Use of Machine Learning to predict and avert attacks on evolutionary Software-Defined-Networks (SDNs)

Praful Ravi
MSc Cybersecurity
Dublin Business School
10570329@mydbs.ie

INTRODUCTION

The significance of security in the digital domain has ever-growing importance due to the fast digitization of personal records. Any intrusion into digital privacy, known by different names, creates widespread destruction of property, wealth, and the right to personal space. Network security is a general term given to proactive prevention of unauthorized and attacks. Such a process occurs at different levels of the digital domain. Putting down a perspective of the threat based on recorded attacks, some of the worst-hit were, a) CAM4 breaching nearly ten billion records, b) Yahoo 2017 breach leaking three billion accounts, c) Aadhaar data breach of 1.1 billion Indian citizens, d) social media platforms - LinkedIn 2018 leak of 700 million users, Facebook data leak of nearly 600 million accounts in 2019, and Sociallarks leak of 200 million records. The leaks artefacts such as personal photographs and videos and financial documents create catastrophic impacts on affected individuals and corporations. These attacks happen on a combinatorial structure starting at network and switching levels, and perimeter security becomes significant. **(Bekker, 2021) (Tunggal, 2021) (Sobers and Varonis, Inc, 2021) (Bland et al., 2020)**

IMPACT OF RESEARCH – APPLICATION IN 5G SYSTEMS

The evidence of skyrocketing data traffic puts the mobile data, in 2020, at 30.6 exabytes per month, nearly 8.5 times that of 2015 figures. **(Cisco, 2020)**. The scope of telecommunication beyond the fourth generation is not a revamp in the infrastructure, like in the earlier migrations. Instead, the infrastructure bases on the catenation (of legacy systems) and augmentation (with evolutionary hardware) and controlling automation software. **(Hill, 2015)** suggested that the top research directions in the fifth generation telecommunication are towards end-to-end QoE/QoS, data mining on-the-go, contextual awareness, and NFV and SDN-assisted Cloud/distributed cloud. Countries and consortiums invest in 5G anticipating the logarithmic rise in bandwidth demand. The interests are towards global cooperation and interoperability (open standardization) through the alliances of ITU, 3GPP, and 5G PPP (Public-Private-Partnership) and the international community. It is a known fact that the future of evolved industries and digitization (referred to as industry 4.0) also rests on 5G success. **(European Commission - 5G Policy International, 2020)**. **(European Commission - Industries, 2020)**. 5G opens up cutting-edge technologies to enhance safety and sustainability, faster deployment of emergency services and connecting remote expertise with patients, improving data-sharing features to connect machines like vehicles and sensors to avoid collision and warn natural disasters early. **(Ericsson, 2020) (Carlson, 2020) (ITU-R, 2015)**. Three different 5G services are proposed based on usage scenarios, grouped as enhanced mobile broadband (eMBB), massive machine-type communications (MMTC), and ultra-reliable and low-latency communications (URLLC). **(ITU-R, 2015)**. The 3rd Generation Partnership Project (3GPP) suggests five service categories (eMBB, critical communications (Cris), massive IoT (m-IoT), network operation (NEO) and eV2X) and 97 use-cases for 5G mobile services. **(Third Generation Partnership Project - 3GPP, 2015)**. Device-to-Device (D2D) connectivity is one of the groundbreaking 5G cellular network innovations that allows mobile devices to communicate directly without the Base Station intervention (BS). One of its key goals is to provide seamless mobility that meets QoS criteria not supported by legacy cellular and LTE networks. D2D enables 'smart homes', networked smart-products, automation, healthcare, and entertainment. Development of ML models should ensure optimized network resources, seamless integration, strict security and privacy at both service providers and customer ends. **(Ouali et al., 2020) (Wang et al., 2018)**. There is a rising popularity of using ML techniques to solve overloaded links with varying link capacities, sporadic network traffic, and topology optimization. Innovative real-time and real-world issues in 5G architecture need ML/AI techniques with SDN/NFV for networks and algorithm scheduling, addressing route information failure in core-network by NFV (network function virtualization), and network perimeter security. **(ITU News, 2020)**.

Machine learning (ML) has substantially increased the capacities in cybersecurity and cyberwarfare. Hackers can use ML to contaminate networks gaining control over the training data in software-defined networks, manipulate the trained classifier, and use the decision-making metadata to exploit the known vulnerabilities. Reinforcement Learning (RL) can act as an autonomous defense in computer networks (especially adversarial machine learning attacks) to ensure essential resources and define optimal actions to secure sensitive servers by isolating the vulnerable nodes. **(Han et al., 2018) (Novaes et al., 2021)**. Challenges in using counter ML techniques in intrusion detection approaches are the sparse databases to train models. The virtualization of the networks only increases vulnerability to newer threats and exploits on the network front. DDoSNet, a deep learning model based on the Recurrent Neural Network (RNN)-Autoencoder with the best assessment measures (Recall, Precision, Accuracy and F1-scores) relative to classical ML techniques, can detect Distributed Denial of Service attacks against a Software Defined Network (SDN). **(Said Elsayed et al., 2020)**. An integration of extension libraries and work-suites could implement the SDN DDoS Intrusion Detection framework in a 5G-like scenario. Results have shown that the SDN DDoS security controller program effectively mitigates SYN flood attacks, man-in-the-middle vulnerability, possible single-point attacks, communication channels problems and proves SDN as a security component. **(Alghamdi and Braun, 2020) (Forland et al., 2019)**. To improve network performance and QoS, self-protection, self-healing, and self-optimization, dynamically, through virtualisation might require virtual machines, metric collection, and training process using datasets containing measurements from previously known operations. **(Ahrens et al., 2018)**. Networking models such as the PetriNet help simulate cyberdefense amidst threat players, combining ML models both from threat and defender levels. CAPEC database with documented cyberattacks and invoked with PetriNet provide a competitive edge. **(Bland et al., 2020)**

PROBLEM STATEMENT

The Internet has become indispensable to humans in the new world, affecting every walk of life. The threat of exposure of data belonging to humans has become more prominent. An example would be the evolutionary sensory networks that have increased importance in numerous field applications ranging from agriculture to defense, undersea to outer space, industry to education, and many more. These systems help plan and intervene in critical systems, such as weather forecast, disaster alert and management, agriculture and ecological eventualities, etc. Attacks on such networks through vulnerable nodes defeat the purpose achieved through the use of these technological advancements. **(Yu, Dong and Kang, 2021) (Li et al., 2019)**. Computer networks, the backbone of data, are heterogeneous and complex, requiring software-defined networking (SDN) paradigm for management and evolution. SDN acquires distributed management on the data forwarding plane into centralizing network management enabling global topology manipulations. Unfortunately, with SDN attackers exploiting vulnerabilities had much more access and control on networks. The use of machine learning to overcome defensive mechanisms of network systems requires the defenders to step up the ante, using better learning algorithms. **(Novaes et al., 2021)**. The problem-set is so diverse, complex and interconnected that finding solutions requires above-human intelligence and computing power. Software automation with artificial intelligence techniques, algorithms and cloud computing seems to provide better solutions in the current times, even with the problem-set growing with each advancement.

RESEARCH QUESTIONS

1) How can ML manage threat management in an evolutionary network?
2) Can ML applications combine with SDN to form perimeter defense against cyber-attacks?

RESEARCH OBJECTIVE

The author seeks to study the scope of network threat management in software-defined-networks (SDN), using machine language (ML) models countering evolving threats in such networks

HYPOTHESIS

1) $H_{0\text{-}1}$ –ML (or other AI) will improve learning and devise solutions to problems encountered in evolving networks, as it surpasses human-level intelligence by many folds.
2) $H_{0\text{-}2}$ – Standard/open datasets assist in a possible benchmarking of machine learning algorithms and techniques and find the best fit for an application or the problem solution.

METHODOLOGY

Use of data sets and ML for parameter detection – The analysis of datasets consumes more resources than the actual selection of model and algorithm. Data engineering and filtering determine the degree of model fitting and generalization of test behavior for the particular experiment confines. The reactive approach to network attacks such as botnets relies on postmortem data in the network. Data analysis identifies the attack strategy correlational to the data packets specific to the attack. The algorithm selection can become a case-dependent ensemble of proven techniques, bettering individual model performances. **(Bijalwan, 2020)**. SDN can classify network traffic from client devices, and traffic classification based on ML on SDN network produces high accuracy in training, validations, and real-

time testing **(Raikar et al., 2020)**. The research dataset collates efforts from various contributors (universities, industry partners, independent researchers, and standardization institutes), and live traffic testing remains limited in this academic research. **(GITHUB -ITU AI Challenge, 2020) (University of New Brunswick, 2020)**.

RESOURCES AND TIME ESTIMATIONS

Using Python programming language, the research posits to model AI-based algorithms to evaluate network threat and security in an evolutionary software-defined network. As the model running poses a challenge and regular computation systems fail to emulate the complexity and due to the process intensity, cloud systems appear the only choice. The popular options of cloud platforms on offer are Kaggle Notebooks, Google Colab (Pro), and AWS, and the research intends to study and choose the best among them. **(Kim, 2020) (AWS, 2020) (Kaggle - Costs, 2019)**. Simulating SDN, VNF requires a software platform, like OvS (free license), and a minimum computational requirement level. **(Ovs- System Requirement, 2016),** The research needs to use datasets for training, validation, and testing designed models to predict the tested parameters and sourced from public data repositories. **(GITHUB -ITU AI Challenge, 2020) (University of New Brunswick, 2020)**.

1) Google Colab / Colab Pro – The cost of the lease is $0-10 per month, depending on the platform subscription. The free version runs only for 12 hours and resets the instance after time expiration. An Estimation of running the project in the platform, including setting up and analysis, is about 80 hours.
2) AWS – Amazon offers cloud products under AWS, and machine language is one of the many services under the brand. Like other cloud offering companies, AWS is also lease-based, with different billing schemes, charging as Compute and Monthly prediction fees ($0.42/hr + $0.1/1000 predictions), having no usage limitations (pay-per-use). An estimation of running the project in the platform, including setting up and analysis, is about 80-120 hours.
3) Kaggle – is an open-source cloud platform dedicated to machine learning and competitive skill exhibitions. The platform has specialised GPU and TPU cores inbuilt to support process-intensive machine learning works. The limitation is that Kaggle allows only 30 hours of running per week. An estimation of running the project in the platform, including setting up and analysis, is about 80 hours.
4) OvS – is a simulation platform to model the virtualization of physical switching and routing stack, supporting most of the commonly used protocols in computer networks. It will suffice the study's requirement of implementing a network environment to invoke machine learning model design. It is free to use under Apache 2.0 license and nominal hardware configuration, Linux OS or requiring Mingw package for Windows, Open Flow Controller library **libssl**, **libcap-ng**, and Python 3.4 or later. The estimation on the segment of the project in the platform, including setting up and analysis, is about 80 – 100 hours.

ASSUMPTIONS MADE IN THE RESEARCH

1) The availability of dataset and formability is a risk assumed necessary for the research, but not a high one as similar sources provide a base for network optimization and threat detection investigations.
2) The finalization of Cloud platforms for running ML models (resource intensive) can happen only after fixing parameters, dataset, and model(s), as the cloud resource involves purchasing services.
3) The research scope necessitates open datasets for the training, validation, and testing of designed models to predict the tested parameters. The correctness of the dataset shall decide the outcome of the research, and the use of the datasets might require conformance to public licenses of data use and reuse.
4) The use of modelling, simulation, validation and testing platforms like OvS, Kaggle Notebooks, Google Colab bind under respective licenses and terms of free restricted/non-profit operation, particular to each case.

REFERENCES

1. Ahrens, J., Strufe, M., Ahrens, L. and Schotten, H. (2018). *An AI-driven Malfunction Detection Concept for NFV Instances in 5G*. [online] Available at: https://arxiv.org/pdf/1804.05796.pdf.
2. Alghamdi, K. and Braun, R. (2020). Software Defined Network (SDN) and OpenFlow Protocol in 5G Network. *Communications and network*, 12(01), pp.28–40.
3. AWS (2020). Services and Costs. [online] Amazon Web Services, Inc. Available at: https://aws.amazon.com/getting-started/projects/build-machine-learning-model/services-costs/.
4. Bekker, E. (2021). 2021 Data Breaches - The Worst Breaches of the Year | IdentityForce®. *We Aren't Just Protecting You From Identity Theft. We Protect Who You Are.* Available at: https://www.identityforce.com/blog/2021-data-breaches [Accessed Aug. 2021].
5. Bijalwan, A. (2020). Botnet Forensic Analysis Using Machine Learning. *Security and Communication Networks*, [online] 2020, pp.1–9. Available at: https://doi.org/10.1155/2020/9302318 [Accessed 2020].

6. Bland, J.A., Petty, M.D., Whitaker, T.S., Maxwell, K.P. and Cantrell, W.A. (2020). Machine Learning Cyberattack and Defense Strategies. *Computers & Security*, [online] 92(101738), pp.1, 2. Available at: https://www.sciencedirect.com/science/article/pii/S0167404818309799 [Accessed Aug. 2021].
7. Carlson, E.K. (2020). What Will 5G Bring? *Engineering*. [online] Available at: https://www.sciencedirect.com/science/article/pii/S2095809920301351?via%3Dihub.
8. Cisco (2020). *Cisco Annual Internet Report - Cisco Annual Internet Report (2018–2023) White Paper*. [online] Cisco. Available at: https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html.
9. Ericsson (2020). *What is 5G? Do you want to know more about 5G?* [online] Ericsson.com. Available at: https://www.ericsson.com/en/5g/what-is-5g?gclid=Cj0KCQiAifz-BRDjARIsAEElyGKUtsGcr57lkkeJixaAPi5EHQ3gXLvCABwgP9JSo3cgLM89r2KsNKAaAj23EALw_wcB&gclsrc=aw.ds.
10. European Commission - 5G Policy International (2020). *International Cooperation on 5G*. [online] Shaping Europe's digital future - European Commission. Available at: https://ec.europa.eu/digital-single-market/en/5G-international-cooperation.
11. European Commission - Industries (2020). *Digitising European Industry*. *Shaping Europe's digital future - European Commission*. Available at: https://ec.europa.eu/digital-single-market/en/digitising-european-industry.
12. Forland, M.K., Kralevska, K., Garau, M. and Gligoroski, D. (2019). *Preventing DDoS with SDN in 5G*. [online] IEEE Xplore. Available at: https://ieeexplore.ieee.org/document/9024497.
13. GITHUB -ITU AI Challenge (2020). ITU-AI-ML-in-5G-Challenge. [online] GitHub. Available at: https://github.com/ITU-AI-ML-in-5G-challenge.
14. Han, Y., I.P. Rubinstein, B., Abraham, T., Alpcan, T., De Vel, O., Erfani, S., Hubczenko, D., Leckie, C. and Montague, P. (2018). *Reinforcement Learning for Autonomous Defence in Software-Defined Networking*. [online] Available at: https://arxiv.org/pdf/1808.05770.pdf.
15. Hill, K. (2015). *Top 10 areas for 5G research*. [online] RCR Wireless News. Available at: https://www.rcrwireless.com/20150721/test-and-measurement/top-10-areas-for-5g-research-tag6.
16. ITU News (2020). AI and machine learning for a 5G world: Meet the champions of the ITU AI/ML in 5G Challenge. *www.itu.int*. [online] 18 Dec. Available at: https://www.itu.int/en/myitu/News/2020/12/18/13/45/ITU-AI-ML-machine-learning-5G-grand-challenge-winners.
17. ITU-R (2015). *M.2083 : IMT Vision - "Framework and overall objectives of the future development of IMT for 2020 and beyond."* [online] www.itu.int. Available at: http://www.itu.int/rec/R-REC-M.2083-0-201509-I/en.
18. Kaggle - Costs (2019). Weekly Maximum GPU Usage | Data Science and Machine Learning. [online] Kaggle.com. Available at: https://www.kaggle.com/general/108481 [Accessed 2020].
19. Kim, B. (2020). Google newly launches Colab Pro! - comparison of Colab and Colab pro · Buomsoo Kim. [online] buomsoo-kim.github.io. Available at: https://buomsoo-kim.github.io/colab/2020/03/15/Google-newly-launches-colab-pro.md/.
20. Li, G., Shen, Y., Zhao, P., Lu, X., Liu, J., Liu, Y. and Hoi, S.C. H. (2019). Detecting cyberattacks in industrial control systems using online learning algorithms. *Neurocomputing*, [online] 364(2019), pp.338–348. Available at: https://www.sciencedirect.com/science/article/pii/S0925231219309762 [Accessed Aug. 2021].
21. Novaes, M.P., Carvalho, L.F., Lloret, J. and ProençaJr., M.L. (2021). Adversarial Deep Learning approach detection and defense against DDoS attacks in SDN environments. *Future Generation Computer Systems*, [online] 125(2021), pp.156–167. Available at: https://www.sciencedirect.com/science/article/pii/S0167739X21002429 [Accessed Aug. 2021].
22. Ouali, K., Kassar, M., Nguyen, T.M.T., Sethom, K. and Kervella, B. (2020). *An Efficient D2D Handover Management Scheme for SDN-based 5G networks*. [online] IEEE Xplore. Available at: https://ieeexplore.ieee.org/document/9045534 [Accessed 2020].
23. Ovs- System Requirement (2016). Open vSwitch on Linux, FreeBSD and NetBSD — Open vSwitch 2.14.90 documentation. [online] docs.openvswitch.org. Available at: https://docs.openvswitch.org/en/latest/intro/install/general/#installation-requirements.
24. Raikar, M.M., S M, M., Mulla, M.M., Shetti, N.S. and Karanandi, M. (2020). Data Traffic Classification in Software Defined Networks (SDN) using supervised-learning. *Procedia Computer Science*, 171(2020), pp.2750–2759.

25. Said Elsayed, M., Le-Khac, N.-A., Dev, S. and Jurcut, A. (2020). *DDoSNet: A Deep-Learning Model for Detecting Network Attacks*. [online] IEEE. Available at: https://arxiv.org/pdf/2006.13981.pdf.
26. Sobers, R. and Varonis, Inc (2021). 56 Must Know Data Breach Statistics for 2019. *Inside Out Security*. Available at: https://www.varonis.com/blog/data-breach-statistics/ [Accessed Aug. 2021].
27. Third Generation Partnership Project - 3GPP (2015). *Specification #: 22.891 - Study on new services and markets technology enablers (Change Control)*. *3gpp.org*. Available at: http://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=2897.
28. Tunggal, A.T. (2021). The 29 Biggest Data Breaches [Updated for 2020]. *Upguard.com*. Available at: https://www.upguard.com/blog/biggest-data-breaches [Accessed Aug. 2021].
29. University of New Brunswick (2020). Datasets | Research | Canadian Institute for Cybersecurity | UNB. [online] www.unb.ca. Available at: https://www.unb.ca/cic/datasets/index.html.
30. Wang, P., Ye, F., Chen, X. and Qian, Y. (2018). Datanet: Deep Learning Based Encrypted Network Traffic Classification in SDN Home Gateway. *IEEE Access*, [online] 6, pp.55380–55391. Available at: https://ieeexplore.ieee.org/document/8473682.
31. Yu, D., Dong, J. and Kang, J. (2021). Service Attack Improvement in Wireless Sensor Network Based on Machine Learning. *Microprocessors and Microsystems*, [online] 80(103637), pp.1–2. Available at: https://www.sciencedirect.com/science/article/pii/S0141933120307845 [Accessed Aug. 2021].