Classifying Social-Media Websites

Mark Miller, Advisor: Prof. Wei Lu
Department of Computer Science, Keene State College, NH.

Abstract

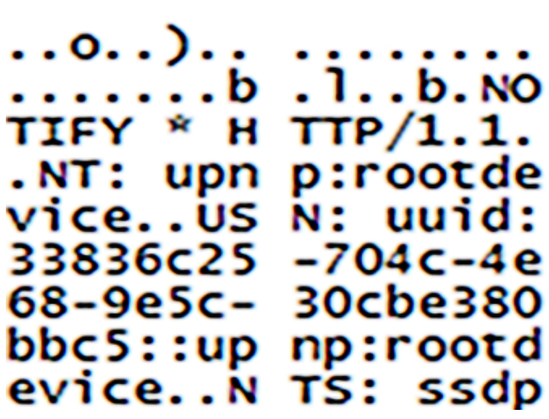
There are over two hundred social-media websites in widespread use, most famously Facebook and Twitter. Differentiating among such sites and interpreting the activities of their users, however, is very challenging and is still an issue to be solved. Previously attempted solutions included looking at information stored in network packets, which provides very limited information nowadays because of the implementation of encryption technologies. In particular, many current social-media websites have already applied HTTPS to all connections by default, rendering inscrutable the application data being transported. Prior to such wide-spread adoption of HTTPS, it was a very simple task to view all of the data inside packets—such as the name of the website, usernames, and even passwords if they were not properly hashed. In this research, we created IP address-matching software to identify the origins of secure packets of Internet traffic. We examine the source address of each packet we have intercepted, and attempt to find a matching IP addresses in our database. Our database contains IP blocks associated with popular social-media websites. Using this method, we can easily identify from which social-media website a packet originated.

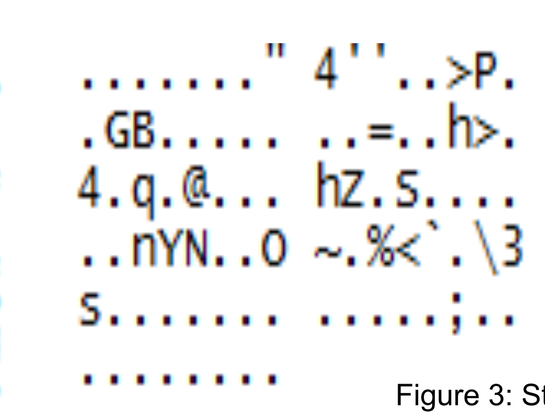
Introduction

Beginning around 2010, many websites began implementing HTTPS, a protocol to encrypt information being sent over the Internet. In March 2011, Twitter added a feature to allow users to encrypt their application data; meaning their usernames, passwords, and any activity the user does on the site can't be seen by any unintended audiences. Facebook made a similar move for security, following the lead of Google and Microsoft's Hotmail. HTTPS offers great security benefit with nearly no downside to the end users.

Packet Structure

All TLSv1.2 application data packets follow the same structure. Each header contains information about this packet. The Ethernet header shows us the hardware address of the sending device and the receiving device. The IPv4 header shows us the Internet addresses of these same devices. We can identify the port the packet was sent from and is received into from the TCP header. Lastly, secure Internet packets contain a secure socket layer (SSL). The SSL encrypts all of the data that lies beneath it, and thus makes it unreadable to humans. Before the implementation of HTTPS, the secure socket layer (SSL) would not be present, and the application data would be readable by a trainer person.





Figures 1-2: Plain text data (left) compared to encrypted data (right)

Packet Collection

To capture Internet packets, we make use of a well-known program called Wireshark. This "packet sniffing" application can capture all traffic going through a given interface, such as a computer's network adapter.

No.	Time	Source	Destination	Protocol	Length	Info
1004	21.986304000	69.196.226.13	158.65.226.179	TLSv1	1414	Application Data
1032	22.004775000	69.196.226.13	158.65.226.179	TLSv1	562	Application Data
1035	22.007116000	158.65.226.179	69.196.226.13	TLSv1	688	Application Data
1037	22.015238000	69.196.226.13	158.65.226.179	TLSv1	1414	Application Data
1043	22.015256000	69.196.226.13	158.65.226.179	TLSv1	1414	Application Data
1051	22.019934000	69.196.226.13	158.65.226.179	TLSv1	1414	Application Data
1052	22.019944000	69.196.226.13	158.65.226.179	TLSv1	1414	Application Data
1053	22.019947000	69.196.226.13	158.65.226.179	TLSv1	120	Application Data
1059	22.043387000	69.196.226.13	158.65.226.179	TLSv1	559	Application Data
1065	22.043406000	69.196.226.13	158.65.226.179	TLSv1	1414	Application Data
1069	22.045804000	69.196.226.13	158.65.226.179	TLSv1	1414	Application Data
1070	22.045809000	69.196.226.13	158.65.226.179	TLSv1	120	Application Data
1073	22.045822000	69.196.226.13	158.65.226.179	TLSv1	1414	Application Data
1078	22.045837000	69.196.226.13	158.65.226.179	TLSv1	363	Application Data
1091	22.047496000	69.196.226.13	158.65.226.179	TLSv1	1414	Application Data
1092	22.047500000	69.196.226.13	158.65.226.179	TLSv1	701	Application Data
1098	22.048152000	69.196.226.13	158.65.226.179	TLSv1	517	Application Data
1104	22.074368000	69.196.226.13	158.65.226.179	TLSv1	1414	Application Data
1105	22.074370000	69.196.226.13	158.65.226.179	TLSv1	120	Application Data

Figure 4: A packet capture shown on Wireshark

Wireshark will show us every packet that travels through an interface. We can then identify information about every packet, such as their source IP address. In this experiment, we captured thirty minutes of Internet usage on a laptop computer. The user was performing normal daily tasks, including using social media. In these thirty minutes, we collected 55,697 data packets. Of those, 42% were encrypted using HTTPS. Among all captured packets (regardless of encryption), we identified 1,560 packets that came from Facebook, and 1,994 packets that were received from Twitter.

IP Address Matching

For our purpose of classifying social-media websites, we created a program called "SocialPacketeer". Our program is written with a Java front-end, and MySQL back-end database. SocialPacketeer is designed to read packet capture (pcap) files obtained through the use of Wireshark. Our database contains IP address blocks. Large web providers, such as social-media websites, will reserve blocks of IP addresses. For example, 199.96.57.0/24 is an IP block. The first 24 bits (199.96.57.__) are reserved by an Internet service provider , and the last 8 bits are then used for the web service's servers. We only add the first 24 bits to the database, as it makes it much easier to match IP addresses. We find these IP blocks by using websites such as Hurricane Electric Internet. Services.

74.119.76.0/22	Facebook, Inc.
69.63.184.0/21	Facebook, Inc.
69.63.176.0/24	Facebook, Inc.
69.63.176.0/21	Facebook, Inc.
69.171.255.0/24	Facebook, Inc.
69.171.240.0/20	Facebook, Inc.
69.171.239.0/24	Facebook, Inc.
69.171.224.0/20	Facebook, Inc.

Figure 5: Some of the IP ranges registered to Facebook, Inc. as seen on bgp.he.net

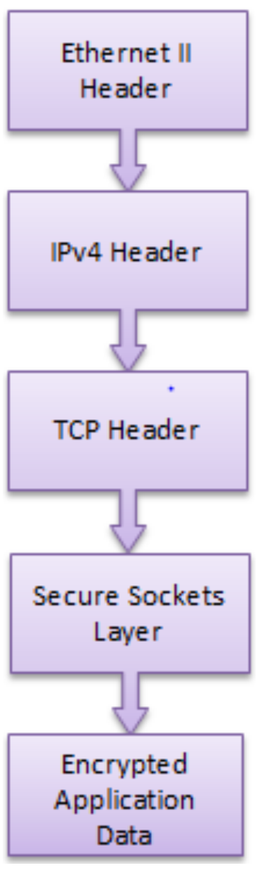


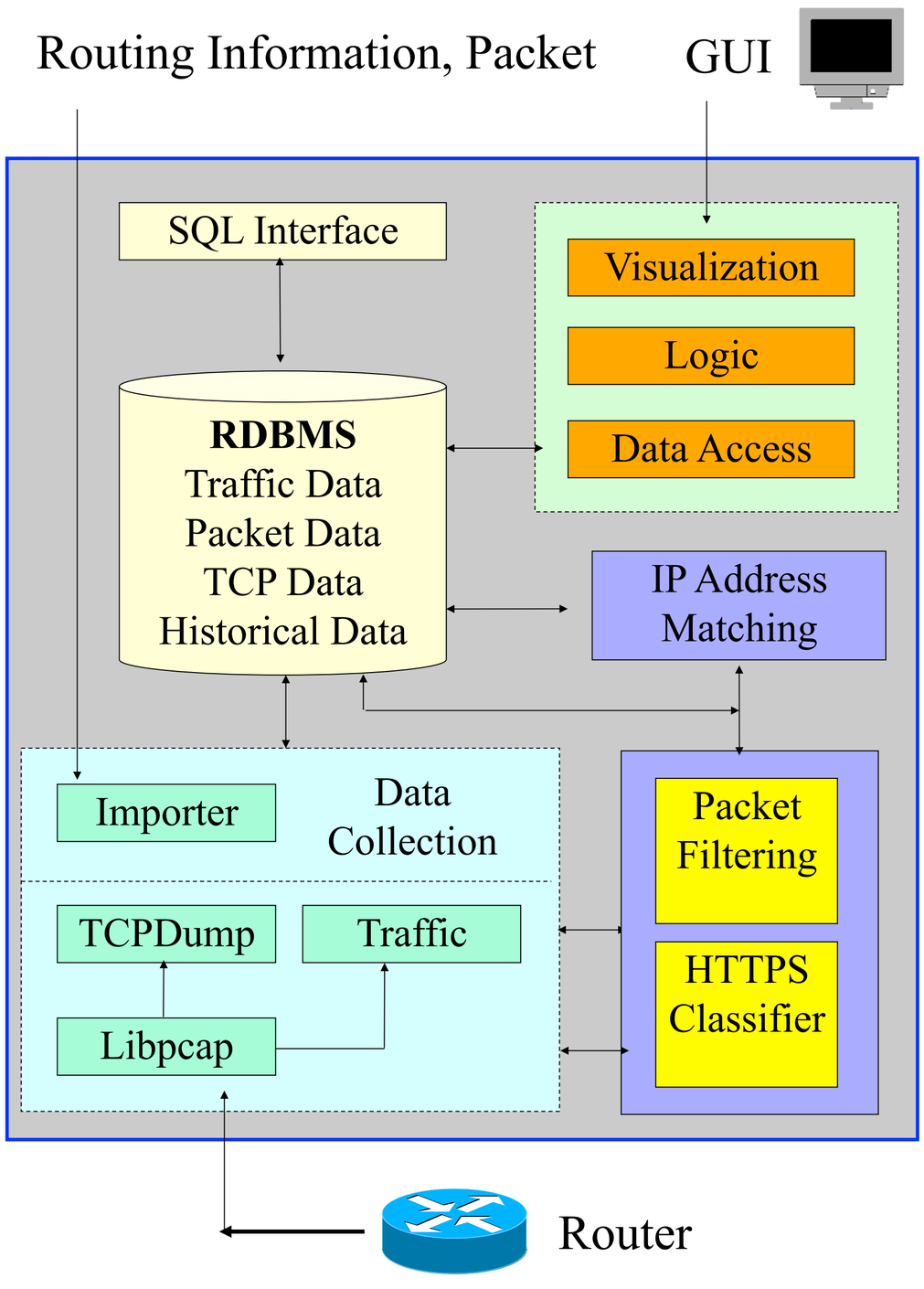
Figure 3: Structure of a TLS Packet

Social-Media Website Name	Social-Media IP Address
Facebook	31.13.71
Twitter	199.96.56
Linkedin	70.42.142

Figure 6: A diagram of what our database looks like.

As we read in a packet-capture (pcap) file, we look at each packet individually and sequentially. We look at the source IP address, and iteratively compare it to the addresses in our database. The source IP address is shortened to the same length as the current database item. For example, the current packet uses HTTPS, and has a source address of "31.13.71.33". The current database item is "31.13.71". Our program shortens the source address to "31.13.71", and then both are compared. In this example, both IP addresses are the same, and we can identify that this packet was received from Facebook.com.

System Framework



Future Work

Potential future enhancements to this project include:

- Increase the database for more websites..
- Create user a graphical interface