# Análise e extração de alertas antecipados sobre ameaças e incidentes de segurança em sistemas computacionais usando fontes de dados não estruturados

Rodrigo Campiolo

Tese apresentada
ao
Instituto de Matemática e Estatística
da
Universidade de São Paulo
para
obtenção do título
de
Doutor em Ciências

Programa: Pós-Graduação em Ciência da Computação

Orientador: Prof. Dr. Daniel Macêdo Batista

Durante o desenvolvimento deste trabalho o autor
recebeu auxílio financeiro da Fundação Araucária

São Paulo, agosto de 2016

# Análise e extração de alertas antecipados sobre ameaças e incidentes de segurança em sistemas computacionais usando fontes de dados não estruturados

Esta é a versão original da tese elaborada pelo
candidato (Rodrigo Campiolo), tal como
submetida à Comissão Julgadora.

# Resumo

CAMPIOLO, R. **Análise e extração de alertas antecipados sobre ameaças e incidentes de segurança em sistemas computacionais usando fontes de dados não estruturados**. 2016. 153 f. Tese (Doutorado) - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2016.

A sofisticação dos ataques, ameaças dia zero e o grande volume de dados em redes de computadores impõem desafios para os mecanismos tradicionais de segurança (sistemas de detecção de intrusão, filtros de acesso, analisadores de fluxo, entre outros). A ineficácia desses mecanismos tem dificultado, para administradores de redes e especialistas de segurança, a proteção dos recursos computacionais das organizações e da Internet. Mesmo quando um ataque é detectado em uma organização, a cooperação limitada com outras organizações e a falta de mecanismos eficientes para a propagação de alertas acabam tornando-se um empecilho para evitar que outros potenciais alvos sejam atacados. Os Sistemas de Alerta Antecipado são sistemas que procuram detectar e prever possíveis ataques ou alvos e, disseminar alertas rapidamente, com o intuito de possibilitar uma reação proativa ou premente aos incidentes de segurança. Neste trabalho, é proposta uma abordagem colaborativa para a identificação de alertas antecipados em fontes de dados não estruturados, em especial, mídias sociais que disponibilizam dados publicamente, por meio da proposta de um arcabouço que descreve um processo de análise e mecanismos para a extração de alertas associados à cibersegurança. Os resultados do uso de técnicas de Recuperação de Informação, Processamento de Linguagem Natural e heurísticas validaram o Twitter, IRC e Facebook como fontes de informações relevantes para a detecção e propagação de alertas. Analisando as mensagens postadas no Twitter potencialmente relacionadas com segurança, verificou-se que cerca de 92% delas abordam tópicos em segurança de redes de computadores e que mais de 50% representam potenciais alertas. No IRC e Facebook, foram identificadas orquestrações de ataques, discussões de potenciais alvos, ataques em andamento ou recém realizados. Também foi comprovado que sistemas de classificação e recomendação são necessários para otimizar a seleção e priorização de alertas coletados nas fontes de dados não estruturados. Além disso, o arcabouço norteou o desenvolvimento de diversos mecanismos que encontram-se em produção em um sistema para detecção antecipada de incidentes de segurança na rede acadêmica brasileira. Esta pesquisa avança o estado da arte em Sistemas de Alerta Antecipado porque descreve e valida um arcabouço para a análise e extração de alertas a partir de fontes de dados heterogêneas e com dados não estruturados, em destaque, pelo uso de mídias sociais como fontes de informação.

**Palavras-chave:** sistemas de alerta antecipado, detecção de intrusão, mídias sociais, mineração de dados.

# Abstract

CAMPIOLO, R. **An approach to cybersecurity early warning systems using unstructured data sources**. 2016. 153 f. Tese (Doutorado) - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2016.

Sophisticated attacks, zero-day threats and the large volume of network data pose challenges to traditional security mechanisms. Early warning systems are based on cooperation and collaboration to address these issues. They can detect and predict threats and disseminate security notifications to partners or publicly. This work proposes a framework to analyze and extract potential alerts of cibersecurity from unstructured data sources. Publicly available sources are analyzed and validated to act as networks sensors aiming more effective detections and disseminations of new security threats. The proposed framework enables data integration from heterogeneous sources and the use of different approaches to aggregate and correlate alerts in distributed and collaborative environments. Results validate Twitter, IRC e Facebook as sources of relevant information to detection and propagation of security alerts. By analysing the messages posted at Twitter pottentially related to security, it was found that about 92% of them address topics related to computer security and more than 50% represent potential alerts. Further more, attacks orquestrations, potential targets, ongoing cyber attacks were identified in Facebook and IRC. It was also verified that classification and recommendation systems are required for selection and prioritization of alerts collected from unstructured data sources. Thus, the unstructured data sources should be explored to act as new network sensors and, thereby, to provide new mechanisms for early detection of incidents. This research proposal advances the state of the art in Early Warning Systems because it explores the recovery of security alerts from publicly available sources, especially from social medias.

**Keywords:** early warning systems, intrusion detection, social medias, data mining.

# Sumário

# Lista de Abreviaturas

| | |
|---|---|
| API | *Application Programming Interface.* |
| APT | *Advanced Persistent Threats.* |
| ATLAS | *Active Threat Level Analysis System.* |
| CAIS | Centro de Atendimento a Incidentes de Segurança. |
| CAP | *Common Alerting Protocol.* |
| CERT.br | Centro de Estudos, Resposta e Tratamento de Incidentes de Segurança no Brasil. |
| CEWS | *Cibersecurity Early Warning System.* |
| CIDS | *Collaborative Intrusion Detection Systems.* |
| CRF | *Conditional Random Field.* |
| CSIRT | *Computer Security Incident Response Team.* |
| CSV | *Comma Separated Values.* |
| CVE | *Common Vulnerabilities and Exposures.* |
| CVRF | *Common Vulnerability Reporting Framework.* |
| CybOX | *Cyber Observable eXpression.* |
| DDoS | *Distributed Denial of Service.* |
| DOMINO | *Distributed Overlay for Monitoring InterNet Outbreaks.* |
| ENISA | *European Union Agency for Network and Information Security.* |
| EWS | *Early Warning Systems.* |
| HIDS | *Host-based Intrusion Detection Systems.* |
| HTTPS | *Hyper Text Transfer Protocol Secure.* |
| IAS | *Internet Analysis System.* |
| IDES | *Intrusion Detection Expert System.* |
| IDMEF | *Intrusion Detection Message Exchange Format.* |
| IFS | *Intrusion Forecasting System.* |
| IMS | *Internet Motion Sensor.* |
| IODEF | *Incident Object Description Exchange Format.* |
| IRC | *Internet Relay Chat.* |
| ISC | *Internet Storm Center.* |
| JSON | *JavaScript Object Notation.* |
| MMEP | Médias Móveis Exponencialmente Ponderadas. |
| NIDS | *Network-based Intrusion Detection Systems.* |
| NSA | *National Security Agency.* |
| NVD | *National Vulnerability Database.* |
| OpenIOC | *Open Indicators of Compromise.* |

| | |
|---|---|
| OWL | *Web Ontology Language*. |
| PLN | Processamento de Linguagem Natural. |
| PoS | *Part-of-Speech*. |
| PRODAM | Processamento de Dados do Amazonas. |
| RDF | *Resource Description Framework*. |
| RNP | Rede Nacional de Pesquisa. |
| STIX | *Structured Threat Information eXpression*. |
| SVM | *Support Vector Machine*. |
| TAXII | *Trusted Automatic Exchange of Indicator Information*. |
| TF-IDF | Term Frequency–Inverse Document Frequency. |
| UFBA | Universidade Federal da Bahia. |
| USP | Universidade de São Paulo. |
| UTFPR | Universidade Tecnológica Federal do Paraná. |

# Lista de Endereços Web

| | |
|---|---|
| Facebook | https://www.facebook.com/ |
| Twitter | https://www.twitter.com/ |
| Google+ | https://plus.google.com/ |
| LinkedIn | https://www.linkedin.com/ |
| MySpace | https://myspace.com/ |
| Sina Weibo | http://sinaweibo.com/ |
| Tumblr | https://www.tumblr.com/ |
| Anonops | https://anonops.com/ |
| Skype | http://www.skype.com/ |
| WhatsApp | https://www.whatsapp.com/ |
| OWASP | https://www.owasp.org/ |
| US-CERT | https://www.us-cert.gov/ |
| Symantec | https://www.symantec.com/security_response/ |
| Cisco | http://blogs.cisco.com/ |
| Microsoft | https://blogs.microsoft.com/cybertrust/ |
| Snort | http://blog.snort.org/ |
| Mcafee | https://blogs.mcafee.com/ |
| KrebsonSecurity | http://krebsonsecurity.com/ |
| Threat Post | https://threatpost.com/ |
| Hack Forums | http://www.hackforums.net/ |
| Anonymous BR | http://forum.anonymousbrasil.com/ |
| Caveira Tech | http://caveiratech.com/forum/ |
| PasteBin | http://pastebin.com/ |
| Slexy | http://slexy.org/ |
| Pastie | http://pastie.org/ |
| Codepad | http://codepad.org/ |
| Github Gist | https://gist.github.com/ |
| YouTube | https://www.youtube.com/ |
| Vimeo | https://vimeo.com/ |

Endereços Web acessados em 01/07/2016.

# Lista de Figuras

# Lista de Tabelas

# Capítulo 1

# Introdução

Independente das suas atividades-fim, pode-se dizer que praticamente todas as organizações hoje em dia dependem da disponibilidade e da normalidade de operação dos sistemas computacionais e da infraestrutura de rede para realizarem seus negócios. Além disso, muitas organizações também utilizam os sistemas computacionais para proteger suas informações sensíveis e sigilosas. Ataques aos sistemas computacionais, à infraestrutura de rede ou o acesso não autorizado às informações implicam em graves prejuízos financeiros. Estima-se que, globalmente, os prejuízos anuais devido ao assim chamado cibercrime estejam na casa dos bilhões de dólares (Anderson *et al.*, 2013).

Apesar de boa parte das corporações investir em mecanismos para prevenção, detecção e reação a ataques, nem sempre eles têm sido suficientes para garantir a segurança das informações e evitar os danos produzidos, por exemplo, por uma invasão ou uma tentativa de invasão. Muitas vezes, o período de tempo necessário para aplicar medidas reativas não é o adequado para obter resultados efetivos contra os ataques. Por exemplo, uma falha em um serviço de rede de uma organização é descoberta por um sistema de detecção de intrusão e, para prevenir a exploração, os administradores da rede ativam um mecanismo para bloquear conexões suspeitas. Em seguida, as autoridades e o fornecedor do produto são contatados para propagar a notícia e prover uma atualização que corrija a falha do serviço. O tempo acumulado entre a detecção, mitigação do problema, aviso às autoridades, propagação da notícia da falha e reação de outros usuários do serviço, nos dias atuais é consideravelmente grande (Shahzad *et al.*, 2012; Trustwave, 2014). Nesse intervalo, o invasor poderia comprometer muitas outras infraestruturas de redes.

Para piorar o cenário, apesar dos mecanismos de detecção de intrusão, *honeypots*, análise de fluxo, *firewalls*, entre outros, terem evoluído e obtido sucesso contra muitas ameaças, inclusive com a detecção de padrões de ataques de negação de serviço distribuídos – *Distributed Denial of Service* (DDoS), os ataques evoluíram também, por exemplo, por meio do uso de *botnets* que mascaram um DDoS como uma atividade de usuário normal (Kirubavathi e Anitha, 2014) e são capazes de sobrecarregar até mesmo sistemas projetados para suportar uma grande quantidade de tráfego e acessos. Além disso, as atividades ilícitas na Internet também estão cada vez mais organizadas e os códigos maliciosos têm se mostrado cada vez mais complexos e persistentes com a finalidade de evadir sistemas de detecção de intrusão (Corona *et al.*, 2013; Virvilis e Gritzalis, 2013).

Apesar de mostrar bons resultados contra certos ataques, a confiança única e exclusivamente em análises de dados coletados por sensores na rede local, como registros de *firewalls*, fluxos de rede e alertas de sistemas de detecção de intrusão pode ser arriscada (Meng *et al.*, 2015). Nem sempre essas fontes possibilitam concluir ou detectar uma ameaça de ataque real ou reagir a tempo de evitar maiores danos. Também não possibilitam notificar a possibilidade real de ataque a outras organizações antes que realmente o ataque tenha sido identificado, pois gerariam muitos falsos positivos. Para contornar essa situação, muitas pesquisas na área de segurança estão sendo direcionadas a proativamente combater e mitigar as chamadas ameaças cibernéticas.

Neste sentido, são propostos os Sistemas de Alerta Antecipado, em inglês, *Early Warning Systems* (EWS), que visam detectar e prever ameaças de ataques a partir do comportamento dos sistemas, gerando alertas de situações que apresentam padrões de risco, com o intuito de desencadear mecanismos reativos antecipadamente, evitando ou diminuindo os danos causados por um ataque. Em síntese, os Sistemas de Alerta Antecipado permitem estabelecer hipóteses e predições correlacionando informações incertas e incompletas providas por sensores em uma rede (Biskup *et al.*, 2008).

Na área de Segurança de Redes de Computadores e Sistemas Computacionais, as arquiteturas de Sistemas de Alerta Antecipado têm sido projetadas baseadas na colaboração e cooperação. Apel *et al.* (2009); Grobauer *et al.* (2006) justificam que a troca cooperativa de informações provê vantagens na detecção de ataques, pois as evidências de ataque em uma rede provavelmente são similares a padrões ou ataques identificados em outras redes. No entanto, limitações em compartilhar informações, principalmente considerando questões de sigilo e credibilidade, continuam como um obstáculo. Logo, uma forma de colaborar na identificação de ameaças à segurança é explorar novas fontes de informações. Neste caso, explorar o uso de informações de origens abertas e não estruturadas, como redes sociais, bases de dados de vulnerabilidades, fóruns, entre outras, que em conjunto com as fontes tradicionais, possibilitariam estabelecer cenários de ameaças mais contextualizados e, dessa forma, prover tempo de reação antecipado ou mais rápido a intrusões.

Nesta tese, é proposta uma abordagem colaborativa e distribuída para a identificação de alertas antecipados em fontes de dados não estruturados, em especial, fontes que disponibilizam dados publicamente (redes sociais, fóruns, blogs, entre outras), por meio da proposta de um arcabouço que descreve um processo de análise e mecanismos para a extração de alertas associados à cibersegurança. Também são analisadas e validadas fontes de dados não estruturados, como sensores de redes, para detectar e disseminar mais efetivamente alertas de ameaças à segurança. Além disso, é mostrado como o arcabouço pode ser implantado em uma arquitetura que integra dados de diferentes origens e explora o uso de diferentes mecanismos para filtrar, agregar e classificar alertas em ambientes distribuídos e colaborativos.

## 1.1 Considerações preliminares

A segurança em redes de computadores não pode apenas depender de alertas gerados por sensores de redes tradicionais, como registros de *firewalls*, sistemas de detecção de intrusão ou monitoramento de tráfego de rede, devido à sofisticação dos ataques, ameaças dia zero e volume massivo de dados. Algumas soluções têm sido desenvolvidas para correlacionar informações de diferentes origens (Elshoush e Osman, 2011; Mirheidari *et al.*, 2013; Salah *et al.*, 2013), mas não são efetivas na disseminação de alertas e muitas vezes limitadas a monitoramento específico ou soluções proprietárias.

Os EWS geram e propagam alertas baseados em padrões de risco com a finalidade de prover mecanismos proativos de reação e mitigar o dano causado por um ataque. Conforme afirmado por Apel *et al.* (2009); Grobauer *et al.* (2006), a colaboração é um fator decisivo na criação de um EWS. No entanto, há muitas questões em aberto nas arquiteturas de EWS. Destacam-se os desafios para a padronização e classificação de grandes volumes de dados coletados em organizações, a garantia de confidencialidade das informações compartilhadas, a disseminação rápida de alertas de novas ameaças e a correlação de diferentes tipos de sensores.

Em diversas partes do mundo, projetos têm sido propostos por organizações governamentais e privadas para detectar antecipadamente potenciais ameaças cibernéticas (Bourgue *et al.*, 2013; CERT.br, 2014; DShield, 2014; Symantec, 2014). A preocupação das nações com essas ameaças tem aumentado principalmente após o Stuxnet (2010), um código malicioso que afeta infraestruturas críticas. Mais recentemente, a segurança dos serviços e dados na Web foi abalada devido à vulnerabilidade denominada de Heartbleed (2014), que afeta a biblioteca criptográfica OpenSSL

e permite o acesso a informações críticas, como chaves criptográficas dos serviços e usuários. O Heartbleed permaneceu não detectado por cerca de dois anos. Por essas razões, a detecção antecipada de ameaças é essencial para combater ameaças cibernéticas e tem se mostrado viável, quando informações de diversas fontes estão disponíveis e são correlacionadas.

O escândalo de espionagem da *National Security Agency* (NSA), afora as questões legais e éticas, mostrou a viabilidade de correlacionar dados coletados da rede global, desde que o acesso às informações seja possível. Com uma infraestrutura organizada e alto poder de computação, foi divulgado pela mídia que a NSA coletava e extraia informações para identificar ameaças à segurança, além de manter uma base de perfis de indivíduos e organizações. Destacam-se o uso de replicadores de tráfego em fibras óticas para copiar tráfego em pontos estratégicos, o acesso a dados armazenados por grandes empresas de informática e telecomunicação (vigilância não autorizada), o acesso (ilegal) a redes e máquinas na Internet. As técnicas consistiam em inspeção detalhada de pacotes, buscando palavras, padrões e conexões suspeitas[1].

A União Europeia conta com a *European Union Agency for Network and Information Security* (ENISA) para atuar proativamente contra as ameaças à segurança na Europa. A ENISA desenvolve importantes ações colaborativas para aprimorar os mecanismos de detecção e reação a ameaças à segurança. Ela também atua na cooperação entre o setor público e privado, facilitando o estabelecimento de parcerias e compartilhamento de informações, e atua na elaboração de legislação visando tratar questões relativas ao ciberterrorismo.

Na América, um dos principais esforços em sistemas de alertas antecipados é o *Internet Storm Center* (ISC)[2], um projeto suportado pelo SANS Institute. No ISC são coletados e analisados registros de mais de 50 países. A detecção e relatórios de ameaças e incidentes é realizado por analistas e métodos automatizados que analisam a base de dados do DShield, um sistema de coleta e detecção de ameaças distribuído. Ao detectar uma ameaça, um grupo de especialistas especifica a prioridade e a forma de propagação do alerta, ou ainda, se necessário, o bloqueio de tráfego nos provedores de serviço de Internet.

No Brasil, destaca-se a iniciativa do Centro de Estudos, Resposta e Tratamento de Incidentes de Segurança no Brasil (CERT.br) com o projeto de *honeypots* distribuídos (honeyTARG)[3]. Neste projeto são usados *honeypots* de baixa interatividade para coletar informações sobre ameaças e *spams* na Internet brasileira. A análise das estatísticas possibilita identificar e relacionar as tendências das ameaças às infraestruturas de redes e organizações. São sumariadas diversas informações de fluxos como país de origem, sistemas autônomos, sistemas operacionais, portas de destino e protocolos.

As organizações privadas de segurança da informação, por exemplo, McAfee, Symantec e Cisco, têm apresentado soluções baseadas em reputação para identificar tráfego de fontes suspeitas, inspeção de tráfego detalhada para encontrar códigos maliciosos e a coleta e análise de relatórios de ameaças de clientes para criarem suas próprias redes de alerta antecipado. Entretanto, são soluções comercializadas para público específico e não abrangem a quantidade e diversidade de dados trafegados na Internet.

Portanto, verifica-se um cenário composto por diferentes visões e soluções, mas considerando um inimigo comum. Há a preocupação em colaboração e cooperação, mas há também a dificuldade em realizar a difusão das informações. Também é notável a quantidade massiva de dados para ser analisada e, por consequência, a impossibilidade de identificar todas as ameaças presentes na rede. A era do *Big Data* viabiliza novas fontes de informações para os especialistas de segurança e, também, para os criminosos. Há a necessidade de colaboração entre os governos, organizações e usuários de sistemas computacionais para aprimorar os mecanismos de detecção de ameaças.

No entanto, limitações em compartilhar informações entre organizações, principalmente con-

[1] https://www.eff.org/nsa-spying/how-it-works/ – Acessado em 01/07/2016.
[2] https://isc.sans.edu/ –Acessado em 01/07/2016.
[3] http://honeytarg.cert.br/ – Acessado em 01/07/2016.

siderando questões de sigilo e credibilidade, continuam como um obstáculo. Logo, uma forma de colaborar na identificação de ameaças à segurança é explorar novas fontes de informações. Neste caso, explorar informações não estruturadas e, em especial, as de origens que disponibilizam dados abertos, como algumas redes sociais, bases de dados de vulnerabilidades, fóruns, entre outras, em conjunto com as fontes tradicionais, possibilitaria estabelecer cenários de ameaças mais contextualizados e prover tempo de reação antecipado a intrusões.

## 1.2 Objetivos

### 1.2.1 Objetivo geral

Nesta tese é proposto um arcabouço para a investigação e avaliação de fontes de dados não estruturados (redes sociais, microblogs, fóruns, entre outras) para uso como sensores de redes e para a extração de alertas cibernéticos, preferencialmente antecipados, de forma colaborativa e distribuída. O arcabouço visa descrever uma abordagem para o desenvolvimento de mecanismos que possibilitem priorizar e disseminar alertas de segurança de redes e sistemas computacionais, de forma antecipada ou mais rapidamente, visando mitigar potenciais ameaças ou minimizar o tempo de resposta a incidentes.

### 1.2.2 Objetivos específicos

Os objetivos específicos são:

- Avaliar diferentes fontes de dados de origem aberta como sensores de redes, sobretudo o uso de mídias sociais, e analisar o impacto dessas fontes em EWS voltados à cibersegurança.
- Especificar e avaliar um arcabouço para a análise e extração de alertas cibernéticos, preferencialmente antecipados, a partir de dados não estruturados.
- Desenvolver e avaliar mecanismos que possibilitem a implementação de EWS a partir de uma abordagem colaborativa e distribuída.

## 1.3 Contribuições

As principais contribuições desta tese são:

1. Avaliação de novos sensores de rede para detecção de ameaças de segurança, em destaque, o uso do Twitter, Facebook e IRC como fontes que possibilitam identificar possíveis ameaças e ataques recém realizados.
2. Evidências empíricas do uso de técnicas de recuperação de informação, aprendizagem de máquina, processamento de linguagem natural e recomendação no auxílio à identificação de alertas a partir de dados não estruturados.
3. Um arcabouço para análise e extração de alertas cibernéticos de fontes de dados não estruturados, que possibilita adaptar e reusar elementos (mecanismos, bases de dados, políticas de monitoramento) para diferentes fontes de dados.
4. A implementação de um sistema de alerta antecipado de cibersegurança que coleta, normaliza, filtra e classifica informações de fontes de dados não estruturados em alertas de segurança e, possibilita a colaboração e disseminação de alertas entre analistas e especialistas em segurança.

Essas contribuições foram disseminadas nas seguintes publicações:

- *Analysis of security messages posted on Twitter*, SBSC 2012 (Santos *et al.*, 2012).
- *Evaluating the utilization of Twitter messages as a source of security alerts*, SAC 2013 (Campiolo *et al.*, 2013).
- *Detecção de alertas de segurança em redes de computadores usando redes sociais*, SBRC 2013 (Santos *et al.*, 2013).
- *Uma arquitetura autonômica para detecção e reação a ameaças à segurança em redes de computadores*, WoSIDA (SBRC 2014) (Santos *et al.*, 2014).
- *Análise de mensagens associadas à cibersegurança em redes IRC*, SBSeg 2015 (Campiolo e Batista, 2015).
- *A Collaboration Model to Recommend Network Security Alerts Based on the Mixed Hybrid Approach*, SBRC 2016 (Esposte *et al.*, 2016).
- *Abordagem autonômica para mitigar ciberataques em LANs*, SBRC 2016 (Santos *et al.*, 2016).
- *GT-EWS: Building a Cybersecurity EWS based on Social Networks*, TNC 2016 (Batista *et al.*, 2016).

Além da relação direta com o grupo de trabalho GT-EWS[4] da Rede Nacional de Pesquisa (RNP), que faz uso do arcabouço e de vários dos mecanismos desenvolvidos durante a pesquisa realizada nesta tese.

## 1.4 Organização do trabalho

No Capítulo 2, são apresentados os conceitos, características, desafios, arquiteturas e trabalhos relacionados a EWS. Também são abordados os conceitos fundamentais de técnicas para a análise e identificação de conteúdos de interesse em dados não estruturados. No Capítulo 3, é apresentado o arcabouço proposto para a análise e extração de alertas cibernéticos em fontes de dados não estruturados. Neste capítulo, são discutidos os conceitos, fluxos de processamento e componentes que especificam o arcabouço e como podem ser usados na análise e geração de alertas. No Capítulo 4, são apresentados os resultados e a avaliação do uso do arcabouço em duas fontes de dados, microblog Twitter e na rede IRC. Também é apresentada a proposta de um modelo de recomendação e classificação de alertas cibernéticos a partir de dados não estruturados obtidos em fontes heterogêneas. No Capítulo 5, é apresentado o desenvolvimento, resultados e discussões de um sistema de alerta antecipado para cibersegurança baseado no arcabouço proposto na tese. Além disso, são apresentados os resultados da extração de alertas do Twitter e Facebook. Finalmente, no Capítulo 6, são apresentadas as conclusões e trabalhos futuros.

[4] https://gtews.ime.usp.br/ – Acessado em 01/07/2016.

# Capítulo 2

# Fundamentação Teórica e Revisão da Literatura

Os Sistemas de Alerta Antecipado (EWS) são sistemas que visam detectar situações de risco ou ameaças e emitir alertas para viabilizar a reação antecipada ou mais rápida possível para prevenir, evitar ou minimizar os danos causados por essas situações. Na área de Segurança de Redes de Computadores e de Sistemas Computacionais, os alertas antecipados visam proporcionar a administradores de redes ou especialistas de segurança uma forma de detectar potenciais riscos e ameaças às infraestruturas computacionais e, dessa forma, reagir proativamente ou mais rapidamente a essas questões. Os riscos e ameaças mais comuns estão associados a ataques e orquestrações de ataque a sistemas computacionais, códigos maliciosos, vazamentos ou roubos de informações e vulnerabilidades de software.

Este capítulo apresenta os conceitos fundamentais usados na tese e discute o estado da arte da pesquisa em Sistemas de Alerta Antecipado e áreas afins. A Seção 2.1 discorre sobre a área de detecção de intrusão, caracterizando as direções e os mecanismos durante sua evolução. A Seção 2.2 apresenta os principais conceitos sobre Sistemas de alerta antecipado e como se relacionam com a área de Segurança. Descreve EWS em sistemas computacionais sob a visão de diferentes pesquisadores (Subseção 2.2.1) e mostra a importância de alertas antecipados sobre ameaças e vulnerabilidades de segurança (Subseções 2.2.2 e 2.2.3). A Seção 2.3 apresenta as características importantes em EWS, como a coleta de informação, gerenciamento e correlação de alertas, detecção e predição de ameaças, resposta a incidentes e disseminação de alertas, e os desafios associados a cada atividade, em especial, a privacidade e confiabilidade dos participantes e das informações compartilhadas. A Seção 2.4 discute os conceitos fundamentais associados à mineração de dados não estruturados que serviram de base para a avaliação da proposta da tese. A Seção 2.5 aborda as arquiteturas, propostas e implementações de EWS na literatura, destacando suas características, vantagens e limitações (Subseção 2.5.1). Também apresenta os trabalhos correlatos diretamente associados à proposta desta tese (Subseção 2.5.2), isto é, que exploram a ideia do uso de fontes de dados não estruturados voltados para a segurança. Por fim, apresenta uma síntese e discussões sobre os principais trabalhos relacionados (Subseção 2.5.3).

## 2.1 Evolução de IDS para EWS

Esta seção aborda uma linha do tempo com algumas importantes contribuições para a detecção de intrusão e que resultaram na detecção antecipada ou proativa de incidentes de segurança. A Figura 2.1 ilustra autores, sistemas e as contribuições ou resultados das pesquisas.

Os avanços significativos na detecção de intrusão em sistemas computacionais tiveram início a partir de (Anderson, 1980), que discutiu a importância de analisar registros de sistema para

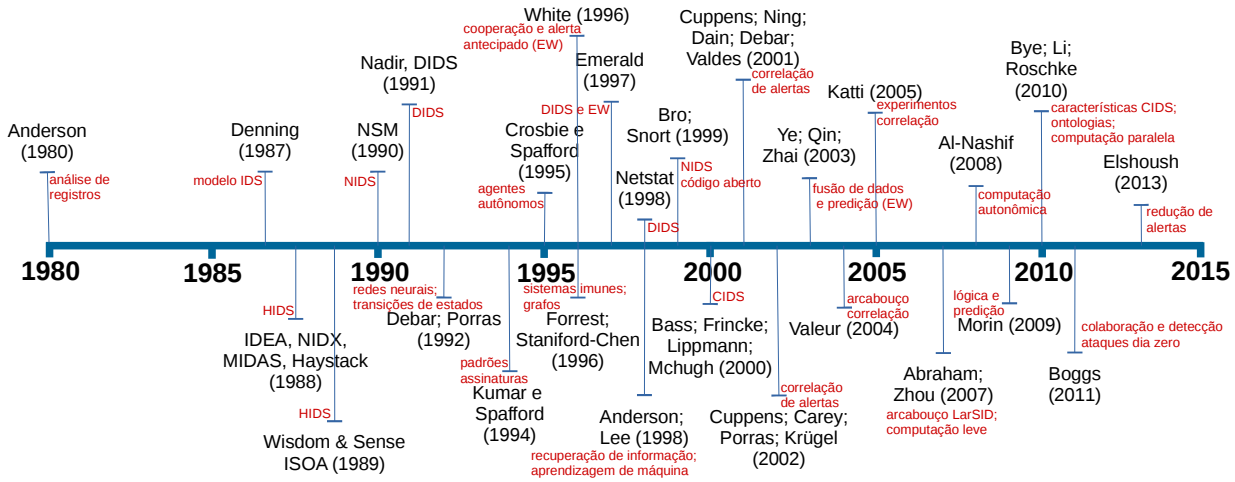


**Figura 2.1:** *Linha do tempo - Evolução IDS para EWS.*

identificar atividades anormais. Anderson propunha o uso de parâmetros estatísticos para definir normalidade no sistema considerando os perfis de usuários legítimos como base. Denning (1987) apresentou um modelo de detecção de intrusão baseado nos padrões anormais de uso do sistema para identificação de possíveis ameaças. O modelo foi implementado no *Intrusion Detection Expert System* (IDES) (Lunt e Jagannathan, 1988; Lunt *et al.*, 1989), um sistema de detecção baseado em hospedeiro - *Host-based Intrusion Detection Systems* (HIDS) - que emprega detecção de anomalias usando abordagem estatística e baseada em regras.

Na mesma época, outros HIDS foram desenvolvidos e tiveram contribuições relevantes para avanços em IDS. O NIDX (Bauer e Koblentz, 1988) é baseado no IDES, contudo é dependente de plataforma, no caso, sistemas UNIX. O NIDX considera que técnicas de invasão e vulnerabilidades são específicas para o sistema alvo. As regras são especificadas conforme o tipo de ação maliciosa que pode afetar cada recurso. O MIDAS (Sebring *et al.*, 1988) foi desenvolvido para o Multics e também é baseado no IDES. Possui heurísticas para monitorar comportamento anormal do usuário e do estado global do sistema. O Haystack (Smaha, 1988) realiza o sumário de registros de auditoria e o modelo de detecção é baseado no comportamento do usuário e grupos de usuários. O Wisdom & Sense (Vaccaro e Liepins, 1989) monitora comportamento anormal examinando os dados históricos de usuário e do sistema. O ISOA (Winkler e Page, 1989) centraliza a análise de registros e correlaciona informações para realizar a detecção. O modelo de agregação de registros de sistema em tópicos elimina a redundância nos dados.

Considerando a evolução e a crescente agregação de computadores em redes, foram propostos os sistemas de detecção de intrusão baseados em rede - *Network-based Intrusion Detection Systems* (NIDS). O NSM (Heberlein *et al.*, 1990) é um NIDS que monitora o tráfego da rede em tempo real e compara com regras e o histórico de uso dos recursos da rede para detectar padrões de ameaças à segurança. O modelo possibilita uma análise hierárquica em três níveis de atividades: máquinas, serviços e conexões. Bro (Paxson, 1999) é um NIDS projetado para monitorar passivamente a rede por meio de uma arquitetura em camadas simples. É composto por monitoramento de rede filtrando fluxos de interesse definidos por uma política. Snort(Roesch, 1999) é um NIDS de código aberto e idealizado para operar em redes menores. Apresenta uma arquitetura composta por três subsistemas: decodificador de pacotes, módulo de detecção, subsistema de registro e alerta. O Snort apresenta simplicidade e flexibilidade na criação e adição de novas regras.

Os primeiros sistemas que coletavam dados distribuídos para realizar detecção tiveram início na década de 90. NADIR (Hochberg *et al.*, 1993; Jackson *et al.*, 1991) é um sistema automatizado de relatórios baseado em regras que analisa a atividade em nível de rede, em estações dedicadas, por meio da coleta de registros dos serviços de rede. Adota uma abordagem em fases constituídas por filtros para a detecção, além de agrupar, reduzir e correlacionar os perfis. Considera importantes

questões de processamento de dados (falta de informação, organização idiossincrática, racionalização, fusão, agregação e redução). O DIDS (Snapp *et al.*, 1991) é um sistema de detecção de intrusão distribuído que agrega e correlaciona dados de múltiplas máquinas e segmentos de rede. A análise de dados é realizada por sistema especialista centralizado baseado em regras e com aprendizagem simples. O modelo de detecção é hierárquico e dividido em seis camadas: registros, eventos independentes, identificação única de sujeitos, contexto espacial e temporal, agregação em tópicos e estado de segurança. NetStat (Vigna e Kemmerer, 1998) adota a abordagem de análise de transição de estados para detectar intrusões em redes por meio do monitoramento e análise de conexões ativas, estados das interações e valores de tabelas usadas no gerenciamento de comunicações em redes, como tabelas de roteamento, tabelas ARP, entre outras. Uma das dificuldades dessa abordagem é a dependência do administrador na descrição dos cenários e ataques.

Aos poucos, novas técnicas de detecção começaram a ser experimentadas nos IDS. Debar *et al.* (1992) utilizam redes neurais como um componente para modelar o comportamento do usuário. Ilgun (1993); Ilgun *et al.* (1995); Porras e Kemmerer (1992) modelam intrusões como uma sequência de transições de estado. A observação de mudança de estados nos sistemas indica possíveis ameaças. A premissa essencial foi avaliar as ações críticas que devem ocorrer para caracterizar um ataque. Os autores afirmam que o modelo é interessante para detectar ataques cooperativos. Kumar e Spafford (1994) apresentam algoritmos para encontrar padrões para alguns tipos de assinaturas usadas por IDS e apresentam um algoritmo baseado em redes de Petri coloridas. Crosbie e Spafford (1995) propõem o uso de agentes autônomos para identificar intrusões. Os agentes monitoram o comportamento do sistema e cooperam entre si para identificar anomalias. Dasgupta (1999); Forrest *et al.* (1996) abordam o uso de sistemas naturais imunes para modelar IDS. Staniford-Chen *et al.* (1996) apresentam uma abordagem de grafos de atividade para representar atividades na rede considerando ambientes de grande escala. Utilizam um esquema de agregação hierárquica e uma heurística de detecção que reconhece grafos como potenciais ameaças quando ultrapassam um limiar. Kosoresow e Hofmeyer (1997) realizam a detecção de comportamentos anômalos monitorando as chamadas de sistema, mapeando para uma representação reduzida e avaliando com um autômato finito determinístico que representa chamadas normais. Anderson e Khattak (1998) usam técnicas de recuperação de informação para detectar intrusões considerando o histórico de comandos dos usuários. Lee e Stolfo (1998) utilizam aprendizagem de máquina para criar classificadores de tráfego normal e malicioso. Verificaram que características estatísticas e temporais, obtidas por algoritmos de associação de regras e de episódios frequentes, podem melhorar a acurácia do modelo de classificação.

A importância da colaboração para a detecção e reação a intrusão foi uma questão natural considerando o aumento da transição de redes locais para redes globais altamente interconectadas. White *et al.* (1996) abordam essa preocupação e propõem um modelo de detecção cooperativa baseado em pares e extensível para limites externos à rede local. Apresentam um protótipo de componente para detectar intrusões localmente e propagar ações suspeitas para outros pares visando uma detecção proativa. Porras e Neumann (1997) descrevem EMERALD, um sistema de detecção distribuído para redes de grande escala. EMERALD provê um arcabouço para detecção e resposta antecipada a ações maliciosas em uma rede corporativa constituída de domínios independentes. Os módulos de monitoramento propagam notificações assíncronas para os assinantes. Utiliza técnicas baseadas em perfis e assinaturas para a detecção de intrusões. Os trabalhos de Porras e Neumann (1997); White *et al.* (1996) disseminam informações para mitigar problemas identificados em outros locais da rede, por isso, podem ser considerados como os primeiros modelos de EWS voltados à cibersegurança.

O surgimento de diversos IDS e a aplicação de diferentes técnicas de detecção de intrusão não foram efetivas para lidar com as questões de alto número de falsos positivos e com a complexidade dos ataques. No final da década de 90 e início do novo século, o foco dos sistemas de detecção de intrusão é voltado para estratégias de colaboração e fusão de dados para lidar com essas questões. Bass (2000) afirma que a taxa de alertas falsos positivos é um problema persistente e preponde-

rante. Discute a importância da fusão de dados de diferentes origens para lidar com esse problema e prover o estado do ciberespaço (*cyberspace situational awareness*). Frincke (2000) afirma que uma vez que um alvo potencial é identificado, o passo seguinte é identificar outros potenciais alvos com as mesmas características. Destaca como essencial o uso de filtros para a cooperação, especialmente para remover informações irrelevantes para a detecção e para evitar vazamento de dados confidenciais da origem. Lippmann *et al.* (2000a,b) verificaram que sistemas que usavam a combinação de informações de diferentes fontes alcançavam melhor desempenho na análise da base de dados DARPA 1998 e 1999. Mchugh *et al.* (2000) afirmam que a combinação e correlação da saída de diferentes sensores em diferentes localizações é vital para aprimorar a detecção e diminuir o número de falsos positivos.

A cooperação de diferentes fontes implica no uso de técnicas para realizar o correlacionamento entre alertas. Os mecanismos de correlação são essenciais para a criação de sistemas de alerta antecipados. Cuppens (2001) propõe uma abordagem de agrupamento de alertas baseado em regras especialistas. Cuppens e Miege (2002) representam os ataques usando a linguagem LAMBDA, pois acreditam que a constante definição dos ataques em bases de assinaturas é um problema. Os ataques são especificados por meio de pré e pós-condições, cenários de ataque, de detecção e de verificação. Adotam a abordagem de correlação explícita e semi-explícita. Ning *et al.* (2001) propõem um modelo hierárquico para abstração de eventos e especificação de ataques. A abordagem possibilita a cooperação na detecção de ataques por dividir assinaturas em partes menores entre diferentes IDS. Ning *et al.* (2002, 2004) propõem uma abordagem baseada em cenários para a correlação de alertas usando pré-requisitos e consequências de intrusões. Assumem que intrusões ocorrem a partir de um conjunto de ações de ataques. Dain e Cunningham (2001) analisaram três modelos de correlação: agrupamento por endereços de origem, heurísticas e probabilidade, e mineração de dados (redes neurais e árvores de decisão). Adotam um modelo denominado átomo que considera que um alerta é adicionado uma única vez em um agrupamento. Os melhores resultados foram obtidos com árvores de decisão. Debar e Wespi (2001) apresentam um modelo baseado em regras, que utiliza os conceitos de consequências e situações para agregar e correlacionar alertas. Consequências correspondem a ligações sequenciais entre alertas em um intervalo de tempo. Situações são definidas por características comuns entre alertas. Valdes e Skinner (2001) abordam a correlação de alertas de sensores heterogêneos por meio de um modelo probabilístico. Um novo alerta é agregado a um grupo de meta-alertas desde que alcance os limiares especificados para a similaridade geral e de cada característica. Abordam também a similaridade entre classes de ataque usando uma matriz de similaridade de incidentes, ou seja, há possibilidade de inferência de ataques de múltiplos passos. Consideram o tipo do sensor e a sua localização como informações importantes na detecção.

Carey *et al.* (2002) propõem uma abordagem simples de agregação, redução e correlação para integrar alertas no formato IDMEF de múltiplos IDS em uma base relacional. Krügel *et al.* (2002) modelam uma intrusão como padrões de eventos em grafos dirigidos e acíclicos. Alertas são emitidos quando um conjunto de eventos preenche as restrições de conteúdo e tempo de um cenário. Porras *et al.* (2002) apresentam uma abordagem baseada em missão para correlação e priorização de alertas. Uma missão consiste em definir os recursos e dados críticos para os usuários da rede e quais são as ameaças que mais preocupam os administradores. Ye *et al.* (2003) apresentam um modelo para identificar intrusões observando as variações significantes na intensidade de eventos em um sistema de informação. Qin e Lee (2003) empregam agregação de alertas, análise estatística e séries temporais para encontrar relacionamentos entre ataques. Usam análise de causalidade em séries temporais para verificar se uma variável em uma série temporal provê informação significante sobre uma variável em outra série temporal. Dessa forma, poderiam identificar estratégias novas e desconhecidas de ataques. Tian *et al.* (2005); Zhai *et al.* (2003) apresentam o uso da teoria Dempster-Shafer para estimar uma situação de intrusão a partir da fusão de dados de múltiplos sensores. Argumentam que os alertas dos IDS são evidências que podem ser combinadas para prover alertas antecipados de ataques.

Valeur *et al.* (2004) apresentam um modelo genérico de correlação de alertas de segurança

obtidos de diferentes sensores de redes. Os autores afirmam que a otimização em cada uma das etapas propostas no modelo colabora para obtenção de melhores resultados na agregação e diminuição no número de alertas. Katti *et al.* (2005) realizaram um estudo empírico sobre correlação de alertas em larga escala. Concluem que IDS devem trocar informações em tempo real e que a comunicação entre poucos IDS em um mesmo grupo apresenta benefícios como diminuição de sobrecarga e mesma taxa de detecção. Abraham *et al.* (2007) definem um DIDS como um IDS distribuído sobre uma rede de grande cobertura, que se comunicam entre si ou com um servidor central para prover monitoramento avançado de rede. Avaliaram que Computação Leve (*Soft Computing*) como uma abordagem a ser considerada em IDS. Zhou *et al.* (2007) propõem um arcabouço denominado LarSID (*Large Scale Intrusion Detection*) para compartilhar informações com evidências de ataques baseado em uma arquitetura publicador/assinante descentralizada. Concluem que a abordagem distribuída é mais eficiente do que a centralizada.

Algumas novas abordagens surgem ao final da primeira década do século. Al-Nashif *et al.* (2008) apresentam o ML-IDS (*Multi-Level IDS*), uma arquitetura baseada em computação autonômica e análise de fluxos de rede, protocolos e pacotes. Morin *et al.* (2009) apresentam uma arquitetura para fusão e correlação de alertas baseada em lógica de primeira ordem. O modelo de dados permite inferir sobre incidentes de segurança por meio da correlação de alertas de diferentes origens na rede. Li e Tian (2010) propõem um modelo de correlação de alertas baseado em ontologias. O modelo agrega correlação por pré-requisitos e consequências, representados por estados de segurança, e correlação por cenário de ataques predefinidos. Uma limitação do uso de bases de conhecimento de ontologias é a necessidade de atualização para contemplar novos ataques. Roschke *et al.* (2010) apresentam uma abordagem baseada em memória e banco de dados orientados a colunas para aumentar o desempenho na correlação de alertas. Dessa forma, é possível usar computação paralela e outras abordagens para lidar com questões de armazenamento, recuperação e processamento de alertas.

Bye *et al.* (2010) sintetizam questões importantes para o desenvolvimento de *Collaborative Intrusion Detection Systems* (CIDS): esquema de comunicação (P2P, agente, *middleware*), formação de grupo (explícita ou implícita), estrutura organizacional (hierárquica ou heterárquica), compartilhamento de informação (privacidade, interoperabilidade, anonimidade), segurança do sistema (disponibilidade, controle de acesso, gerenciamento de confiança). Boggs *et al.* (2011) desenvolveram um arcabouço para detecção de ataques dia zero em requisições Web que consiste em correlacionar alertas gerados por sensores em servidores espalhados em diferentes redes. Verificaram que os mesmos falsos positivos se repetem entre sítios Web, logo, correlacionando informações de diferentes origens, foi possível reduzir significantemente o número de alertas irrelevantes. Elshoush e Osman (2013) propõem um arcabouço para reduzir o número de alertas no processo de correlação por meio da reorganização dos componentes. O objetivo é descartar alertas falsos e de pouco interesse o mais cedo possível.

Há ainda inúmeras pesquisas nos últimos anos que abordam a correlação de alertas (Meng *et al.*, 2015; Vasilomanolakis *et al.*, 2015). Em especial, há diversos trabalhos que começaram a apresentar arquiteturas e protótipos para a detecção antecipada de incidentes em grupos de redes e na Internet. A seção 2.5.1 discute algumas dessas propostas e implementações de EWS.

## 2.2 Sistemas de alerta antecipado

### 2.2.1 Conceitos

O conceito de EWS tem ganhado notoriedade em muitas áreas de vigilância a ameaças, principalmente devido a catástrofes naturais, em especial ao tsunami ocorrido no final de 2004 na Indonésia. As Nações Unidas consideram que os EWS devem ser centrados nas pessoas e integrar conhecimento de risco, serviços de monitoramento e alerta, disseminação de alertas relevantes para situações de

risco, e conscientização e preparação pública para responder às ameaças (United Nations, 2006).

Na literatura, há diversas pesquisas em EWS sob a perspectiva de diferentes áreas, como agricultura, finanças, clima, medicina, engenharia, política, catástrofes naturais, entre outras. Alguns temas recorrentes nessas áreas são modelos conceituais de predição, identificação e avaliação de indicadores, sistemas de classificação e de medidas de risco e descrição de novas arquiteturas (Choo, 2009). Da mesma forma, as premissas dos EWS têm sido usadas em sistemas e redes de computadores para desenvolver sistemas e mecanismos para combater e mitigar as ameaças cibernéticas.

No contexto de segurança de sistemas e redes de computadores, (Bastke *et al.*, 2010) definem três tipos de alerta antecipado:

1. antes do início do ataque: alerta é emitido durante as ações preparativas ou entre os nós intermediários de um ataque.
2. durante o ataque: alerta é emitido durante o ataque, mas antes desse causar danos ou alcançar o ápice.
3. antes de uma possível ameaça: alertas são amplamente propagados em situações de ameaças potenciais e novas vulnerabilidades descobertas, mas o ataque não ocorrerá necessariamente.

(Bastke *et al.*, 2010) argumentam que EWS são limitados para alertas antecipados do tipo 1 e 2, pois o período de tempo para avisar o alvo é muito curto. Por consequência, concluem que EWS devem gerar alertas do tipo 3. Além disso, consideram que um EWS é para ser usado na Internet e definem como sendo uma sêxtupla composta por: rede monitorada, organizações participantes, estrutura organizacional e operacional, aspectos legais, objetivos e componentes técnicos.

(Biskup *et al.*, 2008) definem EWS como sistemas que objetivam estabelecer hipóteses e predições sobre ameaças por meio de informações incertas e incompletas obtidas de sistemas e redes de computadores. Os EWS geram e propagam alertas considerando padrões de risco com o propósito de prover mecanismos proativos de reação para a mitigação de danos causados por ataques. (Apel *et al.*, 2010) afirmam ainda que a automatização da detecção de ameaças e da propagação de alertas de incidentes também são requisitos essenciais para EWS.

(Koch, 2011) afirma que, comparado com IDS, os EWS atuam na Internet, monitorando fontes de dados distribuídas e, a partir de alertas gerados em uma sub-rede, objetivam proteger outras sub-redes não comprometidas. Nesse mesmo sentido, (Engelberth *et al.*, 2010) afirmam que EWS são importantes para evitar a propagação de códigos maliciosos, pois podem detectar a ocorrência de uma ameaça e propagar informações para a tomada de medidas efetivas.

(Freiling, 2010) apresenta uma definição formal para EWS fundamentada em três conceitos bases: informação, espaço e tempo. Deve haver, no mínimo, duas localizações diferentes no ciberespaço, $lugar_1$ e $lugar_2$. Deve haver, no mínimo, duas instâncias de tempo, $tempo_1$ e $tempo_2$, onde $tempo_1 < tempo_2$. A informação deve consistir de relatos úteis, por exemplo, detalhamento da forma, causa e/ou efeito de um ataque, código malicioso ou invasão ocorridos em $lugar_1$ em $tempo_1$. A informação deve ser transferida para o $lugar_2$ antes de $tempo_2$ para efetivamente evitar ou mitigar os danos.

A partir das definições, considerando as características e velocidade de disseminação de alguns ataques, pode ser impossível transferir informações úteis em um curto período de tempo, pois o tempo de análise do ataque pode ser maior que o tempo de transferência do alerta ou o intervalo pode ser pequeno comparado com o tempo de reação e transferência de informações em uma rede. Entretanto, considerando as tendências de ataques modernos, ou seja, ataques avançados e persistentes, os EWS são efetivos para evitar ou mitigar ataques que causam gravíssimos problemas para muitas organizações e usuários de sistemas computacionais.

### 2.2.2 Ameaças e alertas antecipados

Diversas ameaças cibernéticas têm surgido no decorrer dos anos e causado impactos significantes nas organizações. Entre essas ameaças destacam-se os códigos maliciosos e as vulnerabilidades críticas exploradas por esses códigos. São exemplos de códigos maliciosos e vulnerabilidades que marcaram a história: Morris (1988), Code Red (2001), Nimda (2001), Slammer (2003), Conficker (2008), Stuxnet (2010), Flame (2012) e, recentemente, o Heartbleed (2014).

Os alertas em Segurança da Informação (*InfoSec*) são mensagens emitidas por diferentes sensores de redes de computadores e sistemas para notificar sobre possíveis ameaças à segurança (Porras *et al.*, 2002; Qin e Lee, 2003). Alertas antecipados em InfoSec são alertas emitidos por sensores de rede, o mais cedo possível, para que as ameaças possam ser mitigadas em outras localizações. A utilidade do alerta é dependente do tipo de ameaça. Logo, alertas antecipados possuem vantagens e limitações ao lidar com tipos específicos de ataques.

Limmer e Dressler (2008) classificam as ameaças à segurança de redes em três grupos: intenção, escopo e metodologia. Essa classificação provê um indício da utilidade dos alertas antecipados para cada ameaça.

O grupo intenção consiste de varreduras para coletar informações sobre os sistemas, ataques de negação de serviço para derrubar ou sobrecarregar o sistema e a exploração de vulnerabilidades para identificar uma forma de comprometer o sistema. Logo, detectar essas intenções possibilitam aos administradores reagir proativamente contra possíveis ameaças. Receber informações sobre as intenções observadas em outras organizações, comprometidas ou não, possibilita avaliar o risco real da ameaça.

O grupo escopo se divide em ataques direcionados e amplos (não direcionados). Ataques direcionados são classificados como *Advanced Persistent Threats* (APT) e ocorrem sorrateiramente para não serem percebidos. Ataques amplos visam um maior número de alvos e são mais ruidosos. Um EWS não lida bem com ataques direcionados por visarem alvos específicos, logo a detecção é complexa. Mesmo após a detecção, muitas organizações não divulgam o ataque e isso viabiliza ao atacante explorar outras organizações com características similares. Por sua vez, ataques amplos podem ser detectados e propagados mais rapidamente. A questão é que dependendo do ataque, como um *worm* ou mesmo um ataque de negação de serviço, o tempo de reação é insuficiente para emitir alertas antecipados.

A grupo metodologia consiste de ataques diretos ou indiretos, ou seja, considera a forma como o atacante acessa os alvos. No primeiro caso, é fácil identificar a origem e emitir alertas antecipados a partir da confirmação das intensões maliciosas. No segundo caso, o mais comum, o uso de endereços falsos e de clientes comprometidos, torna difícil considerar a origem como uma forma de alerta antecipado.

(Bastke *et al.*, 2010) abordam a importância em caracterizar os tipos de ameaças que podem ou não ser previstas por EWS. Realizaram a análise de cinco cenários diferentes: negação de serviço, códigos de exploração *exploits*, propagação de códigos maliciosos, *botnets* e roteamento.

Nos cenários de negação de serviço, o tempo para reação é muito pequeno, logo não há como enviar mensagens para as vítimas antes da própria vítima detectar o ataque.

Nos cenários com códigos de exploração, são sondados ou comprometidos os cenários com vulnerabilidades, logo um EWS é efetivo para alertar sobre uma possível ameaça por meio da propagação de notificações de novas ameaças e vulnerabilidades. Por sua vez, para uma vulnerabilidade dia zero, pode não ser possível emitir um alerta para todos os potenciais alvos antes de um *honeypot* ou alvo ser comprometido.

Nos cenários de propagação de códigos maliciosos, um alerta antecipado é efetivo antes da propagação alcançar a curva exponencial de propagação, o que no caso, dependendo da ameaça, pode ocorrer em minutos (SQL Slammer, Witty) ou poucas horas (Code Red). Logo, um alerta

mais efetivo é sobre a vulnerabilidade explorada pelos códigos maliciosos.

Nos cenários de *botnets*, as redes têm sido criadas descentralizadas e levam questão de minutos para organizar um ataque. Um alerta antecipado pode ser efetivo se a *botnet* é lenta ou se o comando e controle é centralizado, ou ainda, se é possível identificar a assinatura do comando de controle.

Nos cenários de roteamento, como a propagação de uma rota ocorre em poucos segundos, é difícil um alerta antecipado evitar a propagação de rotas alteradas.

Considerando os cenários descritos, verifica-se que alertas antecipados são eficientes para alertar sobre varreduras, vulnerabilidades e códigos de exploração. Entretanto, trabalhos recentes têm abordado EWS que usam informações coletadas nas próprias organizações e nos sistemas autônomos para predizer a possibilidade de ocorrência de um DDoS, detectar ameaças à infraestrutura de roteamento, a ação de *botnets* e de códigos maliciosos.

### 2.2.3 Vulnerabilidades e alertas antecipados

As vulnerabilidades são falhas ou brechas no projeto, implementação, operação ou gerenciamento de sistemas que apresentam risco de serem exploradas por códigos ou indivíduos maliciosos com o intuito de violar a política de segurança dos sistemas (Shirey, 2000). As vulnerabilidades dia zero são as mais perigosas por serem desconhecidas e ainda não terem sido anunciadas publicamente. Também apresentam grande risco as vulnerabilidades conhecidas que não possuem atualizações ou soluções disponíveis para a correção (Levy, 2004).

Segundo Bastke *et al.* (2010), os alertas antecipados são de suma importância para mitigar a propagação de ameaças que exploram vulnerabilidades. Entretanto, alguns especialistas são contra a divulgação pública antes que a correção seja disponibilizada pelo fabricante. Isso, porque, a vulnerabilidade pode ser amplamente explorada nessa janela de tempo. Independente das visões, compreender o ciclo de vida das vulnerabilidades possibilita identificar o papel dos alertas antecipados na contenção de ameaças que as exploram.

(Frei *et al.*, 2006, 2008) enumeram os riscos à exposição e as fases do ciclo de vida de vulnerabilidades (Figura 2.2).

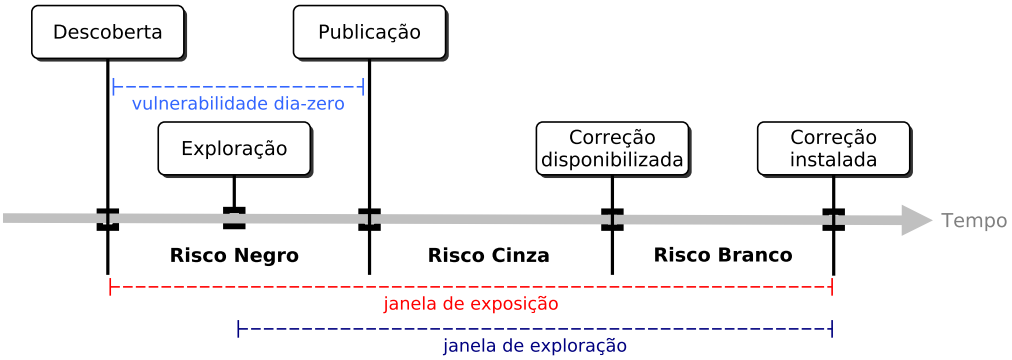


**Figura 2.2:** *Ciclo de vida de uma vulnerabilidade (Adaptado de Frei et al. (2008)).*

O ciclo de vida de uma vulnerabilidade pode ser classificado em:

- Descoberta: descoberta e verificação que a vulnerabilidade representa riscos à segurança.
- Exploração: primeiro uso da vulnerabilidade com a finalidade de comprometimento da segurança de um sistema computacional.
- Publicação: vulnerabilidade documentada por especialistas e divulgada de forma ampla, pública e confiável.

- Correção disponibilizada: disponibilização de correções ou medidas corretivas para a vulnerabilidade fornecida pelo provedor original do software.
- Correção instalada: as correções para a vulnerabilidade são implantadas ou instaladas no ambiente computacional.

O risco de exposição a uma vulnerabilidade ocorre desde a descoberta, seja por um usuário normal ou malicioso, membro de uma organização ou mesmo do provedor do software, até a instalação da correção. A partir da primeira exploração, o risco aumenta devido à existência comprovada de métodos ou códigos maliciosos. Frei *et al.* (2006) classificam os períodos de riscos em negro, cinza e branco.

O período de risco negro ocorre entre a descoberta e divulgação de uma vulnerabilidade. Os ataques que ocorrem são comumente denominados de ataques dia zero. Apenas um pequeno grupo de pessoas tem conhecimento da vulnerabilidade. Entretanto, quando as vulnerabilidades são exploradas por códigos maliciosos, geralmente, causam prejuízos e comprometimento de muitos sistemas. Uma preocupação constante é o comprovado comércio de códigos maliciosos para exploração de vulnerabilidades dia zero (Egelman *et al.*, 2013).

O período de risco cinza ocorre entre a divulgação e a disponibilização da correção de uma vulnerabilidade. Nessa fase, um grande número de pessoas tem acesso à informação e muitos códigos maliciosos e simples de serem usados são criados para possibilitarem a exploração da vulnerabilidade. Bilge e Dumitras (2012) constataram que o número de ataques aumenta em cinco ordens de magnitude após a divulgação de vulnerabilidades. Frei *et al.* (2006); Shahzad *et al.* (2012) verificaram que a criação de códigos maliciosos para exploração de vulnerabilidades é mais rápida do que a disponibilização de correções pelos fabricantes. De qualquer forma, mesmo sem as correções, medidas preventivas ou o uso de soluções de terceiros podem ser implantadas para conter as ameaças.

O período de risco branco ocorre entre a disponibilização e instalação/implantação da correção de uma vulnerabilidade. Nessa fase, somente os alvos que não implantaram as medidas corretivas ou não instalaram uma nova versão ou atualização para o software estão sujeitos a exploração da vulnerabilidade. Infelizmente, pesquisas indicam que mesmo após a disponibilização de atualizações, as ameaças são exploradas em média por quase um ano (Bilge e Dumitras, 2012). Como exemplo, o *worm* Sasser, que tinha uma atualização disponível em abril de 2003 e continuou se propagando até maio de 2004 (Jumratjaroenvanit e Teng-Amnuay, 2008).

Considerando os cenários de risco decorrentes das vulnerabilidades, ou seja, as ameaças dia zero e o período de tempo entre a notificação de vulnerabilidades e medidas corretivas, quanto antes um alerta ou predição de ameaça é recebido por um administrador de sistemas ou um usuário de produto de software, maior a probabilidade de prevenir ou conter uma ameaça por meio de mecanismos reativos, mesmo em situações que não haja uma correção provida pelo fabricante do produto.

## 2.3 Características e desafios

Nesta seção são analisadas diferentes características técnicas e não técnicas presentes em sistemas de alerta antecipado e correlacionados (IDS, DIDS, CIDS, entre outros).

### 2.3.1 Coleta de informações

Há diferentes fontes de informações usadas para prover dados para um EWS. Essa variedade de fontes possibilita a visão mais completa da situação da rede. Entretanto, os mecanismos e procedimentos de coleta são dependentes dos sensores usados para extrair informações relevantes dessas

fontes. Além disso, questões importantes estão relacionadas à disposição e metadados dos sensores, tipo e representação da informação, políticas e mecanismos de coleta, confidencialidade de comunicação e privacidade nas fontes.

A coleta de informações para monitoramento, detecção e geração de alertas de segurança pode ser realizada a partir de pacotes e fluxos de rede, chamadas de sistema, sequência de comandos, registros de auditoria (sistemas, *firewalls*, serviços), informações de roteamento e gerenciamento de rede, informações de hardware de rede, redes sociais (Twitter, Facebook, IRC, fóruns e blogs), endereços IP não usados (*dark address space*), spams, Internet (arquivos, URLs, tráfego). Esse conjunto de informações é responsável por prover dados aos sensores de redes, ou seja, fontes de informações para detecção de ameaças.

Os sensores de redes são os mecanismos que monitoram, extraem ou geram informações relevantes sobre ameaças a partir de fontes de informações. Os principais sensores de redes usados para geração de alertas de segurança são: *honeypots/honeynets*, NIDS, HIDS, antivírus, anti-spams, monitores específicos (serviços, hardware, redes sociais, tráfego, entre outros) e monitoramento de alertas (ataques, vulnerabilidades, atualizações, *worms*). Cada classe de sensor monitora fontes específicas de informações e uma mesma classe pode usar diferentes técnicas de detecção e geração de alertas. Também é importante notar que a informação coletada implica na detecção de ameaças específicas (Ghorbani *et al.*, 2010).

Os *honeypots* e as *honeynets* emulam ou executam sistemas e serviços em uma máquina ou rede específica para serem comprometidos com o intuito de identificar códigos maliciosos ou novas ameaças (Spitzner, 2002). São classificados de baixa interatividade quando emulam infraestruturas computacionais e de alta interatividade quando executam em reais. Por ser um recurso que não tem finalidade de produção, qualquer acesso é considerado suspeito. Logo, são usadas para identificar novas ameaças, assinaturas de códigos maliciosos, monitorar redes de *bots*, identificar URLs comprometidas, fontes de spams, alvos de ataques, entre outros. Essas informações se propagadas são consideradas alertas a ameaças identificadas, tendências de ataques ou novas ameaças. Por essa razão, os *honeypots* são usados em projetos de EWS como HoneyTARG (CERT.br, 2014), Honeynet (Honeynet, 2014), InMAS (Engelberth *et al.*, 2010), entre outros.

Os NIDS analisam e detectam ameaças a partir de dados coletados da rede, ou seja, pacotes e fluxos de rede. Realizam a análise de cabeçalhos, fluxos e, dependendo da implementação, da carga dos pacotes. A principal técnica empregada consiste na detecção de assinaturas de ataques. As principais limitações estão associadas a captura e análise de altos volumes de tráfego e à geração de alertas falsos (Ghorbani *et al.*, 2010). Ainda assim, é um dos principais sensores de redes usados na detecção de intrusões em soluções individuais e colaborativas. Isso se deve ao fato de que os alertas gerados por NIDS possuem classificação, características temporais e espaciais do ataque, e podem ser integrados a outras soluções de segurança. Em soluções colaborativas, possibilita correlacionar os alertas com outros de fontes diferentes. Entretanto, para colaboração externa, mandatória em EWS, acabam sendo utilizados com restrições devido a questões de compartilhamento.

Os HIDS analisam e detectam ameaças a partir de dados coletados em máquinas. As principais fontes de informações são os registros de auditoria, chamadas de sistemas, sequências de comandos, memória e arquivos do sistema. As principais técnicas de detecção são baseadas em detecção de anomalias e padrões. Possuem como vantagens o monitoramento de uma única máquina ou características específicas do sistema. São limitações o comprometimento ou evasão do sistema de detecção, e a instalação de agentes em cada estação monitorada. O compartilhamento externo de informações é perigoso por viabilizar o vazamento de informações sensíveis. O uso em EWS está associado ao compartilhamento da informação de alerta sem prover os detalhes da intrusão.

Os antivírus analisam arquivos e processos à procura de códigos maliciosos. Utilizam assinaturas e padrões para identificar ameaças ou códigos suspeitos. A colaboração é comum entre produtos de uma mesmo provedor do software. Compartilhamento de assinaturas é essencial para mitigar a propagação de novas ameaças em um EWS. No projeto AMSEL (Apel *et al.*, 2009), são criadas

assinaturas de códigos maliciosos e compartilhadas em um repositório. Já os anti-spams possibilitam identificar origens de códigos maliciosos e gerar listas com as informações sobre origem e tipos de ameaças. A análise das ligações externas contidas nas mensagens eletrônicas pode viabilizar a detecção de novas ameaças. Por esse motivo, é um sensor interessante para EWS e, é usado em projetos como InMAS (Engelberth *et al.*, 2010) e Internet EWS (Bastke *et al.*, 2010).

A atividade realizada por monitores específicos possibilita identificar ameaças em serviços e aplicações. Por analisarem um alvo potencial ou focarem em um único ou pequeno conjunto de ataques, são capazes de gerar informações de interesses para EWS. O projeto DShield (DShield, 2014) permite identificar tendências e novas ameaças por meio da correlação de registros de firewalls. Boggs *et al.* (2011) desenvolveram um arcabouço colaborativo para detectar ameaças dia zero em alertas gerados a partir de requisições Web. Bailey *et al.* (2005); Caida (2014); Inoue *et al.* (2009) usam sensores para monitorar os endereços IP não usados na Internet e, dessa forma, caracterizar e monitorar ameaças. Sensores para monitorar redes sociais e encontrar ameaças também têm sido desenvolvidos, como para o Twitter (Lee e Kim, 2013) e o Facebook (Robertson *et al.*, 2010).

Por fim, alertas distribuídos por correspondência eletrônica, blogs, redes sociais, repositórios e sítios especializados, constituem uma forma de antecipar possíveis ameaças que podem comprometer a segurança da infraestrutura de rede. O sensor mais usual é a assinatura de notificações especializadas, por exemplo, bases do CVE e NVD. Algumas ferramentas de segurança[1] integram-se a notificações de alerta para verificar se a infraestrutura computacional apresenta alguma vulnerabilidade descrita nas notificações. As limitações no uso dessas informações são a diversidade de fontes e o período da confirmação da ameaça até sua homologação na base de vulnerabilidades.

### 2.3.2 Gerenciamento de alertas

O gerenciamento de alertas tem por objetivo processar os alertas coletados de diversos sensores e realizar os procedimentos de normalização e filtragem. Normalização é o processo de representar o alerta em um formato padrão e estruturado (Kruegel *et al.*, 2005; Valeur *et al.*, 2004). Muitas vezes, são adicionados ou removidos atributos. Filtragem é o processo de remover alertas sem relevância e redundantes. Muitas vezes, são selecionados os alertas baseados no interesse da aplicação. A normalização e filtragem podem ocorrer diretamente no sensor de rede, em outros componentes de um EWS ou em ambos. As ações de normalização e filtragem são fases de pré-processamento (Ghorbani *et al.*, 2010). Segundo Bass (2000), esses processos caracterizam a transformação de dados em informações, isto é, a transformação de alertas puros em informações identificadas, contextualizadas e estruturadas.

O núcleo da normalização consiste em transformar e validar alertas de sensores heterogêneos em um formato padrão para as próximas fases de processamento. A seção 2.3.6 discute alguns desses formatos. Em especial, destaca-se o uso do IDMEF em muitos projetos (Carey *et al.*, 2002; Feitosa *et al.*, 2012; Grobauer *et al.*, 2006; Qin e Lee, 2003; Roschke *et al.*, 2010). Em outros projetos, os alertas são transformados em padrões específicos, por exemplo, representação usando ontologias (Mathews *et al.*, 2012; More *et al.*, 2012), bases de dados, estruturas específicas (Debar e Wespi, 2001).

No processo de normalização, a ordenação dos eventos é importante. Duas formas de abordar essa questão são o uso de marcações de tempo (*timestamps*) ou de relógios lógicos (Krügel *et al.*, 2002). O uso de marcações de tempo está associado à sincronização com um relógio global. O uso de relógios lógicos permite estabelecer uma relação temporal entre eventos de máquinas ou redes distintas. No entanto, dependendo da fonte da informação, é impossível caracterizar uma ordenação baseada no tempo ou ocorrência do evento. Logo, deve ser identificado o momento da criação ou do recebimento da notificação pelo componente de normalização.

[1] https://cve.mitre.org/compatible/organizations.html

Dentre as inúmeras dificuldades para normalizar alertas, Hochberg *et al.* (1993) já apontavam as questões como falta de informações e organização idiossincrática associadas a alertas de múltiplas fontes. Quanto maior o número de sensores heterogêneos, mais complexa a fase de normalização. O uso de padrões e ontologias nos sensores e por módulos de conversão tem sido alternativas empregadas para sobrepor esses problemas. Apesar do uso de ontologias, muitos sensores similares classificam os mesmos ataques com terminologias diferentes. A padronização da natureza dos atributos dos alertas é essencial para o uso de técnicas de detecção de anomalias (Chandola *et al.* , 2009).

A normalização em EWS ainda pode ter problemas relacionados à estrutura semântica de diferentes línguas, localização do alerta devido a questões de representação e privacidade, tempo de emissão devido a diferentes fusos e atrasos na propagação, entre outros. Algumas dessas questões são abordadas pelo *Common Alerting Protocol* (CAP), um padrão genérico definido para disseminação de alertas em redes referentes a qualquer tipo de ameaça (OASIS, 2010). O RFC 4766 (Wood e Erlinger, 2007) também discute algumas dessas questões considerando alertas emitidos por IDS.

O processo de filtragem consiste em reduzir o número de alertas por meio da filtragem de alertas redundantes e não relevantes para aumentar o desempenho e acurácia do sistema. Os alertas redundantes são alertas duplicados ou com nenhum novo atributo para detalhar uma situação já informada. Os alertas não relevantes são alertas malformados, alertas associados a recursos não presentes na rede ou falsos positivos. Por consequência, a filtragem desses alertas contribui diretamente para a diminuição de recursos computacionais, como armazenamento, rede e memória.

Nos EWS, o processo de filtragem também é responsável por aplicar políticas de compartilhamento para evitar o vazamento de informações sensíveis das organizações. Além disso, protege contra o vazamento de vulnerabilidades, topologia e serviços na rede, os quais podem ser extraídos de alertas compartilhados e, se observadas por um atacante, podem facilitar ou originar um ataque (Porras e Shmatikov, 2006). Por outro lado, os filtros devem balancear a remoção, ocultação ou anonimidade das informações, senão acabam tornando o compartilhamento sem relevância para a correlação.

Há uma diversidade de abordagens e técnicas que podem ser usadas para realizar a filtragem baseada nas características das informações. Por exemplo, o uso de técnicas de agregação permite eliminar alertas redundantes, pois pode comparar atributos comuns ou similares dos alertas. Um conjunto de regras especialista pode ser definido para remover alertas irrelevantes baseado em palavras-chave, impacto do alerta na rede, política de compartilhamento, hardware e software instalados, entre outros. Outra abordagem é a criação de hierarquia de filtros multifases para processar fluxos de interesses com diferentes políticas e detalhamento. Já os filtros adaptativos permitem otimizar o processo de filtragem, por viabilizar a geração de novas regras baseando-se na análise e retroalimentação com informações do sistema monitorado.

Observa-se que filtros desempenham um importante papel em EWS. Elshoush e Osman (2013) afirmam que a seleção de atributos e o descarte de conteúdos irrelevantes agilizam o processo de correlação de alertas. Defendem que a filtragem deve ser realizada nas fases de processamento de alertas o mais breve possível.

### 2.3.3 Correlação de alertas

A correlação de alertas é importante em EWS por possibilitar a associação de alertas de diferentes origens e, dessa forma, alcançar uma visão situacional consistente de cenários de ataques devido ao preenchimento das lacunas entre alertas de diferentes fontes. Como consequência, obtém-se a redução do número de alertas e falsos positivos, devido à fusão e agregação dos alertas. A fusão gera alertas que representem significantemente um grupo de alertas, comumente denominado meta-alerta (Kruegel *et al.*, 2005), e a agregação agrupa alertas comuns a um cenário de ataque.

As abordagens usadas na correlação de alertas segundo (Cuppens e Miege, 2002; Morin *et al.* , 2009) são: explícita, implícita e semi-explícita. A explícita consiste em correlacionar alertas a ataques por meio de padrões conhecidos e predefinidos por um especialista em segurança. A implícita consiste em correlacionar observações implícitas entre alertas por meio de técnicas de aprendizagem de máquina. A semi-explícita consiste em associar as consequências de um ataque às pré-condições de um próximo ataque. As correlações explícita e semi-explícita identificam ataques de múltiplos passos, mas apenas a semi-explícita viabiliza a correlação com cenários não conhecidos de ataques. A correlação implícita simplifica a análise de cenários com muitos alertas, mas restringe a interpretação automática dos cenários.

Diferentes técnicas para realizar a correlação de alertas são abordadas na literatura. Xu e Ning (2008) classificam as técnicas como baseadas em similaridade, cenários de ataques predefinidos, pré-requisitos e consequências, e fontes de informações múltiplas. Ghorbani *et al.* (2010) classificam as técnicas como baseadas em similaridade de atributos, cenários conhecidos, e consequências e pré-requisitos de ataques. Salah *et al.* (2013) consideram várias formas de classificação para as técnicas e, em especial para os métodos de correlação, classificam como baseadas em similaridade, sequência e caso. Nesta tese será considerada a classificação proposta por Salah *et al.* (2013).

As técnicas baseadas em similaridade (Cuppens, 2001; Qin e Lee, 2003; Valdes e Skinner, 2001) correlacionam alertas considerando a similaridade de características, por exemplo, atributos espaciais e temporais de alertas. A similaridade pode ser definida por comparação direta ou por funções de similaridade entre características. Essas técnicas consideram a premissa de que alertas similares compartilham a mesma causa, ou seja, estão associados a um mesmo ataque. Como vantagens, as técnicas são computacionalmente menos intensivas e diminuem o número de alertas. Como desvantagens, as funções de similaridade são complexas de serem estabelecidas entre ambientes heterogêneos e as relações causais entre alertas são difíceis de serem identificadas.

As técnicas baseadas em sequência (Cuppens e Miege, 2002; Ning *et al.*, 2002; Xu e Ning, 2008) correlacionam alertas considerando as relações causais entre alertas, isto é, consideram que alertas resultam de sequências de passos de ataques. Como vantagens, podem identificar novos ataques e detalhar cenários de ataques. Como desvantagens, correlações imprecisas e falsas podem ser geradas devido à dependência dos predicados de correlação e informações dos sensores. Muitas técnicas têm sido usadas para realizar a correlação por sequência, como as baseadas em consequências e pré-requisitos, grafos, redes bayesianas, cadeias de Markov e redes neurais.

As técnicas baseadas em caso (Dain e Cunningham, 2001; Debar e Wespi, 2001; Morin *et al.* , 2009) correlacionam alertas considerando cenários de ataques conhecidos, ou seja, alertas são associados a cenários que poderiam ser prováveis causas do alerta. Os cenários podem ser definidos por especialistas em segurança ou aprendidos em bases de dados de treinamento (Elshoush e Osman, 2011). Como vantagens, são eficientes em identificar cenários de ataques na base de conhecimento. Como desvantagens, não identificam novos cenários de ataques e há o custo de manutenção da base de conhecimento. As principais técnicas usadas são baseadas em sistemas especialistas, cenários predefinidos e por aprendizagem de máquina.

Apesar da diversidade de técnicas de correlação, ainda é um desafio correlacionar alertas de fontes heterogêneas devido às características inerentes a cada fonte.

### 2.3.4 Detecção e predição de ameaças

A detecção e predição de ameaças é importante em EWS por possibilitar a emissão de alertas de segurança com base em informações obtidas a partir da correlação de alertas de diferentes fontes. Detecção significa a identificação de um comportamento anômalo ou de ataque em um evento resultante da correlação ou a partir de uma assinatura de ameaça encontrada por um dos parceiros de um EWS. Predição significa identificar a possibilidade de ameaça através de assinaturas de ataques e correlação de eventos de diferentes fontes. A predição é uma das características que difere

EWS de sistemas tradicionais de detecção de intrusão.

Considere um cenário de varredura de uma porta específica que foi observada em três agências bancárias. O processo de correlação agregou os alertas em um meta-alerta descrevendo as informações espaciais, temporais e características da varredura. A detecção consiste em identificar que o meta-alerta é sobre uma ameaça e a predição consiste em identificar a possibilidade dessa ameaça evoluir para um ataque mais elaborado ou ser realizada em outras agências.

Geralmente, a detecção de ameaças é realizada por métodos baseados em assinatura ou em anomalias. Tipos de ataques específicos exigem métodos de detecção próprios. Muitos EWS detectam códigos maliciosos por meio de *honeypots*, em especial, usando *sandboxes* para realizar a análise dinâmica de códigos e detectar novas ameaças. Se detectada uma nova ameaça, são geradas as assinaturas e emitidos alertas para os parceiros. Outros EWS realizam a detecção de ameaças por meio do monitoramento de rede de *bots* e propagam listas de aviso sobre endereços e URL maliciosas. Ataques de DDoS são detectados, por exemplo, por monitoramento de tráfego em diferentes pontos na Internet. A correlação de alertas de fontes heterogêneas de informação também permite identificar ameaças existentes e até mesmo novas. Na seção 2.3.3 são discutidas diferentes formas de correlação de alertas, vantagens e desvantagens para a detecção de ameaças.

Em muitas situações, a constatação de uma ameaça é definida por limiares e regras dependentes de configuração por especialistas humanos, tipos e característica dos métodos de correlação, confiabilidade das fontes, entre outros. Por sua vez, o nível de predição é controlado pela flexibilização desses limiares e regras para indicar a possibilidade de ameaças. Entretanto, a flexibilização de limiares e regras de predição podem conduzir a uma quantidade expressiva de falsos positivos e, assim, gerar uma quantidade significativa de alertas. Geralmente, soluções em EWS contam com mecanismos de classificação, priorização e análise por especialistas humanos para caracterizar ameaças e realizar predições mais precisas a partir de assinaturas de ataques e correlações de alertas.

A classificação pode considerar diferentes critérios, como características dos ataques (tipo, origem, destino, relevância, impacto, causa), características dos alertas (data, horário, localização física, sensor) e características dos parceiros (interesses, políticas de privacidade, confiabilidade, grau de colaboração). A priorização consiste em destacar os alertas que apresentam maior relevância como um alerta antecipado entre os meta-alertas resultantes da correlação. Novamente, alguns critérios devem ser considerados para julgar a prioridade de alertas sobre outros. Também colabora para descartar falsos positivos quando a quantidade de alertas é muito alta. Por fim, a análise por especialistas humanos é empregada para aumentar o grau de confiabilidade de alertas e para extrair cenários mais detalhados de ameaças à segurança. Dependendo da fonte de informação e complexidade dos ataques, é impossível depender apenas de métodos automatizados para realizar a detecção e predição de ameaças.

Geib e Goldman (2001) enumeram requisitos para a realização de predição e reconhecimento de planos de ataques: inferir ações não notificadas por sensores de redes, analisar as mudanças de estado pois podem indicar ações não observadas, reconhecer planos parcialmente ordenados, identificar os múltiplos objetivos de um ataque, considerar quais ações do atacante geram a possibilidade de múltiplos objetivos, distinguir entre ações isoladas e ataques com múltiplos passos, conhecer a topologia para identificar efetivamente as ameaças, calcular e ordenar as probabilidades de ocorrência das hipóteses.

A literatura apresenta diferentes técnicas usadas em predição de ameaças à segurança que procuram abordar alguns pontos apontados por Geib e Goldman (2001):

- Active LeZi: algoritmo de predição sequencial baseado na teoria da informação usado em domínios estocásticos sem conhecimento específico do domínio de aplicação (Gopalratnam e Cook, 2007). Feitosa *et al.* (2012) implementam uma função de predição e usam o algoritmo para calcular a maior probabilidade de ocorrência de regras (padrões de ameaças).
- Teoria de Dempster-Shafer: possibilita tratar incerteza na análise de intrusões e combinar

crenças de diferentes fontes de evidências. Zomlot *et al.* (2011) estendem a teoria para priorizar alertas em um grafo de correlação de alertas de IDS.

- Modelos Ocultos de Markov: permitem tratar de estados não observáveis por meio da observação de estados observáveis. Haslum *et al.* (2008) utilizam o modelo para predição de intrusões por meio da combinação de alertas de diferentes IDS.
- Médias Móveis Exponencialmente Ponderadas (MMEP): permitem atribuir prioridade alta a observações mais recentes em relação a distantes por meio do ponderamento de pesos. Pontes e Guelfi (2009) combinam com outras técnicas em níveis topológicos hierárquicos para realizar a predição de ameaças à segurança.
- Aprendizagem de Máquina: possibilita a descoberta de padrões de ataques ou anomalias na rede por meio de aprendizagem supervisionada ou não supervisionada e, dessa forma, prediz possíveis situações que caracterizam ameaças à segurança. Cipriano *et al.* (2011) usam técnicas de aprendizagem de máquina para a predição de ameaças por meio do comportamento aprendido do histórico de ataques anteriores.
- Redes Bayesianas ou Causais: são grafos acíclicos que possibilitam a geração de predições mesmo na ausência de informações, ou seja, agrega a incerteza na tomada de decisões. Qin e Lee (2004) preveem ataques futuros por meio da inferência probabilística de atividades totais ou parciais de ataques observados anteriormente.
- Redes Bayesianas Dinâmicas: extensão de redes bayesianas que permite modelar séries temporais ou sequenciais. Feng *et al.* (2009) realizam predição através da combinação de redes bayesianas com estimação de probabilidade analisando as sequências de chamadas de sistema.
- Filtro de Kalman: estima a ocorrência de estados considerando medições ruidosas. Zou *et al.* (2005) usam filtro de Kalman baseado na observação de tráfego ilegítimo de varreduras para prever a tendência de propagação de *worms* em seus estágios iniciais.

Independente da técnica, a predição de ameças à segurança sempre agregará a incerteza aos resultados. No entanto, prevendo uma ameaça, pode-se tomar medidas proativas para aumentar a segurança dos sistemas e das infraestruturas de rede e, assim, minimizar as chances e os prejuízos associados ao comprometimento desses recursos.

### 2.3.5 Resposta a incidentes e disseminação de alertas

As respostas a incidentes são a execução de contramedidas para conter, mitigar ou recuperar de uma violação das políticas de segurança. Em um EWS, após a detecção ou predição de ameaças, é mandatório realizar a disseminação rápida de alertas para aplicar medidas preventivas nos sistemas não atingidos e para auxiliar a recuperação de sistemas parcialmente ou totalmente comprometidos.

Os processos envolvidos em resposta a incidentes são compostos por uma organização complexa e devem respeitar aspectos legais e políticas locais, nacionais e internacionais. Os grupos para lidar com resposta a incidentes são denominados de *Computer Security Incident Response Team* (CSIRT) (West-Brown *et al.*, 2003). A colaboração entre CSIRTs existe, mas ainda é limitada.

EWS devem realizar a resposta em incidentes primeiramente por divulgação de alertas. Entretanto, a resposta a incidentes também contempla contramedidas e correção das ameaças. Diferentes formas de contramedidas e correção a ameaças são a atualização de bases de vulnerabilidades, atualização de bases de assinaturas de novos códigos maliciosos, atualização de listas de endereços e URL suspeitos e/ou que possuem códigos maliciosos, atualização de estatísticas sobre varreduras de portas, notificação a fabricantes sobre possíveis vulnerabilidades identificadas. Além disso, dependendo do nível de reação implementado, podem ser realizadas medidas proativas em diferentes níveis para conter ameaças, como a solicitação de bloqueio de tráfego suspeito em sistemas autônomos.

A disseminação de alertas pode ser realizada por meio de assinatura de notificações ou publicamente. A divulgação em bases de vulnerabilidades é um exemplo de notificação pública e por assinatura, enquanto que bases atualizadas de antivírus são divulgadas para assinantes do produto.

Há diferentes políticas a serem consideradas para a divulgação de alertas de segurança. Por exemplo, o comprometimento ou descoberta de vulnerabilidades em sistemas de agências financeiras não devem ser divulgadas publicamente por motivos de privacidade e reputação das organizações, e também, para evitar ataques massivos e pânico dos clientes. Por outro lado, uma vulnerabilidade em um sistema operacional voltado para usuários domésticos e corporativos deve ser notificada amplamente para acelerar o processo de correção.

A comunicação com parceiros e assinantes em EWS também deve ocorrer respeitando questões de privacidade durante a notificação de ameaças e também o grau de colaboração dos envolvidos. Um EWS depende da colaboração para otimizar os resultados de detecção e predição. Logo, parceiros devem ser priorizados para receber as notificações e medidas reativas. Além disso, as notificações para parceiros podem ser documentadas com mais informações sobre a ameaça e variações. Essa documentação pode ser realizada incrementalmente com a ajuda dos próprios parceiros. Há modelos de EWS que proveem interfaces colaborativas para analistas colaborarem na análise e documentação de ameaças e soluções.

Os interesses durante a disseminação de alertas devem ser considerados a fim de evitar gargalos de propagação e facilitar os mecanismos automáticos de reação instalados nos sistemas de parceiros e usuários. Dessa forma, a resposta a incidentes nos ambientes pode ser realizada baseada em missão e impacto do alerta para a organização ou sistema. Além disso, os recursos humanos focariam nos alertas priorizados para a organização e não despenderiam tempo com alertas sem importância.

A comunicação pública de alertas pode ser realizada em redes sociais, visando disseminar as informações rapidamente. A disseminação de alertas deve considerar o público alvo para o alerta, por exemplo, ameaças globais devem ser divulgadas em diferentes línguas e escrita técnica adequada para o público. Para organizações, a linguagem técnica e estruturada é, independente da forma de divulgação, a mais adequada. Até porque, sistemas de segurança podem consumir os alertas divulgados para verificar automaticamente os sistemas computacionais.

### 2.3.6 Compartilhamento de informações

O compartilhamento de informações entre organizações deve considerar os aspectos legais, procedurais, técnicos, questões de confiança e interesses dos envolvidos (Bourgue *et al.*, 2013). Os aspectos legais e procedurais concentram-se em problemas como tipos de informações que podem ser trocadas e como deve ser realizada entre parceiros devido às restrições legais inerentes a cada organização. Os aspectos de confiança são importantes para a valorização da informação compartilhada e comprometimento das partes, ou seja, deve ser bidirecional entre os pares participantes. Os aspectos técnicos são ligados, principalmente, a questões de compartilhamento e qualidade dos dados.

Bourgue *et al.* (2013) afirmam que mesmo entre os próprios CERTS, há dificuldades técnicas, tais como:

- problemas com notificações (*feeds*): alterações de formatos sem aviso prévio, representação de horários entre zonas, informações atrasadas, dados insuficientes para realizar prevenção.
- divergência na taxonomia usada na informação compartilhada.
- número excessivo de falsos positivos.
- uso de software desenvolvido internamente e não interoperável com outras soluções.
- dificuldades para exportar dados entre sistemas.

- ausência de um serviço centralizado para troca de informações estruturadas.

Para lidar com algumas dessas questões, há diferentes arcabouços e padrões abertos e privados definidos para o compartilhamento de informações de segurança entre organizações, sistemas e divulgação pública. Nenhum dos padrões é efetivamente um padrão de fato, mas destacam-se os arcabouços *Incident Object Description Exchange Format* (IODEF) (Danyliw *et al.*, 2007) e o *Intrusion Detection Message Exchange Format* (IDMEF) (Debar *et al.*, 2007) do IETF Network Working Group[2], os padrões *Cyber Observable eXpression* (CybOX) (MITRE, 2014b), *Structured Threat Information eXpression* (STIX) (MITRE, 2014c) e *Trusted Automatic Exchange of Indicator Information* (TAXII) (MITRE, 2014d) especificados pelo MITRE [3], o *Open Indicators of Compromise* (OpenIOC) (OpenIoC, 2014) definido pela empresa Mandiant[4] e comunidade aberta, e o *Common Vulnerability Reporting Framework* (CVRF) (ICASI, 2014) definido pelo consórcio da indústria ICASI[5].

O IODEF e o IDMEF são arcabouços para compartilhamento de informações de segurança e possuem implementação em XML. O IODEF é usado para compartilhar informações de incidentes e o IDMEF é usado para compartilhar informações para a detecção de intrusão. O CybOX é um arcabouço aberto para especificar atributos e eventos de ambientes computacionais e de rede, o STIX é uma linguagem estruturada para descrever informações sobre ameaças à segurança e o TAXII é um arcabouço para especificar serviços e troca de informações entre organizações e sistemas. O OpenIOC é um esquema XML para compartilhar informações de ameaças em um formato padronizado. Foi desenvolvido para troca de informações entre produtos da empresa MANDIANT. A especificação é aberta e a comunidade de software livre pode contribuir com o projeto. O CVRF é um arcabouço para troca de informações sobre vulnerabilidades e ameaças à segurança apoiado por um consórcio de grandes empresas (Microsoft, Intel, Cisco, IBM, entre outras).

O CVRF na versão 1.1 apresenta um arcabouço estruturado e simplificado em relação a outros padrões para descrever e documentar vulnerabilidades e ameaças à segurança. Além disso, possibilita a realização de ligações com outras fontes de informações, por exemplo, o *Common Vulnerabilities and Exposures* (CVE) e produtos de software. Entretanto, para garantir a consistência dos documentos, foram definidos elementos raízes mandatórios, o que dificulta a geração automatizada do documento. O conjunto CybOX, STIX e TAXII possui o arcabouço mais completo para as necessidades de compartilhamento, entretanto a linguagem é mais complexa do que nos outros padrões. O OpenIoC apresenta uma linguagem mais legível, mas está intrinsecamente associado a uma empresa privada, o que dificulta sua adoção por outras empresas concorrentes. O IDMEF é uma solução simples se o objetivo é apenas descrever e representar alertas de segurança em um formato padronizado.

Comumente, o IDMEF ou extensões são usados em pesquisas e soluções para integração de IDS, por exemplo, no Prelude-IDS. O modelo de dados do IDMEF é composto por uma hierarquia de classes e tem como classe base a *IDMEF-Message* (Figura 2.3). As classes agregadas são a classe *Alert*, que especifica a ocorrência de um ou mais alertas, e a classe *Heartbeat*, que possibilita o envio opcional e periódico de mensagens de estado para os gerenciadores. A classe *Alert* agrega diversas classes e possibilita identificar informações de origem e tempo dos eventos, origem e destino do ataque, acrescentar reações e informações com relação ao evento ocorrido, entre outros.

No compartilhamento público de informações, destacam-se o CVE (MITRE, 2014a) e o *National Vulnerability Database* (NVD) (NIST, 2014). O CVE e o NVD são repositórios que armazenam e descrevem vulnerabilidades e riscos a ameaças computacionais. Viabilizam a automação e acesso por diferentes ferramentas de segurança, além de apresentarem medidas para indicar o nível de criticidade das vulnerabilidades catalogadas.

[2] https://www.ietf.org/
[3] http://www.mitre.org/
[4] https://www.mandiant.com/
[5] http://www.icasi.org/

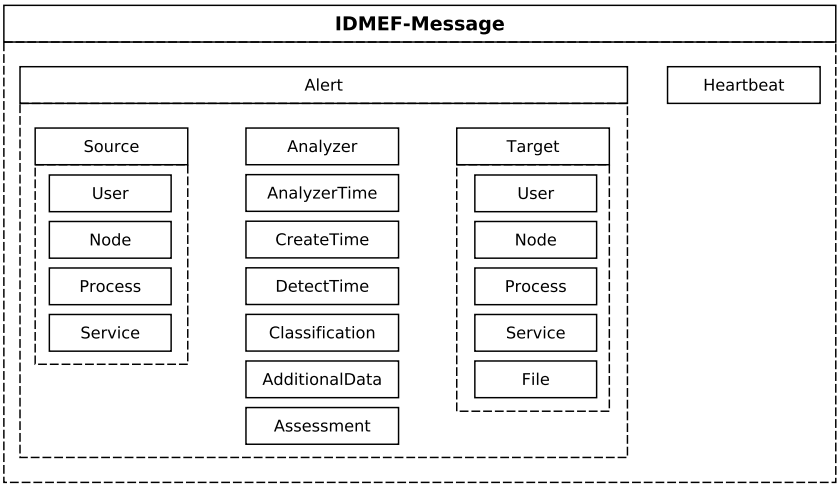


**Figura 2.3:** *Visão geral de uma mensagem IDMEF (Debar et al., 2007).*

### 2.3.7 Privacidade e confiabilidade

O compartilhamento de informações entre participantes de um EWS é essencial para o processo de geração de alertas antecipados relevantes. No entanto, dois pontos são críticos para a realização efetiva do compartilhamento: (i) privacidade e (ii) confiabilidade dos participantes e informações. Esses pontos afetam diretamente a informação a ser compartilhada. O uso de técnicas para garantir a privacidade envolvem a anonimidade da fonte de dados e das informações. As técnicas para garantir a confiabilidade implicam em estabelecer níveis de confiabilidade para os participantes e conhecimento sobre tipos, configurações e localização dos sensores.

A privacidade é comprometida quando dados sensíveis e a identidade dos contribuidores são usados por terceiros para obter vantagens ou realizar atividades maliciosas contra os contribuidores (Porras e Shmatikov, 2006). Além disso, aspectos legais associados aos dados compartilhados podem ser violados (Lincoln *et al.*, 2004). Por exemplo, um arquivo infectado e compartilhado pode conter informações sigilosas de uma organização ou do governo.

Os dados comumente considerados sensíveis são os endereços IP, endereços físicos, números de portas, protocolos, informações temporais e complementares de alertas, como arquivos infectados, registros de sistemas e carga de pacotes. Todas essas informações possibilitam revelar, diretamente ou por meio de associações, a topologia de rede, vulnerabilidades, abrangência de cobertura dos mecanismos de segurança, informações confidenciais de organizações, entre outras informações sensíveis (Farah, 2013; Lincoln *et al.*, 2004). E, por consequência, viabilizam aos atacantes meios para selecionar potenciais alvos e os melhores vetores de ataque.

Os principais ataques usados contra a privacidade de alertas compartilhados são: navegação casual em busca por endereços e sensores conhecidos, análise a respostas de sondagens coordenadas (*watermarking*), ataques de dicionário a alertas específicos, inundação de alertas a um alvo para identificar a variação do número de informações compartilhadas, corrupção de repositórios para ter acesso direto às informações dos contribuidores e dos alertas (Lincoln *et al.*, 2004; Parekh, 2007).

As soluções existentes para proteger a privacidade de alertas compartilhados são baseadas em funções de espalhamento e criptográficas para anonimato de informações sensíveis, na remoção ou cifragem dos dados de pacotes de redes, e no uso de roteamento anônimo para ocultar a origem das fontes (Xu e Ning, 2008). O processo de remover ou ocultar informações sensíveis obtidas por monitoramento de redes e ainda preservar a utilidade dos dados é denominado anonimização ou sanitização.

Crypto-PAn[6], AnonTool[7] e PktAnon[8] são algumas das ferramentas usadas para realizar a anonimização de informações coletadas em sistemas de monitoramento. Essas ferramentas usam diversos algoritmos para ocultar ou remover informações sensíveis, como por exemplo, remover ou substituir uma informação por um valor fixo (*black marker*), funções de espalhamento para substituir uma informação por um resumo criptográfico (*hash*), permutações aleatórias de endereços e preservação de uma parte da informação (*pseudo-anonymization*)(Farah, 2013). Os métodos de anonimização devem ser definidos por uma política de privacidade baseada nos interesses dos contribuidores (Lincoln *et al.*, 2004; Porras e Shmatikov, 2006).

No entanto, os métodos usados na anonimização de alertas interferem na utilidade dos dados no processo de correlação (Lincoln *et al.*, 2004; Parekh, 2007; Porras e Shmatikov, 2006; Xu, 2006). Na literatura, são apresentados alguns métodos para realizar a sanitização e correlação de alertas de forma a preservar a privacidade (Parekh, 2007). Por exemplo, Xu (2006) propõe as abordagens de generalização e pertubação para sanitização de alertas e, a correlação e construção dos cenários de ataques usando similaridade de atributos sanitizados e probabilidade de conexões entre os alertas. Parekh (2007) apresenta uma arquitetura para correlação de alertas focando em privacidade e faz o uso de filtros de Bloom, transformações de frequência e Z-strings. Burkhart *et al.* (2010) utilizam computação multiparte segura aplicada à agregação de eventos de múltiplos domínios de redes.

Apesar da possibilidade de empregar métodos para garantir a privacidade, nos EWS é interessante que as partes possam estabelecer a relação de confiabilidade com um órgão neutro (p. ex. CERT) e, dessa forma, obter predições mais precisas e disseminar rapidamente alertas, de forma anônima, aos interessados. Logo, é uma solução viável alcançar a privacidade garantindo a confiabilidade entre as partes.

A confiabilidade, por sua vez, pode ser comprometida por informações imprecisas e por participantes maliciosos. As informações imprecisas podem ser originadas por diversos fatores, por exemplo, imprecisão nos processos de coleta e detecção, configuração inapropriada dos sensores, restrições da política de compartilhamento, informações incompletas compartilhadas, entre outros. Já os participantes maliciosos podem executar diversos tipos de ataques, como sybil, novatos, traidores, conluio e de inconsistência (Fung, 2013).

Algumas soluções para amenizar o impacto de informações imprecisas consistem em diminuir a importância dessas na fase de detecção e predição de ameaças, corrigir ou validar informações incompletas com informações de outros participantes, classificar a confiabilidade e importância da informação baseado no participante e no tipo de sensor. Soluções para evitar os ataques comuns a confiabilidade incluem o uso de mecanismos de autenticação e de aumento da confiança baseado no tempo de contribuição e qualidade da informação, a diminuição da confiabilidade de participantes que contribuem negativamente para o sistema, intervenção de especialistas na análise das informações compartilhadas, principalmente para ataques mais complexos como de conluio e inconsistência.

Apesar das soluções para aumentar a confiabilidade e a qualidade do processo de correlação de alertas, muitas são conflitantes com os interesses de privacidade das fontes de informações. Além disso, as informações trafegadas em redes de computadores possuem dados pessoais, logo não é fácil realizar o compartilhamento considerando os aspectos legais. De qualquer forma, um EWS é viável somente se há participação e cooperação das partes envolvidas, logo confiança deve ser estabelecida entre as partes e mecanismos de anonimidade e privacidade devem ser empregados de forma a proteger os interesses dos contribuidores, mas sem comprometer o processo de detecção antecipada.

[6] http://www.cc.gatech.edu/computing/Telecomm/projects/cryptopan/
[7] http://www.ics.forth.gr/dcs/Activities/Projects/anontool.html
[8] http://www.tm.uka.de/software/pktanon/

## 2.4 Mineração de dados não estruturados

A mineração de alertas e ameaças à segurança a partir de dados não estruturados obtidos de fontes heterogêneas acaba por gerar uma quantidade enorme de mensagens com os mais diversos conteúdos. Conforme (Santos *et al.*, 2012), a coleta de informações de segurança pode resultar em: (i) alertas relevantes à segurança de sistemas; (ii) alertas específicos para um público (por exemplo, administrador de redes ou usuário doméstico); (iii) mensagens sobre segurança não computacional (por exemplo, mensagens de segurança pública); (iv) mensagens de propagandas (por exemplo, propaganda de antivírus); (v) mensagens de segurança desatualizadas e não mais relevantes como alertas; (vi) outros conteúdos irrelevantes (por exemplo, mensagens mal formadas, rumores ou com outros conteúdos).

Para lidar com essas questões, nesta tese são usadas técnicas de Recuperação de Informação e Processamento de Linguagem Natural (PLN) em conjunto com mecanismos de classificação e recomendação de alertas. Esta seção aborda alguns conceitos bases sobre esses assuntos que foram usados durante o desenvolvimento dessa tese.

### 2.4.1 Classificadores

Os sistemas de classificação objetivam determinar a classe a qual um objeto pertence considerando um conjunto de classes (Manning *et al.*, 2008). Os algoritmos para realizar a classificação são categorizados como estruturais ou estatísticos (Marmanis e Babenko, 2009). Também podem ser categorizados segundo a técnica usada na aprendizagem, nesse caso, como supervisionados e não supervisionados. O problema de classificar textos em alertas de segurança de sistemas ou não alertas, é um problema de classificação textual de duas classes. Os requisitos do sistema de classificação, as características dos textos e a avaliação dos resultados retornados por cada algoritmo são critérios comuns usados para selecionar os algoritmos de classificação.

São exemplos de algoritmos de classificação usados em classificação textual (Feldman e Sanger, 2006):

- Naive Bayes: é um método de aprendizagem probabilístico que identifica se um documento pertence a uma classe pela probabilidade condicional das características ocorrerem no documento segundo uma classe. Assume que as características são condicionalmente independentes.
- Entropia máxima: é um método de aprendizagem probabilístico baseado no princípio de entropia máxima. Identifica se um documento pertence a uma classe pela distribuição condicional de uma classe dado as características. Assume que as características não são condicionalmente independentes.
- Árvores de decisão: é um método de aprendizagem baseado em regras de decisão inferidas a partir das características dos dados. Consiste de um fluxograma em forma de árvore, onde o nó inicial e o os nós intermediários são nós de decisão que consideram as características e os nós folhas são as classes. Define a classe para um documento de entrada ao percorrer um caminho da raiz até uma folha.
- *Support Vector Machine* (SVM): é um método de aprendizagem baseado na construção de hiperplanos em um espaço multidimensional que particiona as classes. Usa uma função matemática, denominada *kernel*, para mapear a entrada para um espaço de características.

A avaliação de desempenho de um classificador pode ser realizada por uma matriz de confusão. Para um classificador de duas classes (binário), é uma tabela 2x2, onde o cabeçalho das linhas indica as classes (sim e não) e das colunas a predição para as classes (Figura 2.4).

| | Predição: Classe A | Predição: Não Classe A |
|---|---|---|
| Classe A | TP | FP |
| Não Classe A | FN | TN |

TP - Verdadeiros positivos
FP - Falsos positivos
TN - Verdadeiros negativos
FN - Falsos negativos

**Figura 2.4:** *Matriz de confusão*

O TP indica o número de instâncias classificadas como positivas corretamente, o TN indica o número de instâncias classificadas como negativas corretamente, o FP indica o número de instâncias classificadas como positivas erradas e o FN indica o número de instâncias classificadas como negativas erradas.

A partir dessa matriz, algumas métricas usadas na avaliação são:

$$precisao = \frac{TP}{TP + FP} \tag{2.1}$$

$$abrangencia = \frac{TP}{TP + FN} \tag{2.2}$$

$$acuracia = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.3}$$

A precisão (Equação 2.1) indica o percentual de acertos para a classe, a abrangência (Equação 2.2) indica o percentual de acertos em relação a todas as instâncias daquela classe, e a acurácia (Equação 2.3) indica o percentual de acertos das duas classes com relação ao total de instâncias.

### 2.4.2 Recomendadores

Os sistemas de recomendação possibilitam recomendar itens para usuários segundo as características desses itens (baseados em conteúdo) ou similaridade entre usuários e/ou itens (filtragem colaborativa) (Rajaraman e Ullman, 2011; Ricci *et al.*, 2011). Esses sistemas têm sido amplamente usados para lidar com a sobrecarga de informações (Ricci *et al.*, 2011). Logo, o uso de sistemas de recomendação é uma forma de colaborativamente viabilizar a priorização de alertas e descarte de falsos positivos baseado na recomendação por especialistas e interesses dos usuários. Como implicações diretas, há a redução do número de alertas e do custo de manter profissionais alocados para realizar a tarefa de análise desses alertas. Além disso, um sistema de recomendação pode selecionar os especialistas baseados no conhecimento para aumentar o grau de confiança na avaliação de alertas.

A filtragem baseada em conteúdo apresenta limitações associadas a análise do conteúdo e do usuário, também não considera a qualidade dos itens. Além disso, há o problema de não recomendar itens que não são similares ao perfil do usuário, mas que mesmo assim poderiam ser de interesse. A filtragem colaborativa apresenta limitações associadas ao tratamento de novos itens e usuários no sistema (problema de início-frio) e espaço de recomendações esparso, isto é, a maioria das entradas são desconhecidas (Feldman e Sanger, 2007). Alternativas para lidar com essas limitações envolvem o uso de abordagens híbridas, por exemplo, a abordagem híbrida mista que combina em uma lista os resultados da filtragem colaborativa e da baseada em conteúdo (Burke, 2002, 2007). É interessante por evitar o problema de recomendação de novos itens, mas ainda sofre com o problema de novos usuários.

A avaliação de recomendadores pode ser realizada por três abordagens: experimentação *offline*,

estudos de usuários e experimentação *online*. A experimentação *offline* usa bases de dados com o histórico de interações dos usuários e itens avaliados. Como ponto positivo, possibilita avaliar e comparar os algoritmos de recomendação. No entanto, não possibilita verificar a influência do sistema em produção nos comportamentos e avaliações dos usuários. O estudo de usuário é realizado por meio da seleção de um grupo de usuários que deve interagir com o sistema de recomendação. Como pontos positivos, coleta de dados qualitativos e a observação da influência da recomendação no comportamento dos usuários. Em contrapartida, precisa-se de um grupo de usuários e o trabalho de observação é oneroso. A experimentação *online* é a avaliação da interação direta dos usuários com o sistema em produção. Como ponto positivo, possibilita avaliar se o modelo de recomendação atende as principais propriedades e outras importantes como a retenção de usuários. Em contrapartida, tem-se a dificuldade de realizar a avaliação de diferentes algoritmos em um sistema em produção.

### 2.4.3 Processamento de Linguagem Natural

O Processamento de Linguagem Natural consiste no uso de modelos e técnicas computacionais, linguísticas e estatísticas para o processamento da língua natural (Feldman e Sanger, 2006; Manning e Schütze, 1999). Pode ser empregado para a análise fonética, fonológica, morfológica, sintática, semântica e pragmática de informações representadas em uma língua natural (Indurkhya e Damerau, 2010; Manning e Schütze, 1999). Nesta tese, os alertas e ameaças à segurança são obtidos de dados textuais, logo o PLN trata apenas dos quatro últimos itens.

A análise morfológica consiste na análise das formas das palavras. Dois conceitos associados à morfologia são a lematização e a radicalização. A análise sintática consiste na análise da estrutura da sentença. Em conjunto com a morfológica, há o conceito de classificação gramatical (*Part-of-Speech* (PoS)). A análise semântica consiste na análise do significado. Alguns conceitos dessa análise são a identificação de entidades nomeadas, relações entre palavras, desambiguação de sentido das palavras, entre outros. E a análise pragmática analisa a sentença dentro de um contexto.

Outro conceito explorado nesta tese é a mineração de associação de palavras, pois possibilita identificar relações paradigmáticas, sintagmáticas e *collocations* entre palavras dentro de um contexto (Manning e Schütze, 1999). As relações paradigmáticas identificam palavras que podem ser substituídas por outras de uma mesma classe e mantêm o significado da sentença; as relações sintagmáticas relacionam palavras semanticamente; e as *collocations* são expressões compostas por dois ou mais termos comumente usados para expressar algo.

Em síntese, com o uso de PLN é possível identificar termos e entidades nos alertas, executar o pré-processamento léxico e sintático de dados coletados, desambiguar termos que não são associados à segurança computacional, extrair características para algoritmos de classificação e categorização de alertas, entre outras tarefas de processamento de texto.

### 2.4.4 Recuperação de Informação

A Recuperação de Informação consiste em encontrar informações de natureza não estruturada, geralmente dados ou documentos textuais, que satisfazem um critério de busca aplicado a um grande volume de informações (Manning *et al.*, 2008). Os conceitos e técnicas de Recuperação de Informação possibilitam identificar informações em fontes de dados não estruturados que caracterizam potenciais alertas, agrupar alertas similares, coletar informações em diferentes fontes, atribuir ponderações a termos em buscas, entre outros.

Dentre esses conceitos, destaca-se o índice invertido, que consiste em mapear os termos para as partes do documento onde ocorrem. O índice invertido possibilita a construção de modelos de recuperação de informação, como o modelo Booleano (*Boolean Model*) e o modelo de Espaço Vetorial (*Vector Space Model*).

O modelo Booleano realiza a busca considerando expressões lógicas com os termos, ou seja, recupera documentos com os termos buscados sem considerar a importância dos termos. O modelo de Espaço Vetorial calcula um escore para classificar quão similar é o documento em relação ao termo utilizado na busca. Utiliza cálculos de similaridade como Term Frequency–Inverse Document Frequency (TF-IDF) e similaridade de cosseno para tal finalidade. Quanto maior o escore, maior a similaridade, dessa forma é possível classificar a importância do documento em relação à busca (Manning *et al.*, 2008).

A avaliação de sistemas de recuperação de informação geralmente consideram a precisão (Equação 2.1) e a abrangência (Equação 2.2). A precisão mede o percentual de documentos devolvidos que são relevantes segundo a informação consultada e a abrangência mede o percentual de documentos relevantes que foram devolvidos considerando todos os possíveis documentos que atenderiam a informação consultada.

## 2.5 Trabalhos relacionados

### 2.5.1 Arquiteturas e sistemas

Nos últimos anos, as arquiteturas e sistemas de EWS começaram a ganhar notoriedade no meio acadêmico e na indústria. Termos associados a EWS, como *Cyber*, *Threat*, *Cyberthreat*, *Security* antecedendo *Intelligence* e/ou após *Open*, têm sido amplamente usados em muitos produtos comerciais e abertos. Já na área acadêmica, propostas de diversas arquiteturas e sistemas procuram abordar as características e desafios discutidos na Seção 2.3. Nesta seção, são apresentadas algumas arquiteturas e sistemas de EWS, procurando destacar as fontes de detecção, os métodos usados para correlação e predição, as vantagens e limitações sob a perspectiva de EWS.

#### 2.5.1.1 AMSEL

O AMSEL (Apel *et al.*, 2009, 2010), desenvolvido na Universidade de Dortmund, é um EWS de códigos maliciosos que não necessita de intervenção humana para coletar, analisar, gerar assinaturas e alertas de novos códigos maliciosos. A arquitetura (Figura 2.5) é composta por quatro componentes: módulo de coleta e aprendizagem (CL), módulo de detecção e alerta (DA), repositório de ameaças e repositório de alertas.

O módulo CL coleta e gera assinaturas de novos códigos maliciosos. A coleta é realizada usando *honeypots*. São aplicadas técnicas de análise estática e dinâmica para identificar e extrair características dos códigos, diferenciando, assim, os códigos benignos e maliciosos. A assinatura é gerada pelo agrupamento de códigos maliciosos com características similares (distância de Manhattan) usando algoritmos de agrupamento hierárquico. A assinatura comportamental pode ser obtida aplicando o algoritmo de Ukkonen (Ukkonen, 1995) em cada grupo de código malicioso. O algoritmo constrói, em tempo linear, uma árvore de sufixos de strings que possibilita identificar padrões. Neste caso, são identificadas sequências de chamadas de sistema idênticas nos elementos dos grupos, descartando as chamadas que fazem parte do grupo de controle de chamadas de sistema seguras. Outra alternativa pode ser o uso do SVM aos grupos e a assinatura corresponderia ao hiperplano resultante. Por fim, a assinatura e o código malicioso são armazenados no repositório local de ameaças.

O módulo DA detecta e alerta sobre incidentes de segurança usando as informações locais e o repositório de ameaças, um repositório centralizado e atualizado remotamente pelos módulos CL distribuídos entre organizações. Os módulos DA devem atualizar o repositório de alertas, um repositório centralizado para armazenar alertas de diversas origens e fornecer um mapeamento geral sobre um incidente. Os módulos CL e DA devem ser operados por organizações, enquanto os repositório de ameaças e de alertas por organizações governamentais, como o CERT, por exemplo.

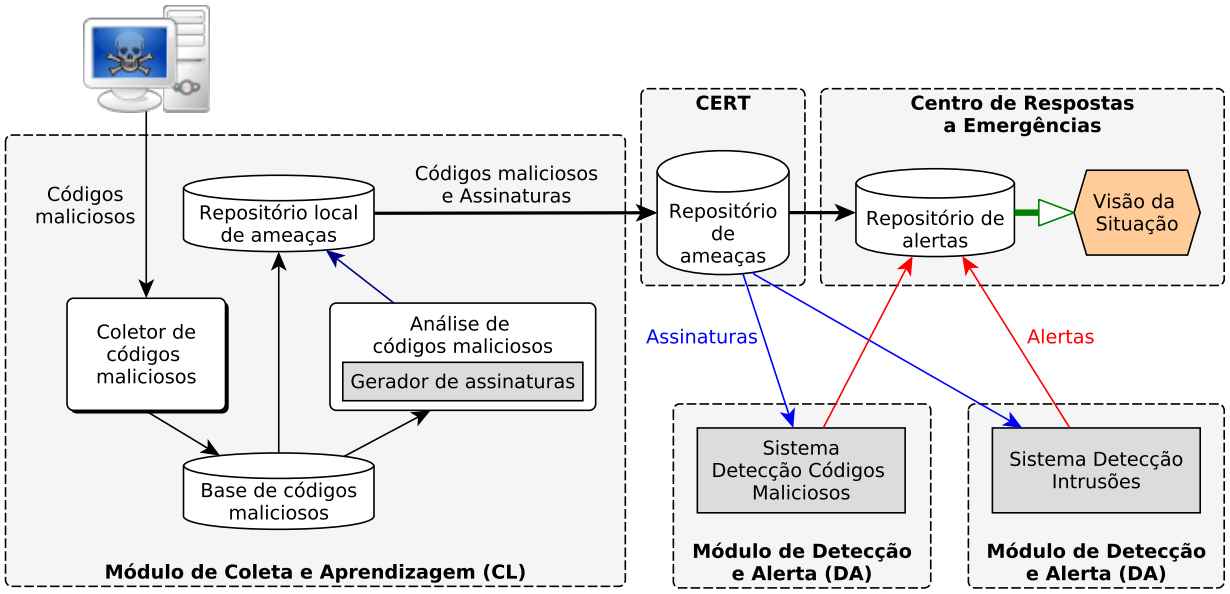


**Figura 2.5:** *Arquitetura do AMSEL (Adaptado de Apel et al. (2010)).*

Podem notificar seus integrantes informando nomes ou portas de serviços explorados, origem de códigos maliciosos, origem de vítimas comprometidas, entre outros.

A arquitetura possui módulos distribuídos, mas os repositórios de ameaças e alertas são centralizados. A centralização visa facilitar a disseminação de alertas e ameaças, além de garantir a proteção e confidencialidade dos dados coletados dos módulos CL e DA. Um dos pontos positivos do AMSEL é o processo automatizado de predição e geração de notificações. Por outro lado, o processo automatizado pode gerar uma grande quantidade de assinaturas ou alertas não relevantes para as organizações.

#### 2.5.1.2 CarmentiS

O projeto CarmentiS (Grobauer *et al.*, 2006), desenvolvido pela Associação dos CERTs da Alemanha[9], provê uma infraestrutura e um arcabouço organizacional de compartilhamento e cooperação para a análise e correlação de dados de diferentes sensores. A arquitetura do CarmentiS depende de três entidades: Parceiros, CERTs e Governo (Figura 2.6). Parceiros são organizações que realizam o monitoramento, enviam dados de sensores e notificam a situação local. CERTs mantêm repositórios centralizados para armazenar e coordenar as informações e análises, além de emissão de alertas antecipados. O Governo é responsável por proteger as infraestruturas críticas, tais como comunicações, transportes e energia.

CarmentiS apresenta uma infraestrutura centralizada no Governo e no CERT para cooperação entre organizações. A cooperação é por meio do compartilhamento, correlação e análise dos dados. Os dados são analisados por especialistas, baseados na seleção por um perfil. É responsabilidade das organizações exportarem os dados para um formato suportado pelo CarmentiS. Os tipos de dados suportados são fluxos de redes e alertas de IDS. Usam padrões e ferramentas existentes para realizar a exportação e transporte dos dados. Um problema é o volume de dados, logo seria interessante a existência de processos automatizados de análise. Em geral, o CarmentiS apresenta-se como uma infraestrutura de colaboração para análise de incidentes.

[9] https://www.cert-verbund.de/

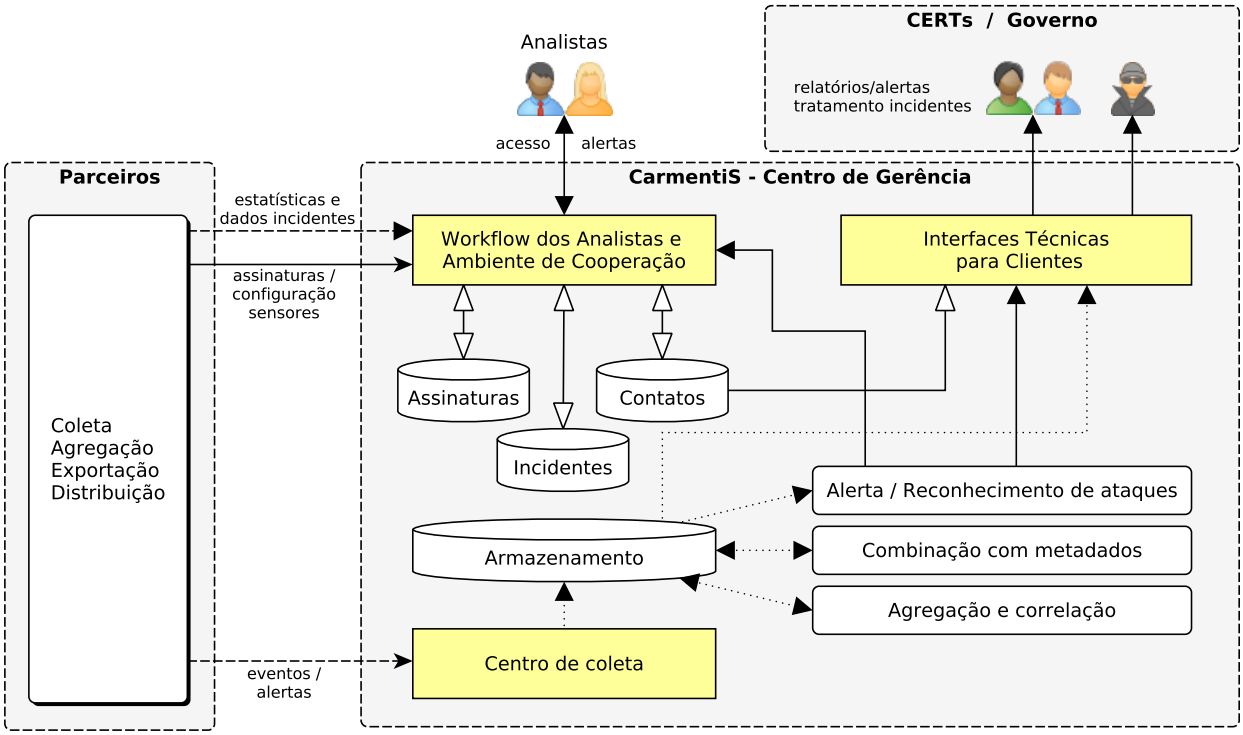


**Figura 2.6:** *Arquitetura do CarmentiS (Adaptado de Grobauer et al. (2006)).*

#### 2.5.1.3 Internet EWS

Bastke *et al.* (2010) propõem uma arquitetura distribuída composta por componentes locais de EWS e centros situacionais. Justificam a escolha de componentes locais devido a diferentes políticas e condições de ambientes, rápida disseminação de ameaças, legislações diferentes entre países e o fato de inspeção em pacotes criptografados só poder ser realizada nos destinos finais. A arquitetura proposta (Figura 2.7) é composta pelos seguintes componentes: sensores, detecção, base de conhecimento, gerenciamento de incidentes, base de evidências e distribuição da informação.

Os sensores mais importantes são os *honeypots*, pois identificam códigos maliciosos e novos *exploits*. Outros sensores podem ser empregados para monitorar tráfego de *botnets*, roteadores, serviços de nomes, spams e *phishing*. Os componentes de detecção devem aplicar técnicas de detecção de anomalias, métodos para correlação de eventos e predição de incidentes, e devem importar informações de origem externa para auxiliar na detecção. A base de conhecimento deve armazenar as informações relevantes para o EWS, por exemplo, assinaturas de códigos maliciosos, identificação de vulnerabilidades, metodologias de ataques, aspectos legais e organizacionais de instituições, entre outros. O gerenciamento de incidentes deve ser um sistema especialista que auxilia na construção da situação e nas medidas reativas. A base de evidências é usada em processos legais e análise mais detalhada do incidente. A distribuição da informação deve especificar o tipo de notificação e ser emitida o mais rápido possível. A comunicação entre as entidades é P2P e cada organização deve possuir um módulo local de EWS. Há os módulos centralizadores, responsáveis por geração de cenários globais e por armazenar informações de ameaças e estatísticas.

A proposta de uma arquitetura P2P é interessante por viabilizar processamento local nas entidades, mas dificulta lidar com as questões de privacidade e confidencialidade das organizações. Os proponentes do Internet EWS discutem conceitualmente os componentes de EWS, mas não apresentam a implementação de um protótipo. A arquitetura depende da interação de especialistas nos sensores e no gerenciamento de incidentes.

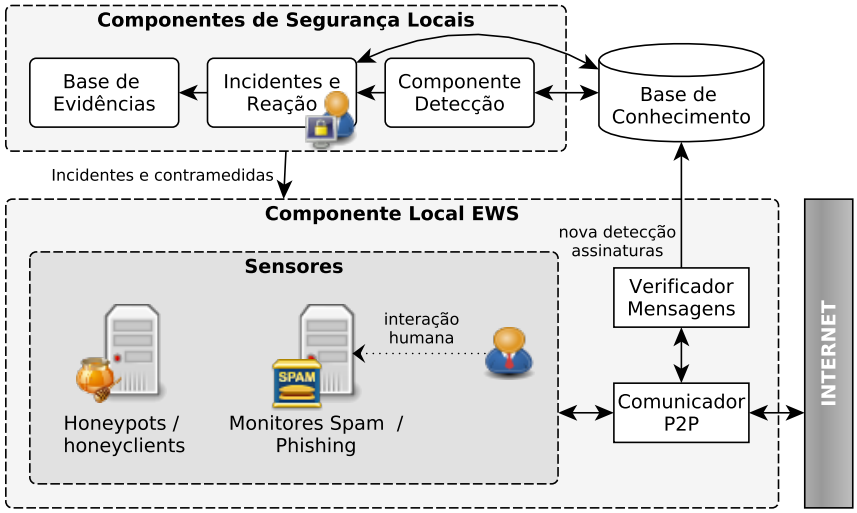


**Figura 2.7:** *Arquitetura do Internet EWS (Adaptado de Bastke et al. (2010)).*

#### 2.5.1.4 InMAS

O *Internet Malware Analysis System* (InMAS) (Engelberth *et al.*, 2010), desenvolvido na Universidade de Mannheim, foi implementado para monitorar códigos maliciosos na Internet. A arquitetura do InMAS possui quatro principais módulos (Figura 2.8): sensores, repositório, análise e interface Web.

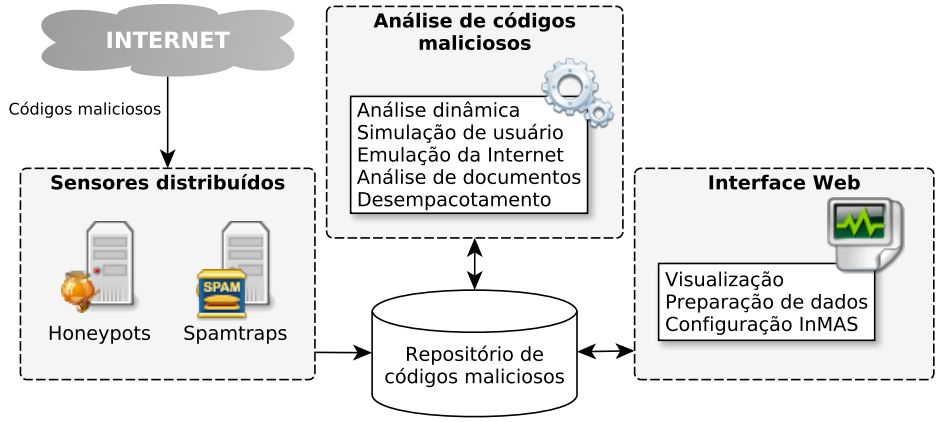


**Figura 2.8:** *Arquitetura do InMAS (Adaptado de Engelberth et al. (2010)).*

A arquitetura é centrada no repositório, responsável por armazenar os códigos maliciosos em conjunto com as análises e assinaturas. A interface Web apresenta os resultados das análise e permite a configuração do InMAS. O módulo de sensores é responsável por coletar os códigos maliciosos por meio de *honeypots* de baixa e alta interatividade. A detecção de códigos maliciosos em emails é realizada por *spamtraps* - contas de email que não deveriam enviar e receber mensagens. O módulo de análise realiza análise dinâmica, análise de documentos e identificação de código ofuscado usando o CWSandbox, um ambiente restrito para a análise dinâmica de códigos maliciosos. Além disso, são usadas simulações de usuários para gerar eventos e detectar códigos maliciosos que operam somente em circunstâncias específicas e a emulação de Internet para evitar o envio de tráfego para a rede real.

O InMAS utiliza um conjunto de ferramentas de código aberto para coleta e análise de códigos

maliciosos na Internet. É apresentado como um sistema de EWS pelos autores, por permitir identificar códigos maliciosos a partir de sensores distribuídos em diversas partes da Internet. Apesar de identificar os códigos, não possui uma estrutura de colaboração para compartilhamento e correlação de dados, logo não permite a propagação de alertas de forma colaborativa e antecipada. Por consequência, o InMAS é mais uma plataforma de monitoramento de códigos maliciosos que um EWS.

#### 2.5.1.5 OSINF

O OSINF (Dorges e Sander, 2010) é um protótipo para coleta, processamento e exibição de informações de segurança obtidas a partir de fontes abertas. O processo de gerenciamento da informação é composto por coleta, normalização, agregação, classificação, análise de relevância e publicação. O processo é apresentado na Figura 2.9. O protótipo objetiva extrair alertas antecipados a partir de fontes como blogs, bases de vulnerabilidades, listas de email, entre outros.

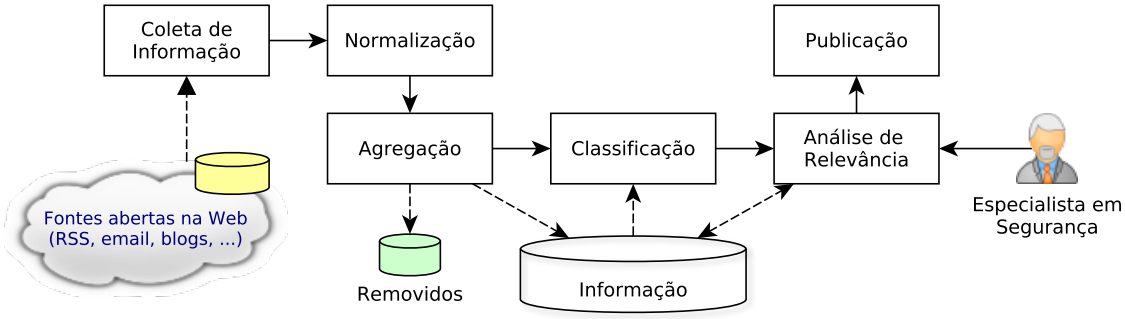


**Figura 2.9:** *Fluxo de operação do OSINF (Adaptado de Dorges e Sander (2010)).*

O protótipo foi desenvolvido usando um módulo para realização da coleta e processamento da informação e um módulo visual para publicação e interação com o usuário. O módulo de coleta e processamento foi desenvolvido em Ruby e o módulo visual foi baseado no OTRS[10] e escrito em Perl e XML.

A construção do protótipo procurou contemplar três requisitos de qualidade: informação, fontes e processo. A qualidade da informação considerou a avaliação da credibilidade, relevância e remoção de dados duplicados por um administrador. A qualidade das fontes considerou a análise do administrador e os resultados da avaliação de qualidade da informação. A qualidade do processo foi assegurada na coleta, normalização e processamento das informações.

Os pontos positivos do trabalho são o uso de informações de fontes abertas, a especificação do processamento da informação e a apresentação de resultados no OTRS. Os pontos negativos são a falta de detalhamento nas fases de normalização, agrupamento e classificação; o protótipo aparentemente exibe apenas a informação coletada; o protótipo não apresenta possíveis fortes relacionamentos entre fontes de diferentes origens sem intervenção humana; e não apresenta como as informações coletadas são relacionadas com as informações dos sensores de rede. O uso de redes sociais como fonte aberta não é discutida e nem implantada no protótipo.

#### 2.5.1.6 DOMINO

*Distributed Overlay for Monitoring InterNet Outbreaks* (DOMINO) é uma arquitetura para CIDS baseada na colaboração entre nós heterogêneos por meio de uma rede hierárquica. Tem por

[10] http://www.otrs.com

objetivo compartilhar informações de intrusão com todos os participantes. A arquitetura é composta por nós eixos, comunidades satélites e contribuidores térreos (Figura 2.10).

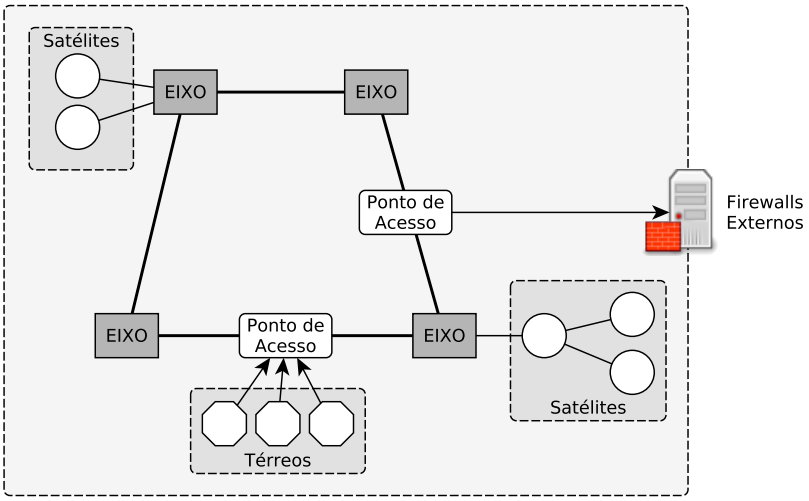


**Figura 2.10:** *Arquitetura DOMINO (Adaptado de Yegneswaran et al. (2004)).*

Os nós eixos são os principais componentes da arquitetura e atuam como pontos de coleta e compartilhamento de dados de intrusão. Cada nó mantém um NIDS e um *active-sink*. Os nós eixos mantêm uma relação de alto grau de confiança e trocam periodicamente informações de intrusão. As comunidades satélites são organizadas de forma hierárquica e são responsáveis por coletar dados de intrusão de redes maiores. Os dados coletados nos nós satélites são encaminhados para os nós eixos e são considerados menos confiáveis. Os contribuidores térreos são nós que coletam dados de participantes externos e possibilitam a integração com qualquer firewall ou NIDS. Fornecem a maioria dos dados, mas em geral, apenas resumos de varreduras de porta. Os pontos de acesso permitem incluir fontes de dados externas fora da infraestrutura do DOMINO. As mensagens em DOMINO são representadas em XML baseado em uma extensão do IDMEF.

Cada nó eixo mantém uma visão local (própria rede e satélites) e global (gerada a partir das mensagens recebidas de seus pares). A geração da visão pode ser implementada sem dependências entre os nós, ou seja, cada nó pode selecionar as estratégias para a agregação. Apresenta como ponto positivo diferentes níveis de confiança conforme o tipo de nó. Foi validado baseado em dados de duas origens: conjunto de registros coletados durante quatro meses de 1.600 redes distribuídas no mundo; implementação de uma rede (*sink*) que monitora 100.000 endereços IP. Os resultados da análise retrospectiva dos *worms* SQL-Snaker e SQL-Saphire, por meio da simulação dos nós de DOMINO, apresentaram a redução na taxa de alarmes falsos e no tempo de reação a epidemias de *worms*.

#### 2.5.1.7 Worminator

Worminator (Locasto *et al.*, 2005; Stolfo, 2004) foi desenvolvido para detecção e disseminação de alertas antecipados sobre *worms* e ameaças dia zero. Também possibilita a verificação da propagação de *worms* pela análise e agrupamento de origens de ataques com comportamentos similares. As informações compartilhadas são os endereços e portas observados por um NIDS em cada nó da arquitetura. A colaboração é realizada por meio de filtros de Bloom, uma estrutura que representa e compacta dados em um vetor de bits por meio de funções de espalhamento. Os filtros garantem a privacidade e diminuição do tráfego entre os pares. O uso de federações entre os participantes possibilita aos nós a inclusão ou remoção da rede a qualquer momento. A forma de organização pode ser centralizada (as informações são agregadas em partes confiáveis) ou descentralizada (os pares se comunicam diretamente).

Boggs *et al.* (2011) estenderam o Worminator para correlacionar solicitações Web anômalas entre múltiplos servidores Web localizados em diferentes domínios administrativos. Cada domínio possui um sensor local de detecção de anomalia para analisar as requisições para o servidor. A troca de informações é realizada em tempo real e com garantias de privacidade por meio de filtros de Bloom. A validação foi realizada usando dados de três domínios durante três meses. Se 80% de correlação é constatado para um alerta consultado por meio de filtro de Bloom, será classificado como um ataque. Como resultado, os autores do Worminator encontraram dois ataques dia zero e diminuíram o número de falsos positivos.

#### 2.5.1.8 Semantic Room

A Semantic Room (SR) (Lodi *et al.*, 2014) é uma abstração para o desenvolvimento de plataformas colaborativas de segurança na Internet. Viabiliza o compartilhamento controlado de dados entre diferentes sistemas e possibilita identificar ameaças e fraudes antecipadas por meio de correlação de eventos compartilhados entre as partes.

As organizações participantes são estruturadas em federações e denominadas membros. A premissa é que organizações com áreas de atuação similares sofrem os mesmos ataques. Uma SR é definida por três elementos principais: objetivo, contrato e implantações. O objetivo especifica o foco da SR, por exemplo, monitorar varreduras. O contrato regulamenta as obrigações e restrições da SR quanto ao serviço e processamento. Por exemplo, proteção de dados, privacidade, isolamento, confiança, segurança, desempenho, entre outros. As implantações são flexíveis para acomodar diferentes tecnologias e arquiteturas de serviço.

A Figura 2.11 apresenta a estrutura geral de uma SR e o relacionamento com as organizações. Os componentes *Processamento de Eventos Complexos e Aplicações* e *Disseminação de Dados* são dependentes da implantação, ou seja, das tecnologias e arquiteturas utilizadas.

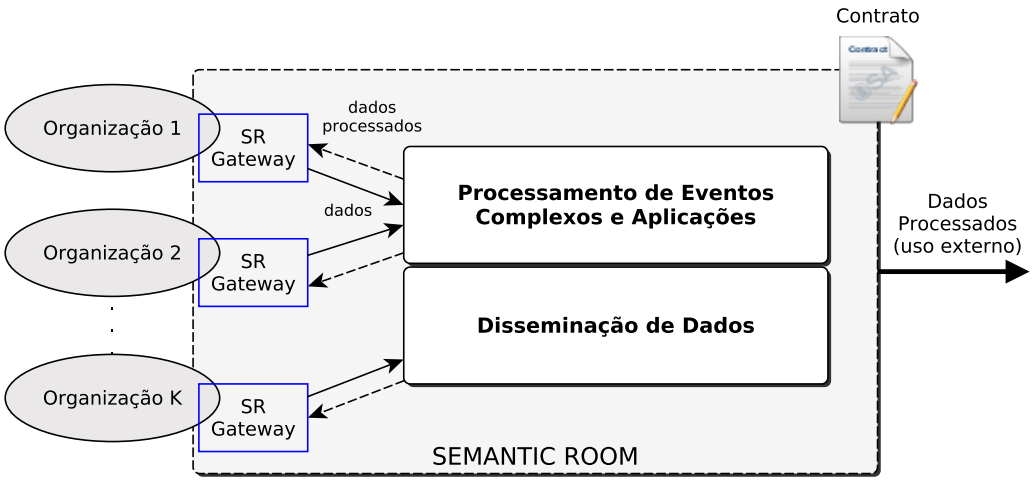


**Figura 2.11:** *Abstração de Semantic Room (Adaptado de Lodi et al. (2014)).*

Na Figura 2.11, observa-se que as organizações enviam dados para serem processados na SR usando os SR Gateways. Os SR Gateways são responsáveis pela coleta e pré-processamento dos dados, bem como pela comunicação com a SR. No pré-processamento, realizam a filtragem, agregação e anonimização dos dados segundo as cláusulas especificadas no contrato. Os dados processados na SR podem ser usados internamente, compartilhados com as organizações membros ou com comunidades externas. Os membros têm acesso a todos os dados puros e os processados especificados por contrato.

Lodi *et al.* (2014) apresentam duas implementações e avaliações usando SR. Uma aplicação para

detectar ataques de varreduras de portas silenciosos contra organizações membros da SR e outra para correlacionar informações de agências bancárias e financeiras relacionadas a fraudes na Itália. Na primeira, obtiveram alta acurácia e pequeno atraso nas detecções. Na segunda, descobriram novos indícios de fraudes e a detecção ocorre a tempo de reagir à propagação.

#### 2.5.1.9 DShield

DShield (DShield, 2014) é um serviço aberto e gratuito para coleta e análise de ameaças que provê alertas antecipados por meio de notificações. É um serviço provido pelo SANS ISC que coleta registros de *firewalls* de mais de 50 países. As ferramentas usadas para enviar os registros removem as informações de identificação e dados sensíveis. A detecção e os relatórios de ameaças e incidentes são realizados por analistas e métodos automatizados que analisam a base de dados de registros. Ao detectar uma ameaça, um grupo de especialistas especifica a prioridade e a forma de propagação do alerta, ou ainda, se necessário, o bloqueio de tráfego nos provedores de serviço de Internet. Apresenta uma arquitetura centralizada e analisa somente um tipo de fonte de informação. Entretanto, conta com a cooperação global de voluntários.

#### 2.5.1.10 Arakis

Arakis (CERT Polska, 2014) é um EWS desenvolvido pelo CERT polonês voltado a detecção e caracterização de novas ameaças que se propagam automaticamente. São usadas informações coletadas de redes distribuídas de *honeypots*, *firewalls*, antivírus e monitores de *darknets*. Os *honeypots* são usados para gerar e identificar assinaturas de potenciais ameaças. Os *firewalls* e monitores de *darknets* são usados como indicadores de anomalias para atividades de varreduras de portas. Os antivírus proveem informações sobre ameaças conhecidas. O principal alerta emitido é composto por uma assinatura para o Snort. Apesar da geração automatizada das assinaturas, há alta probabilidade de alertas não relevantes. Logo, é necessário intervenção humana para validar, refinar e aplicar essas assinaturas em ambientes reais. Por outro lado, o monitoramento de varreduras possibilita identificar tendências de ameaças ou a propagação de *worms*.

#### 2.5.1.11 DeepSight Early Warning Services

DeepSight (Symantec, 2014) é um conjunto de serviços de alerta antecipado proprietário da Symantec. Os dados são coletados a partir dos produtos de segurança da empresa. Permite, por meio de um portal Web, acesso a informações processadas de *firewalls*, IDS e *honeynets* distribuídos globalmente. Também dissemina informações de ameaças por meio de email, RSS, SMS ou XML. Correlaciona e mantém bases de vulnerabilidades, códigos maliciosos, riscos de segurança, endereços e domínios maliciosos, entre outros. Por ser um produto de uma empresa de grande porte e com muitos analistas, viabiliza a avaliação de confiabilidade das fontes de informações e a priorização de alertas. As principais limitações são os produtos voltados a grupos fechados, em especial grandes organizações, e a falta de compatibilidade com outras soluções.

#### 2.5.1.12 Outras arquiteturas

As arquiteturas descritas anteriormente possuem várias características que as associam a EWS. Entretanto, há muitas arquiteturas que propõem a colaboração ou cooperação entre partes para aprimorar o processo de detecção de intrusões. Nesta seção, são descritas outras arquiteturas que também apresentam contribuições para a detecção de ameaças antecipadas.

*Internet Motion Sensor* (IMS) tem por objetivo monitorar, caracterizar e medir o nível de ameaças à segurança em redes por meio de sensores espalhados na Internet(Bailey *et al.*, 2005). Os

sensores monitoram de forma passiva e ativa as faixas de endereços IP não utilizados. Os sensores ativos possuem a limitação de não conseguir interagir e identificar as ameaças que dependem de respostas em nível de aplicação. Utiliza a função de espalhamento MD5 para comparar a carga de dados de pacotes e, assim, realiza o armazenamento uma única vez. O IMS foi desenvolvido na Universidade de Michigan em conjunto com a Arbor Networks e, atualmente, o sistema está englobado no *Active Threat Level Analysis System* (ATLAS).

Zou *et al.* (2005) usam filtro de Kalman para detectar a propagação de *worms* nas fases iniciais. Identificaram que a busca de novos alvos por um *worm* apresenta características similares de tráfego devido ao fato do objetivo nas fases iniciais ser a rápida propagação. Utiliza dois monitores distintos de tráfego: um monitora o tráfego de saída para identificar o comportamento de varredura de um *worm* e o outro monitora o tráfego de entrada para registrar o tráfego com destino a endereços locais não usados. A premissa de detecção é que as varreduras de *worms* causam o aumento exponencial do tráfego dos monitores. Ao correlacionar os dados observados em monitores distribuídos é possível detectar antecipadamente *worms* e calcular a taxa média e a distribuição das varreduras.

Bsufka *et al.* (2006) apresentam uma arquitetura denominada A-EWS, um sistema baseado em agentes para sistemas de alerta antecipado direcionado a infraestruturas críticas. Utiliza sensores distribuídos na infraestrutura monitorada para detectar e disseminar informações de ataques. A distribuição ótima dos sensores tem sido explorada por meio de abordagens da teoria dos jogos. Ataques identificados não ajudam a infraestrutura comprometida, mas são propagados para outros alvos potenciais.

O *Internet Analysis System* (IAS) é um sistema que extrai informações estatísticas a partir de tráfego monitorado na Internet. A ideia é que observando comportamentos e variações de padrões de tráfego é possível correlacionar com informações de outras ferramentas para detectar, em estágios iniciais, ameaças à segurança ou tráfego não desejado (Pohlmann e Proest, 2006). As estatísticas podem ser armazenadas para mostrar a evolução do tráfego. Certa de 870.000 parâmetros e suas combinações são coletados a partir dos cabeçalhos dos pacotes. Os autores do IAS acreditam que visões temporais de fatos locais possibilitam a identificação de ameaças. Por exemplo, o aumento de número de mensagens eletrônicas com conteúdo compactado (Hesse e Pohlmann, 2008).

Kim *et al.* (2008) propõem o *Intrusion Forecasting System* (IFS), um EWS que combina as abordagens de análise de séries temporais, modelagem probabilística e mineração de dados. Um alerta é emitido sempre que duas abordagens acusam um comportamento anômalo. Kim et al. argumentam que a análise de séries temporais permite observar a mudança de variáveis no decorrer do tempo, mas são imprecisas quando há mudanças repentinas no tráfego normal e na definição de limiares, enquanto a modelagem probabilística permite estabelecer facilmente escalas de probabilidades de ataque e níveis de alertas. Por sua vez, métodos de mineração de dados consideram muitas características de múltiplas variáveis, mas proveem resultados complexos de serem compreendidos. Nos experimentos foram usadas MMEP para a análise de séries temporais, Cadeia de Markov para a modelagem probabilística e uma modificação de agrupamento de dados para detecção de DDoS como técnica de mineração de dados. A avaliação experimental foi realizada na base DARPA 2000 e foi verificado que a combinação de diversos métodos apresenta melhores resultados do que se aplicados individualmente.

Pontes e Guelfi (2009) apresentam uma arquitetura colaborativa de detecção de intrusão considerando diferentes técnicas de predição. Para tal, são usadas cinco técnicas de predição, em destaque MMEP, e quatro níveis de análise. O primeiro nível analisa os alertas independentes gerados pelos hospedeiros. O segundo nível realiza a correlação de alertas dos hospedeiros. O terceiro nível analisa a rede e observa as bordas da rede local. O quarto nível ocorre em nível de sistemas autônomos. Os níveis superiores dependem da análise dos níveis anteriores. Pontes e Guelfi realizaram uma prova de conceito considerando os três primeiros níveis e um ambiente composto por três cenários geograficamente distribuídos. Foram usados HIDS e NIDS como sensores.

Meissen e Voisard (2010) propõem uma arquitetura de referência para um EWS genérico com-

posto por três principais subsistemas: monitoramento, detecção e alerta. O subsistema de monitoramento é responsável por monitorar eventos por meio de medições ou estimativas em intervalos de tempo especificados. Os dados devem estar em um formato bem definido. Um componente de gerenciamento deve gerenciar as atividades dos sensores, como comunicação e controle. Além disso, os dados devem ser submetidos aos processos de filtragem e fusão visando a qualidade da informação. O subsistema de detecção é responsável por detectar ameaças e predizer possíveis cenários com base nas informações coletadas. A detecção pode ser realizada por processos automáticos, semi-automáticos ou exclusivamente por especialistas humanos. Em geral, são utilizados limiares, regras, modelos e outras fontes de informação para a realização da detecção. Por sua vez, um componente de projeção e seleção de cenários permite inferir sobre o impacto da ameaça. O subsistema de alerta é responsável por converter informações de ameaças em alertas e disseminar as informações para recipientes de destino. Opcionalmente, pode implementar filtros baseados no conteúdo dos alertas e interesse dos recipientes.

Theilmann (2010) discute que cada rede e administrador têm diferentes características, como requisitos de proteção de dados e privacidade, tráfego, armazenamento e monitoramento. Theilmann defende a criação de federações de EWS para promover a cooperação e propõe o Herold, uma arquitetura para EWS baseada em agentes independentes que cooperam baseados no interesse individual e no contrato de cooperação firmado com outros agentes.

Há ainda outros projetos que abordam a detecção proativa de incidentes. As tabelas A.1 e A.2 (ver apêndice A) apontam alguns desses projetos e suas características. Informações adicionais sobre projetos de detecção proativa podem ser encontradas em (Gorzelak *et al.*, 2011).

### 2.5.2 Trabalhos similares

Os trabalhos de (Joshi *et al.*, 2013; More *et al.*, 2012; Mulwad *et al.*, 2011; Rodrigues, 2012) exploram a Web como fonte de dados para a detecção de ameaças à segurança. Eles afirmam que por meio da mineração e agregação de informações na Web é viável prover alertas antecipados de novas vulnerabilidades e ataques, bem como monitorar e estimar a evolução e distribuição das ameaças à segurança existentes. (Benjamin *et al.*, 2015) exploram fóruns, rede IRC e sítios de comercialização de número de cartões para identificar potenciais ameaças e reagir antecipadamente. (Ritter *et al.*, 2015) exploram o Twitter como fonte de alertas de segurança e destaca a importância de explorar essa fonte para identificação de eventos de segurança. Esses trabalhos estão relacionados diretamente com a natureza da proposta dessa tese.

Mulwad *et al.* (2011) desenvolveram um arcabouço para extrair da Web informações sobre vulnerabilidades e ameaças composto por três componentes: um classificador SVM, um sistema de extração de informação e um sistema para representação do conhecimento. O classificador recebe informações de fontes textuais na Web e seleciona textos relacionados a ameaças à segurança. O sistema de extração identifica textos relevantes baseado no conhecimento da Wikitology[11] e uma ontologia de detecção de intrusão. Os conceitos extraídos são armazenados em uma base de conhecimento usando a ontologia IDS OWL[12]. A avaliação do protótipo foi realizada usando descrições textuais da base de dados NVD[13]. Uma contribuição do trabalho é a viabilidade de extrair conceitos de segurança de textos para processamento automatizado. Como limitações destacam-se: (i) o uso de uma única base para a avaliação e ainda uma base composta exclusivamente por descrições de ameaças à segurança; (ii) o classificador SVM utiliza um conjunto de treino limitado (80 positivos e 75 negativos).

More *et al.* (2012) propõem um modelo de detecção de intrusão baseado na integração de sistemas tradicionais com outras fontes de dados em conjunto com um módulo de decisão auxiliado

[11] http://ebiquity.umbc.edu/project/html/id/83/Wikitology
[12] http://ebiquity.umbc.edu/ontologies/cybersecurity/ids/
[13] http://nvd.nist.gov/

por uma ontologia e uma base de conhecimentos. O modelo visa identificar ameaças que não possuem assinaturas. A arquitetura proposta utiliza monitoramento de máquinas e redes, sensores de hardware e de fontes Web (estruturadas e não estruturadas). Utiliza uma ontologia composta por três classes: formas de ataque, consequências do ataque e informações dos alvos. As entidades extraídas são classificadas em uma das classes considerando as propriedades da classe e o significado da entidade. A base de conhecimento é constituída de triplas (sujeito, predicado, objeto). As triplas são usadas pelo módulo de decisão para indicar um possível ataque a partir das informações coletadas dos sensores de rede. A validação do modelo ocorreu por meio da simulação de um ataque ao Adobe Acrobat em ambiente controlado. O módulo de decisão encontrou, no registro de sistema de uma das máquinas, o serviço com a vulnerabilidade sendo acessado pelo Adobe Acrobat. Nas informações extraídas da Web havia indicação de que deveria haver um acesso remoto, o qual foi identificado pelo módulo via informações do IDS, logo foi gerado um alerta. A principal contribuição é mostrar que informações coletadas na Web auxiliam na detecção de ataques. Como limitações, destacam-se: (i) um único caso de teste; (ii) considera na validação somente um blog e o CVE como fontes Web, logo não explora outras mais complexas; (iii) a base e a ontologia não são atualizadas automaticamente.

Joshi *et al.* (2013) apresentam um arcabouço automático para extrair, relacionar e publicar conceitos, entidades e relações associados a ameaças à segurança. Os conceitos são extraídos usando uma ontologia *Web Ontology Language* (OWL) e são associados ao NVD por meio de ligações *Resource Description Framework* (RDF). A arquitetura proposta contém três principais componentes: reconhecedor de entidades e conceitos, gerador RDF e gerador de ligações. O reconhecedor de conceitos e entidades usa o algoritmo *Conditional Random Field* (CRF) para identificar os conceitos. Sete classes foram definidas considerando blogs, boletins de segurança e descrições do CVE: software, termos de rede, ataque, nome de arquivo, hardware, informação de versão, outros termos técnicos. O conjunto de treinamento para o arcabouço de extração foi constituído por 30 blogs, 240 descrições do CVE e 80 boletins da Microsoft e Adobe. Joshi et al. realizaram validação do classificador por meio de validação cruzada com cinco divisões de partes de mesmo tamanho. A avaliação usou as medidas de precisão, abrangência e escore F1. As classes de nome de arquivo, sistema operacional e software obtiveram valores altos para todas as medidas. Por outro lado, classes de ataque, termos de rede e outros termos técnicos obtiveram valores baixos para todas as medidas. Como contribuições, destacam-se a extração automática de termos e a ligação de informações de segurança por meio de RDF. Como limitações, (i) usa um conjunto pequeno e específico de dados para realizar a validação; (ii) muitas classes apresentam valores baixos para as medidas de precisão e abrangência.

Rodrigues (2012) desenvolveu uma ferramenta que recupera informações de segurança a partir de bases de vulnerabilidades e serviços de alertas de segurança na Web. A ferramenta possibilita realizar consultas especializadas no conteúdo. A estratégia de coleta é baseada em um Web *crawler* que acessa as páginas e indexa localmente o conteúdo. Foram aplicadas diferentes heurísticas e técnicas de mineração de dados para transformar o conteúdo não formatado e incompleto das páginas, em informações estruturadas representadas pelas principais informações da página (denominado pelo autor de *templates*). Os padrões descritos nos *templates* propiciaram a extração apenas das informações de valor como notificações de segurança. Os conteúdos podem ser consultados usando notações específicas para obter informações sobre uma determinada vulnerabilidade, serviço ou alerta. Como contribuição, o trabalho concentra as notificações de segurança mais recentes divulgadas por organizações especializadas em uma única base. Por outro lado, limita-se a realizar a busca em um escopo específico, não contemplando outras fontes de dados.

Benjamin *et al.* (2015) realizam a análise de comunidades hackers em fóruns, canais IRC e sítios Web de comercialização de número de cartões de crédito com o intuito de identificar potenciais ameaças. Propõem um arcabouço baseado em técnicas de Recuperação de Informação para automatizar o processo de identificar ameaças nessas comunidades. A Figura 2.12 apresenta o arcabouço proposto para a coleta e identificação de ameaças em fóruns.

O processo descrito no arcabouço faz uso de busca e ponderação de palavras-chave para iden-

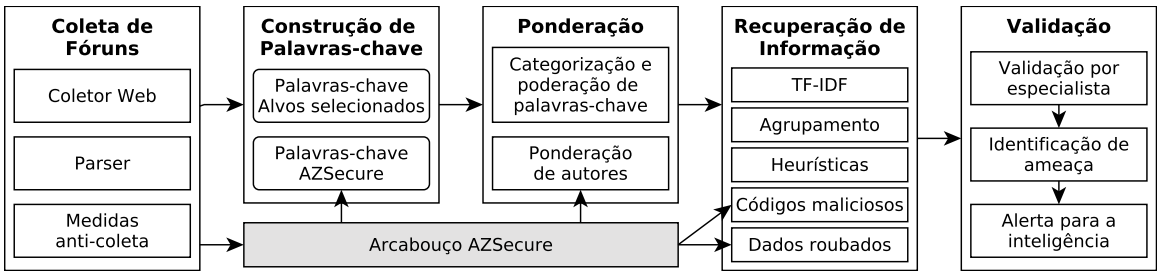


**Figura 2.12:** *Arcabouço para análise de fóruns hackers (Adaptado de Benjamin et al. (2015)).*

tificar as potenciais ameaças nas comunidades *hackers*. Também faz uso do arcabouço AZSecure (Li e Chen, 2014), um arcabouço que analisa indivíduos envolvidos com vendas de códigos maliciosos e fraudes bancária. Como resultado, foram identificadas postagens que contém informações úteis sobre ameaças cibernéticas. Já para as outros tipos de comunidade, foram adaptadas outras abordagens, pois exigem mecanismos específicos da fonte de dados para a coleta e a análise. Em todas as fontes foram identificadas evidências de ameaças cibernéticas. Como pontos positivos, apresentam uma abordagem para a análise de comunidades *hackers* e comprovam a importância de monitorar e identificar potenciais ameaças nessas comunidades. Como pontos negativos, não realizam avaliação do número de falsos positivos e eficácia das abordagens propostas.

Ritter *et al.* (2015) desenvolveram uma abordagem fracamente supervisionada para classificar eventos no Twitter. Realizaram o estudo em três tipos de eventos de segurança de computadores: ataques DDoS, vazamentos de dados e sequestros de contas. Os eventos foram coletados pelo monitoramento das respectivas palavras-chave: *DDoS*, *breach*, *hijacking*. Afirmam que abordagem proposta requer somente de 10 a 20 exemplos etiquetados de treinamento para a detecção de novos eventos em tempo real. Os eventos são representados pela entidade e a data de ocorrência. Dessa forma, novos eventos que se referem a mesma entidade e data são localizados e aumentam o número de instâncias de treinamento. Como pontos positivos, apresentaram uma abordagem para a coleta de novos eventos relacionados a um acontecimento e um método para gerar bases de treinamento. Também demonstraram que há um grande número de eventos associados à segurança no Twitter. Como pontos negativos, a abordagem apresentou um número alto de falsos positivos e a pesquisa limitou-se a explorar um conjunto de apenas três palavras-chave.

### 2.5.3 Síntese e discussões

Essa seção discute os trabalhos relacionados e também apresenta a comparação com a pesquisa realizada nesta tese.

Os projetos CarmentiS (Grobauer *et al.*, 2006), AmSel (Apel *et al.*, 2009, 2010), DShield (DShield, 2014) e Arakis (CERT Polska, 2014) apresentam arquiteturas centralizadas em órgãos governamentais. Nós propomos um arcabouço que possibilita o desenvolvimento de arquiteturas centralizadas ou descentralizadas de um órgão. Compartilhamos a ideia da especificação de uma política compartilhamento assim como em Herold (Theilmann, 2010), CarmentiS e Semantic Room (Lodi *et al.*, 2014). As questões de privacidade devem estar definidas nessas políticas. O arcabouço proposto não especifica um padrão para garantir a privacidade, mas especifica um componente para possibilitar a implantação de um modelo entre as partes. O projeto Worminator (Boggs *et al.*, 2011; Locasto *et al.*, 2005), ao contrário, utiliza filtros de Bloom para garantir a privacidade de compartilhamento.

Assim como em Semantic Room, há a liberdade de uso de diferentes tecnologias e modelos de desenvolvimento para os componentes descritos no arcabouço. O arcabouço não define tipos de sensores para a implementação em uma arquitetura distribuída, mas define o fluxo de processamento e apresenta atributos importantes a serem identificados nas fontes e dados coletados. A proposta

do arcabouço foi descrever uma abordagem que possibilite coletar e processar alertas de diferentes tipos de fontes de dados não estruturados. Por outro lado, há projetos que fazem o uso de um único tipo de fonte, como o uso exclusivo *honeypots* (Apel *et al.*, 2009; Bastke *et al.*, 2010) ou registros de *firewall* (DShield, 2014).

Os projetos DOMINO (Yegneswaran *et al.*, 2004) e Semantic Room (Lodi *et al.*, 2014) apresentam arquiteturas hierárquicas que possibilitam pré-processamento e a criação de federações. A divisão dos componentes e os fluxos definidos em nossa proposta de arcabouço propiciam a implementação de componentes de pré-processamento distribuídos e também próximos as origens de dados.

A coleta de informações de fontes abertas foi inspirada por OSINF (Dorges e Sander, 2010) que também faz o uso de mecanismos de recuperação de informação para extrair notificações de segurança de fontes abertas. Nossa proposta apresenta como inovação a exploração de mídias sociais como fonte de alertas antecipados e o uso de sistemas de recomendação para incluir o auxílio de especialistas de segurança. Vários dos projetos (CarmentiS, OSINF, DeepSight (Symantec, 2014), DShield, entre outros) que propõem detecção antecipadas de incidentes, há a participação obrigatória de especialistas no fluxo de processamento de alertas.

Quanto as propostas de arcabouços para extração de informações de segurança na Web, Joshi *et al.* (2013); Mulwad *et al.* (2011) usam bases de dados que armazenam apenas notificações de segurança, enquanto nosso arcabouço propõe uma abordagem para obter informações de qualquer tipo de fonte de dados não estruturados e contempla pré-processamento para filtrar conteúdo indesejado. Já (Benjamin *et al.*, 2015) propõem um arcabouço para a análise e extração de eventos de segurança, mas não possui uma abordagem que viabilize a adaptação para qualquer tipo de fonte. No entanto, faz uso de técnicas de Recuperação de Informação da mesma forma que é especificado em nosso arcabouço. Rodrigues (2012) também mostrou como é possível usar Recuperação de Informação para construir uma base de consulta para notificações de segurança relevantes publicadas em sítios Web.

Os trabalhos de Joshi *et al.* (2013); More *et al.* (2012) fazem o uso de ontologia para extrair informações de interesse de alertas de segurança. Nosso arcabouço não define uma forma de usar ontologia, mas especifica um componente que viabiliza a integração dessas abordagens. More *et al.* (2012) também apresentam uma abordagem para comparar alertas de fontes tradicionais com os alertas extraídos de fontes na Web. O nosso arcabouço está direcionado principalmente a extração de alertas de dados não estruturados, mas também oferece uma forma de agregar mecanismos de comparação com outras fontes, como as tradicionais. Ritter *et al.* (2015) apresentam uma abordagem para classificação de eventos de segurança no Twitter e constata a importância de monitorar esses eventos. Essa constatação apenas ratifica a motivação do desenvolvimento do arcabouço proposto, pois já havíamos comprovado a importância das mídias sociais para geração de alertas antecipados de ameaças cibernéticas.

## 2.6 Considerações finais

Este capítulo apresentou os conceitos bases e trabalhos relacionados a pesquisa realizada nesta tese. Verificou-se a evolução das pesquisas na área de Detecção de Intrusão, que partiu de sistemas localizados apenas em máquinas ou em pontos específicos de tráfego na rede, para sistemas distribuídos, colaborativos e que visam a detecção antecipada de potenciais ameaças. Também foi abordado os conceitos associados a EWS e como podem ser usados na detecção e disseminação de alertas sobre diferentes tipos de ameaças. Verificou-se que mesmo que um ataque tenha obtido sucesso em comprometer uma infraestrutura, o quanto antes essa informação for disseminada, mais rápido podem ser tomadas as medidas preventivas em outras localizações. Foram discutidas características e desafios para a implementação de EWS. Essas questões foram discutidas sob a perspectiva de cinco processos comuns a EWS: (i) coleta de informações, (ii) gerenciamento de

alertas, (iii) correlação de alertas, (iv) detecção e predição de ameaças e (v) resposta a incidentes e disseminação de alertas. Também foram discutidos dois pontos importantes e desafiadores de serem estabelecidos: (i) compartilhamento de informações e (ii) privacidade e confidencialidade. Por fim, para finalizar a parte de fundamentação, foram apresentados os conceitos associados à mineração de informações de dados não estruturados. Foram destacados a importância e os conceitos de classificadores, recomendadores, processamento de linguagem natural e recuperação de informação. A análise da características e desafios, associados com a implementação de técnicas para a exploração de grandes conjuntos de dados textuais, serviram de alicerce para a criação do arcabouço para a análise e extração de alertas cibernéticos de fontes de dados não estruturados que é apresentado no próximo capítulo. Quanto a revisão da literatura, foram identificados os principais trabalhos que motivaram a pesquisa, além da discussão de pontos similares e distintos que se relacionam com a proposta desta tese.

# Capítulo 3

# Arcabouço para Análise e Extração de Alertas em Fontes de Dados Não Estruturados

Este capítulo apresenta um arcabouço para a análise e extração de alertas de cibersegurança em fontes de dados não estruturados, em especial, fontes que disponibilizam dados publicamente. O arcabouço visa extrair alertas associados à cibersegurança, principalmente alertas antecipados. Deste capítulo em diante, o arcabouço é denominado arcabouço EWS ou simplesmente arcabouço, e as fontes de dados não estruturados são denominadas apenas fontes de dados.

No arcabouço, são discutidas e categorizadas as principais fontes de dados relevantes para a construção de um EWS, como microblogs, redes sociais, listas de e-mails, páginas Web, bases de vulnerabilidades, entre outras. São apresentados e detalhados os componentes, processos, conceitos e entidades do arcabouço proposto. Também são descritos mecanismos para a extração de alertas, preferencialmente antecipados, em cada componente do arcabouço. Por fim, apresenta-se como o arcabouço pode ser implantado em uma arquitetura distribuída e como ele propicia a colaboração entre administradores de redes e/ou especialistas em segurança por meio do compartilhamento de informações e recomendação de alertas de segurança.

## 3.1 Visão geral do arcabouço EWS

O arcabouço EWS é um arcabouço voltado para identificação, extração e notificação de alertas de cibersegurança a partir do processamento de dados não estruturados. Ele foi elaborado a partir da investigação de mecanismos para a identificação e extração de alertas de cibersegurança em diferentes fontes de dados, em especial, microblogs, redes sociais online, redes IRC, blogs e boletins de notificações de segurança. O arcabouço tem como entrada dados não estruturados e como saída potenciais alertas de segurança. É constituído por quatro componentes principais, por um processo de análise de dados, por entidades que interagem com os componentes do arcabouço e por bases contendo as especificações de coleta, as políticas de privacidade, os dados coletados nas fontes e as informações de inteligência e alertas produzidos pelos componentes do arcabouço.

A Figura 3.1 apresenta uma visão geral dos elementos do arcabouço, que são:

- Fontes de Dados: proveem informações de interesse para um EWS voltado à cibersegurança, no caso, dados não estruturados que podem conter informações sobre atividades hackers, ameaças à segurança de redes ou sistemas computacionais, orquestrações de ataques, vazamento de dados, novas vulnerabilidades ou ataques, atualizações de software e usuários suspeitos de ações cibernéticas que comprometam a segurança.

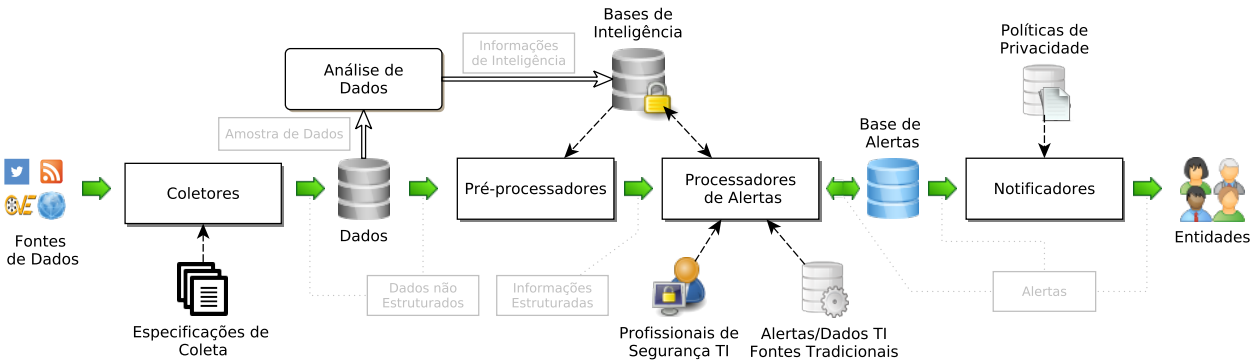


**Figura 3.1:** *Arcabouço EWS - Uma visão geral*

- Coletores: realizam o monitoramento das fontes, coletam e armazenam os dados em bases de dados segundo as especificações de coleta.
- Especificações de Coleta: definem parâmetros de operação e expressões de consultas a serem usadas pelos coletores.
- Análise de Dados: processa amostras de dados coletados para a geração de informações de inteligência a serem usadas pelos componentes de pré-processamento e processamento de alertas. A análise de dados é um processo estático que deve ser realizado periodicamente e preferencialmente antes da adição de novas fontes.
- Informações de Inteligência: são características identificadas nas fontes e nos dados, por exemplo, vocabulário, tipos de alertas, requisitos e limitações de monitoramento, entre outros, que auxiliam no desenvolvimento de mecanismos para a extração de alertas.
- Amostras de Dados: são conjuntos de dados selecionados para o processo de análise de dados que visa identificar novas informações de inteligência.
- Pré-processadores: processam os dados não estruturados visando remover informações duplicadas e irrelevantes, expandir informações e gerar informações estruturadas segundo um padrão.
- Informações Estruturadas: são informações organizadas em uma estrutura padrão que foram pré-processadas e podem conter potenciais informações de interesse para serem identificadas como alertas.
- Processadores de Alertas: classificam as informações estruturadas em alertas ou não. São auxiliados por bases de inteligência, profissionais de segurança TI e informações de fontes tradicionais. Também produzem informações de inteligência usadas na atualização das bases de inteligência.
- Profissionais de Segurança TI: são responsáveis pela supervisão do processamento de alertas e pelo compartilhamento de novas notificações de segurança.
- Fontes Tradicionais: são alertas de sensores de redes tradicionais (p. ex. IDS, honeypots, outros) e informações de infraestrutura (p. ex. software e hardware) usadas para identificar alertas importantes por meio de correlação.
- Alertas: são informações estruturadas que contém como atributo principal a informação coletada e identificação da fonte, além de outros atributos que podem auxiliar na mitigação ou reação proativa a ataques e ameaças à segurança.
- Base de Alertas: armazena os alertas identificados pelos processadores de alertas.

- Notificadores: propagam os alertas para as entidades de interesse segundo as políticas de privacidade.
- Políticas de privacidade: especificam as regras e restrições de notificação considerando os atributos dos alertas, restrições das fontes ou das entidades de interesse.
- Entidades: são os recursos humanos e sistemas que são notificados dos alertas extraídos pelo arcabouço.

O fluxo principal do arcabouço é iniciado pelos coletores que em tempo real coletam dados não estruturados das fontes segundo as especificações de coleta para cada fonte. Esses dados são processados pelos pré-processadores que filtram, expandem e normalizam os dados para uma estrutura padronizada. Os pré-processadores usam as informações de inteligência e encaminham informações estruturadas para os processadores de alertas. Por sua vez, os processadores processam essas informações com auxílio de informações de inteligência, alertas externos e interação de especialistas de segurança, identificando os potenciais alertas. Os alertas são encaminhados para os notificadores que notificam as entidades de interesse respeitando as definições das políticas de privacidade.

A Figura 3.2 apresenta a visão detalhada dos elementos do arcabouço. Nessa visão, destacam-se os componentes e suas especializações. As seções seguintes apresentam e discutem individualmente os elementos que compõem o arcabouço.

## 3.2 Fontes de dados não estruturados

As fontes de dados abordadas nesta tese são as que representam e disponibilizam seus conteúdos na Web ou em infraestruturas de redes fechadas sem uma estrutura formalmente definida, comumente na forma de texto. Redes sociais, microblogs, blogs, bases de vulnerabilidades, listas de correspondência eletrônica (e-mail), sítios especializados, motores de busca, serviços de compartilhamento de texto, fóruns Web são exemplos de fontes de dados não estruturados. Nesta seção, são descritas e categorizadas as principais fontes de dados de interesse para EWS.

No contexto do arcabouço, foram identificadas 13 categorias de fontes de dados. A seguir são apresentadas essas categorias, exemplos de fontes e como elas se relacionam com alertas antecipados. As considerações sobre cada categoria foram decorrentes da análise especialista de dados resultados do monitoramento de pelo menos um exemplo de fonte.

- Redes sociais: são estruturas sociais compostas geralmente por pessoas ou organizações conectadas entre si que compartilham relações ou interesses em comum. São exemplos de redes sociais: Facebook, Google+, LinkedIn e o MySpace. Proporcionam a obtenção de informações sobre alertas de forma rápida devido à estrutura das redes e à disseminação rápida de conteúdo. Nessas redes é possível monitorar perfis e publicações que contenham conteúdo associado a incidentes, ameaças ou notificações de segurança. No entanto, muitas comunidades são fechadas, o que inviabiliza o monitoramento e acesso ao conteúdo trafegado.
- Microblogs: são serviços para publicação de elementos curtos na Web, tais como mensagens curtas, fotos e *links*. São exemplos de microblogs: Twitter, Sina Weibo e Tumblr. Proporcionam a obtenção de informações sobre alertas de forma rápida devido à abertura para o estabelecimento de relações e à facilidade para a disseminação rápida de imagens e textos curtos. Por outro lado, o texto curto e a informalidade das mensagens dificultam a identificação de conteúdos relevantes como alertas.
- Chats: são serviços para conversações em tempo real, tanto em grupos como privadas. São exemplos de serviços de chat a rede IRC, sítios Web de conversações e mensageiros instantâneos. São exemplos de serviços de chat: rede IRC Anonops, Skype e o WhatsApp. Proporcionam a obtenção de alertas a partir do conteúdo publicado nas comunidades (segurança ou

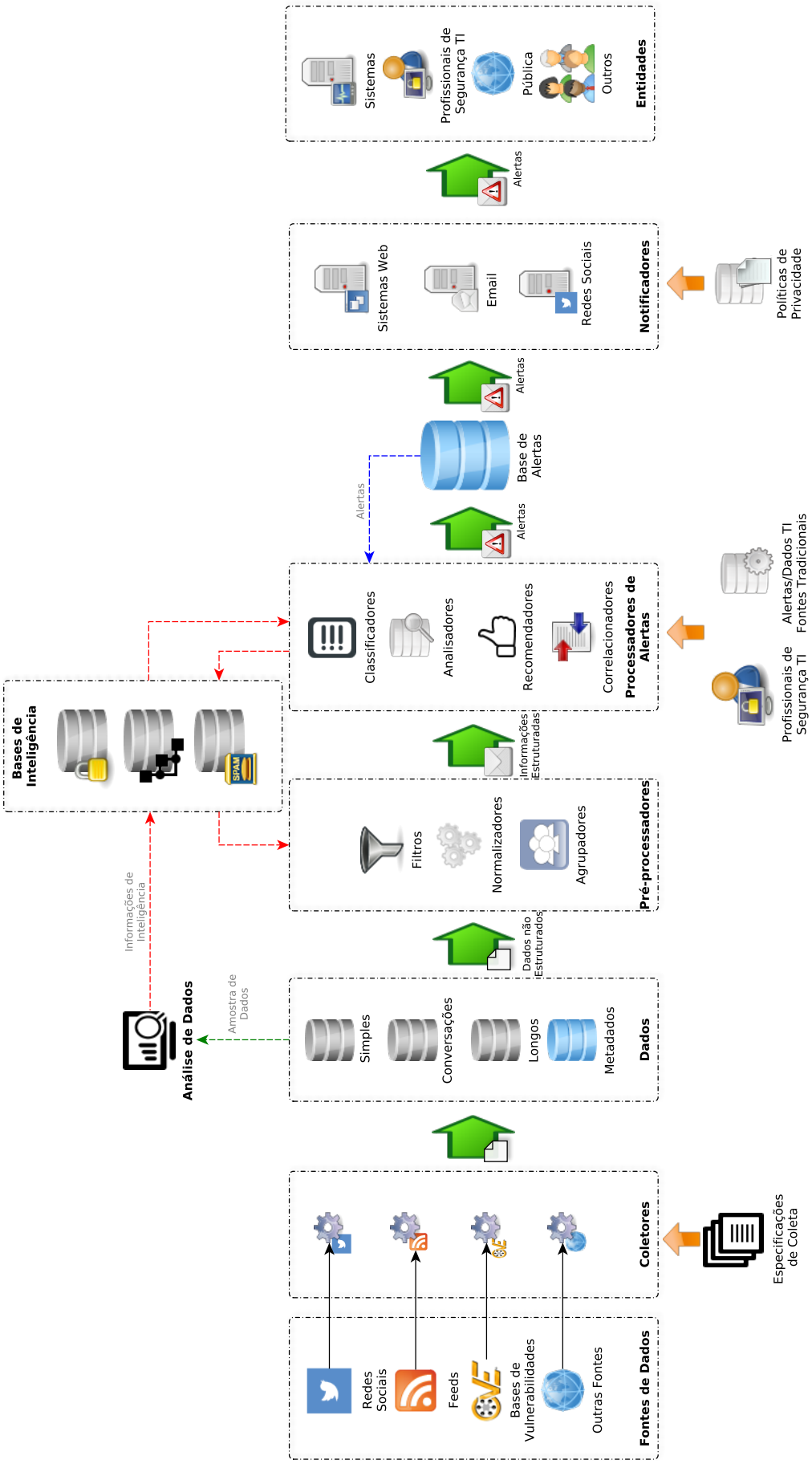


**Figura 3.2:** *Arcabouço EWS*

hackers) ou monitoramento de usuários suspeitos. As redes de IRCs, em especial, são interessantes devido ao uso por comunidades especializadas em segurança e por comunidades que trocam informações de atividades ilícitas (Décary-Hétu e Dupont, 2012). As principais limitações estão associadas às restrições de monitoramento das interações privadas e de grupos fechados ou com poucos usuários.

- Sítios especializados: são páginas na Web que divulgam conteúdo associado à segurança de redes e de computadores. São exemplos de sítios especializados: OWASP, US-CERT, Symantec, entre outros. Também é comum grandes organizações manterem sítios para divulgarem boletins de segurança sobre atualizações e correções de vulnerabilidades de seus aplicativos. Esses sítios especializados são interessantes devido a confiabilidade da informação publicada. No entanto, os alertas não são publicados assim que descobertos, somente após a confirmação e a disponibilização da correção do problema alertado.

- Blogs: são páginas na Web, com carácter mais informal, que possibilitam a postagem de publicações por seus usuários. Há blogs de segurança apoiados por organizações e blogs de segurança de entusiastas ou especialistas em segurança. São exemplos de blogs organizacionais de segurança: Cisco, Microsoft, Snort, Mcafee. Esses blogs tem como vantagem a confiabilidade da informação publicada. São exemplos de blogs pessoais ou abertos para postagens sobre segurança: KrebsonSecurity e Threat Post. Esses blogs tem como vantagem a independência para a realização de publicação dos usuários. A limitação em ambos é a dificuldade de distinguir alertas de notícias de segurança.

- Notificações de publicação (*feeds*): são serviços para receber notificações de conteúdo publicados ou atualizados em sítios Web e blogs. Proporcionam acesso à informação assim que publicadas na Web. Os formatos mais comum são os RSS (RSS Advisory Board, 2016) e Atom (Nottingham e Sayre, 2005). Devem ser considerados como fontes importantes de alertas antecipados, pois é comprovado que muitas vulnerabilidades são exploradas até aproximadamente um ano após sua divulgação pública. Os pontos positivos são o acesso a notícia assim que publicada e o acesso é aberto para muitos sítios. Os pontos negativos são a propagação lenta de notificações devido à dependência da assinatura e o formato semiestruturado das notificações.

- Bases de vulnerabilidades: são serviços para documentação e divulgação de vulnerabilidades encontradas em sistemas computacionais. Destacam-se nesse grupo o CVE e o NVD, que apresentam um catálogo padronizado de vulnerabilidades e são endossados pela indústria como padrões para identificação e descrição de vulnerabilidades e ameaças à segurança. Outras fontes são bases oriundas de projetos como DShield[1], MalwareBlacklist.com[2], entre outros. Proporcionam a obtenção de informações detalhadas e comprovadas sobre vulnerabilidades e ameaças à segurança. Os pontos positivos são a confiabilidade das informações e o acesso é aberto para diversas bases. Os pontos negativos são o tempo de atualização das bases e a dependência de assinatura de notificações dessas fontes. Em geral, as bases possuem formatos estruturados para descrever as informações, mas cada base usa um formato próprio.

- Listas de correspondência eletrônica: são formas de organizar grupos com interesses comuns para possibilitar a comunicação por difusão via o serviço de correspondência eletrônica. São exemplos de interesse para um EWS as listas de discussão de segurança e vulnerabilidades de software, por exemplo, a lista de e-mail da comunidade de aplicativos Open Source[3] e listas específicas de produtos como o núcleo e distribuições Linux. Os pontos positivos são o detalhamento das notificações e propagação de potenciais vulnerabilidades que podem ser exploradas. Os pontos negativos são o formato das notificações e a identificação de correspondências relevantes.

[1] https://www.dshield.org/
[2] http://malwareblacklist.com
[3] http://oss-security.openwall.org/wiki/mailing-lists/oss-security

- Fóruns Web: são grupos de discussão na Web para postagem de mensagens entre os participantes. É comum possuírem uma estrutura de tópicos e subtópicos dos assuntos discutidos e manterem o histórico de postagens antigas. Os fóruns hackers são interessantes para monitoramento em um EWS, por exemplo, Hack Forums, anonymous BR, Caveira Tech. Os pontos positivos são a possibilidade de identificar orquestrações de ataques e alvos. O principal ponto negativo é a limitação de monitoramento devido a necessidade de autenticação de muitos fóruns.

- Motores de busca: são serviços que realizam busca na Web por meio de expressões de busca e palavras-chave. Os motores de busca podem ser usados para encontrar vulnerabilidades em sistemas Web, atividade denominada de Google Hacking quando usado o motor de busca Web da Google. Há bases de dados para compartilhar sentenças de busca, como é o caso da Google Hacking Database[4] (GHDB). Como ponto positivo tem-se a identificação de potenciais alvos de ataques. Como ponto negativo tem-se que nem sempre as expressões de busca produzem o resultado esperado.

- Serviços de compartilhamento de texto: são serviços para o compartilhamento de texto puro ou com formatação simples. O compartilhamento é realizado comumente pela divulgação de uma URL da postagem. São exemplos de serviços interessantes para um EWS: PasteBin, Slexy, Pastie, Codepad, Github Gist, entre outros. Os pontos positivos são a identificação de vazamentos de informações, alvos de ataque e compartilhamento de códigos de exploração. O principal ponto negativo é diferenciar automaticamente os conteúdos relevantes dos irrelevantes, por exemplo, um código fonte malicioso de um código de programação normal.

- Serviços de compartilhamento de vídeo: são serviços para o compartilhamento e divulgação de vídeos. Podem ser usados para compartilhar técnicas de invasão ou promover uma invasão realizada com sucesso. São exemplos os serviços: YouTube, Vimeo, entre outros. Apesar de terem como vantagem o detalhamento de técnicas de invasão, esses serviços possuem como desvantagem a dificuldade de monitoramento e principalmente alto número de falsos positivos. São de interesse para um EWS quando é detectado um vídeo compartilhado por um grupo ou usuário suspeito de atividades ilícitas.

- Deep Web: são sistemas interconectados cujas informações não podem ser acessadas por motores de busca devido não serem indexadas, mas que podem ser acessadas diretamente por quem disponibilizou ou um grupo de indivíduos com acesso privilegiado. Como costuma ser usada para realização de atividades ilícitas, torna-se uma fonte que pode conter informações para um EWS. No entanto, por não possuir indexação e formatos padrões, acaba se tornando uma fonte de difícil monitoramento.

Mesmo fontes em uma mesma categoria diferenciam entre si devido a questões de formato, forma de monitoramento e restrições de acesso a informações. Por exemplo, a fonte de dados Facebook e Google+ estão na categoria redes sociais, no entanto, os atributos e a política de acesso aos dados dessas fontes são diferentes. É comum fontes terem políticas que restringem o monitoramento, limitando assim o volume e os dados que podem ser acessados. Portanto, o conhecimento das fontes é importante para decisões e implementações do monitoramento realizado no arcabouço pelos componentes Coletores (ver seção 3.3).

## 3.3 Coletores

Os coletores são responsáveis por monitorar e capturar o conteúdo das fontes de dados não estruturados. Os coletores atuam em fontes que transitam conteúdos associados à segurança da

[4] https://www.exploit-db.com/google-hacking-database/

informação de TI e em fontes que transitam conteúdos associados a atividades suspeitas. Conteúdos de segurança estão geralmente associados à divulgação de vulnerabilidades, atualizações de software, vetores de ataques, novas ameaças dia zero. Conteúdos de atividades suspeitas estão geralmente associados a orquestrações de ataques, desfiguração de páginas, vazamento de informações, divulgação de alvos, de ferramentas e de métodos de ataque.

A forma de implementação dos coletores é dependente das fontes, isto é, deve-se considerar os recursos e restrições de monitoramento inerentes ou ofertados por cada fonte. Há fontes que disponibilizam bibliotecas ou serviços para a realização do monitoramento. Por exemplo, o Twitter e o Facebook disponibilizam uma API Restful para o monitoramento. O Twitter ainda oferece um serviço pago de monitoramento que garante disponibilidade e acesso a todas as mensagens postadas segundo o critério de busca. Por outro lado, há fontes que necessitam que sejam desenvolvidos códigos para coletar, realizar o processamento e armazenar as informações. Por exemplo, monitorar e recuperar informações de sítios Web é usado um Web Crawler. Há fontes ainda mais restritas, que necessitam de autorização para ter acesso aos dados ou participação de um usuário como um cliente da fonte. Por exemplo, nos canais IRC e grupos do Facebook. Há ainda as fontes que tornam inviável o acesso, como é o caso do Skype, que os dados podem ser acessados somente pelos participantes de uma sessão ou pelos provedores de serviço.

A operação dos coletores é definida pelas especificações de coleta. Podem ser especificados parâmetros como periodicidade de coleta, quantidade máxima de dados, forma de operação, entre outros. No entanto, as principais especificações de coleta são as expressões para acessar o conteúdo de interesse.

No arcabouço, são definidas os seguintes grupos para a construção de expressões de monitoramento:

- perfis: especifica o monitoramento de perfis de autores suspeitos ou de autores que divulgam notícias importantes de segurança. Os perfis podem ser identificados por endereços específicos (páginas, fóruns, canais IRC, serviços), em campos de autoria ou em referências no conteúdo textual da mensagem.
- termos de cibersegurança: especifica o monitoramento de termos usados em cibersegurança.
- atributos: especifica o monitoramento de atributos definidos nas fontes ou que podem ser identificados. Por exemplo, monitoramento de idioma, de mensagens longas, de mensagens criptografadas, de mensagens com imagem, entre outros.
- ferramentas: especifica o monitoramento de ferramentas usadas para a realização de atividades hackers.
- infraestrutura: especifica o monitoramento de software e hardware de infraestruturas de TI.
- entidades: especifica o monitoramento de entidades, por exemplo, organizações, pessoas, IP, URL, entre outros.
- novos termos: especifica o monitoramento de novos termos que são usados temporariamente. Enquadram-se nesse grupo as palavras do momento que descrevem um ataque específico (p. ex. HeartBleed), identificadores de atividades suspeitas (p. ex. #ddosenem).

Nas expressões de busca usadas no monitoramento, pode-se realizar combinações com os elementos de um mesmo grupo ou outros. Atributos disponibilizados pelas fontes também podem ser usados para a realização de consultas mais especializadas, por exemplo, localização geográfica e data de postagem. Nas fontes que são categorizadas em conversações, recomenda-se capturar todo o conteúdo para contextualizar os potenciais alertas.

Os coletores têm ainda que lidar com as limitações das fontes e artifícios dos clientes. Alguns dos problemas são a detecção de monitoramento, geração de rumores falsos, uso de truques para ocultar a informação (imagens, vídeos, *links*, linguagem alternativa, criptografia), criptografia, canais privados, discussões em grupos privados. Algumas dessas questões, como a linguagem alternativa e *links* podem ser tratadas com o estudo da fonte e desenvolvimento de métodos para explorar características do vocabulário e dos destinos apontados pelos *links*. Outras, no entanto, dependem da infiltração de usuários em grupos privados ou da intervenção humana para entender o significado do conteúdo.

## 3.4 Dados

Os dados são os dados coletados pelos coletores durante o monitoramento das fontes. Nesta fase do arcabouço ainda são dados brutos, contendo alertas de cibersegurança, antecipados ou não, em conjunto com informações não relevantes como alertas.

No arcabouço, os dados são classificados considerando a forma da postagem da fonte, que são: simples, conversações e longas. A Figura 3.3 apresenta o agrupamento das fontes nessas 3 classificações.


**Figura 3.3:** *Classificação das fontes segundo a forma de postagem.*

As postagens simples são caracterizadas por apresentarem uma postagem geralmente curta, independente e informativa. As postagens em conversações são postagens curtas, mas que dependem do conhecimento do contexto, no caso, mensagens anteriores e posteriores, assunto discutido, autores da postagem. As postagens longas são caracterizadas por textos longos que detalham o assunto publicado. Algumas fontes disponibilizam ainda metadados e/ou atributos que favorecem o pré-processamento e processamento de alertas.

Na Figura 3.3, observa-se que algumas fontes estão classificadas em mais de uma categoria, como é o caso de fóruns Web e listas de e-mail. Os fóruns Web e listas de e-mail são, em geral, postagens auto explicativas, mas há situações que é necessário extrair o contexto da publicação e, para tal, é necessário ter acesso a outras informações de contexto, principalmente mensagens anteriores e posteriores. As postagens simples, especialmente as que agregam metadados (redes sociais, microblogs e notificações de publicações), ampliam as possibilidades de mecanismos para a extração de conteúdo relevante devido a agregação de informações e por conter o assunto em postagens individuais. As postagens em conversações, especialmente em chats, dependem do acesso ao conteúdo publicado e identificação do contexto. As postagens longas são difíceis de serem processadas quando são apenas texto (blogs, sítios especializados), mas quando possuem um formato semiestruturado (bases de vulnerabilidades) acabam por propiciar a identificação de atributos relevantes para uma notificação de segurança.

Independente da forma de postagem, o atributo principal para a identificação de potenciais alertas é o texto. O texto é o atributo que contém a descrição de um possível alerta. Nas postagens

simples, o texto é a mensagem publicada nas redes sociais, microblogs ou fóruns, ou ainda, a descrição ou título das notificações de publicações ou serviços de compartilhamento de vídeo. Nas postagens em conversações, o texto é o conjunto de mensagens enviadas a um outro indivíduo ou a um grupo. Nas postagens longas, o texto é o corpo principal das páginas em sítios especializados, blogs, algumas bases de vulnerabilidades, ou o título e conteúdo de serviços de compartilhamento de texto ou listas de e-mail, ou ainda, o resultados de motores de busca e pesquisas na Deep Web.

Em relação aos outros atributos/metadados que são de interesse para um EWS, destacam-se:

- autor: identifica o autor dos dados coletados e é usado na credibilidade da informação, investigação de suspeitos, filtros e priorização de alertas.
- data de publicação: identifica a data da publicação e é usada na verificação da informação como alerta antecipado.
- título: identifica o assunto da postagem e é usado para auxiliar na extração dos principais termos.
- idioma: identifica o idioma da publicação e é usado para a especificação dos algoritmos para o processamento e uso de mecanismos de extração de alertas dependentes de língua e de vocabulário.
- localização: identifica a localização geográfica da postagem e é usada para a identificação de origem e correlação com outras fontes.
- recomendação: identifica o grau de credibilidade e aceitação por usuários da informação e é usada em filtros e priorização de alertas.
- disseminação: identifica o grau de importância do conteúdo por usuários e é usada para definir a credibilidade da informação e em filtros e priorização de alertas.
- palavras-chave: identificam os principais termos do conteúdo e são usadas para descrever o alerta e em filtros e priorização de alertas.
- referências: descrevem referências para assuntos ou autores relacionados ao conteúdo e são usadas para a correlação com outros perfis, fontes ou dados, priorização de alertas, complementação de informações.
- entidades: descrevem entidades (p. ex. IP, URL, software, organização) relacionadas ou que estão no próprio texto da postagem e são usadas para expansão e complementação dos dados, filtros e priorização de alertas.
- alvo: identifica a entidade alvo de um ataque (p. ex. IP, URL, software, organização) e é usada para avisar aos interessados ou responsáveis
- criticidade: identifica o grau de severidade do alerta e é usada na priorização dos alertas.
- tipo de ataque/vulnerabilidade: identifica o ataque ou a vulnerabilidade e é usado em filtros, priorização e notificação do alerta.

Em muitas fontes, esses atributos estão apenas no texto da publicação, logo não são evidenciados durante a coleta. Os componentes de pré-processamento e processamento de alertas são os responsáveis por identificar esses atributos para o EWS.

## 3.5 Bases de Inteligência

As bases de inteligência armazenam informações estruturadas para auxiliar nos pré-processamentos dos dados e nos processamentos de alertas. São constituídas por termos irrelevantes usados por filtros de remoção, por termos de cibersegurança usados por filtros de priorização e classificação de alertas, por padrões usados para ambos filtros de remoção e priorização e na classificação de alertas. São construídas a partir da análise de amostras de dados das fontes, por definições de especialistas em segurança, por um processo de retroalimentação a partir do processamento de alertas e por informações de bases consolidadas de cibersegurança.

Algumas bases de inteligência são voltadas para fontes específicas, por exemplo, bases de perfis suspeitos de usuários de microblogs. Outras são usadas para fins específicos, por exemplo, bases com URLs usadas para filtrar alertas irrelevantes em canais de notícias de segurança. No entanto, o arcabouço viabiliza o reaproveitamento ou adaptação de bases para diferentes fontes, por exemplo, base de termos irrelevantes criada e atualizada por termos comuns do microblog Twitter, que pode ser usada para filtrar as postagens obtidas no Facebook ou outra rede social.

No arcabouço, define-se algumas bases que auxiliam na detecção de alertas antecipados associados à cibersegurança:

- Bases de padrões: identificam padrões para identificar alertas relevantes ou não relevantes. Os padrões são combinações de atributos ou termos que possibilitam fazer essa distinção entre alertas e não alertas.
- Bases de filtros: identificam termos e entidades que possibilitam realizar a remoção ou priorização de informações. Essas bases geralmente podem ser usadas para a construção de listas brancas e negras para o uso em filtros básicos.
- Bases de treinamento: armazenam atributos e itens classificados usados em treinamentos de algoritmos de classificação supervisionados.
- Bases de reputação: armazenam a reputação de entidades, em especial, IP e URL, usados para identificar potenciais alertas. A reputação de autores também é usada para incluir ou descartar alertas.
- Bases de recomendação: armazenam a recomendação de alertas realizada por usuários e são usadas para remover, priorizar ou classificar alertas.
- Bases de estatísticas: armazenam estatísticas sobre autores, mensagens e fontes que auxiliam na remoção, priorização ou classificação de alertas.

Algumas bases de inteligência devem ser temporais, ou seja, possuírem a flexibilidade de manter itens durante um período finito, pois em mídias sociais é comum novos termos para filtros, especialmente de remoção, terem uma vida curta. Outras precisam ser adaptativas e retroalimentadas pelos algoritmos ou especialistas em segurança para contemplar novas informações.

## 3.6 Pré-processadores

Os pré-processadores são responsáveis por processar os dados coletados, remover as entradas irrelevantes e redundantes, priorizar as entradas segundo a relevância, e, para cada entrada, estender as informações e padronizar para um esquema estruturado. No arcabouço, os módulos que realizam esse processamento são divididos em três classes: Filtros, Normalizadores e Agrupadores. As subseções seguintes abordam individualmente cada classe.

### 3.6.1 Filtros

Os filtros são responsáveis por selecionar os dados na fase de pré-processamento. Eles são necessários para reduzirem a carga de dados para os próximos processos, pois o processamento de dados textuais exige uso extensivo de recursos computacionais, especialmente se houver muitos dados. Dessa forma, remover as informações irrelevantes e redundantes, e também, priorizar as informações que possuem maior probabilidade de serem alertas, são responsabilidades atribuídas aos filtros. No arcabouço, esses filtros são denominados de filtros de remoção e filtros de priorização, respectivamente.

Os filtros também são classificados pelo modo de operação: especialista, adaptativo e hierárquico. Os filtros especialistas usam listas e expressões para remover ou priorizar entradas, os filtros adaptativos possibilitam a retroalimentação e adaptação automatizada, e os filtros hierárquicos possibilitam a estruturação de diferentes filtros para gerir a sequência de processamento.

Os filtros básicos do arcabouço devem filtrar entradas textuais com conteúdos irrelevantes. Neste aspecto, os filtros especialistas e/ou adaptativos que usam listas negras são interessantes por possibilitarem a atualização de termos nas listas. Em geral, os filtros usam as bases de inteligência para gerenciar os termos e regras de filtragem. Como as bases são atualizadas na fase de processamento, os filtros adaptativos podem fazer uso dessas informações para otimizar os resultados.

O arcabouço define um conjunto de filtros base para um EWS:

- filtros de termos: removem ou priorizam entradas considerando listas brancas e negras de termos.
- filtros de entidades: removem ou priorizam entradas considerando listas de brancas e negras de entidades como URL, IP, autores, entre outros.
- filtros de combinações: removem ou priorizam entradas considerando combinações de termos, de entidades e/ou de padrões.
- filtros de características: removem entradas considerando características como tamanho, forma de escrita, número de elementos, tipos de elementos, entre outras, que geralmente acabam por indicar informações irrelevantes.
- filtros de tempo: removem entradas considerando dados de marcação de tempo que indiquem alertas antigos.
- filtros de duplicidade: removem mensagens iguais obtidas de diferentes sensores.

Independente se o filtro for de remoção ou de priorização, eles podem ser implementados como especialistas, adaptativos e hierárquicos. Os filtros hierárquicos, por exemplo, viabilizam a especificação de conjunto sequencial de filtros que atenda uma fonte de dados específica.

### 3.6.2 Normalizadores

Os normalizadores são responsáveis por estender os dados de entrada e por padronizar em uma estrutura bem definida. No arcabouço, tanto os componentes de padronização quanto os de expansão de dados são denominados de normalizadores.

O arcabouço define um conjunto básico de normalizadores para um EWS:

- normalizadores de idioma: identificam o idioma a partir do texto da publicação.
- normalizadores de entidades: identificam as entidades, tais como, URLs, IP, organizações, software, hardware, entre outras.

- normalizadores sintáticos: identificam os substantivos e verbos em sentenças, possibilitando assim, a identificação de sujeitos, ações e predicados.
- normalizadores de termos: identificam os termos chaves associados à cibersegurança, por exemplo, os termos de busca.
- normalizadores de URL: transformam URLs curtas para longas.
- normalizadores de categorias: identificam tópicos nas sentenças.
- normalizadores de formato: representam os dados em uma estrutura padronizada, isto é, normalizam e expandem os dados em informações estruturadas.

As informações estruturadas produzidas pela normalização de formato, devem procurar contemplar a identificação única para a informação, identificação da fonte, descrição da informação e do contexto associados à informação, data e hora de criação e coleta dos dados, idioma do texto, descrições do autor da publicação, identificação e localização de entidades no texto da publicação e dos termos associados à cibersegurança, entre outras informações agregadas ao texto da publicação.

No arcabouço, recomenda-se implementar filtros de remoção para diminuir o número de informações antes de executar as normalizações. No entanto, alguns filtros de remoção são dependentes de informações providas pelos normalizadores, como é o caso de filtros que usam URLs e dependem do idioma. Logo, somente filtros básicos, como por características e por listas brancas e negras são executados antes dos normalizadores.

### 3.6.3 Agrupadores

Os agrupadores são responsáveis por juntar informações de uma única fonte e evitar que informações duplicadas ou similares sejam enviadas para as fases seguintes. Em um EWS, é importante considerar que informações recentes sejam encaminhadas, logo informações antigas não precisam necessariamente serem enviadas para os processadores de alertas, exceto se agregam novidades, como as atualizações de alertas.

O fator mais importante para os agrupadores é a janela temporal. A janela temporal define o tempo máximo entre a primeira e última informação similares pertencerem ao mesmo grupo. Agrupadores baseados somente em similaridade devem avaliar o tamanho da janela temporal. Se a janela for curta, informações similares não serão agrupadas, se for longa, informações distintas serão agregadas. Agregadores que consideram heurísticas, por outro lado, podem flexibilizar a janela temporal. Por exemplo, considere a avaliação de uma entrada que possui uma URL que aponta para o mesmo destino. Nessa situação, a probabilidade é alta que sejam assuntos diretamente relacionados, principalmente em se tratando de microblogs e redes sociais. A janela temporal deve ser avaliada através da análise de agrupamentos (seção 3.11.8) para cada fonte.

## 3.7 Processadores de Alertas

Os processadores de alertas são responsáveis por processar e classificar as informações selecionadas e normalizadas na fase de pré-processamento como alertas. Também podem realizar a análise e correlação com alertas de fontes de dados tradicionais ou outros alertas gerados pelo próprio arcabouço. Os processadores são divididos em quatro classes: Classificadores, Analisadores, Recomendadores e Correlacionadores. As subseções seguintes abordam individualmente cada classe.

### 3.7.1 Classificadores

Os classificadores são responsáveis por processar e determinar para uma entrada a classe que pertence em um conjunto de classes predefinido. No caso do EWS, as classes são alerta ou não. E, no caso de um alerta, averigua-se o potencial de ser um alerta antecipado.

Conforme experimentos (seção 4.1.3.4), a coleta de informações de cibersegurança pode resultar em alertas relevantes à segurança de sistemas ou específicos para um público (p. ex. administrador de redes ou usuário doméstico), ou alertas irrelevantes, como no caso de informações sobre segurança não computacional (p. ex. mensagens de segurança pública), propagandas ou dicas (p. ex. propaganda de antivírus), informações desatualizadas e não mais relevantes como alertas ou outros conteúdos (p. ex. mensagens mal formadas, rumores, entre outros).

A escolha do classificador deve considerar as possibilidades citadas, pois podem conduzir a geração de muitos falsos positivos. Para tal, recomenda-se avaliar a fonte e as mensagens para identificar outras características além do texto, visando assim, selecionar os melhores classificadores para cada fonte específica.

### 3.7.2 Analisadores

Os analisadores são responsáveis por realizar a análise e extração de novos conhecimentos a partir do processamento de uma entrada ou conjunto de entradas, composta por informações estruturadas ou por alertas.

As ações executadas por analisadores podem ser: agregar atributos de mensagens padronizadas, identificar novos termos, construir o histórico de alertas, complementar informações, por exemplo, identificar o alvo ou o tipo de alerta. Dependendo do objetivo, os analisadores operam independente das outras classes de processadores ou sequencialmente após a confirmação de alerta. Também podem ser usados na tomada de decisão em incerteza sobre um alerta.

### 3.7.3 Recomendadores

Os recomendadores são sistemas de recomendação que possibilitam recomendar alertas para usuários segundo as características desses alertas (baseados em conteúdo) ou similaridade entre usuários e/ou alertas (filtragem colaborativa). Possibilita a colaboração entre especialistas e usuários do EWS para viabilizar a priorização de alertas e descarte de falsos positivos. Devido a natureza não estruturada dos alertas, há alertas que precisam ser analisados por humanos para serem identificados como relevantes e, muitos, mesmo quando analisados, são difíceis de se distinguir como antecipados ou não devido a forma como foram escritos.

Os recomendadores também são interessantes para tratar sobre rumores e relevância de alertas. Os resultados dos recomendadores podem ser usados para filtrar conteúdos irrelevantes em fases anteriores ao processamento e para aumentar a credibilidade e importância de alertas na base de alertas. Auxiliam também os sistemas de notificação a encaminharem os alertas para o grupo de interesse correto.

O uso de recomendadores possibilita a colaboração entre entidades do EWS. Como são abordados alertas de fontes de dados não estruturados, por meio da colaboração, especialistas de segurança podem detalhar os alertas do sistema e também publicar anonimamente alertas que podem ser de interesse de outros especialistas ou administradores de redes. O arcabouço procura automatizar todas as fases de um EWS, mas considera essencial a participação e colaboração de especialistas para um EWS produzir minimamente falsos positivos e compartilhar informações úteis como alertas.

### 3.7.4 Correlacionadores

Os correlacionadores são responsáveis por associar ou relacionar alertas de fontes de dados não estruturados com informações estruturadas ou alertas providos por outros sensores de redes, por exemplo, de IDS, sistemas de inventário, *honeypots*, entre outros. Apesar de ser interessante realizar essa atividade, os componentes dessa classe de processadores são complexos de serem implementados.

No arcabouço, são enumerados alguns tipos de correlacionadores que podem ser desenvolvidos:

- correlacionar a infraestrutura de software para identificar alertas de interesse para um perfil de usuário do EWS.
- correlacionar alterações no tráfego da rede e identificar se algum alerta pode estar associado a essa alteração.
- correlacionar o tráfego de saída e verificar se há pacotes que correspondem a algum alerta que identifica IP, URL ou alvo.
- associar o aumento de tráfego em um servidor com informações coletadas pelo EWS, por exemplo, identificar constantes tentativas de conexões em um servidor Web e postagens em uma rede social sobre DDoS contra a infraestrutura dessa rede, pode indicar um ataque.
- associar acessos novos ou mudanças de páginas em um servidor Web em horários não usuais, por exemplo, finais de semana, e tentar verificar se correspondem a desfigurações de páginas divulgadas em redes sociais.
- identificar ameaças recentes e relacionar com serviços de uma infraestrutura, ou seja, identificar os serviços altamente visados e aumentar o monitoramento ou sensibilidade de alertas nesses serviços.
- correlacionar informações coletadas em *honeypots* com alertas recentemente publicados em bases de vulnerabilidades ou de códigos de exploração.

## 3.8 Base de Alertas

A base de alertas armazena os alertas produzidos pelos processadores de alerta. Logo, deve usar um formato padronizado e que possibilite relacionar informações adicionadas por analisadores, correlacionares e recomendadores.

Apesar do arcabouço não especificar um formato padrão, pois há vários formatos para a representação de alertas, conforme discutido na revisão da literatura (Seção 2.3.6), são destacados atributos que devem ser identificados quando possível e que são compatíveis com formatos já consolidados:

- identificação: identificação única para o alerta.
- fonte: identificação da fonte de dados.
- sumário: descrição curta do alerta, geralmente no formato textual.
- descrição: descrição completa do alerta, geralmente o texto original e informações de contexto.
- data de criação: indicação temporal de quando a informação foi publicada/divulgada.
- data de coleta: indicação temporal de quando a informação foi coletada.
- confiança: indicação de confiança que a informação é um alerta.

- severidade: indicação do grau de criticidade/severidade da notificação descrita no alerta. Pode ser preenchido automaticamente por analisadores ou diretamente por especialistas de segurança com a finalidade de elevar a atenção para o alerta.
- categoria: indicação de categoria do alerta, por exemplo, desfiguração de página, orquestração de ataque, vulnerabilidades, atualizações de software, vazamento de dados.
- termos: identificação e localização dos termos associados à cibersegurança que aparecem na descrição do alerta.
- autor: identificação de autoria do ataque ou do provável autor do ataque quando refere-se a ameaças à segurança.
- alvo: identificação dos alvos descritos no ataque, por exemplo, software, sistema operacional, serviço, URL, empresa, hardware, pessoa.
- entidades: identificação de entidades do alerta, por exemplo, URL, IP, palavras-chave, software, entre outros.
- idioma: identificação do idioma na descrição do alerta.
- associações: referência para informações associadas ao alerta. Pode ser usado para indicar o agrupamento de mensagens/ informações similares de uma mesma fonte ou fontes distintas.
- relacionamentos: referência para outros alertas, especialmente para alertas de fontes tradicionais (p. ex. IDS). Deveria ser preenchido somente por um especialista ou por correlacionadores.
- adicionais: estruturas adicionais para descrever o alerta, por exemplo, imagens, trechos de códigos maliciosos, indicações de resolução do problema, outros.

Os atributos citados são importantes para a realização da classificação e para a notificação de alertas, no entanto, como potenciais alertas proveem de dados não estruturados, não é trivial identificá-los nas informações coletadas diretamente da fonte. A Tabela 3.1 apresenta exemplos de alertas antecipados identificados nas categorias de fontes definidas para o EWS. Como pode ser observado, cada fonte apresenta vocabulários específicos (formais ou informais), variações no idioma (inglês e português), e claro, um formato não estruturado. Logo, as implementações dos normalizadores e processadores devem contemplar algoritmos para extrair esses atributos geralmente direto do texto da publicação.

## 3.9 Notificadores

Os notificadores de alertas são responsáveis por disseminar os alertas para o grupo de interesse. Podem ser disponibilizados como sistemas Web ou local, usar tecnologias já consolidadas como a correspondência eletrônica ou ainda divulgar em redes sociais. Apesar de não precisar de auxílio humano para executar suas operações, os notificadores podem necessitar de revisão de alerta por especialista em segurança caso o alerta apresente um nível de confiabilidade baixo.

Nos notificadores é indicado o uso mecanismos para ocultação ou ofuscação de informações do alerta com o intuito de garantir a privacidade do alvo ou evitar o vazamento de informações sensíveis para entidades que não deveriam ter acesso. Logo, é necessário definir uma política de privacidade para os notificadores segundo a forma de publicação.

Em situações em que o EWS propicia o compartilhamento de informações identificadas pelos próprios administradores nas infraestruturas que administram, deve-se prover mecanismos para ocultar informações sensíveis da infraestrutura de rede e também anonimizar o remetente da mensagem.

**Tabela 3.1:** *Exemplos de alertas das fontes de dados não estruturados.*

| Fontes | Exemplos de Alertas |
|---|---|
| Redes sociais | (Facebook) é isso mesmo ?? 0day no cms do gov do RS?? https://www.facebook.com/ http://www.susepe.rs.gov.br/<br>(Facebook) #Hacked :D? #N0V3??Site: http://icaro.defesasocial.mg.gov.br/?-.... ?Mirror: http://www.zone-h.org/mirror/id/24970628<br>(Facebook) Como prometido no post anterior!??Relatórios de 2003 à 2015 da Aneel??Download aqui: https:... |
| Microblogs | (Twitter) https://t.co/bwhHgO562F?CENTRO DE PAGAMENTO DO EXÉRCITO?#Pwned https://www.cpex.eb.mil.br/<br>(Twitter) #tangodown http://t.co/QPaM9o4qgl #anonymous #lafirmasec<br>(Twitter) WarezTuga alvo de ataque e milhares de passwords são divulgadas: ... |
| Chats | (IRC) TARGET: DDoS www.brasil.gov.br & www2.brasil.gov.br<br>(IRC) hey guys, i'm looking for a trojan backdoor i can email a target ...<br>(IRC) link to the 0day? |
| Sítios especializados | (US-CERT) Alert (TA15-105A), Simda Botnet, ... April 15, 2015, Systems Affected Microsoft Windows... Overview...<br>(Symantec) Backdoor.Voldat, Risk Level ... Discovered: ... November 17, 2015 ...Updated: November 19, ... Systems Affected: Windows...<br>(Zero Day Initiative) Adobe Flash TextField autoSize Use-After-Free Remote Code Execution Vulnerability |
| Blogs | (Blog Sektioneins) OS X 10.10 DYLD_PRINT_TO_FILE Local Privilege Escalation Vulnerability, posted: 2015-07-07 17:30 by Stefan Esser<br>(TheHackersNews) 13-year-old SSL/TLS Weakness Exposing Sensitive Data in Plain Text...Saturday, March 28, 2015 Swati Khandelwal<br>(Blog auth0) Critical vulnerabilities in JSON Web Token libraries ... Tim McLean ... March 31, 2015 |
| Notificações de publicação | (RSS) Zero Day Initiative - Upcoming Advisories (Notificações de vulnerabilidades que estão para ser anunciadas publicamente)<br>(RSS) Defacements RSS - Zone-H.org (Notificações de desfigurações que acabaram de ser compartilhadas no site) |
| Bases de vulnerabilidades | (CVE) CVE-2015-8043 ... 10-11-2015 22:00 ... Use-after-free vulnerability in Adobe Flash Player before...<br>(OSVDB) 119873...2015-02-19...Linux Kernel Intel Microcode Loader Local Stack Buffer Overflow Weakness<br>(Exploit Database) 2015-12-08 ...phpFileManager 0.9.8 Remote Code Execution...php...metasploit |
| Listas de e-mail | (Seclists.org - Full Disclosure) BF and CE vulnerabilities in ASUS RT-G32...From: "MustLive" ... Date: Mon, 30 Nov 2015 23:51:40 +0200<br>(Seclists.org - Bugtraq) [SECURITY] [DSA 3409-1] putty security update |
| Motores de busca | (Google) site:br index of "/application/configs/" (vulnerabilidade no Zend Framework)<br>(SqlMap) sqlmap -u "www.target.com/vuln.php?id=1" (busca por base de dados vulneráveis) |
| Serviços compartilhamento de texto | (PasteBin) [Not Patched] 0-day XenonLegend Exploit Hack V.1 [02-17-15]<br>(PasteBin) Infoleakdatabase Prefeitura Municipal de Natal by: a guest on Feb 18TH, 2015<br>(GitHub Gist) googleinurl/facecheck2.0.php |
| Serviços compartilhamento de vídeos | (YouTube) https://www.youtube.com/watch?v=DOpryDLFSxM... Joomla SQL Injection Vulnerability in Full Administrative Access...<br>(YouTube) How to find lots of SQL vulnerable sites |
| Fóruns Web | (AnonymousBrasil) Tópico: #Hacked,... estou vendo que todos têm atacado os .gov ... em abril vamos ter uma maratona de invasões... |

## 3.10 Entidades

As entidades são responsáveis por receber as notificações de alerta e executar os procedimentos de contramedidas. Também podem auxiliar na identificação de informações irrelevantes e na confirmação de rumores de ameaças.

No arcabouço, são definidas as seguintes classes:

- Sistemas: são sistemas computacionais que irão consumir os alertas diretamente. Pode ser um sistema passivo, que simplesmente recebe a notificação e aguarda interação de um administrador, ou pode ser um sistema ativo, que recebe a notificação e ativa mecanismos de reação.
- Profissionais de Segurança TI: são administradores de redes e/ou especialistas em segurança que usarão a informação do alerta para responder a um ataque caso a infraestrutura de TI tenha sido comprometida ou usarão a informação para proativamente proteger a infraestrutura de TI.
- Pública: são entidades públicas que divulgam informações de segurança, por exemplo, sítios ou blogs de boletins de segurança. Em geral, essas entidades detalham o alerta e a forma de mitigação antes de publicar a informação.
- Outros: são entidades interessadas em notificações específicas do sistema, por exemplo, em notificações de um software ou uma infraestrutura. Por exemplo, entidades interessadas em notificações de orquestrações para roubo de identidade.

As entidades em nível nacional, como o Governo, estão interessadas em ameaças à infraestruturas críticas e que acarretam prejuízos a toda a população. Nesse caso, as políticas de privacidade e os mecanismos de notificação devem seguir também políticas de privacidade estritas a essas organizações.

## 3.11 Análise de Dados

O processo de análise de dados é o alicerce para a identificação de características usadas na construção dos algoritmos de monitoramento, pré-processamento e processamento de alertas. Também é fundamental para a construção das bases de inteligência usadas por esses algoritmos. Nesta seção, são apresentados processos e suas flexibilizações para a realização da análise em fontes de dados não estruturados visando o desenvolvimento de sistemas de alerta antecipado. O próprio processo de análise pode ser considerado um arcabouço voltado à análise de dados de cibersegurança.

A Figura 3.4 apresenta o processo de análise de dados de cibersegurança. A análise consiste em processar como entrada uma amostra de dados coletados de fontes de dados não estruturados que propagam informações associadas à segurança e a ameaças computacionais e devolver como saída características que compõem as informações de inteligência, isto é, as informações para a criação de bases de inteligência e algoritmos para o processamento e identificação de alertas, preferencialmente antecipados.

O processo de análise é constituído por um subconjunto de processos (retângulos) que podem ou não ser executados dependendo da fonte analisada. A entrada (cilindro) é uma amostra de dados coletados em uma fonte de dados não estruturados e a saída (losangos) são as características usadas para a construção das bases de inteligência e dos algoritmos a serem implementados no arcabouço EWS. Nas próxima subseções, são discutidas individualmente a entrada, os processos de análise e a saída. Pretende-se descrever como o processo de análise provê o alicerce para a construção dos algoritmos e bases de inteligência usadas na extração de alertas antecipados.

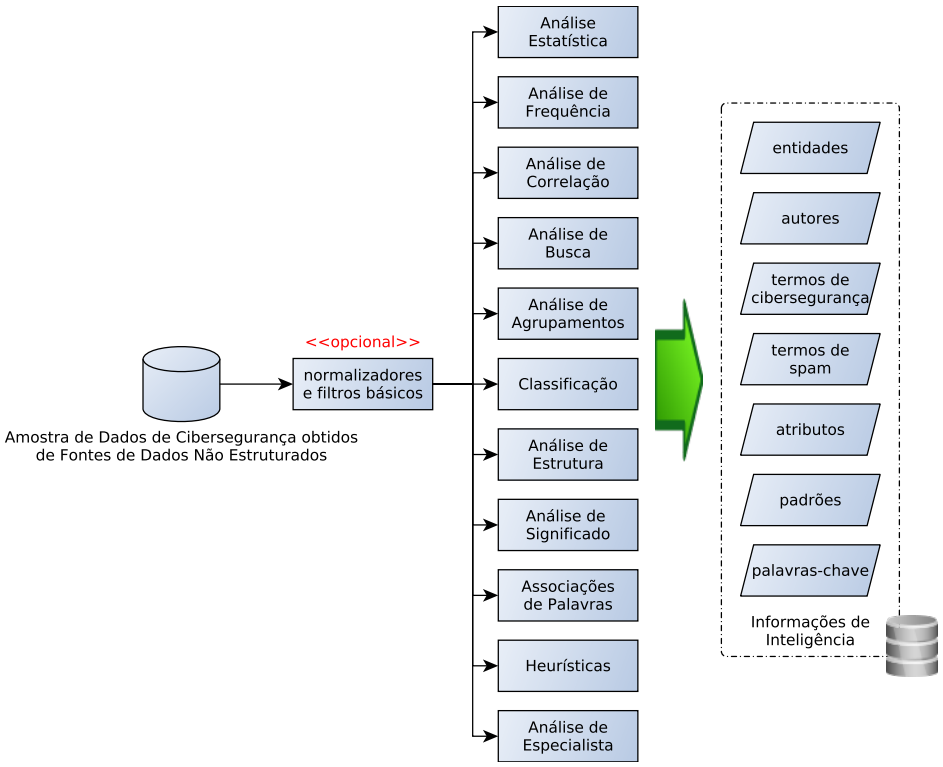


**Figura 3.4:** *Análise de dados de cibersegurança obtidos de fontes de dados não estruturados.*

### 3.11.1 Amostra de dados das fontes de dados não estruturados

A entrada do processo de análise é uma amostra de dados coletadas da fonte a ser analisada. Os tipos e características das fontes estão descritas na seção 3.2. Considerando que os dados já foram coletados, uma questão importante é determinar o tamanho da amostra, ou seja, se representa de forma fidedigna o conteúdo propagado na fonte.

Determinar o tamanho da amostra, isto é, a quantidade de informação para obter uma amostra significativa, é uma questão complexa e, em geral, envolve a experiência e conhecimento do analista sobre o que é discutido na fonte. Pode ser realizada de forma incremental, analisando-se resultados coletados segundo um tempo de monitoramento até observar que os resultados não trazem novas informações ou descobertas; por um valor empírico, considera-se as experiências anteriores (M mensagens durante D dias de monitoramento); por uma base com muitas mensagens que garanta a inclusão de uma variabilidade significativa de informações; ou por uma amostra probabilística.

Independente do mecanismo de amostragem, os dados publicados em muitas fontes de segurança cibernética são influenciados por acontecimentos pontuais e ações de interesses de grupos hackers, o que pode levar uma amostra ter uma representatividade significativa apenas em um determinado período temporal. Logo, o processo de análise de dados deve ser realizado sazonalmente para possibilitar a identificação de novas características para serem incorporadas ao arcabouço de análise e extração de alertas.

### 3.11.2 Informações de inteligência

A saída do processo de análise são as informações de inteligência, isto é, atributos e valores de interesse para o EWS observados nas fontes de dados, que são usados para a construção de bases de inteligência e algoritmos de extração de alertas. As saídas dos processos de análise propostos no arcabouço são categorizadas em sete grupos de interesse: entidades, autores, termos de cibersegurança, termos de spam, atributos, padrões e palavras-chave.

As entidades são agregados sequenciais de texto, compostos por uma ou mais palavras, que possuem um significado dentro de um contexto ou categoria. No caso, especificamos um conjunto de categorias de interesse para um EWS, que são: URL, IP, Localização, Acrônimo, Software, Hardware, Organização, Pessoa, Marcação de tempo, entre outros. Essas entidades possibilitam identificar alertas relevantes, identificar alvos ou origens de ataque, validade do alerta como antecipado, estruturação do alerta, elaboração de algoritmos mais precisos para filtragem e classificação dos alertas.

Os autores são os responsáveis pela postagem da informação em uma fonte de dados. A enumeração de autores possibilita inferir sobre a confiabilidade da origem da informação, identificar responsáveis por ataques ou orquestrações, relacionar perfis para monitoramento ou para a exclusão (autores automatizados, por exemplo). Dependendo do meio a ser monitorado, por exemplo, canais e fóruns hackers, os autores são identificados por pseudônimos, que podem ou não variar em postagens. O monitoramento de nomes de autores referenciados em mensagens também viabiliza a identificação de novos perfis a serem monitorados no processo de coleta. Esses perfis também podem ser usados nos algoritmos de filtragem e classificação dos alertas.

Os termos de cibersegurança são os termos (palavras, siglas, símbolos) que se relacionam com segurança computacional. Alguns exemplos de termos são DDoS, ataque, vulnerabilidade, patch, entre outros. Muitos desses termos já fazem parte da base padrão de monitoramento, mas é importante identificar os mais frequentes em uma fonte específica. O principal uso são no processo de monitoramento e nos algoritmos de identificação de potenciais alertas.

Os termos de spam são os termos que se relacionam com outros assuntos e não à segurança computacional. Alguns exemplos são termos de segurança pública, ataques terroristas, propagandas, vídeos, notícias publicadas por portais jornalísticos, entre outros. Como essas informações não são de interesse para um EWS, os termos de spam são comumente usados como filtros para a remoção e desqualificação de postagens como alertas.

Os atributos são características observadas nas fontes e dados que podem ser usados para filtrar, agrupar e classificar alertas. Alguns exemplos de atributos que podem ser identificados em postagens são a idade da mensagem, confiabilidade da informação, severidade do alerta, presença de referências, se a mensagem tem origem de um autor monitorado, número de pessoas que recomendaram a postagem, entre outros. Os atributos podem variar de fonte para fonte, por isso recomenda-se o estudo da fonte e das informações que ela propaga em conjunto com as mensagens, especialmente os metadados.

Os padrões são combinações de informações de possibilitam identificar postagens como alertas ou como não alertas. Um exemplo de padrão para um alerta, seria, por exemplo, a presença da palavra DDoS em conjunto com um IP e a informação foi postada a partir de um perfil monitorado. Um exemplo de padrão para um não alerta seria a palavra ataque combinada com um termo de futebol.

As palavras-chave são palavras que possibilitam aumentar o interesse em mensagens monitoradas dependendo da fonte ou do período de monitoramento. Por exemplo, dependendo do período, o aparecimento de palavras como Olimpíadas e ENEM, em mensagens de cibersegurança, podem indicar ataques ou orquestrações. Além disso, há palavras-chave usadas em determinadas comunidades, como “nova” , “crítico”, “urgente”, que dependendo de associações e padrões de uma fonte, podem indicar, por exemplo, “nova vulnerabilidade”, “bug crítico” ou “atualização urgente”.

### 3.11.3 Normalizadores e filtros básicos

Os normalizadores são responsáveis por padronizar e expandir os dados para os próximos subprocessos de análise. Os filtros básicos são filtros para remover conteúdos irrelevantes, como mensagens automatizadas ou pequenas. Para alguns processos, os filtros não devem ser usados, exceto para

remover redundância de monitoramento. Logo, no processo de análise, deve-se considerar todo o conteúdo da fonte analisada para descobrir informações de inteligência que possibilitem não apenas evidenciar alertas, mas excluir informações irrelevantes.

O processo de normalização deve realizar a formatação dos dados para um padrão (p. ex. JSON, CSV ou outros) ; selecionar os metadados de interesse (p. ex. marcação de tempo, origem, e outros) e expandir os dados se necessário (p. ex. expandir URLs, identificar idioma). Os filtros devem realizar processamentos básicos como remover mensagens que sejam curtas, malformadas, repetidas, de usuários automatizados ou com conteúdos irrelevantes.

### 3.11.4 Análise estatística

A análise estatística, em especial, a análise estatística descritiva, proporciona avaliar o volume de informações coletado em uma fonte, definir critérios para o monitoramento da fonte e dos dados, analisar o comportamento de postagem dos usuários na fonte. Isso é possível pelo sumariação dos dados em variáveis quantitativas e qualitativas, e pela visualização desse conjunto de informações por gráficos.

No processo de análise estatística pode ser observado o número mínimo, máximo e médio de postagens diárias, tamanho das postagens, número de usuários, as entidades presentes nas mensagens, o tempo ininterrupto de monitoramento, entre outras variáveis. A análise pode ser realizada em todos os dados e nos dados filtrados. Com os dados filtrados, verifica-se a carga de processamento para o sistema de classificação de alertas.

### 3.11.5 Análise de frequência

A frequência é usada como uma medida da estatística para indicar a repetição de ocorrências de um evento. Os tipos de mais comuns são a frequência absoluta, quantidade de vezes que um elemento ocorreu na amostra ou população, e a frequência relativa, percentual de vezes que um elemento ocorreu em relação à amostra ou população. Neste arcabouço, é usada identificar os termos mais frequentes associados ou não a cibersegurança presentes nos dados capturados das fontes de dados e, possibilitar dessa forma, a construção de bases de inteligência compostas por termos de cibersegurança e irrelevantes (spam).

A análise de frequência pode considerar a frequência de palavras e sentenças, palavras mais comuns, frequência de tamanho de palavras e sentenças, frequência de sílabas por palavra e sentença, frequência de mensagens segundo a expressão de busca e/ou autores. A análise de frequência requer que um especialista selecione as palavras a serem usadas para a elaboração das bases de inteligência. Também recomenda-se remover as palavras comuns irrelevantes (*stop words*) antes de realizar essa análise.

### 3.11.6 Análise de correlação

A análise de correlação no arcabouço corresponde a encontrar associações de elementos textuais por similaridade e não à correlação de variáveis estatísticas. A correlação de informações de fontes previamente conhecidas associadas à cibersegurança possibilita qualificar a fonte analisada como relevante para alertas ou ainda identificar padrões e termos nos dados para auxiliar na extração de alertas.

Bases de dados consolidadas de ameaças podem ser usadas para realizar a correlação, tais como CVE, boletins de notificações de segurança, bases de software, listas enumeradas de ameaças, ontologias de segurança. Se dados em uma fonte foram similares a mensagens publicadas nesses meios, há assuntos para serem monitorados. Outro uso para as bases consolidadas consiste em verificar se

há alertas que foram primeiramente notificados por uma fonte antes de meios mais tradicionais. Em geral, bases consolidadas mantém a data e hora da divulgação pública das notificações de segurança.

### 3.11.7 Análise de busca

A análise de busca é baseada no processo de indexação e busca em documentos textuais. Esse processo consiste em selecionar um conjunto de documentos, dividir cada documento em termos, processar cada termo por meio de operadores linguísticos (lematização, radicalização, remoção de termos comuns não relevantes à busca), indexar os documentos baseados na ocorrência de termos e possibilizar a realização de consultas por meio de expressões de busca. Essas expressões de busca podem ser constituídas por palavras-chave, frases, proximidade de termos, intervalos, campos e outras combinações.

O processo de busca devolve como resultado um conjunto de documentos classificados e priorizados segundo a expressão de busca. Dessa forma, pode-se realizar a análise de busca para identificar os documentos que se relacionam com cibersegurança em uma coleção de documentos. O resultado devolvido pode ser usado para a identificação de vocabulário e padrões usados na fonte. Além disso, a própria expressão de busca pode ser usada na extração de alertas de segurança. A indexação de datas possibilita restringir a busca a informações recentes, o que é importante para um EWS.

### 3.11.8 Análise de agrupamentos

O agrupamento de dados consiste em dividir os dados em grupos de forma que as informações internas de cada grupo sejam similares entre si. Aplicando esse conceito em uma coleção de documentos, é possível gerar grupos associados e não associados à cibersegurança. Dentro dos grupos associados, teremos outros grupos que estão relacionados a notificações específicas.

A análise dos grupos possibilita observar o comportamento temporal de informações de cibersegurança, identificar padrões de agrupamento de informações relevantes e irrelevantes, priorizar mensagens que são muito difundidas, criar filtros para a remoção de mensagens sem relevância, entre outros. Logo, realizar os processos de análise de frequência e análise por especialista aplicados aos grupos direcionam a extração de informações de inteligência.

### 3.11.9 Classificação

A classificação de dados consiste em etiquetar recursos associados à cibersegurança com o intuito de gerar bases de treinamento a serem usadas por algoritmos de aprendizagem supervisionada. Em geral, o processo de classificação de texto como alertas é realizado manualmente sobre uma amostra de dados. Também pode ser automatizada desde que as notificações sejam todas associadas à segurança.

Por meio dos algoritmos de classificação, é possível enumerar características que influenciam na identificação positiva de alertas. Para uma precisão na classificação manual, recomenda-se que seja realizada por dois ou mais especialistas para solucionar as disputas de conteúdos que são ou não alertas. Para realizar a classificação, é de interesse identificar os tipos de informações que são considerados alertas antecipados e quais não são.

A Tabela 3.2 apresenta os critérios para a classificação de informações como alertas ou não alertas.

**Tabela 3.2:** *Critérios para classificação manual de alertas.*

| Alertas | Não Alertas |
|---|---|
| - Desfiguração de páginas, publicações de notícias de páginas desfiguradas visualmente.<br>- Vazamento de dados, publicações que contêm nomes de usuários, senhas e/ou dados pessoais.<br>- Ferramentas de ataque, publicações que abordam atualizações ou novos aplicativos maliciosos.<br>- Orquestração de ataques, organização de grupos para atacar algum alvo ou publicações de potenciais alvos vulneráveis.<br>- Vulnerabilidades e correções de software, publicações sobre correções ou vulnerabilidades de software. | - Notícias sobre hackers presos.<br>- Pessoas dizendo que são hackers.<br>- Contas comprometidas em redes sociais.<br>- Notificações de segurança ou notícias de outras áreas.<br>- Dicas ou propagandas de produtos sobre segurança de computadores. |

### 3.11.10 Análise de estrutura

A análise de estrutura compreende a análise morfológica e sintática de notificações de segurança. Com a análise morfológica é possível determinar variações de palavras de interesse que estão associadas à cibersegurança. Por exemplo, identificar as variações da palavra "hacker", como "hackers", "hackeando". Esse tipo de análise possibilita criar novas buscas e criar regras genéricas para o processamento de notificações de segurança com as flexões morfológicas de palavras. A análise sintática avalia a gramática do conteúdo textual e pode ser usada para identificar o papel das palavras em sentenças de segurança. Por exemplo, pode-se identificar o sujeito e o predicado da ação em sentenças. Assim, é possível identificar importantes conceitos que podem ser úteis na automatização para identificação de autores, alvos e formas de ataque.

### 3.11.11 Análise de significado

A análise de significado compreende a análise semântica e pragmática de notificações de segurança. Com a análise semântica é possível fazer a identificação de entidades, eliminar ambiguidade e auxiliar na análise pragmática em sentenças. Pode ser usada na criação de ontologia para uma fonte em análise ou mesmo uma ontologia de cibersegurança. Com a análise pragmática pode-se confirmar o contexto da sentença.

### 3.11.12 Associações de palavras

A mineração de associações de palavras possibilita identificar relações entre palavras dentro de um contexto. São interessantes para prover variações em consultas em recuperação de texto e para auxiliar na identificação de tópicos e entidades no contexto de cibersegurança. No arcabouço recomenda-se realizar a busca por relações sintagmáticas e paradigmáticas. Em especial, a identificação de *collocations* auxilia na elaboração de expressões de busca por alertas de cibersegurança.

### 3.11.13 Heurísticas

O uso de heurísticas possibilita identificar ou descartar notificações como alertas ou alertas antecipados. A avaliação de heurísticas viabiliza a geração de conhecimento para a implementação de mecanismos que auxiliam na criação de filtros e classificadores de alertas.

Neste arcabouço são sugeridas a análise das seguintes heurísticas:

- Identificar URLs que geralmente apontam para conteúdos associados ou não à cibersegurança.
- Verificar os autores que geralmente postam notificações de interesse.

- Observar a presença de data nas notificações visando aceitar ou excluir a mensagem como alerta antecipado.
- Observar influência da estrutura da mensagem como notificação (tamanho, número de palavras, frequência de postagem).
- Observar o vocabulário usado nas mensagens das notificações e não notificações (p. ex. presença de emojis em alertas).
- Analisar as mensagens que são perguntas para verificar interesse em atividades ilícitas.
- Observar as categorias de notificação que uma mensagem de segurança pode pertencer (p. ex. segurança computacional, segurança física, segurança criminal).
- Procurar por heurísticas específicas de domínio (p. ex. excesso de hashtags na mensagem via Twitter).

A análise dos resultados produzidos pelas heurísticas deve ser traduzida para algoritmos que otimizem filtros, normalizadores e classificadores do arcabouço. Por exemplo, mensagens com poucos caracteres não são notificações de segurança, logo essa heurística pode ser usada para implementar um filtro que remova mensagens com poucos caracteres.

### 3.11.14 Análise de especialista

A análise de especialistas corresponde ao processo manual de verificação de dados em uma fonte de dados por especialistas em segurança com a intenção de observar padrões e comportamentos de propagação de informações de cibersegurança.

A análise de dados *offline* é usada para a construção das bases de conhecimento e ocorre em todas as fases de análise. Já a análise de dados *online*, com a participação do analista diretamente na fonte, propicia a identificação de outras características ou situações relevantes, até porque há a interação com a fonte e seus usuários.

## 3.12 Em direção à uma arquitetura distribuída e colaborativa

O arcabouço EWS foi construído visando a sua implementação em arquiteturas distribuídas e possibilitando a colaboração entre especialistas de segurança e/ou administradores de redes. Nesta seção, é mostrado como o arcabouço pode ser implantado em uma arquitetura distribuída e como a colaboração entre os parceiros pode ser alcançada nessa arquitetura.

No arcabouço EWS há quatro componentes principais, que são: coletores, pré-processadores, processadores e notificadores. A Figura 3.5 mostra a distribuição desses componentes em uma arquitetura distribuída.

Os coletores são independentes entre si e podem ser instalados em máquinas individuais ou não Podem ser implementados com redundância para garantir a coleta de todas as informações. Podem ainda dividir as expressões de busca para não sobrecarregar as consultas, principalmente porque há serviços que limitam o número de consultas por autenticação de sessão, usuário ou máquina.

Os pré-processadores podem ser agregados diretamente nos coletores ou em máquinas remotas. Se agregados nos coletores, evitam a transferência de informações irrelevantes na rede. Por outro lado, geram uma carga de processamento nos coletores que deve ser considerada na implantação do coletor.

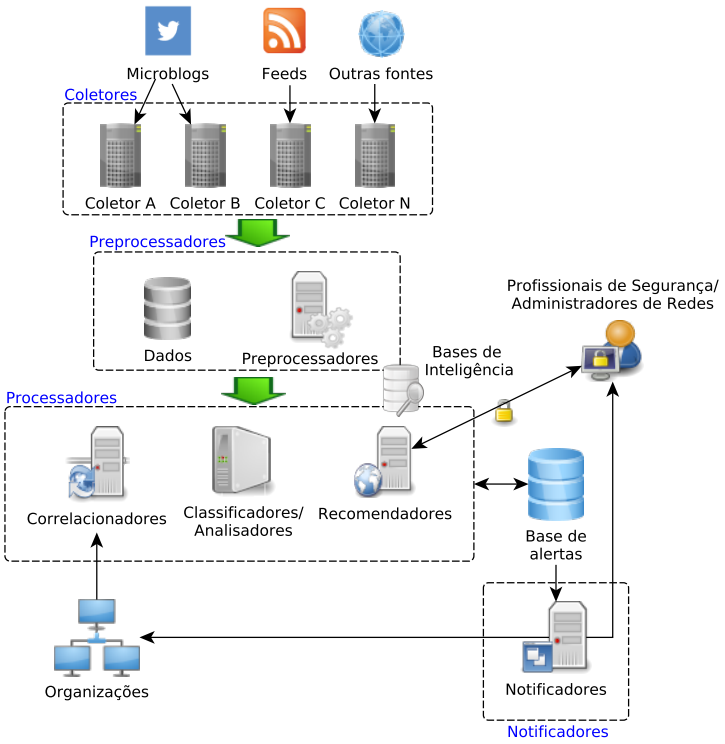


**Figura 3.5:** *Arquitetura para um EWS baseado em fontes de dados não estruturados.*

Os processadores podem ser divididos em múltiplas máquinas, mas devem estar localizados em um órgão de confiança devido à importância das informações compartilhadas entre as entidades envolvidas. A comunicação entre os componentes deve ser realizada por um canal confiável e criptografado. Somente as entidades cadastradas e autenticadas podem enviar informações. Podem ser adotados padrões com o STIX para a representação das informações e TAXII para a troca de mensagens, ou ainda, implementadas estruturas mais simples para incentivar e facilitar a coleta e transferência de informações.

Ao centralizar os processadores em um órgão de confiança, por exemplo, órgão governamental, ganha-se em gerenciamento e segurança das informações. Como os correlacionadores também recebem alertas de sensores tradicionais de organizações parceiras, a arquitetura deve prover anonimidade e proteção às informações sensíveis dos parceiros. A desvantagem da centralização dos processadores é a sobrecarga na coleta e processamento das informações. No entanto, há tecnologias como nuvens de armazenamento e processamento que conseguem lidar com grandes volumes de informações.

Não é definido um formato para envio de informações de sensores tradicionais para o EWS, mas como há padrões para representação de alertas, como o IDMEF, prefere-se que sejam empregados. As organizações que participam do EWS, ou seja, organizações parceiras, devem para garantir a anonimidade dos alertas e prover metadados para descrever e contextualizar os alertas processados. Exemplos de metadados são a localização, serviço e dimensão da rede que originou o evento de alerta. Todas as informações a serem compartilhadas devem estar de acordo com uma política de compartilhamento e serem enviadas por um canal seguro.

Os recomendadores possibilitam a colaboração de especialistas de segurança na qualificação dos alertas e compartilhamento de novas informações. Foi realizada uma pesquisa por meio de questionário para verificar o interesse e participação de administradores de redes em um sistema de recomendação sobre notificações de segurança (ver Apêndice D). Constatou-se que todos os respondentes (40 especialistas) compartilhariam informações em um sistema de recomendação, com ressalvas apenas para dados sensíveis. Além disso, a maioria tem interesse em notícias de ciberse-

gurança para proteger seus sistemas de forma antecipada.

Os notificadores são implementados respeitando as políticas de privacidade dos parceiros para a divulgação das informações. Em geral, os alertas devem ser encaminhados para os interesses das entidades segundo um perfil predefinido no sistema. Notificações genéricas podem ser enviadas a todos os parceiros e, se de interesse global, divulgadas publicamente.

A colaboração na arquitetura é garantida pela participação de organizações disponibilizando informações de sensores tradicionais, mas principalmente pelos recomendadores. Os recomendadores são os componentes que possibilitam a supervisão e complementação dos alertas obtidos de fontes de dados não estruturados por especialistas de segurança. Sendo assim, o uso de um canal seguro e mecanismos de autenticação são necessários para evitar problemas como conluio ou falsos rumores de ameaças.

A arquitetura que pode ser construída sobre os conceitos do arcabouço EWS, diferencia-se das demais e avança o estado da arte em EWS, por viabilizar a exploração e disseminação de alertas por meio de fontes de dados não estruturados e heterogêneas, por possibilitar o compartilhamento da análise de especialistas por meio de sistemas de recomendação, e por considerar o uso de diferentes componentes de processamento e pré-processamento que podem ser adaptados segundo as definições do próprio arcabouço.

## 3.13 Considerações finais

Este capítulo apresentou o arcabouço para a análise e extração de alertas de cibersegurança, preferencialmente antecipados, em fontes de dados não estruturados. O arcabouço possui quatro principais componentes: coletores, pré-processadores, processadores e notificadores. Todos os componentes possuem suas próprias especializações. Os coletores são componentes construídos para monitorar fontes de dados heterogêneas, logo para cada fonte é necessário um coletor especializado. Os pré-processadores são responsáveis por normalizar, expandir e priorizar as informações coletadas, mas também por diminuir a carga para os processadores por meio de filtros de remoção. Os processadores são responsáveis por analisar e classificar os potenciais alertas, correlacionar com outros alertas e realizar a recomendação de alertas. Uma vez identificado os alertas, esses são armazenados em uma base de alertas e também encaminhados para os notificadores, que são responsáveis de transmitir os alertas para as entidades de interesse (parceiros). Um processo importante do arcabouço é a análise de dados, que consiste em investigar amostras de dados coletados para produzir as bases de inteligência usadas no desenvolvimento de novos algoritmos e fornecimento de informações de inteligência para os pré-processadores e processadores. No final do capítulo, discutiu-se como os conceitos do arcabouço propiciam a construção de uma arquitetura distribuída e colaborativa para EWS.

O Capítulo 4 apresenta dois estudos em diferentes fontes e a avaliação do arcabouço para a análise e extração de alertas cibernéticos nessas fontes. Também são apresentadas a criação das bases de inteligência a partir das informações analisadas nas fontes do estudo. Já o Capítulo 5 apresenta a implementação de um EWS segundo os componentes e fluxos definidos no arcabouço.

# Capítulo 4

# Experimentos e Resultados

Este capítulo apresenta os experimentos e resultados com fontes de dados não estruturados para avaliar a análise e a extração de alertas de segurança, preferencialmente alertas antecipados, no arcabouço EWS. São apresentados experimentos com duas fontes distintas: microblogs e redes IRC. Cada uma dessas fontes apresenta características diferentes quanto a forma de publicação, que são respectivamente, postagens simples e postagens em conversações. Para cada fonte é mostrado como arcabouço é flexibilizado para a realização da análise e extração de alertas. Os elementos do arcabouço são avaliados e comparados com trabalhos relacionados durante a investigação de cada fonte. Também é apresentado a especificação de um modelo de recomendador e de um classificador para alertas de segurança. Por fim, apresenta-se uma discussão geral do arcabouço, procedimentos e resultados dos estudos em cada fonte e dos modelos de recomendação e classificação.

## 4.1 Estudo 1: Microblogs

Nesta seção, são descritos os experimentos e resultados com o microblog Twitter. Avalia-se o arcabouço EWS na análise e extração de alertas de segurança nas mensagens postadas no Twitter. São usadas quatro bases de dados distintas para avaliar diferentes elementos do arcabouço. Parte dos experimentos e resultados já foram publicados em (Campiolo *et al.*, 2013; Santos *et al.*, 2012, 2013). A avaliação foi dividida em subseções que abordam diferentes conceitos e elementos do arcabouço.

### 4.1.1 Fonte de dados

Este estudo investiga o arcabouço EWS usando como fonte de dados o microblog Twitter. O Twitter provê um serviço de microblog onde usuários (seguidores) acompanham postagens públicas de até 140 caracteres adicionados com outras informações (links, imagens, comentários, emojis) de outros usuários (Kwak *et al.*, 2010; Russell, 2011). Há recursos para a troca de mensagens privadas entre os usuários, mas o principal recurso é a disseminação de mensagens entre todos os seguidores. As mensagens recebidas podem ser replicadas ser replicadas, esta prática é conhecida como *retweet* (Ye e Wu, 2010). Alguns autores consideram que mensagens que sofrem *retweet* tem grande chance de serem importantes (Morris *et al.*, 2012).

Nas mensagens do Twitter, o termo 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*. O Twitter é muito utilizado para disseminação e buscas de informações sobre acontecimentos atuais (Phuvipadawat e Murata, 2010; Russell, 2011). Morris *et al.* (2012) 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.

O Twitter é uma fonte relevante para um EWS devido a publicação de alertas de segurança que tendem serem anunciados assim que detectados e de forma rápida, pois seus usuários colaboram e disseminam informações de segurança (Santos *et al.*, 2012). Além disso, indivíduos e grupos que realizam ciberataques costumam disseminar suas ações em mídias sociais, como o Twitter, para se autopromoverem (Holt *et al.*, 2012).

### 4.1.2 Coletores e bases de dados

A coleta de mensagens do Twitter foi realizada por meio de um software desenvolvido a partir da API Twitter4j[1] que faz uso das APIs públicas REST e Streaming do Twitter[2]. Essas APIs possibilitam o monitoramento de mensagens recentes e o monitoramento em tempo real de *tweets*, respectivamente.

Nos experimentos foram usadas quatro bases de dados, denominadas de bases A, B, C e D. As bases foram armazenadas no formato *JavaScript Object Notation* (JSON), formato padrão retornado pelas APIs do Twitter.

As bases A, B e C contêm *tweets* em inglês e foram coletadas com a API REST. Nas consultas foram usadas expressões com variações de operações lógicas OR e AND com os seguintes termos associados à segurança: security, virus, worm, attack, intrusion, invasion, ddos, hacker, cracker, vulnerability, exploit, patch, malware, zero-day, infosec, trojan, botnet, hijack, backdoor, spyware, security alert. Também foram realizadas consultas com expressões que associavam simultaneamente ameaças a algum dos termos: windows, linux, apple, android, java, adobe, google, amazon, microsoft, firefox, chrome, ie, internet explorer.

A base D contém *tweets* em português e foi coletada com a API Streaming. Nas consultas foram usadas palavras-chave e perfis associados a ações *hackers* e segurança cibernética. Foram monitoradas as mensagens com as seguintes palavras-chave: ataque pagina, ataque alvo, problema rede, ataque servidor, servidor derrubado, rede lenta, internet lenta, ataque site, problema site, target, hacker, ddos, deface, redeface, pwned, fuckgovbr, hacked, owned, antisec, choragov, defaced, sqli, xss, priv8, anon, exposed, dox, doxed, leaked, pwn3d, fuckgov, ownz, cyberwar, rooted, bypass, exploit, tangodown.

A base A contém 11492 *tweets* que foram coletados de 28/04/2012 a 19/05/2012. A base B contém 143172 *tweets* que foram coletados de 28/04/2012 a 05/12/2012. A base C contém 3687720 *tweets* que foram coletados de 01/01/2015 a 31/12/2015. A base D contém 2010 *tweets* que foram coletados de 15/05/2015 a 30/10/2015. A Tabela 4.1 apresenta um resumo sobre as bases e mensagens coletadas.

**Tabela 4.1:** *Bases de dados coletados no Twitter.*

| Bases | Dias de Coleta | Tweets | Usuários | com URL | # | @ |
|---|---|---|---|---|---|---|
| A | 21 | 11492 | 7778 | 10104 | 4218 | 4109 |
| B | 222 | 143172 | 70262 | 121546 | 54127 | 62320 |
| C | 365 | 3687720 | 554430 | 3204035 | 2594748 | 1498577 |
| D | 168 | 2010 | 1589 | 834 | 492 | 773 |

### 4.1.3 Análise de dados e bases de inteligência

Na análise de dados é realizado o processamento de amostras de dados coletados no Twitter para extrair as características usadas na construção de algoritmos e das bases de inteligência. As amostras de dados usadas na análise foram as coletas providas pelas bases A, B, C e D. A análise

[1] http://twitter4j.org
[2] https://dev.twitter.com/

de correlação foi realizada somente sobre a amostra da Base A e a classificação somente na amostra D.

Os normalizadores usados na análise realizaram a resolução das URL curtas para URL longas e os filtros apenas removeram as mensagens mal formadas e de outros idiomas diferentes do padrão de coleta. No entanto, na base D, as mensagens foram normalizadas, filtradas e priorizadas como potenciais alertas.

#### 4.1.3.1 Análise estatística

Na análise estatística são considerados os resultados apresentados na Tabela 4.1. Observa-se nas bases A, B e C que a maioria dos *tweets* possui URLs enquanto na base D, menos da metade. Isto ocorre devido ao tipo e expressões de monitoramento usadas na construção da Base D, que consistiu de monitoramento em tempo real e uso de expressões com perfis de grupos *hackers*, logo acaba-se recuperando muitas mensagens que fazem partes de conversações, as quais não são comuns a presença constante de URLs.

A média diária de *tweets* nas bases A, B e C foram respectivamente 547, 645 e 10103. A base C apresentou um número médio elevado de *tweets* em relação as outras bases devido o aperfeiçoamento das expressões de monitoramento. Logo, pode-se assumir que há muitas mensagens de segurança sendo postadas diariamente e que, com o uso de mais expressões ou perfis associados ao vocabulário de cibersegurança, aumenta-se o número de informações que pode ser recuperado.

A média diária de usuários que postaram mensagens nas bases A, B e C foram respectivamente 370, 316 e 1519. Novamente, a base C apresentou um número médio elevado de usuários que postam mensagens associadas à segurança. A investigação desses usuários, bem como as menções a usuários nas mensagens, pode auxiliar a identificar entusiastas em segurança ou grupos *hackers* que compartilham ameaças a sistemas ou vulnerabilidades de software.

#### 4.1.3.2 Análise de frequência

A análise de frequência foi usada para identificar os termos para o uso futuro em filtros de remoção e priorização. Também foi investigado o uso do vocabulário em mensagens relevantes e irrelevantes.

A Tabela 4.2 apresenta os exemplos de termos frequentes associados à cibersegurança nas mensagens coletadas nas diferentes bases de dados.

**Tabela 4.2:** *Termos frequentes associados à cibersegurança nos tweets.*

| Bases | Termos Frequentes |
|---|---|
| A | malware, attack, hacker, exploit, virus, cyber, infosec, new, site, anti, android, firm, apple |
| B | exploit, ddos, zero(day), vulnerab(ility), trojan, bot(net), bug, hijack, inject, backdoor |
| C | infosec, attack, cybersecurity, malware, spyware, tech, vulnerability, hack(*), bug, exploit, privacy |
| D | hacke(*), ataque, ddos, anonymous, alvo, target, exploit, dados, servidor, exposed, deface, dox |

Os termos apresentados na Base A foram investigados e usados para aprimorar as consultas do monitoramento. Constatou-se, 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. 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* pode ser indicativo do aumento do uso de dispositivos móveis e possíveis ameaças à segurança ao sistema Android. O termo *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.

Os termos frequentes na Base B foram analisados e selecionados para construir duas listas: uma lista branca e uma lista negra. A lista branca contém os termos associados à cibersegurança e a lista negra contém os termos associados a mensagens irrelevantes como alertas. Foram extraídos 300 termos associados à cibersegurança e 120 termos associados a mensagens irrelevantes. O critério adotado para a geração da lista branca foi a seleção de termos frequentes e específicos à área de segurança e redes de computadores e, para a lista negra, os termos com mais de 10 ocorrências e que estavam presentes em mensagens irrelevantes.

A Base C apresenta os termos presentes nas bases A e B para uma coleta de 365 dias, logo verifica-se que mesmo variando o período de coleta e o volume de dados, mantêm-se os termos frequentes. No entanto, realizando uma análise mensal, observou-se termos sazonais, como *venom* e *poodle*. O *venom* denomina uma vulnerabilidade que possibilita explorar o código do dispositivo virtual de disquete e, assim, viabilizar o acesso ao hospedeiro e outras máquinas virtuais (ver CVE-2015-3456). O *poodle* denomina um ataque que usa artifícios para ativar versões obsoletas de protocolos criptográficos na comunicação, como foi o caso para o SSLv3 (ver CVE-2014-3566). Além de termos sazonais de segurança, observou-se termos sazonais em mensagens irrelevantes. Portanto, esses termos podem ser usados para selecionar, priorizar ou remover mensagens em determinados períodos.

Como a Base D contém dados coletados de perfis e de palavras-chave associados a atividades *hackers*, verifica-se referências a grupos *hackers*, como é o caso de *anonymous*, e a presença de termos como *ddos*, *alvo*, *target* e *dox*, que é próprio do vocabulário usado por esses grupos. O termo *deface* está relacionado a uma prática comum de grupos brasileiros, que consiste em desfigurações de páginas de instituições, do governo, entre outros, e a respectiva divulgação nas mídias sociais.

#### 4.1.3.3 Análise de correlação

A análise de correlação foi usada para verificar a existência de mensagens de interesse como alertas antecipados nos *tweets*. Foi realizada a correlação entre *tweets* da Base A e notícias de segurança de sítios especializados, conforme apresentado na Figura 4.1.

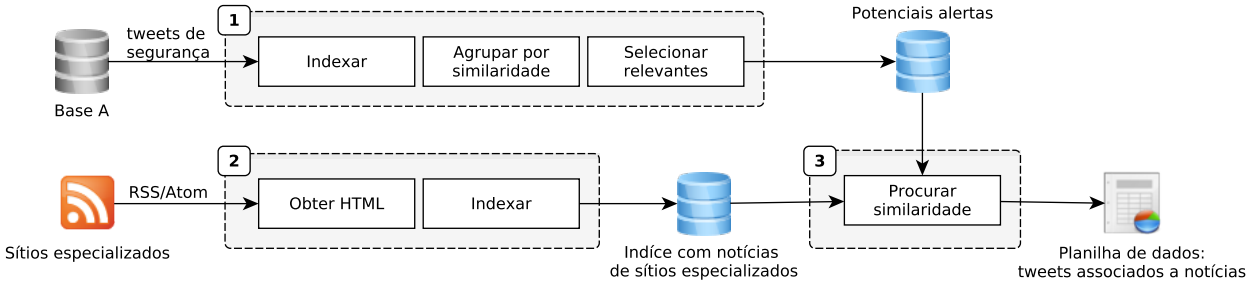


**Figura 4.1:** *Análise de correlação entre tweets e notícias de sítios especializados.*

A coleta de notícias de segurança foi realizada por um software desenvolvido usando a API de manipulação de *feeds* Informa[3]. 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 (Tabela C.1).

Durante um mês, foram coletados 3.988 *feeds* sobre segurança. As informações essenciais para análise são o título, a descrição, a fonte (URL) e a data de publicação da notícia. Inicialmente,

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

pretendia-se encontrar a similaridade entre os *tweets* com o título e descrição dos *feeds*. Entretanto, constatou-se que a quantidade de dados provida pelos *feeds* era insuficiente para obter uma correlação entre ambos.

Para contornar a situação, foi desenvolvido 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*. O conteúdo das páginas associado com a data de publicação foi indexado para a realização de buscas. Esse processo corresponde a Etapa 2 na Figura 4.1.

Por sua vez, na Etapa 1, os *tweets* foram indexados e agrupados usando o Apache Lucene. 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* abordam o mesmo assunto ou similar. Após experimentos, adotou-se empiricamente o valor de similaridade 0,5. Dessa forma, evita-se que assuntos de contextos diferentes sejam agrupados, mas ainda mantém um número significativo de grupos, apesar da geração de grupos distintos que abordam o mesmo assunto.

Para identificar a importância do *tweet* como alerta de segurança, os *tweets* agrupados foram submetidos como termo de consulta para as notícias de segurança indexadas, que corresponde a Etapa 3 na 4.1. Após experimentos, foi adotado empiricamente o valor 0,2 como fronteira entre baixo e alto grau de similaridade. 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.

Como resultados, o processo de agrupamento de *tweets* gerou 278 mensagens de alertas. Dessas mensagens, após a classificação manual, 119 foram identificadas relevantes como alertas. Ao comparar com as notícias, 69 (57%) apresentaram grau de similaridade acima de 0,2. Verificou-se também que 45% dos *tweets* apresentam data mais recente que uma notícia publicada em uma mídia tradicional, logo há mensagens no Twitter propagadas antes ou na mesma data de alguns sítios especializados. No entanto, o Twitter apresenta a vantagem da rápida disseminação de informações. A correlação também possibilitou inferir que há informações de segurança no Twitter e que essas indicam ameaças potenciais e que usuários se preocupam em alertar outros usuários sobre problemas de cibersegurança. Os detalhes dos procedimentos e avaliações são descritos em (Santos *et al.*, 2012).

#### 4.1.3.4 Análise de agrupamento

A análise de agrupamento foi realizada na Base A e consistiu da realização dos passos descritos na Etapa 1 da Figura 4.1.

Os *tweets* foram submetidos a um processo de filtragem para remover conteúdo irrelevante, por exemplo, *tweets* com menos de três palavras. Utilizou-se um filtro simples a fim de identificar a influência de ruído nos dados.

A indexação e o 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* abordam o mesmo assunto ou similar. Um assunto foi considerado relevante quando gerasse um grupo com vários elementos (Morris *et al.* , 2012). Foi usado um grau de similaridade fixo de 0,5 para considerar que dois *tweets* similares.

Esse valor foi definido empiricamente por meio de testes na base e analisando os grupos gerados.

A Tabela 4.3 apresenta uma amostra com a ordenação, a quantidade de elementos no grupo e as mensagens selecionadas para discutir os pontos positivos e limitações dos resultados.

**Tabela 4.3:** *Amostra de grupos de tweets relevantes*

| 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... |

No universo de mensagens do Twitter escritas 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 metodologia de agrupamento das mensagens. Basicamente, o que difere as mensagens são as URLs e as 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 4.3) 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. 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 incorretamente (*Relgious*), fazendo com que o grau de similaridade do grupo tenha diminuído em relação ao primeiro. Mesmo com o erro de escrita, o grupo obteve 195 *tweets* similares (também escritos incorretamente). 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*", essa mensagem possui o mesmo conteúdo das anteriores e só foi resumida. Como os três grupos tratam do mesmo problema podemos somá-los, o que resulta em 555 mensagens.

Observa-se na Tabela 4.3 que há 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 guerra cibernética. 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 24 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. O grupo 90 relata 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.

Para analisar os procedimentos e a qualidade dos resultados, foram selecionadas 60 amostras de grupos e realizada a verificação de um *tweet* representante do grupo com o conteúdo de notícias de sítios especializados, 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 notícias sobre ferramentas de segurança, como antivírus, 7% são informações relacionadas à segurança, mas não de computadores. E a maioria, 62%, são informações relevantes à segurança e poderiam ser utilizadas por muitos administradores como alertas de possíveis ameaças.

A Figura 4.2 mostra a média do tempo de propagação das mensagens apresentadas na Tabela 4.3. Pode-se observar que a maioria das mensagens tem seu ápice no segundo dia a partir de sua publicação (Lerman e Ghosh, 2010). Algumas mensagens ainda são mencionadas brevemente alguns dias depois. O tempo médio de propagação das mensagens é de 12 dias.

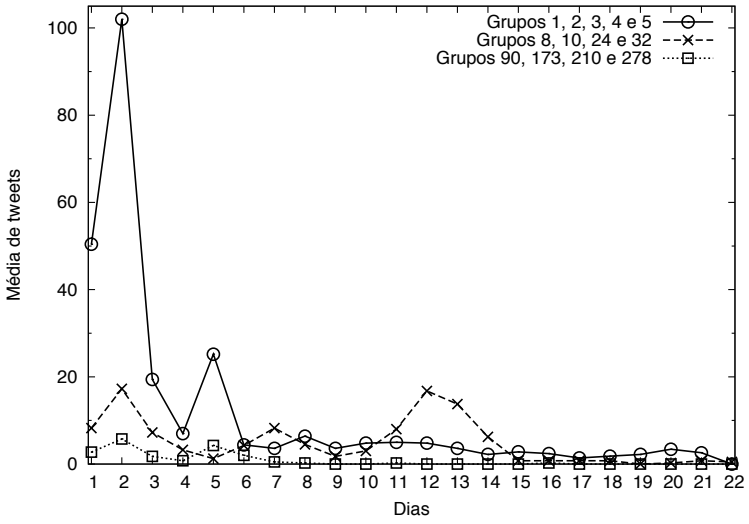


**Figura 4.2:** *Linha de tempo (em dias) da propagação dos tweets.*

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 4.3) 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 (Kwak *et al.*, 2010)). Então, a mensagem com mais menções, postada por 347 usuários do Twitter, pode ter alcançado aproximadamente 340.000 pessoas.

Ao explorar os grupos, verificou-se que mensagens de segurança de computadores são propagadas por muitos usuários, o que acaba por priorizar mensagens que são importantes. No entanto, o mesmo pode acontecer com mensagens irrelevantes como alerta, logo há necessidade de filtros e classificadores para identificar e excluir grupos e mensagens irrelevantes. Neste sentido, a análise de agrupamento possibilitou identificar os tipos de mensagens de segurança de computadores mais propagadas na fonte e o comportamento de propagação dessas mensagens. Assim, é possível definir um intervalo para comparar e agrupar mensagens relevantes como alerta nos pré-processadores e processadores de alertas, e também usar os dados de propagação para filtros de mensagens periódicas, por exemplo, para a remoção de propagandas de antivírus.

#### 4.1.3.5 Classificação

A classificação manual de mensagens foi aplicada a Base D para caracterizar as mensagens que são alertas e as que geram falsos positivos, visto que essa base apresenta um conjunto de dados normalizados, filtrados e priorizados. O resultado da classificação foi usado para identificar padrões de alertas e de falsos positivos, bem como para criar uma base de inteligência para treinamento de algoritmos de classificação supervisionados.

Os alertas foram classificados como notificações de segurança, não necessariamente antecipadas, e não notificações. De 2010 mensagens, 780 foram positivas ou mereciam ser notificadas como potenciais alertas, e 1230 indicavam falsos alertas.

A Tabela 4.4 apresenta uma categorização e mensagens de interesse como alertas. Observa-se orquestração ou intenção de ataque, possíveis vazamentos de informações, comercialização de ataque, rumor de ataque a um sítio Web, identificação de perfis para monitoramento, notícias de segurança (infosec) que podem auxiliar administradores de rede na proteção de suas infraestruturas de TI, anúncios de vulnerabilidades e de ferramentas usadas para realização de ataques, ataques de desfiguração e vazamento de dados recém realizados.

**Tabela 4.4:** *Classificação de notificações de segurança.*

| **Padrões** | **Amostra de Mensagem** |
|---|---|
| orquestração de ataque | ... se for pra deface e melhor inurl:.php *gov.br |
| rumor | ... que estranho a capa da época sumiu do seu blog ... será que houve algum ataque hacker? |
| comercialização de ataque | Galera, to vendendo ataques ddos, quem quiser, so pedir! 3 reais a cada ataque! |
| vazamento | to pra fazer exposed do site [redacted] , só que to com preguiça de arrumar tudo certinho e colocar no pastebin :( |
| oportunismo | Essa é a hora de lançar um link com exploit falando que vazou o tema da redação. |
| perfil para monitorar | É dificil ficar de boa com todos seus amigos sendo blackhat , vc pensa, nao agr vou parar... dai PLAU deface, dai vc para e PLAU exposed |
| infosec | Detectado aumento ataques DDOS por routers e impresoras ... |
| vulnerabilidade | Exploit 0day CMS HB 1.5 http://t.co/vIf76QV99Q #0day #exploit #php #web #pentest #sec #vulner http://t.co/hkugXsD33K" |
| desfiguração | #pwned #4fun nem sei o eh isso, mas ta owned, eh noix [redacted] |
| ferramentas de invasão | Tool lfiINURL - exploring LFI http://t.co/QKRKTNVCDh #php #exploit #hacking #infosec#websec #lfi #web #security #tool #linux #inurlbr |

A Tabela 4.5 apresenta uma categorização e mensagens que podem auxiliar a identificar falsos positivos no EWS. Por essas mensagens possuírem termos comuns usados em notificações de alertas cibernéticos, como *ataque*, *hacker*, *ddos* e *vazou*, podem acabar sendo categorizadas como alertas. São comuns mensagens que abordam outros ataques (p. ex. ataque do coração, de futebol, contra a dengue), uso em ofensas ou brincadeiras, palavras dependentes de contexto (p. ex. target e anon), ataques a redes sociais, entre outros.

A classificação da Base D foi usada a implementação dos componentes de filtros e classificação de alertas notificados em língua portuguesa.

#### 4.1.3.6 Heurísticas

Identificar características que possibilitem evidenciar mensagens como potenciais alertas é importante para a implementação de filtros, classificadores e analisadores. Nesta seção, é mostrado como o uso de heurísticas auxiliou na geração de bases de inteligência e em algoritmos no contexto de EWS.

A Base B foi analisada as seguintes dimensões: URLs, tamanho da mensagem, número de palavras, número de *hashtags* e número de menções a outros usuários.

A análise de URLs consistiu em expandir e agrupar as URLs curtas presentes nos *tweets* e as mais

**Tabela 4.5:** *Classificação de falsos positivos em notificações de alertas.*

| **Padrões** | **Amostra de Mensagem** |
|---|---|
| série TV | @HELOVEBRASIL com o vingador, ai salva a hacker, e o cara vai preso, o robs sai ... |
| besteira | Eu sou o hacker (hacker), hacker (hacker), hacker do amor pra você surfar na net e ... |
| outros ataques | Petrobras continua alvo de ataque especulativo do mercado para favorecer capital externo, ... |
| usar hacker | Quando não é Russo [redacted] usando hacker, é um sem mãe [redacted] ... |
| cantada | @reportermilgrau dessa vez o unico hacker foi você que hackeou meu coração |
| pergunta | Twitch tá sofrendo ataque de DDoS? |
| propaganda | Ninguém,está livre de contaminar sua máquina com algum vírus e do ataque Hacker. Ferramentas de Segurança de Redes.: http://t.co/6jurbFlOch |
| jogos online | wall dos hacker da rodada abre o olho esses 2 aqui tão de hacker.....modo pistola ... |
| ofensa | Vou dar Ddos no [redacted],ddos no [redacted],ah meu nome Pikachu |
| falsos hackers | Hacker? miga pf hacker invade sistema do governo da nasa ou slá, vc n é hacker vc é só uma otária que trocou a senha no tt da amiguinha bjs |
| música | ...se vazarem ela completa a Britney tem fazer com esse hacker [redacted] igual ... vazou o RH |
| pseudo ameaça | @Springlees @pedrost_ mas tu sabe que o pedrinho é hacker ddos 1337 vai conseguir essa fotinha |
| nunca hackeou | Paga de hacker mais nunca hackeou um LINUX u-u |
| esporte | Não sou lá muito fã do Barcelona mas esse trio de ataque é hacker |
| acidente | Dia finalizado terminei minha aplicação de serviço de background, ataque DDos sem querer... |
| agradecimento | queria deixar aqui o meu muito obrigado ao hacker que vazou The Witcher III |
| comentário | Hoje pode se comprar serviço de ataque DDoS por cinco, dez dólares :-o |
| anon | ai que bom nao fui a unica atacada pelo anon brabao h... ? ... |
| estudo | JB Filho já emulou o ataque hacker? |
| hacker celular | Meu crush e um hacker filho da mae que hackeou meu celular e leu minhas conversas #askcrush |
| vazamento falso | Juro que foi o hacker que vazou minha [redacted]. http://t.co/A0GbXdufto |
| oficina hacker | Assista ao debate da oficina Ateliê Hacker sobre apropriação de ferramentas para produção ... |
| gírias | Nosso amg [redacted] vazou de ddos @andersonendres @Junior_Andraade |
| desejo hacker | @todokicker SE EU FOSSE HACKER EU JÁ TINHA HACKEADO A ONE DIRECTION ... |
| divulgação grupo | @googleinurl #php #exploit #0day #web #vulnerabilidade #vuln #desenv #cms #inurlbr |
| outros hackers | Hacker cria robô com impressora 3D que consegue destravar cadeados de senha em 5 minutos... |
| favor hacker | ... amiga pegou de volta pq ela é hacker, vc pode tentar pedir p ela |
| facebook hacker | ... Estou reconstruindo meu fc que foi destruído por um hacker. Obrigada (desculpa se já pedi) |
| twitter hacker | eu fui hackeado e a hacker bloqueou mais de 1.000 pessoas so pra mim perder meus seguidores |
| nome jogo | Biker Exploit \| UOL Jogos Online... (http://t.co/tctulP5JQu) |
| evento hacker | 3 Encontrão Hacker começa nesta sexta (15), em Fortaleza http://t.co/NJfJO2B9Q1 |
| hacker social | ou a katie foi hackeada ou não sei deve ter sido mas esse hacker nem pra tweetar meu nome |
| prisão hacker | Bonner disse que um hacker abordado ontem no JN tinha “cara de maluco”. Foi o ... |
| lições ataque | http://t.co/p87rOyxUYw - O que aprender com o ataque hacker ao app do Starbucks ... |
| zoação | HAHAHA Caíram no ataque do hacker! ... |
| rumor falso | ddos ou dox? O polvo disse ddos. #pergunteaopolvo http://t.co/cWr68UcEwF |
| proteção | RT @revista__super: Qual dieta manter para evitar um ataque hacker do Estado Islâmico? |
| pedidos ensinar | Nego ama falar que exposed ddos é coisa de fracassado mais pra ficar pedindo pra ensinar ... |
| ataque coração | Quase ter um ataque do coração, ... recebeu um e-mail de hacker, não tem preço |
| loja Target | RT @HilaryDuffBR: A edição exclusiva do #BIBO da Target já está em pré-venda ... |
| brincadeiras | the 1975 em: como promover sua banda: delete o site oficial... |
| cidade | Conselho de Habitação faz visita a conjuntos residenciais de Ddos \| Prefeitura de Dourados ... |
| invenção ataque | Planned Parenthood inventa ataque em seu site: http://t.co/yPtvOKXchE |
| quase ddos | É sempre assim nas primeiras 2h de venda. O servidor não aguenta, é quase um ataque DDoS. |
| vazamento antigo | @bruvmartins @DayrelGodin Ano passado um hacker vazou fotos de toneladas de famosas nuas, lembra das fotos da J-Law? Foi isso. Esse evento + |
| curso | RT @SecurityTube: Supercharge your XSS demos - go beyond Alert(XSS)! ... |

frequentes foram analisadas manualmente. Verificou-se que muitas apontavam para sítios de notícias tradicionais que não abordavam especificamente cibersegurança, enquanto outras apontavam para sítios que não publicariam alertas antecipado. A partir dessas URLs, foi criada uma base de domínios irrelevantes como fontes de alertas no contexto do EWS. No total foram selecionados 160 domínios irrelevantes (ver amostra em Tabela 4.6).

**Tabela 4.6:** *Amostra de domínios irrelevantes.*

| Ocorrências | Domínio |
|---|---|
| 1117 | washingtonpost.com |
| 484 | forbes.com |
| 428 | amazon.com |
| 405 | computerworld.com |
| 356 | pcworld.com |
| 27 | infowars.com |
| 26 | inquisitr.com |

A análise de tamanho de mensagens consistiu em gerar uma distribuição com intervalo de 10 caracteres e verificar manualmente as mensagens na distribuição. Foram consideradas mensagens de 0 a 60 caracteres. O limite superior de 60 foi definido baseado em (Morzy, 2011) que afirma que o tamanho médio das palavras em inglês são 5.10 letras, logo com 60 caracteres é possível escrever cerca de 10 palavras considerando o espaço entre elas. A Tabela 4.7 apresenta o número de ocorrências e exemplos de mensagens para os intervalos. Verificou-se que entre 10 e 20, 97% eram URLs curtas que precisariam ser expandidas e recuperado texto da página apontada, como o título da página. Entre 20 e 30 o padrão mais comum foi uma palavra e uma URL. Um padrão de interesse foi o uso de termos associados à cibersegurança seguido da URL, como *virus*, *malware*, *cracker*. Entre 30 e 40 o padrão se mantém como o de 20 e 30 mais com uma descrição com uma ou duas palavras a mais para a palavra que acompanha a URL. Em geral, nos mensagens de 0 a 40 caracteres, a quantidade de texto torna difícil dizer se a mensagem é relevante ou não como alerta sem uma investigação do destino apontado pela URL. Nos casos acima de 40 caracteres já temos mensagens bem formadas e com texto que possibilite realizar processamento direto na mensagem.

**Tabela 4.7:** *Análise do tamanho de mensagens dos tweets.*

| Intervalo | Ocorrências | Exemplo |
|---|---|---|
| 0 – 10 | 0 | – |
| 10 – 20 | 141 | http://t.co/uPU5iLig |
| 20 – 30 | 52 | Malware: http://t.co/qNSTvyx3 |
| 30 – 40 | 261 | Skype users alert! http://t.co/3bWhCJpP |
| 40 – 50 | 662 | Adobe hacked by SQL Injection:http://t.co/fo8yLGjb |
| 50 – 60 | 1762 | Successful side channel attack: http://t.co/ydG5C2Nc |

A análise de número de palavras consistiu em gerar uma distribuição de 1 a 5 palavras com intervalo simples. Foram removidos os termos com menos de 3 caracteres das mensagens antes de contabilizar o número de palavras. A Tabela 4.8 apresenta a distribuição de mensagens e exemplos *tweets* por número de palavras. Verificou-se que mensagens com 1 palavra, o padrão foi postagens de URLs únicas. Com 2 palavras, verificou-se o padrão de duas URLs e menção a usuário e URL. Com 3 palavras, verificou-se o uso de duas palavras para descrever a fonte seguido de URL. Para 4 palavras, as mensagens são maiores, mas poucas possibilitam identificar o conteúdo pelo texto. Logo, mensagens com 4 ou menos palavras precisam explorar o conteúdo apontado pela URL para aferir a importância do conteúdo como alerta. Já com 5 ou mais palavras, já é possível apresentar além da URL um sujeito e/ou predicado e uma ação, por exemplo, o *Watch out* que indica a ação de ficar atento, nesse caso ilustrado, ficar atento a um *worm*. Logo, mensagens com 5 ou mais palavras são bem formadas e possibilitam o processamento direto do texto.

A análise de *hashtags* consistiu em investigar se um número excessivo de *hashtags* indica uma mensagem mal formada ou irrelevante. Foram investigadas mensagens com 3, 4, 5 e 6 ou mais *hashtags*. Como resultado, obteve-se respectivamente: 8528, 4801, 2785 e 2544 mensagens. Verificou-se que nas mensagens de segurança as *hashtags* são usadas para descrever o assunto apontado e

**Tabela 4.8:** *Análise do número de palavras dos tweets.*

| Palavras | Ocorrências | Exemplo |
|---|---|---|
| 1 | 142 | http://t.co/uPU5iLig |
| 2 | 129 | http://t.co/kHE3A95p #security. |
| 3 | 174 | #securityalert Metasploit http://t.co/x0nJ7vfu. |
| 4 | 394 | not ddos, people http://t.co/aDOPbUga. |
| 5 | 775 | Watch out for WORM_VOBFUS - http://t.co/1o1wgxLj. |

para indicar outras palavras-chave do texto apontado por URLs. Mesmo em situações com muitas *hashtags*, o texto se apresentava com mais palavras e mantinha o sentido. Logo, tentar explorar as *hashtags* pode ajudar a definir o tópico da notícia quando há URLs. Por outro lado, deve-se tratar mensagens irrelevantes que postam sempre as mesmas *hashtags*.

A análise de menções de usuários consistiu em investigar a se havia uma relação entre o número de menções e a importância da mensagem. Foram investigadas mensagens com 1, 2, 3 e 4 ou mais menções. Como resultados, obteve-se respectivamente 50823, 9902, 1212 e 383 mensagens. Verificou-se que nas mensagens de segurança que a menção, nas maiorias das vezes, indica a retransmissão de mensagens. Observou-se que a indicação de retransmissão direta (RT) ocorreu com os respectivos percentuais: 70%, 84%, 72% e 63%. Logo, em geral, as menções são usadas em mensagens de segurança retransmitidas.

### 4.1.4 Normalizadores e filtros

Nesta seção, são apresentados o desenvolvimento e avaliação dos pré-processadores para o EWS, principalmente para a diminuição de falsos positivos. Avalia-se o uso dos normalizadores e filtros na Base B. A Figura 4.3 descreve os normalizadores e filtros de remoção encadeados usados no processamento de mensagens em inglês.

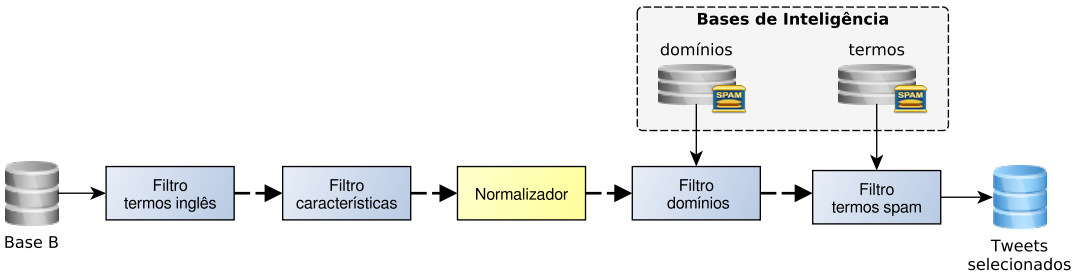


**Figura 4.3:** *Processo de filtragem e normalização de tweets.*

O *Filtro termos inglês* remove mensagens que não são da língua inglesa. Apesar do uso de termos em inglês na coleta, há mensagens em outras línguas que, se não tratadas, acabam gerando carga de trabalho inútil para as próximas fases. Logo, adotou-se a política de remover da base todos os *tweets* com o código de linguagem ou escrita diferente do inglês. A base usada nos experimentos já era uma base normalizada com mensagens em inglês, logo não apresentou redução.

O *Filtro características* remove mensagens com as seguintes características: menos de quatro palavras, tamanho inferior a quarenta caracteres, mais de três URLs, menção a usuários e número de *hashtags* superior a metade do número de termos na mensagem. Essas características foram identificadas na fase de análise e indicam prováveis mensagens irrelevantes como alertas. Foram identificadas e removidas apenas 292 mensagens com alguma dessas características.

O *Filtro domínios* remove as mensagens irrelevantes como alertas considerando a URL presente nas mensagens dos *tweets*. Aproximadamente 84,9% dos *tweets* coletados possuíam URL, pois *tweets* de notificação geralmente contêm referência para uma descrição mais detalhada. Usando a base de domínios irrelevantes gerada pela análise de dados, as mensagens que continham URLs com algum dos domínios foram descartadas. Nesse processo, foram identificadas e removidas 29,2% das

mensagens. Esse filtro também remove mensagens de interesse como alertas, no entanto, são alertas obtidos de sítios de notícias que replicam informações de outras fontes mais especializadas.

O *Filtro termos spam* remove as mensagens considerando a base de termos irrelevantes que foi gerada pela análise de dados. Logo, mensagens que possuam qualquer um desses termos é considerada irrelevante e é removida. Nesse processo, foram identificadas e removidas 29,7% das mensagens. Em uma amostra de 100 *tweets*, 6% abordavam informações de segurança de computadores. Logo, há uma chance de descartar mensagens que poderiam ser de interesse como alerta.

O *Normalizador* executa duas atividades: (i) expande as URLs curtas para longas e (ii) separa os termos em uma lista. Ambas são para facilitar e otimizar desempenho dos filtros.

No geral, verificou-se que filtros e normalizadores são essenciais para diminuir o número de mensagens para a realização do processamento e extração de alertas. Também verificou-se que, dependendo do nível de filtragem, há a possibilidade de descarte mensagens relevantes como alertas.

### 4.1.5 Analisadores

Esta seção apresenta um método para evidenciar notificações de segurança publicadas no microblog Twitter. O método visa descobrir tendências de ameaças e mensagens que identificam potenciais riscos a infraestruturas de redes de computadores e sistemas computacionais. Para tal, emprega-se a filtragem de notícias de segurança, seguida de agrupamento e evidenciação de alertas usando um conjunto de inteligência gerado na fase de análise de dados. No arcabouço, todas as implementações de métodos para investigar um conjunto de dados ou notificações específicas pertencem a classe Analisadores. A Figura 4.4 apresenta a proposta atualizada do método que já foi publicado em (Santos *et al.*, 2013).

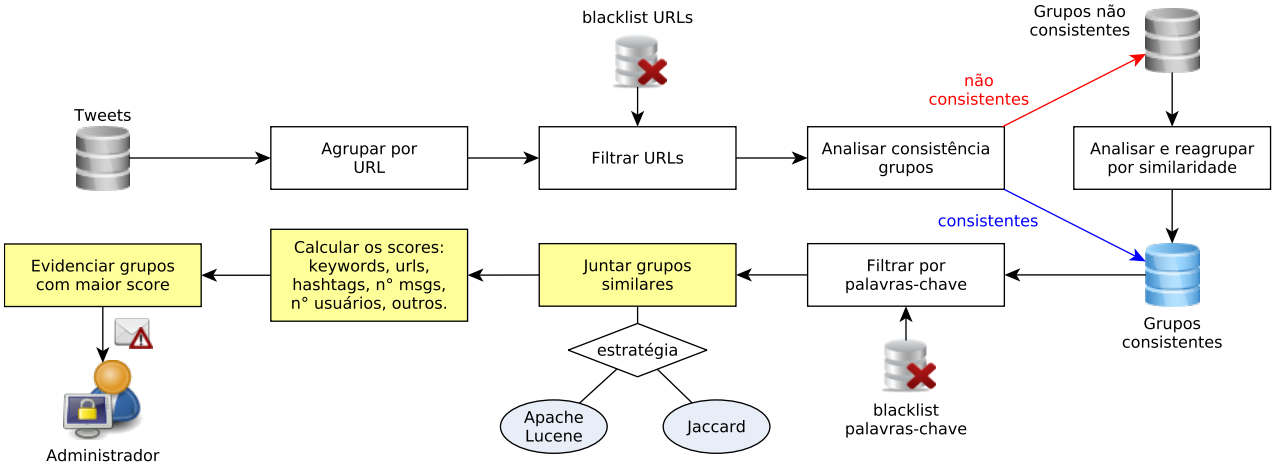


**Figura 4.4:** *Método para extrair e evidenciar alertas postados no Twitter.*

No fluxograma, as caixas brancas representam fases de pré-processamento e as caixas amarelas a implementação do Analisador. São omitidos alguns detalhes, como a normalização das URLs curtas para longas e remoção de *tweets* em outras línguas. A base usada no experimento é a base C, que contém cerca de três milhões e meio de *tweets* coletados no ano de 2015.

No pré-processamento, há uma intercalação de processos de filtros e agrupadores. Inicialmente é realizado um agrupamento direto baseado na URL, isto é, todos os *tweets* que apontam para a mesma URL pertencem ao mesmo grupo. Além disso, o agrupamento considera um intervalo de trinta dias para realizar a junção. Esse intervalo foi definido para possibilitar a análise mensal dos resultados.

Após o processo "Agrupar por URLs", aplica-se o "Filtrar URLs" que consiste em remover os grupos que contém URLs de sítios Web que geralmente obtém notícias de outros sítios especializados

ou são sítios voltados para publicação de outros tipos de notícias de segurança. A "blacklist URL" foi construída usando a base de inteligência "domínios spam".

No processo "Analisar consistência grupos", é realizada uma amostragem de *tweets* para cada grupo e verificada a similaridade entre eles por meio de cálculo de similaridade de cadeias de caracteres. O método usado foi a distância de Jaccard que devolve valores entre 0 e 1 indicando o percentual de similaridade. Foi estabelecido um limiar variável que possibilita indicar se dois grupos são similares ou não. Nos experimentos foi adotado um limiar de 0.2 e uma amostra de 4 elementos do grupo. Caso um par *tweets* de uma amostra não atinja o limiar, o grupo é considerado não consistente e é enviado para o processo "Analisar e reagrupar por similaridade" que é responsável por desagrupar e criar grupos consistentes usando uma abordagem baseada em similaridade de cadeia de caracteres.

Uma vez que os grupos são considerados consistentes, é aplicado o último processo de pré-processamento, o "Filtrar por palavras-chave". Assim, são removidos todos os grupos que a ponderação entre palavras-chave de segurança e não segurança computacional seja negativo. A lista negra (*blacklist*) de palavras-chave foi construída considerando a base de inteligência "termos spam".

No analisador propriamente dito têm-se três processos: "Juntar grupos similares", "Calcular os scores" e "Evidenciar grupos com maior score". O primeiro consiste em aplicar um algoritmo de agrupamento visando juntar os grupos que abordam o mesmo assunto. Há várias estratégias que podem ser empregadas para realizar o agrupamento. Nesta pesquisa, foram avaliadas a estratégia usando busca (via Apache Lucene), isto é, cada *tweet* é usado como termo de busca e os *tweets* retornados que atingirem um limiar são considerados similares (Manning *et al.*, 2008), e a estratégia usando a distância de Jaccard em conjunto com o número de termos (*tokens*) e a data da publicação. O segundo visa calcular pontuações baseados em heurísticas que possibilitem identificar os grupos mais relevantes e descartar as mensagens que não são de interesse e não foram eliminadas nos processos anteriores. O terceiro realiza uma ponderação sobre o que é mais relevante para o administrador, e assim, evidenciar as tendências ou alertas importantes para diminuir o número de notificações a ser exibida.

Na estratégia de agrupamento com o Apache Lucene deve-se especificar um limiar ou grau de similaridade para considerar que os *tweets* pertencem ao mesmo grupo. Inicialmente foi utilizado um grau de similaridade fixo, normalmente 0,5 ou 0,75, mas depois de experimentos preliminares, chegou-se a conclusão que usar um grau de similaridade fixo não produzia bons resultados devido ao tamanho variável do texto dos *tweets*. O problema é que quando é submetida uma mensagem pequena - menos que 70 caracteres, por exemplo - há uma grande chance dessa combinar com algumas palavras de outras mensagens maiores, assim o *tweet* pequeno parece falsamente ter um alto grau de similaridade com a mensagem maior e por consequência é dado como importante, sendo que na maioria das vezes não é importante. Para resolver esse problema, o grau de similaridade é calculado pela Equação 4.1, que permite que o valor seja calculado em função do tamanho do *tweet*.

Na Equação 4.1, $\delta$ representa o grau de similaridade mínima aplicada a uma mensagem de tamanho máximo; $x$ é o número de caracteres da mensagem que está sendo analisada; $\alpha$ é usado como um fator extra de crescimento para o grau de similaridade.

$$GrauSimilaridadeExigido = ((\delta - \frac{x * \delta}{160}) * \alpha) + \delta \tag{4.1}$$

Quanto menor o texto submetido para a consulta, mais rígidas são as regras de agrupamento aplicadas pela Equação 4.1, pois o grau de similaridade exigido aumenta, o que valoriza a importância da informação e evita que assuntos de contextos diferentes sejam agrupados.

Essa estratégia considera uma premissa básica entre os usuários do Twitter, que é: uma mensagem importante tem a tendência de ser retransmitida entre os usuários do Twitter, ou seja, sofre um *retweet* (Morris *et al.*, 2012). Assim, considerando os *retweets* e as mensagens similares, é con-

cluído se um grupo é importante ou não. No entanto, após experimentos preliminares, foi notado que normalmente mensagens irrelevantes continham um grande número de *tweets* similares e comprometiam os resultados na busca por notificações de segurança. Também foi notado que essas mensagens irrelevantes tinham origem de alguns poucos usuários do Twitter, então a importância do grupo foi redefinida para o número de usuários distintos que postaram as mensagens.

O Algoritmo 1 apresenta o pseudocódigo para a estratégia do Apache Lucene adaptada para a entrada com um pré-agrupamento. Os resultados dessa abordagem foram publicados em (Santos *et al.*, 2013) e não são descritos nesta tese.

**Algoritmo 1:** Agrupamento por busca usando o Apache Lucene

```
   Entrada: Lista grupos de tweets - gruposTweet, Número de usuários - N
   Saída: Grupos tweets importantes e não importantes
1  ApacheLucene.indexa(gruposTweet)
2  α ← 2
3  δ ← 0,75
4  enquanto gruposTweet ≠ ϕ faça
5      tweet ← gruposTweets.proximo().obtemTweet()
6      x ← tweet.textoMensagem.tamanho()
7      grauSimilaridadeExigido ← ((δ − (x∗δ)/160) ∗ α) + δ
8      tweetsLucene ← ApacheLucene.obterTweetsSimilares(tweet.textoMensagem)
9      para cada tweetLucene ← tweetsLucene faça
10         se tweetLucene.score ≥ grauSimilaridadeExigido então
11             tweetsSimilares.adiciona(tweetLucene)
12     numeroUsuarios ← removeTweetsComUsuariosRepetidos(tweetsSimilares).tamanho()
13     se numeroUsuarios ≥ N então
14         gruposImportantes.adiciona(numeroUsuarios, tweet.textoMensagem)
15     senão
16         gruposNaoImportantes.adiciona(numeroUsuarios, tweet.textoMensagem)
17     gruposTweet.remove(tweetsSimilares)
18 retorna gruposImportantes, gruposNaoImportantes
```

Já a estratégia de agrupamento por Jaccard, considera a distância de Jaccard em conjunto com o número de termos e o período de postagem. Antes de realizar a comparação, também são aplicadas heurísticas para simplificar as mensagens, tais como: remover RT, URLs, menções a usuários e *hashtags* que estão no início e final da mensagem. O Algoritmo 2 apresenta o pseudocódigo usado nesta estratégia.

**Algoritmo 2:** Agrupamento por Jaccard e número de termos (*tokens*)

```
   Entrada: Lista grupos de tweets - gruposTweet
   Saída: Grupos de tweets
1  ndias ← 30
2  para i ← 0 até gruposTweet.tamanho() faça
3      msgGroup ← limpaMsg(gruposTweet[i].textoMensagemGrupo)
4      numTokensMsgGroup ← extraiNumeroTokens(msgGroup)
5      dataLimite ← gruposTweet[i].obtemData + ndias
6      para cada grupo ← gruposTweet faça
7          se grupo.obtemData ≤ dataLimite então
8              msg ← limpaMsg(grupo.textoMensagemGrupo)
9              numTokensMsg ← extraiNumeroTokens(msg)
10             menorNumTokens ← minino(numTokensMsgGroup, numTokensMsg)
11             jaccardSim ← calculaSimilaridadeJaccard(msgGroup, msg)
12             se menorNumTokens > 6 então
13                 result ← jaccardSim ≥ 0.4
14             senão
15                 result ← jaccardSim ≥ 0.95 − (menorNumTokens ∗ 0.05)
16             se result então
17                 gruposTweet[i].adicionaGrupo(grupo)
18                 gruposTweet.remove(grupo)
19 retorna gruposTweet
```

Observa-se que no Algoritmo 2 também é considerado o tamanho do texto para a realização do agrupamento. No entanto, diferente do Algoritmo 1, considera-se o número de termos no texto e é realizada a sanitização da mensagem. Logo, quanto menor o número de termos, maior a similaridade mínima exigida. Esses resultados se baseiam no que foi investigado na fase análise de dados do arcabouço EWS. Outra diferença, é a saída devolvida por cada estratégia, visto que no Algoritmo 1 são devolvidas duas listas de grupos e enfatizado a priorização por número de usuários e no Algoritmo 2 é devolvida uma única lista com todos os grupos. Dessa forma, tenta-se priorizar não somente mensagens muito difundidas, mas mensagens com poucas postagens que possam indicar potenciais ameaças ou serem de interesse para um administrador em específico.

O processo "Calcular os scores" é responsável por expressar numericamente características que possibilitem evidenciar ou descartar mensagens no contexto de notificações de segurança. Algumas dessas características foram:

- número de mensagens: identificar os grupos com muitas postagens. É importante para evidenciar o grupo e para comparar com o número de usuários que postaram as mensagens.
- número de usuários: identificar se as mensagens de um grupo foram difundidas por muitos ou poucos usuários. No caso de poucos usuários, foi verificado que geralmente se trata de mensagens irrelevantes.
- número de *hashtags*: identificar a confiabilidade da mensagem, visto que mensagens com muitas *hashtags* são usadas com finalidade de divulgação em relação a notificação. Também são comuns muitas *hashtags* em propagandas de produtos de segurança ou notícias sensionalistas.
- ponderamento de palavras-chave: atribuir pesos positivos a termos que comumente estão associados à segurança computacional (p. ex. zero-day) e pesos negativos a termos que estão associados a outros tipos de notificações (p. ex. arrested). Dessa foram, grupos com poucas mensagens podem ser evidenciados. Além disso, os pesos podem ser definidos pelo interesse de um administrador.
- URLs: diminuir ou aumentar a importância da mensagem considerando a URL postada na mensagem. Por exemplo, notificações com URLs para sítios de compartilhamento de texto puro podem indicar vazamento de dados ou compartilhamento de código malicioso, enquanto notificações com URLs para sítios de notícias genéricas geralmente indicam notificações já publicadas de sítios especializados ou de outros contextos associados à segurança.
- período de propagação: identificar mensagens que apresentam muitos dias de propagação, pois podem indicar que são não relevantes. Em geral, notícias de segurança cibernética tem poucos dias de propagação.
- frequência de palavras: identificar se palavras nas mensagens obtiveram aumento de um dia para outro, pois podem indicar um novo tipo de ameaça.
- ocorrência de palavras raras: identificar o número de palavras raras nas mensagens. No caso de notificações de segurança cibernética, geralmente há uma ou duas palavras raras.

O processo "Evidenciar grupos com maior score" considera os resultados do processo anterior para decidir o que apresentar para o administrador. A combinação dos atributos pode ser realizado por importância ou ponderamento. Para observar tendências de ameaças e descartar mensagens, o número de usuários e o período de propagação devem ser considerados. Para observar ameaças que foram poucas difundidas, o ponderamento de palavras-chave e a URL devem ser considerados.

As Tabelas 4.9 e 4.10 apresentam amostras de alertas e mensagens irrelevantes obtidos com a execução do Analisador em conjunto com o Algoritmo 2.

**Tabela 4.9:** *Amostra de alertas evidenciados no período de Jan/2015 a Dez/2015.*

| Mês | Msgs | Usu | Dias | Kwd | URL | Htag | Texto da mensagem |
|---|---|---|---|---|---|---|---|
| Jan | 5090 | 5090 | 0 | 4 | 0 | 0 | RT @dancohens: Found a backdoor into Windows Operating Systems. How cool - Microsoft gave me credit ... |
| | 574 | 574 | 9 | 5 | 0 | 0 | A #0Day vulnerability in #Adobe Flash Player is being exploited to install malware ... |
| Fev | 879 | 17 | 27 | 1 | 1 | 0 | http://t.co/Xb6SXigEYE Emails: 101 Hashes: 8 E/H: 12.63 Keywords: -0.09 #infosec #dataleak |
| | 528 | 496 | 8 | 3 | 0 | 0 | Serious bug in fully patched #InternetExplorer puts user credentials at risk. http://t.co/53nxILExna |
| Mar | 685 | 675 | 4 | 4 | 0 | 0 | Facebook worm spreads by leveraging cloud services http://t.co/Rr8a0SZ4cF #infosec |
| | 497 | 496 | 4 | 3 | 0 | 0 | http://t.co/Ivq2rFWux2 - Mass infection malware attack targets @Android #malware #mobile |
| Abr | 889 | 466 | 6 | 5 | 0 | 0 | RT TimeWaster: Critical HTTPS bug may open 25,000 iOS apps to eavesdropping attacks ... |
| | 547 | 507 | 7 | 5 | 0 | 0 | Wi-Fi software security bug could leave Android, Windows, Linux open to attack: ... |
| Mai | 836 | 817 | 11 | 4 | 0 | 0 | HTTPS-crippling attack threatens tens of thousands of Web and mail servers https://t.co/CWu7SmTqxw |
| | 346 | 277 | 16 | 3 | 0 | 0 | 'Venom' vulnerability: Serious computer bug shatters cloud security #smallbizIT http://t.co/BN1JY8u169. |
| Jun | 1666 | 1597 | 21 | 3 | 0 | 0 | New exploit leaves most Macs vulnerable to permanent backdooring ... #netsec #devops #security ... |
| | 463 | 463 | 11 | 3 | 0 | 0 | RT nmap: Warning: sourceforge is STILL distributing spyware from their fake Nmap page: ... |
| Jul | 870 | 757 | 24 | 4 | 0 | 0 | Second Flash Player zero-day exploit found in Hacking Team's data - BITSPY SOLUTION By ... |
| | 718 | 718 | 11 | 3 | 0 | 0 | 30K MongoDB instances are accessible from the Internet without authentication, containing 595Tb of data ... |
| Ago | 937 | 913 | 18 | 4 | 0 | 0 | 0-day bug in fully patched #OSX comes under active exploit to hijack Macs http://t.co/S5ge13OjTU |
| | 710 | 695 | 23 | 5 | 0 | 0 | Zero-Day Attack on Firefox Users Stole Password and Key Data https://t.co/3lxGKllT6Q #Z3r0 |
| Set | 761 | 622 | 12 | 2 | 0 | 0 | Active malware campaign uses thousands of WordPress sites to infect visitors: 15-day-old campaig... |
| | 671 | 643 | 10 | 4 | 0 | 0 | #FireEye reports attackers install highly stealthy backdoors in #Cisco routers. Not found in US yet ... |
| Out | 777 | 739 | 5 | 3 | 0 | 0 | In a flash: New #0day exploit hits fully patched Adobe Flash http://t.co/LRt9jPpYoW ... |
| | 363 | 348 | 5 | 5 | 0 | 0 | Patch Report: All Versions of #Windows affected by Critical #Vulnerability ... #infosec #cybersecurity |
| Nov | 523 | 509 | 9 | 4 | 0 | 0 | Hackers use anti-adblocking service to deliver nasty malware attack https://t.co/MlsQ14auw9 |
| | 1 | 1 | 0 | 4 | 1 | 0 | Cross Site Scripting (XSS) 0day in SimpleViewer all versions: https://t.co/2k1H6E8CnB #0day #infosec #hacking |
| Dez | 1565 | 1561 | 12 | 3 | 0 | 0 | Woah! Juniper discovers a backdoor to decrypt VPN traffic (and remote admin) has been inserted into ... |
| | 182 | 175 | 6 | 4 | 0 | 0 | All Windows users should patch these critical security flaws \| ZDNet https://t.co/fCJhV0dfg0 |

* Msg: mensagens; Usu: usuários; Dias: propagação; Kwd: palavras(peso); URL: *links*(peso); Htag: *hashtags*(peso).

Observa-se na Tabela 4.9 que, em geral, os alertas apresentam uma relação baixa entre número de mensagens e de usuários, o número de dias de propagação inferior a 20 dias, o ponderamento das palavras-chave acima de 3 e ponderamento de URL e *hashtags* igual ou superior a 0. A amostra de Fev/2015 contém uma propagação de 27 dias, ponderamento de palavra 1 e contém poucos usuários para um elevado número de usuários, o que indicaria uma provável mensagem irrelevante. No entanto, esse grupo contém peso 1 na URL, o que indica que está na lista branca, ou seja, pode indicar vazamento de informações. O mesmo acontece com uma amostra de Nov/2015, que possui 1 único usuário, mas recebeu ponderamento alto para palavras e URL.

Já na Tabela 4.10, observa-se que, em geral, que muitas mensagens irrelevantes apresentam uma relação alta entre número de mensagens e de usuários. Logo, a possibilidade de ser um "spam" aumenta, visto que poucos usuários retransmitiram a mesma mensagem muitas vezes. Como exemplo, a segunda mensagem da amostra de Fev/2015 que faz a propaganda de um software para remoção de *spyware*. Outra característica comum em mensagens irrelevantes é o número de dias de propaga-

ção, visto que se tratam de mensagens sobre política, economia, guerras, entretenimento e acabam sendo retransmitidas por usuários diversas vezes. O ponderamento das palavras também resulta em um valor inferior ou igual a dois e, nos eventos que é alto, acabam tendo outras propriedades negativas, como é o caso de muitas *hashtags*, que indicam notícias sencionalistas ou propagandas.

**Tabela 4.10:** *Amostra de mensagens irrelevantes no período de Jan/2015 a Dez/2015.*

| Mês | Msgs | Usu | Dias | Kwd | URL | Htag | Texto da mensagem |
|---|---|---|---|---|---|---|---|
| Jan | 667 | 658 | 25 | 0 | 0 | 0 | Two 'Lizard Squad' Hackers Arrested After Christmas DDoS Attacks: A 22-year-old man linked to the ... |
| | 1003 | 892 | 24 | 2 | 0 | 0 | On Day One, the new Congress launches an attack on Social Security: submitted by nirad ... |
| Fev | 645 | 73 | 3 | 1 | 0 | 0 | 'Tehran: Vendors were attack and their properties looted by the Security and City agents ...' |
| | 393 | 1 | 27 | 1 | 0 | 0 | SuperAntiSpyware Remove spyware, Not just the easy ones! Over 45 Million Downloads Worldwide ... |
| Mar | 401 | 26 | 30 | 5 | 0 | -1 | The Long List of Password Breaches ...URL #infosec #security #encryption #privacy #password #TrueCrypt |
| | 490 | 71 | 30 | 1 | 0 | 0 | Did Robert Thurman solicit cyber crime? http://... #LamaGate #tibetbenefit2015 #DalaiLama ... |
| Abr | 4305 | 56 | 29 | 1 | 0 | 0 | ... passwordrandom: #TweetMovie Watch #Infosec #Malware unfold at http://t.co/W4gIPxkrFx ... |
| | 988 | 21 | 29 | 2 | 0 | -1 | RT ... Easily Memorize Strong Passwords URL ... #infosec #security #password #memorization #musclememor... |
| Mai | 208 | 168 | 30 | 0 | 0 | 0 | ... #weathers Noadware - spyware/adware #remover: http://t.co/d6ArOhN8cV Promote The Top Anti-spyw ... |
| | 829 | 21 | 29 | 4 | 0 | -1 | Tutorial: Easy Encryption with TrueCrypt ... #infosec #security #encryption #privacy #password #truecrypt |
| Jun | 477 | 142 | 20 | 2 | 0 | 0 | Deals Today ... Computer/Laptop Network Firewall Security Anti-Malware Appliance ... |
| | 411 | 13 | 29 | 1 | 0 | -1 | Why #password #manager are free? by philanthropy? #Password #infosec #authentication ... |
| Jul | 1624 | 66 | 25 | 3 | 0 | -1 | Legislative Briefs on #Cybersecurity and #Critical #Infrastructure. #IoT #Cyberinsurance #infosec ... |
| | 1000 | 44 | 30 | 1 | 0 | -1 | #ICIT Fellow Insights: How #Legislation can Minimize #Cyber Impact. #infosec #Congress ... |
| Ago | 418 | 309 | 27 | 0 | 0 | 0 | RT NorseCorp: Bypassing Antivirus with Shellter 4.0 on Kali Linux http://t.co/5Ll4vcl0Cb... |
| | 298 | 25 | 30 | 1 | 0 | 0 | #infosec From The Cyber Law Library: WikiLeaks files suggest US spied on Japan, Japanese companies ... |
| Set | 540 | 508 | 7 | 1 | 0 | 0 | Germany Blocks All Trains from Austria to Stem Migrant Overflow: Germany has halted train traffic from Austri... |
| | 268 | 1 | 29 | 2 | 0 | 0 | Best Cell Phone Mobile Spy Software Remotely Read SMS, Check Call ... remote spyware monitoring |
| Out | 159 | 32 | 6 | 3 | 0 | 0 | Adware Cleaner - Remove Adware, Spyware, and Restore Your... http://t.co/9SRLkJTmRp... |
| | 45 | 39 | 30 | 1 | 0 | 0 | Don't get blacklisted by Google - - find security threats before hackers do. http://t.co/8Tfz4s3BeF |
| Nov | 974 | 906 | 7 | 1 | 0 | 0 | Pope Francis: Paris attacks were part of a 'piecemeal Third World War' ... |
| | 599 | 534 | 18 | 1 | 0 | 0 | TERROR ALERT: 3,000 Islamist extremists in Britain 'ready to ATTACK the UK in weeks' ... |
| Dez | 1506 | 1353 | 27 | 1 | 0 | 0 | Obama: 'America Is Safe from Islamic State' Before Attack HIS TIMING IS AWESOME https://t.co/xTEauDr4Tr |
| | 163 | 158 | 0 | 0 | 0 | 0 | NY: Man Arrested for Plotting New Year's Islamic State Attack https://t.co/NkSTtqoybe |

* Msg: mensagens; Usu: usuários; Dias: propagação; Kwd: palavras(peso); URL: *links*(peso); Htag: *hashtags*(peso).

### 4.1.6 Trabalhos relacionados

No Twitter, as mensagens sobre eventos são propagadas em tempo real e podem atingir uma alta taxa de disseminação em um curto período de tempo (Kwak *et al.*, 2010; Lerman e Ghosh, 2010; Ye e Wu, 2010). Essas características são úteis para identificar eventos associados à cibersegurança antes da divulgação em mídias especializadas ou identificar tendências de ameaças de forma mais rápida. Nesta seção, são analisados e comparados trabalhos que investigam o Twitter como fonte de informação para a geração de alertas antecipados.

Sakaki *et al.* (2010) utilizaram mensagens postadas no Twitter para a predição e aviso antecipado sobre terremotos no Japão. A abordagem consistiu em monitorar postagens usando palavras-chaves no contexto de terremotos e um classificador que considera os termos, as palavras-chaves, o número de palavras, e as palavras antes e após a palavra-chave. Desenvolveram um modelo temporal e espacial para verificar a confiabilidade e localidade do evento respectivamente, para só então, propagar a notificação. Como resultados, detectaram 96% de terremotos iguais ou superiores a escala 3, e que o tempo de entrega das notificações foi muito mais rápido (no mínimo seis vezes) que a difusão da Agência Meteorológica Japonesa. Da mesma forma, monitoramos palavras-chave associadas à cibersegurança no Twitter, mas usamos métodos não supervisionados e baseados em heurísticas para a classificação.

Al-Qasem *et al.* (2013) desenvolveram uma abordagem para gerar alertas em tempo real sobre a propagação de códigos maliciosos. A abordagem consistia na coleta, filtragem e detecção na variação do número de postagens para indicar um evento significante. Realizaram o monitoramento de três palavras-chave: malware, backdoor e cyber attack. Usaram um filtro para selecionar as postagens que continham algum dos seguintes termos: computer security, new, discover, hit, infect, warn e watch out. Para identificar um evento, observaram o aumento do número de postagens considerando a variância móvel exponencialmente ponderada. Realizaram a avaliação em uma coletânea de cinco dias e encontraram dois eventos de códigos maliciosos. Nosso trabalho foi mais amplo, pois além de um conjunto maior de palavras-chave e tempo de monitoramento, não estabelecemos um limiar rígido para gerar a alerta, visto que o quanto antes uma ameaça fosse identificada, tão logo deveria ser evidenciado o alerta.

Avvenuti *et al.* (2014) propuseram um arcabouço para um sistema de alerta antecipado voltado à detecção de terremotos na Itália que usa o Twitter como fonte de informação. O arcabouço consistia das seguintes fases: aquisição de dados, filtragem, detecção de evento, estimativa de danos e alerta antecipado. Foram usadas duas palavras-chave para monitoramento e, após a filtragem, 1412 *tweets* foram classificados manualmente. No Weka, foram testados algoritmos de classificação, e o J48 obteve os melhores resultados com a validação cruzada de 10-fold. A detecção do evento consistiu, além da classificação, de análise temporal e espacial. Na análise temporal foi considerado o crescimento das mensagens em um período de tempo. Como resultado, o atraso médio das notificações foi de 1 minuto contra 15 minutos do órgão oficial. Comparado a nossa proposta, nós realizamos uma análise detalhada das mensagens antes de realizar o processo de extração. Também consideramos a análise de propagação para identificar a importância da mensagem como notificação.

Kostkova *et al.* (2014) demonstraram como o Twitter pode ser usado para detecção e alerta antecipado de epidemias antes dos sistemas de vigilância dos órgãos oficiais. Selecionaram um subconjunto de 25K *twees* de 3M que foram obtidos com o termo "flu" (gripe) durante aproximadamente 7 meses e meio no ano de 2009. Agruparam os *tweets* que continham referências que o próprio usuário estava com gripe. Em seguida, cruzaram com informações dos órgãos oficiais. Os picos de postagem se relacionavam com o aumento do número de consultas. Por meio de correlação dos dados do Twitter e órgãos oficiais, verificou-se que antecipava a identificação de epidemias de uma a três semanas. Apesar do escopo de nosso trabalho ser diferente, também consideramos o aumento de notificações sobre um evento para mostrar a relevância como alerta. No entanto, para evitar que fossem evidenciadas mensagens fora de contexto, implementamos diferentes tipos de filtros.

Ritter *et al.* (2015) desenvolveram uma abordagem fracamente supervisionada para classificar eventos no Twitter. Realizaram o monitoramento de três tipos de eventos de cibersegurança: ataques DDoS, vazamentos de dados e sequestros de contas. A estratégia proposta consiste em identificar as entidades e a data de ocorrência do evento, agrupar os eventos similares considerando essas entidades e as palavras-chave, e extrair características baseadas nas entidades, posição das palavras-chave, categoria gramatical das palavras e contexto da mensagem. Esse conjunto de dados é usado para treinamento de algoritmos de classificação supervisionados. Como resultado, encontraram eventos de segurança que são de interesse como alertas de segurança. Em nossa pesquisa já havíamos realizado essa constatação (Campiolo *et al.*, 2013; Santos *et al.*, 2012) e também implementamos um método

que identifica a relevância dos alertas baseado em um conjunto de constatações sobre notificações de cibersegurança e mensagens irrelevantes.

A principal inovação em relação aos trabalhos relacionados é o pioneirismo no uso do Twitter para a extração de alertas na área de Cibersegurança. Para tal, foi conduzida uma investigação dos tipos de mensagens que poderiam representar alertas e não alertas de segurança, além da identificação de características para a construção de heurísticas e técnicas de aprendizagem de máquina e recuperação da informação que atendessem as especificidades dessa área.

### 4.1.7 Discussão e síntese dos resultados

O estudo do microblog Twitter envolveu a análise e processamento de quatro bases de dados. Em todas as bases foram identificados notificações de interesse como alerta antecipado, em especial, na base D, que trata de alertas na língua portuguesa. Esses resultados podem ser observados nas Tabelas 4.3, 4.4 e 4.9.

As análises realizadas no processo de Análise de Dados possibilitou identificar o vocabulário associado e não associado à cibersegurança; comprovar a relevância do Twitter como fonte de informação para alertas de cibersegurança e a construção de bases de inteligência que foram usadas nas fases de pré-processamento e processamento. Também possibilitou a elaboração de listas negras e brancas de URLs e termos, que foram importantes para reduzir o número de tweets para o processamento de potenciais alertas. Esses resultados são apresentados na Seção 4.1.3.

Há limitações quanto ao processo de análise devido a definição do que é um alerta de segurança, pois, dependendo do ponto de vista, uma mensagem pode não ser considerada um alerta de interesse. Apesar dessa questão, nesta tese, na análise não foi considerado o público alvo de uma notificação de segurança, logo pode ser um administrador de redes ou um usuário de um dispositivo móvel.

No estudo do Twitter, foi usado a normalização apenas para remover caracteres indesejáveis e para a tradução das URLs curtas para longas. Além de padronizar as mensagens, a normalização expande informações para o uso nas fases seguintes. No arcabouço EWS é proposto o uso de PLN para o enriquecimento dos dados na normalização, mas não foi aplicado a esse estudo.

Os filtros reduzem significantemente o número de *tweets* para a fase de processamento como pode ser observado na Seção 4.1.4. Em ambientes de produção, seria interessante que os filtros fossem adaptativos, ou seja, pudessem usar termos ou entidades identificadas em mensagens irrelevantes para realizar a remoção.

Os analisadores possibilitaram destacar tendências de ataques e remover mensagens irrelevantes não filtradas anteriormente. O uso de informações da Análise de Dados possibilitou a implementação de heurísticas para extrair características que possibilitem categorizar os grupos como alertas e não alertas. Esses resultados são apresentados nas Tabelas 4.9 e 4.10.

Quanto a detecção de alertas antecipados, os melhores resultados são na base D (alertas em português). Nas bases A, B e C, foram identificadas ameaças emergentes e novas vulnerabilidades, que se notificadas a administradores de redes, ajudariam a tomar medidas preventivas mais rapidamente.

Resumindo, foi avaliada e constatada a importância do Twitter como uma fonte de informação importante para um sistema de alerta antecipado voltado à cibersegurança. Além disso, a partir da seleção de processos propostos no arcabouço EWS, foram geradas bases de inteligência e heurísticas que possibilitam extrair e evidenciar alertas em *tweets*. Também foi desenvolvida uma abordagem não supervisionada para a identificação notificações de interesse para alertas antecipados sobre ameaças ou tendências de ameaças.

## 4.2 Estudo 2: Redes IRC

Nesta seção, são descritos os experimentos e resultados com as redes *Internet Relay Chat* (IRC). Avalia-se o arcabouço EWS na análise e extração de alertas de segurança nas mensagens postadas nos canais abertos de redes IRC. São analisados dois tipos de canais: canais usados para discutir sobre segurança de computadores e canais usados por indivíduos ou grupos hackers para compartilhar informações de alvos, ferramentas de ataque, vulnerabilidades e/ou alvos comprometidos. São usadas duas bases de dados para avaliar diferentes elementos do arcabouço. Os experimentos e resultados foram publicados em (Campiolo e Batista, 2015). A avaliação foi dividida em subseções que abordam diferentes conceitos e elementos do arcabouço.

### 4.2.1 Fonte de dados

A rede IRC (Oikarinen e Reed, 1993) foi uma rede social usada intensivamente por inúmeros usuários para a troca de mensagens instantâneas até meados de 2000. Fundamenta-se no protocolo IRC que é responsável pela comunicação privada ou em grupos usando mensagens textuais para controle e dados (Oikarinen e Reed, 1993). A arquitetura é baseada no modelo cliente-servidor usando o protocolo TCP/IP. Uma rede IRC é composta por um único ou vários servidores conectados entre si (Kalt, 2000a). Os clientes IRC conectam-se aos servidores e associam-se a canais (Kalt , 2000c). Os canais agregam usuários e são identificados com um nome para descrever um tópico específico (Kalt, 2000b). Qualquer tipo de conteúdo pode ser transferido na rede IRC, no entanto, os mais comuns são mensagens e arquivos entre os usuários.

Devido ao surgimento de outras redes, o número de usuários nas redes IRC diminuiu consideravelmente. No entanto, por possuir uma estrutura de rede descentralizada, a rede IRC acaba por ser usada para a realização de atividades ilegais, já que a arquitetura da rede dificulta o monitoramento do conteúdo por especialistas em segurança. Há diversos casos em que a rede é usada para controle de *botnets* e troca de conteúdos maliciosos (códigos de exploração, vulnerabilidades e organização de ataques) (Michels, 2012). Logo, monitorar e identificar essas ameaças é importante para a geração de alertas antecipados e para a mitigação de ações maliciosas antecipadamente, já que mesmo antes de um ataque começar, ele pode ser discutido inicialmente no IRC.

Na rede IRC também há canais que abordam aspectos de segurança de aplicativos, sistemas operacionais e redes de computadores. Nestes canais é comum a participação de especialistas ou entusiastas em segurança, administradores de redes, entre outros, que colaboram e trocam conhecimento entre si. O monitoramento destes canais é interessante para identificar possíveis vulnerabilidades, códigos de exploração e notícias sobre segurança da informação. Apesar da diminuição de usuários, há canais com centenas de usuários, como é o caso do canal *security* no servidor freenode[4].

### 4.2.2 Coletores e base de dados

A coleta de dados consistiu em capturar todas as mensagens postadas nos canais monitorados e armazenar em um formato padrão na base de dados. Os canais monitorados foram selecionados a partir da popularidade do servidor IRC e pela quantidade de usuários. A seleção ocorreu por pesquisa Web e observações preliminares em servidores de IRC. Os canais com quantidade pequena de usuários foram descartados devido à facilidade na identificação do software de coleta, o que provavelmente levaria a um banimento e à falha no monitoramento. Os canais monitorados foram de dois servidores de IRC: freenode (irc.freenode.org) e anonops (irc.anonops.com). No freenode, foram monitorados os canais com tópicos relacionados à segurança: security, networking, owasp e oss-security. No anonops, foram monitorados os canais relacionados a atividades suspeitas: anonops, hackers, ddos, opnewblood, defacement e opferguson.

[4] irc.freenode.org

O monitoramento foi realizado a partir de duas redes distintas para dificultar a identificação do software de coleta. Os resultados de cada canal foram armazenados em arquivo no formato *Comma Separated Values* (CSV) contendo as informações: horário, operação, identificação do usuário e mensagem. As operações monitoradas foram JOIN (usuário entra no canal), LEAVE (usuário deixa o canal) e MESSAGE (mensagem postada publicamente no canal). As mensagens privadas não são possíveis de monitorar, exceto as enviadas diretamente ao software de coleta.

O software de coleta foi desenvolvido em Ruby 1.9.1 usando a biblioteca cinch[5]. O software possibilita a configuração dos servidores IRC e canais a serem monitorados. Também foi desenvolvido um software para coleta em Java usando a biblioteca pircbotx[6]. No entanto, devido a identificação do robô (*bot*) em alguns servidores, acabou sendo usado apenas para monitorar os usuários em canais específicos.

Os dados coletados são de diferentes períodos e resultaram em duas bases: Base A e Base B. A Base A foi coletada no período de novembro/2014 a dezembro/2014 durante duas semanas. Nessas duas semanas foram selecionados de 8 a 12 dias com dados completos para a análise. A Base B foi coletada no período de março/2015 a maio/2015, com a finalidade de avaliação dos componentes do arcabouço EWS. Nesses meses foram selecionados 60 dias de coleta que se apresentava completa. Durante o período mais longo, o monitoramento foi suspenso temporariamente (10 dias) devido à identificação do software de coleta e, por consequência, ao bloqueio de acesso ao servidor.

### 4.2.3 Análise de dados e bases de inteligência

Na análise de dados é realizada a caracterização e análise dos dados coletados, principalmente usando os dados da base menor, visando extrair e priorizar informações de interesse como alertas em redes IRC. As mensagens foram investigadas por análise ....

#### 4.2.3.1 Análise estatística

A Tabela 4.11 apresenta a estatística descritiva dos dados coletados especificamente para a análise, isto é, os resultados da média de 10 dias de coleta. Os canais *oss-security* e *owasp* não foram considerados por contabilizarem apenas 17 mensagens no período e os canais *defacement* e *hackers* foram monitorados somente na constituição da Base B.

**Tabela 4.11:** *Caracterização da coleta de dados no IRC (Base A)*

| Canais | Período (dias) | Total de mensagens | Média diária de mensagens | Desvio padrão (mensagens) | Média diária de usuários ativos | Desvio padrão (usuários) |
|---|---|---|---|---|---|---|
| anonops | 12 | 39921 | 3326,75 | 740,44 | 162,25 | 17,09 |
| ddos | 12 | 5890 | 490,83 | 230,67 | 50,33 | 12,09 |
| opferguson | 8 | 28225 | 3525,13 | 2295,01 | 143,13 | 82,24 |
| opnewblood | 12 | 14578 | 1224,83 | 409,54 | 100,42 | 14,18 |
| security | 8 | 11352 | 1419,00 | 528,86 | 95,25 | 13,73 |
| networking | 8 | 22452 | 2806,50 | 495,82 | 112,00 | 10,90 |

Na Tabela 4.11, observa-se que o maior desvio padrão proporcional ao número médio de mensagens e usuários está associado ao canal *opferguson*. Isso acontece pelo fato do canal estar associado a um conjunto de ações hackers em protesto ao assassinato de um jovem negro desarmado por policial em Ferguson (EUA)[7]. Logo, um assunto não apenas de interesse de um grupo menor, mas relacionado à sociedade. Em canais do servidor anonops, é comum o uso do prefixo "op" para indicar uma orquestração de ataque visando um alvo ou em favor de causas políticas, sociais e/ou

[5] https://github.com/cinchrb/cinch
[6] http://code.google.com/p/pircbotx/
[7] http://rt.com/usa/179532-anonymous-op-ferguson-missouri/

econômicas. Logo, é comum no IRC encontrar canais que começam com esse prefixo, como é o caso do *opnewblood* voltado à orientação de iniciantes, embora este último seja um caso a parte já que é um canal que permanece ativo no servidor.

Na maioria dos canais monitorados, foi observado um número elevado de mensagens diárias que, em sua grande maioria, são sobre tecnologia ou assuntos do cotidiano dos participantes. No entanto, é interessante notar que mesmo nestes canais, é possível acessar tutoriais e ferramentas para execução de ataques através de um simples comando. Neste caso, destacam-se os canais *ddos* e *opnewblood* que possuem usuários automatizados para indicar ferramentas e tutoriais didáticos para ataques. Já nos canais de segurança/rede, os usuários acabam comentando sobre assuntos relacionados a vulnerabilidades e formas de exploração.

A Tabela 4.11 não apresenta o número total de usuários diários devido à alta flutuação de entrada e saída de usuários nas salas. Em geral, a maior parte dos usuários não interage publicamente, especialmente nos canais de segurança. No canal *security*, o número de usuários ultrapassa 1000, mas a interação ocorre em média entre um décimo dos participantes.

A Tabela 4.12 apresenta a estatística descritiva dos dados coletados para análise e também para a avaliação do arcabouço EWS, isto é, os resultados da média de 60 dias de coleta. Os canais *hackers* e *defacement* começaram a ser monitorados 10 dias após os outros, logo apresentam um número reduzido de dias de coleta. O monitoramento do canal *opferguson* foi descontinuado devido ao número pequeno de usuários e mensagens.

**Tabela 4.12:** *Caracterização da coleta de dados no IRC (Base B)*

| Canais | Período (dias) | Total de mensagens | Média diária de mensagens | Desvio padrão (mensagens) | Média diária de usuários ativos | Desvio padrão (usuários) |
|---|---|---|---|---|---|---|
| anonops | 58 | 253522 | 4405,5 | 1202,22 | 162,05 | 22,47 |
| ddos | 58 | 22695 | 391,29 | 231,7 | 32,74 | 10,51 |
| defacement | 49 | 5350 | 109,18 | 58,51 | 8,35 | 3,51 |
| hackers | 50 | 19911 | 398,22 | 314,2 | 30,64 | 10,58 |
| opnewblood | 58 | 48369 | 833,95 | 322,44 | 73,57 | 15,81 |
| security | 68 | 118360 | 1740,58 | 592,74 | 90,04 | 15,17 |
| networking | 58 | 131310 | 2263,97 | 579,23 | 108,40 | 14,75 |

Na Tabela 4.12, o desvio padrão dos canais *ddos*, *defacement* e *hackers* é proporcionalmente maior devido à existência de usuários automatizados para fornecer ajuda por meio de tutoriais e indicação de ferramentas de ataques. O número de usuários ativos indica o interesse nos canais suspeitos de atividades maliciosas, mas também nos canais de segurança. Por meio de um histograma de frequências relativas, na prática, verificou-se que nesses canais, 20% dos usuários são responsáveis por mais de 70% das mensagens.

Comparando os dados das tabelas 4.11 e 4.12, foi observado que as relações médias entre mensagens e usuários dos canais comuns mantiveram-se uniformes, mesmo considerando um período de coleta maior.

Visando entender o comportamento dos usuários e as mensagens postadas nos canais, foram gerados gráficos para observar a relação entre o período de coleta e as mensagens diárias e entre o período de coleta e o número de usuários ativos. Os dados apresentados nos gráficos são da Base B devido ao número maior de dias consecutivos de coleta. As Figuras 4.5 e 4.6 apresentam os gráficos gerados para o canal *hackers*.

As análises de pico e vale do gráfico possibilitam identificar a tendência de maior número de mensagens e usuários no decorrer da semana e queda nos finais de semanas. Esse comportamento se repete em todos os outros canais. Nos canais de segurança, acreditamos que o motivo é a presença de administradores em horário de trabalho. Nos canais suspeitos, ficamos intrigados por não haver alta atividade nos fins de semana por pensar que seria o período mais propício para ataques a sítios Web devido à ausência dos administradores e à disponibilidade dos atacantes. Nenhuma dessas

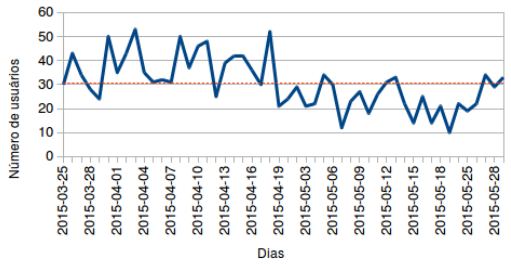


**Figura 4.5:** *Usuários ativos diários.*

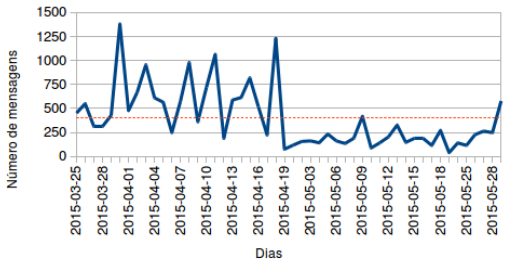


**Figura 4.6:** *Mensagens diárias.*

observações foi explorada nesta pesquisa usando métodos científicos.

#### 4.2.3.2 Análise de frequência

A análise de frequência de palavras teve por objetivo caracterizar o vocabulário usado nos canais. Foram identificados termos para a remoção de mensagens irrelevantes e priorização das mensagens relevantes como alertas.

*Cloud tags* foram geradas para confirmar a suspeita do uso de termos mais informais nos canais suspeitos e de uma linguagem mais formal nos canais de segurança. Nos canais de segurança receberam destaque termos tecnológicos e assuntos relacionados à segurança de redes e computadores. Nos canais de atividades suspeitas receberam destaque ofensas, gírias e acrônimos. Os resultados também auxiliaram na definição de categorias de interesse para os termos. Foram definidas as seguintes categorias:

- acrônimos/gírias: acrônimos e gírias usados na Internet.
- segurança: termos associados à segurança de sistemas.
- atividades maliciosas/ameaças: termos associados às atividades suspeitas e ameaças à segurança de sistemas.
- termos frequentes (hackers): termos frequentes não comuns relacionados aos tópicos nos canais suspeitos.
- termos frequentes (security): termos frequentes não comuns relacionados aos tópicos nos canais de segurança.
- ofensas: termos com caráter ofensivo.

A categorização foi realizada manualmente e considerou os primeiros 3000 termos mais frequentes em cada canal. A Tabela 4.13 apresenta uma visão geral dos termos em cada categoria:

**Tabela 4.13:** *Classificação dos termos em categorias*

| Categoria | Total | Amostra de termos |
|---|---|---|
| acrônimos/gírias | 48 | lol, xd, lmao, ur, pm, wtf, idk, ... |
| segurança | 32 | security, patch, cve, firewall, ssl, pentesting,... |
| atividades maliciosas / ameaças | 124 | ddos, dos, down, bonet, attack, target, injection, ... |
| termos frequentes (hackers) | 401 | off, police, top, site, windows, linux, server, government, ... |
| termos frequentes (segurança) | 436 | new, over, windows, server, linux, ip, problem, nsa, ... |
| ofensas | 35 | shit, idiot, bitch ... |

O acrônimo "pm" significa "private message" e ocorreu algumas vezes em intenções de atividades suspeitas. Os termos "patch" e "cve" ocorreram em discussões de vulnerabilidades recentes. Os

termos “ddos”, “down” e “target” ocorreram em situações de orquestração de ataques e notificação de sucesso de ataque. O termo “ddos” também ocorreu em mensagens irrelevantes, como em solicitações de informações sobre ferramentas. Os termos “police” e “government” estão associados a causas que o grupo *Anonymous* estava atuando na época da coleta. O termo “new” esteve associado a mensagens de novos códigos de exploração ou novas vulnerabilidades. Os termos de ofensas, em geral, estavam associados a mensagens irrelevantes como alertas.

#### 4.2.3.3 Análise de correlação

A correlação de informações de diferentes fontes possibilitou identificar informações relevantes como alertas. No nosso estudo, foi realizada a correlação com dados dos boletins da Microsoft e CVE. A Tabela 4.14 destaca algumas das mensagens relevantes como alertas.

**Tabela 4.14:** *Correlação com outras fontes*

| | Mensagens |
|---|---|
| 1 | Anyone have an exploit yet for **MS14-066**? |
| 2 | Anyone know what the exploit behind **MS14-068** is? |
| 3 | anyone experiment with the recent **CVE-2014-6352** exploit? |
| 4 | So, has there been a public exploit for **CVE-2014-6321** developed yet? |

Observando as mensagens é possível identificar o interesse em obter códigos de exploração para novas vulnerabilidades. Em um dos casos analisados, foi divulgada uma URL para um código de exploração, mas infelizmente ela não estava mais disponível ao processar os registros. Foi observado que em canais de segurança é comum especialistas postarem notícias sobre novas vulnerabilidades. Logo, essa informação pode ser útil para a prevenção antecipada. A identificação de usuários interessados em códigos maliciosos para novas vulnerabilidades auxilia na escolha de usuários a serem supervisionados.

#### 4.2.3.4 Associações de palavras

A identificação de associações de palavras ocorreu usando os métodos de extração de bigramas e trigramas. Foram selecionados os 200 melhores escores para três métricas de pontuação: Frequência, *Pointwise Mutual Information* (PMI) e Razão de Verossimilhança (*Likelihood Ratio*). Cada uma das associações foi analisada manualmente por meio de consulta às mensagens que as continham e próximas, no caso, as N anteriores e N posteriores (na pesquisa foi usado N=2). A Tabela 4.15 apresenta algumas das associações de interesse para o monitoramento.

**Tabela 4.15:** *Amostra de associações relevantes para monitoramento*

| | Expressão | Exemplo de mensagem |
|---|---|---|
| 1 | take down | lets **take down** www.gadgetwide.com |
| 2 | taking down | I need help **taking down** this site but I think there is not enough people here |
| 3 | get involved | anyone want to **get involved**? |
| 4 | tango down | **tango down** target www.infotec.be |
| 5 | new target | send a **new target** |
| 6 | .check <URL> | **.check** *www.slmpd.org* / **.check** *www.micheldestot.fr* |
| 7 | .dns <URL> | **.dns** *thegrapevine.cc* |
| 8 | 0day ...<SOFTWARE> | ...play around with this new super-**0day** for *Windows*?" |

As linhas de 1 a 5 da Tabela 4.15 apresentam *collocations* que, ao serem pesquisadas, devolveram informações relevantes como alertas nas próprias mensagens ou próximas. Outros bigramas (linhas 6 e 7) estavam sempre relacionando os termos *.check* e *.dns* com diferentes URLs. Logo, caracterizaram uma relação sintagmática entre o termo e uma URL. No entanto, como a URL pode ser substituída

por qualquer outra da mesma classe, também caracterizou uma relação paradigmática. A linha 8 foi obtida a partir de um trigrama, mas tratada como um bigrama, pois o "for" pode ser substituído por outros termos ou vazio.

É importante ressaltar que nem sempre uma busca por uma associação devolve bons resultados. Por exemplo, ao buscar por *"take down"* são retornadas mensagens fora do contexto como *"... if you wanna take down a country"* ou *"Anything going on about TPB being take down ..."*. Logo, não basta apenas verificar a presença da associação para filtrar efetivamente mensagens de interesse como alertas. No entanto, são de grande auxílio para diminuir o escopo de análise para outros mecanismos ou para um analista.

#### 4.2.3.5 Análise de significado

Na análise de significado realizamos a extração de tópicos e entidades para identificar mensagens associadas à e segurança de redes e alvos. A Tabela 4.16 apresenta uma síntese dos resultados da análise de extração de tópicos e entidades nos canais *security* e *ddos*.

**Tabela 4.16:** *Extração de tópicos e entidades*

| Canal | Mensagens analisadas | Mensagens c/ tópicos | Mensagens c/ entidades | Tópicos frequentes | Exemplos de entidades |
|---|---|---|---|---|---|
| security | 10401 | 1019 | 1696 | Technology Internet, Human Interest, Entertainment Culture, Business Finance | (Windows XP, Operating System), (Java, Technology), (Austria, Country) |
| ddos | 2677 | 110 | 534 | Technology Internet, Human Interest, Entertainment Culture, Politics | (JSON, Technology), (Google, Company), (United Nations, Organization) |

Como pode ser observado na Tabela 4.16, o número de mensagens com tópicos e entidades em relação ao total de mensagens é pequeno. Esse resultado é devido à linguagem informal usada na rede IRC e ao tamanho das mensagens. Após uma análise por inspeção manual dos tópicos e entidades, verificou-se que apenas essas informações são insuficientes para dizer se uma mensagem é relevante ou não como alerta. No entanto, podem ser usadas como uma característica para auxiliar na priorização ou descarte de mensagens. Por exemplo, pode-se criar regras para priorizar mensagens com o termo "exploit" e a entidade "Operating System".

Há também a questão das identificações erradas de tópicos e entidades, como no caso de "atm" (*at the moment*) como Technology e "dos" (*deny of service*) como Operating System. Logo, não basta apenas identificar o tópico para priorizar ou descartar uma mensagem individual, mas analisar o contexto das mensagens anteriores e posteriores.

#### 4.2.3.6 Heurísticas

Na análise de potenciais heurísticas para identificar alertas foram investigadas mensagens interrogativas, a presença de ofensas e as URLs.

A investigação de mensagens interrogativas consistiu em selecionar mensagens que caracterizavam questões e verificar se apresentavam padrões para identificar usuários com intenções de realizar ataques e potenciais alvos. A seleção de amostra para a análise ocorreu usando um algoritmo simples baseado na presença de ponto de interrogação (?) nas mensagens. Em seguida, foram consultadas as seguintes características nas mensagens selecionadas: (a) URLs (b) IPs (c) termos de segurança (d) termos *hackers* (e) ofensas. Se mensagens de potenciais alertas ou mensagens irrelevantes se destacavam no conjunto com essas características, as características e as mensagens foram anotadas para a implementação de filtros de priorização e remoção de mensagens. A Tabela 4.17 apresenta os resultados da relação entre características e mensagens interrogativas postadas nos canais.

**Tabela 4.17:** *Análise de mensagens interrogativas (Base A)*

| Canais | Número de questões | com URL | com IP | com termos de segurança | com termos *hackers* | com ofensas |
|---|---|---|---|---|---|---|
| anonops | 3796 | ◐ | ● | ○ | ● | ○ |
| ddos | 773 | ● | ● | ○ | ● | ○ |
| opferguson | 3329 | ○ | - | ○ | ● | ○ |
| opnewblood | 1819 | ○ | - | ○ | ◐ | ○ |
| security | 1848 | ◐ | - | ◐ | ◐ | ○ |
| networking | 3447 | ○ | ○ | ○ | ○ | ○ |

● característica relevante para identificação de alerta.
◐ característica com restrições para identificação de alerta.
○ característica não relevante para identificação de alerta.
- característica não presente na amostra.

Observa-se na Tabela 4.17 que mensagens interrogativas acompanhadas com termos *hackers*, na maioria dos canais, possibilita identificar potenciais alertas. Por exemplo, as mensagens "*Who is getting Ddos today?*" e "*wanna dDos something?*" indicam a intenção de realização de ataques ou procura de alvos. No entanto, nos canais *opnewblood* e *security* há poucas mensagens, como do canal *security*, "*Don't 0day for sale?!*". O uso de ofensas em mensagens interrogativas sem qualquer outra característica indica mensagens irrelevantes. O mesmo ocorre com o uso de apenas termos de segurança em mensagens, que geralmente indicam questões sobre ferramentas de segurança, como conectar em redes virtuais privadas, entre outros. No entanto, no canal *security*, a combinação com determinadas palavras (p. ex. patch, vulnerabilities e bugs) possibilitam encontrar potenciais alertas. O uso de IP e URL no canal *ddos* geralmente indicam potenciais alertas. Já no canal *security*, as URLs podem indicar páginas que descrevem novas ameaças à segurança.

A investigação de mensagens com ofensas consistiu em selecionar mensagens com insultos ou linguagem imprópria para verificar se essa característica pode ser usada como uma heurística para a remoção de mensagens irrelevantes. A seleção da amostra ocorreu por meio de lista com as ofensas mais comuns nas mensagens postadas nos canais analisados. A lista foi construída a partir da análise de frequência das palavras nas mensagens. A Tabela 4.18 apresenta a relação entre mensagens com ofensas e número de potenciais alertas por canal.

**Tabela 4.18:** *Análise de mensagens com ofensas (Base A)*

| Canais | Mensagens com ofensas | Potenciais alertas |
|---|---|---|
| anonops | 2254 | 10 |
| ddos | 239 | 5 |
| opferguson | 1391 | 9 |
| opnewblood | 300 | 5 |
| security | 269 | 6 |
| networking | 634 | - |

Observa-se na Tabela 4.18 que há um número reduzido de mensagens que possuem ofensas e podem se tornar candidatas a alertas nos canais. Mesmo nas mensagens potenciais, há outras características além da presença de ofensa, por exemplo, a presença de termos de cibersegurança ou URLs, como nas mensagens "*everyone attack this shit: habbin.biz*" e "*is shit still going down on the 6th?*". Por outro lado, mesmo mensagens com termos de cibersegurança podem ser irrelevantes como alertas, como na mensagem "*i will backdoor your* [redacted]". Nesses casos, é interessante analisar o contexto, ou seja, as mensagens anteriores e posteriores. Entretanto, de forma geral, o uso de ofensas está associado a mensagens sem relevância como alertas.

A investigação de URLs foi usada para identificar potenciais alvos, orquestrações de ataques e compartilhamento de informações nos canais de IRC. Observou-se que usuários postam URLs para indicar alvos de ataque, para compartilhar novas técnicas e códigos, e também para apontar a existência de vulnerabilidades documentadas. Por outro lado, também são usadas para propagar informações sem relevância, por exemplo, notícias, vídeos, imagens e ofensas. Logo, uma heurística

interessante para implementação de filtros, consiste em explorar as URLs para reduzir o escopo de busca por potenciais ameaças ou alertas.

O processamento das URLs envolveu os seguintes passos: (1) extração e normalização das URLs; (2) agrupamento de URLs para identificar domínios e recursos relevantes e irrelevantes; (3) remoção das URLs irrelevantes; (4) uso de outros mecanismos para evidência ou exclusão de URLs; (5) inspeção por especialista da lista final de potenciais URLs com informações relevantes.

No passo 1, a extração e normalização das URLs foi realizada com o auxílio da API twitter-text[8]. No passo 2, sítios Web de notícias, entretenimento, pornografia e outros foram destacados e incluídos em uma lista de URLs para remoção. Sítios Web de compartilhamento, em especial de texto puro, foram destacados e incluídos em uma lista de URLs relevantes. Nesse passo também foram definidas heurísticas para a remoção de URLs, como URLs com data ou termos "news", "article", "story" e nomes longos separados por hífen, pois geralmente não caracterizam alertas ou ameaças. No passo 3, foram removidas as URLs irrelevantes e armazenadas em uma base para inspeção e análise de falsos negativos. No passo 4, foram aplicadas as seguintes heurísticas: acesso ao título do sítio e verificação por termos associados à segurança ou atividades suspeitas; uso de serviço de categorização[9] de sítios Web para identificar sítios irrelevantes; e análise do contexto das mensagens que contém a URL e das mensagens próximas. Esse passo não foi explorado intensivamente, apenas foi realizado como prova de conceito para futuras implementações. O passo 5 consistiu em averiguar os resultados (Tabela 4.19).

**Tabela 4.19:** *Resumo do processamento de URLs (Passos 1 a 3)*

| Canal | Total de URLs | URLs - Inspeção | URLs - Relevantes |
|---|---|---|---|
| anonops | 1229 | 233 | 11 |
| ddos | 219 | 155 | 8 |
| opferguson | 856 | 311 | 15 |
| opnewblood | 156 | 66 | 2 |
| security | 404 | 247 | 12 |
| networking | 667 | 332 | 15 |

Nos resultados da aplicação dos passos 1 a 3, observamos reduções significativas no número de URLs em alguns canais. Por exemplo, no canal *anonops* a redução foi de aproximadamente 81%. No entanto, considerando a média de 10 dias de coleta, se totalizarmos o número de URLs para a inspeção, o valor resultante diário de URLs (134) pode ser considerado uma carga de trabalho alta para um especialista. O cenário pioraria se considerássemos mais canais a serem monitorados. Devido a experimentos preliminares seguindo o passo 4, acreditamos que é viável reduzir ainda mais o número de URLs para a inspeção.

### 4.2.4 Pré-processadores e processadores de alertas

Esta seção descreve o uso dos componentes de pré-processamento de dados e processamento de alertas do arcabouço EWS para a extração de alertas no IRC usando as informações e bases de inteligência originadas da etapa de análise de dados. Também apresenta a avaliação dos resultados produzidos pela implementação desses componentes.

A Figura 4.7 apresenta os componentes especificados para a identificação de alertas em mensagens coletadas na rede IRC.

Nos filtros foram especificados e implementados filtros de remoção e de priorização para reduzir o número de mensagens para o processamento de alertas. As informações usadas nos filtros foram obtidas na análise de dados e armazenadas nas bases de inteligência.

[8] https://github.com/twitter/twitter-text
[9] https://developer.similarweb.com/website_categorization_API

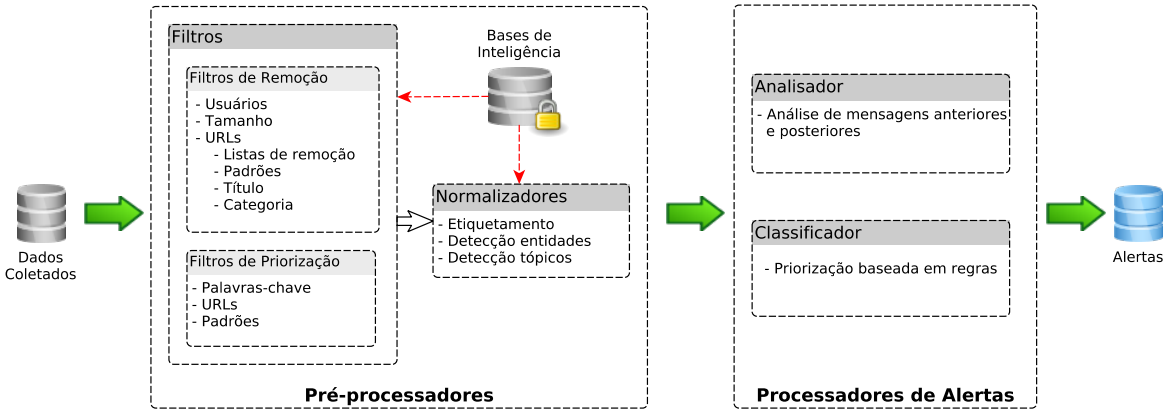


**Figura 4.7:** *Pré-processamento e processamento de alertas de mensagens do IRC.*

Os filtros de remoção implantados para o IRC foram:

- Usuários: remove usuários que publicam informações automatizadas e usuários identificados na análise de dados como propagadores de informações irrelevantes.
- Tamanho: remove as mensagens compostas por poucos caracteres. Nos experimentos foi usado o valor de corte de 12 caracteres.
- URLs: remove mensagens que possuem URLs que conduzem a conteúdos irrelevantes. As técnicas usadas para o filtro foram remoção por lista negra e baseado em padrões. A análise do título e da categoria do sítio obtido pela inspeção do conteúdo da URL não foram implementadas no experimento, apenas avaliadas em algumas situações.

Os filtros de priorização implantados para o IRC foram:

- Palavras-chave: selecionam mensagens que possuem palavras-chave associadas à cibersegurança.
- URLs: selecionam mensagens que apresentam URLs usadas para compartilhamento de informações, principalmente URLs que apontam para sítios de compartilhamento de texto puro.
- Padrões: selecionam mensagens que apresentam padrões de notificações de ataques identificados na análise, como um endereço IP em conjunto com termos usados para indicar ataques (p. ex. tango down, down, .check).

Os normalizadores especificados tinham como objetivo enriquecer o conjunto de dados por meio da expansão das informações.

Os normalizadores especificados para o IRC foram:

- Etiquetamento: identifica e etiqueta termos segundo as categorias especificadas na Tabela 4.13 e outros por expressões regulares (URLs e IPs). Foi efetivamente implementado para possibilitar a análise e classificação de potenciais alertas.
- Detecção de entidades: identifica entidades como pessoas, locais, organizações, entre outros. Foi apenas avaliado para um conjunto de dados restrito e manualmente inspecionado.
- Detecção de tópicos: pode ser usado para identificar tópicos em sentenças e usar a informação na classificação de análise. Foi apenas avaliado para um conjunto de dados restrito e manualmente inspecionado.

O analisador consistiu em adicionar pesos para as mensagens anteriores e posteriores considerando a presença de padrões usados em mensagens de cibersegurança. Em seguida, na análise de uma mensagem, considerava-se como uma característica de contexto, a análise das N mensagens anteriores e posteriores para tentar identificar em que contexto estava inserida a mensagem: potencial alerta ou irrelevante. O classificador usou as informações de contexto para ordenar as mensagens para a inspeção de um analista. Os resultados obtidos são apresentados na Tabela 4.20.

**Tabela 4.20:** *Resultados do uso do arcabouço EWS em diferentes canais*

| Canais | Período (dias) | Total de mensagens | Mensagens destacadas | Mensagens priorizadas | Precisão (1) | Precisão (2) |
|---|---|---|---|---|---|---|
| anonops | 58 | 253522 | 7281 | 626 | 0,50 | 0,37 |
| ddos | 58 | 22695 | 3100 | 892 | 0,84 | 0,90 |
| defacement | 49 | 5350 | 96 | 96 | 0,76 | 0,77 |
| hackers | 50 | 19911 | 1681 | 203 | 0,50 | 0,64 |
| opnewblood | 58 | 48369 | 3266 | 217 | 0,47 | 0,52 |
| security | 68 | 118360 | 14885 | 434 | 0,31 | 0,36 |
| networking | 58 | 131310 | 8913 | 177 | 0,18 | 0,18 |

(1) amostra de 200 primeiras mensagens priorizadas (2) amostra aleatória de 50 mensagens priorizadas

Observa-se na Tabela 4.20 que o número de mensagens destacadas é bem menor que o número total de mensagens. Isso ocorreu devido aos processos de filtros, seleção e classificação. Esse resultado também deve-se à remoção de mensagens que não foram marcadas normalização de dados, logo não apresentam entidades, tópicos ou palavras-chave de interesse como alertas.

Na análise de precisão realizada em amostras, verifica-se que os canais *ddos* e *defacement* apresentam resultados melhores que os outros. Isso ocorreu devido a facilidade de identificar padrões e entidades nesses canais, que estão associadas a URLs, IPs, questionamentos sobre novos ataques e alvos. Nos outros canais, as mesmas expressões produzem uma quantidade maior de falsos positivos. Além disso, os canais *anonops* e *opnewblood* têm a política de não permitir a publicação de alvos, apesar de alguns usuários mesmo assim publicarem. Os canais *security* e *networking*, apesar de possuírem muitas mensagens associadas à cibersegurança, apresentaram baixa precisão pois essas mensagens eram associadas a pedidos de auxílio ou discussões de segurança.

A classificação de entidades possibilitou identificar alvos que não seriam detectados por IP ou URL, ou mesmo priorizados só por palavras-chave, como foi o caso da mensagem *"Anyone interested in a DDoS attack on Dolce & Gabbana?"* e notícias associadas a um sítio web do governo *"brasil website hard to knock out using torshammer"*. No entanto, os melhores resultados com relação a alertas antecipados, são devido ao esquema de etiquetamento baseado nos grupos extraídos pela análise de frequência de palavras e análise de bigramas. Com esses grupos foi possível construir expressões regulares para identificar usuários interessados em realizar atividades maliciosas, potenciais alvos e ataques sendo realizados no momento.

Mesmo conseguindo classificar as informações relevantes, um especialista humano e o uso de filtros adaptativos são indispensáveis para minimizar os falsos positivos e, dessa forma, aumentar a precisão dos resultados. Verificamos que, somente a partir da análise e caracterização do canal a ser monitorado, é possível conduzir a extração de mensagens associadas à cibersegurança. Logo, comprova-se que a análise de canais individuais como uma contribuição importante do arcabouço EWS em relação a outros trabalhos associados ao monitoramento de IRC que visam a segurança de sistemas e alertas antecipados.

### 4.2.5 Trabalhos relacionados

Nesta seção são analisados e comparados trabalhos que investigam o IRC como fonte de informação para identificação de atividades ilícitas e de usuários suspeitos.

Brown (2007) propõe a arquitetura de uma ferramenta automatizada para investigações de roubo de identidade na rede IRC. O trabalho estrutura a arquitetura em cinco módulos (coleta, armazenamento, análise, alerta e localizador) e discute princípios para a implementação de cada módulo. No módulo de análise é proposto o uso de algoritmos de mineração de dados e análise de palavras-chave, frases-chave e expressões regulares. Como apenas modela a ferramenta, apesar de destacar as dificuldades e detalhes de cada fase, o autor não apresenta prova de conceito da arquitetura. Em contrapartida, o uso do arcabouço EWS, proposto nesta tese, possibilitou identificar pontos importantes para a implementação de mecanismos e, além disso, foi realizada a implementação de um protótipo para a extração dos alertas e bases de inteligência para realizar a mineração das informações relevantes.

Michels (2012) implementa e analisa uma ferramenta automatizada para auxiliar investigadores na análise de mensagens em tempo real no IRC. A coleta de informações é realizada durante 1 minuto em cada canal considerado suspeito e foram utilizadas 5 palavras-chave. O processo consiste em identificar as palavras-chave, encontrar tópicos de interesse e realizar a análise de categorias. O investigador é alertado pela ferramenta se for encontrado conteúdo suspeito. Segundo o autor, a parte mais complexa foi identificar canais abertos que estejam abordando atividades suspeitas. Infelizmente, essa abordagem aplicada em canais organizados falharia devido as políticas de bloqueio de aplicativos automatizados e, por consequência, a coleta seria interrompida. Além disso, apenas capturar mensagens em um curto período de tempo é uma abordagem limitada, pois não há garantia que a informação relevante seja publicada no exato momento que a coleta esteja sendo realizada. Inicialmente, na fase de coleta de dados do arcabouço EWS, também foi testada uma abordagem similar que consistia em armazenar somente as mensagens com termos relevantes, porém essas mensagens dispersas, mesmo com a análise de especialista, não possibilitavam confirmar uma ameaça em potencial. Além disso, diferente de Michels, no arcabouço EWS foram monitorados canais abertos e que apresentavam potenciais mensagens de segurança.

Gainaru *et al.* (2010) usam processamento de linguagem natural, clusterização e análise de conhecimento para analisar sessões de conversação. O trabalho aborda a detecção de tópicos, identificação de cadeias léxicas e resolução de correferência. Concluem que os resultados podem ser melhorados por uso de heurísticas, em especial, a detecção de tópicos. Em contraste, com o arcabouço EWS foram analisados vários canais e também explorado o uso de heurísticas identificadas a partir dessa análise.

Iqbal *et al.* (2012) realizam a extração de associações e identificação de tópicos a partir da análise dos registros de sessões de chat obtidas a partir de máquinas apreendidas para investigação criminal. Para tal, realizam a divisão dos registros em períodos temporais para identificação de associações entre os participantes e, em seguida, a identificação de tópicos nesses períodos. Os autores destacam a dificuldade na análise das sessões devido ao tamanho e informalidade das mensagens, além dos erros de grafia. Com o arcabouço EWS, foi realizada a extração considerando bases de dados resultantes de monitoramento contínuo durante dois períodos distintos e, por devido a caracterização das mensagens dos canais de segurança, foram identificados e usados padrões, por exemplo, ocorrência de gírias, acrônimos e ofensas, para auxiliar no processo de filtro e classificação de relevância.

Décary-Hétu e Dupont (2012) identificam potenciais suspeitos de atividades hackers pela análise de sessões IRC obtidas a partir de computadores apreendidos. Apesar de analisarem comunicações privadas, a análise de redes sociais possibilita eleger os grupos que merecem mais atenção no monitoramento. Com os resultados obtidos com o arcabouço EWS, foi possível identificar usuários suspeitos ou com intenções de executar ataques, mas considerando o monitoramento de canais. Essa abordagem usada em tempo real, pode facilitar para investigadores identificar potenciais ameaças antes que ocorram ou mais rápido que depender de informações de outros meios.

Benjamin e Chen (2014) direcionam seus esforços em uma abordagem proativa e na proposta de novas metodologias para compreender a ação dos hackers e ameaças emergentes. Reforçam a

ideia de que hackers visitam diversas comunidades para melhorar suas habilidades simplesmente consumindo os recursos compartilhados nessas comunidades. Os canais IRC e fóruns são os principais locais usados por comunidades hackers para divulgar suas ações e recursos. Eles analisaram os usuários do canal *anonops* durante seis meses e constataram que as atividades ilegais são discutidas em canais privados enquanto a maior parte das mensagens são relacionadas a tecnologia em geral. Em contrapartida, por seguirmos a abordagem de analisar diferentes canais com o arcabouço EWS, nossos resultados mostraram que há informações relevantes para identificação de ameaças emergentes, por exemplo, alvos de ataques, postagem de novas ferramentas e notificações de vulnerabilidades.

Benjamin *et al.* (2015) propõem também um arcabouço para identificar as ameaças e vulnerabilidades em fóruns Web, nos canais IRC e sítios de compras de dados de cartões. Utilizam abordagens baseadas em técnicas de aprendizagem de máquina e recuperação da informação. Em especial, no estudo do IRC, identificam o vocabulário usando análise de frequência, uma das abordagens que usamos no nosso estudo. Em contrapartida, focam apenas em identificar palavras-chave associadas à cibersegurança e não analisam palavras e padrões que caracterizam mensagens irrelevantes e nem associações de palavras comuns em canais hackers. Os resultados que apresentam apenas apontam que existem informações relevantes como alertas, mas não mostram a precisão de sua solução por apresentarem somente amostras de ameaças. No nosso estudo também verificamos os canais que abordam segurança de computadores.

Nossa proposta inova ao investigar características para a identificação de informações que podem ser usadas como alertas, preferencialmente antecipados, nas mensagens postadas em canais de IRC. Difere-se das demais propostas similares pelo monitoramento constante de diferentes canais e pela proposta de um arcabouço de extração de informações associadas à cibersegurança, que considera a combinação de técnicas de processamento de linguagem natural e recuperação da informação adaptadas segundo os resultados da análise dos canais. Também reaproveitamos bases de inteligências de outros estudos, como a base de URLs que apontam para conteúdos irrelevantes obtida na análise de dados do Twitter.

### 4.2.6 Discussão e síntese dos resultados

Na análise de canais do IRC foi verificado que há informações que podem ser usadas como alertas por administradores de redes, como pode ser observado nas Tabelas 4.15 e 4.14. Verificou-se que há várias mensagens identificando ações maliciosas, como definição de alvos para ataque e a procura por códigos de exploração. Também foram obtidas URLs para recursos suspeitos, como arquivos e descrição de alvos de ataques.

Quanto a alertas antecipados sobre alvos de ataques em canais *hackers*, foram identificados alvos ou rumores de alvos compartilhados abertamente nesses canais, como pedidos de auxílio para executar um ataque. Foi confirmado que há alvos discutidos em sessões privadas e não podem ser observados, mas isso possibilita identificar usuários suspeitos que precisam ser monitorados. Além disso, observou-se tendências de ações e as ferramentas usadas para a execução de ataques. Nos canais do anonops, uma operação a alvos é facilmente identificada pelo prefixo “op”.

O processo de identificação de alertas pode ser automatizado segundo o arcabouço EWS, mas os métodos precisam ser refinados nos diferentes componentes do arcabouço. A supervisão humana é estritamente necessária para evitar problemas no monitoramento e, se possível, para garantir acesso a salas fechadas. Em nossa prova de conceito, ainda foram identificadas muitas mensagens irrelevantes para um analista fazer a conferência manual. No entanto, os métodos propostos, se refinados, podem conduzir a uma diminuição nos falsos positivos.

Resumindo, nas redes IRC foi confirmado a existência de informações relevantes como alertas e que essas redes podem ser usadas como fontes de sistemas de alertas antecipados, mesmo considerando o fato de que o IRC teve uma queda na sua utilização nos últimos anos. Nos canais de

segurança foi averiguado a discussão de novos códigos de exploração e vulnerabilidades, enquanto que nos canais de atividades suspeitas foi averiguado a discussão de ferramentas e alvos de ataques. O arcabouço EWS possibilitou a identificação e a extração de mensagens associados à cibersegurança, mas ainda há várias dificuldades com o monitoramento e processamento da relevância das mensagens retornadas como alertas. Verificou-se a diminuição de mensagens e evidência de informações importantes devido às técnicas e características extraídas da fase de análise de dados. Acredita-se que os resultados podem ser melhorados, principalmente se empregados novos métodos para os componentes de processamento de alertas.

## 4.3 Classificador

Nesta seção, são apresentados o desenvolvimento e avaliação de classificadores voltados a alertas de cibersegurança extraídos de fontes de dados não estruturados. O modelo de classificação proposto considera características comuns de diferentes fontes, logo, não abrange características de uma fonte específica. Não se objetiva realizar um estudo aprofundado para otimizar os classificadores para identificação de alertas e não alertas, mas sim, mostrar como são importantes no contexto do arcabouço EWS. A Figura 4.8 apresenta um fluxograma de desenvolvimento e avaliação do classificador.

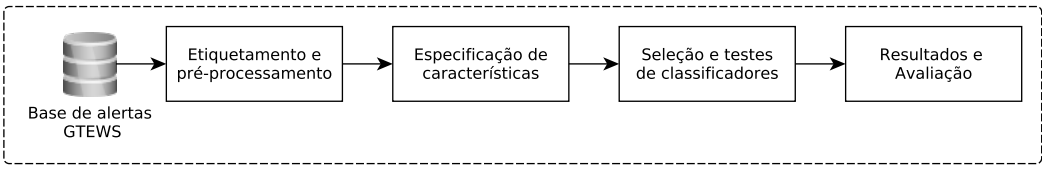


**Figura 4.8:** *Métodos para o desenvolvimento do classificador.*

As próximas subseções abordam cada uma das fases e resultados ilustrados na Figura 4.8.

### 4.3.1 Dados

A base de dados usadas para o desenvolvimento e avaliação do classificador corresponde a base D (Tabela 4.1) adicionada de novas mensagens coletadas no Twitter e Facebook. Essas mensagens foram apresentadas como alertas pelo sistema desenvolvido no GTEWS[10] e foram coletadas no período de março/2015 até junho/2016. A base contém 4925 mensagens em português que já foram filtradas e pré-classificadas pelo algoritmo especialista implantado no sistema. Como há falsos positivos, essas mensagens foram etiquetadas manualmente por dois especialistas em segurança. A Tabela 4.21 apresenta amostras de mensagens contidas na base.

Na Tabela 4.21, as linhas 1 a 6 apresentam exemplos de mensagens que não são interessantes como alertas. A linha 1 contém uma notícia de segurança (infosec); As linhas 2 e 3 são piadas sobre *hackers*; a linha 4 é uma propaganda sobre ferramenta de segurança; a linha 5 é um agradecimento para um *hacker* que vazou um jogo; e a linha 6 é sobre segurança física. Já as linhas 7 a 10 contém amostras de mensagens que são alertas de segurança. As linhas 7 e 8 são sobre desfiguração de páginas; a linha 9 sobre ataque DDoS em andamento; e a linha 10 sobre um vazamento de dados. O principal desafio é classificar as mensagens que não são notificações de cibersegurança mesmo considerando os casos que há muitos termos indicando o contrário.

[10] https://gtews.ime.usp.br

**Tabela 4.21:** *Amostra de mensagens da base de dados.*

| Linha | Alerta | Mensagem |
|---|---|---|
| 1 | x | 2015 já é ano recorde para **DDoS** #**segurança** http://... |
| 2 | x | meudeusssssss ela eh **hacker** nao tenho mais nos meus contatos aaaa que medo... |
| 3 | x | você descobriu a **senha** socorro merece o oscar de melhor **hacker** parabéns... |
| 4 | x | Ninguém,está livre de **contaminar** sua **máquina** com algum **vírus** e do **ataque Hacker**. Ferramentas de **Segurança** de **Redes**.: http://... |
| 5 | x | queria deixar aqui o meu muito obrigado ao **hacker** que **vazou** The Witcher III |
| 6 | x | **Hacker** cria **robô** com impressora 3D que consegue destravar cadeados de **senha** em 5 minutos - Adrenaline http://... |
| 7 | ✓ | **Hacked** - É **admin**,sua proteção não adiantou muito. **http://...** |
| 8 | ✓ | Continua **defaceada https://...** |
| 9 | ✓ | ...Caros a **rede** Brasil esta **inoperante** há alguns minutos, a mesma esta sobre forte **ataque Ddos**. O mesmo esta sendo **mitigado**. |
| 10 | ✓ | WarezTuga **alvo** de **ataque** e milhares de **passwords** são divulgadas: O conhecido WarezTuga foi **vitima** de um **ataque**... http://... |

### 4.3.2 Etiquetamento e pré-processamento

O pré-processamento visa eliminar as mensagens similares, remover informações irrelevantes nas mensagens e realizar o enriquecimento dos dados. O etiquetamento consiste em identificar a qual classe a mensagem deve ser atribuída. No caso, como há duas classes, consistiu em marcar se a mensagem é relevante ou não como um alerta de cibersegurança.

O etiquetamento foi realizado por dois especialistas em segurança por meio da análise individual de cada mensagem. Os conflitos nas decisões de classificação foram resolvidos após a discussão e acordo entre os especialistas. No final foram identificadas 3450 mensagens como alertas e 1475 como não alertas.

O pré-processamento realizado consistiu de:

- sanitizar o início e final das mensagens (remover RTs, símbolos, menções, caracteres de controle, entre outros).
- remover as palavras definidas na lista de palavras irrelevantes (*stop words*).
- realizar a classificação gramatical das palavras (*part-of-speech (PoS)*).
- identificar entidades (URLs, IP, usuário).
- identificar os termos de cibersegurança.
- traduzir URLs curtas para longas.

Após a realização do pré-processamento, as mensagens foram submetidas ao processo de extração de características.

### 4.3.3 Especificação de características

As características consideradas para o treinamento, teste e construção do modelo do classificador foram selecionadas a partir de atributos comumente presente nas fontes. A Tabela 4.22 apresenta o conjunto de características.

Selecionou-se quatro conjuntos de características que combinavam os atributos apresentados na Tabela 4.22. O conjunto de características F1, F2, F3 e F4 são respectivamente compostos pela combinação de atributos das linhas: 1 a 11; 12 e 13; 1, 3, 4, 14 e 15; 12 a 15.

**Tabela 4.22:** *Seleção de atributos para a classificação de alertas.*

| L | Característica | Tipo | Descrição |
|---|---|---|---|
| 1 | usuario_mon | booleana | perfil do usuário é monitorado na coleta. |
| 2 | mencao_usuario_mon | booleana | usuário monitorado é mencionado no texto. |
| 3 | palavras_whitelist | numérico | número de palavras associadas à cibersegurança. |
| 4 | palavras_blacklist | numérico | número de palavras comumente associadas a outros contextos. |
| 5 | ip_privado | numérico | número de endereços IP redes privadas ou não válido na Internet |
| 6 | ip_publico | numérico | número de endereços IP válidos na Internet |
| 7 | urls_blacklist | numérico | número de URLs de sítios irrelevantes para notificação. |
| 8 | urls_whitelist | numérico | número de URLs de sítios públicos e instituições de ensino. |
| 9 | urls_paste | numérico | número de URLs para sítios de divulgação de compartilhamento de texto ou desfigurações de páginas. |
| 10 | bow_t | numérico | atributos de combinação de palavras com janela 5 e termo de cibersegurança no centro. |
| 11 | bow_p | numérico | atributos de combinação de PoS com janela 5 e termo de cibersegurança no centro. |
| 12 | bow_u_c | numérico | combinação de palavras unigramas de termos de cibersegurança. |
| 13 | bow_u_s | numérico | combinação de palavras unigramas de termos "spam". |
| 14 | bow_b_c | numérico | combinação de palavras bigramas de termos de cibersegurança. |
| 15 | bow_b_s | numérico | combinação de palavras bigramas de termos "spam". |

### 4.3.4 Seleção e testes de classificadores

Os classificadores selecionados para a realização de testes foram: Entropia Máxima, Naive Bayes, Random Trees, J48 e o SVM. Todos os algoritmos foram executados usando a API do Weka[11], exceto o de Entropia Máxima, que foi implementado com a API Apache openNLP[12] e o SVM que usou a API LibSVM (Chang e Lin, 2011).

Os experimentos foram executados em uma máquina com processador Intel i5, 12 GiB de memória, sistema operacional Ubuntu 14.04, Ambiente de execução Java 1.8.0_101, Weka 3.8.0 e openNLP 1.6.0.

A Tabela 4.23 apresenta os parâmetros usados na execução dos classificadores. Os parâmetros foram os mesmos para todos os conjuntos de características.

**Tabela 4.23:** *Configuração e resultados dos testes de classificadores.*

| Classificador | Parâmetros | Teste |
|---|---|---|
| Entropia Máxima | - | Validação cruzada 10-folds |
| J48 | -C 0.25 -M 2 | Validação cruzada 10-folds |
| RandomForest | -P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1 | Validação cruzada 10-folds |
| Naive Bayes | - | Validação cruzada 10-folds |
| SVM | -S 0 -K 2 -D 3 -G 0.0 -R 0.0 -N 0.5 -M 40.0 -C 1.0 -E 0.001 -P 0.1 | Validação cruzada 10-folds |

Na próxima seção são apresentados os resultados dos testes.

### 4.3.5 Resultados e avaliação

A Tabela 4.24 apresenta os resultados da execução de cinco classificadores diferentes nos conjuntos de características F1, F2, F3 e F4. O algoritmo de Entropia Máxima extrai as características diretamente do texto a ser classificado, logo os conjunto F1 a F4 não se aplicam a esse modelo de classificador.

[11] http://www.cs.waikato.ac.nz/ml/weka/
[12] https://opennlp.apache.org/

**Tabela 4.24:** *Medidas de precisão(P), abrangência(R), acurácia(A) e taxa de falsos positivos(F).*

| | F1 | | | | F2 | | | | F3 | | | | F4 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **P** | **R** | **A** | **F** | **P** | **R** | **A** | **F** | **P** | **R** | **A** | **F** | **P** | **R** | **A** | **F** |
| J48 | 0,79 | 0,84 | 0,74 | 0,53 | 0,73 | 0,97 | 0,73 | 0,82 | 0,71 | 0,95 | 0,70 | 0,89 | 0,73 | 0,97 | 0,73 | 0,82 |
| RandomForest | **0,80** | **0,88** | **0,76** | **0,52** | 0,74 | 0,96 | 0,73 | 0,81 | 0,72 | 0,93 | 0,70 | 0,82 | 0,74 | 0,96 | 0,73 | 0,81 |
| Naive Bayes | 0,75 | 0,78 | 0,66 | 0,61 | 0,74 | 0,93 | 0,72 | 0,77 | 0,70 | 0,98 | 0,70 | 0,97 | 0,74 | 0,93 | 0,72 | 0,77 |
| SVM | 0,72 | 0,99 | 0,72 | 0,91 | 0,74 | 0,96 | 0,73 | 0,81 | 0,70 | 1 | 0,70 | 1 | 0,74 | 0,97 | 0,73 | 0,81 |
| Entropia Máxima | P=**0,81**; R=**0,95**; A=**0,81**; F=**0,51** | | | | | | | | | | | | | | | |

Observa-se que o classificador de Entropia Máxima apresentou o melhor resultado ao alcançar a precisão 81%, no entanto, ainda apresenta uma alta taxa de falso alarme (51%). Outro classificador que apresentou resultado razoável foi o de RandomForest, com a precisão de 80% e uma taxa de falso alarme de 52%. Em todas os testes o número de falsos alarmes foi acima de 50%, o que mostra que o modelo de classificação precisa ser melhorado. Em especial, os modelos que fazem o uso de características.

Há várias razões que conduziram a resultados não tão satisfatórios na tarefa de identificar alerta de não alerta, são elas: (i) características não foram suficientes para identificar padrões para alertas falsos; (ii) conteúdo de alertas são informais, o que dificulta padronizar a entrada e extrair atributos relevantes automaticamente; (iii) a diferença entre alertas e não alertas é muito sutil, principalmente em textos reduzidos; (iv) o número de falsos positivos na base de treinamento é pequeno; (v) a qualidade da base é dependente de quem classificou manualmente; e (vi) os parâmetros dos algoritmos de classificação não foram otimizados.

Esses resultados mostram a importância da supervisão ou colaboração de especialistas em segurança para evidenciar alertas que são relevantes. Também mostram que é necessário investigar novas formas para classificar automaticamente alertas cibernéticos que foram obtidos de fontes de dados não estruturados.

## 4.4 Recomendador

Nesta seção, são apresentados o processo de modelagem, o desenvolvimento e a avaliação de um modelo de recomendação voltado a alertas de cibersegurança extraídos de fontes de dados não estruturados. O modelo proposto apresenta uma abordagem híbrida para possibilitar a recomendação de alertas de baseado nas preferências anteriores do usuário (filtragem baseada em conteúdo) e na similaridade de preferências de outros usuários (filtragem colaborativa). Além disso, o modelo é usado na redução de falsos positivos via a colaboração de especialistas em segurança durante a recomendação dos alertas.

### 4.4.1 Metodologia

A metodologia para a construção do modelo de recomendação para alertas de cibersegurança adota o processo proposto por Picault (Picault *et al.*, 2011). Nesse processo, o problema de recomendação é abordado em três dimensões: usuários, dados e aplicação.

No presente escopo, os usuários são administradores de redes, especialistas em segurança, entre outros com perfis ou interesses similares, que objetivam proteger suas infraestruturas de redes e sistemas computacionais. Os dados são não estruturados e abordam conteúdos textuais relevantes ou irrelevantes como alertas. A aplicação consiste na extração e notificação de alertas de cibersegurança, preferencialmente antecipados, de modo a minimizar os falsos positivos e notificar alertas ou potenciais alertas de interesse.

Considerando a importância em compreender os usuários, dados e a aplicação, foi adotado os seguintes processos para o desenvolvimento e avaliação do modelo: (i) análise de requisitos e

caracterização de usuários e itens do modelo; (ii) especificação do modelo; e (iii) avaliação do modelo. As próximas subseções descrevem cada um desses processos.

#### 4.4.1.1 Requisitos de usuários e itens

A compreensão das entidades em um sistema de recomendação deve ser a atividade inicial para o desenvolvimento de um modelo de recomendação (Ricci *et al.*, 2011). A caracterização dos usuários possibilita identificar perfis, atributos e como o usuário interage com os itens. Já a caracterização dos itens, no caso, alertas de segurança, possibilita identificar atributos-chave para a construção do modelo.

A análise de requisitos de usuários e itens foi realizada por meio de um questionário enviado a dezenas de administradores/profissionais de rede de computadores. Esses profissionais atuam em organizações privadas e públicas do Brasil. O questionário foi disponibilizado online no ESurv[13], uma ferramenta aberta para a criação de questionários. O Apêndice D apresenta a versão do questionário usado na pesquisa.

Com o objetivo de explorar o perfil dos usuários, características das notificações de segurança, interesse em colaboração em um sistema de recomendação, o questionário é composto por 26 questões e foi dividido em quatro partes:

- Levantamento de perfil: identificar características dos potenciais usuários do sistema de recomendação.
- Conhecimentos sobre cibersegurança: identificar o nível de conhecimento dos usuários sobre cibersegurança e quais canais usam para se manterem atualizados sobre segurança e a potenciais ameaças.
- Avaliação de interesse: identificar os principais requisitos para o modelo do ponto de vista do usuário do sistema, como os atributos-chave que são relevantes para uma notificação de segurança e também as potenciais relações entre usuários e itens de recomendação.
- Situações de uso: avaliar situações hipotéticas de uso de um sistema de recomendação para analisar mais precisamente as potenciais interações dos usuários.

A síntese e análise das respostas do questionário são apresentadas na Seção 4.4.2. Os resultados foram usados para a especificação do modelo de recomendação.

#### 4.4.1.2 Especificação do modelo

A especificação do modelo consistiu inicialmente em caracterizar os usuários e itens do sistema. Para tal, foram usados os resultados obtidos do questionário (Seção 4.4.2). A partir dessas informações foram definidas as estruturas de dados para representar usuários e itens. Em seguida, foram especificadas as transações, ou seja, como os usuários interagem com os itens. Novamente, pela análise do questionário, foi definido um conjunto de interações que seriam relevantes para um sistema de recomendação de alertas de segurança. Por fim, foram selecionados e adaptados os algoritmos que atendem os requisitos para a implementação do recomendador.

A Seção 4.4.3 apresenta a especificação do modelo contendo a estrutura dos usuários e itens, as transações e fluxos de processamento do sistema, os algoritmos e fórmulas para a realização da recomendação.

[13] http://esurv.org/

#### 4.4.1.3 Avaliação do modelo

Como descrito em (Gunawardana e Shani, 2011), a avaliação do modelo de recomendação consiste em identificar o conjunto de propriedades que podem influenciar o sucesso do sistema de recomendação no contexto da aplicação, no caso, realizar a recomendação de alertas de interesse para usuário de um sistema de alerta antecipado.

Há três abordagens usadas para a avaliação de sistemas de recomendação: experimentos offline, estudo de usuário e experimentos online. A experimentação offline usa bases de dados com o histórico de interações dos usuários e itens avaliados. O estudo de usuário é realizado por meio da seleção e estudo de um grupo de usuários que interage com o sistema de recomendação. A experimentação online é a avaliação da interação direta dos usuários com o sistema em produção (Gunawardana e Shani, 2011).

A abordagem utilizada para a realização da avaliação foi o experimento offline. Nessa abordagem é necessário o uso de uma base de dados contendo avaliações de alertas de cibersegurança por usuários. Como não há uma base com essas informações, no presente trabalho foi realizada uma avaliação com uma base de dados de filmes, apenas para verificar a operação dos algoritmos, e com uma base sintética de notificações de alertas de segurança e de avaliações de usuários.

A base sintética foi gerada com o uso de funções estatísticas que procuram imitar o comportamento de usuários com o perfil de administradores de rede ou especialistas em segurança. Os itens avaliados correspondem aos alertas publicados no CVE em 2015. As preferências dos usuários foram geradas baseadas na proporção de alertas de aplicativos e fornecedores. Por ser sintética, a base tem a finalidade apenas de suportar a avaliação dos algoritmos, mas não possibilita analisar a influência do sistema de recomendação nos usuários.

As métricas de avaliação usadas foram a precisão e a abrangência. A precisão mede a habilidade de recomendar somente o que é de interesse para o usuário (Equação 2.1). A abrangência mede a habilidade de recomendar tudo o que é relevante para o usuário (Equação 2.2). Ambas as métricas produzem valores entre 0.0 e 1.0, onde o valor mais alto representa melhor desempenho do algoritmo segundo a propriedade.

Os detalhes das bases de dados, resultados dos experimentos e discussões são apresentados na Seção 4.4.4.

### 4.4.2 Análise de requisitos do recomendador

A análise de requisitos consistiu em analisar e definir os usuários e os itens para a construção do modelo do recomendador. Ambos elementos podem ser definidos como estruturas de dados que representam respectivamente os administradores e alertas no mundo real. Nesta seção, são apresentadas a compilação das respostas do questionário e a análise de atributos e relações entre usuários e itens.

O questionário foi respondido por 44 indivíduos que possuem o perfil de administrador de redes e trabalham em organizações privadas ou públicas. Verificando as respostas associadas ao perfil do usuário, observou-se que a experiência na administração e segurança de redes é de aproximadamente 9 anos. A maioria dos administradores de redes opera infraestruturas de redes com servidores Linux e clientes Windows, permitem o uso de dispositivos móveis do usuário na rede da organização, no entanto, não realizam o monitoramento desses dispositivos. Esses resultados indicam que essas infraestruturas acabam sendo alvos dos mais diversos tipos de ataques devido a heterogeneidade de dispositivos e aplicativos, mas principalmente pela liberdade de acesso e falta de monitoramento dos dispositivos. Isso reforça a necessidade de uma forma de evidenciar notificações de segurança aos administradores, alertando a criticidade de uma notificação específica para as infraestruturas geridas por eles.

Nas respostas associadas ao conhecimento sobre cibersegurança, observou-se que a maioria dos administradores possuem conhecimento sobre cibersegurança, porém, apenas um terço realiza a leitura de notícias de segurança diariamente (Figura 4.9). Além disso, a minoria das notícias são relevantes para os administradores (Figura 4.10). Verificou-se que a forma mais popular para a obtenção de informações de segurança é via correspondência eletrônica e blogs. No entanto, apenas metade trocam informações com outros administradores e, quando o fazem, realizam geralmente por e-mail.

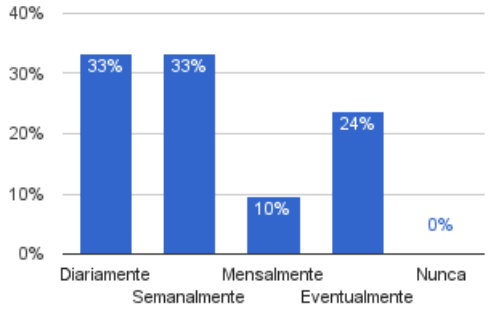


**Figura 4.9:** *Frequência de leitura de notificações de segurança.*

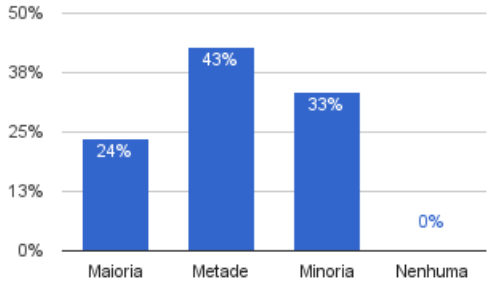


**Figura 4.10:** *Notificações lidas de interesse.*

As Figuras 4.9 e 4.10 evidenciam novamente a necessidade de um sistema de recomendação para filtrar notícias de segurança segundo os interesses dos administradores e diminuir o número de notificações irrelevantes. Além das formas populares, notícias publicadas em outras fontes poderiam ser evidenciadas rapidamente por recomendação de outros administradores. Há a possibilidade de promover uma maior interação e colaboração entre administradores, visto que apenas metade dos entrevistados relacionam-se com seus pares.

Nas respostas de avaliação de interesse, 98% dos entrevistados assinalaram a importância de um sistema para compartilhar informações de segurança, no entanto, todos usariam e colaborariam com um sistema que compartilhasse e recomendasse informações segundo o perfil do administrador. Essas respostas mostraram que além de recomendar notícias de outras fontes, os administradores também colaborariam acrescentando novas notificações, o que propicia o aumento das chances de identificação de alertas antecipados.

Na identificação de atributos para o compartilhamento, o gráfico ilustrado na Figura 4.11 mostrou que a maioria dos administradores não proveria informações pessoais como o nome pessoal e da instituição. Atributos relevantes para os algoritmos de recomendação, isto é, sistemas operacionais, aplicativos, equipamentos e notificações de segurança seriam anotados em uma notificação.

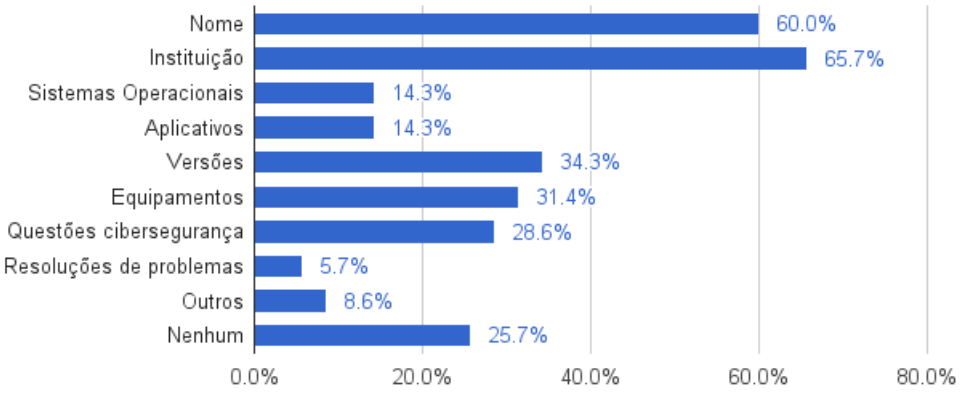


**Figura 4.11:** *Informações que os usuários não forneceriam a um sistema de colaboração.*

Os atributos apontados como mais relevantes em notificações foram: aplicativos envolvidos, tipo

da ameaça, impacto e origem da notificação. Essas informações podem ser usadas para determinar a preferência do usuário em um item (notificação). Para a composição do item, ou seja, os alertas de segurança, os resultados do questionário evidenciaram, da mais importante para a menos importante, as seguintes informações: título, software/hardware relacionado ao alerta, origem do alerta, nível de criticidade, categoria da notícia, palavras-chave (tags). Essas informações devem ser anotadas, processadas ou extraídas de textos que descrevem alertas de segurança obtidos em fontes não estruturadas para atender os requisitos dos usuários de um sistema de recomendação de alertas de cibersegurança.

A Figura 4.12 apresenta as preferências dos usuários para a avaliação de itens.

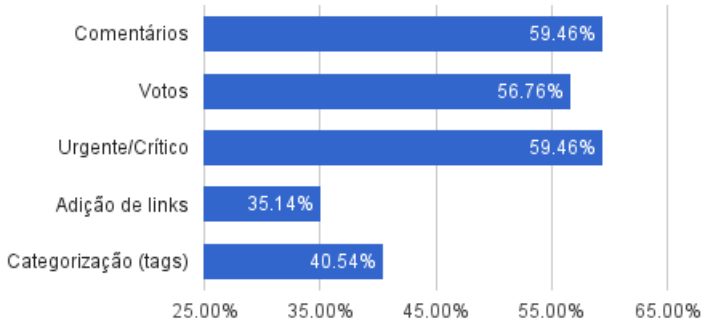


**Figura 4.12:** *Mecanismos de avaliação.*

Como pode ser observado na Figura 4.12, os mecanismos mais populares são de assinalar um alerta como urgente/crítico, comentários e votos. A urgência ou criticidade de um alerta representa quão relevante ele é para um indivíduo. Os votos possibilitam identificar se um notícia ou notificação é relevante ou não e pode ser influenciado por diversas razões, por exemplo, tipos de aplicativos de interesse do usuário, propagandas ou notificação genéricas, entre outros. Já os comentários são interessantes por possibilitar a interação dos usuários com a notificação, mas para um sistema de recomendação para dados não estruturados acabam por gerar novos dados não estruturados. Por esse motivo, é preferível o uso do mecanismo de categorização, que possibilitam caracterizar com poucos termos as partes relevantes e o contexto dos alertas, além de descreverem as preferências dos usuários (software, sistemas operacionais, tipo do ataque, entre outros).

Para finalizar, nas respostas de situação de uso, observou-se que 60% dos administradores seguiriam uma recomendação com menos de 10 votos positivos. Além disso, independente do número de votos, mais de 90% consultariam outras fontes antes de realizar algum tipo de ação. Isso mostra que é importante o sistema realizar uma ligação direta com a fonte original da notícia e que novas informações possam ser adicionadas ao alerta para aumentar o seu nível de confiança.

### 4.4.3 Especificação do modelo

Esta seção especifica o modelo de transações, o fluxo de processamento de alertas, e os algoritmos de recomendação usados no projeto do recomendador.

O modelo de transações do recomendador de alertas cibernéticos considera cinco transações, conforme apresentado na Figura 4.13.

A transação *visualiza* registra a visualização de um item por um usuário. No presente modelo, é usada para estimar os alertas mais populares e/ou para evitar que ocorra a recomendação de um alerta que já foi visualizado por um usuário.

A transação *categoriza* registra uma nova categoria para um alerta. Essa transação viabiliza a adição de termos para melhorar a contextualização do alerta. Pode ser realizada pelo usuário ou por técnicas de mineração de textos. No presente modelo, é usada para otimizar a filtragem baseada

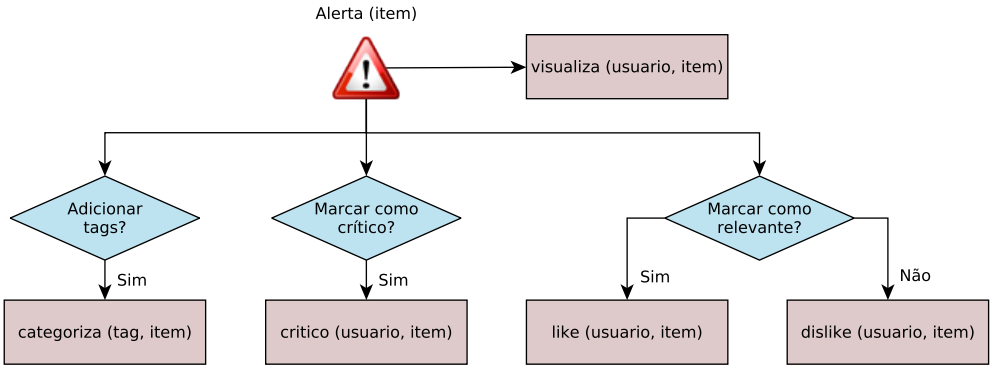


**Figura 4.13:** *Modelo de transações do recomendador.*

em conteúdo.

A transação *critico* registra a opinião do usuários sobre a urgência ou severidade de um alerta. No presente modelo, é usada para a geração da classificação geral e, segundo o resultado do questionário, deve possuir um peso maior que os outros mecanismos de avaliação.

A transação *like* registra uma relação positiva entre o usuário e o item. Essa transação indica que o alerta é relevante para o administrador ou foi ratificado como um alerta. No presente modelo, é usada para: (i) estimar os alertas mais populares na classificação dos usuários, ou seja, na recomendação geral; (ii) recomendar novos alertas usando filtragem colaborativa, pois possibilita inferir similaridade entre os usuários; e (iii) recomendar novos alertas usando a filtragem baseada em conteúdo, pois possibilita definir os interesses de um usuário por meio das categorias do alerta.

A transação *dislike* registra uma relação negativa entre o usuário e o item. Essa transação indica que o alerta não é relevante para o administrador ou simplesmente não é um alerta. No presente modelo, é usada para: (i) calcular a classificação geral; e (ii) remover falsos alertas da lista de relevantes. Desempenha um papel importante no fluxo de extração de alertas antecipados de fontes de dados não estruturados por possibilitar a eliminação de falsos positivos que não foram removidos nas fases de pré-processamento e na classificação.

A Figura 4.14 apresenta o fluxo de processamento do recomendador quando novos alertas são apresentados ou avaliados por um usuário.

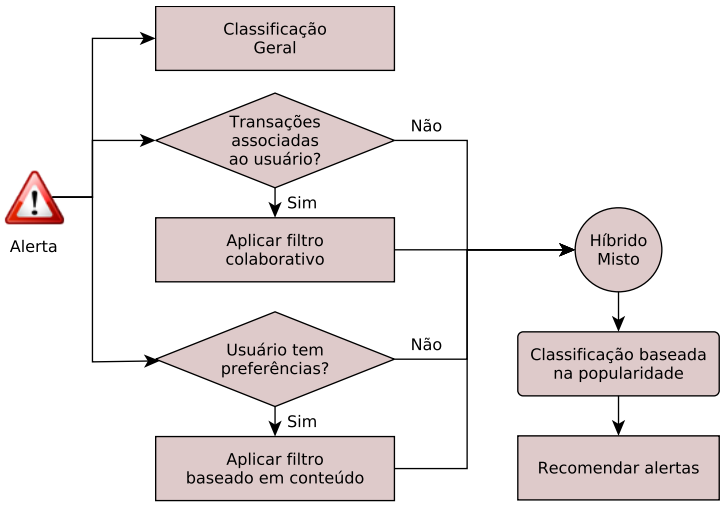


**Figura 4.14:** *Fluxo de processamento do recomendador.*

Como pode ser observado na Figura 4.14, sempre que temos itens avaliados, deve-se executar três processamentos separados: (i) classificação geral; (ii) filtragem colaborativa; e (iii) filtragem baseada em conteúdo. Os resultados de (ii) e (iii) são usados na recomendação híbrida mista.

A classificação geral define a posição de um alerta em relação a todos os outros alertas sem considerar informações de usuários. No presente modelo, é baseada na média Bayesiana, como definida na Equação 4.2, a qual incorpora informações existentes associadas aos alertas para minimizar o impacto de variações grandes provenientes das votações. Quanto maior o valor de $ranking\_score(i)$, melhor classificado será um item $i$.

$$ranking\_score(i) = \frac{(\bar{v} \cdot \bar{r}) + (v_i \cdot r_i)}{\bar{v} + |r_i|} \tag{4.2}$$

onde $\bar{v}$ é a média do número de votos (like, dislike ou critico) entre todos os alertas de um período; $\bar{r}$ é a média de avaliação de todos os alertas; $v_i$ é o número de votos de um item $i$; e $r_i$ é a avaliação do item $i$ segundo a Equação 4.3.

$$r_i = \alpha c_i + \beta l_i - \phi d_i \tag{4.3}$$

onde $c_i$ é o número de votos que indicam a criticidade do alerta, $l_i$ é o número de votos que indica a aprovação (like) do alerta, e $d_i$ é o número de votos que indica desaprovação (dislike) do alerta. As constantes $\alpha$, $\beta$ e $\phi$ são pesos que podem variar conforme a implementação do modelo. Para manter a consistência com os resultados obtidos com a análise do questionário (Seção 4.4.2), sugere-se o maior peso para $\alpha$ que está associada a criticidade e, em seguida, para $\phi$, que está associada a alertas irrelevantes.

Quando não há itens avaliados no sistema, a classificação geral apresenta sempre os itens mais recentes.

A filtragem colaborativa consiste em determinar a similaridade entre usuários e recomendar itens ao usuário que foram avaliados positivamente por outros usuários com perfil similar.

Primeiramente, calcula-se a similaridade entre usuários para encontrar os vizinhos (*neighbors*), que são os usuários que apresentam preferências passadas similares a um usuário $u$. Foi escolhido o coeficiente de Jaccard (Equação 4.4) para calcular a similaridade entre dois usuários $u$ e $v$, pois é adequado para votos binários.

$$J(u,v) = \frac{|\ U_l \cap V_l\ |}{|\ U_l \cup V_l\ |} = \frac{|\ U_l \cap V_l\ |}{|\ U_l\ | + |\ V_l\ | - |\ U_l \cap V_l\ |} \tag{4.4}$$

onde $U_l$ são os itens recomendados pelo usuário $u$ e $V_l$ são os itens recomendados pelo usuário $v$.

O resultado da Equação 4.4 é um valor entre 0 e 1. Quanto maior o coeficiente de Jaccard, maior a similaridade entre dois usuários. Seleciona-se os $k$ vizinhos de maior similaridade com o usuário $u$, o que é denotado por $Neighbor(u)$. Em seguida, combina-se as preferências dos vizinhos para recomendar os itens para o usuário $u$. Para tal, calcula-se uma pontuação para medir se o usuário $u$ pode ter interesse em um item $r$, conforme a Equação 4.5 (Zheng e Li, 2010). Na Equação 4.5 não são considerados os itens já visualizados pelo usuário $u$.

$$score(u,r) = \frac{\sum_{v \in Neighbor_{(u)}} R_{v,r} \times J(u,v)}{|\ \sum_{v \in Neighbor_{(u)}} J(u,v)\ |} \tag{4.5}$$

onde $R_{v,r}$ é a avaliação do item $r$ pelo usuário $v$.

A filtragem baseada em conteúdo considera a relação entre os pesos das categorias (*tags*) que indicam as preferências do usuário e as categorias dos itens. Sempre que um usuário recomenda um item, as categorias associadas ao item são adicionadas aos interesses do usuário e seu peso é alterado para 1, se é uma nova categoria, caso contrário, é incrementado em 1. A Equação 4.6, adaptada do

método de Zheng e Li (2010), calcula a relação das categorias de interesse de um usuário $u$ com relação as categorias que descrevem um item $i$, isto é, mede o interesse de um usuário $u$ por um item $i$. A Equação 4.6 resulta em um valor entre 0 e 1, onde o valor mais alto indica maior interesse de um usuário $u$ por um item $i$.

$$w_{tag}(u,i) = \frac{\sum_{t_i \in tag(i)} weight(u,t_i)}{\sum_{t_j \in tag(u)} weight(u,t_j)} \tag{4.6}$$

onde *tag(i)* representa o conjunto de categorias especificadas para um item $i$; *tag(u)* representa o conjunto de categorias de interesse de um usuário $u$; $weight(u,t_i)$ indica o peso da tag $t_i$ para o usuário $u$, e $weight(u,t_j)$ indica o peso da tag $t_j$ para o usuário $u$.

A filtragem colaborativa deve ser executada periodicamente em um intervalo de tempo pequeno ou sempre que um alerta é avaliado. A filtragem baseada em conteúdo deve ser executada sempre que um novo alerta é gerado no sistema ou novas categorizações são adicionadas a um alerta. A abordagem híbrida mista é responsável por combinar a saída das duas categorizações para identificar alertas recentes e os alertas considerados relevantes segundo ambas as abordagens. Para evidenciar os alertas importantes, a Equação 4.7 combina o resultado de todas as saídas.

$$mixed\_score(u,i) = 2^{\gamma} \cdot ranking\_score(i) \tag{4.7}$$

onde $\gamma$ é dois se o item $i$ é recomendado em ambas abordagens, um se o item $i$ é somente recomendado por uma das abordagens e zero se não é recomendado por nenhuma das abordagens.

Dessa forma, os alertas com maior pontuação na classificação são os selecionados para serem recomendados. O número de alertas a serem recomendados pode ser um parâmetro definido pelo usuário.

### 4.4.4 Avaliação dos algoritmos e do modelo de recomendação

Os algoritmos para as abordagens de filtragem colaborativa e baseada em conteúdo foram avaliados separadamente usando a base do MovieLens (Harper e Konstan, 2015), obtida em uma sítio Web dedicado a pesquisa em sistemas de recomendação para filmes baseado nas avaliações de usuários. Mesmo não sendo uma base de alertas, possibilitou imitar a ideia de categorias e votos para avaliar as fórmulas de recomendação.

Foram selecionados aleatoriamente um conjunto de dados com 207 usuários, 1287 itens e 11088 classificações para compor a base de avaliação. O conjunto de treinamento correspondia a 80% da base e o conjunto de testes correspondia a 20% da base. Os algoritmos foram avaliados considerando o problema top-N, que corresponde identificar os N itens que serão de interesse para um usuário (Karypis, 2001). Na avaliação foram considerados $N \in \{3, 5, 10, 15\}$.

Um software foi desenvolvido para a realização dos experimentos, denominado de Konsilo, que está disponível sob a licença AGPL v3 em https://gitlab.com/konsilo/konsilo. As Figuras 4.15 e 4.16 mostram os resultados segundo as métricas de precisão e abrangência.

Como pode ser observado na Figura 4.15, a filtragem colaborativa obteve uma precisão aceitável com pouco variação sob diferentes valores de $N$. Por outro lado, a filtragem baseada em conteúdo teve uma redução ao aumentar o valor de N. Acredita-se que apenas a categorização de filme não seja um atributo que consiga capturar o interesse do usuário, ou seja, não é porque um usuário tem interesse em filme de terror que gostará de todos os filmes na categoria terror. Porém, para alertas, a premissa de categorias é mais influente, pois um administrador de redes que possui um software

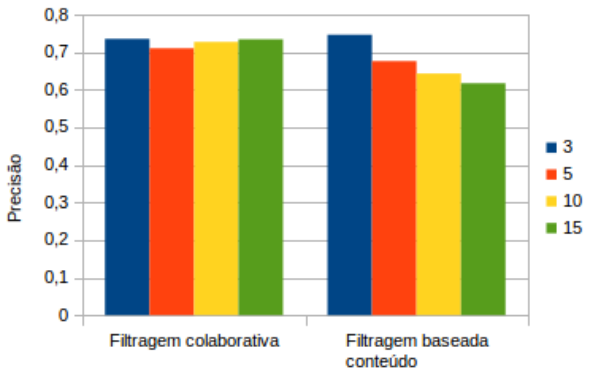


**Figura 4.15:** *Medições de precisão segundo o número de recomendações (N).*

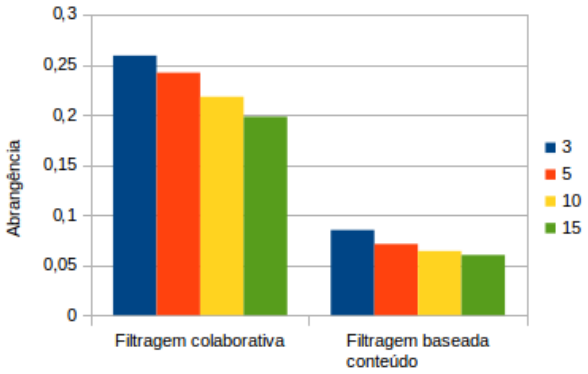


**Figura 4.16:** *Medições de abrangência segundo o número de recomendações (N).*

Y, está interessado em alertas que estão associadas a ameaças e atualizações sobre esse software Y.

Quanto a abrangência, observa-se na Figura 4.16 que o desempenho foi baixo. Acredita-se que se deve a quantidade de itens disponíveis que poderiam ser recomendados no experimento, desde que resultados similares foram encontrados em outros experimentos relacionados ao problema top-N (Cremonesi *et al.*, 2008; Herlocker *et al.*, 2004). Isso não invalida o modelo, pois estamos interessados em recomendar alertas recentes que sejam de interesse para o usuário, logo o número de alertas de interesse deve ser reduzido para cada usuário.

## 4.5 Discussões sobre o arcabouço

Nesta seção, são discutidas a avaliação de cada um dos componentes do arcabouço EWS segundo os experimentos e resultados das investigações realizadas neste capítulo.

**Fontes de dados** As fontes de dados investigadas mais detalhadamente foram o Twitter e as redes IRC. No Twitter, foram identificadas tendências de ameaças, como no caso da vulnerabilidade Venom, que rapidamente foi destacada pelos métodos no início de sua divulgação. Também foram identificados divulgação de atualizações críticas de software e possíveis vulnerabilidades dia zero. Em especial, em língua portuguesa, foram encontrados alertas antecipados associados a orquestrações de ataques, vazamentos, vulnerabilidades e ferramentas de exploração. Nas redes IRC, foram encontrados compartilhamento de informações de segurança entre especialistas que auxiliam na proteção de ameaças emergentes, mas principalmente, nos canais *hackers*, a divulgação de potenciais alvos e ataques que estavam sendo realizados. Outras fontes também foram investigadas, Facebook, Pastebin, Zone-h, Blogs, CVE, Fóruns Hackers e WhatsApp, mas não foram realizados experimentos específicos. No entanto, em todas essas fontes foram encontradas evidências concretas que indicam ameaças à segurança de sistemas (ver Tabela 3.1).

**Coletores** Os coletores tiveram de ser personalizados para cada fonte de dados e necessitam se adaptar constantemente as mudanças de *Application Programming Interface* (API), regras de monitoramento e medidas anti-coleta (*anti-crawling*). Em nossos experimentos, o coletor do Twitter foi modificado três vezes nos últimos quatro anos. Constatamos que as coletas em comunidades *hackers* são mais complexas, principalmente por operações de ataques serem discutidos em canais privados e pelo uso de medidas anti-coleta, por exemplo, detecção de monitoramento, autenticação, desafios (captcha) e ausência de interação. Essas questões foram observadas no monitoramento das redes IRC. Como lições aprendidas: (i) no momento da coleta, é necessário também fazer o mapeamento de recursos ligados com o conteúdo da mensagem, por exemplo, o registro forense de

vazamentos de dados, páginas desfiguradas ou de código malicioso; (ii) uso de medidas anti-coleta desde que não infrinjam questões de privacidade da fonte; (iii) rever constantemente as especificações de coleta e adicionar novos perfis e palavras-chave para aumentar a qualidade dos dados.

**Normalizadores** Os normalizadores implementados adicionaram informações aos dados coletados, como a identificação de entidades e termos em mensagens nos experimentos com a rede IRC. O principal componente de normalização implementado foi o de tradução de URLs curtas para longas, que possibilitou a realização de agrupamentos e filtros no pré-processamento e a evidenciação de alertas no processamento de alertas. No entanto, também verificou-se que expandir informações a partir do linguagem informal nos canais, pode gerar informações inconsistentes (ver último parágrafo na Seção 4.2.3.5). Como lições aprendidas: (i) identificar entidades (software, URL, IP) são características importantes para a identificação de mensagens relevantes como alertas, principalmente se associadas com termos de cibersegurança; (ii) armazenar a mensagem original e a normalizada na construção de bases de dados facilita o processamento dos módulos seguintes e evita a perda de informação associada ao contexto no instante da coleta.

**Filtros** Os filtros implementados visaram principalmente a remoção de mensagens irrelevantes dos dados. Foi verificado que usar as URLs de sítios de notícias genéricos e termos que são comuns em notícias de segurança de outras áreas, possibilita o descarte significativo de mensagens a partir de filtros com listas negras. No Twitter, foi verificado que há diversas características que podem ser usadas para descartar mensagens: forma de propagação, tamanho da mensagem, número de caracteres ou termos, entre outros. No IRC, os filtros de priorização são mais efetivos que os de remoção, visto que em conversações a maioria das mensagens são irrelevantes, exceto as que apresentam evidências de assuntos sobre cibersegurança. Como lições aprendidas: (i) filtros construídos a partir da análise da fonte reduzem significativamente a quantidade de dados a serem transmitidos ou processados; (ii) filtros precisam ser adaptativos em mídias sociais para evitar a sobrecarga de mensagens inúteis quando acontece eventos sem importância (p. ex. invasão de contas de artistas famosos).

**Agrupadores** Os agrupadores implementados diminuíram o número de mensagens para o processamento, mas também geraram vários grupos contendo o mesmo assunto. É difícil lidar com essa questão devido a natureza não estruturada dos dados e a forma que cada indivíduo e mídia publicam a informação. Os mecanismos de agrupamento no pré-processamento foram simples, em geral, considerando similaridade e a URL. Não foi avaliada uma janela temporal ideal para a junção de mensagens no mesmo grupo, até porque uma notificação de ataque similar pode ser publicada várias vezes em um curto período. Apesar de não terem sido realizados testes, a identificação das entidades e o contexto das mensagens de cibersegurança são pontos que poderiam ter sido explorados para otimizar o agrupamento.

**Analisadores** Os analisadores foram providos no arcabouço para estender as funcionalidades de processamento de alertas para fins específicos, como identificar tendências de ameaças no Twitter ou análise de contexto nas redes IRC. O analisador implementado no Twitter utilizou uma abordagem de agrupamento com finalidade de identificar notícias de cibersegurança que ganharam notoriedade rapidamente e de filtrar de notícias irrelevantes por meio de heurísticas. O analisador implementado nas redes IRC gerou características de contexto que foram usadas por um algoritmo de classificação baseado em ponderamento e, assim, foi possível identificar conversações de interesse. No contexto de alertas antecipados, os analisadores podem ser usados para confirmar ataques em andamento, rumores de ameaças e prover processamento específico para informações de uma dada fonte, como foi o caso dos analisadores apresentados nos Estudos 1 e 2.

**Recomendadores** O uso de recomendação procurou contemplar a colaboração direta de especialistas em segurança na identificação de ameaças e, também, o compartilhamento de informações para que um outro administrador não sofra um ataque já relatado por outros. No modelo de recomendação proposto, procurou-se agregar os interesses individuais de cada administrador e também possibilitar o destaque de ameaças relatadas como importantes por administradores de interesses comuns. Indiretamente, a abordagem de recomendação também auxilia na remoção de informações inúteis como alertas ou alertas antecipados. A construção do modelo procurou consultar a comunidade de segurança no Brasil, para identificar como seria a interação com um sistema de recomendação. Apesar do número restrito de respostas (44), foi evidenciado o interesse da comunidade nacional em aumentar a colaboração para a proteção das infraestruturas de redes. Já na avaliação do modelo, foi necessário a criação de uma base sintética visto que não há bases de alertas cibernéticos com essa finalidade. Nesta tese, a recomendação de alertas é defendida como um ponto chave para sucesso de um sistema de alerta antecipado de cibersegurança.

**Classificadores** O estudo conduzido para a implementação de classificação teve a finalidade de mostrar como um modelo genérico pode ser usado para decidir se deve ser gerado um alerta ou não. Como resultado, um modelo simples foi construído e alguns algoritmos avaliados. Os algoritmos de Árvore de Decisão (RandomForest) e Entropia máxima obtiveram os melhores resultados. No entanto, verificou-se uma alta taxa de falsos positivos na maioria dos classificadores e combinações de características avaliados. De certa forma, não foi um resultado inesperado, pois pela amostragem de dados apresentada na Tabela4.21, verificou-se que há muitas mensagens que possuem muitos termos associados à cibersegurança, mas o contexto delas não é relevante como um alerta.

**Correlacionadores** Os correlacionadores desempenham papel fundamental no arcabouço para possibilitar a ligação entre alertas de sensores tradicionais com os alertas obtidos de informações não estruturadas. Não foram avaliados por meio de implementação nesta tese, devido a complexidade de prover tal mecanismo. No entanto, um exemplo simples de correlação é relacionar os alertas para aumentar a prioridade de alertas emitidos por sistemas de detecção de intrusão. Por exemplo, se um alerta de tráfego ou acesso suspeito é realizado em um servidor Web K, e há notificações indicando vulnerabilidades novas que ameaçam o servidor Web K, logo o alerta deve ser priorizado. Os trabalhos de (More *et al.*, 2012; Rodrigues, 2012) corroboram a importância de minerar informações na Web e relacionar a alertas tradicionais e mostram a viabilidade de implementação de módulos para essa finalidade.

**Notificadores** A proposta de notificadores engloba a definição de políticas de privacidade para compartilhamento dos alertas. Essas políticas são importantes especialmente se o alerta afeta a reputação dos parceiros ou software de terceiros. Isso porque a notificação de novas vulnerabilidades de um produto de software não é de interesse dos provedores do produto e organizações não querem admitir comprometimento de suas infraestruturas de tecnologia da informação. No entanto, há relatos que muitas notificações de vulnerabilidades relatadas diretamente a fabricantes continuam sem correção por longos períodos. No arcabouço EWS, há a preocupação com a política de notificação e privacidade, mas não é descrito um processo para a realização de tais notificações. Acredita-se que o interesse dos envolvidos, a criticidade do alerta e o nível de colaboração dos parceiros é que afetam diretamente esses componentes. Em geral, para evitar falsas notificações, no arcabouço foi proposto o uso de recomendação e a revisão de especialistas.

**Análise de dados** A análise de dados consistiu em estabelecer um conjunto de processos para a identificação de informações de inteligência para viabilizar a extração de alertas das fontes. Apesar de um processo de análise de dados ser algo genérico a qualquer área e livros textos da área de Mineração de Texto fornecerem uma quantidade extensiva de formas de ser realizada, no arcabouço

procurou-se identificar e contextualizar processos básicos voltados à cibersegurança. Nesse contexto, verificou-se que a análise de frequência e a associação de palavras devolvem termos para elaboração de vocabulário na área. Também foi verificado que a análise de agrupamento possibilita isolar termos de cibersegurança quanto termos associados a mensagens sem relevância como alertas. Outros processos, como o uso de heurísticas e análise por especialista, possibilitam identificar padrões para a implementação de algoritmos usados para filtrar ou priorizar informações. A análise de dados como um processo é um dos pontos positivos do arcabouço EWS.

**Bases de dados** A geração de bases de dados para a pesquisa na área de cibersegurança é uma das contribuições desta tese, visto que não há bases de mídias sociais para essa finalidade. Foram coletadas e geradas bases de fontes como Twitter, redes IRC, Facebook e de blogs de segurança. Nesta tese, foram exploradas as duas primeiras. O Capítulo 5 apresenta a prova de conceito do arcabouço por meio da implementação de um sistema e, uma das consequências, é a geração de bases etiquetadas com informações de cibersegurança coletadas na Web. O arcabouço está sendo empregado em duas pesquisas que também estão gerando novas bases de dados, no caso, para o Pastebin e Zone-h.

**Bases de inteligência** As bases de inteligência constituem outra contribuição direta desta tese, em especial, as bases para extração de alertas no idioma português. Na pesquisa do Twitter, foram geradas listas brancas e negras de URLs e termos que, quando usadas por filtros, diminuem significantemente o número de mensagens a serem processadas como potenciais alertas. O agrupamento de informações possibilitou identificar um conjunto de características para a implementação de analisadores e classificadores para o Twitter e o IRC. Foi verificado que bases produzidas a partir de uma fonte podem ser usadas em outras, como foi o caso do uso da base de termos spam para filtrar URLs no IRC.

**Bases de alertas** Quando se trata de alertas antecipados, a principal preocupação é descobrir a validade da informação como alerta antecipado, lembrando que o antecipado significa avisar parceiros para não serem comprometidos por uma mesma ameaça ou reagir proativamente a uma ameaça eminente. Em nossa pesquisa não realizamos esse tipo de distinção, logo, a base contém ambos. Outra questão é quanto a precisão dos alertas, ou seja, quantos são verdadeiros e falsos na base. Pelos experimentos, verificamos que há ainda um número moderado de alertas, até pela dificuldade de discernir notificações de segurança de outras mensagens. Uma questão observada em experimentos é identificação de notícias antigas como alertas. Logo, a colaboração por administradores e o uso de mecanismos para identificar quão antiga são essenciais em um EWS.

## 4.6 Considerações finais

Este capítulo apresentou a aplicação do arcabouço EWS para a investigação de duas fontes distintas de dados: microblog Twitter e a rede IRC. Foi verificado que ambas as fontes contém informações relevantes para a geração de alertas antecipados. No Twitter, foi proposta uma abordagem que possibilita a agregação de notificações de segurança para identificar ameaças e tendências de ameaças. Nas redes IRC, foi constatado que são usadas para a disseminação de conhecimento *hacker*, potenciais alvos e orquestrações de ataques. Foram propostos modelos de recomendação e classificação direcionados a seleção e classificação de alertas. Procurou-se implementar cada uma das propostas do arcabouço, no entanto, não foi apresentada uma prova de conceito sobre os correlacionadores e notificadores. No capítulo 5, os resultados obtidos com os experimentos são aplicados para a construção de um sistema de alerta antecipado voltado à cibersegurança, que implementa os componentes e o fluxo de processamento definidos no arcabouço EWS.

# Capítulo 5

# Um Sistema de Alerta Antecipado de Cibersegurança

Este capítulo apresenta o desenvolvimento, implantação, resultados e avaliação de um Sistema de Alerta Antecipado de Cibersegurança - CEWS - que monitora e processa informações coletadas de fontes de dados não estruturados e identifica potenciais alertas associados à segurança de redes de computadores e sistemas computacionais. O sistema foi desenvolvido como prova de conceito para o arcabouço proposto nesta tese, resultado da aplicação direta dos conceitos e processos definidos pelo arcabouço EWS. Além disso, o sistema foi desenvolvido com o apoio financeiro e de infraestrutura da RNP e está sendo implantado como um projeto piloto para o Centro de Atendimento a Incidentes de Segurança (CAIS), que é o grupo de resposta a incidentes da rede acadêmica da RNP.

## 5.1 Histórico

A proposta de desenvolvimento do CEWS foi aprovada nos Grupos de Trabalhos da RNP 2014-2015, Fase 1, com a temática *Mecanismos para um Sistema de Alerta Antecipado*, e o grupo foi denominado de *GT-EWS*. Os membros do grupo eram compostos por membros das instituições parceiras no projeto: Universidade de São Paulo (USP), Universidade Tecnológica Federal do Paraná (UTFPR) e Universidade Federal da Bahia (UFBA). Nesta fase, foi desenvolvido um protótipo de serviço para a detecção de alertas de novas vulnerabilidades publicados no Twitter e de ameaças à infraestrutura de rede das instituições parceiras da RNP divulgadas no Twitter e Facebook. Também foi desenvolvido um estudo de integração de informações de ferramentas de monitoramento tradicionais de redes.

Na Fase 1, foram identificados alertas publicados em redes sociais que divulgavam desfigurações de páginas e vazamento de dados de várias instituições brasileiras. Como esses alertas possibilitavam identificar o problema no momento que se tornavam públicos, foram importantes para agilizar o processo de resposta a incidentes. Também foi observado no sistema, a evidenciação e priorização de informações de segurança que representavam alto grau de risco às infraestruturas de redes. Isso foi possível devido ao mecanismo que priorizava mensagens de segurança muito discutidas na comunidade de segurança.

Devido aos resultados obtidos na Fase 1, O GT-EWS foi aprovado para a Fase 2 dos Grupos de Trabalhos da RNP 2015-2016, que consiste na implantação de piloto em instituições parceiras. O projeto foi redefinido para atender a necessidade imediata de monitoramento de fontes de dados não estruturados do CAIS. Além disso, novos parceiros se interessaram em implantar o projeto: Processamento de Dados do Amazonas (PRODAM) e a Polícia Federal.

Na Fase 2, o desenvolvimento e implantação do piloto exigiram a reestruturação do projeto para

possibilitar a agregação de qualquer novo tipo de fonte de dados não estruturados, a elaboração de uma nova interface gráfica e a implementação de mecanismos para a diminuição de falsos positivos. Considerando esses requisitos, o arcabouço EWS foi importante para o desenvolvimento e adaptação da arquitetura, especificações dos padrões de mensagens e protocolos de comunicação entre os componentes e, principalmente, para nortear o processo de desenvolvimento de mecanismos para a diminuição de falsos positivos.

## 5.2 Arquitetura do Sistema

O CEWS é um sistema de software voltado à detecção antecipada de orquestrações de ataques, vazamentos de dados, desfigurações de páginas Web, vulnerabilidades de software, novas ameaças (ferramentas, códigos de exploração, entre outros) por meio do monitoramento de fontes de dados não estruturados, em especial, redes sociais e microblogs.

A Figura 5.1 mostra a distribuição e interação dos componentes de software por meio de uma arquitetura orientada a serviço.

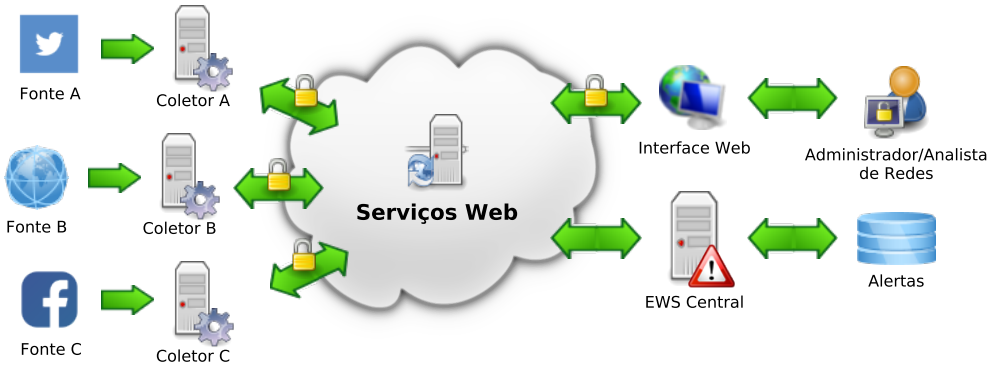


**Figura 5.1:** *Visão geral da arquitetura do CEWS*

Os coletores são independentes e implementados para monitorar fontes de dados específicas. Comunicam-se com o sistema central (EWS Central) por meio de serviços Web usando um canal criptografado e um protocolo especificado pelo CEWS. O monitoramento de cada fonte de dados ocorre pela instanciação de um serviço de coleta (Coletor) específico para a fonte. Também é possível a instanciação de serviços independentes para monitorar a mesma fonte. Isso provê um meio de atender monitoramentos específicos ou sazonais sem modificar a configuração dos coletores em execução. Outra característica é a possibilidade de replicação os coletores, garantindo disponibilidade de serviço e evitando pontos de falhas.

Os serviços Web proveem a interface para a comunicação entre os componentes do sistema. Toda a comunicação é criptografada e realizada por estruturas e protocolos padronizados. Os principais fluxos de comunicação são entre os coletores e o EWS central e o EWS central e a Interface Web. Os fluxos secundários são serviços que podem ser agregados ao sistema para prover novas funcionalidades. Esses serviços podem ser usados pelos coletores para realização de pré-processamento de informações e pelo EWS Central para a realização de processamento de alertas.

O EWS Central centraliza os serviços e realiza a orquestração do processamento dos alertas. Possui módulos para gerenciamento dos coletores remotos, processamentos dos alertas e notificações para as entidades de interesse. Novos componentes do EWS, como analisadores, classificadores ou notificadores, podem ser agregados como serviços independentes ou módulos no EWS Central. Os alertas possuem um formato padronizado e são armazenados em um banco de dados relacional.

A Interface Web é responsável por prover uma interface gráfica para as funções de gerenciamento e visualização de alertas, e também para a configuração e gerenciamento dos componentes

distribuídos do sistema. O administrador de redes ou especialista em segurança interage e colabora diretamente com o sistema via a Interface Web. O administrador pode acrescentar ou corrigir informações de alertas e também colaborar com a evidenciação de alertas ou descarte de falsos positivos.

Os coletores enviam para o EWS Central as informações coletadas e pré-processadas em um formato estruturado. Também enviam periodicamente informações do estado de operação. O EWS Central interage com os coletores provendo autenticação e configurações para o monitoramento e módulos dos coletores. A interface Web acessa periodicamente o EWS Central para atualizar os alertas e acessar informações sobre o estado do sistema. O EWS Central pode notificar alertas por outros meios, como correspondência eletrônica. Os alertas são estruturados em um formato padrão com atributos resultantes dos processamentos desde a coleta até a interação com os usuários do sistema.

## 5.3 Componentes do Sistema

Esta seção apresenta individualmente os componentes do sistema e como foram desenvolvidos considerando a estrutura do arcabouço EWS.

### 5.3.1 Coletor TwitterSearch

O coletor TwitterSearch monitora e coleta as mensagens públicas postadas no microblog Twitter. Também é responsável por realizar o pré-processamento das mensagens coletadas, por exemplo, filtragem, normalização e priorização. As mensagens caracterizadas como potenciais alertas são enviadas para o EWS Central em um formato padronizado. A Figura 5.2 ilustra o fluxo de processamento realizado pelo coletor.

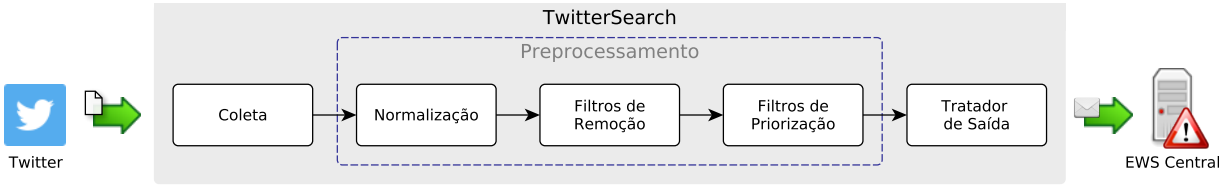


**Figura 5.2:** *Módulos e fluxos de processamento do TwitterSearch*

O processo Coleta realiza o monitoramento baseado em expressões de busca compostas por palavras-chave e/ou perfis combinadas por operadores lógicos ou de atributos. As palavras-chave e os perfis foram obtidos a partir da análise de mensagens associadas à cibersegurança postadas no Twitter e em colaboração com especialistas de segurança. O software foi desenvolvido em Java 1.8 com a API Twitter4J[1], que provê uma interface para acessar via a linguagem de programação Java as APIs do Twitter.

A coleta possui dois modos de execução: *basic* e *stream*. No modo *basic*, as consultas são realizadas usando a API Rest do Twitter, logo devolvem um conjunto de tweets previamente indexados e dentro de período de busca específico. No modo *stream*, as consultas são realizadas usando a API Stream do Twitter, logo devolvem um percentual de *tweets* que estão sendo publicados no momento da consulta. O modo stream possibilita monitorar em tempo real os perfis de usuários associados a atividades hackers.

O processo Normalização consiste na remoção de ruídos (p. ex. caracteres não ASCII), tradução de URLs curtas para longas, identificação dos termos detectados pela consulta, identificação de entidades nomeadas, marcação da estrutura sintática das mensagens, criação e organização das

[1] http://twitter4j.org

informações normalizadas e estendidas para uma estrutura de dados padronizada. Novos módulos de normalização, isto é, novos algoritmos ou técnicas de processamento, podem ser acoplados em tempo de execução e podem ser estruturados de forma linear (pipeline) ou hierárquica.

O processo Filtros de Remoção consiste na remoção de mensagens irrelevantes como alertas. Há alguns filtros implementados por padrão que usam informações obtidas da análise da fonte, por exemplo, filtros que usam listas negras para URLs e palavras-chave. Outros filtros podem ser personalizados, por exemplo, filtros que consideram um conjunto de entidades ou termos detectados e realizam a ponderação para identificar mensagens irrelevantes. O processo Filtros de Priorização consiste na priorização de mensagens e pré-classificação de potenciais alertas. O filtro padrão realiza uma avaliação ponderada das entidades e termos detectados na mensagem, por meio de uma árvore de decisão simples, que avalia a mensagem e remove se não atingir um limiar predefinido. Novos filtros de remoção e priorização podem ser acoplados em tempo de execução e estruturados de forma linear ou hierárquica.

O processo Tratador de Saída é responsável por realizar o empacotamento e envio de informações para o EWS Central, outros serviços do sistema e/ou armazenamento de mensagens. Além do importante papel de comunicação entre as partes do sistema, pode implementar uma política local de privacidade, isto é, pode avaliar se há informações sensíveis que estão sendo encaminhadas para o EWS Central que deveriam ser anonimizadas ou descartadas.

Como pode ser observado na Figura 5.2, o fluxo de informações e módulos do TwitterSearch são resultados da aplicação direta do arcabouço EWS. Observa-se os componentes de coleta e pré-processadores que objetivam além de monitorar mensagens de interesse, diminuir o número de mensagens a serem tradadas pelo EWS Central e inspeção de especialistas em segurança. Além disso, a implementação segue as recomendações do arcabouço de possibilitar a adição e organização de novos módulos de filtros, isto é, aplicação de filtros adaptativos, especialistas e hierárquicos.

Há duas instâncias do TwitterSearch em execução no CEWS, a TwitterSearch(en) que monitora termos em inglês associados à cibersegurança usando o modo *basic* e a TwitterSearch(ptbr) que monitora termos em português e perfis associados a atividades hackers no Brasil usando o modo *stream*. Essa última instância é a principal provedora de dados para a atual implementação do CEWS.

### 5.3.2 Coletor FacebookSearch

O FacebookSearch monitora e coleta as mensagens postadas na rede social Facebook. São monitoradas postagens em páginas, grupos e usuários. Também realiza o pré-processamento das postagens da mesma forma que o TwitterSearch. A Figura 5.3 ilustra a operação do coletor FacebookSearch.

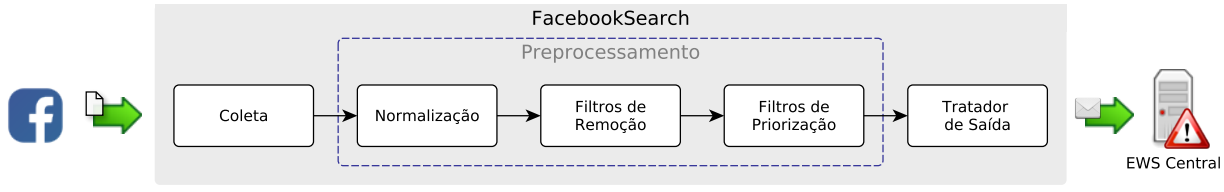


**Figura 5.3:** *Módulos e fluxos de processamento do FacebookSearch*

Os módulos e processos do FacebookSearch são idênticos ao TwitterSearch, exceto pelo processo Coleta e adaptações realizadas no processo Normalização, em específico, o mapeamento inicial dos dados coletados para a estrutura padronizada.

O processo Coleta realiza o monitoramento de páginas, grupos e usuários suspeitos de postagens associadas a atividades hackers. Todas as postagens são capturadas em um intervalo periódico para serem avaliadas nos próximos processos. A coleta é realizada por um software desenvolvido em

linguagem Java e com o auxílio da API RestFB[2], que provê acesso a API Graph do Facebook.

O pré-processamento das mensagens reaproveitaram as bases de inteligência obtidas da análise do microblog Twitter e também os métodos para normalização e filtros com pequenas modificações. Isso foi possível devido a estruturação do código segundo as especificação do arcabouço EWS. No entanto, as restrições de monitoramento do Facebook são mais rígidas, o que inviabiliza um monitoramento global das postagens e, muitas vezes, forçam modificações no código do coletor para se adequar as políticas de privacidade da API provida pelo Facebook.

Há uma única instância do FacebookSearch que monitora páginas de organizações, grupos de discussão e usuários suspeitos de atividades hackers no Brasil. Apesar de usar apenas um único coletor, outras instâncias poderiam facilmente ser instanciadas devido a modularidade do software coletor.

### 5.3.3 EWS Central

O EWS Central é responsável por centralizar os serviços do CEWS e orquestrar o processamento e notificações de alertas. Processa as informações estruturadas e pré-processadas enviadas pelos coletores e devolve notificações de alertas que são enviadas para a Interface Web ou para e-mail das entidades de interesse. A Figura 5.4 ilustra os módulos e fluxos de processamento do EWS Central.

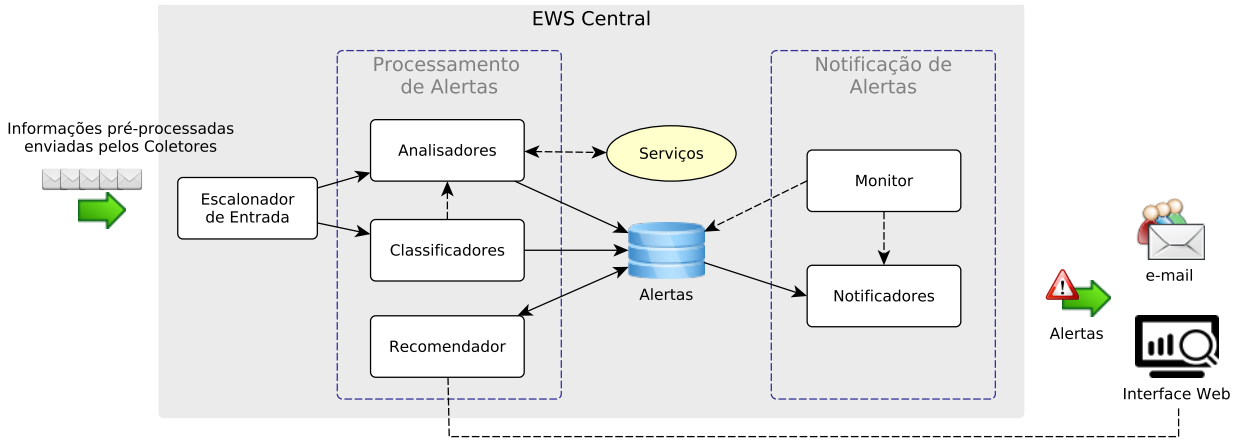


**Figura 5.4:** *Módulos e fluxos de processamento do EWS Central*

O processo Escalonador de Entrada implementa uma fila de entrada para receber as informações pré-processadas enviadas pelos coletores. Essas informações são os potenciais alertas descritos em uma estrutura padronizada. Por meio de uma função de decisão que considera a fonte e o tipo de informação, o escalonador seleciona se a estrutura padronizada deve ser encaminhada para os módulos de classificação (Classificadores) ou para os módulos de análise (Analisadores).

O processo Classificadores consiste de um conjunto de algoritmos de aprendizagem de máquina que analisa os alertas, classifica como alerta de cibersegurança ou não. Caso seja um alerta, atribuí um nível de confiança por um mecanismo de votação baseado no resultado da classificação de três algoritmos distintos. Há classificadores que podem solicitar uma análise do alerta, como recuperar informações do alvo ou outros tipos de dados. Por isso, nesses casos, um classificador pode invocar uma funcionalidade de um analisador.

O processo Analisadores consiste de um conjunto de módulos acopláveis ou de serviços independentes que realizam processamentos específicos nas informações. No CEWS há analisadores que agregam notícias de segurança publicadas no Twitter para evidenciar as que estão sendo mais comentadas e as que apresentam alto grau de risco. Também são implementados analisadores que

[2] http://restfb.com/

avaliam URLs que podem ter sido alvos de ataques e, realizam o registro no caso de uma desfiguração de página ou averiguam se endereço apontado está ativo. Outros analisadores podem avaliar se o conteúdo de uma URL pode ser um potencial vazamento de dados.

O processo Recomendador consiste de um sistema de recomendação de alertas de segurança para possibilitar a filtragem colaborativa de alertas e também para evidenciar os alertas mais importantes por meio da colaboração. Além disso, alertas podem ser modificados para incluírem novas informações ou para a remoção de informações inconsistentes resultantes das fases de pré-processamento de dados. Visando aumentar o grau de colaboração, uma funcionalidade prevista é possibilitar a postagem de alertas pelos próprios usuários do CEWS.

A base Alertas armazena os alertas padronizados que foram selecionados pelos analisadores e classificadores. Além disso, persiste as estruturas usadas na recomendação e os dados associados aos alertas, como capturas de telas de desfigurações de sítios Web e dados de vazamento.

O processo Notificadores é acionado após a identificação de novos alertas pelo processo Monitor. O processo Monitor inspeciona a base de alertas periodicamente para identificar a adição ou edição de novos alertas. Ao receber uma notificação de novo alerta, os notificadores verificam os interesses dos assinantes e, dependendo do nível de confiabilidade e da entidade interessada na notificação, disparam um processo de notificação por e-mail. Já a Interface Web só é atualizada após a solicitação de atualização pela própria interface, a qual também ocorre periodicamente.

Toda a orquestração e interação de serviços remotos é realizado por serviços Web. O EWS Central define as interfaces para acesso aos alertas e à fila de entrada do escalonador. Os módulos do EWS Central são responsáveis por orquestrar os diferentes serviços de análise e classificação de alertas. Além das funções de gerenciamento dos alertas, também provê mecanismos para possibilitar a expansão do sistema, como a adição de correlacionadores com fontes tradicionais.

Observa-se na figura 5.4, que o EWS Central implementa dois componentes do arcabouço EWS que são os processadores e os notificadores de alertas. Verificou-se na evolução do sistema que a estrutura e fluxo de informações que foram propostos no arcabouço EWS viabilizam a implementação e flexibilidade de adição de novos classificadores e analisadores.

### 5.3.4 Interface Web

A Interface Web possibilita a visualização e gerenciamento de alertas, personalização e monitoramento de operação dos sensores e gerenciamento dos usuários do sistema. A interface foi desenvolvida em Javascript e acessa as informações via serviços Web. Toda a interação com os outros componentes do sistema é realizada por canais criptografados por meio do protocolo *Hyper Text Transfer Protocol Secure* (HTTPS).

A Figura 5.5 apresenta uma visão da interface Web. É apresentada a visualização detalhada de um alerta. Verificam-se os termos detectados na mensagem, as opções de recomendação e os atributos do alerta. Também são mostradas as funcionalidades da interface dispostas no menu lateral esquerdo, por exemplo, resumo do sistema (Dashboard), configuração e operação dos coletores (Sensors), visualização de localidade das postagens que são alertas (Geolocation), entre outras.

### 5.3.5 Estruturas do sistema

As principais estruturas do sistema são a EWSMessage, que representa uma informação estruturada após a fase de pré-processamento e a EWSAlert, que representa um alerta identificado pelo sistema. Ambas as estruturas foram baseadas nos itens abordados nas Seções 3.4 e 3.8.

A estrutura EWSAlert possui todos os campos do EWSMessage acrescida com informações sobre agrupamento. No caso, um EWSAlert agrega uma lista de EWSMessage e novos atributos:

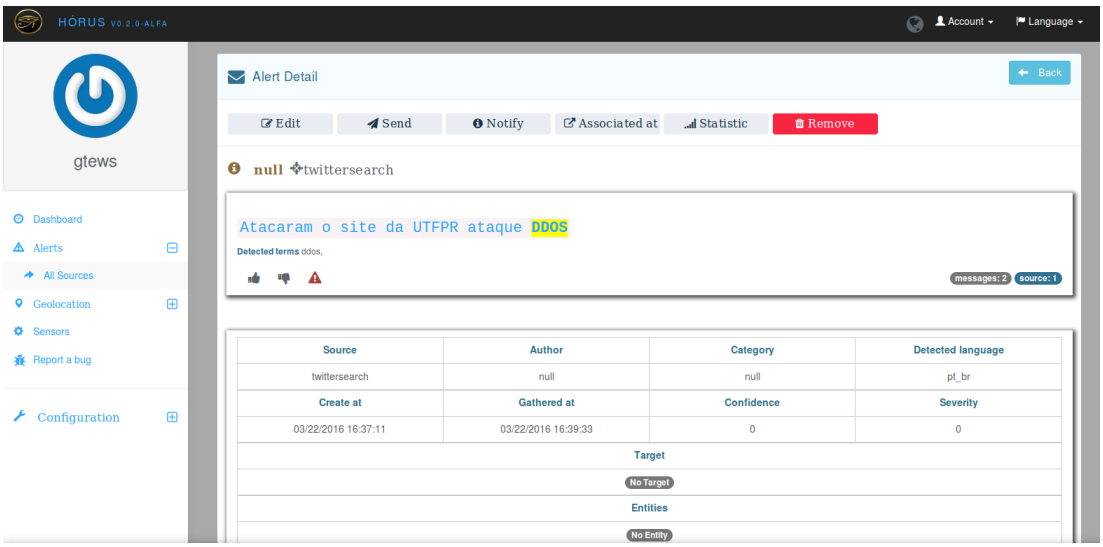


**Figura 5.5:** *Interface Web do CEWS*

número de fontes, número de mensagens e data do primeiro e último alerta. A Tabela 5.1 apresenta os atributos de EWSAlert.

A estrutura do EWSAlert foi concebida baseada e visando a integração com padrões já consolidados, como o STIX, IDMEF e VERIS.

O atributo *entities* é obtido por identificação de entidades nomeadas usando algoritmos de Processamento de Linguagem Natural. O atributo *category* é obtido por meio de heurísticas e aprendizado de máquina. O atributo *severity* é obtido pela colaboração entre os usuários e o *confidence* por classificação supervisionada. O atributo *associated_ at* usa algoritmos de similaridade textual para agrupar alertas que abordam o mesmo assunto e que estão dentro de uma janela temporal predefinida. O atributo *additional_ data* é obtido explorando as URLs no texto do alerta.

## 5.4 Resultados e Discussões

Esta seção apresenta e discute os resultados do monitoramento de mensagens postadas no Twitter e Facebook realizado pelo sistema CEWS que está em operação desde março/2015. Inicialmente, são apresentados alertas divididos em categorias, em seguida, as limitações do sistema e, por fim, melhoramentos futuros.

**Tabela 5.1:** *Estrutura do EWSAlert*

| Atributo | Descrição |
|---|---|
| id | Identificação interna do sistema. |
| alert_id | Identificação única para o alerta de uma fonte específica. Se possível, recomenda-se usar o identificador original da fonte. |
| information_source | Identificação única para a fonte de dados primária do alerta. |
| alert_summary | Descrição textual resumida do texto do alerta. |
| alert_description | Descrição textual completa do texto do alerta com contexto. |
| first_alert_at | Indicação temporal de quando a primeira informação foi publicada/divulgada. |
| last_alert_at | Indicação temporal de quando a última informação agregada ao alerta foi publicada. |
| confidence | Indicação de confiança que a informação é um alerta. Os valores são estão no intervalo inteiro de 0 a 100 e, indicam um percentual de confiança. |
| severity | Indicação do grau de criticidade/severidade do alerta. Os valores estão no intervalo inteiro de 1 a 10 e, quanto maior o valor, maior o nível de criticidade. |
| category | Indicação de categoria do alerta. As categorias predefinidas são: desfiguração de página, orquestração de ataque, ataque ddos, vulnerabilidades, vazamento de dados, infosec, desconhecida. |
| classification | Indicação interna da classificação primária do alerta como potencial alerta, potencial spam, informações insuficientes, entre outros a definir. Atributo usado comumente para processamento interno. |
| detected_terms | Indica as palavras-chave detectadas no monitoramento e a posição no texto do alerta (alert_description). |
| author | Identificação de autoria do ataque ou do provável autor do ataque. |
| target | Identificação dos alvos descritos no ataque. Atributo composto por uma lista de estruturas de *entities*. |
| entities | Identificação de entidades do alerta. Uma lista polimórfica para armazenar entidades nomeadas presentes no texto do alerta, por exemplo, URLs, IPs, organizações, software, hardware, pessoas e outros. |
| lang | Idioma do texto original do alerta. |
| associated_at | Referência a informações associadas ao alerta. Corresponde ao conjunto de EWSMessage que referem-se a um mesmo alerta. |
| related_to | Referência a alertas de sensores tradicionais. Preenchido somente por um especialista ou por um módulo de correlação. Atributo não implementado. |
| additional_data | Referência a estruturas adicionais usadas para descrever o alerta, como capturas de tela, trechos de códigos maliciosos, indicações de resolução do problema, entre outros. |
| number_sources | Indica o número de fontes que compõem o alerta. |
| number_messages | Indica o número de EWSMessage agregadas no alerta. |

### 5.4.1 Alertas de orquestrações de ataque

Uma orquestração de ataque corresponde a organização de indivíduos na realização de ataque ou indicação de alvos cibernéticos que serão atacados. Idealmente, essa é a situação onde um alerta antecipado pode propiciar a um administrador de redes proteger sua infraestrutura ou dados antes da realização do ataque, ou seja, o sistema detecta o incidente antes mesmo de sua ocorrência.

Uma das primeiras ocorrências desse tipo de ação detectado pelo CEWS foi a desfiguração de páginas em massa realizada contra o domínio ufmg.br. A primeira mensagem foi publicada em 21/03/2016 e indicava a desfiguração do domínio dee.ufmg.br. No dia seguinte, foi publicado um agradecimento ao grupo de atacantes e mais 16 domínios desfigurados (Figura 5.6).

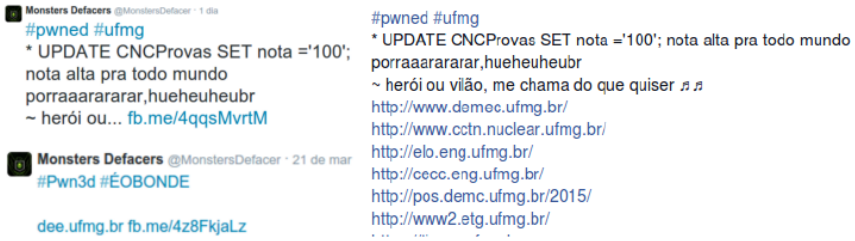


**Figura 5.6:** *Orquestração para desfiguração de páginas.*

Por meio de análise temporal das publicações e dos ataques realizada pela equipe do GT-EWS, acredita-se que um grupo se organizou e realizou o ataque após a primeira ocorrência de sucesso.

Foi encontrado indícios de injeção SQL que apontavam para diferentes localidades. Se um alerta antecipado, contendo a primeira mensagem, tivesse sido encaminhado para o administrador de redes, os outros subdomínios poderiam ter sido protegidos.

Em outro exemplo de orquestração, um alerta de 16/01/2016 indicava o recrutamento de novos membros para um grupo hacker via uma competição de desfiguração de páginas. Uma vaga na equipe era oferecida para o indivíduo que realizasse 15 desfigurações com a marca do grupo (Figura 5.7).

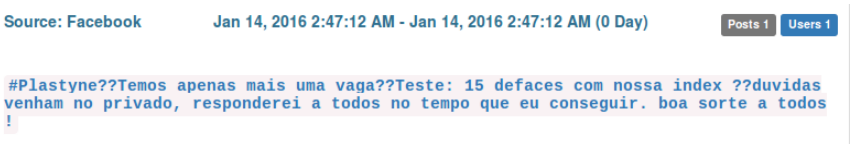


**Figura 5.7:** *Teste de admissão para novos membros de um grupo hacker.*

Observa-se na mensagem que os detalhes não são explicados abertamente, mas sim em uma discussão privada. Esse tipo de informação é interessante para situações de investigação, onde um agente infiltrado poderia identificar os alvos futuros ou os principais membros do grupo. Essas informações poderiam ser usadas para amplificar o monitoramento e para mitigar proativamente os ataques.

Um terceiro exemplo de orquestração foi a indicação de possíveis alvos em forma de pergunta para o usuários (Figura 5.8). Depois de alguns minutos, a página de um dos órgãos apontados como alvo foi desfigurada.


**Figura 5.8:** *Alerta com indicação de ataque a possíveis alvos.*

Nessa situação, o alerta antecipado poderia ter sido enviado para as entidades envolvidas na mensagem para que as mesmas verificassem os acessos de rede e também realizassem atualizações de software. Em último caso, mesmo após a desfiguração, o problema de segurança poderia ser identificado e corrigido mais rapidamente.

Considerando a análise das mensagens, verifica-se que o modelo de extração de alertas do sistema, que foi baseado no arcabouço EWS, consegue evidenciar eventos de orquestrações que poderiam ser usados como alertas antecipados por instituições para a contenção proativa de potenciais incidentes.

### 5.4.2 Alertas de DDoS

Os ataques de negação de serviço, principalmente os distribuídos, visam causar a indisponibilidade de acesso a serviços. Alertas antecipados para esse tipo de situação implicam em avisar os responsáveis o mais rápido possível que o serviço está indisponível. Outra possibilidade é avisar sobre orquestrações de ataques DDoS. Nesse caso, o alvo poderia se prevenir ou mitigar o tráfego malicioso.

Em outubro de 2015, o sistema detectou um ataque de DDoS publicado no Twitter, que alertava sobre um ataque a um sítio Web de partido político (Figura 5.9).

Esse tipo de alerta pode ser usado para mitigar os fluxos maliciosos que estavam inviabilizando o acesso, identificar a causa do ataque e também prevenir novos ataques. Por outro lado, os me-

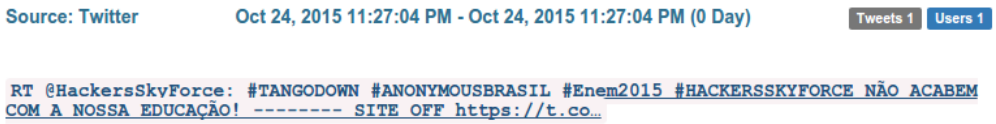


**Figura 5.9:** *Alerta de ataque DDoS ao sítio Web de partido político.*

canismos reativos a ataques de DDoS são complicados, pois dependem do bloqueio dos fluxos do ataque alcancem o alvo. De qualquer forma, monitorar e identificar antecipadamente intenções de ataques DDoS propiciam a minimização dos danos causados por esses ataques.

A Figura 5.10 apresenta outro exemplo de alerta de DDoS que indica que um sítio ficou indisponível durante algumas horas devido a um grupo hacker. O ataque foi realizado em 02/12/2016 e a mensagem publicada por volta das 5 horas.

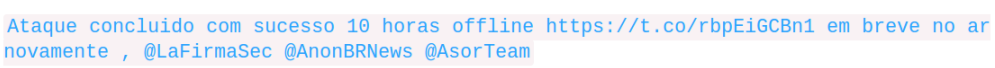


**Figura 5.10:** *Alerta de ataque DDoS ao sítio Web da Polícia Militar.*

Nesta situação, fica claro a impossibilidade de reação imediata pois a orquestração não foi detectada pelo sistema. Acredita-se que tenha ocorrido de forma privada entre os membros do grupo. Novamente, destacam-se a vantagem da identificação dos autores do ataque e o registro de evidência forense, no caso, o texto da publicação. É comum a divulgação das ações hackers em redes sociais, mas em seguida, o texto da publicação ser removido da linha do tempo do autor.

Para identificar ataques de DDoS não anunciados diretamente, também foram monitorados eventos que envolvem publicações com os termos "rede lenta", "site indisponível", "internet lenta" e outros similares, para identificar possíveis DDoS por meio de mensagens postadas por usuários leigos. Essa abordagem não produziu alertas relevantes, mas ainda necessita ser mais explorada.

Avaliando a extração de alertas antecipados para DDoS, verificou-se que há um número reduzido de anúncios nas redes sociais Twitter e Facebook comparado as estatísticas divulgadas pelo CERT.br. No entanto, os alertas registrados no sistema podem ser interessantes para registro de atividades de grupos hackers e identificação dos envolvidos.

### 5.4.3 Alertas de desfiguração de páginas

Os ataques de desfiguração de páginas Web visam descaracterizar a página alvo e publicar informações de autopromoção dos autores do ataque ou manifesto/protesto associados aos alvos. Alertas antecipados contra esses ataques consistem em notificar rapidamente aos responsáveis que a página foi atacada. Outro tipo de alerta antecipado depende da colaboração do alvo em compartilhar informações da infraestrutura atacada para avisar outros prováveis alvos com as mesmas características.

As redes sociais são comumente usadas para disseminar a desfiguração de páginas no Brasil, pois alcançam um número elevado de pessoas e o objetivo principal desse ataque é justamente divulgar a desfiguração. Logo, a maior parte dos alertas detectados no sistema são associados à desfiguração de páginas. A Figura 5.11 apresenta um padrão de publicação de desfiguração.


**Figura 5.11:** *Alerta de desfiguração de página.*

No contexto do CAIS, os alertas de desfiguração contra a infraestrutura da RNP tem auxiliado a reagir mais rapidamente a esse tipo de incidente. O sistema tem notificado o ataque no momento que é divulgado nas redes sociais, o que possibilita reagir no instante da divulgação do ataque. Mesmo quando as páginas desfiguradas são espelhadas em sistemas usados por hackers para registrar seus ataques, como o Zone-h, o sistema acaba sendo mais efetivo que esses sítios de espelhamento, pois esses sítios analisam a desfiguração antes de avisar os alvos.

Analisando os alertas de desfiguração de páginas, percebe-se que é necessário prover e promover a colaboração entre os administradores de redes. O arcabouço EWS descreve uma forma de alcançá-la por recomendação e também por compartilhamento de informações. No entanto, devido a vários fatores que envolvem desde privacidade, interesse e cultura dos envolvidos, as notificações antecipadas de desfiguração são apenas aquelas que auxiliam a reagir mais rapidamente e não aquelas que auxiliam a reagir proativamente.

### 5.4.4 Alertas de códigos de exploração

Os códigos de exploração (*exploits*) possibilitam atacantes explorarem vulnerabilidades conhecidas ou não conhecidas (vulnerabilidades dia zero) para comprometer ou obter acesso ao alvo. Alertas antecipados sobre novos códigos de exploração, principalmente os que exploram vulnerabilidades recentes ou desconhecidas, são importantes para um administrador proativamente proteger a infraestrutura de redes. Esses alertas possibilitam os administradores realizarem testes de penetração em suas infraestruturas e também procurar por correções ou atualizações de software.

A Figura 5.12 e 5.13 apresentam respectivamente divulgação de código de exploração a módulos do Wordpress e a um sistema de gerenciamento de conteúdo.

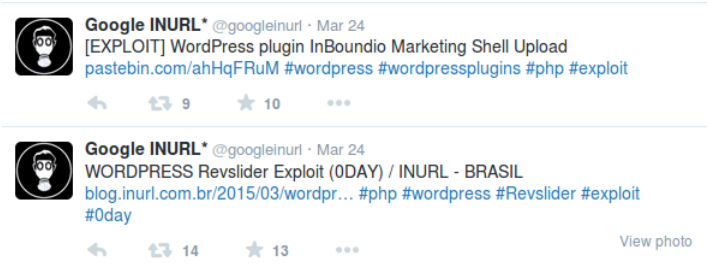


**Figura 5.12:** *Alerta de código de exploração contra módulos do Wordpress.*

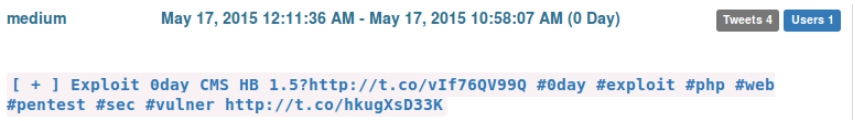


**Figura 5.13:** *Alerta de código de exploração contra sistema de gerenciamento de conteúdo.*

Esses tipos de alertas antecipados possibilitam uma reação preventiva por parte do administrador de redes e também possibilitam a fornecedores do software identificar problemas o quanto antes para prover uma correção. Do ponto de vista de regionalidade, como o sistema monitora ações de grupos hackers no Brasil, mensagens com esse teor são importantes para que os administradores sejam notificados o quanto antes de potenciais futuras ameaças.

Avaliando o arcabouço EWS, esse tipo de mensagem é identificado por uso de termos de impacto como “0day” e “exploit”. No entanto, também é comum esses termos gerarem uma quantidade de mensagens irrelevantes. Por isso, averiguações de URL, rotinas de priorização baseadas nos autores e número de postagens, e mecanismos de recomendação, possibilitam distinguir as mensagens irrelevantes das relevantes como alerta.

### 5.4.5 Alertas de vazamento de dados

Os vazamentos de dados consistem na divulgação pública de documentos ou informações privadas que foram obtidas e/ou publicadas sem permissão dos autores ou donos. Como consequência, implicam em diversos prejuízos a organizações, indivíduos e sistemas computacionais. Alertas antecipados de vazamentos consistem em avisar a entidade prejudicada o mais rápido possível assim que as informações são publicadas ou identificar ameaças de vazamento que, dependendo das circunstância, dos dados a serem disponibilizados e do conhecimento do alvo, podem possibilitar uma reação proativa.

No caso de vazamentos sem aviso prévio, foram detectados vários casos no sistema que possibilitaram avisar ou identificar rapidamente o compartilhamento de informações. A Figura 5.14 apresenta o vazamento de usuários e os resumos criptográficos de senhas obtidos de uma base de dados. Esse tipo de alerta é importante para que o administrador de rede faça a correção e solicite a alteração das senhas o mais rápido possível para evitar que indivíduos usem os dados para comprometer a segurança do sistema. Além disso, possibilita aos usuários trocarem suas senhas de outros serviços, pois é comum usuários fazerem uso de uma mesma senha em diversos serviços.


**Figura 5.14:** *Alerta de vazamento de usuários e senhas de um sistema.*

A Figura 5.15 apresenta o caso de vazamento de dados de um órgão do Governo do Amazonas e que foi identificado pelo sistema e, em seguida, notificado aos responsáveis para iniciarem o processo de resposta a incidentes. Nesse caso, os documentos estavam disponíveis para acesso em um serviço de armazenamento de arquivos. Esse alerta possibilitou a reação mais rápida ao ocorrido, visto que o vazamento ocorreu no final de semana.


**Figura 5.15:** *Alerta de vazamento de documentos do Governo do Amazonas.*

Outras situações de vazamento ocorrem quando os atacantes avisam que disponibilizarão informações privadas de organizações momentos ou horas antes da publicação efetiva. A Figura 5.16 apresenta a ameaça de vazamento da Agência do Nacional de Energia Elétrica (ANEEL) e, no dia seguinte, a mensagem de vazamento com a URL para os dados.

Nessa situação, o primeiro alerta poderia ter ajudado o administrador de redes identificar se o ataque estava em andamento ou já tinha ocorrido. No mínimo, essa informação indicaria ao administrador a necessidade de corrigir a falha explorada. O segundo alerta ajudaria nas medidas de contenções por parte dos administradores da ANEEL, que poderiam se preparar mais rápido para os impactos da divulgação da informação.

Outro alerta de ameaça de vazamento seguido de vazamento efetivo, os atacantes apenas anunciaram que publicariam informações privadas de quatro instituições governamentais, mas não mencionaram os nomes na publicação (Figura 5.17). No dia seguinte, publicaram os nomes e bases de dados dessas organizações. Esse tipo de notificação, possibilita apenas manter atenção no grupo


**Figura 5.16:** *Alertas de ameaça e vazamento de relatórios da ANEEL.*

que está realizando a ameaça e tomar as medidas de resposta a incidentes o mais rápido possível após a publicação do vazamento.

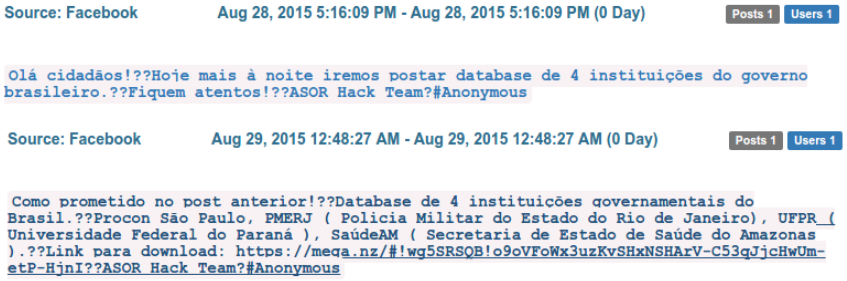


**Figura 5.17:** *Alertas de ameaça e vazamento de base de dados de instituições governamentais.*

O uso do sistema para detectar vazamentos tem proporcionado a identificação de alertas antecipados que possibilitam reagir mais rápido aos ocorridos, mesmo que os dados já estejam disponibilizados. Comprovou-se que o mecanismo de notificação é mais rápido que outras fontes e, dessa forma, possibilita organizar a resposta a incidentes o quanto antes.

### 5.4.6 Alertas de rumores

Os alertas de rumores consistem de alertas de possíveis ataques que precisam ser investigados para a confirmação. Esses alertas possibilitam a organizações aumentarem o nível de proteção de suas infraestruturas de redes e de seus sistemas, caso informações sensíveis (nomes, IP, URLs, outros) de organizações sejam identificados como alvos em mensagens. Também possibilitam investigar supostos vazamentos de informações e invasões de sistemas que podem causar diferentes dimensões de danos se foram concretamente realizados.

No sistema foi identificado um alerta de um possível vazamento de informações do Exame Nacional do Ensino Médio (ENEM), usado por muitas Universidades públicas como critério de seleção para entrada em seus cursos. A Figura 5.18 apresenta a mensagem da publicação.

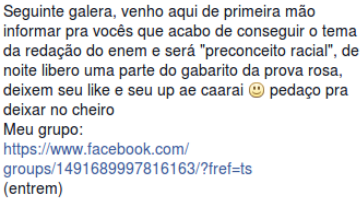


**Figura 5.18:** *Alerta sobre possível vazamento de informações e gabarito do ENEM.*

Essa mensagem serviu como um alerta antecipado de um potencial vazamento de informações que implicaria diretamente nos resultados de um exame importante no Brasil. Logo, acabou re-

sultando em uma investigação para comprovar se as informações eram verídicas ou não. A partir desse alerta, verificou-se que o sistema baseado no arcabouço EWS poderia ser usado para alertar e direcionar a investigação de outros tipos de crimes que são divulgados em redes sociais ou em outras fontes de dados não estruturados.

A identificação de informações que possam indicar a possibilidade de um crime cibernético ou intenção de ataque é uma característica importante para a composição de um EWS. Mesmo quando o rumor não é verdadeiro, não é desperdício de tempo a investigação desse tipo de alerta, pois há casos que o rumor pode causar danos a nível nacional, como o caso do ENEM.

## 5.5 Limitações

As limitações do CEWS estão associadas as limitações das implementações dos seus componentes. O principal problema está associado ao número de alertas que são falsos positivos. A implementação atual apresenta um número moderado de falsos positivos que ainda possibilita que um administrador/supervisor realize o descarte das mensagens irrelevantes, mas inviabiliza o processamento automático.

Nos coletores, a principal limitação é manter um conjunto de perfis e palavras-chave que consiga capturar mensagens importantes como alertas. Devido a linguagem informal, siglas para representar palavras e escrita errada, ocasiona a perda de alguns alertas. Por exemplo, para referir-se ao termo "hackeada", um usuário usou o termo "récked", logo a mensagem não foi capturada. Além disso, outro problema é manter as listas de perfis de grupos ou hackers a serem monitorados. Alguns grupos se dissolvem em grupos menores e não postam mais informações relevantes nos perfis antigos. Outros grupos criam novas páginas ou perfis sociais para divulgar suas ações e o monitoramento é prejudicado. Por fim, as APIs públicas também restringem o número de buscas e fornecem monitoramento limitado às redes sociais.

Nos filtros, a principal limitação está no estabelecimento de um limiar de corte e das palavras-chave para as listas negras usadas na remoção de alertas irrelevantes. Um limiar baixo implica em um número maior de mensagens para a classificação e, consequentemente, acaba por gerar um número maior de falsos positivos. Um limiar alto implica na remoção de mensagens que poderiam ser potenciais alertas. O mesmo princípio ocorre com o uso de poucas e muitas palavras-chave e entidades usadas em listas negras.

Nos normalizadores, o reconhecimento de entidades nomeadas que são usadas em filtros e classificadores são em algumas circunstâncias identificadas erradas pelos algoritmos, principalmente nos textos obtidos de redes sociais, e isso implica em categorização errada de ataques ou classificações equivocadas. Nos agrupadores, devido a variedade e forma na escrita dos textos, há ocorrências de mais de um grupo para alertas que deveriam ser agrupados juntos.

Nos componentes de processamento de alertas, a principal preocupação é diminuir os falsos positivos. Neste sentido, o arcabouço tem usado medidas de confiança para alertas e o sistema de recomendação para melhorar os resultados dos classificadores e analisadores. Há muitas mensagens que são difíceis de identificar se são alertas relevantes ou não, até porque em algumas situações abordam ataques ou ameaças de ataques, mas que podem se mostrar como alertas, mas tem valor apenas para alguns indivíduos. É comum encontrar notícias de ataques DDoS em servidores de jogos, notícias de contas de Facebook e Twitter comprometidas, usuários ameaçando outros de invasão em um contexto particular, entre outros. Para dificultar ainda mais a classificação, há os casos onde siglas como "ddos" não se refere a ataques, mas sim a cidade de Dourados. Uso de ataques em outros contextos: "ataque político", "ataque organização", "ataque do coração", "ataque ao governo", entre outros. Mensagens com esses termos e um contexto associado à Internet, dificultam a classificação correta do alerta. Para lidar com essa limitação, a solução são os filtros adaptativos e o uso de recomendação para filtrar os alertas.

Os componentes de notificação, na implementação atual, são limitados a um limiar de certeza para automaticamente propagar um alerta e também a supervisão e encaminhamento de alertas somente após a revisão por um especialista em segurança.

## 5.6 Considerações Finais

Este capítulo apresentou o CEWS, um sistema de alerta antecipado de cibersegurança, que foi construído baseado no arcabouço EWS desenvolvido nesta tese. O desenvolvimento do CEWS contou com o apoio da RNP que financiou o projeto em duas fases. A arquitetura do sistema foi baseada em serviços para facilitar a distribuição e comunicação remota dos componentes. Os principais coletores implementados realizam o monitoramento de palavras-chave e perfis no Twitter e Facebook. O módulo de processamento central implementa analisadores, classificadores e recomendadores de alertas. A notificação é realizada por meio de e-mail estruturado segundo as necessidades dos parceiros e também por meio de uma interface Web. Há padrões para adição de módulos acopláveis, novos mecanismos de pré-processamento e processamento, além de formatos para a representação estruturada das informações e alertas. Durante a produção do sistema, foram identificados alertas de orquestrações de ataque, desfiguração de páginas, ataques de DDoS, vazamentos de dados e códigos de exploração. Alguns alertas possibilitaram reação mais rápida ao incidente e outros, se tivessem sido encaminhados para as unidades responsáveis, poderiam até mesmo evitar o incidente. A implementação também possibilitou identificar pontos que precisam ser melhorados, como a diminuição de falsos positivos e categorização dos alertas. No entanto, mesmo considerando as limitações, comprovou-se que a implementação dos componentes do arcabouço, uso de bases de inteligência e de mecanismos de classificação e recomendação possibilitam mitigar ameaças cibernéticas proativamente ou mais rapidamente.

# Capítulo 6

# Conclusões

Nesta tese, foi proposto e validado um arcabouço para a análise e extração de alertas, preferencialmente antecipados, de fontes de dados não estruturados. Foi verificado em duas mídias sociais com características diferentes, microblog Twitter e a rede IRC, que há informações relevantes associadas à cibersegurança, e que os mecanismos propostos no arcabouço possibilitam a extração dessas informações e a geração de alertas. Por consequência, confirmou-se a relevância do desenvolvimento de Sistemas de Alertas Antecipados visando aumentar o nível da segurança de redes de computadores e de sistemas computacionais, principalmente possibilitando ações proativas ou medidas reativas mais rapidamente.

Na análise e extração de alertas no microblog Twitter, foi identificado que usuários colaboram com a segurança cibernética difundindo mensagens associadas a vulnerabilidades e falhas de software. Por outro lado, foi verificado que há usuários que usam o microblog para organizações de ataques, divulgar ataques recém realizados, dados de terceiros ou ferramentas de ataque. Pela análise da difusão dessas mensagens, foi possível criar mecanismos para identificar e evidenciar tendências e impactos de novas vulnerabilidades, potenciais ameaças e notícias de segurança (infosec).

Na análise e extração de alertas nas redes IRC, foram investigados os canais de segurança e de atividades *hackers*. Nos canais de segurança, verificou-se a discussão de novos códigos de exploração e vulnerabilidades. Nos canais atividades *hackers*, verificou-se a discussão de ferramentas e alvos de ataques. O uso dos processos definidos no arcabouço EWS possibilitou a identificação e a extração de mensagens associados à cibersegurança, mas ainda há várias dificuldades com o monitoramento e processamento da relevância das mensagens retornadas como alertas. No entanto, o processo de análise de dados proposto no arcabouço, possibilitou a implementação de técnicas e extração de características que auxiliaram na diminuição de mensagens irrelevantes e na evidência de alertas.

No arcabouço foram apresentados componentes para o processamento de alertas a partir de informações pré-processadas por filtros, normalizadores e agrupadores. Duas principais funções investigadas por serem essenciais na evidência de alertas foram a classificação e recomendação de alertas. Usando uma abordagem de classificação não supervisionada, analisadores filtraram notificações de interesse pelo número de usuários que postaram uma mensagem. Esse tipo de abordagem mostrou-se adequada para identificar tendências de ataques ou nível de criticidade de ameaças cibernéticas. Já usando classificadores supervisionados, foi verificada a dificuldade de evidenciar alertas relevantes devido a natureza informal e não estruturada das mensagens.

O mecanismo de recomendação proposto mostrou como a colaboração auxilia na filtragem e priorização de mensagens segundo o interesse de cada usuário. Apesar do modelo não ter sido validado com experimentação online, foi possível verificar como, por meio de uma abordagem mista, pode-se atender os interesses particulares de cada administrador de redes e, ao mesmo tempo, identificar alertas relevantes e não relevantes por meio de filtragem colaborativa.

A validação do arcabouço de forma geral ocorreu pela implementação de um sistema de alerta antecipado de cibersegurança. No desenvolvimento do sistema, os conceitos e componentes do arcabouço foram mapeados em implementações de software. Além disso, o sistema foi construído baseado em uma arquitetura distribuída. Também foi comprovado que as bases de inteligência produzidas no decorrer da pesquisa possibilitaram o desenvolvimento de mecanismos que retornam diferentes tipos de alertas antecipados.

As principais limitações da pesquisa são associadas à otimização dos métodos implementados para validação do arcabouço. Em geral, os mecanismos implementados apresentaram alertas antecipados, no entanto, ainda são restritos quanto ao número de falsos positivos.

Em relação a trabalhos futuros, a investigação de sistemas de alerta antecipado, o desenvolvimento do arcabouço e a implementação dos componentes gerou uma base sólida para novas pesquisas associadas à cibersegurança:

1. Investigação de novas fontes: cada fonte de dados tem sua particularidade e, baseado nos processos de análise e extração de alertas do arcabouço, podem ser conduzidas investigações em outras fontes de dados não estruturados. Além disso, a investigação pode ajudar a evoluir os processos do próprio arcabouço.

2. Classificação de alertas: avaliar novas características e algoritmos para a construção de modelos de classificação otimizados para categorizar os alertas. A classificação pode considerar classes como: relevante e não relevante; antecipado e não antecipado; tipo de notificação (DDoS, vulnerabilidade, vazamento de dados, orquestração, código de exploração, entre outros).

3. Filtros adaptativos: desenvolvimento de filtros adaptativos para a remoção e priorização de alertas, principalmente nas fases de coleta e pré-processamento.

4. Mecanismos de predição: desenvolver mecanismos que correlacionem informações de diferentes fontes, tradicionais e não tradicionais, para a predição de ameaças.

5. Colaboração e privacidade: desenvolver um modelo de privacidade para a colaboração e disseminação de alertas entre parceiros, inclusive considerando diferentes entidades, como o Governo, empresas públicas e privadas, entre outras.

6. Novas bases de inteligência: realizar pesquisas para a construção de bases de inteligência (p. ex. vocabulário, identificação de entidades, ontologia de cibersegurança, entre outros) considerando a língua Portuguesa.

No decorrer desta tese, mostramos como são importantes os Sistemas de Alerta Antecipado para mitigar ameaças em infraestruturas computacionais. Logo, do ponto de vista de pesquisa, nossa principal contribuição foi investigar e mostrar como podemos extrair informações de dados não estruturados, em especial das mídias sociais, para a criação de novos serviços de segurança e, também, como a colaboração indireta desses meios pode ser usada para identificar a organização de grupos *hackers* e as constantes vulnerabilidades de software, alertando organizações, administradores de redes e usuários sobre ameaças reais e tendências de ataques.

# Apêndice A

# Projetos para a detecção antecipada

**Tabela A.1:** *Projetos associados à detecção proativa de ameaças (parte 1)*

| Projeto | Site |
|---|---|
| DNS-BH - Malware Domain Blocklist | http://www.malwaredomains.com |
| Malware URL | http://www.malwareurl.com/ |
| Dshield | https://www.dshield.org/ |
| Honeyspider Network 2 | http://www.honeyspider.net |
| Cert.br Honeytarg | http://honeytarg.cert.br/ |
| FIRE (Finding Rogue nEtworks) | http://www.maliciousnetworks.org |
| EXPOSURE | http://exposure.iseclab.org/ |
| Zeus tracker / SpyEye tracker / Feodo tracker | http://www.abuse.ch/ |
| Malware Domain List | http://www.malwaredomainlist.com/ |
| Spamhaus | http://www.spamhaus.org/ |
| Shadowserver Foundation | http://www.shadowserver.org |
| ARAKIS | http://arakis.pl |
| Malc0de database | http://malc0de.com/database/ |
| MalwareBlacklist.com | http://www.malwareblacklist.com |
| Arbor ATLAS | http://atlas.arbor.net/ |
| Composite Blocking List (Spamhaus) | http://cbl.abuseat.org/ |
| Serviços do Team Cymru | http://www.team-cymru.org/ |
| Project Honeypot | http://www.projecthoneypot.org |
| Malware Patrol | http://www.malwarepatrol.net |
| Zone-H | http://www.zone-h.org |
| SenderBase | http://www.senderbase.org/ |
| Deepsight | http://tms.symantec.com/ |

**Tabela A.2:** *Projetos associados à detecção proativa de ameaças (parte 2)*

| Projeto | Sensores | Técnica | Alerta | Formato |
|---|---|---|---|---|
| DNS-BH - Malware Domain Blocklist | Google Safe Browsing, malcde.com database, phishtank e outros | Agregação de fontes | Listas de domínios maliciosos | arquivos de zonas para Bind e Windows, AdBlock, ISA |
| Malware URL | VirusTotal, Wepawet, Anubis, Threat Expert | Agregação de fontes e técnicas de análise | novas URL observadas, endereços IP relacionados com a resolução DNS, número de AS relacionados com servidor. | RSS, CSV |
| Dshield | Registros de firewalls | Agregação e análise de especialistas | Listas de bloqueios de endereços IP | RSS,TSV |
| Honeyspider Network 2 | Honeypots | Varreduras de URL | Notificações de URL maliciosas | e-mail, rss, WAPI (API), texto puro |
| Cert.br Honeytarg | Spampots, Honeypots | Agregação e estatísticas | Notificações CSIRT, estatísticas IP, portas, protocolos, spam (membros) | Web, e-mail |
| FIRE (Finding Rogue nEtworks) | Anubis, Wepawet, PhishTank, HoneySpider Network | Agregação de fontes | números de AS de redes maliciosas, endereços de servidores que espalham ameaças | - |
| EXPOSURE | Anubis, Wepawet | Análise passiva de DNS | lista de bloqueio de nomes de domínio | - |
| Zeus tracker / SpyEye tracker / Feodo tracker | - | Monitoramento de servidores C&C | Lista de bloqueio de nomes e endereços IP | texto, RSS |
| Malware Domain List | colaboração usuários, outros sites | Análise e acompanhamento de códigos maliciosos | lista de bloqueio de nomes de domínio, endereços IP e número de AS | CSV, texto, RSS |
| Spamhaus | - | Monitoramento de serviços e propagadores de spam | Lista de bloqueio de spam, exploit, domínio, políticas | - |
| Shadowserver Foundation | - | Monitoramento de redes | - | CSV, HTML, XML, texto |
| ARAKIS | Honeypots, Firewalls, antivírus, darknets | Monitoramento de varreduras por ameaças (principal) | Detalhes de conexões que executaram varreduras maliciosas ou suspeitas, conexões que acionaram regras do Snort | CSV (download via https para parceiros) |
| Malc0de database | - | Monitoramento de URL maliciosas | Lista de bloqueio URL, IP, ASN, MD5 do malware. | RSS, Twitter, HTTP (Tabela), texto |
| MalwareBlacklist | Honeypots, sandbox, usuários submetem URL para verificação | Monitoramento de URL maliciosas e download de códigos maliciosos | Lista de bloqueio URL, IP, ASN. | API, HTTP, RSS |

**Tabela A.2:** *Projetos associados à detecção proativa de ameaças (parte 2)*

| Projeto | Sensores | Técnica | Alerta | Formato |
|---|---|---|---|---|
| Arbor ATLAS | Honeypots, IDS, Scan logs, estatísticas DoS, relatórios vulnerabilidades, amostras de malware, monitoramento de botnets | Monitoramento de malware, exploits, DdoS, varreduras, phishing. | Gráficos e listas de ameaças no portal | XML/IODEF, CSV |
| Composite Blocking List (Spamhaus) | Servidores de email, spamtraps | Dados providos por servidores de e-mail e spamtraps | Lista de bloqueio de DNS para suspeitos de spam | texto, sincronização com rsync |
| Serviços do Team Cymru | - | - | - | E-mail, web |
| Project Honeypot | Honeypots | Monitoramento de spammers e spambots | Endereços, URLs spammers | RSS e outros |
| Malware Patrol | Colaboração usuários, crawlers url suspeitas | Monitoramento de códigos maliciosos | Lista de bloqueio | diversos formatos |
| Zone-H | Colaboração anônima e outros | Monitoramento de sítios Web atacados (deface) | Lista de IP e URL | RSS, e-mail |
| SenderBase | Web, Email, Firewall e IPS | Monitoramento de tráfego Web e e-mail. | Listas de reputação | Serviço de busca, ranking |
| Deepsight | - | Monitoramentos diversos | Listas de reputação | E-mail, web |

Nota: (-) informações não identificadas ou pendentes.

# Apêndice B

# Análise de fontes abertas

**Tabela B.1:** *Resumo de fontes abertas para EWS (parte 1)*

| Fonte | Classe | Formato | Abrangência | Privacidade | Divulgação | Usuários ativos | Confiabilidade | Autor da publicação | Data da publicação |
|---|---|---|---|---|---|---|---|---|---|
| Microblogs | mídia social | semi-estruturado | global | conteúdo público | anonimizar a fonte | Twitter (271 milhões) | média | sim | sim |
| Blogs | mídia social | não estruturado | global, regional (língua) | conteúdo público | referência ao autor ou fonte | - | alta (blogs especializados), baixa (outros) | sim | sim |
| Redes de relacionamento | mídia social | semi-estruturado | global, comunidades | conteúdo público e privado | referência ao autor ou fonte | Facebook (1.3 bilhões), Google+ (343 milhões) | baixa (publicações), alta (extração de códigos maliciosos) | sim | sim |
| Redes IRC | mídia social | semi-estruturado | global, comunidades | conteúdo público e privado | anonimizar a fonte e autor | 400 mil | média (canais de segurança, hackers) | sim | sim |
| Web Feeds | mídia social | semi-estruturado | global | conteúdo público | referência ao autor ou fonte | - | alta (blogs especializados), baixa (outros) | sim | sim |
| Bases de vulnerabilidades | web | semi-estruturado | global | conteúdo público | referência à base | - | alta | sim | sim |
| Fóruns | web | não estruturado | comunidades | conteúdo público e privado | anonimizar a fonte e autor | - | média | sim | sim |
| Listas de e-mail | web | não estruturado | comunidades | conteúdo privado | referência ao autor ou fonte | - | alta | sim | sim |
| Motores de busca | web | não estruturado | global | conteúdo público | referência ao autor ou fonte | - | baixa (depende de análise especialista) | às vezes | não, geralmente |
| Deep web | web | não estruturado | global | conteúdo público | fonte anônima | TOR (cerca de 2 milhões) | baixa (depende de análise especialista) | não | não, geralmente |

Nota: (-) item não identificado na fonte

**Tabela B.2:** *Resumo de fontes abertas para EWS (parte 2)*

| Fonte | Acesso | Disseminação | Vantagens | Limitações | API | Precauções | Formato |
|---|---|---|---|---|---|---|---|
| Microblogs | público | rápida e ampla | acesso a metadados e outras informações, URL nas mensagens | consultas limitadas, tamanho das mensagens, contexto | sim | datas desatualizadas, mensagens falsas. | JSON |
| Blogs | público | lenta e limitada | selecionar os sítios relevantes para segurança, conteúdo mais detalhado | uso de lista de blogs, muitas notícias na mesma localização (URL), organização idiossincrática dos blogs | sim (crawlers) | datas desatualizadas, contexto da notícia (patch, vulnerabilidade, exploit, propaganda, ...) | texto |
| Redes de relacionamento | público e restrito | rápida e ampla | número de usuários, comunidades de segurança, possibilita implementar mecanismos para detectar códigos maliciosos | heterogeneidade de conteúdo e mídias (vídeos, áudios, textos, imagens), consultas limitadas | sim | contexto da notícia, credibilidade do usuário | JSON |
| Redes IRC | público e restrito | lenta e limitada | acesso a informações de ataques, vulnerabilidades | salas protegidas, número de usuários reduzido | sim | identidade do crawler, contexto das mensagens | - |
| Web Feeds | público | lenta e limitada | assinaturas de assuntos específicos (patches, vulnerabilidades, ataques), estrutura RSS ou Atom, facilidade de verificar atualizações | ameaça descrita sucintamente, depende de assinatura | sim | data desatualizada | JSON e outros |
| Bases de vulnerabilidades | público | lenta e limitada | confiabilidade, estrutura de publicação, detalhes técnicos | atrasos na publicação, incompletude dos registros sobre a ameaça | sim (xml) | descrição não completa dos registros, data de divulgação | XML e outros |
| Fóruns | público e restrito | lenta e limitada | comunidades específicas | muitos tópicos, organização idiossincrática, assinatura do fórum | - | processamento complexo das informações | - |
| Listas de e-mail | restrito | lenta e limitada | descrições detalhadas para administradores | dificuldades para processamento automatizado, assinatura, voltado a administradores | - | processamento complexo das informações | texto |
| Motores de busca | público | lento e limitado | viabiliza pesquisa por dados complementares | contexto das mensagens e datas | sim | escopo de busca | HTML e outros |
| Deep web | público | lento e limitado | usado para divulgar informações de cibercrimes | dificuldades para processamento automatizado, em especial, realização de busca | - | conteúdo acessado, escopo de busca | HTML e outros |

Nota: (-) item não identificado na fonte

# Apêndice C

# Dados usados nos experimentos

**Tabela C.1:** *Lista de endereços dos feeds usados nos experimentos (maio/2012)*

| URL - Feeds |
|---|
| http://www.us-cert.gov/channels/techalerts.atom |
| http://www.us-cert.gov/channels/current.atom |
| http://www.symantec.com/xml/rss/sepr.jsp |
| http://www.symantec.com/xml/rss/listings.jsp?lid=mixedsecurityrisks |
| http://www.symantec.com/xml/rss/listings.jsp?lid=latestthreats30days |
| http://www.symantec.com/xml/rss/listings.jsp?lid=advisories |
| http://www.securelist.com/en/rss/latestnews |
| http://www.securelist.com/en/rss/descriptions |
| http://www.h-online.com/security/atom.xml |
| http://www.ehackingnews.com/feeds/posts/default |
| http://threatpost.com/en_us/taxonomy/term/9/0/feed |
| http://threatpost.com/en_us/taxonomy/term/6/0/feed |
| http://threatpost.com/en_us/taxonomy/term/3/0/feed |
| http://threatpost.com/en_us/taxonomy/term/10/0/feed |
| http://threatpost.com/en_us/rss.xml |
| http://technet.microsoft.com/en-us/security/rss/comprehensive |
| http://technet.microsoft.com/en-us/security/rss/bulletin |
| http://technet.microsoft.com/en-us/security/rss/advisory |
| http://securityaffairs.co/wordpress/feed |
| http://seclists.org/rss/cert.rss |
| http://rssnewsapps.ziffdavis.com/eweeksecurity.xml |
| http://rss.techtarget.com/160.xml |
| http://news.cnet.com/8300-1009_3-83.xml |
| http://feeds2.feedburner.com/HelpNetSecurity |
| http://feeds.pheedo.com/techtarget/Searchsecurity/SecurityWire |
| http://feeds.pheedo.com/SecurityBytes |
| http://feeds.pheedo.com/SearchsecurityThreatMonitor |
| http://feeds.feedburner.com/scmagazinenews |
| http://feeds.feedburner.com/SansInstituteAtRiskAll?format=xml |
| http://blogs.mcafee.com/consumer/consumer-threat-alerts/feed |

# Apêndice D

# Questionário - Sistemas de Recomendação sobre Notícias de Cibersegurança

O questionário é composto de 26 questões divididos em quatro seções:

- Levantamento de perfil: identificar o perfil dos administradores que responderam o questionário.
- Conhecimentos sobre segurança cibernética: identificar o grau de conhecimento da sua infraestrutura e de atualização sobre notícias de segurança.
- Avaliação de interesse: identificar o interesse dos administradores e atributos para um sistema de recomendação de notícias de cibersegurança.
- Situações de uso: situações práticas para verificar se refletem com a avaliação de interesse.

## D.1 Levantamento de perfil

1. Há quanto tempo você atua como administrador de redes? (Em anos)
   ABERTA
2. Quais são os sistemas operacionais que você administra/usa para serviços/servidores?
   (a) Linux (b) Windows Server (c) Mac OS X Server (d) FreeBSD (e) Outro
3. Quais são os sistemas operacionais que você administra/usa para clientes DESKTOPs da rede?
   (a) Linux (b) Windows (c) Mac OS (d) BSD (e) Outro
4. Na rede que você administra, as pessoas levam os seus próprios dispositivos (ex: laptops, celulares, tablets) e possuem permissão para usar esses dispositivos na rede da organização?
   (a) Sim (b) Não
5. Na rede que você administra, faz parte do seu trabalho monitorar dispositivos móveis (ex: laptops, celulares, tablets) e atualizá-los caso alguma vulnerabilidade seja descoberta?
   (a) Sim (b) Não

## D.2 Conhecimentos sobre segurança cibernética

1. Como você classificaria o seu conhecimento sobre Segurança Cibernética?
   (a) Excelente (b) Bom (c) Regular (d) Fraco
2. Quais os tipos de ataques que a infraestrutura que você administra sofre frequentemente? (Deixar em branco caso não saiba)
   (a) Negação de serviços (b) Software maliciosos (c) Roubos de senha ou sessão (d) Exploração de dados ou quebra de confidencialidade (e) Outros

3. Você acompanha algum tipo de canal de notícias ou comunicação para se atualizar sobre as principais questões relacionadas à Segurança Cibernética? Caso sim, ordenar pelo canal mais utilizado.
   (a) Lista de e-mails (b) Facebook (c) Twitter (d) IRC (e) Blogs (f) Feeds RSS (g) Outro
4. Com qual frequência você lê/obtém essas informações sobre Segurança Cibernética?
   (a) Diariamente (b) Semanalmente (c) Mensalmente (d) Eventualmente (e) Nunca
5. Qual a quantidade de notícias que você lê sobre Segurança Cibernética realmente te interessam?
   (a) Maioria (b) Metade (c) Minoria (d) Nenhuma
6. Você usa algum canal de comunicação para trocar conhecimento e informações com outros Administradores de Rede? Se sim, qual?
   (a) Sim (b) Não

## D.3 Avaliação de interesse

1. Você acha que seria útil ter um sistema que mostrasse/compartilhasse as informações de segurança que outros administradores julgam importantes para manter a segurança de seus sistemas?
   (a) Sim (b) Não
2. Você usaria um sistema para colaboração com outros administradores com o perfil semelhante ao seu (utilizam equipamentos ou softwares semelhantes ou compartilham os mesmos interesses)?
   (a) Sim (b) Não
3. Você faria recomendações de alertas de segurança em um sistema para colaboração com outros administradores de rede?
   (a) Sim (b) Não
4. Você compartilharia alertas de segurança em um sistema para colaboração com outros administradores de rede?
   (a) Sim (b) Não
5. Ordene as informações que você considera relevantes ao ler uma notícia sobre atualização de software, ameaça ou vulnerabilidade: (Ordenar do item Mais importante para o Menos importante).
   (a) Tipos de ameaças (b) Software envolvidos (c) Casos conhecidos (d) Como explorar falhas e vulnerabilidades (e) Origem do conteúdo (f) Impacto (financeiro/perdas) (g) Dicas de proteção
6. Ordene os fatores mais importante que fazem com que você se interesse por uma notícia sobre Segurança Cibernética: (Ordenar do item Mais importante para o Menos importante).
   (a) Título (b) Resumo (c) Imagem (d) Palavras-chave (e) Categoria da notícia (patch, ameaça, vazamento, ataque) (f) Software/hardware afetados (g) Grau de criticidade (h) Fonte de notícia (i) Aceitação/feedback da comunidade (curtidas, comentários, retweets)
7. Você já compartilhou informações sobre problemas de segurança? Caso sim, conte sobre alguma vez que o fez (onde e como):
   (a) Sim (b) Não
8. Quais as informações você NÃO forneceria para o sistema considerando manter a confidencialidade da sua rede? (Lembrando que o sistema não compartilharia informações do perfil para ninguém, somente utilizaria para filtrar as recomendações).
   (a) Nome (b) Instituição/Empresa onde você trabalha (c) Sistemas operacionais (d) Nomes de software que utiliza (e) Versões de software que utiliza (f) Equipamentos utilizados (g) Problemas de segurança cibernética que você já sofreu (h) Como você solucionou um problema de segurança
9. Quais os tipos de classificações você faria em uma recomendação? (Selecione os tipos de interações que você utilizaria para dar feedback ao sistema sobre um determinado conteúdo).
   (a) Comentários abertos (b) Voto (Verdadeiro/Falso ou +1/-1) (c) Urgente/Crítico (Sim/Não) (d) Complementar com links externos (e) Categorizar (Colocar tags sobre o tipo ameaça)

## D.4 Situações de uso

1. Se alguns administradores de redes recomendassem uma atualização do Servidor Web Apache, você seguiria a recomendação se ela fosse apoiada por, no mínimo, quantos administradores?
   (a) de 1 a 5 (b) de 6 a 10 (c) de 11 a 20 (d) mais de 20
2. Se um administrador Y, que possui um perfil semelhante ao seu (utiliza os mesmos softwares, equipamentos e políticas de segurança), fizesse uma recomendação em uma notícia de possível vulnerabilidade em um serviço ou software, você despenderia tempo analisando e comentando o caso?
   (a) Sim (b) Não
3. Que situações levaria você gerar uma recomendação de um alerta?
   ABERTA
4. Você tomou conhecimento sobre a vulnerabilidade POODLE SSLv3? Se sim, se você recordar, poderia informar quando e a fonte de informação em que leu a notícia?
   (a) Sim (b) Não
5. Se um administrador com alto grau de conhecimento em segurança sobre o Sistema Operacional usado na sua infraestrutura de rede marcasse uma notícia de atualização urgente, você atualizaria?
   (a) Pesquisaria a respeito e tomaria minha decisão depois (b) Sim, sem nem pesquisar a respeito (c) Não, sem nem pesquisar a respeito
6. Se um administrador com grau de conhecimento moderado ou baixo sobre segurança no Sistema Operacional usado na sua infraestrutura de rede marcasse uma notícia de atualização urgente, você atualizaria?
   (a) Pesquisaria a respeito e tomaria minha decisão depois (b) Sim, sem nem pesquisar a respeito (c) Não, sem nem pesquisar a respeito

# Referências Bibliográficas


**Abraham** *et al.* **(2007)** Ajith Abraham, Ravi Jain, Johnson Thomas e Sang Yong Han. D-scids: Distributed soft computing intrusion detection system. *Journal of Network and Computer Applications*, 30(1):81 – 98. ISSN 1084-8045. doi: http://dx.doi.org/10.1016/j.jnca.2005.06.001. URL http://www.sciencedirect.com/science/article/pii/S1084804505000421. Citado na pág. 10

**Al-Nashif** *et al.* **(2008)** Y. Al-Nashif, A.A. Kumar, S. Hariri, Guangzhi Qu, Yi Luo e F. Szidarovsky. Multi-level intrusion detection system (ml-ids). Em *Autonomic Computing, 2008. ICAC '08. International Conference on*, páginas 131–140. doi: 10.1109/ICAC.2008.25. Citado na pág. 10

**Al-Qasem** *et al.* **(2013)** I. Al-Qasem, S. Al-Qasem e A. T. Al-Hammouri. Leveraging online social networks for a real-time malware alerting system. Em *Local Computer Networks (LCN), 2013 IEEE 38th Conference on*, páginas 272–275. doi: 10.1109/LCN.2013.6761247. Citado na pág. 84

**Anderson (1980)** James P. Anderson. Computer security threat monitoring and surveillance. Relatório técnico, James P. Anderson Company, Fort Washington, Pennsylvania. Citado na pág. 6

**Anderson e Khattak (1998)** Ross Anderson e Abida Khattak. The use of information retrieval techniques for intrusion detection. Em *Proceedings of First International Workshop on the Recent Advances in Intrusion Detection (RAID)*. Citado na pág. 8

**Anderson** *et al.* **(2013)** Ross Anderson, Chris Barton, Rainer Böhme, Richard Clayton, Michel J.G. van Eeten, Michael Levi, Tyler Moore e Stefan Savage. Measuring the cost of cybercrime. Em Rainer Böhme, editor, *The Economics of Information Security and Privacy*, páginas 265–300. Springer Berlin Heidelberg. ISBN 978-3-642-39497-3. doi: 10.1007/978-3-642-39498-0_12. URL http://dx.doi.org/10.1007/978-3-642-39498-0_12. Citado na pág. 1

**Apel** *et al.* **(2009)** Martin Apel, Joachim Biskup, Ulrich Flegel e Michael Meier. Towards early warning systems - challenges, technologies and architecture. Em Erich Rome e Robin E. Bloomfield, editors, *CRITIS*, volume 6027 of *Lecture Notes in Computer Science*, páginas 151–164. Springer. ISBN 978-3-642-14378-6. Citado na pág. 2, 15, 28, 39, 40

**Apel** *et al.* **(2010)** Martin Apel, Joachim Biskup, Ulrich Flegel e Michael Meier. Early warning system on a national level - project amsel. Em *Proceedings of the European Workshop on Internet Early Warning and Network Intelligence, EWNI*. Citado na pág. 11, 28, 29, 39

**Avvenuti** *et al.* **(2014)** M. Avvenuti, S. Cresci, M. N. La Polla, A. Marchetti e M. Tesconi. Earthquake emergency management by social sensing. Em *Pervasive Computing and Communications Workshops (PERCOM Workshops), 2014 IEEE International Conference on*, páginas 587–592. doi: 10.1109/PerComW.2014.6815272. Citado na pág. 84

**Bailey** *et al.* **(2005)** Michael Bailey, Evan Cooke, Farnam Jahanian, Jose Nazario e David Watson. The internet motion sensor: A distributed blackhole monitoring system. Em *In Proceedings of Network and Distributed System Security Symposium (NDSS'05)*, páginas 167–179. Citado na pág. 16, 35

**Bass (2000)** Tim Bass. Intrusion detection systems and multisensor data fusion. *Communications of the ACM*, 43(4):99–105. ISSN 0001-0782. doi: 10.1145/332051.332079. URL http://doi.acm.org/10.1145/332051.332079. Citado na pág. 8, 16

**Bastke** *et al.* **(2010)** Sascha Bastke, Mathias Deml e Sebastian Schmidt. Internet early warning systems - overview and architecture. Electronic document repository of the University of Dortmund, 2010. Disponível em https://eldorado.tu-dortmund.de/bitstream/2003/26690/1/4_deml.pdf. Acessado em 29/10/2014. Citado na pág. 11, 12, 13, 16, 30, 31, 40

**Batista** *et al.* **(2016)** Daniel Batista, Luiz A. F. Santos, Rodrigo Campiolo, Wagner Monteverde, Marlon F. Antonio, Thiago L. Vieira, Eder Ferreira, Rafael Silvério e Fausto Vetter. Gt ews: Building a cybersecurity ews based on social networks. TNC16 Networking Conference, June 2016. Disponível em https://tnc16.geant.org/getfile/2691. Citado na pág. 5

**Bauer e Koblentz (1988)** D.S. Bauer e M.E. Koblentz. Nidx - an expert system for real-time network intrusion detection. Em *Proceedings of the Computer Networking Symposium*, páginas 98–106. doi: 10.1109/CNS.1988.4983. Citado na pág. 7

**Benjamin e Chen (2014)** V. Benjamin e Hsinchun Chen. Time-to-event modeling for predicting hacker irc community participant trajectory. Em *IEEE Joint JISIC, 2014*, páginas 25–32. doi: 10.1109/JISIC.2014.14. Citado na pág. 96

**Benjamin** *et al.* **(2015)** V. Benjamin, W. Li, T. Holt e H. Chen. Exploring threats and vulnerabilities in hacker web: Forums, irc and carding shops. Em *Intelligence and Security Informatics (ISI), 2015 IEEE International Conference on*, páginas 85–90. doi: 10.1109/ISI.2015.7165944. Citado na pág. x, 37, 38, 39, 40, 97

**Bilge e Dumitras (2012)** Leyla Bilge e Tudor Dumitras. Before we knew it: an empirical study of zero-day attacks in the real world. Em *Proceedings of the 2012 ACM conference on Computer and communications security*, CCS '12, páginas 833–844, New York, NY, USA. ACM. ISBN 978-1-4503-1651-4. doi: 10.1145/2382196.2382284. URL http://doi.acm.org/10.1145/2382196.2382284. Citado na pág. 14

**Biskup** *et al.* **(2008)** J. Biskup, B.M. Hämmerli, M. Meier, S. Schmerl, J. Tölle e M. Vogel. 08102 working group - early warning systems. Em *Perspectives Workshop: Network Attack Detection and Defense*, volume 08102. Dagstuhl Seminar Proceedings. Citado na pág. 2, 11

**Boggs** *et al.* **(2011)** Nathaniel Boggs, Sharath Hiremagalore, Angelos Stavrou e Salvatore J. Stolfo. Cross-domain collaborative anomaly detection: So far yet so close. Em Robin Sommer, Davide Balzarotti e Gregor Maier, editors, *RAID*, volume 6961 of *Lecture Notes in Computer Science*, páginas 142–160. Springer. ISBN 978-3-642-23643-3. Citado na pág. 10, 16, 33, 39

**Bourgue** *et al.* **(2013)** Romain Bourgue, Joshua Budd, Jachym Homola, Michal Wlasenko e Dariusz Kulawik. Detect, share, protect - solutions for improving threat data exchange among certs. Relatório técnico, ENISA. Citado na pág. 2, 21

**Brown (2007)** Dugald A. Brown. Architecture for an automated irc investigation tool. Dissertação de Mestrado, West Virginia University. Citado na pág. 95

**Bsufka** *et al.* **(2006)** Karsten Bsufka, Olaf Kroll-Peters e Sahin Albayrak. Intelligent network-based early warning systems. Em Javier Lopez, editor, *Critical Information Infrastructures Security*, volume 4347 of *Lecture Notes in Computer Science*, páginas 103–111. Springer Berlin Heidelberg. ISBN 978-3-540-69083-2. doi: 10.1007/11962977_9. URL http://dx.doi.org/10.1007/11962977_9. Citado na pág. 36

**Burke (2002)** R. Burke. Hybrid recommender systems: Survey and experiments. Em *UMUAI 12*, páginas 331–370. Citado na pág. 26

**Burke (2007)** R. Burke. Hybrid web recommender systems. Em *The Adaptive Web*, páginas 377–408. Citado na pág. 26

**Burkhart** *et al.* **(2010)** Martin Burkhart, Mario Strasser, Dilip Many e Xenofontas Dimitropoulos. Sepia: Privacy-preserving aggregation of multi-domain network events and statistics. Em *Proceedings of the 19th USENIX Conference on Security*, USENIX Security'10, páginas 15–15, Berkeley, CA, USA. USENIX Association. ISBN 888-7-6666-5555-4. URL http://dl.acm.org/citation.cfm?id=1929820.1929840. Citado na pág. 24

**Bye** *et al.* **(2010)** Rainer Bye, Seyit Ahmet Camtepe e Sahin Albayrak. Collaborative intrusion detection framework: Characteristics, adversarial opportunities and countermeasures. Em *Proceedings of the 2010 International Conference on Collaborative Methods for Security and Privacy*, CollSec'10, Berkeley, CA, USA. USENIX Association. URL http://dl.acm.org/citation.cfm?id=1929808.1929810. Citado na pág. 10

**Caida (2014)** Caida. The ucsd network telescope, 2014. Disponível em http://www.caida.org/projects/network_telescope/. Acessado em 24/06/2014. Citado na pág. 16

**Campiolo e Batista (2015)** Rodrigo Campiolo e Daniel Macêdo Batista. Análise de mensagens associadas à cibersegurança em redes IRC. Em *Anais do XV Simpósio Brasileiro em Segurança da Informação e Sistemas Computacionais (SBSeg 2015)*, páginas 114–127, Florianópolis. Citado na pág. 5, 86

**Campiolo *et al.* (2013)** Rodrigo Campiolo, Luiz Arthur F. Santos, Daniel Macêdo Batista e Marco Aurélio Gerosa. Evaluating the utilization of twitter messages as a source of security alerts. Em *Proceedings of the 28th ACM SAC*, páginas 942–943. ISBN 978-1-4503-1656-9. Citado na pág. 5, 67, 84

**Carey *et al.* (2002)** Nathan Carey, Andrew Clark e George M. Mohay. Ids interoperability and correlation using idmef and commodity systems. Em *Proceedings of the 4th International Conference on Information and Communications Security*, ICICS '02, páginas 252–264, London, UK, UK. Springer-Verlag. ISBN 3-540-00164-6. URL http://dl.acm.org/citation.cfm?id=646280.756798. Citado na pág. 9, 16

**CERT Polska (2014)** CERT Polska. Arakis, 2014. Disponível em http://honeytarg.cert.br/. Acessado em 11/07/2014. Citado na pág. 35, 39

**CERT.br (2014)** CERT.br. honeytarg honeynet project, 2014. Disponível em http://honeytarg.cert.br/. Acessado em 22/06/2014. Citado na pág. 2, 15

**Chandola *et al.* (2009)** Varun Chandola, Arindam Banerjee e Vipin Kumar. Anomaly detection: A survey. *ACM Comput. Surv.*, 41(3):15:1–15:58. ISSN 0360-0300. doi: 10.1145/1541880.1541882. URL http://doi.acm.org/10.1145/1541880.1541882. Citado na pág. 17

**Chang e Lin (2011)** Chih-Chung Chang e Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm. Citado na pág. 100

**Choo (2009)** Chun Wei Choo. Information use and early warning effectiveness: Perspectives and prospects. *Journal of the American Society for Information Science and Technology*, 60(5):1071–1082. ISSN 1532-2890. doi: 10.1002/asi.21038. URL http://dx.doi.org/10.1002/asi.21038. Citado na pág. 11

**Cipriano *et al.* (2011)** Casey Cipriano, Ali Zand, Amir Houmansadr, Christopher Kruegel e Giovanni Vigna. Nexat: A history-based approach to predict attacker actions. Em *Proceedings of the 27th Annual Computer Security Applications Conference*, ACSAC '11, páginas 383–392, New York, NY, USA. ACM. ISBN 978-1-4503-0672-0. doi: 10.1145/2076732.2076787. URL http://doi.acm.org/10.1145/2076732.2076787. Citado na pág. 20

**Corona *et al.* (2013)** Igino Corona, Giorgio Giacinto e Fabio Roli. Adversarial attacks against intrusion detection systems: Taxonomy, solutions and open issues. *Information Sciences*, 239:201 – 225. ISSN 0020-0255. doi: http://dx.doi.org/10.1016/j.ins.2013.03.022. URL http://www.sciencedirect.com/science/article/pii/S0020025513002119. Citado na pág. 1

**Cremonesi *et al.* (2008)** P. Cremonesi, R. Turrin, E. Lentini e M. Matteucci. An evaluation methodology for collaborative recommender systems. Em *Proceedings of the International Conference on Automated solutions for Cross Media Content and Multi-channel Distribution*, páginas 224–231. Citado na pág. 109

**Crosbie e Spafford (1995)** Mark Crosbie e Eugene H Spafford. Defending a computer system using autonomous agents. Relatório técnico, Purdue University. Citado na pág. 8

**Cuppens (2001)** F Cuppens. Managing alerts in a multi-intrusion detection environment. Em *Annual Computer Security Applications Conference*. doi: 10.1109/ACSAC.2001.991518. Citado na pág. 9, 18

**Cuppens e Miege (2002)** F. Cuppens e A. Miege. Alert correlation in a cooperative intrusion detection framework. Em *Security and Privacy, 2002. Proceedings. 2002 IEEE Symposium on*, páginas 202–215. doi: 10.1109/SECPRI.2002.1004372. Citado na pág. 9, 18

**Dain e Cunningham (2001)** Oliver Dain e Robert K Cunningham. Fusing a heterogeneous alert stream into scenarios. Em *Proceedings of the 2001 ACM Workshop on Data Mining for Security Applications*, volume 13. Philadelphia, PA. Citado na pág. 9, 18

**Danyliw *et al.* (2007)** R. Danyliw, J. Meijer e Y. Demchenko. The Incident Object Description Exchange Format. RFC 5070 (Proposed Standard), Dezembro 2007. URL http://www.ietf.org/rfc/rfc5070.txt. Citado na pág. 22

**Dasgupta (1999)** Dipankar Dasgupta. Immunity-based intrusion detection system: a general framework. Em *Proc. of the 22nd NISSC*, volume 1, páginas 147–160. Citado na pág. 8

**Debar *et al.* (1992)** H. Debar, M. Becker e D. Siboni. A neural network component for an intrusion detection system. Em *IEEE Computer Society Symposium on Research in Security and Privacy*. doi: 10.1109/RISP.1992.213257. Citado na pág. 8

**Debar *et al.* (2007)** H. Debar, D. Curry e B. Feinstein. The Intrusion Detection Message Exchange Format (IDMEF). RFC 4765 (Experimental), Março 2007. URL http://www.ietf.org/rfc/rfc4765.txt. Citado na pág. 22, 23

**Debar e Wespi (2001)** Hervé Debar e Andreas Wespi. Aggregation and correlation of intrusion-detection alerts. Em *Proceedings of the 4th International Symposium on Recent Advances in Intrusion Detection*, RAID '00, páginas 85–103, London, UK, UK. Springer-Verlag. ISBN 3-540-42702-3. URL http://dl.acm.org/citation.cfm?id=645839.670735. Citado na pág. 9, 16, 18

**Décary-Hétu e Dupont (2012)** David Décary-Hétu e Benoit Dupont. The social network of hackers. *Global Crime*, 13(3):160–175. Citado na pág. 46, 96

**Denning (1987)** Dorothy E. Denning. An intrusion-detection model. *IEEE Transactions on Software Engineering*, 13(2):222–232. ISSN 0098-5589. doi: 10.1109/TSE.1987.232894. URL http://dx.doi.org/10.1109/TSE.1987.232894. Citado na pág. 7

**Dorges e Sander (2010)** Till Dorges e Jurgen Sander. Integrating open source information - rumors and facts in early warning. Slides in First European Workshop on Internet Early Warning and Network Intelligence (EWNI), Janeiro 2010. Citado na pág. 32, 40

**DShield (2014)** DShield. Dshield, 2014. Disponível em http://dshield.org/. Acessado em 23/06/2014. Citado na pág. 2, 16, 35, 39, 40

**Egelman *et al.* (2013)** Serge Egelman, Cormac Herley e Paul C. van Oorschot. Markets for zero-day exploits: Ethics and implications. Em *Proceedings of the 2013 Workshop on New Security Paradigms Workshop*, NSPW '13, páginas 41–46, New York, NY, USA. ACM. ISBN 978-1-4503-2582-0. doi: 10.1145/2535813.2535818. Citado na pág. 14

**Elshoush e Osman (2011)** Huwaida Tagelsir Elshoush e Izzeldin Mohamed Osman. Alert correlation in collaborative intelligent intrusion detection systems - a survey. *Applied Soft Computing*, 11(7):4349 – 4365. ISSN 1568-4946. doi: http://dx.doi.org/10.1016/j.asoc.2010.12.004. URL http://www.sciencedirect.com/science/article/pii/S156849461000311X. Citado na pág. 2, 18

**Elshoush e Osman (2013)** HuwaidaTagelsir Elshoush e IzzeldinMohamed Osman. Intrusion alert correlation framework: An innovative approach. Em Gi-Chul Yang, Sio-long Ao e Len Gelman, editors, *IAENG Transactions on Engineering Technologies*, volume 229 of *Lecture Notes in Electrical Engineering*, páginas 405–420. Springer Netherlands. ISBN 978-94-007-6189-6. doi: 10.1007/978-94-007-6190-2_31. URL http://dx.doi.org/10.1007/978-94-007-6190-2_31. Citado na pág. 10, 17

**Engelberth *et al.* (2010)** Markus Engelberth, Felix C. Freiling, Jan Göbel, Christian Gorecki, Thorsten Holz, Ralf Hund, Philipp Trinius e Carsten Willems. The inmas approach. Em *Proceedings of the European Workshop on Internet Early Warning and Network Intelligence (EWNI)*. Citado na pág. 11, 15, 16, 31

**Esposte *et al.* (2016)** Arthur Esposte, Rodrigo Campiolo, Fabio Kon e Daniel Batista. A collaboration model to recommend network security alerts based on the mixed hybrid approach. Em *Anais do SBRC 2016*, páginas 586–599, Salvador, Bahia. Citado na pág. 5

**Farah (2013)** Tanjila Farah. Algorithms and tools for anonymization of the internet traffic. Dissertação de Mestrado, Simon Frase University. Citado na pág. 23, 24

**Feitosa *et al.* (2012)** Eduardo Feitosa, Eduardo Souto e Djamel H. Sadok. An orchestration approach for unwanted internet traffic identification. *Computer Networks*, 56(12):2805–2831. ISSN 1389-1286. doi: 10.1016/j.comnet.2012.04.018. URL http://dx.doi.org/10.1016/j.comnet.2012.04.018. Citado na pág. 16, 19

**Feldman e Sanger (2007)** R. Feldman e J Sanger. *he Text Mining Handbook: Advances Approaches in Analyzing Unstructured Data*. Cambridge University Press. Citado na pág. 26

**Feldman e Sanger (2006)** Ronen Feldman e James Sanger. *Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press, New York, NY, USA. ISBN 0521836573, 9780521836579. Citado na pág. 25, 27

**Feng** *et al.* **(2009)** Li Feng, Wei Wang, Lina Zhu e Yi Zhang. Predicting intrusion goal using dynamic bayesian network with transfer probability estimation. *Journal of Network and Computer Applications*, 32(3):721 – 732. ISSN 1084-8045. doi: http://dx.doi.org/10.1016/j.jnca.2008.06.002. URL http://www.sciencedirect.com/science/article/pii/S1084804508000659. Citado na pág. 20

**Forrest** *et al.* **(1996)** S. Forrest, S.A. Hofmeyr, A. Somayaji e T.A. Longstaff. A sense of self for unix processes. Em *Security and Privacy, 1996. Proceedings., 1996 IEEE Symposium on*, páginas 120–128. doi: 10.1109/SECPRI.1996.502675. Citado na pág. 8

**Frei** *et al.* **(2006)** Stefan Frei, Martin May, Ulrich Fiedler e Bernhard Plattner. Large-scale vulnerability analysis. Em *ACM Sigcomm Workshop on Large-Scale Attack Defense*, Pisa, Italy. Sigcomm. Citado na pág. 13, 14

**Frei** *et al.* **(2008)** Stefan Frei, Bernhard Tellenbach e Bernhard Plattner. 0-day patch - exposing vendors (in)security performance. Em *BlackHat Europe 2008*, Amsterdam. Citado na pág. 13

**Freiling (2010)** Felix C. Freiling. What is a early warning system? Slides in First European Workshop on Internet Early Warning and Network Intelligence (EWNI), Janeiro 2010. Citado na pág. 11

**Frincke (2000)** Deborah Frincke. Balancing cooperation and risk in intrusion detection. *ACM Trans. Inf. Syst. Secur.*, 3(1):1–29. ISSN 1094-9224. doi: 10.1145/353323.353324. URL http://doi.acm.org/10.1145/353323.353324. Citado na pág. 9

**Fung (2013)** Carol Fung. *Design and Management of Collaborative Intrusion Detection Networks*. Tese de Doutorado, University of Waterloo. Citado na pág. 24

**Gainaru** *et al.* **(2010)** Ana Gainaru, Stefan Daniel Dumitrescu e Stefan Trausan-Matu. Toolkit for automatic analysis of chat conversations. Em *8th COMM 2010 (IEEE)*, páginas 99–102. Citado na pág. 96

**Geib e Goldman (2001)** C.W. Geib e R.P. Goldman. Plan recognition in intrusion detection systems. Em *DARPA Information Survivability Conference amp; Exposition II, 2001. DISCEX '01. Proceedings*, volume 1, páginas 46–55 vol.1. doi: 10.1109/DISCEX.2001.932191. Citado na pág. 19

**Ghorbani** *et al.* **(2010)** AliA. Ghorbani, Wei Lu e Mahbod Tavallaee. Concepts and techniques. Em *Network Intrusion Detection and Prevention*, volume 47 of *Advances in Information Security*. Springer US. ISBN 978-0-387-88770-8. Citado na pág. 15, 16, 18

**Gopalratnam e Cook (2007)** K. Gopalratnam e D.J. Cook. Online sequential prediction via incremental parsing: The active lezi algorithm. *Intelligent Systems, IEEE*, 22(1):52–58. ISSN 1541-1672. doi: 10.1109/MIS.2007.15. Citado na pág. 19

**Gorzelak** *et al.* **(2011)** Katarzyna Gorzelak, Tomasz Grudziecki, Paweł Jacewicz, Przemysław Jaroszewski, Łukasz Juszczyk e Piotr Kijewski. Proactive detection of network security incidents. Relatório técnico, ENISA. Citado na pág. 37

**Grobauer** *et al.* **(2006)** Bernd Grobauer, Jens Ingo Mehlau e Jürgen Sander. Carmentis: A co-operative approach towards situation awareness and early warning for the internet. Em Oliver Göbel, Dirk Schadt, Sandra Frings, Hardo Hase, Detlef Günther e Jens Nedon, editors, *IMF*, volume 97 of *LNI*, páginas 55–66. GI. ISBN 978-3-88579-191-1. Citado na pág. 2, 16, 29, 30, 39

**Gunawardana e Shani (2011)** Asela Gunawardana e Guy Shani. *Recommender Systems Handbook*, chapter Evaluating Recommender Systems, páginas 265–308. Springer US, Boston, MA. ISBN 978-1-4899-7637-6. doi: 10.1007/978-1-4899-7637-6_8. URL http://dx.doi.org/10.1007/978-1-4899-7637-6_8. Citado na pág. 103

**Harper e Konstan (2015)** F. Maxwell Harper e Joseph A. Konstan. The movielens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems*, 5(4):19:1–19:19. ISSN 2160-6455. doi: 10.1145/2827872. URL http://doi.acm.org/10.1145/2827872. Citado na pág. 108

**Haslum** *et al.* **(2008)** K. Haslum, A. Abraham e Svein Knapskog. Fuzzy online risk assessment for distributed intrusion prediction and prevention systems. Em *Computer Modeling and Simulation, 2008. UKSIM 2008. Tenth International Conference on*, páginas 216–223. doi: 10.1109/UKSIM.2008.30. Citado na pág. 20

**Heberlein** *et al.* **(1990)** L.T. Heberlein, G.V. Dias, K.N. Levitt, B. Mukherjee, J. Wood e D. Wolber. A network security monitor. Em *Proceedings. IEEE Symposium on Research in Security and Privacy*, páginas 296–304. doi: 10.1109/RISP.1990.63859. Citado na pág. 7

**Herlocker** *et al.* **(2004)** J. Herlocker, J. Konstan, L. Terveen e J. Riedl. Evaluating collaborative filtering recommender systems. Em *ACM Transactions on Information Systems (TOIS)*, páginas 5–53. Citado na pág. 109

**Hesse e Pohlmann (2008)** M. Hesse e N. Pohlmann. Internet situation awareness. Em *eCrime Researchers Summit*, páginas 1–9. doi: 10.1109/ECRIME.2008.4696966. Citado na pág. 36

**Hochberg** *et al.* **(1993)** Judith Hochberg, Kathleen Jackson, Cathy Stallings, J.F. McClary, David DuBois e Josephine Ford. Nadir: An automated system for detecting network intrusion and misuse. *Computers & Security*, 12(3):235 – 248. ISSN 0167-4048. doi: http://dx.doi.org/10.1016/0167-4048(93)90110-Q. URL http://www.sciencedirect.com/science/article/pii/016740489390110Q. Citado na pág. 7, 17

**Holt** *et al.* **(2012)** Thomas J Holt, Deborah Strumsky, Olga Smirnova e Max Kilger. Examining the social networks of malware writers and hackers. *International Journal of Cyber Criminology*, 6(1):891–903. Citado na pág. 68

**Honeynet (2014)** Honeynet. The honeynet project, 2014. Disponível em http://www.honeynet.org/. Acessado em 22/06/2014. Citado na pág. 15

**ICASI (2014)** ICASI. The common vulnerability reporting framework, Agosto 2014. Disponível em http://www.icasi.org/cvrf. Acessado em 13/08/2014. Citado na pág. 22

**Ilgun (1993)** K. Ilgun. Ustat: a real-time intrusion detection system for unix. Em *Research in Security and Privacy, 1993. Proceedings., 1993 IEEE Computer Society Symposium on*, páginas 16–28. doi: 10.1109/RISP.1993.287646. Citado na pág. 8

**Ilgun** *et al.* **(1995)** K. Ilgun, R.A. Kemmerer e P.A. Porras. State transition analysis: a rule-based intrusion detection approach. *Software Engineering, IEEE Transactions on*, 21(3):181–199. ISSN 0098-5589. doi: 10.1109/32.372146. Citado na pág. 8

**Indurkhya e Damerau (2010)** Nitin Indurkhya e Fred J. Damerau. *Handbook of Natural Language Processing*. Chapman & Hall/CRC, 2nd ed. ISBN 1420085921, 9781420085921. Citado na pág. 27

**Inoue** *et al.* **(2009)** Daisuke Inoue, Mio Suzuki, Masashi Eto, Katsunari Yoshioka e Koji Nakao. Daedalus: Novel application of large-scale darknet monitoring for practical protection of live networks. Em *Proceedings of the 12th International Symposium on Recent Advances in Intrusion Detection*, RAID '09, páginas 381–382, Berlin, Heidelberg. Springer-Verlag. ISBN 978-3-642-04341-3. doi: 10.1007/978-3-642-04342-0_33. Citado na pág. 16

**Iqbal** *et al.* **(2012)** Farkhund Iqbal, Benjamin C. M. Fung e Mourad Debbabi. Mining criminal networks from chat log. Em *IEEE/WIC/ACM*. Citado na pág. 96

**Jackson** *et al.* **(1991)** Kathleen A Jackson, David H DuBois e Cathy A Stallings. An expert system application for network intrusion detection. Relatório técnico, Los Alamos National Lab., NM (United States). Citado na pág. 7

**Joshi** *et al.* **(2013)** Arnav Joshi, Ravendar Lal, Tim Finin e Anupam Joshi. Extracting cybersecurity related linked data from text. Em *Proceedings of the 7th IEEE International Conference on Semantic Computing*. IEEE Computer Society Press. Citado na pág. 37, 38, 40

**Jumratjaroenvanit e Teng-Amnuay (2008)** Amontip Jumratjaroenvanit e Yunyong Teng-Amnuay. Probability of attack based on system vulnerability life cycle. Em Fei Yu, Qi Luo, Yongjun Chen e Zhigang Chen, editors, *ISECS*, páginas 531–535. IEEE Computer Society. ISBN 978-0-7695-3258-5. Citado na pág. 14

**Kalt (2000a)** C. Kalt. Internet Relay Chat: Architecture. RFC 2810 (Informational), Abril 2000a. URL http://www.ietf.org/rfc/rfc2810.txt. Citado na pág. 86

**Kalt (2000b)** C. Kalt. Internet Relay Chat: Channel Management. RFC 2811 (Informational), Abril 2000b. URL http://www.ietf.org/rfc/rfc2811.txt. Citado na pág. 86

**Kalt (2000c)** C. Kalt. Internet Relay Chat: Client Protocol. RFC 2812 (Informational), Abril 2000c. URL http://www.ietf.org/rfc/rfc2812.txt. Citado na pág. 86

**Karypis (2001)** G. Karypis. Evaluation of item-based top-n recommendation algorithms. Em *10th Conference of Information and Knowledge Management*, páginas 247–254. Citado na pág. 108

**Katti *et al.* (2005)** Sachin Katti, Balachander Krishnamurthy e Dina Katabi. Collaborating against common enemies. Em *Internet Measurement Conference*. Citado na pág. 10

**Kim *et al.* (2008)** Sehun Kim, Seong-Jun Shin, Hyunwoo Kim, Ki Hoon Kwon e Younggoo Han. Hybrid intrusion forecasting framework for early warning system. *IEICE Transactions on Information and Systems*, 91-D:1234–1241. doi: 10.1093/ietisy/e91-d.5.1234. Citado na pág. 36

**Kirubavathi e Anitha (2014)** G. Kirubavathi e R. Anitha. Botnets: A study and analysis. Em *Computational Intelligence, Cyber Security and Computational Models*, volume 246 of *Advances in Intelligent Systems and Computing*, páginas 203–214. Springer India. ISBN 978-81-322-1679-7. doi: 10.1007/978-81-322-1680-3_23. Citado na pág. 1

**Koch (2011)** R. Koch. Towards next-generation intrusion detection. Em *Cyber Conflict (ICCC), 2011 3rd International Conference on*, páginas 1–18. Citado na pág. 11

**Kosoresow e Hofmeyer (1997)** A.P. Kosoresow e S.A. Hofmeyer. Intrusion detection via system call traces. *Software, IEEE*, 14(5):35–42. ISSN 0740-7459. doi: 10.1109/52.605929. Citado na pág. 8

**Kostkova *et al.* (2014)** Patty Kostkova, Martin Szomszor e Connie St. Louis. #swineflu: The use of twitter as an early warning and risk communication tool in the 2009 swine flu pandemic. *ACM Trans. Manage. Inf. Syst.*, 5(2):8:1–8:25. ISSN 2158-656X. doi: 10.1145/2597892. URL http://doi.acm.org/10.1145/2597892. Citado na pág. 84

**Kruegel *et al.* (2005)** Christopher Kruegel, Fredrik Valeur e Giovanni Vigna. Alert collection. Em *Intrusion Detection and Correlation - Challenges and Solutions*, volume 14 of *Advances in Information Security*, páginas 35–42. Springer US. ISBN 978-0-387-23398-7. doi: 10.1007/0-387-23399-7_4. Citado na pág. 16, 17

**Krügel *et al.* (2002)** Christopher Krügel, Thomas Toth e Clemens Kerer. Decentralized event correlation for intrusion detection. Em Kwangjo Kim, editor, *Information Security and Cryptology - ICISC 2001*, volume 2288 of *Lecture Notes in Computer Science*, páginas 114–131. Springer Berlin Heidelberg. ISBN 978-3-540-43319-4. doi: 10.1007/3-540-45861-1_10. URL http://dx.doi.org/10.1007/3-540-45861-1_10. Citado na pág. 9, 16

**Kumar e Spafford (1994)** Sandeep Kumar e Eugene H. Spafford. An application of pattern matching in intrusion detection. Relatório técnico, Purdue University. Citado na pág. 8

**Kwak *et al.* (2010)** Haewoon Kwak, Changhyun Lee, Hosung Park e Sue Moon. What is twitter, a social network or a news media? Em *Proceedings of the 19th international conference on World wide web*, WWW '10, páginas 591–600, New York, NY, USA. ACM. ISBN 978-1-60558-799-8. doi: 10.1145/1772690.1772751. Citado na pág. 67, 73, 83

**Lee e Kim (2013)** Sangho Lee e Jong Kim. Warningbird: A near real-time detection system for suspicious urls in twitter stream. *Dependable and Secure Computing, IEEE Transactions on*, 10(3):183–195. ISSN 1545-5971. doi: 10.1109/TDSC.2013.3. Citado na pág. 16

**Lee e Stolfo (1998)** Wenke Lee e Salvatore J. Stolfo. Data mining approaches for intrusion detection. Em *Proceedings of the 7th Conference on USENIX Security Symposium*, volume 7 of *SSYM'98*, Berkeley, CA, USA. USENIX Association. URL http://dl.acm.org/citation.cfm?id=1267549.1267555. Citado na pág. 8

**Lerman e Ghosh (2010)** K. Lerman e R. Ghosh. Information contagion: An empirical study of the spread of news on digg and twitter social networks. Em *Proceedings of 4th ICWSM*. Citado na pág. 73, 83

**Levy (2004)** E. Levy. Approaching zero [attack trends]. *Security Privacy, IEEE*, 2(4):65–66. ISSN 1540-7993. doi: 10.1109/MSECP.2004.1281250. Citado na pág. 13

**Li e Chen (2014)** W. Li e H. Chen. Identifying top sellers in underground economy using deep learning-based sentiment analysis. Em *Intelligence and Security Informatics Conference (JISIC), 2014 IEEE Joint*, páginas 64–67. doi: 10.1109/JISIC.2014.19. Citado na pág. 39

**Li e Tian (2010)** Wan Li e Shengfeng Tian. An ontology-based intrusion alerts correlation system. *Expert Syst. Appl.*, 37(10):7138–7146. ISSN 0957-4174. doi: 10.1016/j.eswa.2010.03.068. URL http://dx.doi.org/10.1016/j.eswa.2010.03.068. Citado na pág. 10

**Limmer e Dressler (2008)** Tobias Limmer e Falko Dressler. Survey of event correlation techniques for attack detection in early warning systems. Technical Report 01/08, University of Erlangen, Dept. of Computer Science. Citado na pág. 12

**Lincoln *et al.* (2004)** Patrick Lincoln, Phillip Porras e Vitaly Shmatikov. Privacy-preserving sharing and correlation of security alerts. Em *In USENIX Security Symposium*, páginas 239–254. Citado na pág. 23, 24

**Lippmann *et al.* (2000a)** Richard Lippmann, Joshua W Haines, David J Fried, Jonathan Korba e Kumar Das. The 1999 darpa off-line intrusion detection evaluation. *Computer networks*, 34(4):579–595. doi: 10.1016/S1389-1286(00)00139-0. Citado na pág. 9

**Lippmann *et al.* (2000b)** R.P. Lippmann, D.J. Fried, I. Graf, J.W. Haines, K.R. Kendall, D. McClung, D. Weber, S.E. Webster, D. Wyschogrod, R.K. Cunningham e M.A. Zissman. Evaluating intrusion detection systems: the 1998 darpa off-line intrusion detection evaluation. Em *DARPA Information Survivability Conference and Exposition, 2000. DISCEX '00. Proceedings*, volume 2, páginas 12–26. doi: 10.1109/DISCEX.2000.821506. Citado na pág. 9

**Locasto *et al.* (2005)** Michael E Locasto, Janak J Parekh, Angelos D Keromytis e Salvatore J Stolfo. Towards collaborative security and p2p intrusion detection. Em *Information Assurance Workshop, 2005. IAW'05. Proceedings from the Sixth Annual IEEE SMC*, páginas 333–339. IEEE. Citado na pág. 33, 39

**Lodi *et al.* (2014)** Giorgia Lodi, Leonardo Aniello, Giuseppe A. Di Luna e Roberto Baldoni. An event-based platform for collaborative threats detection and monitoring. *Information Systems*, 39:175 – 195. ISSN 0306-4379. doi: http://dx.doi.org/10.1016/j.is.2013.07.005. URL http://www.sciencedirect.com/science/article/pii/S0306437913001014. Citado na pág. 34, 39, 40

**Lunt e Jagannathan (1988)** T.F. Lunt e R. Jagannathan. A prototype real-time intrusion-detection expert system. Em *Security and Privacy, 1988. Proceedings., 1988 IEEE Symposium on*, páginas 59–66. doi: 10.1109/SECPRI.1988.8098. Citado na pág. 7

**Lunt *et al.* (1989)** T.F. Lunt, R. Jagannathan, R. Lee, A. Whitehurst e S. Listgarten. Knowledge-based intrusion detection. Em *Proceedings of the Annual AI Systems in Government Conference.*, páginas 102–107. doi: 10.1109/AISIG.1989.47311. Citado na pág. 7

**Manning e Schütze (1999)** Christopher D. Manning e Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, USA. ISBN 0-262-13360-1. Citado na pág. 27

**Manning *et al.* (2008)** Christopher D. Manning, Prabhakar Raghavan e Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, UK. ISBN 978-0-521-86571-5. Citado na pág. 25, 27, 28, 79

**Marmanis e Babenko (2009)** Haralambos Marmanis e Dmitry Babenko. *Algorithms of the Intelligent Web*. Manning Publications Co., Greenwich, CT, USA, 1st ed. ISBN 1933988665, 9781933988665. Citado na pág. 25

**Mathews *et al.* (2012)** M.L. Mathews, P. Halvorsen, A. Joshi e T. Finin. A collaborative approach to situational awareness for cybersecurity. Em *Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom), 2012 8th International Conference on*, páginas 216–222. Citado na pág. 16

**Mchugh *et al.* (2000)** John Mchugh, Alan M. Christie e Julia Allen. Defending yourself: The role of intrusion detection systems. *IEEE Software*, 17:42–51. doi: 10.1109/52.877859. Citado na pág. 9

**Meissen e Voisard (2010)** Ulrich Meissen e Agnes Voisard. Towards a reference architecture for early warning systems. Em *Proceedings of the 2010 International Conference on Intelligent Networking and Collaborative Systems*, INCOS '10, páginas 513–518, Washington, DC, USA. IEEE Computer Society. ISBN 978-0-7695-4278-2. doi: 10.1109/INCOS.2010.81. URL http://dx.doi.org/10.1109/INCOS.2010.81. Citado na pág. 36

**Meng *et al.* (2015)** Guozhu Meng, Yang Liu, Jie Zhang, Alexander Pokluda e Raouf Boutaba. Collaborative security: A survey and taxonomy. *ACM Computing Surveys (CSUR)*, 48(1):1:1–1:42. ISSN 0360-0300. doi: 10.1145/2785733. URL http://doi.acm.org/10.1145/2785733. Citado na pág. 1, 10

**Michels (2012)** Marvin O. Michels. Real time text analysis on internet relay chat conversations. Dissertação de Mestrado, Purdue University. Citado na pág. 86, 96

**Mirheidari *et al.* (2013)** SeyedAli Mirheidari, Sajjad Arshad e Rasool Jalili. Alert correlation algorithms: A survey and taxonomy. Em Guojun Wang, Indrakshi Ray, Dengguo Feng e Muttukrishnan Rajarajan, editors, *Cyberspace Safety and Security*, volume 8300 of *Lecture Notes in Computer Science*, páginas 183–197. Springer International Publishing. ISBN 978-3-319-03583-3. doi: 10.1007/978-3-319-03584-0_14. URL http://dx.doi.org/10.1007/978-3-319-03584-0_14. Citado na pág. 2

**MITRE (2014a)** MITRE. Common vulnerabilities and exposures, Agosto 2014a. Disponível em https://cve.mitre.org/. Acessado em 13/08/2014. Citado na pág. 22

**MITRE (2014b)** MITRE. Cyber observable expression, Agosto 2014b. Disponível em http://cybox.mitre.org/. Acessado em 13/08/2014. Citado na pág. 22

**MITRE (2014c)** MITRE. Structured threat information expression, Agosto 2014c. Disponível em https://stix.mitre.org/. Acessado em 13/08/2014. Citado na pág. 22

**MITRE (2014d)** MITRE. Trusted automated exchange of indicator information, Agosto 2014d. Disponível em http://taxii.mitre.org/. Acessado em 13/08/2014. Citado na pág. 22

**More *et al.* (2012)** Sumit More, Mary Matthews, Anupam Joshi e Tim Finin. A knowledge-based approach to intrusion detection modeling. Em *IEEE Symposium on Security and Privacy Workshops*, páginas 75–81. IEEE Computer Society. ISBN 978-1-4673-2157-0. Citado na pág. 16, 37, 40, 111

**Morin *et al.* (2009)** Benjamin Morin, Ludovic Mé, Hervé Debar e Mireille Ducassé. A logic-based model to support alert correlation in intrusion detection. *Information Fusion*, 10(4):285 – 299. ISSN 1566-2535. doi: http://dx.doi.org/10.1016/j.inffus.2009.01.005. URL http://www.sciencedirect.com/science/article/pii/S1566253509000177. Special Issue on Information Fusion in Computer Security. Citado na pág. 10, 18

**Morris *et al.* (2012)** Meredith Ringel Morris, Scott Counts, Asta Roseway, Aaron Hoff e Julia Schwarz. Tweeting is believing?: understanding microblog credibility perceptions. Em *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, CSCW '12, páginas 441–450, New York, NY, USA. ACM. ISBN 978-1-4503-1086-4. doi: 10.1145/2145204.2145274. Citado na pág. 67, 71, 79

**Morzy (2011)** Mikolaj Morzy. Internet forums: What knowledge can be mined from online discussions. Em A.V. Senthil Kumar, editor, *Knowledge Discovery Practices and Emerging Applications of Data Mining: Trends and New Domains*, chapter 15, páginas 315–336. IGI Global. doi: 10.4018/978-1-60960-067-9.ch015. Citado na pág. 76

**Mulwad *et al.* (2011)** Varish Mulwad, Wenjia Li, Anupam Joshi, Tim Finin e Krishnamurthy Viswanathan. Extracting information about security vulnerabilities from web text. Em *Proceedings of the 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology - Volume 03*, WI-IAT '11, páginas 257–260, Washington, DC, USA. IEEE Computer Society. ISBN 978-0-7695-4513-4. doi: 10.1109/WI-IAT.2011.26. URL http://dx.doi.org/10.1109/WI-IAT.2011.26. Citado na pág. 37, 40

**Ning *et al.* (2001)** Peng Ning, Sushil Jajodia e Xiaoyang Sean Wang. Abstraction-based intrusion detection in distributed environments. *ACM Transactions on Information and System Security*, 4:407–452. doi: 10.1145/503339.503342. Citado na pág. 9

**Ning *et al.* (2002)** Peng Ning, Yun Cui e Douglas S. Reeves. Constructing attack scenarios through correlation of intrusion alerts. Em *Proceedings of the 9th ACM Conference on Computer and Communications Security*, CCS '02, páginas 245–254, New York, NY, USA. ACM. ISBN 1-58113-612-9. doi: 10.1145/586110.586144. URL http://doi.acm.org/10.1145/586110.586144. Citado na pág. 9, 18

**Ning *et al.* (2004)** Peng Ning, Yun Cui, Douglas S. Reeves e Dingbang Xu. Techniques and tools for analyzing intrusion alerts. *ACM Transactions on Information and System Security*, 7:274–318. doi: 10.1145/996943.996947. Citado na pág. 9

**NIST (2014)** NIST. National vulnerability database, Agosto 2014. Disponível em http://nvd.nist.gov/. Acessado em 13/08/2014. Citado na pág. 22

**Nottingham e Sayre (2005)** M. Nottingham e R. Sayre. The Atom Syndication Format. RFC 4287 (Proposed Standard), Dezembro 2005. URL http://www.ietf.org/rfc/rfc4287.txt. Updated by RFC 5988. Citado na pág. 46

**OASIS (2010)** OASIS. Common alerting protocol version 1.2, 2010. Citado na pág. 17

**Oikarinen e Reed (1993)** J. Oikarinen e D. Reed. Internet Relay Chat Protocol. RFC 1459 (Experimental), Maio 1993. URL http://www.ietf.org/rfc/rfc1459.txt. Updated by RFCs 2810, 2811, 2812, 2813. Citado na pág. 86

**OpenIoC (2014)** OpenIoC. The openioc framework, Agosto 2014. Disponível em http://www.openioc.org/. Acessado em 13/08/2014. Citado na pág. 22

**Parekh (2007)** Janak J. Parekh. *Privacy-Preserving Distributed Event Corroboration*. Tese de Doutorado, Columbia University. Citado na pág. 23, 24

**Paxson (1999)** Vern Paxson. Bro: A system for detecting network intruders in real-time. *Computer Networks*, 31(23-24):2435–2463. URL http://www.icir.org/vern/papers/bro-CN99.pdf. Citado na pág. 7

**Phuvipadawat e Murata (2010)** S. Phuvipadawat e T. Murata. Breaking news detection and tracking in twitter. Em *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*, volume 3, páginas 120 –123. doi: 10.1109/WI-IAT.2010.205. Citado na pág. 67

**Picault *et al.* (2011)** Jérome Picault, Myriam Ribière, David Bonnefoy e Kevin Mercer. *Recommender Systems Handbook*, chapter How to Get the Recommender Out of the Lab?, páginas 333–365. Springer US, Boston, MA. ISBN 978-0-387-85820-3. doi: 10.1007/978-0-387-85820-3_10. URL http://dx.doi.org/10.1007/978-0-387-85820-3_10. Citado na pág. 101

**Pohlmann e Proest (2006)** Norbert Pohlmann e Marcus Proest. Internet early warning system: The global view. Em *ISSE 2006 - Securing Electronic Busines Processes*, páginas 377–386. Vieweg. ISBN 978-3-8348-0213-2. doi: 10.1007/978-3-8348-9195-2_40. URL http://dx.doi.org/10.1007/978-3-8348-9195-2_40. Citado na pág. 36

**Pontes e Guelfi (2009)** E. Pontes e A.E. Guelfi. Ifs - intrusion forecasting system based on collaborative architecture. Em *Digital Information Management, 2009. ICDIM 2009. Fourth International Conference on*, páginas 1–6. doi: 10.1109/ICDIM.2009.5356772. Citado na pág. 20, 36

**Porras e Kemmerer (1992)** P.A. Porras e R.A. Kemmerer. Penetration state transition analysis: A rule-based intrusion detection approach. Em *Computer Security Applications Conference, 1992. Proceedings., Eighth Annual*, páginas 220–229. doi: 10.1109/CSAC.1992.228217. Citado na pág. 8

**Porras e Shmatikov (2006)** Phillip Porras e Vitaly Shmatikov. Large-scale collection and sanitization of network security data: Risks and challenges. Em *Proceedings of the 2006 Workshop on New Security Paradigms*, NSPW '06, páginas 57–64, New York, NY, USA. ACM. ISBN 978-1-59593-923-4. doi: 10.1145/1278940.1278949. URL http://doi.acm.org/10.1145/1278940.1278949. Citado na pág. 17, 23, 24

**Porras e Neumann (1997)** Phillip A. Porras e Peter G. Neumann. Emerald: Event monitoring enabling responses to anomalous live disturbances. Em *In Proceedings of the 20th National Information Systems Security Conference*, páginas 353–365. Citado na pág. 8

**Porras *et al.* (2002)** Phillip A. Porras, Martin W. Fong e Alfonso Valdes. A mission-impact-based approach to infosec alarm correlation. Em *Proceedings of the 5th International Conference on Recent Advances in Intrusion Detection*, RAID'02, páginas 95–114, Berlin, Heidelberg. Springer-Verlag. ISBN 3-540-00020-8. URL http://dl.acm.org/citation.cfm?id=1754701.1754710. Citado na pág. 9, 12

**Qin e Lee (2003)** Xinzhou Qin e Wenke Lee. Statistical causality analysis of infosec alert data. Em *In Proceedings of The 6th International Symposium on Recent Advances in Intrusion Detection (RAID 2003*, páginas 73–93. Citado na pág. 9, 12, 16, 18

**Qin e Lee (2004)** Xinzhou Qin e Wenke Lee. Attack plan recognition and prediction using causal networks. Em *Computer Security Applications Conference, 2004. 20th Annual*, páginas 370–379. doi: 10.1109/CSAC.2004.7. Citado na pág. 20

**Rajaraman e Ullman (2011)** Anand Rajaraman e Jeffrey David Ullman. *Mining of massive datasets*. Cambridge University Press, Cambridge. ISBN 9781139157926 1139157922 9781107015357 1107015359. Citado na pág. 26

**Ricci *et al.* (2011)** Francesco Ricci, Lior Rokach e Bracha Shapira. Introduction to recommender systems handbook. Em Francesco Ricci, Lior Rokach, Bracha Shapira e Paul B. Kantor, editors, *Recommender Systems Handbook*, páginas 1–35. Springer US. ISBN 978-0-387-85819-7. doi: 10.1007/978-0-387-85820-3_1. URL http://dx.doi.org/10.1007/978-0-387-85820-3_1. Citado na pág. 26, 102

**Ritter *et al.* (2015)** Alan Ritter, Evan Wright, William Casey e Tom Mitchell. Weakly supervised extraction of computer security events from twitter. Em *Proceedings of the 24th International Conference on World Wide Web*, WWW 15, páginas 896–905, New York, NY, USA. ACM. ISBN 978-1-4503-3469-3. doi: 10.1145/2736277.2741083. URL http://doi.acm.org/10.1145/2736277.2741083. Citado na pág. 37, 39, 40, 84

**Robertson *et al.* (2010)** M. Robertson, Yin Pan e Bo Yuan. A social approach to security: Using social networks to help detect malicious web content. Em *Intelligent Systems and Knowledge Engineering (ISKE), 2010 International Conference on*, páginas 436–441. doi: 10.1109/ISKE.2010.5680839. Citado na pág. 16

**Rodrigues (2012)** Thiago Gomes Rodrigues. Araponga: Uma ferramenta de apoio a recuperação de informação na web voltado a segurança de redes e sistemas. Dissertação de Mestrado, Universidade de Pernambuco. Citado na pág. 37, 38, 40, 111

**Roesch (1999)** Martin Roesch. Snort - lightweight intrusion detection for networks. Em *Proceedings of the 13th USENIX Conference on System Administration*, LISA '99, páginas 229–238, Berkeley, CA, USA. USENIX Association. URL http://dl.acm.org/citation.cfm?id=1039834.1039864. Citado na pág. 7

**Roschke *et al.* (2010)** S. Roschke, Feng Cheng e C. Meinel. A flexible and efficient alert correlation platform for distributed ids. Em *Network and System Security (NSS), 2010 4th International Conference on*, páginas 24–31. doi: 10.1109/NSS.2010.26. Citado na pág. 10, 16

**RSS Advisory Board (2016)** RSS Advisory Board. Rss 2.0 specification, Maio 2016. URL http://www.rssboard.org/rss-specification. Acessado em 01 de agosto de 2016. Citado na pág. 46

**Russell (2011)** Matthew A. Russell. *Mining the Social Web - Analyzing Data from Facebook, Twitter, LinkedIn, and Other Social Media Sites*. O'Reilly. ISBN 978-1-449-38834-8. Citado na pág. 67

**Sakaki *et al.* (2010)** Takeshi Sakaki, Makoto Okazaki e Yutaka Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. Em *Proceedings of the 19th international conference on World wide web*, WWW '10, páginas 851–860, New York, NY, USA. ACM. ISBN 978-1-60558-799-8. doi: 10.1145/1772690.1772777. Citado na pág. 83

**Salah *et al.* (2013)** Saeed Salah, Gabriel MaciÁ-FernÁndez e Jesús E. DÍaz-Verdejo. A model-based survey of alert correlation techniques. *Computer Networks*, 57(5):1289 – 1317. ISSN 1389-1286. doi: http://dx.doi.org/10.1016/j.comnet.2012.10.022. URL http://www.sciencedirect.com/science/article/pii/S1389128612004124. Citado na pág. 2, 18

**Santos *et al.* (2012)** L. A. F. Santos, R. Campiolo, M. A. Gerosa e D. M. Batista. Analysis of security messages posted on twitter. Em *Collaborative Systems (SBSC), 2012 Brazilian Symposium on*, páginas 20–28. doi: 10.1109/SBSC.2012.10. Citado na pág. 5, 25, 67, 68, 71, 84

**Santos** *et al.* **(2013)** L. A. F. Santos, R. Campiolo, M. A. Gerosa e D. M. Batista. Detecção de alertas de segurança em redes de computadores usando redes sociais. Em *Anais do XXXI SBRC*, páginas 791–804. Citado na pág. 5, 67, 78, 80

**Santos** *et al.* **(2014)** Luiz Arthur F. Santos, Rodrigo Campiolo e Daniel Macêdo Batista. Uma arquitetura autonômica para detecção e reação a ameaças de segurança em redes de computadores. Em *Anais do 4o Workshop em Sistemas Distribuídos Autonômicos, SBRC 2014*, páginas 45–48. Citado na pág. 5

**Santos** *et al.* **(2016)** Luiz Arthur Feitosa Santos, Rodrigo Campiolo, Wagner Monteverde e Daniel Batista. Abordagem autonômica para mitigar ciberataques em lans. Em *Anais do SBRC 2016*, páginas 600–613, Salvador, Bahia. Citado na pág. 5

**Sebring** *et al.* **(1988)** Michael M Sebring, Eric Shellhouse, Mary Hanna e R Whitehurst. Expert systems in intrusion detection: A case study. Em *Proceedings of the 11th National Computer Security Conference*, páginas 74–81. Citado na pág. 7

**Shahzad** *et al.* **(2012)** Muhammad Shahzad, Muhammad Zubair Shafiq e Alex X. Liu. A large scale exploratory analysis of software vulnerability life cycles. Em *ICSE'12*, páginas 771–781. Citado na pág. 1, 14

**Shirey (2000)** R. Shirey. Internet Security Glossary. RFC 2828 (Informational), Maio 2000. URL http://www.ietf.org/rfc/rfc2828.txt. Obsoleted by RFC 4949. Citado na pág. 13

**Smaha (1988)** S.E. Smaha. Haystack: an intrusion detection system. Em *Aerospace Computer Security Applications Conference, 1988., Fourth*, páginas 37–44. doi: 10.1109/ACSAC.1988.113412. Citado na pág. 7

**Snapp** *et al.* **(1991)** Steven R. Snapp, James Brentano, Gihan V. Dias, Terrance L. Goan, L. Todd Heberlein, Che-Lin Ho, Karl N. Levitt, Biswanath Mukherjee, Stephen E. Smaha, Tim Grance, Daniel M. Teal e Doug Mansur. Dids (distributed intrusion detection system)-motivation, architecture, and an early prototype. Em *Proceedings of the 14th National Computer Security Conference*, páginas 167–176. Citado na pág. 8

**Spitzner (2002)** Lance Spitzner. *Honeypots: Tracking Hackers*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA. ISBN 0321108957. Citado na pág. 15

**Staniford-Chen** *et al.* **(1996)** Stuart Staniford-Chen, Steven Cheung, Richard Crawford, Mark Dilger, Jeremy Frank, James Hoagland, Karl Levitt, Christopher Wee, Raymond Yip e Dan Zerkle. Grids-a graph based intrusion detection system for large networks. Em *Proceedings of the 19th national information systems security conference*, volume 1, páginas 361–370. Baltimore. Citado na pág. 8

**Stolfo (2004)** Salvatore J. Stolfo. Worm and attack early warning. *IEEE Security and Privacy*, 2(3):73–75. ISSN 1540-7993. doi: 10.1109/MSP.2004.28. URL http://dx.doi.org/10.1109/MSP.2004.28. Citado na pág. 33

**Symantec (2014)** Symantec. Deepsight early warning services. Portal Web, Agosto 2014. Disponível em: https://tms.symantec.com/. Acessado em 01/08/2014. Citado na pág. 2, 35, 40

**Theilmann (2010)** A. Theilmann. Beyond centralism: The herold approach to sensor networks and early warning systems. Em *First European Workshop of Internet Early Warning and Network Intelligence (EWNI 2010)*. Citado na pág. 37, 39

**Tian** *et al.* **(2005)** Junfeng Tian, Jianqiang Zhai, Ruizhong Du e Jiancai Huang. Early warning model of network intrusion based on d-s evidence theory. *Journal of Electronics (China)*, 22(3):261–267. ISSN 0217-9822. doi: 10.1007/BF02687981. URL http://dx.doi.org/10.1007/BF02687981. Citado na pág. 9

**Trustwave (2014)** Trustwave. 2014 trustwave global security report. Relatório técnico, Trustwave. Citado na pág. 1

**Ukkonen (1995)** E. Ukkonen. On-line construction of suffix trees. *Algorithmica*, 14(3):249–260. ISSN 0178-4617. doi: 10.1007/BF01206331. URL http://dx.doi.org/10.1007/BF01206331. Citado na pág. 28

**United Nations (2006)** United Nations. *Global Survey of Early Warning Systems: an Assessment of Capacities, Gaps and Opportunities Toward Building a Comprehensive Global Early Warning System for All Natural Hazards: a Report Prepared at the Request of the Secretary-General of the United Nations*. United Nations. URL http://books.google.com.br/books?id=q-PpjwEACAAJ. Citado na pág. 11

**Vaccaro e Liepins (1989)** H. S. Vaccaro e G.E. Liepins. Detection of anomalous computer session activity. Em *Proceedings of IEEE Symposium on Security and Privacy*, páginas 280–289. doi: 10.1109/SECPRI.1989.36302. Citado na pág. 7

**Valdes e Skinner (2001)** Alfonso Valdes e Keith Skinner. Probabilistic alert correlation. Em *Recent Advances in Intrusion Detection (RAID 2001)*, number 2212 in Lecture Notes in Computer Science. Springer-Verlag. URL http://www.sdl.sri.com/papers/raid2001-pac/. Citado na pág. 9, 18

**Valeur *et al.* (2004)** F. Valeur, Giovanni Vigna, C. Kruegel e R.A. Kemmerer. Comprehensive approach to intrusion detection alert correlation. *Dependable and Secure Computing, IEEE Transactions on*, 1(3):146–169. ISSN 1545-5971. doi: 10.1109/TDSC.2004.21. Citado na pág. 9, 16

**Vasilomanolakis *et al.* (2015)** Emmanouil Vasilomanolakis, Shankar Karuppayah, Max Mühlhäuser e Mathias Fischer. Taxonomy and survey of collaborative intrusion detection. *ACM Computing Surveys (CSUR)*, 47(4):55:1–55:33. ISSN 0360-0300. doi: 10.1145/2716260. URL http://doi.acm.org/10.1145/2716260. Citado na pág. 10

**Vigna e Kemmerer (1998)** Giovanni Vigna e R.A. Kemmerer. Netstat: a network-based intrusion detection approach. Em *Computer Security Applications Conference, 1998. Proceedings. 14th Annual*, páginas 25–34. doi: 10.1109/CSAC.1998.738566. Citado na pág. 8

**Virvilis e Gritzalis (2013)** N. Virvilis e D. Gritzalis. The big four - what we did wrong in advanced persistent threat detection? Em *Availability, Reliability and Security (ARES), 2013 Eighth International Conference on*, páginas 248–254. doi: 10.1109/ARES.2013.32. Citado na pág. 1

**West-Brown *et al.* (2003)** Moira J. West-Brown, Don Stikvoort, Klaus-Peter Kossakowski, Georgia Killcrece, Robin Ruefle e Mark Zajicek. *Handbook for Computer Security Incident Response Teams (CSIRTs)*. Handbook (Carnegie Mellon University. Software Engineering Institute). Carnegie Mellon University, Software Engineering Institute, 2 ed. URL http://books.google.com.br/books?id=jogknwEACAAJ. Citado na pág. 20

**White *et al.* (1996)** G.B. White, E.A. Fisch e U.W. Pooch. Cooperating security managers: a peer-based intrusion detection system. *Network, IEEE*, 10(1):20–23. ISSN 0890-8044. doi: 10.1109/65.484228. Citado na pág. 8

**Winkler e Page (1989)** J. R. Winkler e W. J. Page. Intrusion and anomaly detection in trusted systems. Em *Proceedings of Fifth Annual Computer Security Applications Conference*, páginas 39–45. doi: 10.1109/CSAC.1989.81023. Citado na pág. 7

**Wood e Erlinger (2007)** M. Wood e M. Erlinger. Intrusion Detection Message Exchange Requirements. RFC 4766 (Informational), Março 2007. URL http://www.ietf.org/rfc/rfc4766.txt. Citado na pág. 17

**Xu (2006)** Dingbang Xu. *Correlation Analysis of Intrusion Alerts*. Tese de Doutorado, North Carolina State University. Citado na pág. 24

**Xu e Ning (2008)** Dingbang Xu e Peng Ning. Correlation analysis of intrusion alerts. Em *Intrusion Detection Systems*, volume 38 of *Advances in Information Security*, páginas 65–92. Springer US. ISBN 978-0-387-77265-3. doi: 10.1007/978-0-387-77265-3_4. URL http://dx.doi.org/10.1007/978-0-387-77265-3_4. Citado na pág. 18, 23

**Ye *et al.* (2003)** Nong Ye, Sean Vilbert e Qiang Chen. Computer intrusion detection through ewma for autocorrelated and uncorrelated data. *IEEE Transactions on Reliability*, 52:75–82. doi: 10.1109/TR.2002.805796. Citado na pág. 9

**Ye e Wu (2010)** Shaozhi Ye e S. Felix Wu. Measuring message propagation and social influence on twitter.com. Em *Proceedings of the Second international conference on Social informatics*, SocInfo'10, páginas 216–231, Berlin, Heidelberg. Springer-Verlag. ISBN 3-642-16566-4, 978-3-642-16566-5. Citado na pág. 67, 83

**Yegneswaran *et al.* (2004)** V. Yegneswaran, P. Barford e S. Jha. Global intrusion detection in the domino overlay system. Em *In Proceedings of Network and Distributed System Security Symposium (NDSS)*. Citado na pág. 33, 40

**Zhai** *et al.* **(2003)** Jian-Qiang Zhai, Jun-Feng Tian, Rui-Zhong Du e Jian-Cai Huang. Network intrusion early warning model based on d-s evidence theory. Em *Machine Learning and Cybernetics, 2003 International Conference on*, volume 4, páginas 1972–1977 Vol.4. doi: 10.1109/ICMLC.2003.1259825. Citado na pág. 9

**Zheng e Li (2010)** N. Zheng e Q. Li. A recommender system based on tag and time information for social tagging systems. Em *Expert Systems with Applications*, páginas 4575–4587. Citado na pág. 107, 108

**Zhou** *et al.* **(2007)** C.V. Zhou, S. Karunasekera e C. Leckie. Evaluation of a decentralized architecture for large scale collaborative intrusion detection. Em *Integrated Network Management, 2007. IM '07. 10th IFIP/IEEE International Symposium on*, páginas 80–89. doi: 10.1109/INM.2007.374772. Citado na pág. 10

**Zomlot** *et al.* **(2011)** Loai Zomlot, Sathya Chandran Sundaramurthy, Kui Luo, Xinming Ou e S. Raj Rajagopalan. Prioritizing intrusion analysis using dempster-shafer theory. Em *Proceedings of the 4th ACM Workshop on Security and Artificial Intelligence*, AISec '11, páginas 59–70, New York, NY, USA. ACM. ISBN 978-1-4503-1003-1. doi: 10.1145/2046684.2046694. URL http://doi.acm.org/10.1145/2046684.2046694. Citado na pág. 20

**Zou** *et al.* **(2005)** Cliff C. Zou, Weibo Gong, Don Towsley e Lixin Gao. The monitoring and early detection of internet worms. *IEEE/ACM Trans. Netw.*, 13(5):961–974. ISSN 1063-6692. doi: 10.1109/TNET.2005.857113. URL http://dx.doi.org/10.1109/TNET.2005.857113. Citado na pág. 20, 36