KF7004 – MComp Computing Research Project
MComp Research Proposal

16018262

Sept 22

Word count: 1584

Contents

1 Research question

A huristic analysis of website log files to detect attacks using a formulaic approach.

2 Literature review

There have been multiple studies looking at website attacks but typically these have focused on single variable analysis e.g. CPU depletion. A good 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.

Due to high volume of research in this area, high rate attacks don't pose the threat they used to, showing 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. 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 their 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, the cumulative total over time should be considered. 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.

An anomaly detection technique was proposed by Tripathi that attempted to detect attacks by measuring the Chi squared (X^2) differential value between the expected traffic pattern, the result suggested that 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 it cannot be said 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.

It has been suggested 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." (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 do so. Most websites 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.

Previous research 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 documented attempt to look at multiple data points when detecting low rate attacks. However in Smith's conclusions he states that the network that the IP address comes from could potentially have a greater impact on the risk as due to VPN technology can change the country. Furthermore, the risk of a particular country only looks at the total number of attacks per country 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. In additon to 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 from an individual country, the values used in the formula are only based on the historical 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 Cloudflare DDoS threat report 2022 Q3, it was identified that the number of attacks coming from China-registered IP addresses increased by 29% from the previous year. India has the second- largest source of HTTP DDoS attack traffic with an increase of 61% (Yoachimik 2021).

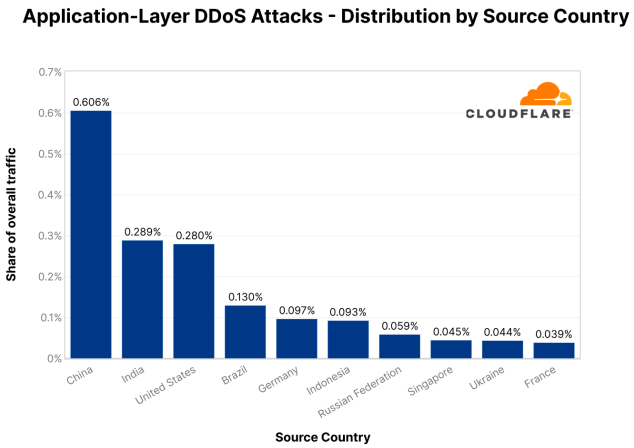


Figure 1: Attacks by country according to Cloudflare

3 Aims

This research aims to develop a novel formula that will result in a new appoarch to attack detection, as was shown in the literature review. There has been very litttle work done with real data to detect attacks on websites, therefore one of the main aims of the present study is to develop the formula, and test its effectiveness using real data, instead of simulated data that other studies have used. Should the aims of this novel piece of research be achieved, it will set a new precedent for the future of cybersecurity and expose weaknesses in current methodologies.

3.1 Technical approaches and datasets

In order to achieve these aims, the approaches used will be combining existing techniques in a new way. There are websites online that track ip adresses that are attacking websites (such as https://www.abuseipdb.com/) however as they put the ip addresses online, attackers can see when the ip address has been detected. This work intends to hide this infomation from users also, only 10% of the risk comes from previous knoweldge. There will be a backend database that will hold data about the IP adresses of known bots so that these can return a risk of 0.

Another table will hold risks of http status codes as seen in appendix B, by keeping the values in the database table, it is easier to modify the risk. Furthermore another table holds signatures known attacks so that this can be run against the data. This makes it very quick to add new attacks to the formula.

The data analysed will be the apache common log format, this format contains a lot of technical information that will be beneficial to the formula. This data is stored in various data structures that can then be accessed to determine risk. In addition to this, the data from the database, for example the fragments of known bots are getting stored so that they can be cross referenced.

To be able to identify the country of origin IP address, the Maxmind geoip database is used. Then each country is assigned risk wieghting within a seperate database. The same process is used for the network risk by using the maxmind databases, it ensures that the IP information is always up to date.

4 Legal, social and ethical considerations

An area of concern that may be a legal, social and ethical issue is whether the users of a website should be informed that their IP address is being collected for analysis and be aware of its repercusions, as collecting IP addresses could lead to individuals being identified, although the database used for the location of an IP states that 'IP geolocation is inherently imprecise. Locations are often near the center of the population' (*GeoLite2 Free Geolocation Data*). However most privacy policies on websites states that the IP will be collected to analyse trends, including one of the websites that has a agreed to donate data (Peters-Web 2022) and the system will only look at the network and country the IP belongs to.

Whilst the consent of the website owners were obtained, the one ethical issue could be that the research has not recieved consent from each of the individuals whose data will be analysed. However anyone attempting malicious activity on a website may want their data excluded from the analysis. Therefore it will be difficult to prove if the forumla can pick up malicious activity. One social and ethical issue could be that the entire countries are given a risk, however a user in that country may have a legitimate reason to access a website. Therefore, to mitigate the risk of a social issue, the country that an IP belongs to will be given a minimal weight in the overall formula.

A potential issue with all security both digital and physical is that people will always try and circumvent the measures put in place, therefore people might try to reverse engineer the analytics. This is not unique to this research as Schneier states that "no software is secure against reverse-engineering."(Schneier 1998)

A full ethics form for this research can be found in appendix A

5 Scope

The scope of this work is to build a small programme to analyse the data in website log files, which determines if attacks can be detected. The main points of the work are:

- identify how attacks are evading current techniques
- understand attack characteristics and refine the formula
- develop an updated formula that can detect and determine risk

This work will not automatically block IP addresses from accessing the website as that this may cause IP addresses to be incorrectly blocked. It has been pointed out that the unique 'human' ability to appraise the contextual 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 making the decision for them.

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). Using this as a framework, this research is focused on proving if a formula can identify attack traffic. If user testing was done, it may be difficult to determine if it was the formula or the user that identifies attacks.

Due to the limitations of simulated samples, identified in Tripathi's paper, this work will not be looking to generate its own data as it may be easier to prove the accuracy of the formula on real data sets and means the formula is written in ways that it will be possible to interpret real data.

6 Risks

One of the potential pit falls is to differentiate potential attacks with genuine user error. For example with low rate DDOS attacks, a login page could be repeatedly loaded, however this could also be due to a user forgetting their password.This is why the research needs multiple indication of intent before classing this an attack.

Due to the sensitive nature of the website logs, website owners maybe unwilling to give these logs for analysis, however early feedback seems to indicate people are willing to have their website data analysed. In addition to this, it may be difficult to determine the efficacy of the formula due to the fact that there may not be any attacks in the selcted data set.

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/.

MaxMind. *GeoLite2 Free Geolocation Data*. Online; accessed 22/11/2022. URL: https://dev.maxmind.com/geoip/geolite2-free-geolocation-data?lang=en (visited on 2018).

Peters-Web (Nov. 2022). *Privacy Policy Peters Web*. Online; accessed 08/11/2022. URL: https://www.petersweb.me.uk/privacy-policy/.

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.

Schneier, B. (1998). "Cryptographic design vulnerabilities". In: *Computer* 31.9, pp. 29–33. DOI: 10.1109/2.708447.

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

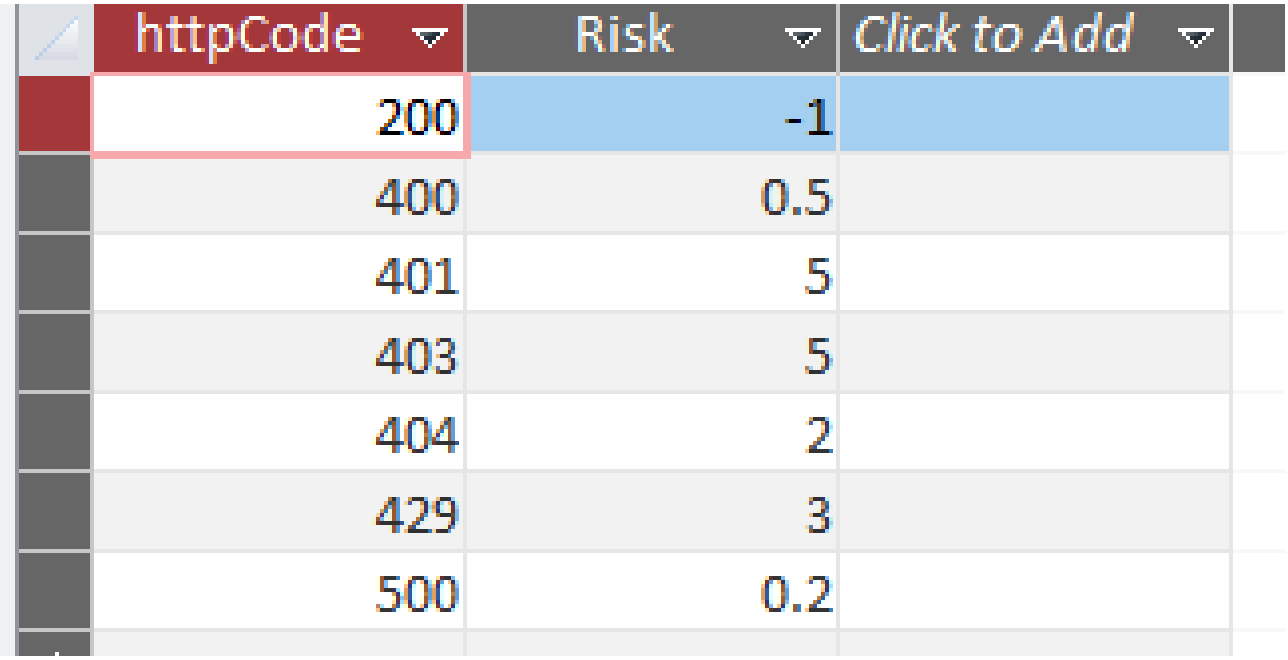
Staniford, Stuart, James A Hoagland, and Joseph M McAlerney (2002). "Practical automated detection of stealthy portscans". In: *Journal of Computer Security* 10.1-2, pp. 105–136.

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



httpCode	Risk	Click to Add
200	-1	
400	0.5	
401	5	
403	5	
404	2	
429	3	
500	0.2	

Figure 2: HTTP Status codes

B HTTP Status codes