Analys av The Onion Router

Axel Lönnqvist (ax2568lo-s@student.lu.se) och Jakob Nabialek (ja7506na-s@student.lu.se)

***Abstract*—The Onion Router (Tor) is a distributed network enabling anonymous internet communication by re-routing encrypted TCP data through relays hosted by over 7000 volunteers throughout the world. In this study, Wireshark is used to examine the data being relayed through a Tor exit node hosted on a Google Cloud server. The result show that web traffic makes up the vast majority of the bandwidth and connections while keeping the throughput rather volatile. Further, the connections' country of origin is being investigated and a method to determine if a connection is established by a user or in fact by another relay is being presented. The security concern is also being addressed and sensitive information is identified in the collected data.**

I. INTRODUKTION

Idag har hela 95% av Sveriges befolkning tillgång till internet i hemmet [1]. Parallellt med denna utveckling lyfts frågor om säkerhet och integritet på internet i allt större utsträckning och många oroar sig för vem som faktiskt ges tillgång till ens digitala fotavtryck. Anonymiseringstjänster ämnade att försvåra spårning av internetaktivitet växer i användarantal och tjänsten Tor, "The Onion Router", uppskattas i dagsläget ha uppemot 2 miljoner användare dagligen [2]. Nätverkets ursprungliga syfte var att skydda hemlig kommunikation mellan regeringar och journalister men handlar idag främst om att stödja och upprätthålla yttrandefrihet och bistå människor i länder med stark politisk censur [3]. Antalet användare som anslutit till tjänsten de senaste åren syns i figur 1 och tros öka ytterligare framöver. Tor är fritt tillgängligt och går att komma åt från såväl den egna webbläsaren Tor Browser som operativsystemet Tails. Vidare förlitar sig nätverket på volontärer som delar med sig av sina anslutningar vilket är en förutsättning för att tekniken ska fungera och för att användarnas anonymitet ska kunna garanteras.

För att få ökad förståelse för hur anonymiseringstjänster används och fungerar och för att utvärdera hur säkra de faktiskt är har vi valt att studera trafiken som flödar genom en Tor relay och kort därpå når internet.

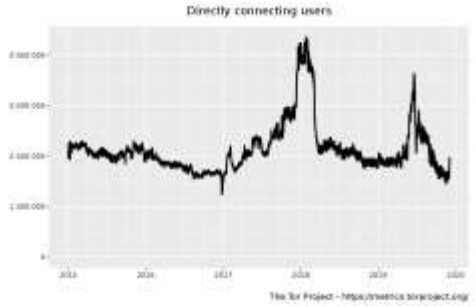

Fig. 1. Graf över antal användare som anslutit till Tor 2015–2019.

II. BAKGRUND

Under mitten av 90-talet tog amerikanska flottan fram en ny teknik för anonym kommunikation över offentliga nätverk kallad Onion Routing där meddelanden kapslades in i flera krypteringslager likt lagren som omsluter en lök. Knappt tio år senare lanserades andra generationens Onion Router och Tor, ett fritt tillgängligt distribuerat overlay-nätverk ämnat att anonymisera TCP-baserad webbsurfning och kommunikation och skydda användarnas integritet online genom att dirigera internettrafik över nätverkets idag drygt 7000 relays drivna av volontärer runt hela världen [4].

Principiellt handlar Tor om att överföra meddelanden inkapslade i flera krypteringslager genom en serie relayer (även kallade onion routers) i ett nätverk. Varje relay dekrypterar ett lager åt gången vilket avslöjar var informationen ska skickas vidare. När det sista lagret dekrypterats görs det ursprungliga meddelandet synligt och kan överföras till slutdestinationen. Den ursprungliga avsändaren förblir anonym då varje relay i nätverket endast har kännedom om följande och föregående relayer i kedjan (bortsett från den första relayen som vet vem avsändaren är men inte känner till vilken slutdestinationen är).

Figur 2 visar schematiskt hur ett meddelande tar sig från relay till relay genom nätverket i vad som kallas en "circuit". Notera att det finns flera typer av relays mellan användarens klient och destinationen, dessa vidarebefordrar meddelanden mellan varandra men kan även konfigureras så att de tillåts vidarebefordra trafik utanför Tor-nätverket till internet och kallas då för "exit relays".

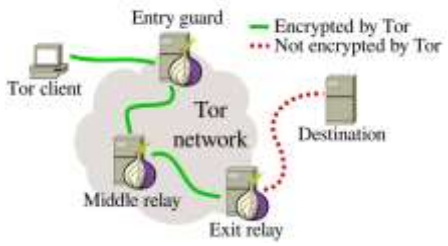

Fig. 2. Schematisk bild av hur Tor fungerar

När en användare upprättar en anslutning till en server via Tor väljs tre slumpmässiga relayer ut och bildar en slags krypterad tunnel, en circuit, som meddelandet kan ledas igenom. Inledningsvis tas information, som kan användas för att identifiera användaren, bort med hjälp av applikationsfilter varpå data överförs över nätverket genom TLS-krypterade anslutningar. Tor bygger stegvis upp en circuit och samlar in nycklar från samtliga relays innan data börjar skickas. När alla nycklar tagits emot bryts meddelandet ned i så kallade fixed size cells om 512 bytes och krypteras därefter iterativt i tre lager med hjälp av nycklarna tillhörandes respektive relay i kretsen.

Notera att krypteringen sker i omvänd ordning mot hur meddelandet rör sig genom kretsen, alltså krypteras meddelandet först med nyckeln som delas med den sista relayen följt av den mellersta och första [4]. Tekniken illustreras i figur 3 och förklarar varför en mellanliggande relay aldrig känner till mer än vilka de föregående och nästkommande relayerna är.

När servern mottagit meddelandet, som nu bär exit-relayens IP-adress, skickas svaret tillbaka genom tunneln på motsvarande sätt utan att användarens IP-adress exponeras och denne förblir anonym. När användaren vill besöka en annan sida väljs en ny väg genom Tor-nätverket [5].

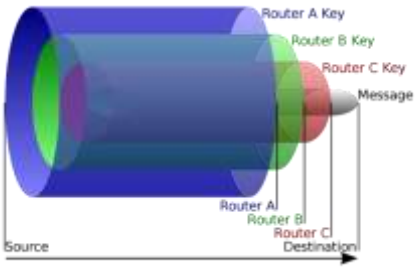

Fig. 3. Övergripande förklaring av hur meddelanden krypteras av Tor

Notera att trafiken som lämnar exit-relayen inte är krypterad utan skickas i samma format som den inledningsvis levererades i till Tor, detta gör det möjligt att studera infångade data och identifiera vilka meddelanden en användare skickar till en webserver.

Vi är införstådda med att det finns allvarlig integritetsproblematik som måste adresseras vid insamlandet av uppgifter från ett anonymt nätverk likt Tor som ju är utformat för att förhindra trafikanalys och övervakning. Denna integritetsaspekt övervägdes noggrant när vi valde ut vilken information som skulle presenteras och analyseras. Värt att nämna i sammanhanget är återigen att all kommunikation mellan relayerna, såväl som länken mellan användaren och den första relayen i kedjan, är krypterad. Förvisso är länken mellan vår exit relay och webbservern som användaren vill besöka ej nödvändigtvis krypterad men det är trots detta inte troligt att användarens identitet avslöjas genom den typ av avlyssningsteknik som tillämpades vid detta försök.

III. Frågeställning

Syftet med arbetet är att analysera och studera trafik som skickas och tas emot av en exitkonfigurerad relay som del av anonymiseringsnätverket Tor.

Intressant att undersöka är vilka protokoll som används, hur genomströmningen ser ut, hur god säkerheten är samt om det går att dra några intressanta slutsatser kring de relayer som ansluter till vår relay vad gäller vad för typ av relay de är och hur de är spridda rent geografiskt. Frågorna vi avser besvara är följande:

- Vilka applikationsprotokoll används av Tor-nätverket?
- Hur varierar genomströmningen över tid?
- Kan vi hitta känslig information i den okrypterade trafiken?
- Vilka ansluter till vår relay och var befinner de sig?

IV. Metodbeskrivning

Av säkerhetsskäl ville vi inte driva en exit relay lokalt från eget nätverk då detta är förknippat med många risker. För att på ett säkert, enkelt och smidigt sätt kunna köra och kontrollera en exit relay iordningställdes istället en virtuell server genom att utnyttja Google Cloud, vilken kan liknas vid en vanlig virtuell maskin i molnet. Linux-distributionen Ubuntu 16.04 LTS installerades varpå vi kunde ansluta till servern genom Secure Shell (SSH) och styra den med kommandon via terminal.

På detta sätt kunde vi fjärrstyra en dator i molnet och därigenom installera och konfigurera Tor. Vi satte vår OR-port (Onion Routing-port), genom vilken all trafik skickas vidare, till 9001 (standard). Efter att konfigurationen av relayen färdigställts genom att definiera parametervärden kopplade till mängd bandbredd vi vill uppgiva (5 MiB/s) samt vilken typ av trafik vi vill tillåta kunde relayen slutligen göras tillgänglig för anslutningar på Tor-nätverket och vi kunde se trafikflödet tillta.

Fig. 4. Första tecknet på liv – data flödar genom relayen

För att fånga in och spara ner de paket som gick genom relayen till en fil användes verktyget `tcpdump` och kommandot `sudo tcpdump -i ens4 -s 65535 -w <filnamn>.pcap` exekverades. Efter att ha fångat in trafik i nära tre timmar avbröts dumpen och filen kunde laddas ned för vidare analys i Wireshark. Utöver enklare portfiltrering (tcp.port==9001) användes även verktygen I/O graph (för att visualisera bandbreddsanvändningen över tid), Endpoints (med MaxMind GeoLite2 som tillägg för att lokalisera IP-adresser) och Protokollhierarki för att skapa ett underlag att basera resultatet på.

För att avgöra om en anslutning gjorts från en annan relay i nätverket eller från en användare som använt vår relay som första steg i en circuit (Entry Guard) kontrollerades porttillgängligheten genom att använda verktyget `nmap`. Vanligtvis sätts OR-porten på en relay till port 9001 eller 443, varför man rimligtvis bör kunna anta att en anslutande part med någon av dessa portar öppna är en relay.

För att testa denna teori valdes två IP-adresser tillhörande klienter som konverserat med vår OR-port slumpmässigt ut från dumpen i Wireshark varefter vi med hjälp av `nmap` kontrollerade om port 9001 och 443 var öppna. Resultatet kunde sedan kontrolleras genom att utnyttja tjänsten ExoneraTor som tillhandahåller ett register över alla relayer som varit aktiva ett givet datum.

Notera att all trafik som samlades in under försöket enbart gick genom servern, vår egen setup påverkade därmed ej hur försöket fortskred och all datainsamling kan sägas ha ägt rum isolerat i molnet utan att ha påverkats av andra anslutningar som gjorts till och från vår egen dator.

V. VALIDITETSDISKUSSION

Resultatet som presenteras i denna studie är inte på något sätt menat att tolkas som en skildring av typisk dataaktivitet som pågår på Tor. Det finns alltför många variabler i ett distribuerat nätverk för att något av resultatet ska kunna anses representativt. Denna begränsning har främst att göra med det faktum att genomströmning, bandbredd, geografisk plats och politiska avgränsningar påverkar den data som går genom en relay. Det är därmed svårt att dra generella slutsatser kring användandet av nätverket speciellt då datainsamlingen sker under en så pass begränsad tidsperiod som vår.

Konfigurationen av den enskilde användarens Tor-klient varierar även i hög grad då användaren själv kan styra hur dennes trafik ska hanteras av nätverket och kan bland annat välja genom vilka länder trafiken ska gå.

Exempelvis är en användare baserad i ett land vars regering censurerar internet och därmed begränsar dennes rätt till yttrandefrihet mer benägen att konfigurera sin Tor-klient så att trafiken går genom länder utan censur. Kort sagt är Tor ett nätverk som går att konfigurera och använda på mycket mer än ett givet sätt samtidigt som det är högst dynamiskt och beroende av många faktorer varför vårt resultat inte bör tolkas alltför ordagrant.

VI. RESULTAT

De applikationsprotokoll som TCP-strömmarna bestod av varierade något men utgjordes främst av HTTP och TLS. Protokollfördelningen går att skåda i tabell I nedan.

TABELL I
PROTOKOLLFÖRDELNING FÖR ANSLUTNINGAR GJORDA TILL RELAY

Protokoll	Andel av totalt antal TCP-anslutningar	Andel av total datamängd
HTTP	42,4 %	34,8%
TLS	34%	43,6%
SSH	12,3%	2,1%
BitTorrent	6%	7,2%
Osäkra (ftp, email, telnet)	0,12%	0,14%
Övrigt	1,18%	4,3%
Okända	4%	7,86%

Vad gäller genomströmning kan man genom att studera I/O-grafen i figur 4 konstatera att den högsta uppnådda hastigheten låg på ca 4,1 MiB/s vilket kan jämföras med den övre gräns på 5 MiB/s som vår server var konfigurerad till. Notera dock att medel låg på betydligt lägre än 4,1 MiB/s.

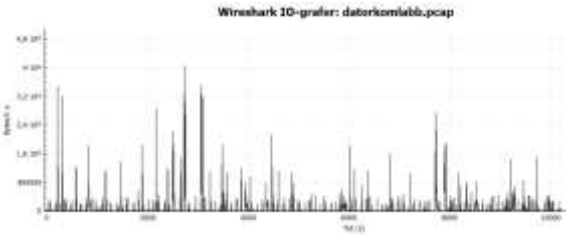

Fig. 5. I/O-graf som visar genomströmning under försöket

Vad gäller säkerhetsaspekten som utgör en ytterst kritisk faktor för Tor har vi dessvärre kunnat konstatera att flertalet känsliga uppgifter kunnat identifieras i klartext vid studerandet av dekrypterade pakets data. I övrigt har vi även haft insyn i vilka webbplatser användare som använt vår relay som exit relay besökt.

Resultatet från den kontroll som gjordes med hjälp av `nmap` av två slumpvis utvalda IP-adresser som anslutit till vår relay för att besvara huruvida en anslutning gjorts av en användare eller en annan relay visas i figur 6 nedan. I de fall som presenteras tros den ena anslutningen vara från en annan relay (IP-adress 82.64.163.188) och den andra från en användare (IP-adress 175.119.91.147).

Fig. 6. Nmap resultat för två anslutningar som gjorts till vår OR-port

Våra misstankar bekräftades av ExoneraTor som mycket riktigt indikerade att den ena IP-adressen var registrerad som relay det datum försöket gjordes (2019-12-04) medan den andra adressen inte var det.

Baserat på IP-adresserna för alla anslutningar som gjordes till vår OR-port under försöket kunde vi också klassificera dem efter ursprungsland, anslutningarna tycks domineras av framförallt amerikanska och tyska IP-adresser vilket återges i diagrammet i figur 7.

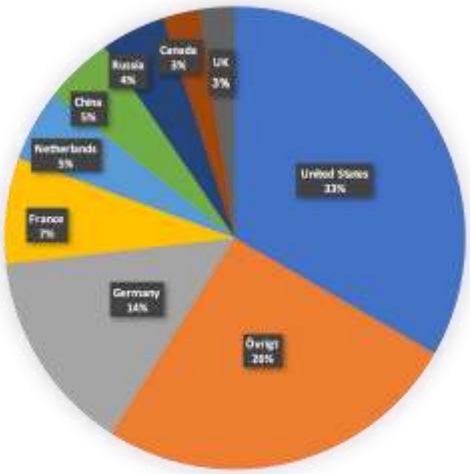

Fig. 7. Överblick av anslutningarnas ursprung

VII. DISKUSSION

Protokollen för de TCP-anslutningar som gjordes till vår relay består till största del av HTTP och TLS. Av TLS anslutningarna utgörs rimligtvis en stor del av krypterade HTTP-anslutningar (HTTPS). Övriga protokoll utgjorde mindre andelar men mer påtagliga anslutningar gjordes med hjälp av bland annat BitTorrent och SSH. Då BitTorrent i vanliga fall används av privatpersoner vid illegal fildelning är det rimligt att Tor utnyttjas i syfte att dölja fildelares identitet under denna process.

Genomströmningen varierade relativt mycket över tid och trots högsta uppnådda hastighet på 4,1 MiB/s förekommer

perioder då ingen trafik registrerades. Den höga variationen tros främst bero på att nya Tor relays sammankopplas med relays som har liknande historik av trafik [6]. Inledningsvis är denna trafik begränsad på grund av att Tor genomför kontroller för att verifiera att relayen fungerar som den ska och trafiken förblir låg under 3-8 dagar framöver då nya kretsar upprättas. Vår relay hade endast varit i drift i ungefär en dag då studien genomfördes så detta skulle mycket väl kunna förklara den förhållandevis låga genomströmningen.

Huruvida Tor är säkert att använda eller ej är till stor del upp till den enskilde användaren. För vanlig websurfning fungerar tjänsten relativt riskfritt och användarens anonymitet garanteras av flera lager kryptering och servern man ansluter till ser endast IP-adressen tillhörandes exit-relayen. Tidigare har emellertid sårbarheter i systemet uppenbarats vilka i enskilda fall utnyttjats för att avslöja användares identitet. Detta är dock ingenting som valts att undersökas i vårt arbete. Vad vi däremot kunnat konstatera är att man som användare bör vara försiktig med att uppge känsliga uppgifter om man inte säkert vet att ens anslutning till en webserver är krypterad. Om okrypterade meddelanden kommuniceras mellan en avlyssnad exit relay och en server på internet kan människor med onda avsikter få tag i känsliga uppgifter och göra stor skada. Eftersom en Tor-användares meddelanden är dekrypterade då de når exit-relayen kan de åskådliggöras förutsatt att ingen annan kryptering används i det sista steget till servern. Besöker man en sida som använder sig av exempelvis HTTP och inte HTTPS bör man därför vara uppmärksam på detta för att inte råka illa ut. I figur 8 syns ett exempel på känsliga uppgifter i form av inloggningsuppgifter i klartext som kunde identifieras i den data som samlats in under denna studie.

Fig. 8. Inloggningsuppgifter i klartext vid osäker HTTP-anslutning

Ur den geografiska spridningen framgår att de enskilt största länderna sett till antal anslutningar var USA med 33 %, och Tyskland med 14 %, av anslutningarna. Detta anses rimligt av demografiska anledningar då båda dessa länder är välrepresenterade sett till antal internetanvändare relativt andra länder i världen. Man kan även spekulera i att många amerikanska anslutningar etablerades på grund av att vår server var baserad i USA och därmed borde ha gett dessa användare snabbare anslutningstider men detta är högst osannolikt.

VIII. SLUTSATS

Som väntat tycks Tor, av protokollanvändningen att döma, i huvudsak användas för vanlig websurfning och till viss del även för fildelning och annan internetbaserad kommunikation.

Genomströmningen varierade väldigt mycket över tid, mellan 0 och 4 MiB/s. Då detta kan bero på kalibrering kan vi inte dra några generella slutsatser för hur genomströmning genom en godtycklig relay ser ut. Säkerhetsmässigt bör man vara försiktig då man använder sig av Tor, mer specifikt när man uppger känslig information. Detta främst eftersom det finns en risk att ens kommunikation avlyssnas. Man kan exempelvis motverka denna problematik genom att inte logga in någonstans då man använder sig av Tor. Den geografiska fördelningen av anslutningar visar att det främst är människor från större länder i västvärlden som använder sig av Tor och driver relays. Vidare vore det intressant att undersöka hur många av dessa anslutningar som görs av vanliga användare respektive av andra relays. Möjligtvis skulle detta gå att genomföra genom att utnyttja den nmap-metod som presenterades i detta arbete.

REFERENSER

[1] SCB, “Andel personer som har tillgång till internet i hemmet”. Hämtad 6 januari 2020. https://www.scb.se/hitta-statistik/statistik-efter-amne/levnadsforhallanden/levnadsforhallanden/befolkningens-it-anvandning/pong/tabell-och-diagram/andel-personer-som-har-tillgang-till-internet-i-hemmet/

[2] Tor Project, “Tor Metrics – Users”. Hämtad 6 januari 2020. https://metrics.torproject.org/userstats-relay-country.html

[3] Chaabane, A., Manils, P., and Kaafar, M.A., “Digging into Anonymous Traffic: a deep analysis of the Tor anonymizing network”, *2010 Fourth International Conference on Network and System Security*, DOI 10.1109/NSS.2010.47, september 2010. Hämtad 6 januari 2020.

[4] Dingledine, R., Mathewson, N., and Syverson, P., “Tor: The Second-Generation Onion Router”, *Proceedings of the Usenix Security Symposium*, 2004. Hämtad 6 januari 2020. https://svn-archive.torproject.org/svn/projects/design-paper/tor-design.pdf

[5] Barker, J., Hannay, P., and Szewczyk, P., “Using traffic analysis to identify The Secon Generation Onion Router”, *2011 Ninth IEEE/IFIP International Conference on Embedded and Ubiquitous Computing*. DOI 10.1109/EUC.2011.76, mars 2011. Hämtad 6 januari 2020.

[6] Tor Project, “The lifecycle of a new relay-relay”. Hämtad 6 januari 2020. https://blog.torproject.org/lifecycle-new-relay