

Graph Data Science for improved Financial Fraud Detection

Dr. S Babu¹, Mr. V Rama Narayanan²

¹ Department of Computer Science and Application, SCSVMV, Kanchipuram, India, babulingaa@kanchiuniv.ac.in

² Department of Computer Science and Application, SCSVMV, Kanchipuram, India, vgram31@hotmail.com

Abstract—Financial fraudsters use Gen AI, digital channels, global networks, and synthetic identities making it complex to identify the fraudulent activities. Traditional rule-based systems relying on traditional methods do not identify frauds which use multi-step transaction routing with multiple institutions and across borders.

Graph database using Labelled Property Graphs, represents customers, accounts, and transactions as interconnected nodes and edges. By ingesting live transaction data, they apply pattern-matching and community-detection to expose suspicious subgraphs. Money-laundering rings or collusive clusters—and let investigators trace multi-hop links to “hub” accounts with clear visual audit trails.

Machine learning models trained on vast historical datasets, using supervised classifiers (e.g., gradient boosting) and unsupervised anomaly detectors. Features like transaction amounts, geolocation consistency, device fingerprints, and temporal sequences feed these models, while recurrent architectures capture evolving fraud tactics. Yet they often suffer from concept drift, require extensive labelled data, underperform on imbalanced cases, and behave as opaque black boxes, generating false positives and hampering trust.

A hybrid framework combines relational graph insights with statistical scoring, boosting detection accuracy, reducing false alarms, and enhancing investigators’ confidence in fraud detection and prevention.

Keywords; Graph; Labelled Property Graph; Graph Data Science; Fraud Analytics.

1 Introduction to Graph Database and Graph Data Science

In today’s data driven landscape, Relational Database Management Systems (RDBMS) faces significant limitations. RDBMS has rigid schema requirements which hinders the evolution of data structures which are common in modern applications. RDBMS are not suitable for massive datasets with high traffic data load which are the basic characteristics of big data of the modern applications. Performance bottlenecks are created in distributed systems when we strictly follow the ACID properties. Unstructured data like JSON, XML or multimedia cannot be handled easily with RDBMS. RDBMS’s table-based structure is not flexible to handle complex relationships that’s created in social networks, IoT systems and real-time analytics. This has resulted in many organizations moving towards other alternative solutions which cater to the requirements of today’s systems.

Graph database stores the data in the form of Nodes and Edges. Data elements are represented in Labelled Property Graph (LPG) which helps to represent complex, interconnected data efficiently. Data Schemas in graph databases are flexible and easily adapt to changing business requirements.

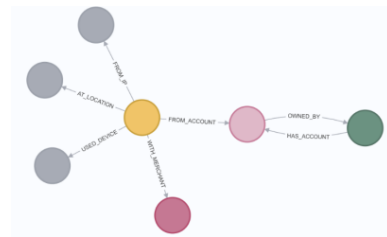


Fig. 1. Graph Database with Nodes and Relationships

Labelled Property Graph and Graph Data Science algorithms can be used as a sophisticated analytics tools to understand the data within the network. Exploring the connected entities can reveal patterns and structures that uncovers the insights that are hidden in the traditional RDBMS. Graphs provide structured way to navigate and understand connected networks, whether that’s a underground time table, a map of roads and highways, an organizational chart, or how parts of a system work together.

Relational databases are good at retrieving structured lists, they often fall short to describe the relationship and context. To understand the direct and hidden relationships, we need Graph Data base structure. Graph algorithms are used for this: traversing relationships between data points to uncover paths, critical hubs, natural clusters, and meaningful patterns. By applying Graph Data Science algorithms, raw business data can be transformed into richly connected intelligence, providing deeper insights that can help to take more impactful decisions.

Applying Graph Data Science algorithms to the connected data results in enriching the data with critical information that can be used for improved results for fraud investigation and detection. Graphs provide unique way of visualizing the connected data and relationships instead of tabular format which requires specific knowledge of RDBMS SQL to write complex queries with multi-level joins. With Graphs, business users will be able to spend more time on the fraud investigation independently and conclude the case investigation with verifiable and explainable results.

2 Shortcomings of Traditional Methods

There are various limitations of Relational Database Management Systems (RDBMS). Today's data driven applications require agile and flexible data modeling features to handle the dynamic nature of application requirements.

Limitations of RDBMS – Table based architecture of RDBMS is very rigid

Performance Bottlenecks for Traversal – Fraud detection requires tracing connections or funds across multiple hops, example Person A to Person B to Person C in Bank XYZ, Account ABC1234

Recursive Joins : In RDBMS, each hop requires a JOIN operation which is resource intensive.

Exponential Slowdown : Performance in RDBMS can degrade by 85% for each additional level of relationship depth.

Throughput Gap : Modern Financial networks need to analyse upto 50,000 transactions per second, using RDBMS based systems support only upto 3000 transactions per second.

Inability to Detect Complex Patterns– Fraud detection requires tracing connections or funds movements across the customers with multiple hops, example Person A to Person B to Person C in Bank XYZ, Account ABC1234

Table Silos : In RDBMS, each hop requires a JOIN operation which is resource intensive. Performance in RDBMS can degrade by 85% for each additional level of relationship depth resulting in latency bottleneck

Static Rule Engines : Traditionally Banks and Financial Institutions deploy static rule engines for transaction monitoring any transaction failing to satisfy the rules will be put for additional level of human intervention. Fraudsters often by-pass these rules by changing the transaction amount, transaction date, days between transactions etc., to escape the monitoring and continue with the fraud transaction.

Limitations of Machine Learning Algorithms – Financial Fraud is comparable to finding needle in haystack. Imbalanced dataset provided to any of the Machine Learning algorithm will not provide accurate results and hence the comprehensive Enhanced Minority Oversampling TEchnique (EMOTE) can be used as input to the classifier by balancing the dataset. The key idea of the EMOTE method is to balance the dataset by tuning the misclassified instances of the minority classes into correctly classified instances through oversampling their nearest neighbor.

Traditional analytics work on tabular, disconnected data. While the Graph Data Science reveals relationship-driven insights impossible to detect with SQL

Table 1. Traditional SQL Vs Graph Database

Aspect	Traditional SQL	Graph Database
Data Model	Tables and Rows	Nodes and Relationships
Focus	Entities	Connections and Structure
Example	“Who bought what”	“Who influences others to buy”
Strength	Aggregation	Pattern discovery

3 Fraud Detection Approach with Graph Database

Graph Algorithms are powerful tools for detecting fraud because fraudulent activities often involve patterns of relationships that are difficult to detect using traditional methods. Here are some of Graph Data Science Algorithms that can be used for Fraud detection:

3.1 Community Detection - Find groups or clusters of nodes in the graph database. Identify sets of nodes that are more densely connected to each other.

Key Questions Answered:

- Are there different customer segments in the database based on behaviour or connections?
- Identify the potentially coordinated groups, teams, or fraud rings?
- Which are the subdivisions or clusters within the network?
- Is the data in the graph connected or isolated?

Key Algorithms:

Table 2. Community Detection Algorithms

Algorithm Name	Purpose of the Algorithm
Louvain Modularity	Hierarchical modularity optimization that identifies dense communities by measuring internal connection strength
Leiden Modularity	An improvement on Louvain, offering better-defined and more stable community detection.
Label Propagation	Fast structure-based detection where node labels spread through the network until stable communities emerge
Strongly Connected Components	Finds maximal sets of nodes in directed graphs where every node can reach every other node in the set through directed paths
Weakly Connected Components	Groups nodes into components where any path exists between nodes, ignoring edge direction essential for understanding overall graph structure

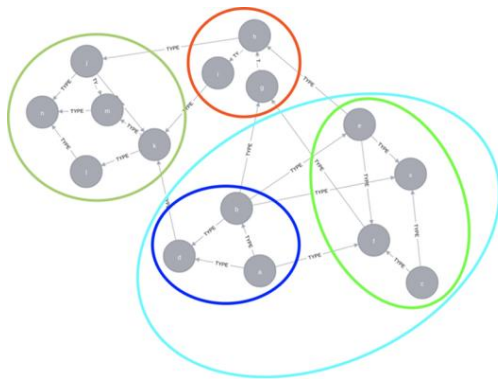


Fig. 2. Sample Community Detection Algorithm Results

Community detection reveals the hidden structure of networks by identifying nodes naturally cluster together into tightly connected groups. These algorithms help us understand organizational patterns, social dynamics and structural relationships within complex graphs.

3.2 Centrality Algorithms help to measure the importance or influence of individual nodes within the network based on their position and connections

Key Questions Answered:

- Who are the most influential or critical persons in the network?
- Who are the that act as key connectors, bridges, or hubs?
- Which are the potential bottlenecks or centres of activities?
- Are the nodes reachable by other nodes?

Key Algorithms:

Table 3. Centrality Algorithms

Algorithm Name	Purpose of the Algorithm
Degree Centrality	Measures: Direct relationship count Use Case : Find popular people or frequently purchased products Example: Social network users with most friends – potential influencers
Closeness Centrality	Measures: Average distance to all other nodes Use Case : Detect individuals who reach everyone quickly Example: In communication networks, identifies who spreads information fastest
Betweenness Centrality	Measures: Frequency on shortest paths between others Use Case : Identify connectors or brokers Example: Major airport hubs connecting other airports in logistics networks
Page Rank	Measures: Importance based on incoming links Use Case : Ranking web pages Example: Websites with quality backlinks rank higher in search

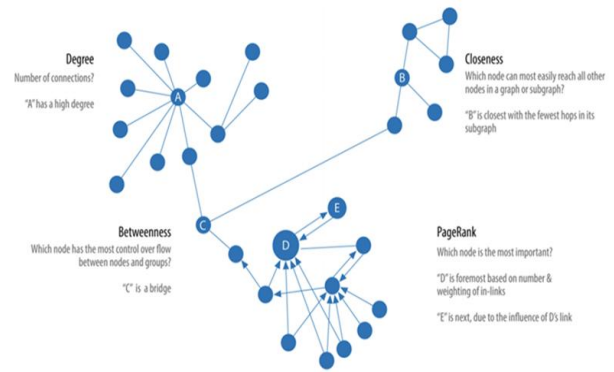


Fig. 3. Centrality Algorithms

Centrality algorithms help to identify the most influential and connected nodes in the network. The Centrality algorithms help identify highly active or isolated accounts, understand the overall connectedness of the transaction network.

3.3 Other Graph Data Science Algorithms include Path Finding, Pattern Matching, Similarity, Anomaly Detection and Link Prediction, algorithms can be used separately or with combinations to get more accurate results for the given use case

Table 4. Fraud Type and Graph Algorithms

Fraud Type	Algorithm to be used	Purpose of the Algorithm
Organized Fraud Rings	Louvain, Page Rank	Find Connected groups
Individual Fraudsters	Centrality, Anomaly Detection	Identifies Outliers
Staged Accidents	Pattern Matching	Detects Specific schemes
Identity Theft	Similarity, Path Finding	Find duplicates and connections
Repair Shop Fraud	Degree Centrality	High Claim Volume
Money Laundering	Shortest Path	Tracers Fund Flow
Collusion	Triangle Counting	Finds Closed Groups
Emerging Fraud	Link Prediction, GNN	Anticipates New Patterns

4 Experiments and Results

While imbalanced datasets are curated using EMOTE technique, there is scope for further improving the accuracy of prediction using Graph Data Science. Highly interconnected datasets can be represented in Graph Database as Nodes and Relationships, Data represented in Graph Database are optimized with deep, multi-level connections such as bank transactions, social networks and network infrastructure.

Table 5. Fraud Analytics Datasets

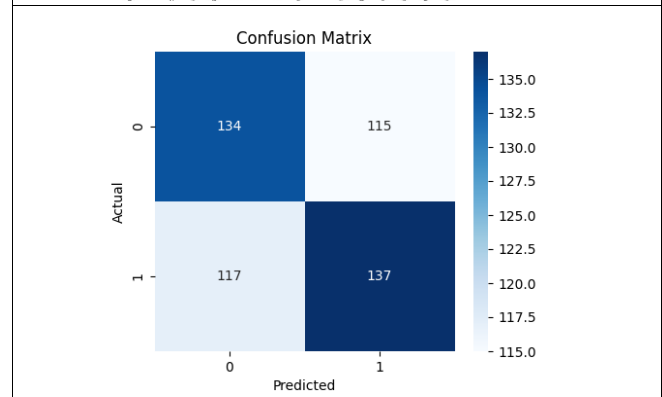
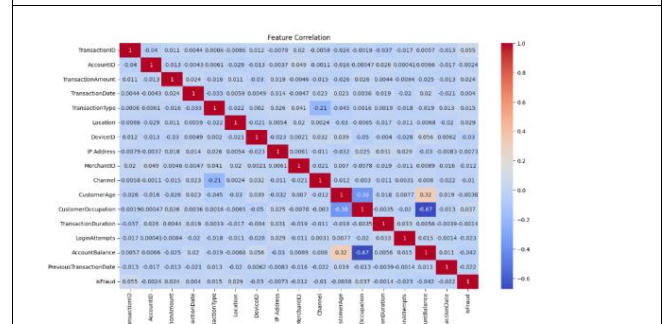
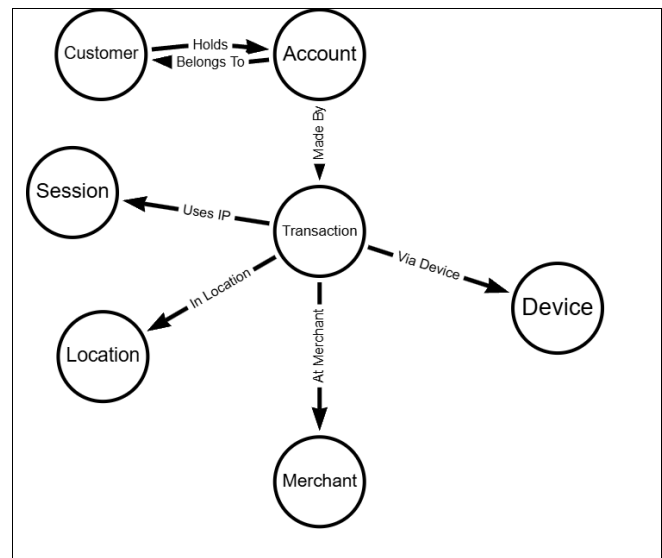
Kaggle Dataset Name	Author	# of rows	# of columns
UPI Payment Transactions Dataset	Rugved Patil	1000	8
Bank Transaction Dataset for Fraud Detection	Vala Khorasani	2512	16
Bank Transaction Fraud Detection	Sagar Maru	200K	24
Fraudulent Financial Transaction Prediction	Younus Mohamed	227K	28
Bank Account Fraud Dataset Suite(Neurl IPS 2022)	Sergio Jesus	1 M	32
Fraud Detection Dataset	Aman Ali Siddiqui	6.3 M	11

Experiment was conducted on Bank Transaction Dataset for Fraud Detection (Vala Khorasani – Highlighted in Table 5)

Below table illustrates the Nodes and Relationship of the experiment data set

Table 6. Fraud Analytics Datasets – Vala Khorasani Dataset Graph

Dataset Details	Table Column Head		
	Table Columns	Nodes	Relationships
Financial Fraud Dataset	TransactionID	Transaction (2,512)	Account ownership and transaction flows (FROM_ACCOUNT, OWNS, HAS_ACCOUNT) Transaction context (AT_LOCATION, USED_DEVICE, FROM_IP, WITH_MERCHANT, VIA_CHANNEL) Session tracking (DURING_SESSION) Pattern detection (SHARES_DEVICE, SHARES_IP, SHARES_MERCHANT, SHARES_LOCATION) Sequential tracking (NEXT_TRANSACTION) Geographic locations (SHARES_LOCATION) Channel (7) - Transaction channels (likely mobile, web, ATM, etc.)
	AccountID	Account (2,512)	
	TransactionAmount	Financial transactions	
	TransactionType	LoginSession (2,512) - User login sessions	
	Location	Customer (498)	
	DeviceID	Customer (498)	
	IP Address	Customer (498)	
	MerchantID	Account (498) - Bank accounts	
	Channel	Device (684) - Devices used for transactions	
	CustomerAge	Devices used for transactions	
	CustomerOccupation	IpAddress (595) - IP addresses	
	TransactionDuration	Merchant (103)	
	LoginAttempts	Merchant/vendor information	
	AccountBalance	Location (46) - Geographic locations	
	TransactionDateNew	Channel (7) - Transaction channels (likely mobile, web, ATM, etc.)	



Graph based features are applied to the data set to increase the performance of the classification model, the below set of Graph Data Science Algorithms are applied to the dataset to obtain the additional feature elements

Table 7. Graph features enriched in Fraud Analytics Dataset

Community Detection WCC	Network in the dataset is identified by direct transfer from User A to User B and indirect connections between users is identified by shared device, IP Address or location.
Closeness Centrality	The Closeness centrality algorithm evaluates how close a node is to all the other nodes
Page Rank	PageRank algorithm is commonly used to find the most important or influential nodes in the network

By Combining the Graph Based features to the baseline features improves the predictability of the Machine Learning Model. While more Graph Data Science algorithms can be applied to the dataset, this paper used only the Centrality and Community detection algorithms.

Table 8. Traditional Vs Graph Features

Traditional ML Algorithm Feature Importance of the Model (Baseline)		Graph Data Science – Enriched Feature Importance of the Model (Baseline + Graph) Combined	
Feature	Value	Feature	value
numberOfDevices	0.360835	numberOfDevices	0.105482
incomingTransactions	0.193416	incomingTransactions	0.060731
maxIncomingAmount	0.106009	maxIncomingAmount	0.027178
totalIncomingAmount	0.074445	totalIncomingAmount	0.028057
numberOfCCs	0.072848	numberOfCCs	0.013917
avgIncomingAmount	0.068528	avgIncomingAmount	0.022201
maxOutgoingAmount	0.025771	maxOutgoingAmount	0.005899
avgOutgoingAmount	0.025304	avgOutgoingAmount	0.005938
totalOutgoingAmount	0.025192	totalOutgoingAmount	0.005224
numberOfFlps	0.024789	numberOfFlps	0.009248
outgoingTransactions	0.022863	outgoingTransactions	0.005958
		componentSize	0.37182
		part of community	0.301243
		Closeness	0.026716
		Pagerank	0.010389

5 Conclusion

Graph Databases and Graph Data Science algorithms offer a good solution to provide insights that will help the financial institutions to overcome the present challenges. By using graph data science algorithms along with proven machine learning methods improves the overall results of predicting fraud, reducing the false positives. Furthermore, complex and suitable Graph Data Science algorithms can be applied to different set of fraudulent activities to uncover the hidden patterns, identify the missing links. A hybrid framework combines relational graph insights with statistical scoring, boosting detection accuracy, reducing false alarms, and enhancing investigators' confidence in fraud detection and prevention.

References

- [1] Corydon Baylor, Enzo Htet, <https://neo4j.com/blog/graph-data-science/graph-algorithms/#:~:text=Graph%20algorithms%20are%20powerful%20analytics,More%20in%20this%20guide>:
- [2] Patil, A., Mahajan, S., Menpara, J., Wagle, S., Pareek, P., & Kotecha, K. (2025). Enhancing fraud detection in banking by integration of graph databases with machine learning., *MethodsX* (Elsevier).
- [3] Paul, S., Mitra, A., & Koner, C. (2019). A review on graph database and its representation. *Proceedings of the International Conference on Recent Advances in Energy-efficient Computing and Communication (ICRAECC)*, 2019, 1-6.
- [4] [5] Babu, S., & Ananthanarayanan, N. R. (2017). EMOTE: Enhanced Minority Oversampling TEchnique. *Journal of Intelligent & Fuzzy Systems*, 33(1), 67-78.
- [5] [6] B. Can, A.G. Yavuz, E.M. Karsligil, M. Amac Guvensan, A closer look into the characteristics of fraudulent card transactions, *IEEE Access* 8 (2020) 166095–166109.
- [6] https://www.researchgate.net/publication/393746965_Comparative_Performance_Analysis_of_Four_RDBMS_Systems_Integrated_with_Django's_ORM
- [7] <https://backstick2.rssing.com/chan-1994510/article5443.html?nocache=0>
- [8] https://www.researchgate.net/profile/Babu-Santhalingam/publication/317852573_EMOTE_Enhanced_Minority_Oversampling_TEchnique/links/621333ea4be28e145ca639ae/EMOTE-Enhanced-Minority-Oversampling-TEchnique.pdf
- [9] <https://neo4j.com/blog/graph-data-science/graph-algorithms/>
- [10] <https://neo4j.com/blog/developer/using-neo4j-graph-data-science-in-python-to-improve-machine-learning-models/>
- [11] <https://www.kaggle.com/datasets/valakhorasoni/bank-transaction-dataset-for-fraud-detection?resource=download>