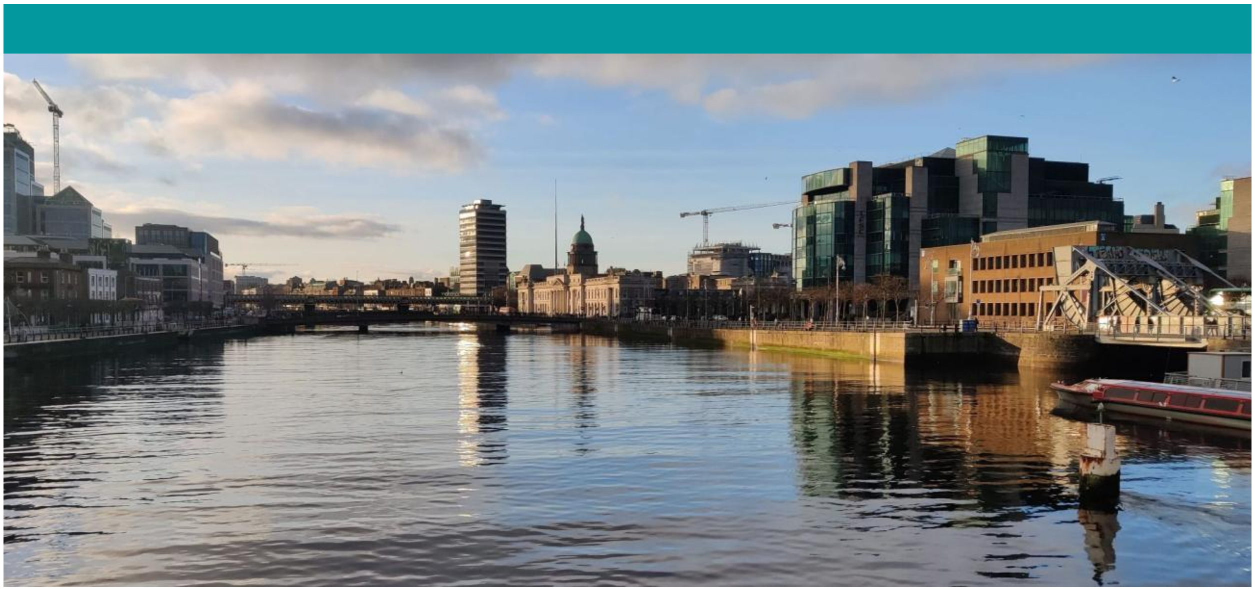
NEWS LETTER

20

NOVEMBER | 2025

CLOUD STARS



Insights on Outlier Identification through Data Provenance for Vulnerabilities Towards Membership Inference Attacks

by Gabriele Padovani (University of Trento)

Outlier detection plays a central role in assessing the privacy and security of machine learning models, particularly in the context of membership inference attacks (MIAs), where adversaries exploit distributional discrepancies to infer whether a given data point was part of a model's training set. Outliers -samples that deviate significantly from the data distribution- pose a heightened privacy risk due to their increased susceptibility to memory and inference. This risk is particularly apparent in large datasets and in models trained on heterogeneous data sources, such as large language models (LLMs), where manual outlier identification is infeasible and unreliable. Despite the growing importance of this threat, the existing literature provides limited guidance on systematic and automated methods for detecting and mitigating outliers in high-dimensional settings.

Gabriele Padovani visited IBM Dublin for three months in 2025, to work on investigating the presence of outliers using side-channel methods, such as inspecting loss and gradient discrepancies, with the aim of achieving a fully automated identification approach. The proposed novel technique works by identifying synthetic outliers and isolating samples most susceptible to inference attacks through a ranking system. The approach is validated through both controlled experiments and real-world use cases, and all runs performed are stored in provenance JSON files, which can be inspected in an interoperable manner.

CLOUD STARS

cloudstars.eu | twitter.com/Cloudstars_2023 | github.com/cloudstars-eu

CLOUDSTARS project has received funding from the European Union's Horizon research and innovation programme under grant agreement No 101086248