Bayesian Network Application on Information Security

Isai Rojas González, José Omar García Gallardo

Instituto de Investigaciones Eléctricas, Gerencia de Tecnologías de la Información,
Av. Reforma 113, 62490 Cuernavaca, Morelos, México.
{irojas,gallardo}@iie.org.mx

Abstract. This article presents a bibliographic analysis of main applications of Bayesian networks in terms of security. We described an overview of the use of artificial intelligence techniques to create mechanisms that protect information, emphasized in the use of Bayesian logic as the basis for the development of Intrusion Detection Systems (IDS) and predict threats and security risks. We show an oriented probabilistic model of computer security, which had has as main objectives to detect anomalies and to provide predictive information attacks.

Keywords: Bayesian network, computer security, treats predictive, intrusion detection.

1 Introduction

Nowadays, computer systems are of great importance in our lives, one way or another We interact with them and we use data that they provided to us. These systems interconnected through global networks, opening doors for access to the information contained therein. Data is critical, secrets and / or necessary, when this kind of information stolen by an attack, or used improperly, which can cause damage ranging from minor alterations innocuous information to real disaster of global proportions. Therefore, becomes crucial protect the information.

There are several security mechanisms that have developed along the evolution of information, however, computer systems are larger, and complex, it requires to implement more advanced security techniques. Artificial intelligence is a science that emulates human intelligence and learning based on use of computers, this science offers several methods of processing complex information using techniques that developed on robust computer security mechanisms.

Bayesian networks are an artificial intelligence technique, used to determine the probability of occurrence of an event; the determination made through the analysis of relationships between different variables, which influence the occurrence of the event. According to [1], Bayesian networks are "graphical structures to represent the probability relations between a large number of variables and doing probabilistic inference with these variables."

Special Issue in Advances in Artificial Intelligence and Applications
Research in Computing Science 51, 2010, pp. 87-96

Computer security Bayesian networks have many different applications, the most commonly used as part of expert systems are focused to detect intruders in computer systems, are also widely used to determine the likelihood of threats or attacks on computer systems.

2 Conceptual Framework

2.1 Information Security

Computer security known as the set of techniques, policies, and controls that allow adequately protect data information systems. We protect information through virtual computer programs such as antivirus, antispyware, firewalls, intrusion detection system (IDS), systems analysis, and prediction of seizures, among others. Information systems often exposed to risks and threats that may affect its availability, reliability, and integrity. The origin of threat usually classified as accidental or natural or provenance deliberate or intentional, in both cases can manifest as disruptions in the continuity of service, interception, modification, and / or data generation. The Bayesian network techniques are use to create computer programs that offer security through classification and prediction of future events using Bayesian logic.

2.2 Artificial Intelligence

Artificial intelligence and intelligent behavior have as main goals produce useful and intelligent machines. We could summarize as the principal goal of artificial intelligence in the "study of the process of thought and intelligent behavior of humans and produce machines and systems that represent these processes" [2]. Then, represent Artificial intelligence into various techniques, each simulating a different section of the reasoning process:

- Fuzzy logic.- Form of multi-valued logic derived from fuzzy set theory to deal with reasoning that is approximate rather than precise and accept values as "certainly true" and "half truly"
- Genetic algorithm.- Heuristic search that mimics the process of natural evolution and genetic raised by J. Holland at 70's. This heuristic routinely used to generate useful solutions to optimization and search problems.
- Expert systems.- Is software that attempts to provide an answer to a problem, or clarify uncertainties where normally one or more human experts would need to be consulted, Expert systems simulate the behavior of a human expert learning, memorizing, reasoning and communicating in a specific knowledge domain used to support take decisions.
- Intelligent Agent. - Autonomous entity, which observes and acts upon an environment and directs its activity towards achieving goals, intelligent agents may also learn or use knowledge to achieve their goals.

- Network neural. – The term used to refer to a network or circuit of biological neurons and interconnect between them; each neuron has an established specific value and way out functions, those factors determine behavior of neural network.
- Bayesian networks. – It's a model based on fundamentals described in 1763 by Thomas Bayes, based on conditional probabilities and statistics to predict event future.

In [2] details on these and other artificial intelligence techniques applied to computer security issues.

2.3 Bayes Theorem

Bayesian logic is basis on Thomas Bayes' theorem, so it is important to understand it's operation and implementation. Probabilistic models that aim to predict future events based on known statistics and management of uncertainty variables:

Given two variables X and Y, such that P(x) > 0 for all x P(y) > 0 for all met:

$$P(x_i|y) = \frac{P(x_i) * P(y|x_i)}{\sum_{j=1}^{k}[P(x_j) * P(y|x_j)]} \quad (1)$$

Where k: total number elements x

"In practice, it is used to determine the posterior probability of some variable of interest given a set of findings [4].

2.3.1 Example

Three machines, A, B, and C, producing 800 beach balls, the production is distributed as follows:

A: 300, 30% reds, 30%blues, 15 % blacks and 25 % whites.
B: 360, 25% reds, 18%blues, 20 % blacks and 37 % whites.
C: 140, 10% reds, 15%blues, 40 % blacks and 35 % whites.

All the balls discharged into a common container without any ranking. Taking a ball at random from the container:

a) ¿Which is the probability that have made for machine C?
b) ¿Which is the probability that besides the ball will be white?

Solution:

a) Probability that: $P(x_i) = P(C) = \frac{140}{800} = 0.175 = \mathbf{17.5\%}$ P
b) Order the data to has a better vision:

Table 1. Occurrences probabilities distribution

P(y|x)	A (x_1)	B (x_2)	C (x_3)
Red (y_r)	0.30	0.25	0.10
Blue (y_a)	0.30	0.18	0.15
Black (y_n)	0.15	0.20	0.40
White (y_b)	0.25	0.37	0.35

Applying Bayes' Theorem has that:

$$P(x_i|y) = P(x_a|y_b) = \frac{P(x_a) * P(y_b|x_a)}{\sum_{j=1}^{a}[P(x_j) * P(y_b|x_j)]}$$

$$P(x_a|y_b) = \frac{0.175 * 0.35}{(0.375 * 0.25) + (0.45 * 0.37) - (0.175 * 0.35)}$$

$$P(x_a|y_b) = \frac{0.06125}{0.09375 + 0.1665 + 0.06125} = \frac{0.06125}{0.3215} = 0.1905 = \mathbf{19.05\%}$$

2.4 Security Ontology

In philosophy, ontology is the study of the nature of being, existence, or reality in general, as well as the basic categories of being and their relations. For example the relationship that exists in a universe (red), and an individual (the apple), or the relationship between an event and its participants.

In computing, the term adopted to refer to the formulation of a detailed and rigorous conceptual schema within one or multiple domains, in order to facilitate communication and interchange of information between different systems and entities.

Today, the concept of computing is widely used ontology in artificial intelligence. In some applications, several schemes are combining in a full facto structure of data containing all relevant entities and their relationships within the domain, under this approach, software used for purposes such as inductive reasoning, data classification, and various problem-solving techniques.

The ontology information is use to feed data to patterns of Bayesian networks. The data provided deducted under the assumptions of ontology, analysis of the participating entities and their relations of domination.

3 Bayesian network applications

Bayesian networks use Bayesian logic to represent the model as directed acyclic graphs. Bayesian classification is a technique unsupervised data classification [2]. The prediction methods Bayesian networks are widely used in computer security

issues, for example, to filter spam e-mails messages, are often used Bayesian networks to determine whether the words contained in the message have a high probability of belonging to a type email spam.
In computer security issues is crucial to anticipate risks and threats to computer systems. Bayesian networks are a powerful tool to determine the probability of attack on information systems. The analysis of these variables is essential for decision-making and adjusts of safety measures.
In general, preventive measures are vital, they seek to anticipate an unwanted event to prevent this happen; however, we must remember that there is no foolproof security as there are sometimes inappropriate or malicious intrusions on information systems. When an intrusion occurs, our security scheme should be able to identify and correct it. In this manner, Bayesian networks are very important to analyze the situation and establish the likelihood of an atypical situation that indicates a possible intrusion.

3.1 Prediction of treats and risks

The National Institute of Standard and Technology (NIST) defines risk management information security as the process that allow to the CIO (Chief Information Officer) balancing operational and economic costs of implementing protection measures and the advantage to have protected information systems and data they contained, which contribute to the mission of the organization [5].
Managing security risks is priority to have identified the risk involved. Understanding as a risk, the probability of an unwanted event occurs affecting the suitable functioning of information system and cause damage to the organization.
The risks that menace the computer systems can be frame in unauthorized access, disclosure of important information, denial of service, loss of resources, vandalism, and sabotage. Threats mainly affect the hardware, software, and data.
Therefore, these phenomena are due to interruption, interception, modification, and generation. The threats in computer security issues are situations that may affect the suitable functioning of an information system, causing inherent risks consequently.
Bayesian networks are widely used to establish and determine the levels of risk information systems. The method used to model the variables identified as important aspects of the process you want to ensure, should consider factors that correlated and infer the probability of risk.
The software can use the point of view of ontology for a variety of purposes, including inductive reasoning, classification, and a variety of problem-solving techniques [6].
In [5] proposes a model to determine the likelihood of threats where T is considered as a set of variables (T1, T2, ..., Tn) and each of them represents a threat whose probability of occurrence should be determined. The proposal is premised on each threat has exactly one value from a finite set of possible values. The figure below shows the proposed model.

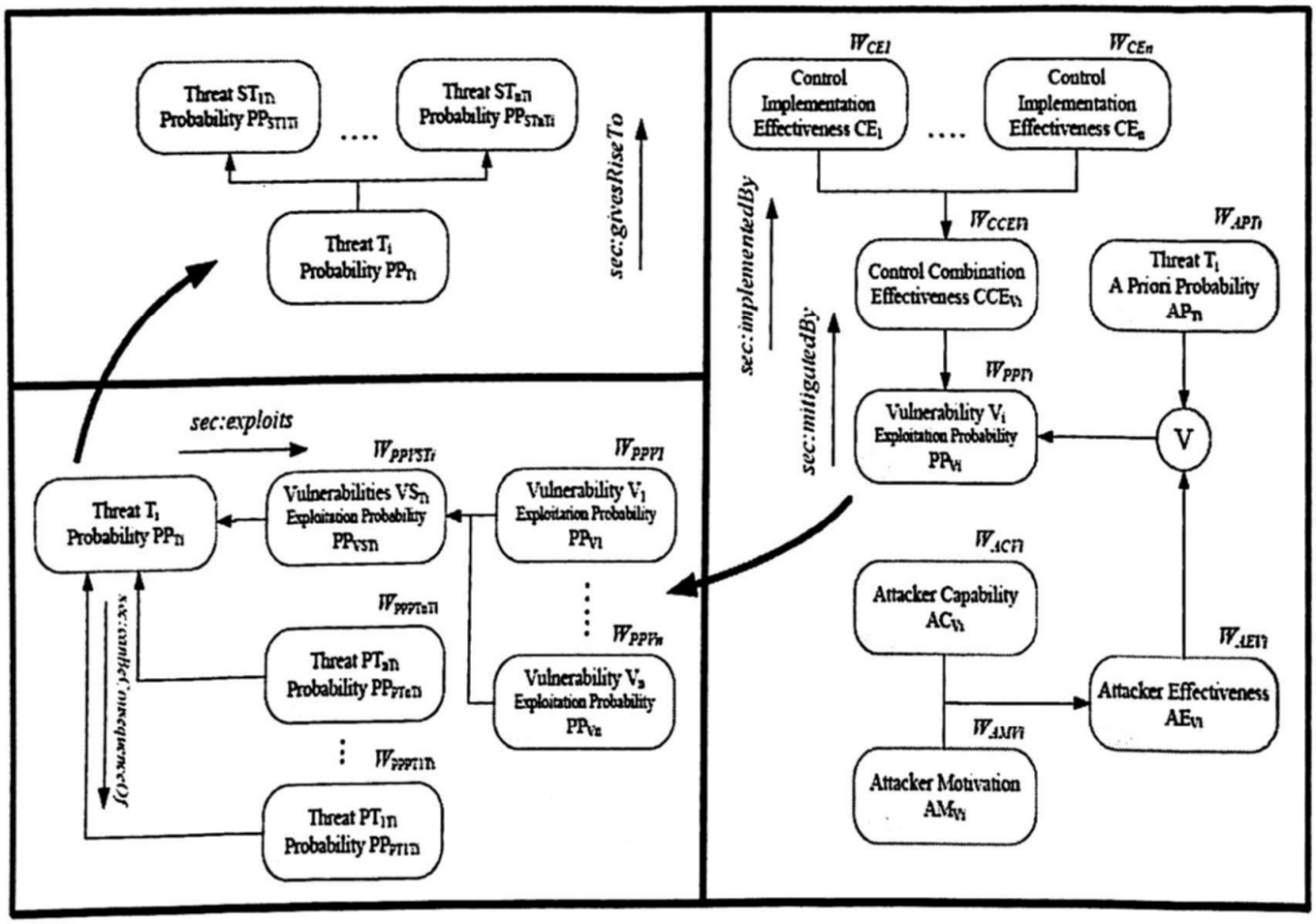

***Figure 1** This diagram utilizes security ontology for the Bayesian threat probability determination.*

The probability of a threat and the influential factors are not determined quantitatively, therefore using a qualitative rating scale, for example, high probability, medium and low. This allows a better understanding human language, providing precise information on model inputs, and obtained results are easier to interpret. Determine the probability of the threat should consider various factors of influence, including the following that have been identified: threats predecessors (PT ¬ 1Ti, ..., PTnTi) that influence the threat in question (Ti) which turn influences threats successors (ST1Ti, ..., STnTi). Each threat (Ti) should include one or more vulnerabilities to become effective, concluding that the probability of the threat has seen significantly affected by the existence of the mentioned vulnerabilities. You can use security controls to mitigate identified vulnerabilities, the degree of mitigation depends on the effectiveness of the possible combination of security controls (CCEVi) which depend on the effectiveness of security controls that are part of this combination (CE1, ..., CEN). We must consider also the threat of deliberate origin, when exist, the likelihood of exploitation of vulnerability is determined by the effectiveness of a potential attacker and this in turn is determined by motivation and capabilities of the attacker. In cases of threats of accidental or natural origin, the probability of exploitation of vulnerability (LVPP) is determined by the prior probability (APTI) of the threat (Ti) corresponding.

Figure 1 shows a proposed model to determine the likelihood of threats using Bayesian networks in combination with ontology in computer security. It's note that the risk managers only need to classify the nodes of motivation (GAVI) and capacities (ischemic stroke) of the attacker. The rest of the inputs and intermediate nodes are deriving by security ontology. The results of each extinction risk interpreted as a distribution of values chosen as a rating scale, for example, high probability, medium, and low.

The concept of *sec: Probability* and *property sec: security* ontology probability distribution is relations that connect each threat of a particular physical location with its prior probability. The specific gravities of all threats equally distributed according to threats and vulnerabilities influential.

The advantage of Bayesian determination of the probability of threat is that it provides the manager of risk management, a structured methodology to determine the likelihood of the threat, adding security ontology [6].

3.2 Intrusion detection

Computer systems are susceptible to attacks and security breaches, this happens despite preventive safety measures have been implemented. This kind of situation is imperative to detect the anomaly as soon as possible, to make good decision. The intrusion detection systems (Intrusion Detection System, IDS) are widely used to address security threats in computer networks. We could define Intrusion detection as the process to identify malicious behavior that threatens the computer network and its resources [7].

There are various types of intrusion detection systems, some are based on "misuse" (Misuse-based), this type of system uses a list of alerts of attacks, the descriptions are tested against data obtained from audits and monitoring systems, what is done in the comparison is to look for evidence that matches any of the attacks previously cataloged. The disadvantage of this method is that it can detect unknown attacks, because the detection based on documented information known attacks.

Another type of IDS's are those based on anomaly detection, this kind of initial information systems use is a reference to determine the normal behavior of users, network traffic, and applications. When there is a deviation in the behavior specified then the event should be classify as an anomalous situation, however, remains to be determined whether this really an attack is.

Systems based on anomaly detection, have the advantage of anticipating attacks by unknown information inference, on the other hand, a drawback of this technique is that it usually has a high rate of false positives, for example: situations identified as intrusions unusual but completely legitimate.

In [7] proposes a model of intrusion detection that combines the techniques of systems based on misuse (Misuse-based) and those based on anomalies (anomaly-based) with the use of Bayesian networks and mechanisms learning. The model describes an operation in which there is an initial registration of known attacks and the events raised these patterns are compared against known,

complemented by the use of anomaly detection techniques, this would cover known attacks and attempts to prevent unknown attacks.

Bayesian networks used in the detection mechanism and classification of anomalies, the heuristic technique used to determine, with any degree of accuracy, the possible existence of a security violation. Together with an additional learning mechanism, analyzes the events that occur and the results of their classification, positive and false positives are stored in a dynamic framework that serves for future comparisons. All this reduces the rate of false positives and thus have a much more robust IDS and effective. This type of dynamic schema naturally becomes a work environment adaptive and evolutionary. The intrusion detection model proposed in [8] consists of six modules as shown in Figure 2. Data gathering module (sensors) and parsing, is responsible for collecting data and analyzing the monitored network connections. The connection understood as a session between two hosts on the network. The Bayesian Network Inference module is the IDS analysis engine responsible for processing the information gathered by the sensor. The knowledge base (Knowledge base) contains an intelligent model (Bayesian Model) which learns from observing the traffic and has the ability to predict when a network connection is an attack. The system configuration (System Configuration) provides information about the current state of IDS. The component of response (Response) triggers actions when an intrusion detected. The Bayesian Network Learning module used to generate knowledge from training data set that is not online.

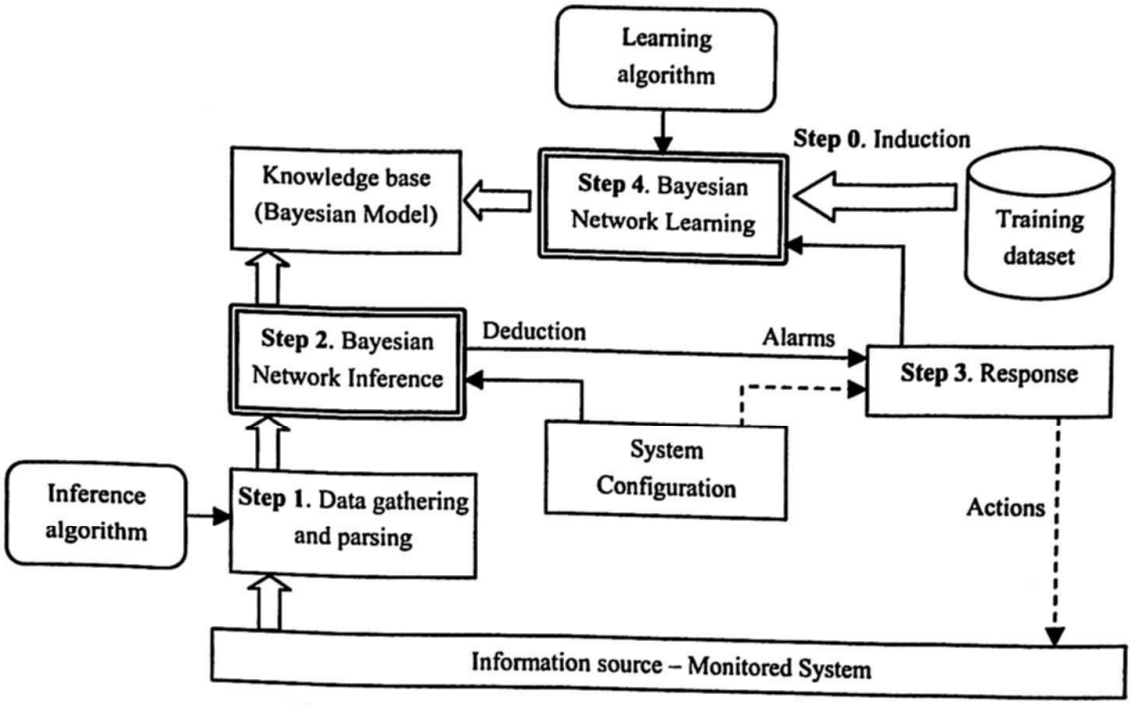

***Figure 2** Bayesian Network-based IDS architecture.*

4 Conclusions

Throughout this article, we noted that several artificial intelligence techniques used to implement powerful computer security mechanisms. These systems, which simulate human intelligence and reasoning, may be as large and complex as needed; This must determined by the balance cost - benefit to be made for each organization. Bayesian networks are a very valuable contribution of artificial intelligence, the method is widely used in various fields of knowledge [9] [10].

The application of Bayesian networks, security systems, provides mechanisms that are more robust and efficient to establish appropriate measures of protection of information. Based systems using Bayesian networks, are often highly effective and have a very low error rate, however, consider that the accuracy of such systems directly related to the clarity, precision and accuracy of data input, which used as initial variables and influence. The use of qualitative ratings in the range provides better management of both the input information, and the interpretation of results. Finally, we note that, due to the high dependence of the input values should be further research on more accurate methods for the collection, storage and delivery of the initial information required by the systems, the purpose is to ensure maximum accuracy and precision possible in the results of probability calculations that perform Bayesian networks.

References

[1] R. Neapolitan. Learning Bayesian networks. Prentice Hall,2003.

[2] Nestor Dario Duque Mendez, Julio Cesar Chavarro Porras, Ricardo Moreno Laverde. Seguridad Inteligente. Scientia et Technica Año XIII, No. 35. Universidad Tecnológica de Pereira. 2007.

[3] Diccionario Oxford Complutense de Física. Página 306. Editorial Complutense. 2007.

[4] (1) Álvaro Marín Illera. Sistemas Expertos, Redes Bayesianas y sus aplicaciones. Semana ESIDE, Universidad de Deusto. 2005.

[5] Stefan Fenz, Thomas Neubauer. How to determine threat probabilities using ontologies and Bayesian networks. Vienna University of Technology, Secure Bussiness Austria. 2009.

[6] Stefan Fenz, A Min Tjoa. Ontology-based generation of Bayesian networks. International Conference on Complex, Intelligent and Software Intensive Systems. Institute of Software Technology and Interactive Systems Vienna University of Technology & Secure Business Austria.

[7] M. Mehdi, S. Zair, A. Anou and M. Bansebti. A Bayesian Networks in Intrusion Detection Systems. Journal of Computer Science 3 (5): 259-265. Electronics Department, University of Blida, Algeria. 2007.

[8] Alma Cemerlic, Li Yang, Joseph M. Kizza. Network Intrusion Detection Based on Bayesian Networks. Department of Computer Science Engineering, University of Tennessee at Chattanooga. 2008.

[9] A. Stein, M.A.J.S. van Boekel and A.H.C. van Bruggen. Bayesian Statistics and Quality Modelling in the Agro-Food Production Chain. Cap. 9. 2004.

[10] Gustavo Arroyo-Figueroa, Luis Enrique Sucar. A Temporal Bayesian Network for Diagnosis and Prediction. Uncertainty in Artificial Intelligence. IIE – USP, ITESM – Campus Morelos. 2000.

[11] Krister Johansen, Stephen Lee. Network Security: Bayesian Networks Intrusion Detection (BNIDS). 2003.

[12] Salem Benferhat, Tayeb Kenaza, Philippe Leray. Data Mining and Detecting Complex Attacks. Universit´e d'Artois, rue Jean Souvraz, Ecole Militaire Polytechnique, Site de PolytechNantes, rue Christian Pauc. 2006.

[13] Xiangdong An and Dawn Jutla, Nick Cercone. Privacy Intrusion Detection Using Dynamic Bayesian Networks. Finance and Management Science Department, Saint Mary's University. Faculty of Computer Science Dalhousie University. 2006.

[14] Peng Xie, Jason H Li, Xinming Ou, Peng Liu, Renato Levy. Using Bayesian Networks for Cyber Security Analysis. Intelligent Automation Inc. Rockville, Kansas State University, Penn State University. 2010.

[15] Marcel Frigault, Lingyu Wang. Measuring Network Security Using Bayesian Network-Based Attack Graphs. Annual IEEE International Computer Software and Applications Conference. Concordia Institute for Information Systems Engineering. 2008.

[16] Marcel Frigault and Lingyu Wang, Anoop Singhal, Sushil Jajodia. Measuring Network Security Using Dynamic Bayesian Network. Concordia Institute for Information Systems Engineering Concordia University, Computer Security Division National Institute of Standards and Technology, Center for Secure Information Systems George Mason University. 2008.