Lessons Learned from Honeypots - Statistical Analysis of Logins and Passwords

Pavol Sokol, Veronika Kopčová

Lessons learned from honeypots - statistical analysis of logins and passwords

Pavol Sokol[1] and Veronika Kopčová[2]

[1] Institute of Computer Science,
Faculty of Science, Pavol Jozef Safarik University in Kosice,
Jesenna 5, 040 01 Kosice, Slovakia
pavol.sokol@upjs.sk
[2] Institute of Mathematics,
Faculty of Science, Pavol Jozef Safarik University in Kosice,
Jesenna 5, 040 01 Kosice, Slovakia
veronika.kopcova@student.upjs.sk

Abstract. Honeypots are unconventional tools to study methods, tools and goals of attackers. In addition to IP addresses, timestamps and count of attacks, these tools collect combinations of login and password. Therefore, analysis of data collected by honeypots can bring different view of logins and passwords. In paper, advanced statistical methods and correlations with spatial-oriented data were applied to find out more detailed information about the logins and passwords. Also we used the Chi-square test of independence to study difference between login and password. In addition, we study agreement of structure of password and login using kappa statistics.

Keywords: honeypot,login,password,spatial data, Chi-square test,kappa statistic

1 Introduction

In current information society we deal with an increasing security threat. Therefore, an important part of information security is protection of information. Common security tools, methods and techniques used before are ineffective against new security threats. Therefore, it is necessary to choose other tools and techniques. It seems that the network forensics, especially honeypots and honeynets, are very useful tools. The use of the word "honeypot" is quite recent [1], however honeypots have been used for more than twenty years in computer systems. It can be defined as a computing resource, whose value is in being attacked [2]. Lance Spitzner defines honeypot as an information system resource whose value lies in unauthorized or illicit use of that resource [3].

The most common classification of honeypot is classification based on the **level of interaction**. The definition of level of interaction is the range of possibilities the attacker is given after attacking the system. Honeypots can be divided into low-interaction and high-interaction. Example of this type of honeypots is

Dionaea [4]. On one hand, low-interaction honeypots emulate the characteristics of network services or a particular operating system. On the other hand, a complete operating system with all services is used to get more accurate information about attacks and attackers [5]. This type of honeypot is called high-interaction honeypot. Example of this type of honeypots is HonSSH [6].

Concept of honeypot is extended by **honeynet** - a special kind of high-level interaction honeypot. The honeynet can be also referred to as "a virtual environment, consisting of multiple honeypots, designed to deceive an intruder into thinking that he or she has located a network of computing devices of targeting value" [7]. Four main parts of the honeynet architecture are known, namely data control, data capture, data collection and data analysis [2, 7].

The main reason to use these tools is collection and analysis of data captured using honeypots and honeynets. Learning new unconventional information about the attacks, attackers and tools is involved in the protection of the network services and computer networks of organizations. Each honeypot collects the IP addresses of attackers and special data according to type of honeypot. In paper we use the low-interaction honeypots Kippo [8], which collect timestamps, IP address of attacker, type of SSH clients and combination of logins and passwords. For purpose of this paper we focus on logins, passwords and their combinations.

This paper is a sequel to the analysis of data collected from honeypots and honeynets. In paper [9] authors focus on **automated secure shell (SSH)** bruteforce attacks and discuss the length of passwords, password composition compared to known dictionaries, dictionary sharing, username-password combination, username analysis and timing analysis.On the other hand, the main aim of this paper is to provide light on attackers' behaviour, and provide recommendations for SSH users and administrators. In this paper we focus on two main statistical analyses. Firstly, chi-square test of independence that analyzes group of differences. Secondly, Kappa statistics that measures agreement between observes.

To formalize the scope of our work, authors state two research questions:

- What attribution of logins, passwords and their attribution are significant for security of systems?
- What is the relationship between the logins and passwords and origin of attacks?

This paper is organized into seven sections. Section II focuses on the review of published research related to lessons learned from analysis in the honeypots and honeynets. Section III outlines the dataset and methods used for experiment. Sections IV-VI focus on statistical and spatial analysis of logins, passwords and combination of them. The last section contains conclusions, discussion and our suggestions for the future research.

2 Related works

As it was mentioned before, the main task of honeypots and honeynet is in analysing the captured data and searching for new knowledge about the attacks

and attackers. This section provides overview of papers that focus on lessons learned from honeypots and honeynets data.

Analysis of data collected by **high-interaction honeypots** are discussed in Nicomette et. al. [10] and Alata et. al. [11]. [10] concentrate on the attacks executed by the SSH service and the activities executed after attackers gain access to the honeypot. Attackers and their activities after logging in are discussed in [11]. Authors correlated their findings with the results from distributed low-interaction honeypots.

But then, **low-interaction honeypots** are discussed in Sochor and Zuzcak in papers [12, 13]. In [12] data show currently spreading threats caught by honeypots. But then, the thorough interpretation of lessons learned from using the honeypots was outlined. Principal results are shown in [13], in addition they underline the fact that the differentiation between honeypots according to their IP address is quite rough (e.g. differentiation for academic and commercial network).

SGNET was used by [14] as a **distributed system of honeypots**. They doubt the floatation of representative malware samples datasets. They claim that the false negative alerts differ from what they are allowed to be. Additionally, there is occurrence of false positive alerts on abrupt places. Clustering attack patterns with a suitable similarity measure are discussed in [15]. The results of this study allow identification of the activities of several worms and botnets in the collected traffic.

Time-oriented data were of interest in [16]. Visualization of this data in honeypots and honeynets was outlined. In addition, the authors provide results based on heatmaps that is special visualisation. It was proved that the time is an important aspect of attacks. Attackers are mainly active at night (according to the honeynets time zone analysis).

Next example of using low-interaction honeypots (Dionaea) in order to studying is in [17]. It presents the results of nearly two years operation of honeypot systems, installed on unprotected research network. The paper focuses on the information about the life time of malware programs and the long-time malware activity.

3 Data collection and analysis methodology

The data were collected from the honeynet located in the campus network. The honeynet that runs on port 22 consists of SSH honeypots Kippo [8] in low-interaction mode. The honeypots do not allow attackers to log into shell in this mode, they only capture data about network flows entering the honeynet. The honeypots have collected authentication attempts from 3rd August 2014 to 24th December 2015. During this period **1 391 746 records** were collected. Each record contains username and password used in an attempt, as well as IP address and version of client of attacker, beginning and end of sessions. Dataset contain **unique 5 488 logins**, **unique 205 477 passwords** and **unique 212 687 combinations** of login and password.

For spatial analysis, each record was competed with spatial data using the **IP-API.com** service [18]. This service provides free use of its Geo IP API through multiple response formats. Each record was supplemented with time zone, country, region, city, Internet service provider (ISP), and global positioning systems (GPS) coordinates.

Data cleaning and analysing was performed using, **the HoneyLog framework** [19]. This framework for analysing honeypots and honeynets data is based on a PHP framework of FuelPHP and JavaScript libraries. It has two main segments: a client part and a server part.

For purpose of paper, important part of dataset consists of combination of logins and passwords. Since the logins and passwords are the qualitative data it needed to be converted into quantitative data. For each login and password, we assigned following attributes:

- **contains only lowercases** login or password contains only lowercase characters (ASCII codes between 97 and 122);
- **contains only uppercases** - login or password contains only capital characters (ASCII codes between 65 and 90);
- **contains only numbers** - login or password contains only numbers (ASCII codes between 65 and 90);
- **contains number** - login or password contains at least one number;
- **contains year** - login or password contains year (2014 or 2015) and
- **contains special character** - login or password contains at least one special character (ASCII codes 32-47,58-64,91-96 and 123-127);

In paper we use two statistical methods: chi-square test of independence and kappa statistics. **The Chi-square test of independence**, also known as **the Pearson Chi-square test** [20], is one of the most useful tools for testing hypotheses when the variables are nominal. It is a non-parametric tool designed to analyse group differences. Each non-parametric test has its own specific assumptions as well. The assumptions of the Chi-square include:

1. The data in the cells should be frequencies, or counts of cases.
2. The categories of the variables are mutually exclusive.
3. Each subject may contribute data to one and only one cell in the Chi-square.
4. The study groups must be independent.
5. While Chi-square has no rule about limiting the number of cells (by limiting the number of categories for each variable), a very large number of cells (over 20) can make it difficult to meet assumption #6 below, and to interpret the meaning of the results.
6. The value of the cell expected should be 5 or more in at least 80% of the cells, and no cell should have an expected of less than one (3). This assumption is most likely to be met if the sample size equals at least the number of cells multiplied by 5.

On the other hand, **Kappa** [21] is intended to give the reader a quantitative measure of the magnitude of agreement between observers. Interobserver variation can be measured in any situation in which two or more independent observers are evaluating the same thing.

4 Logins

The first observed aspect of analysis is **login**. Top 10 logins are shown in Fig. 1(left). This diagram shows that the most tested login is **root**. According to other logins, attackers test default logins for different systems (admin, user, PI, Oracle, etc.). Also attacker is often trying the same login and password combination. In this paper we focus on analysis of login with the largest number of unique passwords. Top 10 logins with unique passwords are shown in Fig. 1 (right). From this perspective, the most tested login is root. Attacker also tests following logins with large number of unique passwords: user, test, nagios, mysql.

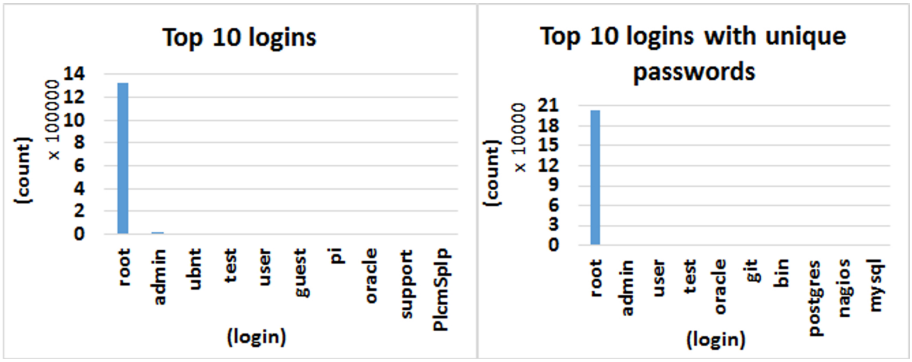

Fig. 1. Top 10 logins and top 10 logins with unique passwords

4.1 Attributes of logins

According to Linux documentation for tool useradd [22], Unix/Linux's username (login) equals regular expression **[a-z_][a-z0-9_-]*[$]?$.** This expression means that the first character of login is lowercase and other characters are lowercases or numbers. Also capital letters are not allowed. Moreover, logins must neither start with a dash nor contain a colon or a whitespace, end of line and tabulation etc. Documentation notes that using a slash may break the default algorithm for the definition of the user's home directory.

As we can see in Fig. 2, the largest group of logins is logins containing **only lowercases (88,47 %).** A slight amount of logins contains a **number (7,89 %)** or **special character (4,46 %)**. According to our opinion, logins, which contain capital letters or special character are tested by special group of attackers - script kidies or attacks were directed to other systems like UNIX/LINUX.

Another studied aspect is the length of logins (Fig. 3). According to above mentioned Linux documentation [22], logins may only be up to 32 characters long. The length of tested logins is in range from 1 to 50 characters. The logins with length between 33 and 50 are a sign of incorrect use of automated

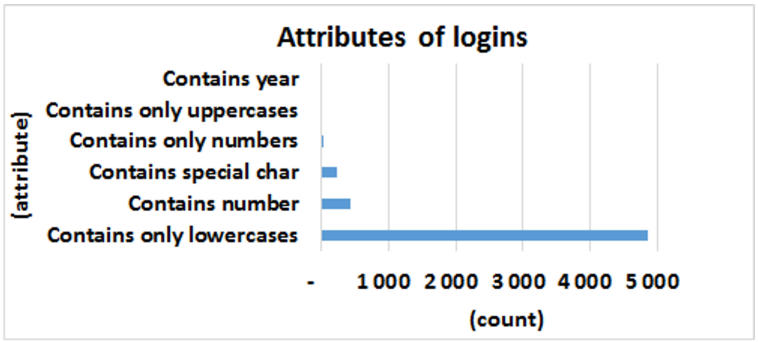

Fig. 2. Attributes of logins

programs. For example root1a1O0GlNs$KPwONdPK6G5KqjsVNNOyb. The largest group of logins contains six characters. The largest amount of logins has number of characters in range from 3 to 14.

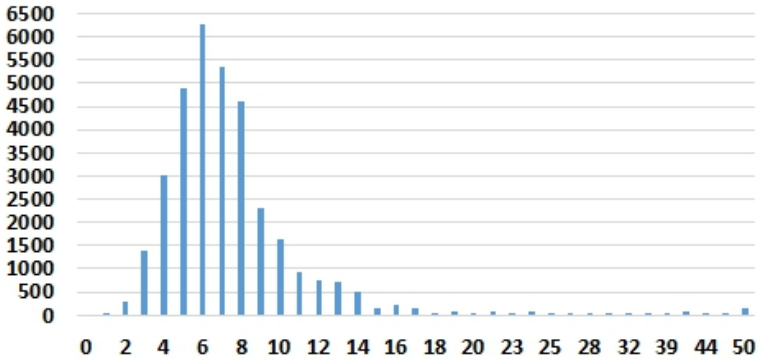

Fig. 3. Length of logins

4.2 Frequency of ASCII characters in logins

For purpose of the frequency of ASCII characters in logins we created frequency table (Fig. 4). This table takes into account the frequency of at least one occurrence of a given character within a login. ASCII character with the highest occurrence is **lowercase a**. **Lowercase e**, which is the most frequent character in many alphabets (e.g. English, French and German alphabet), is in the 2nd place. On the other hand, **lowercase q and x** have the lowest occurrence. The most used number is **1** and **2**. On the other hand, **6** and **8** are used at least. In the most cases the login contain special character /. In contrast to this, pass-

words do not contain this character. According to our opinion, it is again sign of **incorrect use of automated programs**.

NULL	0	DLE	0	(space)	0,24	0	2,19	@	0,09	P	0,42	`	0,04	p	12,28
SOH	0	DC1	0	!	0,22	1	3,86	A	0,38	Q	0,09	a	50,98	q	1,75
STX	0	DC2	0	"	0,26	2	3,12	B	0,33	R	0,26	b	12,79	r	36,79
ETX	0	DC3	0	#	0,09	3	2,39	C	0,29	S	0,35	c	19,42	s	29,12
EOT	0	DC4	0	$	0,2	4	1,53	D	0,2	T	0,29	d	19,08	t	28,43
ENQ	0	NAK	0	%	0,15	5	1,22	E	0,27	U	0,2	e	43,31	u	16,93
ACK	0	SYN	0	&	0,02	6	0,8	F	0,15	V	0,13	f	7,38	v	7,45
BEL	0	ETB	0	'	0,13	7	0,87	G	0,09	W	0,15	g	13,23	w	6,27
BS	0	CAN	0	(	0,07	8	0,8	H	0,13	X	0,16	h	14,83	x	4,06
HT	0	EM	0	)	0,07	9	1,04	I	0,38	Y	0,15	i	37,26	y	10,79
LF	0	SUB	0	*	0,16	:	0,15	J	0,15	Z	0,15	j	6,49	z	4,28
VT	0	ESC	0	+	0,04	;	0,35	K	0,09	[	0,07	k	9,91	{	0,05
FF	0	FS	0	,	0,93	<	0,22	L	0,29	\	0,04	l	25,07	|	0
CR	0	GS	0	-	0,42	=	0,11	M	0,27	]	0,09	m	17,62	}	0
SO	0	RS	0	.	2,02	>	0,15	N	0,27	^	0,09	n	34,17	~	0,02
SI	0	US	0	/	3,39	?	0,07	O	0,22	_	0,18	o	31,01	DEL	0

Fig. 4. Frequency table of ASCII characters in logins

4.3 Logins and origin of attacks

Tab. 1 shows top 20 countries, which are origin of attacks. For each country, table shows the count of attacks, top login and its count and percentage and the top three logins, which are tested by attackers from country. The login **root** is the most tested login from **each top 20 country**. The interesting fact is that percentage of tested login root to all tested passwords from country is different. On one hand, there is high percentage in countries such as China, Hong Kong, France, Hungary etc. On the other hand, there is low percentage in countries such as Argentina or Singapore. The most tested group of logins are **root/admin/ubnt**, **root/admin/test** and **root/admin/user**. Based on this it can be concluded that groups of tested logins, considering origin of attacks, can be interesting **indicator for finding group of attackers**.

5 Passwords

The second observed aspect is **password**. Compared to logins the types of passwords are pronounced. The most commonly used password is **admin**. Top 10 the most used passwords (123456, password, root, 1234, etc.) is shown in Fig. 5 (left). Like in login, we focus on the passwords that are used with the most unique logins. In this regard, the most used login is password **(none)**. Other most used passwords with the most unique logins are shown in Fig. 5 (right).

5.1 Attributes of passwords

In this section we focus on **attributes of passwords**. These attributes are shown in Fig. 6. Compared to the login, Linux documentation does not restrict

Table 1. Logins and top 20 countries

Country	Count of attack	Top login	Count (percent) of top login	The 2nd and 3rd login
China	895 945	root	873 321 (97,47 %)	admin/ubnt
Hong Kong	219 621	root	219 025 (99,73 %)	admin/ubnt
France	123 430	root	122 889 (99,56 %)	admin/developer
United States	92 721	root	81 381 (87,77 %)	admin/ubnt
Hungary	6 952	root	6 820 (98,10 %)	deployer/ubuntu
Rep. of Korea	5 459	root	4 074 (74,63 %)	admin/test
Germany	2 872	root	804 (27,99 %)	admin/test
Russia	2 851	root	1 848 (64,82%)	admin/user
Brazil	2 609	root	868 (33,27%)	admin/ubnt
Argentina	2 131	root	87 (4,08%)	mysql/jboss
Singapore	2 113	root	188 (8,90%)	admin/test
Vietnam	2 021	root	1 095 (54,18%)	admin/test
UK	1 536	root	472 (30,73%)	admin/ubnt
Poland	1 358	root	641 (47,20%)	admin/user
Netherlands	1 343	root	437 (32,54%)	admin/user
Canada	1 276	root	597 (46,79%)	admin/test
Spain	1 142	root	467 (40,89%)	admin/ubnt
Japan	1 127	root	697 (61,85%)	admin/ubnt
Ukraine	1 124	root	642 (57,12%)	admin/user
Turkey	939	root	712 (75,83%)	admin/ubnt

password from the perspective of characters (no security). It is due to the fact that system stores hash of password (no clear password). According to Fig. 6 the most frequently used passwords **contain numbers** (50,36 %). A slightly smaller number of the passwords containing only lowercase (45,24 %). In contrast, entries containing only a number occur almost three times less often. An interesting fact is that among the top 10 passwords were four passwords containing only numbers (123,1234,12345,123456) (9,9 %) and the only one password containing only lowercase characters (test) (0,83 %).

Another attribute of password is its **length**. The length of the password is in the range between **0** and **98**. The most passwords contain **8 characters**. The largest number of length of passwords is in the range between **3 and 20 characters**. It is worth mentioning that passwords with 32 characters are hashes (e.g. 706e642a056c7e894ed5a01e55700004). Number of characters of passwords is shown in Fig. 7 (left). Passwords with 33 characters and more are a sign of **incorrect using of tool** (e.g. #files th a:hover { background:transparent; border...) or **manual attack by script-kidies** (e.g. roooooooooooooooooooooooooooooo-ooooooooooooooooooooooooooot)

We also focus on the largest group of passwords that contain only numbers. In this group the largest subgroup of passwords contains 8 respectively 6 digits. Number of length of passwords, which contain only numbers, are shown in Fig. 7 (right).

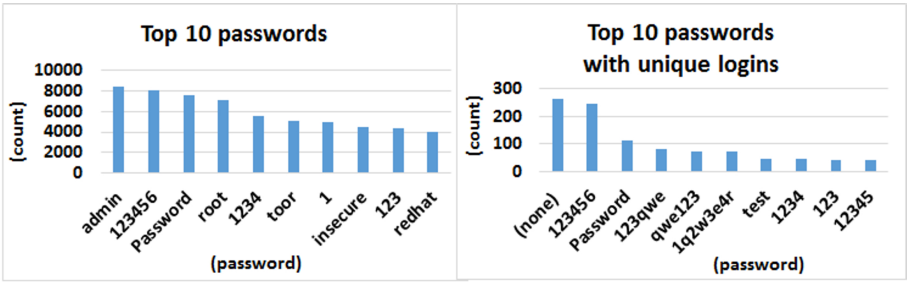

Fig. 5. Top 10 passwords and top 10 passwords with unique logins

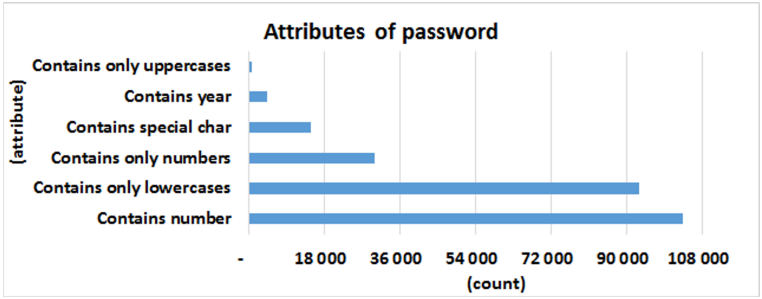

Fig. 6. Attributes of passwords

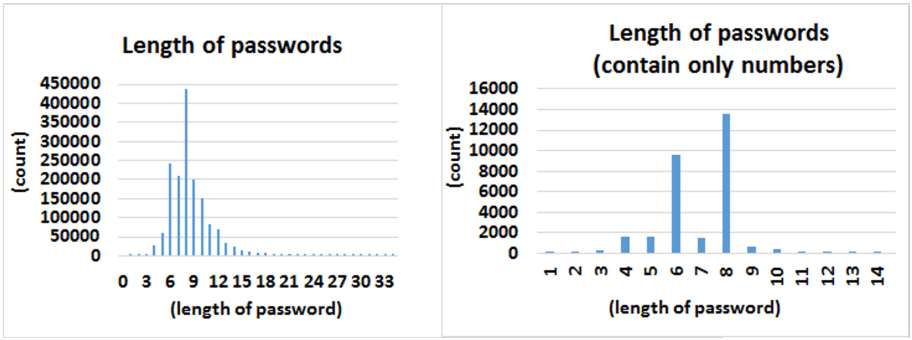

Fig. 7. Length of passwords

5.2 Frequency of ASCII characters in passwords

Like for a login, the frequency tables of ASCII characters in passwords were created (Fig. 8). This table takes into account the frequency of at least one occurrence of a given character within a password. ASCII character with the highest occurrence is lowercase **a**. Lowercase **e**, which is the most frequent character in many alphabets (e.g. English, French and German alphabet), is in the 2nd place. On the other hand, capital V and **capital K** have the lowest occurrence. Similar to login, the most used number is **1** and **2**. On the other hand, **6** and **7** are used the least. In the most cases the passwords contain special characters **@** and **!**. Interesting fact is occurrence of characters **Horizontal Tab** (ASCII code 9) and **Device control 1-4** (ASCII codes 17-20) in passwords (e.g. %username DC1 3!@, %username DC2 34567890-=). These codes are used for software flow control (e.g. DC 1 for quit application). These codes are not visible in logs. Passwords with these codes begin with special characters **!, %** or **@** and they are linked to login root. According to our opinion, passwords with these codes are used in incorrect using of a tool by script-kidies.

NULL	0	DLE	0	(space)	0,15	1	33,69	A	0,9	Q	0,47	a	43,33	q	3,03
SOH	0	DC1	0	!	2,83	0	16,71	@	3,59	P	0,59	`	0,04	p	10,59
STX	0	DC2	0,01	"	0,03	2	24,9	B	0,4	R	0,51	b	11,5	r	30,18
ETX	0	DC3	0	#	2,15	3	18,54	C	0,6	S	0,78	c	18,59	s	27,1
EOT	0	DC4	0	$	1,32	4	11	D	0,63	T	0,5	d	19,29	t	23,73
ENQ	0	NAK	0	%	1,04	5	10,11	E	0,65	U	0,4	e	38,49	u	16,24
ACK	0	SYN	0	&	0,33	6	9,31	F	0,38	V	0,24	f	8,44	v	6,9
BEL	0	ETB	0	'	0,04	7	9,43	G	0,35	W	0,56	g	12,01	w	8,1
BS	0	CAN	0	(	0,24	8	10,9	H	0,41	X	0,41	h	15,71	x	3,74
HT	0,03	EM	0	)	0,21	9	13,38	I	0,56	Y	0,28	i	31,88	y	11,31
LF	0	SUB	0	*	0,44	:	0,06	J	0,27	Z	0,35	j	4,79	z	4,53
VT	0	ESC	0	+	0,11	;	0,23	K	0,26	[	0,06	k	10,22	{	0,05
FF	0	FS	0	,	0	<	0,03	L	0,49	\	0,08	l	22,2	|	0,02
CR	0	GS	0	-	0,39	=	0,07	M	0,46	]	0,04	m	17,94	}	0,05
SO	0	RS	0	.	1,34	>	0,04	N	0,53	^	0,84	n	30,25	~	0,07
SI	0	US	0	/	0	?	0,07	O	0,39	_	0,25	o	28,98	DEL	0

Fig. 8. Frequency table of ASCII characters in passwords

5.3 Passwords and origin of attacks

Tab. 2 shows top 20 countries, where attacks originated. For each country, table shows the count of attacks, the most used passwords with their count and percentage and the top three logins, which were tested by attackers from country. In table (none) means that password without chars was inputted. The password **123456** is the most tested from 7 top countries. An interesting finding is password **weubao** in Hong Kong. In case of logins, there is similar the most tested groups of logins considering the origin of attacks. In case of passwords, there are no similar groups with top 3 passwords. Based on this it can be concluded that **there is relationship** between passwords and origin of attacks.

Table 2. Passwords and top 20 countries

Country	Count of attack	Top password	Count (percent) of top password	The 2nd and 3rd password
China	895 945	admin	6 384 (0,71%)	password/123456
Hong Kong	219 621	wubao	188 (0,09%)	jiamima/(none)
France	123 430	(none)	112 (0,09%)	fff1fff/password
United States	92 721	default	895 (0,97%)	(none)/admin
Hungary	6 952	123456	14 (0,20%)	(none)/raspberry
Rep. of Korea	5 459	123456	90 (1,65%)	admin/default
Germany	2 872	ADMIN	164 (5,71%)	123456/password
Russia	2 851	password	58 (2,03%)	admin/123456
Brazil	2 609	123456	48 (1,84%)	default/admin
Argentina	2 131	123456	44 (2,06%)	password/server
Singapore	2 113	123456	23 (1,09%)	1234/test
Vietnam	2 021	(none)	181 (8,96%)	password/admin
UK	1 536	123456	114 (7,42%)	password/admin
Poland	1 358	123456	28 (2,06%)	password/12345
Netherlands	1 343	(none)	144 (10,72%)	admin/1234
Canada	1 276	password	43 (3,37%)	123456/ C@r*i%n$t#o!(s
Spain	1 142	admin	8 (0,70%)	ubnt/12345
Japan	1 127	default	36 (3,19%)	admin/ubnt
Ukraine	1 124	(none)	80 (7,12%)	admin/root
Turkey	939	admin	23 (2,45%)	123456/123456789

6 Combination of logins and passwords

In previous sections we focus on logins and passwords. Since attacker test combinations of login and password, we focus on this aspect. The most tested combination of login and password, which are used by attackers, are following: **root/admin**, **root/root**, root/Password, root/123456, root/toor, root/1234, root/1 etc. In the following sections we focus on relationship between logins and passwords.

6.1 Association between passwords and logins and their attributions

For purpose of association between passwords and logins and their attributions the **Chi-square test of independence** [20] is used. In our case study, there are two groups: passwords and logins. The independent variable is login/password and dependent variable is its attribution: special char, only number, number, only uppercase. Our goal is to find out, whether login and password differ. Tab. 3 shows our data where marginals were calculated.

The formula for calculating Chi-Square values is: $\chi^2 = (O - E)^2/E$, where O is observed and E is expected value. Chi-Square expecteds are calculated as follows: $E = Mr * Mc/n$. Table 4 provides the results of this calculation for each cell. Expected value (chi square value).

Table 3. Calculation of marginals

	special char	only number	number	only uppercase	Marginals Mr
Password	41 623	177 543	442 514	3 862	665 542
Login	989	226	1933	50	3 198
Marginals Mc	42 612	177 769	444 447	3 912	**668 740**

Table 4. Cell expected values and (cell Chi-square values)

	special char	only number	number	only uppercase
Password	42408,22 (14,54)	176918,89 (2,20)	442321,60 (0,08)	3893,29 (0,25)
Login	203,78 (3025,76)	850,11 (458,20)	2125,40 (17,42)	18,71 (52,34)

Table 5. Examples of logins and passwords in Chi-square test of independence

	special char	only number	number	only uppercase
Password	garland!@#	30011970	itac2014	GENGISHAN
Login	root!"?$%&	12345678	Aa12345root	NASA

Now we sum cell chi square values to obtain chi square statistic for the table. In this case it is 3571. The chi square table requires knowledge of degrees of freedom to determine the significance level of the statistics. It holds: $df = (numberofrows - 1) * (numberofcolumns - 1) = 1 * 3 = 3$. The critical value for chi square distribution with $df = 3$ is **7,815**. So our calculated value is bigger than critical value: $3571 > 7,815$ and we can conclude that null hypothesis is rejected, which means that there is a relationship between login and password. However, this result does not specify what impact on this relationship. It can be seen in Tab. 4. The largest values of cell chi square values can be seen in a special char for login. It means that number of logins that contain special char is significantly greater than expected value. On the other hand, cell chi square values less than 1 means that number of observed cases is equal to number of expected cases. So there is no effect on password for number and only uppercase.

Based on the above mentioned, it can be concluded that there is a relationship between the login and password. Especially if the password contains a special character or number. Logins typically contain only lowercases. Therefore, if it contains special characters, numbers, at least one number or all capital characters, there is a **relationship between the login and password**. In the greatest extent it occurs in case of login with **special character** (e.g. password garland!@# for login root). Another example is the login root!"?$%& with password (none). In these cases, it can be concluded that it is not a dictionary attack, respectively brute force attack, but a **manual attack** or **automated attack by script-kidies**.

Table 6. Kappa statistics

Login/Password	**special char**	**only number**	**number**	**only uppercase**	**Total**
special char	547	38	22	0	607
only number	0	218	0	0	218
number	11	98	1088	0	1197
only uppercase	1	4	27	13	45
total	559	358	1137	13	2067

Table 7. Examples of logins and passwords in Kappa statistics

Login/Password	**special char**	**only number**	**number**	**only uppercase**
special char	*root/*, . **kl;iop890**	*root/* − ∗ **123456**	root-***123456**	-
only number	-	123456 **123456**	-	-
number	*rootzo*9 ∗?*qp*	*tom*6*bj* **278497**	*r*00*t* **loler11q**	-
only uppercase	NASA **N.A.S.A**	SZIM **888888**	USERID **passw0rd**	CSICI **CCC**

6.2 Agreement of structure of password and login

For study agreement of structure of password and login, we use kappa statistics. The data were collected in Table 6.

We can simply calculate the percentage of agreement as a sum of diagonals divided by number of observations, we have **90,3%** agreement. But that measure does not take into account the random chance of agreement. We calculate expected agreement that is $Pe = 0,416$. Formula for kappa: $K = (Po - Pe)/(1 - Pe) = 0,834$. Using table in [21] we can conclude that **agreement of login and password is substantial.**

7 Conclusions, recommendations and future works

Attacks collected by honeypots are interesting source for further analysis. In paper we focus on logins, passwords and their combination. We outline statistical analysis of collected data. General rules for passwords creating state that password should contain lowercase, capital letter, number and special character. Length of password should be **8 or more**. According to above mentioned, we propose to use **capital V, capital K** and number **6 and 7** in passwords. We recommend avoiding the following lowercases: **a,e,i,n,r,o,s** and following numbers: **1,2,3 and 9**. To strengthen password it is recommended to use password with length **10 or more** and special characters: **[,],{ and }**.

Since the combination of login and password is used in attack, it is needed to deal with the strength of login. General safety rules state that default passwords

and root should not be used. We agree with these rules, but above mentioned we propose the following rules for login creating. The first character of password must be lowercase. Lowercase q or x look like the best choice. The login must have length between 1 and 32 characters. We recommend use the login with length **between 12 and 32 characters**. We recommend avoiding the following lowercases: **a,e,i,r,n,o,s,t,l,c** and following numbers: **1,2,3 and 0**. In general, using the numbers increase the security of the password, especially numbers: **6,7 and 8**.

As we showed before, Chi-square test of independence and Kappa statistics show that there is relationship between logins and passwords. On the basis of these tests, attacks can be divided into manual attacks and automated attacks.

In the future, the research in field of analysis of collected data will continue. We will primarily focus on types of clients and time-oriented analysis from the perspective of logins and passwords.

Acknowledgments

We would like to thank colleagues from the Czech chapter of The Honeynet Project for their comments and valuable input. This paper is funded by the Slovak Grant Agency for Science (VEGA) grants under contract No. 1/0142/15 and No. 1/0344/14, VVGS projects under contract No. VVGS-PF-2016-72610 and No. VVGS-PF-2016-72616 and Slovak APVV project under contract No. APVV-14-0598.

References

1. Pouget, F., Dacier, M., *et al.*: Honeypot-based forensics. In: AusCERT Asia Pacific Information Technology Security Conference (2004)
2. Spitzner, L.: The honeynet project: Trapping the hackers. IEEE Security and Privacy **1**(2), 15–23 (2003)
3. Spitzner, L.: Honeypots: Tracking Hackers. Addison-Wesley Reading, Boston (2003)
4. Dionaea project. Accessed: 20th August 2016. `https://github.com/rep/dionaea`
5. Joshi, R., Sardana, A.: Honeypots: a New Paradigm to Information Security. CRC Press, Boca Raton (2011)
6. HonSSH project. Accessed: 20th August 2016. `https://github.com/tnich/honssh/wiki`
7. Abbasi, F.H., Harris, R.: Experiences with a Generation III virtual Honeynet. In: Telecommunication Networks and Applications Conference (ATNAC), 2009 Australasian, pp. 1–6 (2009). IEEE
8. Kippo project. Accessed: 27th August 2016. `https://github.com/desaster/kippo`
9. Abdou, A., Barrera, D., Van Oorschot, P.C.: What lies beneath? analyzing automated ssh bruteforce attacks. In: International Conference on Passwords, pp. 72–91 (2015). Springer

10. Nicomette, V., Kaâniche, M., Alata, E., Herrb, M.: Set-up and deployment of a high-interaction honeypot: experiment and lessons learned. Journal in computer virology **7**(2), 143–157 (2011)
11. Alata, E., Nicomette, V., Dacier, M., Herrb, M., et al.: Lessons learned from the deployment of a high-interaction honeypot. arXiv preprint arXiv:0704.0858 (2007)
12. Sochor, T., Zuzcak, M.: Study of internet threats and attack methods using honeypots and honeynets. In: International Conference on Computer Networks, pp. 118–127 (2014). Springer
13. Sochor, T., Zuzcak, M.: Attractiveness study of honeypots and honeynets in internet threat detection. In: International Conference on Computer Networks, pp. 69–81 (2015). Springer
14. Canto, J., Dacier, M., Kirda, E., Leita, C.: Large scale malware collection: lessons learned. In: IEEE SRDS Workshop on Sharing Field Data and Experiment Measurements on Resilience of Distributed Computing Systems (2008). Citeseer
15. Thonnard, O., Dacier, M.: A framework for attack patterns' discovery in honeynet data. digital investigation **5**, 128–139 (2008)
16. Sokol, P., Kleinová, L., Husák, M.: Study of attack using honeypots and honeynets lessons learned from time-oriented visualization. In: EUROCON 2015-International Conference on Computer as a Tool (EUROCON), IEEE, pp. 1–6 (2015). IEEE
17. Skrzewski, M.: Network malware activity–a view from honeypot systems. In: International Conference on Computer Networks, pp. 198–206 (2012). Springer
18. IP-API.com service. Accessed: 20th August 2016. `http://ip-api.com`
19. Sokol, P., Pekarcik, P., Bajtos, T.: Data collection and data analysis in honeypots and honeynets. In: Proceedings of the Security and Protection of Information (2015). University of Defence
20. McHugh, M.L.: The chi-square test of independence. Biochemia Medica **23**(2), 143–149 (2013)
21. Viera, A.J., Garrett, J.M., *et al.*: Understanding interobserver agreement: the kappa statistic. Fam Med **37**(5), 360–363 (2005)
22. Linux documentation for useradd. Accessed: 22th August 2016. `http://www.unix.com/man-page/all/0/useradd`