NSERC-CSE Research Communities: Robust, Secure and Safe Artificial Intelligence and Exploratory Analysis of Unstructured Data

Communautés de recherche CRSNG-CSE : Intelligence artificielle robuste, sécurisée et sûre et analyse exploratoire des données non structurées

(Org: **Camille Archambault** (McGill University), **Steven Ding** (McGill School of Information Studies) and/et **David Thomson** (Tutte Institute for Mathematics and Computing))

CAMILLE ARCHAMBAULT, McGill University

An Agentic Pipeline Combining GraphRAG and UMAP for Explainable Vulnerability Discovery in Low-Level Code.

Identifying the root cause of vulnerabilities in low-level code is difficult, time-consuming, and requires expert knowledge. Understanding the cause, not just the visible symptom, is essential for patching and analyzing security impact. However, low-level code provides little structural context: during compilation, programs are re-organized into many small blocks, and a vulnerability may appear in one location even though its true cause lies elsewhere. Existing tools typically highlight where the problem is detected but cannot trace the underlying source-sink chain that leads to it. With the rise of Large Language Models (LLMs), new opportunities emerge for automating vulnerability discovery while improving transparency in vulnerability analysis.

To address this challenge, we treat the binary and its low-level code as a searchable knowledge base that the LLM can query during analysis. However, because vulnerability causes span long chains across multiple functions, standard RAG is insufficient. We therefore turned to GraphRAG, which incorporates graph relationships between code elements but is computationally expensive on large graphs and still lacks a global semantic view of the program. Our pipeline therefore also leverages UMAP to organize code embeddings into a compact semantic space. This combination allows the agent to quickly identify relevant blocks of code before performing focused graph traversal, enabling more efficient discovery of source-sink paths and producing more interpretable explanations of low-level vulnerabilities.

This work is part of an ongoing master's thesis project and presents a research direction for explainable vulnerability discovery in assembly code, laying the foundation for future implementation and evaluation.

BENJAMIN COOKSON, University of Toronto

Unifying Proportional Fairness in Centroid and Non-Centroid Clustering

Proportional fairness criteria inspired by democratic ideals of proportional representation have received growing attention in the clustering literature. Prior work has investigated them in two separate paradigms. Chen et al. (2019) study centroid clustering, in which each data point's loss is determined by its distance to a representative point (centroid) chosen in its cluster. Caragiannis et al. (2024) study non-centroid clustering, in which each data point's loss is determined by its maximum distance to any other data point in its cluster.

We generalize both paradigms to introduce semi-centroid clustering, in which each data point's loss is a combination of its centroid and non-centroid losses, and study two proportional fairness criteria, the core and, its relaxation, fully justified representation (FJR). Our main result is a novel algorithm which achieves a constant approximation to the core, in polynomial time, even when the distance metrics used for centroid and non-centroid loss measurements are different. We also derive improved results for more restricted loss functions and the weaker FJR criterion, and establish lower bounds in each case.

Based on joint work with Nisarg Shah and Ziqi Yu

SANJEENA DANG, Carleton University

Clustering compositional data with a logistic normal multinomial mixture model with an underlying latent factor structure

The human microbiome plays a crucial role in health and disease. Advances in next-generation sequencing technologies have made it possible to quantify microbiome composition with high resolution. Clustering microbiome data can uncover meaningful

patterns across samples, offering insights into biological variability and disease mechanisms. However, this task presents several challenges. Microbiome data are typically high-dimensional, over-dispersed, and compositional, reflecting relative abundances. As such, analyzing such compositional data presents many challenges because they are restricted to a simplex, which complicates standard statistical analysis. Here, we develop a family of logistic normal multinomial factor analyzers (LNM-FA) by incorporating a factor analyzer structure. The family of models is suitable for high-dimensional microbiome data, as the number of parameters in LNM-FA can be greatly reduced by assuming that the underlying latent factors are small. Parameter estimation is done using a computationally efficient variant of the alternating expectation conditional maximization algorithm that utilizes variational Gaussian approximation. The proposed method is illustrated using simulated and real datasets.

STEVEN DING,
Transforming Generic Coder LLMs to Effective Binary Code Embedding Models for Similarity Detection

Compiled programs appear as opaque binary instructions, yet many security tasks require deciding when two such binaries implement the same underlying computation. In this talk, I'll show how large language models can vectorize binary code in a way that captures this hidden structure: similar programs map to nearby points in embedding space, even when the binaries have been transformed by optimization, architecture changes, or obfuscation.

We introduce several simple training strategies—data augmentation, translation-style learning, improved embedding extraction, and a cumulative metric-learning loss—that greatly strengthen these representations. The result is a general-purpose model that learns stable, invariant embeddings of program behavior and outperforms specialized tools for binary similarity.

BENOIT HAMELIN, Tutte Institute for Mathematics and Computing
Representation of cyber defense telemetry for exploration tasks

Cyber defense of networks relies on the acquisition of large quantities of system telemetry, providing visibility into events that reveal intrusions. We present here a simple methodology for building a representation of salient objects that enables identifying *interesting* activity through an explorative lens. This approach organizes anomalies along similarity axes, while emphasizing features that distinguish objects from others. The methodology leverages labelling of routine activity, providing factual documentation of the baseline of systems as observed by sensors. Anomalous objects, among which lie traces of intrusions, are thus expressed through a vocabulary of modes of normal behaviour they are similar to, facilitating their interpretation.

JOHN HEALY, Tutte Institute for Mathematics and Computing
Exploiting distortions in clustering and dimension reduction for unstructured data

Understanding data is crucial for generating useful hypotheses and identifying various problems and biases that may be present in a dataset. Unstructured data can make this exploration challenging, but recent advances in vectorization methods have enabled the conversion of unstructured data into meaningful vector representations. Dimension reduction and clustering algorithms are powerful techniques for exploring and gaining insights from such vectorized data. However, these techniques inherently distort data—sometimes in useful ways, and sometimes problematically. In this talk, we examine the strengths, assumptions, and distortions inherent in some of the most popular clustering and dimension reduction techniques with a focus on some of the methods developed at the Tutte Institute.

TORYN QWYLLYN KLASSEN, University of Toronto
Remembering to Be Fair: Non-Markovian Fairness in Sequential Decision Making

Fair decision making has largely been studied with respect to a single decision. Here we investigate the notion of fairness in the context of sequential decision making where multiple stakeholders can be affected by the outcomes of decisions. We observe that fairness often depends on the history of the sequential decision-making process, and in this sense that it is inherently non-Markovian. We further observe that fairness often needs to be assessed at time points within the process, not just at the end of the process. To advance our understanding of this class of fairness problems, we explore the notion of non-Markovian

fairness in the context of sequential decision making. We identify properties of non-Markovian fairness, including notions of long-term, anytime, periodic, and bounded fairness. We explore the interplay between non-Markovian fairness and memory and how memory can support construction of fair policies. Finally, we introduce the FairQCM algorithm, which can automatically augment its training data to improve sample efficiency in the synthesis of fair policies via reinforcement learning.

This is joint work with Parand A. Alamdari, Elliot Creager, and Sheila A. Mcilraith.

PAUL MCNICHOLAS, McMaster University
Clustering and Dimension Reduction

An overview of some important concepts in clustering and dimension reduction. Some recent research and examples are discussed.

GERALD PENN, University of Toronto
Predicting Levenshtein Edit Sequences for Fine-Grained Estimation of Automatic Speech Recognition Error

The predominant method for scoring the quality of automatic speech recognition (ASR) transcripts when ground-truth labels are not available is to predict the word error rate (WER) from the corresponding audio segment. We propose WAV2LEV, a novel paradigm for WER estimation which predicts the underlying sequences of Levenshtein edit operations (substitutions, deletions, insertions and matches) from which the WER can be computed. This approach offers more fine-grained token-level error estimation in comparison to previous work without compromising on performance for WER estimation. To support this investigation, we present Mini-CNoiSY (Miniature Clean-Noisy Speech from YouTube), a bespoke 353 hour noisy speech corpus which ensures confidence in ground-truth labeling and captures a diverse range of noise artifacts which degrade ASR performance. Our results show that WAV2LEV achieves near state-of-the-art performance for the task of WER estimation with a root mean square error (RMSE) of 0.1706 and a Pearson correlation coefficient (PCC) of 87.42%, while generating predictions of ASR error that are more informative and fine-grained than that of direct WER estimators.

OPENING REMARKS, Session Organizing Committee
Opening Remarks

We will begin this session with some opening remarks from the organizers at 08:15, with the talks beginning on schedule at 08:30.

KALEB RUSCITTI, University of Waterloo
Modifying Mapper for Temporal Topic Modelling

In this talk, I will discuss how the Mapper algorithm can be used to model the evolution of topics in a corpus of documents over time. Many real-world corpora have document publication frequency that varies locally in semantic space, and this makes it difficult to select appropriate parameters for Mapper. I will describe my proposed modification of Mapper that removes the assumption of a single resolution scale across semantic space and improves the robustness of the results under change of parameters, as well as how this improves Mapper's utility for temporal topic modelling of real-world datasets.