Homework 5: Anonymity and Privacy

This homework is due Thursday, November 29 by 6 p.m. and counts for 5% of your course grade. Late submissions will be penalized by 10% plus an additional 10% every 5 hours until received. Late work will not be accepted after 20.5 hours past the deadline. If you have a conflict due to travel, interviews, etc., please plan accordingly and turn in your homework early.

We encourage you to discuss the problems and your general approach with other students in the class. However, the answers you turn in must be your own original work, and you are bound by the Honor Code. Solutions must be submitted electronically via Gradescope in PDF format. Remember to select pages for each question in Gradescope.

Concisely answer the following questions.

1. **Tor and Using Tor**
   We advise you to visit the Tor Project website: https://www.torproject.org/ and conduct your own research to answer the following questions.

   (a) Define the following terms and how they are used in the context of Tor (1-2 sentences each):

   i. Guard
   ii. Bridge
   iii. Exit
   iv. meek
   v. bwauth
   vi. Shared Randomness

   (b) Below are listed some activities that a user could complete while using Tor. For each of the following, what is the potential leak of information?

   i. Logging into a social media website
   ii. Making a Bitcoin payment
   iii. Changing the size of the browser window
   iv. Conducting research for a term paper

   (c) You want to buy a copy of the anarchist cookbook anonymously (knowledge is power, man) using Tor. What are the steps you would take to remain anonymous in order to receive your shipment? (Hint: include at least how to handle payment, receipt of goods and communication)

2. **ZMap and Censys.** Censys was created in 2015 at the University of Michigan by the security researchers who developed ZMap, the most widely used tool for Internet-wide scanning. Over the past five years, the team has performed thousands of Internet-wide scans, consisting of trillions of probes, and has played a central role in the discovery or analysis of some of the most significant Internet-scale vulnerabilities: FREAK, Logjam, DROWN, Heartbleed, and the Mirai botnet.

   (a) ZMap is one of the tools used in the Censys pipeline. ZMap is a fast single packet network scanner designed for Internet-wide network surveys. However, other tools also exist for scanning the internet. Masscan is another TCP scanning tool developed around the same time by Robert David Graham of Errata Security. Nmap was developed in 1997 by Gordon Lyon. These three programs are used by researchers around the world to conduct measurement studies. However, there are quite a few differences in the programs and their capabilities. A paper describing a small sample of these differences is available here: https://goo.gl/T94Dbd.

      i. Compare and contrast these three tools. Your comparisons should include technical items such as (but not limited to) hit rate, speed, memory usage, and other general capabilities.

   (b) Censys provides the ability to run search queries on the scans that it has collected. For example, entering 'location.country_code: US' in the search bar provides all IPv4 hosts that are located in the United States. For more help on developing queries you should visit: https://censys.io/domain/help. (If you get stuck or have questions, you should use Piazza and the course instructors, do **not** contact the Censys team with questions.)
   For free users, Censys currently has a restriction of 10 queries per day. You may create a free account if you need to (which updates the amount of queries to 250 per month) but you do not have to. As you will be conducting large searches through data, you should think through queries carefully before running them.
   For each question below, provide the solution **and** the query that you used.

      i. What is the number of hosts that are vulnerable to Heartbleed?
      ii. What is the number of certificates that have an RSA public key exponent equivalent to 3 and are trusted?
      iii. We saw in Project 1 how trusting keys like this could be dangerous. Why do you think they would still be used and trusted? (hint: why were you able to perform the Bleichenbacher attack, and what would have stopped you?)
      iv. What is the number of self-signed certificates that are unexpired?
      v. How many servers support SMB?
      vi. Why is SMB an important protocol to scan for in regards to computer security? (Hint: SMBv1)
      vii. How many of the Alexa Top 1 Million Websites are located in the US and are untrusted by browsers?

(c) Censys (and Shodan, another internet scanning service) offer free tiers of usage. This means any user or attacker may use the service to search and download results of queries.

  i. Argue (in 1-3 sentences) that these companies are violating their ethical obligations.

  ii. Argue (in 1-3 sentences) that these companies are not violating their ethical obligations.

3. **Net Neutrality and Censorship** China has implemented a firewall to prevent Internet traffic from outside the country from entering, and to prevent traffic from inside the country from exiting, depending on the traffic, its source, and its destination. Due to historical reasons and the amount of resources required to firewall a country of over 1 billion people, this firewall is often referred to as the Great Firewall (GFW). The GFW blocks traffic in a variety of ways, and understanding exactly how it chooses which traffic to block is the subject of a significant amount of research. A link to a popular paper on the subject can be found here: https://ensa.fi/papers/Ensafi2015a.pdf. Let's consider blocking that occurs due to the GFW dropping packets in the TCP handshake between a client and a blocked server.

(a) Recall from lecture that TCP connections are initiated by the client by sending a SYN packet to a server, and then waiting for a SYN/ACK packet from the server before sending an ACK. Clients will wait `timeout` amount of time for the server to respond before giving up on the connection. If you are a client in China trying to establish a connection to a server that is blocked by the GFW, what behavior would you expect to see?

(b) The Linux kernel's implementation of TCP attempts to handle many new, simultaneous TCP connections by maintaining a backlog of half-open connections. When a SYN is received, the server responds with SYN/ACK and adds that connection to the backlog. A half-open connection will stay in the backlog until the final ACK is received, at which time it becomes a fully-established connection. The backlog is maintained in memory and typically has a fixed maximum size, with a default size of 256 pending connections. In order to handle network unreliability, servers will retransmit SYN/ACK packets to clients that haven't responded. Typically, the server will attempt to retransmit up to five times. If the GFW only blocks SYN/ACK packets from blacklisted sites, what would happen to servers who receive many concurrent connections from Chinese users?

(c) To handle performance degradation while handling many connections at once, the Linux kernel's implementation of TCP begins pruning the SYN backlog once it becomes half full. This is done by reducing the number of SYN/ACK retransmissions for each pending connection before flushing it from the backlog. What information can we gain from this side-channel?

(d) You are a researcher attempting to identify what part of the TCP handshake the GFW drops. You set up an experiment where you send 200 SYN packets from a client in China to a server that is blocked. Immediately after, you send a SYN packet to the server from a US measurement machine. If your measurement machine receives a

SYN/ACK, it does not send an ACK packet back to the server. Using your knowledge of the difference between normal TCP timeout behavior and backlog pruning, what behavior would you observe from your measurement machine if the GFW drops SYN packets to blocked servers?

(e) What would you observe if the GFW only dropped SYN/ACK packets from blocked servers?

(f) If the GFW were only able to inject spoofed packets rather than dropping or blocking real packets, how could the GFW still block connections using TCP?

(g) What is another protocol that the GFW could attack to implement Internet censorship?