Coleta e Caracterização de um Conjunto de Dados de Tráfego Real de Redes de Acesso em Banda Larga*

Martin Andreoni Lopez[1], Renato Souza Silva[2], Igor Drummond Alvarenga[1], Diogo Menezes Ferrazani Mattos[1] e Otto Carlos Muniz Bandeira Duarte[1]

[1]Grupo de Teleinformática e Automação (GTA)
{martin,alvarenga,menezes,otto}@gta.ufrj.br
[2]Laboratório de Redes de Alta Velocidade (RAVEL)
renato@ravel.ufrj.br
Universidade Federal do Rio de Janeiro (UFRJ)
Rio de Janeiro – RJ – Brasil

Resumo. *A segurança do acesso à Internet banda larga reside na implantação de políticas de perímetro e na adoção de listas de controle de acesso. Essas medidas são precárias, pois se baseiam em perfis comuns e pouco atualizados de ameaças aos usuários residenciais. Este artigo analisa e caracteriza o tráfego de usuários residenciais de redes fixas de acesso à Internet em banda larga de uma grande operadora de comunicações, por um período de uma semana, e obtém o perfil dos alarmes gerados por um sistema de detecção de intrusão sobre esse tráfego. Os resultados demonstram que a caracterização proposta permite classificar os fluxos, com uma sensibilidade de* 93% *a alertas, diferenciando os fluxos legítimos dos fluxos que geram alarmes, validando o conjunto de dados coletado, e permite reduzir em* 73% *o tráfego direcionado ao analisador de tráfego, tornando a segurança da rede de acesso mais dinâmica e eficiente.*

Abstract. *Broadband Internet access security lies in the implementation of perimeter policies and in the adoption of access control lists. These measures are precarious because they are based on common and poorly updated profiles, lacking residential users threat information. This article analyzes and characterizes residential user traffic from fixed broadband Internet access networks of a large communications operator, for a period of one week, and obtains the profile of the security alarms generated by an intrusion detection system on this traffic. The results show that the proposed characterization allows classification of the flows, with an alert sensitivity of 93% in the differentiation of the legitimate flows and the alarm generating flows, thus, validating the collected dataset, and allows a 73% reduction for the traffic directed to the traffic analyzer, enabling more dynamic and efficient access network security.*

1. Introdução

O acesso à Internet está presente em 45,5% dos domicílios brasileiros [IBGE 2016]. No entanto, os provedores de infraestrutura de acesso à Internet e os órgãos reguladores encaram a segurança desse serviço de forma precária e com ações paliativas para mitigar possíveis prejuízos. A segurança das redes de acesso à Internet

*Este trabalho foi realizado com recursos do CAPES, CNPq e FAPERJ.

é implementada de forma coletiva, por perímetros de segurança [Bhatt et al. 2014]. Geralmente, não existem medidas individuais por parte dos provedores para garantir a segurança dos elementos das redes residenciais. As restrições instaladas ficam obsoletas conforme o uso da rede evolui, como por exemplo, com a adoção da computação em nuvem e da Internet das coisas que trazem novas ameaças à segurança do usuário residencial [Puthal et al. 2016]. Assim, há a necessidade de analisar sistematizadamente alertas de segurança para, então, atualizar estas restrições de perímetro de segurança da rede.

A definição de novas políticas de segurança depende do conhecimento das ameaças de segurança que estão presentes na rede para garantir maior precisão no combate e na mitigação dos efeitos dos ataques [Heidemann e Papadopoulos 2009]. Uma das maiores dificuldades nas pesquisas relacionadas à análise de tráfego na Internet e à extração de conhecimento de situações de ameaças é a obtenção de bases de dados reais (*datasets*). Apesar de existirem algumas bases de dados disponíveis para pesquisa [CAIDA 2007, D. Kotz e Abyzov 2004, NSL-KDD 2009], essas bases, muitas das vezes, não são atuais ou foram criadas a partir de padrões de ataques e ameaças sintéticas. Outras limitações ao acesso de dados reais recaem sobre questões regulatórias e de sigilo dos dados coletados, já que os dados contem informações sensíveis ou confidenciais.

Este artigo analisa e a caracteriza o tráfego em uma rede de acesso à Internet banda larga, diferenciando o tráfego normal do tráfego que gera alertas de segurança. Emprega-se um conjunto de dados real e anonimizado de uma importante operadora de telecomunicações. O conjunto de dados criado é baseado na captura de 5TB de dados de acesso de 373 usuários residenciais de banda larga na cidade do Rio de Janeiro. O tráfego é tratado e analisado para a criação do conjunto de dados contém tráfego legítimo, ataques e outras ameaças de segurança. O tráfego foi analisado em um sistema de detecção de intrusão (*Intrusion Detection System* - IDS) e, então, resumido em um conjunto de dados de características de fluxos associadas a uma classe de alarme do IDS ou à classe de tráfego legítimo. Por fim, o artigo avalia o desempenho de um classificador de tráfego, baseado no algoritmo de árvore de decisão, para identificar em tempo real os fluxos suspeitos.

Em trabalhos anteriores [Andreoni Lopez et al. 2017, Lobato et al. 2016], foi estudada a adequação de classificadores e da computação por fluxo na identificação de anomalias em rede. Outras propostas criam conjuntos de dados baseados em ataques sintéticos e já desatualizados para o estudo da segurança em redes [NSL-KDD 2009]. Há ainda propostas de criação de potes de mel [Song et al. 2013] e de previsão do comportamento de atacantes baseada em processos estocásticos [Chen et al. 2015]. Este artigo, por sua vez, analisa e caracteriza um conjunto de dados reais constituído de pacotes de dados capturados, no período de uma semana, de usuários residenciais de banda larga fixa de uma importante operadora de telecomunicações. A caracterização provê a diferenciação de dois tipos de tráfegos: um tráfego normal e um tráfego que gera alertas de segurança quando tratados por uma ferramenta de análise tráfego. Os resultados obtidos mostram que, observando somente as estatísticas do fluxo, é possível identificar, com 93% de precisão, fluxos que geram alertas de segurança. A adoção da classificação proposta tem o potencial de reduzir em até 73% do tráfego encaminhado às ferramentas de análise de pacotes, inclusive àquelas que verificam camadas superiores.

O restante do artigo está organizado da seguinte forma. A Seção 2 aborda os trabalhos relacionados. O problema de processamento, caracterização de fluxos, caracte-

rização de alertas e o procedimento de coleta dos dados são apresentados na Seção 3. A Seção 4 explicita o procedimento de análise e apresenta os seus resultados. A Seção 5 conclui o trabalho.

2. Trabalhos Relacionados

Construir um conjunto de dados que permita a análise do perfil de uso fiel da rede é um desafio, pois as redes estão em constantes mudanças e os dados podem ser capturados em diferentes locais da rede. Heidemann e Papadopoulos evidenciam quais são os locais ideais da rede para a captura dos dados, abordam as dificuldades em anonimizar e extrair conhecimento de dados anonimizados e discutem os principais conjuntos de dados disponíveis [Heidemann e Papadopoulos 2009]. Dentre os conjuntos de dados para pesquisa em segurança, o mais utilizado é o KDD [NSL-KDD 2009], cuja caracterização foi realizada por Tavallaee *et al.* [Tavallaee et al. 2009].

Além da localização do ponto de coleta, outros fator importante para evitar a contaminação tendenciosa da base de dados é o tamanho da amostra. Shiravi *et al.* discutem a criação e a análise de uma base de dados real, voltada para a detecção de intrusões [Shiravi et al. 2012]. Os autores também geraram três categorias de bases de dados com tráfego real que se encontram disponíveis para a comunidade acadêmica. A principal desvantagem dessas bases de dados reside no fato de que todo o tráfego anômalo foi coletado a partir de ataques simulados em ambientes controlados. Este artigo, por sua vez, utiliza um IDS para identificar o tráfego malicioso e anômalo. Embora não haja a garantia de completude nessa abordagem, evitam-se os problemas oriundos da inserção de ataques artificiais, como a criação de um conjunto dados muito tendenciosos.

Com o surgimento de novos serviços e diferentes tipos de tráfegos, os ataques se modernizam. Bases de dados antigas como em [NSL-KDD 2009] já não podem ser utilizadas para aferir sistemas de detecção de intrusões mais modernos. A abordagem proposta em [Shiravi et al. 2012] analisa diferentes amostras de serviços TCP (HTTP, SSH, FTP, SMTP, IMAP e POP3) em relação ao número de requisições no tempo e compara as curvas obtidas com distribuições de probabilidade conhecidas. As distribuições similares são então utilizadas para montar perfis de ataques que podem ser posteriormente reproduzidos sinteticamente. O conjunto de dados do presente artigo contempla, entre outros, os mesmos serviços e pode ser utilizado para geração de perfis de ataques atuais.

A utilização de potes de mel (*honeypots*) para a montagem de bases de dados reais é explorada em alguns trabalhos de pesquisa. Chen *et al.* analisam uma base de dados criada a partir de 491 *honeypots* TCP, para estudar as características estocásticas dos ataques [Chen et al. 2015]. Song *et al.* propõem modificar os potes de mel para emular sistemas proativos que, inclusive, visitam páginas maliciosas e se juntam a *botnets* [Song et al. 2013]. Além de aumentar o realismo dos dados coletados, as propostas buscam superar a dificuldade de diferenciação tráfego legítimo e do malicioso, considerando que todo o tráfego coletado nos potes de mel é proveniente de ataques. No entanto não é possível garantir, ou verificar, essa premissa. A utilização de IDS para identificação de tráfego malicioso é utilizada no presente artigo como solução para o mesmo problema. A principal diferença entre os métodos reside na probabilidade de ocorrência de falsos positivos e falsos negativos no conjunto de dados, respectivamente prevalentes em [Song et al. 2013, Chen et al. 2015], e no presente trabalho.

Drapper-Gil *et al.* apresentam a caracterização do tráfego de redes privadas virtuais baseada em características temporais dos fluxos [Draper-Gil et al. 2016]. Para tanto, os autores propõem o uso de 8 características para a classificação dos fluxos em 14 diferentes tipos, incluindo fluxos de VPN e não VPN. Os resultados mostram que o algoritmo de aprendizado de máquina por árvores de decisão apresenta um desempenho um pouco melhor do que o algoritmo de k-vizinhos mais próximos. No entanto, os autores não avaliam se as características usadas para a classificação são as que melhor descrevem o conjunto de dados, nem avaliam o perfil de uso da rede. A redução de dimensionalidade do conjunto de dados pode ser realizadas através de métodos de seleção de características, selecionando as que melhor descrevem os dados [Andreoni Lopez et al. 2017], ou através de técnicas que levam os dados a outro espaço vetorial [Pascoal et al. 2012]. O presente artigo utiliza as técnicas descritas nesses trabalhos para realizar uma avaliação de relevância de atributos em relação à identificação de anomalias, resultando também em um conjunto de oito características principais.

Kato *et al.* propõem o uso de aprendizagem profunda (*deep learning*) para realizar a caracterização do tráfego da rede, já que advogam que a aprendizagem profunda é capaz de extrair padrões mais complexos do que outras técnicas [Kato et al. 2017]. Por sua vez, Nie *et al.* usam uma rede bayesiana para detectar anomalias e modelam a matriz de tráfego da rede [Nie et al. 2016]. Os resultados mostram que a previsão de tráfego é acurada, mas concentram-se no ambiente de computação em nuvem e não visam a predição de tráfego de usuários finais. O presente artigo oferece um conjunto de dados complementar ao utilizado por Kato *et al.*, uma vez que é composto somente por tráfego de usuários finais.

O conjunto de dados deste artigo é real e corresponde à captura, durante uma semana entre os dias 24 de fevereiro e 4 março de 2017, de pacotes de acesso à Internet de 373 usuários de banda larga fixa de uma importante operadora de telecomunicações na Zona Sul da cidade do Rio de Janeiro. São observadas as recomendações sobre a localização dos pontos de captura em [Heidemann e Papadopoulos 2009], bem como adotada como base a metodologia de caracterização de [Tavallaee et al. 2009]. Os alarmes de segurança são identificados pela análise dos dados através de uma ferramenta de detecção de intrusão, em contrapartida às abordagens sintéticas [Shiravi et al. 2012] e baseadas em potes de mel [Song et al. 2013, Chen et al. 2015]. Dessa forma, pretende-se possibilitar o estudo de métodos de classificação em dados atualizados, de forma a complementar e atualizar estudos como [Shiravi et al. 2012, Kato et al. 2017].

3. A Análise do Tráfego e o Conjunto de Dados de Acesso de Usuários Residenciais à Internet

A caracterização do perfil dos alertas em uma rede de acesso exige o monitoramento, o processamento e o gerenciamento de grandes volumes de dados gerados em tempo real. Esse grande volume de dados é processado por sistemas de detecção de intrusão (*Intrusion Detection System* - IDS) e, também, pela correlação com as informações de fluxos na rede [Wu et al. 2014, Lobato et al. 2016]. Ferramentas conhecidas como Gerenciamento e Correlação de Eventos de Segurança (*Security Information and Event Management*- SIEM) realizam este tipo de tarefa a um custo econômico elevado e ainda assim podem gerar atrasos. Em média, a reação a ameaças de segurança é tomada após 123 horas da ocorrência e, no caso de detecção do vazamento de informações, a demora na identificação dessa falha de segurança chega a 206 dias [Clay 2015]. Ao conhecer o

perfil mais comum dos alertas de segurança, é possível identificar mais rapidamente o vazamento de informações e as vulnerabilidades exploradas.

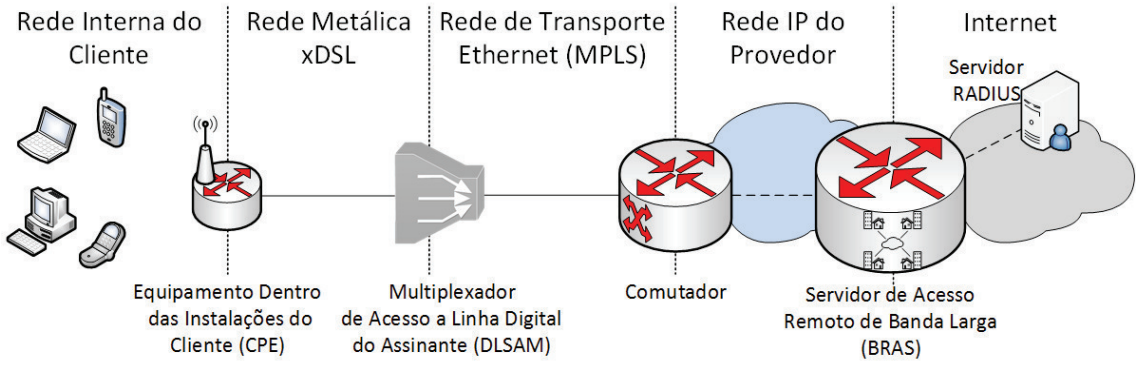


Figura 1. Topologia típica da rede de acesso banda larga. A conexão entre o *Home Gateway* e a Internet é autenticada, contabilizada e registrada pelo servidor *Radius*. O tráfego é encapsulado em sessões PPPoE (*Point-to-Point Protocol over Ethernet*) entre a casa do usuário e o BRAS (*Broadband Remote Access Server*). A inspeção e a coleta do tráfego ocorre após o BRAS.

A Figura 1 mostra uma topologia de acesso típica para o serviço de banda larga composta por um equipamento dentro das instalações do usuário, *home gateway* ou CPE (*Customer Premises Equipment*), ligado a um multiplexador de acesso (*Digital Subscriber Line Asymmetric Multiplexer* - DSLAM), uma rede de transporte, como por exemplo uma rede MPLS (*Multiprotocol Label Switching*), e um agregador de seções (*Broadband Remote Access Server* - BRAS) que autentica a sessão dos usuários através de um servidor RADIUS, responsável também pela auditoria de uso da rede. Assim, em uma rede de acesso para usuários de banda larga fixa, a inspeção é realizada somente após a agregação do tráfego, já que não há nós que permitam a inspeção dos dados nas premissas do usuário ou no perímetro mais próximo dos usuários.

O tráfego a ser analisado é composto pelo tráfego agregado proveniente da alta capilaridade de diferentes usuários, com uma grande variedade de perfis de serviços acessados por cada usuário e gerando um grande volume de dados. Portanto, o problema de caracterização do perfil de alertas consiste um problema complexo de análise de grandes massas de dados (*big data*) [Costa et al. 2012], que requer ferramentas de processamento apropriadas. A ideia central deste artigo é gerar, analisar e caracterizar um conjunto de dados que represente o mais fielmente possível o perfil de uso dos usuários de banda larga fixa residencial com a finalidade de treinar classificadores de tráfego. Como foi mencionado anteriormente, o conjunto de dados corresponde ao tráfego de acesso de 373 usuários de banda larga fixa de uma grande operadora de telecomunicações na Zona Sul da cidade do Rio de Janeiro. Os dados foram coletados nas premissas da operadora e foram anonimizados para garantir o sigilo e a privacidade dos usuários. As análises realizadas descartam verificações do conteúdo dos pacotes. A base de dados analisada foi criada a partir da captura de pacotes brutos, contendo informações reais de tráfego IP (*Internet Protocol*) dos usuários residenciais. O tráfego foi coletado e gravado de forma ininterrupta por uma semana através do *software* `tcpdump`[1]. O processo de coleta e gravação

[1]Disponível em http://www.tcpdump.org.

dos arquivos não utilizou quaisquer filtros de pacotes e, portanto, todos os pacotes da rede foram gravados na sua forma bruta (*raw data*) diretamente na base de dados. A estrutura física de coleta foi configurada espelhando o tráfego agregado de um DSLAM em uma outra porta do comutador *metro-ethernet* da rede de transporte. O espelhamento da porta do DSLAM no comutador permite que todo o tráfego originado ou destinado para o DSLAM seja clonado para um computador executando o sistema operacional Linux Ubuntu, o qual foi conectado a esta segunda porta para coletar e gravar em formato *pcap* todos os pacotes. Para garantir o armazenamento em alta velocidade e para permitir o fácil transporte dos dados, a base de dados foi gravada em um disco rígido externo com interface USB 3.0. A Figura 2 mostra a topologia básica e a estrutura montada para a coleta dos dados. Vale ressaltar que, embora a rede da operadora considerada seja muito maior que a amostragem tomada neste trabalho, os dados coletados representam o consumo real de usuários residenciais e não é o escopo do trabalho analisar todo o tráfego da operadora.

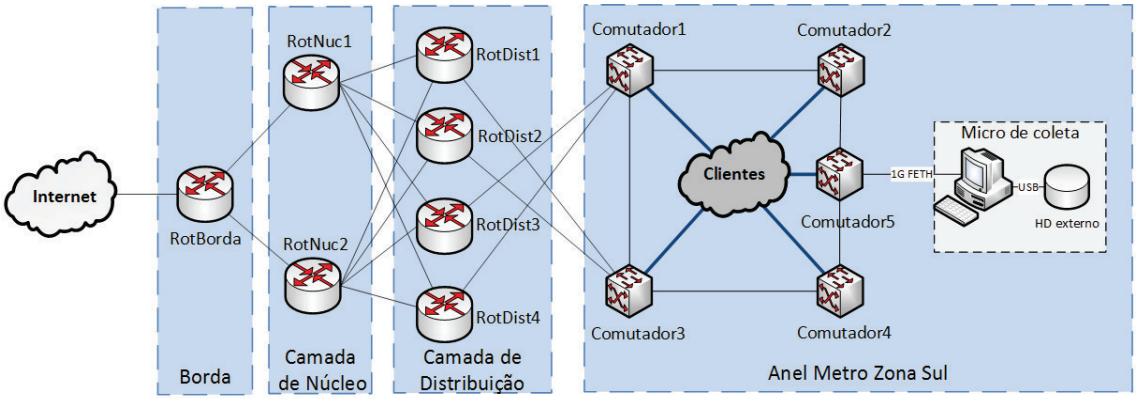


Figura 2. Topologia da estrutura de coleta de dados da porta principal do DSLAM com 373 clientes de banda larga.

O procedimento usado na captura dos dados garantiu não haver perdas de pacotes no espelhamento de portas a 1Gb/s montado para coleta e gravação dos pacotes. Assim, 100% do tráfego gerado pelos 373 clientes foi coletado e gravado na base de dados, totalizando 5TB de informações. Apesar da velocidade média disponível em cada porta do DSLAM ser de aproximadamente 12Mb/s, gerando um tráfego agregado hipotético superior a 4Gb/s, foi verificado que durante todo o processo de captura, o tráfego real agregado não superou a taxa de 800Mb/s. O tráfego agregado conta com tráfegos de ida e volta (*uplink* e *downlink*). Vale notar que os dados capturados são somente de usuários residenciais, portanto todo tráfego é proveniente de sessões banda larga fixa.

4. A Análise dos Dados

A análise dos dados capturados da rede da operadora de telecomunicações foi dividida em três etapas. A primeira etapa trata os arquivos de captura de dados brutos através de um sistema de detecção de intrusão (*Intrusion Detection System* - IDS) de rede e, posteriormente, gera um resumo dos dados na forma de fluxos. A segunda etapa analisa a distribuição das principais características do conjunto de dados, evidenciando diferenças entre o tráfego normal e o tráfego que gera alerta. Por fim, a terceira etapa consiste em comparar o uso de classificadores para verificar a acurácia do classificador em separar tráfego normal do tráfego gera que alertas. Ao realizar a classificação é possível

direcionar somente a parcela do tráfego que pode gerar alerta dentre todo o tráfego da operadora de telecomunicações para o IDS, diminuindo a carga de tráfego analisada.

A primeira etapa de análise dos dados baseou-se na extração das características dos fluxos representados pelos pacotes capturados, assim como na verificação de possíveis alertas através do IDS. Por se tratar de um tráfego de clientes residenciais com acesso ADSL (*Asymmetric Digital Subscriber Line*), o tráfego capturado é encapsulado em sessões PPPoE (*Point-to-Point Protocol over Ethernet*), o que dificulta a análise dos pacotes, já que alguns IDS não realizam a inspeção desse tipo de tráfego, como, por exemplo, o SNORT [Roesch et al. 1999]. Portanto, a fim de executar a classificação do tráfego em diferentes tipos de alertas, foi usado o IDS Suricata[2], versão 3.2, com a sua base de assinaturas atualizada. Vale ressaltar que a classificação entre tráfego normal e alerta foi realizada somente com base nas assinaturas do Suricata, pois não havia conhecimento prévio sobre a origem dos dados coletados. Como os dados são reais, não é possível assegurar que todos os fluxos são legítimos ou, mesmo após a classificação do IDS, que todos os fluxos de alerta são de fato maliciosos. No entanto, a classificação gerada pelo IDS é utilizada como referência e considerada como definitiva e correta no contexto do artigo.

Paralelamente à classificação dos pacotes pelo IDS, os pacotes capturados foram desencapsulados da sessão PPPoE, usando a ferramenta `stripe`[3], e foram resumidos em fluxos, através da ferramenta `flowtbag`[4]. Ademais, foi desenvolvida uma aplicação Python que processa a saída do IDS, o relatório de fluxos que geram alertas, e correlaciona os alertas com a informação de fluxos resumida. Assim, foi possível obter um conjunto de dados de fluxos com a marcação da classe a qual pertencem. O conjunto de dados apresenta 42 características de cada fluxo e mais a classe a que pertence cada fluxo. A classe de saída, característica 43, é dada pelo tipo de alerta gerado pelo IDS, no caso de um fluxo que dispara um alerta, ou o fluxo é marcado com a classe 0 indicando que é um fluxo normal. No conjunto de dados não se adicionam os endereços IP de origem e de destino dos fluxos para garantir a *"anonimização"* dos dados.

A segunda etapa consiste em extrair conhecimento dos dados. Para tanto, foi utilizada a plataforma de análise de dados gratuita e de código aberto KNIME[5], versão 3.3.1. Em um primeiro momento, a análise dos dados foca na Análise das Componentes Principais (*Principal Component Analysis* - PCA) com o intuito de verificar quais são as características do conjunto de dados que carregam mais informação. Assim, a Figura 3 mostra as seis componentes principais do conjunto de dados. Foram escolhidas seis componentes principais, pois são aquelas em que o autovalor absoluto é muito maior que 0, o que garante a preservação de 99% da informação do conjunto de dados. As componentes foram ordenadas em relação ao autovalor associado a cada autovetor que define a componente. Tendo em vista as componentes principais, destaca-se que as características mais relevantes para a caracterização do tráfego são: porta de origem, porta de destino, volume total do fluxo, quantidade de pacotes no fluxo, volume dos subfluxos de ida e de volta e volume de dados em cabeçalhos nos fluxos de ida e de volta. A partir dessas características, analisou-se o comportamento do tráfego normal e dos alertas.

[2]Disponível em https://suricata-ids.org.
[3]Disponível em https://github.com/theclam/stripe.
[4]Disponível em https://github.com/DanielArndt/flowtbag.
[5]Disponível em https://www.knime.org/

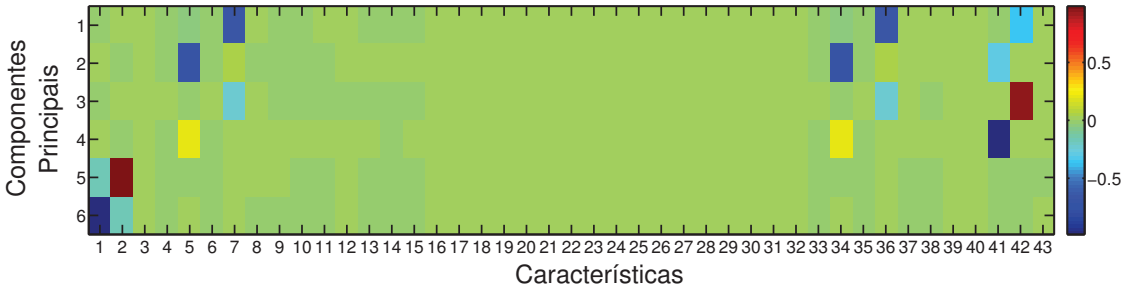
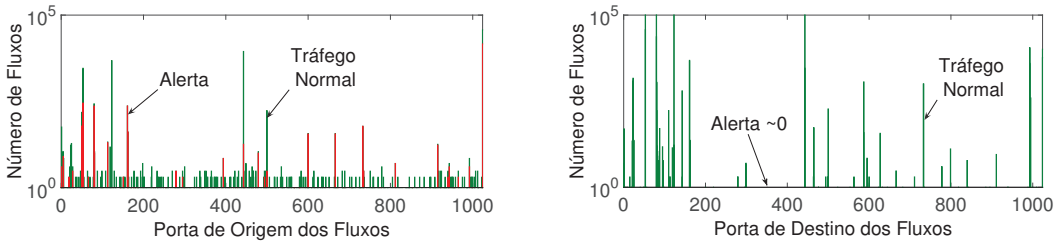


Figura 3. Visualização das seis componentes obtidas com a Análise de Componentes Principais (PCA) sobre o conjunto de dados, em ordem decrescente de autovalor. As oito características que mais se destacam nas componentes são: porta de origem (1), porta de destino (2), volume total do fluxo (5), quantidade de pacotes no fluxo (7), volumes dos subfluxos de ida (34) e de volta (36), volume de dados em cabeçalhos nos fluxos de ida (41) e de volta (42).

A primeira característica analisada foi a porta em que o tráfego ocorre. A Figura 4 apresenta as portas de origem e destino dos fluxos. A figura concentra-se sobre as 1024 primeiras portas (de 0 a 1023), pois são as portas restritas. Usualmente, essas portas são usadas por *daemons* que executam serviços com privilégios de administrador do sistema. A definição de fluxo usada considera como porta de origem a porta que inicia a conexão TCP. Como o conjunto de dados retrata usuários residenciais, é esperado que a maior parte das conexões seja destinada a portas restritas e não originadas dessas. Assim, verifica-se que o número de alertas provenientes de conexões que usam as portas restritas como destino é baixo em relação ao total de conexões nessas portas, Figura 4(b). Contudo, ao se considerar os fluxos que usam as portas restritas como porta de origem, quase a totalidade dos fluxos é marcada como alerta pelo IDS, mostrado na Figura 4(a). Outro fato marcante é que se observa que grande parte dos fluxos analisados refletem o uso do serviço de DNS (UDP 53) e os serviços HTTPS e HTTP (TCP 443 e 80). O predomínio do uso de serviços HTTPS sobre os HTTP reflete a mudança de que os principais provedores de conteúdo da Internet, tais como Google e Facebook, têm passado a usar o serviço criptografado por padrão para garantir a segurança e privacidade dos usuários.

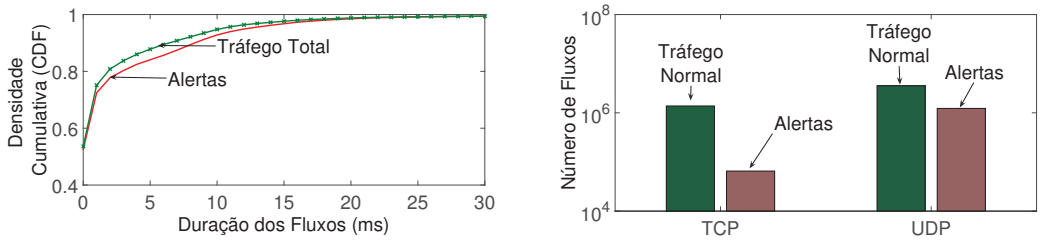


(a) Distribuição das portas de origem dos fluxos. (b) Distribuição das portas de destino dos fluxos.

Figura 4. Portas usadas nos fluxos. Comparação do uso das 1024 portas mais baixas (portas restritas) nos fluxos avaliados. Por se tratarem de usuários domésticos, o maior número de fluxos com origem nessas portas são fluxos que geram alertas.

Os resultados dos serviços mais comumente acessados na rede tornam-se mais claros ao serem comparados com a duração e os protocolos usados nos fluxos, mostrados

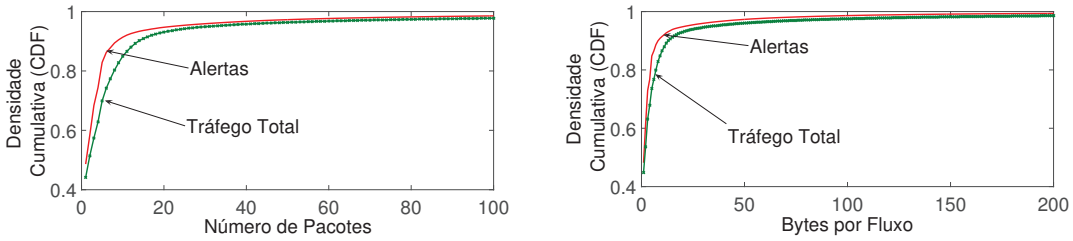
nas Figuras 5(a) e 5(b). A duração dos fluxos analisados majoritariamente é menor que 30 ms, caracterizando o uso dos serviços DNS, HTTP e HTTPS. Uma boa aproximação para duração média dos fluxos é a distribuição Erlang, com $k = 1$ e $\lambda = 3.7$. A aproximação foi calculada através da adequação da distribuição aos dados e através da minimização do erro quadrático médio entre a distribuição e os dados obtidos. Contudo, pelo teste hipótese de Kolmogorov-Smirnov[6], a hipótese de que os dados seguem tal distribuição deve ser rejeitada. Em relação aos protocolos usados, é evidente o predomínio de fluxos UDP, referentes a consultas DNS. Vale ressaltar que o número de alertas gerados por fluxos UDP é mais de 10 vezes superior ao número de alertas gerados por fluxos TCP. Outro ponto importante é que o número de fluxos que geram alertas é de aproximadamente 26% dos fluxos totais.



(a) Duração dos Fluxos. (b) Protocolos de transporte usados.

Figura 5. Função de Densidade de Probabilidades Cumulativa (CDF) para a distribuição da duração dos fluxos em *ms* e número de fluxos por protocolos de transporte. a) Os fluxos que geram alertas tendem a ser de menor duração que os fluxos totais. b) Os fluxos legítimos com UDP são numerosos em função do DNS (porta 53 UDP). O número de alertas em UDP é mais de 10 vezes maior que em fluxos TCP.

A Figura 6 mostra a caracterização do número de pacotes por fluxo, em ambos os sentidos da comunicação, ida e volta. Em ambos os sentidos, a comunicação ocorre com até 32 pacotes em 95% dos casos e com até 100 pacotes em 98, 5%. O resultado mostra que as conexões no cenário residencial são em sua grande maioria conexões com poucos pacotes. Nota-se ainda que os fluxos que geram alertas têm, em geral, menos pacotes do que fluxos do tráfego normal, pois 95% dos fluxos de alerta apresentam até 22 pacotes.

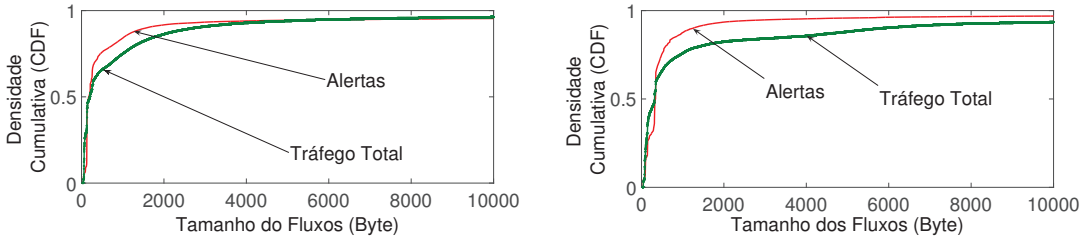


(a) Fluxos no sentido de ida. (b) Fluxos no sentido de volta.

Figura 6. Função de Densidade de Probabilidades Cumulativa (CDF) para a distribuição do número de pacotes por fluxo. Fluxos que geram alertas tendem a ter menos pacotes.

[6]O teste de Kolmogorov-Smirnov verifica se uma das distribuições de probabilidade difere da distribuição em hipótese com base em um número finito de amostras.

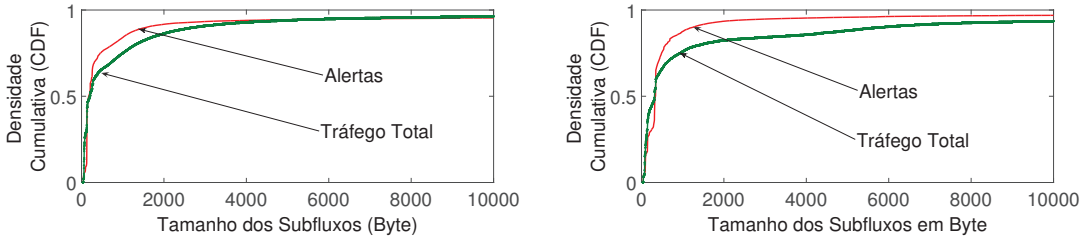
Considerando-se a quantidade de dados trafegada em cada fluxo, a Figura 7 compara os fluxos de ida e volta em relação ao volume em *bytes* trafegados. É visível a disparidade do volume de tráfego nos dois sentidos da comunicação. Enquanto, no sentido de ida, 95% do tráfego apresenta no máximo o volume de $6,4$ kB, no sentido de volta, a mesma parcela de tráfego apresenta até 18 kB. A distribuição que melhor se adequa ao volume de dados, verificando-se a que minimiza o erro quadrático médio, é distribuição lognormal, com $\mu = 5.67$ e $\sigma = 31.68$. A validação da adequação à distribuição lognormal foi realizada através do teste de Kolmogorov-Smirnov sobre uma subamostragem aleatória dos dados para uma significância estatística de 95%. Esse resultado demonstra que o perfil do usuário de banda larga residencial é o de um consumidor de conteúdo. Outro ponto interessante é que os fluxos que geram alertas têm um perfil de volume de tráfego semelhante nos sentidos de ida e de volta. Tráfegos assimétricos são mais característicos do usuário classificado como legítimo.



(a) Fluxos no sentido de ida. (b) Fluxos no sentido de volta.

Figura 7. Função de Densidade de Probabilidades Cumulativa (CDF) para a distribuição do volume em *bytes* por fluxo. Fluxos que geram alertas tendem a ter menor volume em *bytes* trafegados.

Já a Figura 8 mostra o comportamento dos subfluxos gerados em cada conexão. Tal característica foi apontada pelo método PCA como uma das características mais importantes para se descrever o conjunto de dados. No entanto, o comportamento estatístico do volume de dados dos subfluxos é o mesmo do fluxo total. Isso ocorre, pois os fluxos são majoritariamente de curta duração, evidenciado na Figura 5(a), e assim não geram subfluxos. A análise dos dados mostrou que os fluxos não passam ao estado *idle*.

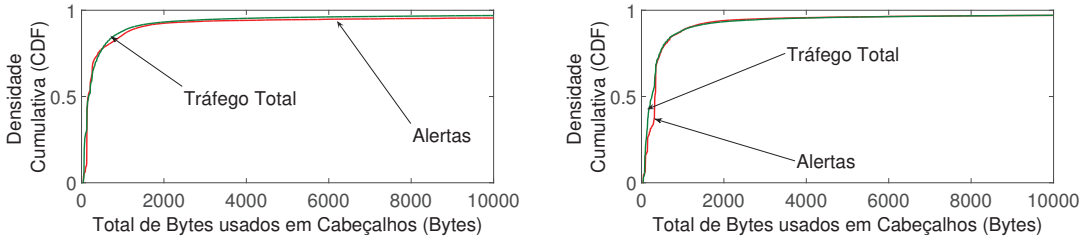


(a) Fluxos no sentido de ida. (b) Fluxos no sentido de volta.

Figura 8. Função de Densidade de Probabilidades Cumulativa (CDF) para a distribuição do volume em *bytes* por subfluxo em cada fluxo. Fluxos que geram alertas tendem a ter menor volume em *bytes* trafegados em subfluxos.

Outra característica importante é a quantidade total de dados trafegada nos cabeçalhos dos pacotes. A Figura 9 evidencia que, em ambos os sentidos dos fluxos, tanto o tráfego marcado como alerta quanto o total apresentam o mesmo comportamento. Em

especial, percebe-se uma simetria no tráfego de ida e de volta quanto ao volume de dados nos cabeçalhos.



(a) Fluxos no sentido de ida. (b) Fluxos no sentido de volta.

Figura 9. Função de Densidade de Probabilidades Cumulativa (CDF) para a distribuição do volume em *bytes* dos dados trafegados em cabeçalhos. O comportamento do tráfego que gera alertas é muito semelhante ao tráfego total.

Ao fim da etapa de extração de conhecimento dos dados, foi analisado o perfil dos alertas gerados pelo IDS. A Figura 10 mostra quais são as principais classes de alertas disparados pelo IDS. Destacam-se primeiramente os alertas de ataques ao HTTP. Nessa classe de alertas enquadram-se os ataques de injeção de SQL através de chamadas HTTP e ataques XSS (*cross-site scripting*). Tais ataques são plausíveis de serem executados por usuários residenciais, pois usam os parâmetros das chamadas HTTP para inserir algum código malicioso nos servidores e, portanto, não são barrados por regras de acesso. Outros alertas importantes são os de escaneamento de portas e vulnerabilidade (`scan`) e os de execução de aplicativos maliciosos (`trojan` e `malware`). Os escaneamentos visam, no geral, identificar portas abertas e vulnerabilidades nas premissas do usuário (*gateway* doméstico). Os alertas referentes a `trojan` e `malware` identificam atividades características de aplicativos maliciosos conhecidos que visam criar e explorar vulnerabilidades nos dispositivos dos usuários residenciais. Os demais alertas são referentes a mecanismos de roubos de informação e, também, a assinaturas de ataques bizantinos em protocolos comuns, como IMAP e Telnet[7].

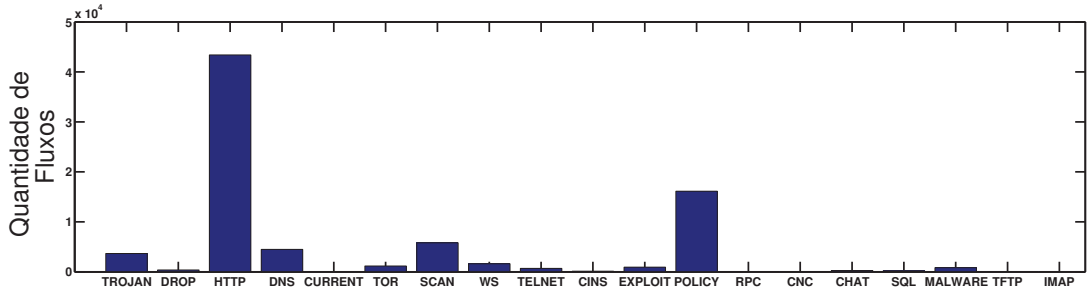


Figura 10. Distribuição dos principais tipos de alertas no tráfego analisado.

A terceira etapa da análise dos dados consiste em projetar um classificador para separar os fluxos em normais ou naqueles que podem gerar alarmes. O objetivo é realizar a classificação dos fluxos com uma precisão considerável para otimizar a análise de trafego realizado pela operadora de telecomunicações, pois somente o tráfego classificado como

[7]Principalmente usado para a configuração remota de equipamentos de rede.

suspeito deverá ser desviado para ser tratado pelo serviço de IDS. Para tanto, o foco do artigo é realizar a classificação supervisionada, em que o treinamento do classificador é feito com um conjunto de dados marcados com uma classe de saída. As métricas de avaliação do classificador verificam o quanto o classificador proposto se aproxima da classe de saída pré-determinada para cada fluxo. No caso deste artigo, a classe de saída usada no treinamento e avaliação do classificador é o resultado da análise pelo IDS.

A primeira abordagem de classificador considera uma rede neural com duas camadas ocultas com 20 neurônios em cada, com seis neurônios na camada de entrada e um neurônio na camada de saída. O algoritmo de aprendizado usado foi o *Multilayer Perceptron* (MLP) com o ajuste de pesos através do algoritmo de *backward propagation*. A modelagem da rede neural leva em consideração que a entrada dos dados é formada pelas seis componentes principais calculadas pelo PCA. A saída da rede é a probabilidade de o fluxo ser marcado como suspeito. Considera-se fluxo suspeito todo aquele que tiver saída maior que 0,5 e fluxo normal aqueles cujo valor final esteja abaixo desse limiar. A Tabela 1 mostra o desempenho da rede neural em uma avaliação cruzada em 10 rodadas[8]. Verifica-se que a acurácia da rede neural foi de $0,847$, com sensibilidade de $0,625$ na classe de alerta.

Tabela 1. Classificação com Rede Neural com 2 camadas ocultas.

	VP	FP	VN	FN	Precisão	Sens.	Espec.
Alerta	809626	396162	3234673	485174	0.671	0.625	0.891
Normal	3234673	485174	809626	396162	0.870	0.891	0.625

Tabela 2. Classificação com Árvore de Decisão.

	VP	FP	VN	FN	Precisão	Sens.	Espec.
Alerta	1209238	87499	3543336	85562	0.933	0.934	0.976
Normal	3543336	85562	1209238	87499	0.976	0.976	0.934

Em uma segunda abordagem, avaliou-se o uso do classificador baseado em Árvore de Decisão. A entrada do classificador foram as seis componentes principais extraídas do PCA. A saída do classificador é a marcação em uma das duas classes possíveis, alerta ou normal. A Tabela 2 mostra o resultado da classificação usando-se a Árvore de Decisão em uma avaliação cruzada em 10 rodadas. A acurácia atingida pelo classificador foi de $0,956$. A sensibilidade na classe de alertas foi de $0,934$ e de $0,976$ na classe normal. Esse resultado mostra que o uso desse classificador como um pré-tratamento dos fluxos reduz em até 73% a carga no analisador de tráfego da operadora de telecomunicações, com uma sensibilidade de $0,934$ no tráfego suspeito.

5. Conclusão

O conhecimento do perfil de uso da rede é importante para melhor dimensionar a rede e identificar os principais serviços utilizados. A identificação dos principais alertas de segurança na rede, por sua vez, permite conhecer quais são as principais ameaças e projetar possíveis contramedidas. Esse artigo apresentou a criação de um conjunto de

[8]VP: Verdadeiro Positivo; FP: Falso Positivo; VN: Verdadeiro Negativo; FN: Falso Negativo; Sens.: Sensibilidade; Espec: Especificidade.

dados de alertas de segurança em uma rede real de uma operadora de telecomunicações na cidade do Rio de Janeiro[9]. O conjunto de dados representa o uso do serviço de acesso banda larga fixa de 373 usuários residenciais. A análise dos dados permite identificar que os principais serviços acessados são os de DNS e serviços *web*. O perfil dos fluxos é caracterizado por conexões rápidas, de até 30 ms, com a transferência de até 18 kB em 95% dos casos. A fim de validar o conjunto de dados coletado quando à acuidade da marcação de alertas por fluxo, foram empregados métodos de aprendizado de máquina para classificação destes fluxos em uso normal e alertas. Os resultados obtidos através da aplicação de rede neural e árvore de decisão demonstram que a marcação aplicada ao conjunto de dados é consistente com padrões observáveis. Com base na caracterização do uso normal e dos alertas, esse artigo propôs o uso de um classificador baseado em árvore de decisão para identificar os fluxos suspeitos e reduzir a carga no analisador de tráfego. Os resultados mostram que o uso de um classificador simples é capaz de reduzir em até 73% o tráfego enviado ao analisador de tráfego, com a capacidade de identificar até 93% dos fluxos que geram alertas na rede.

Como trabalhos futuros, pretende-se utilizar este conjunto de dados na validação de uma arquitetura de processamento por fluxos para a classificação por árvores de decisão, verificando a precisão da análise de tráfego em tempo real. Ademais, será avaliado o desempenho de outros algoritmos de classificação e de treinamento em tempo real.

Referências

Andreoni Lopez, M., Lobato, A., Mattos, D. M. F., Alvarenga, I. D., Duarte, O. C. M. B. e Pujolle, G. (2017). Um algoritmo não supervisionado e rápido para seleção de características em classificação de tráfego. Em *XXXV Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos (SBRC'2017 - a ser apresentado)*.

Bhatt, S., Manadhata, P. K. e Zomlot, L. (2014). The operational role of security information and event management systems. *IEEE Security Privacy*, 12(5):35–41.

CAIDA (2007). Supporting research and development of security technologies through network and security data collection. http://www.caida.org/funding/predict/. Acessado 28-03-2017.

Chen, Y.-Z., Huang, Z.-G., Xu, S. e Lai, Y.-C. (2015). Spatiotemporal patterns and predictability of cyberattacks. *PLOS ONE*, 10(5):1–19.

Clay, P. (2015). A modern threat response framework. *Network Security*, 2015(4):5–10.

Costa, L. H. M. K., de Amorim, M. D., Campista, M. E. M., Rubinstein, M. G. e Duarte, O. C. M. B. (2012). Grandes massas de dados na nuvem: Desafios e técnicas para inovação. Em *Minicursos do XXX Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos (SBRC'2012)*.

D. Kotz, T. H. e Abyzov, I. (2004). CRAWDAD: A community resource for archiving wireless data at dartmouth. http://crawdad.cs.dartmouth.edu/. Acessado 28-03-2017.

Draper-Gil, G., Lashkari, A. H., Mamun, M. S. I. e Ghorbani, A. A. (2016). Characterization of encrypted and vpn traffic using time-related features. Em *Proceedings of the*

[9]Os dados anonimizados podem ser consultados através de contato com os autores por e-mail.

2nd International Conference on Information Systems Security and Privacy, páginas 407–414.

Heidemann, J. e Papadopoulos, C. (2009). Uses and challenges for network datasets. Em *Proceedings of the IEEE Cybersecurity Applications and Technologies Conference for Homeland Security (CATCH)*, páginas 73–82, Washington, DC, USA. IEEE.

IBGE (2016). *Síntese de indicadores sociais : uma análise das condições de vida da população brasileira*, volume 36. IBGE, Rio de Janeiro.

Kato, N., Fadlullah, Z. M., Mao, B., Tang, F., Akashi, O., Inoue, T. e Mizutani, K. (2017). The deep learning vision for heterogeneous network traffic control: Proposal, challenges, and future perspective. *IEEE Wireless Communications*, PP(99):2–9.

Lobato, A., Andreoni Lopez, M. e Duarte, O. C. M. B. (2016). Um sistema acurado de detecção de ameaças em tempo real por processamento de fluxos. Em *XXXIV Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos (SBRC'2016)*, Salvador, Bahia.

Nie, L., Jiang, D. e Lv, Z. (2016). Modeling network traffic for traffic matrix estimation and anomaly detection based on bayesian network in cloud computing networks. *Annals of Telecommunications*, páginas 1–9.

NSL-KDD (2009). Nsl-kdd data set for network-based intrusion detection systems. http://iscx.cs.unb.ca/NSL-KDD/. Acessado 28-03-2017.

Pascoal, C., de Oliveira, M. R., Valadas, R., Filzmoser, P., Salvador, P. e Pacheco, A. (2012). Robust feature selection and robust PCA for Internet traffic anomaly detection. Em *2012 Proceedings IEEE INFOCOM*, páginas 1755–1763.

Puthal, D., Nepal, S., Ranjan, R. e Chen, J. (2016). Threats to networking cloud and edge datacenters in the internet of things. *IEEE Cloud Computing*, 3(3):64–71.

Roesch, M. et al. (1999). SNORT: Lightweight intrusion detection for networks. Em *Lisa*, volume 99, páginas 229–238.

Shiravi, A., Shiravi, H., Tavallaee, M. e Ghorbani, A. A. (2012). Toward developing a systematic approach to generate benchmark datasets for intrusion detection. *Comput. Secur.*, 31(3):357–374.

Song, J., Takakura, H., Okabe, Y. e Nakao, K. (2013). Toward a more practical unsupervised anomaly detection system. *Inf. Sci.*, 231:4–14.

Tavallaee, M., Bagheri, E., Lu, W. e Ghorbani, A. A. (2009). A detailed analysis of the kdd cup 99 data set. Em *2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications*, páginas 1–6.

Wu, K., Zhang, K., Fan, W., Edwards, A. e Yu, P. S. (2014). RS-Forest: A rapid density estimator for streaming anomaly detection. Em *2014 IEEE International Conference on Data Mining*, páginas 600–609.