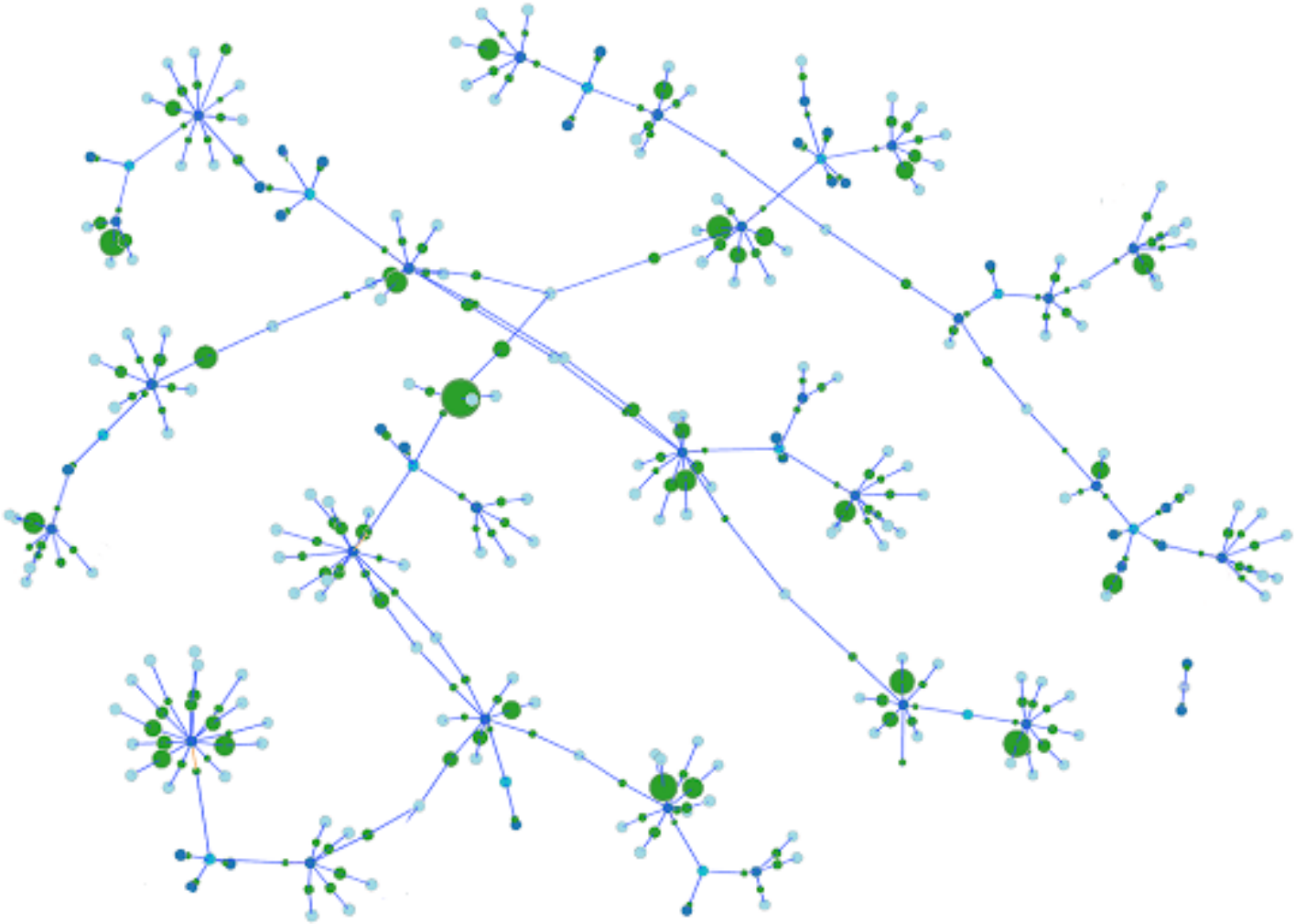
KINEVIZ

White Paper

Accelerating Fraud Detection with Visualization-First Analytics



Contents

Overview

What is fraud?

Fraud is a broad category of financial crime in which money is stolen using deceptive identity and/or misrepresentation of products, services, and financial transactions. Fraud is also an element of other types of cybercrime, especially money laundering. Perpetrators can be single individuals, but high-value fraud increasingly involves groups of individuals or business entities.

Impact of Fraud

As financial transactions transitioned from mostly in-person to a majority made online, fraudsters were quick to exploit that opportunity. By 2018, global annual losses from online credit card fraud alone was estimated to cost businesses $27.8 billion per year, and these losses are projected to rise to over $40 billion by 2028.

So far, recovery rates are low. Losses are often written off, ultimately increasing the cost to the entire customer base of the goods or services involved. For example, in the United States, insurance fraud alone costs at least $80 billion annually, much of which is passed on to consumers in the form of an estimated $400 to $700 in annual premiums.

Fraud Detection and Prevention

Businesses large and small have responded by developing processes and installing sophisticated fraud detection systems to detect fraud quickly and, if possible, to prevent it from occurring. Today all financial institutions and many businesses must include fraud detection capability as part of normal operations.

Important general requirements of a fraud detection system include capabilities to:

- Analyze and visualize transactions and user identities over time and location.
- Distinguish fraud events and patterns quickly enough to prevent completion of fraudulent transactions or to aid in recovery.
- Respond quickly to ever-evolving fraud tactics as new threat patterns are identified.

Additional business requirements include:

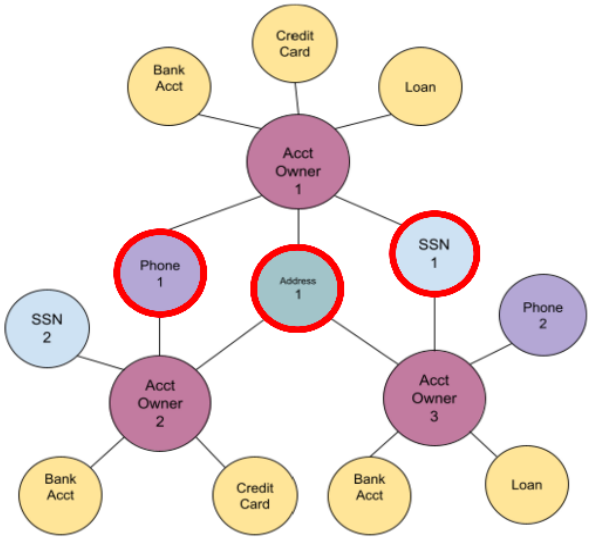
- Flag only actual fraudulent activities and not legitimate ones, to avoid losing customers and to maintain good customer relations.

- Support audit trails and logging of results. This is essential for enterprises that manage sensitive data in industries such as health care, financial services, insurance, and law enforcement.
- Support regulatory requirements to maintain detailed information on customer identity (also known as Know Your Customer or KYC), report large transactions, and report suspicious activity to regulators and law enforcement.

As a first step, fraud detection systems have incorporated business rules on top of existing relational databases. Examples of such rules include flagging high-value credit card charges or cash transfers, or implementing restrictions based on user identity and past transaction history. Once these rules are incorporated as standard practice, they can be quite effective at preventing monetary loss.

However, since the potential rewards of defrauding institutions are so great, fraudsters are continually deploying new tactics. At present, most of these involve networks of individual or business accounts (both real and fake) that were specifically created to evade detection. Individually, the accounts or contacts may look legitimate but through the combined action of the fraud network or ring, activity can be fraudulent, potentially resulting in large financial losses.

Expressed as a graph, a fraud ring might look like the schematic illustration below, in which the accounts (Bank, Credit Card, and unsecured Loans), are connected in a probable fraud ring through the account owners' shared social security numbers, phone numbers, and/or addresses.



When used as designed, relational or NoSQL data stores provide fast, secure access to data. But they are poorly suited to find hidden or rapidly changing patterns in data. What's increasingly needed are ways to analyze the connections between transactions and to visually highlight networks and clusters that have a high likelihood of being fraudulent.

Furthermore, to keep up with ever-evolving fraudster tactics, the analytic framework must be both flexible and extensible—difficult to achieve in the context of any relational database or document data store. Fortunately, these are the very problems that a graph data model, analytics, and rapid visualization can effectively address.

Types of Fraud

Detection and investigation of fraud will differ depending on how the fraud is perpetrated and the assets that are being stolen and/or hidden. General types of online fraud include:

- Credit Card / E-Commerce (In Person or Card-Not-Present (CNP). This can include purchases, loans, chargebacks, and/or ATM funds transfers, and can be characterized as first-, second- or third- party fraud, depending on whether the accounts are nominally controlled by their owners. Regardless of type, phone numbers, social security numbers, and/or addresses that may be shared among account holders can be visualized as a graph, in which patterns of suspicious connection and transfer of funds can be visualized.
- Insurance (Medical, Auto/Accident, Loss from Theft). This can involve one or a few individuals, or a highly sophisticated fraud ring involving many people who do not seem to be connected. Medical fraud can involve fraudulent billing for medical services not rendered. Insurance fraud may be online only, or staged to get photographs and other supporting evidence. A fake auto accident requires the collusion of accident Participants and Providers of services. Participants in this type of fraud ring will often take on different roles (Driver, Passenger, Witness, etc.) so as to increase the number of unique fake claims the fraud ring can submit. This is a complicated problem, however graph data can be examined for social connections and financial transfers between individuals involved in any given set of accidents.
- Sports (Horse Racing, Football, Soccer, etc. and associated Online Gambling). Personal and financial connections between the people who bring the events to the public, as well as those who regulate the terms of competition can reveal suspicious patterns of collusion or fraudulent activity.
- Money Laundering. Those involved in fraud rings do need to hide the source of transferred money or assets to make income seem legitimate, and patterns of those transfers can reveal how and when this is done. In general, money laundering is also involved in a much broader range of crimes such as securities fraud, environmental crime (illegal logging & mining etc.), tax fraud, sales of illegal drugs, human trafficking, and more.

Model and Analytics for Fraud Detection

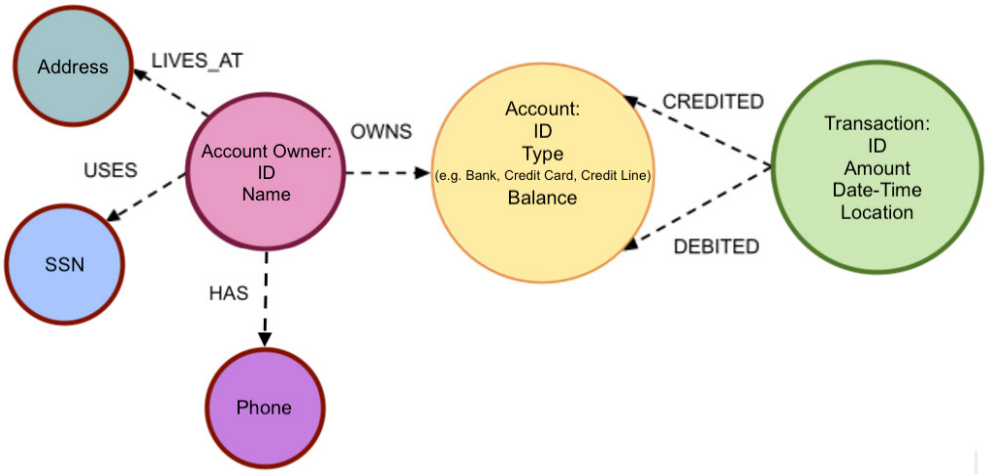
Detecting fraud quickly and more easily comes down to visualization and analysis of the connections between individual financial transactions with respect to:

- Identity of account holders (fraudsters as well as victims).
- Connections between unique accounts and account holders.
- Flow of assets or communication among accounts.

Although fraud strategies are ever-evolving and can be highly sophisticated, transaction data can be modeled as a graph to discover such connections quickly.

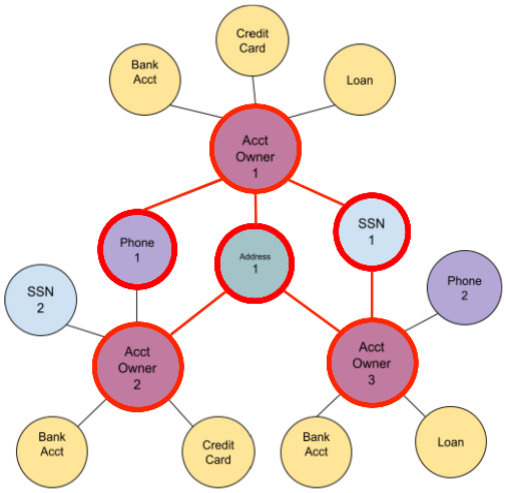
Transaction Data Model

A relatively simple data model can be used to visualize connections between accounts and their account holders in basic transaction data. With this model, the types of accounts and any account owners who are using the same social security number (SSN), phone or address can be visualized quickly and directly.



Graph Visualization

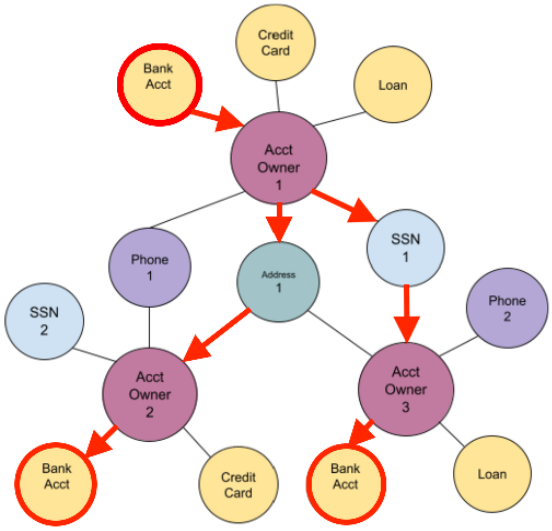
You can begin by viewing the entire graph and then exploring any strongly connected subgraphs that might signal fraudulent activity. As we've seen, such a graph visualization can quickly reveal key connections that flag potential fraud rings.



Focusing on the identity of the people involved, one can explore questions such as:

- Which account owners are the most strongly connected to each other?
- Who are these actors? Persons? Business entities? Are they real or fake?

A potentially more powerful visualization strategy is to focus on an entity such as a specific account or account holder, and expand the graph to explore connections from that starting point to other accounts, account holders, and transactions.



Focusing on assets and activities involved, one can explore questions such as:

- How much money (or unused credit) is represented in the connected accounts?
- Where did the transactions occur (IP address, physical Address information)?

Regardless of focus, the graph data model is unquestionably effective since it supports rapid exploration and visualization. And because the schema for graph data is extensible, you can build up a model from a simpler starting point and add to it as additional questions arise. A prime example is money laundering, which typically involves a variety of assets, large numbers of accounts, and complicated chains of transactions. These may include cash, credit cards, gift cards, and virtual assets (VA) such as Bitcoin. The assets may be transferred among accounts owned by many different individuals and businesses. Specific money laundering patterns differ—for example, laundering online proceeds from credit card identity theft or securities fraud, as opposed to cash from criminal activities like narcotics trafficking.

Visualization can be effective even with messy or incomplete data. A human investigator can often make sense of partial patterns, easily identify anomalies and outliers, and quickly focus on specific subgraphs for more detailed exploration. In addition, unlike black box graph analytics, exploratory analysis is transparent in that it can be explained and reproduced fairly easily.

Graph Analytics

Graph analytics can be efficient and effective, especially when investigating overall patterns of connection. Centrality or community detection graph algorithms provide measures of subgraphs or clusters in the data that can help highlight the presence and nature of fraud rings.

Algorithms such as Louvain provide a generalized measure of connectedness, while Centrality Degree indicates the actual number of connections in a cluster. Path finding algorithms can also illustrate specific chains of connection in graph data.

However, graph algorithms do rely on clean, more or less complete data. Also, while algorithms that can automate discovery of subgraphs are powerful, they are essentially black box processes whose results are not necessarily easy to explain or communicate. This is especially true of algorithms that have been developed for highly specific purposes.

Transaction Data for Testing

So where do we get data to test models and workflows for fraud detection? There is a vast amount of transaction data but in general it is private and not freely available. Transaction data can be anonymized, and this is an active area of research. However, it is a difficult problem, and current processes do not guarantee that anonymity can be maintained.

Simulated or generated data can be a useful alternative, and datasets of that type are increasingly available for testing. Often-used sources include:

- PaySim. Simulated transaction data for mobile money transfer, available as a Kaggle dataset. The Neo4j Fraud Detection Sandbox uses this dataset and recasts it as graph data. The sandbox example also shows how to use community detection algorithms to group and flag persons that may be participating in fraud rings.
- Machine Learning for Credit Card Fraud Detection. Describes elements of a typical Fraud Detection System (FDS). Discusses issues, and provides a Python notebook to simulate and work with transaction data.

Conclusion

Graph visualization can become an important ally in addressing the ever-changing challenges of combating fraud. The graph data model is designed to be flexible and extensible, which enables fast iteration of the cycle of exploration and analysis. In addition, a graph provides a rapid view of the connections and pathways between individuals and the assets they control— which is difficult or impossible with document or relational data stores. Visualization and analysis of graph patterns, whether clusters of connected entities in an entire graph, or an expanded network of connections from an individual entity of interest, can bring a more pinpoint focus to detection and deterrence on the ground.

References and Further Reading

Coalition Against Insurance Fraud. Anti-fraud alliance which reports fraud statistics by industry sector.
https://insurancefraud.org/fraud-stats/

Insurance Fraud Costs. 2021 review of insurance fraud costs to individuals.
https://www.nerdwallet.com/article/insurance/insurance-fraud-cost

Credit Card Fraud Statistics. Digest of worldwide losses and other statistics.
https://dataprot.net/statistics/credit-card-fraud-statistics/

Fraud Detection with Neo4j Sandbox. Developer blog exploring fraud detection.
https://neo4j.com/developer-blog/exploring-fraud-detection-neo4j-graph-data-science-part-1/

Protecting Privacy with Synthetic Data. May, 2022 article on technical challenges in ensuring privacy in anonymized and synthetic data.
https://sinews.siam.org/Details-Page/protecting-privacy-with-synthetic-data