Restitution de la journée de travail sur les jeux de données du 14 janvier 2025
&
Roadmap Lot 5

Plénière SuperviZ - 11/03/2025

Objectifs de la journée de travail

Les jeux de données sont au cœur de l'évaluation des systèmes de détection d'intrusions (IDS)

- Explorer la qualité, la sélection et l'utilisation des jeux de données pour les IDS
- Identifier les meilleures pratiques, les lacunes et les axes d'amélioration
- Discuter des méthodes de pré-traitement, des représentations et des évaluations de la qualité et de sélection des jeux de données

Problématique abordée

- Les datasets sous forme de snapshots provenant des testbed [1]
- Fréquemment obsolètes et ne couvrant pas les attaques récentes
- Erreurs courantes : mauvais étiquetage, mauvaise représentation des features (eg., corrélations non pertinentes) [2] [3]

Ces limitations rendent les expérimentations moins fiables, d'où l'importance d'évaluer rigoureusement les jeux de données.

Déroulement de la journée

- Présentation invitée par Philippe Owezarski du LAAS-CNRS
  - Métrologie réseau et corrélations temporelles (e.g., paramètre de Hurst)
  - IA générative pour la génération de trafic d'attaque diversifié (NetGlyphs)
- Session 1 : Usages des jeux de données
- Session 2 et 3 : Evaluation et sélection des jeux de données

Session 1 (1/2)

- Représentation des données NIDS
  - **Flux** (e.g., NetFlow), tout en conservant les fichiers PCAP
  - Granularité des captures, observations temps-réel, et prise de décision
- Pré-traitement des jeux de données
  - Nettoyage, normalisation, validation croisée
  - Traitement des valeurs aberrantes selon l'objectif
- Extraction des features
  - Sélection guidée (expertise) ou automatique (representation learning [4])
  - Explicabilité et importance des caractéristiques (e.g., Info Gain, Gini)

Session 1 (2/2)

- Domaine et couverture des données
  - Besoin d'intégrer des scénarios réels et modernes pour mieux couvrir les attaques actuelles (les TTPs de MITRE ATT&CK)
  - Inclure des comportements légitimes variés
- Le concept de généralisation
  - La généralisation doit être abordée sous différents angles : échantillons, distributions, et tâches [5]

Session 2-3 (1/2)

- **Caractérisation** plutôt que qualification
- **Mesures intrinsèques** des jeux de données
  - Corrélations entre couches, spatiales et temporelles
  - Entropies et mesures de diversité des jeux de données
  - Identifier les limites intrinsèques des jeux de données
- Évaluation des jeux de données : méthodologies complémentaires
  - **Tests de substitution** pour comparer différents jeux de données
  - Utilisation de DeepCrime pour tester la robustesse face aux jeux de données intentionnellement corrompus

Session 2-3 (2/2)

- Détection et gestion des corruptions
  - Anomalies structurelles : incohérences entre features, duplications non intentionnelles
- Qualité et augmentation des jeux de données
  - Importance de générer des données cohérentes tout en augmentant la diversité des scénarios
    - ★ NetGlyph [4], FlowChronicle [6], ...

Enseignements clés et Actions Futures

Enseignements clés

- Les jeux de données IDS sont souvent biaisés par leur dépendance aux environnements de test, limitant leur représentativité [1]
- L'évaluation des jeux de données doit inclure des mesures intrinsèques et leur adéquation avec la tâche de détection
- Augmenter la diversité des jeux de données avec des scénarios réalistes et des TTPs (MITRE ATT&CK) [4], [6]

Actions Futures

- Identifier et corriger les biais structurels et contextuels dans les jeux de données [7]
- Intégrer des scénarios augmentés et des vecteurs d'attaque modernes
- Concevoir de nouvelles mesures pour caractériser la qualité des jeux de données IDS [8]

Roadmap Lot 5 et Suite des Travaux

1. **Fiabilité et validité des mesures de diversité** → Avril 2025
   - Fiabilité en lien avec différents jeux de données
   - Validité en lien avec la détection
2. **Mesures de complexité de classification** → Septembre 2025
   - Tests sur environ 10 jeux de données existants
   - SoTA pour la détection des erreurs et biais dans les jeux de données
3. **Framework de caractérisation des jeux de données** → Déc. 2025
   - Intégration des aspects de diversité et de complexité
4. **Génération de données pilotée** → Mars 2026
   - Satisfaction des caractéristiques souhaitées
5. **Méthodologie d'évaluation des IDS** → Juin 2026
   - Intégration avec le framework FREIDA
   - Evaluation des IDS SuperviZ

References I

[1] G. Apruzzese, P. Laskov, and J. Schneider, "Sok: Pragmatic assessment of machine learning for network intrusion detection," in *2023 IEEE 8th European Symposium on Security and Privacy (EuroS&P)*, pp. 592–614, IEEE, 2023.

[2] A. Kenyon, L. Deka, and D. Elizondo, "Are public intrusion datasets fit for purpose characterising the state of the art in intrusion event datasets," *Computers & Security*, vol. 99, p. 102022, 2020.

[3] M. Lanvin, P.-F. Gimenez, Y. Han, F. Majorczyk, L. Mé, and E. Totel, "Errors in the cicids2017 dataset and the significant differences in detection performances it makes," in *International Conference on Risks and Security of Internet and Systems*, pp. 18–33, Springer, 2022.

[4] G. Noblet, C. Lefebvre, P. Owezarski, and W. Ritchie, "Netglyph: Representation learning to generate network traffic with transformers," in *2024 20th International Conference on Network and Service Management (CNSM)*, pp. 1–9, IEEE, 2024.

[5] C. Rohlfs, "Generalization in neural networks: A broad survey," *Neurocomputing*, vol. 611, p. 128701, 2025.

[6] J. Cüppers, A. Schoen, G. Blanc, and P.-F. Gimenez, "Flowchronicle: Synthetic network flow generation through pattern set mining," *Proceedings of the ACM on Networking*, vol. 2, no. CoNEXT4, pp. 1–20, 2024.

[7] R. Flood, G. Engelen, D. Aspinall, and L. Desmet, "Bad design smells in benchmark nids datasets," in *2024 IEEE 9th European Symposium on Security and Privacy (EuroS&P)*, pp. 658–675, IEEE, 2024.

References II

[8] M. Mitchell, A. S. Luccioni, N. Lambert, M. Gerchick, A. McMillan-Major, E. Ozoani, N. Rajani, T. Thrush, Y. Jernite, and D. Kiela, "Measuring data," *arXiv preprint arXiv:2212.05129*, 2022.