Computer Science and Information Systems 18(4):1197–1217 https://doi.org/10.2298/CSIS200715040Z

Using Honeynet Data and a Time Series to Predict the Number of Cyber Attacks

Matej Zuzčák and Petr Bujok

Department of Informatics and Computers, Faculty of Science, University of Ostrava
30. dubna 22, 701 03 Ostrava, Czech Republic
{matej.zuzcak,petr.bujok}@osu.cz

Abstract. A large number of cyber attacks are commonly conducted against home computers, mobile devices, as well as servers providing various services. One such prominently attacked service, or a protocol in this case, is the Secure Shell (SSH) used to gain remote access to manage systems. Besides human attackers, botnets are a major source of attacks on SSH servers. Tools such as honeypots allow an effective means of recording and analysing such attacks.However, is it also possible to use them to effectively predict these attacks? The prediction of SSH attacks, specifically the prediction of activity on certain subjects, such as autonomous systems, will be beneficial to system administrators, internet service providers, and CSIRT teams. This article presents multiple methods for using a time series, based on real-world data,to predict these attacks. It focuses on the overall prediction of attacks on the honeynet and the prediction of attacks from specific geographical regions. Multiple approaches are used, such as ARIMA, SARIMA, GARCH, and Bootstrapping. The article presents the viability, precision and usefulness of the individual approaches for various areas of IT security.

Keywords: cyber attacks, honeynet, honeypot, SSH, time series, prediction.

1. Introduction

Besides common users, servers providing various services are the target of virtually unceasing cyber attacks. These servers are most commonly managed using the SSH protocol. SSH provides the administrator with a remote access console offering the same functionality as if they were at the server site. It is one of the most commonly attacked protocols, both by human attackers and by automated bots that are a part of extensive botnets. The SSH protocol was selected as it is among the most frequently attacked protocols, according to the following reports: F-Secure Attack landscape H2 2018[1], Akamai - The State of the Internet Q4 2014[2]. Botnets most commonly use the computers of unaware users, connected to the internet via various technologies and internet service providers across the world.

Server administrators must inevitably protect their systems from a variety of attacks. To do so effectively, they must know and analyse the threats and use that knowledge

[1] F-Secure Attack landscape H2 2018 – https://blog.f-secure.com/attack-landscape-h2-2018/

[2] Akamai - The State of the Internet Q4 2014 – https://www.akamai.com/us/en/multimedia/documents/state-of-the-internet/akamai-state-of-the-internet-report-q4-2014.pdf

to, for example, expand the databases of IDS [3]/IPS [4] systems. Honeypots, which can be likened to lures or traps, are an ideal tool for this task. Besides being able to analyse historical attacks, a certain foresight of what to expect is also useful to administrators. This will allow them to prepare appropriate protective measures in advance and estimate what types of attacks are going to be prevalent.

In addition to server administrators, ISPs should also be aware of botnets and any infected computers on their networks. These companies would also benefit from an effective estimation of attack rates on their networks. It would allow them to more effectively deploy countermeasures to such attacks, such as dynamic IP address management, since the reputation of these addresses could suffer damage if they were assigned to an infected device conducting malicious activities. Therefore an overview of situational development enables ISPs to make appropriate decisions.

Other groups that can benefit from such foresight are the Computer Security Incident Response Teams (CSIRTs) and researchers to whom the ability to predict potential attacks is imperative. For instance, CSIRTs can effectively predict from which autonomous systems or IP address ranges intensive attacks can be expected, and how attacks with certain identifying features will progress. This would allow CSIRTs to prepare countermeasures and contact the operators of the affected autonomous systems ahead of time. Predicting specific details of how a threat spreads through the world, such as which RIR or country it will likely spread from, allows researchers to deploy monitoring tools appropriately to gain as much data as possible.

Various methods can be used to make predictions, with a time series being one of the most commonly used. There are multiple approaches to setting the necessary parameters. This paper analyses and compares the approaches with the goal of identifying the most effective one in predicting the attacks on a system over time. Predictions by every approach were made over the same period of time and compared with real-world data collected over the same period, a period of approximately one year. The real-world data was collected by the author's honeynet.

2. Honeypot and Honeynet

A honeypot [1] is a system for analysing activity taking place within itself. The activity is commonly malicious, with the goal of using the infected system to spread itself or other threats and conducting other malicious activity such as DDoS attacks or sending spam. A honeypot can consist of software, hardware, or an entire network [2]. Such a system is usually made intentionally vulnerable, and it provides no real-world services. It is usually operated with the intention of analysing and assessing the activity taking place within it. Such a system has to be closed insofar as no activity taking place within it could possibly negatively influence other systems or spread via LAN, WAN, or the internet in general. At the same time, the system must be sophisticated enough to allow the minimal possible contact between the attacker and the outside. The goal here is to give the attacker the impression of a real-life system it can conduct its activity within, without realising it is actually restricted. A compromise between the security and realism of the system has to be achieved, depending on what specific threats the honeypot is focused on.

[3] IDS – instruction detection system

[4] IPS – instruction prevention system

The term ”honeynet” [6] tends to be context-dependent. It is commonly used in reference to a honeypot with a high level of interaction. In this case, it means a specific type of network that, besides the honeypot, may contain other components, such as a special firewall called a honeywall, an IDS/IPS system, and various database systems for data collection, etc.

An additional meaning for a honeynet, is a system of honeypots forming a logical but non-physical system. This meaning is commonly applied to collections of honeypots with a low to medium level of interaction. The use of special tools such as firewalls is not necessary in this case. Data from all the honeypots in a honeynet are commonly collected into a single database. A honeynet can provide a large amount of threat data for analysis.

3. Related Work

The prediction of the development of attacks using time sets, applying various algorithms and methodologies is the focus of several papers. The paper [8] directly deals with predicting attacks detected by honeypots. It uses data from the CZ NIC honeynet that is composed of Kippo honeypots running on port 22. The paper proposes a model that predicted attacks on an emulated SSH protocol, providing the attacker with the ability to log in to the shell and execute some commands. Overall, 179 540 records from the period between 2.11.2014 and 8.5.2016 were analysed. Data from 75 weeks was used to train the model, and data from 5 weeks was used to compare the prediction of that period with the real data from it. An AR(1) - AR model of the 1st order time series with bootstrap point prediction was used. The paper concludes by stating the model is viable for predicting future attacks based on the demonstration.

In the paper [9], a large series with a large amount of data from security incidents is used. It compares the possible ways to predict attacks using a model based on a time series and using the Non-Homogeneous Poisson Process (NHPP) software reliability growth model.

The paper [10] proposes a prediction of the intensity of attacks based on known data regarding the number of attacks per day using the ARIMA model. Four types of attacks are identified: Denial of Service (DoS), malicious emails, malicious URLs, and attacks on the Internet facing service (AOIFS).

In the paper [11], an IDS system for wireless networks for process automation (WIA-PA) is proposed. It is based on recorded network traffic, processed using a model based on the ARMA time series.

In the paper [12], a framework for the prediction of vulnerabilities based on a statistical analysis using a time series between January 1999 and January 2016 is introduced. The ARCH, GARCH, and SARIMA models were used. The data was taken from the National Vulnerability Database (NVD) [5] in its 2016 state. The results of the predictions were mainly useful for the risk management of vulnerabilities.

In the paper [13], predictions based on two approaches, the Extreme Value Theory (EVT) and Time Series Theory (TST) are presented. The TST used the FARIMA + GARCH model. It concludes that EVT is more effective for predicting a longer time period, 24 hours and more, whereas the TST is better for immediate threats, such as within

[5] National Vulnerability Database – `https://nvd.nist.gov`

the hour. It uses data from a honeypot recording the network activity during five time periods in 2010 and 2011. The honeypot emulated several services using the following solutions: Amun [6], Dionaea, Mwcollector [7]a Nepenthes [8]. Data was extracted from the PCAP files generated by capturing network traffic, where every TCP connection, including an unsuccessful TCP handshake, was considered an attack.

In the paper [14], the prediction of attacks based on past event logs is studied. Various methods are applied and evaluated, such as the historical communication between the attacker and the victim, models for neighbour searching, techniques searching for global patterns using Singular Value Decomposition (SVD), and a time series using the Exponential Weighted Moving Average (EWMA) model. Logs from the Dshield project[9] over a period of one month, formed the data set used. The solution was proposed as a framework for a Blacklisting Recommendation system (BRS) as a linear combination of three approaches, namely a time series and two approaches from a neighbouring model area - K-Nearest Neighbor (KNN) algorithm and a co-clustering algorithm. Using SVD showed no significant improvement in the predictions, and it was therefore not included in the proposed solution. However, the solution could be useful for improving the generation of lists of the addresses of attackers.

The content of the paper [15] is not directly concerned with predicting attacks, but a time series is used to represent captured attacks and to demonstrate analytical outputs. Specifically, it proposes a framework for clustering captured attacks on honeypots to as many similar clusters as possible. Symbolic aggregate approximation (SAX) is the technique used for clustering, providing the ability to reduce the dimensionality of the data, therefore ignoring insignificant details. As a result, a cluster may contain attacks against different ports but represents the same network worm that spreads by using multiple ports.

In the paper [16], various aspects of prediction methods used in cyber security are analysed. The methods are divided into three categories: data mining, dynamic network entity reputation, and the use of time series'. The time series methods used were: ARIMA models, exponential smoothing models, "naive approach", and the average of the ARIMA and exponential smoothing models. The paper also looks at and evaluates the accuracy of the categories from the point of view of blacklisting. The data used in the paper was acquired from the SABU platform, which gathers intrusion detection alerts. The data covers a period of seven days. The authors conclude that attack prediction is an approach useful for estimating the number of attacks in the near future and can be used by the given system's operator to optimise its countermeasures. We consider the time period of seven days to be too short.

In the paper [17], a deep, state-of-the-art overview of the current approaches, taxonomy, and methods used for cyber security attack prediction are presented.

The content of the paper [18] is focused on "data-driven incident prediction" methods, and the shift from reactive to proactive approaches of protection.

The authors of the paper [19] propose an IACF framework focused on alert aggregation and correlation, and attack prediction and detection.

---

[6] Amun honeypot – `https://github.com/zeroq/amun`

[7] Mwcollector part of – `https://sourceforge.net/projects/honeybow/files/honeybow/0.1.0/`

[8] Nepenthes honeypot – Deprecated honeypot solution. It is no longer developed nor supported.

[9] Dshield project – `https://www.dshield.org/`

Only the papers [8], [13], and [15] contain data gathered by honeypots.

The aim of paper [8] is conceptually the closest to this one, due to the chosen approach, but it uses very few methods of prediction, only AR(1) and Bootstrapping. It also only predicts the number of attacks on honeypots, and does not deal with predicting the behaviour of individual attackers over time, and the relationship to geographical location and autonomous systems.

The authors of the paper [13] analysed a time series and an EVT approach. They only used a single time series method, and by considering every TCP connection to be an attack, it is arguably too broad a definition of an attack.

In the paper [15], honeypot gathered data is used, but the attacks are not directly predicted, but rather clustered using a time series.

In papers [9], [12] attacks are not predicted, but vulnerabilities and security incidents are predicted using the ARIMA, SARIMA, ARCH, and GARCH approaches.

In papers [10], [11], [16], [18], and [19], the intensity of cyber attacks in a wider context is predicted, for example, DoS attacks or malicious emails. It uses the ARIMA and ARMA approaches.

The analysis in the paper [14] is specific, as it analyses event logs using the SVD and EWMA approaches.

None of the available related research uses an approach utilising a range of time series based prediction methods and also do not focus on predicting attacks based on the geographical location or other clustering variables of the attackers, such as address ranges. Due to this fact, this paper focuses on these specific aspects.

4. The Honeynet Used and Delineation of the Relevant Time Period

Individual honeypots are based on various types of networks, with the captured connections, and the potential attacks, being sent to a central server where they are saved in a central MySQL[10] database for further analysis. The honeypots are presented in table 1. Each node, or sensor, is running an instance of the Cowrie honeypot emulating an SSH server.

Table 1. Honeypots composing the honeynet.

Sensor ID	Network type	Port
HP1	CESNET - Czech academic network	22
HP2	Czech VPS hosting - grey zone – grey zone	22
HP3	Regular Czech VPS hosting	22
HP4	Czech ISP	22
HP5	Slovak ISP - dynamic IP	2222
HP5-B	Slovak ISP - dynamic IP	22
HP6	VPS hosting - India	22

[10] MySQL– `https://www.mysql.com/`

4.1. Analysed Data

The honeynet captured all connections heading mostly to port 22, an SSH shell emulation. In one case port 2222 (Tab. 1) was used. Every connection established between a honeypot and a potential attacker is called a session. If during a session the potential attacker logs into the shell and conducts additional activity by inputting commands, such as downloading files and executing them, or uploading files from the emulated system, such a session is considered to be an attack in the context of this paper.

The article focuses on two main areas. The first is predicting the overall number of attacks against the honeynet in the given time period, described in Chapter 6.1. The second is predicting attacks based on their source, or from the point of view of their source, and is subdivided into three areas: Regional Internet Registry (RIR) in Chapter 6.2, country of origin (Chapter 6.3), and the activity of the autonomous systems (AS), specifically, where the attack originated from (Chapter 6.4).

A detailed analysis of the sources of attack from a geographical and analytical point of view is considered important. This is due to the needs of AS administrators often only concentrating on gathering and estimating the development of attacks in the area relevant to them. Therefore assessing the effectiveness of predicting attacks from the point of view of the source area is one of the main goals of this paper.

Before any prediction took place, the Kwiatkowski–Phillips–Schmidt–Shin (KPSS) test [24] was applied to the data to verify its stationarity. The null hypothesis of the test is as follows: the data is stationary around a deterministic trend, as opposed to the alternative where they are not stationary. Table 2 presents the calculated p-values, in the case of p being < 0.05 it means a rejection of the null hypothesis. For every evaluated aspect, such as the overall number of sessions or the source of attacks, a search for the best available prediction method was conducted, based on the evaluation of the Mean absolute percentage error (MAPE) in Chapter 5. Individual aspects were modelled by using a variety of established approaches: the Holt-Winters algorithm, ARIMA, SARIMA, GARCH models (for some situations where the data allowed it), and Bootstrapping models. With Bootstrapping, the predictions were always made using three approaches: stationary, fixed, and one based on a model. The one based on a model was not always viable. The tables in the following chapters only present the Bootstrapping model, which achieved the lowest MAPE error.

4.2. Time Period Used for the Prediction

Real data captured by the author's honeynet in the time period between 30.7.2017 and 7.11.2018 was used for predicting and for training the methods. Considering the time period is 16 months, an accumulation of the daily data had to be considered, mainly for reducing the zero value observation for some days. (i.e. there are 466 daily-measured values). Cumulating it into weeks (i.e. seven daily-measured values into one) seems appropriate, as months or quarters of the year would result in too few data points for prediction, therefore, dramatically decreasing the accuracy.

As a result of this, a time period of seven days was chosen to accumulate the measured values to a weekly aggregate. This resulted in 66 weekly data points, out of which 58 weeks were used to train the models, and the last eight weeks were used to test the accuracy of the predictions.

5. Methods Used for Prediction

There is an entire gamut of methods for predicting future observations using a time series. Traditional approaches are mainly based on the decomposition of values in a time series, or by using the Box-Jenkins methodology. Besides the more traditional approaches, other, less conventional ones are available, for example, those based on Bootstrapping. This paper applies several approaches to obtain as accurate a prediction of future observations for a time series as reasonably possible, while also demonstrating the robustness of the methods used. To predict the future values of a time series, $Y_{t+\tau},\ \tau = 1, 2, \ldots$, with sufficient accuracy, a standard deviation of prediction from the real value, an error, has to be introduced:

$$e_t = Y_t - Y_t^{'}(t-1)\ , \tag{1}$$

where Y_t is the value of the time series in the time t, and $Y_t^{'}(t-1)$ is the prediction of that value from the value of the time series in the time $t-1$. Using the error, we can evaluate the quality of the predictive model based on the values of the time series using the *Mean absolute percentage error* (MAPE):

$$\mathit{MAPE} = \frac{100}{n}\sum_{t=1}^{n}\frac{|e_t|}{Y_t} \tag{2}$$

5.1. Holt-Winters

In 1957 Holt introduced a general algorithm of exponential smoothing [25], which was subsequently expanded by Holt and Winters three years later [26]. The Holt-Winters algorithm is based on three components of a time series: level, trend, and a seasonal component. Based on the application of the components, there are two variants of the algorithm, *additive* and *multiplicative*. In the additive model, the components add up, with each being measured in the same units as the time series itself. In the multiplicative model, only the level is in the time series' units, the trend and the seasonal component are factors within the interval $(0,\ 1)$. Even though the Holt-Winters algorithm is rather simple, the results show it achieves very accurate predictions in many different contexts and areas.

5.2. ARIMA

The stationary mixed model of Box-Jenkins methodology ARIMA(p,d,q) was introduced in 1970 [23] and it can be symbolically expressed using the following equation:

$$\phi(B)(1-B)^d Y_t = \theta(B)\varepsilon_t \tag{3}$$

where ϕ is the autoregressive (AR) process, θ is the process of the moving average (MA), B is the lag operator, d is the differentiation operator, and ε_t is white noise.

Besides the autoregressive process AR(p) and the moving average process MA(q), the model also contains the differentiation operator I, which is used to stationarise the non-stationary time series.

The ARIMA model can be calibrated by adjusting the values of the parameters p, d, and q. Setting a parameter value to zero leaves the parameter out, so for example, if $d = 0$ the model is ARMA(p,q) and so on.

5.3. SARIMA

SARIMA is a variant of the ARIMA model expanded to include the seasonal part, allowing it to model a time series influenced by a seasonal component. The model is inscribed as SARIMA(p,d,q)(P,D,Q)$_{Sz}$, where the symbols in the first pair of brackets represent the parameters of the standard ARIMA, while those in the second pair represent the seasonal variants. The Sz parameter is the number of seasons per year. The SARIMA model can also be inscribed using a lag operator:

$$\phi(B)\Phi(B^{12})\Delta^d\Delta_{12}^D Y_t = \theta(B)\Theta(B^{12})\varepsilon_t \tag{4}$$

As with ARIMA, the model can be calibrated by adjusting the values of the parameters (p, d, q, P, D, Q, Sz), with ε_t again being white noise. Setting a parameter to zero omits it.

5.4. GARCH

The GARCH (Generalized Autoregressive Conditionally Heteroscedastic) model was introduced in 1982 [27] and is a generalisation of the ARCH model. GARCH assumes variable volatility, the heteroscedasticity, of a time series. The value of the series in time t can be inscribed using the GARCH(m, s) model as:

$$Y_t = \mu_t + \varepsilon_t\sqrt{\sigma_t} \tag{5}$$

with

$$\sigma_t^2 = \alpha_0 + \alpha(B)Y_t^2 + \beta(B)\sigma_t^2 \tag{6}$$

where $\alpha_0 > 0$, $\alpha_i \geq 0$, $\beta_j \geq 0$ are the parameters of the model.

5.5. Bootstrapping

The Bootstrapping technique was introduced in detail in paper [28] by Bradley Efron. Bootstrapping is a very straightforward technique. In order to calculate the confidence interval *CI* for a statistic $T = t(X)$ with a set of n elements $X = x_1, x_2, ..., x_n$, it just repeats the following scheme R times:

- For the i-th iteration: sample n elements from the available sample, while allowing for the repeated choice of the same elements.
- Based on the sample created in the previous step X_i, calculate the new statistics $T_i = t(X_i)$.

There are several modifications of the Bootstrapping method, one is known as block Bootstrapping. Here, the data vector $(X_1, ..., X_n)$ is divided into blocks of length l.

$$Y_1 = (X_1, \ldots, X_l), Y2 = (X_{l+1}, \ldots, X_{2l}), \ldots Y_k = (X_{(k-1)l+1}, \ldots, X_n) \tag{7}$$

This is followed by an independent random sampling from the population of vectors $Y_1, \ldots, Y_N$, providing the sampled vectors $Y_1^*, Y_2^*, \ldots, Y_N^*$. A vector of random variables $(X_1^*, \ldots, X_n^*) = (Y_1^*, \ldots, Y_k^*)$ is considered a Bootstrap selection.

For fixed block sampling (BS_fixed), the date of the beginning of the block is generated first, followed by the date of the point chosen, this allows the time series to have the same length as the block.

For stationary block sampling (BS_stationary), the date of the beginning of the block is generated first using geometric distribution. The new block of data is then drawn onto a new time point and added to the series. This process is also repeated until the new series has the same length as the original one.

In this paper, Bootstrapping methods from the R[11] [30] language and $tsboot$[12] function containing several methods for the resampling of a time series were used. The function $auto.arima$[13] was used, with the parameters: $max.p = 25$, $max.d = 0$, $max.q = 0$, $max.P = 0$, $max.Q = 0$, $max.D = 0$, $ic = 'aic'$, $max.order = 25$, $seasonal =$ *TRUE*. Resampling was set to the value of 100. More detail about the residual bootstrap method is found in [29] and a description of auto.arima in R is found in [31].

6. Prediction of Attacks on the Honeynet

This paper applies various models of time series to predict either the overall number of attacks, or the behaviour from individual sources of attacks. The stationarity hypothesis of the data was tested using the standard Kwiatkowski–Phillips–Schmidt–Shin test (KPSS) described in Chapter 4.1 for every specific time series. The most accurate predictions of the given aspect evaluated by the lowest Mean absolute percentage error (MAPE) are plotted as a graph, effectively demonstrating the predictions through real time. In all figures except Fig. 5, the real time series is presented by the green plot, with the orange plot representing the teaching run, and the yellow plot illustrating the predicted values of the time series provided by the most accurate method (the model with the lowest MAPE error).

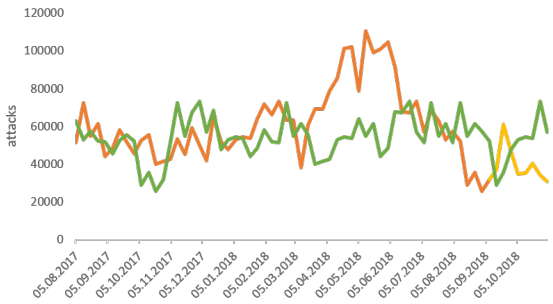

Fig. 1. Overall number of attacks in the given time period

[11] R language – https://www.r-project.org/

[12] Tsboot function – https://www.rdocumentation.org/packages/boot/versions/1.3-23/topics/tsboot

[13] Auto.arima function – https://www.rdocumentation.org/packages/forecast/versions/8.9/topics/auto.arima

6.1. Prediction of the Total Number of Attacks

The first aspect to be analysed was the overall activity of the attacks and sessions directed at the honeynet. This data shows the activity to be quite variable and unstable, as the sources are often home computers being used as part of a botnet. A more detailed analysis is available in [22]. The KPSS test applied to the measured data achieves a significance level of 0.02, see Table 2. Therefore the null hypothesis for the stationarity of the data is rejected. The results obtained from applying each individual method are presented in Table 3. The ARIMA(1,1,0) method achieves the lowest error and thus also the lowest deviation from the real time data of the predicted time period. It is represented in Fig. 1. The results of the predictions can be assessed, given the dynamic behaviour of the attackers, as satisfactory. Analysts can use this to predict trends in the future activity of attacks on the honeynet. However, for a more detailed analysis, the predictions must be made in shorter time frames, as highlighted in the following chapters.

Table 2. KPSS for all time series data.

data	**All_attacks**	**AFRINIC**	**APNIC**	**ARIN**	**LACNIC**
KPSS	0.02	0.05	0.07	>0.10	<0.01
RIPENCC	**AS4134**	**AS4837**	**AS16276**	**AS14061**	**AS45899**
>0.10	>0.10	>0.10	0.04	>0.10	>0.10
China	**Russia**	**Netherlands**	**USA**	**France**	
>0.10	<0.01	>0.10	<0.01	<0.01	

Table 3. An overview of the MAPE values when predicting the overall number of attacks on the honeynet.

Attacks on honeynet	**MAPE (%)**
Holt-Winters$_A$	50.3
SARIMA(1,1,0)(2,0,0)	43.5
SARIMA(0,1,0)(2,0,0)	42.1
SARIMA(1,1,0)(0,1,0)	22.4
ARIMA(1,0,0)	36.2
ARIMA(1,1,0)	22.0
GARCH(2,2)	26.9
BS_fixed	50.6

The values of significance for the KPSS test relating to the individual time series modelled, are presented in Table 2. If the value of significance is less than 0.05, the assumption of the stationarity of the time series is rejected. Based on the KPSS test, the following time series are stationary: AFRINIC, APNIC, ARIN, RIPENCC, AS4134, AS4837, AS14061, AS45899, China, and the Netherlands.

6.2. Activity of Attackers from the Point of View of RIR

When considering RIR, the predictions were very close to the geographical distribution of the continents, allowing the prediction of attacking trends from certain regions. As shown in Table 4 and Fig. 2, which present the best models, attacks from each RIR were best predicted by a different approach.

Table 4. Overview of MAPE values when predicting attacks on a honeypot from individual RIRs.

AFRINIC	**MAPE**	**APNIC**	**MAPE**	**ARIN**	**MAPE**
Holt-Winters$_A$	109.2	Holt-Winters$_M$	31.2	Holt-Winters$_A$	82.4
SARIMA(1,0,0)(0,1,0)	87.3	SARIMA(1,1,0)(1,0,0)	41.4	SARIMA(2,1,0)(0,1,0)	104.8
SARIMA(1,1,0)(2,0,0)	39.7	SARIMA(1,1,0)(2,0,0)	30.3	SARIMA(2,1,0)(1,0,0)	81.9
ARIMA(1,0,0)	32.3	ARIMA(2,0,0)	110.8	ARIMA(1,1,0)	34.7
GARCH(1,1)	35.3	GARCH(1,1)	130.1	GARCH(1,1)	55.7
BS_stationary	64.9	BS_stationary	73.1	BS_stationary	33.6
LACNIC	**MAPE**	**RIPENCC**	**MAPE**		
Holt-Winters$_M$	57.3	Holt-Winters$_A$	41.7		
SARIMA(1,0,0)(0,1,0)	37.7	SARIMA(0,1,0)(1,0,0)	28.0		
SARIMA(1,0,0)(2,0,0)	31.9	SARIMA(1,0,0)(1,0,0)	26.7		
ARIMA(1,0,0)	33.6	ARIMA(1,0,0)	27.0		
BS_fixed	35.6	BS_stationary	30.0		

From the graphs representing the individual aspects of attacks, it is apparent these are not easily predictable variables, as their expected value and variance change over time.

The best models were able to predict the development over time with a MAPE error of roughly 30%. The most accurate prediction with the lowest error of 26.7% was achieved by RIPENCC. Again, this suggests the prediction of incoming attack trends from individual RIRs can be rather easily predicted. The accuracy of individual RIR predictions does have a 30% error, although the results are still accurate enough to be useful to researchers for analysis.

According to KPSS, the AFRINIC time series is stationary, resulting in very good results for the ARMA(1,0) model with only a 32% error of prediction. With the APNIC time series, the rather simple Holt-Winters approach was able to achieve a very good level of prediction with an error of only 31%. This success can also be attributed to the series being stationary according to KPSS. Even though the ARIN series is stationary, the ARIMA(1,1,0) stationary model, has the best results here.

The model of non-stationary series LACNIC - SARIMA(1,0,0)(2,0,0) with an error of 30%, being non-stationary, is also surprising.

The last stationary region is RIPENCC, with very similar results achieved by the two models of the Box-Jenkins methodology. However, the seasonal SARIMA(1,0,0)(1,0,0) achieves an even smaller error of 26.7%.

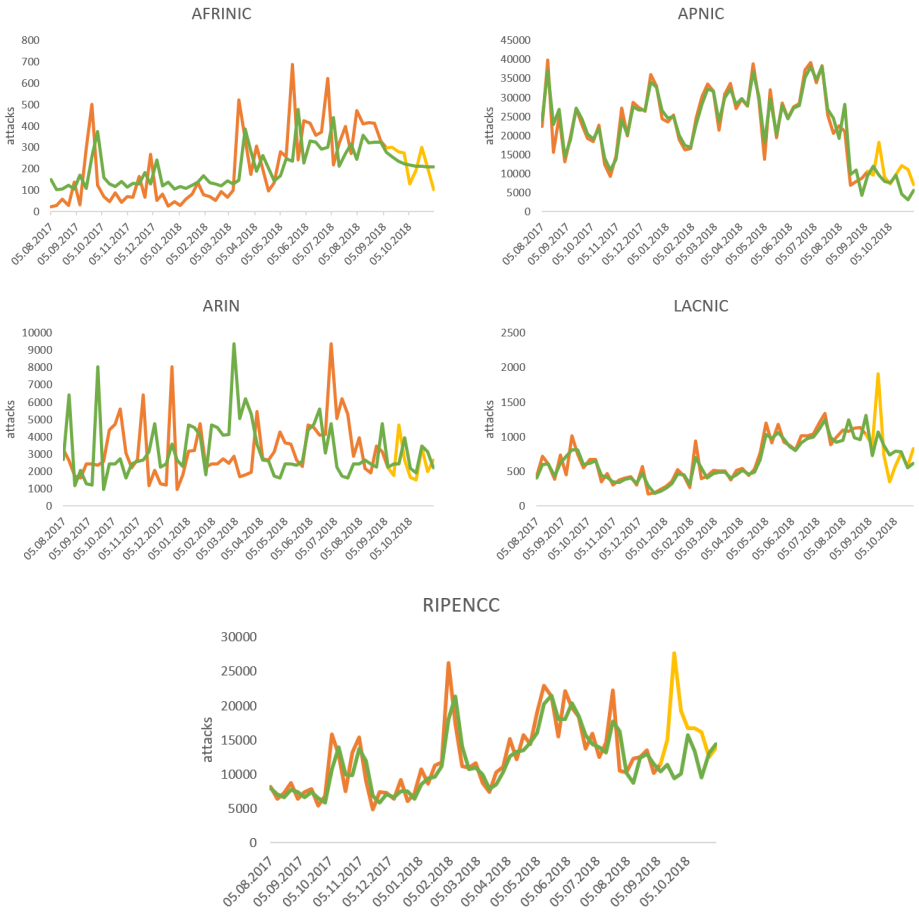

Fig. 2. Prediction of attacks from individual RIRs. The model with the lowest MAPE error according to Table 4 is always presented. The green plot represents the real progression of attacks over time, with the orange representing the teaching run and the yellow in the foreground representing the model.

6.3. Activity of Attackers from the Point of View of Individual Countries

The prediction of attacks based on the country of origin proved to be the most accurate in this research. Given the limited extent of this text, the five most active countries are presented in Table 5 and Figure 3, with the graphs of the best models. The MAPE error of the best models is between 20% (China) and 54% (France), and compared with the errors for RIR, they more accurately predict attacks. The errors for variants of the Bootstrapping model are in the 5th and the 10th line of Table 5.

Analysing the models of prediction and their errors in detail, it shows that three out of the five states obtained their lowest error by using ARIMA. In the case of attacks from China, the ARIMA(1,1,0) model with differentiation has the best result, which is surprising since this time series had weak stationarity according to KPSS. For comparison, ARIMA(1,0,0) a non-stationary model has a 4% higher error. The same model, ARIMA(1,1,0), was also the best for predicting attacks from Russia, which is more fitting since the series is not stationary according to KPSS. The error of prediction for the most successful model for the USA, ARIMA(2,1,0), is 21.4%. It is a very accurate model meant for non-stationary series, which the US one is, according to KPSS. The least accurate of the five countries are the models predicting attacks from France, with the most accurate one being the seasonal SARIMA(0,1,0)(1,0,0) model with differentiation, with an error of 54%. The only model based on Bootstrapping achieving the highest attack prediction accuracy, was for the Netherlands, with an error of 36%.

Apart from the pure research aspect, this information can be very useful to national CSIRT teams, allowing them to prepare appropriate countermeasures in their country well in advance. From a global point of view, predicting attacks based on their source, especially a country based prediction, seems to be the most accurate. This is probably influenced by each country having its own specific predictable variables, like the number of connected computers for common users, the number of servers, habits of the users, and security standards. The predictions were most successful for China, the USA, and Russia. The reason is probably because the USA and China have a proportionally large number of devices connected to the internet. The USA belongs to the ARIN RIR, and China to the APNIC RIR. As shown here, both of these regions obtain a rather good prediction level with an error of roughly 30%. These two countries are also major parts of their regions, allowing for the successful prediction for their RIRs as a whole.

The predictions for the cases of European countries, France and the Netherlands, is less accurate. The activity is more dynamic, and in the case of the Netherlands, it is also influenced by a disproportionately large number of data centres being located there, yet managed from other countries. A detailed analysis of this issue with the Netherlands is found in the paper [22]. France, the Netherlands and Russia belong to the RIPENCC RIR. As mentioned above, the prediction for this region was the most accurate of all the RIRs. In the case of RIPENCC compared to ARIN and APNIC, the main reason for the high accuracy is probably because it contains a large number of small countries and the impact of these is not as large individually as China or the USA in their respective RIRs.

Table 5. An overview of MAPE results for the prediction of attacks from the five most active countries.

China	**MAPE**	**Russia**	**MAPE**	**USA**	**MAPE**
Holt-Winters$_M$	32.6	Holt-Winters$_M$	34.0	Holt-Winters$_M$	59.7
SARIMA(0,0,0)(1,1,0)	29.6	SARIMA(1,1,0)(1,0,0)	47.8	SARIMA(0,0,0)_(1,1,0)	52.1
ARIMA(1,1,0)	20.1	ARIMA(1,1,0)	27.2	ARIMA(2,1,0)	21.4
BS_autoARIMA(1,0,0)	61.5	BS_fixed	33.4	BS_stationary	40.6
France	**MAPE**	**Netherlands**	**MAPE**		
Holt-Winters$_M$	64.4	Holt-Winters$_A$	61.8		
SARIMA(0,1,0)(1,1,0)	54.0	SARIMA(1,0,0)(0,1,0)	52.8		
ARIMA(2,1,0)	87.9	ARIMA(1,1,1)	44.0		
BS_fixed	66.5	BS_fixed	36.2		

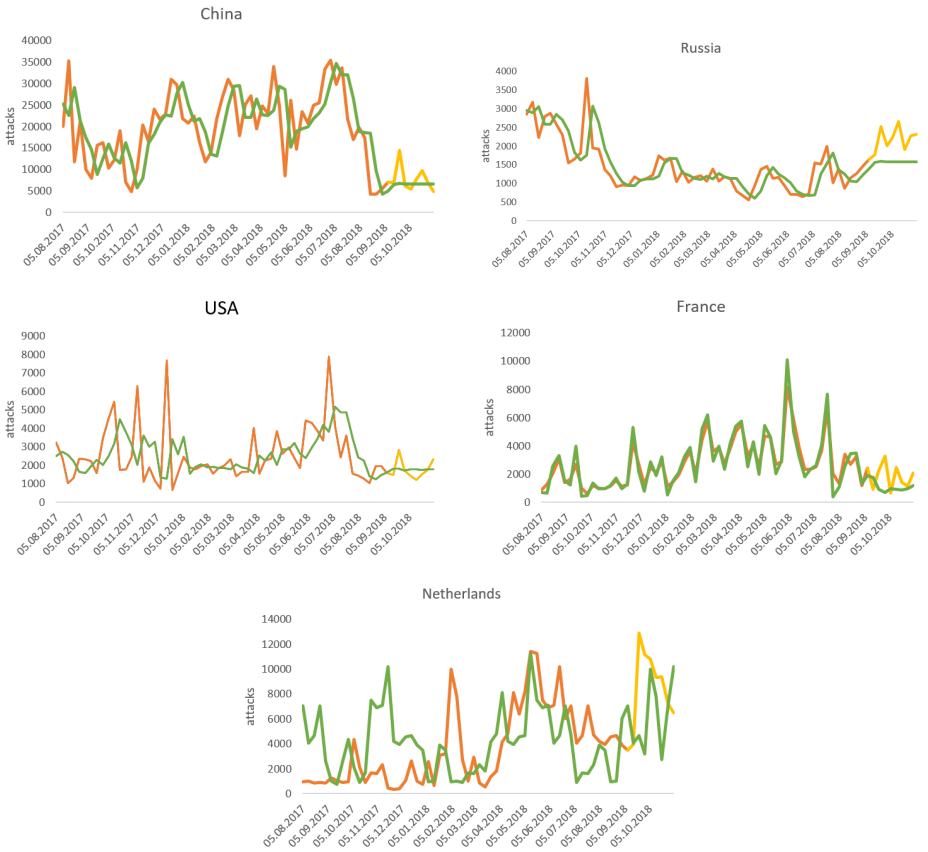

Fig. 3. An overview of MAPE results for predicting the number of attacks from the five most active countries. The green plot represents the real progression of attacks over time, with the orange representing the teaching run and the yellow in the foreground representing the model.

6.4. Activity of Attackers from the Point of View of Autonomous Systems

The prediction of attacks ascertained by the autonomous systems[14] was shown to be the least useful. Table 6 and Figure 4 show that the MAPE error for the five most active autonomous networks varies significantly.

The best error values were between 49.7% and 58%. These are not very accurate estimates, considering most of the series for autonomous systems are stationary. Considering the worst MAPE errors, values higher than 1000% were obtained. With three out of the five autonomous system models, the best predictions were achieved using ARIMA. With the stationary series AS4134 the model ARIMA(1,0,0), achieved the most accurate prediction, with the seasonal SARIMA model achieving more than four times the standard error of about 220%. Large differences in the accuracy of the models are also shown in the case of AS4837, for which ARIMA(1,1,0) was the best, achieving 58% accuracy, even though it is a stationary series, while Bootstrapping predicted values with an error of over 1060%. It should be added that with all the methods of prediction, the best ones were chosen based on the analysis of the settings of the model. System AS16276 achieved a reasonable error level ranging from 51.6% with SARIMA, to 69.2% with the multiplicative Holt-Winters algorithm. The graph for this model shows a very accurate approximation of attacks for the teaching part of the model, with a noticeable reduction in accuracy for the verification part. The next chapter presents the difference between the numerical and factual accuracy of a prediction. Rather balanced errors (compared to errors of other autonomous systems) were also achieved in the case of AS14061, from 49.7% for Bootstrapping, to 76.1% for SARIMA. The stationary series AS45899 was predicted the most accurately by an ARIMA model using a differencing step, with an error of 57.8%, and it was predicted the least accurately by the Bootstrapping model, with an error of 116.4%.

The instability of systems connected within specific autonomous systems is high, whether they are home computers, workstations, or IoT devices. Users turn them on and off at various times, for variedly long periods, with the ISP often mitigating DDoS and spam activity. The common dynamic of assigning addressees should be considered as well. Based on the predictions obtained, it can be concluded that predicting attacks based on autonomous systems is not very effective, with the ISP and AS providers being better off choosing a different approach to predict attacks on their infrastructure.

6.5. Representation of Accuracy and Applicability of the Predictions

The previous chapter presents models for the prediction of attacks using the most accurate methods, specifically, the methods that achieved the lowest MAPE error. When analysing the difference between the predicted and real values of attacks, it appears that even when the model has the lowest error, the prediction is often not very reliable when compared to real values. This is because it predicts the values by a linear or an exponential curve. Therefore, graphs displaying the prediction using the various methods for the chosen models were created, as presented in Figure 5.

[14] Autonomous system (AS) – is a collection of connected Internet Protocol (IP) routing prefixes under the control of one or more network operators on behalf of a single administrative entity or domain that presents a common, clearly defined routing policy to the internet. RFC 1930 - `https://tools.ietf.org/html/rfc1930`

Table 6. An overview of MAPE results for the prediction of attacks from autonomous systems.

AS4134	**MAPE**	**AS4837**	**MAPE**	**AS16276**	**MAPE**
Holt-Winters$_M$	159.4	Holt-Winters$_A$	146.7	Holt-Winters$_M$	69.2
		GARCH(1,1)	696.2		
SARIMA(0,0,0)(1,1,0)	219.5	SARIMA(0,0,0)(1,1,0)	216.6	SARIMA(1,1,0)(1,1,0)	51.6
ARIMA(1,0,0)	52.6	ARIMA(1,1,0)	58.0	ARIMA(1,0,0)	60.8
BS_autoArima(1,0,0)	90.4	BS_fixed	1061.8	BS_fixed	57.6
AS14061	**MAPE**	**AS45899**	**MAPE**		
Holt-Winters$_A$	74.1	Holt-Winters$_M$	64.5		
SARIMA(1,1,0)(0,1,0)	76.1	SARIMA(0,0,0)(1,1,0)	76.5		
ARIMA(1,1,0)	56.2	ARIMA(1,1,0)	57.8		
BS_fixed	49.7	BS_stationary	116.4		

Fig. 4. An overview of MAPE results for predicting the number of attacks from individual autonomous systems. The green plot represents the real progression of attacks over time, with the orange representing the teaching run and the yellow in the foreground representing the model.

The lowest error of prediction, 27% for Russia, was achieved with the ARIMA model, and the curve is nearly constant. Conversely, the second most accurate, the Bootstrapping model, with 33% of error, or the Holt-Winters multiplicative model, with 34% of error, have variable curves over time. It is evident with certain predicted values that these two models are further away from the real values than ARIMA, however, in some other examples, they very accurately predict the values of the real series.

A similar situation occurred with the number of attacks from China, where the most accurate model is ARIMA, with 20% of error, predicting almost constant values. In contrast, SARIMA, with 30% of error, and the Holt-Winters multiplicative model, with 33% of error, are often far closer to the values of the real series.

The prediction of the France time series is the most accurate with the SARIMA model, with 54% of error. Even though the predicted values are not in a linear nor an exponential curve, it is evident it mostly covers the bottom peaks of the real values. The model with the second lowest error of 64%, the Holt-Winters multiplicative algorithm, predicts very similarly. However, while the Bootstrapping model achieves a large percentage error of 66%, it is evident that besides the first value, the prediction is rather close to the real values of the series.

In the case of the APNIC time series, the ARIMA and GARCH models achieve the worst predictions with the largest errors, 110% and 130% respectively, and their curves show no relation to the real values. Alternatively, both the lowest errors, 30% and 31%, and the closest curves were achieved by SARIMA and the Holt-Winters multiplicative model.

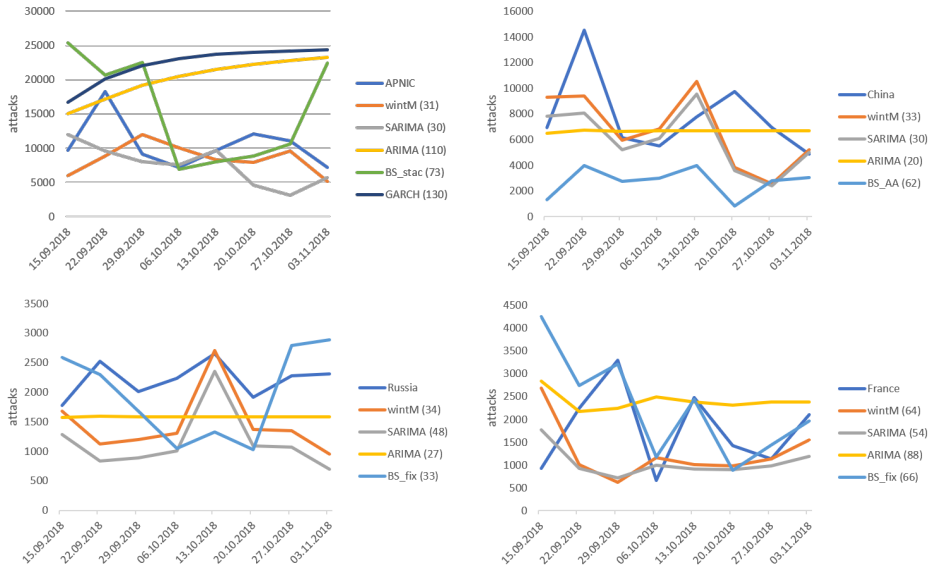

Fig. 5. Representation of accuracy and applicability of prediction.

6.6. Overall Evaluation

Overall, it can be concluded that the most effective predictions of attacks on the honeynet were achieved with a time series predicting the number of attacks from individual countries (i.e. the lower error values were achieved mostly for the data from individual coutries). Such predictions are relatively accurate and can be useful to national CSIRT teams as well as researchers. However, the least effective prediction was achieved with a time series predicting attacks from autonomous systems. With the RIR time series, the predictions can be accurate using an appropriate method for the given RIR. The overall prediction of attacks on a honeynet, regardless of the source, provides a reasonably accurate prediction and potentially useful prediction of attacking trends.

The influence of user and provider behaviour on systems located in a specific country is conclusively strong. Aspects such as user behaviour, the number of provided services such as VPS servers, and security measures are very specific to individual countries, so that when they are grouped, as in a RIR for example, the similarities are not sufficient to increase prediction accuracy. For example, RIPENCC consists of countries from both Western and Eastern Europe, with rather large differences in the behaviour of users and IT services. Even though the predictions from some individual countries such as Russia or France are not very accurate, their influence on the entire model is not sufficient to counter the better predictions of attacks from RIPENCC as a whole. In the case of the USA and ARIN, the prediction is good in both cases, since the USA is a major part of ARIN, reflecting and influencing the predictions of the entire RIR.

While predicting based on countries can be useful to researchers and CSIRT teams, it has its limits. It is important to realise it can only accurately predict for a relatively stable time period, a period during which no new rapidly spreading threat emerges, for example due to a newly found vulnerability. In such a case, it would cause a rapid drop in accuracy. With this in mind, it signifies the necessity to use rather short time periods for prediction to both maximise prediction accuracy and minimise the impact potentially new, rapidly spreading malware will have. Fortunately, the emergence of such new malware is not a very common situation.

From a statistical point of view, it is valid to say that in the case of some of the time series, the assumption that a non-stationary series is best predicted by methods using a differencing step, and vice versa was proven incorrect. This is probably caused by the unpredictable changes in the number of attacks, even though the prediction of the constant expected value, the variability of the series in time, and its weak stationarity, was often confirmed.

It was also established that even though some methods of prediction have a lower error value, such a prediction is less useful than another model with a higher error that more closely matches the curve of the real values over time. The latter models may be more successful with further application, as future attacks will likely not be constant either. The error of prediction in this experiment is highly dependent on the particular series, ranging from 20% to more than 1000%.

When it is considered that nearly all known approaches to time series prediction were applied, with many different settings, it is safe to conclude that the number of attacks is a rather hard series to predict, especially with an economic time series, for example. Another reason for the low achievement in the accuracy of the predictions, is due to the short length of the time series, not allowing for very accurate estimates of the parameters

for the used models. Namely those using the seasonal character of the data. Despite this, the series' APNIC, LACNIC, RIPENCC, France or AS16276 achieved the most accurate predictions using the seasonal model SARIMA. What was also surprising was the very good results obtained by the simple Holt-Winters algorithm, achieving the lowest error with the APNIC and China series.

Despite the large error values of some models, the results of the analysis of this experiment can be considered successful, as they helped to reveal other potential areas that should be researched further.

7. Conclusion and Further Research

The paper shows the possibilities and reliability of predicting attacks on a honeynet based on real-world data. The prediction was analysed as the overall attacks, and based on the source of the attacks from specific geographic locations. From a usability point of view, it could provide an analyst with useful predictions and information. It can also provide valuable, directly applicable information to CSIRT teams, mainly at a national level. In most cases, it will provide at least a useful short term prediction of the trends of attacks, often providing accurate predictions. The most accurate predictions were achieved with individual countries used as the source of attacks. The predictions with RIRs as sources and for the overall number of attacks on the honeynet were also acceptably accurate. The predictions with autonomous systems as the source were the least accurate.

The results of the analysis show that even despite using multiple methods and calibrating them, it is impossible to reach acceptable accuracy for all observed aspects. In most cases, the prediction accuracy is acceptable, given the length of the time series used. The methods of prediction using a seasonal component of the time series increase their efficiency with the growing number of seasons they have at their disposal. In the end, it is safe to conclude that using a time series to predict future attacks on a honeynet has proven to be beneficial and, in some cases, effective.

Further research in this area will be focused on the application of soft-computing methods for the prediction of a time series in the area of cyber-security, such as with neural nets.

References

1. Spotzner, L., Honeypots: Tracking Hackers, Addison Wesley Longman Publishing Co., Inc., USA (2002)
2. Joshi, C. R. and Sardana, A., Honeypots A New Paradigm to Information Security, Science Publishers, USA (2011)
3. Provos N., Holz T., Virtual Honeypots: From Botnet Tracking to Intrusion Detection, Addison Wesley Professional, USA (2007).
4. Ligh Hale M., Adair S., Hartstein B., Matthew R. Malware Analyst's Cookbook and DVD - Tools and Techniques for Fighting Malicious Code, Wiley Publishing, Inc, USA (2011)
5. Grudziecki T., Jacewicz P., Juszczyk Ł., Kijewski P., Pawliński P. and ENISA editors, Proactive Detection of Security Incidents Honeypots, ENISA publication, Greece (2012)
6. Abbasi, F.H., Harris, R.J. Experiences with a generation III virtual honeynet, Australasian Telecommunication Networks and Applications Conference, ATNAC 2009 - Proceedings, art. no. 5464785 (2009)

7. Balas, E., Viecco, C., Towards a third generation data capture architecture for honeynets, Proceedings from the 6th Annual IEEE System, Man and Cybernetics Information Assurance Workshop, SMC 2005, art. no. 1495929, pp. 21-28 (2005)
8. Sokol P., Gajdoš A., Prediction of Attacks Against Honeynet Based on Time Series Modeling, Silhavy R., Silhavy P., Prokopova Z. (eds) Applied Computational Intelligence and Mathematical Methods. CoMeSySo 2017, Advances in Intelligent Systems and Computing, vol 662, Springer (2017)
9. Condon, E., He, A., Cukier, M., Analysis of computer security incident data using time series models, 19th International Symposium on Software Reliability Engineering, ISSRE 2008, pp. 77–86, IEEE (2008)
10. Werner, G., Yang, S., McConky, K., Time series forecasting of cyber attack intensity, Proceedings of the 12th Annual Conference on Cyber and Information Security Research, p. 18. ACM (2017)
11. Wei, M., Kim, K., Intrusion detection scheme using traffic prediction for wireless industrial networks. J. Commun. Netw. 14(3), 310–318 (2012)
12. Tang, M., Alazab, M., Luo, Y., Exploiting vulnerability disclosures: statistical framework and case study, Cybersecurity and Cyberforensics Conference (CCC), pp. 117–122. IEEE (2016)
13. Zhan, Z., Xu, M., Xu, S., Predicting cyber attack rates with extreme values, IEEE Trans. Inf. Forens. Secur. 10(8), 1666–1677 (2015)
14. Soldo, F., Le, A., Markopoulou, A., Blacklisting recommendation system: using spatio-temporal patterns to predict future attacks, IEEE J. Sel. Areas Commun, 29(7), 1423–1437 (2011)
15. Thonnard O. snf Marc D., A framework for attack patterns' discovery in honeynet data, Digital Investigation, Volume 5, Supplement, S128-S139, ISSN 1742-2876 (2008)
16. Husák M., Bartoš V., Sokol P., Gajdoš A., Predictive methods in cyber defense: Current experience and research challenges, Future Generation Computer Systems, Volume 115, 517-530 (2021)
17. Husák M., Komárková J., Bou-Harb E., Čeleda P., Survey of attack projection, prediction, and forecasting in cyber security, IEEE Commun. Surv. Tutor. 21 (1) 640–660 (2019)
18. Sun N., Zhang J., Rimba P., Gao S., Zhang L. Y., Xiang Y., Data-driven cybersecurity incident prediction: A survey, IEEE Commun. Surv. Tutor. 21 (2) 1744–1772 (2019)
19. Zhang K., Zhao F., Luo S., Xin Y., Zhu H., An intrusion action-based IDS alert correlation analysis and prediction framework, IEEE Access 7 150540–150551 (2019)
20. Sokol, P. and Zuzčák, M. and Sochor, T., Definition of attack in context of high level interaction honeypots, Advances in Intelligent Systems and Computing, vol. 349, pp. 155-164 (2015)
21. Sokol, P. and Zuzčák, M. and Sochor, T., Definition of attack in the context of low-level interaction server honeypots, Lecture Notes in Electrical Engineering, vol. 330, pp. 499-504 (2015)
22. Zuzčák M. and Bujok P., Causal analysis of attacks against honeypots based on properties of countries, IET Information Security (2019)
23. Box, G. and Jenkins, G., Time Series Analysis: Forecasting and Control. Holden-Day, San Francisco (1970)
24. Kwiatkowski, D., Phillips, P.C., Schmidt, P., Shin, Y.: Testing the null hypothesis of stationarity against the alternative of a unit root: how sure are we that economic time series have a unit root? J. Econom. 54(1–3), 159–178 (1992)
25. Holt, C. C., Forecasting seasonal and trends by eponentially weighted moving averages. Res. Mem. no. 52. Pittsburg: Carnagie Institute of Technology (1957)
26. Winters, P. R., Forecasting sales by exponentially weighted moving averages. Management Science, vol. 6, p. 324-342 (1960).
27. Engle, R. F., Autoregressive conditional heteroscedasticity with the estimates of the variance of United Kingdom inflations. Econometrica, vol. 50, p. 987-1007 (1982)

28. Efron B. Bootstrap Methods: Another Look at the Jackknife. In: Kotz S., Johnson N.L. (eds) Breakthroughs in Statistics. Springer Series in Statistics (Perspectives in Statistics). Springer, New York, NY (1992)
29. Härdle W., Horowitz J., Kreiss J.-P., Bootstrap methods for time series, International Statistical Review 71, 435 – 459 (2003)
30. Chernick, M.R., LaBudde, R.A.: An Introduction to Bootstrap Methods With Applications to R. Wiley, Hoboken (2014)
31. Lahiri S. N., Resampling methods for dependent data, Springer-Verlag, New York (2003).

Matej Zuzčák earned his PhD in 2020 and he is currently working as a assistant professor and a researcher at the Department of Informatics and Computers of the Faculty of Science at University of Ostrava. His scientific research is focused mainly on honeypots, honeynets, network security, expert systems and data analysis. He has been the head of university CSIRT team - CSIRT OU since 2017. He is also a member of The Honeynet Project in Czech chapter.

Petr Bujok works as an Associate professor at the Department of Informatics and Computers of the Faculty of Science at University of Ostrava. His research area is mainly focused on the development and application of Evolutionary algorithms for global optimisation and applied statistics.

Received: July 15, 2020; Accepted: May 10, 2020.