Time Pattern Analysis of Malware by Circular Statistics

Liuxuan Pan
Royal Holloway, University of London
Egham Hill, Egham
Surrey, UK
liuxuan.pan.2013@live.rhul.ac.uk

Allan Tomlinson
Royal Holloway, University of London
Egham Hill, Egham
Surrey, UK
allan.tomlinson@rhul.ac.uk

Alexey A. Koloydenko
Royal Holloway, University of London
Egham Hill, Egham
Surrey, UK
alexey.koloydenko@rhul.ac.uk

ABSTRACT

Circular statistics present a new technique to analyse the time patterns of events in the field of cyber security. We apply this technique to analyse incidents of malware infections detected by network monitoring. In particular we are interested in the daily and weekly variations of these events.

Based on "live" data provided by Spamhaus, we examine the hypothesis that attacks on four countries are distributed uniformly over 24 hours. Specifically, we use Rayleigh and Watson tests. While our results are mainly exploratory, we are able to demonstrate that the attacks are not uniformly distributed, nor do they follow a Poisson distribution as reported in other research. Our objective in this is to identify a distribution that can be used to establish risk metrics.

Moreover, our approach provides a visual overview of the time patterns' variation, indicating when attacks are most likely. This will assist decision makers in cyber security to allocate resources or estimate the cost of system monitoring during high risk periods.

Our results also reveal that the time patterns are influenced by the total number of attacks. Networks subject to a large volume of attacks exhibit bimodality while one case, where attacks were at relatively lower rate, showed a multimodal daily variation.

Keywords

Circular statistics; malware; time patterns; uniformity hypothesis test

1. INTRODUCTION

Circular statistics have been applied to analyse time patterns in diverse areas such as public disorder, wind direction, and the turning patterns of elephant movement [3, 12, 13]. Brunsdon and Corcoran show how this technique may be applied to the time pattern analysis of certain types of public disorder and thus assist the police in prioritising resources and improving targeting [3]. Our aim is similar but within the field of network monitoring and cyber security.

To our knowledge, circular statistics are rarely used in cyber security research. Analysis usually focuses on linear statistics applied to the distribution of malware over IP address space [14,16]. Analysing time patterns events by circular statistics is a new approach, providing visual methods to identify the distribution of attacks over a fixed time period.

In the following, we restrict our notion of an "attack" to an automated opportunistic attack on a network or host and we consider the distribution of such attacks over fixed time periods. Dedicated targeted attacks are beyond the scope of this work.

A typical opportunistic attack is an infection by malware. In studies of malware, it is conventional to use standard histograms over the entire time course of data and view the distribution of attacks over long periods such as several months or years. By focusing on linear distributions over long time periods in their statistical analysis, most studies of malware do not consider the viewpoint of time-of-day or day-of-week relating to cyber risk.

In practice, our observations show that attack data appear to be distributed non-uniformly over 24-hour periods. For instance, the frequency of malware communicating with a sinkhole may appear to follow a Poisson distribution from 01:00 to 15:00 (the results come from section 3.2.1), but a different distribution outside these hours. Maillart and Sornette consider the cyber-risk of personal identity losses to follow a power-law tail distribution, which is related to the size of organizations [10]. But in their study, the exact time of the identity theft is not considered when quantifying the distribution.

Our argument is that certain types of attacks are more likely to occur at different times. Our goal in this work is to be able to characterise distributions of attacks on enterprise networks over fixed time periods. If we can do this we will be in a better position to carry out a risk assessment for such networks. Current approaches to risk assessment either classify the likelihood of attack as, for example, low/medium/high; or use probability figures estimated by experienced security professionals. Thus the ability to provide such estimates based on the identification of a known probability distribution will improve the overall risk assessment.

Moreover, by identifying distributions and visualising time patterns we can help security managers to allocate and adjust monitoring resources and firewall and intrusion detection rules to target malware.

This paper will study the time patterns of events caused by malware and analyse the daily and weekly variation by

978-1-5090-6386-4/17 $31.00 © 2017 IEEE
DOI 10.1109/ANCS.2017.26

applying circular statistical methods. Our approach is to describe the time variation patterns of malware events on a circle. We base our analysis on data provided by The Spamhaus Project[1]. Spamhaus has a long history of providing network monitoring and traffic analysis services and has a vast collection of data which is updated in real time. We used a dataset based on around 1000 users from commercial (business) organizations. The complete dataset was just over 1GB and contained around 9 million records, each one containing a source and destination IP address, time of observation, and type of traffic.

To narrow things down to a manageable size, we chose to focus on the conficker virus [6] and corresponding botnet traffic, and in particular traffic sending data to a sinkhole. Statistical analysis of conficker was presented by Shin et al. [16] but they looked at the distribution over the IP address space and domain name by linear statistics and did not investigate the distribution of events around the clock.

For the dataset selected, we test a uniformity hypothesis about the malware activity, and potential daily and weekly relationships. Rayleigh and Watson's tests are two common methods to test the uniformity hypothesis in circular statistics [13]. These tests help us identify time patterns of the conficker attacks around the 24 hour clock. Furthermore, we use the Mardia-Watson-Wheeler test [13] to compare distributions of two different botnets.

In the following, we demonstrate the time variations in the malware attacks using three different types of graphs: ordinary histograms, rose diagrams and helix graphs. In addition to the above uniformity tests and visualisation, we analyse the conficker daily and weekly cycles by country. We conclude our analysis by comparing the time patters between confiker and the other botnets and test whether these patterns have a common distribution using the Mardia-Watson-Wheeler test. The combination of analyses provide us with an overview of the time pattern variations under different conditions or sub-datasets.

The rest of the paper is organized as follows. Section 2 briefly introduces the background. Section 3 presents the theory of circular statistics. In section 4, we will discuss the source of the data and how we extract relative data as the analysis basis of this paper. Section 5 analyses the time patterns of conficker in four countries using daily circle models, helix diagrams, uniformity hypothesis testing and weekly circles. Section 6 discusses the time distributions of conficker and other botnets in the same domain and uses the Mardia-Watson-Wheeler test to test the hypothesis of a common distribution between two botnets. Finally, we conclude our findings in section 7.

2. BACKGROUND

The data we analyse is captured as IP packets enter and exit monitoring points (taps) on networks. The data recorded contains the source and destination IP address of the packet, the time recorded, and a diagnostic message identifying the type of malware data that was observed. It also identifies the ASN number, domain name, and geographical region associated with the source IP address.

We therefore have a set of records describing malware traffic as it enters and exits points on the networks under observation.

[1]www.spamhaus.org

Once infected, in general, a host will do a number of things:

1. It will propagate the virus to other hosts.
2. It will often communicate with a controller.
3. It will send data to a receiver, or sinkhole.

The diagnostic data from Spamhaus is able to determine if data is being sent to a sinkhole.

We are therefore able to select records from this dataset for further analysis according to a number of different parameters. For example we may choose to investigate data from a particular domain, e.g. bt.net, or from geographical location, e.g. networks located in the UK or China. We also may select which particular virus, or botnet, to analyse and choose to focus on command and control data or data being sent to the sinkhole.

3. CIRCULAR STATISTICS

Mardia states that circular statistics analyses distributions of random variables that are cyclic in nature [11]. Thus, regarding the time of an attack as a random variable, we map our original data covering several contiguous days to a 24-hour circle. Clearly, the ensuing analysis will be different from that of the original data stretched along its entire time span. Special statistical tools have been developed to assist researchers with circular data analysis [13].

3.1 Circular Mean

To illustrate the importance of circular statistics, Brunsdon and Corcoran [3] provide an example to reveal the misuse of ordinary, or linear, statistics with cyclic data: four disorder incidents recorded at midnight times 23:30, 00:15, 00:30 and 00:45, the mean value of these four times by the ordinary (arithmetic) averaging gives a morning time of 06:15, whereas the true (Fréchet [4]) average is 00:15. There are different ways to average circular data. One can argue that the most natural definition is the Fréchet mean [4] given by:

$$\bar{\theta}_F = \arg\inf_{\theta\in\mathbb{S}} \sum_{i=1}^{n} d^2(\theta_i, \theta) \tag{1}$$

where $\mathbb{S}$ is the unit circle ($\mathbb{S} \in \mathbb{R}^2$) and $d(\cdot,\cdot)$ is the arc length on it. But in practice one often uses an alternative definition [3]:

If

$$A = \sum_{i=1}^{n} \sin(\theta_i), B = \sum_{i=1}^{n} \cos(\theta_i) \tag{2}$$

then

$$\bar{\theta} = \begin{cases} \arctan(A/B) & \text{if } B \geq 0 \\ \arctan(A/B) + \pi & \text{if } B < 0 \end{cases}$$

where $\theta_1, \theta_2, \ldots, \theta_n$ are the n observations of circular data.

This definition can be considered an approximation to the Fréchet sample mean, and its version is implemented in the R package 'Circular' [13], which we use in this paper.

3.2 Distribution Hypothesis Tests

Pewsey et al. proposed that the uniformity hypothesis is the most basic null hypothesis in circular statistics, and its

rejection in favour of a general alternative means that the data provide evidence that the circular distribution in question is non-uniform [13]. Disregarding, for the time being, the issue of periodicity, the theoretical assumption that (unordered) times of attacks are distributed uniformly over a time interval $(a, b]$ of interest suggests that (ordered) times of attacks follow a homogeneous Poisson process. Consequently, the number of attacks in a fixed time interval should have a Poisson distribution with parameter $\lambda(b-a)$, where $\lambda > 0$ and is known as rate or intensity of the Poisson process. This assumption may or may not hold in practice [8].

3.2.1 Tests for a Poisson process

Here, we consider an example of malware received at a botnet sinkhole in the domain bt.net over a period of 15 days. Thus, we consider these BT data to be a random sample from a certain point process and we test the null hypothesis that the process is Poisson.

Without assuming homogeneity of the process, we partition the day into 24 one hour intervals and apply the standard Chi-square goodness-of-fit test to each of these 24 sub-samples, lumping all the days together. Thus, our null hypothesis is that the i-th sub-sample is a random sample from a Poisson distribution with an unspecified parameter λ_i, $i = 0, 1, \ldots, 23$. For example, we observe a total of 63 attacks during 01:00 – 01:59 and out of these 60 minutes, 34 minutes have no attacks, 15 minutes have 1 attack each, 7 minutes have 2 attacks each, 3 minutes have 3 attacks each, and only 1 minute has 4 attacks. Hence, we have five bins labelled by the number of attacks as 0, 1, 2, 3, 4 and more (accounting for the infinite tail of the Poisson distribution). We use the standard implementation of the goodness-of-fit test provided in R ('vcd' package [15]) with its default settings (minimum chi-square estimation of λ_i and the rule to have at least five expected counts disabled).

The results in table 1 show only five hours with p-values below 5%, a common significance level. That is, unlike the other 19 sub-samples, each of these five sub-samples provides significant evidence against the null hypothesis that the frequency of malware observed on bt.net follows a Poisson distribution.

Assuming further that the 24 sub-samples are independent and all the 24 null hypotheses hold true (i.e. the day process is Poisson with intensity varying from one hour to another but constant within each hour), we would expect only $0.05 \cdot 24 \approx 1$ of the 24 tests to reject its null hypothesis. Thus, aggregating these 24 tests into a single binomial test, we would obtain the overall p-value, i.e. the probability to reject five or more (at 5% significance level) out of the total of 24, of 0.006, which is very low. Thus, the sample subjected to these simple tests reveals strong overall evidence against the Poisson hypothesis, which is concentrated in the evening hours (5pm and later).

In our BT dataset, we record the attacks by minutes and merge 15-day data together to do the Poisson distribution test. We find the 15-day graphs of λ from each day have various variations as showed in figure 1. All graphs show the similar variations that morning have less attacks than afternoon and evening.

Given these variations in the process intensity, it appears sensible to also examine the Poisson hypothesis for each hour of each day separately. However, if we were to apply the same Chi-square goodness of fit test within each such time interval, we would have too little data for the test to be meaningful.

Table 1: Chi-square goodness-of-fit tests reveal evidence (at 5% significance level) that malware received at the BOT sinkhole in domain bt.net do not conform to the Poisson assumption. Y is "Yes"; N is "Not", and TNH is the total number of attacks in the given hour; λ is the estimated mean (and the variance) of the number of attacks in the given hour; df is the degree of freedom of the test. The results are based on the minimum chi-square estimates of λ.

Time slots	P-value	Poisson	TNH	λ	df
00:00-00:59	0.5778	Y	126	2.1359	5
01:00-01:59	0.4904	Y	63	1.0771	3
02:00-02:59	0.8509	Y	52	0.8830	2
03:00-03:59	0.2307	Y	42	0.7370	3
04:00-04:59	0.9593	Y	28	0.4812	1
05:00-05:59	0.8745	Y	31	0.5212	2
06:00-06:59	0.9134	Y	33	0.5542	2
07:00-07:59	0.8495	Y	55	0.9307	3
08:00-08:59	0.2407	Y	137	2.3353	5
09:00-09:59	0.4820	Y	178	3.0978	4
10:00-10:59	0.3918	Y	192	3.1835	7
11:00-11:59	0.2525	Y	219	3.8242	9
12:00-12:59	0.7202	Y	186	3.1350	6
13:00-13:59	0.5833	Y	201	3.5358	9
14:00-14:59	0.0781	Y	195	3.7625	10
15:00-15:59	0.6344	Y	190	3.2320	6
16:00-16:59	0.5061	Y	227	3.7705	8
17:00-17:59	0.0012	N	214	4.1493	11
18:00-18:59	0.4774	Y	198	3.3066	6
19:00-19:59	0.0013	N	230	4.6770	12
20:00-20:59	0.0272	N	261	4.8489	12
21:00-21:59	0.0906	Y	242	4.2135	9
22:00-22:59	0.0079	N	230	3.7304	7
23:00-23:59	0.0237	N	176	3.3006	9

If we instead partitioned the data over longer intervals, then this could easily miss the distinct inhomogeneity feature of the process. Fortunately, there are tests that are specifically designed for situations when homogeneity of the Poisson process cannot be assumed [9]. Kim and Whitt [9] study several such tests based on an earlier work of Brown [2]. We also use one such test, referred to as 'Log Test' in the research of Kim and Whitt [9], which is based on pivotal quantities (3). Table 2 shows all test results of every hour of the acceptance or rejection of a Poisson process.

The first part of the idea is as before, i.e. to use a piecewise approximation of the rate, or intensity, function $\lambda(t)$. But the main part of the idea is to convert the attack times into the log transformed inter-arrival times (3) as those (under the null hypothesis) follow the exponential distribution with a constant mean (λ=1). Brown et al. define log transformed inter-arrival times X_{ij} as follows [2]: $i = 1, ..., I$, $j = 1, ..., n_i$,

$$X_{ij} = -(n_i + 1 - j) log \left(\frac{T - T_{ij}}{T - T_{i,(j-1)}} \right), \qquad (3)$$

where

- I: the total number of time intervals;
- n_i: the total number of attacks in the ith interval;

Table 2: Test results for a Poisson process (H1 (0:00-0:59) is the first hour of a day; d1 (7th Aug) is the first day of the tested datasets; A is "Accept" the Poisson hypothesis; R is "Reject"; T is the number of attacks in each hour or each day; "NA" means there is no attack in that hour; all tests are at 5% significance level.)

Hour/Day	d1	d2	d3	d4	d5	d6	d7	d8	d9	d10	d11	d12	d13	d14	d15	T
H1	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
H2	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
H3	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
H4	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
H5	A	A	A	A	A	A	A	A	A	A	R	A	NA	A	A	A
H6	A	A	A	A	A	NA	A	NA	A	A	A	A	R	A	NA	A
H7	A	A	A	A	R	NA	R	A	A	NA	A	A	R	A	A	A
H8	A	A	A	A	R	A	A	A	A	A	A	A	A	A	A	A
H9	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
H10	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
H11	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
H12	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
H13	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
H14	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
H15	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
H16	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
H17	A	A	A	A	A	A	A	A	A	A	A	A	A	A	R	A
H18	A	A	A	A	A	A	A	A	A	A	R	A	R	A	A	R
H19	A	A	R	R	A	A	A	A	A	A	A	A	A	A	A	A
H20	A	A	A	A	A	A	A	A	R	A	A	A	A	R	A	R
H21	A	A	R	A	A	A	R	A	A	A	A	A	A	A	R	R
H22	A	A	R	A	A	A	A	A	A	A	R	A	A	A	R	A
H23	A	A	A	A	A	A	A	R	A	A	A	A	A	A	A	R
H24	A	A	A	R	A	A	R	A	R	A	A	A	A	A	A	R
T	A	R	R	R	R	A	R	R	R	R	R	R	R	A	R	R

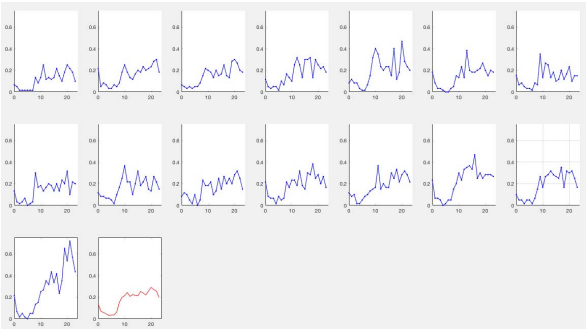


Figure 1: Lambda graphs (1-15 graphs present the variation of λ in each day, the final graph denotes the variation of the 15-day average λ)

- T: the total time (minutes) of an interval(here,1 hour);
- T_{ij}: the jth ordered attack time in the ith interval so that $T_{i1} \leq ... \leq T_{i,n_i}$ and T_{i0} =0.

We transform the original data T_{ij} to X_{ij} and test the variables X_{ij} by the 'Log test' to examine whether X_{ij} are a sequence of i.i.d exponential variables with mean 1 [2]

$$X_{ij} \sim Exp(1)$$

Not only does this make goodness of fit tests such as the Kolmogorov-Smirmov one applicable within each short subinterval, but it also allows us to merge the transformed data from the individual intervals into larger samples to increase the test power.

Our results show that with small samples (individual one hour intervals) the Poisson process hypothesis appears acceptable (see table 2). However, when we start aggregating our data as described above, the large sample tests lead to rejection. An immediate interpretation is that this is entirely due to the increase of the test power (although we do not specify the alternative hypothesis). We then do more experiments to examine this explanation. Thus, we tests sub-samples of increasing size corresponding to the first k hours in a given day. We then also randomly permute our transformed samples and aggregate those. Interestingly, the former experiment shows how the p-value tends to drop as the sample size increases. This supports our conjecture that the actual process is mildly non-Poisson, which was impossible to detect with small samples. The latter experiment shows that the p-value does not drop noticeably as the sample size increases, which suggests that unlike a Poisson process, our attack process may be violating the assumption of independence of inter-arrival times. This may be a sensible explanation as attacks indeed need not be statistically independent.

In the original dataset, we find that there is no attack at some hours. In that case, we consider them missing data. Hence, there is no p-value in these hours. Most of p-values are greater than the 5% significance level. Therefore, we have no evidence against the Poission process hypothesis.

3.2.2 *Uniformity Hypothesis Tests*

It may be interesting to experiment with other partitions

as a way of assessing robustness of the above results. In practice, most of the existing research of malware distributions do not consider the periodicity of the data over 24 hours. For instance, Ramachandran and Feamster illustrate that the vulnerable hosts across IP address space follow a uniform distribution in the assumptions of the worm propagation analysis [14]. But they do not mention the relationship between this uniform distribution and time periods. Therefore, this paper will apply two common circular tests, Rayleigh and Watson tests, to examine the uniformity hypothesis to the selected malware data [13]. Bogdan et al. [1] state that Rayleigh and Watson's tests are the classical methods for testing the uniformity hypothesis in circular statistics. We set the null hypotheses H_0 of the tests as

- H_0: the observations θ of attack times are uniformly distributed around the circle.

where decimals observations θ are converted from the original time. For example, 3:30 is converted to 3.5 by $3 + \frac{30}{60}$. That is, we have an observation $\theta = 3.5$. Thus, if the p-value is less than the 5%, we have significant (at 5% significance level) evidence that the observations are not uniformly distributed around the circle.

3.2.3 *Rayleigh test*

The Rayleigh test is considered to be the most common test for uniformity hypothesis [1]. Given n observations of $\theta_1, ...\theta_n$ [5], calculate the test statistic R:

$$R = (V^2 + W^2)^{\frac{1}{2}} \tag{4}$$

and

$$V = \sum_{i=1}^{n} \cos\theta_i, W = \sum_{i=1}^{n} \sin\theta_i \tag{5}$$

where R is the length of the vector sum, V is the northerly component of the sum (southerly, if negative), and W is the easterly component. Thus, the p-value of the Rayleigh test is $e^{-R^2/n}$. If the p-value is less than given significant level (like 5%), then we have evidence against the Uniformity hypothesis.

3.2.4 *Watson test*

Brunsdon and Corcoran [3] state that Watson test is another common method to the uniformity hypothesis. They demonstrate the differences between Rayleigh and Watson test in the uniformity hypothesis as follows [3]:

- Rayleigh will focus on whether the observations θ are distributed uniformly around the circle;
- Watson pays more attention on whether the observations θ have the same means around the circle.

In later section, we will apply two tests to the samples of attack time.

3.3 Large-sample Mardia-Watson-Wheeler Test

Large-sample Mardia-Watson-Wheeler test is a method to test multiple independent samples for a common distribution [13]. The test statistic W_g is given by Pewsey et al. as follows [13] :

$$W_g = 2\sum_{k=1}^{g} \frac{C_k^2 + S_k^2}{n_k}, \tag{6}$$

where

$$C_k = \sum_{j=1}^{n_k} \cos(\frac{2\pi R_{kj}}{N}), S_k = \sum_{j=1}^{n_k} \sin(\frac{2\pi R_{kj}}{N}) \tag{7}$$

- R_{kj}: the rank of the j^{th} element in the k^{th} sample;
- g: the number of independent samples;
- θ: the vector which is combined by the g sub-samples and ranked by an arbitrary zero direction;
- N: the total number of combined sample of θ;
- n_k: the sub-sample of N with $N = \sum_{i=1}^{k} n_k$;
- $2\pi R_{kj}/N$: the uniform scores of the data in θ.

4. DATASETS AND GEO-LOCATION

4.1 Dataset Source

SPAMTEQ provides the malware datasets which are observations from 1000 hosts on commercial (business) organizations between 8th August and 21st August 2016. The datasets include the following information [17]:

IP address: The address of the host detected originating behaviour.

ASN: The autonomous system number of the IP address via routeviews at the time of file generation and less than 24 hours old.

Country: The Country code where the IP address is geolocated derived partly from private database and partly from RIRs.

Domain: The domain associated with the entity that owns the ASN and derived from private database.

Timestamps: Epoch time of last connection.

Diagnostic: An unformatted raw record as generated from the CBL engine.

4.2 Dataset Selection

In our paper, we focus on the sinkhole class of diagnostic from 8th August 2016 to 21st August 2016. The top three countries, plus UK, are chosen from the whole malware dataset, followed by the sinkhole(s_) data, from the information of the diagnostic of these four countries. Then we select the top 1 domain from the sinkhole datasets, and the top three malware from the top 1 domain. The flow chart of figure 2 illustrates how we selected the datasets.

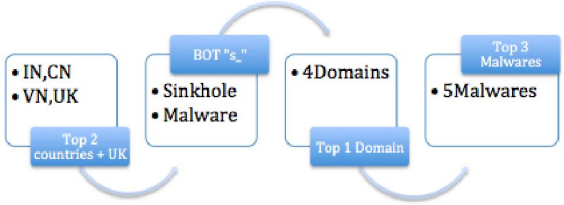


Figure 2: Flow chart for the selection of malware datasets

The top 3 countries are selected: India(IN), Vietnam(VN) and China(CN). UK is ranked 36 and chosen into our datasets. IN, VN and CN have the same top 3 malware. Furthermore, conficker is common malware in these four countries. The attack number of each country is described in table 3.

Table 3 shows that India has 288779 attacks and therein 74% are from the top 1 domain of 'sancharnet.in'. 99821 are conficker attacks in the top 1 domain. There are 110428 worm attacks in the same domain. In fact, conficker accounts for a larger proportion in the top 1 domains of four countries (96% in chinanet.cn.net and 62% in 'vnnic.net.vn').

5. TOP DOMAIN ANALYSIS

Shin and Reddy state the importance of conficker in their research and study the victim distribution patterns to provide better defence against this particular malware [16]. They also illustrate that the current analysis of conficker has two classifications: binary behaviour and internet propagation pattern [16]. However, our paper will apply circular statistics as a new technique to analyse the time patterns of conficker attacks as follows:

- Daily cycles;
- Uniformity hypothesis tests for the daily cycles;
- Weekly circles

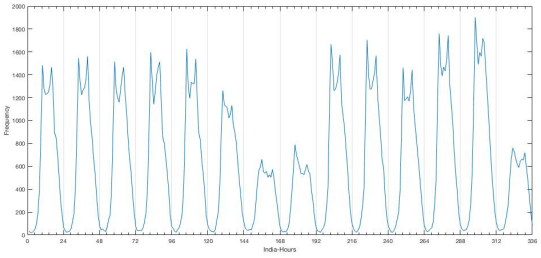
5.1 Linear Histograms, Daily Cycles and Helix Graphs

This section will plot three different graphs to describe the time distributions of malware as follows: linear histograms, rose diagrams and helix graphs. They will present the 14-day data from different perspectives and reveal the time variations of the malware attacks.

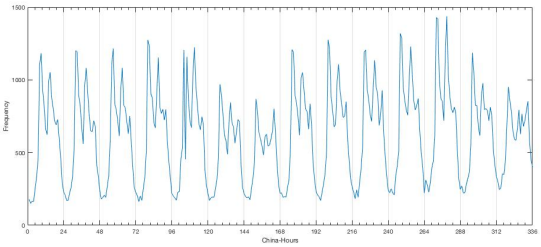
5.1.1 Linear Histograms

Normally, a linear histogram is an universal graph to describe the frequency distribution of attacks in cyber seucrity. Here, we draw the line plots for four countries by converting 14 days into 336 hours as figure 3. The four linear histograms are observed as follows:

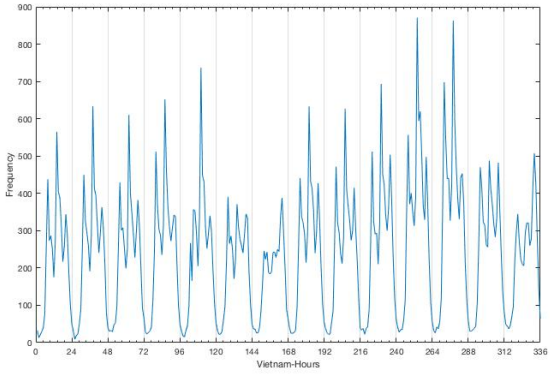
1. IN figure 3a: The second Saturday (20st August) has the highest frequency; the first Sunday (14th August), the second Monday (15th August) and second Sunday (21 August) have the lower frequency of conficker attacks; it is worth to noting that 15th August is the Indian public holiday (Independence Day).
2. CN figure 3b: There are roughly three peaks in each day; two higher peaks appear in the mid morning and mid afternoon; the first peak is generally higher than the second one; the attacks of malware in the working days are more than the weekend as a whole.
3. VN figure 3c: The line histogram of Vietnam also has three peaks in each day; the second peak is generally higher than the other two ones; the attacks of malware at the weekend are less than the weekdays.
4. UK figure 3d: The UK linear histogram shows the relatively irregular changes in the overall trend; it is multimodal variation in each day.



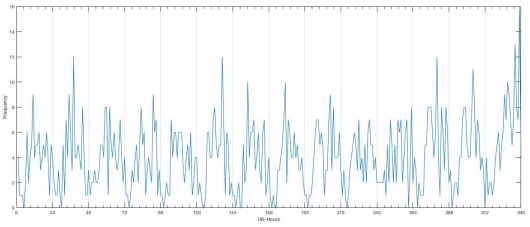
(a) IN



(b) CN



(c) VN



(d) UK

Figure 3: Linear histograms of four countries

Table 3: The number of malware attacks in 4 countries (C.N. is the total attack number of a country; D.N. is the total attack number of top 1 domain; S.M(w)/(t) is the second top malware :worm_dorkbot(w) and tinba(t).)

Country	C.N.	Domain	D.N.	Conficker	S.M(w)/(t)
IN	288779	sancharnet.in	213874(74%)	99821(47%)	110428(52%,w)
CN	302135	chinanet.cn.net	217047(72%)	20763(96%)9	7896(4%,w)
VN	135780	vnnic.net.vn	133097(98%)	82029(62%)	47652(36%,w)
UK	5894	opaltelecom.co.uk	3455(59%)	1352(39%)	1451(42%,t)

Table 4: Mean Times of Countries (Top domains in four countries; Mean time is computed in R 'circular' package [13])

Country	Domain	Mean time
IN	sancharnet.in	13:25
CN	chinanet.cn.net	13:06
VN	vnnic.net.vn	14:03
UK	opaltelecom.co.uk	14:12

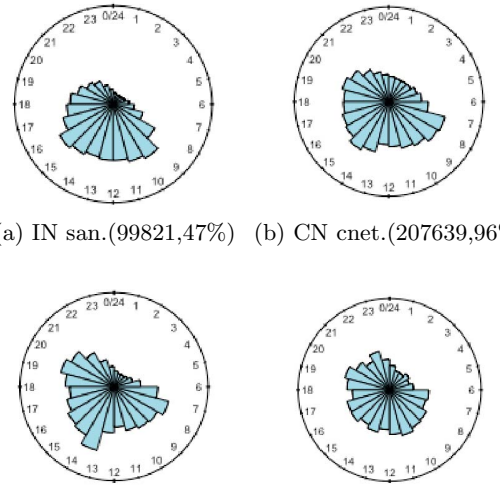
To sum up, the linear histograms show the frequency distribution of malware over hours. There are roughly two peak times in each day, one in the morning and the other in the afternoon. And then the frequency of conficker attacks normally falls at noon and early overnight. Furthermore, it may reasonably infer that the attacks will occur less at Holiday and weekend. However, we still need more data to support this idea in future. However, these findings illustrate that the necessity of considering the relationships between working times and the frequency of attacks.

Although the linear histograms describe the frequency distribution over 336 hours, they do not illustrate the real relationship between the frequency of malware attacks and time-in-day. If the falling time point is consistent with the some human-being living habits like mealtime, and the peak time points are located in the working time, we may infer reasonably the habit will affect the time behaviour of conficker attacks. Therefore, we will apply a daily cycle model to demonstrate the their relationships.

5.1.2 Daily Cycles

In the daily cycles, we aggregate the 2-week data into cycles and observe the active time periods in the 24-hour pattern. As we have mentioned above, the attack times are converted into decimals in modelling a daily cycle of conficker.

Four daily circles described in figure 4 show that the same botnet (conficker) in different countries has different time patterns. Furthermore, the graphs also demonstrate the frequency of conficker attacks is not evenly distributed around the circle. We observe four rose diagrams of attacks respectively and find their bimodality feature. For instance, the IN rose diagram illustrates that the frequency of conficker attacks reaches a peak at around 15:30, falls off at around 2:00 and peaks at around 9:30 again. Table 5 summarizes the frequency of conficker attack in top 1 domain and the active and quiet time. The usual peak hours for the conficker are the working times in the morning and afternoon. The common quiet time is at early overnight. Hence, the daily cycles show the features of bimodality and multimodality. In this case, we need to test the uniformity hypothesis of these four datasets to examine the features in the later section.



(a) IN san.(99821,47%) (b) CN cnet.(207639,96%)

(c) VN vnc.(82029,62%)) (d) UK op.(1352,39%)

Figure 4: Rose diagrams of conficker attacks in the top 1 domain (The top domains in four countries are respectively: sancharnet.in(san.), vnnic.net.vn(vnc.), chinanet.cn.net(cnet.), opaltelecom.co.uk(op.); the number of sub-caption shows the total attack times, followed by the percentages in their domains.)

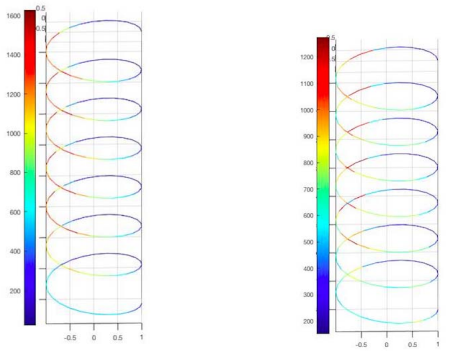
Table 5: Peak and Fall times for conficker

Country	Frequency	Peak time	Fall time
IN	99821	9:30,15:30	2:00,13:30
CN	207639	7:30,14:30,19:30	2:00,12:30
VN	82029	7:30,13:30,19:30	2:00,11:30,17:30
UK	1352	10:30,19:30,22:30	2:00,12:30

5.1.3 Helix Graph

We have aggregated the 2-week data into the daily cycles. However, aggregating the data may miss some information such as the real peak times of conficker attacks in each day. Therefore, we use a helix graph to describe the data by time-in-hour (168 hours for 7 days, from 8th August to 14th August). The helix diagram provides a better view to the active time patterns of attacks. Figure 5 shows the four helix graphs for India, China, Vietnam and UK. The gradual change of colour (the colour bar) of a helix graph suggests the range change of occurrence frequency of mal-

ware. There are 7 cycles in a helix graph and each cycle presents a day.



(a) IN (b) CN

(c) VN (d) UK

Figure 5: 7-day Helix graphs (High frequency is indicated by the lighter colour)

We observe the India helix (figure 5a) that the darker colour appears at overnight in the parts of a cycle. That means, the conficker is not active at overnight in this week. The cycles become lighter colour from the mid morning to afternoon. According to the gradual colour change, we can more intuitively observe the time variations of malware attacks in each day. In principle, the daily cycles and helix diagrams provide a new method to analyse the time patterns and the relationships with frequency of malware attacks instead of linear statistics. Aggregating the data into circles can help us understand the active and quiet time of attacks.

5.2 Results of Uniformity Hypothesis Tests

Although the daily cycles show the the time-in-day attacks are not evenly distributed over 24 hours, we still need to use a statistical test to examine this finding. Thus, we implement two circular statistical methods: Rayleigh and Watson tests to four countries' datasets in R's 'circular' package [13]. The results of two tests are p-value = 0 and p-value < 0.01 respectively for all datasets. Therefore, we have significant (at 1% significance level) evidence that the occurrence times of conficker are not uniformly distributed around a 24-hour circle.

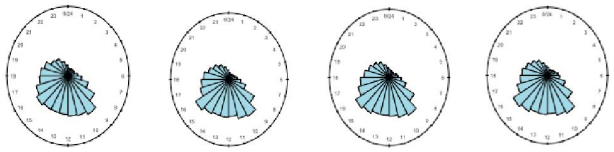
5.3 Weekly Circles

In this section, we calculate the circular mean for the time-in-week datasets and show the results in table 6. Here, the mean time represents the average time that a conficker attack is detected during a day. The mean times of India and China from Monday to Sunday are at around 13:00-14:00. For Vietnam, the mean times of a week are at around 14:00-15:30. 13:30-16:00 is the range of mean times for the UK. Moreover, table 4 suggests that the mean times of Sunday are, in general, later than the other days in theses four countries. That is, the attack from conficker will be detected at later time on Sunday than the other days. Overall, the results of the table 4 show the mean times of attacks for the top 1 domain of four countries are between 13:00 and 14:00.

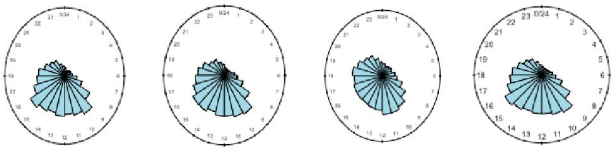
The differences of the mean times for each country help us to observe the daily distribution of each day and the changes of the attack frequency in 24 hours. Therefore, we draw the rose diagrams for time-in-week. In fact, the datasets include 2 weeks data collected from 8th August 2016 to 21st August 2016. Here, We firstly check the Inida histograms for these 2-week data. The frequency of conficker attacks of day-in-week has two peaks in figure 3a. The linear histograms show that the data are not normalized.

Table 6: Mean times of a week

Country	IN	CN	VN	UK
Monday	13:24	13:04	13:56	13:29
Tuesday	13:23	12:58	13:57	13:40
Wednesday	13:25	12:58	14:09	14:04
Thursday	13:22	13:00	13:49	14:23
Friday	13:23	12:59	13:36	13:44
Saturday	13:19	13:21	14:06	13:54
Sunday	13:55	13:43	15:29	15:56



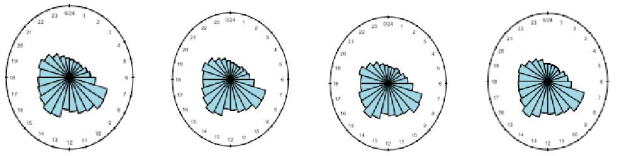
(a) Mon(11050) (b) Tue(16381) (c) Wed(16424) (d) Thu(15426)



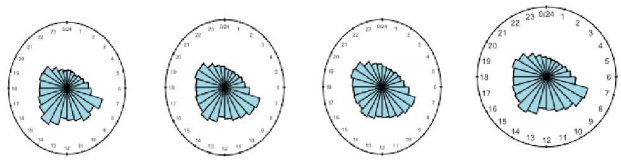
(e) Fri(17377) (f) Sat(16089) (g) Sun(7074) (h) Day(99821)

Figure 6: IN Weekly Circles (The figure is the attack number in two weeks)

Figure 6 illustrates the time variations of time-in-week in India. The India rose diagrams from Monday to Saturday have the similar regulation distribution as the daily cycle. They peak at around 15:30, fall at around 2:00 and peak again at around 9:30. The Sunday rose diagram is different from the other days, it shows that the frequency of conficker attacks peak at round 11:30, and then keep a steady level until the next peaking time at round 18:30.



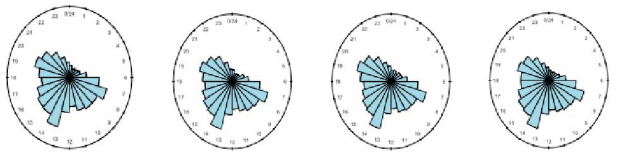
(a) Mon(29638) (b) Tue(29965) (c) Wed(30217) (d) Thu(31562)



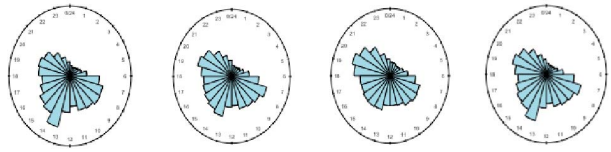
(e) Fri(32446) (f) Sat(27798) (g) Sun(26013) (h) Day(207639)

Figure 7: CN Weekly Circles

The China conficker dataset is a large sample and has 207639 detected attacks. The rose diagrams of Monday to Sunday are similar to the daily circle. They have three peaks respectively are at around 7:30, 14:30 and 19:30, and the falling times at around 14:00 and 12:30. Overall, the China rose diagrams show a regular change in the time pattern variations of conficker.



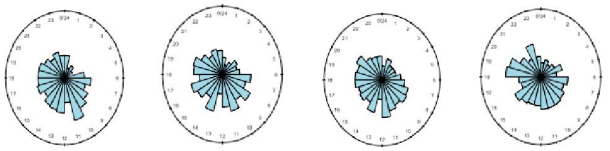
(a) Mon(11179) (b) Tue(11337) (c) Wed(11857) (d) Thu(13519)



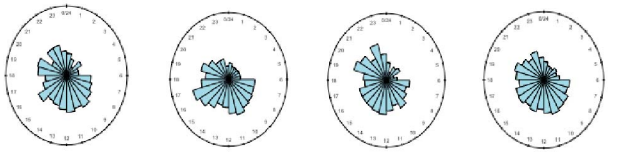
(e) Fri(13537) (f) Sat(10918) (g) Sun(9682) (h) Day(82029)

Figure 8: VN Weekly Circles

The VN weekly circles have a similar variation as India. From Monday to Saturday, the rose diagrams have a high similarity to the daily rose diagram with the same peak and falling times. Three peaks appear at around 7:30, 13:30 and 19:30, and 13:30 has the highest frequency. Nevertheless, the peak time period of the Sunday rose diagram is 18:00-21:00. The frequency of conficker occurring on Sunday afternoon is lower than the other days.



(a) Mon(176) (b) Tue(192) (c) Wed(184) (d) Thu(175)



(e) Fri(197) (f) Sat(200) (g) Sun(228) (h) Day(1352)

Figure 9: UK Weekly Circles

UK has 1352 conficker attacks and its figure 9 presents an irregular time pattern. The time pattern is very different from the other countries. The rose diagrams show the irregular peak and falling times. Therefore, we obtain the following information by observing all rose diagrams:

1. India and Vietnam have the similar patterns from Monday to Sunday. The frequency of conficker attacks from Monday to Saturday has similar variations as the daily rose diagram. However, Sunday afternoon has different time variation.

2. All China cycles show the high similarity in the time variations.

3. The UK cycles have the irregular changing rules from Monday to Sunday and are different from its daily cycle.

4. The time pattern variations of conficker illustrate that it is active during the working hours in working days plus Saturday. Figure 7 and 8 show the regular time variations like India rose diagrams. The rose diagrams of Monday, Tuesday, Wednesday, Thursday, Friday and Saturday in India, China and Vietnam are very similar to the corresponding daily circles. That means, the attacks of conficker in these days will become active from around 7:00, peak at around 13:30 or 14:00 and fall overnight.

5. The Sunday rose diagrams in India, China and Vietnam are different from the other circles. Firstly, the frequency of conficker on Sunday is lower than the other six days. Secondly, the Sunday time patterns show that the frequency of conficker will decrease at around 13:00 to 15:00.

To sum up, the occurrences of conficker in the UK are to a large extent less frequent than India, China and Vietnam. We discuss the reasons with the users' habits of Microsoft Windows application in these countries. The conficker working group illustrates the computers are infected by conficker at around the world, particularly the developing world of Asia such as India [7]. The group infers reasonably the reason why heavily located in the developing world that the computers universally install the pirated Windows operating system software without patching these systems [7].

Therefore, attackers prefer to attack users in these countries. The attacks also display more regular variation from Monday and Saturday as the related daily circles. The UK, with low frequency of conficker, has an irregular weekly time distribution.

The results of time-in-week analysis reveal that the total number of attacks has an effect on the time patterns. Specifically, India, China, and Vietnam, which have noticeably larger numbers of attacks than the UK, exhibit bimodality in the daily distributions of the attack, and little day-to-day variation within the week. The UK instead shows a somewhat multimodal daily variation with irregular peaks and quiet times.

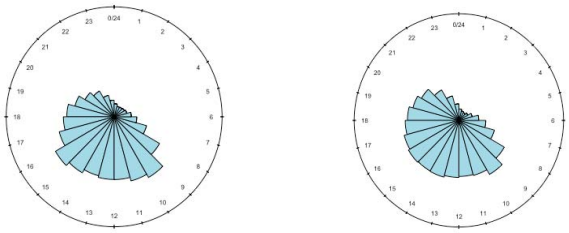
6. COMPARISONS BETWEEN CONFICKER AND THE OTHER MALWARE

In this section, we focus on the time variations of different malware attacks in four countries. The two most frequent malware in each country are compared, and the Mardia-Watson-Wheeler test (the test results are provided by the R 'circular' package [13]) is used to test whether these malware attacks have a common distribution.

We observe that conficker and worm_dorkbot are the top two botnets in India, China and Vietnam. In the UK, conficker and tinba are the top two malware. All the observations of the malware are based on the top 1 domain of each country.

6.1 Comparison in India

In India, 52% attacks are from worm_dorkbot and 47% are from conficker. We observe from figure 10 that the time variations of two botnets are very similar with no big difference in the frequency of attacks. We apply the Mardia-Watson-Wheeler test to these two-botnet attack samples, and find the p-value is nearly 0. In this case, we have evidence against the null hypothesis that attacks from these two botnets have a common distribution.

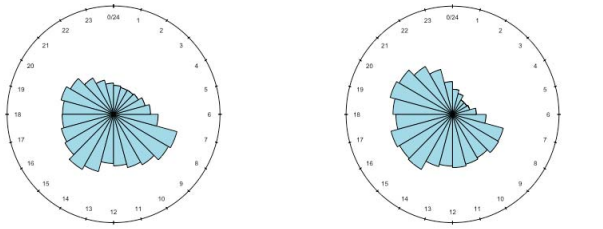


(a) Conficker(99821,47%) (b) Worm_dorkbot(110428,52%)

Figure 10: IN Daily Circles

6.2 Comparison in China

In China, 96% attacks are from the top 1 botnet conficker and only 4% attacks from Worm_dorkbot . Figure 11 shows that the frequency of attacks from these two botnets has similar time variations. The p-value of the Mardia-Watson-Wheeler test is nearly zero. Thus, we have significant (at 1% significance level) evidence that the attacks from two botnets do not have a common distribution.

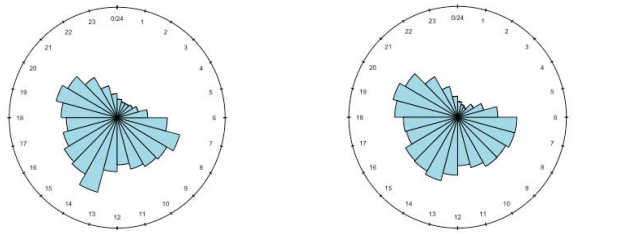


(a) Conficker(207639,96%) (b) Worm_dorkbot(7896,4%)

Figure 11: CN Daily Circles

6.3 Comparison in Vietnam

Conficker is still the top 1 bontnet in domain vnnic.net.vn and 62% attacks from it. 36% attacks are from the second botnet Worm_dorkbot. The time variations of the two botnets are very alike, although the number of attacks in each botnet are rather different as showed in Figure 12. The p-value of Mardia-Watson-Wheeler test is nearly zero. Thus, we have evidence against the common distribution hypothesis. In other words, the attacks from two botnets do not have a common distribution in domain vnnic.net.vn.

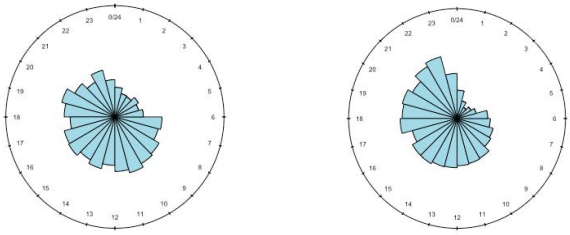


(a) Conficker(82029,62%) (b) Worm_dorkbot(47652,36%)

Figure 12: VN Dialy Circles

6.4 Comparison in the UK

Tinba is a trojan, rather than a botnet. This malware is the most prevalent in the domain opaltelecom.co.uk with 42% attacks. Conficker has 39% attacks in the same domain. The daily cycles of conficker and tinba shown in figure 13 are very different. We observe that two cycles do not have common regular variations. The result of Mardia-Watson-Wheeler test shows that the p-value is nearly zero. That is, we have significant (at 1% significance level) evidence that two malware do no have a common distribution in domain opaltelecom.co.uk.



(a) Conficker(1352,39%) (b) Tinba(1451,42%)

Figure 13: UK Daily Circles

To sum up, conficker and other malware in India, China, Vietnam and United Kingdom do not have a common distribution regardless of the observation that the daily circle patterns of different malware. The results of the Mardia-Watson-Wheeler test support this conclusion with very small p-values.

7. CONCLUSION

In this paper we have demonstrated how circular statistics may be applied to analyse and visualise the time patterns of malware events. In particular, the daily, weekly cycles and helix diagrams provide a visualisation method. Thus we can monitor the malware behaviour and allocate resources to mitigate attacks in a more efficient manner. The results of our analysis for four countries are also worth noting and are summarized as follows:

1. In India, China, and Vietnam, the active time periods for conficker are approximately from 7:00 to 8:30, 13:00 to 15:00.

2. The frequency of conficker attacks in the UK is lower that the other three countries.

3. The Rayleigh and Watson tests of uniformity illustrate that the active time of conficker is not uniformly distribution around the 24-hour circle.
   Further, the results of time-in-week analysis reveal that the total number of attacks has an effect on the time patterns. Specifically, India, China, and Vietnam, which have noticeably larger numbers of attacks than the UK, exhibit bimodality in the daily distributions of the attack, and little day-to-day variation within the week. The UK shows a somewhat multimodal daily variation with irregular peaks and quiet times.

4. In the UK, the daily and weekly cycles show irregular time variations, but the cycles of India, China and Vietnam follow the similar and regular time pattern variations.

5. In India, China and Vietnam, the Sunday cycles are different from the other weekdays. The occurrences of conficker on Sunday afternoon is less than the same period of time in the other days.

6. The results of the large-sample Mardia-Watson-Wheeler test demonstrate that conficker and the other malware do not have a common distribution, no matter the daily circles are alike or different, or the attacking frequency of malwares are rather high or low.

We have demonstrated that, overall, the data strongly suggest non-uniform distributions of malware. In other words, the attack time of malware is not uniformly distributed over a 24-hour cycle. Malware will be active in some time period and related with the human behaviours. Having said that, we have also illustrated that the malware time patterns of different countries have some common points such as peaking at mid morning and mid afternoon, and falling at early overnight.

We believe that these findings will be helpful to improve the effect of detection systems. The analysis and visualisations should help decision makers in cyber security to efficiently allocate resources or estimate the cost of system monitoring in the different time periods. And if malware activity is observed in an unusual time period, security managers may then investigate further.

One possible limitation of this work is that we only use data in August. Thus we did not consider it would be useful to extend the analysis to monthly or quarterly patterns. But our goal was to examine the application of circular statistical analysis applied to incidents of malware attacks rather than looking at the distributions using linear statistics. Furthermore, in the future research, it is worth noting that closing the time line into a circle results in having infinitely many uniform partitions (of a fixed size, say, hourly). This will provide the possibility of assessing robustness of various statistical tests to time translations.

However, we will keep tracking the time patterns of malware attacks by more real data. Having demonstrated that we can identify probability distributions for malware events, we are hoping to extend this work by investigating the likelihood of cyber attacks in general. If we can quantify this, it will allow us to make a better estimate of the risk an organisation faces in terms of cyber attacks.

8. ACKNOWLEDGMENTS

The authors would like to thank Simon Forster (Spamhaus Technology Ltd) for providing the spam datasets.

9. REFERENCES

[1] M. Bogdan, K. Bogdan, and A. Futschik. A data driven smooth test for circular uniformity. *ACM Trans. Program. Lang. Syst.*, 54(1):29–44, March 2002.

[2] L. Brown, N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, and L. Zhao. Statistical analysis of a telephone call center: A queueing-science perspective. *Journal of the American statistical association*, 100(469):36–50, March 2005.

[3] C. Brunsdon and J. Corcoran. Using circular statistics to analyse time patterns in crime incidence. *Computers, Environment and Urban Systems*, 30(3):300–319, May 2006.

[4] I. L. Dryden, A. Koloydenko, and D. Zhou. Non-euclidean statistics for covariance matrices, with applications to diffusion tensor imaging. *The Annals of Applied Statistics*, 3(3):1102–1123, March 2009.

[5] D. Durand and J. A. Greenwood. Modifications of the rayleigh test for uniformity in analysis of two-dimensional orientation data. *The Journal of Geology*, 66(3):229–238, May 1958.

[6] C. W. Group. Conficker working group: Lessons learned. Technical report, The Rendon Group, June

2010.

[7] C. W. Group et al. Conficker working group: Lessons learned. *Conficker-Working-Group-Lessons-Learned-17-June-2010-final. pdf, published Jan*, January 2011.

[8] N. Heard. 2016. https://www.statslife.org.uk/files/Slides/1-Nicholas-Heard.pdf.

[9] S.-H. Kim and W. Whitt. Choosing arrival process models for service systems: Tests of a nonhomogeneous poisson process. *Naval Research Logistics (NRL)*, 61(1):66–90, January 2014.

[10] T. Maillart and D. Sornette. Heavy-tailed distribution of cyber-risk. *The European Physical Journal B*, 75(3):357–364, April 2010.

[11] K. V. Mardia. Statistics of directional data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 37(3):349–393, March 1975.

[12] R. M. Mutwiri, H. Mwambi, and R. Slotow. Approaches for testing uniformity hypothesis in angular data of mega-herbivores. *International Journal of Science and Research*, 5(3):1202–1207, March 2016.

[13] A. Pewsey, M. Neuhäuser, and G. D. Ruxton. *Circular statistics in R*. Oxford University Press, Oxford, UK, 2013.

[14] A. Ramachandran and N. Feamster. Understanding the network-level behavior of spammers. In *ACM SIGCOMM Computer Communication Review*, pages 291–302. ACM, October 2006.

[15] V. Ricci. Fitting distributions with r. *Contributed Documentation available on CRAN*, 96, February 2005.

[16] S. Shin, G. Gu, N. Reddy, and C. P. Lee. A large-scale empirical study of conficker. *IEEE Transactions on Information Forensics and Security*, 7(2):676–690, December 2012.

[17] SPAMTEQ. 2016. https://www.spamhaustech.com.