

Avaliação da Qualidade de Dados Sintéticos em Cenários de Ameaças Persistentes Avançadas

Murilo H. Block¹, Rafael C. Pregardier¹, Jaime A. Daniel Filho¹,
Carlos R. P. dos Santos¹, Luis A. L. Silva¹

¹Curso de Ciência da Computação – Universidade Federal de Santa Maria (UFSM)
Santa Maria – RS – Brasil

{mhblock, rcpregardier, jafilho, csantos, luisalvaro}@inf.ufsm.br

Abstract. *Advanced Persistent Threat (APT) detection requires representative data, but operational records are scarce and may expose sensitive information. This work presents a quality assessment protocol combining statistical fidelity and network-flow domain validity. Its main contribution is a complementary diagnostic procedure with explicit rules and aggregation criteria. Applied to CTGAN, TimeGAN, and TWCGAN datasets using Unraveled as reference, the protocol identified TWCGAN as having the highest domain validity (0.826) and TimeGAN as having the highest statistical fidelity (0.840). Their different rankings and violations of discrete-count constraints show why statistical similarity alone does not establish suitability for detector training. Detection utility and temporal preservation remain to be evaluated.*

Resumo. *A detecção de Ameaças Persistentes Avançadas (APTs) requer dados representativos, mas registros operacionais são escassos e podem expor informações sensíveis. Este trabalho apresenta um protocolo de avaliação que combina fidelidade estatística e validade de domínio dos fluxos de rede. A contribuição principal é um procedimento de diagnóstico complementar, com regras e critérios de agregação explícitos. Aplicado aos conjuntos de CTGAN, TimeGAN e TWCGAN, com Unraveled como referência, o protocolo identificou maior validade de domínio na TWCGAN (0,826) e maior fidelidade estatística na TimeGAN (0,840). Os rankings distintos e as violações de contagens discretas mostram por que a semelhança estatística, isoladamente, não estabelece a adequação ao treinamento de detectores. A utilidade na detecção e a preservação temporal ainda precisam ser avaliadas.*

1. Introdução

Ameaças Persistentes Avançadas (*Advanced Persistent Threats* – APTs) são ataques direcionados, furtivos e duradouros, conduzidos em múltiplos estágios. Sua detecção exige representar comportamentos e relações entre eventos [Zhang et al. 2025]. Sistemas de Detecção de Intrusão (*Intrusion Detection Systems* – IDSs) baseados em aprendizado de máquina dependem de dados representativos, cuja obtenção é dificultada pela escassez de campanhas documentadas e pela sensibilidade dos registros operacionais [Sommer e Paxson 2010; Ring et al. 2019]. Redes Adversárias Generativas (*Generative Adversarial Networks* – GANs) aprendem a distribuição dos dados por meio da competição entre

gerador e discriminador [Goodfellow et al. 2014]. Sua vantagem potencial neste contexto é ampliar a disponibilidade de amostras preservando relações entre atributos, em vez de apenas duplicar registros. A CTGAN aborda dados tabulares com atributos mistos e categorias desbalanceadas [Xu et al. 2019]; a TimeGAN busca preservar dependências temporais [Yoon et al. 2019]. Essa motivação não assegura que os registros produzidos sejam válidos, diversos ou úteis para um IDS [Theis et al. 2016; Borji 2019].

Um sintético pode reproduzir médias e correlações, mas conter contagens fracionárias, flags incompatíveis com a quantidade de pacotes ou relações inválidas entre mínimos e máximos. Esses defeitos podem introduzir padrões impossíveis no tráfego representado. A questão investigada é, portanto, se os conjuntos que melhor reproduzem as características estatísticas também respeitam as restrições do domínio. A contribuição principal é um protocolo que combina duas dimensões de qualidade, com regras verificáveis, pesos explícitos e parâmetros de amostragem. Sua aplicação a CTGAN, TimeGAN e TWCGAN ilustra como localizar limitações antes do treinamento de detectores. O protocolo pode avaliar saídas de outros geradores; as conclusões experimentais referem-se aos conjuntos disponíveis.

2. Trabalhos relacionados

Myneni et al. [2023] apresentam o Unraveled, um dataset semissintético que reúne tráfego de usuários emulados e ataques executados em um ambiente de rede controlado. Seus rótulos permitem estudar estágios de APT, mas o uso de uma referência controlada limita a extrapolação para redes operacionais [Sommer e Paxson 2010; Ring et al. 2019]. Xu et al. [2019] avaliam geração tabular por CTGAN, enquanto Yoon et al. [2019] propõem a TimeGAN para geração de séries temporais. Esses trabalhos motivam a análise de distribuições e dependências, mas não substituem a verificação das restrições específicas dos fluxos avaliados. O SynthEval reúne métricas para avaliar fidelidade, utilidade e privacidade de dados tabulares [Lautrup et al. 2025]. A MMD, por sua vez, compara distribuições por representações em espaços de características definidos por kernels [Gretton et al. 2012]. Este trabalho utiliza métricas estatísticas existentes e as articula com regras de validade dos fluxos. A contribuição está nessa operacionalização para o diagnóstico de dados de APT; não são propostas novas métricas nem uma nova arquitetura geradora.

3. Metodologia

3.1. Dados e preparação

Foram avaliados fluxos derivados do Unraveled [Myneni et al. 2023] e os conjuntos sintéticos identificados como CTGAN, TimeGAN e TWCGAN. O recorte contém *Benign*, *Reconnaissance*, *Establish Foothold*, *Lateral Movement* e *Data Exfiltration*. A avaliação utiliza os registros disponíveis, sem novo treinamento ou ajuste dos geradores. Assim, compara esses conjuntos específicos, sem isolar o efeito da arquitetura ou de seus hiperparâmetros. Foram removidos registros de *padding* e, na referência, identificadores repetidos por sobreposição de janelas. Identificadores de janela, posição e preenchimento foram excluídos das características. Utilizaram-se as 20 colunas comuns: *Stage* e 19 atributos numéricos *src2dst*, relativos a duração, pacotes, bytes, oito flags TCP e mínimo, média, máximo e desvio padrão dos tamanhos e dos intervalos entre pacotes. A validade utiliza todos os fluxos preparados; a fidelidade utiliza amostras sem reposição.

3.2. Fidelidade estatística

O SynthEval 1.7.2 foi configurado com nove famílias: médias, sobreposição de faixas de dispersão, KS/TVD, Hellinger, quantis, correlação, informação mútua, PCA e MMD. As primeiras medidas examinam aspectos das distribuições individuais; correlação e informação mútua examinam relações entre atributos; PCA e MMD avaliam aspectos da estrutura multivariada. Foram amostrados 10.000 fluxos por conjunto, com sementes 42 para a referência e 43, 44 e 45 para CTGAN, TimeGAN e TWCGAN, respectivamente. Utilizaram-se dez quantis, nível de significância de 5% e 200 permutações no KS/TVD. A sobreposição foi calculada com faixas baseadas no desvio padrão; a PCA utilizou padronização e apenas atributos numéricos. Três flags constantes na referência foram excluídas somente da correlação. A MMD utilizou 2.000 fluxos por conjunto, pelas mesmas sementes, kernel RBF com $\gamma = 1/19$ e apenas atributos numéricos, devido ao custo quadrático do cálculo.

Seja $\mathcal{M}$ o conjunto de famílias. Para cada família m , $S_{d,m}$ é a média das saídas normalizadas pelo SynthEval para o conjunto d , em $[0, 1]$, com valores maiores indicando maior fidelidade. A síntese é:

$$F_d = \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} S_{d,m}. \quad (1)$$

Neste experimento, $|\mathcal{M}| = 9$. A média dentro de cada família impede que famílias com mais saídas recebam maior peso. A média entre famílias resume diferentes aspectos estatísticos, mas pode compensar resultados baixos com altos; por isso, F_d deve ser lido junto às pontuações individuais [Theis et al. 2016; Borji 2019].

3.3. Validade de domínio e aplicação do protocolo

Foram definidas quatro propriedades: (i) mínimo $\leq$ média $\leq$ máximo para tamanhos e intervalos entre pacotes; (ii) contagens de cada flag TCP entre zero e o total de pacotes; (iii) durações, contagens, tamanhos e intervalos finitos e não negativos; e (iv) contagens inteiras de pacotes, bytes e flags, com tolerância 10^{-3} à distância do inteiro mais próximo. Sejam $\mathcal{S}$ as classes, $\mathcal{P}$ as propriedades, $\mathcal{R}_p$ as regras da propriedade p , $N_{d,s}$ o número de fluxos de d na classe s e $\mathbf{x}_{d,s,i}$ o fluxo i . Define-se:

$$V_d = \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} \frac{1}{|\mathcal{R}_p|} \sum_{r \in \mathcal{R}_p} \frac{1}{N_{d,s}} \sum_{i=1}^{N_{d,s}} \mathbf{1}[r(\mathbf{x}_{d,s,i})]. \quad (2)$$

O indicador vale 1 quando o fluxo satisfaz a regra e 0 caso contrário. A equação calcula primeiro a proporção de conformidade de cada regra em uma classe, depois as médias entre regras, propriedades e classes. Isso evita que propriedades com mais regras ou classes mais frequentes dominem o resultado. Neste recorte, $|\mathcal{S}| = 5$ e $|\mathcal{P}| = 4$. V_d mede conformidade média, não a fração de fluxos que satisfazem simultaneamente todas as regras. Para reproduzir a avaliação, devem-se alinhar as colunas e aplicar a limpeza descrita, calcular as regras por classe, amostrar os fluxos com os parâmetros indicados, executar as famílias estatísticas e agregar pelas Equações 1 e 2. Em outros contextos, substituem-se as regras e classes pelas restrições pertinentes, preservando as saídas por família e classe. Propriedades não observáveis devem ser registradas como não avaliadas. Os dois escores permanecem separados para evidenciar suas diferenças.

4. Resultados e discussão

Dos 50.015 registros sem *padding* da referência, restaram 37.227 fluxos após a deduplicação. CTGAN e TimeGAN contêm 48.000 fluxos cada, e TWCGAN contém 160.000. A Tabela 1 apresenta as médias macro das propriedades nas cinco classes.

Tabela 1. Validade de domínio: conformidade média por propriedade e global.

Conjunto	Ordenação	Flags	Não negativos	Inteiros	V_d
Unraveled	1,000	1,000	1,000	1,000	1,000
CTGAN	0,840	0,759	0,872	0,539	0,753
TimeGAN	0,908	1,000	1,000	0,362	0,817
TWCGAN	0,910	0,946	1,000	0,450	0,826

A referência satisfaz todas as regras avaliadas. A TWCGAN obteve maior V_d , mas sua vantagem de 0,009 sobre a TimeGAN é descritiva: não foram estimados intervalos de confiança para esse ranking. A integralidade foi a principal limitação dos três sintéticos. Na TimeGAN, a conformidade média de 0,362 corresponde a 63,8% de violações nas verificações de contagens inteiras, com os pesos do protocolo; não significa que essa porcentagem de fluxos distintos seja inválida. Contagens fracionárias são incompatíveis com a interpretação física de pacotes, bytes e flags. A CTGAN apresentou conformidade de 0,759 nas flags e de 0,872 na propriedade de valores finitos e não negativos. Esses resultados identificam aspectos concretos a corrigir antes de interpretar os registros como tráfego plausível. Eventuais restrições no gerador ou correções posteriores devem ser reavaliadas, pois arredondar contagens ou limitar valores pode alterar distribuições e relações entre atributos.

Tabela 2. Fidelidade das distribuições individuais.

Conjunto	Médias	Dispersão	KS/TVD	Hellinger	Quantis
CTGAN	0,986	0,713	0,318	0,987	0,983
TimeGAN	0,982	0,707	0,351	0,986	0,953
TWCGAN	0,972	0,735	0,295	0,987	0,952

Tabela 3. Fidelidade das dependências e da estrutura multivariada.

Conjunto	Correlação	Info. mútua	PCA	MMD	F_d
CTGAN	0,972	0,928	0,755	0,195	0,760
TimeGAN	0,955	0,953	0,774	0,902	0,840
TWCGAN	0,963	0,959	0,751	0,570	0,798

As Tabelas 2 e 3 mostram maior F_d na TimeGAN, sobretudo pela MMD. A CTGAN preservou melhor médias, correlação e quantis, mas sua MMD de 0,195 mostra que bons resultados nessas medidas não asseguram semelhança multivariada sob o kernel adotado. A TWCGAN liderou em sobreposição das faixas de dispersão e informação mútua, sem liderar a fidelidade agregada. Portanto, a escolha de um único resumo oculta diferenças relevantes entre os conjuntos [Theis et al. 2016; Borji 2019]. No KS/TVD, foram detectadas diferenças significativas em 100%, 85% e 95% das 20 variáveis de CTGAN, TimeGAN e TWCGAN, respectivamente. Esses resultados evidenciam discrepân-

cias apesar das pontuações altas de outras famílias. Contudo, significância estatística depende do tamanho amostral e não quantifica, isoladamente, o impacto dessas diferenças em um IDS [Wasserstein e Lazar 2016]. A menor sobreposição de dispersão da TimeGAN recomenda investigar a cobertura dos comportamentos, mas não comprova perda de diversidade nem *mode collapse*; essa hipótese exige diagnósticos adicionais [Borji 2019].

A implicação central é que a liderança estatística da TimeGAN convive com violações de contagens discretas, enquanto a maior validade da TWCGAN convive com discrepâncias estatísticas. Um IDS pode aprender relações incompatíveis com o tráfego representado ou receber cobertura insuficiente de comportamentos de APT. O protocolo permite localizar essas limitações e orientar melhorias, mas não determina qual conjunto produzirá a melhor detecção. Isso requer treinamento e teste em dados de referência reservados [Sommer e Paxson 2010; Theis et al. 2016]. O estudo limita-se a um recorte semissintético, uma amostragem por conjunto e regras observáveis nos atributos comuns. Não avalia privacidade, preservação das sequências temporais nem todas as relações aritméticas entre atributos. As médias podem ocultar falhas por classe, e os pesos iguais são uma escolha explícita, não uma medida universal de qualidade. A aplicação a outra rede exige adequar as regras e verificar a utilidade no cenário-alvo [Ring et al. 2019].

5. Conclusão

Este trabalho apresentou um protocolo de diagnóstico que combina fidelidade estatística e validade de domínio, com preparação, parâmetros e agregações explícitos. A aplicação mostrou que a TWCGAN lidera a conformidade média (0,826) e a TimeGAN a fidelidade (0,840), sem que qualquer conjunto satisfaça todas as restrições avaliadas. A diferença entre os rankings sustenta a análise conjunta das duas dimensões antes do uso dos dados em IDSs. Como trabalhos futuros, serão avaliadas a preservação temporal, a estabilidade dos resultados entre amostras e a utilidade dos conjuntos na detecção de APTs.

Referências

- Borji, A. (2019). Pros and cons of GAN evaluation measures. *Computer Vision and Image Understanding*, 179:41–65. doi:10.1016/j.cviu.2018.10.009.
- Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. e Bengio, Y. (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27. <https://arxiv.org/abs/1406.2661>.
- Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B. e Smola, A. (2012). A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773. <https://www.jmlr.org/papers/v13/gretton12a.html>.
- Lautrup, A. D., Hyrup, T., Zimek, A. e Schneider-Kamp, P. (2025). SynthEval: A framework for detailed utility and privacy evaluation of tabular synthetic data. *Data Mining and Knowledge Discovery*, 39, artigo 6. doi:10.1007/s10618-024-01081-4.
- Myneni, S., Jha, K., Sabur, A., Agrawal, G., Deng, Y., Chowdhary, A. e Huang, D. (2023). Unraveled – A semi-synthetic dataset for advanced persistent threats. *Computer Networks*, 227:109688. doi:10.1016/j.comnet.2023.109688.
- Ring, M., Wunderlich, S., Scheuring, D., Landes, D. e Hotho, A. (2019). A survey

- of network-based intrusion detection data sets. *Computers & Security*, 86:147–167. doi:10.1016/j.cose.2019.06.005.
- Sommer, R. e Paxson, V. (2010). Outside the closed world: On using machine learning for network intrusion detection. *IEEE Symposium on Security and Privacy*, p. 305–316. doi:10.1109/SP.2010.25.
- Theis, L., van den Oord, A. e Bethge, M. (2016). A note on the evaluation of generative models. *International Conference on Learning Representations*. <https://arxiv.org/abs/1511.01844>.
- Wasserstein, R. L. e Lazar, N. A. (2016). The ASA statement on p-values: Context, process, and purpose. *The American Statistician*, 70(2):129–133. <https://doi.org/10.1080/00031305.2016.1154108>.
- Xu, L., Skoularidou, M., Cuesta-Infante, A. e Veeramachaneni, K. (2019). Modeling tabular data using conditional GAN. *Advances in Neural Information Processing Systems*, 32. <https://arxiv.org/abs/1907.00503>.
- Yoon, J., Jarrett, D. e van der Schaar, M. (2019). Time-series generative adversarial networks. *Advances in Neural Information Processing Systems*, 32.
- Zhang, B., Gao, Y., Kuang, B., Yu, C., Fu, A. e Susilo, W. (2025). A survey on advanced persistent threat detection: A unified framework, challenges, and countermeasures. *ACM Computing Surveys*, 57(3), artigo 62. doi:10.1145/3700749.