An Approach for Estimating Forensic Data Provenance of an Object in the Cloud Environment using One Dimensional Successive Bisection Method

Victor R. Kebande
ICSA Research Lab
Department of Computer Science
University of Pretoria
Private Bag, X20, Hatfield, Pretoria, South Africa

Nickson M. Karie
ICSA Research Lab
Department of Computer Science
University of Pretoria,
Private bag, X20, Hatfield, Pretoria, South Africa

ABSTRACT
The decline in traditional forensics and the emergence of cloud forensics has made it hard for digital forensic investigators to conduct digital investigations due to inadaptability of the cloud to digital forensic processes. Additionally, data that resides in the cloud is normally scattered across constantly changing data centers, hosts and virtual instances are easily destroyed as they are created. On the same note it is not easy to seize physical devices where a particular crime has occurred and locating the origin of an object in the cloud can be a challenge when we do not know where the actual data resides. Data provenance provides a solution whereby one can trace data based on the tag as it keeps changing directions in the cloud. However, if the distance between two particular tagged data is taken into account then the source and destination of an object can be located easily which can then enable digital forensic investigators to locate the physical devices. The problem that this paper is addressing is that, there is no easy way of locating data provenance in the cloud environment without estimating the distance between tagged data that moves in the cloud. Based on this, the authors have proposed a mechanism for locating the provenance of an object using successive bisection method. Similar test have been carried using different intervals in an experiment and the results are very promising.

General Terms
Digital Forensics, provenance, cloud.

Keywords
Forensic, data, provenance, object, cloud, bisection method.

1. INTRODUCTION
The nature of the cloud model presents data as the most important element in the cloud environment because a majority of relevant processes and activities are aimed at managing data. Notwithstanding that, a Cloud Service Provider (CSP) is usually tasked with trustworthiness and transparency on how different aspects of data activities are conducted including the provenance. Data provenance is a mechanism that allows the history or origin of data to be derived so that the physical resources that contain data can be located. According to Reddy and Seltzer [1] it is metadata that provides details of the origins or history of a data object" which describes people, entities and activities that were involved in producing a data object. Nevertheless, a lot of care needs to be taken when tracking data provenance because according to Saad, Jalil and Manaf [2], there can exist provenance integrity if there is false information or tampering by unauthorized party.

The cloud computing paradigm requires new procedures, methods and guidelines for conducting digital forensic investigation which means traditional forensic methods are unadoptable to the new cloud model. At the time of writing this paper, still there exist, no acceptable standards for conducting digital forensic investigations in the cloud environment. As a result of these, the cloud has inherited numerous vulnerabilities in the TCP/IP stack and in the web service protocols [3] and seizing the physical device becomes a big challenge because of constantly changing environment. Therefore, in order for data provenance to provide digital forensic investigators with the physical location we consider the scenario below.

[PXY], [UVW] and [MNO] are data centers across diverse locations which forms the cloud environment. {A, B}, {A,C,DE} and {F,G,H} are data objects from data centers [PXY], [UVW] and [MNO] holding tags D_{ab}, D_{acde} and D_{fgh} respectively. {A,B}, {A,C,DE} and {F,G,H} are randomly moving within the cloud environment and DF who is a digital investigator needs to get the physical location for {ACDE) and {AB} which seem to be compromised.

The contributions of this research have been presented as follows:

- Present an approach for estimating data provenance using one dimensional successive bisection.
- Conduct experiments to proof the concept.
- Provide contextual evaluation for the proposed concepts.

As for the remainder of this paper, Section 2 presents reviews on background. Section 3 presents related work. Thereafter, Section 4 presents a proposed approach for estimating data provenance using successive bisection. This is then followed by Section 5 that presents the experiments that acts as a proof of concept. After this, Section 6 presents a contextual evaluation for the proposed concept. Section 7 concludes the work and mentions a possible future work.

2. BACKGROUND
The section presents the background of the study on the following aspects: DF, cloud computing aspects and successive bisection method. The aforementioned parameters are considered because of the following reasons: DF is discussed to show how the science of investigation can help in excavation of digital evidence using scientific proven

methods. Cloud computing is discussed to show how it is a target for moving data. Successive bisection has been discussed to show how estimation can be done using data provenance.

2.1. Digital Forensics

In the 21st century Digital Forensics (DF) which is relatively a new area has emerged to be the fastest investigative field dealing with cyber-crime, and digital investigation processes. This involves the process of excavating potential digital evidence and using that evidence to create a hypothesis that can be presented in a court of law to try to prove occurrence of a security incident. The extracted potential digital evidence in this context should satisfy admissibility in a court of law based on the legal perspectives on digital evidence depending on a given jurisdiction. At the first Digital Forensic Research Workshop (DFRWS) held in Utica, New York in 2001, Palmer gave a definition for digital forensics as; "The use of scientifically derived and proven methods toward the preservation, collection, validation, identification, analysis, interpretation, documentation and presentation of digital evidence derived from digital sources for the purpose of facilitating or furthering the reconstruction of events found to be criminal, or helping to anticipate unauthorized actions shown to be disruptive to planned operations" [4]. Based on this definition it is evident that the methods used in this process must be acceptable scientifically.

2.2. Cloud Computing

The cloud presents an approach that is aimed at processing data and storing it across data centers. The popularity of the cloud is associated with its effectiveness and robustness [5]. However, the control of data, monitoring and analysis is trusted to CSPs and cloud service administrators. According to the National Institute of Standards and Technology, NIST, defines it as "a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction" [6]. Generally cloud computing is an extension of network architectures that are virtualized to provision resources. Inability to track data in the cloud has raised a number of challenges, for example, data provenance has turned out to be a very big concern because there exist no data tracking tools focused in the cloud at the time of writing this paper. Nevertheless the logging mechanisms that exist are mainly focused on operating in a system-centric perspective [7]. Research on methods of collecting digital evidence in the cloud and potential digital evidence have been proposed by Kebande & venter, where potential evidence can be digitally preserved for purposes of conducting digital investigation [17], [18], [19], [20],[21],[22].

2.3. One-Dimensional Successive Bisection

One-Dimensional successive Bisection which is a root finding method for non-linear equations is an iteration algorithm that is used to generate a number of iterations by choosing an origin state point which is treated as the fixed point. Thereafter an increment is usually given to the fixed point and then a numerical simulation is used to test whether the state point is stable [8],[9],[10]. Based on the traditional bisection method assumes that a continuous function *g(z)* which lies on the intervals [α, β] satisfies the condition $g(\alpha)\, g(\beta)<0$. Figure 1 show the steps involved in estimating and finding non-linear roots using bisection method. Note that the error tolerance for the root given the function *g(z)* is given as ε [11].

Iteration for *Bisection* {g, $\alpha, root, \beta, \varepsilon$ }

1. *Define y={*$\alpha + \beta)/2$ }
2. *If* $\beta - y \le \varepsilon, then$ *accept root=y and exit.*
3. *If g(*$\beta).g(y) \le 0, then \quad \alpha = y, otherwise$ $\beta = y$
4. *Go to step 1.*

Based on the iteration algorithm shown in Figure 1, bisection method convergence speed is given as exponential which means that anytime that the function g(z) has more than one roots the interval $[\alpha, \beta]$ is chosen to get one root in the interval[10][11]. Figure 1 shows the convergence of the graph during interval halving process of bisection.

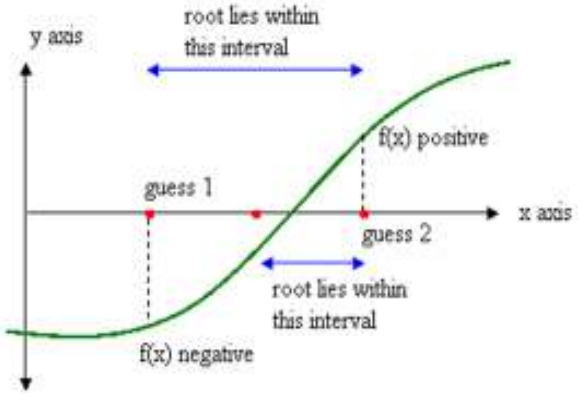

Fig.1: A graphical representation of convergence of bisection/interval halving method (source: [23])

3. RELATED WORK

Previous studies on data provenance in the cloud have emphasized the importance of tracking an object in the cloud in order to locate the physical location of a machine in order to enhance forensic investigation. Nevertheless, none has focused on the use of bisection method. The authors therefore present work that is somewhat used as related work.

A digital forensic model to provide better data provenance by Trenwith and Venter investigates how objects in the cloud can be used in order to provide a trace of the path through which data travels through in the cloud environment. In this research the authors highlights that if the metadata tags cannot be appended to the headers of all or any other file, then the metadata should be stored separately from the data objects [12].

An S2Logger end-to end data tracking mechanism for cloud provenance is a data-centric logger which is able to trace activities like file creation, transfers, duplication, deletion and editions across the cloud servers. The S2Logger was able to detect critical data-related cloud security problems ranging from privacy violations, analysis of data provenance, data leakage and malicious attacks [13].

A Progger or provenance logger which is an efficient logger from cloud data provenance tracking is a mechanism that allows security analysts to collect provenance from a lowest form possible. It also enables other tools to be built for effective end-to end tracking of data provenance. A part from that the Progger was able to address the limitation of other provenance by providing log tamper-evidence, prevention of fake or manual entries, an accurate timestamp synchronisation across several machines, efficient log space growth, and the accurate logging of root usage of the system [13].

A research paper by Kebande and Venter aimed at enhancing detection in the cloud proposed a method of adding event reconstruction to a cloud forensic readiness model. In this paper event reconstruction was proposed as a mechanism that could help detection of security events. This was based on a proposal of using a Non-malicious Botnet (NMB) that could collect digital forensic information [14], [15], [17-19], [21], [22]. Another mechanism that was used in this context to try identifying attacks in the cloud was detection of botnets attacks using Artificial Immune Systems (AIS) [16].

The authors highly acknowledge the aforementioned work that has been done by other researchers. This has not only given an insight on preliminary work but also simplified the approach that the author has to use while giving a contribution. In the next section the reader is introduced to the proposed concept of estimating forensic data provenance in the cloud environment using one dimensional bisection method based on the scenario highlighted.

4. AN APPROACH FOR ESTIMATING FORENSIC DATA PROVENANCE OF AN OBJECT IN THE CLOUD ENVIRONMENT

This section provides an approach for estimating data provenance in the cloud environment during digital forensic investigations as a contribution. The approach uses one-dimension bisection method to do estimation. The authors will constantly refer to the hypothetical scenario that was previously highlighted in Section 1. Based on the scenario the following represents the mapping of the data objects.

Table 1: Tagged Data objects from Data centers

	Data center	Data objects	Data Tag
1	[PXY]	{A,B}	D_{ab}
2	[UVW]	{A,C,D,E}	D_{acde}
3	[MNO]	{F,G,H}	D_{fgh}

A digital forensic investigator needs to get the physical location for {ACDE} and {AB}, the authors apply successive bisection method to identify where the physical location might lie based on the given two objects. Two points from the data objects {A,B},{A,C} and {D,E} with respective data object tag D_{ab} and D_{acde} then the authors let [PXY] and [UVW] be the value of the root. Based on this one-dimensional bisection will be represented using different approaches as follows:

$A \le [PXY] \le B$, $_{Dab}$ and $A \le [UVW] \le C$,$_{Dacde}$ and $D \le [UVW] \le E$, $_{Dacde}$----------(1)

The authors then let {A,B} {A,C} {D,E} be the values of AB, AC and DE on the n^{th} iteration of the bisection algorithm. Therefore the error bound for the respective objects is given as follows:

$$|PXY| \le 1/2^n (B-A) \text{----------(2)}$$

$$|UVW| \le 1/2^n (C-A) \text{----------(3)}$$

$$|UVW| \le 1/2^n (E-A) \text{----------(4)}$$

As a result of the aforementioned errors that is based on the data objects at the data centers, a number of iteration are needed in order to be able to locate the specific physical location of the data center. This can therefore be computed based on total accuracy which is given as follows:

$$n \ge \log_{\log 2} (B-A)/\varepsilon \text{ /----------(5)}$$

$$n \ge \log_{\log 2} (C-A)/\varepsilon \text{ /----------(6)}$$

$$n \ge \log_{\log 2} (E-D)/\varepsilon \text{ /----------(7)}$$

In the next section a series of conducted experiments have been presented which shows how data provenance can be estimated based on the data tag as objects move around the cloud environment.

5. CONDUCTED EXPERIMENTS

The authors conducted experiments to show the easiest way of estimating the provenance of an object using successive bisection method using f(x) (pow(x,3)-18). The following objects were considered in this case: {A,B},{A,C},{D,E},{F,G} and {F,H}. Additionally the following object tags were considered respectively based on the first and second objects: 001, 002,011, 022, 021, 023, 024, 025, 026 and 030 respectively. Random non-negative roots between 1 and 6 were taken for each object to test the estimation and convergence of the objects. The object value was taken by incrementing a single digit in the second value as shown in Table 2. The error bound remained constant across as 0.001. This is shown in Table 2.

Table 2. Data sets for conducted Experiments

Obj1	Obj2	Tag(Obj1)	Tag(Obj2)	Val_1	Val_2	Error
A	B	001	002	1	2	0.001
A	C	011	022	2	3	0.001
D	E	021	023	2	4	0.001
F	G	024	025	2	5	0.001
F	H	026	030	2	6	0.001

A total of 9,10 and 11 iterations were conducted respectively in the experiment and the results have shown very promising results when the roots of the experiments are shown and how they converge.

1.1. Experiment Obj1 {A} and Obj2 {B} with value 1 and 2 and error as 0.001

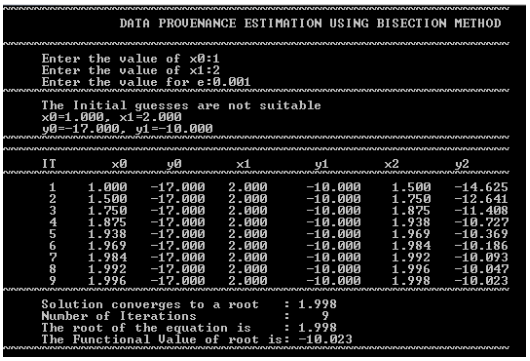

Fig. 2: Data provenance estimation with value 1 and 2

In this case Figure 2 shows the experiment that the initial guess provided is presented as not suitable while the solution converges to 1.998. The functional value is set at -10.023

1.2. Experiment 1 Obj1 {A} and Obj2 {C} with value 2 and 3 and error as 0.001

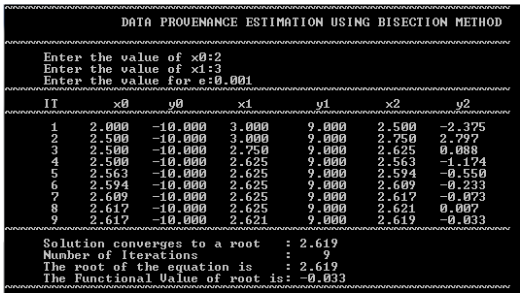

Fig. 3: Data provenance estimation with value 2 and 3

In Figure 3 a total of 9 iterations have been conducted in this experiment the solution converges to a root of 2.619. The functional values provide a negative value of -0.033. Initial guesses are not affected in any way.

1.3. Experiment 2 Obj1 {D} and Obj2 {E} with value 2 and 4 and error as 0.001

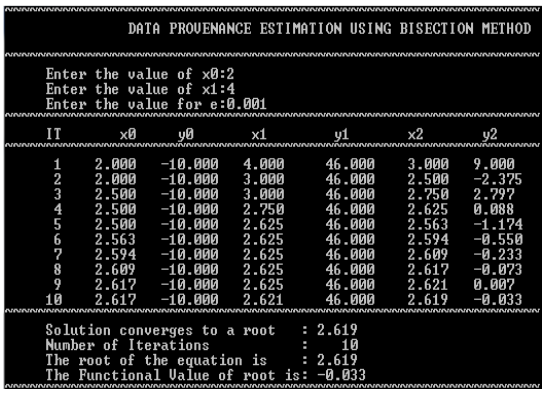

Fig. 4: Data provenance estimation with value 2 and 4

From Figure 4 the solution was tested with root 2 and 4 and the solution converges to 2.619 with 10 iterations. The functional value was set to -0.033.

1.4. Experiment 3 Obj1 {F} and Obj2 {G} with value 2 and 5 and error as 0.001

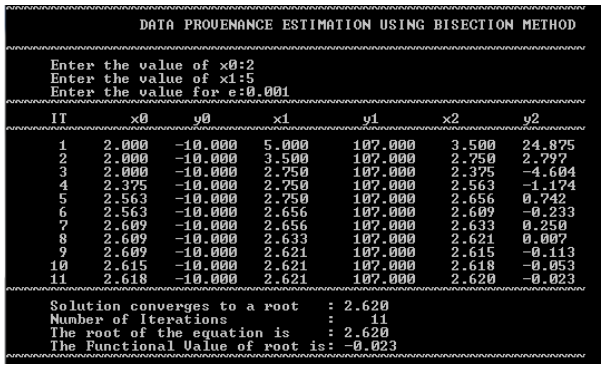

Fig. 5: Data provenance estimation with value 2and 5

In Figure 5, a total of 11 iterations have been conducted in this case where the solution converges to 2.620 and the functional value of root is set at -0.023.

1.5. Experiment Obj1 {F} and Obj2 {H} with value 2 and 6 and error as 0.001

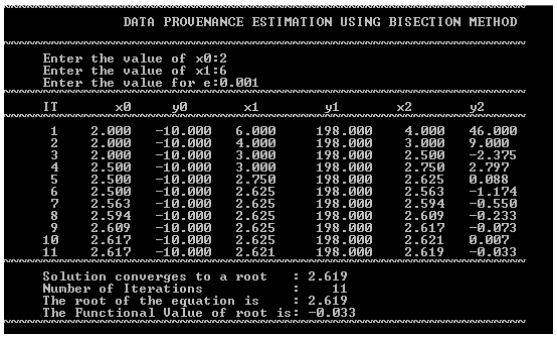

Fig. 6: Data provenance estimation with value 2and 6

In Figure 6, a total of 11 iterations have been conducted in this case and the solution converges to 2.619.The functional value is set at 0.033.

Based on experiments conducted using objects {A,B} and {A,C}, {D,E}, {F,G}, {F,H} the solution converges to different figures when [1,2], [2,3],[2,4],[2,5],[2,6] root interval are used. Table 3 shows the convergence and functional values of the experiments

Table 3. Table that shows convergence and Functional values of the roots

Obj1	Obj2	Iterations	Val 1	Val 2	Convergence	Functional Value
A	B	9	1	2	1.998	-10.023
A	C	9	2	3	2.619	-0.033
D	E	10	2	4	2.619	-0.033
F	G	11	2	5	2.620	-0.023
F	H	11	2	6	2.619	-0.033

From Table 3 a similar convergence can be seen in the [A,C], [D,E] and [F,H] which depicts that there will be a close relationship between the data objects that moves in the cloud environment. This implies that the object tag, convergence and the functional values of the root are factors that contribute to estimating the closeness of data object when detecting

provenance in the cloud. In the next section a critical evaluation of the concept is given.

6. CONCEPT EVALUATION

The testing experiments have generated five different scenarios for the objects {A,B} and {A,C}, {D,E}, {F,G}, {F,H} and the test values [1,2], [2,3], [2,4], [2,5] and [2,6] respectively. The authors chose to carry out these separate tests for two main reasons: To find the convergence of the roots that are treated as data objects that move within the cloud and to find the functional object that will help to show how closeness the objects can be during data provenance.

The results that have been portrayed in Figure 2, Figure 3, Figure 4, Figure 5 and Figure 6 illustrates that the roots of a bisection equation can be used as moving data objects in the cloud environment to compute how data provenance can be estimated. The results that managed to streamline inside had different iterations convergence and functional values.

The author has also noted {A, B} and {F,G} do not have identical convergence and functional value as {F,H}, {A,C} and{D,E} . This implies that given the same error (0.001), there still can exist a variation between this data objects. Because data can be misleading at time the authors have illustrated the scenarios using Table 3 which has shown how convergence and functional values have been achieved. Based on the presumption that objects must be close, similar with relatively identical tags as the move in the cloud in order for the physical devices or data center to be detected in the cloud, the authors conducted the experiments using f(x) (pow(x,3)-18) which enabled fast detection of the data objects which have been represented as roots. Therefore, it is the authors' opinion that during digital forensic investigation, forensic experts can utilize this technique to enable easy tracing the physical device in order to excavate potential evidence effectively.

In order to address the existing research gap that has been generated by the study of presentation of this research paper, the authors have provided a contextual evaluation based on the results of the experiments that have been conducted. The possible applicability of this can only be spearheaded with an effective prototype that will help forensic analyst to be able to reconstruct events by being able to seize the physical device for possible detection of security incidents.

7. CONCLUSION AND FUTURE WORK

The authors have proposed an approach for estimating data provenance of an object in the cloud environment using one-dimensional bisection method during digital forensic process. Based on the principles of digital forensics in the cloud, data might not be located easily but by employing bisection method the authors have deduced that the convergence shows that different data objects that have different tag names may be closely related based on this concept. Additionally, this contributes to the fact that detection of potential events in order to facilitate investigations can only be done if the actual physical machine can be able to be located.

In future the authors aim to develop a prototype which is able to support automatic detection of physical locations and physical devices of data that moves in the cloud environment. Furthermore, this prototype will be aimed at estimating the exact physical machine that dispatched the data object and we hope this will have a cost benefit situation while running the different scenarios.

8. REFERENCES

[1] K. Muniswamy-Reddy and M. Seltzer, "Provenance as First Class Cloud Data," SIGOPS Operating Systems Review, vol. 43, no. 4, pp. 11–16, 2010.

[2] M. I. M. Saad, K. A. Jalil and M. Manaf, "Achieving trust in cloud computing using secure data provenance," *Open Systems (ICOS), 2014 IEEE Conference on*, Subang, 2014, pp. 84-88.

[3] Zafarullah, F. Anwar and Z. Anwar, "Digital Forensics for Eucalyptus," *Frontiers of Information Technology (FIT), 2011*, Islamabad, 2011, pp. 110-116.

[4] Gary L Palmer.(2001). A Road Map for Digital Forensic Research. Technical Report DTR-T0010-01, DFRWS. Report for the First Digital Forensic Research Workshop (DFRWS).

[5] V. M. Katilu, V. N. L. Franqueira and O. Angelopoulou, "Challenges of Data Provenance for Cloud Forensic Investigations," *Availability, Reliability and Security (ARES), 2015 10th International Conference on*, Toulouse, 2015, pp. 312-317.

[6] Mell, Peter, and Tim Grance. "The NIST definition of cloud computing." (2011): 20-23.

[7] Suen, C. H., Ko, R. K., Tan, Y. S., Jagadpramana, P., & Lee, B. S. (2013, July). S2logger: End-to-end data tracking mechanism for cloud data provenance. In *Trust, Security and Privacy in Computing and Communications (TrustCom), 2013 12th IEEE International Conference on*(pp. 594-602). IEEE.

[8] Zhang, P., Tian, Y., & Liu, Z. (2009, February). Bisection method for evaluation of attraction region of passive dynamic walking. In *Autonomous Robots and Agents, 2009. ICARA 2009. 4th International Conference on* (pp. 692-697). IEEE.

[9] K. E. Atkinson, An Introduction to Numerical Analysis, New York: John Wiley and Sons, 1993.

[10] Pendharkar, P. C. (2008). A threshold varying bisection method for cost sensitive learning in neural networks. *Expert Systems with Applications*,*34*(2), 1456-1464.

[11] Trenwith, P. M., & Venter, H. S. (2014, August). A digital forensic model for providing better data provenance in the cloud. In *Information Security for South Africa (ISSA), 2014* (pp. 1-6). IEEE.

[12] Suen, C. H., Ko, R. K., Tan, Y. S., Jagadpramana, P., & Lee, B. S. (2013, July). S2logger: End-to-end data tracking mechanism for cloud data provenance. In *Trust, Security and Privacy in Computing and Communications (TrustCom), 2013 12th IEEE International Conference on*(pp. 594-602). IEEE.

[13] Ko, R. K., & Will, M. A. (2014, June). Progger: An Efficient, Tamper-Evident Kernel-Space Logger for Cloud Data Provenance Tracking. In *Cloud Computing (CLOUD), 2014 IEEE 7th International Conference on* (pp. 881-889). IEEE.

[14] Kebande, V. R., & Venter, H. S. (2015, August). Adding event reconstruction to a Cloud Forensic Readiness model. In *Information Security for South Africa (ISSA), 2015* (pp. 1-9). IEEE.

[15] Kebande, V. R., & Venter, H. S. (2014). A Cloud Forensic Readiness Model Using a Botnet as a Service. In *The International Conference on Digital Security and Forensics (DigitalSec2014)* (pp. 23-32). The Society of Digital Information and Wireless Communication.

[16] Kebande, V. R., & Venter, H. S. (2014, April). A cognitive approach for botnet detection using Artificial Immune System in the cloud. In *Cyber Security, Cyber Warfare and Digital Forensic (CyberSec), 2014 Third International Conference on* (pp. 52-57). IEEE.

[17] Kebande, V. R., & Venter, H. S. (2015, February). Obfuscating a Cloud-Based Botnet Towards Digital Forensic Readiness. In *Iccws 2015-The Proceedings of the 10th International Conference on Cyber Warfare and Security* (p. 434). Academic Conferences Limited.

[18] Kebande, V., & Venter, H. S. (2015, July). A Functional Architecture for Cloud Forensic Readiness Large-scale Potential Digital Evidence Analysis. In *Proceedings of the 14th European Conference on Cyber Warfare and Security 2015: ECCWS 2015* (p. 373). Academic Conferences Limited.

[19] Kebande, V., & Venter, H. (2015, October). Towards a Model for Characterizing Potential Digital Evidence in the Cloud Environment during Digital Forensic Readiness Process. In *ICCSM2015-3rd International Conference on Cloud Security and Management: ICCSM2015* (p. 151). Academic Conferences and publishing limited.

[20] Karie, N. M., & Venter, H. S. (2013, August). Towards a framework for enhancing potential digital evidence presentation. In *Information Security for South Africa, 2013* (pp. 1-8). IEEE.

[21] Kebande, V., & Venter, H. Requirements for Achieving Digital Forensic Readiness in the Cloud Environment Using an NMB Solution. In *11th International Conference on Cyber Warfare and Security: ICCWS2016* (p. 399). Academic Conferences and publishing limited.

[22] Kebande, V., Hermann Stephane Ntsamo & Venter, H. Towards a prototype for Achieving Digital Forensic Readiness in the Cloud using a Distributed NMB Solution. In 15th International Conference on Cyber Warfare and Security: ECCWS2016. Academic Conferences and publishing limited.-To Appear.

[23] Frank Tyger)" CS 211 Lesson 10 "Program Design" [online]-Accessed at "http://cse.unl.edu/~sincovec/Matlab/Lesson%2010/CS211%20Lesson%2010%20-%20Program%20Design.htm