accounts for 97.2% of all flows. Correspondingly, these flows are on average only 20kB large. From this follows, that connections to these countries are very small. Either these are "shorter" protocols (see SSH below), or these countries host smaller webpages (e.g. no videos) – as most traffic is webbrowsing. Why the latter should be true is unclear, so a more detailed investigation would be interesting.

Regarding AS the result is less clear, as traffic is spread over a significantly larger number. While 12 AS account for 50.2% of all traffic (see Table 2), the rest is enormously distributed. With the top 120 AS only 73.7% of cumulated traffic has been reached (compare this to countries: 60 countries sum up to 99.7% of all traffic volume).

One conclusion from these statistics is that almost all of the top AS are not (only) actual "service providers". For example, it is unlikely that all traffic to "Amazon.com, Inc." is directed towards the shop, but rather that a large part is "hosting" (e.g. cloud services) provided to third parties. The same applies to Google. This can be seen from other AS on the list, which are merely providers of networks/hosting/ housing. This means that illegal traffic that might occur there is often not caused by their own business, but rather their customers. This could aid enforcement, as these providers are probably more willing to help the police if they are not directly affected. Compare this e.g. to Facebook, where all traffic is the responsibility of the social network site and any investigation would target the company directly. See also the legal situation in the EU: If hosting providers do not know about illegal content, they are not liable (see the E-Commerce directive 2001/31/EC Art 14). But as soon as they obtain knowledge, they must take immediate action. So to avoid liability, removing content or restricting customers is typically easy to achieve with such providers (e.g. Amazon; for them most individual customers are sufficiently unimportant to not care unduly about restricting or losing them). In contrast to this, Facebook always "knows" about the content of their network and cannot profit from this exemption.

Another, but more surprising, result is the huge disparity in the transfer sizes. Some AS show very large flows, i.e. a lot of traffic but few connections. For example Google has an average size of 94 kB/flow, while Voxility has 3.26 MB/flow (the maximum is AS "S29927" on place 70 with 36.5 MB/flow). It seems therefore that some companies specialize in hosting large files, while others provide for a more balanced set of customers. Note that this need not apply to the company as a whole, merely to that share of their traffic that is accessed via Tor.

There exists only a slight correlation between down-/upload ratio and size per flow, however (correlation coefficient 0.48). So if there is more down- than upload, then the probability of large transfers increases – or in reverse: large transfers are more likely to be down- than uploads, which makes sense for a mostly HTTP environment (see below).

To be able to inspect content for whatever reasons, e.g. to prevent malware (scanning for viruses, trojans, etc.) or detect other illegal content (like brute-force cracking attempts), the communication must take place unencrypted. Within the Tor network itself, all data is encrypted, so at or immediately after exit nodes is the only possible place for detecting such activities (apart from the end-user's client). In our system we assigned ports an (assumed) encryption status according to their protocol. Some ports are typically unencrypted (like 80 – HTTP; but see below), others always encrypted (like 443 – HTTPS), and some cannot be determined reliably without content inspection (e.g. 110 – POP3

Table 2: Traffic per destination AS.

Autonomous System	*Traffic per day [GB]*	*Share of traffic*	*Flows per day*	*Share of flows*	*Avg. traffic per flow [KB]*
Google Inc.	97.44	8.57%	10,859,990	5.12%	94.1
ACOnet Backbone	80.78	7.11%	604,109	2.85%	140.2
Limelight Networks, Inc.	54.68	4.81%	196,640	0.93%	291.6
Highwinds Network Group, Inc.	49.05	4.31%	146,268	0.69%	351.6
Amazon.com, Inc.	48.53	4.27%	1,263,938	5.96%	40.3
Voxility S.R.L.	45.60	4.01%	14,326	0.07%	3,337.3
Webzilla B.V.	42.52	3.74%	298,343	1.41%	149.4
OVH SAS	38.63	3.40%	931,562	4.39%	43.5
LeaseWeb Netherlands B.V.	37.06	3.26%	651,604	3.07%	59.6
Cogent Communications	32.19	2.83%	7,5802	0.36%	445.2
CloudFlare, Inc.	23.01	2.02%	606,039	2.86%	39.8
Facebook, Inc.	21.44	1.89%	727,917	3.43%	30.9
Aggregated rest (all other AS)	579.48	49.78%	14,593,842	68.85%	40.7

is typically unencrypted, but often STARTTLS is used to switch to encrypted data transfer inside).

Unfortunately, the results are only of limited use: While (presumably) unencrypted traffic accounts for 62.4%, encrypted traffic is 37.2% and unknown traffic merely 0.4%, practically all unencrypted traffic (61.9%) is targeting port 80. This leaves 0.5% of other unencrypted traffic. Although this seems little, it still amounts to 5.53 GB/day, so definitely a significant amount of unprotected data exists, regardless of the uncertainty of the actual content within HTTP. Table 3 summarizes the distribution.

Table 3: Traffic categorization.

Kind of traffic	*Traffic per day [GB]*	*Share of traffic*
Port 80 (Unencrypted/ Encrypted Non-HTTP)	703.83	61.9%
Other unencrypted traffic	5.40	0.47%
Encrypted traffic	423.23	37.23%
Unknown traffic	4.36	0.38%

As Tor is a proxy only on the transport layer (in contrast to the application layer) we do not know for certain whether port 80 is actually used for HTTP or for some completely different protocol (e.g. BitTorrent as other studies suggest, OpenVPN, or other encrypted protocols occasionally configured to use port 80 to work around limited firewall configurations). Additionally, while HTTP is always unencrypted, the payload within (e.g. an "upload" by POST request), may very well be encrypted. More detailed information regarding this traffic cannot be derived unless the actual content data would be inspected. Using regular expressions (detecting HTTP verbs) and calculating the entropy to estimate encryption status (would need to take HTTP headers into account to distinguish it from compression) would allow at least some privacy-preserving investigation and is subject to future work.

So while there is still a not insignificant share of unencrypted traffic, a large part is (based on port numbers) encrypted and presumably some part of the unknown traffic is as well. As encryption in the web is increasing, the potential for content inspection can be expected to diminish. This is a problem for law enforcement, but also prevents general malware scanning of exit traffic (e.g. traffic with known attack signatures. This is offset with the obvious gain in privacy and the prevention of modifications – which especially for HTTP traffic is an issue (introducing additional scripts to subvert anonymization or changing the page content).

With regards to used services, our analysis is based on metadata, specifically port numbers. While most traffic is port 80 and 443 (=Web, but see above), one additional port shows up: port 22 (SSH) with 2.88% of all flows, but merely 0.12% of all bytes. The statistics show that while there are many SSH connections, these are mostly very brief (2.4 KB/flow on average, which is a about the size of a connection try with an unsuccessful login); see also Figure 2. Corroboration for this is that "scanning", i.e. trying SSH connections to more or less random IP addresses or brute-force attacks on SSH passwords appear in the abuse reports we receive. Therefore, the exit node is probably also used for malicious exploration and intrusion traffic.

Another port standing out is 43, the WhoIs protocol. A significant number of requests regarding the existence and ownership of domain names take place via Tor. This must not be confused with DNS:

retrieving IP information for domain names or vice versa. Reasons for this are currently unclear, but this might be used by malware to check for the existence of "random" domain names used for control servers in "fast-flux" networks or for gathering contact information, e.g. for later sending spam.

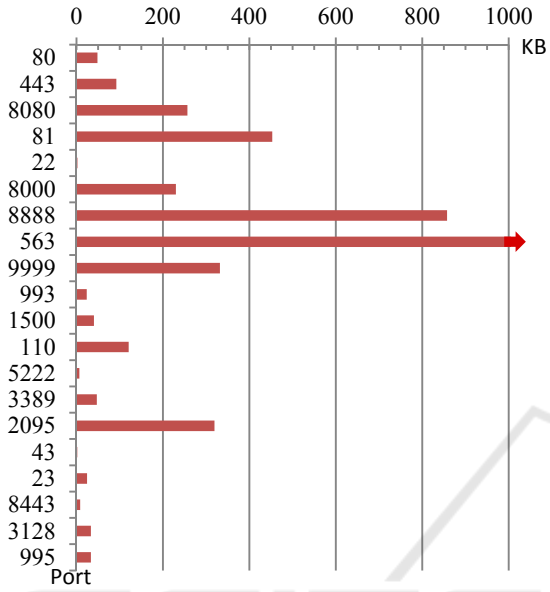

Figure 2: Traffic/flow per port in KB.

Another interesting detail is that the amount of data transferred per connection is varying strongly: at the upper end NNTP (port 563) connections start with approx. 2748 KB/flow, then comes port 8888 (HTTP/S, probably proxies) with 837 KB/flow. Then it drops down to about 442 KB/flow with an exponential decline down to port 5900 (VNC) with 0.9 KB/flow (but still 2500 connections/day; probably mostly just password scanning). Port 80, while covering the largest traffic share, is placed only number 18 in this ranking, with 47.6 KB/flow. Based on the size this looks like actual web traffic (Callahan et al., 2010; a study between 2006 and 2009: average GET transaction size increased from ca. 12 kB to 28 kB) than other tunnelled protocols, like BitTorrent, where individual pieces are often 64 kB or significantly larger. The full distribution is shown in Figure 2 (note that the value for port 563 – NNTP – has been cut and is 2.8 times as long as shown), which is sorted from top to bottom according to the total traffic amount.

Although HTTP(S) traffic is overwhelming concerning the data amount, especially "messaging" (both encrypted&unencrypted: POP, IMAP, NNTP, XMPP) as well as remote access technologies (SSH, Telnet, RDP) are also in significant use.

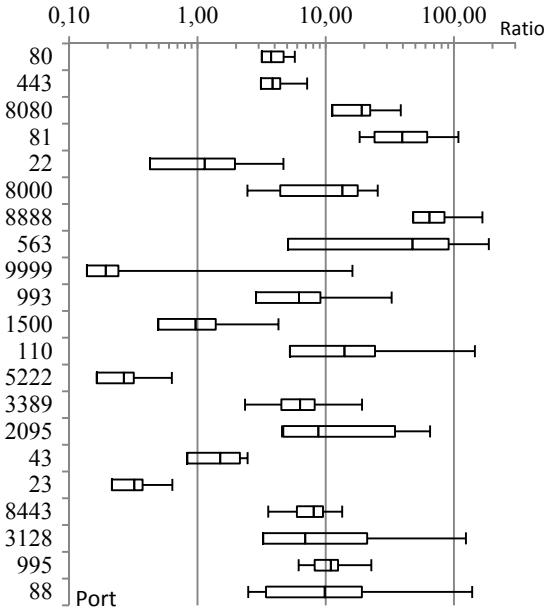

Figure 3: Down-/upload ratio per protocol.

Figure 3 shows the ratio between down- and upload data amount for these protocols (note the logarithmic scale) as a boxplot (minimum, first quarter, median, third quarter, and maximum) as they vary during the hours of one day. So while port 80 is in all hours on average downloads with the download being on average 25 times larger than the upload, SSH, port 9999, Telnet and XMPP are used for many uploads as well.

5 CONCLUSIONS

A limitation of this investigation is that it is based on data from a single exit node in a single country, although with a large bandwidth. Aggregating data from several sources would be technically no problem, but could lead to privacy issues if one or more of the participating nodes were malicious.

Almost all traffic is directed towards few countries, but remaining traffic is widely distributed. For criminal investigations this means that any significant amount of illegal data transferred using Tor (e.g. copyright violations such as movie downloads) can be easily handled with connections to few countries, but if small data transfers are part of the activity (for instance a single high-value fraudulent E-Mail), practically any country might be involved.

Regarding Autonomous Systems the situation is less ideal, as traffic is more widely spread. However,

top targets are typically large network/hosting providers, which are probably willing to assist if approached correctly (e.g. with judicial orders). Still, a significant share of traffic is extremely distributed, and especially for smaller providers the necessary knowledge, resources, and willingness might be limited. So if attacks were detected automatically, notifying targets would result in a perceptible burden, as many organizations need to be contacted.

If data from an "exit" (=uplink) of a normal small/medium-size ISP were available, a comparison to ordinary traffic would become possible. In this way it could perhaps be (dis-)proven that Tor exit traffic closely resembles normal traffic and therefore does not pose a special danger of illegal use.

A large share of traffic could be unencrypted (HTTP), but without content investigation this cannot be guaranteed and remains a task for further investigation - including deep packet inspection, if associated privacy&legal issues can be solved. Still a significant part, (presumably) about one third, is encrypted, and direct content investigation is impossible. While definitively (apart from probably – see above) unencrypted traffic is only a tiny part, this still amounts to a significant amount of data, posing notable risk if a fraudulent exit node were involved.

Some traffic we see on our exit node appears strange already from the outer metadata. While it might be useful to ask for the owner of a domain anonymously, e.g. when considering to buy it, this cannot explain the large number of WhoIs requests. Similarly, part of the SSH traffic is suspicious: While using it to connect to a server does not grant anonymity against this server but only anyone observing the traffic, the tiny average connection size hints at brute-force password cracking.

ACKNOWLEDGEMENTS

We would like to thank both the Johannes Kepler University Linz as well as the AcoNet for supporting this project by granting permission and providing necessary bandwidth. We also thank Heinrich Schmitzberger for patching the Tor source code to enable marking exit traffic for correct monitoring.

REFERENCES

asn, Some statistics about onions, [online] Available at: https://blog.torproject.org/blog/some-statistics-about-onions [Accessed 21.9.2016]

Ailanthus. 2015. Ethical Tor research: Guidelines, [online] https://blog.torproject.org/blog/ethical-tor-research-guidelines [Accessed 21.9.2016]

Akamai, 2015. akamai's [state of the internet] / security Q2 2015 report, [online] https://www.akamai.com/uk/en/multimedia/documents/state-of-the-internet/2015-q2-cloud-security-report.pdf [Accessed 21.9.2016]

Biryukov, A., Pustogarov, I., Thill, F, and Weinmann, R.-P. 2014. *Content and Popularity Analysis of Tor Hidden Services*, ICDCS Workshops 2014, 188-193

Callahan, T., Allman, M., and Paxson, V. 2010. A longitudinal view of HTTP traffic. *Proceedings of the 11th international conference on Passive and active measurement (PAM'10)*, Springer-Verlag, 222-231.

Chaabane, A., Manils, P., and Kaafar, M.2010. Digging into anonymous traffic: A deep analysis of the Tor anonymizing network, *Proceedings of the 4th International Conference on Network and System Security (NSS)*, 2010, 167–174.

INS, 2016. Tor system setup, [online] Available at https://www.ins.tor.net.eu.org/tor-info/index.html [Accessed 21.9.2016]

Jansen, R., Johnson, A., 2016. Safely Measuring Tor. *Proceedings of CCS'16*. To appear

Ling, Z., Luo, J., Wu, K., Yu, W., and Fu, X. 2015. TorWard: Discovery, Blocking, and Traceback of Malicious Traffic Over Tor, *IEEE Tr. on Information Forensics and Security*, Vol 10/12, 2515 - 2530

Loesing, K., Sandmann, W., Wilms, C., and Wirtz, G. 2008. Performance Measurements and Statistics of Tor Hidden Services, *Applications and the Internet. SAINT 2008.* Turku, 2008, 1-7

Loesing, K., Murdoch, S. J., and Dingledine, R. 2010. A case study on measuring statistical data in the tor anonymity network, *Proceedings of the 14th international conference on Financial cryptograpy and data security (FC'10)*, Springer, 203-215

MaxMind, *GeoLite2 Legacy Downloadable Databases*, [online] https://dev.maxmind.com/geoip/legacy/geolite [Accessed 21.9.2016]

McCoy, D., Bauer, K., Grunwald, D., Kohno, T., and Sicker, D. 2008. Shining light in dark places: Understanding the Tor network, *Proceedings of the 8th International Symposium on Privacy Enhancing Technol. (PETS)*, 63–76

Pmacct project, [online] http://www.pmacct.net/ [Accessed 21.9.2016]

Soghoian, C., 2011. Enforced Community Standards For research on Users of the Tor Anonymity Network, *Proc. 2011 International Conference on Financial Cryptography and Data Security*, Springer, 146-153

Sonntag, M., 2015. Rechtsfragen im Zusammenhang mit dem Betrieb eines Anonymisierungsdienstes. *JusIT* 6, 2015, 215-222