Methods

situations, crafted mistakes, and malicious activity.

The software will use the training dataset, to make predictions against similar datasets to verify accuracy. It will then be tested against a human actor to test predictions in a high-fidelity simulation. A testing dataset of 100 commands will be given to the model for batch prediction and will be evaluated for statistical accuracy. The method utilizes the Amazon Machine Learning software. It will be trained on a dataset comprised of normal user

Introduction and Motivation

Insider threats are on the rise. Current commercial software can monitor, log, and prevent access to designated files and directories. However, it remains difficult to predict and prevent unauthorized insider usage. Due to the gaps in research in the area, the focus of this study is to more accurately predict insider threats in a server environment

Connection Setup

Set up Amazon Web Services connection
client = boto3.client('machinelearning')









(0) Proa ear L

Prescott Valley School of Informatics, Computing, Arizona University Northern

and Cyber Systems

terminal? Why the linux

Linux has been chosen because of the ubiquity of the operating systems based on it are connected to the internet in servers around the globe. It is also transferable to all POSIX-compatible terminal environments that are in use today.

Data Sources

Mistake

of

Warn

http://joshuabowen.info

jdb567@nau.edu

Bowen

Joshua

command may mistake, please try agian.') do not execute command
75 and ev >= 0.25:
I think your command may
75 ontain a mistake, pl # if mistake, do r
elif ev <= 0.75 ar
print('\nI thi
be or cc</pre>

Activit Malicious Prevent

session determined malicious, stop s
ev < 0.25 and ev >= 0:
print('\nIntruder detected!')
sys.exit(1)

Use Normal Allow

command if good, allow co
f ev > 0.75:
 os.system(com) iή

There are three main sources of data used for the machine learning model. The first is an un-edited history file from the test system. Second is a list of commands assembled during this research. The third is the UNIX User Dataset from Purdue University. Simple programs were developed to remove unnecessary lines from the two gathered datasets.



User Get

The baat premal

_input(): def

e batch predictions have shown that the model is 100% accurate t predicting normal use and has a %30 rate of false negatives for malicious and mistake commands. The results of the high-fidelity simulation indicate that the software is partially effective at capturing all input. The participants were able to escape it and after that became untraceable to my system. This indicates that further work is needed in this area. Overall the process has indicated that a prototype like this can capture near 80% of the commands and suggests that further work on this model can produce far more reliable results.

Results

terminal from working dir le username, hostname, and work
le = os.environ.get('USER')
le = os.environ.get('HOSTNAME')
os.environ.get('PWD') username hostname Update

str(hostname).rstrip() + # # 0 + + username).rstrip() + + str(wdir).rstrip() prompt mand str(COM 7 # Builc prompt

Get user input
id = input(prompt) cmd

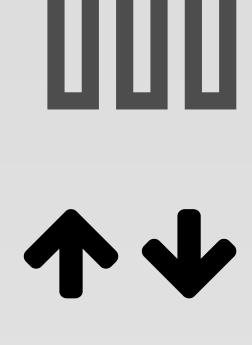
Convert to and understndable
cmds = str(cmd)

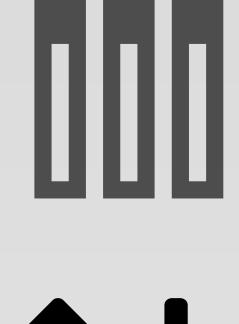
string

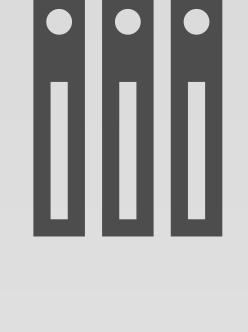
cmds return

(Evaluatio Machine Send

Conclusions







٩ get def

request Send respon:

Cor

The results of the test indicate the model isn't sufficiently accurate in detecting intrusion even when given special weighting. This is likely due to the lack of data in the malicious and mistake categories. Future research in the area needs more open data. One solution is to ask organizations to publish the cleaned logs after an attack.

retur

This study also indicated that further improvements are needed to the program to accurately capture information from a live terminal. While the program created a fake prompt it was severely limited in what it could capture and that skewed the data from the simulation.

Future research is needed in order to determine how to best approach the problem of practical proactive intruder detection.

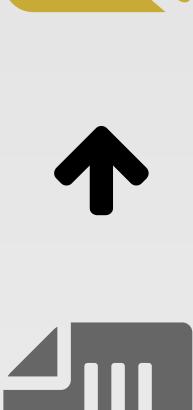
As other researchers have indicated, larger datasets comprising of more points such as timestamps and keyboard information would help better track user behavior in an attempt to establish recognizable patterns by neural networks. This and other research indicate that more data is needed.

atio Evalu and Command 0

%H:%M:%S...)

def









- James A. Subach, Ph. Acknowledgements Faculty Advisor Academic Credit A

credit Advisor - Nancy L. Jensen p - Ian Harvey on Stack Overflow Data Sources - Purdue University - Felix Breuer -Riddle [Ethical] Hackers Club Amazon Web Services The NAU Cyber Security Team Template Poster Embry

and Material Code, Refer

http://github.com/Skraelingjar/cyberml