Anonymous Communication Systems:
Usage Analysis and Attack Mechanisms

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in Computer Science

by

Esra Erdin

Dr. Mehmet Hadi Gunes/Thesis Advisor

May, 2012
We recommend that the thesis prepared under our supervision by

ESRA ERDIN

entitled

Anonymous Communication Systems: Usage Analysis And Attack Mechanisms

be accepted in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Mehmet Hadi Gunes, Ph. D., Advisor

George Bebis, Ph. D., Committee Member

Murat Yuksel, Ph. D., Committee Member

Cahit Evrensel, Ph. D., Graduate School Representative

Marsha H. Read, Ph. D., Dean, Graduate School

May, 2012
Abstract

Privacy in communication has been an essential security requirement with the rapid growth of Internet. This has resulted in growing interest in methods for anonymous communication. Several system designs have been developed with the common aim of preserving communication privacy within the shared network environment. In this study, the structure of these systems is explained and an analysis of current anonymity technologies; proxy servers, remailers, JAP, Tor and I2P, is presented. Because Tor is the largest anonymity system currently in use, a detailed analysis is also presented. Moreover, growing interest in anonymity systems resulted in several de-anonymization techniques developed to exploit these anonymous communication systems. These techniques are also explained to better understand the security threats of anonymous communication.
Acknowledgments

I would like to express my deepest gratitude to my advisor Dr. Mehmet Hadi Gunes for his guidance, understanding and patience he showed during this study.

I also would like to thank to my colleagues BingDong Li and Christopher Zachor for their effort and help to put together this project.

Finally, I could never have been able to finish this thesis without the endless support and love of my family. With all my sincerity, I would like to thank to my family.

This work was supported in part by National Institute of Justice with grant number 2010-DN-BX-K248.

Esra Erdin

University of Nevada, Reno

May
## Contents

Abstract ................................................................. i

Acknowledgments ....................................................... ii

List of Figures .......................................................... v

Chapter 1 Introduction .................................................. 1

Chapter 2 Anonymization Technology Usage ....................... 6
  2.1 Proxy Server ...................................................... 6
  2.2 Remailer ............................................................ 8
  2.3 Mix Network ........................................................ 11
  2.4 Onion Routing ...................................................... 13
    2.4.1 The Onion Routing .......................................... 14
    2.4.2 Invisible Internet Project .................................. 21

Chapter 3 Tor Usage Analysis .......................................... 24
  3.1 Related Work ...................................................... 24
  3.2 Tor Measurement Study .......................................... 25

Chapter 4 Application Perspective .................................. 30
# List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Proxy Server</td>
<td>7</td>
</tr>
<tr>
<td>2.2</td>
<td>Geographic Proxy Distribution (log-scale)</td>
<td>9</td>
</tr>
<tr>
<td>2.3</td>
<td>Geographic Proxy Distribution (cont.) (log-scale)</td>
<td>10</td>
</tr>
<tr>
<td>2.4</td>
<td>Remailer Geo-Distribution (log-scale)</td>
<td>11</td>
</tr>
<tr>
<td>2.5</td>
<td>A Mix Process</td>
<td>12</td>
</tr>
<tr>
<td>2.6</td>
<td>JAP Geo-Distribution (log scale)</td>
<td>13</td>
</tr>
<tr>
<td>2.7</td>
<td>The Onion Router (Tor) communication</td>
<td>15</td>
</tr>
<tr>
<td>2.8</td>
<td>Geographic Tor Server Distribution (log-scale)</td>
<td>17</td>
</tr>
<tr>
<td>2.9</td>
<td>Geographic Tor Server Distribution (cont.) (log-scale)</td>
<td>18</td>
</tr>
<tr>
<td>2.10</td>
<td>Geographic Tor Bandwidth Contribution (log-scale)</td>
<td>19</td>
</tr>
<tr>
<td>2.11</td>
<td>Geographic Tor Bandwidth Contribution (cont.) (log-scale)</td>
<td>20</td>
</tr>
<tr>
<td>2.12</td>
<td>Geographic I2P Server Distribution (log-scale)</td>
<td>22</td>
</tr>
<tr>
<td>2.13</td>
<td>Geographic I2P Bandwidth Contribution (log-scale)</td>
<td>23</td>
</tr>
<tr>
<td>3.1</td>
<td>Tor Usage (log-scale)</td>
<td>27</td>
</tr>
<tr>
<td>3.2</td>
<td>Geographic TOR Server Distribution</td>
<td>28</td>
</tr>
<tr>
<td>3.3</td>
<td>Geographic TOR User Distribution</td>
<td>28</td>
</tr>
<tr>
<td>3.4</td>
<td>Client Version Distribution (log-scale)</td>
<td>29</td>
</tr>
</tbody>
</table>
4.1 Combined Spam Data ........................................... 32

5.1 A link on the website that provides clickjacking ................. 34
5.2 JavaScript Code ................................................. 35
5.3 Facebook Like page ........................................... 36
5.4 Creating Button in Pdf File .................................... 38
5.5 Features of Button Field ...................................... 39
5.6 Embedded Video File ......................................... 40
5.7 Features of Embedding Operation .............................. 41
Chapter 1

Introduction

Privacy in communication has been an essential security requirement with the rapid growth of the Internet. Each time when we visit a website or send an email, our client sends packets of data through the Internet that contain essential information like who is sending the message and who is the receiver of the packet. Although the data transmitting via several hops from source to the destination may be encrypted, IP header is still visible to an observer. Adversaries can access significant information about the traffic carried between the sender and the receiver of the data. By observing communication through a link, third parties can easily identify who is communicating with whom based on source and destination IP addresses.

The term anonymizer refers to various tools that help the users to keep their activities private. Many situations like freedom of speech, anticensorship issues caused a growing interest in anonymous communication through the Internet. There are different situations in which someone wants to communicate anonymously on the Internet. Some organizations such as governments, military services may want some topics to remain secret as they are so critical and significant and their exposure may
be inconvenient or harmful to an investigation. Law enforcement agencies may want to design online forms that people can provide hints or other information they may have without the fear of retribution or punishment. Even private citizens may want to browse the Web without advertisers collecting statistics on their personal browsing habits. There are also people living under oppressive regimes that try to limit the rights of their citizens especially on the Internet [17]. Citizens of oppressive countries can express their thoughts, ideas through anonymizer networks.

Because of all the situations mentioned above, researchers design anonymity systems that build an overlay network running on top of the Internet. By using anonymizer systems, a user can communicate with another one without revealing their identities like IP addresses and location.

Since milestone research on anonymous communications Chaum [12], research on these systems has extended to many areas. Anonymous communication, traffic analysis, provable shuffles, anonymous emails, anonymous publications, private information retrieval, taxonomy, security and improvement, communication censorship, and anonymous voting are some of the areas that researchers have focused on [3, 14, 17, 27]. Several anonymity services are provided by either commercial companies or open source developers. Anonymizer.com is one of the commercial websites that requires its users to pay a subscription fee to keep their identity private while browsing web via Anonymizer.com servers. Moreover, Tor Onion Router (TOR) [16] and Invisible Internet Project (I2P) [4] are open source systems which are hosted by volunteers.

A communication network is composed of messages, senders, that send messages, and recipients, that receive messages. Sender and recipient can be any client, server or peer in a communication network. These subjects exchange messages via
public communication channels. There is also a possibility to be an attacker who may be interested in monitoring the traffic transmitted between sender and recipient. The attacker can be outside of the communication network, or inside of the communication network. In order to better understand anonymous communication systems, it is essential to understand the terms *anonymity*, *unlinkability*, and *unobservability*.

**Anonymity**: In their proposal Pfitzman and Hansen describe anonymity as "the state of being not identifiable within a set of subjects, the anonymity set" [37]. All subjects that can cause an action constitute the anonymity set. In other words, a message can possibly be sent by any of the senders who belong to his/her sender anonymity set and, similarly, a message can be received by any recipient who might be anonymous within a set of recipient anonymity set. Here, we can consider a subject as a client, server or peer in a network. All subjects might have their own anonymity sets.

**Unlinkability**: A subject of the system can make multiple uses of a resource without an adversary being able to link to these uses to the particular subject [41]. In other words, an adversary observing the senders and recipients of a network is unable to identify the communicating parties and the ability of understanding the communication between joined parties does not increase by observing the network.

The relation between anonymity and unlinkability is that anonymity of an item in a system is not linkable to any particular identity and the anonymity of a particular identity is not linkable to any item [41]. More clearly, any particular message sent through a network is not linkable to a particular sender in the sender set which ensures sender anonymity.

**Unobservability**: Observers cannot observe the subjects of interest from any other subjects in the communication network. As anonymity sets, unobservability sets
of subjects are defined with respect to unobservability [37]. Sender unobservability assures that it is not noticable whether any sender within the unobservability set sends the message. Similarly, recipient unobservability assures that it is not noticable whether any recipient within the unobservability set receives the message.

As indicated earlier, there has been considerable research effort on providing anonymous communications; however, not much information is covered on the current utilization of deployed anonymity systems. It is crucial to understand the deployment and usage of anonymous networks. In this thesis, I cover the analysis and usage of current anonymizer technologies, including; proxy servers, remailers, JAP mix network, TOR, and I2P.

Additionally, I analyze Tor usage in detail, as it is currently the largest anonymity system in use. Two data sets are used for Tor analysis: data collected by setting up a Tor relay as an exit node, and historical Tor measurements reported at Tor project website [7].

Moreover, the usage of anonymity technology from the application perspective; specifically spam-email is studied. For this, IP addresses of spam-emails are compared to the collected IP addresses of the anonymity servers during our studies.

Growing interest in anonymity systems has increased the attention to develop several attack systems to deanonymize these systems. Finally, various mechanisms that can be used to identify a user of an anonymizer network including; clickjacking, website fingerprinting, active documents, URI methods, network-level attacks are discussed.

Main contributions of this thesis is that, in addition to overviewing existing anonymous communication systems, it presents the analysis of these systems and gives a detailed analysis of Tor network, which is the most popular network among
all. Usage and server geo-location distribution of these systems are studied in de-
tail. Moreover, the usage of anonymity technology from the application perspective;
specifically spam-email, is studied which has not been seen in the literature previ-
ously. We also proved the applicability of several de-anonymization mechanisms to
accessing the users of anonymizer systems.
Chapter 2

Anonymization Technology Usage

Anonymous communication designs can be categorized in several categories based on network type, protocol type, anonymity properties, or adversary capability. However, generally, they are often divided into two categories: \textit{high-latency systems} and \textit{low-latency systems}. High-latency anonymity systems provide strong anonymity. These systems are mostly used for non-interactive applications that can tolerate high delays while communicating. On the other hand, low-latency systems often provide better performance as the delays of the communication is low. They are suitable for real-time applications like instant messaging and web browsing. The analysis of popular deployed anonymity systems will be provided in this section.

2.1 Proxy Server

A proxy server simply forwards all incoming traffic without any packet reordering, acting as an intermediary for requests from clients \cite{17}. A client connects to the proxy server requesting some service. The proxy server evaluates the request and relays the requests user to its intended destination as the proxy is requesting and relays the
responses to the user as in Figure 2.1. Significant improvement in the performance of the anonymous communication can be gained by using a single proxy. However, the proxy server knows about both the source and destination which allows it to trace the traffic between the source and destination. The improved performance of these systems can also increase the systems’ susceptibility to certain traffic analysis attacks by third parties [17].

![Figure 2.1: Proxy Server](image)

Web proxies are the most popular among all other proxy servers. Although there are many volunteer-based proxies, some commercial companies also offer proxy service such as Anonymizer.com and GoTrusted.com. Clients pay a subscription fee and then can relay their traffic through the servers operated by the company.

Figure 2.2 and 2.3 present the geographic location of 7,246 public proxy servers obtained on March 10-15, 2011. These proxies were collected from major announcement lists (i.e. anonymousproxylists.net, freeproxy.ru, fresh-proxy-list.net, free-proxy-server-list.com, multiproxy.org, nntime.com, proxies4u.com, proxynet.org, proxy4free.com, proxy.org, proxyliservice.com, tech-faq.com, proxyserverprivacy.com, www.samair.ru, and publicproxyservers.com,) and are a sample of available public proxy systems. Hence, this is not a complete list of public proxy servers but a representative sample for the purpose of this study. Note that, the figure is in logarithmic scale. Among the available public proxy servers from 112
countries, most were located in the U.S. (i.e., 1,869) and in China (i.e., 710). Moreover, 46 countries hosted more than 10 public proxy servers while 21 hosted a single server.

2.2 Remailer

Remailers are the servers that receive messages with several instructions embedded into it and forward the messages to the destination without revealing any information of the sender. In particular, they remove the identifying email headers before sending the message to the destination.

The first widespread public implementation of remailers is Cypherpunk remailers [24,36]. These are also called Type I remailers. These remailers forward messages between several systems before sending to the actual destination. Later, Mixmaster systems, which are also called Type II remailers, are implemented to be more resistant to traffic analysis [13]. These remailers include message padding, message pools and some other features to prevent emails to be monitored by an adversary [41]. Finally, Type III remailer, called Mixminion, was proposed in [15]. Free routing algorithm makes Type III remailers more secure compared to other ones. Every email transmitting through the connection passes several mixes and no single mix can see any other mix besides the adjacent mixes. Hence, no mix can link the message sender with any recipient.
Figure 2.2: Geographic Proxy Distribution (log-scale)
Figure 2.3: Geographic Proxy Distribution (cont.) (log-scale)
During our web/blog search in Oct 2010, we could only identify 15 active re-mailers as shown in Figure 2.4. Due to heavy usage of re-mailers by spammers or scammers, they are no more actively deployed.

2.3 Mix Network

Chaum’s seminal work in [12] was a base system for secure mix systems. Mix systems use mix servers to relay the messages to the destination. The principal idea is that all the messages that is desired to be anonymized relays through these mix nodes.

We can consider mix networks as public-key cryptographic primitive that takes a number of ciphertexts as input encrypted with the public keys of the mix nodes. Each mix has its own RSA public key, and messages are encrypted with these public keys after divided into blocks. The first block mostly includes the address of the other mix which acts as a header of the message. When the first mix node gets the message, it decrypts all the blocks with its public key, extract the first block of remaining blocks which includes the address of the recipient and appends a block of random bits at the end of the message. The random bit block is added to the message to make the size of the data fixed.

Figure 2.5 shows the working mechanism of mix networks. In the figure, path \( P \) of a message \( M \) consists of 3 mix process Mix-1, Mix-2, and Mix-3. The client
builds ciphertext $C$ by encrypting message $M$ along with random text $R$ using each mix’s public key $K$. The ciphertext (e.g., $E_1(A_{Mix-2}, R_1 + E_2(A_{Mix-3}, R_2 + E_3(D, R_3 + M)))$) specifies the exact path the message will take through the mix network. Each mix node (e.g., Mix-1) receives the ciphertext decodes one layer to find next hop destination (e.g, $A_{Mix-2}$) and forwards payload (e.g., $E_2(A_{Mix-3}, R_2 + E_3(D, R_3 + M)))$).

Decryption of ciphertexts and padding functionality are essential for mix networks as they aim to provide bitwise unlinkability. A link between the encrypted message arriving at the mix node and decrypted message departing from the mix should not be related by an attacker of the system. Flushing algorithms are also used in mix networks to enhance anonymity of users. Flushing algorithms determine which message will be forwarded to the other mix node, and when to flush. The algorithm buffer incoming messages into a pool and forward messages in rounds. This also provides unlikability and defends against traffic analysis attacks.
There are different variations of mix networks that were deployed over the time such as Crowds [40], Tarzan [19], and JAP [9].

JAP (Java Anon Proxy) is a low-latency mix cascade that uses servers provided by volunteers. The clients run JAP software to send the packets to the destination, mix cascades within JAP network forwards the traffic coming from the client. JAP cascades encrypted packets through several mixes and keep the traffic in a constant rate to avoid rate-based traffic analysis. The JAP program displays active mixes and users are able to select JAP cascades from those active mixes. Figure 2.6 presents the geographic location distribution of 11 JAP servers that were active on 12-19 Oct 2010. Compared to onion routing based systems Tor and I2P, JAP seems to have minimal usage at the time of our analysis.

2.4 Onion Routing

Onion routing was first developed by Reed, Syverson, and Goldschlag [21, 22, 39]. It is the most prevalent design for low-latency anonymous communication [21]. The structure of these systems is composed of layer upon layer of encryption which encapsulates the message, like an onion. The basic idea of onion routing is similar to the structure of mix cascades which is analogous to message transmission with multiple envelopes constructed by a sender, with the address of the first mix to be the outermost envelope. However; there is such a difference that, onion routing clients do not
necessarily forward the traffic for other clients. In onion routing mechanism, there is a set of servers called onion routers that relay the traffic for clients [17].

Like in mix networks, each node keeps a public and private key. The public key is known by the client to be able to set the path of the communication. The first step a client does in an onion routing system is to construct an encrypted tunnel, which is called circuit, through the network by using public-key cryptography. After setting up the circuit, symmetric key cryptography is used to transfer the data. Another big difference between mix networks and onion routing is that mix networks use flushing algorithms which may cause to keep the packet for an amount of time while waiting to receive adequate number of packets to mix together. The onion routing work real-time mechanism and relays the traffic as it arrives to the node. Although this decreases the delay time in onion routing, it increases the vulnerability of these systems to possible attack mechanisms like timing analysis attacks [26].

There are different variations of onion routers such as The Onion Router (Tor) [16], and Invisible Internet Project (I2P) [4]. These systems differ based on how the routing servers are organized; how the encryption algorithms are applied; how the tunnels are established; whether the transport-layer protocol uses TCP or UDP; or whether the clients relay traffic to other clients or not.

2.4.1 The Onion Routing

Tor, shown in Figure 2.7, is the most popular anonymity system, based on the number of users, as it combines the best parts of the previous methods (e.g., the directory discovery of routing servers for clients, telescopic circuit establishment, and location hiding). Directory servers are responsible for distributing signed information about known relays in the network [17]. Authoritative directory servers, currently 7 systems
trusted by Tor developers [23], are utilized to determine three-hop paths among volunteer servers using secured TCP connections. User messages are then encrypted using symmetric key encryption with a telescoping key establishment scheme. Encrypted messages forwarded through the established circuit to the dedicated exit router, which forwards the message to the final destination and relays replies back to the source.

To reduce latency, onion routers neither add intentional delays nor batch messages. Hence, they are vulnerable to timing analysis attacks in general [26]. Entrance and exit nodes are particularly important as they know the source and the destination of the communication, respectively. Hence, to protect client profiling, authoritative directory servers pick only a subset of existing systems, which seems to be reliable, to become entry nodes. Moreover, packets originate from the exit system from the destination’s perspective and may be questioned regarding user actions. Hence, Tor provides an option of not becoming an exit node in case the user does not want to deal with possible complaints.

Figure 2.7: The Onion Router (Tor) communication
Figure 2.8 and 2.9 presents a snapshot of Tor servers based on their geographic location during Oct 20, 2010 - March 15 2011. The values for each country represent the daily average number of servers with standard deviation value for that country. For this analysis, we monitored the authoritative directory servers to determine the total number and geographical location of Tor servers. During the sampling period, we identified 61,798 unique servers at 103 countries while Tor system has approximately 1,500 active public relays at a given time. Most of Tor relays are located in few countries. Similar to earlier studies [11, 33], United States and Germany had highest number of volunteers and consequently higher daily average number of servers. Considering continents, which are indicated with different lines and colors, Europe had the highest number of servers.

We also present the bandwidth contribution of each country in Figure 2.10 and 2.11. Similar to the geographic server distribution graphs, we calculated daily average bandwidth and standard deviation values for each country. Similar to the number of servers, United States and Germany has the highest bandwidth. Moreover, we observe that a considerable number of countries have a significant variance in contributed bandwidth while the number of servers is stable.
Figure 2.8: Geographic Tor Server Distribution (log-scale)
Figure 2.9: Geographic Tor Server Distribution (cont.) (log-scale)
Figure 2.10: Geographic Tor Bandwidth Contribution (log-scale)
Figure 2.11: Geographic Tor Bandwidth Contribution (cont.) (log-scale)
2.4.2 Invisible Internet Project

Invisible Internet Project (I2P) is another system offering anonymization services that identity-sensitive applications can use. The I2P network is strictly message based (i.e., UDP) but there are libraries that allow reliable streaming communication on top of the I2P network. I2P works by routing traffic through other peers and mainly is a closed system. Many applications can interact with I2P including mail, peer-to-peer, and IRC chat. However, as I2P does not focus on end-to-end delay it is not preferred for low latency applications.

Similar to Tor, I2P announces its peers. To analyze I2P usage, we collected active I2P relays by joining the system during Oct 11, 2010 - March 15, 2011. Figure 2.12 presents the geographic distribution of servers in 44 countries ordered by daily average number of servers. We also calculated standard deviation for each country to show the variance in the number of servers. Even though we had a longer sampling of I2P, we observed fewer servers than the Tor system. Moreover, like to Tor, Germany, U.S. and France are the top three when ordered according to daily average number of volunteers.

Figure 2.13 shows the bandwidth of I2P servers volunteering for each country. Similar to other figures, it is ordered by the daily average bandwidth of active servers we had monitored and also shows standard deviation for each country. We observe a considerable variation in both the number of servers and allocated bandwidth of I2P relays. This indicates instability of the volunteering nodes which may hamper longer communications.

\[1\] Countries were looked up from ipinfodb.com database.
Figure 2.12: Geographic I2P Server Distribution (log-scale)
Figure 2.13: Geographic I2P Bandwidth Contribution (log-scale)
Chapter 3

Tor Usage Analysis

In this chapter, we analyze Tor usage, as it is currently the largest anonymity system. We use two data sets for Tor usage analysis: data collected by setting up a Tor relay as an exit node, and historical Tor measurements reported at Tor project website [7].

3.1 Related Work

There have been some earlier studies that analyzed the Tor usage as it gained popularity [11, 29, 31–33].

One of the earliest studies by McCoy et al. analyzed how Tor is being used, how it is being mis-used, and who are its users [33]. In their experiments, the authors analyzed application-level protocols that use their nodes as exit node. According to their finding, interactive protocols, such as HTTP, make up 92% of the connections and consume 58% of bandwidth. Similarly, bit-torrent traffic consumes 40% of bandwidth even though it accounts for 3.3% of the connections. The authors also pointed to malicious usage of Tor routers and developed a method to detect malicious logging at exit routers. Moreover, they indicated that Tor has a global user base based on
its client distribution. Our results in Chapter 2 also indicate that Tor has the largest volunteer base among the anonymity systems.

Similarly, Chaabane et al. analyzed the applications that use Tor [11]. Authors monitored traffic on six Tor servers which were pairwise located in the U.S., Europe and Asia to inspect geo-diverse relays. Authors analyzed the HTTP and BitTorrent traffic in depth. They pointed out that BitTorrent consumes significant resources both in terms of the number of packets and the carried traffic volume. Finally, authors pointed that some users take advantage of the Tor servers as 1-hop SOCKS proxies and presented a method to detect such use.

Additionaly, Loesing observed the trends in the Tor network by analysing the public directory information [31]. The results of the study indicated that there are some problems in the Tor network that need to be addressed such as the efficient usage of bandwidth capacity, support for dynamic IP addresses, and upgrade of the Tor software version by relays.

Finally, Loesing et al. provided guidelines for a statistical analysis of Tor data focusing on countries of connecting clients and exit traffic by port [32]. Pointing to privacy issues, the authors derived guidelines for measuring sensitive data in anonymity networks. Moreover, they pointed to interesting cases such as the increase in the Tor usage by Iranian IP space in June 2009 after the Iranian elections; and Tor blocking by China and consequent increase in bridge usage by Chinese IP addresses.

### 3.2 Tor Measurement Study

In our analysis, first, with my colleague we set up one Tor exit node with the default policy at a dedicated Windows server using Tor 0.2.3.4-alpha version during September 5 - October 5 2011. We captured packet headers, i.e., IP addresses and
port numbers for both source and destination, and payload size. We used this data to estimate the usage ratio of Tor among Internet users. As indicated earlier, user privacy were protected as no personal information was logged.

Figure 3.1 presents Tor usage of the top 50 countries along with the average number of Tor users and relay servers from these countries. We calculate the usage ratio of Tor in each country as the ratio of the number of observed clients in the data set over the estimated number of the Internet users obtained from http://internetworldstats.com. Then, we normalized the values over the average where a value of $x$ indicates $x$ times the average usage. Interestingly, Italy, Ukraine and Finland had the highest ratios of Tor usage relative to its Internet users while U.S.A and Turkey are at the average.

Next, we used historical Tor measurements reported at Tor project website [7] to analyze the evolution of Tor deployment. Loesing presents statistical analysis of the Tor system [31], but in this study we focused on Tor usage. Figure 3.2 and Figure 3.3 present the number of Tor users and relay servers for the top 10 countries. We excluded other countries as they occluded the figures. We observe that since 2007 most relays are hosted in the United States and Germany. They also have the highest number of clients connecting to the Tor. However, considering Internet users, Tor usage is average for the United States. Moreover, South Korea and Iran have 3rd and 8th, respectively, highest Tor users while they are not among the top 10 Tor relays.
Figure 3.1: Tor Usage (log-scale)
Figure 3.2: Geographic TOR Server Distribution

Figure 3.3: Geographic TOR User Distribution
Finally, we analyzed the versions of Tor clients on June 23, 2011. Figure 3.4 presents the top 25 client versions of the day. As seen in the figure, majority of clients are the most recent stable versions but there are considerable number of earlier versions, some of which has known security vulnerabilities. Likewise, there are many early adapters of the alpha/beta versions of the client.
Chapter 4

Application Perspective

In this section, the usage of anonymity technology from the application perspective; specifically spam-emails is studied. For this, we compared the observed IP addresses of applications to the collected IP addresses of the anonymity servers discussed in Chapter 2.

Spam email data was collected using two approaches. First, we trimmed the IP addresses of spam emails collected from the department during Oct 2010 to November 2011. Then, we gathered publicly available spam email IP addresses of recent spammers from the Internet. An important issue was to obtain recent data sets as anonymizer server IP addresses change over the time (except for the commercial systems).

We collected 58,969 unique IP addresses of spam e-mails from departmental mail servers during Oct 2010 to November 2011. In this data, 154 IP addresses belonged to anonymity servers indicating that the spam e-mail was sent through an anonymity network. Figure 4.1-(a) presents the distribution of the utilized anonymizer technology. In this data set, proxies were utilized as spam relays more than the other
systems. Moreover, Figure 4.1-(e) indicates the geo-location of relays of spam servers, including the public data sets. China and United States are the main relays of spam emails in the local data set.

We also collected IP addresses of publicly available data sets (156,371 IP from address lists) that were recently marked as spam generators by public systems including spam-ip.com, okean.com, projecthoneypot.org, spam-ip-list.com, and blogspot.com. Among these IP addresses, 1,845 belonged to an anonymity network. We analyzed the distribution of utilized anonymizer technology and the server geo-location for spammer IP addresses separately for each data set, and combined all data set from these web sites. Figure 4.1-(b)-(c) represent data sets from spam-ip.com and okean.com, respectively. While most spams reported by spam-ip.com utilized proxies, the spams listed by okean.com utilized Tor.

Moreover, Figure 4.1-e presents the distribution of utilized anonymizer technology and the server geo-location for spammer IP addresses in combining all data from these web sites together with local data set. Germany, China, Canada and U.S. were the four major relay nodes for spammers among the 100 countries observed and account for 66% of all servers. Proxy servers and Tor servers were utilized the most and account for 96.5%. Interestingly, the commercial anonymizer system goTrusted.com was utilized by spammers to send spam e-mails.

According to the figure we can conclude that, overall Tor, proxies and I2P are the three major sources utilized by spammers to relay e-mails about 76.2%, 20.2%, 3.3% respectively and in total 99.7%. Moreover, servers in Germany, China, Canada, and U.S. were the main exit nodes of spam e-mails about 49.3%, 7.9%, 4.5%, and 4.4% respectively and in total 66.1%.
Figure 4.1: Combined Spam Data
Chapter 5

Deanonymization

Growing interest in anonymity systems has increased the attention to develop several attack systems to deanonymize these systems. In this chapter, we analyze various mechanisms that can be used to identify a user of an anonymizer network. In all of these attacks, a global adversar might modify and monitor flow through a subset of anonymizer nodes, so that, the attacker identifies certain anonymous users.

5.1 Clickjacking

Clickjacking is a malicious technique of tricking a web user to click on a button or a link on a different page from what the user perceives he/she is clicking on. Attacker uses transparent layers to achieve this attack. By this method the control of the computer of the user can be taken by the attacker by routing the user to another page owned by another application. An adversary can also direct the request of the user to a malicious server using this method and can get information about the user.

For example, imagine an attacker who builds a web site that has a button on it that says ‘click here for a free iPod’. However, on top of that web page, the
attacker has loaded an iframe with your mail account, and lined up exactly the ‘delete all messages’ button directly on top of the ‘free iPod’ button. The victim tries to click on the ‘free iPod’ button but instead actually clicked on the invisible ‘delete all messages’ button.

An exploitation of this attack can be done by abusing Facebook’s ‘Like’ and Google’s ‘+1’ functionality. We implemented a JavaScript code which seeks to make a user automatically ‘Like’ a unique link on their Facebook profile via clicking on any
link on the website. All we need to do is to include the JavaScript code to the webpage and create a link on the website as seen in the Figure 5.1. A piece of JavaScript code is showed in Figure 5.2. When the user clicks on the link, JavaScript embedded in the webpage automatically runs and makes the user automatically ‘Like’ the website as seen in Figure 5.3. The assumption with this attack is that the user is already signed into his/her Facebook account.

![Image of JavaScript code]

Figure 5.2: JavaScript Code

The same attack can be launched with Google+. A JavaScript code, which automatically makes the user ‘+1’ the website is embedded in the website. With the user’s clicking on any point on the website, JavaScript code runs immediately making the user ‘+1’ the website without the permission of the user.
In either case, malicious website can provide a unique link to each anonymous visitor so that they are identified.

![Facebook Like page](image)

Figure 5.3: Facebook Like page

## 5.2 Code Injection

Code injection is the exploration of a computer bug that is caused by processing invalid data. Code injection technique can be used to get the history and cookies of the web browser by injecting HTML or script to the web browser of the client. An adversary may achieve to get essential information of the user by executing malicious script through a vulnerable website. Moreover, in anonymizer systems, an exit node can inject HTML tags, scripts or flash as a component of the relayed webpage that can help in identifying the user [43]. In either case, a new connection to a predetermined server is initiated when the user runs the injected component on his/her client and the attacker can easily obtain the identity of the user.

## 5.3 Website Fingerprinting

Although the data sent through anonymous systems is encrypted, an attacker can still identify who is communicating with whom by observing how much data is transmitted through the network [25]. If the attacker has a list of guesses as to which site is
targeted, by observing the quantity and length of data packets received in response, he/she can build a fingerprint of what the website’s response traffic will look like. By observing target user’s traffic, attacker can compare observed pattern against stored fingerprint database to identify possible end points. Most modern anonymity systems split messages into equal sized packets to prevent such attacks.

5.4 Active Documents

A webserver owner can identify the clients connecting to that server even if they are using anonymous networks by running active documents on his/her system.

5.4.1 Pdf Documents

Acrobat forms allow one to embed a JavaScript in the Pdf file and can define fields in pdf documents. By getting Field object of these fields, one can trigger JavaScript using these objects and perform the desired action. By using this feature, one can create fields; button in this example, as in Figure 5.4. The user can define desired functions for these buttons by giving directions while defining these fields. In this example, we create a ‘Submit’ button, and give the command ‘Open a Web Link’ as can be seen in Figure 5.5. When the user clicks on ‘Submit’ button, the embedded action runs and directs the user to specified website, ‘www.google.com’ in this example.

Many actions can be defined for these fields, such as ‘Open a file’, ‘Play a sound’, ‘Execute a menu item’. The user is unaware of what will actually happen if he/she clicks on the button which might be named anything. By using this feature an attacker can direct the user coming to his/her website to any other page he/she wants and may infer user identity through the new connection that bypasses proxy settings.
Newer versions of Adobe Acrobat have the functionality to embed flash, sound, video files as can be seen in Figure 5.6. An adversary can attack the users visiting a website, by inserting a pdf file to the website which plays a streaming video that can be located in any server. Figure 5.7 shows the Insert Flash tool in pdf file. The user initiates the connection with the server by accessing the pdf file. However, after starting to play the video, the video hosting server continues to communicate with the user by bypassing established proxy tunnel. Again a webpage can direct the user to any malicious server that may be able to identify the identity and reveal location of the user.

5.4.2 Macro Word Attacks

Macro is a series of commands that are recorded so that they can be played or executed later automatically. Many applications, such as Microsoft Word and Excel, support macro languages. A macro virus attacks these kinds of applications, infecting the
documents and template files. Hence, if your computer has a macro virus, and you use a word template file to make a new document, it will also become infected by the virus. One of the common ways, a macros virus can harm your computer is by replacing the normal functioning macros, and causing a series of automatic actions
that destructs your files. These macros can also be used to send user identity to a malicious server.

In either case, active documents bypass proxy settings and establish a new connection that reveals user identity.

5.5 URI Methods

Uniform Resource Identifier (URI), is the generic term for all types of names and addresses referring to objects on the World Wide Web (WWW). It can be defined as the identification of an object enabling interaction with representations of the resource over WWW. Browsers have URI handling, which enables developers to pass applica-
Figure 5.7: Features of Embedding Operation

tion commands through the browser and install software on the user’s computer. It can be used in many applications, such as establishing a skype call to a person directly from a website. Because many browser implementations do not properly sanitize the data passed via the URIs, with quotes and null bytes, it is possible to circumvent the browsers’ ability to handle the incoming data and slip out of the sandbox. Hence, an attacker may be able to access the local file system of the victim.

Skype [5] is one of the applications that has released a security bulletin to address a remote vulnerability [6]. This vulnerability was due to an error in the
handling of “file:” URIs. By convincing a user to click on a specially crafted “file:” URI, a remote, unauthenticated attacker may be able to execute arbitrary code on victim’s system. This code can help reveal user identity.

Similar to Skype, Apple QuickTime [1] is also prone to a vulnerability that allows remote attackers to launch arbitrary applications and access files [2]. The issue arises because of improper handling of the “file:” URI. The data is passed to ‘url.dll!FileProtocolHander’ and can result in the processing of non-HTTP filetypes, if the file type in a URI supplied via the “qt:next” attribute isn’t recognized by the application. By supplying malicious QuickTime content to be processed by a vulnerable user, an attacker can exploit this vulnerability. The attacker can execute arbitrary malicious codes on the victim’s computer and may gain unauthorized remote access control.

In either case, the attacker can run `ipconfig` or `ifconfig` to obtain anonymous user’s real IP address and transmit it over the anonymizer network.

### 5.6 Network-Level Attacks

There are several attacks that a global adversary, who is able to monitor some portion of the anonymizer network, launch against anonymizer networks.

#### 5.6.1 Timing Attack

An adversary who is able to observe the connections entering and exiting the anonymity network can link inputs and outputs based on packet inter-arrival times. Because low-latency anonymity systems like onion routing do not introduce any delay or alter the timing characteristics of an anonymous system, they are mostly vulnerable to such timing attacks [17, 28, 35, 42, 46].
5.6.2 Predecessor Attack

The attacker tracks a stream of communications over a number of rounds. In each round, the attacker simply logs any node that sends a message that is part of the tracked stream. By this mechanism the attacker may be able to compromise the initiator of the communication [45].

5.6.3 Intersection Attack

By watching the communication between each user and cascades’ first mix, an attacker can see when the user is active and when he/she is idle or off-line. The attacker also watches all messages that leave the cascade. Some of these messages can be linked to previous messages and recognized as belonging to a sessions. This reveals when a session is active or inactive [10]. Using these analysis, the attacker can zoom on certain users to identify targeted anonymous user.

5.6.4 Multiplication Attack

This attack is a variation of injection attack where an exit node modifies replied messages that would trigger several new requests. Then, entry nodes may identify the user using the pattern of requested objects [44]. An attacker running several nodes in an anonymizer network can also accomplish multiplication attack on the network if entry and exit nodes of his/her circuit belong to the same communication circuit which the client builds before sending the message. The exit node controlled by the attacker marks the cells with an identifier and forwards them to the anonymizer network. If the entry node of the circuit is also driven by the attacker, when the node gets the packet it recognizes the packet and may find who is communicating with whom [30]. Similarly, replaying a cell or dropping a cell helps identifying the flow between entry and exit node [20, 38].
5.6.5 Circuit Clogging

To detect anonymizer nodes that are active in a flow destined to the server, circuit clogging can be performed by the malicious web server [18, 34]. By sending bursty traffic to the user, the server can monitor anonymizer nodes using a direct connection to itself. The anonymizer nodes that have higher delay during burst periods but not during silent times are the ones that provide the connection to the end user. This attack can also be done by a third party to observe the traffic between end users [8].
Chapter 6

Conclusions and Future Work

Privacy in communication has been an essential security requirement of communication. The rapid expansion of Internet, resulted in growing interest in methods for anonymous communication. Several system designs have been developed with the common aim of preserving communication privacy within the shared network environment. In this thesis, I explained the structure of anonymizer systems and present an analysis of the state of the current anonymity technologies including; proxy servers, remailers, JAP, Tor and I2P. According to the results, proxy servers and Tor are the most used anonymity systems, while remailers and JAP have minimal usage.

I also analyzed where each system’s infrastructure is located at and where each system’s users originate from. Although these systems have a worldwide usage, there are huge differences of usage ratios between different countries. For example, Germany and U.S. are among the top countries in terms of the number of servers and the number of users in all anonymizer systems. Moreover, the number of servers in these countries is extremely larger than the number of servers in a country that is listed towards the end of the list.
Additionally, because it is the largest anonymity system currently in use, I presented a detailed analysis of the Tor network. In the study, we observed that Internet users at European countries utilize the Tor network more than other countries as evidenced by the top 13 highest usage being from Europe even though they are not the top countries hosting the Tor routers. For example, Italy has the highest ratio of Tor usage even it is not among the top of the list of Tor server.

We observe that, newer anonymity systems have higher popularity because of the increased security mechanisms of these systems. This fact also affects the earlier systems to become extinct as anonymity requires crowds of users.

I also analyzed spam e-mails to observe the anonymity system usage by spammers. Although we do not observe much spam activity through anonymizer networks, it can be concluded that proxy servers are mostly used to send spam e-mails. Moreover, Germany and USA are the top countries these spam e-mails originated from.

Growing interest in anonymity systems also increased the attention to develop de-anonymization techniques. I also presented an overview of these techniques including clickjacking, code injection, website fingerprinting, active documents, URI methods, and network-level attacks.

The rapid development in anonymous systems is a sign of that they will have much wider area of applications in the future. A deeper analysis of these systems helps the Internet users seeking for privacy preserving methods for communication. Because higher popularity of Tor network, it would be beneficial to conduct the application level analysis of Tor network. Also detailed analysis of de-anonymizing techniques helps anonymous system developers to understand and detect possible vulnerabilities that their system has. These systems can be improved as more attack models identified.
Bibliography


