

Comparação entre estratégias de agregação de fluxos para a detecção de APTs com modelos Transformer

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

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

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

Abstract. *Advanced Persistent Threats (APTs) unfold over multiple stages and extended periods, producing sparse malicious activity amid legitimate traffic. This behavior hinders their detection by classifiers that analyze each flow in isolation, since contextual patterns across flows may be lost. This work compares two strategies for incorporating such context (count-based and window-based aggregation) into the classification of individual flows, using the same Transformer-based model and the Unraveled dataset. Results show that both strategies increase PR-AUC, but the reduction in false positives occurs mainly with small contexts and is most pronounced with short time windows, reaching 60%. This indicates that the temporal proximity between flows is more relevant for detection than the context size.*

Resumo. *Ameaças Persistentes Avançadas (APTs) desenvolvem-se ao longo de múltiplos estágios e períodos prolongados, produzindo atividade maliciosa esparsa em meio ao tráfego legítimo. Esse comportamento dificulta sua detecção por classificadores que analisam cada fluxo isoladamente, pois padrões contextuais entre fluxos podem ser perdidos. Este trabalho compara duas estratégias para incorporar esse contexto (agregação por contagem e por janela de tempo) à classificação de fluxos individuais, utilizando um mesmo modelo baseado em Transformer e o conjunto de dados Unraveled. Os resultados mostram que ambas as estratégias aumentam a PR-AUC, mas a redução de falsos positivos ocorre principalmente com contextos pequenos e é mais expressiva com janelas temporais curtas, chegando a 60%. Isso indica que a proximidade temporal entre os fluxos é mais relevante para a detecção do que o tamanho do contexto.*

1. Introdução

Em 2017, a empresa americana Equifax sofreu um vazamento que expôs dados de mais de 145,5 milhões de pessoas [U.S. Government Accountability Office 2018]. Os invasores exploraram uma vulnerabilidade no *framework* Apache Struts (CVE-2017-5638) e, a partir desse acesso inicial, moveram-se lateralmente até dezenas de bancos de dados, extraíndo informações em pequenas porções. Eles permanecerem assim por 76 dias sem serem detectados. Em 2020, o ataque foi atribuído a membros do exército chinês. O incidente ilustra um padrão conhecido como Ameaça Persistente Avançada (*Advanced Persistent Threat*, APT): ataques altamente direcionados, conduzidos por grupos com amplos recursos que operam de forma silenciosa e prolongada. Ao longo de

semanas ou meses, esses grupos percorrem múltiplos estágios, como reconhecimento, comprometimento inicial, movimentação lateral, exfiltração de dados e ocultação de rastros [Alshamrani et al. 2019].

Os fluxos de rede associados a uma APT são raros e se misturam a um grande volume de tráfego legítimo, o que dificulta sua detecção tanto por sistemas tradicionais baseados em assinaturas ou regras quanto por classificadores de aprendizado de máquina que analisam cada fluxo isoladamente [Myneni et al. 2023]. Em ambos os casos, padrões temporais e contextuais que caracterizam uma APT podem se perder. Nesse sentido, trabalhos recentes têm explorado a agregação de fluxos vizinhos, seja por um número fixo de fluxos, seja por uma janela de tempo fixa, para fornecer esse contexto a classificadores [Nguyen and Watabe 2023, Martin-Reizabal et al. 2026], mas ainda não foram avaliados no cenário específico de APTs nem comparados a um classificador de fluxo único. Este trabalho busca preencher essa lacuna, comparando essas duas estratégias de agregação ao fluxo único na detecção de uma APT, utilizando o conjunto de dados Unraveled [Myneni et al. 2023], sob um mesmo modelo Transformer. Os experimentos demonstram que o contexto de fluxos vizinhos pode melhorar a detecção, mas o ganho em termos de redução de alarmes falsos depende de sua proximidade temporal.

O restante do artigo está organizado da seguinte forma: a Seção 2 apresenta os conceitos e a literatura relacionados; a Seção 3, o conjunto de dados, as estratégias de agregação, o modelo e os experimentos; a Seção 4, os resultados obtidos; e a Seção 5 conclui o trabalho, apontando direções futuras e limitações.

2. Fundamentação teórica

Um fluxo agrega os pacotes trocados entre dois *hosts* que compartilham endereços, portas e protocolo, resumindo-os em atributos como duração, contagens de pacotes e bytes e estatísticas de tempo entre pacotes [Nguyen and Watabe 2023]. A maioria dos Sistemas de Detecção de Intrusão trata cada fluxo como uma amostra independente, embora fluxos quase idênticos possam ser benignos ou maliciosos conforme o que ocorre ao seu redor [Nguyen and Watabe 2023, Martin-Reizabal et al. 2026]. Nesse sentido, a arquitetura baseada em Transformer se destaca, uma vez que o seu mecanismo de atenção permite calcular a representação de cada fluxo em função de todos os outros, incorporando o contexto de fluxos vizinhos à classificação individual. Dois trabalhos recentes exploram essa arquitetura para detecção de tráfego malicioso, utilizando estratégias diferentes para definir quais fluxos vizinhos devem ser considerados. [Nguyen and Watabe 2023] agrupa N fluxos consecutivos em blocos de tamanho fixo, formando a sequência de entrada de um modelo BERT. [Martin-Reizabal et al. 2026], por outro lado, agrupa todos os fluxos ativos dentro de uma janela de tempo fixa, de modo que a atenção capture as características de comunicações simultâneas.

Ambos os trabalhos mencionados foram avaliados em conjuntos de ataques convencionais, o CIDDS-001 e o CICIoT2023. Esses conjuntos contêm ataques em rajada, como negação de serviço e varreduras de portas, que produzem fluxos maliciosos concentrados no tempo. Nessas condições, a vizinhança de um fluxo de ataque tende a ser formada por outros fluxos do mesmo ataque, o que facilita a detecção. No entanto, em uma APT, as atividades maliciosas ocorrem em baixa frequência [Alshamrani et al. 2019]. O Unraveled [Myneni et al. 2023] foi construído para criar um conjunto de dados represen-

tativo para esse padrão de ataque, simulando uma rede corporativa com três departamentos e uma rede de produção, protegida por um IDS Snort e por uma equipe de defesa que bloqueia o que consegue detectar. 15 funcionários emulam a utilização da rede, enquanto atacantes experientes executam, ao longo de cinco semanas, uma APT que segue táticas de grupos catalogados no MITRE ATT&CK, cobrindo cinco estágios: reconhecimento (*Reconnaissance*), estabelecimento (*Establish Foothold*), movimentação lateral (*Lateral Movement*), exfiltração de dados (*Data Exfiltration*) e ocultação de rastros (*Cover Up*).

Este trabalho se diferencia ao aplicar e comparar, sob o mesmo modelo e protocolo de avaliação, estratégias de agregação de contexto inspiradas em [Nguyen and Watabe 2023] e [Martin-Reizabal et al. 2026] à detecção de uma APT completa, usando o cenário representado pelo Unraveled.

3. Metodologia

Esta seção descreve os dados, as agregações, o modelo e a política de avaliação.

3.1. Conjunto de dados, pré-processamento e divisão

Os experimentos utilizaram os 6.877.157 fluxos do Unraveled, organizados em arquivos por dia e por sub-rede. Todos os fluxos rotulados com estágio diferente de benigno foram classificados como APT. Assim como em [Martin-Reizabal et al. 2026], foram mantidos apenas os 57 atributos numéricos, e endereços, portas e protocolo foram descartados. Os instantes de tempo foram mantidos para a agregação dos fluxos, mas não foram utilizados como entrada do modelo. Sobre os atributos, foi aplicada a normalização *z-score* ajustada no conjunto de treino sem agregação.

Os fluxos foram particionados temporalmente em treino, validação e teste. Isso se deu por meio da divisão de cada dia em intervalos de 30 minutos, dos quais se escolhe dois blocos contíguos, cada um com 10% dos intervalos, para formar a validação e o teste, enquanto os 80% restantes compõem o treino. Os blocos de validação e de teste são escolhidos de modo que cada um mantenha uma proporção de fluxos APT próxima à do dia e contenha cerca de 10% dos fluxos totais. A Tabela 1 resume a divisão, mostrando que essa estratégia consegue distribuir os fluxos de forma equilibrada.

Partição	Fluxos	% APT	Rec.	Est. Foot.	Lat. Mov.	Data. Exf.	Cover Up
Treino	5.475.677	1,50	33.085	21.092	21.764	6.101	286
Validação	716.208	1,11	1.626	2.847	2.801	685	0
Teste	685.272	1,02	85	3.179	2.880	736	76

Tabela 1. Fluxos por partição.

3.2. Estratégias de agregação

As agregações são aplicadas a cada arquivo após a divisão, com os fluxos ordenados pelo início e tratados como *tokens*, de modo que nenhuma sequência mistura partições, dias ou sub-redes. As seguintes estratégias foram avaliadas:

Fluxo único (*baseline*). Cada fluxo forma uma sequência de tamanho 1.

Contagem. Os fluxos formam blocos consecutivos e sem sobreposição de N fluxos, como em [Nguyen and Watabe 2023]. Dado o tamanho fixo do contexto, a duração

temporal desses blocos flutua, ou seja, restringe-se a poucos segundos em períodos movimentados ou estende-se por vários minutos em períodos ociosos. Os valores testados foram $N \in \{8, 32, 128, 512, 2048\}$.

Janela. Seguindo a abordagem de [Martin-Reizabal et al. 2026], o tempo é dividido em janelas fixas de W segundos, e cada fluxo é copiado, sem alterações, para todas as janelas que se sobrepõem à sua duração. Como os fluxos do Unraveled chegam a 30 minutos, e não a 5 s como em [Martin-Reizabal et al. 2026], janelas de 5 s gerariam em média 11 cópias por fluxo, que encobririam o efeito do contexto. Por isso, avalia-se $W \in \{15, 30, 60, 150, 900, 1800\}$ s, sem limitar o número de fluxos por janela, pois o truncamento aproximaria as janelas longas da agregação por contagem.

A Tabela 2 resume as configurações no teste. O contexto indica quantos fluxos a atenção tipicamente considera ao classificar um fluxo, sendo calculado como a mediana, sobre todos os fluxos e suas cópias, do tamanho da sequência em que cada um está.

Configuração	Sequências	Fluxos por sequência (mediana)	Contexto	Cópias por fluxo
Fluxo único	685.272	1	1	1,00
$N = 8$	85.731	8	8	1,00
$N = 32$	21.489	32	32	1,00
$N = 128$	5.440	128	128	1,00
$N = 512$	1.412	512	512	1,00
$N = 2048$	425	2.048	2.048	1,00
$W = 15$ s	130.055	9	70	4,34
$W = 30$ s	66.475	10	91	2,67
$W = 60$ s	33.830	12	130	1,84
$W = 150$ s	13.626	19	245	1,33
$W = 900$ s	2.336	86	1.049	1,05
$W = 1800$ s	1.198	164	1.909	1,03

Tabela 2. Configurações de agregação na partição de teste.

3.3. Modelo e treinamento

O modelo reproduz a arquitetura de [Martin-Reizabal et al. 2026]: projeção linear para dimensão 64, codificação posicional senoidal e quatro camadas *encoder* do Transformer com 4 cabeças de atenção, *feed-forward* de dimensão 128 e *dropout* de 0,1. Porém, como em [Nguyen and Watabe 2023], uma camada linear classifica cada fluxo, em vez de a janela inteira por um *token* [CLS]. Assim, fluxos benignos não herdam o rótulo de um ataque simultâneo, e a comparação com fluxos isolados é possível. O treinamento usa entropia cruzada ponderada pelo inverso da frequência das classes e AdamW com taxa inicial 10^{-3} . A taxa é multiplicada por 0,5 sempre que a perda de validação não melhora por 2 épocas consecutivas (ReduceLROnPlateau). O treino dura no máximo 30 épocas, com *early stopping* após 6 épocas sem melhora do F1 de validação; os pesos avaliados são os da melhor época. Os *batches* reúnem sempre 32.768 fluxos, para que todas as estratégias usem a mesma quantidade de fluxos por atualização. Como as sequências variam de comprimento, cada *batch* é preenchido com zeros até a maior delas, e uma máscara mantém essas posições fora da atenção e da perda.

3.4. Protocolo de avaliação

Todas as configurações são avaliadas sobre os mesmos fluxos de teste. Nas janelas, cada fluxo recebe a maior probabilidade entre suas cópias, ou seja, basta que uma janela o aponte como APT para gerar um alerta. São reportados *precision*, *recall* e F1 da classe APT com limiar de 0,5 e, dado o desbalanceamento, a PR-AUC, indicada por [Myneni et al. 2023] para esse cenário. O FP@R99 complementa a análise com o menor número de falsos positivos entre os limiares que detectam ao menos 99% dos fluxos APT. Os valores são médias de 8 *seeds*, e uma diferença só é considerada quando os intervalos entre mínimo e máximo não se sobrepõem.

4. Resultados e discussão

A Tabela 3 e a Figura 1 resumem os resultados. Com limiar de 0,5, o *recall* supera 0,99 em todas as *seeds*, de forma que as diferenças recaem sobre a *precision*, isto é, nos falsos positivos. Toda agregação eleva a PR-AUC, de 0,945 para ao menos 0,982, com intervalos disjuntos, indicando que os fluxos vizinhos ajudam a separar a APT do tráfego benigno. Os falsos positivos, porém, não acompanham esse ganho, uma vez que caem apenas com contextos pequenos e voltam a crescer conforme a agregação aumenta. Na contagem, $N = 8$ e $N = 32$ reduzem o FP@R99 de 698 para 454 e 508, também com intervalos disjuntos, e $N = 8$ supera o fluxo único em F1 e *precision*. A partir de $N = 128$ a vantagem se desfaz e, com $N \geq 512$, o FP@R99 volta ao patamar do fluxo único.

Configuração	<i>Precision</i>	<i>Recall</i>	F1 (mín.–máx.)	PR-AUC (mín.–máx.)	FP@R99
Fluxo único	0,839	0,997	0,912 (0,906–0,918)	0,945 (0,936–0,964)	698
$N = 8$	0,885	0,996	0,937 (0,925–0,942)	0,982 (0,979–0,984)	454
$N = 32$	0,868	0,996	0,927 (0,911–0,943)	0,986 (0,983–0,990)	508
$N = 128$	0,853	0,996	0,919 (0,904–0,930)	0,987 (0,978–0,990)	599
$N = 512$	0,848	0,995	0,916 (0,895–0,938)	0,988 (0,984–0,994)	752
$N = 2048$	0,842	0,995	0,912 (0,889–0,936)	0,982 (0,972–0,993)	827
$W = 15$ s	0,904	0,998	0,949 (0,938–0,959)	0,995 (0,993–0,997)	269
$W = 30$ s	0,892	0,997	0,941 (0,921–0,949)	0,994 (0,992–0,996)	369
$W = 60$ s	0,863	0,998	0,925 (0,880–0,944)	0,993 (0,986–0,997)	405
$W = 150$ s	0,867	0,997	0,928 (0,918–0,937)	0,992 (0,986–0,998)	498
$W = 900$ s	0,846	0,994	0,914 (0,886–0,938)	0,986 (0,978–0,994)	790
$W = 1800$ s	0,839	0,994	0,910 (0,889–0,925)	0,985 (0,979–0,990)	804

Tabela 3. Resultados na partição de teste.

As janelas de 15 e 30 s obtêm os melhores resultados em todas as métricas. Com $W = 15$ s, a PR-AUC chega a 0,995 e o FP@R99 cai para 269, cerca de 60% abaixo do fluxo único. Com 30 e 60 s, a redução é de 47% e 42%, e essas três janelas tampouco se sobrepõem ao fluxo único. No limiar de 0,5, os falsos positivos caem 44% com 15 s. O ganho diminui à medida que W cresce e desaparece com 900 e 1.800 s.

Os resultados sugerem, portanto, que o contexto temporal é importante para a detecção de APTs. Isso fica evidente ao comparar $W = 15$ s com $N = 128$, visto que, mesmo com um contexto menor, 70 fluxos contra 128, a janela obtém uma PR-AUC superior, sem sobreposição, e reduz o FP@R99 médio de 599 para 269. A explicação provável é que uma janela curta reúne fluxos ativos no mesmo momento, tendendo a capturar uma mesma etapa do ataque, enquanto um bloco de contagem, definido pela

ordem de início, pode juntar fluxos separados por minutos. Com contextos longos, como $N = 2048$ e $W = 1800$ s, o $FP@R99$ se aproxima do fluxo único, possivelmente porque a atividade maliciosa se mistura a mais tráfego benigno.

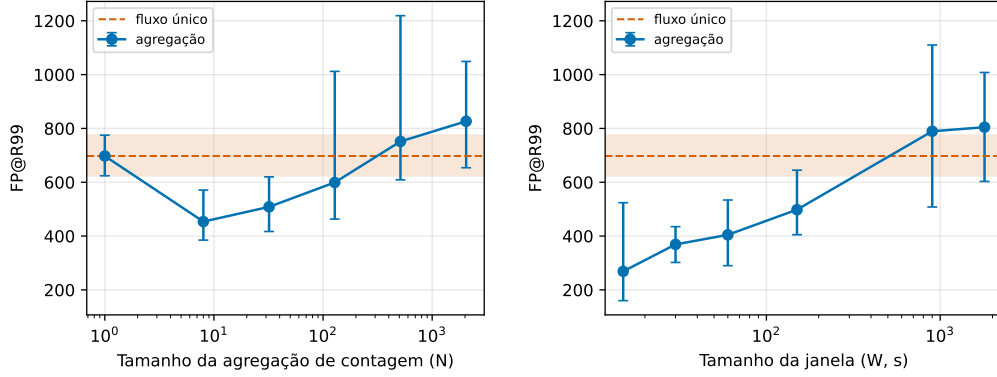


Figura 1. $FP@R99$ em função de N e W . A faixa laranja é o intervalo mín.-máx. das execuções do fluxo único.

5. Conclusão

Este trabalho comparou a agregação de fluxos por contagem e por janela de tempo à classificação de fluxos isolados na detecção de APTs. Toda agregação elevou a PR-AUC, mas a redução dos falsos positivos ocorreu apenas com contextos pequenos e foi maior nas janelas curtas, o que indica que a proximidade temporal dos fluxos importa mais que o tamanho do contexto. A janela de 15 s obteve o menor $FP@R99$, cerca de 60% abaixo do fluxo único, ao custo de 4,34 cópias por fluxo, enquanto a de 30 s preserva a maior parte do ganho com 2,67 cópias. Por reunirem poucos fluxos, essas janelas tendem a poupar recursos e a antecipar alertas, vantagens ainda não medidas. Trabalhos futuros devem quantificar esse custo e esse atraso e avaliar as estratégias em outros cenários de APT, já que os resultados se limitam a um único conjunto de dados.

Referências

- Alshamrani, A., Myneni, S., Chowdhary, A., and Huang, D. (2019). A survey on advanced persistent threats: Techniques, solutions, challenges, and research opportunities. *IEEE Communications Surveys & Tutorials*, 21(2):1851–1877.
- Martin-Reizabal, S., Caballero-Quiroga, A., Gil-Arroyo, B., Basurto, N., and Ruiz-Gonzalez, R. (2026). Transformer-based classification of iot network traffic with flow-to-window aggregation. *Internet of Things*, 36:101879.
- Myneni, S., Jha, K., Sabur, A., Agrawal, G., Deng, Y., Chowdhary, A., and Huang, D. (2023). Unraveled — a semi-synthetic dataset for advanced persistent threats. *Computer Networks*, 227:109688.
- Nguyen, L. G. and Watabe, K. (2023). A method for network intrusion detection using flow sequence and bert framework. In *ICC 2023 - IEEE International Conference on Communications*, pages 3006–3011.
- U.S. Government Accountability Office (2018). Data protection: Actions taken by equifax and federal agencies in response to the 2017 breach. Technical Report GAO-18-559, U.S. Government Accountability Office, Washington, DC.