Proceedings of the 2nd International Conference on Science, Technology and Social Sciences 2016 (ICSESS2016)

Real-Time Malware Uniform Resource Locator Detection: Identification of Novel Discriminative Features through Manual Examination and Empirical Analysis

Morufu Olalere[1], Mohd Taufik Abdullah[2*], Ramlan Mahmod[2], Azizol Abdullah[2]

[1]*Cyber Security Science Department Federal University of Technology Minna, 920102, Niger, Nigeria*
[1,2]*Information Security research Group Faculty of Computer Science and information Technology, Universiti Putra Malaysia, 43400, Selangor, Malaysia*

**Corresponding author email: lerejide@futminna.edu.ng*

ABSTRACT. Gone are the days when attackers used to introduce malware into enterprise network through storage devices. With the rapid proliferation of internet technologies and web applications, attackers now use web as a means of introducing malware into enterprise network. This development has forced many enterprises to subscribe to manually created blacklist of malware Uniform Resource Locator (URLs). Manually created blacklist is faced with challenges of wrong detection due to human error and inability to detect newly created malware URL that has not been added to the blacklist. This make blacklisting approach inadequate for detection of any malware URL encountered. Therefore, a real-time malware URL detection that is based on machine learning is required. To achieve this, there is a need to identify discriminative features of malware URL. This need motivated this study. Consequently, the authors of this study identified novel discriminative lexical features of malware URL and study the prevalence of these features. To identify discriminative lexical features, two methods including manual examination of malware URL and empirical analysis were employed. Manual examination of malware URLs was carried out using existing blacklist of malware URLs. This allowed the authors to identify discriminative lexical features. To determine whether there is consistency in the way the attackers craft malware URLs, empirical analysis was carried on both the existing blacklisted malware URLs and newly collected malware URLs. Empirical analysis revealed that there is consistency in the way malware URLs is crafted by the attackers. Therefore, these features can be used to build real-time malware URLs detection.

Keywords. Attackers, Lexical Features, Malware URL, Blacklist, Rea-time Malware URL Detection.

1. Introduction

Gone are the days when a malware infection on an enterprise network occurred only through external storage devices such as external hard disks and flash drives. With the rapid proliferation of Internet technologies, mobile devices, and web applications, attackers now use the Web as a vector for introducing malware into enterprise networks through employee's mobile devices in an environment such as Bring Your Own Device (BYOD). No wonder the Malware challenge remains the topmost challenge facing BYOD[1]. The personal mobile device is used to access a web application through the Internet either by typing a URL in the web browser or by clicking a

URL link to the web application. In any case, URLs serve as a means of obtaining access to web applications, thus making it an exploitable tool for attackers to infect malware into the device of their victim.

However, this change in attack vector has forced many organisations to subscribe to blacklisting services of malware URLs which are provided by a range of techniques including manual submission of suspected malware URLs and honeypots. With 571 new websites available on the Internet per minute[2], the blacklist approach to detect malware URLs is no longer sufficient as many new malware URLs are not blacklisted immediately they are launched on the Internet. More so, since the blacklist is created by volunteer experts, human error in detection is unavoidable. Exact matching in blacklisting also renders it easy to be evaded[3].

To address blacklisting challenges, a real-time anomaly based detection of malware URLs is necessary. This approach relies on a machine learning detection model that detects malware URLs as soon as they are encountered, without having to visit the blacklist server. To build such a machine learning detection model, the features of malware URLs play an important role. The selection of discriminative features for any detection algorithm determines the performance of the algorithm. The need for the selection of discriminative features for a malware URL detection model motivated this study. It should be mentioned here that recent studies of other researchers have used different categories of features for the detection of malicious URL (especially phishing and spam). To the best of the knowledge of the authors, little work has been done in the area of malware URL detection or classification. A recent survey[4], concerning malicious URLs (phishing, spam and malware) detection techniques reported works of Choi et al[3] and Eshete et al[5] as the only malware URL detection studies. Previous studies[3,6,7] used lexical features (textual properties) of URLs as discriminative features for malware URL detection in the case of Choi et al[3] and phishing URL detection in the case of Blum et al[6] and Le et al[7]. Similarly, our study identifies discriminative lexical features of malware URL through manual examination of blacklisted malware URLs. Also, to determine the level of consistency in the way attackers craft malware URLs, empirical analysis was carried out.

2. Methodology

The selection of a relevant feature set for any detection model is a process that requires careful attention. In practice, detection models tend to degrade in performance when faced with many features that are not necessary for predicting the correct label. In a situation where there are hundreds or thousands of features, the problem of selecting a subset of a relevant feature set for the best prediction accuracy is always a challenge for detection models. The detection model for malware URL is not left out of this challenge. To address this issue, we used two processes for selecting discriminative lexical features for malware URL detection. These processes include manual examination of URLs in an existing blacklist of malware URLs for identification of discriminative lexical features, and empirical analysis for studying the prevalence of identified features.

A malware patrol blacklist[8] was used to carry out a manual examination and empirical analysis. Malware patrol is a community of security experts that started operation in 2005 and it is a platform where anyone can submit a suspicious URL that may carry malware, viruses, or Trojans, or ransomware. When a URL is submitted, it is verified by security experts before it is added to the blacklist. The blacklist is updated every 1 hour for subscribers with a monthly payment subscription and every 48 or 72 hours for subscribers with a free subscription. Apart

from the fact that the malware patrol blacklist was used by previous studies[7,9], the hourly update is also a factor we considered before choosing the malware patrol blacklist as a source of malware URL data for our study. These processes are discussed in the next subsections.

2.1 Manual Examination of Malware URLS Blacklist

To carry out the manual examination, we downloaded the malware URL blacklist from malware patrol website on the 4th August 2015. On this day, a total of 62015 malware URLs were available on the blacklist. The URLs on the blacklist were manually examined in order to identify discriminative lexical features that make the blacklist URLs different from benign URLs. The discriminative lexical features were identified from three main components (protocol, hostname, and path) of a URL as shown in Figure 1.

<insert figure 1 here>

Based on these components, the feature set is grouped into three groups. Each group comprises of two or more features. The groups are URL to Path features, hostname features, and path features. It is important to note that the technicality behind the lexical structure of malware URL is beyond the scope of this study. Hence, the reason(s) behind the way malware URLs are crafted is/are not discussed in this study. Based on the feature set groups, the feature set identified during manual examination of the blacklisted malware URLs are presented below.

2.1.1 URL to Path Features Group

Two features were identified from this feature set group. These features include the following:

i. Length of URL from protocol to the path end

When we examined the URLs on the blacklist, we observed that some of the URLs have long character strings from the protocol to the end of path. Some URLs on the blacklist have as long as 250 characters. The following URL is an example of URL with long characters from the malware URL blacklist.

dde.integration.storage.conduit-services.com/39/233/ct2331539/cbdebcb46b4149109bd1ed6efbe14178/downloads/prod/dde1.3.8.4_perion.131024.04/13-11-05-21.50.02.936/

ii. Length of URL from protocol to the path end

Our manual examination revealed that many URLs have IP addresses in their hostname (the hostname is either replaced by the IP or the IP is added to hostname), path, and in some cases, both. This implies that the occurrence of the IP address in any part of the URL is a strong indication that the URL is a malware URL.

120.198.196.101/sanlixop/sanlix_data/sample/unknown/2013-03/2013-03-14/106714/

2.1.2 Hostname Features Group

Our manual examination of the blacklisted malware URLs revealed that the hostname of the malware URL is crafted in a form that is different from the hostname of benign URLs.

Consequently, five discriminative lexical features were identified. These features are described below.

i. Length of Hostname

During manual examination, it was observed that many URLs have long character strings which make them different from benign URLs. Example of this type of URL from the malware URL blacklist is given below.

dde.de.resource-efiles-drive.com/29/773/ct7739229/4caee31a80f04d0a83e40d536dba48eb/Downloads/Prod/SmallStub 1.3.9.0.140504.01/15-01-25-09.49.18.728/

ii. The Presence of www

Manual examination of malware URL blacklist revealed that many URLs from the blacklist do not have www. Very few URLs on the malware URL blacklist have www. All the examples of URL given above have no www. The following URL is another example of URL from the malware URL blacklist that has no www.

download2.77169.com/soft/hacrktools/attack/200906/

iii. The Presence of a Third Level Domain (TLD)

Manual examination of the malware URL blacklist revealed that many URLs on the blacklist have TLD. Example of this type of URL from the malware URL blacklist is as follows:

dl-2.one2up.com/onetwo/content/2014/6/12/

iv. The Presence of a Decimal Number in the Second Level Domain (SLD)

Many URLs on the malware URL black list have decimal number in their SLD. It was observed that some URLs SLDs have combination of decimal number(s) and alphabet(s). While some URLs on the malware URL blacklist have only decimal number(s) as their SLDs. Example of URL in this category is given below.

download5.77169.com/soft/other/2006/200612/

v. The Presence of a Decimal Number in the TLD

During manual examination of the malware URL blacklist, it was observed that many URLs on the blacklist have decimal numbers in their TLDs. Some of the URLs on the malware URL blacklist have only decimal number(s) as their TLDs. While some of the URLs have combination of decimal number(s) and alphabet(s) as their TLDs. URL below is an example of this category of malware URL from the blacklist.

56ffec5e.dl-one2up.com/onetwo/content/2015/9/27/

2.1.3 Path Features Group

The path features group represents features identified from the path of the URL. We identified five features from the URL path. These features are described below.

i. Length of the path

The length of the path of the malware URL was observed to be long in most of the blacklisted URLs. Example of this type of URL from the malware URL blacklist is given below.

s.ddirectdownload-about.com/82/288/ct2888182/67b7b53e3fc449c8a73307c88c60bb39/Downloads/Prod/DDE1.4.0.5.150121.02/15-02-17-18.05.10.828/

ii. Number of Subdirectories in the Path

When the malware URL blacklist was examined, it was observed that many of the URLs on the blacklist have two or more subdirectories in their paths. The URL below is an example of this type of URL from the malware URL blacklist. The URL has 8 subdirectories.

s.ddirectdownload-about.com/95/242/ct2427695/ea7f8d9e06d64be6b9730677d138730f/downloads/prod/dde1.4.0.5.150121.02/15-03-07-05.40.59.238/

iii. Length of Longest Subdirectory

During manual examination, it was observed that some of the URLs on the malware URL blacklist have one or more of their subdirectories very long. Below is an example of malware URL with the longest length of its subdirectory equal to 32.

218.207.102.106/1Q2W3E4R5T6Y7U8I9O0P1Z2X3C4V5B/dlsw.baidu.com/sw-search-sp/2015_05_08_20/bind1/36561/

iv. The Presence of a Date in the Path

Many URLs on the malware URL blacklist have a date in their path. It was observed that presence of a date in the path takes different formats. Some of the URLs on the malware URL blacklist have full date format (with month, day and year), while some have only year. Example of URL with dates is given below.

60.10.0.246/1103esv2013/files/322500000016514D/dlsw.baidu.com/sw-search-sp/2015_05_08_22/bind1/11006/

v. The Presence of Hexadecimal String in the Path

The last feature identified under this group is whether there is a hexadecimal character string in the path or otherwise. We observed that many URLs on the malware URL blacklist have a

hexa-decimal character string. The URL below is an example of URL with hexadecimal string in the path.

cdn1.mydown.yesky.com/55a6673a/df3b2fe23a66e96894a7ad6e3f5ddbd3/soft/200807/

2.2 Empirical Analysis

Some of the identified features are categorical (present or not present) while others are not. These categorical features include the presence of an IP, presence of www, presence of a date, whether the hostname has a TLD or otherwise, presence of a decimal number in a SLD, presence of a decimal number in the TLD, and whether a hexadecimal character string is present in the path or not. To study the prevalence of these features, we carried out an empirical analysis of 62103 malware URLs on the blacklist and on the newly collected (as the blacklist is updated) malware URLs. The purpose of this empirical analysis was to determine the level of consistency in the way attackers craft malware URLs. Details of the empirical analysis are described in the following subsections.

2.2.1 Analysis of 62013 URLs

Under this analysis, we extracted the total number of URLs having each of the categorical features. Then, the percentage of each feature appearance in the 62103 malware URLs was computed. Table 1 shows the result of the percentage appearance of each of the categorical features in the 62103 malware URLs.

2.2.1 Analysis of Newly Collected 18015 URLs

To study the prevalence pattern in which malware URL was crafted, we collected newly added malware URLs from [8]. This collection took place from 5th August 2015 to 13th October 2015 and resulted in a total of 18015 malware URLs in 30 rounds. Table 2 summarises the details of how the URLs were collected. While in all the 30 rounds, Tables 3, 4, 5, 6, 7, 8 and 9 show the percentage of the URLs with IP address, without www, with a date, whit a TLD, with a decimal number in the SLD, whit a decimal number in the TLD and with hexadecimal character string in the path respectively. Meanwhile, Table 10 shows the result of the percentage appearance of each of the categorical features in the 18015 malware URLs.

3. Results and Discussion

Figure 2 shows comparison of percentages of each of the categorical features in the 62103 and 18015 URLs. The percentage of the presence of decimal numbers in the TLD in the 62103 URLs was the same as the percentage of the presence of decimal numbers in the TLD in the 18015 URLs. The presence of www, presence of TLD, and presence of hexadecimal numbers in the path have almost the same percentage in both cases. Also, the percentages of the presence of an IP, presence of a date, and presence of decimal numbers in the SLD were slightly higher in the 62103 URLs than in the 18015 URLs. The implication of this is that the attackers tend to use to the same pattern of crafting malware URLs.

<insert figure 2 here>

However, Figure 2 shows that more than 80 % of the 62103 and 90 % of the 18015 URLs contain the TLD. This implies that many malware URLs are crafted to include the TLD. Our analysis revealed that many URLs with a decimal number in the SLD also have a decimal number in the TLD. The SLD and TLD belong to the same part (hostname) of the URL. We therefore combined the presence of a decimal number in the SLD and TLD to form a single feature. We refer to this feature as the presence of a decimal number in the hostname. Table 11 shows a summary of all features with their value type. It is important to note that these features are novel features for malware URL detection, although some of the features have been used for phishing or/and spam URL detection in previous studies. All the categorical features identified in this study with the exception of the presence of IP address are novel features which have not been used for any malicious URL detection in previous studies.

4. Conclusion and Future Work

In this paper, novel discriminative lexical features of malware URL's are identified and consistency in the way malware URLs are crafted by the attackers was also investigated. Our first step was to manually examine blacklisted malware URLs. This step led to the identification of 12 discriminative lexical features. The second step was an empirical analysis of the identified features of existing blacklisted malware URLs and newly collected malware URLs. Empirical analysis was carried out to determine whether there was consistency in the way malware URLs are crafted by the attackers. The results of our empirical analysis revealed that there is indeed consistency.

However, for the purpose of evaluation, the identified features in this study can be used to train any machine leaning algorithm for real-time detection of malware URL. Performance in term of accuracy and time to build detection model with both the previously used features and novel features identified can be compare with a view to identify best set of features for real-time detection of malware URL.

Table 1. Percentage of each of the categorical features in 62103 malware URLs.

Total URL	**62103**		
No.	**Features**	**No. of URL**	**% in Total URL**
1	Presence of IP address	11422	**18.39**
2	Presence of www	57296	**92.26**
3	Presence of a date in the path	27388	**44.10**
4	Presence of TLD	49815	**80.21**
5	Presence of a decimal number in the SLD	17233	**27.75**
6	Presence of a decimal number in the TLD	19218	**30.95**
7	Presence of hexadecimal in path	7988	**12.86**

Table 2. Details of how URLs were collected

Collection round	**Date interval**	**No. of days**	**No. of URL**
Round1	05-07/08/2015	3	**205**
Round2	08-09/08/2015	2	**149**
Round3	10-11/08/2015	2	**184**
Round4	12-14/08/2015	3	**177**
Round5	15-16/08/2015	2	**100**

Round6	17-18/08/2015	2	**47**
Round7	19-21/08/2015	3	**127**
Round8	22-23/08/2015	2	**1330**
Round9	24-25/08/2015	2	**978**
Round10	26-28/08/2015	3	**1783**
Round11	29-30/08/2015	2	**1329**
Round12	31-01/09/2015	2	**1400**
Round13	02-04/09/2015	3	**925**
Round14	05-06/09/2015	2	**457**
Round15	07-08/09/2015	2	**222**
Round16	09-11/09/2015	3	**464**
Round17	12-13/09/2015	2	**1451**
Round18	14-15/09/2015	2	**529**
Round19	16-18/09/2015	3	**1649**
Round20	19-20/09/2015	2	**329**
Round21	21-22/09/2015	2	**301**
Round22	23-25/09/2015	3	**583**
Round23	26-27/09/2015	2	**351**
Round24	28-29/09/2015	2	**368**
Round25	30-02/10/2015	3	**1018**
Round26	03-04/10/2015	2	**594**
Round27	05-06/10/2015	2	**114**
Round28	07-09/10/2015	3	**94**
Round29	10-11/10/2015	2	**71**
Round30	12-13/10/2015	2	**686**
TOTAL		**70**	**18015**

Table 3. The percentage of the URLs with IP address

Round	Total URL collected per round	Presence of IP	% of presence of IP
Round1	205	14	6.83
Round2	149	24	16.11
Round3	184	4	2.17
Round4	177	26	14.69
Round5	100	11	11.00
Round6	47	21	44.68
Round7	127	6	4.72
Round8	1330	115	8.65
Round9	978	31	3.17
Round10	1783	110	6.17
Round11	1329	150	11.29
Round12	1400	101	7.21
Round13	925	14	1.51
Round14	457	5	1.09
Round15	222	23	10.36
Round16	464	14	3.02
Round17	1451	44	3.03
Round18	529	30	5.67
Round19	1649	120	7.28
Round20	329	50	15.20
Round21	301	20	6.64
Round22	583	41	7.03

Round23	351	11	3.13
Round24	368	10	2.72
Round25	1018	80	7.86
Round26	594	89	14.98
Round27	114	15	13.16
Round28	94	10	10.64
Round29	71	5	7.04
Round30	686	104	15.16
TOTAL		18015	1298

Table 4. The percentage of the URLs without www

Round	Total URL collected per round	URLs without www	% URLs without www
Round1	205	195	**95.12**
Round2	149	144	**96.64**
Round3	184	121	**65.76**
Round4	177	142	**80.23**
Round5	100	83	**83.00**
Round6	47	39	**82.98**
Round7	127	122	**96.06**
Round8	1330	1287	**96.77**
Round9	978	956	**97.75**
Round10	1783	1721	**96.52**
Round11	1329	1371	**103.16**
Round12	1400	1371	**97.93**
Round13	925	906	**97.95**
Round14	457	449	**98.25**
Round15	222	217	**97.75**
Round16	464	456	**98.28**
Round17	1451	1404	**96.76**
Round18	529	507	**95.84**
Round19	1649	1613	**97.82**
Round20	329	315	**95.74**
Round21	301	291	**96.68**
Round22	583	546	**93.65**
Round23	351	320	**91.17**
Round24	368	344	**93.48**
Round25	1018	989	**97.15**
Round26	594	579	**97.47**
Round27	114	103	**90.35**
Round28	94	81	**86.17**
Round29	71	66	**92.96**
Round30	686	632	**92.13**
TOTAL	**18015**	**17370**	**96.42**

Table 5. The percentage of the URLs with a date in the path

Round	Total URL collected per round	URLs with a date	% of URLs with date
Round1	205	52	**25.37**
Round2	149	25	**16.78**

Round3	184	35	**19.02**
Round4	177	42	**23.73**
Round5	100	33	**33.00**
Round6	47	15	**31.91**
Round7	127	27	**21.26**
Round8	1330	157	**11.80**
Round9	978	135	**13.80**
Round10	1783	751	**42.12**
Round11	1329	497	**37.40**
Round12	1400	451	**32.21**
Round13	925	125	**13.51**
Round14	457	85	**18.60**
Round15	222	79	**35.59**
Round16	464	80	**17.24**
Round17	1451	135	**9.30**
Round18	529	111	**20.98**
Round19	1649	420	**25.47**
Round20	329	44	**13.37**
Round21	301	64	**21.26**
Round22	583	47	**8.06**
Round23	351	36	**10.26**
Round24	368	63	**17.12**
Round25	1018	109	**10.71**
Round26	594	101	**17.00**
Round27	114	17	**14.91**
Round28	94	11	**11.70**
Round29	71	27	**38.03**
Round30	686	121	**17.64**
TOTAL	**18015**	**3895**	**21.62**

Table 6. The percentage of the URLs with a TLD

Round	Total URL collected per round	URLs with TLD	% of URLs with TLD
Round1	205	181	**88.29**
Round2	149	129	**86.58**
Round3	184	96	**52.17**
Round4	177	122	**68.93**
Round5	100	74	**74.00**
Round6	47	36	**76.60**
Round7	127	114	**89.76**
Round8	1330	1242	**93.38**
Round9	978	908	**92.84**

Round10	1783	1653	**92.71**
Round11	1329	1259	**94.73**
Round12	1400	1341	**95.79**
Round13	925	853	**92.22**
Round14	457	434	**94.97**
Round15	222	214	**96.40**
Round16	464	451	**97.20**
Round17	1451	1323	**91.18**
Round18	529	486	**91.87**
Round19	1649	1552	**94.12**
Round20	329	286	**86.93**
Round21	301	249	**82.72**
Round22	583	468	**80.27**
Round23	351	256	**72.93**
Round24	368	293	**79.62**
Round25	1018	945	**92.83**
Round26	594	555	**93.43**
Round27	114	86	**75.44**
Round28	94	70	**74.47**
Round29	71	57	**80.28**
Round30	686	547	**79.74**
TOTAL	**18015**	**16280**	**90.37**

Table 7. The percentage of the URLs with a decimal number in the SLD

Round	**Total URL collected per round**	**URLs with a decimal No. in SLD**	**% of URL with a decimal No. in SLD**
Round1	205	23	**11.22**
Round2	149	16	**10.74**
Round3	184	21	**11.41**
Round4	177	35	**19.77**
Round5	100	26	**26.00**
Round6	47	7	**14.89**
Round7	127	27	**21.26**
Round8	1330	283	**21.28**
Round9	978	215	**21.98**
Round10	1783	350	**19.63**
Round11	1329	238	**17.91**
Round12	1400	261	**18.64**
Round13	925	201	**21.73**
Round14	457	47	**10.28**
Round15	222	33	**14.86**

Round16	464	95	**20.47**
Round17	1451	252	**17.37**
Round18	529	197	**37.24**
Round19	1649	345	**20.92**
Round20	329	92	**27.96**
Round21	301	89	**29.57**
Round22	583	197	**33.79**
Round23	351	33	**9.40**
Round24	368	39	**10.60**
Round25	1018	233	**22.89**
Round26	594	147	**24.75**
Round27	114	25	**21.93**
Round28	94	13	**13.83**
Round29	71	14	**19.72**
Round30	686	65	**9.48**
TOTAL	**18015**	**3619**	**20.09**

Table 8. The percentage of the URLs with a decimal number in the TLD

Rounds	**Total URLs collected per round**	**URLs with a decimal No. in TLD**	**% of URLs with a decimal No. in TLD**
Round1	205	21	**10.24**
Round2	149	21	**14.09**
Round3	184	32	**17.39**
Round4	177	34	**19.21**
Round5	100	35	**35.00**
Round6	47	12	**25.53**
Round7	127	56	**44.09**
Round8	1330	728	**54.74**
Round9	978	542	**55.42**
Round10	1783	451	**25.29**
Round11	1329	341	**25.66**
Round12	1400	356	**25.43**
Round13	925	320	**34.59**
Round14	457	142	**31.07**
Round15	222	45	**20.27**
Round16	464	111	**23.92**
Round17	1451	346	**23.85**
Round18	529	108	**20.42**
Round19	1649	434	**26.32**
Round20	329	131	**39.82**
Round21	301	115	**38.21**
Round22	583	107	**18.35**

Round23	351	70	**19.94**
Round24	368	54	**14.67**
Round25	1018	557	**54.72**
Round26	594	138	**23.23**
Round27	114	26	**22.81**
Round28	94	36	**38.30**
Round29	71	24	**33.80**
Round30	686	195	**28.43**
TOTAL	**18015**	**5588**	**31.02**

Table 9. The Percentage of the URLs with hexadecimal characters string in the path

Round	Total URL collected per round	URLs with hexadecimal in path	% of presence of URLs with hexadecimal in path
Round1	205	71	**34.63**
Round2	149	57	**38.26**
Round3	184	24	**13.04**
Round4	177	22	**12.43**
Round5	100	17	**17.00**
Round6	47	14	**29.79**
Round7	127	25	**19.69**
Round8	1330	662	**49.77**
Round9	978	612	**62.58**
Round10	1783	637	**35.73**
Round11	1329	531	**39.95**
Round12	1400	723	**51.64**
Round13	925	599	**64.76**
Round14	457	272	**59.52**
Round15	222	101	**45.50**
Round16	464	259	**55.82**
Round17	1451	379	**26.12**
Round18	529	211	**39.89**
Round19	1649	497	**30.14**
Round20	329	40	**12.16**
Round21	301	35	**11.63**
Round22	583	116	**19.90**
Round23	351	46	**13.11**
Round24	368	51	**13.86**
Round25	1018	105	**10.31**
Round26	594	63	**10.61**
Round27	114	19	**16.67**
Round28	94	13	**13.83**
Round29	71	11	**15.49**

Round30	686	61	**8.89**
TOTAL	**18015**	**6273**	**34.82**

Table 10. Percentage of each of the categorical features in 18015 malware URLs

Total URL	**18015**		
No.	**Features**	**No. of URL**	**% in Total URL**
1	Presence of IP	1298	**7.21**
2	Presence of www	17370	**96.42**
3	Presence of a date in the path	3895	**21.62**
4	Presence of TLD	16280	**90.37**
5	Presence of a decimal number in the SLD	3619	**20.09**
6	Presence of a decimal number in the TLD	5588	**31.02**
7	Presence of hexadecimal in the path	6273	**34.82**

Table 11. Summary of the proposed features with their value type

Feature groups	Features	Value type
URL to path	Length of URL to the path end	Integer
	Presence of IP address	Binary
Hostname	Length of the hostname	Integer
	Presence of www	Binary
	Presence of a TLD	Binary
	Presence of a decimal number in the hostname	Binary
Path	Length of the path	Integer
	Number of Subdirectory in the path	Integer
	Length of longest subdirectory in the path	Integer
	Presence of a date in the path	Binary
	Presence of Hexadecimal in the path	Binary

Figure 1. Components of a URL considered for feature set identification.
Figure 2. Comparison of percentage of each of the categorical features in both 62103 URLs and 18015 URLs.

5. References

1. Olalere M, Abdullah MT, Ramlan M, Abdullah A. A review on bring your own device on security issues. Sage Open. 2015; 05(02): 1-11. doi: 10.1177/2158244015580372 Published 10 April 2015
2. Ever wondered how many websites are created every minut? [Internet] 2014 June 11 [updated 2016 Jan 1; cited 2016 Apr 5]. Available from: http://www.designbyconet.com/2014/06/ever-wondered-how-many-websites-are-created-every-minute/.
3. Choi HS, Zhu BB, Lee H. Detecting malicious web links and identifying their attack types. Proceedings of the 2nd USENIX Conference on Web Application Development; 2011 USENIX Association Berkeley, CA, USA. ACM Digital Library; 2011. p. 1-11.
4. Patil DR, Patil JB. Survey on malicious web pages detection techniques. International Journal of U- and E- service, Science and Technology. 2015; 08(5):195-206.

5. Eshete B, Villafiorita A, Weldemariam K. BINSPECT: Holistic analysis and detection of malicious web pages. Proceedings of 8th International ICST Conference, SecureComm 2012; 2012 Sep. 3–5; Padua, Italy. Berlin: Springer; 2013. p. 149-166.
6. Blum A, Wardman B, Solorio T, Warner G. Lexical feature based phishing URL detection using online learning. Proceedings of the 3rd ACM workshop on Artificial Intelligence and Security; 2010 October 04 – 08; Chicago, Illinois, USA. ACM; 2010. p. 54-60.
7. Le A, Markopoulou A, Faloutsos M. PhishDef: URL names say it all. Proceedings of IEEE INFOCOM, 2011; 2011 April 10-15; Shanghai, China. IEEE; 2011.p. 191–195.
8. Malwarepatrol. Available at http://www.malwarepatrol.net.
9. Kalafut AJ, Shue CA, Gupta M. Malicious hubs: detecting abnormally malicious autonomous systems. Proceedings of IEEE INFOCOM, 2010 Conference. 2010 March 14-19; San Diego, CA, USA. IEEE; 2010. p. 1-5.