A method for exploring traffic passive traces and grouping similar urls

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
15 March 2018 (15.03.2018)

W!P O ¦ PCT

(10) International Publication Number
WO 2018/047027 Al

(51) International Patent Classification:
H04L 29/06 (2006.01)

(21) International Application Number:
PCT/IB20 17/054786

(22) International Filing Date:
04 August 2017 (04.08.2017)

(25) Filing Language: Italian

(26) Publication Language: Englíśh

(30) Priority Data:
102016000091521 12 September 2016 (12.09.2016) IT

(71) Applicant: **POLITECNICO DI TORINO** [IT/IT]; C.so Duca degli Abruzzi, 24, 10129 Torino (IT).

(72) Inventors: **MELLIA, Marco;** c/o POLITECNICO DI TORINO, C.so Duca degli Abruzzi, 24, 10129 Torino (IT). **METWALLEY, Hassan;** c/o POLITECNICO DI TORINO, C.so Duca degli Abruzzi, 24, 10129 Torino (IT). **BOCCHI, Enrico;** c/o POLITECNICO DI TORINO, C.so Duca degli Abruzzi, 24, 10129 Torino (IT). **MORICHETTA, Andrea;** c/o POLITECNICO DI TORINO, C.so Duca degli Abruzzi, 24, 10129 Torino (IT).

(74) Agent: **LISA, Elisabetta** et al; c/o PRAXI INTELLECTUAL PROPERTY S.p.A., Corso Vittorio Emanuele II, 3, 10129 Torino (IT).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available):* AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available):* ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Declarations under Rule 4.17:
— *of inventorship (Rule 4.1 7(iv))*

Published:
— *with international search report (Art. 21(3))*

WO 2018/047027 A1

(54) Title: A METHOD FOR EXPLORING TRAFFIC PASSIVE TRACES AND GROUPING SIMILAR URLS

(57) Abstract: Computer security method for the analysis of passive traces of HTTP and HTTPS traffic on the Internet, with extraction and grouping of similar Web transactions automatically generated by malware, malicious services, unsolicited advertising or other, comprises at least the following processing and control steps: a) URLs extraction from an operational network, using passive exploration of the HTTP e HTTPS traffic data and subsequent collection into batches of the extracted URLs; b) detection of similar URLs, by metrics calculation based on the distance among URLs, namely based on a measure of the degree of diversity among pairs of character strings composing the URLs; c) activation of one or more clustering algorithms used to group the URLs based on the similarity metrics and to obtain, within each group of URLs, elements with similar/homogeneous features, adapted to be analyzed as a single entity; d) visualization of elements according to a sorting based on the degree of cohesion of the URLs contained in each grouping.

A METHOD FOR EXPLORING TRAFFIC PASSIVE TRACES AND GROUPING SIMILAR URLS

DESCRIPTION

The present invention relates to a method of computer security for the analysis of traces of HTTP traffic on the Internet (Hyper Text Transfer Protocol - a standard application protocol used as the main system for the transmission of information on the Web), finalized to the extraction and grouping of similar Web transactions generated in an automatic way by malware, malicious services, unsolicited advertising or other. With Web transactions are intended HTTP and HTTPS requests and responses containing within them the URL (Uniform Resource Locator - a unique address of a resource present on the Internet, by which transactions are identified).

In the current state of the art there are some prior documents, US7680858, US7962487, US7376752, EP2291812, WO2013009713, but none of these documents uses the innovative features of the present invention described below, which allow to obtain better performances and greater benefits.

Specifically, US7680858: it performs a normalization of URLs (unique address of a resource present on the Internet) by dividing them into "levels" of information; the measure of the variation between two URLs is calculated on the basis of the "differences" of keywords (search keys); it also uses information about the "content" of the page.

US7962487: it is oriented only towards the improvement of search engines; it relies on clustering (grouping) of tokens (categorized text blocks) associated with search queries (questions).

US7376752: it divides the URL into two parts; the distance among URLs is calibrated so as to recognize typing mistakes.

EP2291812: it is based on the page "content"; it creates a set of features from every page, on which it calculates the "distance" among URLs.

WO2013009713: it aims to the recognition of phishing pages; it searches "relationships" among phishing page files to determine their similarity.

In the scientific literature, there are thus two types of works relating to the subject matter of the present invention, in the first of which are included all those jobs that aim to classify a page processing only the "content" present in it, or the Web address of a page (URL). In this case, thus, only algorithms of "text recognition" are used, that represent only a part of the present invention. The methodologies present in this type of works, however, require a high computational cost for processing the text of billions of Web pages, and aim to recognize the "subject" of each page, consequently their objectives are totally different from those of the present invention.

The second type, on the other hand, comprises all those works that apply data-mining techniques (data extraction and processing) in URLs to detect only "some types" of cyber-attacks, such as phishing or spam.

Therefore, the present invention is much more complete and universal if compared to the current state of the art. Actually, utilizing various and suitably adapted/edited "text recognition" algorithms and "clustering" algorithms (not supervised techniques developed in the field of data-mining to extract information from large amounts of data), a quantity decidedly greater of "artificial" and/ or "malicious" traffic may be detected.

Therefore, the present invention comes to help network administrators and/ or computer security analysts to extract information from the Web traffic generated by networks having thousands of computers. Without tools that could help analysts, actually, detecting problems or faults becomes very difficult when considering data blocks including billions of Web transactions.

The present method inspects traces of Web traffic generated by real users or automatic bots. For each pair of network transactions present in a trace, the "degree of lexical similarity" is then calculated, and "similar" transactions are subsequently "grouped" together to form homogeneous groups that are presented to the network analyst or to the security expert, sorted by "importance".

The present method, particularly, allows to detect automatically and make easily visible all that traffic that is not generated by human users, but by "automatic systems", also called bots (robots) in the technical jargon. This type of traffic, actually,

is often generated by malware or other malicious services, thus a methodology of this kind can be crucial for reducing the time that passes between a cyber-attack and its discovery (on average, about 150-180 days) or for recognizing faults that cause malfunctions in networks.

The present invention differs from the prior art for the following reasons:

- it is based solely on the analysis of URLs and their syntax (address of an Internet resource), ignoring the "content" of the page or other information;

- it does neither analyze nor use particular structural features of URLs, but maintains a neutral point of view, checking only the "similarity" among pairs of URLs;

- it uses techniques based on "non-supervised algorithms", and therefore, a priori, it does not require the use of any kind of knowledge or information;

- it is based solely on the calculation of the "syntactic similarity" among the various URLs, avoiding the need to have a set of pre-labelled elements, and preventing, in this way, also problems of excessive adaptation of the used algorithm.

Inspired by text-mining algorithms (text extraction and processing), the concept of "distance" among URLs is introduced, used to compose "groups" of URLs by means of the well-known clustering algorithm DBSCAN *(Density-Based Spatial Clustering of Applications with Noise),* based on the "density" since it connects regions of points with a sufficiently high density.

In order to better illustrate how the clustering algorithms based on "density" work, a set of points in a sample space to cluster should be considered. D(xl,x2) is the distance between the points xl and x2. Now the sphere having a radius *E* centred in xl is considered. If at least a minimum number of points *(minPoints)* is within the distance *E* from xl, the point xl is classified as "central point". Formally, a given point xl is a "central point" if at least a minimum number of points *(minPoints)* is within the distance *E* from it. These points are defined as "directly reachable" by xl. A generic point xk is "reachable" by xl if there is a path xl,x2,...,xk so that xi+1 is directly reachable by xi. The points reachable by xl form a "cluster", i.e. a "dense" region. The points that are not reachable by xl are called "anomalous values", and may form a separate cluster, if they belong to another dense region, or be included in the so-called

"noise" region. The parameters *minPoints* and *E* are adjustable and can be set by an expert in domains. The parameter *minPoints* defines the minimum size of a cluster and has little impact on final results. The parameter *E*, on the other hand, is a fundamental parameter. If it is set to a too small value, it leads to a high number of small groups and to many points which cannot be clusterized/ grouped. On the other hand, if it is set to a too high value, it leads to a few groups with a multitude of heterogeneous points. A sensitivity analysis is therefore essential to correctly choose the value of the radius *E*.

The groupings thus generated are subsequently sorted to help the visualization for the network administrator or the security expert. The sorting is done by considering the cohesion degree of the elements inside each grouping.

The present invention therefore solves the problem of processing the input data, of aggregating them syntactically and semantically, and showing them to the analyst coherently and consistently, and sorted by importance.

The subject method of the present invention is also capable of offering an aggregate analysis tool of Web traffic, allowing to detect in a simple and direct way Web transactions linked to malicious services, or supplied by automatic systems such as those generating advertising, tracking systems, or in general, interesting for the network administrator or the security expert.

The above and other objects and advantages of the invention, as will appear from the following description, are achieved with the method described in claim 1.

Preferred embodiments and non-trivial variations of the present invention form the subject matter of the dependent claims.

It is understood that all the appended claims form an integral part of the present description.

It will immediately appear obvious that numerous variations and modifications to what described could be made, without departing from the scope of protection of the invention, as it results from the appended claims.

The invention relates to a method of computer security for the analysis of traces of HTTP and HTTPS traffic on the Internet, finalized to the extraction and grouping of

"similar" Web transactions generated in an "automatic" way by malware, malicious services, unsolicited advertising or other.

The main objectives of the present method are essentially:

- reducing the number of elements that the analyst should visualize and process, from hundreds of millions of single transactions to a few hundreds of clusters (groups with similar/ internally homogeneous elements);
- identifying the transactions generated "automatically", for example transactions generated by advertising platforms, polymorphic malware and/ or systems of the wiki-like type.

Specifically, the subject method of the invention comprises at least the following steps of processing and control:

a) extraction of transactions from an operational network, by means of exploration of the HTTP and HTTPS traffic data, and subsequent collection into batch (groups of elements) of the extracted transactions;

b) detection of similar transactions, by metrics calculation based on the "similarity" among pairs of transactions, namely based on a measure of the degree of "diversity" among pairs of character strings composing the URLs;

c) activation of one or more "clustering" algorithms, used to group the transactions on the basis of a similarity metrics, obtaining in this way, within each group of transactions, elements with similar/ homogeneous features, which can thus be analyzed as a "single" entities, considerably reducing the number of elements to be analyzed, facilitating and accelerating the work of analysis and research of the malicious and/or unwanted Internet traffic generated artificially/ automatically;

d) sorting of the transaction groups on the basis of their importance, i.e. of the degree of cohesion of the transactions contained in groupings.

The extraction of transactions takes place via the network/ passive probe for extraction and filtering of traffic, located in a specific link, which processes the data packets in real time, extracts the transactions and then groups them in specific batches for subsequent processing.

Once a batch of transactions is formed, the "distance" among all transactions pairs is then calculated, i.e. the level of likelihood/ similarity, such distance being calculated considering the entire URL as a single string of characters, composed of both "hostname" (identifier name of a device within a network of computers), and "path" (path).

To detect similar URL, a distance among pairs of strings is used, belonging to the "edit-distance" class, suitable to calculate the dissimilarity of pairs of strings of characters composing the URLs, being considered the "distance" among pairs of strings of characters as the minimum number of steps required to convert one of the two strings into the other.

In the state of the art, the most popular technique is the so-called distance of Levenshtein, that assigns a unit value to all editing operations, i.e. insertion, deletion and substitution of one character. It calculates an absolute distance among pairs of strings, that is equal to the length of the longest string at max. This, however, makes the technique of the distance of Levenshtein scarcely convenient when comparing a short URL and a long one (in this case, the URL length may extend from a few to hundreds of characters).

Unlike various known techniques, in the present method, for calculating the "distance" among strings of characters composing URLs, the following conditions apply:

- the "insertion" of a character has a value of 1;
- the "deletion" of a character has a value of 1;
- the "substitution" of a character has a value of 2, the substitution being equivalent to a deletion plus an insertion;
- the obtained value is normalized in the range between 0 and 1 by adding all the previous operations necessary to match the two strings (i.e. insertions, deletions and/ or substitutions) and dividing this value by the sum of the lengths of the two strings;
- the similarity of two strings of URLs characters thus varies in a normalized range of values comprised between 0 and 1, so as to obtain that a pair of identical strings has a distance equal to 0, and a pair of completely different strings has a distance equal to 1.

A pair of similar URLs has a small distance, while a pair of different URLs has a great distance.

Said one or more "clustering" algorithms, used for grouping the URLs on the basis of similarity metrics, group the URLs in a same set when these have a high value of similarity (i.e. low distance).

For the purposes of the present invention, the known clustering algorithm called DBSCAN is preferably used, based on the calculation of the "density" of the elements present within a certain area.

Then, the network administrator or the security expert are provided with a visualization of these groupings of transactions, sorted according to the degree of cohesion, starting from the most cohesive grouping.

In detail, for this task an analysis tool called "coefficient of silhouette" is used. This coefficient, which is based on the concepts of cohesion and separation, provides that a cluster is identified as cohesive if the elements therein are mutually very close. In addition, a cluster is well separated if its points are distant from those of other clusters. Thus, with the coefficient of silhouette how well each point is included in a cluster is evaluated.

Given a point i, a(i) is the average distance among that point and all the other points of the cluster they belong to. In this way, how well the point i is included in its grouping is calculated. On the other hand, with b(i) the average of the lowest distances among i and all the other points of the remaining clusters is defined. The silhouette is thus defined as the ratio between the difference between b(i) and a(i) and the maximum value between a(i) and b(i), obtaining consequently values included the range between 0 and 1. The higher is s(i), the more i is similar to its own cluster. In particular, if the value of silhouette is > 0, it means that the mean distance among i and the other objects in its grouping is lower than the minimum average distance with respect to the elements of all other clusters. For $s(i) < 0$ the opposite of what has just been specified above applies.

The method of the present invention is therefore based solely and advantageously on the URLs "syntax", ignoring the "content" of pages or other information.

CLAIMS

1) Computer security method for the analysis of passive traces of HTTP and HTTPS traffic on the Internet, with extraction and grouping of similar Web transactions automatically generated by malware, malicious services, unsolicited advertising or other, characterized in that it comprises at least the following processing and control steps:

a) URLs extraction from an operational network, using passive exploration of the traffic data and subsequent collection into batches of the extracted URLs;

b) detection of similar URLs, by means of metrics calculation based on the similarity among URLs, namely based on a measure of the degree of diversity among pairs of character strings composing said URLs;

c) activation of one or more clustering algorithms used to group the URLs based on a similarity metrics, and to obtain, within each group of URLs, elements with similar/ homogeneous features adapted to be analyzed as a single entity;

d) sorting said URLs into groups according to their importance, namely the degree of cohesion of the URLs contained in said groupings.

2) Method according to claim 1, characterized in that said extraction of URLs is performed by network/ passive probe for exploration and filtering, located in a specific link, adapted to process the data packets in real time, for extracting and downloading the URLs in specific batches for subsequent processing.

3) Method according to claim 2, characterized in that when an HTTP/ HTTPS transaction is detected, the contained URL is recorded in a specific file.

4) Method according to claims 2 and 3, characterized in that, once a lot of URLs is formed, the distance between all pairs of the various URLs is calculated, namely the level of likeness/ similarity, said distance being calculated considering the entire URL as a single string of characters, composed by both hostname and path.

5) Method according to one or more of the preceding claims 1 to 4, characterized in that to detect similar URLs a similarity metrics among pairs of strings is used adapted to calculate the dissimilarity of pairs of strings of characters composing the URLs, the

distance among pairs of strings of characters as the minimum number of steps needed to convert one of the two strings into the other being considered.

6) Method according to one of the preceding claims 1 to 5, characterized in that, for the calculation of the distance among the pairs of strings of characters composing the URLs, the following conditions apply:

- the insertion of a character has a value of 1;
- the deletion of a character has a value of 1;
- the substitution of a character has a value of 2, the substitution being equivalent to a deletion plus an insertion;
- the normalization between 0 and 1 of the previous value obtained by the sum of the operations to match the two strings divided by the sum of the lengths of the two strings;
- the similarity of a pair of strings of URL characters varying in a normalized range of values between 0 and 1, obtaining consequently that a pair of identical strings has a distance equal to 0, and a pair of completely different strings has a distance equal to 1.

7) Method according to one of the preceding claims 1 to 6, characterized in that a pair of similar URLs has a small distance, while a pair of different URLs has a great distance.

8) Method according to claim 1, characterized in that said one or more clustering algorithms are adapted to be used for grouping the URLs based on a similarity metrics.

9) Method according to claim 8, characterized in that preferably a clustering algorithm DBSCAN is used, based on the calculation of density of elements present within a certain area.

10) Method according to claim 9, characterized in that said groupings generated using the clustering algorithm DBSCAN are sorted according to the degree of cohesion among the URLs contained therein.

11) Method according to claim 10, characterized in that a coefficient of silhouettes is used, based on the calculation of both cohesion and degree of separation for all the elements of each grouping.

12) Method according to one or more of the preceding claims 1 to 11, characterized in that it is based solely on the URLs syntax.

5

INTERNATIONAL SEARCH REPORT

International application No
PCT/IB2017/054786

A. CLASSIFICATION OF SUBJECT MATTER
INV. H04L29/06
ADD.

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
H04L G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
EPO-Internal , WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	ROBERTO PERDISCI ET AL: "Behavioral Cl ustering of HTTP-Based Malware and Si gnature Generati on Using Mal i ci ous Network Traces" , USENIX, , 18 March 2010 (2010-03-18) , pages 1-14, XP06101Q768, abstract Secti ons 1, 3-5	1-12
X	US 2011/283361 Al (PERDISCI ROBERTO [US] ET AL) 17 November 2011 (2011-11-17) abstract paragraph [0008] - paragraph [0046] paragraph [0070] - paragraph [0137] paragraph [0145] - paragraph [0165] f i gures 1-6	1-12

-/ -

[X] Further documents are listed in the continuation of Box C. [X] See patent family annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) orwhich is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
17 October 2017	25/10/2017

Name and mailing address of the ISA/	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016	Bertol i ssi , Edy

INTERNATIONAL SEARCH REPORT

International application No
PCT/IB2017/054786

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	Anthony Verez : "On the Use of Data Mining Techniques for the Clustering of URLs Extracted from Network-based Malware Traces" , , 18 February 2014 (2014-02-18) , XP055386323, Retrieved from the Internet: URL: http://verez.net/docs/malwurl_paper.pdf [retrieved on 2017-06-29] abstract Sections 5 and 6	1-12
X	Piotr Kijewski : "Automated Extraction of Threat Signatures from Network Flows" , 18th Annual FIRST Conference, 25 June 2006 (2006-06-25) , XP055386193, Baltimore, Maryland Retrieved from the Internet: URL: https://www.first.org/resources/papers/conference2006/kijewski-piotr-papers.pdf [retrieved on 2017-06-28] Introduction , Architecture, Classifying the extracted signature	1-12
A	US 2014/297640 Al (DUFTLER MATTHEW J [US] ET AL) 2 October 2014 (2014-10-02) abstract paragraph [0046] - paragraph [0050]	1-12
T	ANDREA M0RICHETTA ET AL: "CLUE: Clustering for Mining Web URLs" , 2016 28TH INTERNATIONAL TELETRAFFIC CONGRESS (ITC 28) , 12 September 2016 (2016-09-12) , pages 286-294, XP055386135 , D0I : 10. 1109/ITC-28. 2016. 146 ISBN : 978-0-9883045-1-2 the whole document	

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No
PCT/IB2017/054786

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2011283361 Al	17-11-2011	US 2011283361 Al	17-11-2011
		US 2015026808 Al	22-01-2015
US 2014297640 Al	02-10-2014	US 2014297640 Al	02-10-2014
		US 2014298341 Al	02-10-2014