Annonce de soutenance: Thèse

Délivré par : INSA Toulouse

École Doctorale : MITT

Unités de recherche : sara / CUSTOCY

Génération de données d'intrusion réseau par apprentissage profond de représentations

Network Intrusion Data Generation via Deep Representation Learning

Gabin Noblet

Date : 3 avril 2026 10:00

Lieu : LAAS-CNRS - Salle de Conférences 7 avenue du colonel Roche 31077 TOULOUSE Cedex 4

Composition du jury

Directeurs

Philippe Owezarski, Directeur de recherche, LAAS-CNRS

William Ritchie, CTO, CUSTOCY

Rapporteurs

Sandrine Vaton, Professeure, IMT Atlantique

Yacine Ghamri-Doudane, Professeur, Université de La Rochelle

Examinateurs

Grégory Blanc, Maitre de conférence, Télécom sud Paris

Pierre-François Gimenez, Chargé de recherche, INRIA

Vincent Nicomette, Professeur, INSA Toulouse

Résumé

Cette thèse propose une solution au problème de génération de données synthétiques de trafic réseau labellisées, essentielles pour l'entraînement et l'évaluation des systèmes de détection d'intrusions fondés sur l'apprentissage automatique. L'absence de jeux de données variés, actuels et annotés entrave leur performance, notamment face à des attaques sophistiquées, tel que les Advanced Persistent Threats (APTs). Nous développons NetGlyphizer, une méthode d'apprentissage de représentation discrète, inspirée du VQ-VAE, convertissant le trafic réseau en séquences de symboles discrets (NetGlyphs). Pour cela, nous proposons Nexus, un outil permettant de représenter le trafic réseau sous format Nxcap, un format de données réseau minimaliste représentant les flux comme séquences de paquets. Ce format permet de capturer la distribution temporelle et structurelle des paquets au sein d'un flux, avec plus de détail que des ensembles de statistiques descriptives. Le NetGlyphizer convertit ces flux en NetGlyphs, servant d'entrée à un modèle génératif fondé sur une architecture Transformer et conditionné par un label, afin de générer des séquences labellisées. Ces séquences sont ensuite décodées en flux réseau et exportées au format Pcap via l'outil Nexus. Le conditionnement par labels permet de générer du trafic spécifique pour divers scénarios ou classes de trafic, tout en préservant les propriétés statistiques et protocolaires des données originales. Les résultats confirment que le trafic synthétique reproduit fidèlement les caractéristiques du trafic réel, tant au niveau des distributions statistiques qu'au respect des protocoles. Ce travail introduit : (1) le format Nxcap et l'outil Nexus, (2) NetGlyphizer, un mécanisme d'apprentissage de représentation discrète pour le trafic réseau, et (3) la génération contrôlée, basée sur l'architecture Transformer, de trafic synthétique compatible avec les outils d'analyse existants.

Mots-clés : Apprentissage automatique, Sécurité des réseaux, Apprentissage de représentation, Ingénierie des caractéristiques, Modélisation générative, Evaluation expérimentale

Abstract

This thesis addresses the challenge of generating labeled synthetic network traffic data, essential for the training and evaluation of machine learning-based intrusion detection systems. The scarcity of diverse, contemporary, and labeled datasets limits their effectiveness, particularly against sophisticated attacks such as Advanced Persistent Threats (APTs). We develop NetGlyphizer, a discrete representation learning method inspired by VQ-VAE, that converts network traffic into sequences of discrete tokens (NetGlyphs). To achieve this, we propose Nexus, a tool that represents network traffic in Nxcap format, a minimalist network data format representing flows as packet sequences. This format captures the temporal and structural distribution of packets within a flow, with greater detail than sets of descriptive statistics. NetGlyphizer encodes these flows into NetGlyphs, which a label-conditioned generative model based on the Transformer architecture uses to produce labeled sequences. These sequences are then decoded back into network flows and exported to Pcap format via the Nexus tool. Label conditioning enables the generation of specific traffic for various scenarios or traffic classes while preserving the statistical and protocol properties of the original data. Results confirm that the synthetic traffic faithfully reproduces the characteristics of real traffic in terms of statistical distributions and protocol compliance. This work introduces: (1) the Nxcap format and the Nexus tool, (2) NetGlyphizer, a discrete representation learning mechanism for network traffic, and (3) a Transformer-based controlled generation of synthetic traffic compatible with existing analysis tools.

Keywords : Machine Learning, Network Security, Representation learning, Feature engineering, Generative modeling, Experimental evaluation