*s*Gradd: Towards RELIABLE S.t..r...e....ami..ng Graph Analytics

Aida Sheshbolouki

David R. Cheriton School of Computer Science
University of Waterloo
aida.sheshbolouki@uwaterloo.ca

1 Transient Concepts in Graphs

The continuous generation of relations among entities and the growing need to run queries over them necessitate Streaming Graph Management Systems (SGMS, such as *s-graffito*[1]). These systems deal with dynamic and high velocity and volume of the data arrivals while generating reliable outputs. SGMS generates unreliable outputs when input data is new, temporal, incomplete, or adversely manipulated and the system does not recognise and manage it properly [2, 3]. A main cause identified for this problem is Concept Drift (CD), which occurs when a change in a hidden context induces changes in a target concept [4]. Understanding, detecting, and adaptation to CD in streaming data is (i) challenging due to stateful and blocking operations, and (ii) impactful in a variety of practical scenarios [5].

The literature is mostly focused on black-box drift detection and adaptation integrated within supervised learning systems, assume independence of data instances, and the target concepts defined as class labels. These assumptions and design choices do not always work. In this talk, I will discuss the challenges of designing a CD detection framework and introduce *sGradd*, a streaming graph framework for drift ddetection.

2 Challenges

CD in data streams is commonly considered as a change in underlying probability distribution of data points, which are generated independently [1]. However, streaming graph records are usually interconnected and dependent. Moreover, current definition implies detection based on comparing data distribution over window samples using fixed size sliding windows. When the streaming rate is highly dynamic with significant rises, adaptive window sizes are more efficient. Thus, we need a CD definition to enable any detection solution in streaming graphs.

The challenges are designing an unsupervised CD detection method and also an effective performance evaluation. Real-world streaming data with drift labels are not easy to acquire. Moreover, the accuracy and latency of the detection are tightly bounded together and different drift patterns require different examinations. I will discuss how we approached these challenges and what tools and techniques we developed to address them.

[1] dsg-uwaterloo.github.io/s-graffito/

3 sGradd

I will introduce *sGradd*, which detects drift in the generative sources. This detection is the first operation when data becomes available in input monitor of SGMS. *sGradd* has two main components: one for data management and the second for drift detection constructed based on multidisciplinary techniques. I will explain how *sGradd* ingests data, updates analytic primitives, performs CD detection, and streams out drift signals with descriptions about their occurrence to inform the next analytics.

References

[1] Daniel Kifer, Shai Ben-David, and Johannes Gehrke. Detecting change in data streams. In *Proc. 30th Int. Conf. on Very Large Data Bases*, pages 180–191, 2004.

[2] Jie Lu, Anjin Liu, Fan Dong, Feng Gu, João Gama, and Guangquan Zhang. Learning under concept drift: A review. *IEEE Trans. Knowl. and Data Eng.*, 31(12):2346–2363, 2019.

[3] Navid Malekghaini, Elham Akbari, Mohammad A Salahuddin, Noura Limam, Raouf Boutaba, Bertrand Mathieu, Stephanie Moteau, and Stephane Tuffin. Data drift in dl: Lessons learned from encrypted traffic classification. In *2022 IFIP Networking Conf. (IFIP Networking)*, pages 1–9, 2022.

[4] Gerhard Widmer and Miroslav Kubat. Learning in the presence of concept drift and hidden contexts. *Machine Learning*, 23(1):69–101, 1996.

[5] Indrė Žliobaitė, Mykola Pechenizkiy, and Joao Gama. An overview of concept drift applications. In *Big Data Analysis: New Algorithms for a New Society*, pages 91–114. 2016.