Zalando
Secure Systems Group, Aalto University

Sebastian Szyller and Samuel Marchal

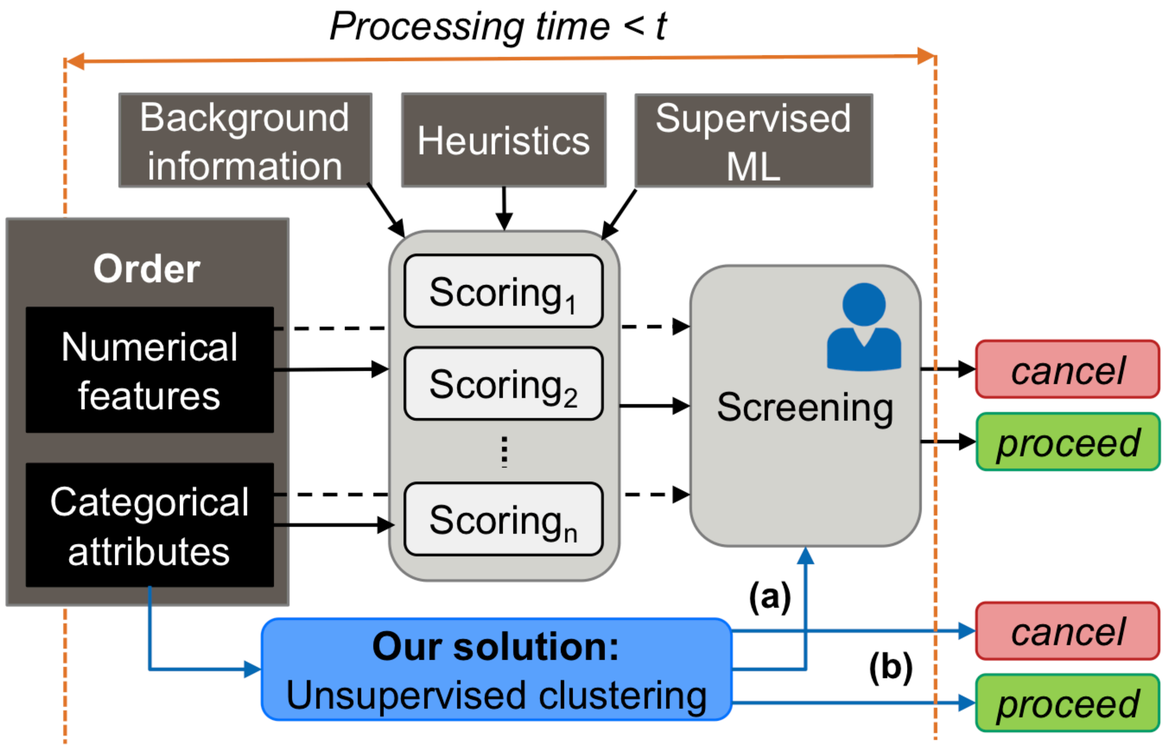
Detecting E-commerce Fraud with Large Scale Categorical Clustering

Motivation

- Online fraud constitutes 1-3% of all orders
- Total global loss over $50 billion a year
- Fraud prevention / income trade-off
- Detecting fraud is costly (manual)
- Fraud detection is time-constrained (within hours) and requires automation

Current Fraud Detection

- Leverages numerical features only
- Analyses orders in isolation
- Often relies on hand-crafted heuristics
- Final decision made by human screener



Identifying Fraud Campaigns

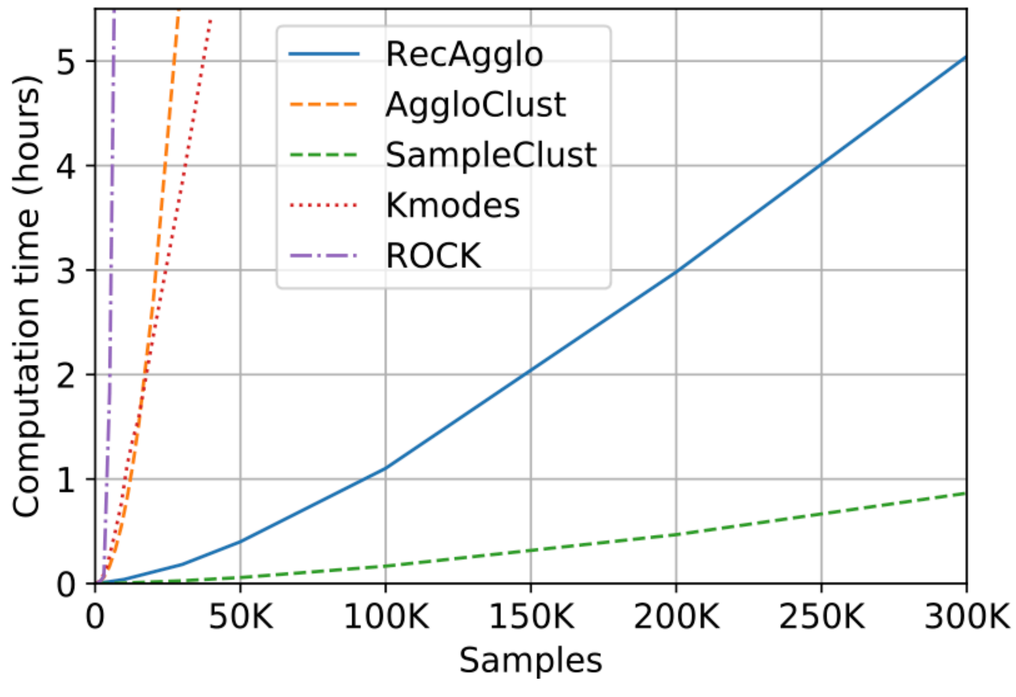
- Significant portion of fraud is organised
- Orders in the same campaign share many similarities, e.g. delivery, payment
- Group similar orders to identify campaigns
- Cancel fraudulent orders in bulk

Our Approach

- Algorithm for grouping categorical data
- Novel hierarchical (agglomerative) clustering
- Generate many small clusters
- Improved scalability through sampling & recursion
- Goals of clustering:
  - Minimise *cluster impurity (CI)*
  - Maximise *clustered fraud rate (CFR)*

Results - Clustering

- Scalable: process 300K orders in 5h
- Significant CFR > 40% & Low CI < 1%



Results – Fraud Detection

- Evaluated on 6 million orders from Zalando
- Cluster new orders with older known frauds
- Extend fraud label to the entire cluster
- Effectiveness and accuracy:
  - Detected fraud (recall): *26.4%*
  - False detection (FPR): *0.1%*
  - Precision: *35.3% (96.9%*)*

* Including returned, cancelled and partly unpaid orders

Aalto University

zalando