ID2221 Data-Intensive Computing

Final project

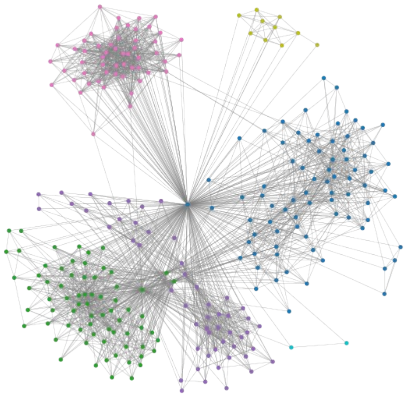
Fraud Detection with Neo4j

Barth Niklas, Camerota Fabio, Castellano Giovanni, Santoro Matteo

October 31, 2022

Introduction

- Electronic money transactions generate **large amount of data**
- Network of transactions as a **graph**
- Banks must detect **fraudsters**

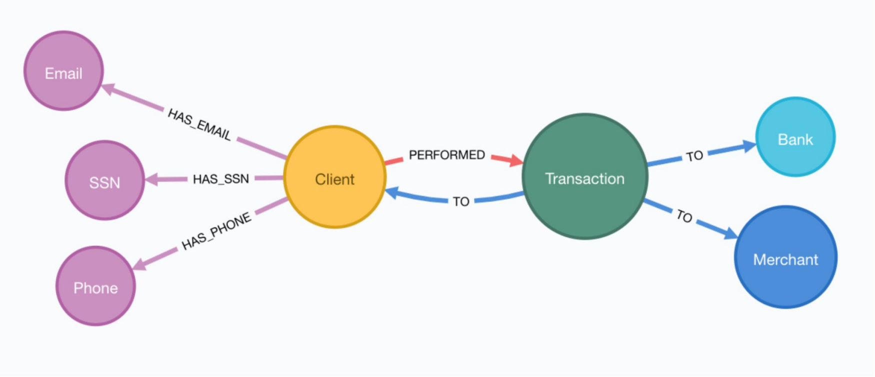


Goal

- **Neo4j** to perform **graph analysis**
- **Identify fraudsters** in a network of transactions
- How the amount of data affects the **performance**

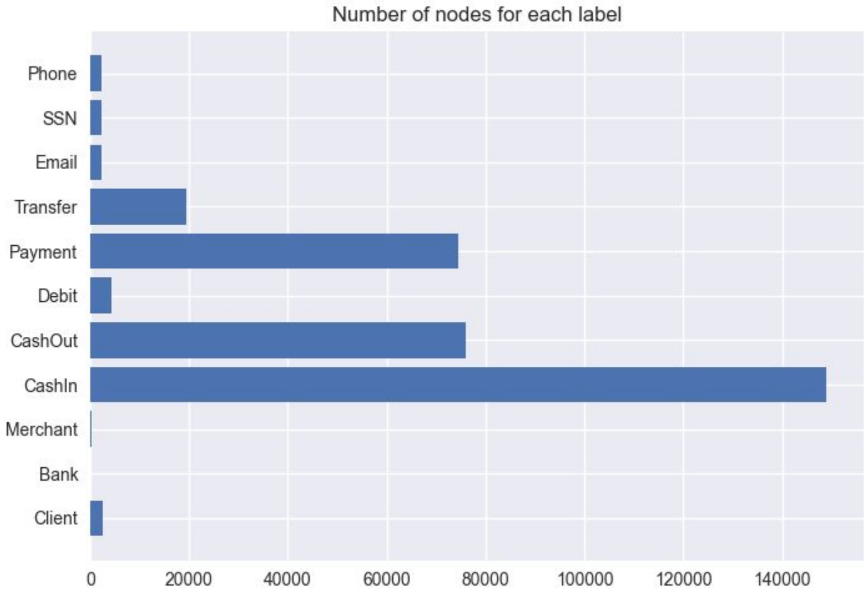
The PaySim Dataset

- **PaySim**: money transactions dataset
- 3 types of nodes
  - **Agents:** Clients, Merchants, Banks
  - **Transactions:** CashIn, CashOut, Debit, Transfer, Payment
  - **Identifiers:** Phone number, Email, SSN



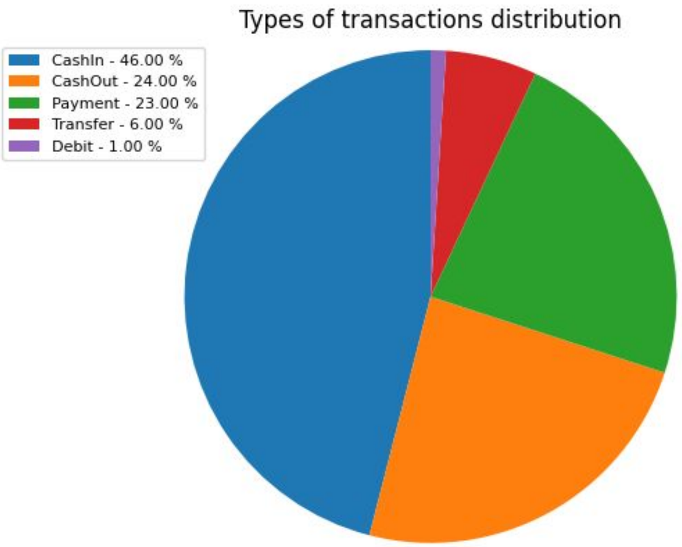
Source: https://www.sisu.io/posts/paysim/

The PaySim Dataset - Node Labels



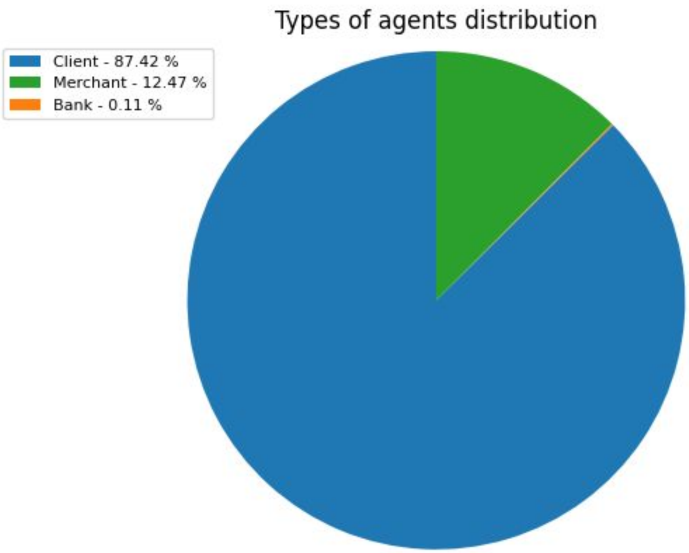
Most of the **nodes** in the graph are **transactions**

The PaySim Dataset - Transactions



CashIn and **CashOut** account for **70%** of **transactions**

The PaySim Dataset - Agents



Most of the **agents** are **clients**

Project Setup

- Installation of **Neo4j Desktop**
- Loading of **PaySim Dataset**
- **Neo4j Python Driver** to interface with database
- **Neo4j Browser** to visualize the results of the queries
- **GDS** to perform graph analysis

Two Types Of Fraudsters

First-Party Fraudsters

Second-Party Fraudsters

First-party Fraudster

- **First-party Fraudster**: a client who gives **false information** about his identity
- If two clients **share identifiers** one of them probably is a First-party Fraudsters

Identification

- New relationship between clients who share identifiers
- Weakly connected component
- Jaccard similarity score
- Degree centrality

Second-party Fraudster

- **Second-party Fraudster**: a client who help a First-party Fraudster
- Clients who **exchange money** with First-party Fraudsters are likely to be Second-party Fraudsters

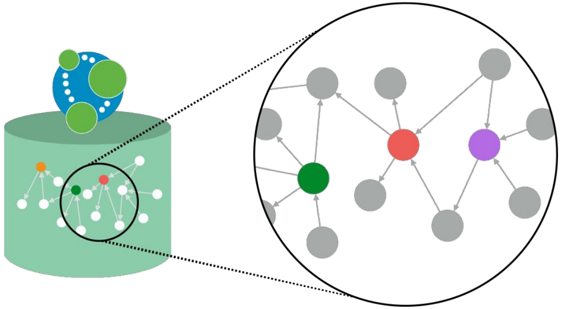
Identification

- New relationship between suspects and First-party Fraudsters
- Weakly connected component
- PageRank

How the amount of data affects the analysis?

Problem

- Small amount of data available
- Low computational capacity



Solution

- Downsample the graph
- Dropping transactions is the best way

Results

	First-party	Second-party
0%	17	46
5%	17	43
15%	17	37
30%	17	36

- The number of **First-party Fraudsters** does not depend on the number of transactions
- **Second-party Fraudsters** decrease as we remove transactions

Thank You