KF7004 – MComp Computing Research Project
MComp Research Proposal

16018262

Sept 22

Word count

Contents

1 Research question

A hursic analysis of website log files to detect attacks

2 Aims

The aim of this research is to build upon work carried out as part of a 2019 Study by Smith P. looking at a formula approach to detecting risks posed by website traffic. The work done by Smith attempted to use website log files to detect suspicious activity on a website. This work will collect more data to prove the accuracy of this approach. As well as expanding the number of data points to detect attacks for example who the network an IP belongs to and the user agents. The study by Smith P. only has a relatively small data set to it is hard to draw any wider conclusions about the accuracy of the technique proposed.

3 literature review

There has been multiple studies looking at website attacks however typically these have focused on single variable analyses for example CPU depletion. For example, Erwin Adi has done a lot of research into Low-rate Denial of Service (LDoS) attacks. His primary paper looks at CPU depletion as an indicator of attack. In the same paper, Adi himself admits that this maybe a flawed technique for attack detection. Adi et al. 2016 Most previous studies into detecting Low Bandwidth attacks only look at a single data point such as CPU. The present research proposes to look across multiple data points to detect attacks.

Research in the area of high rate attacks don't pose the threat they used due to high volume of research in this area, meaning that preventative measures have been made. As Agrawal & Tapaswi state "due to this huge volume of malicious traffic, such attacks can be easily detected. Thus, attackers are getting attracted towards the low-rate DDoS attacks, slowly. Low-rate DDoS attacks are difficult to detect due to their stealthy and low-rate traffic" (Agrawal and Tapaswi 2019). Futhermmore in a study by Zhijun et al. in 2020, they state that "Low-rate Denial of service (LDoS) attacks has become one of the biggest threats to the Internet, cloud computing platforms, and big data centers" (Zhijun et al. 2020) showing the need for an effective attack detection tool.

Cloudflare statistics indicate that high rate ddos attacks are increasing yearly however due to the protocols in place high rate ddos attacks aren't successful with there effects being easily mitigated (Yoachimik 2021). Whereas low rate attacks can be more easily disguised alongside more genuine use of the website thereby bypassing the protocols in place for high rate attacks. It's important to develop a system that is able to detect a wide variety of attacks in order to build a more comprehensive picture of low rate attacks and keep websites safer.

The rate of request is not necessarily the only signal of an attack. If for example if an IP address is constantly searching for login pages or back up files this could be an indication of suspicious behaviour. There is not a lot of data on low rate attacks and this could be a cause of concern and if we don't know how many attacks are happening it may indicate there is no detection method for the attacks.

Tripathi proposed an anomaly detection technique that attempted to detect attacks by measuring the Chi squared (X^2) differential value between the expected traffic pattern, the result suggested attacks could be found with high accuracy. Tripathi Collected 14 hours of HTTP/2 traffic, a fundamental issue with this being that the researcher simulated the data,(Tripathi and Hubballi 2018) whilst theoretically sound there are no real world examples showing this to be effective. Therefore the data is not reliable because of the small time frame in which it was carried out and you cant say for sure whether the changes in traffic would've occurred naturally making it hard to differentiate peaks in traffic or an attack. For this reason, Rajbahadur et al. advocates for "the increased utilization of real-world data instead of simulated data." (Rajbahadur et al. 2018). Therefore the present paper looks to use real data from multiple websites over an increased time-window, using data sciences techniques, to evaluate its efficiency.

Staniford, Hoagland and McAlerney suggest that storing large amounts of network traffic may be impractical Staniford, Hoagland, and McAlerney 2002. However Zhijun et al. states that "A huge amount of network traffic can be collected, stored, organized and classified by big data analysis. Moreover, the detection judgement and defense decision can be achieved by analyzing unknown patterns" Zhijun et al. 2020 Therefore if there is a need for a large amount of data, that may be impractical to store. One solution may be to look at the data already available to analyse and designing a suitable way to analyse that data. Most website keep log files of who is accessing the website and activity history. Therefore theoretically the log files negate the need to collect extra data and may provide sufficient information to detect attacks. This method was tried by Smith (2020). However this only used a small data sample, but did have promising results as it was able to distinguish good traffic from bad traffic.

All the research done to date looks at traffic flow in various ways, however it fails to take into account where that traffic is coming from, for example, most cyber attacks emanate from Russia and China, so the research is ignoring a key area that needs to be explored.

The work done by Smith proposed a formula that took many factors into account. The overall formula was defined as

$$risk = (orrcancesOfipLog \times 0.6) + ((requestRisk + responseRisk) \times 0.3) + (countryRisk \times 0.1)$$

this formula was the first to look at multiple data points when detecting low rate attacks however in Smith's conclusions they state that the network that the IP address comes from could potentially have a greater impact on the risk as due to VPN technology the country can change. Furthermore, the risk of a particular country only looks at the total number of attacks and Smith points out that ”the values used in the software are only based on the number of attacks per country” (Smith 2020) therefore the overall calculation will be changed and the underlying risk assigned to each country will be assessed looking at attacks per head of population. Also, as well as the country, it would be useful to look at the network the IP comes from and build that into the calculation. When looking at the risk of an individual country, the values used in the software are only based on the number of attacks per country. While this is a good way to assess the risk of a country, this methodology could potentially have issues, for example, larger countries will statistically have more attacks than smaller countries. According to the Nexusguard 2018 2 threat report, it was identified that 23.34% of attacks came from China, with a further 14.90% coming from the USA; this is said to be expected, due to the internet presence of their citizens parentage per population, with China's being over one billion users. Therefore, it would be good to look at the total number of attacks per country that are reported to the software for instance, compared to the size of the population.

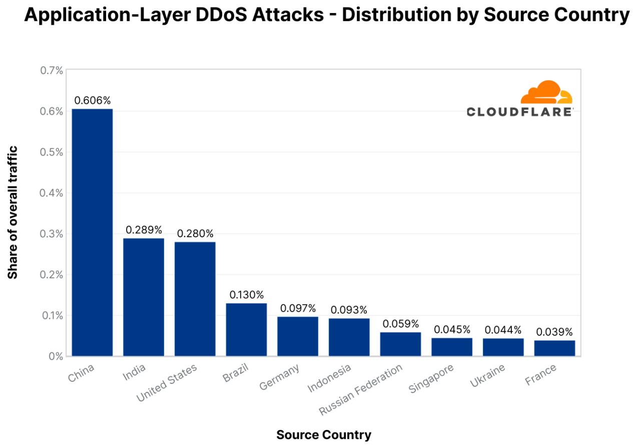

Figure 1: Attacks by country according to Cloudflare

According to Cloudflare in their DDoS threat report 2022, application layer DDoS attacks are mainly comming from China (Yoachimik 2021)

3.1 revlantcy

The work is relevant due to the increasing number of websites increasing as nearly all businesses have a website. In the aftermath of remote working and lock downs E-commerce sites increased (CITE NEEDED) the types of attacks the study aims to detect can be hard to identify. Cloudflare notes in August 2022 that

As websites become part of daily life the number of sites online is increasing as such there are more and more potential targets for attackers and more exploits are discovered, as attacks get more complex the is a need for different detection techniques.

4 ethics

4.1 Research Ethics

When thinking about an ethical way of collecting data one of the key questions is should the website data is collected say the data is being used for this?

The main ethical issue is around collecting IP addresses this could potentially lead to individual being identified however the system will only look at the network the IP belongs to.

Whilst the consent of the website ownes was obtained the one ethical issue could be that havent recieved consent from each of the individuals whose data I will be analysises. However if someone is trying to dos malicious activity on a website they may want their data excluded from the analysis. Therefore it will be difficult to prove if it can pick up malicious activity.

A full ethics form can be found in appendix A

4.2 Wider Ethics

One wider ethical issue of software like this could be if an attacker was able to figure out the formula they could work out how to get around the formula

5 Scope

The scope of this work is to build a samall programme to analyse the data in website log files to determine if attacks can be detected. The mains points of the work are:

- understand attack characteristic
- identify how attack are evading traditional techniques
- develop a formula that can detect and determine risk

This work will not automatically block IP addresses from accessing the website due to the fact that this may cause IP addresses to be incorrectly blocked Global, pointed out that the unique 'human' ability to appraise the conitextual features of a potential threat means that removing them from the loop of a security methodology is inadvisable. (Global 2018) So this work should be seen as a way to aid the decision making of website owners rather than make the decision for them.

Also this work will not check the ability of people to use the software due to the fact it may be difficult to determine if it was the formula or user that identifies attack Bryman states that "If we suggest that X causes Y, can we be sure that X is responsible for the variation for the Y and not something else". (Bryman 2016;41). The work is fouced on proving if a formula can idenify attack traffic.

This work will also not be looking to generate its own data due to the fact that it may be easier to prove the accuracy of the formula on real data sets and will mean that the formula is written in ways that it can interpret real data.

6 Risks

One of the potential pit falls is differentiating potentials attacking with genuine user error. For example with low rate ddos attacks a login page could be repeatudly loaded however this could also be due to a user forgetting their password this is why the research needs multiplce indication of intent before classing this a attack.

People may be un willing to give data

References

Adi, Erwin et al. (2016). “Distributed denial-of-service attacks against HTTP/2 services”. In: *Cluster Computing* 19.1, pp. 79–86. ISSN: 1573-7543. DOI: 10.1007/s10586-015-0528-7. URL: https://doi.org/10.1007/s10586-015-0528-7.

Agrawal, Neha and Shashikala Tapaswi (2019). “Defense Mechanisms Against DDoS Attacks in a Cloud Computing Environment: State-of-the-Art and Research Challenges”. In: *IEEE Communications Surveys & Tutorials* 21.4, pp. 3769–3795. DOI: 10.1109/COMST.2019.2934468.

Bryman, Alan (2016). *Social research methods*. Oxford University Press.

Global, F-Secure (Jan. 2018). *How to Detect Targeted Cyber Attacks: The Importance of Context*. Online; accessed 16/01/2020. URL: https://blog.f-secure.com/detect-targeted-cyber-attacks-importance-context/.

Rajbahadur, Gopi Krishnan et al. (2018). “A Survey of Anomaly Detection for Connected Vehicle Cybersecurity and Safety”. In: *2018 IEEE Intelligent Vehicles Symposium (IV)*, pp. 421–426. DOI: 10.1109/IVS.2018.8500383.

Smith, Peter (2020). “Formulaic approach of detecting website attack traffic with critical evaluation of current methodology”. Unpublished.

Staniford, Stuart, James A Hoagland, and Joseph M McAlerney (2002). “Practical automated detection of stealthy portscans”. In: *Journal of Computer Security* 10.1-2, pp. 105–136.

Tripathi, Nikhil and Neminath Hubballi (2018). “Slow rate denial of service attacks against HTTP/2 and detection”. In: *Computers & security* 72, pp. 255–272.

Yoachimik, Omer (Oct. 2021). *Cloudflare DDoS threat report 2022 Q3*. Online; accessed 28/10/2020. URL: https://blog.cloudflare.com/cloudflare-ddos-threat-report-2022-q3/.

Zhijun, Wu et al. (2020). “Low-Rate DoS Attacks, Detection, Defense, and Challenges: A Survey”. In: *IEEE Access* 8, pp. 43920–43943. DOI: 10.1109/ACCESS.2020.2976609.

A Ethics