ACME! Computer Security Research

Um modelo para filtragem de mensagens aplicado a uma arquitetura de combate a SPAMs.

WTICG - Workshop de Trabalhos de Iniciação Científica e de Graduação

VIII SBSeg
UFRGS – Gramado – RS
1 de setembro de 2008

Isabela Liane de Oliveira

Adriano Mauro Cansian

Representados por Jorge Luiz Corrêa

UNESP – Universidade Estadual Paulista – Instituto de Biociências, Letras e Ciências Exatas (IBILCE) - Campus de São José do Rio Preto, SP

ACME! Computer Security Research

Agenda

- Introdução e objetivos
- Novo modelo de filtragem
- Arquitetura de implantação do modelo
- Resultados obtidos
- Conclusões e trabalhos futuros

Agenda

- **Introdução e objetivos**
- Novo modelo de filtragem
- Arquitetura de implantação do modelo
- Resultados obtidos
- Conclusões e trabalhos futuros

Introdução

- O protocolo responsável pelos *e-mails* trafegados na Internet foi inicialmente proposto em 1980 como MTP [RFC 772].
- Desde então tem sido aprimorado e é atualmente representado pelo SMTP [RFC 2821].
- No entanto, o problema de abuso no serviço de correio foi identificado muito antes do primeiro padrão MTP:
  - Em 1975, Jon Postel atentou para o problema no documento *On the junk mail problem* [RFC 706] (referindo-se aos primeiros protocolos de troca de mensagem na antiga ARPANET).
  - Em 2008, 33 anos depois, o problema continua e pesquisas estimam que neste ano, cerca de 90% das mensagens eletrônicas serão SPAMs
- Assim, os abusos têm sua origem ligada ao próprio surgimento do protocolo de troca de mensagens.

Introdução

- SPAMs são mensagens enviadas em massa, compostas predominantemente de conteúdo comercial.
- Uma outra modalidade é denominada *Phishing Scam* e tem como objetivo aplicar golpes nos destinatários.
- SPAMs significam:
  - Perda de produtividade, tempo e dinheiro;
  - Necessidade de investimento para sua contenção;
  - Caixas lotadas (implica o não recebimento de novos *e-mails*);
  - Sujeição às tentativas de fraudes;
  - Mau uso da banda da instituição;
  - Mau uso dos equipamentos de servidor de correio eletrônico;
  - Etc.

Introdução

- Para lidar com este problema existem atualmente diversas metodologias:
  - *Greylisting*;
  - SPF (*Sender Policy Framework*);
  - *Domain Keys* (*Yahoo*);
  - Razor;
  - *Distributed Cheksum Clearinghouse*;
  - RBL (*Real Time Blacklist*);
  - *Rule-Based Filter* (filtro por pontuação, com análise de conteúdo).
- Cada uma possui suas vantagens e desvantagens.
- Um fator muito importante atualmente é o desempenho destes filtros.

Objetivos

- **Os objetivos deste trabalho são:**
  - Classificar uma mensagem como SPAM ou não;
  - Permitir uma arquitetura de compartilhamento de informações;
  - Conseqüentemente, diminuir problemas relativos a SPAMs comerciais e minimizar as tentativas de fraudes.
- Como?
  - Utilizando assinaturas (conjunto de informações);
  - Usando a menor carga de processamento possível;
  - Rapidamente;
  - Buscando alto índice de eficiência.

Agenda

- Introdução
- **Novo modelo de filtragem**
- Arquitetura de implantação do modelo
- Resultados obtidos
- Conclusões e trabalhos futuros

Novo modelo de filtragem

- Visa realizar a análise de uma mensagem com base em seu conteúdo.
- Utiliza uma assinatura que descreve uma mensagem.
- Outras características do sistema:
  - Capacidade de analisar o conteúdo de uma mensagem sem infringir sua privacidade;
  - Analisar imagens.
- Para um melhor entendimento, o conceito de codificação MIME é importante.

Novo modelo de filtragem

- Mensagem SMTP e padrão MIME
  - O padrão MIME (*Multipurpose Internet Mail Extensions*) provê mecanismos para conversões de representações:
    - O SMTP é limitado à representação ASCII.
    - O MIME permite que outros tipos de dados sejam trocados, como imagens, binários etc.
  - Mensagens que utilizam MIME possuem um cabeçalho indicando quais os tipos de conteúdo.
- O sistema considera mensagens do tipo:
  - *multiparted* com textos e imagens;
  - os tipos *text* e *image;*
  - e as mensagens em texto plano, que não utilizam o padrão MIME.
- Estes tipos são declarados no cabeçalho MIME dentro do campo *Content-Type*.

Novo modelo de filtragem

```
From spammer@spam.com  Sat Aug 20 00:00:00 2008
Return-Path: <otherspammer@spam.com>
... Outros campos
MIME-Version: 1.0
Content-Type: multipart/related;
   type="multipart/alternative";
   boundary="----=_NextPart_000_0000_BDE06FC4.E1911767"
... Outros campos
This is a multi-part message in MIME format.

----=_NextPart_000_0000_BDE06FC4.E1911767
Content-Type: text/plain; charset="iso-8859-1"
Content-Transfer-Encoding: 7bit
... Corpo da mensagem
```

Novo modelo de filtragem

- A assinatura é uma distribuição de freqüência (contagem) dos caracteres que ocorrem em uma mensagem, considerando 68 possíveis ASCII (do ASCII 33 ao 96 e do 123 ao 126).
- Compõem ainda a assinatura, uma soma do total de caracteres e quantos caracteres diferentes ocorrem.
- Representam tanto texto quanto imagem:
  - Texto: assinatura construída com base nas partes textuais da mensagem (plain, html, etc);
  - Imagem: assinatura construída com base na codificação de uma imagem.

Novo modelo de filtragem

Dec	Hx	Oct	Char	Dec	Hx	Oct	Html	Chr	Dec	Hx	Oct	Html	Chr	Dec	Hx	Oct	Html	Chr
0	0	000	NUL (null)	32	20	040		Space	64	40	100	@	@	96	60	140	`	`
1	1	001	SOH (start of heading)	33	21	041	!	!	65	41	101	A	A	97	61	141	a	a
2	2	002	STX (start of text)	34	22	042	"	"	66	42	102	B	B	98	62	142	b	b
3	3	003	ETX (end of text)	35	23	043	#	#	67	43	103	C	C	99	63	143	c	c
4	4	004	EOT (end of transmission)	36	24	044	$	$	68	44	104	D	D	100	64	144	d	d
5	5	005	ENQ (enquiry)	37	25	045	%	%	69	45	105	E	E	101	65	145	e	e
6	6	006	ACK (acknowledge)	38	26	046	&	&	70	46	106	F	F	102	66	146	f	f
7	7	007	BEL (bell)	39	27	047	'	'	71	47	107	G	G	103	67	147	g	g
8	8	010	BS (backspace)	40	28	050	(	(	72	48	110	H	H	104	68	150	h	h
9	9	011	TAB (horizontal tab)	41	29	051	)	)	73	49	111	I	I	105	69	151	i	i
10	A	012	LF (NL line feed, new line)	42	2A	052	*	*	74	4A	112	J	J	106	6A	152	j	j
11	B	013	VT (vertical tab)	43	2B	053	+	+	75	4B	113	K	K	107	6B	153	k	k
12	C	014	FF (NP form feed, new page)	44	2C	054	,	,	76	4C	114	L	L	108	6C	154	l	l
13	D	015	CR (carriage return)	45	2D	055	-	-	77	4D	115	M	M	109	6D	155	m	m
14	E	016	SO (shift out)	46	2E	056	.	.	78	4E	116	N	N	110	6E	156	n	n
15	F	017	SI (shift in)	47	2F	057	/	/	79	4F	117	O	O	111	6F	157	o	o
16	10	020	DLE (data link escape)	48	30	060	0	0	80	50	120	P	P	112	70	160	p	p
17	11	021	DC1 (device control 1)	49	31	061	1	1	81	51	121	Q	Q	113	71	161	q	q
18	12	022	DC2 (device control 2)	50	32	062	2	2	82	52	122	R	R	114	72	162	r	r
19	13	023	DC3 (device control 3)	51	33	063	3	3	83	53	123	S	S	115	73	163	s	s
20	14	024	DC4 (device control 4)	52	34	064	4	4	84	54	124	T	T	116	74	164	t	t
21	15	025	NAK (negative acknowledge)	53	35	065	5	5	85	55	125	U	U	117	75	165	u	u
22	16	026	SYN (synchronous idle)	54	36	066	6	6	86	56	126	V	V	118	76	166	v	v
23	17	027	ETB (end of trans. block)	55	37	067	7	7	87	57	127	W	W	119	77	167	w	w
24	18	030	CAN (cancel)	56	38	070	8	8	88	58	130	X	X	120	78	170	x	x
25	19	031	EM (end of medium)	57	39	071	9	9	89	59	131	Y	Y	121	79	171	y	y
26	1A	032	SUB (substitute)	58	3A	072	:	:	90	5A	132	Z	Z	122	7A	172	z	z
27	1B	033	ESC (escape)	59	3B	073	;	;	91	5B	133	[	[	123	7B	173	{	{
28	1C	034	FS (file separator)	60	3C	074	<	<	92	5C	134	\	\	124	7C	174			|
29	1D	035	GS (group separator)	61	3D	075	=	=	93	5D	135	]	]	125	7D	175	}	}
30	1E	036	RS (record separator)	62	3E	076	>	>	94	5E	136	^	^	126	7E	176	~	~
31	1F	037	US (unit separator)	63	3F	077	?	?	95	5F	137	_	_	127	7F	177	&#127;	DEL

Novo modelo de filtragem

line: uma linha da mensagem;
text_sig.counter[i]: estrutura que guarda as ocorrências dos caracteres;

```
for (i=0; i<strlen(line); i++) {
      if (line[i] >= 33 && line[i] <= 96)
          text_sig.counter[line[i]-31]++;
      else {
          if (line[i] >= 97 && line[i] <=122)
              text_sig.counter[line[i]-63]++;
          else
              if (line[i] >= 123 && line[i] <= 126)
                  text_sig.counter[line[i]-31]++;
       }
}
```

Novo modelo de filtragem

- Possibilidade de falso-positivo (classificar como SPAM o que não é SPAM) existe, mas é bastante remota.
- Considerando apenas as possíveis combinações de caracteres diferentes, o modelo representa 2^{68} combinações;
- Neste ponto, haverá erro se duas mensagens possuírem *a mesma combinação* de caracteres;
- A probabilidade deste caso é baixa, e varia de acordo com a quantidade de caracteres diferentes na mensagem:

$$P = \frac{1}{C_{68}^{x}}$$

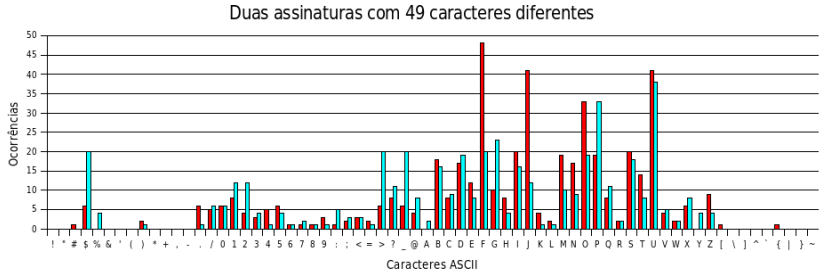
- Onde *x* é a quantidade de caracteres diferentes.

Novo modelo de filtragem

- Calculando P para todos os valores de *x* (1 a 68) teremos valores muito baixos no intervalo de 2 a 66 caracteres:
  - A probabilidade de duas assinaturas com 2 caracteres diferentes possuírem exatamente os mesmos caracteres é $1{,}9 \times 10^{-7}$.
  - Para mensagens com 1 ou 67 caracteres diferentes: 0,000216263.
  - Para mensagens com 68 caracteres diferentes: 1.
- Apesar destes extremos serem de difícil ocorrência, o modelo de assinatura tenta contornar esta possibilidade.
- Para cada caractere diferente, a assinatura armazena seu número de ocorrências.

Novo modelo de filtragem

- Duas assinaturas com a mesma quantidade de caracteres (49) e com quantidades totais próximas (449 e 474) são comparadas:



- Embora exista uma semelhança quanto aos caracteres que ocorrem na mensagem, há uma discrepância quanto ao número de ocorrências de cada um deles.

Novo modelo de filtragem

- **Comparação:**
- As assinaturas comparadas são:
  - a gerada em tempo real pelo sistema;
  - as indexadas de uma base de dados (SPAMs).
- Para cada caractere é determinada uma distância relativa de ocorrências:

$$\sum_{i=1}^{68}\left(1-\frac{menor}{maior}\right)\times\frac{1}{68}$$

- Cada caractere contribuirá para um valor final de distanciamento entre as assinaturas comparadas.
- Este valor final é comparado a um limiar escolhido para classificação (o limiar é escolhido com base em testes).

Novo modelo de filtragem

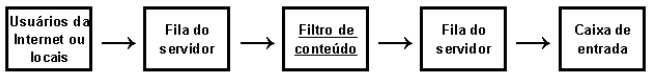
- **Comparação**:
  - Mensagem recebida é analisada e uma assinatura é gerada;
  - Busca-se no banco de dados as assinaturas que possuem 20% mais e 20% menos caracteres, com base no total de caracteres;
  - Cada assinatura buscada é comparada a da nova mensagem, sendo gerada uma distância;
  - Se para alguma assinatura a distância é maior que o limiar escolhido, a mensagem é considerada SPAM;
  - Uma marcação é inserida no cabeçalho da mensagem para que ela seja diferenciada durante a entrega pelo MDA;

Agenda

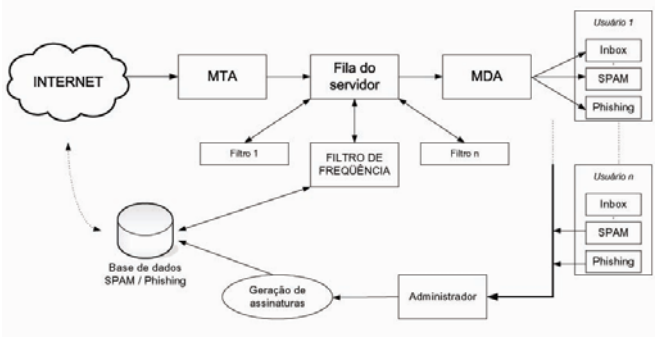
- Introdução
- Novo modelo de filtragem
- **Arquitetura de implantação do modelo**
- Resultados obtidos
- Conclusões e trabalhos futuros

Arquitetura de implantação do modelo

- O modelo é classificado como um filtro após enfileiramento (*after queue*):



- A arquitetura proposta para utilização do modelo é a seguinte:



Agenda

- Introdução
- Novo modelo de filtragem
- Arquitetura de implantação do modelo
- **Resultados obtidos**
- Conclusões e trabalhos futuros

Resultados obtidos

- Protótipo desenvolvido em C e base de dados no MySQL.
- Atualmente o sistema trabalha integrado ao Postfix (MTA bastante utilizado mundialmente).
- Testes quanto ao desempenho computacional:
  - Leitura de um conjunto de mensagens, geração das assinaturas e inserção em uma base de dados.

Conjuntos					
	Total de mensagens	Inválidas	Válidas	Com imagens	Tamanho
SPAM	4545	4545	0	1175	41 MB
INBOX	1376	7	1369	25	102 MB

	Velocidade do processador	Memória dos sistemas	Tipo de disco (E/S)	Tempo médio de inserção da caixa SPAM	Tempo médio de verificação da caixa SPAM	Tempo médio de verificação da caixa INBOX
Hardware 1	800 Mhz	758 MB	ATA 100	23s	3min48s	1min15s
Hardware 2	3 Ghz	1 GB	SATA	13s	1min24s	30s

Resultados obtidos

- Testes quanto às taxas de erro:
  - Taxa de falso-negativo no conjunto SPAM:
    - Das 4545 mensagens avaliadas no conjunto SPAM, 6 falso-negativos, o que representa uma taxa de erro de 0.13%.
    - Casos de mensagens de corpo vazio.
  - Taxa de falso-positivo no conjunto INBOX:
    - Das 1376 mensagens analisadas obtivemos 9 falso-positivos, representando uma taxa de 0.65%.
    - Destes erros, 7 foram pela classificação errada de imagens e 2 de parte textual da mensagem;
    - Interessante: detectou as 7 mensagens inválidas que estavam junto com as válidas, sem que tivessem assinaturas específicas.

Agenda

- Introdução
- Novo modelo de filtragem
- Arquitetura de implantação do modelo
- Resultados obtidos
- **Conclusões e trabalhos futuros**

Conclusões e trabalhos futuros

- Os principais objetivos do modelo foram alcançados:
  - Baixa taxa de erros na classificação;
  - Baixa carga computacional (velocidade de classificação).
- A arquitetura de implantação é interessante quanto à possibilidade de compartilhamento de informações.
- Esta arquitetura pode diminuir a janela de vulnerabilidade de usuários de e-mails contra *Phishings* populares.
  - Baseada no compartilhamento rápido de informações.
- Possibilidade de lidar com imagens e outros padrões MIME;

Conclusões e trabalhos futuros

- Trabalhos futuros:
  - Aprimoramento dos algoritmos de comparação.
    - Utilização de sistemas inteligentes para lidar com “semelhanças”.
  - Criação das metodologias e regras para compartilhamento de informações.
  - Utilização de um decodificador MIME externo.

Obrigado!

- Isabela Liane de Oliveira
  isabela@acmesecurity.org
  PGP KeyID: 0xDA503117
- Adriano Mauro Cansian
  adriano@acmesecurity.org
  PGP KeyID: 0x3893CD2B
- Jorge Luiz Corrêa
  jorge@acmesecurity.org
  PGP KeyID: 0x1BCB7255