Analysis of Security Messages Posted on Twitter

Luiz Arthur F. Santos, Rodrigo Campiolo
Universidade Tecnológica Federal do Paraná
Campo Mourão - PR, Brasil
luizsan@ime.usp.br
campiolo@ime.usp.br

Marco Aurélio Gerosa, Daniel Macêdo Batista
IME - Universidade de São Paulo
São Paulo - SP, Brasil
gerosa@ime.usp.br
batista@ime.usp.br

***Abstract*—The fast spread of computer security alerts, like vulnerabilities notifications, applications updates and threats of attacks, is essential to efficient reactive measures against security incidents. This paper presents an empirical study that evaluates the efficiency of using twitter posts, related to computer security, as notifications of security alerts. The justification to this study is the fact that posts from microblogs have a very fast spread over the Internet. By using similarity analysis and clustering between twitter posts and alerts from specialized sites, it was verified that microblogs spread computer security alerts in a reliable way, reach a high level of dissemination and inform about security threats before some specialized sites.**

***Keywords*—information retrieval; social network analysis; microblogs; security.**

I. Introdução

Um problema na área de segurança da informação é a demora na propagação de informações sobre novas ameaças. Em contrapartida, as redes sociais são famosas pela rápida propagação de informações. Nesse contexto, o presente artigo apresenta um estudo empírico para verificar se as mensagens postadas no Twitter[1] apresentam indicativos de possíveis ameaças à segurança.

Os problemas que afetam a segurança da informação normalmente estão ligados à disponibilidade, confidencialidade e integridade dos dados [1]. Tais problemas podem ser obra de *hackers* ou códigos maliciosos que exploram vulnerabilidades presentes nos sistemas operacionais, aplicativos, hardware e/ou pessoas que utilizam os computadores. Há diversos aplicativos especializados na área de segurança para corrigir, detectar e avisar sobre as ameaças de segurança [2], [3], [4]. Mesmo assim, não são totalmente eficazes contra as novas ameaças que são descobertas constantemente. O problema é amenizado se alertas de segurança são informados em tempo compatível com o surgimento de uma ameaça. Uma solução é o uso de redes sociais para propagação de alertas.

Segundo Russell [5] as redes sociais são um fenômeno global e estão alterando o modo como as pessoas interagem dentro e fora da rede. As redes sociais possibilitam aos seus usuários o compartilhamento de diversos tipos de informações, desde relatos sobre o cotidiano das pessoas, fofocas sobre celebridades, até notícias importantes e de última hora [6], tal como a ocorrência de desastres naturais [7]. Por exemplo, o microblog Twitter não atua somente como rede social, mas também como fonte de notícias [8]. É claro que da mesma forma que os usuários das redes sociais podem disseminar informações relevantes eles também podem utilizar a popularidade das redes para espalhar notícias falsas e *spams* [9].

Neste artigo é analisado um conjunto de mensagens do microblog Twitter para inferir se mensagens de redes sociais ajudam na identificação e alerta antecipado de possíveis problemas de segurança em ambientes computacionais. Como contribuições têm-se a verificação de colaboração em mídias sociais em relação à segurança computacional e a caracterização das mensagens de segurança. Em um contexto mais amplo, essas contribuições são importantes para implementar mecanismos automatizados de alertas sobre possíveis ameaças à segurança de computadores.

II. Trabalhos Relacionados

As redes sociais possibilitam a colaboração e a difusão rápida de informações entre os usuários. Lerman e Ghosh [10] analisaram como as redes sociais afetam o fluxo da informação. Verificaram que histórias propagam-se lentamente no início devido à estrutura de rede esparsa do Twitter, mas mantêm taxa de propagação constante e, consequentemente, podem atingir alta taxa de disseminação. Da mesma forma, Ye e Wu [11] constataram que as mensagens do Twitter são replicadas rapidamente e alcançam um alto nível de propagação, embora a maioria das conversas durem um curto período de tempo. Kwak et al. [12] verificaram que cada mensagem retransmitida no Twitter atinge 1000 usuários independente do número de seguidores do usuário original. Além disso, uma vez retransmitida, as demais retransmissões alcançam rápida difusão. Essa característica das redes sociais as tornam uma possível ferramenta para a disseminação de alertas de segurança.

[1] http://www.twitter.com

978-0-7695-4890-6/12 $26.00 © 2012 IEEE
DOI 10.1109/SBSC.2012.10

É importante que os alertas de segurança sejam gerados prematuramente e por isto não sejam dependentes apenas de mídias oficiais. Pesquisas já avaliaram a propagação de informações em mídias mais modernas, tal como redes sociais, em outras áreas como no caso de terremotos, maremotos e acidentes. Sakaki et al. [13] analisaram mensagens do Twitter como um mecanismo para detecção e alerta de terremotos. Verificaram que as notificações enviadas pelo Twitter são mais rápidas que as repassadas por difusão via Agência Meteorológica. Cha et al. [14] após analisar dados do Twitter identificaram que as mídias de massa possuem um alto índice de audiência. Usuários comuns ou especialistas em algum assunto também propagam notícias mesmo antes de serem notificadas por mídias oficiais.

No mesmo contexto de nosso trabalho, há estudos que abordam o processamento de mensagens do Twitter como forma de obter conhecimento sobre uma determinada área, como por exemplo, detecção de *malwares*, *spams*, recomendações de *hashtags* e predição de eventos como eleições. Phuvipadawat et al. [6] propuseram um método para agrupar e monitorar notícias de última hora no Twitter baseado na similaridade definida pelo esquema TF-IDF (*Term Frequency-Inverse Document Frequency*), aperfeiçoado por um fator de identificação de substantivos importantes nas mensagens, o que possibilita obter melhores resultados no agrupamento de mensagens com pouco texto. Stringhini [9] coletou e agrupou dados de campanhas de *spam* considerando várias redes sociais, usando perfis preparados como alvos de *spams*. As características analisadas nas mensagens foram a taxa de solicitação de amizade, quantidade de URL (*Uniform Resource Locator*) nas mensagens, similaridade de mensagens enviadas por um usuário, número de mensagens enviadas e o número de amigos. O resultado foi um classificador que identificou um número significativo de perfis comprometidos.

Do ponto de vista da segurança da informação, os microblogs são um meio eficiente de propagação de códigos maliciosos e *spams*, visto que há uma relação de confiança entre os usuários. Grier et al. [15] analisaram 25 milhões de URLs postadas no Twitter e descobriram que 8% apontavam para *phishing*, *malware* e *spams*, já identificados em listas de bloqueio tradicionais. Entretanto, o uso de listas de bloqueio é uma solução lenta para identificar novas ameaças. Um resultado interessante para nossa pesquisa é que 0,14% das mensagens de *spam* são de antivírus, ou seja, poderão gerar falsos positivos de alertas contra vírus. Nosso trabalho se difere dos existentes por caracterizar mensagens como alertas de segurança e não apenas verificar se as mensagens são *spams* ou direcionam a códigos maliciosos.

Normalmente, a geração de alertas na área de segurança de computadores é responsabilidade de sistemas de detecção de intrusão, os quais têm se aperfeiçoado para alcançarem um grau de excelência na identificação de invasões. Um fator importante para gerar alertas é obter informações confiáveis e completas sobre atividades maliciosas no sistema ou na rede. Entretanto, coletar dados confiáveis é uma atividade complexa [16]. Baseado na necessidade de um conjunto de dados suficiente e confiável para realizar análise e correlação com possíveis ameaças de segurança, algumas abordagens atuais de detecção de intrusão têm investido na colaboração entre entidades[17]. Pesquisas defendem a troca de informação entre organizações como forma de prevenção e detecção antecipada de ameaças à segurança, propondo uma arquitetura que viabiliza cooperação em nível de coleta e compartilhamento de repositório de alertas entre as organizações [18], [19], [20]. As abordagens apresentadas são direcionadas à colaboração entre organizações e não trata as informações e as ameaças a usuários finais. Todas as abordagens dependem de um contrato prévio entre as organizações para a geração de alertas.

Nosso trabalho propõe um estudo empírico das mensagens de segurança nas redes sociais para verificar se constituem alertas de segurança e se são amplamente difundidas ou detectadas nos canais de informações tradicionais.

III. Referencial Teórico

O presente artigo aborda o uso do Twitter e de notícias de mídias especializadas como fontes para coleta de informação de alertas de segurança. Além disso, realiza a mineração dos dados usando técnicas de agrupamento e similaridade. Nesta seção, são apresentados conceitos, recursos e terminologias usados em nossos estudos.

A. Twitter

O Twitter provê a ideia de rede social por meio de um serviço de microblog, permitindo aos seus usuários postarem mensagens (*tweets*) de até 140 caracteres [12], [5]. No Twitter as mensagens recebidas podem ser replicadas, esta prática é conhecida como *retweet* [11]. Mensagens que são reenviadas ganham notoriedade, pois alcançam mais pessoas que uma mensagem comum, já que os seguidores do usuário que postou originalmente a mensagem e os seguidores da pessoa que reenviou recebem a mensagem automaticamente. Alguns autores consideram que mensagens que sofrem *retweet* tem grande chance de serem importantes [8]. Nas mensagens, o texto após o símbolo @ indica menção a um usuário e após o símbolo #, conhecido como *hashtag*, é utilizado para enfatizar o assunto da mensagem. O Twitter também organiza uma lista com os assuntos mais discutidos, denominada de *trend topics*.

Atualmente, o Twitter é muito utilizado para disseminação e buscas de informações sobre acontecimentos atuais [6], [5]. Morris et al. [8] mostraram que o Twitter é fonte de informação principalmente sobre notícias da atualidade e não deve ser desconsiderado quando o assunto é busca por informação.

B. RSS e Atom

Na Web, os sítios divulgam seus conteúdos recentes usando recursos como RSS (*Really Simple Syndication/Rich Site Summary*) [21] e Atom [22]. Ambos são arquivos no formato XML utilizados para descrever conteúdos publicados em sítios Web, também denominados de *feeds*. Usuários interessados em acessar o conteúdo de um sítio, tornam-se assinantes usando a URL do *feed* para monitorar a publicação de novos conteúdos. No RSS ou Atom, as principais informações disponibilizadas são título, descrição, data de publicação e link para a notícia. Um software cliente, denominado agregador, é responsável por visitar periodicamente os *feeds* assinados e verificar se há atualizações. Sítios especializados em segurança usam *feeds* para notificar a publicação de notícias recentes de segurança, como alertas contra vírus, vulnerabilidades, atualizações de segurança, entre outros assuntos.

C. Agrupamento e Similaridade com Apache Lucene

O Apache Lucene[2] é um software que realiza processamento de documentos por meio de similaridade. O funcionamento básico do Lucene consiste em processar e indexar qualquer tipo de texto. O processo de indexação armazena o texto para que possa ser recuperado posteriormente por um motor de busca do próprio Lucene. A base de dados dos textos indexados pelo Lucene ajuda a recuperar informações referentes a similaridades de textos, que é uma prática muito utilizada em mineração de dados e recuperação de informação [5].

O cálculo para similaridade dos textos indexados no Lucene combina o Modelo *Booleano* (*Boolean Model*) com o Modelo de Espaço Vetorial (*Vector Space Model*). Nesse caso, um documento aceito pelo Modelo *Booleano* por possuir os termos buscados, será submetido ao Modelo de Espaço Vetorial, que por sua vez devolverá um escore (um número) dizendo o quão similar é o documento em relação ao termo utilizado na busca. O Modelo de Espaço Vetorial utiliza cálculos de similaridade como TF-IDF e similaridade de cosseno (*Cosine Similarity*) para fornecer o escore [23]. Na fórmula utilizada pelo Lucene é empregado o cálculo simplificado de similaridade de cosseno e variáveis de ajustes de pontuação, como por exemplo, pesos para termos de busca, amenização de parágrafos repetidos, entre outros. Ao submeter uma consulta ao motor de busca do Lucene, obtém-se as tuplas: documento e escore. O escore elevado representa alto grau de similaridade e o escore baixo representa pouca similaridade dos documentos em relação à busca. Dessa forma, podemos agrupar documentos com alta similaridade entre si [24].

IV. Método da Pesquisa

A pesquisa tem por objetivo analisar as mensagens de redes sociais, neste caso o Twitter, para verificar se as informações de segurança postadas viabilizam a identificação de possíveis problemas de segurança em ambientes computacionais.

Baseado no objetivo foram definidas as seguintes hipóteses:

- H1. Há informações sobre segurança de computadores nas mensagens do Twitter.
- H2. As mensagens do Twitter com conteúdo sobre segurança de computadores indicam ameaças potenciais.
- H3. O Twitter informa antes de sítios especializados os problemas relacionados à segurança da informação.
- H4. Os usuários no Twitter se preocupam em alertar outros usuários sobre problemas de segurança.

O processo de investigação e processamento das informações é apresentado na Figura 1. Basicamente, a arquitetura é composta por quatro etapas:

1) Capturar, armazenar e indexar no Lucene as mensagens sobre segurança de computadores postadas no Twitter;
2) Capturar, armazenar e indexar as notícias sobre segurança de computadores contidas em sítios especializados da Internet;
3) Agrupar os *tweets* por similaridade e selecionar os grupos com grandes quantidades de *tweets*, visando obter as mensagens mais relevantes;
4) Correlacionar os *tweets* relevantes com o conteúdo dos sítios especializados, apontando as mensagens mais importantes.

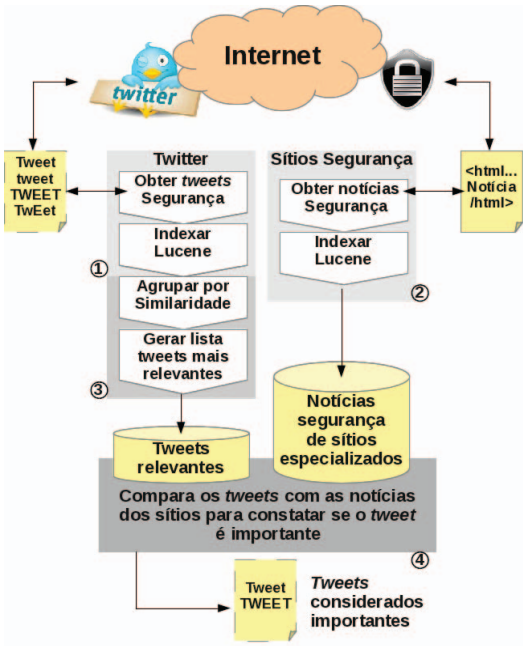

Figura 1. Arquitetura proposta para identificar *tweets* importantes sobre segurança de computadores.

[2]http://lucene.apache.org/core/

As etapas descritas na arquitetura são detalhadas nas subseções seguintes.

A. Coleta de dados no Twitter

A coleta de mensagens do Twitter foi realizada por meio de um software desenvolvido pelos autores a partir de uma API (*Application Programming Interface*) do Twitter[3]. As consultas no Twitter são restritas e devolvem mensagens de até uma semana anterior à data da busca. Então, optou-se por realizar a consulta em intervalos periódicos de 1 minuto. O software realiza a busca, descarta os *tweets* repetidos e armazena os novos *tweets* em uma base de dados. As buscas por *tweets* de segurança foram realizadas através de duas consultas, sendo a primeira utilizando os termos: *"seguranca AND (virus OR worm OR ataque OR intrusao OR invasao OR ddos OR hacker OR cracker OR exploit OR malware)"*, a segunda busca é um subgrupo da primeira, pois inicialmente não se conhecia qual seria o volume de *tweets* retornados e temendo um volume muito grande decidiu-se reduzir ainda mais o universo de busca, utilizando os termos: *"seguranca AND virus"*. Visando comparar regionalidade e ainda preocupando-se com o volume de *tweets* retornados (neste caso se fosse muito pequeno) repetimos as duas consultas utilizando os mesmos termos em inglês. Estas buscas foram: *security AND (virus OR worm OR attack OR intrusion OR invasion OR ddos OR hacker OR cracker OR exploit OR malware)* e *security AND virus*. Basicamente os termos de busca obrigam que todos os *tweets* capturados tenham a palavra segurança seguida de um ou mais termos comuns a área de segurança de computadores. É importante observar que a busca utilizando a API do Twitter ignora acentos. Portanto, uma busca por "seguranca" retornará *tweets* escritos corretamente utilizando o termo "segurança".

O software de coleta de *tweets* sobre segurança foi executado de 28/04/2012 a 19/05/2012, ou seja, durante 21 dias. No processo de coleta foram obtidos 12.309 *tweets* sobre segurança (ver Tabela I), sendo que esses *tweets* foram gerados por 8.376 usuários. Aproximadamente 88% dos *tweets* continham URLs, 35% continham *hashtags* (#) e 36% mencionavam (@) outros usuários no corpo da mensagem.

Tabela I
DADOS DOS TWEETS OBTIDOS DE 28/04/2012 A 19/05/2012

Busca	tweets	usuários	com link	#	@
Vírus (pt)	223	198	177	46	96
Vírus (eng)	2.070	1.473	1.690	587	452
Termos (pt)	817	666	708	161	400
Termos (eng)	11.492	7.710	10.104	4.218	4.109
Total*	**12.309**	**8.376**	**10.812**	**4.379**	**4.509**

* Soma de termos de segurança (pt) e termos de segurança (eng)

Ao final da coleta de dados no Twitter foi constatado que existem mensagens relacionadas ao assunto segurança de computadores.

[3]foi utilizada a API Twitter4j - http://twitter4j.org

B. Coleta de notícias de segurança

Para avaliar nossas hipóteses precisamos correlacionar os dados coletados no Twitter com as informações disponibilizadas nas mídias tradicionais. Para tal, foi realizado um processo de coleta e armazenamento de notícias a partir de *feeds* de sítios especializados em segurança.

A coleta das notícias de segurança foi realizada por um software desenvolvido pelos autores usando a API de manipulação de *feeds* Informa[4]. Foram selecionados 30 sítios especializados em notificações de segurança, de fontes como Centros de Estudo e Tratamento de Incidentes de Segurança (CERTs), fabricantes de sistemas operacionais e de antivírus, e sítios tradicionais em notificações de segurança. As informações coletadas foram armazenadas em uma base de dados, pois alguns sítios mantém apenas os *feeds* mais recentes.

Durante um mês foram coletados 3.988 *feeds* sobre segurança (Tabela II). As informações essenciais para análise são o título, a descrição, o link e a data de publicação da notícia. A princípio, pretendia-se encontrar a similaridade entre os *tweets* com o título e descrição dos *feeds*. Entretanto, experimentos realizados constataram que a quantidade de dados provida pelos *feeds* era insuficiente para obter uma correlação entre ambos.

Tabela II
FEEDS COLETADOS DE MÍDIAS ESPECIALIZADAS

	Total	Ausência de descrição	Ausência de data
Feeds	3988	31	121

Para contornar a situação, desenvolvemos um *Web crawler* para recuperar a notícia completa a partir dos *feeds*. O software recupera a notícia, elimina o conteúdo indesejado, como marcações, links e imagens, e por fim, armazena o conteúdo de cada página em um arquivo nomeado com a identificação do *feed*.

Há dois fatores negativos nos dados coletados. O primeiro é a ausência de data de publicação (Tabela II), o que causou problemas de processamento e comparação com as datas dos *tweets*. Adotou-se a convenção de usar uma data comum de 01 de janeiro de 2000 nos *feeds* sem data. O segundo é que algumas notícias coletadas não são relevantes como alertas de segurança. A decisão foi mantê-las na base de dados e analisar em conjunto com os *tweets* para verificar sua influência nos resultados, visto que para aplicações futuras o processo deve se manter automatizado.

C. Filtragem, indexação e agrupamento dos dados

Nesta seção é abordado o processamento dos dados coletados para a realização da análise. O conteúdo obtido do Twitter foi filtrado, indexado e agrupado. As notícias foram apenas submetidas ao processo de indexação, visto que o conteúdo relevante das páginas foi selecionado durante a coleta.

[4]http://informa.sourceforge.net

Os *tweets* foram submetidos a um processo de filtragem para remover conteúdo irrelevante, como por exemplo, *tweets* com menos de três palavras. O filtro implantado foi simples pois nesta pesquisa é importante verificar a influência de ruído nos dados.

A indexação e agrupamento foram realizados usando o Apache Lucene. A indexação considerou apenas o corpo das mensagens; os metadados, como identificação e data, foram apenas armazenados. Após a indexação dos *tweets* foi possível realizar buscas por similaridade na base de dados coletada. A estratégia de agrupamento consistiu em comparar cada *tweet* na base com todos os outros, ou seja, o termo de busca é a mensagem de cada *tweet*. Dessa forma, a partir de um determinado grau de similaridade é possível considerar que os *tweets* abordem um mesmo assunto ou similar. Um assunto foi considerado relevante caso gerasse um grupo com vários elementos [8].

O software de agrupamento por similaridade produz uma lista de *tweets* que representam assuntos relevantes. Como resultado tem-se duas variáveis importantes: grau de similaridade e o número de *tweets* similares. A variável de grau de similaridade indica o nível de agrupamento por similaridade. Se ajustada para um valor alto tende a produzir mais grupos e aumenta a probabilidade de grupos distintos abordando o mesmo assunto. Se ajustada para um valor baixo tende a produzir menos grupos e aumenta a probabilidade de mensagens de contextos diferentes no mesmo grupo.

Após a realização de vários experimentos, foi considerado um grau de similaridade médio de 0,5. Dessa forma, evita-se que assuntos de contextos diferentes sejam agrupados, mas ainda mantém um número de grupos significativo, apesar da geração de grupos distintos que abordam o mesmo assunto.

Os *tweets* foram agrupados conforme o mesmo assunto e os mais comentados foram armazenados em uma base de dados. O próximo passo consistiu em analisar se os *tweets* selecionados são realmente importantes quando o assunto é segurança em computadores.

D. Comparação dos tweets com as notícias

A comparação entre os *tweets* com as notícias de sítios especializados é a parte mais importante do processo, pois possibilita identificar se as mensagens podem ser consideradas alertas. O cruzamento de informações também possibilita comparar se os *tweets* sobre segurança são publicados antes de algumas mídias oficiais. Além disso, possibilita avaliar a qualidade do agrupamento e similaridade com as notícias.

Para identificar a importância do *tweet* como alerta de segurança, os *tweets* agrupados na fase anterior foram submetidos como termo de consulta para as notícias de segurança indexadas pelo Apache Lucene. Para cada *tweet* foram geradas três avaliações de grau de similaridade e uma amostra foi selecionada para análise e classificação manual. A amostra é composta por 7,2% das mensagens com maior número de menções, 7,2% das mensagens com menor número de menções e 7,2% das mensagens intermediárias. As mensagens intermediárias foram selecionadas por uma função de geração de números pseudo-aleatórios.

A verificação de *tweets* publicados antes de sítios especializados foi realizada comparando a data do *tweet* com a data da notícia associada ao *tweet*. Foi gerado um arquivo contendo a data mais recente de um *tweet* de cada grupo e a data da notícia de maior similaridade com o *tweet*. O arquivo foi importado para uma planilha eletrônica para a contabilização e verificação dos dados.

Por fim, para analisar de forma mais ampla os resultados, correlacionamos os títulos das notícias com os *tweets* agrupados, a fim de verificar se *tweets* pouco mencionados são importantes.

V. Análise dos Dados

Nesta seção é apresentada a caracterização dos resultados, a discussão dos valores obtidos por meio de uma amostragem, a avaliação dos procedimentos e das hipóteses.

A. Caracterização dos resultados

A partir do universo de *tweets* (Tabela I) foi gerada uma lista dos 10 termos mais frequentes (Tabela III) em português, em inglês e os mais frequentes desconsiderando os termos de busca.

Tabela III
Palavras mais usadas pelos tweets de segurança

Português*		**Inglês****		***Principais Termos***	
Qtd	**Termos**	**Qtd**	**Termos**	**Qtd**	**Termos**
219	hacker	3.459	malware	704	cyber
147	vírus	3.078	attack	702	infosec
120	invasão	1.392	hacker	673	new
108	malware	1.188	exploit	632	sites
95	ataque	1.076	virus	590	anti
89	mil	704	cyber	550	android
81	_info	702	infosec	470	firm
79	twitter	673	new	465	amp
76	falsos	632	sites	457	apple
73	dados	590	anti	457	flash

* *tweets* em português; ** *tweets* em inglês.

Observa-se na Tabela III que o termo *hacker* é o mais citado, o que representa 27% dos *tweets* em português. Em inglês, o termo mais usado é *malware*, representando 27% dos *tweets*. Os primeiros cinco termos apresentados nos *tweets* em português e inglês são os mesmos utilizados para a busca no passo de coleta de *tweets* sobre segurança.

Na coluna *Principais Termos* da Tabela III não são considerados os termos usados na coleta. Investigando os termos, foi constatado que *cyber* é usado em muitos contextos relacionados ao mundo virtual, por exemplo, *cyber*-guerra. O termo *infosec* é uma contração de *information security* (segurança da informação), usado frequentemente como *hashtag* para mensagens com conteúdo sobre segurança de computadores ou da informação, um possível termo para ser adicionado em coletas futuras. O termo *new* indica assuntos como o surgimento de um novo código malicioso. O termo

site frequentemente indica a presença de URL nas mensagens. O termo Android[5] pode ser indicativo do aumento do uso de dispositivos móveis e possíveis ameaças de segurança para eles. A palavra *firm* aparece em muitos *tweets* como empresas alertando sobre problemas de segurança. O termo *apple* aparece nos principais tópicos devido aos comentários sobre um vírus para os sistemas da Apple.

Cruzando informações da Tabela I e Tabela III observa-se que há muitos *tweets* que abordam termos comuns a segurança de computadores ou segurança da informação. Analisando os números de *tweets* em inglês, verifica-se uma média de 547 mensagens por dia, enquanto em português verifica-se uma média de 39. Baseado nesse indicativo, realizando a filtragem e agrupamento considerando grau de similaridade 0,5 e grupos com 10 ou mais elementos, obtém-se o resultado apresentado na Tabela IV.

Tabela IV
TWEETS RELEVANTES

segurança (en)	segurança (pt)	vírus (en)	vírus (pt)
278	19	44	4

Devido ao baixo número de *tweets* em português, foi utilizado como base de análise das hipóteses deste trabalho apenas os *tweets* gerados pela busca por termos de segurança em inglês.

O processo de agrupamento e filtragem de *tweets* gerou 278 mensagens de alertas. Para propósitos de análise, essas mensagens foram classificadas manualmente pelos autores em:

- Importantes: mensagens consideradas alertas.
- Irrelevantes: mensagens sobre segurança em geral.
- *Spams*: mensagens e propagandas comerciais.
- Outros: mensagens fora de contexto.

A Tabela V apresenta a classificação manual de todos os 278 agrupamentos de *tweets* e o resultado da correlação por similaridade entre um *tweet* representante de cada grupo com as mídias especializadas. Foi adotado empiricamente o valor de 0,2 como fronteira entre baixo e alto grau de similaridade.

Tabela V
CLASSIFICAÇÃO DOS TWEETS APÓS AGRUPAMENTO

	Tweets	**Similaridade acima de 0,2**	**Similaridade abaixo de 0,2**
Importantes	119	69	50
Irrelevantes	88	31	57
Spams	30	15	15
Outros	41	8	33
Total	278	123	155

Na Tabela V observa-se que foram classificados corretamente 69 de 119 alertas considerados importantes e 31 de 88 que abordavam segurança e por isso tinham relação com as notícias. Os falsos positivos, no caso, os *spams* e outros com similaridade acima de 0,2, representam menos de 20%. Esse resultado é influenciado pelo fato da base de *feeds* conter referências a fabricantes de sistemas operacionais e antivírus e por limitações de filtragem. Considerando os *tweets* importantes com similaridade abaixo de 0,2, verifica-se que 50 de 119 (42%) não foram correlacionados com notícias. Esse resultado é influenciado pelo fato da notícia não estar cadastrada na base de *feeds*, por características das mensagens do Twitter ou porque o alerta foi divulgado no Twitter e não por sítios especializados tradicionais.

[5]http://www.android.com/

B. Discussão dos resultados

Nesta seção são discutidos os pontos positivos, limitações e exemplos de mensagens por meio de amostras extraídas dos 278 grupos relevantes.

A Tabela VI apresenta a ordenação, a quantidade de elementos no grupo e as mensagens selecionadas para discussão.

Tabela VI
AMOSTRA DE TWEETS RELEVANTES (INGLÊS)

Pos[*]	***Tweets***	**Trechos da mensagem**
1	**347**	...Religious Sites Carry More Malware Than Porn Sites...
2	**266**	Adobe releases Flash exploit. Update yours now!...
3	**263**	...ARE WE PREPARED FOR CYBERWAR?...
4	**229**	Adobe issues security update for Flash player, warns against IE exploit...
5	**205**	Flashback malware exposes big gaps in Apple...
8	**146**	The Pirate Bay hit by DDoS attack...
10	**134**	About AVG...Anti-Virus Software...
24	**84**	Android Trojan copies PC drive-by malware attack...
32	**61**	Obama Defends Attack On Romney...
90	**31**	Oracle discloses new zero day exploit...
173	**18**	...New Zeus malware scam promises rebates...
210	**15**	...Microsoft Patch Tuesday Swats 23 Security Bugs, Including Duqu Exploit...
278	**10**	...Ancient Microsoft Word malware threat returns from the grave...

[*] Posição do grupo ordenada pela quantidade de *tweets* similares.

No universo de mensagens do Twitter escritos em inglês o maior grupo obtido é composto de 347 *tweets* semelhantes. Esse grupo alerta que sítios religiosos podem ser tão perigosos para a segurança quanto sítios pornos, pois podem conter *malwares*. A primeira mensagem do grupo é: "*@Ketan Chopda|Religious Sites Carry More Malware Than Porn Sites, Security Firm Reports: The annual Internet Security Threat R... http://t.co/Mrw6iOWT*". A última mensagem é: "*@Darren Anthony|Religious Sites Carry More Malware Than Porn Sites, Security Firm Reports http://t.co/VJMBxUxR @pcworld*". Pode-se notar que o conteúdo das mensagens são muito similares, o que indica coerência na nossa metodologia de agrupamento das mensagens, basicamente o que difere as mensagens são URLs e menções a usuários (@) do Twitter. Os grupos

também foram inspecionados manualmente e foi verificado que as mensagens descrevem o mesmo assunto.

Os *tweets* do grupo 1 (Tabela VI) começaram a ser propagados em 30/04/2012, perderam sua intensidade em 07/05/2012 e apareceram pela última vez em 13/05/2012, ou seja, a mensagem que sítios religiosos podem possuir *malwares* se propagou por 14 dias (ver Figura 2). A mesma mensagem sobre sítios religiosos e pornográficos também ficou como o sexto maior grupo, com a mensagem: "*Relgious Sites Are Greater Security Risk Than Porn Destinations [STUDY] - Mashable http://t.co/Dy9XNnMP*". Isto ocorreu porque a palavra *Religious* está escrita errada (*Relgious*), fazendo com que o grau de similaridade do grupo tenham diminuído em relação ao primeiro. Mesmo com o erro de escrita, o grupo obteve 195 *tweets* similares (também escritos errados). Isto mostra que alguns usuários de redes sociais podem propagar mensagens com erros de escrita. Ainda sobre o assunto de sítios religiosos e pornográficos com *malware* foi encontrado um grupo com 13 mensagens, representado pela mensagem: "*Religion, porn and malware: Behind the headline http://t.co/4JsYV90i*", esta mensagem possui o mesmo conteúdo das anteriores e só foi resumida. Como os três grupos tratam do mesmo problema podemos somá-los, o que dá 555 mensagens. Utilizando um escore de 0,15, em testes posteriores, conseguiu-se agrupar essas mensagens, mesmo com erro de escrita.

Observa-se na Tabela VI que existem diversos tipos de mensagens. Há mensagens de alerta, segurança, *spams* e assuntos fora do contexto. As mensagens dos *tweets* dos grupos 2 e 4 abordam um problema com o Adobe Flash alertando para que os usuários atualizem o software. O grupo 3 indaga se estamos preparados para uma possível *cyber*-guerra. O grupo 5 alerta sobre um *malware* chamado Flashback que contamina computadores da Apple. O grupo 8 informa sobre um ataque DDoS. O grupo 10 é uma propaganda sobre antivírus e pode ser considerada *spam*. O grupo 84 relata a ação de um *trojan* que afeta o Android. Já o grupo 32 diz respeito a segurança, mas não de computadores, pois replica uma informação do presidente dos Estados Unidos da América. No grupo 90 temos o relato sobre problemas de segurança de uma grande empresa de informática. Os grupos 173, 210 e 278 alertam sobre *malwares* e correções de segurança. É importante notar que mesmo com um número reduzido de *tweets* similares, se comparados aos demais, a mensagem ainda pode ser relevante.

As mensagens selecionadas mostram a existência de alertas de segurança postadas no Twitter e que esses alertas podem ser utilizados para reforçar as informações sobre possíveis ameaças para ambientes computacionais.

C. Avaliação dos procedimentos

Para analisar os procedimentos e a qualidade dos resultados, foram selecionadas 60 amostras de 278 *tweets* e realizada a verificação de cada *tweet* com o conteúdo das notícias na base, ou quando não encontrada, por meio de busca na Internet.

Verificou-se que 22% dos *tweets* analisados são informações irrelevantes, por exemplo, com mensagens mal formadas. 10% são *spams* sobre ferramentas de segurança, tal como antivírus. 7% são informações relacionadas a segurança, mas não de computadores. E a grande maioria, 62%, são informações relevantes à segurança e poderiam ser utilizadas por muitos administradores como alertas de possíveis ameaças.

D. Avaliação das hipóteses

Depois de coletar e analisar os dados é possível avaliar as hipóteses levantadas e verificar se as mensagens sobre segurança postadas no Twitter podem ser utilizadas como alertas de segurança.

1) Avaliação da Hipótese H1: A hipótese H1 afirma que existem informações sobre segurança de computadores nas mensagens do Twitter. A hipótese foi confirmada pela análise dos dados apresentados nas tabelas I e V. A Tabela I apresenta o resultado da coleta de mensagens no Twitter. Como resultado, tem-se 12.309 *tweets* sobre segurança coletados em 21 dias, ou seja, uma média de 586 mensagens por dia. A Tabela V apresenta o resultado da correlação das notícias de segurança de sítios especializados com os conteúdos dos *tweets*. Como resultado, tem-se o indicativo de que 74% dos *tweets* abordam assuntos de segurança. Um resultado inesperado foi a baixa quantidade de *tweets* de segurança no idioma português.

2) Avaliação da Hipótese H2: A hipótese H2 afirma que as mensagens do Twitter com conteúdo sobre segurança de computadores indicam ameaças potenciais. A hipótese foi avaliada considerando a similaridade entre as mensagens do Twitter com o de sítios especializados. A Tabela V indica que 42% dos *tweets* se relacionam com alertas de segurança. Foram encontrados grupos de mensagens pouco relevantes, pois não abordam diretamente segurança de computadores ou são *spam*. Aperfeiçoando a filtragem e utilizando sistemas anti-*spam* pode-se evitar estas mensagens.

Para verificar *tweets* importantes que não apresentaram muitas menções, o título dos *feeds* foi comparado com a base de *tweets*. Como resultado foi verificado que há elementos que se relacionam diretamente com as notícias.

Dessa forma, confirma-se a hipótese H2 de que há mensagens importantes em redes sociais para indicar ameaças de segurança.

3) Avaliação da Hipótese H3: A hipótese H3 afirma que o Twitter informa sobre alertas de segurança antes de sítios especializados. A hipótese foi avaliada correlacionando os dados dos *tweets* agrupados com o conteúdo das notícias obtidas a partir dos *feeds*. As datas do Twitter e dos *feeds* foram comparadas considerando apenas as mensagens classificadas como importantes, ou seja, que caracterizam um alerta de segurança.

Verificou-se que 45% dos *tweets* apresentam data mais recente que uma notícia publicada em uma mídia tradicional. Devido a falta de datas em algumas notícias de sítios/*feeds* não é possível afirmar com certeza que esses alertas foram notificados primeiro no Twitter. Entretanto, é um indicativo positivo para afirmar que a propagação dos *tweets* de alerta de segurança ocorrem antes que algumas mídias tradicionais.

Para apoiar esse indicativo, monitoramos uma notificação de alerta recente sobre a vulnerabilidade *"PHP-CGI query string parameter vulnerability"* publicada em 03/05/2012. No Twitter, o primeiro comentário capturado sobre o assunto teve início em 04/05/2012. A vulnerabilidade foi catalogada na base de dados do NIST[6] apenas em 11/05/2012.

Baseado nos dados analisados, há mensagens em redes sociais propagadas antes ou na mesma data de alguns sítios especializados.

4) Avaliação da hipótese H4: A hipótese H4 afirma que os usuários do Twitter se preocupam em alertar outros usuários sobre problemas de segurança. A hipótese foi avaliada baseada em dois parâmetros: o número de usuários que uma mensagem de alerta atinge e a forma como é propagada a mensagem.

A Figura 2 mostra a média do tempo de propagação das mensagens apresentadas na Tabela VI. Pode-se observar que a maioria das mensagens tem seu ápice no segundo dia a partir de sua publicação [10]. Algumas mensagens ainda são brevemente mencionadas alguns dias depois. O tempo médio de propagação das mensagens é de 12 dias.

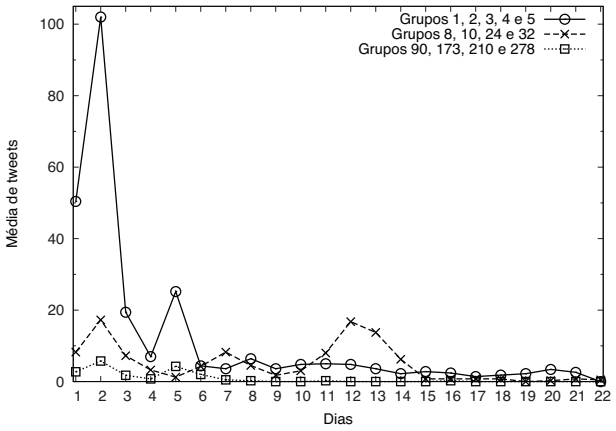

Figura 2. Linha de tempo (em dias) da propagação dos *tweets*.

O índice de propagação de mensagens entre as pessoas que utilizam o Twitter é expressivo. Por exemplo, considerando a mensagem da posição 278 (ver Tabela VI) que tem o menor número de *tweets* semelhantes dentre as amostras, a mensagem alcançou 9.291 usuários do Twitter, pois esse é o número total de seguidores dos usuários que postaram essa mensagem (verificando o que é comprovado por [12]). Então, a mensagem com mais menções, postada por 347 usuários do Twitter, pode ter alcançado aproximadamente 340.000 pessoas.

Dessa forma, confirma-se a hipótese H4 de que os usuários preocupam-se em avisar outros usuários sobre problemas de segurança.

[6]http://www.nist.gov

E. Limitações

Há fatores que influenciaram os resultados, mas não prejudicaram a análise das hipóteses. No processo de coleta de dados de mensagens do Twitter, outros termos de pesquisa por alertas de segurança foram identificados e poderiam ter retornado mais resultados se utilizados, como por exemplo, *infosec* e *patch*. O agrupamento dos *tweets* e a correlação com as notícias foram influenciados pelo texto das mensagens, devido ao uso de abreviações e erros de digitações. A classificação manual das mensagens em alertas foi realizada por apenas dois especialistas na área, no caso, autores do artigo, o que pode implicar em alguns vieses. Uma filtragem de *tweets* mais rigorosa diminuiria os *spams* e provavelmente melhoraria o resultado do agrupamento. Para confirmar que essas questões não afetariam os resultados, amostras foram selecionadas e analisadas minuciosamente pelos autores.

VI. Conclusões

A análise das mensagens sobre segurança de computadores postadas no Twitter confrontadas com o conteúdo de sítios especializados da Internet confirmou que redes sociais são capazes de gerar alertas de segurança. Os alertas de segurança gerados por redes sociais têm a tendência de serem noticiados assim que detectados e de forma rápida, pois em redes sociais as pessoas colaboram e disseminam informações. O formato, tamanho das mensagens e *spams* coletados no Twitter geraram alguns agrupamentos não uniformes e dificultaram a identificação de alertas. Mesmo assim, os resultados revelaram um alto índice de alertas de segurança. Como trabalhos futuros pretende-se: efetuar novas consultas usando outros termos da área de segurança, realizar a avaliação de alertas de segurança em outras redes sociais e desenvolver um sistema automatizado de alertas antecipados de segurança baseado em redes sociais.

Referências

[1] A. Tanenbaum, *Computer Networks*, 4th ed. Prentice Hall Professional Technical Reference, 2002.

[2] OSSEC. (2012, Junho) Open source host-based intrusion detection system. Acessado em 20 de junho de 2012. [Online]. Available: http://www.ossec.net/

[3] Snort. (2012, Junho) Snort. Acessado em 20 de junho de 2012. [Online]. Available: http://www.snort.org/

[4] Tenable Network Security. (2012, Junho) Nessus. Acessado em 20 de junho de 2012. [Online]. Available: http://www.tenable.com/products/nessus

[5] M. A. Russell, *Mining the Social Web - Analyzing Data from Facebook, Twitter, LinkedIn, and Other Social Media Sites*. O'Reilly, 2011.

[6] S. Phuvipadawat and T. Murata, "Breaking news detection and tracking in twitter," in *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*, vol. 3, 31 2010-sept. 3 2010, pp. 120 –123.

[7] Y. Qu, C. Huang, P. Zhang, and J. Zhang, "Microblogging after a major disaster in china: a case study of the 2010 yushu earthquake," in *Proceedings of the ACM 2011 conference on Computer supported cooperative work*, ser. CSCW '11. New York, NY, USA: ACM, 2011, pp. 25–34.

[8] M. R. Morris, S. Counts, A. Roseway, A. Hoff, and J. Schwarz, "Tweeting is believing?: understanding microblog credibility perceptions," in *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, ser. CSCW '12. New York, NY, USA: ACM, 2012, pp. 441–450.

[9] G. Stringhini, C. Kruegel, and G. Vigna, "Detecting spammers on social networks," in *Proceedings of the 26th Annual Computer Security Applications Conference*, ser. ACSAC '10. New York, NY, USA: ACM, 2010, pp. 1–9.

[10] K. Lerman and R. Ghosh, "Information contagion: an empirical study of the spread of news on digg and twitter social networks," *CoRR*, vol. abs/1003.2664, 2010.

[11] S. Ye and S. F. Wu, "Measuring message propagation and social influence on twitter.com," in *Proceedings of the Second international conference on Social informatics*, ser. SocInfo'10. Berlin, Heidelberg: Springer-Verlag, 2010, pp. 216–231.

[12] H. Kwak, C. Lee, H. Park, and S. Moon, "What is twitter, a social network or a news media?" in *Proceedings of the 19th international conference on World wide web*, ser. WWW '10. New York, NY, USA: ACM, 2010, pp. 591–600.

[13] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes twitter users: real-time event detection by social sensors," in *Proceedings of the 19th international conference on World wide web*, ser. WWW '10. New York, NY, USA: ACM, 2010, pp. 851–860.

[14] M. Cha, F. Benevenuto, H. Haddadi, and K. Gummadi, "The world of connections and information flow in twitter," *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, vol. PP, no. 99, pp. 1 –8, 2012.

[15] C. Grier, K. Thomas, V. Paxson, and M. Zhang, "@spam: the underground on 140 characters or less," in *Proceedings of the 17th ACM conference on Computer and communications security*, ser. CCS '10. New York, NY, USA: ACM, 2010, pp. 27–37.

[16] R. Kemmerer and G. Vigna, "Intrusion detection: a brief history and overview," *Computer*, vol. 35, no. 4, pp. 27 –30, apr 2002.

[17] M. Locasto, J. Parekh, A. Keromytis, and S. Stolfo, "Towards collaborative security and p2p intrusion detection," in *Information Assurance Workshop, 2005. IAW '05. Proceedings from the Sixth Annual IEEE SMC*, june 2005, pp. 333 – 339.

[18] M. Apel, J. Biskup, U. Flegel, and M. Meier, "Towards early warning systems - challenges, technologies and architecture," in *CRITIS*, 2009, pp. 151–164.

[19] U. Flegel, J. Hoffmann, and M. Meier, "Cooperation enablement for centralistic early warning systems," in *Proceedings of the 2010 ACM Symposium on Applied Computing*, ser. SAC '10. New York, NY, USA: ACM, 2010, pp. 2001–2008.

[20] B. Grobauer, J. I. Mehlau, and J. Sander, "Carmentis: A co-operative approach towards situation awareness and early warning for the internet," in *IMF*, 2006, pp. 55–66.

[21] RSS Advisory Board. (2012, Maio) Rss 2.0 specification. Acessado em 10 de maio de 2012. [Online]. Available: http://www.rssboard.org/rss-specification

[22] M. Nottingham and R. Sayre, "The Atom Syndication Format," RFC 4287 (Proposed Standard), Internet Engineering Task Force, Dec. 2005, updated by RFC 5988. [Online]. Available: http://www.ietf.org/rfc/rfc4287.txt

[23] lucene.apache.org. (2012, Maio) Similarity. Apache Software Foundation. Acessado em 12 de maio de 2012. [Online]. Available: http://lucene.apache.org/core/3_6_0/api/core/org/apache/lucene/search/Similarity.html

[24] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge, UK: Cambridge University Press, 2008.