CSCI 445:
Mobile Application Security

Lecture 13

Prof. Adwait Nadkarni

Derived from Slides by William Enck, Micah Sherr, and Patrick McDaniel

Announcements

- Project Part 2 **(MILESTONE 3 and MILESTONE 4)** assigned.
  - Due Dates:
    - April 11th, **Analysis Plan (Milestone 3)**
    - May 2nd, **Project Report (Milestone 4)**

How do we study apps?

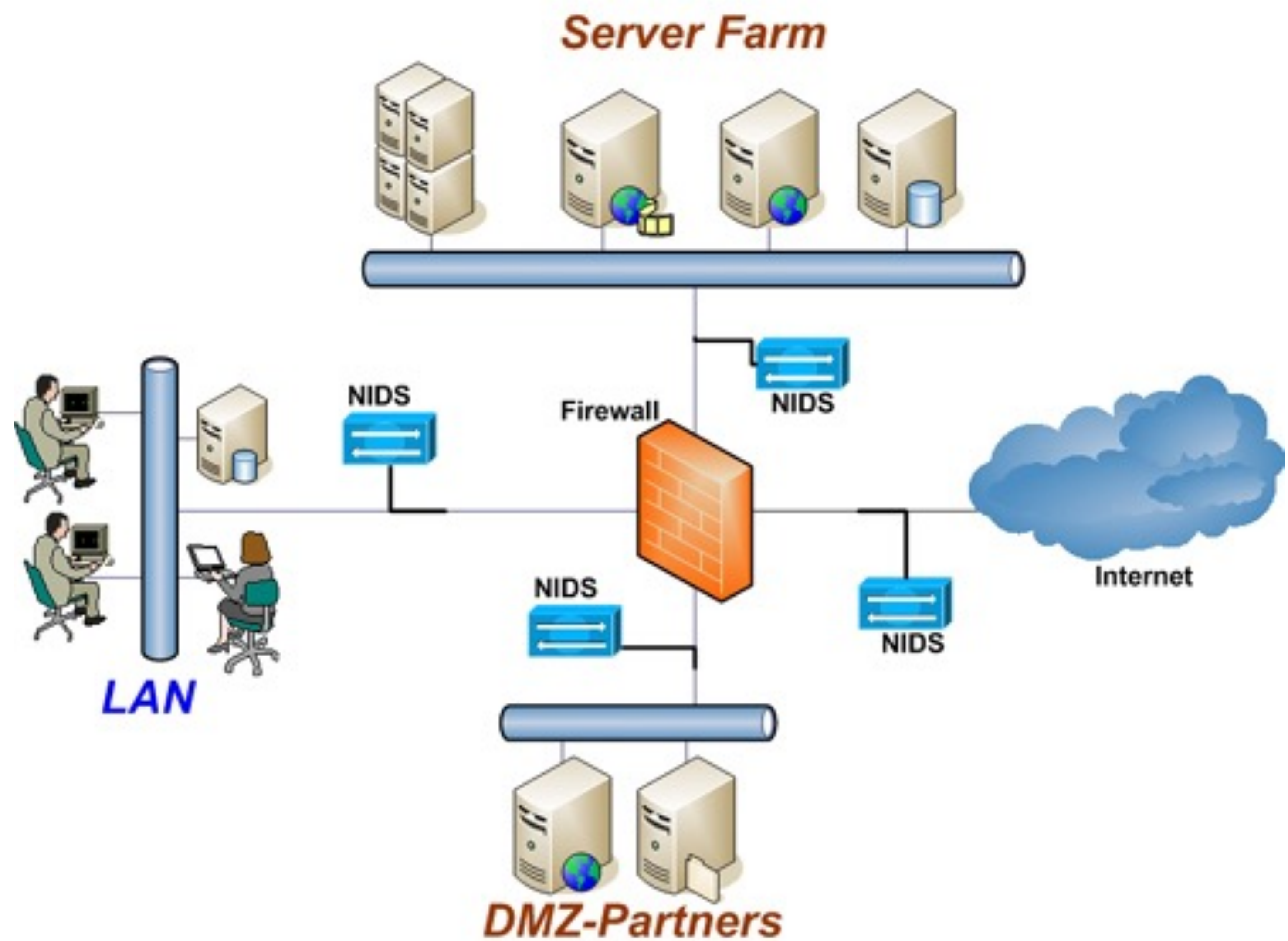
- *Generally, two ways to do this:*
- *Static analysis tells you want can potentially happen.*
  - *Getting source code: ded, dex2jar, androguard*
  - *Extend existing analysis tools (e.g., Fortify)*
  - *Frameworks: Flowdroid, Amandroid, DroidSafe*
- *Dynamic analysis tells you what actually happens given a specific runtime environment*
  - *TaintDroid, DroidScope*
  - *Derivative environments: Droidbox, andrubis, MarvinSafe*
- *Note: dynamic analysis is hard to automate*

Evaluating Analyses

Example: Intrusion Detection Systems

- **Authorized eavesdropper** that listens in on network traffic
- Makes determination whether **traffic contains malware**
  - usually compares payload to virus/worm signatures
  - usually looks at only incoming traffic
- If malware is detected, IDS somehow raises an alert
- Intrusion detection is a **classification problem**

Example Setup



Detection via Signatures

- Signature checking
  - does packet match some signature
    - suspicious headers
    - suspicious payload (e.g., shellcode)
  - great at matching known signatures
  - Low *false positive* rate: **Q: WHY?**
  - Problem: not so great for zero-day attacks -- **Q: WHY?**

Anomaly Detection

- *Learn what ”normal” looks like.*
- Frequently uses ML techniques to identify malware
- Underlying assumption: malware will look different from non-malware
- **Supervised learning**
  - IDS requires learning phase in which operator provides pre-classified ***training data*** to learn patterns
  - {good, 80, “GET”, “/”, “Firefox”}
  - {bad, 80, “POST”, “/php-shell.php?cmd=’rm -rf /’”, “Evil Browser”}
  - ML technique builds model for classifying never-before-seen packets
  - Problem: *False Learning*

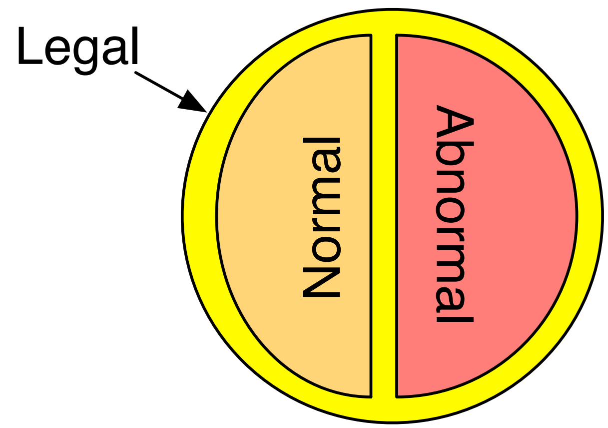
False Alarms

Analysis with *only* False Alarms

- Analysis raises *alarms* (e.g., found data leak, I think this binary is malicious): Require manual effort to validate/confirm
- A useful analysis technique must raise few false alarms, i.e., *truly* reduce manual effort, while also being effective (i.e., few false negatives).

Confusion Matrix

- What constitutes an intrusion/anomaly is really just a matter of definition
  - A system can exhibit all sorts of behavior
- Quality determined by the consistency with a given definition
  - Context-sensitive (i.e., what is “positive/true”?)



		Detection Result: T	Detection Result: F
Reality	T	True Positive	False Negative
Reality	F	False Positive	True Negative

Metrics

- **True positives** (TP): number of correct classifications of malware
- **True negatives** (TN): number of correct classifications of non-malware
- **False positives** (FP): number of incorrect classifications of non-malware as malware
- **False negatives** (FN): number of incorrect classifications of malware as non-malware

Metrics

(from perspective of detector)

		Detection Result: T	Detection Result: F
Reality	T	True Positive	False Negative
	F	False Positive	True Negative

- **False positive rate**:

$$FPR = \frac{FP}{FP + TN} = \frac{\#\text{ benign marked as malicious}}{\text{total benign}}$$

- **True negative rate**:

$$TNR = 1 - FPR = \frac{TN}{FP + TN} = \frac{\#\text{ benign unmarked}}{\text{total benign}}$$

- **False negative rate:**

$$FNR = \frac{FN}{FN + TP} = \frac{\#\text{ malicious not marked}}{\text{total malicious}}$$

- **True positive rate:**

$$TPR = 1 - FNR = \frac{TP}{FN + TP} = \frac{\#\text{ malicious correctly marked}}{\text{total malicious}}$$

Base Rate Fallacy

- Occurs when we assess P(X|Y) without considering prior probability of X and the total probability of Y
- Example:
  - Intrusion detection system is 99% accurate (given known samples)
    - 1% false positive rate (benign marked as malicious 1% of the time)
    - 1% false negative rate (malicious marked as benign 1% of the time)
  - *Base rate* of malware is 1 packet in a 10,000
  - Packet X is marked by the NIDS as malware. *What is the probability that packet X actually is malware?*
    - Let's call this the "true alarm rate," because it is the rate at which the raised alarm is actually true.

Bayes' Rule

- Pr(*x*) function, probability of event *x*
  - Pr(sunny) = .8 (80% of sunny day)
- Pr(x|y), probability of x given y
  - Conditional probability
  - Pr(cavity|toothache) = .6
    - 60% chance of cavity given you have a toothache
- Bayes' Rule (of conditional probability)

$$Pr(B|A) = \frac{Pr(A|B) \cdot Pr(B)}{Pr(A)}$$

- Assume: Pr(cavity) = .5, Pr(toothache) = .1
- What is Pr(toothache|cavity)?

Base Rate Fallacy

- How do we find the true alarm rate? [i.e., Pr(IsMalware|MarkedAsMalware)]

$$Pr(IsMalware|MarkedAsMalware) = \frac{Pr(MarkedAsMalware|IsMalware) \cdot Pr(IsMalware)}{Pr(MarkedAsMalware)}$$

- We know:
  - 1% false positive rate (benign marked as malicious 1% of the time); TNR= 99%
  - 1% false negative rate (malicious marked as benign 1% of the time); TPR= 99%
  - *Base rate* of malware is 1 packet in 10,000
- What is?
  - Pr(MarkedAsMalware|IsMalware) = $\boxed{\text{TPR = 0.99}}$
  - Pr(IsMalware) = $\boxed{\text{Base rate = 0.0001}}$
  - Pr(MarkedAsMalware) = ?

$$Pr(MarkedAsMalware|IsMalware)$$
$$= \frac{\#\ \text{malicious correctly marked}}{\text{total malicious}}$$
$$= \frac{TP}{FN + TP} = TPR$$

Base Rate Fallacy

- How do we find Pr(MarkedAsMalware)?

= Pr(MarkedAsMalware|IsMalware)Pr(IsMalware) + Pr(MarkedAsMalware|IsNotMalware)Pr(IsNotMalware)

- So what is?
  - Pr(IsMalware) = base rate = 0.0001
  - Pr(IsNotMalware) = 1 – Pr(IsMalware) = 0.9999
  - Pr(MarkedAsMalware|IsMalware) = TPR = 0.99
  - Pr(MarkedAsMalware|IsNotMalware) = FPR = 0.01

Pr(A|!B) = 1 – Pr(!A|!B)
Pr(A|B) = 1 – Pr(!A|B)

$$\begin{aligned} &Pr(MarkedAsMalware|IsNotMalware) \\ &= \frac{\#\text{ benign marked as malicious}}{\text{total benign}} \\ &= \frac{FP}{FP+TN} = FPR \end{aligned}$$

- So Pr(MarkedAsMalware) = 0.99 * 0.0001 + 0.01 * 0.9999 ~= 0.01

Base Rate Fallacy

- How do we find the true alarm rate? [i.e., Pr(IsMalware|MarkedAsMalware)]

$$Pr(IsMalware|MarkedAsMalware) = \frac{Pr(MarkedAsMalware|IsMalware) \cdot Pr(IsMalware)}{Pr(MarkedAsMalware)}$$

$$= \frac{0.99 \cdot 0.0001}{0.01} = 0.0099$$

- Therefore *only about 1% of alarms are actually malware!*
  - What does this mean for security analysts?

Base Rate Fallacy

(summary)

- Let Pr(M) be the probability that a packet is actually malware (the base rate)
- Let Pr(A) be the probability that that the IDS raises an alarm (unknown)
- Assume we also know for the IDS
  - Pr(A|M) = TPR = 1 - FNR
  - Pr(A|!M) = FPR
- Then the true alarm rate is

$$Pr(M|A) = \frac{Pr(A|M) \cdot Pr(M)}{Pr(A|M) \cdot Pr(M) + Pr(A|!M) \cdot Pr(!M)}$$

- **Note the strong influence of Pr(M)**

Base-rate Fallacy in the real world

Health Nerd
@GidMK

Follow

So, according to this, the false positive rate for the Apple Watch in detecting atrial fibrillation is 0.04% (99.6% correct)

This means that, on average, Apple Watches will be wrong more than 80% of the time

Sound counterintuitive? This is the issue with population screening

> **STAT** @statnews
> Apple submitted two studies to FDA to get clearance for the new Apple Watch EKG app. Here's the data. buff.ly/2QuhGmG

5:28 AM - 14 Sep 2018

Where is Anomaly Detection Useful?

System	Intrusion Density P(Malware)	Detector Alarm Pr(Alarm)	Detector Accuracy Pr(Alarm|Malware)	True Alarm P(Malware|Alarm)
A	0.1		0.65	
B	0.001		0.99	
C	0.1		0.99	
D	0.00001		0.99999	

$$\Pr(B|A) = \frac{\Pr(A|B)\Pr(B)}{\Pr(A)}$$

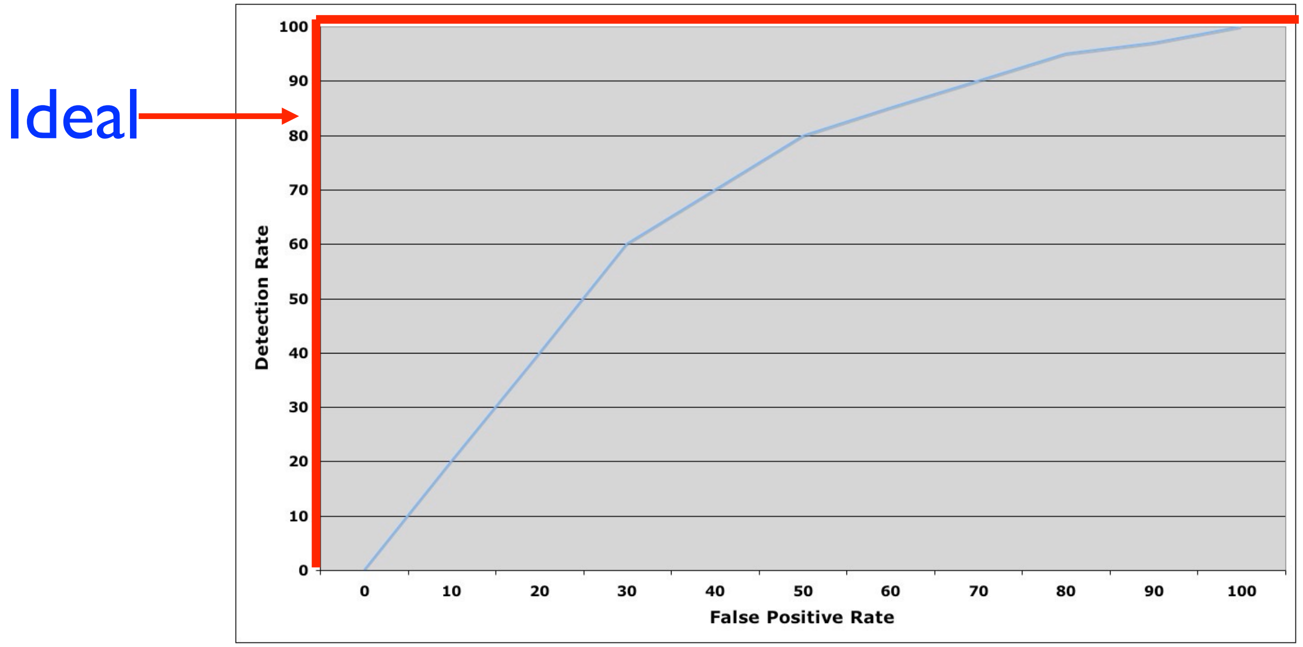
Where is Anomaly Detection Useful?

System	Intrusion Density P(M)	Detector Alarm Pr(A)	Detector Accuracy Pr(A|M)	True Alarm P(M|A)
A	0.1	0.38	0.65	0.171
B	0.001	0.01098	0.99	0.090164
C	0.1	0.108	0.99	0.911667
D	0.00001	0.00002	0.99999	0.5

$$\Pr(B|A) = \frac{\Pr(A|B)\Pr(B)}{\Pr(A)}$$

The ROC curve

- Receiver Operating Characteristic (ROC)
  - Curve that shows that detection/false positive ratio (for a binary classifier system as its discrimination threshold is varied)



- Axelsson talks about the real problem with some authority and shows how this is not unique to CS
  - Medical, criminology (think super-bowl), financial

Example ROC Curve

- You are told to design an intrusion detection algorithm that identifies vulnerabilities by solely looking at transaction length, i.e., the algorithm uses a packet length threshold T that determines when a packet is marked as an attack (i.e**., less than or equal to** length T). More formally, the algorithm is defined:

$$\mathrm{D(k,T)} \rightarrow [0, 1]$$

- where k is the packet length of a suspect packet in bytes, T is the length threshold, and (0,1) indicate that packet should or should not be marked as an attack, respectively. You are given the following data to use to design the algorithm.

  attack packet lengths: 1, 1, 2, 3, 5, 8

  non-attack packet lengths: 2, 2, 4, 6, 6, 7, 8, 9

- Draw the ROC curve.

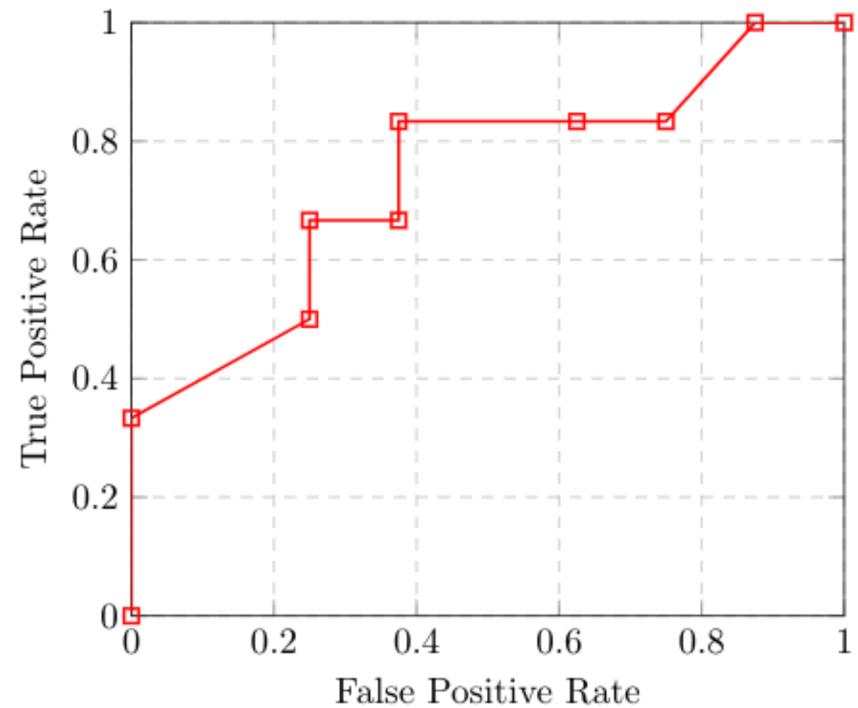
Solution

attack packet lengths: 1, 1, 2, 3, 5, 8

non-attack packet lengths: 2, 2, 4, 6, 6, 7, 8, 9

$$TP\% = TPR = \frac{TP}{TP + FN}$$

$$FP\% = FPR = \frac{FP}{FP + TN}$$



T	0	1	2	3	4	5	6	7	8	9
TP	0	2	3	4	4	5	5	5	6	6
TP%	0.00	33.33	50.00	66.67	66.67	83.33	83.33	83.33	100.00	100.00
FP	0	0	2	2	3	3	5	6	7	8
FP%	0.00	0.00	25.00	25.00	37.50	37.50	62.50	75.00	87.50	100.00

Security Research Methods 1

Reading papers …

- What is the purpose of reading papers?
- How do you read papers?

Understanding what you read

- Things you should be getting out of a paper
  - What is the central idea proposed/explored in the paper?
    - Abstract
    - Introduction
    - Conclusions

    *These are the best areas to find an overview of the **contribution***

  - How does this work fit into others in the area?
    - Related work - often a separate section, sometimes not, every paper should detail the relevant literature. Papers that do not do this or do a superficial job are almost sure to be bad ones.
    - An informed reader should be able to read the related work and understand the basic approaches in the area, and how they differ from the present work.

Understanding what you read (cont.)

- What scientific devices are the authors using to communicate their point?
- Methodology - this is how they evaluate their solution.
  - Theoretical papers typically validate a model using mathematical arguments (e.g., proofs)
  - Experimental papers evaluate results based on test apparatus (e.g., measurements, data mining, synthetic workload simulation, trace-based simulation).
  - Empirical research evaluates by measurement.
  - Some papers have no evaluation at all, but argue the merits of the solution in prose (e.g., design papers)

Understanding what you read (cont.)

- What did they find?
  - Results - statement of new scientific discovery.
    - Typically some abbreviated form of the results will be present in the abstract, introduction, and/or conclusions.
    - Note: *just because a result was accepted into a conference or journal does necessarily not mean that it is true. Always be circumspect.*
- What should you remember about this paper?
  - Take away - what general lesson or fact should you take away from the paper.
  - Note that really good papers will have take-aways that are more general than the paper topic.

The best papers are the ones that teach you something

Tips for reading a paper

- Everyone has a different way of reading a paper.
- Here are some guidelines I use:
  - Always have a copy to mark-up. Your margin notes will serve as invaluable sign-posts when you come back to the paper (e.g., "here is the experimental setup" or "main result described here")
    - Digitally: Zotero, Mendeley
  - After reading, write a summary of the paper containing answers to the questions in the preceding slides. If you can't answer (at least at a high level) these questions without referring to the paper, it may be worth scanning again.
- Over the semester, try different strategies for reading papers and see which one is the most effective for you.

Reading a systems security paper

- What is the security model?
  - Who are the participants and adversaries
  - What are the assumptions of trust (trust model)
  - What are the relevant risks/threats
- What are the constraints?
  - What are the practical limitations of the environment
  - To what degree are the participants available
- What is the solution?
  - How are the threats reasonably addressed
  - How do they evaluate the solution
- What is the take away?
  - key idea/design, e.g., generalization (not solely engineering)
- Hint: I will ask these questions when evaluating course project.

The End