Zero Trust – Malicious Traffic Detector

Amie Davis
Spring 2021
https://amodavis.github.io/

Which Domain?

This data will come from the Cybersecurity realm supporting the Zero Trust model, in which the mantra is "Never trust, always verify."

There are multiple avenues that can be explored with Zero Trust. Identity and Asset Management is required to manage and secure human and non-human identities. Threat Protection is needed to provide automated security, stopping attacks before they begin. Information Protection is essential to protect sensitive data, and Cloud security is necessary to protect assets stores in the Cloud.

For this project, I will focus on Threat Protection. Incoming and outgoing traffic is reviewed to identify suspicious network activity.

1) This article speaks to the philosophy of Zero Trust. There was a time when networks were blocked at the perimeter only. Once credentials were authenticated, users were allowed in. This is no longer sufficient. It is essential to review data at various checkpoints to review for malicious activity. The days of reviewing logs manually after a breach are no longer feasible. Machine learning is necessary to make sense of the vast amount of logs being collected and to make decisions in real time. It also highlights the use of machine learning to help determine behavioral patterns of users. This is necessary to identify anomalies.

   Anchan, P. (2019, December 6). *Achieve Smart Security Using Machine Learning for Zero Trust: Ilantus Blog*. Ilantus Technologies. https://www.ilantus.com/blog/machine-learning-for-zero-trust-how-can-it-be-done/.

2) Zero Trust security is a framework that is "built on four main pillars: (1) verify the user, (2) validate their device, (3) limit access and privilege, and (4) learn and adapt." Columbus also speaks to the advantage Zero Trust brings in which machine learning is not only used to identify normally accepted behavior but to allow users to "apply for conditional access without impacting user experience." This is considered ideal since all anomalies are blocked, but users can be provided automated temporary access to move forward. This is the tricky part: automating approvals.

   Columbus, L. (2018, May 11). *Three Ways Machine Learning Is Revolutionizing Zero Trust Security*. Forbes. https://www.forbes.com/sites/louiscolumbus/2018/05/11/three-ways-machine-learning-is-revolutionizing-zero-trust-security/?sh=4a4d654481ed.

3) This Joint Cybersecurity Advisory from the Cybersecurity and Infrastructure Security Agency (CISA) advises against Tor traffic. "Tor (aka The Onion Router) is software that allows users to browse the web anonymously by encrypting and routing requests through multiple relay layers or nodes."

"While Tor can be used to promote democracy and free, anonymous use of the internet, it also provides an avenue for malicious actors to conceal their activity because identity and point of origin cannot be determined for a Tor software user." They advise you to block incoming Tor traffic.

Alert (AA20-183A): Defending Against Malicious Cyber Activity Originating from Tor. Cybersecurity and Infrastructure Security Agency CISA. (2020, October 24). https://us-cert.cisa.gov/ncas/alerts/aa20-183a.

4) Tor was created for a common cause, that "internet users should have private access to an uncensored web." Since its origins in the 1990s, Tor has been used to "route traffic through multiple servers and encrypt it each step of the way."

   Since the Tor browser is available to anyone, it can easily attract malicious cyber actors. The disguise makes it difficult to determine good traffic versus bad traffic. Therefore, it is best to detect and prevent all Tor traffic.

   The Tor Project: Privacy & Freedom Online. Tor Project. (n.d.). https://www.torproject.org/about/history/.

5) Ferron describes deep web as the second network layer on the internet and a third layer called the darknet "where hackers congregate and facilitate illegal meetings."

   Ferron, J. (2017, May 4). *The Darknet and Deep Web: What Are They, and Why Should I Care?* ISACA. https://www.isaca.org/resources/news-and-trends/isaca-now-blog/2017/the-darknet-and-deep-web-what-are-they-and-why-should-i-care.

6) Zero trust architecture focuses on protecting resources, as opposed to the network. These resources include hardware, applications, data, and users. This architecture assumes there is an intruder inside the network and ensures resources cannot be taken or misused.

   Rose, S., Borchert, O., Mitchell, S., & Connelly, S. (2020, August 10). *Zero Trust Architecture*. NIST. https://www.nist.gov/publications/zero-trust-architecture.

7) This article with embedded video demonstrates how Tor traffic can be detected manually by reviewing Wireshark logs.

   Reese, S. (2016, January 16). *Detecting Tor traffic with Bro network traffic analyzer*. rsreese.com. https://www.rsreese.com/detecting-tor-traffic-with-bro-network-traffic-analyzer/.

8) Redekop discusses how the traditional approach of blocking suspected traffic has changed to allowing only expected traffic under the Zero Trust model and demonstrates how this can be enforced. However, doing this without impacting user experience is the challenge.

   Redekop, D. (2017, July 28). *Using a Zero Trust Model to block outbound VPN, Proxy, TOR, and P2P*. DNSthingy. https://www.dnsthingy.com/2017/07/using-a-zero-trust-model-to-block-outbound-vpn-proxy-tor-and-p2p/.

9) This paper highlights another project performed with the CICDarknet2020 dataset. Although it strives to tackle a similar problem set, it focused on the purpose of the traffic and includes Virtual Private Network (VPN) data.

   Demertzis, K., Tsiknas, K., Takezis, D., Skianis, C., Iliadis, L (2021). *Darknet Traffic Big-Data Analysis and Network Management for Real-Time Automating of the Malicious Intent Detection Process by a Weight Agnostic Neural Networks Framework*. Electronics 2021, 10, 781. https://www.researchgate.net/publication/350382621_Darknet_Traffic_Big-Data_Analysis_and_Network_Management_for_Real-Time_Automating_of_the_Malicious_Intent_Detection_Process_by_a_Weight_Agnostic_Neural_Networks_Framework

10) This technical tip demonstrates how to configure Fortinet to block known Tor browser signatures. It describes concern with the Tor browser as it "allows users to bounce communication traffic around a distributed network of relays located around the world."

    Technical Tip: Blocking and monitoring Tor traffic. Fortinet. (n.d.). https://kb.fortinet.com/kb/documentLink.do?externalID=FD36379.

Which Data?

I will be using the DarkNet 2020 dataset from the Canadian Institute for Cybersecurity at the University of New Brunswick. This dataset expands on Wireshark data previously collected in 2016, during which the data were analyzed and labeled as either malicious (Tor) or benign (non-Tor). What makes this dataset unique is that it also includes the category of the traffic: Audio-Stream, Browsing, Chat, Email, P2P, Transfer, and Video-Stream. This provides a more readable dataset, as opposed to deciphering ports and protocols. In addition to source and destination ip addresses, port and protocol are included. There are also approximately 75 numeric features describing the traffic, such as packet size and speed.

CIC-Darknet2020. University of New Brunswick. (n.d.). https://www.unb.ca/cic/datasets/darknet2020.html.

Research Questions? Benefits? Why analyze these data?

I will analyze the dataset by reviewing subsets and traffic categories. Identifying patterns in the data that are malicious will provide insight as to other risks that may need to be mitigated. It may also shed light on possible user impacts.

- How do you know traffic is malicious?
- What features contribute the most to Tor traffic?
- From what ip subnet is Tor traffic most likely to come?
- How much Tor traffic is detected by each category?
- Is Tor traffic faster or slower than non-Tor traffic?
- Is Tor traffic larger or smaller than non-Tor traffic?

What Method?

I will review features for dependence and multicollinearity, removing features as needed to maintain independence and minimize correlation. Numeric features will be assessed for distribution.

I will use Logistic Regression to solve a Binary Classification problem. Is the traffic suspected to be malicious, yes or no? If the traffic has a high probability of being suspect, an alert will be produced, and the traffic can be blocked.

Potential Issues?

I am concerned that the dataset will be imbalanced. The original data were compiled from Wireshark logs. With real data, most of the traffic would likely be benign, with a few exceptions. However, these data were generated, so there is hopefully a better balance.

I am also concerned that the model will be overfitted to the training data. Since the data were generated, are they too specific to the problem set? Were the data generated for specific ip addresses? I will try to mitigate overfitting by breaking up the ip addresses into subnets. This will also allow me to assess them as numeric data fields.

Concluding Remarks

As the amount of data we share increases, cybersecurity becomes more critical to protect it. We have also seen the need for accessibility increase during the pandemic. Users are connecting from multiple locations on multiple devices. It is necessary to provide this flexibility, while still securing our networks. To enable more access, we must review and validate user access and network access continually. Machine learning is critical to achieving this goal.