Geração de *workload* para treinamento, teste, avaliação e comparação de sistemas de detecção de intrusão baseado na replicação de características estatísticas

Moisés da S. Rodrigues, Sidnei Barbieri, José M. Parente de Oliveira
César A. C. Marcondes, Lourenço Alves Pereira

[1]Divisão de Ciência da Computação
Instituto Tecnológico de Aeronáutica—ITA

{moises,sidnei,parente,cmarcondes,ljr}@ita.br

Resumo. *Este artigo apresenta o resultado de um trabalho de geração de datasets para sistemas de detecção de intrusão. Este processo utiliza datasets existentes constituídos por pacotes, assim como suas características estatísticas. A modelagem da distribuição de variáveis como, por exemplo, o número de pacotes contíguos de mesmo rótulo, é responsável pela geração de novas estruturas, as quais preservam características estatísticas semelhantes. Como resultado, obtivemos uma nova base de dados, que pode ser utilizada não apenas para o treinamento e teste de novos sistemas de detecção, mas também para a reavaliação de outros sistemas previamente desenvolvidos. Isto possibilita a comparação com os resultados de novos trabalhos.*

Abstract. *This paper presents the results of a dataset generation work for intrusion detection systems. This process uses existing datasets made up of packets, as well as their statistical characteristics. The modeling of the distribution of variables, such as the number of contiguous packets of same label, is responsible for the generation of new structures, which preserve similar statistical characteristics. As a result, we have obtained a new database, which can be used not only for the training and testing of new detection systems, but also for the re-evaluation of other previously developed systems. This enables comparison with the results of new jobs.*

1. Introdução

Considerando a relevância de sistemas de detecção de intrusão (IDS), tanto os baseados em uso indevido quanto os baseados em anomalias, para a segurança cibernética, verifica-se a necessidade da disponibilidade de bases de dados confiáveis, compostas por tráfego normal (benigno), e também por tráfego de ataque (maligno), as quais são denominadas *datasets* [Lippmann et al. 2000]. O principal problema, contudo, reside na identificação de bases disponíveis, uma vez que tal disponibilização implica em considerações relacionadas à segurança da informação.

Partindo da premissa de que os IDS, em sua finalidade principal, destinam-se à detecção de tráfego real, vislumbramos que o ideal seria que os *datasets* fossem os mais realistas possível, representando ambientes operacionais de larga escala [Sommer and Paxson 2010]. Sob essa ótica, verificamos que os *datasets* deveriam ser construídos

exclusivamente a partir de tráfego real coletado diretamente em ambientes de redes em produção. Infelizmente tal abordagem tem como principal óbice a dificuldade de obtenção desses dados, uma vez que não é comum o compartilhamento de dados reais, o que também afeta, ainda que parcialmente, os critérios da completude das capturas.

Ao considerarmos *datasets* baseados em tráfego real, devemos considerar que além dos dados desejados para a pesquisa, há, também, informações relevantes para as organizações, como informações sobre sua estrutura de rede ou, ainda, sobre o negócio. Desse modo, não é comum que dados sejam efetivamente compartilhados, o que se justifica pelo argumento da garantia da privacidade. Convém, ainda, ressaltar que, mesmo nas situações em que o acesso é franqueado, muitas vezes tal acesso e a utilização está condicionada à aprovação dos custodiantes, mediante o preenchimento de formulários com a assinatura de concordância dos termos de uso aceitável [Nehinbe 2011], o que, eventualmente, pode impactar na celeridade da pesquisa. Há, ainda, situações em que os dados, antes de serem disponibilizados, são submetidos a um processo anonimização, o que, por vezes, chega a desfigurar a informação, tornando-a inutilizável.

Em virtude das dificuldades inerentes à obtenção de *datasets* baseados na coleta de tráfego real, é fato que existe uma carência de *datasets* que sejam atuais, contenham representação de tráfego real e que seja coletado em redes modernas [Zuech et al. 2015]. Em consequência, verificamos que a alternativa que se apresenta mais viável consiste na geração de tráfego simulado para a composição de *datasets*, definidos como *benchmamarks datasets* [Rao and Naidu 2017]. Na verdade, tal conclusão não é recente, [Scott and Wilkins 1999] já apontavam a necessidade da geração de *datasets* artificiais, sujeitos, contudo, à manutenção de determinadas características de *datasets* reais, como os tipos de padrão identificados no tráfego real.

Gerar *datasets* artificiais, todavia, não é uma tarefa simples, uma vez que deve-se garantir que eles sejam completos, válidos e adequados [Sharafaldin et al. 2017]. Deve-se, também, considerar que, à medida que ataques mais sofisticados são desenvolvidos, algumas atividades maliciosas podem se tornar obsoletas, sendo, portanto, desejável que os datasets sejam dinâmicos, de maneira que possam ser modificáveis, extensíveis e reprodutíveis. Ainda quanto ao tamanho, deve-se ter um cuidado especial para que as bases tenham um tamanho tal que possibilite sua utilização. A quantidade excessiva de registros pode inviabilizar a a utilização do *datasets* como um todo, implicando na necessidade da extração de amostras que, por mais que decorram da aplicação de princípios estatísticos, aumentam a probabilidade de descartar tráfego relevante, tanto sob o ponto de vista de protocolos quanto de ataques [Tavallaee et al. 2009].

A geração de *datasets* artificiais deve garantir que eles sejam completos, válidos e adequados [Sharafaldin et al. 2017]. Na medida que ataques mais sofisticados são desenvolvidos, algumas atividades maliciosas podem se tornar obsoletas, sendo desejável que os datasets sejam dinâmicos, adaptáveis, extensíveis e reprodutíveis. O tamanho das bases de dados também é significativo. A quantidade excessiva de registros pode inviabilizar a utilização do *datasets*, exigindo a extração de amostras que podem implicar na perda de tráfego relevante, tanto sob o ponto de vista de protocolos quanto de espécies de ataque [Tavallaee et al. 2009].

Diversos trabalhos sobre técnicas de avaliação de IDS e de detecção de intrusão vêm

sendo desenvolvidos nos últimos anos, muitos deles baseados em *datasets* distintos, o que dificulta sua comparação com os demais trabalhos desenvolvidos com a mesma finalidade. Assim, incentivado pela carência de *datasets* atuais, contendo tráfego real e coletado em redes modernas [Zuech et al. 2015], o presente trabalho visa propor um método para geração de tráfego simulado, a partir do qual são constituídos *datasets* aplicáveis, entre outros aspectos, na avaliação comparativa de IDS.

Em face do exposto, este artigo propõem e demonstra a aplicação de um método para a geração artificial de *workload* aplicável ao treinamento, teste, avaliação e comparação de sistemas de detecção de intrusão, baseado na replicação de características estatísticas. A seguir, na Seção 2 são discutidos trabalhos relacionados. Já a Seção 3 é dedicada à descrição do método desenvolvido, enquanto que a Seção 4 detalha a sua aplicação no estudo de caso proposto. Por fim, na Seção 5 são apresentadas as conclusões, as limitações e a proposta de trabalhos futuros.

2. Trabalhos relacionados

A geração de *datasets* para sistemas de detecção de intrusão tem sido objeto de trabalhos há mais de 20 anos, sendo, segundo [McHugh 2000], o DARPA 98/99 o primeiro digno de avaliação. Desde então, outros trabalhos propuseram a avaliação ou a geração de novos conjuntos de dados, visando sempre à garantia de bases consideradas realistas e válidas. Em que pese a antiguidade dessa base bem como do KDD Cup 99, nela baseado, esses *datasets* vêm sendo utilizados em diversas pesquisas. O principal impedimento a eles relacionado está na sua desatualização, uma vez que não refletem a evolução do tráfego nem dos ataques. [Tavallaee et al. 2009] propuseram uma análise de problemas estatísticos relacionados ao KDD 99 gerando um novo conjunto de dados, o NSL-KDD, consistindo em uma otimização do conjunto original. Apesar da melhoria alcançada, por ter como base dados já obsoletos, o *dataset* resultante também padeceu das mesmas limitações.

Tendo como referência a busca pelo realismo, [Shiravi et al. 2012] apresentaram 5 prerrequisitos preliminares a serem considerados quando da geração de um *dataset*, a saber: conter tráfego de rede realista, ser rotulado, conter interações entre todas as redes, conter capturas de dados completas e contemplar múltiplos cenários de ataque. Em consequência, seu trabalho culminou na geração de *datasets* baseados na descrição de cenários de ataque e na distribuição matemática do comportamento das entidades presentes em redes computacionais. A despeito dos bons resultados obtidos, conforme destacado pelos próprios autores tal abordagem se limita a contextos particulares, com requisitos e aplicações específicos.

[Gharib et al. 2016] apresentaram um *framework* para avaliação de *datasets* composto por 11 características que as bases deveriam apresentar para que fosse considerados válidas, a saber: diversidade de ataques, anonimato, protocolos disponíveis, completude da captura, completude da interação, completude da configuração de rede, completude do tráfego, conjunto de *features*, heterogeneidade, rotulação e metadados. Adicionalmente, realizaram uma comparação entre 5 abordagens para obtenção de *datasets*, com suas vantagens e desvantagens: *replay* de tráfego de *datasets* públicos, geração artificial de tráfego, desenvolvimento de estratégias de arquitetura para geração de tráfego, simulação e representação física de uma rede real. Com base em seu *framework*, eles avaliaram 11 *datasets*, com *scores* que variam entre 0 e 1, porém não geraram nenhum *dataset* capaz de

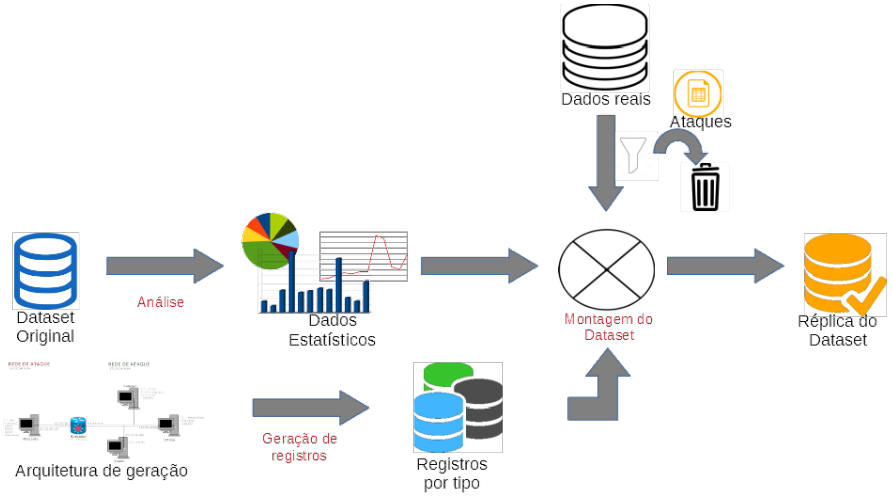

Figura 1. Estrutura interna de um *dataset*.

se adequar ao *framework* proposto.

Na tentativa de garantir a validade e o realismo de um *dataset* artificial [Haider et al. 2017] desenvolveram uma abordagem para avaliação de *datasets* mediante uma modelagem baseada em lógica *fuzzy*, o que lhe permitiu comparar alguns *datasets* em abordagem semi-quantitativa e gerar um novo conjunto de dados, o NGIDS-DS. Com intenções semelhantes, tendo por base os onze critérios de [Gharib et al. 2016], [Sharafaldin et al. 2017] geraram uma nova base de dados valendo-se da abstração do comportamento das interações humanas e simulação de cenários de ataque, o CIC-IDS. Muito embora tenham obtido bons resultados, tais abordagens não viabilizam a comparação de trabalhos voltados ao desenvolvimento de IDS obtidos anteriormente, dificultando a percepção dos avanços alcançados.

Embora existam diversos métodos para geração de *datasets*, a justa comparação entre sistemas de detecção de intrusão antigos e contemporâneos é prejudicada pela evolução do tráfego das redes de computadores. Assim, o presente trabalho oportuniza tanto a geração de *datasets* atuais quanto a extração de subconjuntos com características comparáveis às de *datasets* utilizados em avaliações antigas.

3. Método para Geração de *Workload*

Em linhas gerais, em nosso método focamos em características de datasets baseados em pacotes que pudessem ser reproduzidas, sob o ponto de vista estatístico. Cientes, contudo, das dificuldades associadas à análise temporal, optamos por não realizar tal análise. Assim, de maneira resumida, o método foi desenvolvido de acordo com as seguintes etapas, que serão abordadas nas próximas subseções: análise do *dataset*, visando à identificação de características estatísticas reprodutíveis, seguida da geração do esqueleto do novo *dataset*; obtenção de pacotes para preenchimento do esqueleto, realizada mediante a geração de tráfego maligno e a coleta de tráfego benigno; e montagem do novo *dataset*. A Figura 1 ilustra as etapas realizadas.

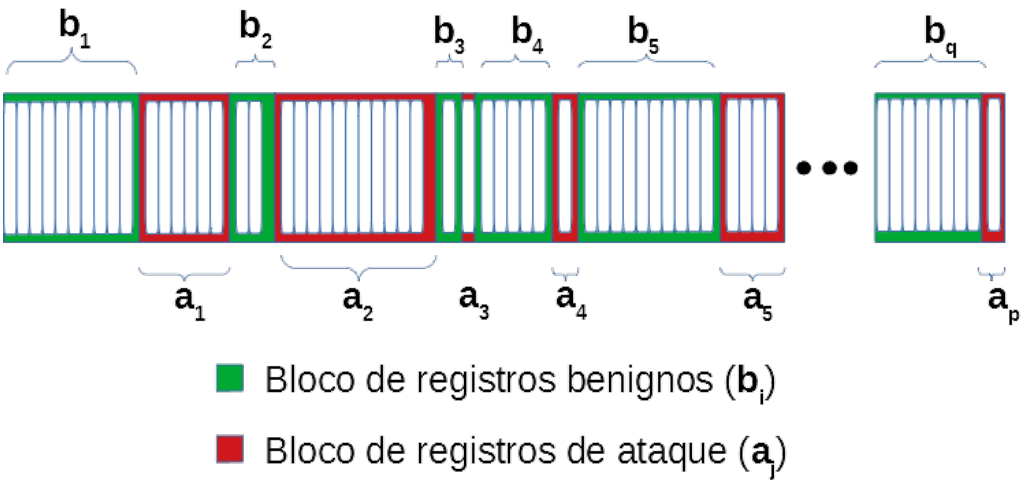

Figura 2. Estrutura interna de um *dataset*

3.1. Análise Estatística do *Dataset*

Ao olhar para um *dataset* baseado em pacotes, observamos que sua estrutura se apresenta em blocos, cada bloco com k registros com mesmo rótulo (benigno/maligno), $k \geq 1$. Assim a cada vez que um registro de um tipo distinto do anterior surge, conforme a sequência de pacotes no *dataset*, considerado o rótulo, inicia-se um novo bloco, conforme podemos ver na Figura 2. Tal constatação nos permitiu identificar como possível variável a ser analisada o tamanho de cada bloco dentro do universo de blocos com pacotes com mesmo rótulo, ou seja, o número de registros consecutivos com mesmo rótulo até que surja um registro com rótulo distinto. Identificamos, ainda, outras possíveis características passíveis de análise, como os percentuais correspondentes aos pacotes de cada serviço ou mesmo os associados aos diferentes protocolos nas diferentes camadas. Verificamos, entretanto, que todas a adoção dessas outras características para a finalidade de referência para a replicação demandaria uma análise adicional vinculada a outras características, o que, neste trabalho, nos levou a focar apenas no tamanho do bloco.

Considerando dois conjuntos distintos de blocos, a saber blocos compostos por pacotes benignos e blocos compostos por pacotes malignos, a análise em cada conjunto da variável tamanho do bloco nos levou a uma distribuição de frequências, cuja submissão ao processo de *fitting*, conforme o gráfico de Cullen e Frey, nos permitiu enquadrá-la em uma função de distribuição de probabilidade, a qual serviu de base para replicação. O ambiente R, utilizado por nós para essa finalidade, apresenta funções adequadas a tal tarefa, como exemplifica a Figura 3. Assim, uma vez definida a função para replicação da distribuição de frequência das variáveis tamanho do bloco com pacotes benignos e tamanho do bloco com pacotes malignos, passamos a montar um esqueleto para a recepção de novos pacotes, ou seja, definimos os tamanhos de cada bloco, sem populá-los com novos pacotes. Para executar tal tarefa, foi necessário definirmos o número de blocos a ser gerado $2h$ (h blocos com registros benignos e h blocos com registros malignos, o que foi feito em função do número n de registros (pacotes) do *dataset* original e das médias das funções de distribuição, μ_a no caso do tamanho dos blocos de ataque e μ_b no caso do tamanho dos blocos benignos, conforme a Equação 1.

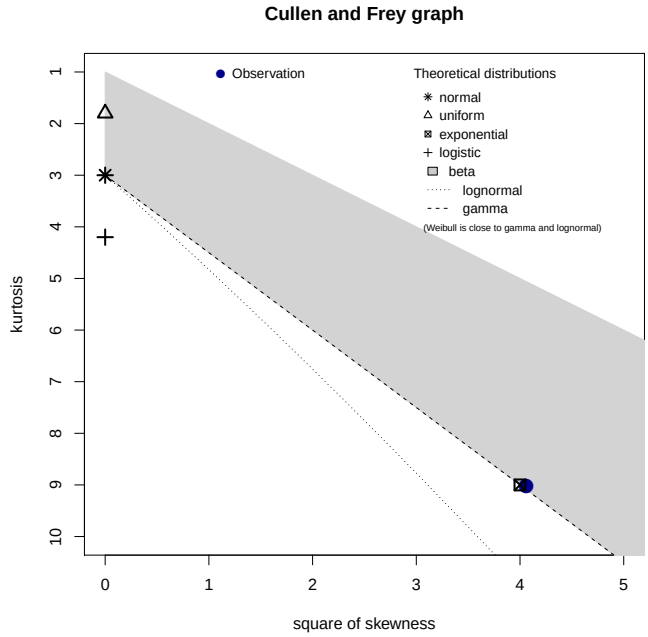

Figura 3. Seleção no R da distribuição associada ao tamanho dos blocos benignos com base na assimetria (*skewness*) e curtose (*Kurtosis*). (Fonte: Autor)

$$h = \frac{n}{\mu_b + \mu_a} \tag{1}$$

Uma vez montado o esqueleto do novo *dataset*, passamos, então à obtenção de novos pacotes para populá-lo.

3.2. Obtenção de pacotes e Montagem do *dataset*

Para obter os dados para preencher o esqueleto criado conforme a Seção 3.1, dividimos nosso trabalho em duas etapas: geração de tráfego malicioso e coleta de tráfego benigno. A geração do tráfego malicioso, considerando a intenção de replicar os tipos de ataques sem focar nos aspectos temporais, se baseou em um *testbed* simples, composto por duas redes, uma denominada rede de ataque e a outra denominada rede de defesa, ambas segregadas entre si por um roteador baseado no sistema operacional FreeBSD, conforme a Figura 4. Para gerar os ataques, utilizamos o sistema operacional Kali Linux, criado especificamente com a finalidade de realizar testes de penetração. Para o papel de *host* vítima, selecionamos o sistema Metasploitable 2, concebido a partir do sistema operacional Ubuntu, com vulnerabilidades conhecidas e intencionais. Inserimos, também *hosts* denominados zumbis, os quais poderiam ser utilizados pelos atacantes como pontes para a realização de ataques e, por fim, coletamos o tráfego gerado por meio de um *host* com sua placa de rede configurada em modo promíscuo, com o sistema Ubuntu e a ferramenta TShark, vinculada à ferramenta de captura de pacotes Wireshark.

Com relação ao tráfego de benigno, cientes de que sua geração não costuma ser intencionalmente planejada, o que dificulta sua replicação em ambiente de laboratório, adotamos como linha de ação a busca de dados em fontes reais com nível de confiabilidade

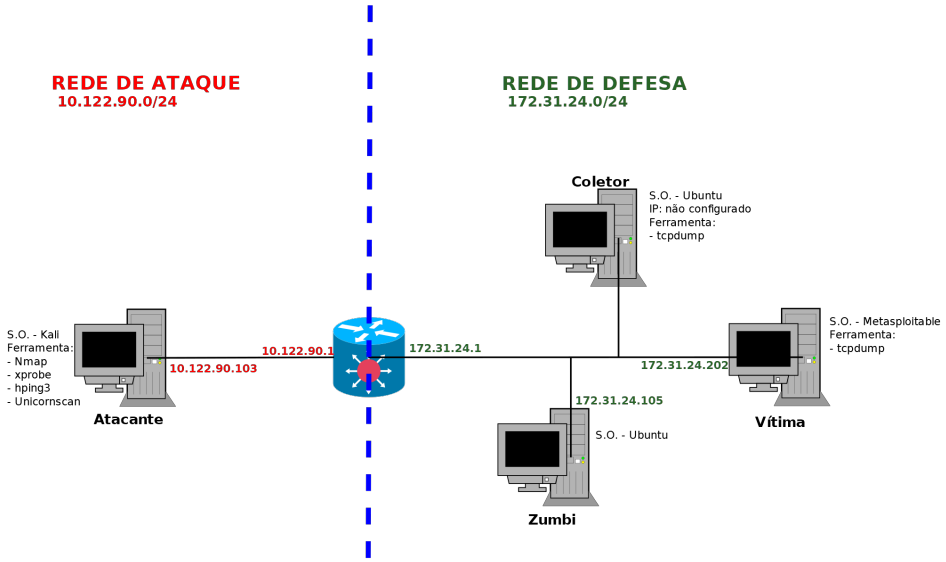

Figura 4. *Testbed* para geração dos ataques do *dataset*.

julgado aceitável. Assim, para complementar o tráfego malicioso gerado, nos valemos dos dados disponibilizados pelo Mawilab [Fontugne et al. 2010] composto por pacotes de múltiplos protocolos e serviços. De uma forma simplificada, o Mawilab disponibiliza dados reais coletados em um *link* transpacífico, acrescidos de um arquivo com os resultados de heurísticas de detecção aplicadas que possibilitam a identificação dos ataques presentes nos dados originais. Assim, utilizando os arquivos que identificam os ataques como filtro, conseguimos obter tráfego genérico, benigno e, consideravelmente, confiável quanto à ausência de possíveis ataques.

Por fim, a última etapa realizada foi a população do esqueleto do *dataset*, preenchendo cada um dos blocos, tanto os de dados malignos quanto os de benignos, respectivamente, com os pacotes maliciosos gerados conforme o *testbed* e os pacotes benignos filtrados a partir daqueles coletados do Mawilab, os quais estão dispostos em uma estrutura de fila. Tal composição garante que, ao final do processo, obtenhamos um novo conjunto de dados, com pacotes completamente distintos dos originais, mantendo, contudo a mesma característica estatística relativa à quantidade de pacotes benignos e malignos contíguos.

4. Estudo de Caso

Com o intuito de demonstrar na prática a aplicação do método de geração de *datasets*, selecionamos o *dataset* desenvolvido no trabalho de [Barbieri 2019], voltado à avaliação do método determinístico por ele proposto para detecção de ataques de TCP *scanning*. Por ser baseado em pacotes e não em fluxos e por sua simplicidade, contendo ataques de um único tipo, julgamos adequado utilizar esse *dataset* como entrada para o nosso método. Segundo [Barbieri 2019], a motivação para a geração de uma nova base de dados foi a ausência de *datasets* que contivessem ataques específicos e identificados de TCP *scanning*, responsáveis por $2,1\,\%$ dos pacotes do tráfego global anual, que, até o final de 2020, pode chegar ao volume anual de $2,3$ Zettabyte [Glatz and Dimitropoulos 2012, Idzikowski et al. 2018, Lee and Lee 2012]. Quanto à composição, o *dataset* apresentado por Barbieri continha $1.603.761$ registros, sendo $1.239.293$ pacotes benignos

Tabela 1. Comparação entre os *datasets* de Barbieri e nossa réplica

Característica	Barbieri	Réplica	Diferença
Total de registros	1.603.761	1603761	—
registros de ataque	364.468	334.660	$-8,18\%$
registros benignos	1.239.293	1.269.101	$2,41\%$
blocos de ataque	281.662	268.637	$4,98\%$
blocos benignos	281.662	268.637	$4,98\%$

e 364.468 pacotes maliciosos. O principal ponto a se destacar no conjunto original apresentado é a forma como os ataques foram obtidos, a saber, mediante a utilização de múltiplas técnicas e *scripts* disponíveis na ferramenta Nmap.

Considerando o conjunto original de dados, identificamos 281.663 blocos com registros normais ou benignos e a mesma quantidade de blocos com pacotes de ataque ou malignos. Submetendo as distribuições de frequência das variáveis tamanho dos blocos ao processo de *fitting*, primeira etapa de nosso método de geração, nos 281.663 blocos com pacotes malignos identificamos uma função Beta, enquanto para a variável tamanho dos blocos relativos ao tráfego normal, com valores coletados nos 281.663 blocos benignos identificamos uma função Gamma, mais especificamente uma função Exponencial. As médias das funções de distribuição identificadas foram, respectivamente, $\mu_a = 1,57$ e $\mu_b = 4,40$, de maneira que, por meio da Equação 1, apresentada na seção 3, aproximadamente, definimos o número de blocos $2h = 537.274$, correspondente a 268.637 blocos com pacotes benignos e 268.637 blocos com pacotes referentes a ataques de TCP *scanning*.

A geração de ataques foi realizada segundo o *testbed* apresentado na Figura 4 utilizando as seguintes ferramentas: Nmap, Angry IP Scan, Dmitry, Hping3, Masscan, Netcat e Unicornscan. No caso do Nmap, foram utilizadas as seguintes técnicas: *ack scan*, *connect scan*, *custom scan*, *decoy scan*, *fin scan*, *fragmented connect scan*, *maimon scan*, *protocol scan*, *stealth syn scan*, *windows scan* e *xmas scan*. Ainda nessa etapa, o tráfego malicioso foi complementado pelo tráfego benigno coletado a partir do Mawilab submetido ao filtro para expurgar os ataques, conforme já explicado na subseção 3.2. Os pacotes de tráfego normal e de ataque obtidos foram utilizados para mobiliar os blocos, conforme a última etapa, chegando a um conjunto final com $1.269.101$ registros benignos e 334.660 registros de ataque do tipo TCP *scanning*. A Tabela 1 compara características gerais do *dataset* original e do *dataset* que geramos, denotando a semalhança entre eles. Complementarmente, aplicamos os critérios de [Gharib et al. 2016] para avaliação da validade de um *dataset*, tendo obtido para ambos os *datasets* o score $0,68$ o que reforça a similaridade entre eles, desta vez sob o ponto de vista da validade.

Repetindo os experimentos realizados no trabalho de [Barbieri 2019], cuja comparação está disposta na Tabela 2, a despeito da similaridade estatística entre a distribuição do tamanho dos blocos, observamos uma forte queda no desempenho do método de detecção proposto, com o *True Positive Rate* passando de 89% para 55%. Tal discrepância, em nossa análise, pode ser um indício de que, talvez, os resultados originais tenham sido enviesados em consequência da ferramenta utilizada, dado que essa foi a principal alteração nos dois conjuntos de dados utilizados para avaliação, sendo o método de detecção baseado em

pacotes e *stateless*.

Tabela 2. Comparação do desempenho do método de detecção proposto por Barbieri frente ao *dataset* original e o que geramos

Dataset utilizado	CC[1] (%)	TPR[2] (%)	TNR[3] (%)
Dataset original	97,59%	89,39%	100,00%
Nossa réplica	82,11%	55,24%	100,00%

[1]CC: Acurácia [2]TPR: *True Positive Rate*. [3]TNR: *True Negative Rate*

Apesar da diferença dos resultados obtidos na avaliação do método de detecção, verificamos utilidade da geração de uma nova base de dados semelhante à original, uma vez que viabiliza a reavaliação e a comparação com novos trabalhos.

5. Conclusões e Trabalhos Futuros

Sistemas de detecção desenvolvidos ou adquiridos não podem ter sua eficácia garantia senão pela submissão a testes e avaliações, o que pressupõe a existência de bases de dados capazes de oferecer o desafio necessário à avaliação. Assim, o presente trabalho teve como objetivo principal desenvolver um método capaz de gerar novos *datasets* a partir da identificação e replicação de características estatísticas presentes em *datasets* já existentes.

A replicação do *dataset* de Barbieri possibilitou verificar que é possível reproduzir as características de um *dataset* para gerar um conjunto de dados distinto, mas que preserva semelhanças estatísticas. Desta forma, pesquisas que poderiam ter a avaliação de seu desempenho prejudicado em virtude da utilização de *datasets* obsoletos, podem obter resultados mais realistas e significativos, comparáveis aos resultados de outras pesquisas.

Além disso, a diversidade conseguida com a construção de novos *datasets* favorece a avaliação de sistemas de detecção de intrusão. O experimento realizado demonstrou que a replicação de um *dataset*, apesar de mantidas algumas de suas propriedades estatísticas, pode alterar fortemente os resultados da avaliação de sistemas de detecção pela inserção de novos tipos de tráfego ou pela exclusão de tipos presentes no *dataset* de origem. Tal constatação aponta para uma aplicação adicional não prevista da replicação, a saber, a possível correção de eventuais falhas na avaliação de um *dataset* gerado artificialmente e com limitação no emprego de ferramentas de ataque.

Por outro lado, a grande diversidade de características existente entre os *datasets* identificados é desfavorável à extração de características comuns e, consequentemente, à sua generalização. Alguns trabalhos relacionados apresentam critérios bem definidos para a avaliação de *datasets*, os quais foram úteis para a identificação das caraterísticas mais significativas para o processo de generalização. Com efeito, o método proposto ficou restrito à reprodução de *datasets*. Não há qualquer impedimento à sua utilização para a extensão de bases já existentes, acrescentando novos dados ou mesmo alterando algumas das características estatísticas identificadas. Tal trabalho, visando à criação de novas bases em contextos específicos, pode ser objeto de novas pesquisas.

Como trabalhos futuros, pretendemos realizar a análise de outros *datasets* que contemplam mais tipos de ataques. Outro ponto a ser abordado é a construção de *testbed* realístico, que permita experimentação com configuração mais flexível, sendo possível

gerar tráfego benigno sintético e mesclá-lo com o tráfego maligno. Uma especialização deste método pode ser focada em Internet das Coisas—IoT. Ademais, a análise temporal também será objeto de futuros trabalhos.

Agradecimentos

Agradecemos Exército Brasileiro e ao Comando de Defesa Cibernética, ITA/PPG em Aplicações Operacionais pelo apoio financeiro.

Referências

[Barbieri 2019] Barbieri, S. (2019). Método para a detecção de pacotes produzidos por *scanning* tcp. Mestrado em engenharia eletrônica e computação, Instituto Tecnológico de Aeronáutica, São José dos Campos, SP.

[Fontugne et al. 2010] Fontugne, R., Borgnat, P., Abry, P., and Fukuda, K. (2010). Mawilab: Combining diverse anomaly detectors for automated anomaly labeling and performance benchmarking. In *Proceedings of the 6th International COnference*, Co-NEXT '10, pages 8:1–8:12, New York, NY, USA. ACM.

[Gharib et al. 2016] Gharib, A., Sharafaldin, I., Lashkari, A. H., and Ghorbani, A. A. (2016). An evaluation framework for intrusion detection dataset. In *2016 International Conference on Information Science and Security (ICISS)*, pages 1–6.

[Glatz and Dimitropoulos 2012] Glatz, E. and Dimitropoulos, X. (2012). Classifying internet one-way traffic. In *Proceedings of the 2012 Internet Measurement Conference*, IMC'12.

[Haider et al. 2017] Haider, W., Hu, J., Slay, J., Turnbull, B., and Xie, Y. (2017). Generating realistic intrusion detection system dataset based on fuzzy qualitative modeling. *J. Netw. Comput. Appl.*, 87(C):185–192.

[Idzikowski et al. 2018] Idzikowski, F., Chiaraviglio, L., Liu, W., and van de Beek, J. (2018). Future internet architectures and sustainability: An overview. In *2018 IEEE International Conference on Environmental Engineering (EE)*, pages 1–5.

[Lee and Lee 2012] Lee, Y. and Lee, Y. (2012). Toward scalable internet traffic measurement and analysis with hadoop. *SIGCOMM Comput. Commun. Rev.*, 43(1):5–13.

[Lippmann et al. 2000] Lippmann, R., Haines, J. W., Fried, D. J., Korba, J., and Das, K. (2000). The 1999 darpa off-line intrusion detection evaluation. *Comput. Netw.*, 34(4):579–595.

[McHugh 2000] McHugh, J. (2000). Testing intrusion detection systems: A critique of the 1998 and 1999 darpa intrusion detection system evaluations as performed by lincoln laboratory. *ACM Trans. Inf. Syst. Secur.*, 3(4):262–294.

[Nehinbe 2011] Nehinbe, J. O. (2011). A critical evaluation of datasets for investigating idss and ipss researches. In *2011 IEEE 10th International Conference on Cybernetic Intelligent Systems (CIS)*, pages 92–97.

[Rao and Naidu 2017] Rao, C. M. and Naidu, M. M. (2017). A model for generating synthetic network flows and accuracy index for evaluation of anomaly network intrusion detection systems. *Indian Journal of Science and Technology*, 10(14).

[Scott and Wilkins 1999] Scott, P. D. and Wilkins, E. (1999). Evaluating data mining procedures: techniques for generating artificial data sets. *Information & Software Technology*, 41(9):579–587.

[Sharafaldin et al. 2017] Sharafaldin, I., Gharib, A., Lashkari, A. H., and Ghorbani, A. A. (2017). Towards a reliable intrusion detection benchmark dataset. *Software Networking*, 2017(1):177–200.

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

[Sommer and Paxson 2010] Sommer, R. and Paxson, V. (2010). Outside the closed world: On using machine learning for network intrusion detection. In *2010 IEEE Symposium on Security and Privacy*, pages 305–316.

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

[Zuech et al. 2015] Zuech, R., Khoshgoftaar, T. M., Seliya, N., Najafabadi, M. M., and Kemp, C. (2015). A new intrusion detection benchmarking system. In *The Twenty-Eighth International Flairs Conference*.