Information Leakage Detection & Prevention

[1]Prof. Pate Sumeet, [2]Mr. Bhide Arun, [3]Ms. Paul Esther, [4]Ms. Sawardekar Shruti

[1]Asst. Professor, [2,3,4]UG Student, [1,2,3,4]Computer Engg. Dept. Shivajirao S. Jondhle College of Engineering & Technology, Asangaon, Maharshatra, India.

[2]*arunbhide031996@gmail.com*, [3]*estherpaul0407@gmail.com*, [4]*shrutisawardekar1@gmail.com*

Abstract—In today's regulated business surroundings, the loss of information records carries significant penalties which may be expressed in terms of lost time, money, client relations and ultimately lost profits. It is obvious that security product should increase data protection and not decrease it; so it's vital to notice the knowledge run and additional vital to forestall it. A comprehensive data leak protection system is as very important for anyone to UN agency must forestall the information from unauthorized persons.

Keywords—Certificates, Man-in-the-middle Attack, Information leakage probabilistic language, Non-termination.

I. INTRODUCTION

In the course of doing business, sometimes sensitive data must be handed over to supposedly trusted third parties. As an example, a hospital may give patient records to researchers who will devise new treatments. Similarly, a company may have partnerships with other companies that require sharing customer data. Another enterprise may outsource its data processing, so data must be given to various other companies. It calls the owner of the data the distributor and the supposedly trusted third parties the agents. The goal is to detect when the distributor's sensitive data has been leaked by agents, and if possible to identify the agent that leaked the data. It considers applications where the original sensitive data cannot be perturbed. Perturbation is a terribly helpful technique wherever the information is changed and created "less sensitive" before being bimanual to agents. As an example, one will add random noise to sure attributes, or one will replace actual values by ranges. However, in some cases it's vital to not alter the first distributor's information. For example, if associate outsourcer is doing our payroll, he should have the precise pay and client checking account numbers. If medical researchers are treating patients (as hostile merely computing statistics), they'll want correct information for the patients. Historically, run detection is handled by watermarking, e.g., a novel code is embedded in every distributed copy. If that duplicate is later discovered within the hands of associate unauthorized party, the source are often known. Watermarks are often terribly helpful in some cases, but again, involve some modification of the first information.

Furthermore, watermarks will generally be destroyed if the information recipient is malicious. During this paper it studies retiring techniques for detective work run of a group of objects or records. Specifically, The study of the subsequent scenario: when giving a group of objects to agents, the distributor discovers a number of those self same objects in associate unauthorized place. (For example, the information is also found on an internet web site, or is also obtained through a legal discovery method.) At now the distributor will assess the probability that the leaked information came from one or additional agents, as hostile having been severally gathered by alternative suggests that exploitation associate analogy with cookies purloined from a jar, if we have a tendency to catch Freddie with one cookie, he will argue that an acquaintance gave him the cookie. However if we have a tendency to catch Freddie with five cookies, it'll be abundant tougher for him to argue that his hands weren't within the jar. If the distributor sees "enough evidence" that associate agent leaked information, he might stop doing business with him, or might initiate legal proceedings. During this paper it's been develop a model for assessing the "guilt" of agents. It conjointly gift algorithms for distributing objects to agents, during a approach that improves the possibilities of characteristic a source. Finally, it contemplate the choice of adding "fake" objects to the distributed set. Such objects don't correspond to real entities however seem realistic to the agents. In a sense, the faux objects acts as a sort of watermark for the whole set, while not modifying anyone members. If it seems associate agent was given one or additional faux objects that were leaked, then the distributor are often additional assured that agent was guilty. The quantification of knowledge run provides a quantitative analysis of the safety of a system. This report propose the usage of Markovian processes to model settled and probabilistic systems. By employing a methodology generalizing the lattice of knowledge approach we have a tendency to model refined attackers capable to watch the inner behavior of the system, and quantify the knowledge run of such systems.

II. RELATED WORK

A. *The SSL Protocol*

The Secure Socket Layer (SSL) protocol was designed to ensure secure communications between 2 entities over untrusted networks. The SSL protocol provides authentication based on the X.509 public key infrastructure, protects information
confidentiality exploitation regular cryptography, and ensures information integrity with cryptanalytic message digests. SSL is often used for securing websites and mail servers, preventing passive network attackers from eavesdropping or replaying the client's messages, and is usually thought of security best observe for websites. By enabling cryptography, websites will simply forestall the eavesdropping of unencrypted confidential information.

B. *The SSL Man-in-the-Middle Attack*

The SSL man-in-the-middle (MITM) attack could be a type of active network interception wherever the wrongdoer inserts itself into the communication between the victim shopper and also the server (typically for the aim of eavesdropping or manipulating personal communications). The wrongdoer establishes 2 separate SSL connections with the shopper and also the server, and relays messages between them, during a approach such each the shopper and also the server area unit unaware of the middleman. This setup allows the wrongdoer to record all messages on the wire, and even by selection modify the transmitted information.

III. LITERATURE SURVEY

Data security is important for many businesses and even data processor users. consumer info, payment info, personal files, checking account details - all of this info may be exhausting to interchange and probably dangerous if it falls into the incorrect hands. knowledge lost because of disasters like a flood or fireplace is crushing, however losing it to hackers or a malware infection will have a lot of larger consequences.

[i] ALGORITHM1: WATERMARK (DIRECTORY D)

1: while D has children
2: di←child I of D
3: if di is a directory then
4: Watermark(di)
5: else
6: Boolean w=DetectWatermark(di)
7: if w=TRUE then
8: compare watermark of di with permissions tag
9; if Watermark does not match tag then
10: Quarantine or securely Remove di
11: end if
12: else
13: Watermark di with signature=permissions tag
14: end if
15: end if
16: end while
17: return

Mointor():

1: W=inotify event descriptor
2: for all Target directories di do
3: Add inotify watch descriptor for "write" and "create" operations within di
4: end for
5: loop
6: f =Read event from event descriptor W
7: pass f to Watermarking Agent for Analysis
8: end loop.

[ii] ALGORITHM 2: FAKE OBJECT

Distributor may be able to add fake objects to the distributed data in order to improve his effectiveness in detecting the guilty agents. However, fake objects may impact the correctness of what agents do, so they may not always be allowable.

Allocation for explicit data requests(EF)
Input: R1,R2,……Rn, cond1,…..condn,b1,……..bn,B
Output: R1,…Rn, F1,….Fn

1. R ← 0 Agents that can receive fake objects
2. fori← 1,…..; n do
3. if bi>0 then
4. R←R U {i}
5. Fi ← 0;
6. while B>0 do
7. i← SELECTAGENT(R1,R2,……Rn)
8. F←CREATEFAKEOBJECT()
9. Ri←RiU{f}
10. Fi←FiU{f}
11. bi←bi-1
12. if bi= then
13. R←R\{Ri}

IV. EXISTING SYSTEM

In existing system, there's no risk to distributed and perturb the sensitive and secured knowledge that is capable of constructing the information less sensitive. Within the existing systems, the information of the initial distributor are going to be modified that is understood as knowledge leak that is incorporated with the embedded data copies. The information leakers may be known only if the information copy is among uncertified user hands and also the watermarks can even be non continuous by the information recipients United Nations agency flip the information into malicious data.

- historically, leak detection is handled by watermarking, e.g., a singular code is embedded in every distributed copy.
- If that replicate is later discovered within the hands of associate unauthorized party, the source may be known.

ALGORITHM MARKOV CHAIN

Data: A markov chain C= (S,s_o,P), the set $T_A \subseteq$ S of states to hide.

1. *Add to S the divergence state* $\uparrow$ *with* $P_{\uparrow,\uparrow}$ *=1 and* $\Pi_{\uparrow}^{(0)}$ *=0;*
2. *While* $T \neq \emptyset$ *do*
3. *Choose a state* $t \in T_A$;
4. *If* $P_{t,t}$*=1 then*
5. $\pi_{\uparrow}^{(0)} \leftarrow \pi_{\uparrow}^{(0)} + \pi_t^0$
6. *For each* $s \in S$ *do*
7. $P_{s,\uparrow} \leftarrow P_{s,\uparrow} + P_{s,t}$
8. $P_{s,t} \leftarrow 0$
9. *end*
10. *Else*
11. *For each* $u \in S$ *do*
12. $P_{t,u} \leftarrow \frac{P_{t,u}}{1-P_{t,t}}$
13. $\pi_u^{(0)} \leftarrow \pi_u^{(0)} + \pi_u^{(0)} P_{t,u}$
14. *For each* $s \in S$ *do*
15. $P_{s,u} \leftarrow P_{s,u} + P_{s,t} P_{t,u}$
16. $P_{s,t} \leftarrow 0$
17. *end*
18. *End*
19. *End*
20. $S \leftarrow S \backslash \{t\}$
21. $T_A \leftarrow T_A \backslash \{t\}$
22. *End*

V. TIME COMPLEXITY

Algorithm 1 finds agents that are eligible to receiving fake objects in O (n) time. Then, in the main loop, the algorithm creates one fake object in every iteration allocates it to random agent. The main loop takes O(B).

VI. PROBLEM STATEMENT

In the business process, sometime owner of data gives set of sensitive data to trusted agents for performing some operation on it. This type of data is very sensitive and leakage of this type of data happens when confidential business data is leaked out, if that data leaked and found in some unauthorized place, it leaves the company unprotected and destroys the image and customers trust and goes outside the jurisdiction of the corporation. This uncontrolled data leakage puts business in a vulnerable position. If this data is no longer within the domain, the company is at serious risk hence distributor must find out the guilty agent if the leaked from one or more agents, as op-posed to having been independently gathered by other means. Here the data allocation strategies (across the agents) that improve the probability of identifying guilty agent are pro-posed. This method works if leaked data is obtained as it was distributed or if fake records are deleted

Suppose a distributor owns a set *T* = {t1 ,*tm*} of valuable data objects. The distributor wants to sharesome of the objects with a set of agents *U*1,*U*2,… ,*U*n but does wish the objects be leaked to otherthird parties. An agent *Ui*receives a subset of objects *Ri*which belongs to *T*, determined either by a sample request or an explicit request, Sample Request *Ri*= SAMPLE (*T*,*mi*) : Any subset of *mi* records from *T* can be given to *Ui*.Explicit Request *Ri*= EXPLICIT (*T*,*condi*) : Agent *Ui*receives all the *T* objects that satisfy *condition* . The objects in *T* could be of any type and size, e.g., they could be tuples in a relation, or relations in a database. After giving objects to agents, the distributor discovers that a set *S* of *T* has leaked. This means that some third party called the target has been caught in possession of *S*. For example, this target may be displaying *S* on its web site, or perhaps as part of a legal discovery process, the target turned over *S* to the distributor. Since the agents *U*1,*U*2,… ,*U*n, have some of the data, it is reasonable to suspect them leaking the data. However, the agents can argue that they are innocent, and that the *S* data was obtained by the target through other means.

ENTITY & AGENTS

A distributor owns a set T = ft1; : : : ; tmg of valuable data objects. The distributor wants to share some of the objects with a set of agents U1;U2; :::;Un, but does wish the objects be leaked to other third parties. The objects in T could be of any type and size, e.g., they could be tuples in a relation, or relations in a database. An agent Ui receives a subset of objects Ri _ T, determined either by a sample request or an explicit request:

_ Sample request Ri = SAMPLE(T;mi): Any subset of mi records from T can be given to Ui.

_ Explicit request Ri = EXPLICIT(T; condi): Agent Ui receives all the T objects that satisfy condition.

VII. PROPOSED SYSTEM

In the proposed system, distributors of data can identify the unauthorized users and their locations where their original data is been changed. In this system, if the unauthorized user is identified, then the distributor can stop distributing their data with the agents and even can legally penalize them for data leakage cases. This project includes the development of a specific model by using different types of algorithms that supports the distributors to identify the unauthorized data users and it can be used to assess the faults done by the third party agents by using fake objects. In a sense, the fake objects acts as a type of watermark for the entire set, without modifying any individual members.

A. Probability Spaces and Discrete-Time Markov Chains

Probability Spaces:

Non-terminating programs may produce an infinite number of observable values; therefore, we cannot represent the probability of observing values as a simple mapping from iparticular values to a number between 0\ and 1. Instead we must represent the probability of observing values as a *probability space*, which is a triple (, B, P). Here, is the set of all possible *events* (i.e., secret and observable values), which is often infinite. B is a _*-algebra* over the set , which is a set B ⊆ 2 of subsets of that contains ∅ and is closed under complement and countable unions. For a set G ⊆ 2, we say that _-algebra B is *generated* by G if it is the smallest _-algebra containing G. P : B → [0, 1] is a *probability measure* over (, B); for each member of B it gives the probability of an event from that set occurring. We call (, B) a *measurable space* and sets b ∈ B are said to be *measurable*.

Let (X, BX) be another measurable space. A *random variable* X defined on (, B) and taking values in (X, BX) is a function X : → X such that, for each bX ∈ BX, X−1(bX) ∈ B where X−1(bX) = {! ∈ | X(!) ∈ bX}.

This means that the random variable X has an associated probability distribution PX : BX → [0, 1] giving the probability
of each bX ∈ BX: PX(bX) = P(X−1(bX)).

Finally, for two (probability) measures P1 and P2, we say that P1 is *absolutely continuous* with respect to P2, if P1(b) = 0 for every set b for which P2(b) = 0.

VIII. SYSTEM ARCHITECTURE

Information security has been researched to considerable depth in the ongoing quest to provide users and corporate entities a more secure computing environment. Although an extraordinary range of effective approaches have been developed to mitigate threats to information security, new threats appear daily. Within the realm of such threats, among the most difficult to detect and prevent involve covert channel, or side channel, attacks.

The biggest DLP challenge lies in protecting the large amounts of sensitive data which exist in unstructured form (e.g., various types of intellectual property like source code, customer lists, and product designs). Therefore, DLP solution providers are continuously improving their data discovery methods using approaches such as fingerprinting and natural-language processing.

CH-IMP MODEL

Command C Conforming the grammar

ρ ranges over probability distribution on arithmetic expressions

B is Boolean expression

1. Initialize a variable V
2. New V:=ρ
3. If (B) {C} else {C}
4. While (B) {C}
5. C;
6. Check the randomness

   Case 1: new rand={0→0.5,1→0.5};
   Observe rand;
   New rand={0→0.5,1→0.5};
   Secret sec;
   New out=sec XOR rand;
   Observe out;

7. Case 2:consider the possible leakage from sets of secrets

   New sec 1 = {0→.0.5,1→.0.5};

   New sec 2 = {0→0.5,1→0.5};

   Secret sec 1;

   Secret sec 2;

   New out =sec 1 Xor sec 2;

   Observe out;

8. New result =0;
9. New i=0; new sec=0
10. While (i<4){
11. Observe result;
12. Sec={1→0.0625,…16→0.0625};
13. Secret sec;
14. If (i==2){
15. Result= sec;
16. }
17. I=i+1;
18. }

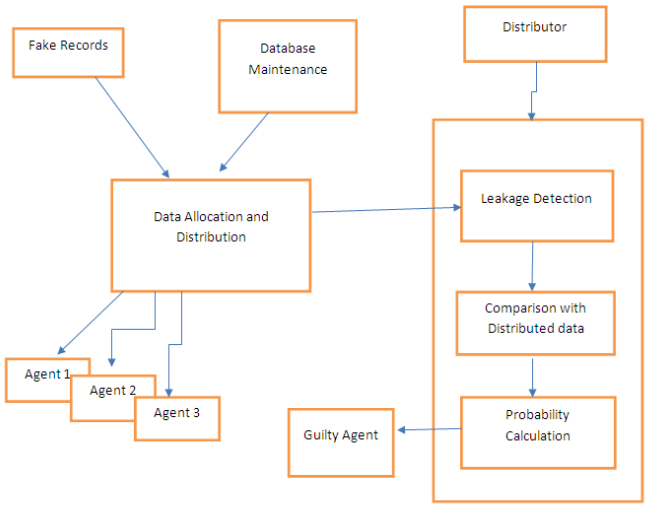

Fig.1 System Architecture

IX. DESIGN DETAILS

Fig.2 Home page of Information leakage detection and prevention

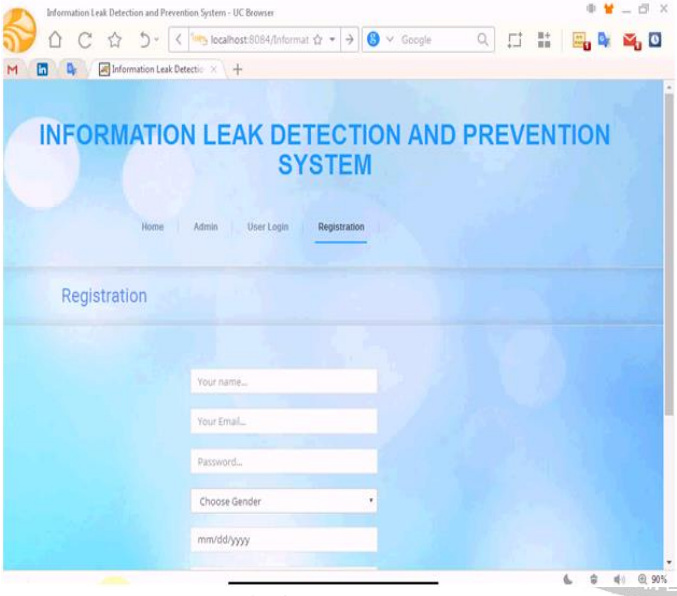

Fig.3 Registration page 1 of Information leakage detection and prevention

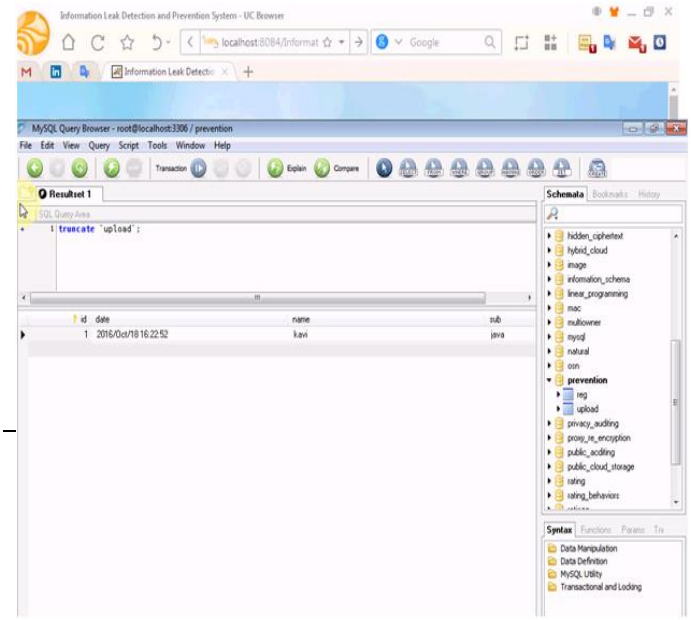

Fig.4 Registration page 2 of Information leakage detection and prevention

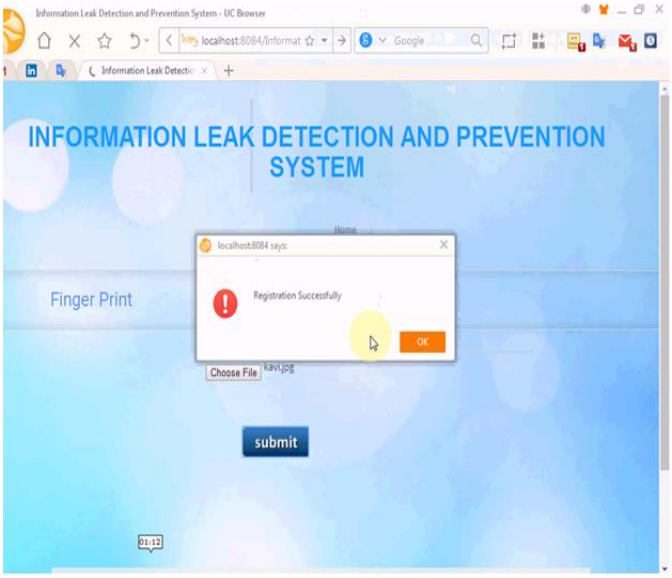

Fig.5 Registration page 3 of Information leakage detection and prevention

X. CONCLUSION

Thus we've got tried to implement the paper Tom Chothia", Yusuke Kawamoto", Chris Novakovic" and David Parker" "Probabilistic Point-to-point data Leakage", IEEE 2013.and Lin-shung huang", Alex Rice", Erling Ellingsen" and Collin Jackson" "Analyzing solid SSL Certificates within the Wild", IEEE 2014 and from implementation we have a tendency to got the conclusion as , It will discover and forestall the information from the leak by exploitation some algorithms and techniques. in an exceedingly excellent world there would be no ought to turn in sensitive information to agents which will inadvertently or maliciously leak it. And notwithstanding it had handy over sensitive information, it may watermark every object in order that it may trace its origins with absolute certainty. However, in several cases it should so work with agents which will not be 100 percent sure, Associate in Nursingd it's going to not be sure if a leaked object came from an agent or from another supply, since sure information cannot admit watermarks. In spite of those difficulties, it's shown it's attainable to assess the probability that Associate in Nursing agent is to blame for a leak, supported the overlap of his information with the leaked information and also the information of different agents, and supported the likelihood that objects are often "guessed" by different means that. This model is comparatively straightforward, however it captures the essential trade-offs. The algorithms have given implement a spread of knowledge distribution ways that may improve the distributor's probabilities of distinctive a source. The distributing objects judiciously will build a major distinction in distinctive guilty agents, particularly in cases wherever there's massive overlap within the information that agents should receive. the longer term work includes the investigation of agent guilt models that capture run eventualities that don't seem to be studied during this paper. A preliminary discussion of such a model is obtainable in another open drawback is that the extension of the allocation ways in order that it will handle agent requests

in an internet fashion (the given ways assume that there's a hard and fast set of agents with requests proverbial in advance). it's been given a framework that may be wont to live data leaks between whimsical points in an exceedingly program. To do so, it's been introduced CH-IMP, a language that enables variables' values to be Associate in Nursing notated as either secret or evident by an offender, which provides a mechanism for quantifying data leaks from secret to evident values. This model is comparatively straightforward, however it believes that it captures the essential trade-offs. The algorithms that have given implementing a spread of knowledge distribution ways that may improve the distributor's probabilities of distinctive a source. It are shown that distributing objects judiciously will build a major distinction in distinctive guilty agents, particularly in cases wherever there's massive overlap within the information that agents should receive. This future work includes the investigation of agent guilt models that capture run eventualities. the information run Detection Model provides security to the information throughout its distribution or transmission method

REFERENCES

[1] http://www.cs.bham.ac.uk/research/projects/infotools

[2] D. Clark, S. Hunt, and P. Malacaria, "*Quantified interference for a whileLanguage*,"Electron. Notes Theory. Computer. Sci., vol. 112, pp. 149–166, 2005

[3] A. Almeida Matos and G. Boudol, "*On declassification and the nondisclosure policy*," Journal of Computer Security, vol. 17, no. 5, pp.549–597, 2009

[4] K. Chatzikokolakis, C. Palamidessi, and P. Panangaden, "*AnonymityProtocols as Noisy Channels*,"Information and Computation, vol. 206,no. 2–4, pp. 378–401, 2008.

[5] F. Biondi, A. Legay, P. Malacaria, and A. Wasowski, "*Quantifying Information leakage of randomized protocols*," in Proc. VMCAI'13,2013, pp. 68–87.

[6] A. McIver and C. Morgan, "*Programming methodology*," A. McIverand C. Morgan, Eds. New York, NY, USA: Springer-Verlag New York,Inc., 2003, ch. A probabilistic approach to information hiding, pp.441–460.[Online]. Available: http://dl.acm.org/citation.cfm?id=766951.766972

[7] A. O. Freier, P. Karlton, and P. C. Kocher, "*The Secure Sockets Layer(SSL) Protocol Version 3.0*," RFC 6101 (Historic), Internet EngineeringTask Force, Aug. 2011

[8] Electronic Frontier Foundation, "*The EFF SSL Observatory*," https://www.eff.org/observatory.

[9] VASCO, "*DigiNotar reports security incident*," http://www.vasco.com/company/about vasco/press room/news archive/2011/news diginotarReports security incident.aspx, Aug. 2011.

[10] TURKTRUST, "*Public announcements*," http://turktrust.com.tr/en/ kamuoyu-aciklamasi-en.html, Jan. 2013.

[11] S. E. Schechter, R. Dhamija, A. Ozment, and I. Fischer, "*The emperor's new security indicators*," in Proceedings of the IEEE Symposium on Security and Privacy, 2007

[12] *A Role-Based Trusted Network Provides Pervasive Security and Compliance* - interview with Jayshree Ullal, senior VP of Cisco

[13] Dave Dittrich, *Network monitoring/Intrusion Detection Systems (IDS)*, University of Washington.

[14] *"Dark Reading: Automating Breach Detection For The Way Security Professionals Think"*. October 1, 2015.

[15]"'*Honeypots, Honeynets*'". Honeypots.net. 2007-05-26.

[16]Wright, Joe; Jim Harmening (2009) "15" *Computer and Information Security* Handbook Morgan Kaufmann Publications Elsevier Inc p. 257

