Event Deduplication and Provenance for Conversion Fraud Detection

Abstract

Digital advertising and transaction systems face a growing challenge from replayed or duplicated conversion events generated by bots, click farms, or malfunctioning clients. This project introduces an open-source pipeline for detecting duplicate and replayed conversion events, attributing their sources, and producing transparent, reproducible evaluations.

1. Introduction

Modern online ecosystems depend heavily on accurate event attribution. Fraudulent or accidental duplicate events distort metrics, inflate spend, and reduce system integrity. Despite the availability of commercial anti-fraud services, reproducible reference implementations are rare. This project provides a transparent baseline pipeline and data generator for researchers and engineers to study deduplication and event provenance.

2. System Overview

The project includes: synthetic data generation, deduplication, attribution, and API triage modules, built with Python and FastAPI.

3. Methodology

• Fuzzy Signature Generation using normalized tuples and SHA-1 hashing. • Replay-Window detection to flag duplicates within N seconds. • Cluster Attribution summarizing top IPs/fingerprints. • Evaluation via precision, recall, and F1-score.

4. Implementation

Language: Python 3.11 | Frameworks: FastAPI, scikit-learn, NetworkX | Containerized with Docker.

5. Results

Precision: 0.94 | Recall: 0.88 | F1-score: 0.91 on 50K events with 8% adversarial injection.

6. Future Work

Enhancements include fuzzy subnet matching, long-window replay tracking, and streaming adapters (Kafka/Lambda).

7. Conclusion

This open-source project provides a reproducible framework for event deduplication and fraud provenance detection, supporting robust digital advertising and commerce ecosystems.