IET Information Security

Research Article

IET Journals
The Institution of Engineering and Technology

Causal analysis of attacks against honeypots based on properties of countries

ISSN 1751-8709
Received on 18th April 2018
Revised 5th February 2019
Accepted on 25th February 2019
E-First on 2nd April 2019
doi: 10.1049/iet-ifs.2018.5141
www.ietdl.org

Matej Zuzčák[1] ✉, *Petr Bujok*[1]

[1]*Department of Informatics and Computers, Faculty of Science, University of Ostrava, 30 dubna 22, 701 03 Ostrava, Czech Republic*
✉ *E-mail: matej.zuzcak@osu.cz*

Abstract: This study studies the influence of country attributes on the number of secure shell attacks originating from it detected by the author's honeynet. Four statistical models are described, based on three sources of data from various countries. The studied attributes of the countries can be broadly divided into demographic, technological, and economic, with each source providing a slightly different set of attributes. Statistical methods such as partial least-squares path modelling are used, clustering countries by their assessed similarity. The population size has the greatest effect on the number of attacks, as expected, though it has to be noted that developing countries did not provide relevant data to the sources used and thus were not included. The following influential attributes were technical such as the access to information and communication technologies (ICT), and the use of ICT, with the economic influence being notable only in rather small countries. The Netherlands was an interesting anomaly, being clustered alongside large countries, even though its country attributes were very much like those of its neighbours.

1 Introduction

There is a current trend of constant increase in the number of attacks on servers providing Internet services, though many attacks are also aimed at home users, businesses, factories, and enterprises [1]. These detected attacks are aimed at various services directly accessible via the Internet, be it web pages and reduction systems, various Windows operating systems services such as the server message block (SMB) protocol, or protocols providing remote access to the system management. One of those is the secure shell (SSH) protocol. It allows the system to be controlled by anyone who passes the authentication and authorisation processes. This process is typically still based on verification of access name and password. On the basis of the device of the user whose name and password have been used to log into the appropriate user group, the system will allow it to execute various operations on the system. The focus is on the SSH protocol due to its widespread use. A lot of malware uses SSH. It is an easily exploitable protocol if the system administrator sets up a weak password. Commonly repeated input patterns via SSH shell can be used to identify spreading threats. Similar classification using other protocols is much more difficult or impossible due to the different nature of the information they carry.

Part of the aforementioned trend is attacked attempting to guess the access logins and passwords in order to gain access to the system, which can then be used for malicious activities. These activities vary, from data theft to abuse of the server as a proxy for other activities, namely as a part of a botnet used to conduct distributed denial of service attacks, spam spreading etc.

To learn more about these attacks and how to defend against them, we can use tools whose primary goal is to detect and analyse attacks rather than prevent them. Such tools are called honeypots. During an analysis, sooner or later a question of why attackers from certain geographical locations attack with certain frequency arises. This paper attempts to clarify the relation between the number of attacks and their countries of origin. It analyses the degree of impact of various indicators characterising the given country and how do these indicators influence the number of attacks on the given honeynet. It also examines the differences and similarities of the countries.

1.1 Honeypots and honeynets

The following section describes the basic information about the used tools, honeypots, and honeynets used to gain information about the attacks.

1.1.1 Honeypots: A honeypot represents a system whose primary goal is to analyse the activity conducted within it [2–4]. A honeypot can be implemented by any system, hardware component, or even an entire network. Usually, such a system is purposely left vulnerable, and it does not provide any real value to the attacker [5].

The system must be closed to such an extent that any activity undertaken within it cannot negatively influence its surroundings, or spread further over local area network, wireless area network, or the Internet. On the other hand, the system must also be sophisticated enough to allow strictly limited, controlled contact of the attacker with the outside. This is important since one of the characteristics of a good honeypot is to convince the attacker that the system is real, and he has to be able to use it without realising its security constraints. Finding a compromise between realism and security constraints is an important task when designing a honeypot. It is based mainly on what threats are the honeypot supposed to analyse. Therefore, honeypots further divide into groups by their interaction with the attacker [2, 3, 6]:

- *Low interaction*: emulation of vulnerable services. Single predefined action is executed after a connection.
- *Medium interaction*: honeypot offers more actions than a low interactive one. For example, it provides complex emulation of the offered protocol. Both categories are commonly joined into one in publications, based on the author's preference.
- *High interaction*: the attacker has access to the entire system with all its services and applications. It is most commonly implemented using virtualisation, allowing it to be restored to a default state and used again quickly.

On the basis of activity, they are divided into two types. The first is active (client) – searching for vulnerable systems and interacting with them. The second one is a server (passive) – passively providing vulnerable services on certain ports.

1.1.2 Honeynets: When using the term honeynet [6], it is important to specify the context in which it is used, as the term is also commonly used to describe high interaction honeypots as well. In this case, it means a specific network which, besides honeypot, also contains additional components such as a specialised firewall called honeywall. In this context, honeynet is a set of all these parts.

Another meaning of the term honeynet is a group of honeypots forming a logical network. This meaning of the term is most commonly used to describe groups of low- and medium-interaction honeypots, and it does not require the use of additional components such as firewalls. It is a common practise to gather data from all the honeypots of a honeynet into a central database. The added value of joining honeypots into a honeynet is in a higher relevance and volume of acquired data about the attacks. This data can be subsequently compared and analysed.

The main goal of this paper is to analyse what factors, whether geographical (size of countries, positions of countries, their population size etc.), social (e.g. how educated is the population), knowledge related [mainly in the information and communication technologies (ICT) area, but also the extend, and the type of the achieved education], or the approach of the users to solve ICT security problems and its effect, and the effect of these factors on the number of attacks detected by our honeypot. Gaining better image about the level of influence of the factors on the number of attacks allows designing and implementation of better defensive mechanisms, determining which areas to further analyse more deeply, or which areas to change (e.g. expanding education and security incident reporting, or increasing proactive activities. We would also like to determine countries that are more susceptible to specific attacks than others, with higher attack success rate, so an analysis of the causes can be carried out and the local Computer Security Incident Response Teams (CSIRTs) can be alerted to the situation and assisted with employing countermeasures.

2 Related works

There are many papers that either directly or indirectly deal with the issue of country indicators.

Paper [7] deals with the relation and correlation between the detected attacks and the indicators of their countries of origin. It divides the indicators into several categories such as the population view or the economic aspects. In conclusion, the magnitude of the correlation is divided into three primary levels: weak, moderate, and strong. Neither individual countries nor the reasons for the influence of the indicators are further explored. It only uses correlation and mainly uses population and economic indicators acquired from The World Bank (WB) database.

Lessons learned in the field of malware detection and analysis is the focus of another paper [8]. It deals with the acquisition of representative data to analyse spreading threats. Databases of VirusTotal, ANUBIS, and SGNET were used. The conclusion is that it is nearly impossible to acquire a representative sample of malware due to a dynamic increase in the number of samples, as well as due to errors in their detection by antiviruses. This paper also emphasises having well-defined data to work with.

Another paper [9] introduces a framework providing the methods for finding patterns and correlation between the attacks captured in the extensive data acquired by honeypots. Several techniques are described, primarily methods of cluster analysis for finding similarities, or 'Attack patterns', and for possible applications of time series to predict attacks.

A few other papers [10–12] primarily deal with a behavioural analysis of the detected attacks. The focus is on the analysis of the attacker's input such as input commands, or downloaded and executed scripts. Other papers attempt to predict an attacker's activity using time series such as in [13], which also adds the neighbourhood model and the singular value decomposition. Paper [14] analyses time periods such as days of the week or the hours of the day when the attackers are most active.

Most of the above-mentioned papers are analysing factors influencing attacks aimed at information systems in some way. Essentially only the papers [7, 14] connect them directly to honeypots. None of the papers, however, deals with the influence of the statistical factors in depth. The paper [7] does map various statistical indicators, but only from The WB database that only includes a limited number of countries and it lacks detailed overviews of indicators such as ICT use. Only correlation was used to evaluate the influence of the statistical indicators on the attacks, which is insufficient to conduct a detailed analysis. Moreover, the attacks used were only captured in a single campus network.

This paper uses data about attacks from honeypots deployed in various types of networks. Besides geographical indicators, it also analyses indicators pertaining to societal makeup and ICT use and solutions of the population. It also employs a more advanced statistical method, partial least-squares path modelling (PLS-PM), and constructs several models. Statistical indicators are drawn from the WB, Organisation for Economic Co-operation and Development (OECD), and Eurostat databases.

3 Technical solution

The technical solution is based on the implementation of a honeynet, followed by an analytical assessment of the gathered data.

3.1 Honeynet

During the measurement time, the honeypot consisted of seven sensors (honeypots). The honeypot implementation used in these sensors was Kippo, specifically the CZ.NIC modification. [15]. It is a medium-interaction honeypot emulating a typical Linux shell that can be accessed remotely. The emulated Linux distribution was Debian 7. Besides basic commands, Kippo also emulates a set of commands likely to be used by an attacker such as commands wget and secure copy protocol (SCP) for downloading files. Every file downloaded by the attacker is automatically sent to the VirusTotal service to be tested. [16]. Owing to the high level of emulation, it is rather easy for a human attacker to note that the system is emulated, which is a general issue with low- and medium-interaction honeypots. For instance, a human attacker can note that not all commands are implemented exactly as they are in a real Linux system. Therefore, it is a prime assumption that the vast majority of the detected attacks was carried out by robots or botnets. To gain access to the system, an attacker had to break a very simple password. It was always a typical combination of login and a password such as a root/toor. Every sensor was placed in a different type of network within Slovakia or the Czech Republic such as an academic network, or an internet service provider (ISP). The sensors were emulating mainly on the standard port 22, with an exception of one emulating on port 2222 which was placed within an autonomous system alongside a standard sensor. Specifics about the honeynet.

Honeynet consists of seven honeypots (sensors) emulating SSH server on port 22, with the exception of honeypot HP5-B where port 2222 is used instead:

- HP1: Czech academic network.
- HP2: Slovak academic network.
- HP3: Common Czech virtual private hosting (VPS) hosting grey zone.
- HP4: Common Czech VPS hosting.
- HP5: Czech ISP network.
- HP5-B: Czech ISP network.
- HP6: Slovak ISP network – dynamic Internet protocol.

All the sensors were accessible directly from the Internet, with no filtration by the Internet access providers. The measurement took place between February 2015 and February 2016, a period of 1 year.

The period was chosen because The WB, OECD, and Eurostat databases have data from this period available. It is prudent to mention these institutions only gather data in certain time periods, thus one with the most overlap among them and with our own data was chosen.

3.2 Captured data

The following attributes characterising the attack were used in evaluating attacks on the honeynet:

- *Sessions*: the term means the establishment of a connection at port 22 (2222 as explained above).
- *AllLogins (AllLog)*: after the session has been established, an attempt to log in can happen, which is measured in AllLogins quantity (number of all attempts to log in).
- SuccessLogins (SucLog): the login attempt was successful.
- UnsuccessLogins (UnSLog): the login attempt was unsuccessful.
- Inputs: any successful login can be followed by entries in the system, i.e. the number of all entered commands and other texts. Here, any text entered into the shell console is collected. Most commonly Linux commands, file downloading, and script execution can happen here.
- UniqInputs (UniqInp): a unique input of a specific command.
- *Files*: if a file or a script is downloaded (e.g. using wget command or SCP).
- UniqFiles (UniqFil): the number of unique downloaded files is evaluated based on the hash.
- Scripts: the number of Files or scripts execution.
- UniqScripts (UniqScr): the number of unique Files or scripts execution. The identity of files is evaluated here based on its name, not hash.

There were several cases when technical problems caused a longer period of probe inactivity. In these cases, such periods (e.g. the periods from 3 May 2015 to 27 June 2015 for HP2 and from 22 May 2015 to 21 July 2015 for HP6 were excluded from processing (not considered to produce zero activity). Detailed information about a behavioural analysis of the data can be found in the author's paper [17].

Basic characteristics of the above-mentioned quantities over all measurements taken are listed in Table 1 in our previous work [17].

3.3 Databases used to define the country indicators

To find relations between attackers and the number of their attacks necessary to find and process data characterising individual attackers. This paper approaches the attackers according to their country of origin. The countries are characterised by various statistical indicators. These indicators are usually tracked by international organisations, making them readily available for analysis.

Three databases were used in this paper: the WB statistical data [18], OECD [19], and Eurostat [20]. Subsequently, it was necessary to choose indicators with at least some assumed relation with the attacks captured by the honeynet and to confirm their validity using statistical methods and models.

3.3.1 World bank: The indicators are divided into three primary groups: population, technical, and economic. Table 1 lists the indicators used in models after they have been validated by the PLS-PM statistical method.

There were several reasons for the discarding. The first problem was missing data from some countries, with their derivation from other countries not being possible due to the differences between countries. Another problem was with the values of some indicators not fitting the model. The last major problem was with the values for indicators not having been recorded during the measurement time. All reviewed and discarded indicators are described in Section 6.

3.3.2 Organisation for Economic Co-operation and Development: Members of OECD provide more accurate and, more importantly, less missing prone data than The WB. However, many of the countries the honeynet was attacked from are either not OECD members or do not provide appropriate data, which prevents them from being analysed. The analysed indicators are listed in Table 2. Data for some indicators were not available from 2015; therefore, data from the closest available time spans were used. The table contains the three highest and several of the lowest values of the indicators. The number of the lowest values is given by the number of the same least non-zero values. At most, six of the lowest values of various countries are selected. In the case of the lower number of the lowest values, second and third least values of indicators are selected. The same rule is applied to Table 3.

Table 1 WB country indicators

Population aspects	Labels
population (2016)	pop
depth of credit information indexes 0–8 (low to high) – 2016	InfIndex
economic aspects	—
goods, exports % of total goods exports – 2015	GoodExp
goods, imports % total goods imports – 2015	GoodImp
technical aspects	—
fixed telephones – subscriptions per 100 people – 2015	FixTel
Internet use individuals using the Internet % of the population – 2015	IntUseIndiv
fixed broadband Internet subscriptions per 100 people – 2015	FixBroadIntSub

3.3.3 Eurostat: Most of the EU member countries provide very accurate and up-to-date statistical data to the Eurostat. While these are not countries with the highest number of attacks, they are also interesting for an analysis of the influence of other indicators related to organisations with defined ICT security measures. Specifically, these are those in ICT security in enterprises analysis [21]. The indicators of interest are:

- Enterprises having a formally defined ICT security policy, by economic activity, 2015 (% enterprises).
- Enterprises addressing all ICT security risks, for selected economic activities, 2015 (% enterprises).

4 Current knowledge of the data captured by honeynet

To arrive at valid conclusions about the influence of the statistical indicators of the countries on the attacks that originate from them, it is necessary to know the current trends of attacks and the similarity of the countries.

4.1 Similarity of countries based on the number of attacks that originates from them

The first, and rather straightforward, analysis of the captured data is a comparison of the number of attacks. Table 3 shows the activity of the countries by categories. The numbers in brackets represent the number of events recorded in our measurement. It is clear, most of events are for the United states, China, Canada, and the Netherlands. The results for the first three countries are rather expected, whereas high values of attacking indicators for proportionally smaller Netherlands are surprising. The names of countries are in the Alpha-2 format as per ISO 3166-1.

Three countries are at the top of nearly all of the indicators: USA, China, and the Netherlands. European countries, except the aforementioned Netherlands, are usually clustered around similar values in the middle. Less developed countries such as Egypt are at the bottom of the graph.

4.2 Similarity of countries based on the measured indicators by an application of cluster analysis

On the basis of the measured indicators, the countries from the paper [17] were subjected to two methods of cluster analysis, [22]:

Table 2 Values of OECD country variables

Variable	Min	Max
PopUpp	PT (2.38), DK (2.43), SK (3.77)	US (143.78), DE (48.13), FR (29.1)
PopTer	SK (1.19), DK (2.18), PT (2.45)	US (147.66), UK (30.18), KR (24.01)
secIncExp3m	NL (8.69), SK (9.09), AT (14.95)	MX (28.38), FR (28.36), DK (26.77)
caughtVirus3m	NL (5.57), SK (7.34), AT (12.25)	FR (25.32), PT (23.24), DK (22.4)
profiNets3m	MX (1.47), TU (2.25), SK (3.04)	NL (28.94), DK (28.47), SE (17.63)
eBanking3m	MX (6.96), GR (13.87), TU (15.04)	DK (84.89), NL (84.53), SE (79.64)
storSpace3m	PL (13.68), GR (17.56), KR (18.6)	DK (43.25), UK (42.85), SE (40.84)
purchase12m	MX (7.06), TU (15.39), IT (26.39)	UK (81.08), DK (78.88), DE (73.08)
transFiles12m	TU (25.66), IT (42.88), PL (43.4)	DK (70.86), NL (63.39), SE (63.08)
downSoft12m	TU (16.38), GR (19.76), SK (23.19)	DK (71.46), SE (67.99), NL (67.48)
BERD-GDP	MX (0.17), GR (0.24), PL (0.33)	KR (3.4), SE (2.31), DE (2.02)
GERD-GDP	MX (0.53), TU (0.88), GR (0.97)	KR (4.23), SE (3.28), AT (3.12)
GDP-per-head	MX (18.2), TU (24.6), GR (26.8)	US (56.4), NL (50.6), AT (49.2)
FixBroadSub	MX (5.31), SK (22.48), PL (27.86)	NL (200), DK (194.09), KR (181.32)
EgovReaIndx	MX (35.26), SK (37.18), PL (42.66)	KR (200), NL (187.68), UK (174.82)
WirBroadSub	TU (10.31), BE (23.72), DE (65.62)	KR (200), SE (189.62), DK (153.27)
AccCompHome	MX (44.91), TU (50.56), GR (68.57)	NL (96.2), DK (92.28), DE (90.99)
ICTAccInd	TU (53.75), MX (57.21), IT (67.86)	DK (96.57), NL (94.26), UK (92.71)
IntBroadAct	MX (33.7), GR (67), TU (68)	KR (98.5), NL (94), UK (90)

Table 3 Outliers of indicators of individual countries measured by the honeynet

Variable	Min	Max
sessions	DZ, BH, BE, DK, KW, LB (1)	US (11,971), CN (11,239), NL (11,224)
AllLog	DZ, BH, BE, DK, KW, LB (1)	CN (16,189), US (12,100), NL (10,865)
SucLog	DZ, BH, BE, DK, KW, LB (1)	US (8092), CN (6157), NL (5536)
UnSLog	AT (1), PT, AR, MX, SK (3)	CN (10,032), NL (5329), US (4008)
Inputs	KW (9), DZ, BH, BE, DK, LB (10)	US (53,491), CN (37,581), NL (35,057)
UniqInp	EC, MN, PY, PE, VN (3)	CN (2797), US (2527), NL (1291)
Files	PT (2), MK, ES (3), HK (5)	US (4073), NL (2262), LT (1105)
UniqFil	HK, MK, ES, TR (1)	US (329), NL (180), CA (52)
Scripts	SE (2), SK (6), BR (8)	US (13,097), NL (6175), CN (4509)
UniqScr	SE, TR (1), SK (2), IT, PT, UA (4)	US (579), CN (376), NL (314)

a hierarchical approach, so-called dendrogram with the single-linkage method, and a non-hierarchical approach, known as the *k*-means [23] algorithm. Task dimension reduction using the principal component analysis (PCA) [24] was applied for better visualisation. The result was the division of countries into three clusters:

- USA, China, and the Netherlands.
- Canada, Germany, United Kingdom, South Korea, Lithuania, and Russia.
- Remaining countries.

On the basis of these findings and from the results from a lower level, meaning the autonomous system data, it is possible to identify the common steps taken by the attacker inside the attacked system, prevalent trends of botnet connecting etc.

4.3 Similarity of countries based on the spreading of the threat

Paper [25] analyses the similarity of attackers based on the spreading of the thread, malware in this case. Several algorithms of binary clustering were used in the analysis. For the hierarchical approach, a dendrogram with Jaccard coefficient [26] was used, followed by a robust clustering using links (ROCK) algorithm for categorical attributes [27] and Proximus [28] algorithms. Owing to this approach, it was possible to observe specific malware and their trends, for instance, malware campaigns in certain countries spreading only within certain geographic locations, with the most significant being the combination of China and Taiwan. Malware executable and linkable format (ELF)*: ChinazN* was only identified in attacks originating in China and the USA. Questions of why certain countries are more active than others, why certain countries are so similar, and the question that the population has the most influence arise. The following section attempts to answer these questions.

5 Statistical processing methods

The first method applied for analysis of the relations between the statistical country indicators and the honeynet data was the canonical correlation analysis (CCA). However, it was later found out it is not suitable for this research. Therefore, the PLS-PM was used. Beside these methods, another two well-known statistical approaches are used to depict results. The first one is the PCA which was introduced by Hotelling [24]. It reduces the number of original variables into a smaller number of new components (typically two components for a scatter plot). The second method serves to cluster similar objects into independent groups (clusters). This algorithm is also known as the Harrigan's algorithm and as *K*-means [23].

5.1 Canonical correlation analysis

CCA is a statistical exploratory method for detecting correlations between two data sets [29, 30]. For two matrices (datasets) Y and X of order $n \times p$ and $n \times q$. The number of both matrices' rows (n) is equal and has to be greater than the number of columns $n \geq p$ and $n \geq q$. Columns of $\boldsymbol{Y}$ and $\boldsymbol{X}$ are standardised, i.e. mean value is 0 and the standard deviation is 1.

The task of the CCA lies in the iterative approach of finding vectors $\boldsymbol{a}^i = (a_1^i, a_2^i, \ldots, a_p^i)^{\mathrm{T}}$ and $\boldsymbol{b}^i = (b_1^i, b_2^i, \ldots, b_q^i)^{\mathrm{T}}$ that

maximise the correlation between the following linear combinations:

$$\boldsymbol{U}^i = \boldsymbol{X}\boldsymbol{a}^i = a_1^i\boldsymbol{X}^1 + a_2^i\boldsymbol{X}^2 + \cdots + a_p^i\boldsymbol{X}^p \quad (1)$$

and

$$\boldsymbol{V}^i = \boldsymbol{Y}\boldsymbol{b}^i = b_1^i\boldsymbol{Y}^1 + b_2^i\boldsymbol{Y}^2 + \cdots + b_q^i\boldsymbol{Y}^q \quad (2)$$

i.e.

$$\rho_i = \mathrm{cor}(\boldsymbol{U}^i, \boldsymbol{V}^i) = \max(\mathrm{cor}(\boldsymbol{X}\boldsymbol{a}, \boldsymbol{Y}\boldsymbol{b})) \quad (3)$$

where $i = 1, 2, \ldots$ and $\boldsymbol{U}^i$, $\boldsymbol{V}^i$ are called ith *canonical variables*.

Correlation between the canonical variables $\boldsymbol{U}^i$ and $\boldsymbol{V}^i$ is from [0, 1], where the best model is evaluated by value 1 and vice versa. In the case of our data sets, the correlation between the first canonical variables ($\boldsymbol{U}^1$, $\boldsymbol{V}^1$) was $\rho_1 = 0.9985$ and for the second ones $\rho_2 = 0.9536$. Although obtained ρ_1, ρ_2 values are very promising for the quality of CCA model, many researchers and statisticians recommended studying correlations between the individual canonical variable (i.e. $\boldsymbol{U}$ and $\boldsymbol{V}$) and the original variables [see (1) and (2)].

The original variables (i.e. $\boldsymbol{X}$s and $\boldsymbol{Y}$s) which correlation with their canonical correlation is not significant (the null hypothesis H_0: cor($\boldsymbol{U}$, $\boldsymbol{X}$) = 0 is not rejected) have to be removed from the CCA model. This phenomenon was the main problem of using the results of the CCA model because most of the correlations between the canonical variable and the original variables were not significant. In the case of the $\boldsymbol{Y}$ dataset, the only variable *Files* was not significantly correlated with its canonical variable. However, in the second ($\boldsymbol{X}$) dataset, only for the *Pop* variable the null hypothesis of 'correlation equal to zero' was rejected.

This was the reason to apply another – very similar method to detect the relation (dependency) between two data sets, called PLS-PM.

5.2 Partial least-squares PM

PLS could be considered as a family of various methods and approaches. A first analytic tool based on PLS was introduced in 1970 by Herman Wold [31] at the University of Uppsala in Sweden. Subsequently, Svante Wold continued the development of PLS methods and their applications in chemistry data analysis [32]. PLS approach can be mentioned as a set of methods used discern relations between blocks of variables. Usually, independent variables are located on one side and dependent variables which would be influenced by them on the other side. This categorisation is not essential and necessary. Currently, approaches of PLS are divided into two fields:

- Regression-based approach.
- PLS-PM-based approach (PLS-PM).

The PLS-PM was not very popular in the 1990s, and it was used primarily in the USA. At the end of the 1990s, the popularity of PLS-PM in Europe increased due to a group around Michel Tenenhaus resulting in various publications such as [33]. The popularity of PLS-PM in statistical research remains high to these days. Gaston Sanchez from the University of Berkeley [34] with extensive experience with PLS-PM uses it for regression in the fields of chemometrics, sensometrics, and biometrics. PLS-PM is also prevalent in social sciences such as psychometrics, marketing, economics, and information technologies.

When we analysed both of these methods (CCA and PLS-PM), we decided to apply the PLS-PM approach. The main reason was that the data was from the field of information technologies, social, and economic aspects. Also, application of PLS-PM principles based on the measured manifest and latent variables of our data appears to be a good choice, further described in Section 6. Another benefit is in the advanced methods of data visualisation.

This section is only a brief introduction to PLS-PM describing keywords used in the further text. More details of PLS-PM approach are available in [34]. At first, variables inputted into our data model represented real, measured values. The measured variables are denoted *manifest* variables (or *indicators*) and are located in 1, 7, and the list in Section 3.3.3. Beside manifest variables, *latent* variables are used in PLS-PM approach. These variables are not directly measured, but they are influenced by the directly measured variables. For example, the success of an army depends on attack and defence. Note, the attack and the defence are called latent variables because they are not measured directly, and they are loaded by the manifest variables. In the case of an army operation, for instance in the case of Suppression of Enemy Air Defenses, the main role will be played by the number and the types of aeroplanes, their appropriate armaments and tactics on the attacking side, and by the number and the disposition of radars and surface-to-air missile sites available for defence. In our model, latent variables will be constructed in various ways depending on the currently designed model (Section 6). Latent variables are specified in two different ways: formative indicators (its indicators are considered to be the causation of latent variable) and reflective indicators (they are consequences and reflect impacts of its responding indicators).

Several blocks of manifest and latent variables are designed and they comprise the *path model*. Generally, full path model is composed of two sub-models: from the *inner model* (structural model) and *outer model* (measurement model). The inner model is composed of the relations between the latent variables. The outer model is more complex and it adds a view of relations between each latent variable and its corresponding block of indicators (manifest variables). The inner model describes structural relations between latent variables in a linear way. Mathematically, these relations are represented as follows (4) in a recursive form:

$$\boldsymbol{LV}_j = \beta_0 + \sum_{i \to j} \beta_{j,i}\boldsymbol{LV}_i + \mathrm{error}_j \quad (4)$$

$\boldsymbol{LV}_j$ is the latent variable; $\boldsymbol{LV}_i$ represents all latent variables for prediction of LV_j in current inner model (called *predictor*); $\beta_{j,i}$ are *path coefficients* representing degree and direction of dependence between a given $\boldsymbol{LV}_j$ and predictors $\boldsymbol{LV}_i$; β_0 represents intercept; and error_j means residuals.

For an estimate of the structural coefficients $\beta_{j,i}$ has used ordinary least-squares method of the multiple regression of $\boldsymbol{Y}_j$ on the $\boldsymbol{Y}_i$'s related with it as below:

$$b_{j,i} = (\boldsymbol{Y}'_i\boldsymbol{Y}_i)^{-1}\boldsymbol{Y}'_i\boldsymbol{Y}_j \quad (5)$$

An important aspect of the inner is that relations are modelled by a standard linear regression as below:

$$E(\boldsymbol{LV}_j \big| \boldsymbol{LV}_i) = \beta_{0,i} + \sum_{i \to j} \beta_{j,i}\boldsymbol{LV}_i \quad (6)$$

if and only if as below:

$$\mathrm{cov}(\boldsymbol{LV}_j, \mathrm{error}_j) = 0 \quad (7)$$

In general, the outer model can be based on the reflexive or formative type of latent variables. Details are available in Section 5.2. In this paper, all outer models are constructed in a reflexive design. Relations between variables in the outer model are considered in a linear-regression way as in inner model and in the case of reflexive latent variables are defined as (8)

$$\boldsymbol{X}_{j,k} = \lambda_{0,j,k} + \lambda_{j,k}\boldsymbol{LV}_j + \mathrm{error}_{j,k} \quad (8)$$

$\boldsymbol{X}$ denotes the matrix of n observations with p variables (matrix of $n \times p$), which is into j mutual blocks, where each block consists of k variables. Specification $\boldsymbol{X}_{j,k}$ poses kth variable from jth block of the origin matrix $\boldsymbol{X}$. In contrary to CCA 5.1, matrix $\boldsymbol{X}$ here contains

all variables from both datasets. Coefficients $\lambda_{j,k}$ are called loadings and λ_0 represents intercept of the model.

Loadings, for convenience and simplicity, represent correlations between the latent variable and its corresponding indicators as below:

$$\widehat{\lambda_{j,k}} = \mathrm{cor}(\boldsymbol{i}_{j,k}, \boldsymbol{Y}_j) \tag{9}$$

where linear relations are designed using standard linear-regression model as below:

$$E(\boldsymbol{X}_{j,k} \big| \boldsymbol{LV}_j = \lambda_{0,j,k} + \lambda_{j,k}\boldsymbol{LV}_j) \tag{10}$$

A very important role in PLS-PM modelling is played by *scores*. The latent variables are estimated as a linear combination of its corresponding indicators – manifest variables. The estimate of the latent variable $\boldsymbol{LV}_j$ is called *score* and is denoted $\boldsymbol{Y}_j$ as below:

$$\widehat{LV_j} = Y_j = \pm f_j \sum_k \tilde{w}_{j,k} X_{j,k} \tag{11}$$

where $\tilde{w}_{j,k}$ represents *outer weights* and f_j represents the scalar necessary for standardisation of $\boldsymbol{Y}_j$. Outer weights are initialised randomly and subsequently, the outside approximation of the latent variables is performed which initial weights are in the order to approximation latent variables as a linear combination of their indicators. Thereafter, inner relations between the latent variables are used to obtain inside approximations. This process could be performed by three different scenarios [33].

A basic idea of PLS-PM approach is to combine manifest variables from each block corresponding to a certain latent variable and compute, or estimate, the corresponding latent variable. After the estimation of the scores, the path coefficients and the loadings are computed. PLS-PM is an iterative approach divided into three independent phases:

- 1 Computation of weights and scores for latent variables.
- 2 Estimation of path coefficients (inner model).
- 3 Computation of loadings (outer model).

5.2.1 Indicators of PLS-PM results: All results in this paper were obtained by R programming language [35] specifically the plspm package [36]. With respect to the blocks of reflexive latent variables tests of unidimensionality in blocks are performed, where indicators have to be in a space of single dimension while they partially indicate the same latent variable. Verification of the designed models is realised by three classification indicators:

- Cronbach's alpha: block is considered unidimensional if $\alpha_C > 0.7$.
- Dillon–Goldstein's ρ: block is considered unidimensional if $\rho_{DG} > 0.7$.
- First eigenvalue (eig.1st) of the indicator's correlation matrix: block is considered unidimensional if the first eigenvalue is bigger than 1 and the second eigenvalue (eig.2nd) is <1.

Values of estimated loadings 5.2 and communalities (squared correlations) are also verified. The main effort is to obtain as big values of correlations as possible. Generally speaking, loading higher than 0.7 is considered acceptable, though of course, the decision depends on the current model and subjective view of the analysts.

Further indicator for evaluating models is *weighed*. The higher the value of this indicator is achieved, the better the model is. The proper values are analysed in order according to the model. Note that weights of indicators of one latent variable in the same outer model should not be negative. Cross-loadings are the loadings of a certain indicator with the rest of the latent variables.

The idea of cross-loadings is to compare the shared variance between a construct and its indicators with the shared variance with other constructs. Each indicator should load highest on the construct it intends to measure.

Table 4 Dependency based on R^2

Dependency	R^2
low	$R^2 < 0.30$
moderate	$0.30 > R^2 < 0.60$
high	$R^2 > 0.60$

An inner model is suitable to assess the whole internal part of the model. The most important is the coefficient of determination R^2 of such a latent variable which indicates the dependent part of the model by another group of latent variables. The value of R^2 is typically shown in Table 4.

The GoF index is a pseudo goodness of fit measure that accounts for the model quality at both the measurement and the structural models. GoF is calculated as the geometric mean of the average communality and the average R2 value. There is no given limit of a proper GoF value, but within the PLS community, the most common GoF values should be over 0.7 [34].

Bootstrapping is a non-parametric technique for validation of the model. The bootstrap procedure is the following: M samples are created in order to obtain M estimates for each parameter in the PLS model. Each sample is obtained by sampling with replacement from the original data set, with a sample size equal to the number of cases in the original data set [34].

6 Result analysis

As mentioned previously, this analysis works with multiple sources of statistical indicators of countries which were the origin of the attacks against the honeynet. Each of the sources contains a set of different countries with different indicators that partially overlap. Therefore, an adequate model for each database was made to ease access to its data. Latent variables were then derived from manifest variables for each model, with the effect of these latent variables on the latent variable Attack being most important. The latent variable attack is derived from data about attacks captured by the honeynet further described in Section 3.2, and it is used in all models. It is derived from the same manifest variables in every model, with the latent variables of each model having a different effect on it.

The manifest variables are grouped into logical sets, with each set representing a latent variable. The grouping was done by moving manifest variables between groups or making new latent variables in a way to maximise the influence of the latent variables on the attack variable. It is imperative to achieve the highest possible values of loadings, weights, and indicators in the unidimensionality test and the overall R^2 of the given model. The goal is to find the best models representing the number of attacks in relation to their countries of origin.

The models were reduced or expanded in a way to get the best indicator results. It is a rather tedious task requiring recalculation of each new model. Originally, all the relevant indicators from the databases were added according to the logic of the aforementioned groupings, though the values of loadings and weights were usually very low. The assessment was done as described in Section 5.2.1. It was attempted to have all the result values above 0.7, which is a recommended value as mentioned in Section 5.2.1, though some were deemed relevant with lower values, and they are pointed out in further sections. Besides loadings and weights, unidimensionality and R^2 of models were assessed. After acceptable values of loadings, weights, unidimensionality, and R^2 were achieved, additional tests as described in Section 5.2.1 were carried out and conclusions based on models were reached.

6.1 The WB model

The first assessed model was the one analysing relations between the latent variables derived from The WB data, described in Table 1, and the attack latent variable.

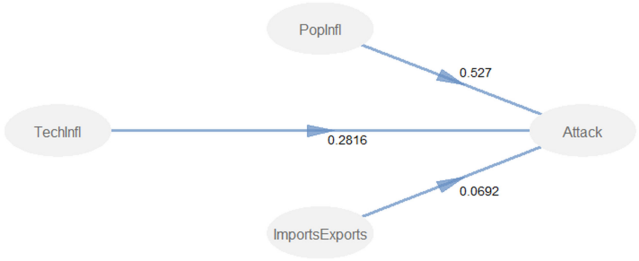

Fig. 1 *PLS-PM model based on the WB data*

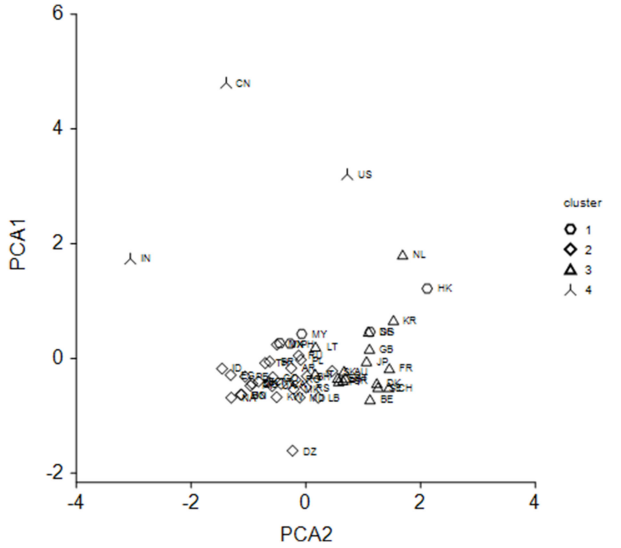

Fig. 2 *Visualisation of individual countries using the two main components of the PCA methods and the k-means division based on the scores of the WB model*

Through gradual testing of various combinations of manifest variables to form the most appropriate latent variables, three latent variables emerged, see Fig. 1. The coefficient of determination of the model, which is a dependence of the explained variable on the variables explaining it, is 0.329. The value is acceptable with moderate dependency according to Table 4. The population influences (PopInfl) of the given country have the highest impact on the model, followed by technical influences (TechInfl), and by the import and export influences (ImportsExports). Loadings of all the manifest variables forming the latent variables are higher than 0.75, see Table 5, with the exception of the information index (InfIndex), which was added by the author's discretion, due to the importance of ICT knowledge of the population. Therefore, all the used manifest variables sufficiently fulfil the corresponding latent variables. In regards to population influences, the size of the country's population clearly has a dominating influence. The weight values of the PLS-PM model are shown in Table 5. GoF index value is 0.529. A non-parametric test (bootstrapping test) with 100 iterations always had loading values higher than 0.75, except for the aforementioned information index. The unidimensionality tests of the latent variables were over the 0.70 marks, except the population influences for α_C with the value 0.211, and for ρ_{DG} with the value 0.717. The first eigenvalue was always above 1 and the second eigenvalue was always below 1.

Fig. 2 shows the incorporation of countries into clusters after the application of the *k*-means algorithm on the scores of the countries gained from the application of the PLM-PS method. The analysis shows how China, India, and the USA have distinctively higher values than other countries. Fig. 3 shows the countries that were the point of origin for most of the attack. Fig. 4 is a result of the *k*-means algorithm application, with visualisation done by applying PCA on the manifest variables of the attack variable. The three prominent countries also commonly form a group after application of the *k*-means algorithm on the manifest variables of the rest of the latent variables. Most of the countries in this figure and also in Figs. 5 and 6 are located in a very small area.

Table 5 Loadings and weights values for 6.1 model

ID	Manifest variables	Latent variables	Weight	Loading
1	Pop	PopInfl	0.874	0.920
2	InfIndex	PopInfl	0.392	0.496
3	IntUseIndiv	TechInfl	0.246	0.897
4	FixBroadIntSub	TechInfl	0.519	0.975
5	FixTel	TechInfl	0.297	0.911
6	GoodsExp	ImportsExports	0.455	0.979
7	GoodsImp	ImportsExports	0.560	0.986
8	AllLog	attack	0.129	0.973
9	SucLog	attack	0.114	0.998
10	UnSLog	attack	0.135	0.883
11	Sessions	attack	0.119	0.989
12	Inputs	attack	0.114	0.995
13	UniqInp	attack	0.124	0.962
14	Files	attack	0.058	0.816
15	UniqFil	attack	0.068	0.874
16	Scripts	attack	0.083	0.933
17	UniqScr	attack	0.104	0.983

Therefore, the same pictures without outlying countries are inserted into these figures to increase of readability of this paper.

Besides assessing scores, it was also possible to apply the response-based unit segmentation (REBUS) [37] algorithm. REBUS is used to find so-called latent classes within the main/global PLS-PM model. Unlike a score-based analysis, REBUS also considers structural relations within the PLS-PM model. It can find classes of objects on which is it more appropriate to base a specific local model with better loadings, weights, and R^2 values. For specifics, see [37].

The algorithm allowed the creation of two local models by dividing countries into two subgroups, increasing the mutual influence of latent variables of the model. The first local model, shown in Fig. 7, achieves a much better coefficient of determination than the global model, specifically a value of 0.816. It includes the following countries: Canada, China, India, Lithuania, Netherlands, and the USA. Fig. 7 shows that in the countries with high population, the influence of the technical and economic influences rises. From an analytical point of view, the addition of Lithuania is rather surprising, though when looking at Fig. 3 it becomes apparent that it was among the highest number of attacks originating from it. While analysing loadings, a high increase of the loading value for the information index in relation to the population count (0.998 and 0.292) stands out. This means that the latent variable is influenced much more by the country's information index than its population.

The second sub model achieves the coefficient of determination of 0.402 and it includes all the remaining countries. The model is represented by Fig. 8. In this case, the population size is again dominant, though the information index still maintains some influence, while economic aspects have a negative influence.

Using The WB model, which as the only one at least partially covers Asian and African countries, this paper concludes that population count is the prime factor influencing the number of attacks. However, at a closer look at the high population countries, technical aspects, and access to ICT resources increase their influence. This makes logical sense since the third world countries are not sources of a massive number of attacks due to their lack of lack of access to ICT. It is unfortunate that The WB does not track additional statistical indicators such as the behaviour of Internet users. There is also quite a high amount of missing values in the analysed data.

6.2 Model based on the OECD data

Data of OECD member countries (Table 6) contain more accurate indicators than The WB. Nonetheless, even some of these countries do not provide all the information, even developed countries such as the USA. The following countries are OECD members that were

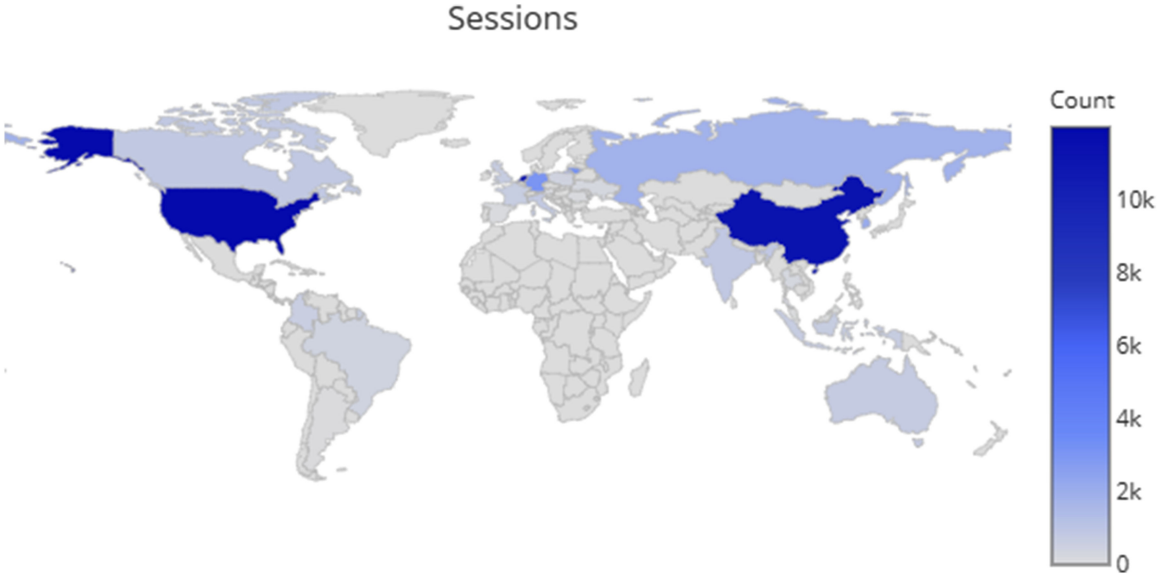

Fig. 3 *Number of attacks against the honeynet as per their countries of origin*

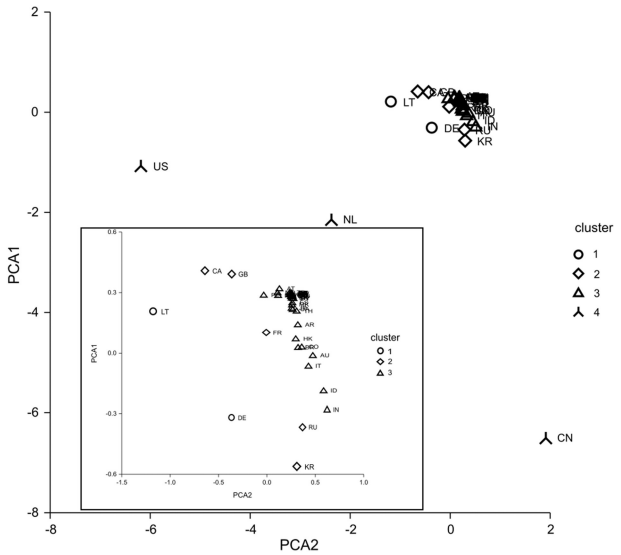

Fig. 4 *Visualisation of countries by manifest variables of the attack latent variable using the two main components of the PCA methods and the k-means division*

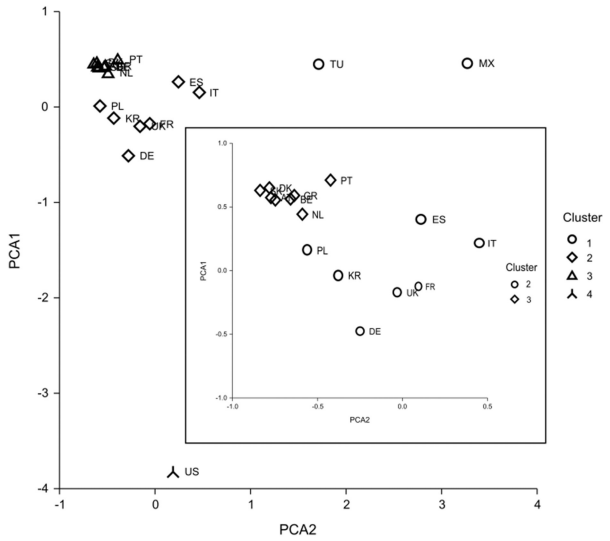

Fig. 5 *Visualisation of individual countries by manifest variables of population influence latent variable based on the two main components of the PCA method using k-means*

also countries of origin for the attacks: Austria, Belgium, Denmark, France, Germany, Greece, Italy, Korea, Mexico, Netherlands, Poland, Portugal, Slovakia, Spain, Sweden, Turkey, United Kingdom, and the USA.

Model represented by Fig. 9 shows selected loading and weight values Table 7. The only manifest variable with loading lower than 0.75 is PopBell, though it was still used since, in this case, the population was divided by education to not to lose the complex image of the population division. The coefficient of determination had a value of 0.690, which means a rather high influence of the latent variables of the model on the attack variable. The GoF index had a value of 0.7422. The unidimensionality tests all have values above 0.75, the first eigenvalues are all above 2, and the second eigenvalues have a maximum value of 0.83, which is overall a good result. Bootstrapping test of loadings with 100 iterations achieved average values above 0.70.

The model available in Fig. 9 shows the population has the highest influence on the attack variable. The economic aspects have low negative influence. The access to ICT such as access to the Internet have a rather low, but notable influence.

The PLS-PM model (Fig. 9) assigned a score to every country, which was then used as the input of the *k*-means algorithm to generate the clusters in Fig. 10. PCA was used for visualisation. There are four visible clusters. The USA is alone in the first, most of the western European countries and South Korea in the second, the rest of the western European countries in the third and the southern and eastern European countries and Turkey in the fourth. The country clusters are very similar to those resulting from *k*-means application on the latent variable Economics as seen in Fig. 11, with the exception of the USA where the population influence set it aside as seen in Fig. 5.

6.3 OECD member countries without the USA and Mexico

OECD provides data about the use of ICT (ICTUse) such as the use of e-banking, downloading files etc. It also provides the experience level of the population with security aspects such as with captured computer viruses. Unfortunately, this data is not available for the USA and Mexico, which is why they are not used in this model. The model, as seen in Fig. 12, contains additional latent variables SecInfl and ICTUse. Table 7 shows the selected manifest variables they are derived from. The coefficient of determination has a value of 0.564, which is moderate according to Table 4. Unidimensionality tests achieved values higher than 0.75. The first eigenvalue was always above 1 and the second had a maximum of 0.80. GoF has a value of 0.6969. Bootstrapping test of 100 iterations achieves loadings higher than 0.75 on average. The model shows that removing the USA and Mexico decreased the influence of population, while also made the SecInf negative,

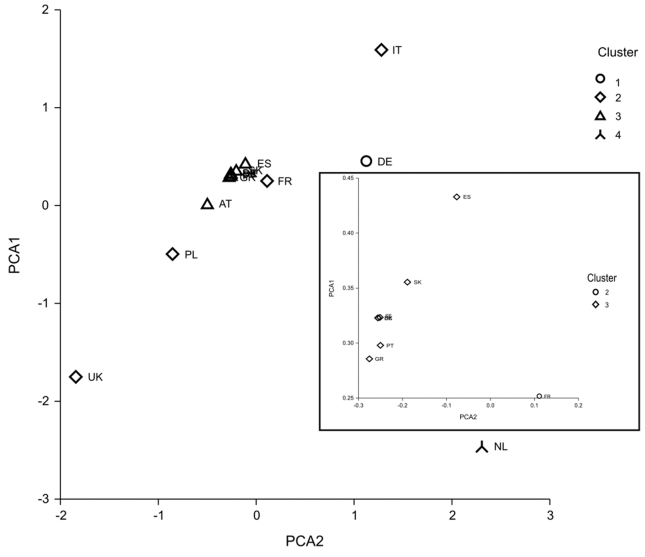

Fig. 6 *Visualisation of individual countries by manifest variables of attack latent variable based on the two main components of the PCA method using k-means*

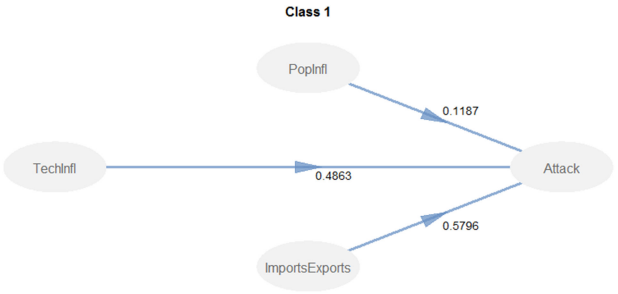

Fig. 7 *First local model based on WB data – Class 1*

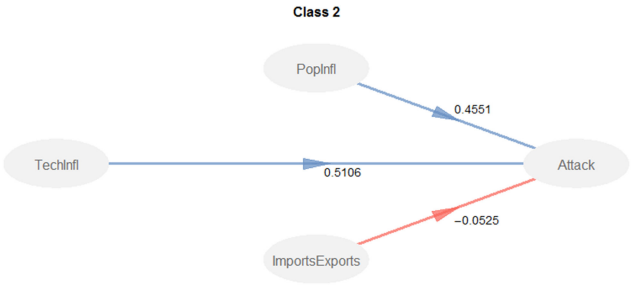

Fig. 8 *Second local model based on WB data – Class 2*

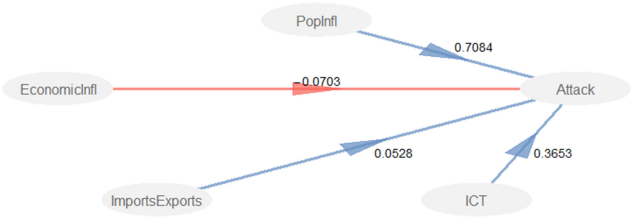

Fig. 9 *PLS-PM model based on OECD data*

meaning the users realise the security risks of the ICT use. If the user detects a computer virus, he is also likely aware of the PC's overall state, and since the attacks are mostly from botnets, is also quite likely to prevent the PC to be used by a botnet. High ICTUse also has a high influence, since the higher the number of ICT users, the more likely their personal computers (PCs) are to be infected, increasing the rate of attacks of the country. The economic aspects also have a rather high negative influence on the attack. The real ICTUse has a much higher influence than the availability of ICT (manifest variable ICT).

Table 6 OECD country indicators

Population aspects	Labels
below (2016)	PopBell
upper (2016)	PopUpp
tertiary (2016)	PopTer
economic aspects	—
business enterprise expenditure on R&D – % of GDP (2015)	BERD.GDP
gross domestic expenditure on research and development – % of GDP (2015)	GERD.GDP
gross domestic product per head (2015)	GDP.per.head
imports/exports aspects	—
total import computers – mil. USD (2015)	TotImpComp
total export computers – mil. USD (2015)	TotExpComp
imports goods – growth in % (2015)	ImpGoods
export goods – growth in % (2015)	ExpGoods
ICT aspects	—
ICT access and usage by households and individuals in the last 12 m, 2015 (%)	ICTAccInd
access to computers from home % of all households (2015)	AccCompHome
E-government readiness index 0–1 (2012)	EgovReaIndx
networks autonomous systems by population – is expressed per million inhabitants (2012)	NAS
wireless broadband subscribers – by population (2012)	WirBroadSub
fixed broadband subscribers – per population (2012)	FixBroadSub
Internet broadband access in % (2015)	IntBroadAct
ICT security aspects % (2015)	—
% of individuals have experienced security incidents in the last 3 months	secIncExp3m
% of individuals have caught a virus or other computer infection in the last 3 months	caughtVirus3m
ICTUse aspects % (2015)	—
% of individuals participating in professional networks in the last 3 months	profiNets3m
% of individuals using Internet banking in the last 3 months	eBanking3m
% of individuals using the Internet for playing networked games – last 3 months	gamePlay3m
% of individuals who have installed or replaced an operating system – last 12 months	instOS12m
% of individuals who have used spreadsheet advanced functions – last 12 months	advSprShFo12m
% of individuals using Internet storage space in the last 3 months	storSpace3m
% of individuals have purchased online in the last 12 months	purchase12m
% of individuals have transferred files in the last 12 months	transFiles12m
% of individuals download and install software from the Internet in the last 12 months	downSoft12m

The PLS-PM model based on the OECD data without the USA and Mexico assigned a score to every country, which was then used as the input of the *k*-means algorithm to generate the clusters in Fig. 13. PCA was used for visualisation. Within Europe, the Netherlands stand out, as it was the origin of a high number of attacks throughout the entire measurement time. The most attacks of all of Europe originated from the Netherlands, as seen in Fig. 3. There is also, again, the division of most of western Europe and South Korea in the first cluster, the rest of western Europe in the second cluster, and southern and eastern Europe in the third cluster alongside Turkey. The Netherlands is truly an anomaly in the number of attacks since most of the OECD indicators are just like those of other western countries.

Table 7 Three biggest and lowest weight and loading values for models 6.2, 6.3, and 6.4

m	ID	Manifest variables	Latent variables	Weight	Loading
6.2	6	GDP.per.head	EconomicInfl	0.598	0.924
	8	TotExpComp	ImportsExports	0.598	0.888
	3	PopTer	PopInfl	0.501	0.991
	20	UnSLog	Attack	0.092	0.917
	17	IntBroadAct	ICT	0.089	0.789
	1	PopBell	PopInfl	0.047	0.362
	19	SucLog	Attack	0.104	0.999
	22	Inputs	Attack	0.103	0.998
	24	Files	Attack	0.106	0.992
	15	WirBroadSub	ICT	0.118	0.717
	7	TotImpComp	ImportsExports	0.178	0.521
	1	PopBell	PopInfl	0.047	0.362
6.3	13	GDP.per.head	EconomicInfl	0.752	0.958
	4	caughtVirus3m	SecInfl	0.622	0.928
	1	PopUpp	PopInfl	0.612	0.96
	28	Scripts	Attack	0.093	0.986
	29	UniqScr	Attack	0.088	0.984
	16	WirBroadSub	ICT	0.017	0.512
	21	SucLog	attack	0.104	0.997
	23	Sessions	attack	0.103	0.996
	27	UniqFil	attack	0.103	0.995
	12	GERD.GDP	EconomicInfl	0.166	0.812
	11	BERD.GDP	EconomicInfl	0.186	0.769
	16	WirBroadSub	ICT	0.017	0.512
6.4	16	GDP.per.head	EconomicInfl	0.81	0.992
	22	ICTSecPol	OrgSec	0.615	0.988
	1	PopUpp	PopInfl	0.592	0.96
	15	GERD.GDP	EconomicInfl	0.09	0.895
	8	storSpace3m	ICTUse	0.084	0.864
	13	instOS12m	ICTUse	0.024	0.808
	25	SucLog	Attack	0.101	0.999
	31	UniqFil	Attack	0.104	0.997
	24	AllLog	Attack	0.1	0.997
	11	transFiles12m	ICTUse	0.095	0.846
	13	instOS12m	ICTUse	0.024	0.808
	10	advSprShFo12m	ICTUse	0.129	0.75

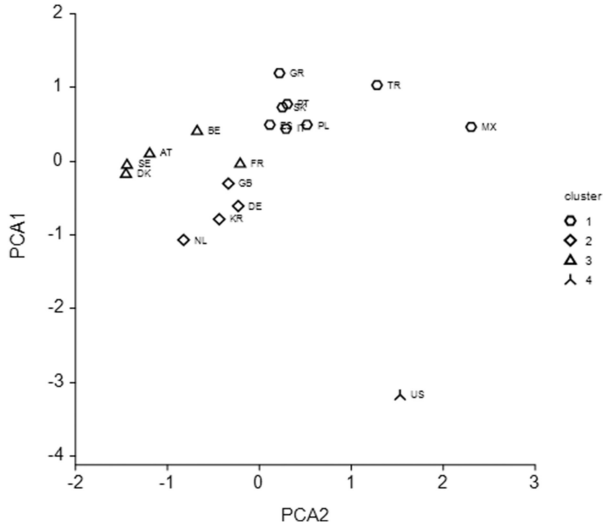

Fig. 10 *Visualisation of the country clustering based on the two main components of the PCA method using k-means based on the scores of the OECD model*

There is a single latent variable that deviates significantly for the Netherlands, the SecInfl variable. The difference between the Netherlands and the rest of Europe can be seen in Fig. 14 representing a *k*-means application on manifest variables *SecIncExp3m* and caughtVirus3m from which SecInfl is derived. The difference is also apparent in Table 2, where, in the number of malware captured by the population (*caughtVirus3m*) and the experience of the population with security incidents (*secIncExp3m*), the Netherlands achieves the lowest values of these indicators. This suggests different habits in dealing with cyber threats, which may be a part of the explanation of the anomaly.

6.4 Extension of the model for the EU member states (Eurostat)

The last model shown in Fig. 15 only operates with the EU countries. Compared to the previous model, it adds a new latent variable, OrgSec, described in Section 3.3.3, based on the added Eurostat variables. The coefficient of determination is 0.641. Unidimensionality tests are all above 0.75, most above 0.9. The first eigenvalue is always above 1 and the second has a maximum of 0.45. GoF has a value of 0.7558. Bootstrapping test with 100 iterations always averages over 0.80.

The new latent variable OrgSec allows analysis of cyber security in European organisations. Its effect is negative, as it logically should be, since decreasing the number of threats, in general, is a goal of any security organisation.

The effect of the population is rather low, while security experience (SecInfl) is moderately negative, and the ICTUse has the highest influence, making the conclusions of this model very similar to those of the last one.

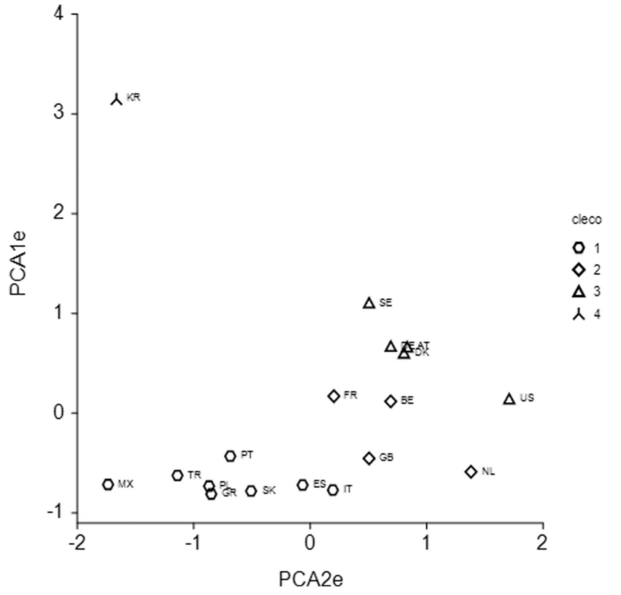

Fig. 11 *Visualisation of individual countries by the manifest variables of the economy latent variable based on the two main components of the PCA method using k-means*

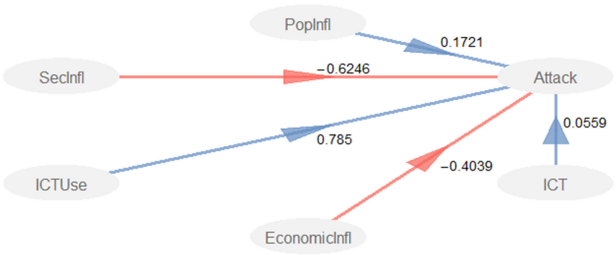

Fig. 12 *PLS-PM model based on OECD data without the USA and Mexico*

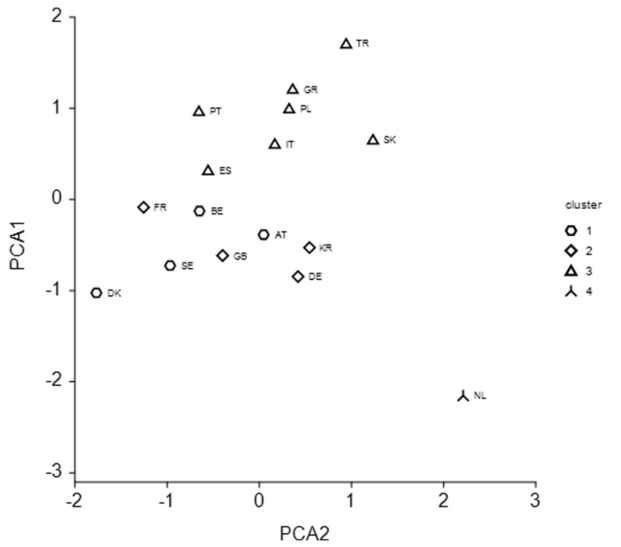

Fig. 13 *Visualisation of individual countries based on the two main components of the PCA method using k-means using the scores of the OECD model without the USA and Mexico*

Economic development has a major negative influence, likely caused by the simple correlation between economic development and general education of the population such as how to safely use a computer.

Compared to the previous model, the ICT variable has a slight negative influence. Therefore, the main influence on attack in this model comes from the ICTUse variable. Loading and weight values are available Table 7. Another difference from the previous models is the PopInfl variable, which is derived only from PopUpp and PopTer, as the other variables had a negligible effect.

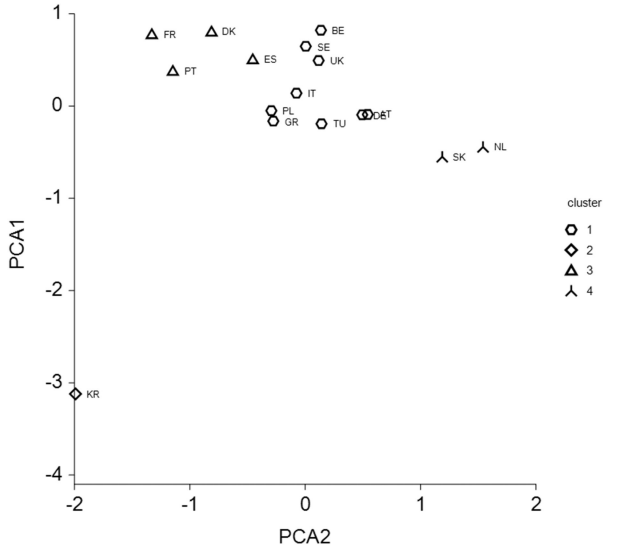

Fig. 14 *Visualisation of individual countries by manifest variables of SecInfl latent variable based on the two main components of the PCA method using k-means*

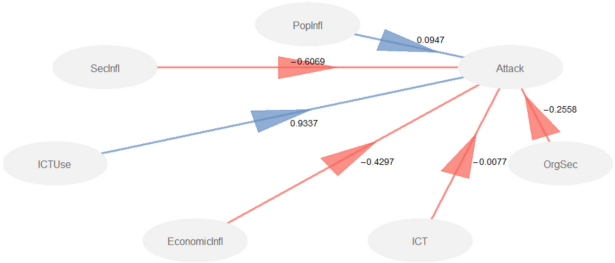

Fig. 15 *PLS-PM model based on OECD and Eurostat data only for EU countries*

Looking at the application of *k*-means and its PCA visualisation, available in Fig. 16, the clustering of countries including the anomalous Netherlands is nearly identical to the previous model. However, taking a closer look at the *k*-means and PCA applied to the manifest variables of attack, as seen in Fig. 6, the division of West, East, and South of Europe did change. It can be concluded that the number of attacks on honeynet from given countries is a combination of various aspects, though the ICT use and the population size are the deciding factors.

6.5 Comparison and conclusions of the models

The WB model provides an overview of countries such as China. However, it does not provide enough data for a detailed analysis, especially since the set of usable data is further limited by missing values. When the countries with high population such as China or the USA, are included, the population influence dominates. When local models grouping populational similar countries are used, the influence of technical aspects such as the number of Internet users, are more prominent. Economic aspects such as imports and exports, also have some small influences in the case of countries without a huge population. This was also the first model to, surprisingly, clustered countries such as the Netherlands and Lithuania with China, the USA, and India.

The OECD model allows us to add additional variables. When the USA is left in the model, the population influence dominates, followed by the ICT availability. As for the clustering, the USA is its own cluster, though it would likely be clustered with China that had to be removed from analysis due to missing data. The Netherlands is clustered with the rest of western Europe despite its anomalous attack count. As for the rest of the countries, the first cluster consists of the western European countries and the second one of eastern and southern European countries along with Turkey and Mexico.

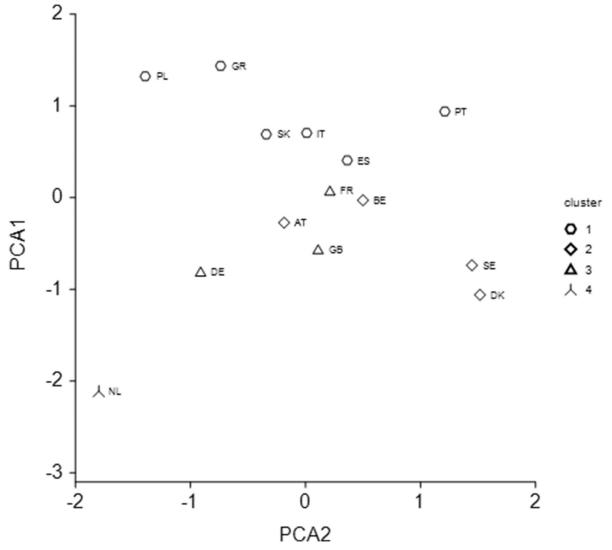

Fig. 16 *Visualisation of individual countries based on the two main components of the PCA method using k-means using the scores of OECD model extended with Eurostat data*

To improve the results, the USA and Mexico were removed in another round of analysis due to a significant amount of missing data. This allowed addition of a new latent variable focused on specific uses of ICT and the experience of the population with cybersecurity. With this new addition, the Netherlands stood out, suggesting the high amount of attacks is at least partially explained by the variable. In general, it can be concluded that among European countries, ICT use has a major influence on attacks, while the influence of population dramatically reduced. Experience with cybersecurity and economic power of a country have a significant negative influence on the attacks, as expected.

Exclusively for the EU member countries, it was possible to expand the model by adding data about how the countries approach cybersecurity of private organisations, companies etc. The overall conclusion was not changed by the new variable very much, but it did confirm that the security measures taken by organisations do indeed have a negative influence on the number of attacks.

The Netherlands is a rather peculiar anomaly in virtually all models. It is clustered along with countries with magnitudes higher population such as China and the USA, due to its high number of attacks, even though its variables from databases are very much like those of its European neighbours, with the notable exception of the SecInfl variable explained in Section 6.3. The Netherlands is shown to also have similarly high results in a report published on the Honeynet Project [38] webpage. The authors of this paper have also contacted the national CSIRT of the Netherlands [39], which suggested that the high number of attacks may be a result of a relatively large number of server farms and virtual private servers which are often infected and become a part botnets. This makes sense since the number of server farms and virtual private servers do not necessarily correlate with the population size.

7 Conclusion

Medium-interaction honeypot has proven to be an appropriate tool for mapping the currently spreading cyber threats. Use of available databases collecting statistical data about individual countries has also proven to be very useful in deducing the influence of the statistical data on the attacks originating from the given countries. The PLS-PM method has shown to be appropriate to process the available data. In general, it can be concluded that the population size has by far the highest influence. Therefore, it makes sense to cluster high population countries together so the influence of other variables between lower population countries can be compared, which is what the models did. Besides population, the ICT use had the highest influence, even higher than ICT availability, which makes sense, since it has to be used to be a potential source of attacks. Another important finding is that the experience of users, organisations, and companies with cybersecurity has a real effect on reducing attacks. The one exception was, again, the Netherlands, where it seems that slightly different habits of the population regarding cybersecurity and the server farms infrastructure providing VPS cause a radically higher number of attacks.

The research in this area is planned to be expanded from only the SSH protocols to other services of operating systems, possibly even server services. This data may be able to further specify the influence of the statistical indicators on the number of attacks from countries. Compared to the results of these papers described in section 2, this paper analyses more available databases of country indicators, and it uses societal indicators and ICT use and ICT problem solving indicators. It presents multiple statistical models of influence on the number of attacks from specific countries that suggest a course further research may take. The results are also useful for communication with national CSIRT teams of the affected countries, as they may use in deploying measures to limit the spread of infections and in developing proactive measures.

Further research will focus on attacks on other protocols besides SSH such as attacks aimed at Windows services, especially SMB, the most commonly targeted by malware. It will be interesting to see if the new data corroborates the results of the SSH attacks' analyses.

8 Acknowledgment

This paper was supported by the University of Ostrava from the project SGS03/PřF/2019.

9 References

[1] Safa, N.S., Maple, C., Watson, T., *et al.*: 'Information security collaboration formation in organisations', *IET Inf. Sec.*, 2018, **12**, (3), pp. 238–245(7)

[2] Spotzner, L.: '*Honeypots: tracking hackers*' (Addison Wesley Longman Publishing Co., Inc., USA, 2002)

[3] Joshi, C.R., Sardana, A.: '*Honeypots a new paradigm to information security*' (Science Publishers, USA, 2011)

[4] Kim, I.S., Kim, M.H.: 'Agent-based honeynet framework for protecting servers in campus networks', *IET Inf. Sec.*, 2012, **6**, (3), pp. 202–211(9)

[5] Grudziecki, T., Jacewicz, P., Juszczyk, Ł., *et al.*: '*Proactive detection of security incidents honeypots*' (ENISA Publication, Greece, 2012)

[6] Balas, E., Viecco, C.: 'Towards a third generation data capture architecture for honeynets'. Proc. from the Sixth Annual IEEE Systems, Man and Cybernetics (SMC) Information Assurance Workshop, West Point, NY, USA, 2005, pp. 21–28

[7] Sokol, P., Kopcova, V.: 'Lessons learned from correlation of honeypots' data and spatial data'. Eighth Int. Conf. Electronics, Computers and Artificial Intelligence (ECAI), Ploiesti, Romania, 2016, pp. 1–8

[8] Canto, J., Dacier, M., Kirda, E., *et al.*: 'Large scale malware collection: lessons learned'. IEEE SRDS Workshop on Sharing Field Data and Experiment Measurements on Resilience of Distributed Computing Systems, Napoli, Italy, 2008

[9] Thonnard, O., Dacier, M.: 'A framework for attack patterns' discovery in honeynet data'. Digital Investigation, Baltimore, USA, 2008, pp. 128–139

[10] Tang, M.J., Alazab, M., Luo, Y.: 'Exploiting vulnerability disclosures: statistical framework and case study'. Cybersecurity and Cyberforensics Conf. (CCC), Amman, Jordan, 2016, pp. 117–122

[11] Skrzewski, M.: 'Network malware activity – a view from honeypot systems'. Computer Networks, Communications in Computer and Information Science, Szczyrk, Poland, 2012, pp. 198–206

[12] Sochor, T., Zuzčák, M., Bujok, P.: 'Analysis of attackers against windows emulating honeypots in various types of networks and regions'. Eighth Int. Conf. Ubiquitous and Future Networks (ICUFN), Vienna, Austria, 2016, pp. 863–868

[13] Soldo, F., Le, A., Markopoulou, A.: 'Blacklisting recommendation system: using spatio-temporal patterns to predict future attacks', *IEEE J. Sel. Areas Commun.*, 2011, **29**, (7), pp. 1423–1437

[14] Sokol, P., Kleinova, L., Husak, M.: 'Study of attack using honeypots and honeynets lessons learned from time-oriented visualization'. IEEE EUROCON 2015 – Int. Conf. Computer as a Tool (EUROCON), Salamanca, Spain, 2015, pp. 1–6

[15] 'CZ-NIC LABS CSIRT.CZ – Kippo fork'. Available at https://gitlab.labs.nic.cz/honeynet/kippo, accessed April 2018

[16] 'Service VirusTotal.com'. Available at https://virustotal.com, accessed April 2018

[17] Sochor, T., Zuzčák, M., Bujok, P.: 'Statistical analysis of attacking autonomous systems'. Int. Conf. Cyber Security and Protection of Digital Services (Cyber Security), 2016, pp. 1–6

[18] 'The World Bank'. Available at http://www.worldbank.org/, accessed April 2018

[19] 'Organisation for Economic Co-operation and Development (OECD)'. Available at http://www.oecd.org/, accessed April 2018

[20] 'Eurostat'. Available at http://ec.europa.eu/eurostat/, accessed April 2018

[21] 'Eurostat: ICT security in enterprises'. Available at http://ec.europa.eu/eurostat/statistics-explained/index.php/ICT_security_in_enterprises, accessed April 2018
[22] MacQueen, J.: 'Some methods for classification and analysis of multivariate observations'. Proc. Fifth Berkeley Symp. Mathematical Statistics and Probability, Berkeley, 1967, 1: Statistics, pp. 281–297
[23] Hartigan, J.A., Wong, M.A.: 'A *K*-means clustering algorithm', *J. R. Stat. Soc. Ser. C (Appl. Stat.)*, 1979, **28**, (1), pp. 100–108
[24] Hotelling, H.: 'Analysis of a complex of statistical variables into principal components', *J. Educ. Psychol.*, 1933, **24**, pp. 417–441
[25] Zuzčák, M., Sochor, T.: 'Behavioral analysis of bot activity in infected systems using honeypots'. Computer Networks, Communications in Computer and Information Science, Lądek Zdrój, Poland, 2017, pp. 118–133
[26] Fichet, B.: 'Distances and Euclidean distances for presence–absence characters and their application to factor analysis'. Proc. Workshop Multidimensional Data Analysis, Cambridge, 1986, pp. 23–46
[27] Guha, S., Rastogi, R., Shim, K.: 'ROCK: a robust clustering algorithm for categorical attributes'. Proc. 15th Int. Conf. Data Engineering, Sydney, NSW, Australia, 1999, pp. 512–521
[28] Koyuturk, M., Grama, A., Ramakrishnan, N.: 'Compression, clustering, and pattern discovery in very high-dimensional discrete-attribute data sets', *IEEE Trans. Knowl. Data Eng.*, 2005, **17**, (4), pp. 447–461
[29] Hardoon, D.R., Szedmak, S., Shawe-Taylor, J.: 'Canonical correlation analysis: an overview with application to learning methods', *Neural Comput.*, 2004, **16**, (12), pp. 2639–2664
[30] González, I., Déjean, S., Martin, P.*, et al.*: 'CCA: an R package to extend canonical correlation analysis', *J. Stat. Softw.*, 2008, **23**, (12), pp. 1–14
[31] Wold, H.: 'Models for knowledge', in Gani, J. (Ed.): '*The making of statisticians*' (Springer-Verlag, New York, USA, 1982), pp. 189–212
[32] Geladi, P.: 'Notes on the history and nature of partial least squares (PLS) modelling', *J. Chemometr.*, 1988, **2**, (4), pp. 231–246
[33] Tenenhaus, M., Vinzi, V.E.: 'PLS regression, PLS path modeling and generalized Procrustean analysis: a combined approach for multiblock analysis', *J. Chemometr.*, 2005, **19**, pp. 145–153
[34] Sanchez, G.: 'PLS path modeling with R', Trowchez Editions, Berkeley, 2013
[35] Ihaka, R., Gentleman, R.: 'A language for data analysis and graphics', *J. Comput. Graph. Stat.*, 1996, **5**, (3), pp. 299–314
[36] 'Introduction to the R package plspms'. Available at https://cran.r-project.org/web/packages/plspm/vignettes/plspm_introduction.pdf, accessed April 2018
[37] Zanin, L.: 'Detecting unobserved heterogeneity in the relationship between subjective well-being and satisfaction in various domains of life using the REBUS-PLS path modelling approach: a case study', *Soc. Indicators Res.*, 2011, **110**, (1), pp. 281–304
[38] 'HoneyNED chapter had a busy 2017'. Available at http://www.honeynet.org/node/1365, accessed April 2018
[39] 'National Cyber Security Centre'. Available at https://www.ncsc.nl/, accessed April 2018