COMP8210/COMP7210
Big Data Technologies

Assignment 2

Semester 2, 2024
Macquarie University, School of Computing

Due: Friday 27 September, at 5pm
Total Mark: 100
Weighting: 25%

This Assessment Task relates to the following Learning Outcomes:

- Obtain a high level of technical competency in standard and advanced methods for big data technologies
- Understand the current status of and recognize future trends in big data technologies
- Develop competency with emerging big data technologies, applications, and tools

Background. With scams and fraud costing Australians over $3B in 2022, increasingly fraud actors are becoming more sophisticated and difficult to detect with traditional techniques. Graph data science powers a new generation of fraud solutions, designed to leverage connections in data to detect behavioural patterns and trace actions through complex networks of operators.

Dataset. We have supplied a number of data files on iLearn representing a set of accounts and money transfers across a fictitious set of individuals and retailers/sellers. There are a series of CSV formatted files as follows :

clients.csv - contains the account details for the individual account holders in the system. This includes synthetically generated id, name, contact details and tax file number.

stores.csv - id and names off the sellers of stores that users are purchasing from.

purchase.csv - amount and time of a purchase made by an individual from a seller

xfer.csv - amount and time of a money transfer made between two users in the system.

You will use this data to construct a graph of interconnections between users and stores in the system, you will also use a graph to analyse the account details of users in the context of fraud.

Part 1. Initial Graph Data Model (20%)

Create a base graph data model for the payments system by importing CSV data into Neo4j. The initial database should follow the model shown in Figure 1, however the assignment will require that we evolve this data model in stages.

Data can be imported to Neo4j using the process shown in the lecture for CSV. Be sure to create appropriate constraints for unique identifiers and appropriate indexes to support the questions in Parts 2 and 3. Note also that the timeOffset field in the data specifies seconds since midnight 12 May 2024, be sure to represent them as absolute values (a date and time) in the database using appropriate conversions and types. Purchase and Transfer are individual labels, but both of these are types of Transaction so use appropriate labels for this.

Figure 1. Graph Data Model for payments system.

Part 2. Initial Queries (35%)

Write Cypher statements, in Neo4j, for each of the following problems:

- Problem 1 (5%):

  Which client spent the most on purchases between 10am and 2pm on May 12, 2024.

  Example result row :

  ```
  name                total
  "John Citizen"      450.12345
  ```

- Problem 2 (10%):

  Find the top 5 clients with the worst negative cashflow overall where incoming money is less than outgoing.

  Example result row :

  ```
  name                balance        big_spend
  ```

```
"John Citizen"    -200.1234     50.1234
```

- Problem 3 (10%):

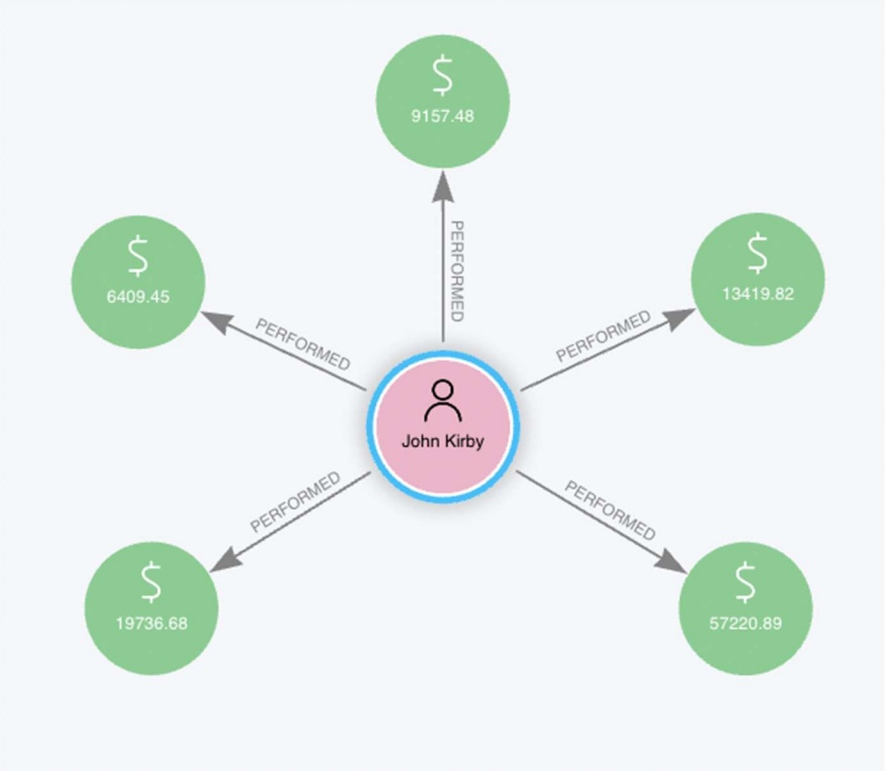
  We have reason to suspect that the Seller named 'Woods' is receiving money stolen in scam activity. Use GQL to discover where money transfers are sent to clients that in turn purchase goods from the suspect Seller. Return the name of these "pass on" clients and the amounts they receive through transfers prior to making transactions and the total amounts of the purchases themselves. Only show results where at least 5% of the money received is used to purchase from the suspected seller.

  Example result row :

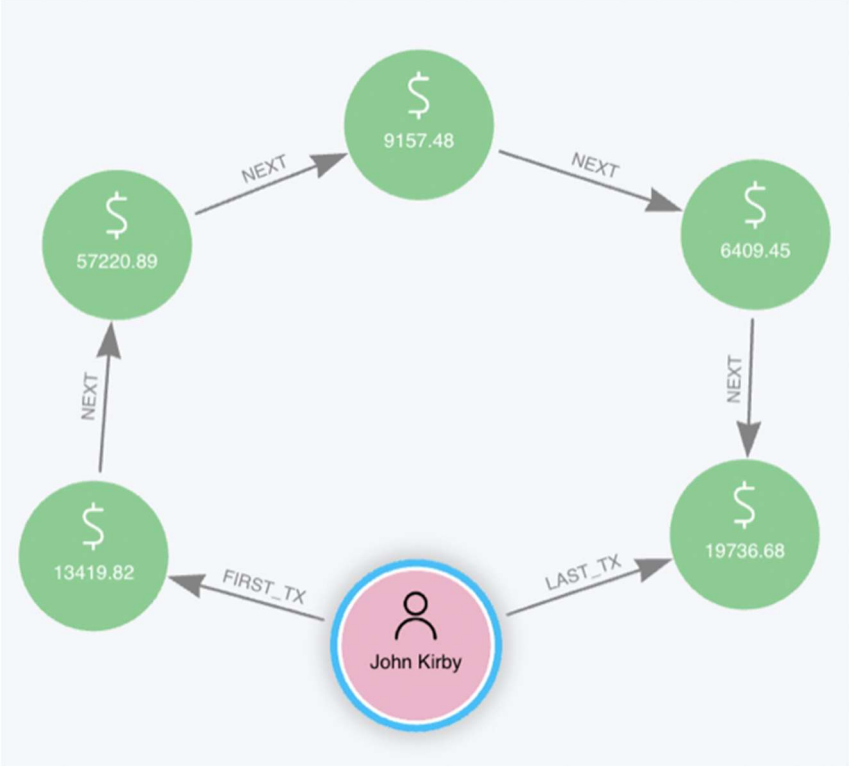
name	percentage	total_xfer	*total_purchase*
"John Citizen"	8.29	5000.1234	4000.1234

- Problem 4 (10%):

  We think that there are repeated patterns of transactions that might serve as a signature for fraudulent behaviour. To support future use cases like this we would like to add a new layer of relationships to the data set that creates an ordering of transactions. Currently, Clients are related to transactions in the following way :



  Write GQL/Cypher statements to create new relationships that connect ordered transaction as follows :



Part 3. Graph Data Science (45%)

Now we will use Neo4j Graph Data Science (GDS) to analyse features of the client - transaction graph.

- Part A (20%):

  From the original data model, we can see that user account information such as email address, phone number etc has been structured such that clients using shared details will be connected in the graph by shared nodes that represent these values. When stolen information is used to create fraudulent accounts, often this forms tightly connected groups in the graph we can detect with algorithms designed for that purpose.

  i) Use Neo4j GDS to create a projection of this portion of the graph and use appropriate algorithms to detect these groups of accounts using shared identifying information.

  ii) Identify the larger groups (of at least 5 members) and tag these Clients with a group ID by writing back to the original Neo4j dataset during execution of a GDS algorithm.

  iii) Create a visualisation in Neo4j Browser or Bloom to show the largest of these groups including members and all of the common shared identifiers connected together.

- Part B (25%):

  While finding, identifying and removing fraudulent accounts is a great use case for graph analytics, the real power comes from being able to dig deeper and further into connections using multiple datasets assembled into a powerful representation of connections in your data.

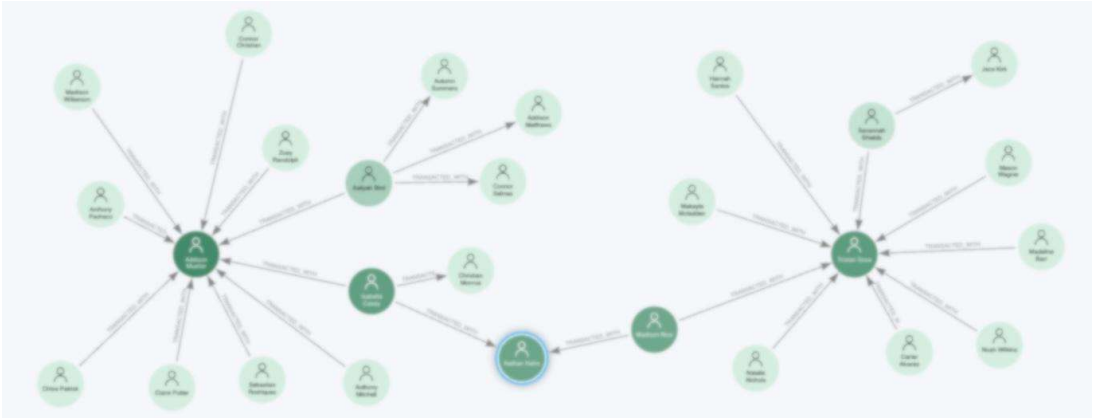
  Now that we have suspected fraudulent accounts identified, what can we learn from any transaction activity they have been able to perform. There must be a way for members to profit from these accounts and looking deeper as connections between the groups might lead us to central players in a larger fraud operation.

i) Write a GQL/Cypher statement that identifies transactional relationships that members of larger fraud groups (more than 5 members) have with accounts outside of their immediate group. Obviously transfers within the group are expected but looking at how money moves out of the group is a key to finding the central actors in a larger organisation.

ii) Create a GDS projection consisting of only these client accounts within larger initial fraud groups and the Clients they have transacted with outside of these groups. Use an appropriate GDS algorithm to now find tightly connected groups within this subset of Clients, these should represent some of the accounts that are being used to move money out of the fraud cells we initially detected.

iii) Use the largest of the groups detected above and perform analysis to try and identify any central client accounts (that is accounts that are the center of the network), find the appropriate GDS algorithms for this kind of analysis. The central account of this network is likely to be a key suspect in funneling scammed funds away from these fraudulent accounts.

iv) Create a Neo4j Bloom visualisation of this result and use value based styling on the nodes to highlight the key suspects. This result will look similar to the following :



Evaluation and Marking

- This is an individual assignment worth 25%.
- All assignments will be submitted using iLearn. The results of all assignments will be available via iLearn.
- You will need to create a video (max 10 minutes) and upload it on YouTube. Then, share the YouTube link in your assignment submission.
- The submission will be a zip file including the source code for Part 1 and 2, the YouTube Link, and the queries for part 3. You do not need to include the Graph database in the submission file.
- Students will demonstrate their assignments **during week 9** in person (in Practical and SGTA sessions).
- No late submissions will be accepted, unless a Special Consideration is Submitted before the assessment submission deadline, and Granted. Your assignment will be evaluated by the tutor independently.
- **If you have any questions related to this assignment, please submit them on the Discussion Forum on iLearn.**