Advances In Multidisciplinary & Scientific Research
A Multidisciplinary & Interdisciplinary Journal

Vol. 2, No. 2, June, 2016

Improving Digital Forensics Analysis in Federated Domains through Estimator Analysis and Network Flow Optimization

Ojeniyi, J.A.
Department of Cyber Security Science,
School of ICT
Federal University of Technology,
Minna, Nigeria
ojeniyija@futminna.edu.ng

Okunoye, A.
Williams College of Business
Xavier University
Cincinati, Ohio, USA
okunoye@xavier.edu

Longe, O.B
Department of Computer Science & Mathematics
Adeleke University, Ede Osun State, Nigeria
longeolumide@fulbrightmail.org

Oguntade, E.S.
Department of Statistics,
University of Abuja
Abuja, Nigeria
oguntadeemmanuel@yahoo.com

ABSTRACT

Digital Forensics is a field that deals with safe and unaltered collection of vital data from the scene of crime incidence for the purpose of investigation and prosecution. Different tools have been developed to help in analysing or estimating the degree or extent of the criminality. However, the exponential growth and expansion being experienced in field of computing and networking is making these estimations or forensic analysis more or less accurate. Some of the reasons militating against effective analysis are attributed to various inhibiting policies across different platforms, routers, domains of networking. In this paper, some tools used for forensics analysis or estimating the probative values of digital evidence are referred to estimators. Three of these estimators are selected and tested in a simulated environment. Analysis of three digital forensics estimators (EnCase, Safeback and TootKit) is carried out in this paper. This is experimentally aided by simulation of heterogeneous domain-based network and packet analyzer is used to collect probability reading in the packet option field at each hop along the communication path between an attacker and the victim. The graphical analysis with varied initial values shows that estimation accuracies of the estimators reduce irrespective of initial values. With the developed model, the router could be configured for packet boosting at the point of dwindling probabilities using Maximum Network Flow Algorithm.

Keywords- Digital Forensics Estimators; Grid Computing; Cloud Computing; Packet Tracking Model; Probability Distribution; Maximum Network Flow..

Aims Research Journal Reference Format:
Ojeniyi, J A., Longe, O.B & Ogutade, E.S. (2016): Improving Digital Forensics Analysis in Federated Domains through Estimator Analysis and Network Flow Optimization. Advances in Multidisciplinary Research Journal. Vol 2, No. 2 Pp 1-12.

1. INTRODUCTION

The extensibility of web-based applications and open source software has given room for improved collaborative efforts. The globalization growth in Information and Communication Technology(ICT) has drawn great number of geographically dispersed digital users. The cyber space is being populated with many contributors of varied motives.[4,11] This trend has, however, given ample opportunities to malicious actors or experts to carry out their unethical activities without being discovered and if possible signalling the indictment to another digital user(s).[15,18,20]. In order to avoid or reduce the rate of unjust punishment on the innocent cyber users, the field of digital forensics came up with the aim of identifying, preserving, extracting and documenting digital evidence for the purpose of successful prosecution and conviction [5,9,13,22]. Many tools have been developed for the estimation and analysis of digital crime incidents. Since we have two categories of digital forensics (computer and network forensics), some of the estimators are computer forensics-oriented, some are network forensics-oriented while some function in both.[2,6,15,16,23]

The exponential growth in the networking of digital devices and the associated increased crime activity make it complex for the conventional network forensic-oriented digital estimators to carry out their forensics analysis[7,10,11]. Peculiar feature of grid computing of spanning across heterogeneous domains has particularly led to dwindling efficiency of digital forensics estimators. In addition, though data communication in cloud computing is within homogeneous domains the virtual boundaries existing created loose coupling between the digital components thereby introducing administrative bottlenecks in the forensic analysis. Different approaches have tried to profer solutions to some of these challenges. Some of the approaches are traceback of active attack flows, out-of-band, router based, deterministic packet marking, probabilistic packet marking and so on[28]. The major setbacks of these approaches are increased overhead thereby leading to slow bandwidth, blocking of ICMP messages, lacking general applicability or extensibility.

Another approach applied the concept of artificial intelligence. Semantic Web Language OWL was used to represent domain-specific event based knowledge. This was later extended to events sourced from new domains to support reasoning across multiple heterogeneous domains[29]. However, expression of correlation rules by forensic expert would rather be complex and difficult. Three Digital Forensics Estimators are analyzed across multiple domains to determine their probabilistic effectiveness with increasing hop counts. In this line, a model is proposed and developed. The essence of this model with the aid of graphical analysis is to have the picture of the point of attenuation beyond which estimation value loses significance.[12,14]

1.1 Maximum Network Flow

It was developed by two Mathematicians by name Lester Randolph Ford and Delbert Ray Fulkerson in 1956. The suitability of its use in this paper is because, the algorithm works when there are algorithms with different running times. It is used to solve Maximum Flow Problem of the nature below: Given a flow network G is defined by two-tuple (V,E), where V is Vertex and E is Edge. In order to determine the greatest possible flow $f(u,v)$ from the source (s) to sink (t) without violating capacity $c(u,v)$, the problem can be represented diagrammatically as seen in Fig. 1.

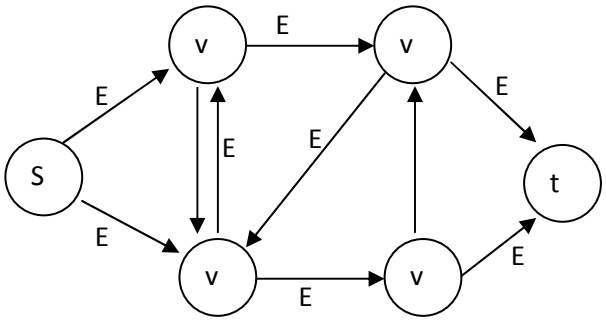
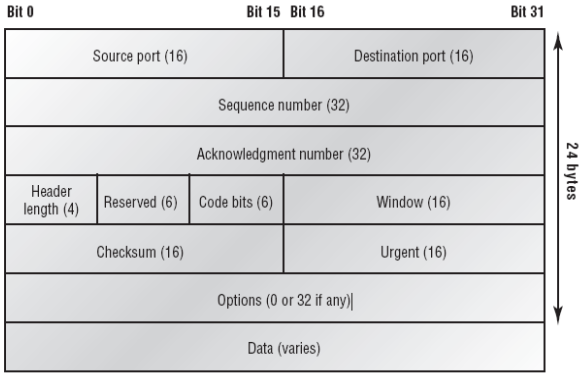


Figure 1: Flow Network *G = (V,E)* with flow *f(u,v)* and capacity *c(u,v)* where *E* consists of *f(u,v)/c(u,v)*

2. SIMULATION

TCP/IP protocol suite constitutes the larger percentage of network data communication. Basically, it has four layers: application layer, transport layer, network layer and data link layer. Encapsulation of data takes place from application layer to data link into frame, which will later be de-encapsulated in reverse order at the receiver's end. Encapsulation entails addition of headers and trailers. At the network layer, IP header is added to the segment to form packet. So, data communication at Network layers involves transmitting of IP packets [24].

The architecture of IP packet includes a field called "option" field, as shown in Fig. 2.



Fgure 2: IP header (Sybex, 2008)

During data communication, each hop encountered writes its probability in the option field of the IP header. Since each packet has source and destination addresses, it is easy to track the packet distances from its source and from its destination [8,27]. RouterSim was used to simulate heterogeneous network and Packet Analyzer was used to get the probabilities at various hops along the path of crime incident.

2.1 Packet Tracking Model

The Packet Tracking Model abbreviated as *PTM* is the probability model of attacker's discovery. In a situation of Digital Forensic Analysis in order to uncover the source of digital crime, this model tends to capture the efficacies of Digital Forensic Estimators in tracking the attacker as hop count increases.
The simulated heterogeneous network assumes the diagrammatic view in Fig. 3.

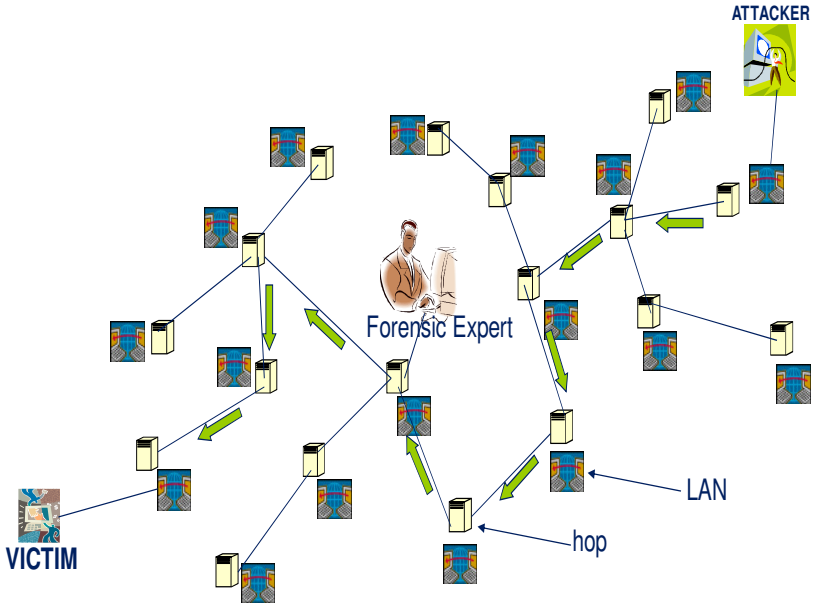


Figure 3. Heterogeneous Network for Digital Forensic Analysis

3. MODEL DEVELOPMENT

Variable declaration
Let *p* be the probability written by any hop in IP header option field
Let *d* be distance between a hop and any victim
The packet analyzer shows that there is a declining probability with increase in *d*,

Therefore, it holds that,

$$p \alpha \frac{1}{\sqrt{d}} \quad (1)$$

Due to the discrete trials and dichotomous qualitative variable, i.e. success of forensics/marking versus failure of forensics/marking, this necessitated the choice of the probability distribution function, *p.d.f* in (2). Since we are interested in one of the dichotomous variable (i.e. the success of each estimator tracking the crime committed), then the binomial distribution becomes important in the analysis of the data[1].

$$p.d.f = {}^{n}C_{x}\, p^{x}\, (1 - p)^{n-x} \quad (2)$$

There are two factors in this pdf:
(i) ${}^{n}C_{x}$ represents the number of different possible attack paths while
(ii) $p^{x}(1-p)^{n-}$x represents the probability of any attack path
Since, 1 attack path of crime incident is being modelled in the simulated network; the first factor [${}^{n}C_{x}$] is discarded, while the second factor is considered.

Packet Tracking Model (*PTM*) given as,

$$(PTM) = p^{x}(1-p)^{n-x} \quad (3)$$

Considering an attack from router 1, by IVP (i.e. Initial Value Problem)

$PTM = p^{x_0}(1-p)^{n-x_0}$ (4)

Taking $x_0 = 1$, $n = hops$

$PTM = p^1(1-p)^{d-1}$ (5)

Putting $d = d - 1$ in (1) (Iterative process), we have

$$p \alpha \frac{1}{\sqrt{d-1}}$$

$$p = \frac{k}{\sqrt{d-1}} \quad (6)$$

$$p\sqrt{d-1} = k$$

Therefore,

$PMT = p(1-p)^{\sqrt{d-1}}$ (7)

Equation (7) is the model for finding the probability of discovering an attacker at a given hop at distance, d, from the crime scene, where p stands for the probability of the hop written in the option field of IP header.

3.1 Model Validation

There are several ways to validate the authenticity of a model depending on its nature. Reliability or validity theory could be adapted in this particular model because of its dichotomy.In order to validate PTM, it can be viewed as a system consisting of many components. The functionality of some or all of the system components has bearing on the validity of the system model. Adapting Reliability theory to validate Packet Tracking Model, fundamentally, it says, 'a series system will function if and only if all its components are functioning, while a parallel system will function if and only if at least one of its components is functioning' [26]. From the principle of logic computation, logic-AND works based on a series system while logic-OR based on parallel system.

Model *PTM* can be iteratively expressed as,

$$PTM = p*\left[\{1-p\}_1 * \{1-p\}_2 * \{1-p\}_3 \cdots \{1-p\}_{n=\sqrt{d-1}}\right] \quad (8)$$

From the Table 1, within the same router, i.e. when the *hop count* is 1, the three forensics estimators are functioning at their best with the highest possible initial probabilities.

If we take $d = 1$, then the exact initial probabilities are expected for the estimators.

When $d = 1$, $\Rightarrow n = 0$

$PTM = p*\left[\{1-p\}_0\right]$ (9)

$PTM = p$ (10)

Clearly, (10) gives exact initial probabilities of the estimators. This implies that the Forensics Estimators will function at their best when tracking criminality within a hop count.

3.2 Maximum Network Flow

In order to enhance the effectiveness of forensic tools across the multiple routers, Maximum Network Flow Optimization is used to serve as boaster at the point of attenuation.

From Figure 1,

$$E = \frac{f(u,v)}{c(u,v)} \qquad (11)$$

The algorithm to be used is as follows:

FORD-FULKERSON-METHOD(G, s, t)
initialize f to 0
while there exists an augmenting path p
do augment flow f along p
return f
where,
G is the flow network *G = (V,E)*
s is the source of the flow
t is the sink
f is the flow from s to t
p is the path

4 Algorithm Explanation

- Iterative process with the flow *f* initial value set to 0
- On each iteration, increase flow f by finding an "augmenting path" and augmenting the flow along this path
- Repeat process until no augmenting path can be found, process terminates yielding maximum flow

5 Flow Properties

1. Capacity Constraints:
   A flow is less than or at most equal to the capacity of the edge,
   i.e. $f(u,v) \leq c(u,v)$
2. Skew symmetry
   Given a flow $f(u,v)$ means a flow from vertex *u* to vertex *v*, the inverse direction gives negative
   i.e. $f(u,v) = - f(v,u)$
3. Flow Conservation
   At a vertex, the total flow entering it is equal to the total flow leaving it
   i.e. $\sum f(u,v) = 0$, at a given vertex

6 Residual Network

Residual network has edges that can admit more flow. This implies that it has residual capacity.
The extra flow that can be pushed to an edge without exceeding the capacity is termed as residual capacity, *cr(u,v)*
i.e.
$cr(u,v) = c(u,v) - f(u,v)$ (12)
Then,
Residual network Gr is a flow network with capacities cr.

7 Augmenting Path

A path *p* from *s* to *t* in the residual network *Gr* is referred to as *Augmenting Path*.
The maximum amount the flow can be increased on each edge in the augmenting path *p* is called the residual capacity of *p*, i.e.
$c_r = min\{c_r(u,v):(u,v) \text{ is on } p\}$ (13)

Residual Algorithm
Pseudo code

```
Ford-Fulkerson(G,s,t)
1 for each edge (u,v) Є E[G]
2    do f[u,v] ←0
3         f[v,u] ← 0
4 while there exists a path p from s to t in the residual  network Gf
5    do cf(p) ←min{cf(u,v) : (u,v) is in p}
6           for each edge (u,v) in p
7                do f[u,v] ←f[u,v] + cf(p)
8                      f[v,u] ← -f[u,v]
```

4. RESULTS DISCUSSION

This model is used to generate the graphical analysis of three Digital Forensic Estimators with increase hop counts.

Assumptions

Initial values: for *SafeBack(S) = 0.9, Encase (E) = 0.7, ToolKit(T) = 0.5*
With the above assumed probability Initial values, knowing that the probability values are between 0 and 1, the data was generated and corresponding graphs were plotted. The marking probabilities of the three estimators with S=0.9, E=0.7 and T=0.5 within 21 hops are given in Table I and the corresponding graphs plotted in Fig. 3.

Fig. 3 clearly shows that Packet Tracking Model is experiencing a sharp decline from 2 to 3 hop counts. This implies that the probability of identifying an attacker using the conventional forensic estimator witnesses a sharp decline of efficiency especially within the first four routers. The figure further shows that the Packet Marking Probabilities decrease asymptotically as from the fourth hop count. Safeback with highest initial value records least marking beyond 3 hops while Toolkit with least initial value has relatively highest marking probabilities.

TABLE I. PROBABILITIES OF ESTIMATORS AGAINST HOP COUNTS WITH INITIAL VALUES, *S=0.9, E=0.7, T=0.5*

Hop Count	S	E	T
1	0.900	0.700	0.500
3	0.875	0.675	0.475
5	0.850	0.650	0.425
7	0.825	0.625	0.375
9	0.800	0.600	0.325
11	0.775	0.575	0.275
13	0.750	0.550	0.225
15	0.725	0.525	0.175
17	0.700	0.500	0.125
19	0.675	0.475	0.075
21	0.650	0.450	0.025

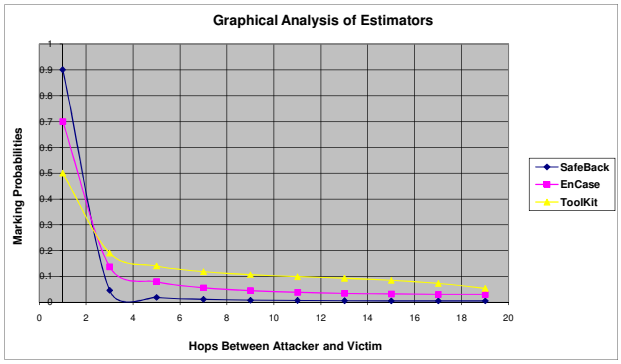


Figure 4. Graphical Analyses of Digital Forensic Estimators with initial values, *S=0.9, E=0.7, T=0.5*

On the other hand, Table II shows initial values of *S=0.9, E=0.5, T=0.7.* In line with previous pattern, EnCase with least initial value records relatively highest marking probabilities beyond three hops as shown in Fig. 4.

TABLE II. PROBABILITIES OF ESTIMATORS AGAINST HOP COUNTS WITH INITIAL VALUES, *S=0.9, E=0.5, T=0.7*

Hop Count	S	E	T
1	0.900	0.500	0.700
3	0.875	0.475	0.675
5	0.850	0.450	0.625
7	0.825	0.425	0.575
9	0.800	0.400	0.525
11	0.775	0.375	0.475
13	0.750	0.350	0.425
15	0.725	0.325	0.375
17	0.700	0.300	0.325
19	0.675	0.275	0.275
21	0.650	0.250	0.225

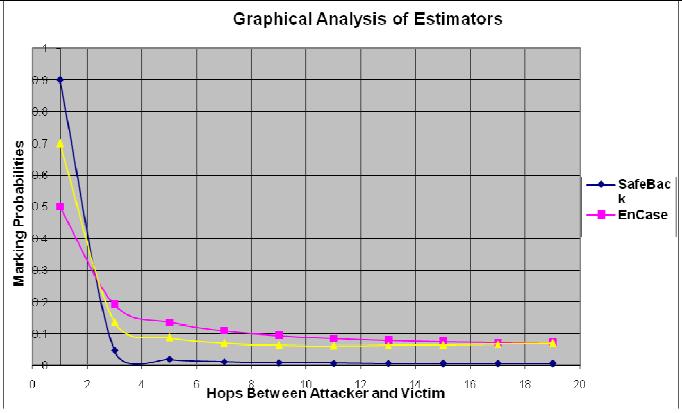


Figure 5. Graphical Analyses of Digital Forensic Estimators with initial values, *S=0.9, E=0.5, T=0.7*

In Table III, the initial values used are *S=0.5, E=0.7, T=0.9.* In agreement with observed pattern, Toolkit with higest initial value of 0.9 records least marking probabilities beyond three hops as shown in Fig. 5.

TABLE III. PROBABILITIES OF ESTIMATORS AGAINST HOP COUNTS WITH INITIAL VALUES, *S=0.5,* *E=0.7, T=0.9*

Hop Count	S	E	T
1	0.500	0.700	0.900
3	0.475	0.675	0.875
5	0.450	0.650	0.825
7	0.425	0.625	0.775
9	0.400	0.600	0.725
11	0.375	0.575	0.675
13	0.350	0.550	0.625
15	0.325	0.525	0.575
17	0.300	0.500	0.525
19	0.275	0.475	0.475
21	0.250	0.450	0.425

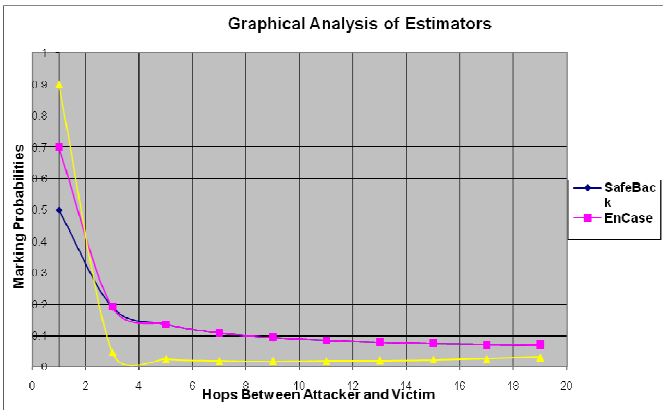


Figure 5. Graphical Analyses of Digital Forensic Estimators with initial values, *S=0.5, E=0.7, T=0.9*

4.1 Maximum Network Flow Optimization

From the analysis above, it is clear that as the number of hops is increasing, the probative values are decreasing. The algorithm of Ford-Fulkerson Method can be employed before the forensic tool is used. In this case,

Our network is $G = (V,E)$

Where V = each node in the network
E = each attack path
s = the source of the attack
t = the point of attenuated signal
f = direction of flow from s to t

5. CONTRIBUTION TO KNOWLEDGE

In this research work, probability density function-based packet tracking model was developed. Also, maximum network flow-based approach was developed to mitigate the challenges of attenuation.

6. CONCLUSION

It is clear from the Fig. 3 to Fig. 5 that irrespective of the initial probability values in the option field of the packet header as seen in Fig. 1, the marking probabilities of the three estimators keep on decreasing across multiple heterogeneous domains. The practical implication of this is that accuracy and reliability of digital forensic results keep reducing. From the three graphs in Fig. 3, Fig. 4 and Fig.5, there is a point of intersection, just before asymptotic movement between the third and fourth routers. This is the point of attenuation. Since routers are intelligent, configurable and customizable, they can be configured for packet boosting at this point of attenuation. The values on the Table I, Table II and Table III can be used to configure the hop at any desired hop count number.

This approach introduces decentralization of administration thereby reducing administrative bottlenecks experienced in other approached or overhead cost incurred. More so, the Maximum Network Flow Algorithm will then serve as booster to every attenuation.

7. FUTURE WORK

The work was tested and evaluated on a simulated testbed. The direction of the future work is carrying out the implementation on physical network infrastructure. Scalability of network devices is also recommended to further evaluate the performance of the estimators.

REFERENCES

1. C.A. Awogbemi, and E.S. E.S.Oguntade, 2010. Elements of Statistical Methods, Suntos Books, Ibadan.
2. J.J. Barbara, 2005. Digital evidence accreditation in the corporate and business environment. Digital Investigation, vol. 2, no. 2, pp. 137-146.
3. Barros, 2008. A proposal for ICMP traceback messages. Internet Draft http://www.research.att.com/lists/ietfitrace/ 2000/09/msg00044.html.
4. Baryamureeba, and F. Tushabe, 2004. The enhanced digital investigation process model. Proceedings of the fourth Digital Forensic Research Workshop.
5. N. Beebe, and J. Clark, 2 005. Dealing with terabyte data sets in digital investigations. Advances in Digital Forensics, pp. 3-16. Springer.
6. N.L. Beebe, and J.G. Clark, 2004. A hierarchical, objectives-based framework for the digital investigations process. Proceedings of the fourth Digital Forensic Research Workshop.
7. Berendt, 2000. Web usage mining, site semantics, and the support of navigation. Proceedings of the Second International Workshop on Visualizing Software for Understanding and Analysis.
8. K . Boundaoud, and F. LeBorgne, 2008. “Towards an Efficient Implementation of Traceback Mechanisms in Autonomous Systems”, University of Nice Sophia Antipolis – 13S-Laboratory – CNRP.
9. R. Brown, B. Pham, and O. de Vel, 2005. Design of a digital forensics image mining system. Proceedings of the International Workshop on Intelligent Information Hiding and Multimedia Signal Processing.
10. F. Buchholz, and C. Falk, 2005. Design and implementation of Zeitline: a forensic timeline editor. Proceedings of the fifth Digital Forensic Research Workshop.
11. F. Buchholz, and E. Spafford, 2004. On the role of file system metadata in digital forensics. Digital Investigation, vol. 1, no. 4, pp. 298-309.
12. Carney, and M. Rogers, 2004. The Trojan made me do it: a first step in statistical based computer forensics event reconstruction. International Journal of Digital Evidence, vol. 2, no. 4.
13. Carrier, and E.H. Spafford, 2003. Getting physical with the digital investigation process. International Journal of Digital Evidence, vol. 2, no. 2.
14. B.D. Carrier, and E.H. Spafford, 2005. Automated digital evidence target definition using outlier analysis and existing evidence. Proceedings of the fifth Digital Forensic Research Workshop.
15. Casey, 2002. Handbook of computer crime investigation: forensic tools and technology. Academic Press.
16. Casey, 2004. Digital evidence and computer crime: forensic science, computers and the Internet. Academic Press.
17. Casey, 2004. Network traffic as a source of evidence: tool strengths, weaknesses, and future needs. Digital Investigation, vol. 1, no. 1, pp. 28-43.
18. M. Eirinaki, and M. Vazirgiannis, 2003. Web mining for Web Personalization. ACM Transactions on Internet Technology, vol. 3, no. 1, pp. 1-27.
19. A.P. Engelbrecht, 2003. Computational intelligence: an introduction. Wiley.
20. B.K.L. Fei, J.H.P. Eloff, M.S. Olivier, H.M. Tillwick, and H.S. Venter, 2005. Using self-organising maps for anomalous behaviour detection in a computer forensic investigation. Proceedings of the Fifth Annual Information Security South Africa Conference.
21. Habib, S. Fahmy, S. R. Avasarala, V. Prabhakar, and. B. Bhargava, 2003. On detecting service violations and bandwidth theft in QoS network domains. Journal of Computer Communications..
22. I.I. Kruse, and J.G. Heiser, 2002. Computer forensics: incident response essentials. Addison-Wesley.
23. C.V. Marsico and M.K. Rogers, 2005. iPod forensics. International Journal of Digital Evidence, vol. 4, no. 2.

24. G. Michael, 2008. “Probabilistic Packet Marking for Large-Scale IP Traceback”, IEEE.
25. K Park, and H. Lee, 2001. A proactive approach to distributed DoS attack prevention using route-based packet filtering. In Proc. ACM SIGCOMM, San Diego, CA.
26. S. M. Ross, 1997. Introduction to Probability Models. 6th ed., Academic Press, Toronto.
27. K. Shaoh-Chen, and C. Yen-Wen, 2007.“An Edge Router-Based Fast Internet Traceback”, Department of Communication Engineering, National Central University, Taiwan, ROC.
28. Wikipedia, the free online encyclopedia
29. Schatz, G. Mohay, and A. Clark, (2004) Generalising Event Forensics Across Multiple Domains, In Asia APIEMS, Brisbane, Australia.
30. K. A. Ravindra, L. M. Thomas, and B. O. James. Network Flows: Theory, Algorithms, and Applications. Prentice Hall, 1993