DECISION MODEL ANALYSIS FOR SPAM

Agustin ORFILA, Javier CARBO, and Arturo RIBAGORDA

Abstract: One of the security challenges in e-Government is to offer a smooth dialogue with citizens, which guarantees the availability, confidentiality and integrity of the information interchanged. Spam jeopardizes the survival of electronic mail as a communication means. Many approaches to tackle the problem with spam have been proposed. This paper shows the necessity of studying the real value of spam filters. Contrary to common belief, false positive rate and false negative rate do not completely reveal to what extent a junk filter is worth using. Very important parameters like the hostility of the environment (summarized by the probability of receiving spam) or the error costs associated with the filter play a decisive role.

Keywords: Network Security, Spam Detection, Decision Analysis, Blacklists, False Positives, False Negatives.

There is no single definition of the term spam.[1,2] The American Heritage Dictionary of the English Language[3] defines it as: "unsolicited e-mail, often of a commercial nature, sent indiscriminately to multiple mailing lists, individuals, or newsgroups." The first problem in identifying spam is that people do not agree on the meaning of the word spam. The barrier between legitimate and illegitimate email is not always clear and depends, in certain cases, on the recipient of the message. For instance, an employee could receive an email from the trade unions in order to sign some proposal. Although this is an unsolicited message, it can be considered legitimate (and even important) for the employee. In fact, Bayesian filters make use of this idea to classify email taking into account the spam concept of the user. However, most users would agree that messages about commercial items sent out in bulk should be filtered.

According to Lambert[4], MSN Hotmail stated in an e-mail sent to all its users in May 2003 that they were blocking more than 2 billion spam e-mails every day. Ferris Research reported that in 2003 the corporate user would waste 15 hours deleting e-mail, compared to 2.2 hours in the year 2000. Gartner Inc. reported that in 2002, 25% of all e-mails were qualified as spam. Brigthmail Inc.[5] points out that in April 2004 64% of total internet email was spam and that these figures are increasing month by month.

There are different techniques to detect spam. Combination of blacklists and whitelists, header analysis, content analysis, Bayesian filters, challenge filters or honeypots.[6,7,8] This article attempts to show the necessity of analyzing the value of spam filters rigorously. Some open questions studied in the paper are the influence of the spam frequency on the value of these tools and how to measure to what extent these filters are worth deploying.

The remainder of the article is organized as follows. The next section reviews related work. The decision model proposed for solving the problem with spam is described after that. Results based on the proposed model and main conclusions finalize the article.

Related work

The related work that tries to evaluate and compare proposals about spam filters uses parameters such as false positive rate and false negative rate. Blacklists[9] and Bayesian filters could be mentioned as examples.[10,11] The authors of this publication only know about an attempt to include the error costs of the filter in the analysis of its performance.[12] Androutsopoulos and co-workers define a quantity called True Cost Ratio (TCR) as a measure of how much time is wasted to delete manually all spam messages when no filter is used, compared to the time wasted to delete manually any spam messages that passed the filter plus the time needed to recover from mistakenly blocked legitimate messages. However they do not take into account the probability of receiving spam in their analysis.

The Receiver Operating Characteristic (ROC) of a detector[13] is a plot of detection probability versus false alarm probability. ROC curves have been extensively used in the characterization of communication systems and radar systems,[14] and in sciences such as Psychology[15] or Meteorology.[16] In all these fields there is a need of decision making under uncertainty in order to classify a data set. Depending on the ROC curves, ROC analysis can be incomplete because it does not take into account neither the error costs nor the hostility of the environment. In Computer Science field, a decision model analysis has been proposed for Intrusion Detection Systems (IDS) establishing that the best operating point for an IDS inherently depends on the error costs and the probability of having an intrusion.[17] They have also demonstrated that for the comparison of IDS' effectiveness these parameters should be taken into account.

Decision Model Analysis for Spam

Independently of its internal nature, we can model a spam filter as a detector. Each email that arrives to the inbox can be considered to be in one of these two states: Spam (S) or Non spam (NS). A spam detector can classify a message as Legitimate

(L) or Illegitimate (I). The main parameters of this binary detector are: the probability of classifying as illegitimate a spam email $P(I \mid S)$, also called hit rate (H), and the probability of classifying as illegitimate a non spam message $P(I \mid NS)$, also known as false positive rate (F).

Let us consider the error costs of misclassifying an email. If the message is spam and the detector does not classify it as illegitimate, there is an associated cost (L) for the recipient that consists of recognizing it and deleting it. Otherwise, if the detector classifies a non spam message as spam then it causes a cost C for the recipient that depends on the kind of filter's response. In general, these costs are asymmetric being the concrete figures related to the consequences of each error. In general, C is greater than L since losing a non spam mail is usually worse than not to filter a spam one. Most email clients with filtering capabilities do not just wipe the messages marked as spam but put them into a trash folder for later deletion. So the user would be able to rescue a badly classified email. C directly depends on the action taken over the filtered mail and the user awareness of the limitations of the current technology. If the trash folder is never consulted each false positive may cause great losses. The costs associated with the response made based on the detector's classification are given in the decision-model contingency matrix shown in Table 1.

The decision model that this article presents combines the detector's report, the response taken, the real email condition and the consequences in order to make the best decision. The best decision should be the one that minimizes expenses.

Figure 1 shows the decision tree with the sequence of actions (squares) and uncertain events (circles) that describes the detector's operation and the response actions (mark the email, move it to a spam folder, erase it, etc.) that can be taken. The costs shown correspond to the consequences of the action taken.

Decision nodes or action nodes (squares) are under the control of the decision system, which will choose which branch to follow. Event nodes (circles) are not under the control of the decision system, but depend on uncertainty. A probability distribution represents the uncertainty about which branch will follow an event node.

Table 1: Decision-Model Contingency Matrix. If a spam email comes and it is not filtered then a cost L is incurred. Otherwise, if a non spam message arrives and the detector filters it then a loss C is incurred.

		Occurs	
		Non spam	spam
Take action	No	0	L
	Yes	C	0

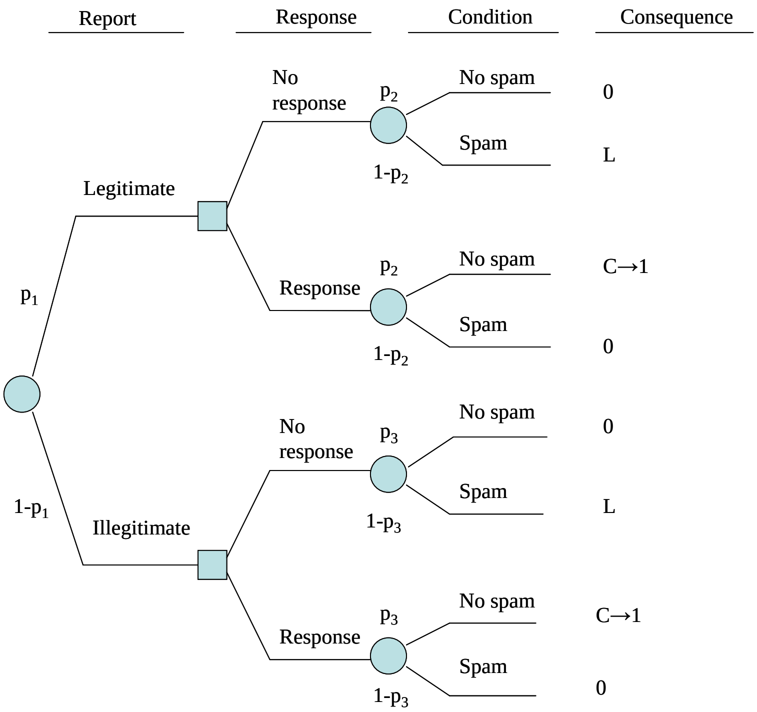

Figure 1: Decision Tree for the Spam Detector. The squares represent the possible actions that can be taken by the decision-making system and the circles describe the uncertain events that can take place.

Each combination of actions and events is characterized by its cost. There is a probability of occurrence associated to each uncertain event. There are three probabilities specified in the tree:

- p_1 : the probability that the detector's report is "legitimate"
- p_2 : the conditional probability of occurring non spam given that the detector's report is "legitimate"
- p_3 : the conditional probability of occurring non spam given that the detector's report is "illegitimate."

The last two probabilities account for both filter errors – falsely reporting that a message is spam (p_3) and falsely reporting that an email is not spam ($1-p_2$).

The expected cost is determined for event nodes by taking the sum of products of probabilities and costs for all of the node's branches. The expected cost at a decision node is the lowest expected cost from among the node's branches. The process is repeated until expected values are determined for all nodes.

Let F and H be the false alarm rate and the hit rate, respectively. An operation point is defined by a pair (F , H). Analyzing Figure 1, it is possible to determine the expected cost associated with an operating point and a specification of the best response decision to make conditional on the detector's report. Let us represent by p the prior probability that a received email is spam. The probabilities of the filter's report are calculated by applying the formula of total probability:

$$\begin{aligned} p_1 &= P(L) = P(L \mid NS)P(NS) + P(L \mid S)P(S) = (1-F)(1-p) + (1-H)p \\ 1-p_1 &= P(I) = P(I \mid NS)P(NS) + P(I \mid S)P(S) = F(1-p) + Hp \end{aligned} \tag{1}$$

The probabilities of the system's state conditional on the detector's report are calculated by applying Bayes' Theorem[18]:

$$\begin{aligned} p_2 &= P(NS \mid L) = P(L \mid NS)P(NS)/P(L) = (1-F)(1-p)/p_1 \\ p_2 &= (1-F)(1-p)/((1-F)(1-p) + (1-H)p) \end{aligned} \tag{2}$$

$$\begin{aligned} 1-p_2 &= P(S \mid L) = P(L \mid S)P(S)/P(L) = (1-H)p/p_1 \\ 1-p_2 &= (1-H)p/((1-F)(1-p) + (1-h)p) \end{aligned} \tag{3}$$

$$\begin{aligned} p_3 &= P(NS \mid I) = P(I \mid NS)P(NS)/P(I) = F(1-p)/(1-p_1) \\ p_3 &= F(1-p)/(F(1-p) + Hp) \end{aligned} \tag{4}$$

$$\begin{aligned} 1-p_3 &= P(S \mid I) = P(I \mid S)P(S)/P(I) = Hp/(1-p_1) \\ 1-p_3 &= Hp/(F(1-p) + Hp) \end{aligned} \tag{5}$$

The expected cost of each response conditional on the detector's response is calculated by taking the sum of the products of the probabilities and costs for the node following the response. The results of the expected costs are shown in Table 2.

Table 2: Expected Costs of Responses Conditional on the Detector's Report.

Detector's report	No response	Response
Legitimate	$L(1-p_2) = L(1-H)p/p_1$	$Cp_2 = C(1-F)(1-p)/p_1$
Illegitimate	$L(1-p_3) = LHp/(1-p_1)$	$Cp_3 = CF(1-p)/(1-p_1)$

The expected cost given the detector's report is the expected cost of the least costly response given the report. So the expected cost given legitimate report from the detector is:

$$Min\{L(1-H)p, C(1-F)(1-p)\}/p_1. \tag{6}$$

Similarly, the expected cost given a report showing illegitimate behaviour is:

$$Min\{LHp, CF(1-p)\}/1-p_1. \tag{7}$$

The expected cost M of operating at a given point (F, H) is the sum of the products of probabilities of the detector's report and the expected costs conditional on the reports. The expected cost of operating at an operating point is:

$$\begin{aligned} M &= p_1 Min\{L(1-H)p, C(1-F)(1-p)\}/p_1 + \\ &+ (1-p_1)Min\{LHp, CF(1-p)\}/(1-p_1) \\ M &= Min\{L(1-H)p, C(1-F)(1-p)\} + Min\{LHp, CF(1-p)\} \end{aligned} \tag{8}$$

Without loss of generality we can rescale costs by defining the cost ratio $L' = L/C$. This substitution results in costs of 1 and L as shown in Figure 1. With this substitution

$$M = Min\{L(1-H)p, (1-F)(1-p)\} + Min\{LHp, F(1-p)\} \tag{9}$$

It is important to mention that this formulation includes the possibility that a decision is made to take action or not regardless of the detector's report. If the expected cost of going against the detector's report is lower than following its indication, the decision maker will decide contrary to the detector. This makes this model stronger than other approaches to decision analysis.[19,20]

For a perfect deterministic forecast $H = 1$, $F = 0$, hence

$$M_{per} = 0. \tag{10}$$

A detector that works based on the knowledge of the spam probability would base its decision on never protecting ($H = 0$ and $F = 0$) or always protecting ($H = 1$ and $F = 1$), which will incur an expected cost of

$$M_{fre} = Min\{Lp, (1-p)\}. \tag{11}$$

We define the value of a spam detector as a measure of the reduction in M over M_{fre}, normalized by the maximum possible reduction associated with a perfect deterministic forecast, i.e.

$$V = (M_{fre} - M)/(M_{fre} - M_{per}). \qquad (12)$$

For a detection system which is no better than the one based on the probability of having spam, $V = 0$; for a perfect detector $V = 1$.

This metric is very useful because it includes all the relevant parameters involved in the evaluation of spam detector effectiveness. The decision maker has to respond in order to maximize V .

Results

The proposed decision model analysis for spam detectors can help us to understand the consequences of deploying spam detection technology in any particular scenario. The following paradigmatic examples try to show the inherent relationship between the spam detector effectiveness, the error costs, and the probability of receiving spam.

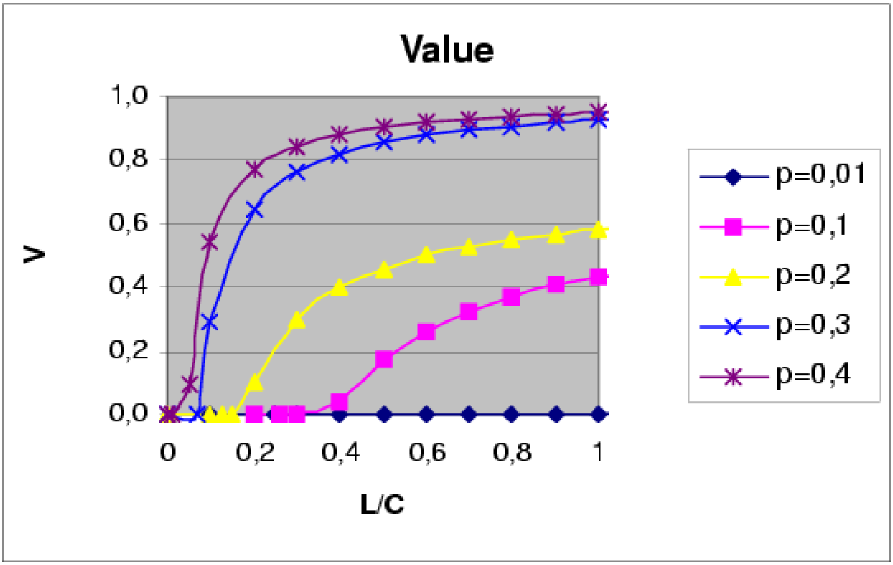

Figure 2: Decision Maker Value as a Function of L/C Relationship. Different curves are plotted for the different probabilities of receiving spam shown at the right part of the figure.

First, let us consider the case where $H = 0.995$ and $F = 0.03$ achieved by a Bayesian filter proposed by Graham.[21] Let us vary the probability of receiving spam. We focus on the cases where $L < C$ because it is more realistic to think that losing non spam email is worse than not to filter spam messages. Results are shown in Figure 2.

The spam filter is more valuable as the probability of receiving spam increases. The range L/C where it has some value is also wider as this probability increases. The value is greater as L becomes closer to C.

It is important to note that for a probability under 1% the proposed filter is worthless. In such a case it is better to base the decision just on the frequency of receiving spam. Furthermore, in the case study, if the error cost of filtering a non spam email is 5 times the cost of not to filter a spam one, then for probabilities of receiving spam under 10% the spam detector would be worthless. But for instance, for $p = 0.4$ the detector would have a 77% of the value of the perfect detector.

Let us now analyze the situation where no false positives are present. Results are shown in Figure 3 and we can see that, in the range of study ($L < C$), the value of the detector is numerically the same with the hit rate H . So, if the cost of losing an email is greater or equal than not filtering a spam one, then a detector that is able to tune it-

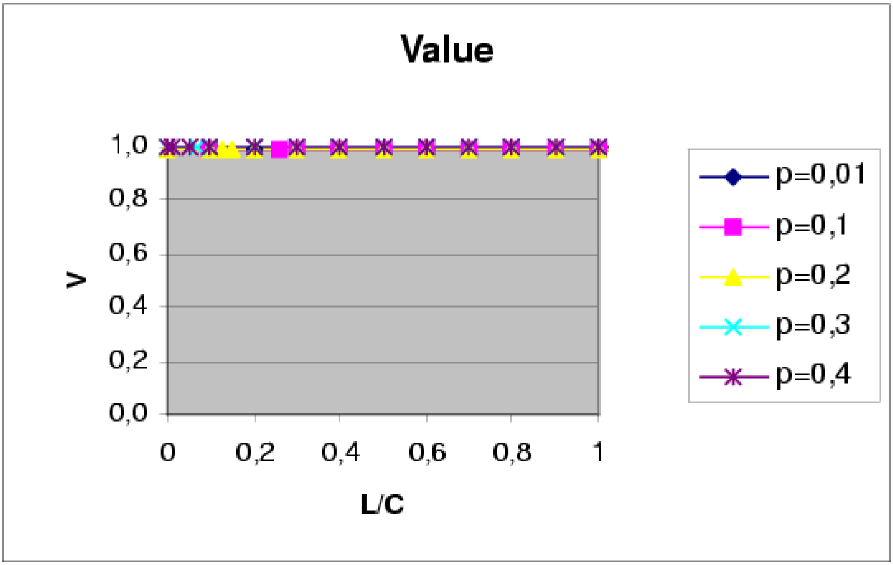

Figure 3: Decision maker value for an ideal detector that does not produce any false positives. The value is numerically equal to the hit rate for L/C range between 0 and 1.

self in order to avoid false positives is clearly superior. In theory, such a detector could exist nowadays. It is the one based on the combination of blacklists and whitelists. The whitelists always have priority over the blacklists, so a zero false positive rate could be achieved. False negatives come from addresses that are not yet in the blacklists or from spammers that have forged the sender address with one that is on the whitelist. The reasons why this is not the definite technology in real world are two. First, blacklisting a single mail address is not useful because sender addresses use to be forged but blacklisting an entire domain usually leads to a set of false positives. Second, there is an operational cost of upgrading both lists that is not considered in the presented decision model.

Conclusions

This paper proposes a new approach to the analysis of the effectiveness of spam filters. As far as the authors are aware, the parameters that have been used to analyze the different techniques have just been based on the number of false positives and false negatives. The inherent relationship between the error costs, the hostility of the environment and the filter capabilities demands a new look. Decision analysis has demonstrated very good performance in other fields and we have pointed out the advantages of applying it to spam detection. The simple decision model proposed has revealed how important would be to ignore false positives and to consider the inherent relationship between the probability of receiving spam and the value of a spam detector. As this probability increases the decision maker's opinion becomes more valuable for the usual range of error costs. The model also increases its value as the cost difference between losing a non spam email gets closer to not filtering an illegitimate one. Whatever spam filter is used and whatever the costs of misclassifying a message are, the decision maker gives a probabilistic measure of the decision value. This article has shown that in order to decide whether to deploy a solution it is not only important to softly increase the hit rate or decrease the false alarm rate but to analyze what the frequency of having spam is and to estimate the expected costs in the particular scenario. The decision model proposed would give a quantitative measure of how valuable a spam filter would become for a specific case.

Notes:

[1] *Definition of Spam*, <http://www.mail-abuse.com/spam_def.html> (26 November 2004).

[2] *Spam Defined*, <http://www.monkeys.com/spam-defined/> (26 November 2004).

[3] *The American Heritage Dictionary of the English Language*, Fourth Edition (Boston: The Houghton Mifflin Company, 2000), <http://www.bartleby.com/61/> (26 November 2004).

[4] Anselm Lambert, *Analysis of Spam*, M.Sc. Dissertation (Department of Computer Science, Trinity College Dublin, September 2003).

[5] *Spam Statistics* (Brightmail Inc., 2004), <http://www.brightlight.com/spamstats.html> (May 2004).

[6] Lambert, *Analysis of Spam*.

[7] Ion Androutsopoulos, Georgios Paliouras, Vangelis Karkaletsis, Georgios Sakkis, Constantine D. Spyropoulos, and Panagiotis Stamatopoulos, "Learning to Filter Spam E-Mail: A Comparison of a Naive Bayesian and a Memory-Based Approach," in *Proceedings of the workshop "Machine Learning and Textual Information Access*," 4th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD-2000), ed. H. Zaragoza, P. Gallinari and M. Rajman (Lyon, France, September 2000), 1-13.

[8] J. Davila, "La Nueva Plaga del Spam," *Newsletter SIC* 57 (November 2003).

[9] Lambert, *Analysis of Spam*.

[10] Patrick Pantel and Dekang Lin, "SpamCop: A Spam Classification & Organisation Program," in *Proceedings of AAAI-98 Workshop on Learning for Text Categorization* (Madison, Wisconsin, 1998), 95-98, <http://www.isi.edu/~pantel/Download/Papers/aaai98.pdf> (26 November 2004).

[11] Paul Graham, *A Plan for Spam* (August 2002), <http://www.paulgraham.com/spam.html> (26 November 2004).

[12] Androutsopoulos, Paliouras, Karkaletsis, Sakkis, Spyropoulos, and Stamatopoulos, "Learning to Filter Spam e-Mail: A Comparison of a Naive Bayesian and a Memory-Based Approach."

[13] Harry L. Van Trees, *Detection, Estimation, and Modulation Theory*, Part I: Detection, Estimation, and Linear Modulation Theory (John Wiley & Sons, Inc., 2001).

[14] James P. Egan, *Signal Detection Theory and ROC-Analysis*, Series in Cognitition and Perception (New York: Academic Press, 1975).

[15] John A. Swets, Robyn M. Dawes, and John Monahan, "Psychological Science Can Improve Diagnostic Decisions," *Psychological Science in the Public Interest* 1, no. 1 (May 2000): 1-26.

[16] Richard W. Katz and Allan H. Murphy, "Forecast Value: Prototype Decision-Making Models," in *Economic Value of Weather and Climate Forecasts*, ed. R.W. Katz and A.H. Murphy (Cambridge, UK: Cambridge University Press, 1997), 183–217.

[17] Jacob W. Ulvila and John E. Gaffney, Jr., "A Decision Analysis Method for Evaluating Computer Intrusion Detection Systems," *Decision Analysis* 1, no. 1 (March 2004): 35-50.

[18] Morris H. DeGroot, *Optimal Statistical Decisions* (New York, NY: McGraw-Hill Book Co., 1970).

[19] John C. Hancock and Paul A. Wintz, *Signal Detection Theory* (New York: McGraw-Hill Book Co., 1966).

[20] Alvin Martin, Mark Przybocky, George Doddington, and Douglas Reynolds, "The NIST Speaker Recognition Evaluation– Overview, Methodology, Systems, Results, Perspectives," *Speech Communications* 31 (2000), 225-254.

[21] Paul Graham, *Better Bayesian Filtering* (January 2003), <http://www.paulgraham.com/better.html> (26 November 2004).

AGUSTIN ORFILA is currently junior lecturer at the Computer Science Department of the Carlos III University of Madrid, Spain. He is a member of the Information Security Group of this department. He obtained a degree in Physics in 1999 and a M.Sc. degree in Computer Science in 2003. Mr. Orfila has several publications in international conference proceedings. His interests and his PhD research are focused on Intrusion Detection Systems and Decision Analysis. *Address for correspondence:* Despacho 22A22, Depto. Informatica, Univ. Carlos III, Av. Universidad 30, Leganes 28911 Madrid (SPAIN); *Phone*: +34 916249422; *E-mail:* adiaz@inf.uc3m.es.

JAVIER CARBO is currently Associate Professor at the Computer Science Department of the Carlos III University of Madrid, Spain. He is a member of the Research Group of Applied Artificial Intelligence at this department. Dr. Carbó has published over 30 papers in international journals and conference proceedings. He has acted as a reviewer of more than 15 international conferences; he has also been invited speaker at several research seminars and has organized 3 workshops and a special session on AI and Information Security. He has participated in a United Nations founded research project, 2 ESPRIT programs and other national projects. His interests include trust issues, automated negotiations, multi-agent systems, electronic payments and fuzzy logic. *Address for correspondence:* Despacho 21B21, Depto. Informática, Univ. Carlos III, Av. Universidad 30, Leganés 28911 Madrid (SPAIN); *Phone*: +34 916249105; *E-mail:* jcarbo@inf.uc3m.es.

ARTURO RIBAGORDA is Professor at the Computer Science Department of the Carlos III University of Madrid, Spain, and he is its current director. He leads the Information Security Group at this department and has over 50 papers in national and international journals. Dr. Ribagorda is also involved in several European and Spanish research projects and his interests are related to Network Security. *Address for correspondence:* Despacho 22A27, Depto. Informática, Univ. Carlos III, Av. Universidad 30, Leganés 28911 Madrid (SPAIN); *Phone*: +34 916249463, *E-mail:* arturo@inf.uc3m.es.