M4: Wikimedia DDoS Detection Project Report

Charel Felten
University of Amsterdam
charel.felten@student.uva.nl

Alex Janczewski
University of Amsterdam
alex.janczewski@student.uva.nl

Gilles Magalhães Ribeiro
Vrije Universiteit Amsterdam
g.magalhaesribeiro@student.vu.nl

1. INTRODUCTION

"The mail servers are down." - is the computer alert that appeared on the night of September 27th, 1996 which signalled one of the first documented "denial of service" (DoS) attacks on Panix – a commercial Internet provider in New York [10]. The 'SYN flood attack', with messages being sent at rate of 210 per second, shut down the internet provider for weeks, thus preventing its customers from accessing their emails.

A distributed denial-of-service, also known as a DDoS attack, is a targeted attack made by a group of, often compromised, hosts with the aim of flooding a target with illegitimate traffic, consuming its resources, and thus denying access of legitimate users to services [18], [2], [12]. DDoS attacks are currently the most common type of cyber-attacks [11]. For example, an average of $500 - 800$ DDoS attacks were recorded per day in Q2 2021 by Kaspersky [8], peaking at 1800 attacks per day in January 2021. [6]. The consequences of DDoS attacks are in most cases reduced or no availability of the target sites. For example, a recent noteworthy volumetric flood took place in early September 2019 and disrupted the services of the Wikimedia Foundation (WMF) on nearly all of the continents [5].

This project report presents our efforts in devising methods to detect DDoS attacks relying on very limited data. We focus our efforts on the WMF page views. In order to do so we first collected and cleaned all relevant publicly available Wikistats data provided by WMF. Next, we performed exploratory data analysis and quantitative investigation of page views distributions using two information-theoretic metrics: Shannon entropy and Kullback-Leibler divergence. The investigation was based on concrete knowledge of a particularly powerful DDoS attack targeted at WMF taking place on September 6th, 2019 starting at 17:00 UTC [7].

2. RELATED WORK

With attacks becoming more complex and detrimental to online communities, a lot of novel techniques and methods are being developed to prevent, detect and mitigate DDoS attacks [11]. In fact, the detection of DDoS attacks poses a lot of challenges, as it is usually difficult to discern a DDoS attack from regular traffic [4]. Attackers either mimic normal network traffic by adjusting the speed of sending packets to reproduce regular Poisson distributions [16] or imitate flash crowds thus making it difficult to efficiently distinguish legitimate surges in traffic from an actual attack [18], [4].

Traditional techniques of detecting DDoS attacks involve monitoring traffic and using statistical divergence to distinguish an attack from legitimate network traffic [18]. Machine learning naturally found its application in supporting DDoS detection based on statistical features [18]. There are various ML-based models developed that make use of e.g. Naive Bayes or K-Nearest neighborhood to categorize the traffic [14]. Furthermore, Yuan, Li and Li also proposed the use of recurrent deep neural networks to learn patterns from sequences of network traffic and further improve the DDoS detection performance [18]. Another approach is proposed by Bhuyan and Kalita who use standard information-theoretic entropy measures such as Shannon entropy and Kullback-Leibler divergence to characterize network traffic data based on IP address and packet size distribution statistics for low-rate and high-rate DDoS attacks [2]. In fact, information entropy- and divergence-based measures are currently widely accepted by the research community as one of the most effective and efficient methods to detect illegitimate network traffic [1], [17], [2], [15].

As previously mentioned in our project proposal, all of the methods that we encountered in our extensive research relied on characterization of traffic flow by tracking source and destination IP addresses, as well as packet size distribution statistics [12], [1], [2], [16], [4]. Therefore, we have not been able to simply base our investigation methods on any particular literature reviewed.

Nevertheless, the method to quantify uncertainty of accesses, Shannon entropy, as well as the method to determine differences between page view probability distributions, Kullback-Leibler divergence, were inspired by the techniques widely used to characterize and compare requested packet size distributions.

3. RESEARCH QUESTION

In view of the fact that the scientific community relies so heavily on the data that we do not have access to, our research aims to determine if it is feasible to detect DDoS attacks with limited data, and especially without knowing the source and destination IP addresses and package sizes requested. The underlying assumption of our research is that if it is possible to detect DDoS attacks within this dataset, then we should be able to observe at least some kind of anomalies or deviations from normal pageview activity on the day of September 6th, 2019 as this is the biggest known DDoS attack recorded on WMF in the available data timespan.

We hypothesize that a DDoS attack leads to an increase in traffic inflow towards individual pages, a whole domain or across all of WMF. As a result, we expect to see both an increase in page views given that there is more traffic than usual to qualify as a DDoS attack, as well as changes to the page view distribution given that this synthetic traffic may not follow the same distributions as normal traffic. More formally, we hypothesise that at the time of the DDoS attack:

1. There will be an increase in page views.
2. There will be an increase in the KL divergence of the page view distribution.
3. There will be an increase in the Shannon Entropy of the page view distribution.

We will test these hypotheses by defining the attack day and time to be September 6th, 2019 at 17:00 UTC onwards, and all other days of days of August, September and part of October 2019 as normal days.

4. PROJECT SETUP

4.1 Data investigation

Various dataset 'dumps' are available from WMF at `https://dumps.wikimedia.org/other/analytics/`. Most relevant to this project are the hourly page view statistics, which contain information on how often every single WMF page was accessed per hour, over several years. This dataset is, however, available in various formats and versions, all slightly different from each other. One version, `pageviews`, is interesting as it contains the response size of hourly requests. It is however only available in a filtered version, with bot and webcrawler traffic removed, making it unsuitable for our research as this filtering may exclude the patterns and artefacts from a DDoS attack. Other versions like `pagecount-raw` and `pagecount-ez` are by now either outdated and no more updated or actually deprecated. Then, there is `wmf.pageview_hourly` which is the most extensive database, available on Hive in Parquet format. Unfortunately, it is private and only available to researchers with a formal application process out of scale (due to 15+ days for the application) for this project. The last remaining option is `pageview-complete`, which is the most extensive publicly available dataset of pageviews and the one we will be focusing our efforts on. We will refer to this dataset simply as the 'pageviews' dataset in the remainder of the report.

The pageviews dataset contains hourly aggregated access to all WMF pages over a 10 year period starting from November 2011 until now. Data is available in either daily or monthly aggregated bzip2-compressed CSV files. The size of all daily files is 2747 GiB and the monthly ones are 591 GiB. The daily aggregated dataset contains requests on a per-hour basis while the monthly aggregated dataset only contains them on a per-day basis[1]. Hence, to achieve the desired granularity of requests per hour, we are forced to work with the daily dataset. The uncompressed CSV files are roughly 5 times larger in size based on our observations. From 2011-2015, there is only one file per day, from 2015-2020 there is a file with bot/spider traffic and one with user traffic and from 2020 onward the user traffic file is further split up into user and automated, the latter being suspicious user traffic classified as bot traffic.

Looking into the files, each row is supposed to contain 6 columns separated by spaces, which are, according to [9]:

1. **Wiki code:** The domain in the shape of `<subproject>.<project>` such as for example `en.wikipedia`.
2. **Page title:** The title of the accessed resource, e.g. "Spark".
3. **Page ID:** Some[2] numeric ID of the accessed resource.
4. **Device type:** A categorical variable that is either "desktop", "mobile-web" or "mobile-app".
5. **Daily total:** Sum of views per day.
6. **Hourly count:** Encoded hourly counts. The letter is used to indicate the hour of the day and the number is the amount of views. For example, "A23" represents 23 views between 00:00-00:59 ('hour 0', represented by the 1st letter of the alphabet, 'A').

There are, however, various inconsistencies and problems with the data. One of the most severe ones is that in some cases, in the same file, there are rows with 5 or 6 columns. This discrepancy stems from the fact that the page ID (column 3) can either be a number, NULL or not even present at all. This problem, though known by WMF, makes it impossible to open the files directly as tables, instead they have to be cleaned before usage. Various other inconsistencies exist as well, an example is shown in Figure 24 (in the Appendix).

4.1.1 Other datasets

Initially, we considered using other datasets provided by WMF, namely 'Unique devices' and 'Clickstream'. However, after later investigations we found that these small datasets could only marginally enrich the pageviews dataset, while requiring a considerable amount of work. Hence, we did not use these datasets and only considered the pageviews dataset for our analysis.

4.2 Data pipeline

In this section, we present the data pipeline we developed during this project. Figure 1 shows the individual steps of our pipeline from the data collection to the production of our data product. For each of the following sections, we have a folder in our repository (under `pipeline/`) with the code that was use to perform that step in the data pipeline.

4.2.1 Collection

As multiple groups needed access to the same raw data for their projects, we discussed with them the range of needed dates and ended up settling on the pageviews data from 2018 until 2021 (including both extremes). This would provide

[1]despite the documentation of `pagecount-ez` suggesting otherwise, which came to a surprise to us after inspecting the data

[2]we say some because there is no documentation on what these IDs represent, nor are they the same as the more commonly encountered Wikidata IDs documented here `https://en.wikipedia.org/wiki/Wikipedia:Finding_a_Wikidata_ID`

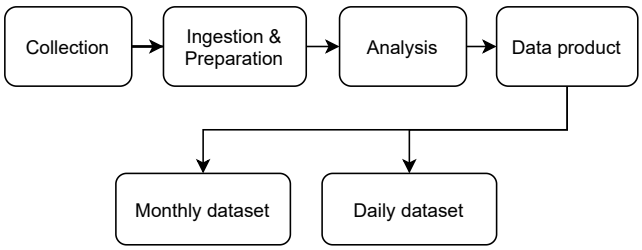

Figure 1: Data pipeline

other groups and us with enough data to do our analysis and more data could always be downloaded later on if necessary (which ended up not being the case).

The raw bz2-compressed pageviews data was downloaded from the `dumps.wikimedia.org` servers using a Python script that was parallelized with Spark running on Databricks. It does exponential backoff once the WMF servers apply rate limiting in order not to overwhelm their server, with a maximum waiting time of 5 minutes to avoid getting our Spark tasks killed. Furthermore, the script is resumable as it checks which files have already been downloaded and skips them if needed. This feature was necessary in order to keep track of the download progress and resume the download if either the script failed or the cluster crashed. The script also does a basic check once all files have been downloaded to make sure that the size of the downloaded files match the advertised size on the WMF website.

The download process took 4 days for the 1.6 TB, using up to 8 worker nodes on the shared Spark cluster[3].

4.2.2 Ingestion & Preparation

With the raw data available in the same S3 region, we could incrementally convert the needed months into a suitable format that would allow us to run the queries either for data analysis or for generating our data product.

Initially, we chose Parquet as our storage format where our Parquet dataset would contain the data bucketed by year, month and day. However, we quickly discovered that it is not possible to incrementally insert new data into an already existing Parquet dataset (as opposed to completely overwriting the dataset or appending to it). Furthermore, we discovered that bucketing the data by day prevented Parquet from applying its compression techniques effectively which made the daily buckets larger than the original bz2 compressed files.

As such, we switched to the Delta Lake format which uses Parquet internally. In addition to having ACID properties, it also supports upsert operations which allow us to insert into the Delta Table without adding duplicates. This is especially useful if our script or the cluster crashes. It also makes our script resumable. Furthermore, we can optimise file management of our table by utilizing z-ordering on the timestamp in order to reduce the number of Parquet files used by the Delta table as we are most interested in querying the data by date which in turn makes our queries faster.

In order to do this, we wrote a Scala program which utilises Spark on Databricks to convert the raw data to a Delta table by processing multiple files at once. Again, to make the script reliable, we read at most one month worth of data by loading in the individual (daily) text files for a particular month (in total around 37 GB for one month), apply data cleaning and insert the DataFrame to the Delta table with a `MERGE INTO` operation. Although our data is supposed to be in CSV format, due to some page titles being inconsistent with the format specified by WMF, we have to load them in as raw text files and match each line using a regular expression to the expected line format. Furthermore, the number of columns per line was also not consistent. Specifically, in some cases, the pageID was missing so we simply used -1 as the missing page id value. In addition to the schema shown in Section 4.1, we added one more columns, namely the traffic type (`user`, `spider` or `automated`) as well as expanding each row with the encoded hourly count into multiple rows with a timestamp that has the file's timestamp (month, day and year) as well as the decoded hour. The files representing the data for one month are roughly 13 GB which is almost 3x smaller as the bz2-compressed CSV files.

4.2.3 Analysis

With the data in an easy-to-query format, we are able to do further analysis. We wanted to understand the data from two different perspectives. Namely, from a domain-level point-of-view (POV) as well as a page-title POV. The domain-level POV analysis was due to the targeted nature of the DDoS attack that occured on September 6th, 2019 on WMF where mostly European subdomains were affected [7]. The page-title POV is useful to understand if there were any particular pages that were particularly hit by this attack. Both POVs of this known attack would help us identify if there are similar patterns for other dates as well. In order to do so, we compute two different metrics, namely the Kullback-Leibler (KL) divergence as well as the Shannon entropy, in order to detect anomalies in our daily data. Section 5 will go into more detail about how these metrics are computed.

To do so, we filter down the data using Spark on Databricks by aggregating or taking a subset of our data, so that it is feasible to run computations locally on our own machines with Pandas and visualise analytical results using Matplotlib as we are more familiar with these tools. In our experience, we found that Pandas works rather well for files around 1GB, any larger file sizes make the visualisations and computations rather slow. For instance, the graphs in this report were generated locally using the data filtered by Spark. In total, we generated 60GB that stored in S3, then queried and downloaded aggregated subsets to analyse locally.

4.2.4 Data product

Our data product is a Next.js[4]-based web application that shows two plots (created using the nivo[5] library) and allows to filter the data shown on the plots with the controls on the right. These controls dynamically update the plots on the left and allow the user to select the months to show in the top plot as well as which days for the selected months to show in the bottom plot. The top plot shows the number of total page views per day whereas the bottom plot shows the distribution of pages sorted by their number of views (in decreasing order). In the latter plot, a page represents a unique title encoded as a number. Furthermore,

[3]depending on the load of the cluster this could be less.

[4]`https://nextjs.org/`

[5]`https://nivo.rocks/`

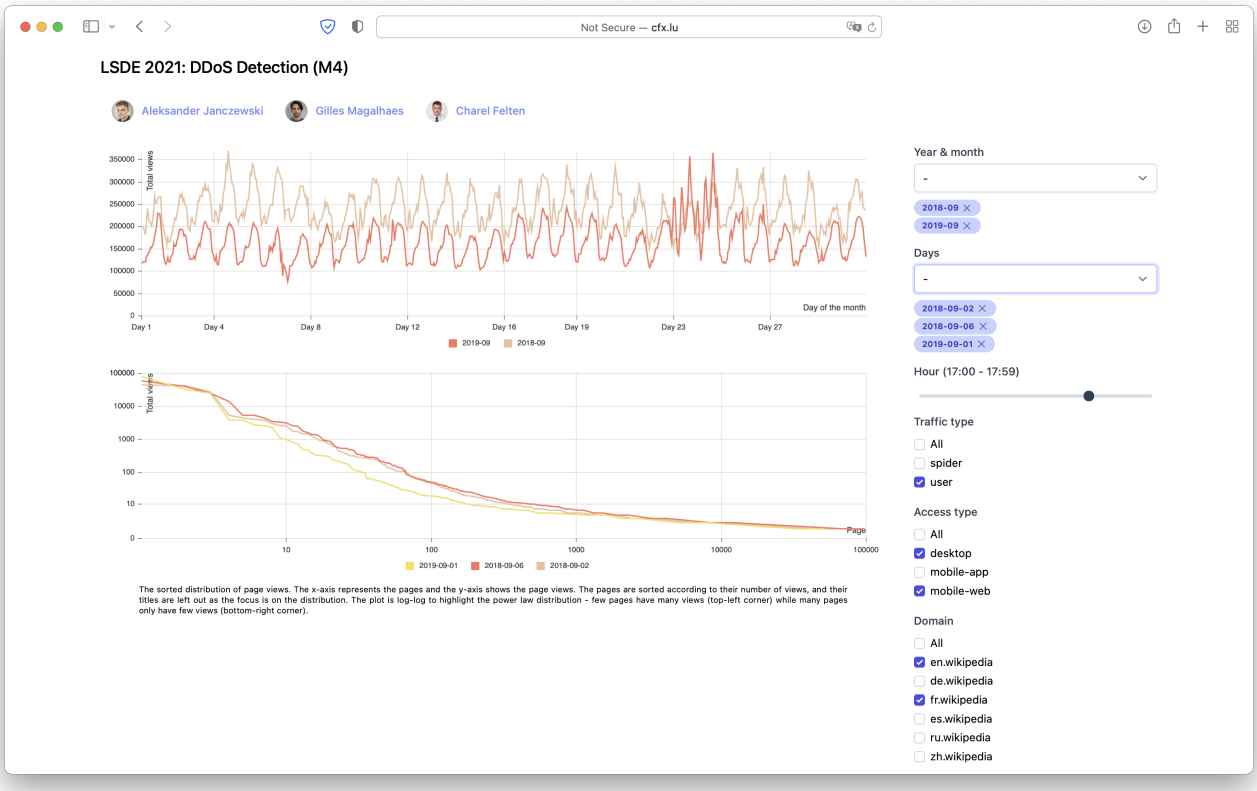

Figure 2: Project visualisation website

with the controls, the user can filter the page views by traffic type (`user`, `spider` or `automated`), access type (`desktop`, `mobile-app` or `mobile-web`) as well as from a list of 6 domains that we found to be relevant for showing the DDoS attack on WMF of September 6$^{\text{th}}$, 2019. These domains are {`en, de, fr, es, ru, zh`}`.wikipedia`. The latter are particularly interesting, because they represent the top 6 most visited domains of WMF, and the European domains were also particularly strongly affected by the DDoS attack of September 6$^{\text{th}}$, 2019. We also investigated the other domains, but did not find any particular outliers.

The same script that we used to generate the data for the plots in the report could also be used to generate the data for the website. We have included three months worth of transformed pageview data on the website which amount to a total of 31.7 MB. Again, this is partly done with Spark and partly with a script that runs locally to generate JSON files containing the plotting data in a directory hierarchy that can be traversed and read by our data product. The reason these files are so small is the following. For the monthly data, we only need to keep the number of page views per hour which for the three months worth of data are 24 (hours) $\cdot$ $(30 \cdot 2 + 31)$ (days) $\cdot 2$ (traffic types) $\cdot 3$ (access types) $\cdot 6$ (domains) $=$ 78624 entries. The monthly data amounts to $\sim$ 2.3 MB. As for the daily data, for each of the previous entries we have a file with $\sim$ 20 entries per hour, so around 480 entries per day which in total amounts to $\sim$ 29 MB.

Figure 2 shows the final version of our website.

5. METHODS

As previously mentioned, our extensive literature review has not allowed us to identify any particular DDoS detection methods that would be applicable to our data. However, we have identified two particularly interesting methods to quantify the differences between packet size distributions, which we applied to our pageviews data.

5.1 Shannon Entropy

Shannon entropy is the fundamental information-theoretic metric introduced by Claude Shannon in his landmark paper "A Mathematical Theory of Communication" [13]. Shannon entropy, intuitively, is a measure of average uncertainty in the random variable [3], and its discrete version is defined in the following manner:

$$H(X) = -\sum_{x \in X} p(x) \log_b p(x) \tag{1}$$

where $p(x)$ stands for the probability mass function of variable $x \in X$. The base b of the logarithms establishes the unit for the entropy (for example, for $b = 2$, the entropy is expressed in bits). In this report, we use logarithmic base of $b = 2$, hence all our information-theoretic metrics are expressed in bits.

5.2 Kullback-Leibler Divergence

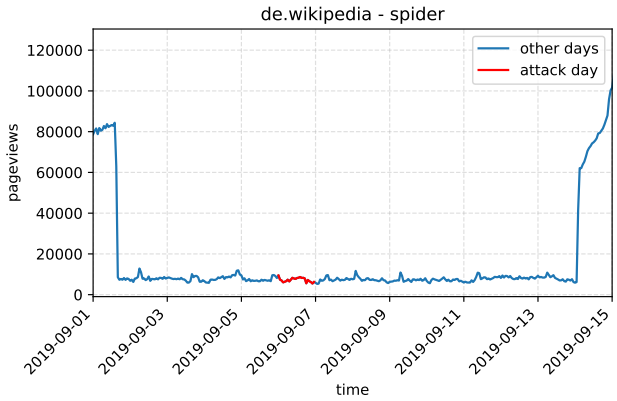

Figure 3: Hourly pageviews over the first half of 2019-09 for all spider traffic on "de.wikipedia".

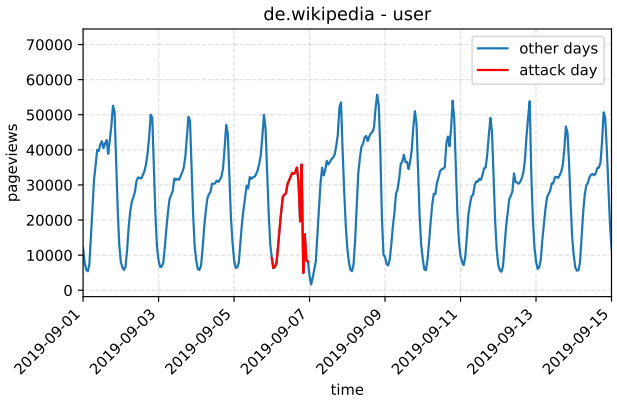

Figure 4: Hourly pageviews over the first half of 2019-09 for all user traffic on "de.wikipedia".

The Kullback-Leibler (KL) divergence also known as relative entropy, is a measure of distance between two probability distributions [3]. This information-theoretic metric for discrete probability distributions p and q is defined as:

$$D(p||q) = \sum_{x \in X} p(x) \log_b \frac{p(x)}{q(x)} \tag{2}$$

where p is the probability distribution of interest, whereas q corresponds to the reference point probability distribution. It is important to note two main properties of this metric. First, the KL divergence is always non-negative, i.e. $D(p||q) \geq 0$ and $D(p||q) = 0 \iff p = q$ [3]. Secondly, the metric is not symmetric, i.e. $(D(p||q) \neq D(q||p))$. Therefore, intuitively, the KL divergence is a directional measure of how probability distribution of interest p is different from the reference probability distribution q. There are many other information-theoretic based interpretations of this metric that we do not discuss, since further discussion of information theory is out of the scope of this report. One of the main drawbacks of KL divergence (in our application of this metric) is the fact that if there is any even $x \in \chi$ such that $p(x) > 0$ and $q(x) = 0$, then $D(p||q) = \infty$.

5.3 Binning Method

Binning method is a standard method used to estimate a probability mass function of a continuous random variable. In essence, binning consists of defining fixed ranges where values can fall into, alike bins, hence the name. As a result, any arbitrarily long sequence of values can be organised into a fixed set of bins. A more detailed application of the binning method in our investigation is further explained in the following section.

6. RESULTS

6.1 Qualitative Investigation

We started our investigation with a brief qualitative investigation of the total number of pageviews per hour over half a month.

Figure 3 and Figure 4 show the user and spider traffic of 2019-09-01 until 2019-09-15. Spider traffic on the day of the attack does not look visibly different to other days surrounding these days. For user traffic, the start of the day looks similar to all other days, however, all other days contain a spike towards the end of the day, which is not present on the day of the attack, most likely due to the DDoS-caused outage. This qualitative analysis just shows that there is no visible increase in traffic at the time of the attack, at least on "de.wikipedia", but we found a similar pattern for other domains. However, we will verify this claim with a more in-depth quantitative analysis.

6.2 Quantitative Investigation

In order to determine whether it is possible to detect DDoS attacks using the pageviews data, and investigate our hypotheses, we perform a quantitative investigation focusing on the days starting from August 1st, 2019 until October 12th, 2019, using the two metrics previously introduced in Section 5. The pageview probability distribution representing the probability of page x having y number of views on a particular domain at a specific hour of the day are produced using the method clarified in the following section.

6.2.1 Binning method to activity profile

In order to better illustrate the method that we use to produce the activity probability distribution, let us assume that the domain "de.wikipedia" has 5 pages ("titles"): A,B,C,D,E . Let us further assume that on January 1st, 2001 at 12pm the number of times each of these five pages were accessed are as follows: A-6, B-4, C-1, D-1, E-1. Now, if we use the following 5 bins: $[6-5]$, $[4-3]$, $[3-2]$, $[2-1]$, we can construct a discrete probability distribution of pages having x number of views. As follows, the probability that a page on "de.wikipedia" domain has 6 to 5 views ($[6-5]$) is $1/5$ (since out of 5 pages only page A had between 6 to 5 views). Analogously, the probability that a page has between 2 to 1 views ($[2-1]$) is $3/5$. In this manner, we produce a probability distribution (activity profile) conveniently representing the activity on a given domain at a particular hour of the day in terms of probabilities. We will further refer to this probability distribution as the activity profile through the remainder of the report.

To illustrate further, on Figure 5 the red plot represents the sorted distribution of page views on the "de.wikipedia" domain on September 6th, 2019 whereas on Figure 6 the

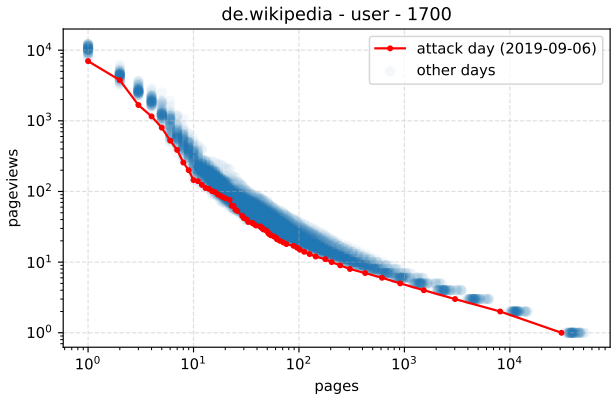

Figure 5: The sorted distribution of page views on "de.wikipedia" accessed by users on 2019-09-06 at 17:00 UTC (red). The pages are sorted according to their number of views, and their titles are left out as the focus is on the distribution. In blue, the sorted distributions of page views on "de.wikipedia" accessed by users from 74 sampled days. The plot is log-log to highlight the power law distribution. The red dots signal the last page with y amount of views. Hence, there are few red points at the beginning due to the logarithmic axis, and few red points at the end as there are thousands of pages having exactly 2 or exactly 1 views, and the dot only represents the last such page.

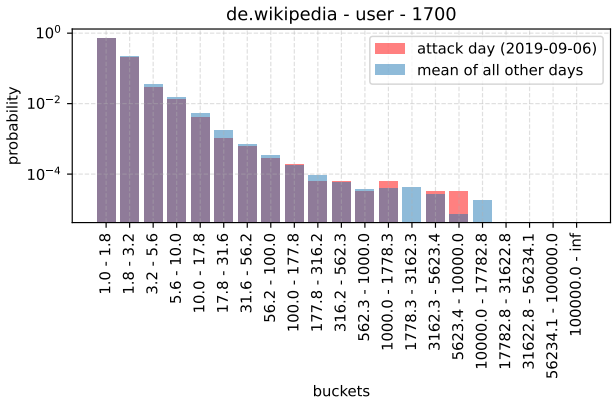

Figure 6: Histogram representing the activity profile of users on the day of the DDoS attack (in red). In blue, the reference activity profile obtained from determining the mean activity profile from a sample of 74 days is shown.

red histogram represents its corresponding activity profile produced with the binning method described in this section.

6.2.2 *Reference activity profile*

KL divergence requires a point-of-reference probability distribution. Therefore, it is necessary to determine an average activity profile (baseline activity) for each domain at a particular time in the day. To produce the reference probability distribution, we first sample 74 days from 2019 (excluding the days of the DDoS attack), then using 20 evenly (in a logarithmic space) spaced bins in range $\left[10^0, 10^5\right]$, we construct the activity profile for each domain per each hour of the day using the method illustrated in Section 6.2.1. The use of evenly-spaced bins in logarithmic space is motivated by the fact that the number of views per page on a domain follow a power law. Next, for each hour of the day, we calculate the average probability of a page having a particular number of views at a particular hour of the day. This represents our reference activity profile (baseline).

To illustrate further, on Figure 5 the blue plot represents the reference distribution of page views on "de.wikipedia" domain produced from a sample of 74 days, whereas on Figure 6 the blue histogram represents its corresponding reference activity profile.

6.2.3 *KL Divergence*

Having established our activity profiles and reference activity profiles, we can now calculate KL divergences. For each hour of each day starting from August 1$^{\text{st}}$, 2019 until October 12$^{\text{th}}$, 2019 we calculate the KL divergence of the spider and user activity profiles as compared to their respective reference activity profiles. The KL divergences of the user and spider activity profiles calculated for each hour of the day are presented in Figures 7 and 8.

On Figures 7 and 8 multiple interesting insights can be observed. First of all, from Figure 7 it appears that the KL

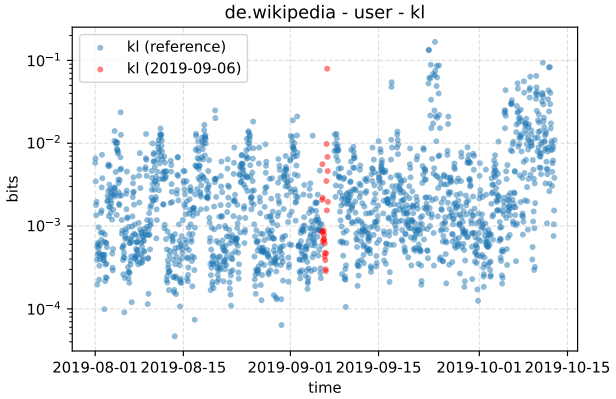

Figure 7: The time evolution of KL divergence of user activity profiles on the "de.wikipedia" domain from 2019-08-01 until 2019-10-12. The values in red represent the DDoS attack day.

divergence of the user activity profiles on the day of attack (2019-09-06 indicated in red) largely falls in the range between 10^{-4} and 10^{-2} bits (alike majority of the computed values). Hence, at first sight, these values do not appear to present any abnormalities in the user activity. However, upon closer inspection one can notice that there is one particular KL divergence (red) value that is significantly further away from the cluster of the rest of the values.

Figure 8 presents the KL divergence of the spider activity profile as compared to its respective reference activity profile. Most of the KL divergence values computed fall in the range between 10^{-2} and 10^{-1} bits. However, the plot also reveals some seasonal spikes in KL divergence of the spider activity profiles. Closer qualitative inspection of the data revealed that these spikes are associated with the crawler activity that significantly disrupts a regular activity of on the domain and hence we observe significant deviations from the reference activity profile. Apart from this particular insight, visual inspection does not reveal any abnormalities in the spider activity profile on the day of the DDoS attack.

In order to quantitatively determine if there is any statistically meaningful abnormal activity on the day of DDoS

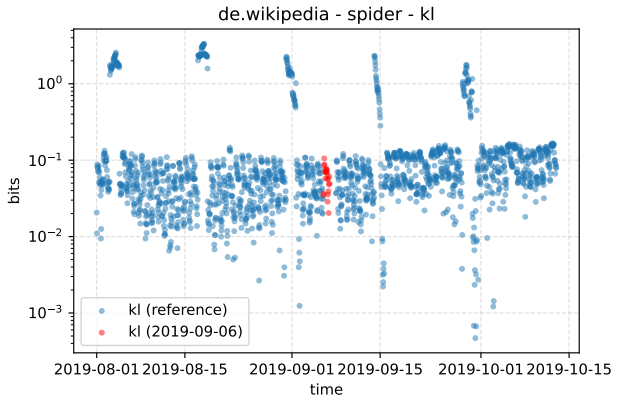

Figure 8: The time evolution of KL divergence of the spider activity profiles on the "de.wikipedia" domain from 2019-08-01 until 2019-10-12. The values in red represent the DDoS attack day.

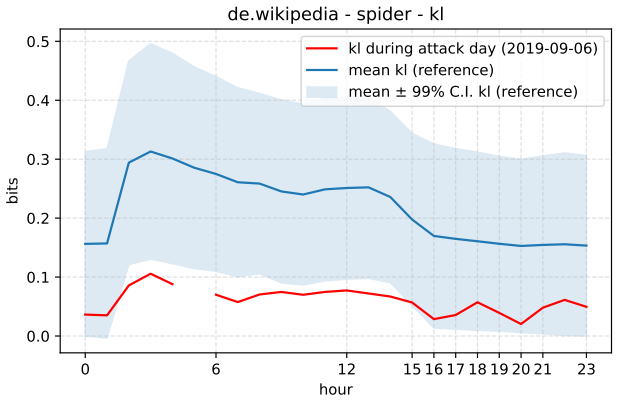

Figure 10: KL divergence of the spider activity profile on the "de.wikipedia" domain on the day of DDoS attack (red). In blue, the mean KL divergence calculated per hour using a sample of 74 days.

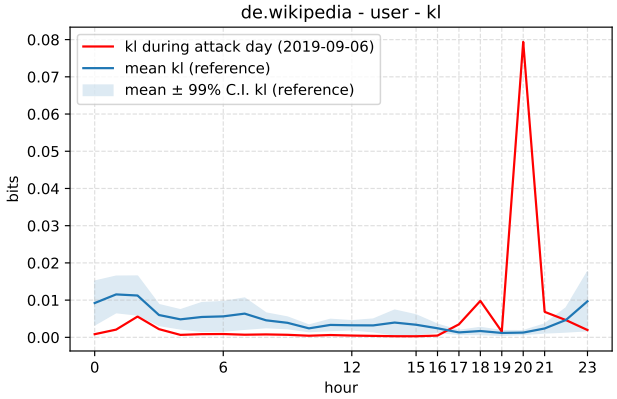

Figure 9: KL divergence of the user activity profile on the "de.wikipedia" domain on the day of DDoS attack (red). In blue, the mean KL divergence calculated per hour using a sample of 74 days.

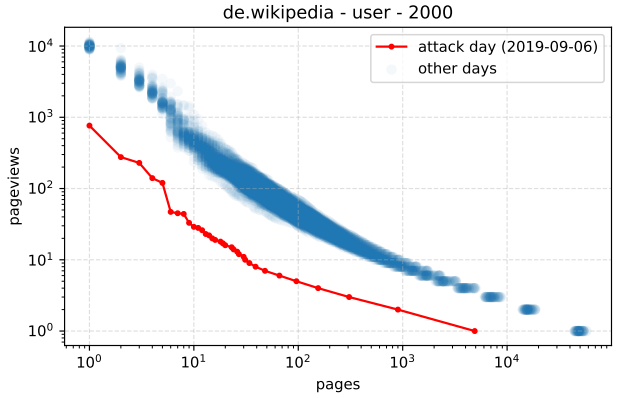

Figure 11: The sorted distribution of page views on "de.wikipedia" accessed by users on 2019-09-06 at 20:00 UTC.

attack, we calculate the mean KL divergence of all 74 days for each particular hour. This yields 24 mean KL divergences, each associated with each hour of the day. Next, we plot the mean KL divergence with their respective 99% confidence intervals, along with the KL divergence of spider and user activity profiles calculated for the day of DDoS attack. In this manner, Figures 9 and 10 are generated.

Figure 9 reveals the KL divergence calculated for the user activity profile on the day of DDoS attack (in red) and the mean KL divergences computed for the user activity profiles from the same hours of all the other 74 days (in blue). The abnormal KL divergence value that was previously identified on Figure 7 is now very clearly depicted on Figure 9 as a spike in KL divergence of the user activity profile on the day of DDoS attack at 20:00 UTC. Closer inspection of the data reveals that the spike is a consequence of a significant drop in the user activity at around 20:00 UTC, which is further depicted on Figure 11. Additionally, it can be observed that the KL divergence significantly falls out of the 99% confidence interval starting from 17:00 UTC until around 21:00 UTC. This finding aligns with the fact that the DDoS attack targeted at WMF started around 17:00 UTC.

Furthermore, Figure 10 reveals that the KL divergence of spider activity on the day of the attack appears to be significantly lower than the mean between 2:00 and 15:00 UTC. This indicates that the activity of spider on the day of attack was simply very similar (and even more similar than on average) to the standard spider activity. On this plot, one can also notice one missing value which is the consequence of both our discretization process and the KL divergence metric drawback previously mentioned in Section 5.2. Thus, this missing value is not indicative of any particular anomaly and should simply be ignored.

6.2.4 *Entropy*

An analogous investigation to the one presented in Section 6.2.3 was also performed using the Shannon entropy. This time, however, the metric itself did not require any point of reference distributions. Therefore, in this section, we simply compare the Shannon entropies of the user and spider activity profiles to their respective average values (obtained from the same sample of 74 days). First, for each hour of each day starting from August 1st, 2019 until October 12th, 2019 we calculate the Shannon entropy of the

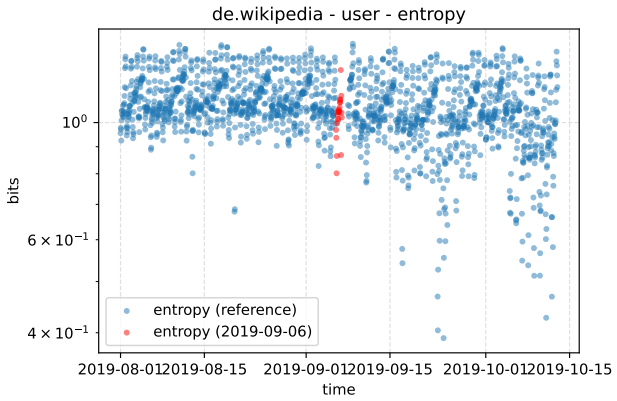

Figure 12: The time evolution of Shannon entropy of the user activity profiles from 2019-08-01 until 2019-10-15. In red, the Shannon entropy of the user activity profiles from the day of the DDoS attack.

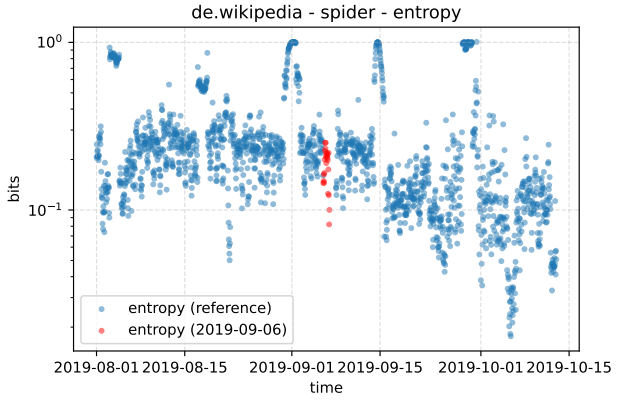

Figure 13: The time evolution of Shannon entropy of the spider activity profiles from 2019-08-01 until 2019-10-15. In red, the Shannon entropy of the spider activity profiles from the day of the DDoS attack.

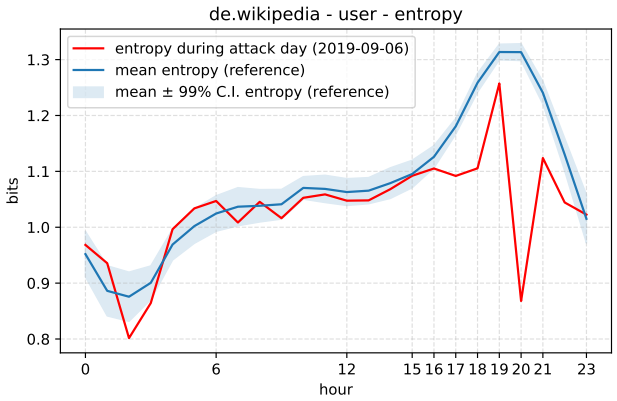

Figure 14: Shannon entropy of the user activity profile on the "de.wikipedia" domain on the day of the DDoS attack and average Shannon entropy of the reference user activity profile computed from a sample of 74 days.

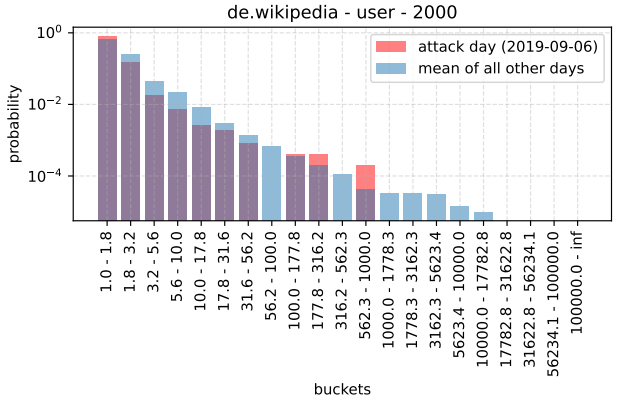

Figure 15: Histogram representing the activity profile of users at 20:00 UTC on the day of the DDoS attack (in red). In blue, the reference activity profile obtained from determining the mean activity profile from a sample of 74 days is shown.

spider and user activity profiles for each domain. The time evolution of Shannon entropy of user and spider activity profiles are presented on Figures 12 and 13.

Figure 12 reveals that most of the Shannon entropy values from the day of the DDoS attack are clustered around 1 bit. However, it can also be noticed that there are 3 (red) values that are particularly low. Besides the fact that Shannon entropies on the day of the DDoS attack appear to be more clustered than on any other day, general visual inspection does not reveal any other anomalies.

Similarly, on Figure 13 no clearly abnormal Shannon entropies are observed for the spider activity profiles. Again, in order to draw statistically significant conclusions, we calculate the mean Shannon entropy for each particular hour, using a sample of 74 days. This yields 24 mean Shannon entropies, each associated with a particular hour of the day. Next, we compare mean Shannon entropies calculated for each hour with the ones calculated for each hour of the day when the DDoS attack took place.

Figure 14 reveals that the Shannon entropy of the user activity profile from the day of the DDoS attack, significantly deviates from the mean starting from 17:00 UTC until around 23:00 UTC. Specifically, it is apparent that there is a large drop in the Shannon entropy at 20:00 UTC which aligns with the findings from Section 6.2.3 and Figure 11.

Intuitively, a decrease in entropy can be associated with a decrease in the uncertainty about how many views a page on the domain has. Therefore, a drop in the Shannon entropy is indicative of an activity profile being truncated. This is in fact the case as depicted on Figure 15 where it can be observed that there are no pages with over 1000 views. Whereas on average at this time of the day, there are pages that have between 10,000 and 17,782 views.

Figure 16 reveals significant deviations of the spider activity profile Shannon entropy at 18:00 UTC and 22:00 UTC. However, closer inspection of the sorted distributions of page views (Figures 17 and 18) also reveals only a slight decrease in the activity on the domain.

7. CONCLUSION

Before we conclude, it is necessary to recall the research question and the main assumptions of this research project.

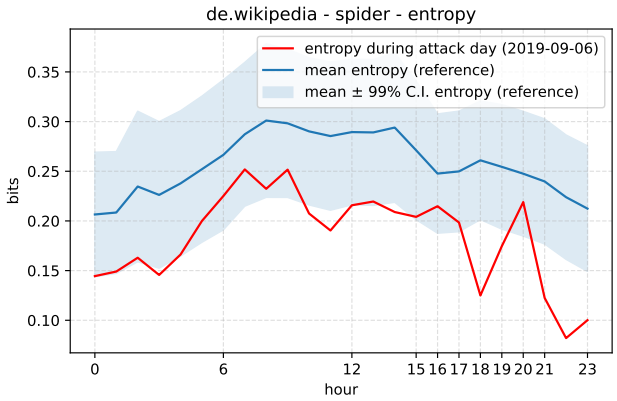

Figure 16: Shannon entropy of the spider activity profile on the "de.wikipedia" domain on the day of the DDoS attack and the average Shannon entropy of the reference spider activity profile computed from a sample of 74 days.

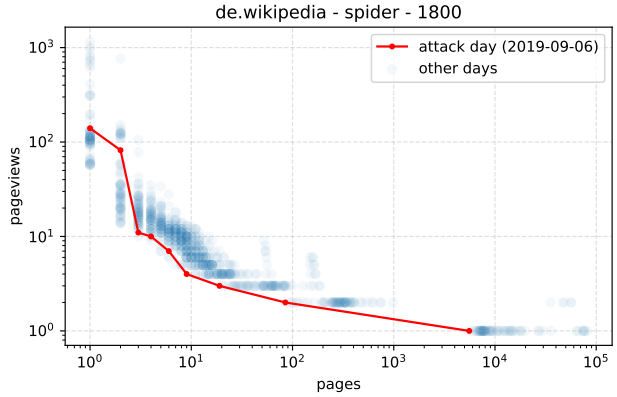

Figure 17: The sorted distribution of page views on "de.wikipedia" accessed by spiders on 2019-09-06 at 18:00 UTC (red). In blue, the sorted distributions of page views on "de.wikipedia" accessed by spiders from 74 sampled days. The plot is log-log to highlight the power law distribution.

The aim of this project was to determine if it is feasible to detect DDoS attacks with the limited data – such as the pageviews data. The underlying assumption of our investigation was that if it is possible to detect DDoS attacks within the pageviews dataset using KL-divergence and Shannon entropy metrics, then we should be able to identify an anomaly in the pageviews activity on the day of DDoS attack that we know of. Finally, we know of the DDoS attack targeted at WMF taking place at 17:00 UTC on the September 6$^{\text{th}}$, 2019.

Our investigation clearly revealed anomalies on the day of the DDoS attack both qualitatively and quantitatively. The anomalies were identified in the user activity at 20:00 UTC on four European domains ("de.wikipedia", "es.wikipedia", "fr.wikipedia", "en.wikipedia"), and a Russian domain "ru.wikipedia" (See Figures 19 to 22 in the Appendix). For comparison, one can see that no anomalies were quantitatively identified for the Chinese domain (see Figure 23 in the Appendix) which was qualitatively verified as well.

KL-divergence uncovered very particular irregularities in

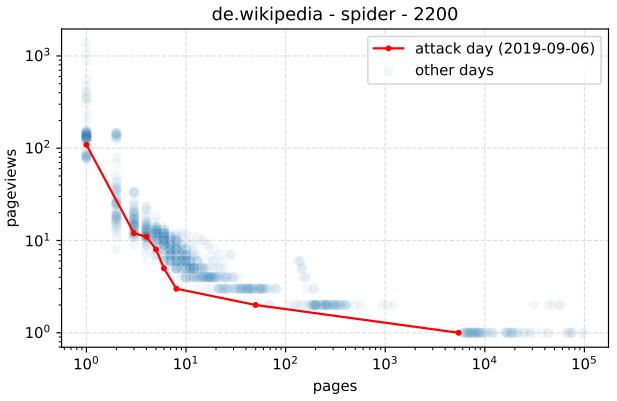

Figure 18: The sorted distribution of page views on "de.wikipedia" accessed by spiders on 2019-09-06 at 22:00 UTC (red). In blue, the sorted distributions of page views on "de.wikipedia" accessed by spiders from 74 sampled days. The plot is log-log to highlight the power law distribution.

the user activity in the European and Russian domains at 20:00 UTC on September 6th, 2019. A very noticeable increase in KL divergence was observed on the day of the DDoS attack, as we hypothesized, however the source of that divergence was different than expected. While we expected to observe an increase in activity on the domain, representing the increase in page views triggered by the large influx of requests sent by a DDoS attack, qualitative investigation revealed that the spike in the KL divergence metric was the consequence of a decline in the user activity on domains due to the outage caused by the DDoS attack. This further aligns with the timeline, since the irregularity was detected at 20:00 UTC, i.e. 3 hours after the apparent start of the DDoS attack at $\sim$ 17:00UTC. Besides that particular irregularity, no other deviations in user or spider activity were detected using the KL divergence.

Moreover, Shannon entropy also exposed the exact same irregularity in the user activity at 20:00 UTC on September 6th, 2019. Since we expected to observe a spike in the activity on the domain, the Shannon entropy was hypothesised to be higher than on any other day. However, since the qualitative investigation revealed an actual decline in the activity on the domain, instead of observing an increase in Shannon entropy, a significant decrease was observed. Analogously to the KL divergence metric, Shannon entropy did not reveal any other anomalies in user or spider activities at the time of the attack.

It is important to note that both quantitative metrics aligned with each other's findings and were confirmed qualitatively. Although, we have not detected any spike in the activity at the time of the attack on any domain, the metrics were able detect a decline in the user activity 3 hours after the apparent start of the attack, which we can associate with the outage of these domains. Therefore, our investigation demonstrated that both the KL divergence and Shannon entropy are capable of detecting anomalies in the activity on the domains using just the pageviews data. However, we have not been able to determine whether these metrics can be used to detect a DDoS attack, simply because we have not found any artefacts from the attack. There are some potential explanations for this.

First of all, we do not know if the data in fact contains the logs of the traffic associated with the attack. It could be that the WMF may have filtered these pageviews and removed the traffic associated with the attack, as they may have considered this traffic not to be interesting for the kind of research they expect this data to be used in.

Another reason could be that the traffic loggers themselves were also targeted by the attack and the data they logged was simply lost and never made it into the dataset.

It is also possible that the requests of the attack were not really page view requests, they could just have been 'SYN' messages to initiate a TCP connection to overload the servers. In that case, the DDoS traffic would not be logged by the pageviews logger as these requests were technically not pageviews.

Lastly, it is always possible that both our qualitative and quantitative analysis did not detect the DDoS attack. However, given the breadth and depth of our investigation, as well as the sheer number of page requests needed to bring down the servers of one the biggest websites worldwide make this explanation very unlikely.

In the end, we were able to detect the outage caused by the DDoS attack, but unable to detect the attack itself. Unfortunately, this means that we are neither able to search for other DDoS attacks in the data, nor would we be able to differentiate between a 'normal' outage and a maliciously caused outage.

8. CONTRIBUTIONS

In Tables 1 to 3, we show the main contributors of each part of the project. Each one of us was involved in all parts of the project, but the main contributors were the ones that were responsible for getting that part of the project completed.

Project part	**Main contributor(s)**
Literature research	Alex
Report writing	Alex & Gilles
Report plots	Charel
Presentation slides	Charel

Table 1: Report & presentation contributions

Project part	**Main contributor(s)**
Data collection	Gilles
Data preparation & cleaning	Gilles & Charel
Qualitative data analysis	Charel & Alex
Quantitative data analysis	Alex
Preparing code for submission	Gilles

Table 2: Code contributions

Project part	**Main contributor(s)**
Website	Gilles
Visualisation data	Charel

Table 3: Visualisation contributions

9. REFERENCES

[1] S. Behal, K. Kumar, and M. Sachdeva. Characterizing ddos attacks and flash events: Review, research gaps and future directions. *Computer Science Review*, 25:101–114, 2017.

[2] M. H. Bhuyan, D. Bhattacharyya, and J. K. Kalita. An empirical evaluation of information metrics for low-rate and high-rate ddos attack detection. *Pattern Recognition Letters*, 51:1–7, 2015.

[3] T. M. Cover. *Elements of information theory*. John Wiley & Sons, 1999.

[4] J. David and C. Thomas. Discriminating flash crowds from ddos attacks using efficient thresholding algorithm. *Journal of Parallel and Distributed Computing*, 152:79–87, 2021.

[5] J. E. Dunn. Wikipedia fights off huge ddos attack, Sep 2019.

[6] A. Gutnikov Oleg Kupreev Gutnikov, O. Kupreev, and E. Badovskaya. Ddos attacks in q1 2021, May 2021.

[7] A. Henthorn-Iwane. Analyzing the wikipedia ddos attack.

[8] Kaspersky. Summer plateau: less ddos attacks and small geographic shifts in q2 2021, Jul 2021.

[9] Wikistats: Pageview complete dumps, Sep 2019.

[10] J. Quittner. Panix attack. *Time*, 148(16):64–64, 1996.

[11] A. Rai and R. K. Challa. Survey on recent ddos mitigation techniques and comparative analysis. In *2016 Second International Conference on Computational Intelligence & Communication Technology (CICT)*, pages 96–101. IEEE, 2016.

[12] P. S. Saini, S. Behal, and S. Bhatia. Detection of ddos attacks using machine learning algorithms. In *2020 7th International Conference on Computing for Sustainable Global Development (INDIACom)*, pages 16–21. IEEE, 2020.

[13] C. E. Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

[14] S. Umarani and D. Sharmila. Predicting application layer ddos attacks using machine learning algorithms. *International Journal of Computer and Systems Engineering*, 8(10):1912–1917, 2015.

[15] Y. Xiang, K. Li, and W. Zhou. Low-rate ddos attacks detection and traceback by using new information metrics. *IEEE transactions on information forensics and security*, 6(2):426–437, 2011.

[16] S. Yu and W. Zhou. Entropy-based collaborative detection of ddos attacks on community networks. In *2008 Sixth Annual IEEE International Conference on Pervasive Computing and Communications (PerCom)*, pages 566–571. IEEE, 2008.

[17] S. Yu, W. Zhou, R. Doss, and W. Jia. Traceback of ddos attacks using entropy variations. *IEEE transactions on parallel and distributed systems*, 22(3):412–425, 2010.

[18] X. Yuan, C. Li, and X. Li. Deepdefense: identifying ddos attack via deep learning. In *2017 IEEE International Conference on Smart Computing (SMARTCOMP)*, pages 1–8. IEEE, 2017.

APPENDIX

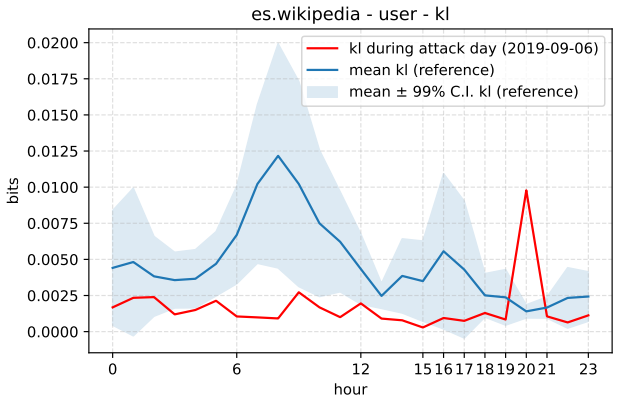

Figure 19: KL divergence of user activity profile on the "es.wikipedia" domain on the day of DDoS attack, as compared to the reference user activity profile.

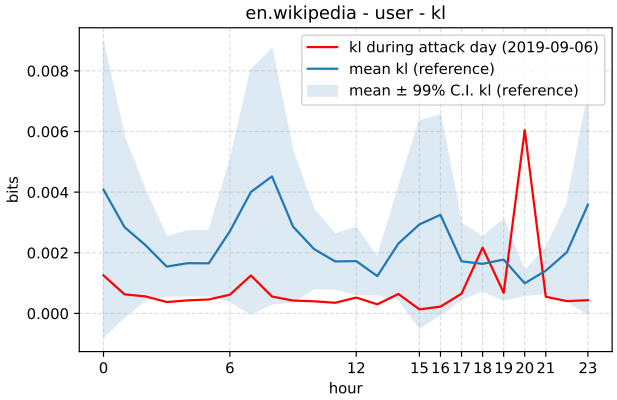

Figure 20: KL divergence of user activity profile on the "en.wikipedia" domain on the day of DDoS attack, as compared to the reference user activity profile.

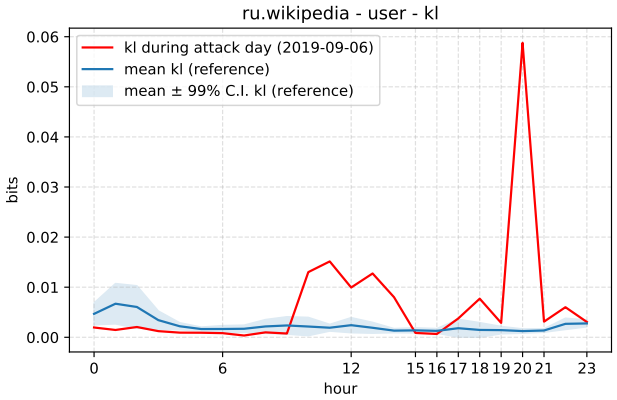

Figure 21: KL divergence of user activity profile on the "ru.wikipedia" domain on the day of DDoS attack, as compared to the reference user activity profile.

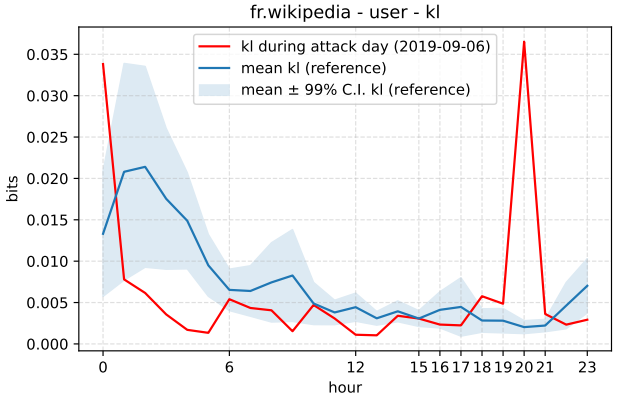

Figure 22: KL divergence of user activity profile on the "fr.wikipedia" domain on the day of DDoS attack, as compared to the reference user activity profile.

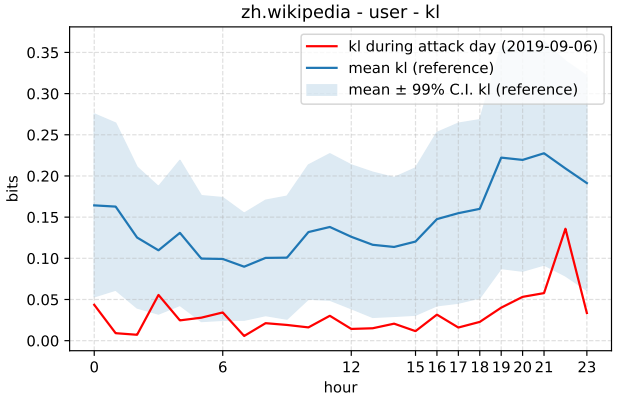

Figure 23: KL divergence of user activity profile on the "zh.wikipedia" domain on the day of DDoS attack, as compared to the reference user activity profile.

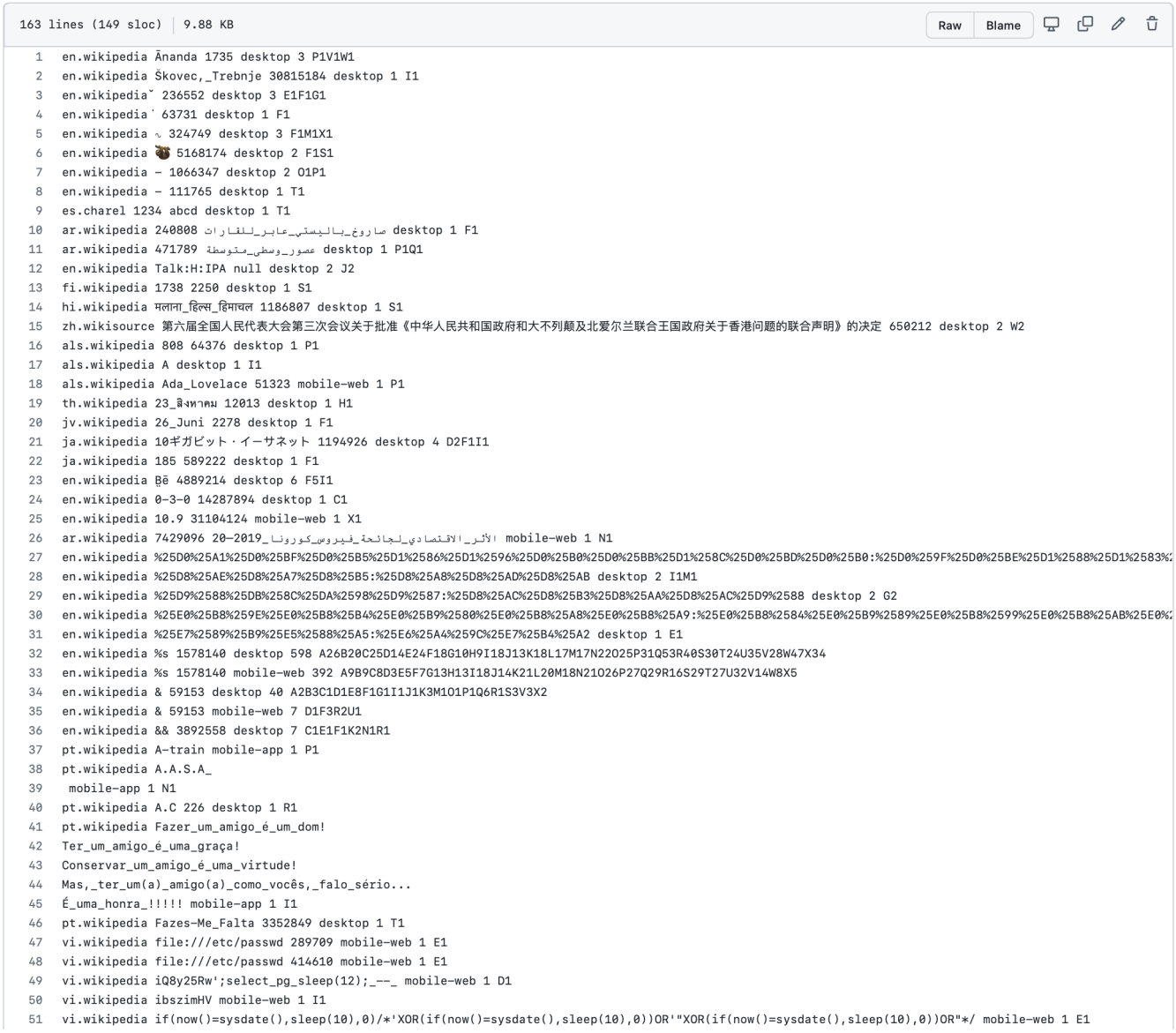

Figure 24: Some lines of pageviews data. Various unicode symbols encountered in the page title column that may cause weird behaviour when rendering, such as consuming preceeding white space (l3,4) or changing the rendering order of title and ID (l10,11). Page titles may not be unique (l7-8,l47-48) and IDs may be null (l12) or omitted (l19). There may also be newline characters in page titles (l38-39, l41-45). Interestingly we also found remnants of either vulnerability testing or injection attacks (l47-49, l51).