

# Efficient STAKCERT KDD Processes in Worm Detection

Madiah Mohd Saudi, Andrea J Cullen and Mike E Woodward

**Abstract**—This paper presents a new STAKCERT KDD processes for worm detection. The enhancement introduced in the data-preprocessing resulted in the formation of a new STAKCERT model for worm detection. In this paper we explained in detail how all the processes involved in the STAKCERT KDD processes are applied within the STAKCERT model for worm detection. Based on the experiment conducted, the STAKCERT model yielded a 98.13% accuracy rate for worm detection by integrating the STAKCERT KDD processes.

**Keywords**—data mining, incident response, KDD processes, security metrics and worm detection.

## I. INTRODUCTION

NOWADAYS, data loss due to worms attack is not rare and a good sound of knowledge in confronting the worms is seen as one of the best solutions to confront the worms attack. On November 2008, Swabey [1] wrote that after the attack of AgentBTZ worm, US Department of Defense had banned the use of the USB unless after it has been scanned and free from any worms. Indeed in the same month of the year, IT system in three hospitals in London went down due to Mytob worm attack which caused the administrative process to be done manually [2]. Furthermore in January 2009, 9 million PCs were infected by Downadup worm or also known as Conficker worm [3]. Indeed, it is not surprising to know that in year 2010, millions of computers all over the world were infected by different type of worms such as the Mariposa, Zeus, Bredolab, TDSS, Koobface, Sinowall and Black Energy 2.0 botnets [4].

Therefore, we need to have a good strategy and technique to identify these worms. The objective of this paper is to introduce new STAKCERT KDD processes that leads to the formation of STAKCERT model for worm detection. STAKCERT stands for starter kit for computer emergency response and KDD stands for knowledge discover databases. Our goal is to achieve the accuracy rate for worm detection more than 91.9%, which was the result, gained by Dai and colleagues [5]. Based on our analysis and experiment, we have identified the integration of STAKCERT KDD processes, which applied the standard operating procedures for worm

incident response, improved KDD processes and data mining technique and resulted the formation of the STAKCERT model for worm detection. It has shown a promising result with 98.13% accuracy rate of worm detection.

The rest of this paper is organised as follows. Next section describes a discussion on previous works on KDD processes, data mining and incident response. Section III explains the research design used and the results are presented in section IV. Section V concludes and discusses the future work for this paper.

## II. RELATED WORKS

Incident response is defined as the process that aims to minimise the damage caused by security incidents and malfunctions; it also monitors and learns from such incidents [22]. The lack of standard operating procedures, in terms of analysing and responding to a worm attack may lead to disaster for both IT personnel and the end user. It is very hard to separate incident response from computer security, as it plays a very important role within such security. Improvements and novel standard operating procedures, particularly within the detection, analysis and disinfection phases, are seen as areas for potential research and exploration. An example of research that proposed a generic incident response process within a corporate environment was that undertaken by [23]. However, research alluding to a combination of worm handling procedures following incidence response has, so far, been scarce. It is suggested here that such research could greatly improve matters by detailing the required procedures for handling a worm incident. This is one of the precepts of the formation of the STAKCERT model for worm detection, of which incident response is a part.

The phrase KDD was first discussed in a KDD workshop in 1989 [7] and ever since the KDD has been successfully applied in different domains all over world. Knowledge discovery in databases (KDD) is defined as an overall process where knowledge or patterns from data are extracted, where the patterns extracted must be valid, novel, useful and understandable. Data mining on the other hand is a specific algorithm to extract the pattern from the data, which is a part of the whole KDD process [8], [9]. Many studies that integrate KDD have been conducted over the past few years and the current research in the year 2010 include [10] in medicine, [11] in financial applications, [12] in intrusion detection and [13] in customer relationship management (CRM).

For this research, KDD is used as a technique to identify the worms' patterns in the dataset. All of the KDD processes are summarised in Fig. 1. The data-preprocessing function is intended to transform the worm's raw data into an appropriate format for the next stage of the analysis, which is data extraction. The steps involved in this phase include feature

Madiah Mohd Saudi is with the School of Computing, Informatics and Media, University of Bradford, UK and attached with Universiti Sains Islam Malaysia (USIM), Malaysia. (email:mbmohdsaudi@bradford.ac.uk /madiah@usim.edu.my).

Dr Andrea J Cullen is a senior lecturer with the School of Computing, Informatics and Media, University of Bradford, UK (e-mail: A.J.Cullen@bradford.ac.uk).

Professor Dr. Mike E Woodward is a professor with the School of Computing, Informatics and Media, University of Bradford, UK (e-mail: M.E.Woodward@bradford.ac.uk).

selection, data cleansing to remove any noise, duplication or outlier and data transformation. The data pattern extraction is achieved by using data mining. The clustering and classification are two of the most common techniques used in data mining. The type of algorithm implemented for our research is *k-means* for clustering and *SMO* is for the worm classification. *SMO* stands for sequential minimal optimisation algorithm for support vector machine classification. Once the patterns are extracted from the data, they will be interpreted to ensure only valid and useful information or knowledge is kept for further exploration. All the KDD processes are iterative to ensure the result achieved is rigorous. Fig. 1 displays common KDD processes involved in developing knowledge.

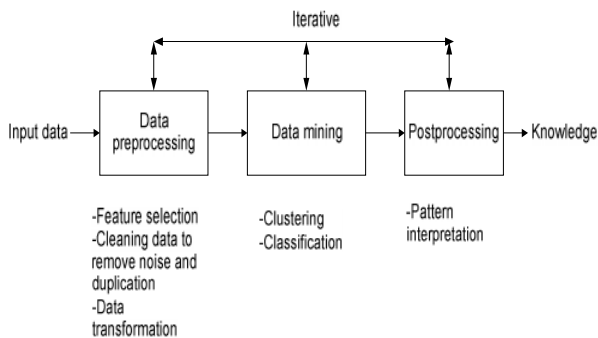


Fig. 1 KDD processes

In order to test the effectiveness of the STAKCERT model for worm detection, a comparison of the work with research conducted by [5] was undertaken. This research used the same datasets as in our research and had the same objective; i.e., to detect worms and increase the worm detection rate. They incorporated dynamic instruction sequence mining techniques involving the runtime features of a worm programme. This research is the most similar to ours and we have bridged the gaps that arose from the aforementioned research by integrating STAKCERT KDD processes within the STAKCERT model.

### III. RESEARCH DESIGN

The datasets in this research consist of different types of worms sourced from VX Heavens [6]. From 66,711 samples downloaded from VX Heavens, 5,614 were identified as worms. Then it was further categorised as email worm with 3.97%, followed by 1.36% for P2P worm, 0.96% represent the IRC worm, 0.81% for the internet worm, 0.42% for the instant messaging worm and 0.86% for worm. The datasets were chosen randomly from all of these worm categories, and 160 variants of the windows worm and benign executables have been used for this research since it is the scope for this research.

The lab used for this testing is illustrated in Fig. 2. It is a controlled lab environment and almost 80 % of the software used in this testing is an open source or available on a free basis. No outgoing network connection is allowed for this architecture. In this lab, the datasets described above were

tested. From these tests, the results can easily be analysed and any flaws found can be fixed immediately.

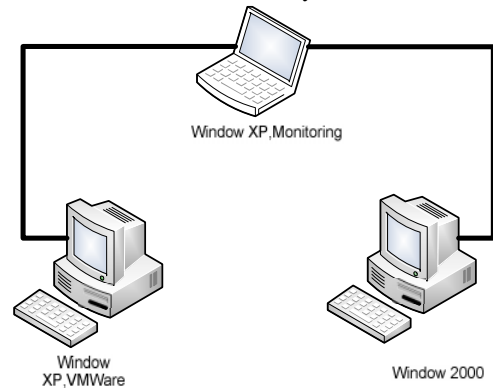


Fig.2 Lab architecture

### IV. EXPERIMENTAL RESULTS

Enhancements have been made to the KDD data-preprocessing and pattern extraction process. Under the data-preprocessing process, the static and dynamic analyses are implemented using the incident response standard operating procedures (SOP). While under the pattern extraction process, statistical methods comprising Chi-square and symmetric measure and security metrics are also introduced, as illustrated in Fig. 3. The details for the Chi-Square and symmetric measure can be referred in [14], which are not discussed in this paper. We only explain in detail of the improvement in the STAKCERT KDD processes and the security metrics in this paper.

#### A. Data-preprocessing

In this research, the datasets from the VX Heavens source were retrieved in multiple formats. In order to use this data, it needs to be transformed into an understandable format; hence the need for feature selection using static and dynamic analysis. It should be remembered that feature selection is a search strategy process where only relevant data is chosen with the goal that the selected data can be valid and useful for the subsequent analysis. In our case, the chosen data was analysed using static and dynamic analysis in a controlled lab environment. The details for static and dynamic analysis can be referred in [15] but how the incident response being integrated with these analyses is explained in this paper.

Before and during the conduction of the static and dynamic analysis, the incident response approach was applied. Standard operating procedures before and during the analysis must be followed and all the related procedures documented. Initially, all the listed software used during the testing were checked to ensure all were already installed and working properly. Secondly, the condition of the testing machines and the network setting for each machine were checked. Thirdly, documentation of all the monitoring and test results were ensured. This was to make certain that there is always documentation if anything needs to be referred to later. With reference to the incident response methodology by [16], as

illustrated in Fig.4, in order to reach any solution which includes recovery steps or to implement security measures, all these seven steps play an important role. However, according to the SANS Institute, six steps are required to handle any incident effectively, namely: preparation, identification, containment, eradication, recovery, and lessons learned [17]. Indeed, MyCERT used the SANS steps to produce the computer worm incident handling standard operating procedures (MyCERT 2002[20] – see Fig.5). Therefore, in the STAKCERT KDD processes, the incident response methodology by [16] and [17], together with the MyCERT SOP for worm handling, are used as a basis and a guide.

How are the incident response methodology and MyCERT SOP in worm handling reflected in our methodology? It is

integrated in ‘*research design*’ and ‘*data-preprocessing*’ accordingly. These are where incident response is integrated. Before data analysis starts, preparation is carried out by examining the checklist to ensure all the installed software is working properly with the right testing lab network settings. Furthermore, all the analyses and findings are documented for all the experiments.

As a whole, in STAKCERT KDD processes, the incident response is already integrated in worm detection. It is hard to separate the incident response since it plays an important role in the security field, especially in responding to the worm incident.

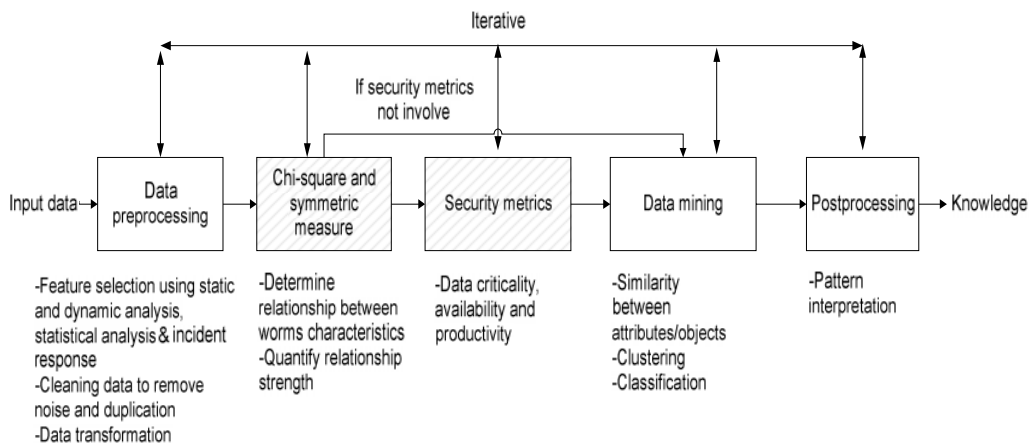


Fig. 3 STAKCERT KDD processes

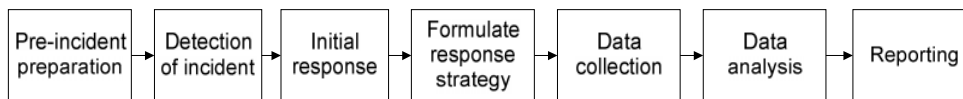


Fig. 4 Incident response methodology

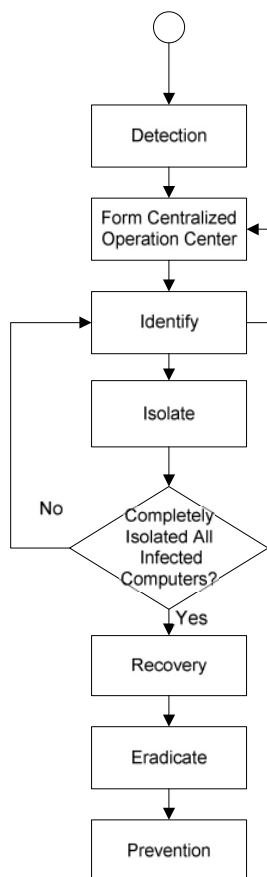


Fig. 5. MyCERT worm IH SOP  
Adapted from MyCERT MA-041.052002:  
computer worm incident handling standard operating procedure  
2002

**B. Security Metrics**

For the security metrics which is integrated in the STAKCERT KDD processes, it helps to quantify, classify and measure information on security operations which leads to the formation of a new STAKCERT rules for worm response. This will be one of our future works, once the STAKCERT model for worm detection is completed. In security metrics, the threats are firstly defined, then the threats are transformed into metrics or representations that can easily be measured. We then seek to understand and identify the vulnerabilities, flaws, problems, weaknesses or damage they can cause to the security infrastructure. Next is to check the existing countermeasure process performance and, if necessary, recommend the improvement of any technology or countermeasure process [18].

The security metrics processes are already being applied in STAKCERT research for worm detection and simplified in Table I. For STAKCERT research, in order to understand the threat posed by a worm, a deep and thorough understanding of worm architecture is necessary; in our case, this led to the formation of STAKCERT worm classification and the STAKCERT worm relational model [24]. Initially, the characteristics that need to be observed are defined. Then,

during the static and dynamic analysis, the worms are analysed and simplified into worm representation, which comprises payload, activation, operating algorithm, infection and propagation. A thorough analysis related to the vulnerabilities, flaws, problems, weaknesses or the damage the worm can cause to the security infrastructure is closely monitored. As a result, weight and severity are chosen as two main attributes in assigning the countermeasure process. Apart from the elements stated in Table I, security metrics can also be measured based on the perimeter defence, control and coverage, availability and reliability and application risks. All these measurements were already taken into consideration when we conducted the worm analysis.

Finally, the main reason why security metrics have been chosen in this research is due to its capability to make the job of defining, understanding, identifying and measuring information security efficient, accurate, measurable and reliable. This is also supported by [19], where they stated that work can be more profitable if it is enhanced using the security metrics and is more efficient if it is measurable.

TABLE I  
SECURITY METRICS IN STAKCERT PROCESSES

Security metrics processes	Applying security metrics from STAKCERT
1) Define worm threats	Yes
2) Represents worm threats into metrics	Yes. -Worm data is represented based on payload, infection, activation, propagation and operating algorithm. -Formation of the STAKCERT worm classification and STAKCERT relational model.
3) Understand and identify the vulnerability, flaw, problem, weakness and damage to security infrastructure	Yes. -Run the static and dynamic analysis. -Identify the need to assign weight and severity value to assign the countermeasure process.
4) Check the performance of the existing countermeasures	Yes. -Integrate and run data mining using JAVA-Weka to check the accuracy rate of weight and severity assigned.
5) Recommend any technology or countermeasure process for improvement.	Yes. -Apoptosis to isolate the most severe worm attacks.

**C. Data mining**

The improvement that has been made in STAKCERT KDD processes, resulted the formation of a new STAKCERT model for worm detection. By integrating the SMO, the STAKCERT model has been simulated using the JAVA-WEKA software. It is an open source JAVA based software and it has a collection of machine learning algorithms to solve data mining problems [21]. Based on the testing conducted, it outperforms the existing work by [5] with 98.13% of our overall accuracy

which is 6.23% higher than theirs. The STAKCERT true positive rate was also higher than in the comparable works and the false negative rate was lower. Furthermore, our false positive rate was also lower. Table II summarises the results. Overall accuracy represents the summation of true positive and true negative, divided by the summation of true positive, true negative, false positive and false negative. TP stands for true positive, TN stands for true negative, FP stands for false positive and FN stands for false negative. The equations used in Table II can be referred in equation 1, 2, and 3.

$$OA = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (1)$$

$$TPR = \frac{TP}{(TP + FN)} \quad (2)$$

$$FPR = \frac{FP}{(FP + TN)} \quad (3)$$

## V. CONCLUSION

TABLE II  
EXPERIMENT RESULTS

Classifier	STAKCERT Result (%)			Comparison work (%)		
	TPR	OA	FPR	TPR	OA	FPR
<i>SMO</i>	98.1	98.13	0.2	93.2	91.9	9.6

TPR=true positive rate, OA= overall accuracy, FPR= false positive rate

We believe that our work offers its own significant contribution towards computer security, KDD processes and data mining, where the novelty of our experiment lies in the method being implemented and the goal achieved by the end of our research. Such implemented methods are an integration of KDD processes, data mining, statistical and incident response techniques. Consequently, our results yielded a better performance than the comparable, existing work. For future work, these results will be used as the input in triggering the worm response where the apoptosis will be integrated.

## REFERENCES

- [1] Swabey,P., "US Department of Defense bans USB drives after worm attack", 20<sup>th</sup> November 2008, Source: Information Age Today, Available from: <http://www.information-age.com/home/information-age-today/814827/us-department-of-defense-bans-usb-drives-after-worm-attack.shtml> [Accessed: 31<sup>st</sup> March 2011].
- [2] Swabey,P. , "Virus takes down three hospitals", 19<sup>th</sup> November 2008, Source: Information Age Today, Available from: <http://www.information-age.com/home/information-age-today/814312/virus-takes-down-three-hospitals-it-systems.shtml> [Accessed: 31<sup>st</sup> March 2011].
- [3] Keizer, G. , "Amazing' worm attack infects 9 million PCs", 19<sup>th</sup> January 2009, Source: Computerworld Security, Available from: [http://www.computerworld.com/s/article/9126205/\\_Amazing\\_worm\\_ack\\_infects\\_9\\_million\\_PCs](http://www.computerworld.com/s/article/9126205/_Amazing_worm_ack_infects_9_million_PCs)[Accessed: 31<sup>st</sup> March 2011].
- [4] Alexander Gostev,"Malware Evolution 2010", Kaspersky Security Bulletin.17 Feb 2011,URL: [http://www.securelist.com/en/analysis/204792161/Kaspersky\\_Security\\_Bulletin\\_Malware\\_Evolution\\_2010#22](http://www.securelist.com/en/analysis/204792161/Kaspersky_Security_Bulletin_Malware_Evolution_2010#22) [Accessed: 31<sup>st</sup> March 2011].
- [5] Dai,J., Guha,R. and Lee,J., "Efficient Virus Detection Using Dynamic Instruction Sequences", *Journal of Computers*, Vol 4, No 5, 2009, pp. 405-414.
- [6] VXHeavens website, "Virus Collection", 2009, Available: <http://vx.netlux.org/vl.php>. [Accessed: 31<sup>st</sup> March 2011].
- [7] Piatetsky-Shapiro, G., "Knowledge Discovery in Real Databases: A Report on the IJCAI-89 Workshop". *AI Magazine* 11(5), 199, pp.68-70.
- [8] Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P., "The KDD Process for Extracting Useful Knowledge", *Volumes of Data. Communications of the ACM*, v. 39(no. 11), 1996, pp. 27-34.
- [9] Maimon,O. and Rokach,L., "Introduction to Knowledge Discovery and Data Mining", In: Maimon, Oded; Rokach, Lior ,eds. *Data mining and knowledge discovery*. 2<sup>nd</sup> edn. New York:Springer, 2010, pp 1-15.
- [10] Lavrac,N. and Zupan,B. " Data Mining in Medicine", In: Maimon, Oded; Rokach, Lior ,eds. *Data mining and knowledge discovery*. 2<sup>nd</sup> edn. New York:Springer, 2010, pp. 1111-1136.
- [11] Kovalerchuk,B. and Vityaev,E., "Data Mining for Financial Applications" , In: Maimon, Oded; Rokach, Lior ,eds. *Data mining and knowledge discovery*. 2<sup>nd</sup> edn. New York:Springer, 2010,pp. 1154-1169
- [12] Singhal,A. and Jajodia,S., "Data Mining for Intrusion Detection", In: Maimon, Oded; Rokach, Lior ,eds. *Data mining and knowledge discovery*. 2nd edn. New York:Springer, 2010, pp.1171-1180.
- [13] Thearling,K., "Data Mining for CRM", In: Maimon, Oded; Rokach, Lior ,eds. *Data mining and knowledge discovery*. 2nd edn. New York:Springer, 2010, pp.1181-1188
- [14] Saudi,M.M, Cullen, A.J. and Woodward, M.E., "Statistical Analysis in Evaluating STAKCERT Infection, Activation and Payload Methods", *Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering 2010, WCE 2010, 30 June - 2 July, 2010, London, U.K.*pp 474-479.
- [15] Saudi,M. M., M.Tamil, E., Cullen,A.J., Woodward, M., IIdris,M.Y., Reverse Engineering: EDOWA Worm Analysis and Classification. In: Ao,S.I.& Gelman,L.,eds. *Advances in Electrical Engineering and Computational Science, Lecture Notes in Electrical Engineering*. Berlin: Springer Netherlands, April 2009, pp. 277-288.
- [16] Prosis, C., Mandia,K. and Pepe,M..Incident Response and Computer Forensics, Second Edition, McGraw-Hill, 2003, p15.
- [17] SANS Institute. "Security 504.1 Incident Handling Step-by-Step and Computer Crime Investigation". SANS Institute,2008..
- [18] Jaquith, A., "Security metrics: replacing fear, uncertainty and doubt". United States of America: Addison-Wesley. 2007, p40.
- [19] Atzeni,A. and Lioy,A., "Why to adopt a security metric? A brief survey". In: Gollmann,D., Massacci,F. and Yautsiukhin,A. (eds.), *Quality of ProtectionSecurity Measurements and Metrics*, USA: Springer, 2006, pp1-12.
- [20] MyCERT website, "Computer Worm Incident Handling Standard Operating Procedure",2002, URL: <http://www.mycert.org.my/en/services/advisories/mycert/2002/main/detail/111/index.html> [Accessed: 31<sup>st</sup> March 2011].
- [21] Hall,M., Frank,E., Holmes,G., Pfahringer,B., Reutemann,P. and Witten,I.H., "The WEKA Data Mining Software: An Update; SIGKDD Explorations", Volume 11, Issue 1,2009.
- [22] BSI. "Information security management, BS7799, part 1: code of practice for information security management", 1999.
- [23] Mitropoulos, S., Patsos, D. & Douligeris, C. "On Incident Handling and Response: A state-of-the-art approach", *Computers & Security*, Volume 25, 2006, pp.351-370. [Accessed: 31<sup>st</sup> March 2011].
- [24] Saudi, M.M, Cullen, A.J. and Woodward, M.E., STAKCERT Worm Relational Model for Worm Detection, *Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering 2010, WCE 2010, 30 June - 2 July, 2010, London, U.K.*pp 469-473