

WHITEPAPER

Preventing identity theft fraud in the financial services industry

How data science can become a detective

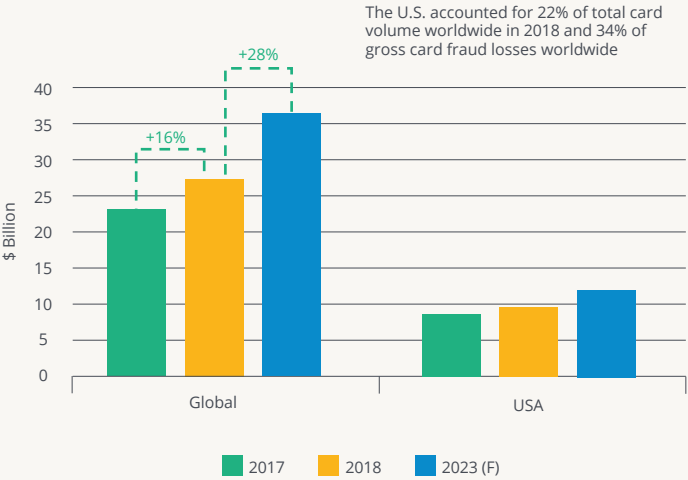
fractal



Digital fraud is growing at a concerning pace, with card issuers, merchants and acquirers facing increasing card fraud year on year.

Global card fraud is expected to increase to \$35B+ by 2023¹ with 72% of total losses experienced by card issuers alone.

Gross Card Fraud Losses



While the smartphone revolution and affordable, high-speed Internet have together transformed the payments market, there's been a simultaneous rise in the tide of "digital fraud", specifically, card payments fraud.

Source: [Nilson report](#)

Card fraud stats you need to know

Within CNP fraud, account takeover and synthetic fraud losses grew in 2018, helped by the easy availability of personally identifiable information (PII) for sale on the dark web¹.

Account takeover increased by 300% from 2017 to 2018, primarily driven by new account creation and existing account fraud².

1. <https://nilsonreport.com/mention/407/1link/>.

2. <https://www.javelinstrategy.com/coverage-area/2018-identity-fraud-fraud-enters-new-era-complexity>

Detecting synthetic fraud

What is synthetic identity fraud?

Synthetic identity fraud is the fastest-growing type of financial crime in the United States¹ and has become a focus for a large number of organized crime rings. In this type of fraud, criminals combine real and fake information to create a new identity. The real information used in this fraud is usually stolen.

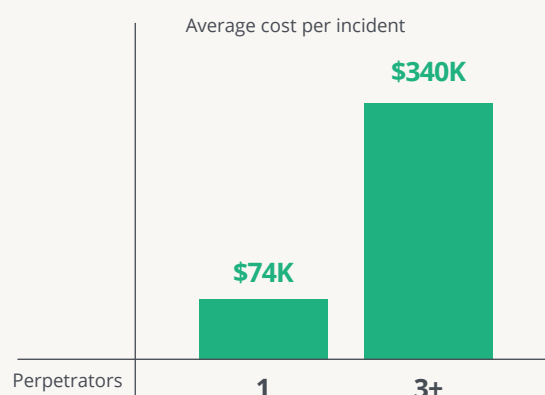
Playing the waiting game

With synthetic fraud, criminals play a waiting game before they hit an issuer with a much larger cost per incident. So how do we tackle this?

Potential data science solutions to identify fraudulent transactions include:

- Predictive analytics such as regression decision trees.
- Machine learning models such as Random Forest, GBM, and XGBoost.
- Deep learning models such as neural networks.

Fraud rings increase loss¹



Source: neo4j.com

A losing battle

However, these techniques that find patterns associated with fraud are of little use. That's because so few cases of synthetic ID fraud have been uncovered on which to train models. In fact, 85-95% of applicants identified as potential synthetic identities are not flagged by traditional fraud models².





While companies are investing millions into increasingly sophisticated fraud detection tools, the dynamic nature of fraud and the recent increase in collusive fraud, such as fraud rings, has led to businesses losing more and more money each year.

1. <https://www.mckinsey.com/business-functions/risk/our-insights/fighting-back-against-synthetic-identity-fraud>

2. https://www.idanalytics.com/wp-content/uploads/2018/11/Synthetic-Identity_Slipping-through-the-cracks_Executive-Summary.pdf

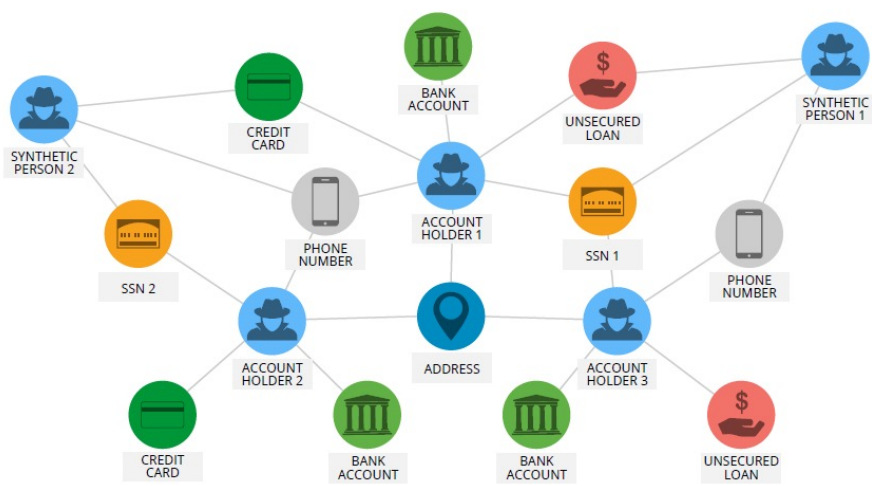
Turning the tide with data science

Graph databases and data science could potentially turn this trend around for several types of organization:

Organization Type		Fraud Type to be Prevented
 Banks, Credit Card Issuers, FinTech firms	→	Payment fraud
 Merchants/Ecommerce Retailers	→	Application fraud
 Digital Payment firms	→	Merchant Fraud
 IRS/Federal Agencies	→	Money Laundering

Why are graph databases important for fraud detection?

Graph databases offer new methods of identifying fraud rings and other advanced fraud scenarios with a high-level of accuracy. Significant insights can be drawn from existing data by transforming the current tabular data as a graph to understand the connections between the data.



Source: neo4j.com

How can graphs be leveraged?

Graphs can be used for analyzing the structure of the data, finding fraudulent patterns, data anomalies and fraud ring discovery among other applications.

When to use graph queries

Graph queries are useful when the intent of the analysis is deterministic, and where output can be determined with 100% certainty. For example:

- Calculating the number of customers with shared identifiers (such as common phone numbers, addresses, Social Security numbers etc.).
- Presence of a known fraudster in a customer's network.
- Multiple parties using the same account.

When to use graph algorithms

Graph algorithms are more relevant in understanding the overall structure of a network. They can also auto-detect suspicious patterns and anomalies in the data where no prior analysis-hypothesis exists, such as the identification of fraud rings.

Graph algorithms are more relevant in understanding the overall structure of a fraud network.

Fraud ring discovery: Four-step approach

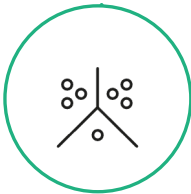
Fraud ring discovery using a graph database is a layered process with multiple steps and algorithms, such as those listed below¹.

1



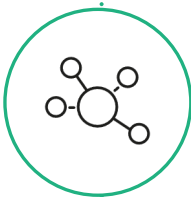
Convert a relational database of customer features such as account number, phone number, IP address and transactional data (such as the number of transactions, purchase volume etc.) into a graph database.

2



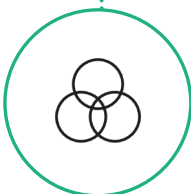
Synthetic fraud rings typically operate in large networks through a mix of real and fake identities shared across the network. Hence, a **community detection algorithm** can be effective in uncovering these networks.

3



A **centrality algorithm** like **PageRank** can then be used to determine nodes central to the fraud ring based on factors such as transaction volume, frequency of transactions and density of connections.

4



Successful isolation of a single fraud ring can help in the identification of other fraud rings based on similar patterns by using **similarity algorithms** such as **Jaccard**.

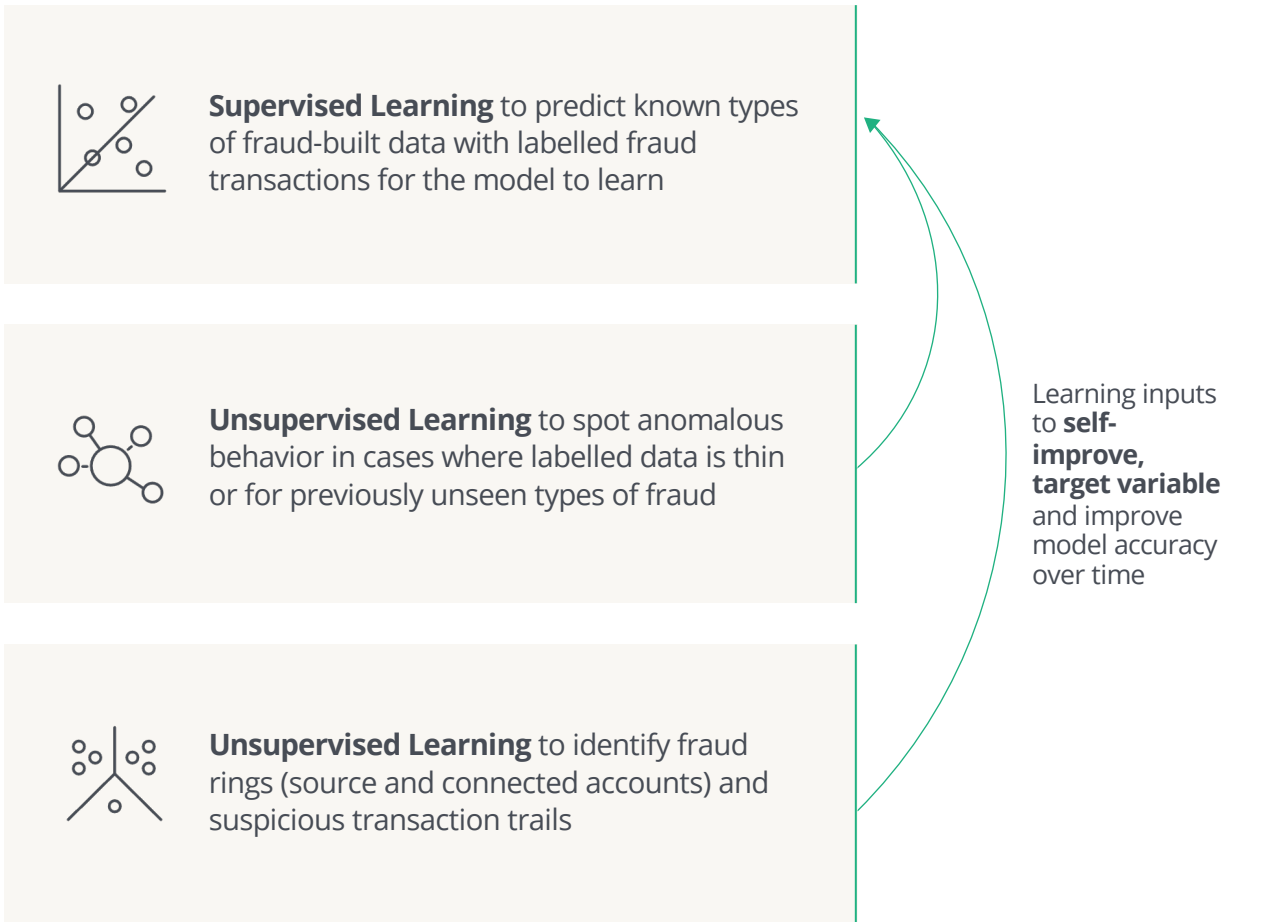
1. <https://neo4j.com/whitepapers/financial-fraud-detection-graph-data-science/>

Can graph algorithms be integrated with existing ML models?

Most organizations continue to rely on traditional ML models to identify fraudulent transactions. However, the pressing need is to integrate graph-based algorithms with existing pipeline of ML models. There are several ways in which this can be achieved without having to change the validated and well-understood approaches of machine learning.

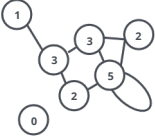

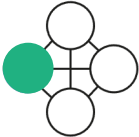

Self-improving labelled fraud data

Once the fraud rings are identified, the transactions can be labelled as fraud, and the fraud data used in supervised models via a self-learning framework. This can improve model accuracy over time.



Graph feature engineering

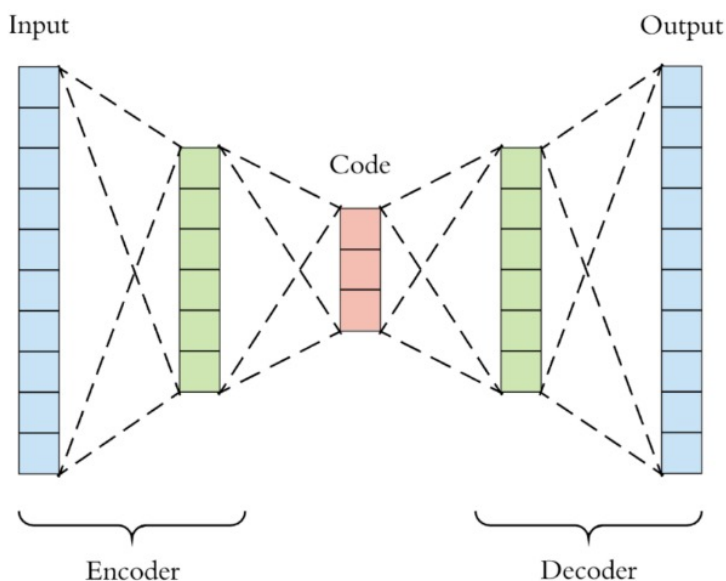
Connected graph features have been particularly helpful in investigating financial crime where fraudsters hide activities under layers of relationships. Some examples are mentioned below.

Feature Name ¹	Feature Description ¹	Use in Card Fraud
 Node Degree	<i>Number of connections to other nodes.</i>	Identifying shared information such as a phone number/SSN etc.
 Betweenness Centrality	<i>Calculates influence of a node (central node location in an account cluster).</i>	Locating potential mules in a fraud ring.
 Clustering Coefficient	<i>Measures density of connection between a group of nodes.</i>	Fully connected nodes could point to collusion. For example, between customer, merchant, and delivery agent.
 Adjacency Matrix	<i>Vectorized representation of a graph.</i>	Commonly used for finding short path transactions (e.g. rapid return of a purchase without reason) or shortest path between two points (e.g. proximity to known fraud accounts).

1. <https://neo4j.com/whitepapers/financial-fraud-detection-graph-data-science/>

Deep Auto-encoder – An unsupervised model for identifying fraud

While graph algorithms help in improving the availability of target fraud variables, the dataset that is available is often heavily unbalanced. This makes it computationally difficult to train an unbiased supervised ML model. This is where unsupervised models, such as Deep Auto-encoder (DAE), prove helpful.

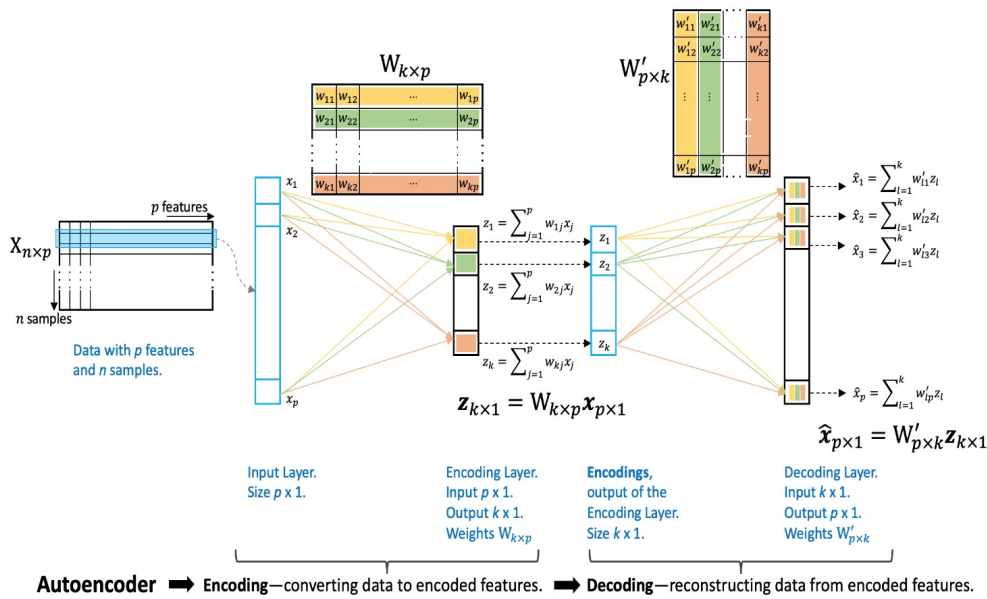


The DAE consists of two parts: the encoder, and the decoder. For any given input vector x , the encoder compresses the input vector to a much smaller latent space vector (z , referred to as code in the above diagram) and the decoder tries to reconstruct the latent space vector into the original input vector \hat{x} .

During the reconstruction, the decoder tries its best to minimize the reconstruction error ($|x - \hat{x}|$). If trained with sufficient data in normal conditions, the DAE can learn to distinguish between an anomaly and normal data by using the reconstruction error as the detection metric. Since DAE has never seen fraud data during the training process, its reconstruction error on Fraud data is significantly greater than 0.

This model is based on two fundamental principles regarding the anomalies:

- 1. A Type I anomaly usually exhibits a high variety of distinct attribute values such as rarely-used account, abnormal posting time and date. These types of anomalies are also referred to as global anomalies.
- 2. Type II anomalies occur when entries exhibit rare combinations of attribute values while their individual values occur quite frequently. These are referred to as local anomalies.



Source: towardsdatascience.com

Every entry x and its corresponding reconstructed output vector \hat{x} , anomaly metric is given by

$$AS(x, \hat{x}) = \alpha RE_{\theta}(x, \hat{x}) + (1 - \alpha)AP(x)..... (1)$$

RE_{θ} is the average reconstruction error obtained from DAE, measured as mean squared error; it is used to measure local anomaly.

$$RE_{\theta} = \frac{1}{n} (x - \hat{x})^T (x - \hat{x})..... (2)$$

$AP(x)$ measures the global anomaly by calculating the normalized probability of each feature constituting input vector x .

$$AP(x^i) = \frac{P(x) - P_{min}}{P_{max} - P_{min}}..... (3)$$

$P(x) = \ln \left(1 + \frac{n^i_j}{N} \right)$ is the sum of individual attributes. As you may notice, the individual features that exhibit high variety (a symptom of a global anomaly) is captured by $P(x)$. As seen in eq. (1) the final anomaly score is the weighted combination of local anomaly metric, eq. (2) and the global anomaly metric eq. (3). Finally, an instance is declared to be an anomaly in the $AS(x, \hat{x})$ exceeds a preset threshold.

Harness the power of connected data

With increasing digitization, fraud is becoming harder to label and detect. While ML models help identify common fraud, newer fraud types need a connected data solution.

A two-step approach is recommended to harness the power of connected data and improve fraud detection:

1 Move from relational to graph databases

Relational databases in tabular formats make it hard to discover the required relationships between data, while graph databases can achieve this with ease.

2 Integrate Graph algorithms with ML models

Graph-based analysis and graph algorithm-driven feature engineering can add powerful dimensions to an organization's available fraud ML pipelines leading to increased model accuracy without disrupting the existing modelling infrastructure.

Our Experts



Arpan Dasgupta
Client Partner, Financial Services



Karan Berry
Senior consultant, Financial Services

Enable better decisions with Fractal

Fractal is one of the most prominent players in the Artificial Intelligence space. Fractal's mission is to power every human decision in the enterprise and bring AI, engineering, and design to help the world's most admired Fortune 500® companies.

Fractal product companies include Qure.ai, Crux Intelligence, Theremin.ai, Eugenie.ai & Samya.ai.

Fractal has more than 2,300 employees across 16 global locations, including United States, UK, Ukraine, India, and Australia. Fractal has consistently been rated as India's best company to work for, by The Great Place to Work® Institute, a 'Leader' by Forrester Research in its Wave™ on Specialized Insights Services, Computer Vision & Customer Analytics and as an "Honorable Vendor" in 2021 Magic Quadrant™ for data & analytics by Gartner.



Corporate Headquarters

Suite 76J,
One World Trade Center, New York,
NY 10007

[Get in touch](#)