



AntiPhish D7.4 : Project Presentation

Patrick Horkan
Symantec Ireland

Version 03

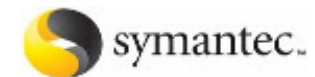
July 2009



EU Project - AntiPhish

- Project Duration: 01/2006 – 06/2009
- Concerned with “the lure” aspect of Phishing (i.e. the email)
- Develop **content-based** phishing filters
- Deploy in realistic workflows
- Trainable and adaptive filters
 - adapt to new phishing formats
 - anticipate new phishing formats

- Consortium
- Fraunhofer IAIS (DE)
 - Symantec (GB, IRL)
 - Tiscali (IT)
 - Nortel (FR)
 - K.U. Leuven (BE)



Presentation Outline

- Phishing – Recent Facts and Figures
- AntiPhish Approach to Phishing Filtering
- Progress on Features
- Results on Benchmark Data
- Real-life Application
 - On real-life 'normal' email stream
 - On general-spam dataset

Phishing : Recent Facts and Figures

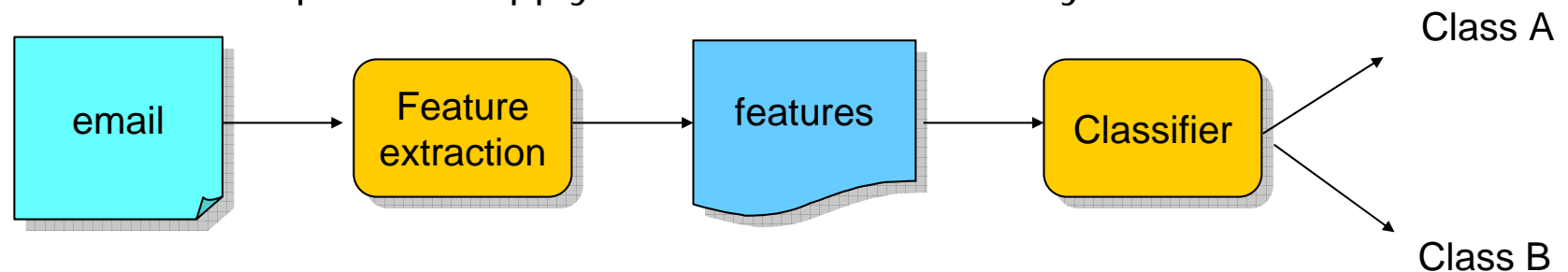
- Sectors Targeted
 - Mainly financial services but also retail, ISP's, internet community (e.g. social networks), governments
- Anti-Phishing Working Group (APWG) – covering 2nd half 2008
 - phishing email reports to them: ~ 28,000 / month
 - phishing sites detected by them: ~ 23,000 / month
 - brands targeted : 247 / month
- Symantec observes (Internet Security Threat Report - ISTR)
 - 85-90% of all email as spam
 - First half of 2009, on average 5% of spam is phishing/fraud
 - detected web hosts up 66% in 2008 (year on year)

Gartner Inc. (2009) - "The War on Phishing Is Far From Over"

- In United States 5 million consumers out of pocket to the tune of \$351 on average during 2008

AntiPhish Approach to Phishing Filtering

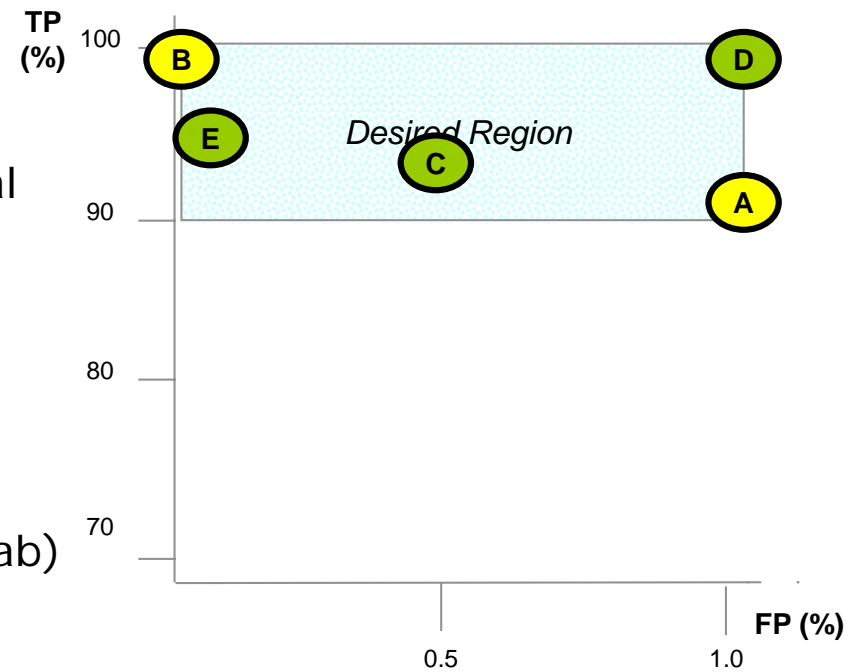
- Machine Learning solution
- Use email features relevant to phishing detection
- Training data: emails labeled with classes ham, spam, phishing
- Train a classifier
 - Select a classification function to optimally predict class
 - Feature weighting according to labelled training data
- Run-time depiction – apply to new emails in binary classification



- Class categories: ham, spam, phishing, non-ham

Summary of Performance

- True Positive and False Positive Accuracy
 - Reference Points
 - A – prototype (APS) goal
 - B – leading commercial general spam filter
 - AntiPhish (APS) results
 - C – non-ham v ham (in field)
 - D – phishing v ham mix (lab)
 - E – phishing v regular spam (lab)
- Processing Rate
 - Filtering: 18-42 msg./sec depends on message size and format



AntiPhish Email Features - Progress

- Valuable Features for real-world deployments
 - Dynamic Markov Chains
 - Latent Topic Model
 - Link – deceptive links [similar to Fette et al]
- Features with a more research perspective
 - Salting – hidden text salting resolution
 1. Resolve known salting tricks to produce perceived text
 2. Detect Unknown salting tricks – using 1. (above) and incorporating Optical Character Recognition (OCR).
Solution tested in demo/lab world
 - Graphical Based – logo detection, image distortion
 - Limited success/progress on these

Dynamic Markov Chains

- Operate on the bit representation of the natural language text of the email
- Model a bit sequence as a stationary and ergodic Markov source with limited memory

01010010100100101110101001011010010101001010100111010010101010101 ...

$$p(\mathbf{x}) \approx \prod_{i=1}^{|\mathbf{x}|} p(x_i | x_{i-k}^{i-1})$$

- Incrementally build such an automaton / Markov chain to model the training sequences
- Train one DMC for each of the classes (i.e., ham, spam, phishing), For a new email look which model fits best
- Has been successfully applied to spam classification [Bratko et al., JMLR 2006]

Dynamic Markov Chains: Details

- States: Two probabilities representing the likelihood that the source emits 1 or 0 as next symbol
- Prediction: Move through automaton, add up likelihoods
- Training (incremental): States are cloned when reached via a frequently used transition
- **Model size reduction:** Use training examples that the model cannot already classify well enough (after some initial training, see also uncertainty sampling in active learning)
- Features: Expected cross entropies of a message for either model (ham and phishing), Boolean membership indicators

Latent Topic Models

Analyze on the co-occurrence of words

- Similar to word clustering: Specify the number of topics in advance
- Common methods: LDA, PLSA
- Probabilistic latent semantic analysis:
Models the probability of each co-occurrence as a mixture of conditionally independent multinomial distributions
- Latent Dirichlet Allocation:
Generative Bayesian version with Dirichlet prior
- Document: Mixture of various topics

Latent Topic Models: Class Specific

- Analyses on the co-occurrence of words Class-Topic Model (CLTOM): Extension of LDA

- Example Topics

<i>T1</i>	<i>T2</i>	<i>T3</i>	<i>T4</i>	...	<i>T10</i>
want	account	order	email	...	Degree
today	security	watches	please	...	Phone
great	bank	watch	message		Outside
free	access	price	online		Leave
body	secure	product	address		Work
acai	accounts	brand	policy		usa
help	services	pills	sent		Allow
start	information	free	view		Few
weight	online	available	contact		Doctorate
diet	member	visit	click		Financial
berry	emails	pharmacy	new		Idea
try	reply	brands	ensure		Sufficient
women	protect	quality	information		Verifiable

- Each topic has a numerical weighting with respect to both phishing and non-phishing
- Words of email determine Probability Distribution over the topics – this acts as the email topic feature

Results on Benchmark Data


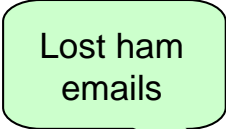
Standard method: **10-fold cross-validation**

- Criteria: Precision, recall, F-measure, false positive rate, false negative rate, accuracy for comparison with related work
- Note: Errors are not of equal importance

Test Corpus: Assembled by [Fette et al., WWW 2007]

- Ham emails: SpamAssassin corpus
- Phishing emails: Collected by Nazario
- Total size: 7808 emails, 6951 ham (89%) and 857 phishing (11%)

Overall result



Features	FPR %	FNR %	F %
Fette et al. 2007	0.13	3.62	97.64
Basic features	0.20	6.39	95.88
Topic features (K=50)	0.20	2.72	97.80
DMC features	0.00	4.02	97.95
DMC + Topic features	0.01	1.89	98.93
All features	0.01	1.30	99.29
Feature Selection	0.00	1.07	99.46

[Bergholz, et al. 2009]

- FPR reduced by 92%, FNR by 64%
- Statistically significant difference to [Fette et al. 07] with less than 1% error
- Feature selection: Better result with fewer features and less training data (20% reserved for validation)

Real-Life Application I – filter ‘normal’ email stream

- AntiPhish training needs to be specific to the stream
- AntiPhish deployed as a non-ham versus ham filter
- Very strict privacy regulations
- Spam filters may be used to aid AntiPhish
- Emails in different languages – but mostly English and Italian
- Experiments: “Almost online”

General Deployment Approach

Start: Initial AntiPhish model M_0

For every day $t \in \{1, \dots, n\}$:

1. Capture a set of emails S_t , sent in real-time through spam filters
2. Select a test subset $T_t \subset S_t$ for evaluation of the current AntiPhish model M_{t-1}
3. Select a subset $A_t \subset S_t$ of emails that are difficult to classify to be used for **active learning**
4. Obtain labels for sets T_t and A_t
5. Evaluate current model M_{t-1} on the set T_t
6. Add set A_t to the training set, train the new model M_t

Details

- AntiPhish is evaluated on arbitrary collected emails.
- Deployment period: $n = 20$ days.
- Used features: unigram, DMC, semantic topics with $k = 25$ topics, link, and lexical features
- Every day a total of $| T_t \cup A_t | = 750$ emails are selected.
- An email is classified as non-ham if and only if it is considered with a probability of at least 95% to be non-ham.

Stratified Evaluation

- T_t : Stratified sample of its underlying base set S_t
- Idea. "Better" represent interesting emails
- Two buckets: Emails that are difficult or easy to classify
- Basic procedure: Over-sample the difficult emails, but give them a lower weight in evaluation
- More specifically: Let $S_t = S_t^{(u)} \cup S_t^{(c)}$, we want to sample k_1 and k_2 emails respectively

- Then

$$w_1 = \frac{s^{(u)}}{|S_t|} * \frac{k_1 + k_2}{k_1} \quad w_2 = \frac{s^{(c)}}{|S_t|} * \frac{k_1 + k_2}{k_1}$$

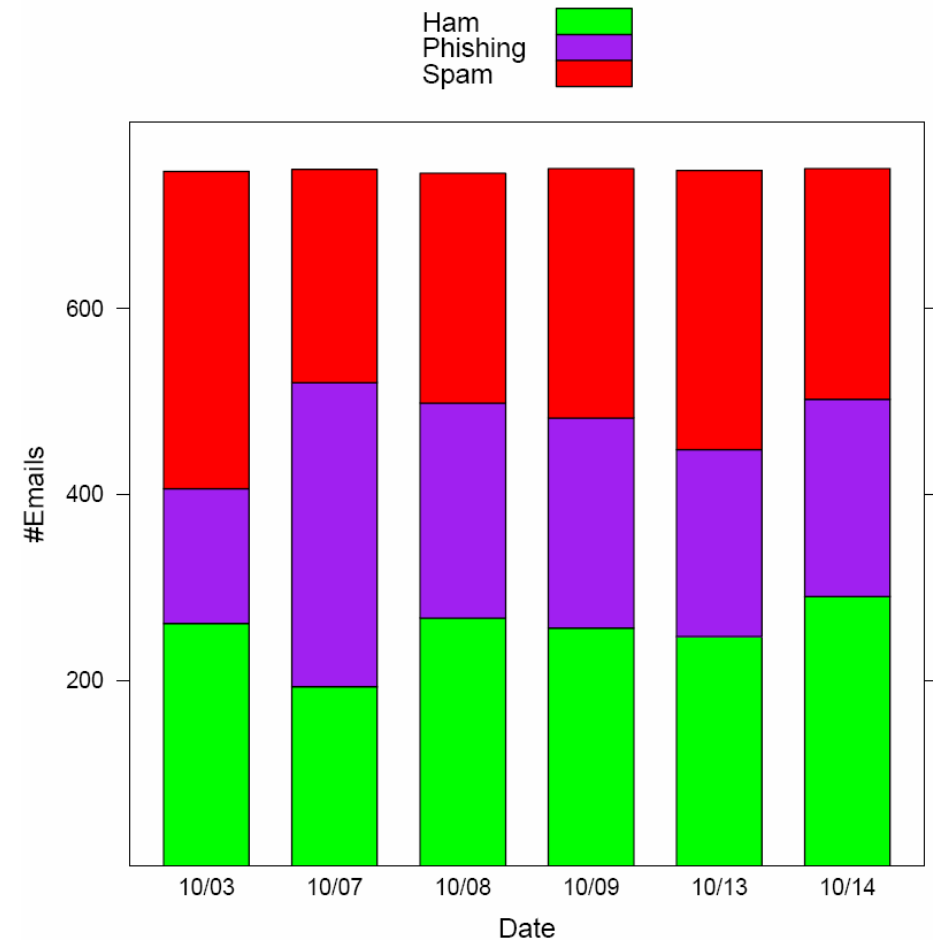
- We use a probability of $p = 95\%$ (for non-ham) as certainty threshold.

Active Learning

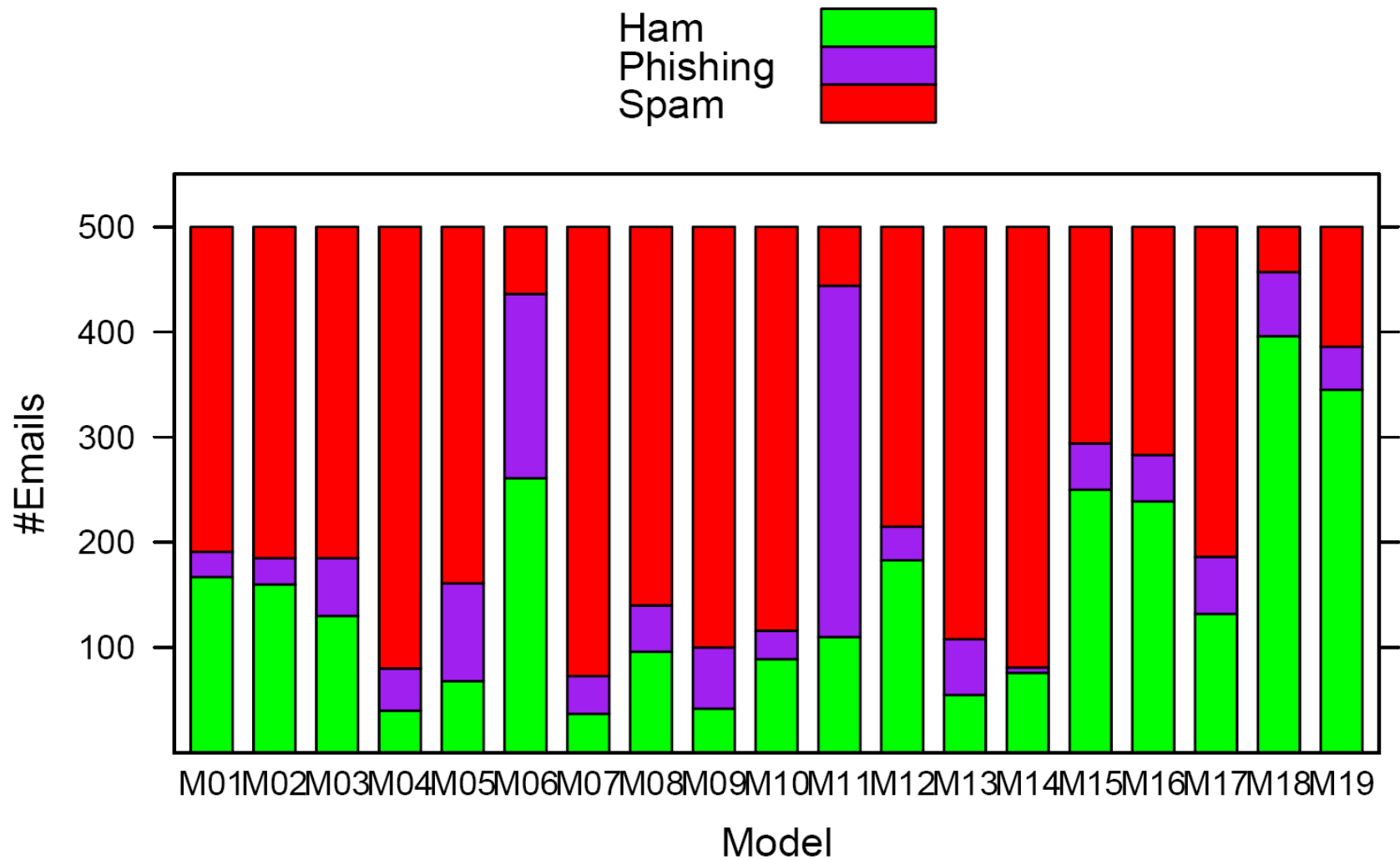
- Set of additional training emails per day A_t , $|A_t| = 500$
- 400 top-ranked emails from S_t having the lowest confidence in classification
- . . . plus 100 emails randomly selected from the rest of S_t
- Minimization of duplicates among the 400 uncertain emails:
Ignore mails with identical score

Initial Dataset

- Initial dataset: Six days of 750 messages each
- Total: 4489 messages
 - Ham: 1514 (34%)
 - Phishing: 1342 (30%)
 - Spam: 1633 (36%)
 - Non-Ham: 2975 (66%)
- Time period for experiment: subsequent 20 days



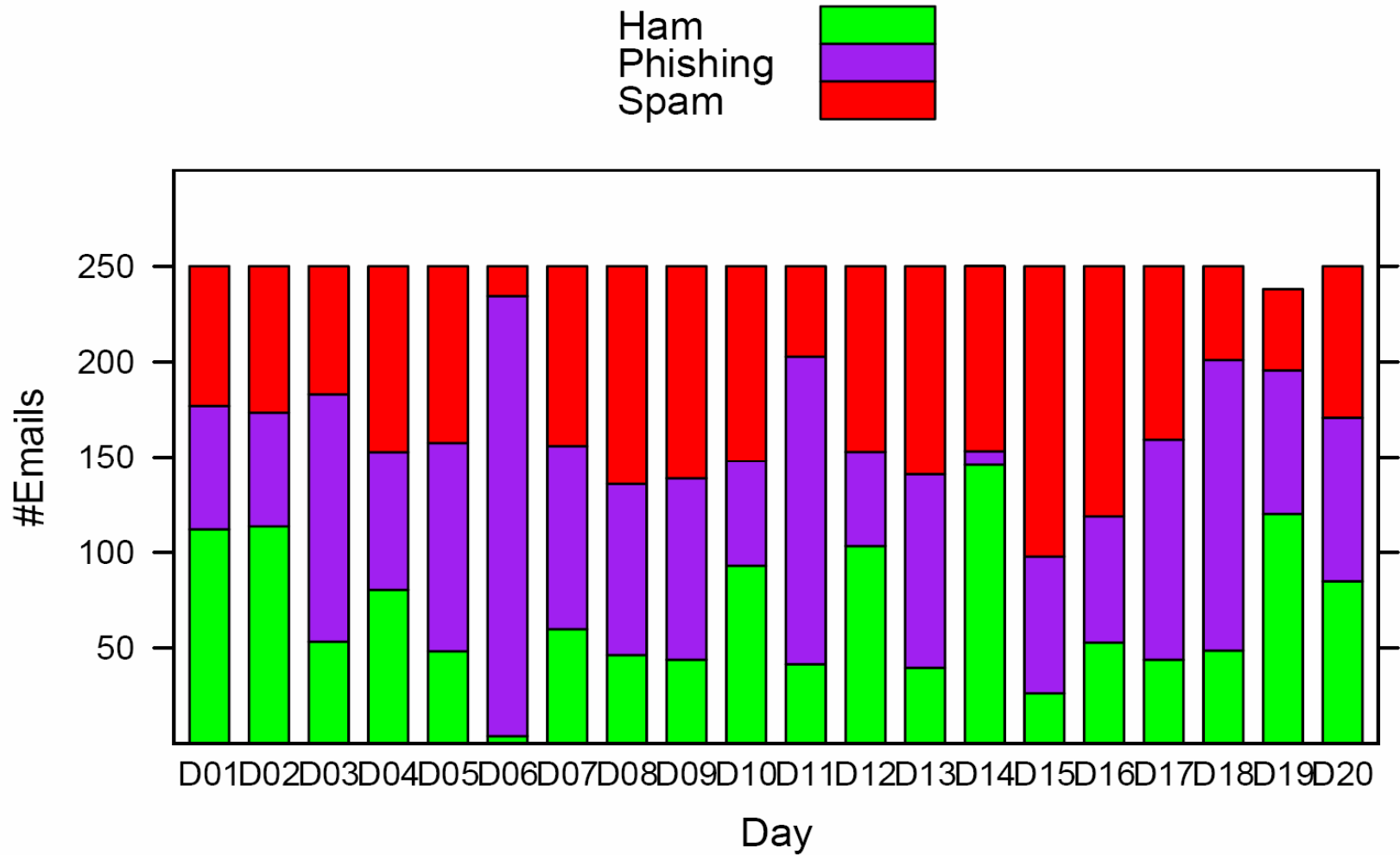
Additional Training Data Through Active Learning



Test Data and Evaluation

- Separate from active learning training data
- 250 messages per day
- $k_1 = k_2 = 125$ difficult and easy messages
- Sometimes less, because not enough difficult emails were found
- Evaluation:
 - False Positive Rate: Proportion of “lost” ham emails in all ham emails
 - False Negative Rate: Proportion of missed non-ham emails in all non-ham emails

Test Data



Baseline Result

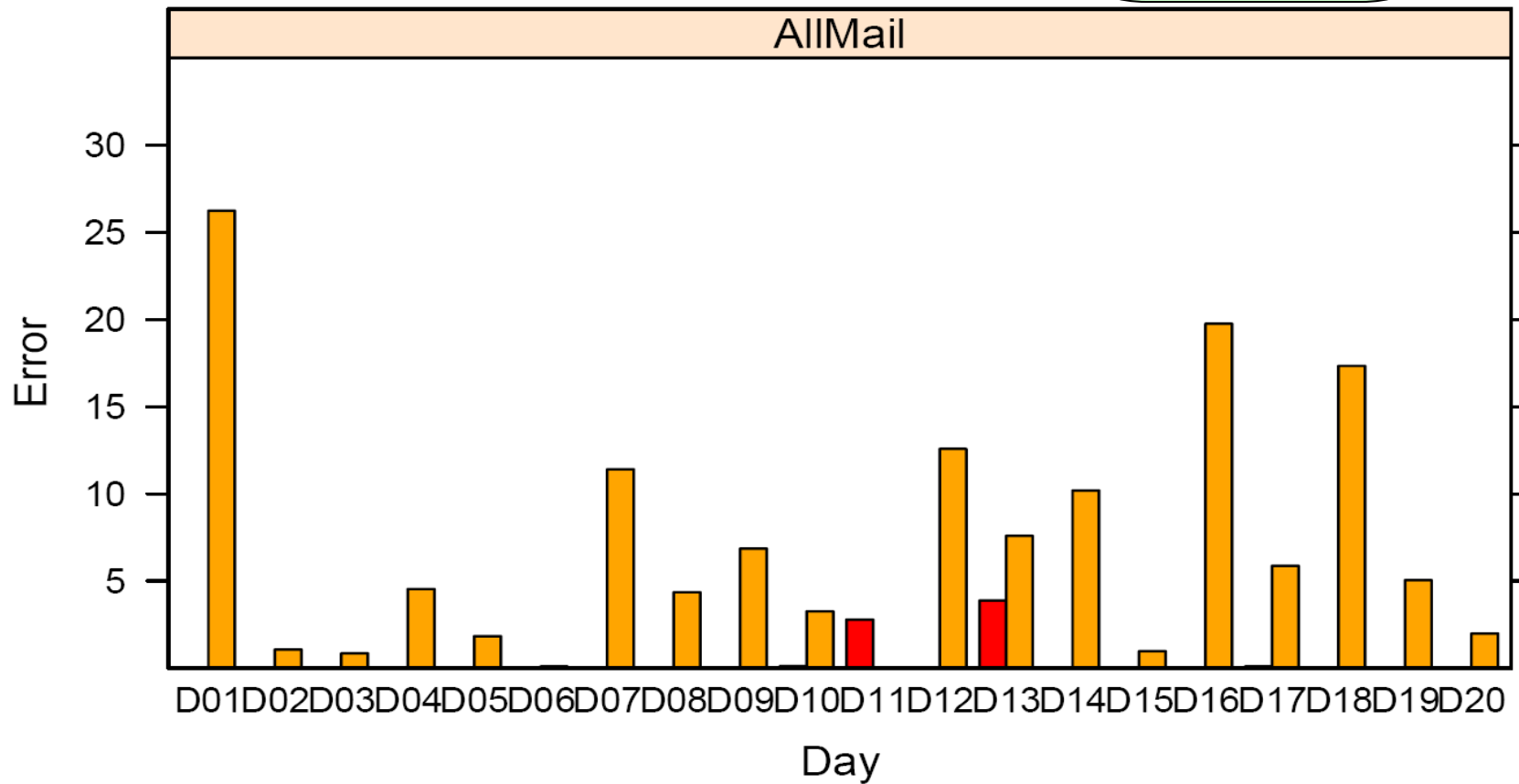
FPR Average: 0.34%

FNR Average: 7.09%

FPR
FNR

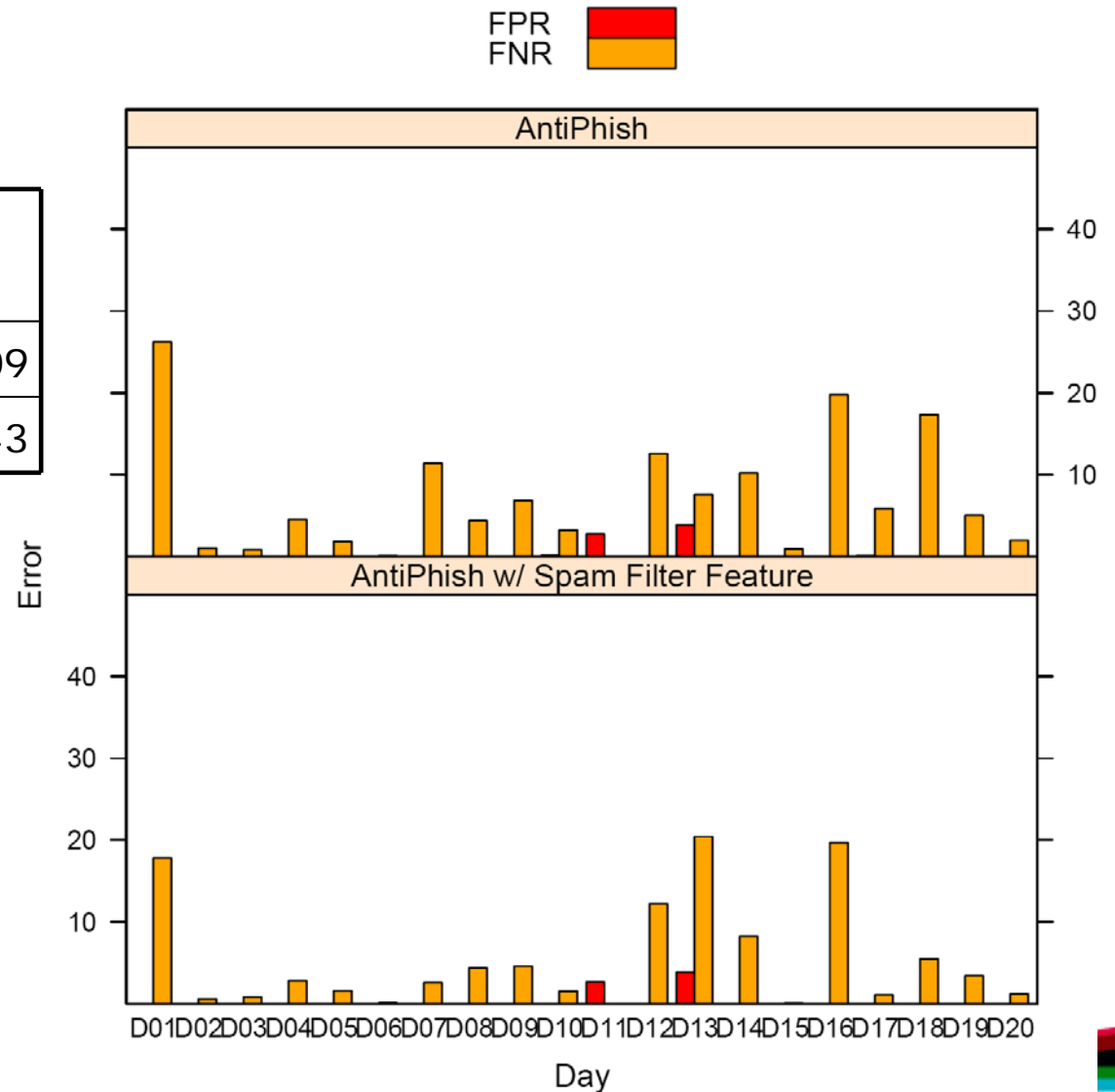
Non-Ham
classified as
Ham

Ham classified
as Non-Ham



Spam Filter Vote as Feature

Spam Filter	FPR%	FNR%
without	0.34	7.09
with	0.33	5.43

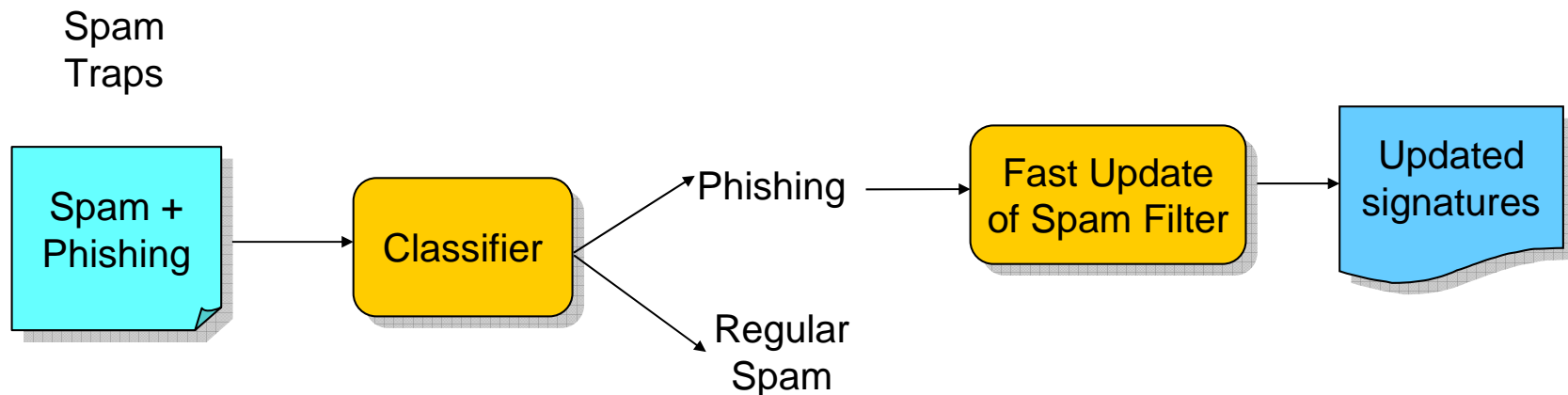


Interpretation of Results

- As stand-alone filter
 - AntiPhish catches 93% of non-ham a good deal of which is phishing
 - FP's were experienced on 4/20 days
- Active-learning improves AntiPhish accuracy
 - We obtained inferior results of (0.43%, 15.47) using M_0
- Using verdict of a spam-filter improves AntiPhish accuracy

Real life Application II – Potential to Identify Phishing in Spam

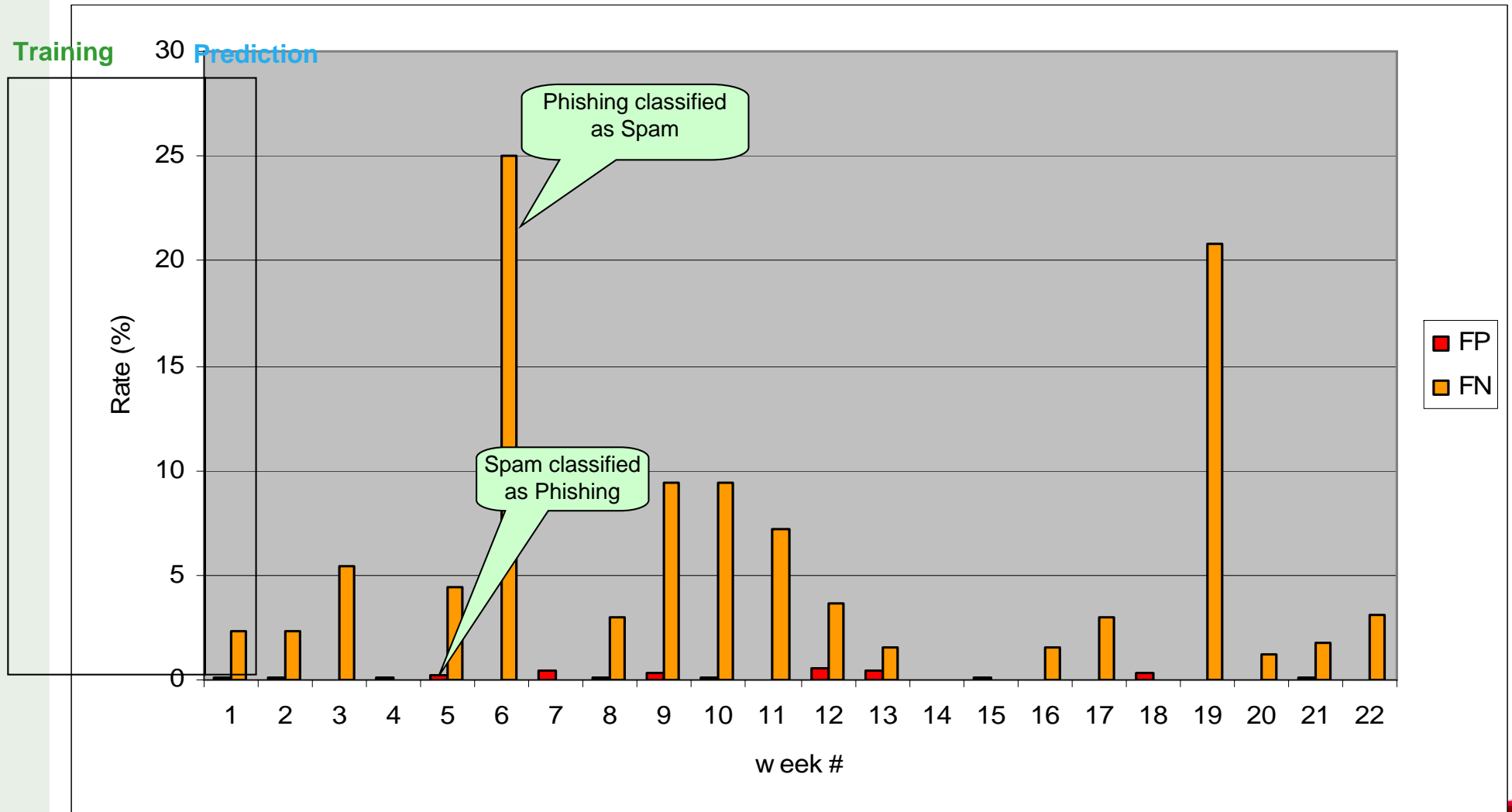
- Anti-spam operations use spam traps to gather the latest spam samples so that these can be better defended against
- The ability to separate out the phishing leads to a quicker defence against such fraudulent activity



Related Laboratory Experiment

- Labelled data – phishing and regular spam from a spam-trap infrastructure
 - Training: 53 phishing vs. 1060 regular spam per week
 - Test: 75 phishing vs. 1443 regular spam per week (on average)
- Duration: June to November 2008 (26 weeks)
- AntiPhish Parameters
 - Features: DMC, semantic (10 topics), unigram, link, lexical
 - Threshold: Neutral (50%)
- Evaluation: Sliding window strategy
 - Each week is filtered on classifier trained on previous N=4 weeks
- Result
 - FPR: Spam classified as Phishing 0.18 %
 - FNR: Phishing classified as Spam 4.89 %

Sliding Window, Training N=4 weeks



Exploitation

- Two possible avenues identified so far:
 1. AntiPhish within an Internet Service Provider (ISP) environment. Account holders may volunteer to adjudicate a handful of their own emails as dictated by the AntiPhish active learning process.
 2. In the context of an anti-spam operations centre AntiPhish being used to filter phishing collected by spam traps – contributing to faster defence against phishing attacks

Dissemination

- Papers accepted for publication include:
 - [G. Paass et al., 2008] "Data Mining for Security and Crime Detection". Security Informatics and Terrorism: Patrolling the Web IOS Press
 - [Bergholz, Paass, Reichartz, Strobel, Moens and Witten 2008] "Detecting Known and New Salting Tricks in Unwanted Emails" CEAS 2008
 - [Bergholz, Chang, Paass, Reichartz, Strobel 2008] "Improved Phishing Detection using Model-Based Features " CEAS 2008
 - [Lioma, Moens, Gomez-Carranza, De Beer, Bergholz, Paass, Horkan 2008] "Anticipating Hidden Text Salting in Emails" RAID 2008.
 - [Bergholz, De Beer, Glahn, Moens, Paass, Strobel 2009] "New Filtering Approaches for Phishing Email" . Accepted for Journal of Computer Security.
- Submitted:
 - [Moens, Boiy, De Beer, Gomez] "Identifying and Resolving Hidden Text Salting". Submitted to Journal IEEE Transactions on Information Forensics & Security.
 - [Paass, Bergholz] „AntiPhish - Machine Learning for Phishing Detection". Project presentation at ECML 2009.
- Other:
 - Panel at ACSAC 2008
 - Invited Talk "AntiPhish – Lessons Learnt"
at KDD Workshop CyberSecurity and Intelligence Informatics (CSI-KDD), Paris 2009