Integrating Graph and Machine Learning for Fraud Detection Use Case

Uri Lapidot[1] and Jay Yu[2,3]

[1]Risk and Fraud Platform, Intuit, Petah Tikva, Israel, uri_lapidot@intuit.com
[2]Technology Futures, Intuit, San Diego, USA, jay_yu@intuit.com
[3]Product and Innovation, TigerGraph, San Diego, USA, jay.yu@tigergraph.com

Abstract

The Risk and Fraud platform team at Intuit relies heavily on graph-based technologies to prevent fraud at scale. One of the challenges we were facing was how to expand the limited capabilities of the traditional ML approach to leverage rich semantics of accounts connected as a graph. In this paper, we will share our approach to integrate graph and machine learning together in an end-to-end risk and fraud platform, including practical solutions to overcome limitations in temporal support and adoption by ML Data Scientists.

Keywords: Fraud Detection, Graph Datab[1]ase, Machine Learning, Artificial Intelligence

1. Introduction

Payment fraud prevention is one of Intuit's top priorities to support the lifeline of our 6M small businesses globally. As fraudsters come up with more and more sophisticated attacks to redirect money flow by setting up fraudulent merchant accounts and faking business transaction activities, we find relying on traditional machine learning data features are not sufficient to detect and stop fraudulent activities. In this talk, we will share our journey to build a graph-based risk and fraud system for fraud detection, investigation and management, our insights from building such a system, challenges encountered and practical solutions to overcome them.

2. Graph-based Features vs. Traditional ML Features

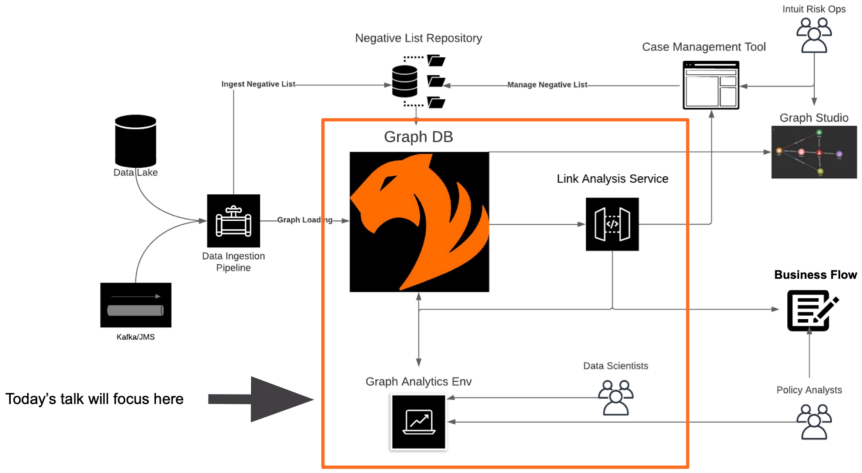
Traditional ML-based features in fraud detection use cases are usually drawn from relational datasets associated with user accounts and interactions. These features can be classified as the following categories:

1. Aggregations: sum of feature data columns. E.g., count of transactions from the same device in X days
2. Ratios: percentage of fraudulent over legitimate transactions. E.g., count of bad transactions from same device, divided by the count of legit transactions from same device in X days
3. Raw: direct feature value comparison. E.g., geo-location mismatch between IP and Zip code

Graph-based features add the new dimension of connectivity between any two user accounts with various degrees of hops on one or multiple paths. These linkages connected overtime offer much more intuitive, context rich, and explainable insights that can be leveraged by machine learning models directly to greatly increase the accuracy of the algorithms. Below is a simple example of how entities are connected via multiple hops in the fraud graph.

3. System Design, Challenges and Solutions

The end-to-end fraud detection and investigation platform is built with a graph database serving as the centerpiece. By modeling and storing in the graph database up-to-date user/merchant account info and their connected paths through contact and online access, we are able to generate on-demand graph-based features to enrich our link analysis ML pipelines and models for Data Scientists. At the same time, the underlying graph database is used to simplify and streamline fraud investigation and management via intuitive graph visualization.



During the implementation of the end-to-end system, we encountered two biggest challenges:

- How to capture the evolving changes of the fraud and risk graph?
- How to allow data scientists to query graph data directly without having to learn a new query language?

To overcome the above challenges, we designed and implemented the following solutions:

- Add time-dimension to all nodes/edges and adopt a hybrid strategy to connect latest snapshot data in graphs with the historical data in relational databases.
- Support de-facto service API standard (GraphQL) as a query language to simplify adoption

4. Integrating Graph Features into ML Models

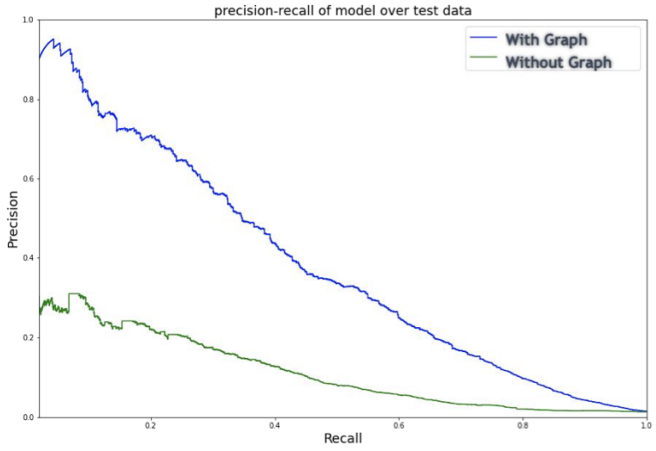
Our graph of business accounts, linked to each other via interactions through shared devices, emails or direct financial transactions, is an optimal representation of entities and relationships that are defined by human experts for the natural and intuitive reflection of the real world. In addition to directly applying graph algorithms to perform unsupervised machine learning directly on graph data without a separate ML pipeline, we explored practical ways to leverage deep insights in the graph to greatly enhance our fraud detection machine learning models.

One such sample insight is the “number of linked closed accounts (related to fraud) in 6 hops”. This graph-based feature is intuitive to get a deeper understanding of the level of risk for the account in review. When graph-based features like this combined with other regular non-graph based features get

fed into a supervised learning process, the resulting model automatically combines human domain knowledge encoded in the graph with the statistical power of machine learning. Thus dramatically increase the effectiveness of the resulting model.

5. Results and Summary

By taking a graph-based approach with seamless integration with machine learning, we are able to improve recall by 50% and precision by 50% for the fraud prediction ML model. In addition, one graph feature rose to the second most important feature for our fraud detection model.



This end-to-end risk and fraud platform built upon the graph and ML integration proved to be a huge success in production, becoming the backbone to fight against payment fraud in Intuit's fast-growing small business payment, capital, and cash capabilities.