Spam recognition by methods independent from text content

Virus Bulletin 2006 Montréal

Ralf Iffert
ISS C-Force

Mark Usher
ISS C-Force

INTERNET | SECURITY | SYSTEMS®

Ahead of the threat.™

Conventional spam filters are ineffective

- Circumvented by random text
- Outsmarted by spams without any text
- RBLs fooled by changing IPs with high frequency

Overview

Introduction of two spam detection methods independent from text analysis

- **Structure Analysis**
  - Analysing the HTML structure of the email
- **Flow Analysis**
  - Analysing the flow of incoming emails

INTERNET|SECURITY|SYSTEMS®

Structure Analysis

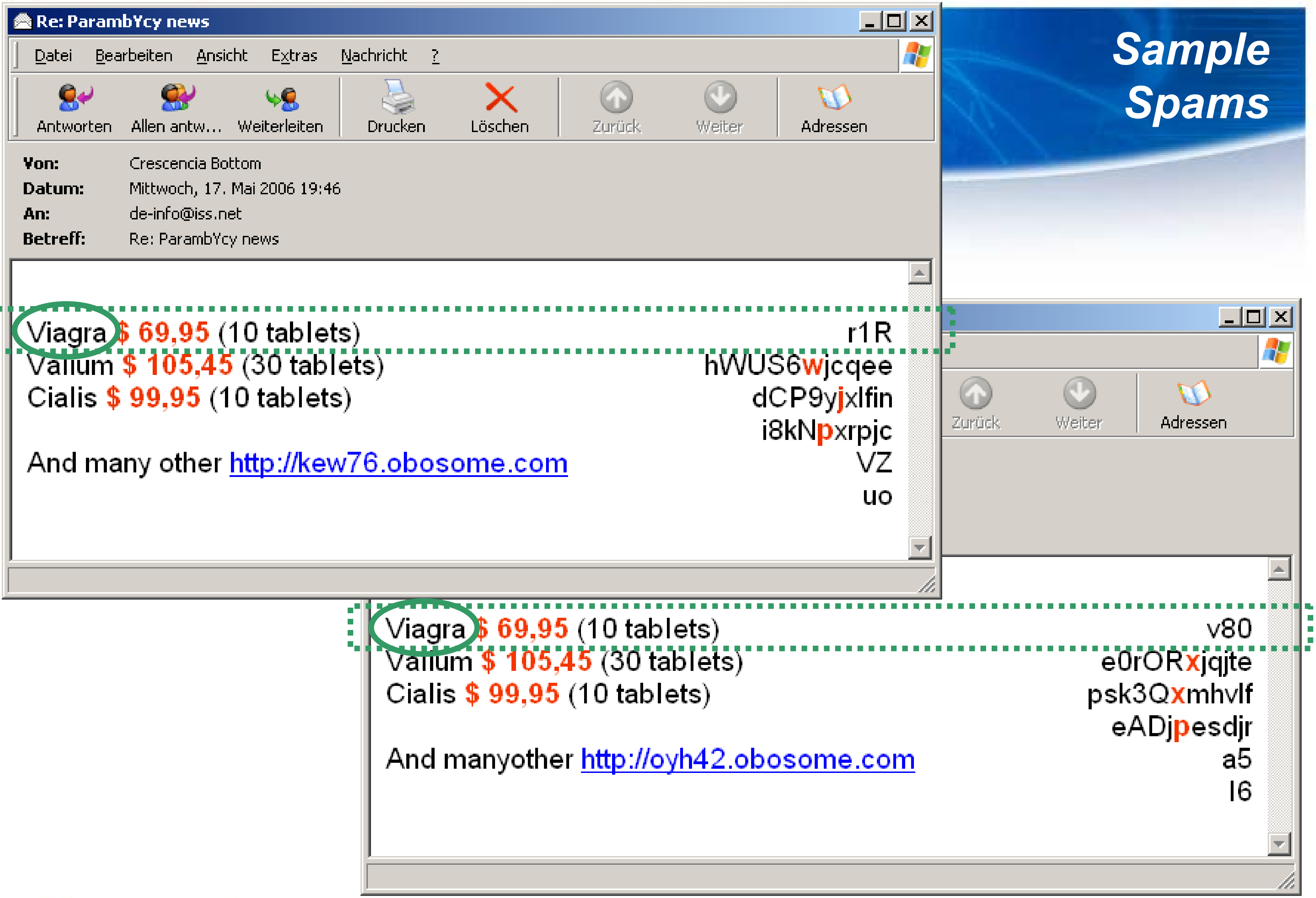
INTERNET | SECURITY | SYSTEMS®

Basic Idea

- Remove all content from HTML part
- Calculate a hash on the remaining HTML structure
- Add hash to a database that is used for spam analysis

INTERNET | SECURITY | SYSTEMS®

Sample Spams



Von: Crescencia Bottom
Datum: Mittwoch, 17. Mai 2006 19:46
An: de-info@iss.net
Betreff: Re: ParambYcy news

Viagra $ 69,95 (10 tablets) r1R
Valium $ 105,45 (30 tablets) hWUS6wjcqee
Cialis $ 99,95 (10 tablets) dCP9yjxlfin
i8kNpxrpjc
And many other http://kew76.obosome.com VZ
uo

Viagra $ 69,95 (10 tablets) v80
Valium $ 105,45 (30 tablets) e0rORxjqjte
Cialis $ 99,95 (10 tablets) psk3Qxmhvlf
eADjpesdjr
And manyother http://oyh42.obosome.com a5
l6

INTERNET | SECURITY | SYSTEMS®

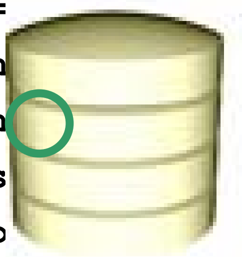
Structure Analysis – Sample Spams: Extract of Source Code

Extract from source code of sample 1

```
<DIV><FONT face=3DArial size=3D3><span style=3D" float : right ">r</span>V<=
span style=3D" float : right ">e</span>i<span style=3D" float : right ">e</=
span>a<span style=3D" float : right ">q</span>g<span style=3D" float : righ=
t ">c</span>r<span style=3D" float : right ">j</span>a <FONT color=3D#EC370=
7><STRONG>$ 69<span style=3D" float : right "> w </span>,95</STRONG></FONT>=
 (1<span style=3D" float : right "> S6 </span>0&nbsp;t<span style=3D" float=
 : right "> WU </span>abIets)</FONT></DIV>
```

Extract from source code of sample 2

```
<DIV><FONT face=3DArial size=3D3><span style=3D" float : right ">v</span>V<=
sig://4A3DDB22F7C25943 (structhash)
span style=3D" float : right ">e</span>i<span style=
span>a<span style=3D" float : right ">j</span>g<span
t ">q</span>r<span style=3D" float : right ">j</span
1><STRONG>$ 69<span style=3D" float : right "> x </s
 (10<span style=3D" float : right "> OR </span>&nbsp
float : right "> 0r </span>s)</FONT></DIV>
```

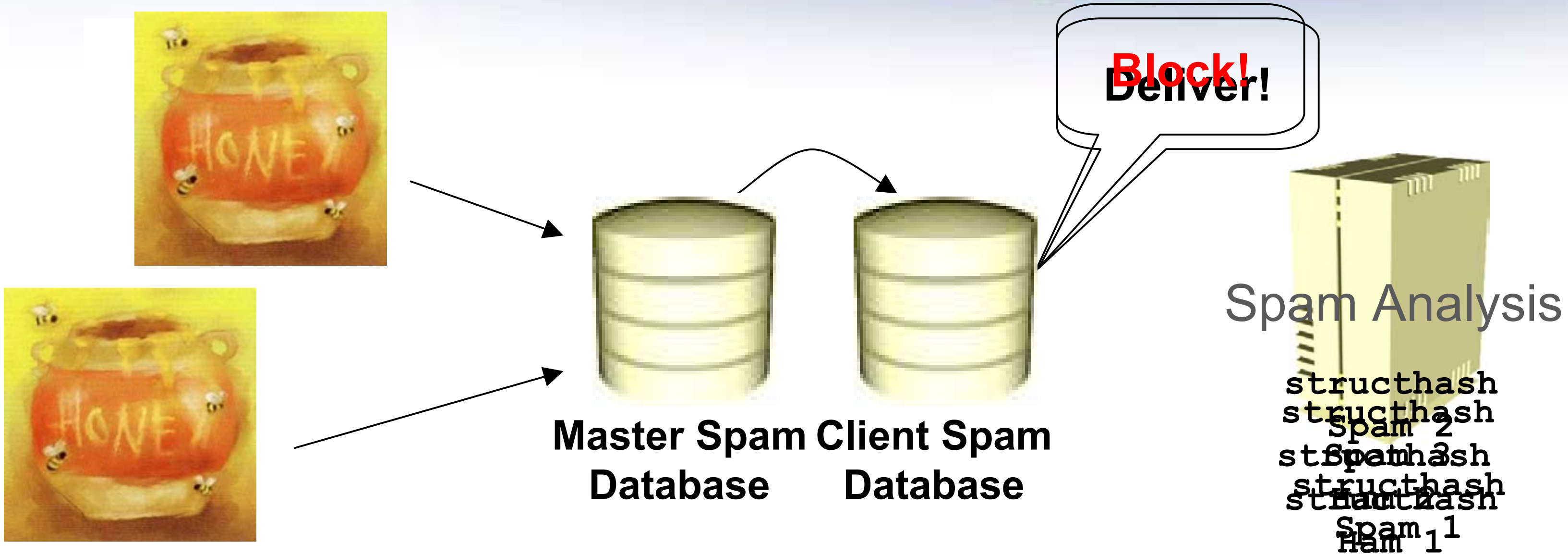


Master Spam Database

INTERNET | SECURITY | SYSTEMS®

Structure Analysis – Workflows

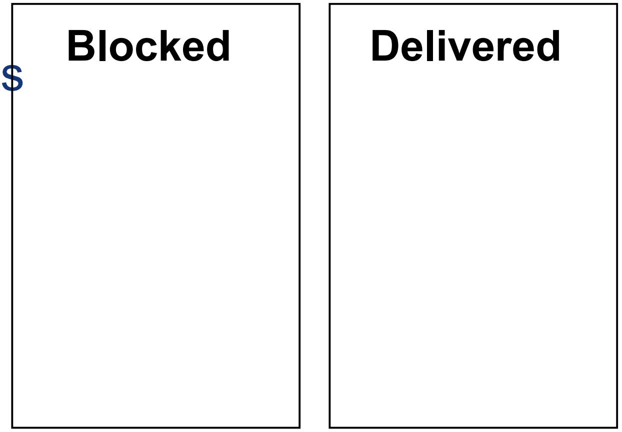
Manufacturer's side	*Client's side*



- Voluminous inflow of spam
- Master Spam Database
  - Fed by hashes delivered from the spam honey pots
- Client Spam Database
  - Exclusively:
  - Used for the local spam analysis
  - Updated by the Master Spam Database

Experimental Results

- Detection ratio: 45.1%
  1.6%
- False positive ratio: 0.010%



INTERNET | SECURITY | SYSTEMS®

Flow Analysis

INTERNET | SECURITY | SYSTEMS®

Flow Analysis

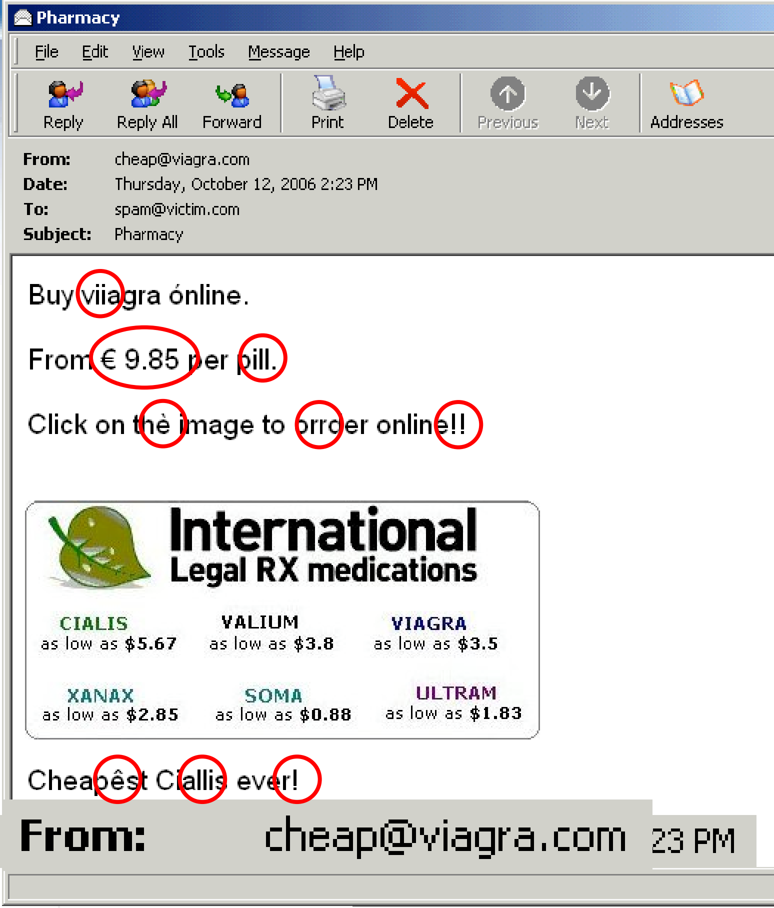
Basic Idea

Identifying “similar“ emails arriving within a small time frame

→Detection of whole spam threads

→Similarity of emails is determined by similarity measures

Flow Analysis



From: cheap@viagra.com
Date: Thursday, October 12, 2006 2:23 PM
To: spam@victim.com
Subject: Pharmacy

Buy viiagra ónline.

From € 9.85 per pill.

Click on thé image to orrder online!!

Cheapêst Ciallis ever!

From: cheap@viagra.com 23 PM

Flow Database

Type	SimilaritySignature	Sender	TimeStamp
1	sig://90445BE9C82FCCAE	sig://E98DBF565BAD2554	2006-10-12 14:23
2	sig://1141FB0CBD68C0F8	sig://E98DBF565BAD2554	2006-10-12 14:23
3	sig://062BAE7F5C13FC22	sig://E98DBF565BAD2554	2006-10-12 14:23
4	sig://155CCDEEEA36B809	sig://E98DBF565BAD2554	2006-10-12 14:23

Similarity Measures

- Simplified body text
- Set of image URLs
- Set of attachments

Further measures are conceivable…

Flow Database

- Administration of similarity measures
- Columns of the Flow Database
  - Type of similarity measure
  - Similarity signature
  - Time when entry was generated
  - Sender of the email

sig://E98DBF565BAD2554
2006-10-12 14:23

```
<META content= name= >
http://enter.your.credit.com/card/number.gif
</HEAD>#<BODY bgColor= > <DIV> <FONT
face= size=2 >
http://and.loose.info/all/your/money.gif
</FONT> <DIV> </BODY> </HTML>
```

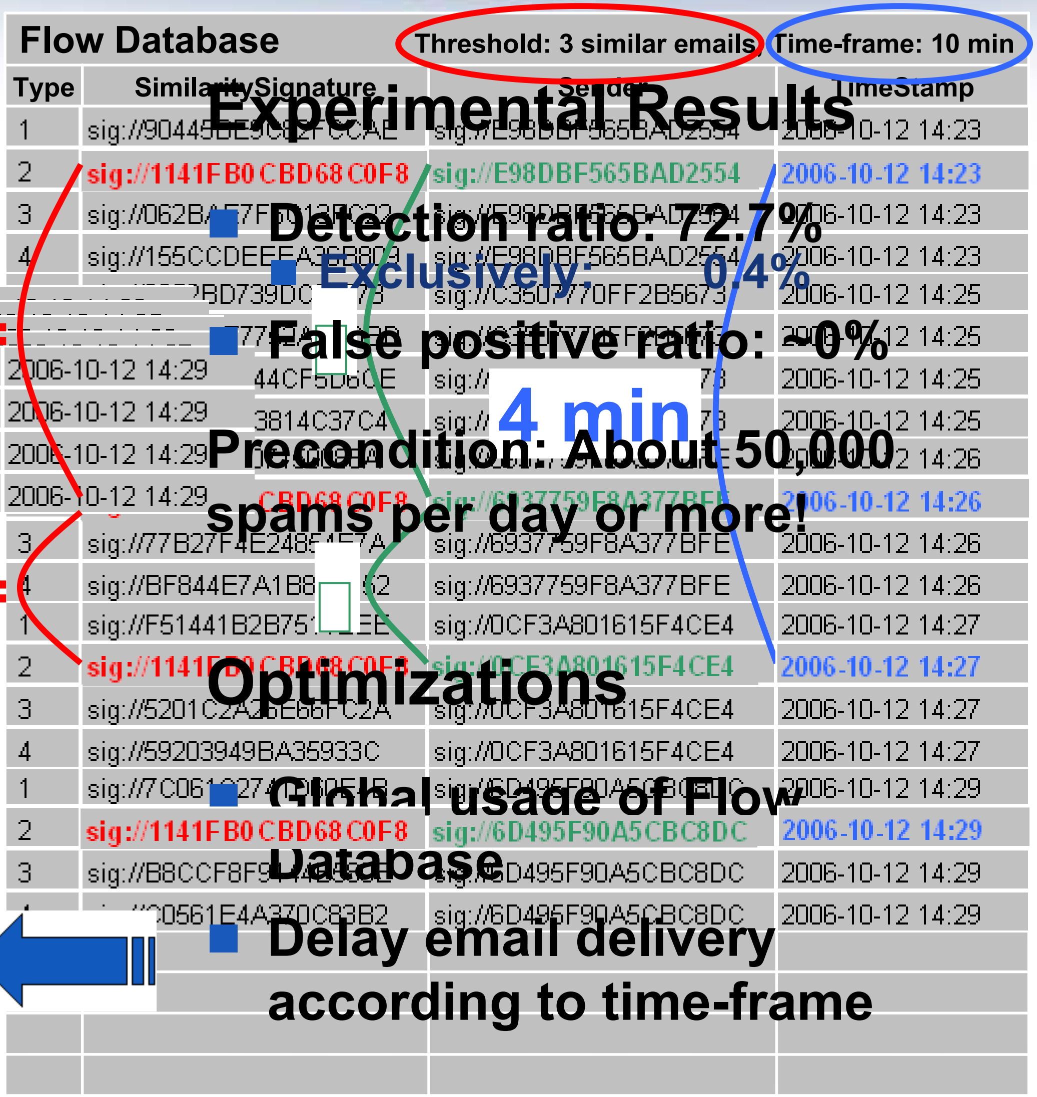
sig://155CCDEEEA36B809

sig://062BAE7F5C13FC22

Flow Analysis – Workflows

Usage of the Flow Database

- Conditioned by the two parameters:
  - Threshold for similar emails
  - Time-frame that is monitored
- Only contains information of emails received within the time-frame



Flow Database Threshold: 3 similar emails, Time-frame: 10 min

Type	SimilaritySignature	Sender	TimeStamp
1	sig://90445[illegible]FCCAE	sig://E98DBF565BAD2554	2006-10-12 14:23
2	sig://1141FB0CBD68C0F8	sig://E98DBF565BAD2554	2006-10-12 14:23
3	sig://062B[illegible]	sig://E98DBF565BAD2554	2006-10-12 14:23
4	sig://155CCDEEEA[illegible]809	sig://E98DBF565BAD2554	2006-10-12 14:23
[illegible]	[illegible]BD739DC[illegible]73	sig://C3507770FF2B5673	2006-10-12 14:25
[illegible]	[illegible]77[illegible]B	[illegible]	2006-10-12 14:25
[illegible]	[illegible]44CF5D6CE	sig://[illegible]73	2006-10-12 14:25
[illegible]	[illegible]3814C37C4	sig://[illegible]73	2006-10-12 14:25
[illegible]	[illegible]	[illegible]	2006-10-12 14:26
[illegible]	[illegible]CBD68C0F8	sig://6937759F8A377BFE	2006-10-12 14:26
3	sig://77B27F4E2485[illegible]7A	sig://6937759F8A377BFE	2006-10-12 14:26
4	sig://BF844E7A1B8[illegible]62	sig://6937759F8A377BFE	2006-10-12 14:26
1	sig://F51441B2B75[illegible]EE	sig://0CF3A801615F4CE4	2006-10-12 14:27
2	sig://1141F[illegible]CBD68C0F8	sig://0CF3A801615F4CE4	2006-10-12 14:27
3	sig://5201C2A[illegible]E66FC2A	sig://0CF3A801615F4CE4	2006-10-12 14:27
4	sig://59203949BA35933C	sig://0CF3A801615F4CE4	2006-10-12 14:27
1	sig://7C061[illegible]	sig://6D495F90A5CBC8DC	2006-10-12 14:29
2	sig://1141FB0CBD68C0F8	sig://6D495F90A5CBC8DC	2006-10-12 14:29
3	sig://B8CCF8F9[illegible]	sig://6D495F90A5CBC8DC	2006-10-12 14:29
[illegible]	[illegible]C0561E4A370C83B2	sig://6D495F90A5CBC8DC	2006-10-12 14:29

2006-10-12 14:29

2006-10-12 14:29

2006-10-12 14:29

2006-10-12 14:29

Experimental Results

- Detection ratio: 72.7%
  - Exclusively: 0.4%
- False positive ratio: ~0%

Precondition: About 50,000 spams per day or more!

Optimizations

- Global usage of Flow Database
- Delay email delivery according to time-frame

INTERNET | SECURITY | SYSTEMS®

Computing time

Computing time depends on

- ***P_i*** Averaged time to extract the similarity data of the ***i**th* similarity measure
- ***M_i*** Averaged hash calculation time on the ***i**th* similarity data
- ***R*** Time for one database request
- ***N*** Number of similarity measures

→ **Computing time per mail:**
of the *i*th similarity measure

$$\sum_{i=1}^{N}(P_i + M_i + R)$$

Memory requirements Flow Database

Required memory depends on

- ***M*** Maximal number of entries in Flow Database
- ***S*** Size in bytes of each Flow Database entry

→ **Required memory: *M*S***

Example

***M* depends on**

- ***N*** Number of similarity measures = 4
- ***E*** Throughput in emails per minute = 600
- ***X*** Time-frame in minutes used for the Flow Analysis = 10

→ ***M = N*E*X***

***S* depends on**

- ***C*** Number of bytes consumed by a Flow Database entry = 44
- ***O*** Memory overhead = 12

→ ***S = C+O***

→ **Required memory: *N*E*X*(C+O)***

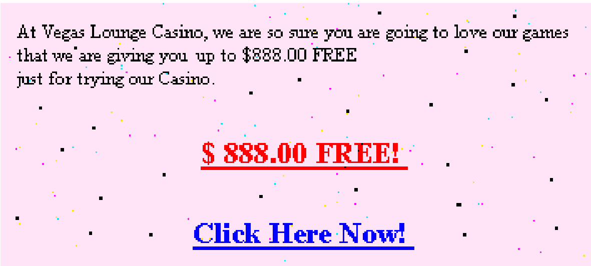
= 1.28 MB

INTERNET | SECURITY | SYSTEMS®

Future Work

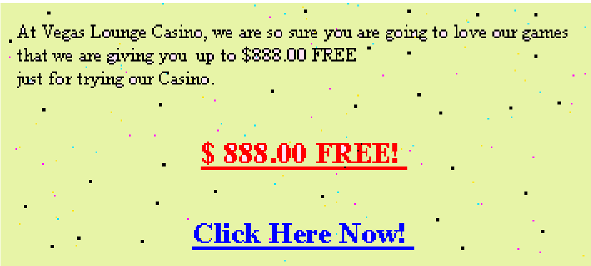
Further approaches

- Automated detection of random text
- Usage of visual features
- Image signatures invariant against random variations


At Vegas Lounge Casino, we are so sure you are going to love our games that we are giving you up to $888.00 FREE just for trying our Casino.

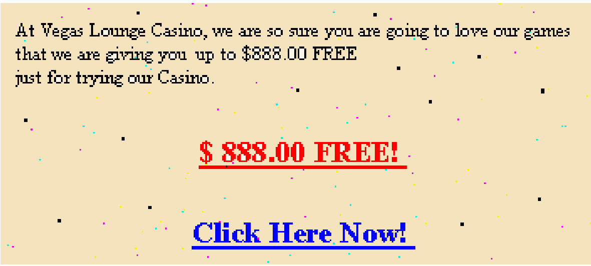
$ 888.00 FREE!

Click Here Now!


At Vegas Lounge Casino, we are so sure you are going to love our games that we are giving you up to $888.00 FREE just for trying our Casino.

$ 888.00 FREE!

Click Here Now!


At Vegas Lounge Casino, we are so sure you are going to love our games that we are giving you up to $888.00 FREE just for trying our Casino.

$ 888.00 FREE!

Click Here Now!

`sig://C3B90474A349823E` `sig://C3B90474A349823E` `sig://C3B90474A349823E`

INTERNET | SECURITY | SYSTEMS®

Many thanks for your attention!

Q & A

Ralf Iffert
ISS C-Force
riffert@iss.net

INTERNET | SECURITY | SYSTEMS®

Ahead of the threat.™