Anomaly Detection in Link Streams

Maroua BAHRI (RO) and Mehdi Naima (ComplexNetworks)

Keywords: Anomaly Detection, Random Sampling, Link Streams, AutoML

1 Context

Anomaly detection of link streams refers to identifying unusual patterns, structures, or suspicious behaviors, in a sequence of edges that arrive continuously over time, see Figure 1 for an example. This research area is crucial in cybersecurity, system failures, social network analysis, or network attacks. Detecting anomalies in link streams presents several challenges, *inter alia*, (i) *dynamic nature*: since the graph evolves over time, algorithms need to operate in an online and real-time manner to process the continuous flow of new edges and nodes, (ii) *scalability*: link streams can be large-scale with billions of edges requiring highly scalable algorithms with low computational and memory overhead, (iii) *lack of labeled data*: in many real-world settings, it is hard to obtain labeled data (ground truth), as result, researchers concentrate on identifying randomly injected links considered as anomalies.

Multiple solutions have been proposed in the literature to detect anomalies [2, 3] in streaming graphs, including traditional graph-theoretic techniques and modern machine learning-based approaches. However, these methods often face limitations in effectively addressing the diverse challenges inherent in link streams, which prevents their ability to accurately detect anomalous node behaviors.

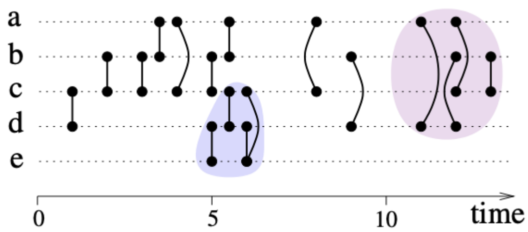

Figure 1: From [1]. A l ink stream: temporal interactions between *a, b, c, d e* over time range $[0, 15]$. Interactions in shaded areas may be traces of frauds or attacks.

2 Objectives

In this project, we propose a novel algorithm to detect both real and artificially generated anomalies in link streams, addressing current limitations and challenges in this domain. An envisaged solution consists of two key components that work in tandem to enhance detection accuracy and manage computational demands:

First, the aim is to use a reservoir sampling technique [4] alongside a sliding window mechanism. The reservoir maintains a fixed-size collection of historical links, preserving a representative subset of the graph's evolution over time. Meanwhile, the sliding window continuously tracks the most recent links in the evolving graph, ensuring that short-term dynamics are captured effectively. This dual approach is inspired by techniques such as the one presented in [5] and enables our model to handle the streaming nature of link data with minimal memory overhead. By combining long-term and short-term views of the graph, the proposed approach preserves essential structural patterns without the need to store the entire graph history, thus ensuring scalability and adaptability in the face of large, evolving link streams (addressing thus the challenges (i) and (ii) previously mentioned).

Second, to tackle the challenge of the scarcity of ground truth data, we would develop an automated framework that leverages multiple state-of-the-art methods for anomaly injection and detection methods, as summarized in [2, 3]. This component of the framework generates both synthetic and real-world-like anomalies, simulating realistic suspicious behaviors to evaluate detection robustness (to address the challenge (iii). For final anomaly scoring of nodes, we propose two aggregation strategies to enhance the accuracy and reliability of detection: (i) a mean-based approach, where scores from various detection methods are averaged to provide a balanced view, capturing general patterns of abnormality, or (ii) a weighted scoring system that dynamically adjusts each method's contribution based on its historical performance, inspired by approaches such as in [6].

Expected outcomes and contributions:

- *A comprehensive research paper:* The proposed framework will be rigorously documented in a scientific publication, including theoretical analysis, algorithmic details, and an extensive experimental study to validate the proposed solution. The paper will also feature a comparative evaluation against current state-of-the-art methods, showcasing the performance gains achieved in terms of detection accuracy, scalability, and robustness to different types of anomalies.
- *Open-source implementation:* To support transparency and further research, a fully documented open-source implementation will be provided. This will include all source code, datasets, and experimental configurations used, allowing researchers and practitioners to reproduce and build upon our work easily.

3 Positioning in regard to LIP6

Project title in French. La détection d'anomalies pour les flots de liens
Coordinators. Maroua Bahri (RO) et Mehdi Naima (ComplexNetworks)
Involved teams and transversal themes:

- Recherche opérationnelle (RO): black-box optimization, incremental algorithms
- ComplexNetworks: Graphs, link streams, dynamic networks

This project fits the strategic AID (Artificial intelligence and data science) research axis of LIP6 , combining the expertise of the two participating teams, ComplexNetwork and RO, and helping to strengthen our position in this emerging field.

4 Internship

The project main funding will be dedicated to a Master's thesis student dedicated to the objectives of the project. The student will do their master's thesis in either of the teams concerned with the project namely ComplexNetworks and RO. The title of their thesis will be the same as the title of this document.

5 Budget

The main aim of our project is to fund a **Master's thesis of 6 months** duration. Additionally, a funding of a **travel to a conference** where the student could present their work in national venues such that AlgoTel, FRCCS, journées MADICS or journées RADIA or international venues such that PAKDD, IEEE Big Data or LoG. Finally, **the student could visit for a week an institute** France or a neighboring country (such that Belgium or Germany) to present their research and gather more ideas around the subject.

Expected budget:

- 6 months internship (3600 euros).
- Conference participation (900 euros).
- Visit of an institute (1000 euros).

In total this would amount to 5500 **euros**.

References

[1] "Labcom fit." `https://www-complexnetworks.lip6.fr/~latapy/fit/`. Accessed: 2024-10-30.

[2] L. Akoglu, H. Tong, and D. Koutra, "Graph based anomaly detection and description: a survey," *Data mining and knowledge discovery*, vol. 29, pp. 626–688, 2015.

[3] G. Pang, C. Shen, L. Cao, and A. V. D. Hengel, "Deep learning for anomaly detection: A review," *ACM computing surveys (CSUR)*, vol. 54, no. 2, pp. 1–38, 2021.

[4] J. S. Vitter, "Random sampling with a reservoir," *ACM Transactions on Mathematical Software (TOMS)*, vol. 11, no. 1, pp. 37–57, 1985.

[5] S. Tabassum and J. Gama, "Sampling massive streaming call graphs," in *Proceedings of the 31st annual ACM symposium on applied computing*, pp. 923–928, 2016.

[6] A. Putina, M. Bahri, F. Salutari, and M. Sozio, "Autoad: an automated framework for unsupervised anomaly detection," in *2022 IEEE 9th International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 1–10, 2022.