Segurança da informação na rede educacional do IFF

André de Azevedo Cunha
Especialista em Criptografia e Segurança de Redes pela Universidade Federal Fluminense (UFF) – Niterói, RJ – Brasil. Gerente de TI da Acesso Total Comércio, Internet e Serviços Ltda. – Brasil.
http://lattes.cnpq.br/1264148750096515
E-mail: aacunha@gmail.com

Simara Netto Martins
Licenciatura em andamento em Matemática pela Universidade Estadual do Norte Fluminense Darcy Ribeiro (UENF) - Campos dos Goytacazes, RJ - Brasil, Brasil.
http://lattes.cnpq.br/4522744124795699
E-mail: simaraiff@gmail.com

Georgia Regina Rodrigues Gomes
Doutora em Informática pela Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio) - Rio de Janeiro, RJ – Brasil. Professora da Universidade Federal Fluminense (UFF) - Icaraí, Niterói - RJ.
http://lattes.cnpq.br/8966061799453364
E-mail: georgia@ucam-campos.br

Submetido em: 03/06/2013. Aprovao em: 08/07/2016. Publicado em: 26/06/2017..

RESUMO

Com o avanço e disseminação dos serviços on-line, as redes de computadores ficaram vulneráveis. Nos dias de hoje, praticamente todos os computadores e dispositivos móveis estão conectados à grande rede mundial, a internet, deixando o maior patrimônio das corporações, a informação, sem a devida proteção, seja por descuido do próprio usuário, deixando sua máquina vulnerável, ou por falha de segurança na rede, com brechas de segurança já conhecidas. Analisar toda essa informação sem uma ferramenta de apoio é inviável, e a aprendizagem de máquina, que provê a base da mineração de dados, é uma das técnicas mais eficientes para este fim. O objetivo deste trabalho é aplicar técnicas de mineração de dados nos logs do tráfego de rede gravados no banco de dados de uma instituição de ensino pública, para identificar tráfegos maliciosos que possam deixar a informação confidencial de pesquisas e do histórico dos discentes vulnerável ou a rede congestionada. Com a extração do conhecimento do tráfego não desejado, é possível bloqueá-lo para aumentar a segurança de toda a rede e otimizá-la para seu devido fim, ou seja, como plataforma de informação para a comunidade. O software utilizado para minerar os dados foi o Weka, no qual foi possível visualizar através do algoritmo a priori, por exemplo, quando o protocolo é UDP e o país de origem o Brasil, a categoria do alerta é trojan-activity. Já com o uso do algoritmo ZeroR, com o atributo IP de origem como entrada, o host 10.12.9.135 teve 22,19% dos alertas.

Palavras-chave: Redes de computadores. Segurança. Mineração de dados. Weka.

Information security in IFF's educational network

ABSTRACT

With the advancement and dissemination of online services, computer networks were vulnerable. Today, almost all computers and mobile devices are connected to world wide web, the internet, leaving the greatest asset of corporations, information, without proper protection. Either by the user's own carelessness, leaving her vulnerable machine, or for safety network failure, with known security holes. Analyze all this information without a support tool is not feasible, and machine learning, which provides the basis of data mining is one of the most efficient techniques for this purpose. The objective of this paper is to apply the network traffic data in the logs mining techniques recorded in the database of an institution of public education, to identify malicious traffic that may leave confidential research information and history of vulnerable students or congested network. With the extraction of knowledge of unwanted traffic can block them to increase the security of the entire network and optimize it for its intended purpose, as information platform for the community. The software used to mine the data was WEKA, which was visible through the a priori algorithm, for example, that when the protocol is UDP and the country of origin Brazil, alert category is trojan-activity. Already using the Zeror algorithm, with the source IP attribute as input, the host 10.12.9.135 had 22.19% of alerts.

Keywords*: Computer networks. Security. Data Mining. Weka.*

Seguridad de la información en la red educacional del IFF

RESUMEN

With the advancement and dissemination of online services, computer networks were vulnerable. Today, almost all computers and mobile devices are connected to world wide web, the internet, leaving the greatest asset of corporations, information, without proper protection. Either by the user's own carelessness, leaving her vulnerable machine, or for safety network failure, with known security holes. Analyze all this information without a support tool is not feasible, and machine learning, which provides the basis of data mining is one of the most efficient techniques for this purpose. The objective of this paper is to apply the network traffic data in the logs mining techniques recorded in the database of an institution of public education, to identify malicious traffic that may leave confidential research information and history of vulnerable students or congested network. With the extraction of knowledge of unwanted traffic can block them to increase the security of the entire network and optimize it for its intended purpose, as information platform for the community. The software used to mine the data was WEKA, which was visible through the a priori algorithm, for example, that when the protocol is UDP and the country of origin Brazil, alert category is trojan-activity. Already using the Zeror algorithm, with the source IP attribute as input, the host 10.12.9.135 had 22.19% of alerts.

Palabras clave*: Redes de ordenadores. Seguridad. Extracción de datos. Weka.*

INTRODUÇÃO

Atualmente, a quase a totalidade das atividades, procedimentos e operações de empresas e instituições, sejam elas públicas ou privadas é feita eletronicamente, ou ao menos o histórico dessas atividades é posteriormente digitalizado. Segundo Capgemini (2014), em seu relatório de pagamentos mundiais, o volume de transações bancárias feitas por meios eletrônicos cresceu 9,4% no primeiro semestre de 2013, chegando a 366 bilhões de operações no período. Já segundo a CAIXA (2014), o uso de smartphones em transações online deve, a partir de 2015, ultrapassar o uso de caixas tradicionais.

Conforme Oliveira (2011), a informação é essencial para a sobrevivência das organizações, devendo ser preservada com o intuito de não ser alterada, evitando que pessoas mal intencionadas obtenham acesso a ela. Já Lopes (2012) ressalta que a informação assumiu um valor vital para as organizações, que até pouco tempo atrás tinham o foco basicamente para os bens tangíveis, e hoje em dia enxergam a informação como principal ativo. Segundo a ABNT (2005), como esse volume de informação é cada vez maior, ela se tornou um ativo muito mais valioso que qualquer bem material, imprescindível para o sucesso do negócio em questão.

Além das instituições, devido à convergência de computação e da comunicação, produziu-se uma sociedade que consome mais informação. Assim, o volume de informação mundial hoje é gigantesco, sendo a sua maioria armazenada em bancos de dados. Como destaca Frank (2005), nos bancos de dados espalhados ao redor do mundo existem muitos dados potencialmente importantes, porém ainda desconhecidos ou não relacionados; no entanto, muitos outros padrões presentes nessas bases são banais e irrelevantes.

Além do mais, há grande apreensão por parte de vários profissionais e gestores em compreender os dados e em utilizar a informação e conhecimento das bases de dados (COSTA, 2004).

Isso provavelmente acontece em decorrência do forte ritmo de geração de dados, o que é devido à incapacidade natural do ser humano de explorar, extrair e interpretar esses dados nesse ritmo para obter conhecimento dessas bases.

Assim, a informática e as tecnologias voltadas para coleta, armazenamento e disponibilização de dados vêm evoluindo muito e disponibilizando técnicas, métodos e ferramentas computacionais automáticas capazes de auxiliar na extração e interpretação de informações úteis contidas em grandes volumes de dados complexos (CARDOSO, 2006; QUONIAM, 2001).

Uma dessas tecnologias é a mineração de dados, utilizada para extrair informações úteis dos dados armazenados em bancos de dados. Informação essa que é expressa de uma maneira compreensível e pode ser usada para vários propósitos (FRANK, 2005).

A ciência da informação é a área que concentra os estudos sobre análise, coleta, classificação, manipulação, armazenamento, recuperação e disseminação da informação (MERRIAM-WEBSTER, 2015). Especificamente na área de segurança de redes de computadores, a análise dos logs de acesso e interpretação dos mesmos, para que seja possível bloquear qualquer tentativa de acesso não autorizado ou tráfego malicioso, é um processo praticamente inviável de ser efetuado sem o auxílio de uma ferramenta de apoio, pois o volume dos dados é muito grande. Ao mesmo passo, diariamente são descobertas novas de brechas de segurança e são criadas ferramentas que possam explorá-las (malwares) (ULBRICH, 2003).

Segundo ULBRICH (2003), muitos bugs (brechas de segurança) que permitem a ação de criminosos poderiam ser facilmente corrigidos, mas as companhias preferem ignorar esses problemas. Ainda conforme o autor, uma pesquisa realizada pela Módulo Security Solutions, empresa especialista em segurança, revelou, pelos dados coletados, que a segurança da informação é fator importante para 45% dos executivos, sendo que 16% a consideram crítica e 32% a classificam

como vital. Mesmo assim, a falta de conscientização dos executivos (45%) e dos usuários (38%) foi apontada como um dos principais obstáculos para a implementação da segurança nas instituições.

Outros dados indicados pela pesquisa que são extremamente preocupantes: 43% das empresas reconheceram ter sofrido ataques nos últimos 12 meses, sendo que 24% dessas ocorrências foram registradas nos últimos seis meses. Porém o mais crítico é que 32% não souberam informar se foram atacadas ou não e, apesar da expectativa de aumento nos problemas de segurança e nos índices de registros de ataques e invasões, a pesquisa mostra que apenas metade das empresas brasileiras (49%) possui planos de ação formalizados em caso de ataques.

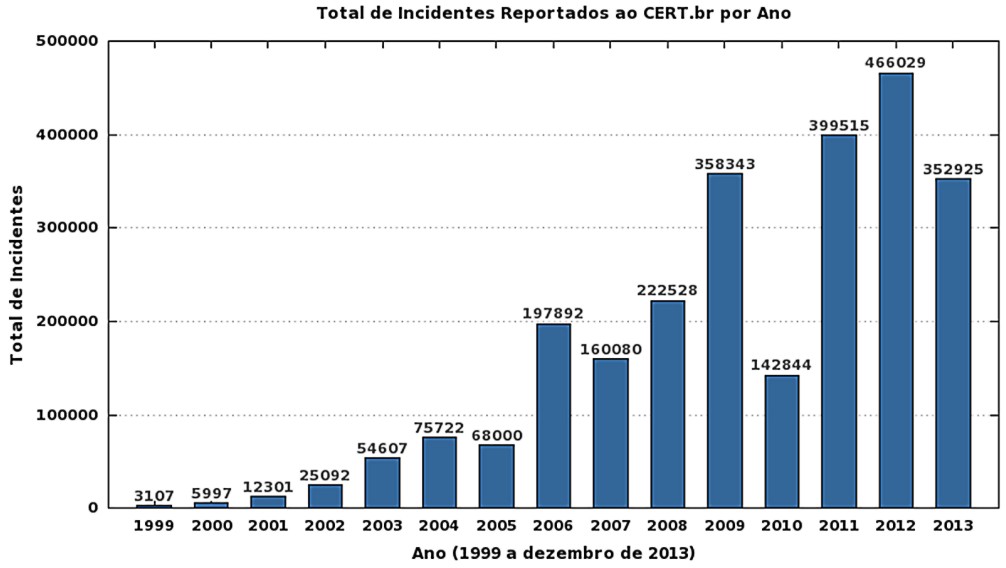
A percepção de falta de segurança continua sendo o maior obstáculo para o desenvolvimento de negócios digitais em escala global. Por exemplo, ao menos 66% dos usuários relatam deixar de realizar transações online por não se sentirem seguros para realizar movimentações financeiras em meios eletrônicos (ULBRICH, 2003).

Este trabalho tem como objetivo utilizar as técnicas de mineração de dados sobre as informações do tráfego de redes do Instituto Federal Fluminense, tanto o tráfego interno entre seus câmpus quanto o tráfego envolvendo a rede mundial de computadores, para extrair regras que apresentem as principais vulnerabilidades e tráfegos considerados maliciosos ainda não identificados pela equipe de tecnologia da informação do instituto. Com base nessas regras, será proposta a devida proteção da informação do instituto.

SEGURANÇA DA INFORMAÇÃO

Existem muitos fatores que contribuem para a grande quantidade de ataques virtuais. Por exemplo, existem muitos sites inseguros. Um estudo da Gartner Group estima que 2/3 dos servidores Web no mundo estão vulneráveis e podem ser invadidos de alguma maneira. Outro ponto que estimula esses ataques é o extenso repositório de ferramentas para ataque disponível na internet. Qualquer pessoa com tempo livre e mesmo com conhecimento técnico médio consegue encontrar informações e os softwares necessários para uma invasão. Mas o principal motivo ainda é a impunidade. Falta uma legislação específica e existem poucos policiais peritos que investigam crimes digitais no Brasil (ULBRICH, 2003).

Figura 1 – Total de incidentes reportados ao CERT.br por ano.

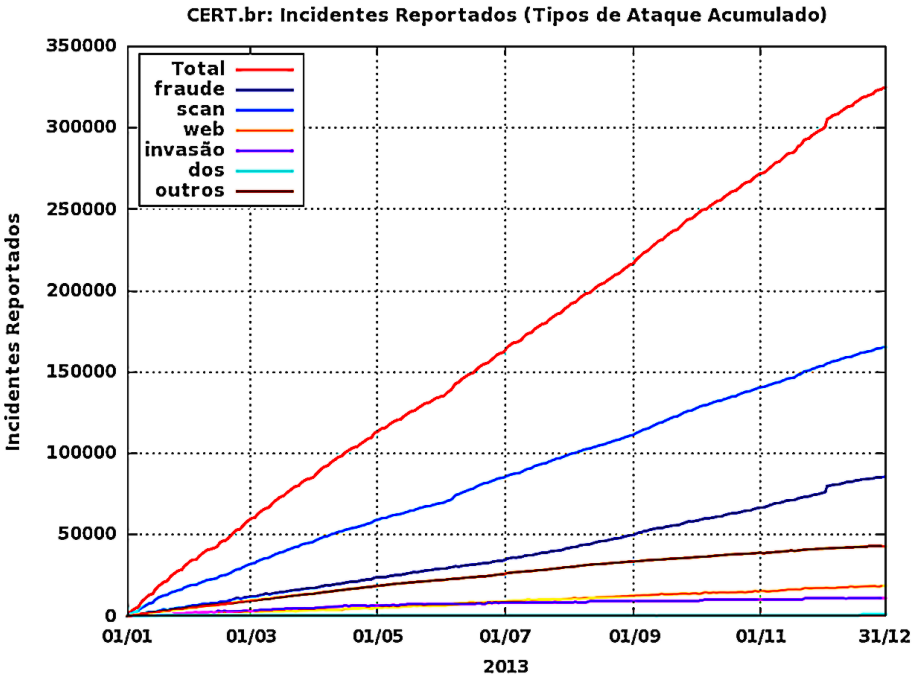


Fonte CERT.br

Baseado nisso, o número de ataques vem crescendo nos últimos anos. De acordo com a CERT.br - Centro de Estudos, Resposta e Tratamento de Incidentes no Brasil (2014), em 2013 houve 352.925 incidentes reportados a esse órgão, que trata os incidentes na Internet no Brasil, ante apenas 3.107 incidentes reportados em 1999. A figura 1 apresenta o gráfico com o total dos incidentes/ano. Incidentes não necessariamente são considerados invasões de fato, pois podem ocorrer falsos positivos na detecção desse tipo de ocorrência. Apenas após uma análise detalhada dos incidentes, eles podem ser confirmados, quando passam a ser considerados invasões.

Para se ter uma ideia da quantidade de ações de hackers, um estudo da Universidade da Califórnia mostrou que eles tentam realizar mais de 4 mil ataques do tipo DoS (Denial of Service) todas as semanas, número bastante elevado e que mostra que é preciso pensar em proteção, quando se está conectado à Web (ULBRICH, 2003). Já um estudo realizado pela CERT.br indica alto número de incidentes. Durante o ano de 2013, foram reportados ao órgão mais de 300.000 dessas ações, estando dentre os tipos de ataques mais comuns a fraude e a invasão de sistemas, conforme apresentado na figura 2.

Figura 2 – Incidentes reportados ao CERT.br.

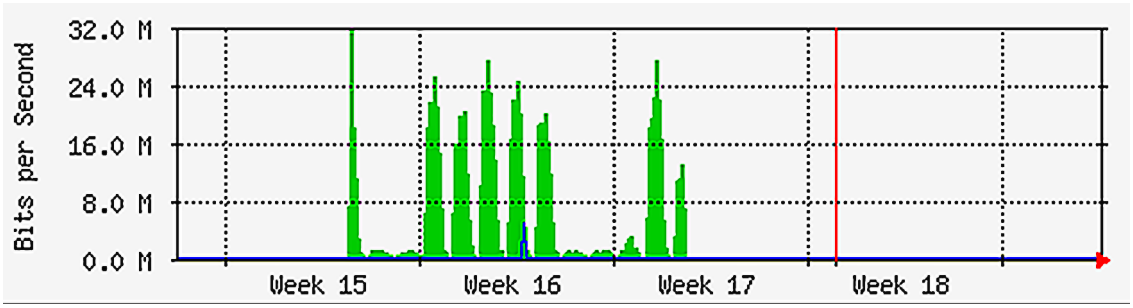


Fonte: CERT.br

Diante de tantas ameaças, é dever da equipe de tecnologia da informação (T.I.) e dos administradores de rede identificar e proteger a rede e seus sistemas, mesmo com o crescente volume de dados. A figura 3 ilustra a média de tráfego de rede do Instituto Federal Fluminense (IFF) durante uma semana (Week 16), na qual o volume de informações trafegadas ficou em torno de 20MBit/s.

Extrair conhecimento útil dessa espessa massa de dados, apontando relações ocultas, padrões e gerando regras para correlacionar dados, é uma alternativa estratégica para as instituições, que pode ajudar na tomada rápida de decisão ou modificar o planejamento da instituição em longo prazo, com maior confiança do que uma decisão tomada apenas pelo sentimento do gestor, o que é possível através das técnicas de mineração de dados (CARDOSO, 2008).

Figura 3 – Média de tráfego diário

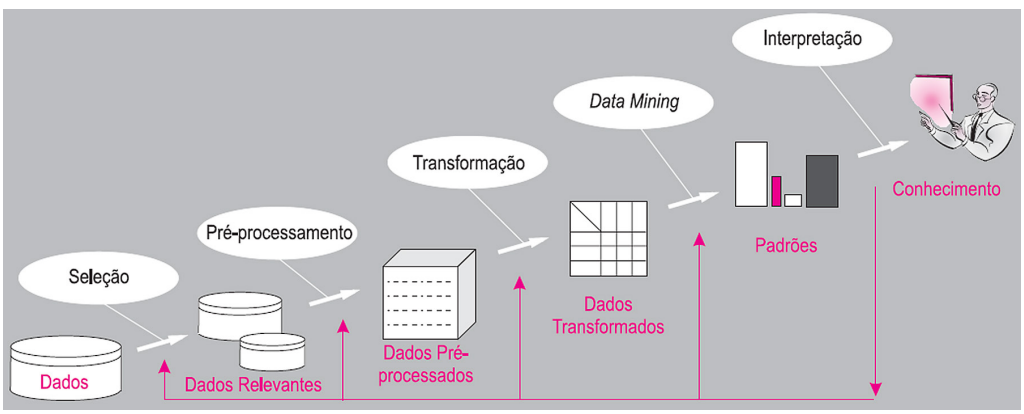


Informação e conhecimento são primordiais, estratégicas e imprescindíveis na vida de qualquer instituição, seja qual for sua área de atuação, pois o mercado exige cada vez mais decisões rápidas e bem embasadas, o que é vital para o sucesso das organizações. Por isso, diversas instituições nacionais e internacionais das mais diversas áreas de atuação já adotaram, em sua rotina, a mineração de dados para monitorar suas atividades, como proteção de redes de computadores (MORAIS, 2010); acessos a servidores Web (CORREIA, 2004); arrecadações; consumo e perfil dos clientes; prevenção de fraudes e riscos do mercado (ARAUJO, 2006), além de evitar evasão escolar (MARTINS, 2011), dentre outras.

KDD

Segundo Martins (2011), o KDD é um processo que busca identificar padrões, associações, modelos ou informações relevantes que estão ocultos em bases, repositórios, dentre outras formas de armazenamento de dados. Tal processo permite identificar padrões válidos, novos e com alta probabilidade de serem úteis e compreensíveis, envolvendo diversas áreas da ciência, como aprendizagem de máquina, banco de dados, estatística, reconhecimento de padrões, dentre outras (MALUCELLI, 2010). Para melhor entendimento, apresenta-se o processo de KDD na figura 4, que se subdivide em 5 etapas: seleção dos dados, pré-processamento, transformação dos dados, mineração dos dados e interpretação do conhecimento extraído.

Figira 4 – Processo do KDD



Fonte: Fayaad, 1996

SELEÇÃO DOS DADOS

De acordo com Dunham (2003), o dado necessário para o processo de mineração de dados pode ser obtido de muitas diferentes e heterogêneas origens. O primeiro passo é obter o dado de várias bases de dados, arquivos e origens não eletrônicas.

PRÉ-PROCESSAMENTO

O dado a ser usado no processo pode ter informações incorretas ou faltantes, podendo ser proveniente de múltiplas fontes anômalas, envolvendo diferentes tipos de dados e métricas. Assim, pode ser necessária a execução de diferentes atividades neste momento. Dados incorretos podem ser corrigidos ou removidos, enquanto os dados faltantes devem ser fornecidos ou estimados (muitas vezes com o uso de ferramentas de mineração de dados) (DUNHAM, 2003).

TRANSFORMAÇÃO

Ainda de acordo com Dunham (2003), dados de diferentes origens devem ser convertidos em um formato comum antes serem processados. Alguns dados pode ser encodados ou transformados em um formato mais usual. A redução dos dados deve ser considerada para diminuir o número de possíveis valores que o dado em questão pode assumir.

MINERAÇÃO DE DADOS (DATA MINING)

O termo Mineração de Dados (DM) surge, inicialmente, como um sinônimo de KDD, mas é apenas uma das etapas da descoberta de conhecimento em bases de dados, do processo global chamado KDD (QUONIAM, 2001). O conhecimento que pode ser obtido através da mineração de dados tem se mostrado bastante útil em diversas áreas de pesquisa e de atuação, como informática, medicina, ensino público e privado, finanças, comércio, marketing, telecomunicações, meteorologia, agropecuária, bioinformáticas, entre outras (JONES, 2000).

A mineração de dados não é um processo simples e trivial, consiste em identificar nos dados os padrões válidos, novos, potencialmente úteis e compreensíveis, envolvendo métodos estatísticos, ferramentas de visualização e técnicas de inteligência artificial (FAYYAD, 1996). Portanto, o processo do KDD utiliza conceitos de base de dados, estatística, ferramentas de visualização e técnicas de inteligência artificial, dividindo-se nas etapas de seleção, pré-processamento, transformação, mineração e interpretação/avaliação dos resultados (FAYYAD, 1996).

Dentre essas etapas, a mais relevante é a mineração de dados, foco de vários estudos em diversas áreas de conhecimento (WICKERT, 2007), que comprovam que a extração da informação a partir de um conjunto de dados, e posteriormente a transformação dessa informação em conhecimento, é imprescindível para o processo de tomada de decisão.

Na mineração de dados, o dado é armazenado eletronicamente e a busca é automatizada, porém este processo não é relativamente novo. Economistas, estatísticos, meteorologistas e engenheiros trabalham com a ideia que padrões nos dados podem ser procurados automaticamente, identificados e validados, sendo usados para previsões (FRANK, 2005).

Com isso, o processo de mineração de dados consiste em extrair dados úteis já presentes em bancos de dados. Por exemplo, uma base de dados das compras dos clientes de algum estabelecimento, juntamente com o perfil desses clientes, é a chave de um problema. Correlacionar essas informações analisando o comportamento desse público traz uma vantagem estratégica, na qual a empresa pode oferecer determinado produto somente a um grupo específico de clientes que realmente se interessam por aquela oferta (FRANK, 2005).

Para tanto, a mineração de dados ou data mining (DM) possui várias etapas: a definição do problema; a seleção de todos os dados e a posterior preparação deles, o que inclui o pré-processamento e a reformatação dos dados (retirada de caracteres e linhas em branco, por exemplo), além de análise dos resultados obtidos do processo de DM (CARDOSO, 2008). A descoberta

do conhecimento deve apresentar as seguintes características: ser eficiente, genérica (possibilidade de ser aplicável a vários tipos de dados) e flexível (facilmente modificável) (STEINER, 2006). O processo de desenvolvimento de DM envolve vários procedimentos, métodos e algoritmos que possibilitam a extração de novos conhecimentos (CARDOSO, 2008). Entre as tarefas de DM, é possível destacar algumas mais utilizadas: associação, classificação, regressão, clusterização e sumarização (GOLDSHMIDT, 2005).

INTERPRETAÇÃO DO CONHECIMENTO

Como os resultados da mineração de dados são apresentados para os usuários, esse é um aspecto básico, porque a utilidade dos resultados é completamente dependente disso. Várias formas de visualização e interfaces estão disponíveis e são usadas nessa etapa (DUNHAM, 2003).

FERRAMENTAS

Uma das aplicações mais utilizadas no processo de mineração de dados é o Weka. O Weka foi desenvolvido na Universidade de Waikato, na Nova Zelândia, e a abreviatura Weka significa Waikato Environment for Knowledge Analysis. Além da ferramenta, Weka, também é uma ave que não voa, com uma natureza inquisitiva, encontrada apenas nas ilhas da Nova Zelândia.

Segundo Frank (2005), o projeto Weka é baseado em uma coleção organizada de algoritmos de aprendizado de máquina e ferramentas de pré-processamento de dados. A ferramenta provê suporte extensivo para todo o processo de mineração de dados, incluindo a preparação dos dados, avaliação estatística dos sistemas de aprendizagem e visualização gráfica dos dados de entrada e do resultado da aprendizagem. A ferramenta possui grande variedade de algoritmos de aprendizagem, incluindo ampla variedade de ferramentas de pré-processamento. Este kit diversificado e abrangente é acessado através de uma interface gráfica, para que seus usuários possam comparar diferentes métodos e identificar aqueles que são mais adequadas para o problema em questão.

METODOLOGIA

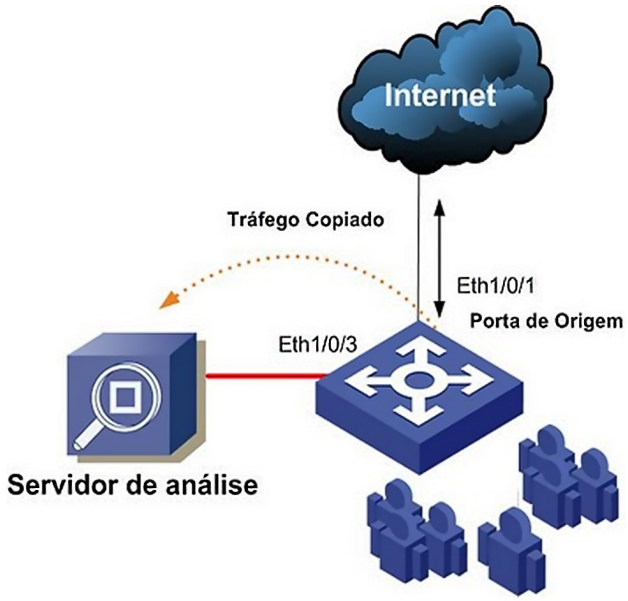
Para que seja possível identificar ameaças em potencial no tráfego de rede do instituto, bem como perceber as vulnerabilidades nos serviços hospedados internamente, foi instalado um sistema de detecção de intrusão (IDS) para logar todo o tráfego de rede de dentro do instituto para a internet e vice-e-versa, e também o tráfego de rede entre os câmpus. Essa ferramenta permite que os logs gerados sejam gravados diretamente em um banco de dados, possibilitando a extração desses dados para utilização posterior no processo de mineração de dados.

Os dados foram extraídos a partir de logs gerados pelo IDS instalado no Instituto Federal Fluminense, localizado em Campos dos Goytacazes, RJ, no período de dezembro de 2011 a novembro de 2012.

Para que a coleta dos dados fosse possível, utilizou-se um servidor de análise, para o qual foi direcionado todo o tráfego de rede do instituto. Nesse servidor foi instalado o sistema operacional Linux Debian 6.0.5 x64 (64 bits), bem como as seguintes ferramentas: Um sistema de detecção de intrusos ou intrusion prevention system (IDS) – Snort 2.9.3, Baynard 2.1.10 (faz a interface do software IDS com o banco de dados), Banco de dados Mysql 14.04 e o Basic Analysis and Security Engine – BASE 1.4.5 (interface web para o IDS).

O direcionamento dos dados a serem monitorados foi estabelecido do seguinte modo: foi configurado em um Switch 3com 4500 a função de espelhamento de porta (port mirroring). Este switch interliga o roteador do provedor de acesso à internet ao firewall da instituição, portanto toda informação a ser monitorada passa pelo equipamento em questão, conforme apresentado na figura 5.

Figura 5 − Estrutura do espelhamento de tráfego



No mesmo *switch* foi configurada outra *Vlan*, para conexão dos hosts internos da rede. Essas portas também foram espelhadas, e assim foi possível avaliar tanto o tráfego interno da rede quanto o externo (com destino à internet). Neste artigo foram analisados os dados capturados dentro da rede.

O IDS instalado no servidor de análise de *logs* inspeciona todo o tráfego, e baseado nas assinaturas das ameaças conhecidas registra no banco de dados toda comunicação suspeita. Para enriquecer a informação disponível, foi desenvolvido um *software* na linguagem PHP, o qual, de posse do endereço IP de origem e destino das comunicações armazenadas no banco de dados, efetua uma consulta *whois*, e armazena no banco de dados o país e o domínio ao qual esses IPs pertencem.

Com os todos os dados armazenados em banco de dados, foi possível a extração deles para a posterior análise no *software* de mineração de dados *Weka*. A seguir, são descritas as etapas da técnica de descoberta de conhecimento em banco de dados, desde a definição do domínio até a extração do conhecimento.

ETAPA DE DEFINIÇÃO E COMPREENSÃO DO DOMÍNIO

Primeiramente foram definidos os dados que seriam relevantes. Analisando a estrutura do banco de dados gerado pelo IDS Snort, além das informações disponibilizadas pelo sistema de consulta *whois*, notou-se que seriam necessários os dados do horário do alerta, qual o alerta e sua categoria, protocolo e portas envolvidas na comunicação, além das portas de origem e destino e os domínios e países envolvidos na comunicação.

ETAPA DE SELEÇÃO DOS DADOS

Após a definição de quais dados seriam relevantes, foi possível a extração deles para a posterior análise no software de mineração de dados Weka.

A extração dos dados foi obtida através de consulta SQL sobre o banco de dados. Após a extração dos dados, eles foram exportados para uma planilha eletrônica.

ETAPA DE LIMPEZA, PREPARAÇÃO E SELEÇÃO DE ATRIBUTOS

Para tratamento e limpeza dos dados, etapa na qual os dados inexistentes foram substituídos por '?', e foram feitos outros ajustes, como por exemplo, o ajuste de casas decimais; em seguida foi gerado um arquivo de texto separado por vírgulas (CSV) para ser tratado, e assim ser possível gerar o arquivo .arff padrão do Weka (FRANK, 2005).

Após a carga dos dados no software de mineração de dados Weka, foram analisados os alertas armazenados na etapa chamada pré-processamento. Nessa etapa foram realizados os ajustes necessários antes da mineração de dados. Dados numéricos foram discretizados, ou seja, os dados relacionados foram agrupados em classes para melhor aproveitamento dos dados, por exemplo, o atributo hora; e atributos redundantes foram eliminados, por exemplo, os atributos mês, ano, minuto e segundo.

ETAPAS DE MINERAÇÃO DOS DADOS E EXTRAÇÃO DO CONHECIMENTO

Na etapa de mineração de dados, utilizando o software Weka, foram utilizadas as tarefas de classificação e associação, sendo selecionados os algoritmos Apriori, ZeroR, J48 e DecisionTable. Os resultados gerados por tais algoritmos e a posterior análise dos resultados são descritos no próximo capítulo.

RESULTADOS E DISCUSSÕES

Ao relacionar o horário do tráfego com a categoria do ataque, tornou-se possível identificar que os alertas de tentativa de acessar *hosts* que distribuem *trojan* acontecem até as 14h, e após esse horário os alertas que predominam são os da categoria de violação de política de acesso, conforme apresentado na figura 6.

Figura 6 – Assinaturas por hora – divididos por categoria do ataque

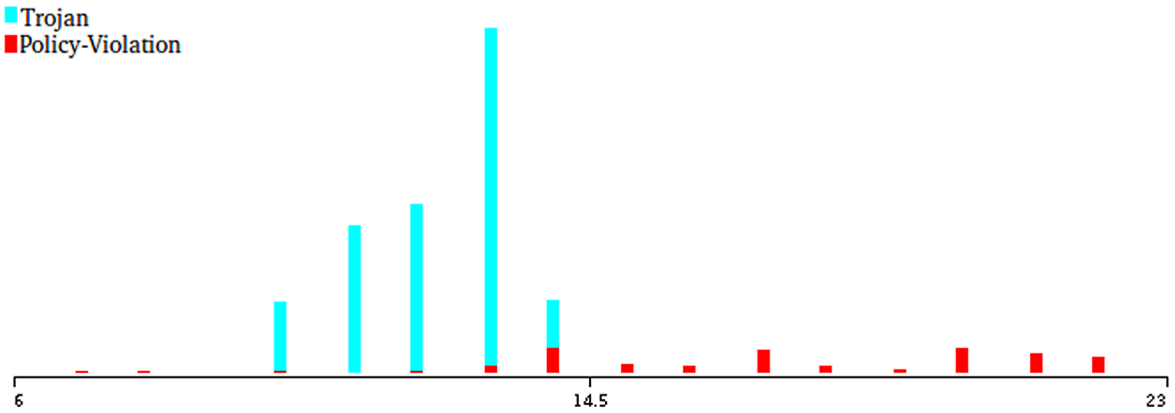
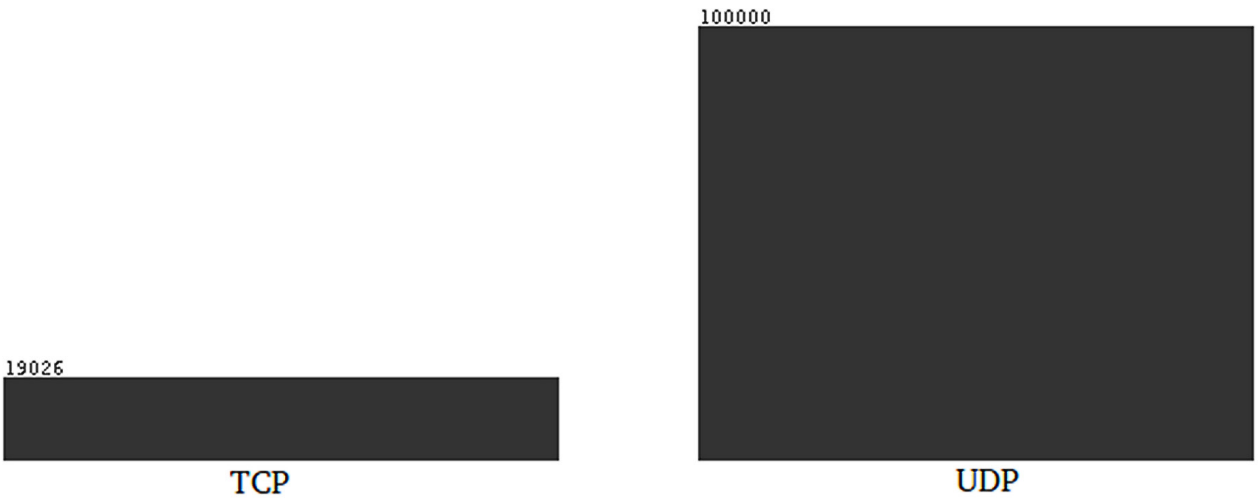


Figura 7 – Assinaturas por protocolo



Fonte: Próprio autor.

Também foram analisados os alertas quanto ao protocolo da comunicação. Apenas dois protocolos geraram alertas no tráfego interno. O protocolo UDP sob o código 17 foi o protocolo com a maioria dos alertas (100.000 alertas). Já o protocolo TCP, código 6 acumulou 19.026 alertas, conforme apresentado na figura 7.

Na etapa de mineração de dados propriamente dita, foram selecionados os classificadores Apriori, um classificador que gera regras de associação e o DecisionTable, que regras de classificação.

As regras formuladas pelo algoritmo Apriori foram expressas no seguinte formato se —> então, e no quadro 1, são apresentadas as regras geradas, que ao todo foram 10. Todas as regras tiveram confiança igual a 1, ou seja, 100%.

Quadro 1– Resultado ao algoritmo Apriori

Se	Então
protocol=UDP	sig_class_name=trojan-activity
sig_class_name=trojan-activity	protocol=UDP
sig_name=SPYWARE-DNS DNS lookup msnsolution.nicaze.net (Malware_Distribution)	sig_class_name=trojan-activity
sig_name=SPYWARE-DNS DNS lookup msnsolution.nicaze.net (Malware_Distribution)	protocol=UDP
sig_name=SPYWARE-DNS DNS lookup msnsolution.nicaze.net (Malware_Distribution) e protocol=UDP	sig_class_name=trojan-activity
sig_name=SPYWARE-DNS DNS lookup msnsolution.nicaze.net (Malware_Distribution) e sig_class_name=trojan-activity	protocol=UDP
sig_name=SPYWARE-DNS DNS lookup msnsolution.nicaze.net (Malware_Distribution)	sig_class_name=trojan-activity e protocol=UDP
owner_dst=iff.edu.br	country_dst=BR
protocol=UDP e country_dst=BR	sig_class_name=trojan-activity
sig_class_name=trojan-activity e country_dst=BR	protocol=UDP

Quadro 2 – Resultado ao algoritmo *DecisionTable* – ip de origem

	TP_Rate	FP_Rate	Precision	Recall	F-Measure	ROC_Area	Class
	0,419	0	0,633	0,419	0,504	0,85	200.143.198.66
	0,374	0	1	0,374	0,544	0,90	200.143.198.49
	1	0,002	0,987	1	0,993	1,00	200.143.198.34
	0,818	0	0,955	0,818	0,881	0,98	200.143.198.110
	1	0	1	1	1	1,00	200.143.198.46
	0,995	0,001	0,996	0,995	0,995	1,00	10.12.9.135
	0,994	0,001	0,984	0,994	0,989	1,00	10.12.8.4
	0,997	0,002	0,993	0,997	0,995	1,00	10.12.8.1
	0,988	0	0,997	0,988	0,992	1,00	10.12.8.5
Weighted_AVG	0,993	0,001	0,993	0,993	0,992	1,00	

Ao observar as regras é possível relacionar que se o protocolo é o UDP, então a categoria dos alertas é o trojan-activity. Bem como se a assinatura do alerta é a consulta dns a uma rede de distribuição de malware, a "DNS lookup msnsolution.nicaze.net", então a categoria é trojan-activity e o protocolo é o UDP. Se o domínio de origem é o "iff.edu.br", o país de destino é o próprio Brasil. Se o protocolo em questão é o UDP e o país de destino é o Brasil, a categoria do ataque é o trojan-activity. Assim como se a categoria do alerta é trojan-activity e o país de destino é o Brasil, significa que o protocolo utilizado na comunicação foi o UDP.

Utilizando o atributo ip_src como valor de entrada para o classificador ZeroR, foi possível identificar que o *host* "10.12.9.135" precisa passar por uma análise detalhada, pois foi identificado como o gerador de 22,19% dos alertas. Já o classificador DecisionTable, com 99,28% das instâncias corretamente classificadas, gerou o quadro 2 com os Ips de origem com maior número de alertas:

Os IPs 200.143.198.34 e 200.143.198.46 tiveram 100% das amostras corretamente classificadas (*recall* =1). Ainda é possível observar que o IP 200.143.198.46 teve a média ponderada entre a precisão e a sensibilidade igual a 100% (F-*Measure* = 1), o que identifica esses endereços como críticos.

CONCLUSÃO

O trabalho apresentado teve como objetivo aplicar mineração de dados em uma base de dados com informações do tráfego de redes do Instituto Federal Fluminense, extraindo conhecimento após a mineração dos *logs* do tráfego de rede, para que seja possível que a equipe de T.I. do instituto bloqueie tráfegos indesejados, quando necessário. Com isso, toda a rede teve seu nível de proteção elevado contra ameaças e ataques, provendo uma melhoria no desempenho geral das comunicações.

Pode-se concluir que a utilização das técnicas mineração de dados para a análise e extração de conhecimento a partir dos *logs* de segurança do tráfego de rede armazenados em banco de dados é de extrema valia para a ciência da informação, com foco especial na segurança da informação, protegendo o ativo principal do instituto, sendo que a aplicação deste trabalho já está sendo realizada com sucesso no Instituto Federal Fluminense.

Já na etapa de pré-processamento, foi possível identificar com maior precisão vários dados, como por exemplo: IPs, domínios, portas de origem e destino, horário dos ataques, protocolos e assinaturas com mais alertas de segurança, comparado com a interface Web de gerência Base do IDS utilizado Snort. Essa interface, devido ao grande volume de dados, tornou-se inutilizável, chegando a não exibir os dados em vários momentos.

Com os resultados dos classificadores na etapa de mineração de dados, pode-se relacionar alguns dados, como horário do alerta com país de origem e tipo do alerta. Esse tipo de informação auxilia o administrador de rede a traçar o perfil das ameaças e proteger a rede bloqueando os tráfegos com tais características no firewall da instituição. Outro conhecimento extraído foi a relação dos tipos de alertas e os domínios envolvidos nas comunicações, o que após a análise do responsável pela rede pode gerar regras de bloqueio para os tráfegos originados por tais domínios. Outro fator importante foi determinar os hosts com mais alertas, e em caso de o host ser uma máquina interna do IFF, ela é encaminhada ao setor de manutenção para análise e correção de suas vulnerabilidades.

REFERÊNCIAS

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS - ABNT. *NBR ISO/IEC 17799:* tecnologia da informação: técnicas de segurança - código de prática para a gestão da segurança da informação. Rio de Janeiro, 2005. 120 p. Disponível em: <http://portal.cjf.jus.br/sigjus/arquivos-diversos/NBR-ISO-IEC-17799-2005.PDF/at_download/file>. Acesso em: 07 mar. 2013.

ARAUJO JÚNIOR R. H.; TARAPANOFF K. Precisão no processo de busca e recuperação da informação: uso da mineração de textos**.** *Ciência da Informação,* v. 35, n. 3, p. 236-247, 2006.

CAIXA ECONÔMICA FEDERAL. *Transações via celular ultrapassarão caixas.* Disponível em: <http://www20.caixa.gov.br/Paginas/NaMidia/Noticia.aspx?inme ID=494>. Acesso em: 27 dez. 2014.

CAPGEMINI. *World payments report 2014.* Disponível em: <https://www.world paymentsreport.com/download>. Acesso em: 05 jan. 2015.

CARDOSO, O. N. P.; MACHADO, R. T. M. Gestão do conhecimento usando data mining: estudo de caso na Universidade Federal de Lavras**.** *Revista de Administração Pública,* v. 42, n. 3, p. 495-528, 2008.

CENTRO DE ESTUDOS, RESPOSTAS E TRATAMENTO DE INCIDENTES DE SEGURANÇA NO BRASIL. *Incidentes reportados ao CERT.br em 2013.* Disponível em: <http://www.cert.br/stats/incidentes/>. Acesso em: 19 dez. 2014.

CORREIA, L. J. *Mineração de dados em arquivos de log gerados por servidores de páginas web.* 2004. 107 f. Monografia (Graduação em Ciência da Computação)– Universidade Regional de Blumenau, Blumenau, 2004.

COSTA, L. D. A. F. Bioinformatics: perspectives for the future. *Genetics and Molecular Research,* v. 3, n. 4, p. 564-74, 2004.

DUHAM, M. H. *Data mining:* introductory and advanced topics. New Jersey: Pearson Education: 2003.

FAYYAD, U. M. et al. *Advances in knowledge discovery and data mining.* Menlo Park: AAAI Press: MIT Press, 1996. 611 p.

FRANK, E.; WITTENM I. H. *Data mining:* practical machine learning tools and techniques. 2nd ed. San Francisco: Morgan Kaufmann Publishers, 2005.

GOLDSCHMIDT, R.; PASSOS, E. *Data mining:* um guia prático, conceitos, técnicas, ferramentas, orientações e aplicações. São Paulo: Elsevier, 2005.

JONES, P. B. C. The commercialization of bioinformatics. *Electronic Journal of Biotechnology,* v. 3, n. 2, p. 33-44, 2000.

LOPES, I. M. *Adopção de políticas de segurança de sistemas de informação na administração pública local em Portugal.* 2012. 437 f. Tese (Doutorado em Engenharia e Gestão de Sistemas de Informação)- Universidade do Minho, Portugal, 2012.

MALUCELLI, A. et al. Classificação de microáreas de risco com uso de mineração de dados. *Revista de Saúde Pública,* v. 44, n. 2, p. 292-300, 2010. Disponível em: <http://www.scielo.br/scielo.php?pid=S0034-89102010000200009&script=sci_abstract &tlng=pt>. Acesso em: 05 jul. 2014.

MARTINS, S. N.; CUNHA, A. A.; GOMES, G. R. R. Aplicação de técnicas de mineração de dados na previsão de evasão escolar em instituição pública. In: SIMPÓSIO DE ENGENHARIA DE PRODUÇÃO, 2011, Bauru. *Anais...* 2011. Bauru, 2011.

MERRIAM-WEBSTER. *Significado de information science.* Disponível em: <http://www.merriam-webster.com/dictionary/information%20science>. Acesso em: 06 jan. 2015.

MORAIS, J. A. S. *Descoberta de conhecimento em logs de tentativas de intrusão:* um estudo de caso em instituições de ensino superior. 2010. 130 f. Dissertação (Mestrado em Engenharia Informática)- Instituto Superior de Engenharia do Porto, Porto, 2010.

OLIVEIRA, A. M. R. de; NOGUEIRA, R. C. S.; LEMES, E. G. Segurança da informação nas empresas: enfocando a engenharia social. In: SEMINÁRIO DE PRODUÇÃO ACADÊMICA DA ANHANGUERA, 2011, Anhanguera. *Anais...* [S.l.: s.n.], 2011.

QUONIAM, L. et al. Inteligência obtida pela aplicação de data mining em base de teses francesas sobre o Brasil. *Ciência da Informação,* v. 30, n. 2, p. 20-38, 2001.

STEINER, M. T. A. et al. Abordagem de um problema médico por meio do processo de KDD com ênfase à análise exploratória dos dados. *Gestão e Produção,* v. 13, n. 2, p. 325-37, 2006.

ULBRICH, H. C.; VALLE, J. D. *Universidade H4ck3r.* 3. ed. São Paulo: Digerati Books, 2003.

WICKERT, E. et al. Nitrogen assimilation in citrus based on CitEST data mining. *Genetics and Molecular Biology,* v. 30, n. 3, p. 810-838, 2007.