You may work in groups of two people
Due date is Wednesday 11/5 11:59pm

Question 1. True/False + Short answer (25 points)

True/False. Answer the following questions with T/F and a short justification.

a. When designing a intrusion detection system, it is fundamentally harder to create a detector with a low rate of false positives than false negatives.
b. Your network logger sits on the channel between clients and a server that serves a webpage. Communication happens using TLS. If two clients request the same webpage, they will get the same set of bytes.
c. Firewalls are signature based intrusion detection systems.
d. According to our guest lecturers, for online web attack detection against their applications, web server logs are the most fruitful (as opposed to logs collected at database, firewall, etc).
e. If you switch your server to service requests over DNSSec, then you are no longer vulnerable to DNS cache poisoning attacks.

Short Answer. Answer the following questions in a few sentences.

f. Let's say Catie is sitting between Alice and Bob. Alice is sending malicious IP packets to Bob. Bob asks Catie to drop any IP packets that match a signature. Catie decides to look at IP packets one at a time and if the payload matches the malicious regex, then Catie drops the packet. Describe a strategy for Alice, so that Bob still gets the malicious payload, but Catie misses it using her regex approach.
g. Describe how defenses at Union Bank (or any financial institution worried about security) would be different if most people in the world, including attackers, used the Tor protocol (onion routing). Specifically describe how Tor would be reflected in the logs and how attack detection would be affected.
h. Traceback is an approach to stop DDoS attacks. Describe the advantages and limitations.
i. Let's say Fred implemented his spam detector using online learning that classifies emails with off-color terms as spam and associates any co-occurring words with spam. Describe how Jill could DoS Mike's communications with Fred.
j. Describe how frog boiling in cooking applies to intrusion detection systems.

Question 2. Touch Biometrics (30 points)

In this question, we explore continuous authentication methods for users who are using a smartphone. Our friends across the bay built 30 features on top of the raw data. You can find the details of the approach and the data at http://www.mariofrank.net/touchalytics/index.html

a. Implement two more features in addition to the 30 found in the database. Do they have positive information gain? That is, are the features useful?
b. Report correlation of these feature to the rest of the implemented features.
c. Train your model on a binary classifier of your choice ("true user" or "false user" classification problem) using the following 4 scenarios in which you use a feature selection method to choose top 10 features. Describe this process. Use 10-fold cross validation to compute precision and recall in the following scenarios. Try to maximize F1 score when optimizing your classifier. Report F1 and any methods you used to optimize your classifier.
    i. 10 top features,
    ii. 10 top features & your features
    iii. 30 computed features,
    iv. 30 computed features & your features
d. Qualitatively describe which family of features are most discriminating in your classifier.

Question 3. Merits of Entropy in Attack Detection/Diagnostics (30 points)

Consider the following dataset: http://web.stanford.edu/class/cs259d/hw/server-log.txt
Two attacks happened this unfortunate day, both somewhere around 8am and ~~8pm~~ noon. Please identify the exact date and time. What approach did the attackers use?
Columns for the server log are the following:

```
Start      Start                    Src    Dest Src          Dest
Date       Time     Duration  Serv  Port   Port IP           IP
```

There has been significant literature discussing how entropy can be used to detect these attacks. To do it effectively, approximation schemes are usually used. You do not have to implement these approximation techniques, but do present an analysis of whether entropy is useful and which combinations you tried, e.g. src ip, dest ip, src-port, dst-port, etc. Do any reveal anomalies when the two attacks happen?

Sources for literature:

- Lall, et all 2013. Data Streaming Algorithms for Estimating Entropy of Network Traffic.
- Clifford, Cosma, 2013. A simple sketching algorithm for entropy estimation over streaming data

Submission

Format an email to cs259d-aut1415-staff@lists.stanford.edu with

Subject: "`Submission HW2: SUNET, SUNET`"

and include an archive with:

README.txt include names, SUNETs, time spent, and thought + frustrations
p1.txt contains answers to problem 1 questions. Any common format (.pdf, .txt, .doc) is fine
p2/README.txt instructions for how to run your code: extraction and classification with 10-fold cross validation
/analysis.pdf f1 scores, precision/recall and analysis for each of the scenarios
/bin/… any code you wrote
p4/README.txt when did the attacks happen and what they were
/analysis.pdf any entropy analysis you did and conclusions
/bin/… any code you wrote