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 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, 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.

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 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 analyse that data. 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).

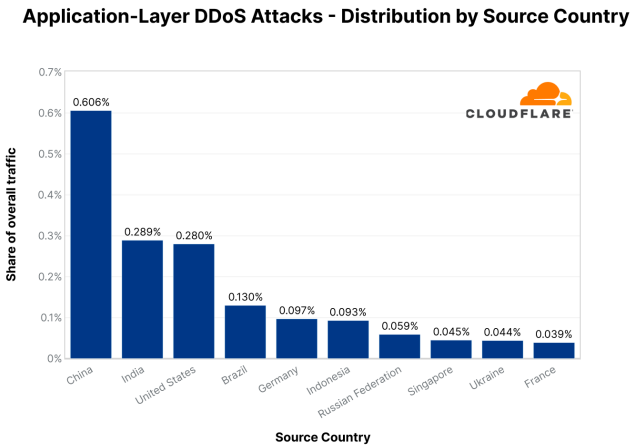

Figure 1: Attacks by country according to Cloudflare

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 aim 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

The approaches used in this work 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 its repercusions as collecting IP addresses could lead to individuals being identified. 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 determine if attacks can be detected. The main points of the work are:

- identify how attack are evading current techniques
- understand attack characteristics and update the fomula
- develop a 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.

The work is focused on proving if a formula can idenify attack traffic. If user testing was done, it may be difficult to determine if it was the formula or the user that identifies attack. As Bryman states "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).

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. If there is lack of data then website could be setup to get log files however this data may be not as desirable as it would be simulated. In addition to this, the websites that are analysed may not have been attacked.Therefore it will be harder to prove if the forumla can detect attacks.

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/.

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

Northumbria University NEWCASTLE

Main Ethics Application Form

Are you a student or member of staff?

- ○ Staff
- ◉ Student

Please confirm your level of study:

- ○ Undergraduate
- ◉ Postgraduate Taught
- ○ Postgraduate Research

Please choose your Faculty or Service from the list:

- ○ Arts Design and Social Sciences
- ○ Business and Law
- ◉ Engineering and Environment
- ○ Health and Life Sciences
- ○ Campus Services
- ○ Finance
- ○ Global Marketing and Business Services
- ○ Human Resources
- ○ IT Services
- ○ Research and innovation Services
- ○ Student, Library and Academic Services
- ○ Vice Chancellors Office

Please choose your Department from the Faculty of Engineering and Environment

- ○ Architecture and Built Environment
- ◉ Computer and Information Sciences
- ○ Geography and Environmental Sciences
- ○ Mathematics, Physics and Electrical Engineering
- ○ Mechanical and Construction Engineering

Does your research require external approval? i.e., NHS, MoD, Social Care

- ○ Yes
- ◉ No

Is this application linked to a Module Level Approval

- ☐ Yes
- ☑ No

Is your application linked to an academic-led project?

- ☐ Yes
- ☑ No

Please input the name of your Module Tutor

First Name

Alan

Surname

Godfrey

Please input the name of your Supervisor

First Name

Alan

Surname

Godfrey

Ethical Review Categories

Does your study involve any of the following: (tick all that apply)

- [ ] Gathering data or information from human participants (e.g. via questionnaire / interview/survey/experiment/ VR)
- [x] Collecting personal data, i.e. name, email, home address, computer IP address, phone number etc.
- [x] Analysis of secondary data either in or outside of the public domain
- [ ] Lab-based research
- [ ] The collection or use of information which is 'commercially sensitive'
- [ ] Financial inducements other than expenses and compensation for time
- [ ] Gathering data/information at a physical location external to Northumbria University campuses, franchised locations, and not your normal place of work
- [ ] Collection of samples such as plants, soils etc, that might disturb the environment or archaeological remains
- [ ] Individuals or groups where permission of a gatekeeper is normally required for initial or continued access to participants (e.g. NGOs, community leaders)
- [ ] Research with potentially vulnerable participants or groups, including people under 18 (which may require DBS clearance)
- [ ] Discussion (e.g. interviews) of highly sensitive topics that may cause undue stress to participants, and researchers, including, but not exclusively: sexual behaviour, drug use; abuse or exploitation; trauma; pornography.
- [ ] Funding from a source that may be controversial (e.g. due to the nature of the funder, or a conflict of interest).
- [ ] Covert methods of investigation or deception.
- [ ] Research with international partners, or research undertaken outside of the UK where there may be issues of local practice and political sensitivities.
- [ ] Access to records of personal or sensitive confidential information, including genetic or other biological information concerning identifiable individuals
- [ ] Intrusive interventions including the use of drugs or other substances (e.g. food, drink, placebos or drugs); and, or, procedures involving physical distress (e.g. prolonged testing) or emotional distress (e.g. stress or anxiety), that are greater than those you would encounter in everyday life.
- [x] Work that involves direct observation of, or participation in, activities during which it is anticipated that illegal activity, or regulatory breach is likely to occur (e.g. hunting, drug dealing, accessing the dark web, hacking).
- [ ] Access to or collection of data, information, materials (e.g. magazines, publications, websites, and social media) relating to extremism, radicalisation or terrorism (including extreme or terror groups).
- [ ] Funding/ sponsorship from, or the involvement of, the UK Ministry of Defence, Military (UK and International), and or, EU Security funding call.
- [ ] The collection of data/information that might be confidential or classified (e.g. protected by the Official Secrets Act) .
- [ ] The funding body e.g. ESRC funded projects require REC review.
- [ ] The collection of bodily tissue e.g. blood, saliva, urine samples from living persons (which may require licence under the HTA and additional training).
- [ ] Culturally sensitive art, artefacts or monuments, or sites.
- [ ] Research with animal subjects

General Aims and Research Design

Project Title

Website risk

Outline the general aims and research objectives of the project.

Collecting logs of website traffic and then analysing them with a formula

Please give a detailed description of your research activities, any ethical issues and how they are addressed:

Information will be collected from various websites. Data will be analyzed by a formula to see if attacks can be detected.
Ethical issues - I will know the IP address of visitors to a website. However, this is only used to interpret the location and networks belongs to. This will not be used to identify individual people.
Another ethical issue is there may be data that was an attack on a website when this is the case I will inform the website owner. This research is not going to cause any more attacks on websites and the research will not attack any websites although they may look at relevant theories.

People and Personal Data

Describe the data pertaining to living individuals you will be collecting

IP address

Describe your sample groups or how you will identify participants

The data is collected automatically by people visiting a website the research will not collect any data that isn't already collected automatically

Legal basis for processing

- ☐ Where the University processes personal data, in most cases, we do so under article 6(1)(e) GDPR, which permits processing that is 'necessary for the performance of a task carried out in the public interest or in the exercise of official authority vested in the controller'
- ☑ Where the University processes special categories of personal data, in most cases, we do so under Article 9(2)(j) GDPR, which permits processing that is necessary for scientific or historical research purposes, providing we have appropriate security safeguards in place

Describe your recruitment process

Not Applicable

Describe any remuneration for participants

Not Applicable

Type of consent

Other

Please provide details of the consent used:

If the study involves participants who lack capacity to consent, procedures in line with sections 30-33 of the Mental Capacity Act will need to be put in place (e.g. NHS SREC review). If you are using alternative formats to provide information and /or record consent (e.g. images, video or audio recording), provide brief details and outline the justification for this approach and the uses to which it will be put.

No consent is required as this is an analysis of data that is already collected

Please upload copies of the consent forms and participant information sheets. Please note you can upload videos and images.

Provide a detailed description of what the participants will be asked to do for the research study, including details about the process of data collection (e.g. completing how many interviews, assessments, when, for how long, and with whom).

Not Applicable

Upload any relevant documentation

Secondary Data

Describe the source of the data and any supplier terms

Data will be given by Peters Web after they have asked website owners. The privacy policy of Peters Web says that they collect the IP addresses and other technical information to analyse trends in data.

Security Sensitve Research

Do you require access to material that is prohibited/restricted (e.g. under Government security classifications or the Official Secrets Act)?

You can access the Government Security Classifications here: Government Security Classifications

- ○ Yes
- ◉ No

Does your research project relate to extremism, radicalisation and or terrorism?

- ○ Yes
- ◉ No

Does your research project activities involve accessing extremist, radical, or terrorist materials?

- ○ Yes
- ◉ No

Data Management

Describe the arrangements for anonymising data, and if this is not appropriate explain why

The IP addresses can not be anonymised as they are used by the analysis software to determine how much of a risk the IP address poses to the website

Describe the arrangements for the storage of any data:

The data will be stored on servers owned by Peter's Web, when the data needs to be analysed for a study, the data will be downloaded off the server for the website that has consented to take part. After the data is analysed, the data will be deleted off the computers that are being used to analyse the data. The the data left on the server will be deleted when the website owner choses to delete the data.

I confirm that I will comply with the University's Research Data Management Policy and data retention schedule and guidance:

- ☑ Yes

Will this research project involve data processing of identifiable high risk special category data?

High risk special category data includes:

i. Racial or ethnic origin.
ii. Political opinions.
iii. Religious and philosophical beliefs.
iv. Trade union membership.
v. Genetic data.
vi. Biometric data for the purpose of uniquely identifying a natural person.
vii. Data concerning health.
viii. Sex life and sexual orientation.

- ○ Yes
- ◉ No

Project Duration

Proposed start date of research activity

01/11/2022

Proposed end date of research activity

01/06/2023

Health and Safety

I confirm that I have read, understood and agree to abide by the University's Health and Safety Policy and arrangements.

- [x] Yes

I confirm that I have read and understood the University's requirements for the mandatory completion of risk assessments in advance of any activity involving potential health and safety risks that are not covered in general University risk assessments. The requirements can be found in the University Risk Assessment Guidance.

- [x] Yes

Please confirm that

- [ ] There are health and safety risks associated with the research project work that are not covered by the general University risk assessments. The correct risk assessment(s) have been attached and are appropriately approved.
- [x] I can confirm that there are no health and safety risks associated with this project, not covered by general University risk assessments and so no project specific risk assessments are required.

Additional Documents

Please upload any additional documents.

Insurance

Click here for University Insurance Questionnaire

- [x] I confirm I have completed and uploaded the University Insurance Questionnaire and that my research project is covered by University Insurance.

If you think your activity may involve a High Risk rating or are unsure how to answer the statements – contact fi.insurance@northumbria.ac.uk for advice, attaching a copy of your research project ethics submission and the University Insurance Questionnaire.

Further information can be accessed here.

I confirm my insurance risk level of:

- [x] Low
- [ ] Medium
- [ ] High

Declaration

I confirm that I have answered all of the sections as fully and accurately as possible.

Once you have signed the form it will automatically be checked for compliance with ethical and governance policies at Northumbria before being allocated for ethical review.

If you are a student once your supervisor has signed the signature request your form will be automatically submitted.

Supervisor Signature

Applicant Signature

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