International Journal of Security and Its Applications
Vol. 12, No. 2 (2018), pp.9-18
http://dx.doi.org/10.14257/ijsia.2018.12.2.02

Classification of Malware Analytics Techniques: A Systematic Literature Review

Nur Farhana Hordri[1,2*], Noor Azurati Ahmad[1], Siti Sophiayati Yuhaniz[1], Shamsul Sahibuddin[1], Aswami Fadillah Mohd Ariffin[3], Nur Afifah Mohd Saupi[3], Nazri Ahmad Zamani[3], Yasmin Jeffry[3] and Mohamad Firham Efendy Md Senan[3]

[1]*Advanced Informatics School, Universiti Teknologi Malaysia, Malaysia*
[2]*UTM Big Data Centre, Universiti Teknologi Malaysia, Malaysia*
[3]*Cyber Security Malaysia, Malaysia*
**nfarhana64@live.utm.my*

Abstract

Context: Malware is a variety of forms of hostile or intrusive software that being thrown around online. Data analytics is the process of examining data sets in order to draw conclusions about information they contain, increasingly with the aid of specialized systems and software. Objectives: The aims of the study are to identify the types of malware analytics and identify the purpose of malware analytics. Method: A Systematic Literature Review (SLR) was carried out and reported based on the preferred reporting items for systematic reviews. 1114 papers were retrieved by manual search in six databases which are IEEE, Science Direct, Taylor and Francis, ACM, Wiley and Springer Link. 53 primary studies were finally included. Results: From these studies, 70% were conference papers and 30% were journal articles. Five classification of malware analytics techniques were identified and analysed. The classifications are (1) descriptive analytics, (2) diagnostic analytics, (3) predictive analytics, (4) prescriptive analytics and (5) visual analytics. Conclusion: This review delivers the evidence that malware analytics is an active research area. The review provides researchers with some guidelines for future research on this topic. It also provides broad information on malware analytics techniques which could be useful for practitioners.

Keywords*: Malware Analytics; Analytics Techniques; Systematic Literature Review*

1. Introduction

Malware is a software that is built to be malicious including computer viruses, worms, trojan horses, ransomware, spyware, adware and many more [1]. Malware is defined by its malicious intent, acting against the requirements of the computer user and does not include software that causes unintentional harm due to some deficiency. However, the web is rich with signals of data breaches, information about newly vulnerable targets, and evidence of pre-planned attacks, but it is nearly impossible to organize all this threat intelligence with manual or ad-hoc systems. The reasons of growing web-based malware in web world are to increase cyber-crimes [2]. Hence, analytics can be intended as intricate procedures running over large scale of data repositories as its main goal is that of mining useful knowledge kept in such repositories [3]. The cyber Threat Intelligence Analytics (TIA) should help organizations discover, visualize, and communicate meaningful insights from a variety of sources. These sources could be from the private feeds listed above, to open-source data, to network logs, enterprise data, and social media [4].

Received (August 23, 2017), Review Result (November 20, 2017), Accepted (November 24, 2017)

ISSN: 1738-9976 IJSIA

In recent years, there has been an increasing interest in malware prediction techniques. In 2008, [5] Konrad and co-workers has aimed to exploit behavioral patterns for classification of malware and proposed a method for learning and discrimination of malware behavior. They have applied Machine Learning (ML) technique which is Support Vector Machine (SVM) to identify the shared behavior of each malware family. In the same vein, Yusoff and Jantan (2011) proposed the usage of Genetic Algorithm to optimize the malware classification system as well as malware prediction. This new malware classification system has an ability to train and learn by itself, so that it can predict the current and upcoming trend of malware attacks [6]. As latest as 2016, an interesting research by [7] attempted to use Deep Learning (DL) to model the malware system call sequences for malware classification. The results of DL outperform Hidden Markov Model and SVM techniques in malware classification.

Apart from that, we have not found any paper that is related to Systematic Literature Review (SLR) on malware analytics techniques. Therefore, the purpose of this paper is to systematically review the current literature on malware analytics techniques. Findings may assist researchers in to design a powerful study that could appropriately determine the potential of the classification malware analytics in cybercrimes.

2. Review Method

In this paper, we have selected SLR as our method to identify the types and purpose of malware analytics. We used SLR guidelines as in research of [8] which SLR is a form of secondary study that uses a well-defined methodology. According to the guidelines, there are three repeated phases in SLR which is planning, conducting and reporting. In this section, we will focus on the planning phase which involves defining the research objectives and how the review is carried out.

2.1. Review Design

This section describes the groundwork of the review by defining the SLR research questions and search keywords.

2.1.1. SLR Research Questions (RQ)

As mentioned above, this paper intended to identify the types for malware analytics. The SLR RQ that we need to answer in this paper is as follows:

"*What are different types and purpose of malware analytics?*"

2.1.2. Search Process

This paper has selected six databases to perform the SLR search process. There are IEEE Xplore, Science Direct, Taylor and Francis, ACM Digital Library, Wiley Online Library and Springer Link. The following search keywords are used to find relevant studies in paper's title, keywords and abstract:

"*Malware AND Analytics AND Techniques*"

2.2. Review Conduction

This section defines the review protocol for conducting the SLR. The SLR review protocol refers to structure and rules of conducting the review.

2.2.1. Inclusion and Exclusion Criteria

According to an SLR from [9], the inclusion criteria are the title and abstract must be written in English, the paper is a full-text article and the focus of the paper is

based on RQ of the SLR. While for the exclusion criteria are the paper is not written in English, cannot access the articles, duplicate studies and short papers.

2.2.2. Quality Assessment (QA)

According to the guidelines of SLR [8], four Quality Assessment (QA) questions must be defined in order to assess the quality of the research of each proposal and to provide a quantitative comparison between them. The scoring procedures are Y (Yes = 1), P (Partly = 0.5) and N (No = 0). The quality assessment questions defined in this SLR were:

1. Was the articles referred?
a. Yes: it either explicitly describe the types of malware analytics
b. Partially: it only mentioned a few either the types of malware analytics
c. No: it neither described nor mentioned types of malware analytics
2. How clearly are the work limitations documented?
a. Yes: it clearly explained the limitation the types of malware analytics
b. Partially: it mentioned the limitation but did not explain why
c. No: it did not mention the limitation
3. Were the findings credible?
a. Yes: the study was methodologically explained so that the finding can be trust
b. Partially: the study was methodologically explained but not in details
c. No: the study was not methodologically explained
4. How well has diversity of perspective and context been explored?
a. Yes: it explicitly explains various perspectives on types of malware analytics
b. Partially: it mentioned the various perspectives on types of malware analytics
c. No: it did not mention the various perspectives on types of malware analytics

2.2.3. Data Extraction

Table 1 indicates the data extraction form that is employed for all selected primary studies to carry out an in-depth analysis.

Table 1. Data Extracted Form

No	Extracted Data	Description	Type
1	Identity of study	Unique identity for the study	General
2	Bibliography references	Authors, year of publication, title and source of publication	General
3	Type of study	Book, journal paper, conference paper, workshop paper, white paper	General
4	The types of malware analytics	Description of the types of malware analytics	RQ
5	Findings/Contributions	Indicating findings and contributions of the study	General

2.2.4. Synthesis

Results from the analysis through SLR revealed 53 studies for further consideration. Figure 1 shows the number of studies after each defined process. While Figure 2 indicates the number of type of study, which stands in selected paper for review and type conference has the higher selected study per type compared to journal. Figure 3 illustrates a temporal trend of the included articles related to malware analytics issues. The linear trend indicates the studies on the malware analytics issues increased steadily throughout the years. As we can see in Figure 3, a surge of publication on malware analytics has been recorded in 2016 with 14 publications.

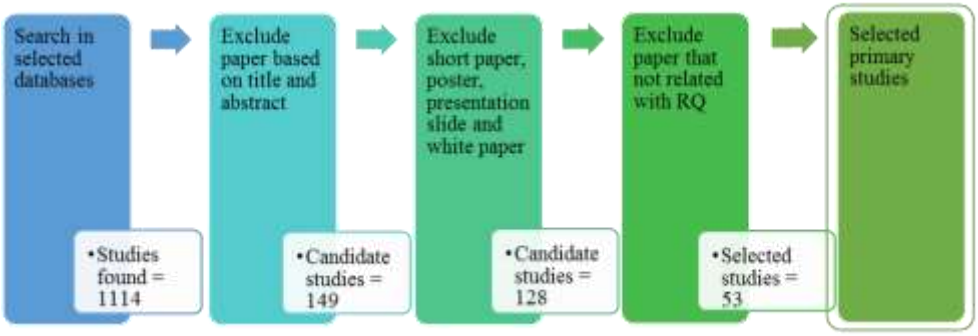

Figure 1. Finding Primary Studies Procedure

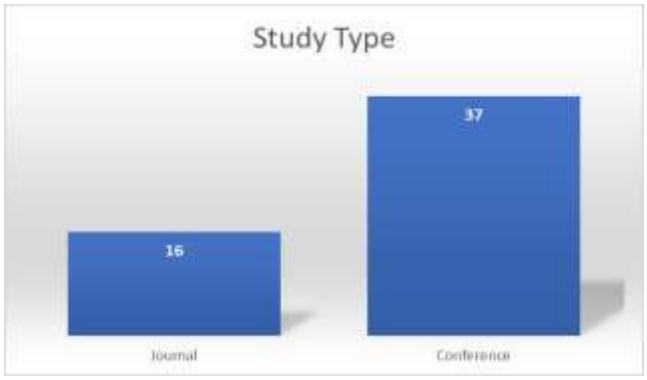

Figure 2. Numbers of Selected Study per Type

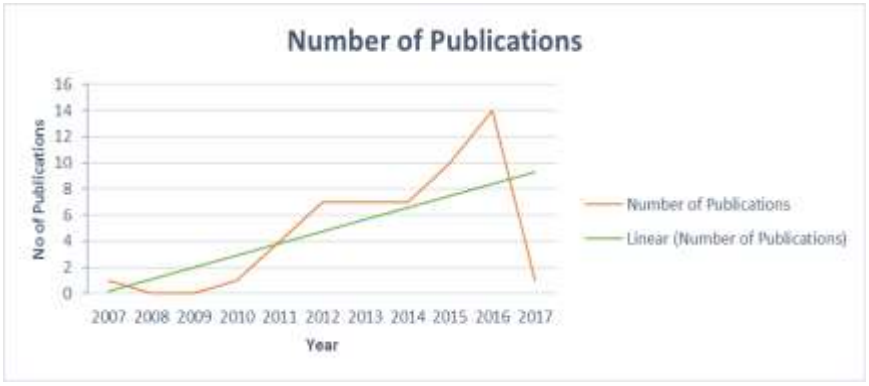

Figure 3. Types of Malware Analytics Publications

3. Results

This section explains the findings and discussion of this review to answer the defined SLR RQ.

3.1. Finding RQ

RQ: What are different types and purpose of malware analytics?

As regards with this RQ, we analyzed the data and as a result, five types of malware analytics are identified. They are (1) Descriptive analytics, (2) Diagnostic analytics, (3) Predictive analytics, (4) Prescriptive analytics and (5) Visual analytics. According to 53 papers that we have read, Figure 4 shows the percentage of each type of malware analytics. Each paper can have one or up to five types of malware analytics that have been

discussed. Table 2 depicts the number of primary studies addressing the identified type of malware analytics. Type (1), (2) and (3) have more than 20 papers addressed the types of malware analytics which is 21, 22 and 28 papers respectively. While Type (4) and (5) have almost same number of papers which are addressed by 8 and 9 papers respectively.

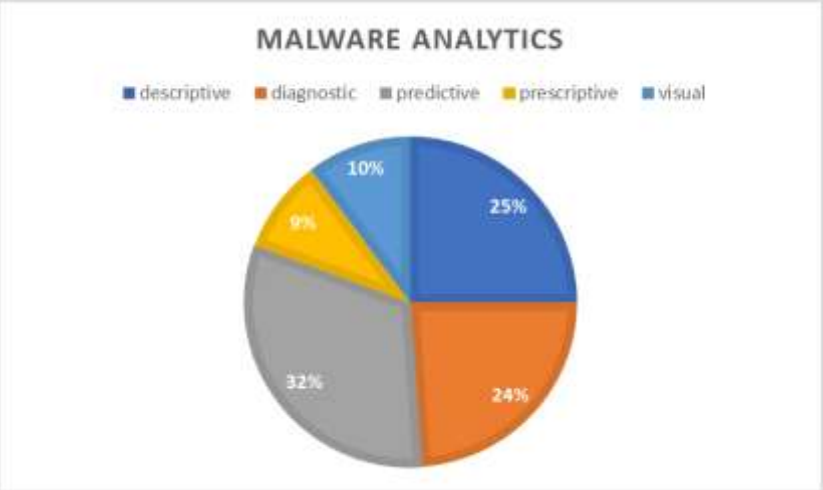

Figure 4. Percentage of Each Type of Malware Analytics over 53 Papers

Table 2. Number of Primary Study Addressing the Identified Types

No	Types	Number of Papers	Study Identifiers
1	Descriptive analytics	22	S1, S5, S6, S7, S8, S13, S16, S19, S20, S21, S27, S28, S32, S35, S38, S44, S45, S47, S49, S50, S51, S53
2	Diagnostic analytics	21	S1, S5, S6, S7, S8, S13, S19, S20, S21, S27, S28, S32, S35, S38, S44, S45, S47, S49, S50, S51, S53
3	Predictive analytics	28	S2, S3, S4, S5, S6, S7, S8, S11, S12, S13, S17, S19, S23, S26, S27, S29, S30, S31, S32, S33, S38, S43, S45, S46, S47, S49, S50, S53
4	Prescriptive analytics	8	S2, S34, S36, S37, S40, S42, S52, S53
5	Visual analytics	9	S9, S14, S15, S22, S25, S39, S40, S41, S45

3.2. Quality Assessment

Once the primary studies of the SLR have been identified, we evaluated them according to the QA questions defined in Section 2.2.2. The score assigned to each study for each question is shown in Table 3.

The row "% total score" shows the percentage of points obtained by all the selected study with regard to the total number of points obtained by all the selected studies in all the QA questions. The last row "% max QA" corresponds to the percentage of points collected by the values assigned for a given QA question over the points that would be collected if every selected study got the highest score. [*i.e* (Total each QA / Total paper) * 100]

The highest score with a score of 3 obtained by S6, S7, S19, S32, S38, S45, S47, S49, S50 and S53 which represents about 75% of the maximum possible. In contrast, 20% of the selected studies obtained a score of 2.5 and representing 62.5 % of the maximum score that one primary study could get. There are 4 studies that could not get any score

which mean that their title and abstract shown that it can give the answer for the RQ for this SLR but after going through the full articles, there is no malware analytics techniques has been discussed. QA3 has the highest percentages of points collected which is 66.04% and follows by QA1, QA2 and QA4 respectively.

Table 3. Quality Assessment of Selected Papers

ID	QA1	QA2	QA3	QA4	Total score	% by Max S
s1	P	P	Y	P	2.5	62.5
s2	P	P	Y	P	2.5	62.5
s3	P	P	P	N	1.5	37.5
s4	P	P	P	N	1.5	37.5
s5	P	P	Y	P	2.5	62.5
s6	Y	P	Y	P	3	75
s7	Y	P	Y	P	3	75
s8	P	P	Y	P	2.5	62.5
s9	P	P	P	N	1.5	37.5
s10	N	N	N	N	0	0
s11	P	P	P	N	1.5	37.5
s12	P	P	P	N	1.5	37.5
s13	P	P	Y	P	2.5	62.5
s14	P	P	P	N	1.5	37.5
s15	P	P	P	N	1.5	37.5
s16	P	P	P	N	1.5	37.5
s17	P	P	P	N	1.5	37.5
s18	N	N	N	N	0	0
s19	Y	P	Y	P	3	75
s20	P	P	P	N	1.5	37.5
s21	P	P	Y	P	2.5	62.5
s22	P	P	P	N	1.5	37.5
s23	P	P	P	N	1.5	37.5
s24	N	N	N	N	0	0
s25	P	P	P	N	1.5	37.5
s26	P	P	P	N	1.5	37.5
s27	N	N	N	N	0	0
s28	P	P	Y	P	2.5	62.5
s29	P	P	P	N	1.5	37.5
s30	P	P	P	N	1.5	37.5
s31	P	P	P	N	1.5	37.5
s32	Y	P	Y	P	3	75
s33	P	P	P	N	1.5	37.5
s34	P	P	P	N	1.5	37.5
s35	P	P	Y	P	2.5	62.5
s36	P	P	P	N	1.5	37.5
s37	P	P	P	N	1.5	37.5
s38	Y	P	Y	P	3	75
s39	P	P	P	N	1.5	37.5
s40	P	P	Y	P	2.5	62.5
s41	P	P	P	N	1.5	37.5
s42	P	P	P	N	1.5	37.5
s43	P	P	P	N	1.5	37.5
s44	P	P	Y	P	2.5	62.5
s45	Y	P	Y	P	3	75
s46	P	P	P	N	1.5	37.5
s47	Y	P	Y	P	3	75
s48	P	P	P	N	1.5	37.5
s49	Y	P	Y	P	3	75
s50	Y	P	Y	P	3	75
s51	P	P	Y	P	2.5	62.5
s52	P	P	P	N	1.5	37.5
s53	Y	P	Y	P	3	75
Total	29.5	24.5	35.0	10.5	99.5	
%Total score	29.7	24.6	35.1	10.6	100.0	
% By max QA	55.7	46.2	66.0	19.8		

4. Discussions

This section provides discussions about this SLR. The discussion is about the RQ mentioned above in Section 2.1.2. The types of malware analytics are as follows:

4.1. Descriptive Analytics

Descriptive analytics is a preliminary stage of data processing that creates a summary of historical data. According to studies in S1, S32 and S44, they must have identified all possible attacks before they need to go through all the security metrics. While for S16, they should have identified the hackers and their specialities. A large number of literatures have explored the use of topic modelling as a technique that can assist experts to analyse malware applications based on their characteristics such as studies in S20, S47, S49 and S53. Data is used to generate the behavioural profile by describing the sensitive application behaviour, S6, S13, S20, S34 and S45. Three studies which are S7, S24 and S51 have pointed out that the use of real name on profile in social media also can attract the attackers.

4.2. Diagnostic Analytics

Diagnostic analytics is a form of advanced analytics which examines data or content to answer the question “Why did it happen?”. Surprisingly, most of the studies that have this Type (2) analytics also have the Type (1) analytics. Many authors have reported that various malicious types will exhibit the same behavior by using similar API calls, S6, S13, S20, S38 and S47. As mentioned in S5, their investigation shows that the attackers choose to attack specific client, so that he can aim at fixed cloud provider. S28 and S45 described malware analysis and attribution using Genetic Information. Study by S35 examined that the attackers usually choose the most popular application because they have higher number of malware-free software.

4.3. Predictive Analytics

Predictive analytics is the branch of the advanced analytics which is used to make predictions about unknown future events. S2 used randomization technique to increase entropy of the system and thwart various attack. On the other hand, S6, S13 and S43 conduct the behavior analysis using Machine Learning (ML) techniques for example SVM, Naïve Bayes and Decision Tree. S7 and S11 have classified users into the vulnerability level using Fuzzy Logic. In addition, S8 rely on Random Forests classifier to perform the classification of the malware types. Studies conducted by S19, S26, S47, S49 and S53 mentioned that ML techniques used for large scale monitoring for malware activities. Meanwhile, S17 said that ML technique has been applied to distinguish spam account from normal one using Bayesian Classification Algorithm. Regarding S32, they have proposed a methodology for predicting attacks rates in the presence of extreme values via two approaches: Time Series Theory and Extreme Value Theory. The sentiment analysis also can be identified by Deep Learning (DL), S23 and S50. Author of S45 has proposed a ContrastMiner to distinguish fraudulent from genuine behavior. On the other hand, Data Mining technique is one of the methods used in S46 to reduce number of attribute on some log dataset to classify the traffic log either normal or suspicious. In the same vein, S5 also used Data Mining to categorized user data, split data into chunks to the proper cloud provider. According to S30 work, they have proposed multilayer hybrid strategy for zero-day filtering and phishing emails.

4.4. Prescriptive Analytics

Prescriptive analytics is the area of business analytics dedicated to find the best course of action for a given situation. S34 has highlighted issues facing cybersecurity domain in big data environment. Then, S40 and S42 also suggested a few solutions to assist the cybersecurity community for better enumeration malicious software. Author in S52 has proposed MinDroid which offers good performance in terms of detection time and execution cost. While S36 proposed Strider Search Ranger which implemented for anti-spam. Research by S53 investigated the accuracy of various machine learning models in the context of known and unknown apps, benign and normal apps so that can possible to detect malicious apps. In S37, they identified threats in Healthcare Information Systems (HIS) and implemented effective security systems and policies in healthcare setting.

4.5. Visual Analytics

Visual analytics is an outgrowth of the fields of information visualization and scientific visualization that focuses on analytical reasoning facilitated by interactive visual interfaces. S9 has visualized the detected compromises, threat and attack using D3 Javascript Library. By using Virtual Worlds, S14 and S48 visualized the network traffic. An interesting study conducted by S15, they enhanced the Neural Network to implement the visualization scheme. Based on S22 and S41, they have visualized the result of captured data using Honeyd-Viz which can be used to visualized logs. S39 said that the important of visualization technique is to allow analyst to quickly assessed and interpreted the generated alerts. Hence, S40 presented their visualization techniques for an interactive hex editor and visualizing packed malware. In contrast, S25 visualized the suspicious email messages based on the information provided using Mapping IP Addresses (MIPA) tool.

5. Conclusions

The goal of the paper is to conduct an SLR on the types of malware analytics techniques. Our aims are to identify the type and purpose of the malware analytics techniques. Our results exposed that there are five types of malware analytics techniques which need to be considered. This study will serve as a base for future studies and we would like to stress an idea that has been presented throughout this study. The results of this study indicate that field of malware analytics techniques is currently an active research area as pointed out in Figure 3. Figure 3 reveals that there has been a gradual increase in the number of study in the malware analytics since 2007 up to now. Type of malware analytics include Descriptive Analytics, Diagnostic Analytics, Predictive Analytics, Prescriptive Analytics and Visual Analytics have been discussed in Section 4. Although ML techniques have been widely used to detect, predict and visualize malicious in cybersecurity domain, the accuracy of ML models to support malware analytics still needs much room for improvement.

APPENDIX: Primary study review (Additional References)

#	Title
S1	A framework for automating security analysis of the internet of things.
S2	A secure architecture design based on application isolation, code minimization and randomization.
S3	Agnostic topology-based spam avoidance in large-scale web crawls.
S4	An approach for detection and family classification of malware based on behavioral analysis
S5	An Approach to Protect the Privacy of Cloud Data from Data Mining Based Attacks
S6	An effective behavior-based Android malware detection system.
S7	An expert system to detect privacy's vulnerability of social networks
S8	Analysis of Malware behavior: Type classification using machine learning
S9	Analysis of Techniques for Visualizing Security Risks and Threats.
S10	Andro-Dumpsys: Anti-malware system based on the similarity of malware creator and malware centric information
S11	Click Trajectories: End-to-End Analysis of the Spam Value Chain
S12	Cloud-based Android botnet malware detection system.
S13	Clustering analysis of malware behavior using Self Organizing Map.
S14	Cognitive cyber situational awareness using virtual worlds
S15	Cyber Incident Response Aided by Neural Networks and Visual Analytics
S16	Descriptive Analytics: Examining Expert Hackers in Web Forums
S17	Don't follow me: Spam detection in Twitter
S18	Dynamic Protection for Critical Health Care Systems Using Cisco CWS: Unleashing the Power of Big Data Analytics.
S19	Enhanced telemetry for encrypted threat analytics
S20	Exploring the Usage of Topic Modeling for Android Malware Static Analysis
S21	Hiding in plain sight: Characterizing and detecting malicious Facebook pages.
S22	Honeypots deployment for the analysis and visualization of malware activity and malicious connections
S23	Identifying Top Sellers In Underground Economy Using Deep Learning-Based Sentiment Analysis.
S24	Intrinsically Secure Next-Generation Networks.
S25	IP geolocation suspicious email messages.
S26	Large-Scale Monitoring for Cyber Attacks by Using Cluster Information on Darknet Traffic Features
S27	MAGMA network behavior classifier for malware traffic

#	Title
S28	Malware Analysis and attribution using Genetic Information
S29	MARFCAT: Fast code analysis for defects and vulnerabilities.
S30	Multilayer hybrid strategy for phishing email zero-day filtering.
S31	PhishAri: Automatic realtime phishing detection on twitter
S32	Predicting Cyber Attack Rates With Extreme Values."
S33	Predictive defense against evolving adversaries
S34	Security Analytics: Big Data Analytics for cybersecurity: A review of trends, techniques and tools
S35	Spotting the Malicious Moment: Characterizing Malware Behavior Using Dynamic Features.
S36	Strider Search Ranger: Towards an Autonomic Anti-Spam Search Engine
S37	Threats to Health Information Security
S38	Towards a Big Data Architecture for Facilitating Cyber Threat Intelligence.
S39	Uncovering periodic network signals of cyber attacks.
S40	Visualization techniques for efficient malware detection
S41	Situational Assessment of Intrusion Alerts: A Multi Attack Scenario Evaluation
S42	Information Management and Sharing for National Cyber Situational Awareness
S43	Machine Learning for the Detection of Spam in Twitter Networks
S44	Determining Risks from Advanced Multi-step Attacks to Critical Information Infrastructures
S45	Effective detection of sophisticated online banking fraud on extremely imbalanced data
S46	Using Data Mining Techniques for Diagnostic of Virtual Systems Under Control of KVM
S47	Comprehensive Behavior Profiling for Proactive Android Malware Detection
S48	MVSec: multi-perspective and deductive visual analytics on heterogeneous network security data
S49	Performance Evaluation of a Natural Language Processing Approach Applied in White Collar Crime Investigation
S50	Improved lexicon-based sentiment analysis for social media analytics
S51	A Model for Identifying Misinformation in Online Social Networks
S52	Preventive Policy Enforcement with Minimum User Intervention Against SMS Malware in Android Devices
S53	Network-based detection of Android malicious apps

Acknowledgements

This work is supported by Integrated Cyber Evidence (ICE) DISTIN Flagship Project under the Ministry of Science, Technology and Innovation (MOSTI) and Cyber Security Malaysia (CSM).

References

[1] B. Thuraisingham, L. Khan, M.M. Masud and K. W. Hamlen, "Data Mining for Security Applications", 2008 IEEE/IFIP International Conference on Embedded and Ubiquitous Computing.
[2] R. M. Yadav and R. K. Bhagel, "Web based Malware Detection using Important Supervised Learning Techniques on Online Web Traffic", International Journal of Computer Applications, vol. 130, no. 17, **(2015)**.
[3] A. Cuzzocrea, I. Y. Song and K. C. Davis, "Analytics over large-scale multidimensional data: the big data revolution!", In Proceedings of the ACM 14th international workshop on Data Warehousing and OLAP, ACM, **(2011)**, pp.101-104.
[4] Ikanow Editorial, "Cyber Threat Analytics versus Threat Intelligence", Retrieved from http://www.ikanow.com/cyber-threat-analytics-versus-threat-intelligence/, **(2015)**.
[5] K. Rieck, T. Holz, C. Willems, P. Düssel and P. Laskov, "Learning and classification of malware behavior", In International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment Springer, Berlin, Heidelberg, **(2008)**, pp. 108-125.
[6] M. N. Yusoff and A. Jantan, "A framework for optimizing malware classification by using genetic algorithm", Software Engineering and Computer Systems, **(2011)**, pp. 58-72.
[7] B. Kolosnjaji, A. Zarras, G. D. Webster and C. Eckert, "Deep Learning for Classification of Malware System Call Sequences", In Australasian Conference on Artificial Intelligence, **(2016)**, pp. 137-149.
[8] B. Kitchenham and S. Charters, "Guidelines for performing systematic literature reviews in software engineering", URL http://www.dur.ac.uk/ebse/resources/Systematic-reviews-5-8. Pdf, **(2007)**.
[9] N. F. Hordri, A. Samar, S. S. Yuhaniz and S. M. Shamsuddin, "A Systematic Literature Review on Features of Deep Learning in Big Data Analytics", International Journal of Advances in Soft Computing & Its Applications, vol. 9, no. 1, **(2017)**.

