



# Deliverable 2.3

## **Yearly Updated Requirements Specification Documents**

Symantec Ireland  
Version 04  
5. February 2009



## AntiPhish

Anticipatory Learning for Reliable Phishing Prevention

Specific Targeted Research or Innovation Project

EU 6th Framework Programme project IST 2006 027600

2.4.3 Towards a global dependability and security framework

Start date of project: 1 January 2006

Duration: 36 months + 6 month extension

## Yearly Updated Requirements Specification Documents

Due date of deliverable: M36 (31 December 2008)

Actual submission date: M37 (DD January 2009)

Lead Contractor for this Deliverable: Symantec Ireland

<b>Responsible author(s):</b>	Patrick Horkan
<b>Co-author(s):</b>	

Revision: Working

<b>Project co-funded by the European Commission within the Sixth Framework Programme (2002-2006)</b>		
<b>Dissemination Level</b>		
<b>PU</b>	Public	X
<b>PP</b>	Restricted to other program participants (including the Commission Services)	
<b>RE</b>	Restricted to a group specified by the consortium (including the Commission Services)	
<b>CO</b>	Confidential, only for members of the consortium (including the Commission Services)	

## Revision history

Deliverable administration and summary		
Project acronym: AntiPhish		ID: IST-2006-027600
<b>Document identifier:</b>	AntiPhish-del-D23-RequirementsSpecification-f-v04	
Leading partner: Symantec Ireland		
Report version: v04		
Report preparation date: 23.Dec.2008		
<b>Classification:</b> Public		
<b>Nature:</b> Report		
<b>Author(s) and contributors:</b> Patrick Horkan in collaboration with all partners		
<b>Status:</b>		Plan
		Draft
		Working
	X	Final
		Submitted
		Approved

The AntiPhish © Consortium has addressed all comments received, making changes as necessary. Changes to this document are detailed in the change log table below.

Date	Edited by	Status	Changes made
23.12.08	Patrick Horkan	Working	Update Reflecting a Year of Progress.
22.01.09	Patrick Horkan	Draft	Update Reflecting a Year of Progress.
04.02.09	Patrick Horkan	Draft	Small edits
05.02.09	Gerhard Paaß	Final	Finalization

Notice that other documents may supersede this document. A list of latest public AntiPhish deliverables can be found at the AntiPhish webpage at [www.AntiPhishResearch.org/publications](http://www.AntiPhishResearch.org/publications).

## Copyright

This report is © AntiPhish Consortium 2006. Its duplication is restricted to the personal use within the consortium, funding agency and project reviewers.

## Acknowledgements

The work presented in this document has been conducted in the context of the EU Framework Programme project IST 2006 027600 AntiPhish. AntiPhish is a 36-month project that started on January 1st, 2006 and is funded by the European Commission as well as by the industrial partners. Their support is appreciated. In 2008 the project was given a six month extension and will now run until mid 2009.

The partners in the project are Fraunhofer Institute for Intelligent Analysis and Information Systems (FHG), Symantec LIRIC Limited (LIRIC), Symantec Ltd. (Symantec Ireland), TISCALI Italia S.p.A. (Tiscali) and K. U. Leuven / Dept. of Computer Science (K.U. Leuven). The content of this document is the result of extensive discussions within the AntiPhish© Consortium as a whole.

## More information

Public AntiPhish reports and other information pertaining to the project are available through AntiPhish public web site under [www.antiphishresearch.org](http://www.antiphishresearch.org).

## Table of contents

Table of contents.....	5
Executive summary.....	6
1 Introduction .....	7
2 Metrics of D2.2 - Progress .....	8
3 Newly Defined Metrics .....	9
4 Requirements Update.....	11
5 Appendix I. Performance Metrics - Defined.....	12
5.1 Metrics from D2.2 .....	12
5.2 Metrics new to D2.3 .....	13

## Executive summary

This document captures the updated requirements for the AntiPhish prototype in the 3<sup>rd</sup> year of its evolution. The aim of AntiPhish is to device improved content filtering techniques to identify phishing email messages. This past year of the project has seen the AntiPhish prototype evaluation in real world situations. As a result, to determine AntiPhish performance, additional performance metrics were required while others needed to be reconsidered. Specifically in 2008 AntiPhish was used in the following real world situations:

1. AntiPhish evaluation in the context of an Internet Service Provider (ISP) - where both legitimate and spam must be handled.
2. AntiPhish evaluation in the context of industrial use - where the prototype is used to identify phishing messages in an email stream comprising general spam emails only.

The performance metrics described in last years Requirements and Specifications deliverable (D2.2) did not cater for the above scenarios as phishing detection effectiveness tended to be defined with respect to legitimate emails only.

This deliverable presents the new performance metrics as well as quoting progress on previously defined metrics.

## 1 Introduction

This document concerns an update to the requirements and specifications for the evolving AntiPhish prototype. The real-world test deployments of AntiPhish during the year meant that new performance metrics were defined. We define and quote performance with regard to these new metrics but also review how AntiPhish has performed with respect to last years metrics.

## 2 Metrics of D2.2 - Progress

The requirements of the AntiPhish prototype as per the corresponding document from one year ago - D2.2, are defined in the appendices. In this section we present updated performance scores of the current prototype with respect to key metrics presented in D2.2.

We focus on AntiPhish accuracy with respect to (a) phishing versus legitimate email message detection and (b) detection of new salting tricks.

The experiments from which these metrics were obtained were conducted in Symantec's laboratory. A couple of factors meant these experiments were of a synthetic nature:

1. Due to privacy concerns the use of real life legitimate emails was not possible in these experiments, therefore the legitimate email content comprised publicly available samples and newsletters.
2. The rarity of new salting tricks in reality meant we needed to devise an experiment that simulates the arrival of such tricks.

With that in mind we present key D2.2 metric scores in the following table.

Metric	Performance
PFM	0.39%
PFN	4.77%
LCPMT (Training)	8.32 msgs/sec
LCPMT (Application)	17.06 msgs/sec
FPNST	13.3%
FNNST	2.1%
LCPMT (NST)	0.247 msgs/sec

**Table 1, AntiPhish Performance in demo-world**

PFM and PFN above, from lab phishing versus ham classifications, satisfy the prototype goals of 1% and 10% respectively.

The FPNST is much inferior to the goal of 1% but the FNNST satisfies the 10% required. The detection of new salting tricks is an ambitious goal and technical considerations with the current approach may prevent us from pursuing improvement of these scores.

### 3 Newly Defined Metrics

The evaluation of AntiPhish in real-world settings where it was exposed to ham, phishing and regular-spam meant that new performance metrics needed to be added or existing ones refined so that regular-spam (i.e. spam of a non-phishing nature) was explicitly accounted for.

Two new performance metric classes were defined so that AntiPhish evaluation at an Internet Service Provider (ISP) could be measured:

**Non-Ham False Negative Rate (NHFN)** - The number of non-ham messages not blocked by AntiPhish divided by the number of non-ham messages sent to AntiPhish.

**Non-Ham False Positive Rate (NHFP)** - The number of legitimate messages blocked (as non-Ham) by AntiPhish divided by the number of legitimate messages sent to AntiPhish.

AntiPhish was evaluated in two scenarios for ISP use; where it acts like a stand-alone general spam filter and where it is positioned as a second line of defense to a commercial spam filter.

Instances of the above metrics were used to capture each scenario, thus: **NHFN\_UF** and **NHFP\_UF** relate to AntiPhish as a standalone solution, while **NHFN\_SF** and **NHFP\_SF** represent where a spam filter is used in front of AntiPhish.

The new performance metrics required for AntiPhish evaluation in the context of an industrial setting namely that of an email security response service, are as follows:

**Phishing False Negative Rate in General Spam Stream (PFN\_GS)** - The number of phishing messages, in the general spam stream, that are not classified as phishing by AntiPhish, divided by the total number of phishing messages, from the same stream, given to AntiPhish.

**Phishing False Positive Rate in General Spam Stream (PFP\_GS)** - The number of regular-spam messages, in the general spam stream, classified as phishing (by AntiPhish), divided by the total number of regular-spam messages, from the same stream given to AntiPhish.

From experiments run at Tiscali and in Symantec lab we obtained the scores presented in the table below.

Metric	Rate (%)	Prototype Goal (%)
NHFP_UF	0.08	1.0
NHFN_UF	6.71	10.0
NHFP_SF	1.76	1.0
NHFN_SF	15.34	10.0
PFP_GS	1.5	0.5
PFN_GS	5	2.5

**Table 2, AntiPhish Performance in real-world environment**

The prototype metric goals have being selected in accordance with the phishing false positive (PFP) and false negative rates (PFN) set at the outset of the project, but also in the knowledge that we need to improve on the obtained rates in the short term.

With regard to the Tiscali experiment, efforts will be made to reduce NHFP\_SF so that AntiPhish catches non-ham missed by conventional spam filters while keeping misclassification of ham emails to a minimum.

With regard to the Symantec lab trial some effort will be made to reduce PFP\_GS and PFN\_GS.

## 4 Requirements Update

	Current Production	Prototype	Production	Goal
<b>PFN</b>	proprietary	0.10*	0.05	0.01
<b>PFP</b>	proprietary	0.01	0.001	1/1M*
<b>RECALL</b>	proprietary	0.90	0.95	0.99
<b>PRECISION</b>	proprietary	0.95	0.99	0.999999
<b>NHFN_UF</b>	proprietary	0.10	TBN	TBN
<b>NHFP_UF</b>	proprietary	0.01	TBN	TBN
<b>NHFN_SF</b>	proprietary	0.10	TBN	TBN
<b>NHFP_SF</b>	proprietary	0.01	TBN	TBN
<b>PFN_GS</b>	proprietary	0.025	TBN	TBN
<b>PFP_GS</b>	proprietary	0.005	TBN	TBN
<b>LCPMT</b>	proprietary	0.08/U	100	1,000
<b>LCPVT</b>	proprietary	0.12	12	100
<b>LATENCY</b>	300	3,000	300*	30
<b>PERSONNEL</b>	proprietary	1	20	5
<b>HCAS</b>	proprietary	TBD	2M	1M
<b>HCPEP</b>	proprietary	10k	1M	10k
<b>PEPMT</b>	proprietary	100,000	100,000	1M
<b>PEPVT</b>	1,000	1,000	1,000	10,000
<b>PPFNST</b>	proprietary	0.01	0.001	0.00001
<b>PFNNTS</b>	proprietary	0.10	0.050	0.0001
<b>SYNTH</b>	No	Goal of Yes	Goal of Yes	Goal of Yes

\* Note: These requirements were proposed for the Prototype in the original proposal to the European Commission.

## 5 Appendix I. Performance Metrics - Defined

### 5.1 Metrics from D2.2

This section revisits the metrics that were defined prior to 2008 and are still relevant to AntiPhish performance measurement.

**Phishing False Negative Rate (PFN)** - The number of phishing messages not blocked at policy enforcement points divided by the number of phishing messages sent through policy enforcement points.

**Phishing False Positive Rate (PFP)** - The number of legitimate messages blocked (as phishing) at policy enforcement points divided by the number of total messages sent through policy enforcement points.

**Recall (RECALL)** – The ratio of phishing messages blocked divided by the number of total phishing messages.

**Precision (PRECISION)** – The ratio of phishing messages blocked divided by the total number of messages blocked.

**Labelled Content Processing Message Throughput (LCPMT)** – The volume of labelled data to be processed by the analysis system in generating blocking criteria, to be disseminated to enforcement points, measured by the number of messages per second through the analysis system.

**Labelled Content Processing Volume Throughput (LCPVT)** – The volume of labelled data to be processed by the analysis system in generating blocking criteria for dissemination to enforcement points measured by megabytes per second through the analysis system.

**Latency (LATENCY)** – The latency from the time a sample of a phishing is acquired from a labelled source to the time an effective rule can be deployed for blocking the phishing messages, measured in seconds. NOTE: LATENCY depends on interactions between the centralized Analysis System and the distributed set of Policy Enforcement Points along with processing times at the Policy Enforcement Points and Analysis System respectively.

**Personnel (PERSONNEL)** – The number of people required to operate the system at any given time to maintain the desired performance.

**Hardware Cost of the Analysis System (HCAS)** – The cost of hardware for the analysis system generating blocking criteria, measured in Euros.

**Hardware Cost Per Enforcement Point (HCPEP)** – The cost of hardware for enforcing blocking criteria at each enforcement point, measured in Euros.

**Projected Enforcement Point Message Throughput (PEPMT)** – The projected volume anticipated for each enforcement point, measured in messages per second. Please note that large customers may operate multiple policy enforcement points in parallel.

**Projected Enforcement Point Volume Throughput (PEPVT)** – The projected volume anticipated for each enforcement point, measured in megabytes per second. Please note that large customers may operate multiple policy enforcement points in parallel.

**Probability of False Positive in Detection of New Salting Techniques (FPFNST)** - The percentage of messages flagged by the Analysis System as having a previously unseen salting technique while lacking such a previously unseen salting technique divided by the number of messages lacking such a previously unseen salting technique sent through the Analysis System specifically to determine FPFNST after conclusion of a training period.

**Probability of False Negative in Detection of New Salting Techniques (PFNNST)** - The percentage of messages with previously unseen salting techniques that are not detected as having previously unseen salting techniques divided by the number of messages with previously unseen salting techniques sent through the Analysis System specifically to determine PFNNST after conclusion of a training period.

## 5.2 Metrics new to D2.3

**Non-Ham False Negative Rate (NHFN)** - The number of non-ham messages not blocked by AntiPhish divided by the number of non-ham messages sent to AntiPhish.

**Non-Ham False Positive Rate (NHFP)** - The number of legitimate messages blocked (as non-Ham) by AntiPhish divided by the number of legitimate messages sent to AntiPhish.

These metrics are further defined to represent AntiPhish processing of a spam-filtered (**SF**) stream and AntiPhish processing of a un-(spam)-filtered (**UF**) stream, to obtain **NHFN\_UF**, **NHFN\_SF**, **NHFN\_UF** and **NHFN\_SF**.

**Phishing False Negative Rate in General Spam Stream (PFN\_GS)** - The number of phishing messages, in the general spam stream, that are not classified as phishing by AntiPhish, divided by the total number of phishing messages, from the same stream, given to AntiPhish.

Phishing False Positive Rate in General Spam Stream (**PFP\_GS**) - The number of regular-spam messages, in the general spam stream, classified as phishing (by AntiPhish), divided by the total number of regular-spam messages, from the same stream given to AntiPhish.