대규모 환경에서의 Tor 트래픽 상관 분석을 위한 근사 최근접 이웃 탐색 기반 기법

홍세연*, 김혜원*, Saidur Rahman Mohammad**, 오세은***

*이화여자대학교 (대학원생), **University of Texas at El Paso, ***이화여자대학교 (교수)

Scalable Tor Traffic Correlation Attack via Approximate Nearest Neighbor Search

Saeyeon Hong*, Hyewon Kim*, Saidur Rahman Mohammad**, Se Eun Oh***

*Ewha Womans University(Graduate student), **University of Texas at El Paso, ***Ewha Womans University(Professor)

요 약

Tor 네트워크에서 종단 공격자는 트래픽 상관 분석을 통해 사용자의 통신 흐름을 역추적할 수 있다. 기존 코사인 유사도 기반 직접 비교 방식인 DeepCoFFEA(DCF) 모델은 효과적이나 플로우 수 n에 대해 $O(n^2)$의 계산 복잡도를 가지므로, 실제 Tor 네트워크 규모에서의 적용이 제한적이다. 본 논문은 실제 Tor 네트워크에서 수집한 GTT 데이터셋 분석을 통해 종단 공격자가 하루에 처리해야 하는 플로우 수가 최대 약 196,000개임을 추정하고, 이 규모에서 DCF의 한계를 실증하고, 효율성을 개선하고자 근사 최근접 이웃 탐색 알고리즘인 CAGRA를 Tor 트래픽 상관 분석에 적용하는 방법론을 제안한다. 실험 결과, 본 논문에서 제안한 기법은 DCF 대비 탐지 성능을 유지 또는 향상시키면서 실행 시간과 연산 비용을 대폭 감소시켜 대규모 실환경에서의 공격 실현 가능성을 높였다.

I. 서론

Tor는 다중 릴레이를 통한 트래픽 암호화를 기반으로 사용자의 익명성을 보장하는 대표적인 익명 통신 네트워크이다. 그러나 Tor 네트워크에서 입구(Entry) 릴레이와 출구(Exit) 릴레이를 동시에 관찰할 수 있는 종단 공격자(end-to-end adversary)는 입구-출구 트래픽 상관 분석을 통해 통신의 양종단을 추적할 수 있다.

이러한 공격을 위해 다양한 딥러닝 기반 트래픽 상관 분석 기법들이 제안되었다. DeepCorr[1]는 CNN 기반으로 트래픽 쌍을 직접 비교하여 $O(n^2)$의 DNN 연산을 요구한다. DeepCoFFEA[2]는 FEN(Flow Embedding Network)을 통해 DNN 연산을 $O(n)$으로 감소시켰음에도, 임베딩 간 전수 코사인 유사도 비교에 따른 $O(n^2)$의 시간 복잡도가 병목으로 남아 대규모 실환경에서의 실시간 적용이 현실적으로 어렵다.

이와 같은 한계는 실제 Tor 환경에서 더욱 두드러지게 나타난다. 본 연구에서는 GTT 데이터셋[3]을 활용하여 종단 공격자가 처리해야 하는 트래픽 규모를 분석하였다. GTT의 확률적 샘플링 비율을 역산하고 웹 트래픽 비율(62.4%)로 보정한 결과, Exit relay 1개 기준 하루 약 47,000개(저부하 시)에서 최대 196,000개(고부하 시)의 플로우가 발생하는 것으로 추정된다. 이러한 규모에서는 기존 $O(n^2)$기반 DeepCoFFEA는 실시간 적용이 어려운 수준의 연산 비용을 요구함을 확인할 수 있다.

따라서 본 논문에서는 이러한 계산 복잡도 문제를 해결하기 위해 근사 최근접 이웃 탐색(Approximate Nearest Neighbor, ANN) 알고리즘인 CAGRA (CUDA Accelerated Graph-based Approximate Nearest

Neighbor)[4]를 Tor 트래픽 상관 분석에 적용하는 방법을 제안한다. 제안하는 기법은 임베딩 간 전수 비교를 하지않고 선택적으로 비교함으로써 연산 효율성을 크게 향상시키면서도 탐지 성능을 유지하는 것을 목표로 한다.

본 연구의 주요 기여는 다음과 같다.

- **[복잡도]** Tor 트래픽 상관 분석 분야에서 최초로 ANN 탐색을 활용하여 유사도 비교 단계의 시간 복잡도를 $O(n^2)$에서 $O(n\log n)$으로 감소시키는 파이프라인을 제안한다.
- **[실규모 분석]** GTT 데이터셋 분석을 통해 실제 종단 공격자가 처리해야 하는 트래픽 규모를 정량적으로 제시한다.
- **[학습]** WGAN Loss (Wasserstein GAN Loss) 에 기반한 WC Loss (Wasserstein Compactness Loss)를 설계하여, ANN 탐색 환경에서의 logAUC 기준 탐지 성능을 향상시킨다.

II. 연구 배경 및 관련 연구

2.1 트래픽 상관 분석

트래픽 상관 분석은 종단 공격자가 Tor 네트워크의 입구 및 출구 트래픽을 동시에 관찰하여 동일 사용자의 플로우를 식별하는 공격 기법이다. DeepCorr[1]는 딥러닝 기반 특징 추출을 통해 기존 통계적 방법 대비 높은 정확도를 달성하였으나, $O(n^2)$의 DNN 연산을 요구한다. 이후 DeepCoFFEA[2]는 FEN을 이용하여 DNN연산을 $O(n)$으로 줄였으나, 임베딩 간 전수 코사인 유사도 비교로 인해 여전히 $O(n^2)$의 계산 복잡도를 가지므로 대규모 환경에서의 적용에 한계가 있다.

2.2 근사 최근접 이웃 탐색

CAGRA[4]는 GPU 병렬 연산을 활용한 그래프 기반 ANN 탐색 알고리즘으로, $O(n\log n)$의 시간복잡도를 보장한다. HNSW 등 CPU 기반 ANN은 GPU 환경에서의 공정한 비교에 적합하지 않으며, IVF는 전체 AUC는 우수하나, 실제 공격 시나리오에서 중요한 low-FPR 구간에서의 탐지 성능(pAUC)가 크게 저하되는 것을 확인하였다. (Table. 2) 이에 본 연구는 GPU 병렬 처리를 활용하며 low-FPR 구간에서 우수한 성능을 보이는 CAGRA를 선택하였다.

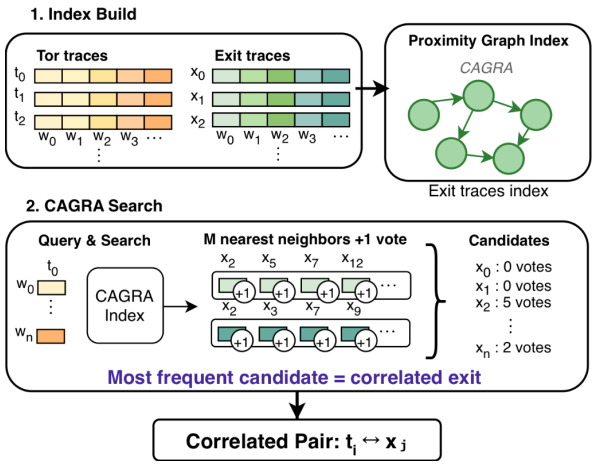

Fig. 1 Overview of the pipeline

III. ANN 기반 Tor Traffic 상관 분석

본 논문에서 제안하는 Tor 트래픽 상관 분석 파이프라인은 총 네 단계로 구성된다. 첫째, 학습 단계에서는 입구(Entry) 및 출구(Exit) 릴레이에서 수집된 플로우를 이용하여 딥 메트릭 러닝을 수행한다. 이때 트리플렛 로스(Triplet Loss)와 WC Loss를 결합하여, 상관된 플로우는 임베딩 공간에서 가깝게, 상관되지 않은 플로우는 멀어지도록 학습한다. 둘째, 학습된 네트워크를 이용하여 Entry 및 Exit 플로우를 각각 임베딩 벡터로 변환한다. 셋째, 변환된 Exit 플로우 임베딩을 ANN 인덱스에 등록하여 탐색 구조를 구성한다. 넷째, Entry 플로우를 쿼리로 사용하여 ANN 인덱스에서 유사한 임베딩을 탐색하고, 상위 M개의 근접 이웃을 반환한다. 각 이웃에는 1표씩 부여되며, 다수의 윈도우에서 얻어진 결과를 집계하여 최종적으로 상관된 플로우를 결정한다.

Fig. 1은 전체 파이프라인 중 ANN 인덱스를 구축한 후 Entry 플로우를 쿼리로 사용하여 근접 이웃을 탐색하고, 이를 기반으로 투표를 수행하여 결과를 집계하는 과정을 나타낸다.

3.1 임베딩 학습과 임베딩 생성

플로우 임베딩 학습은 DeepCoFFEA의 트윈 신경망 구조를 기반으로 수행된다. 기본적으로 Triplet Loss를 사용하여 상관된 플로우는 임베딩 공간에서 가깝게, 비상관 플로우는 멀어지도

록 학습한다. ANN 그래프 기반 탐색에서는 유사한 노드 간 엣지 형성뿐 아니라 상이한 노드 간 탐색 경로(traverse) 확보도 중요하다. 이를 위해 임베딩 공간에서 클러스터 구조가 잘 형성될수록 그래프 탐색 효율이 향상되므로, 매 epoch마다 클러스터링 결과를 손실 함수에 반영하는 Deep Clustering[5]의 접근 방식에 착안하였다. 이를 WGAN Loss[6]에 기반한 WC Loss로 구현하였다. 매 epoch마다 Exit 플로우 임베딩에 대해 k-means 클러스터링을 수행하고, 동일 클러스터에 속한 임베딩 간 유사도는 증가시키고 서로 다른 클러스터 간 유사도는 감소시키도록 학습을 진행한다. 수식은 다음과 같이 정의된다.

$$L_{WC} = \cos(p_i, C_{n(i)}) - \cos(p_i, C_{p(i)})$$

여기서 p_i는 기준 Exit 플로우 임베딩, $C_{p(i)}$는 동일 클러스터 중심(positive centroid), $C_{n(i)}$는 다른 클러스터 중심(negative centroid)을 의미한다. 해당 Loss는 동일 클러스터 중심과의 유사도는 증가시키고, 다른 클러스터 중심과의 유사도는 감소시키는 방향으로 작용한다. 예비 실험 결과, 이러한 WC Loss의 추가는 ANN 기반 탐색 환경에서 logAUC 기준 탐지 성능의 향상에 기여함을 확인하였다.

3.2 ANN 인덱스 생성과 vote 취합

임베딩 변환 이후, Exit 플로우 임베딩은 CAGRA 알고리즘을 이용하여 ANN 인덱스로 구성된다. 이후 Entry 플로우 임베딩을 쿼리로 사용하여 ANN 탐색을 수행하며, 각 쿼리에 대해 상위 M개의 근접 이웃을 반환한다. 여기서 M은 하이퍼파라미터이다. M=8, 16, 32, 64, 128에 대한 예비 실험 결과, M이 작을수록 logAUC는 향상되나 AUC가 저하되는 trade-off가 관찰되었으며, M=32에서 두 지표 간 균형이 최적임을 확인하여 이를 채택하였다.

각 플로우의 윈도우 수는 DeepCoFFEA[2]의 설정을 따라 11개로 고정하였다. 각 윈도우별로 독립적인 ANN 탐색이 수행되며, 결과적으로 윈도우당 M개의 후보 Exit 임베딩이 도출된다. 이렇게 반환된 후보들은 candidate list에 추가되며, 각 후보에 대해 1표씩 부여한다.

Method	Space	Time
DeepCorr	$\mathcal{O}(N^2L)$	$\mathcal{O}(N^2L)$
DeepCoFFEA	$\mathcal{O}(NL/R)$	$\mathcal{O}(N^2L/R)$
Ours	$\mathcal{O}(NL/R)$	$\mathcal{O}(N \log N)$

Table. 1 Time complexity comparison

최종적으로 모든 윈도우에서 얻어진 후보들을 집계하여, 가장 많은 표를 획득한 Exit 임베딩을 해당 Entry 플로우와 상관된 플로우로 판단한다.

3.3 시간 복잡도 분석

기존 DeepCoFFEA는 $O(n^2)$의 전수 비교 방식을 사용하는 반면, 제안 기법은 CAGRA의 인덱스 구축과 탐색 단계로 구성된다. 인덱스 구축은 NN-descent 그래프 생성 $O(n \cdot d \cdot logn)$, Reorder & Prune $O(n \cdot d^3)$, Reverse merge O(nd)로 이루어진다. d에 대한 실험 결과, d가 작을수록 탐색 성능이 저하되고 클수록 처리 시간이 증가하는 trade-off가 관찰되었으며, d=8에서 탐색 성능이 안정적으로 유지되면서 처리 시간이 최소화됨을 확인하였다. d=8 고정 시 지배항은 $O(n\log n)$이며, 쿼리 탐색 $O(\log n)$을 포함한 전체 복잡도는 $O(n\log n)$이다. Table 1은 두 방법의 시간복잡도를 비교한다.

IV. 실험

실험에는 DeepCoFFEA[2]에서 사용된 데이터셋을 활용하였으며, 2021년 6월 수집 데이터를 학습에, 7월 수집 데이터를 평가에 사용하였다. 실행 시간 측정을 위한 대규모 실험(100k~200k)에는 기존 데이터를 복제였으며, ROC 및 AUC 평가에는 실제 트래픽 데이터만을 사용하였다. 기존 DCF의 유사도 계산은 PyTorch 기반 GPU 구현으로 대체하여 동일한 NVIDIA RTX A6000 환경에서 공정하게 비교하였다. AUC만으로는 저 FPR 구간에서의 성능을 정확히 반영하기 어려우므로, logAUC(FPR축을 $\log_{10}$스케일로 변환하여 산출) 및 FPR $\leq$ 1e-6 구간의 pAUC를 추가 평가 지표로 채택하였다.

Fig. 2는 플로우 수 증가에 따른 각 방법의 실행 시간 변화를 나타낸다. 기존 DCF 방식은 n이 증가함에 따라 실행 시간이 급격히 증가하

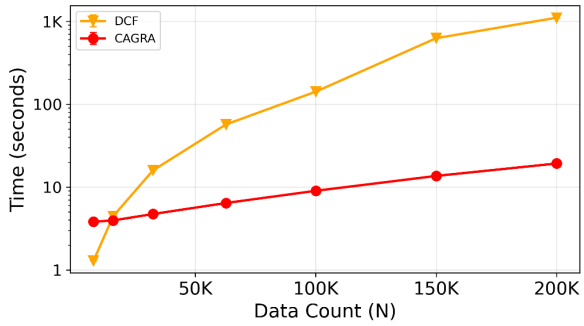

Fig. 2 Running Time Comparison

는 반면, 제안 기법은 완만한 증가 추이를 보이며 실용적인 실행 시간을 유지한다. 특히 실제 Tor 트래픽 규모에 해당하는 $n = 200k$에서 제안 기법은 DCF 대비 약 254배, $n = 60k$에서는 약 125배 빠른 처리 속도를 달성하였다. 한편, $n = 6k \sim 12k$ 구간에서는 ANN 인덱스 구축 초기 비용으로 인해 crossover point 가 관찰되나, 이후 구간에서는 제안 기법이 지속적으로 우수한 성능을 보인다.

Fig. 3은 각 방법의 ROC 커브를 나타내며, Table 2는 AUC, logAUC, pAUC(FPR≤1e-6) 수치를 비교한 것이다. 제안 기법은 DCF 대비 pAUC 기준 0.4145 → 0.4918로 향상되었으며, 실제 공격 환경에서 중요한 low-FPR 구간에서의 탐지 성능이 우수함을 확인하였다. IVF는 전체 AUC(0.9953)는 높으나 pAUC(0.3544)가 낮아 low-FPR 구간에 적합하지 않음을 확인하였으며, 이는 CAGRA 선택의 근거가 된다.

Method	AUC	logAUC	pAUC
DCF	0.9768	0.6171	0.4145
IVF	**0.9953**	0.5715	0.3544
CAGRA	0.9825	**0.6194**	**0.4918**

Table. 2 AUC, logAUC, pAUC of methods

V. 결론

본 논문에서는 Tor 트래픽 상관 분석 문제에 ANN 알고리즘인 CAGRA를 적용하여 대규모 환경에서의 연산 효율성을 향상시키는 방법을 제안하였다. 또한 GTT 분석을 통해 실제 종단 공격자가 처리해야 하는 트래픽 규모를 정량적으로 제시하고, 해당 환경에서 기존 DCF 방식의 한계를 확인하였다.

실험 결과, 제안 기법은 기존 방법 대비 탐지 성능을 유지하면서도 실행 시간과 연산 비용을

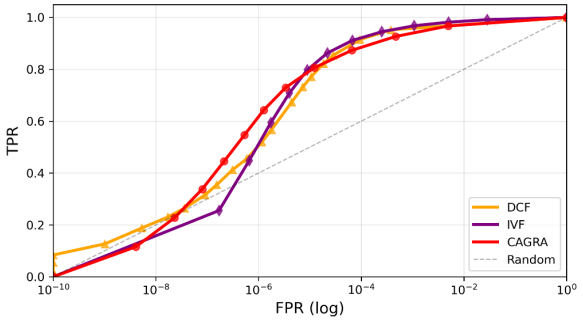

Fig. 3 ROC Curves

크게 감소시키는 것을 확인하였다. 향후에는 보다 세밀한 시간 윈도우 분석과 실시간 탐지 시스템으로의 확장을 검토할 계획이다.

[참고문헌]

[1] M. Nasr, A. Bahramali, and A. Houmansadr, "DeepCorr: Strong Flow Correlation Attacks on Tor Using Deep Learning," in Proc. ACM CCS, 2018.

[2] S. E. Oh, T. Yang, N. Mathews, J. K. Holland, M. S. Rahman, N. Hopper, and M. Wright, "DeepCoFFEA: Improved Flow Correlation Attacks on Tor via Metric Learning and Amplification," in Proc. IEEE S&P, 2022.

[3] R. Jansen, R. Wails, and A. Johnson, "A Measurement of Genuine Tor Traces for Realistic Website Fingerprinting," in Proc. PAM, 2026.

[4] H. Ootomo, A. Naruse, C. Nolet, R. Wang, T. Feher, and Y. Wang, "CAGRA: Highly Parallel Graph Construction and Approximate Nearest Neighbor Search for GPUs," in Proc. IEEE ICDE, 2024.

[5] M. Caron, P. Bojanowski, A. Joulin, and M. Douze, "Deep Clustering for Unsupervised Learning of Visual Features," in Proc. ECCV, 2018.

[6] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein GAN," arXiv:1701.07875, 2017.