HOW MODERN DATA SCIENTISTS IMPROVE PREDICTIONS

A Data Scientist's Guide to Graph Data Science

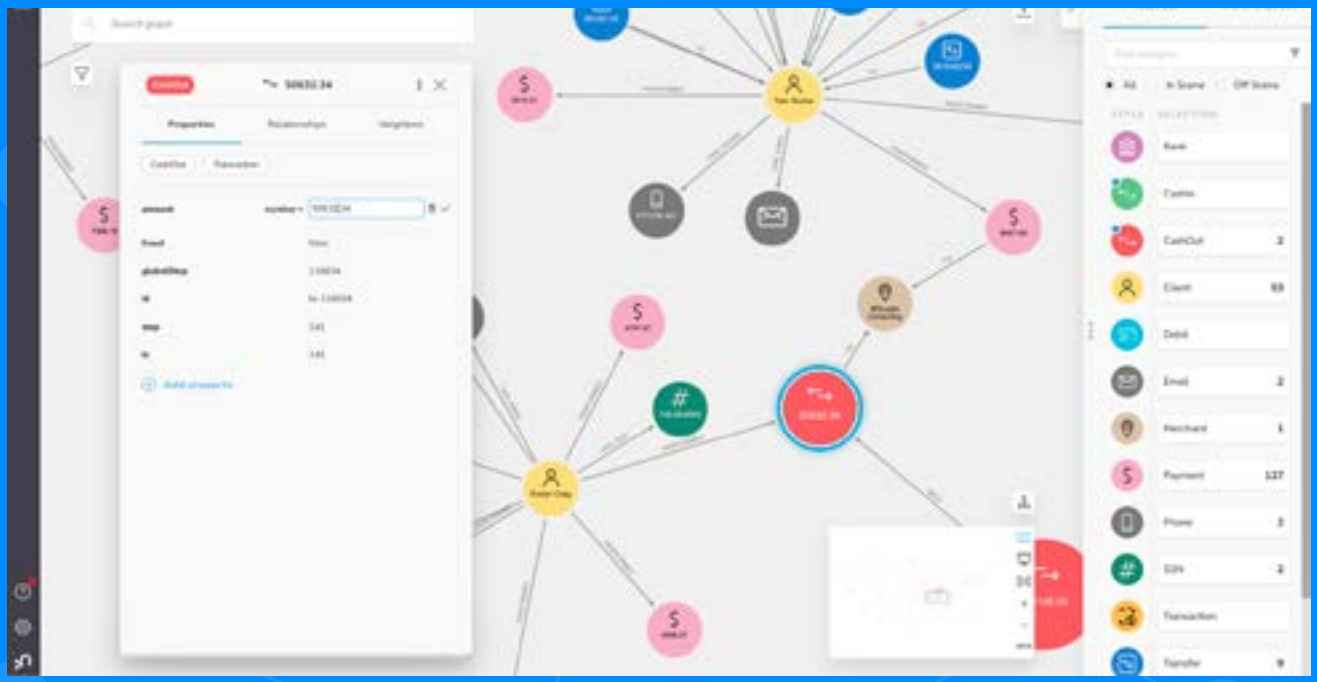
Introduction

You bring value to your organization by improving analytics and ML models so business stakeholders can make better, more informed decisions. Historically, data scientists have structured their data points in a record-oriented or tabular format (rows and columns). However, this approach does not address the connections and relationships between data points that can provide valuable insight and signals.

Unlike relational databases and other tabular formats, a graph gives data scientists the ability to visualize, explore, understand, and analyze connections between data points, providing critical context that is not otherwise captured in a tabular data model.

In a graph, you can see how strong connections are, where groups of connections form, how important each connection is, and how connections influence one another. Graph Data Science offers a fresh way to visualize and explore connections between your data. It also provides a meaningful new way to explain your findings to others.

You can answer a broader spectrum of questions about your data, and analyze the entire graph using graph algorithms that reveal what's unusual, what's important, and what's next. We'll describe the unique value of graph data science to help you decide if it is a good fit for your data science projects.



Explore your data visually to see the most well connected points in your graph.

The Graph Data Science Engine

A graph data science engine is a robust suite of capabilities that make it easier than ever before for you to leverage relationships in analytics and machine learning. This includes:

Graph queries that enable the fast identification of patterns in your data according to subject matter expertise and business rules.

Graph algorithms and unsupervised learning to identify new patterns, similarities, and connections across your entire graph or large portions of it.

Feature engineering and ML capabilities to leverage relationships in predictive inference, as well as identifying missing links and forecasting how relationships will form in the future.

These approaches help you better visualize, explore, and understand the connections in your data and see where new ones may form.

What Makes a Good Graphy Problem?

Effective data scientists know the best way to analyze their data based on the business need (the type of problem) and the data available. But graphs are a valuable tool when stakeholders ask **What's important? What's unusual? What's next?** When relationships provide the context to answer these questions, you have a graphy problem.

What's Important? (Prioritization)

There are numerous examples across departments where stakeholders try to determine project urgency and therefore, prioritization. For example:

- Marketing: What is the most important piece of content, the most important webpage, the most important call to action?
- Product Teams: Where is the most friction?
- Support: Which article is the most important?
- Finance: Which report is most important for leadership teams?

If you hear keywords such as best performing, most important, converting, or challenging, they indicate questions about importance.

What's Unusual? (Anomalies)

It is easy for us to immediately think of fraud detection when we see the question "what's unusual?" But other departments ask their data science counterparts to identify unusual behavior as well, from insider trading, data loss, network attacks, and more. Finding anomalous behavior is one of the most challenging, yet beneficial uses of graphs.

What's Next? (Predictions)

It's hard to predict the future, but Graph Data Science can help you make make the best decisions based on the information you already have. Graph recommendation engines can help determine the next best action, whatever your business need.

Graph predictions can deliver answers to these questions and more.

- Marketing: What email should we send customers next?
- Product Teams: What product should we build next?
- Retailers: What product should we sell next?
- Human Resources: What training should an employee take next?
- Finance: How should we price our products next quarter?
- Operations: What is the fastest path from point a to point b?

Common Uses for Graph Data Science

Next, let's walk through some common use cases for Graph Data Science. Because these are general use cases they can be applied across multiple lines of business. Think broadly about solutions and envision reuse.

Finance
- Fraud Detection
- Pricing Analysis
- Budgeting
- Forecasting

Marketing
- Customer 360
- Influencer Strategy
- Campaign Optimization
- Product Recommendations

Ops
- Product Development
- Pipeline Acceleration
- Supply Chain Optimization
- Infrastructure Planning

IT
- Network Monitoring
- Cybersecurity
- DevOps

HR
- Training
- Upskilling & Retention
- Promotions

> *80% of data and analytics innovations will use graph technologies by 2025.*
>
> *Gartner**

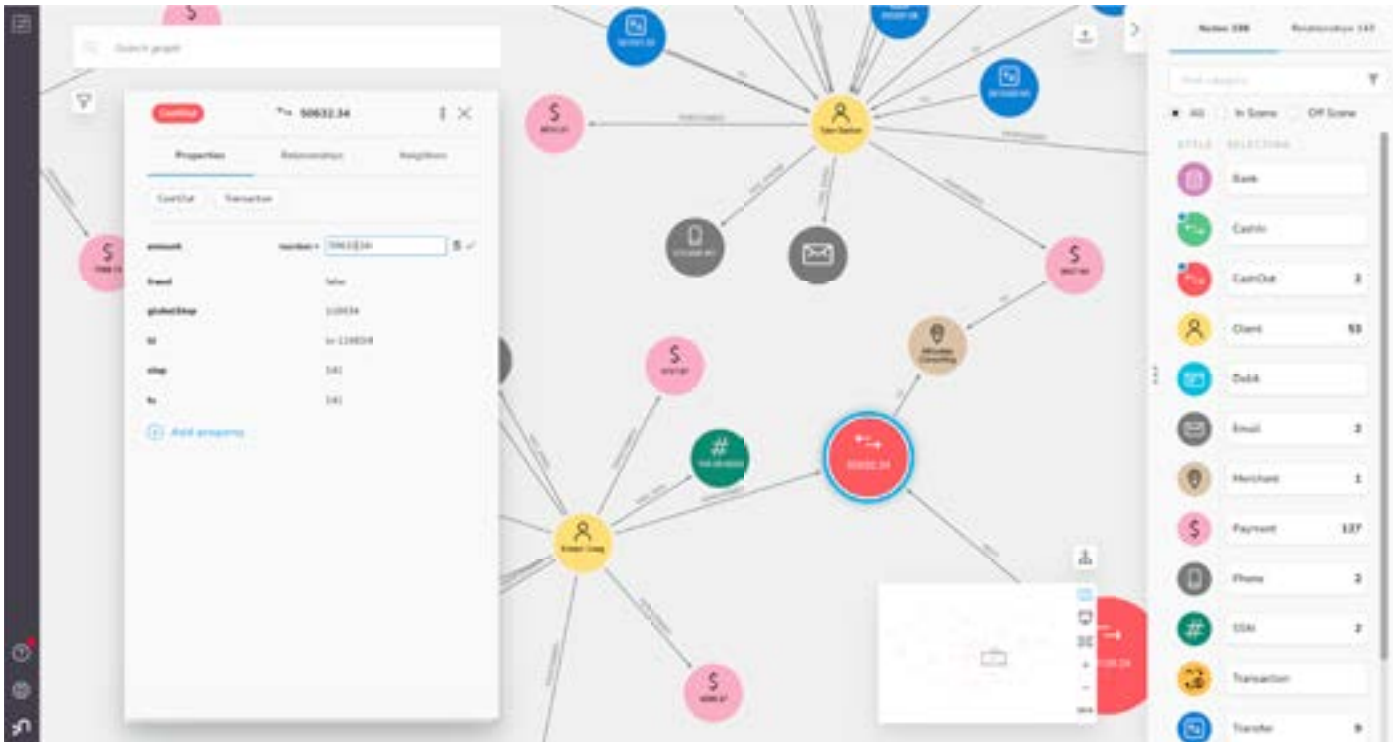
Anomaly and Fraud Detection

Anomaly detection across corporate networks can help to identify cybersecurity attacks and prevent data loss. The same strategy used to identify threat actors in a cybersecurity context can be used to detect fraud in banking, insurance, and government programs by analyzing the relationships and behaviors as a graph.

Effective fraud detection remains one of today's most challenging data science problems. This is because fraud detection has a unique quality – the entity you want to predict is trying assiduously to prevent you from doing so. It's not just a "needle-in-a-haystack" problem. In this case, the proverbial needle is actively trying to hide.

Fortunately, graph-based approaches explicitly model relationships between entities in the data and this, coupled with a graph data science engine, can empower practitioners to rapidly explore, visualize, resolve, and predict fraud entities and patterns. These patterns are obfuscated and challenging to infer in other data models.

*Top 10 Trends in Data and Analytics, 2021, Gartner, Inc.



Easily view relationships to identify communities.

Customer 360

Across the globe, businesses try to better understand their customers and improve lifetime value (LTV). With graph data science, customer knowledge can become more accurate and complete through entity resolution. This process looks at all the database entries and identifies duplicates to create a complete master entry for each customer. Individual customer profiles improve LTV, deepen customer knowledge, and enable marketing organizations to optimize programs and offers.

Whether your work supports retail, media and marketing, or banking and finance, a consistent and accurate view of the people, places, and organizations in your data is central. Because you work with multiple data streams of varying quality, it's important to consolidate and disentangle the data. Entity resolution is an extremely powerful approach to improving customer knowledge by distilling data into unique and valuable master profiles. As a data scientist, entity resolution gives you confidence to better understand the massive amounts of customer data your organization collects to turn insights into actions.

Recommendation Engines

Recommendation engines are well known through Netflix and online shopping experiences. However, recommendation engines have uses across the business. Power the most important parts of your business with graph recommendation engines – from product development to employee retention initiatives, and more.

Companies must be concerned not only about the quality of their recommendations (are we suggesting the right thing?), but also about how quickly they can derive relevant recommendations and serve them to their users (are we suggesting the most up-to-date thing?). No one likes seeing an advertisement for a pair of shoes they bought three weeks ago following them around on the web.

Given the intricacies of recommender systems and the risks of serving the wrong recommendations, it is necessary to leverage advanced analytics to ensure that customers receive the most relevant and timely recommendations. A graph data science engine can analyze the connections in your data to power recommendation engines that

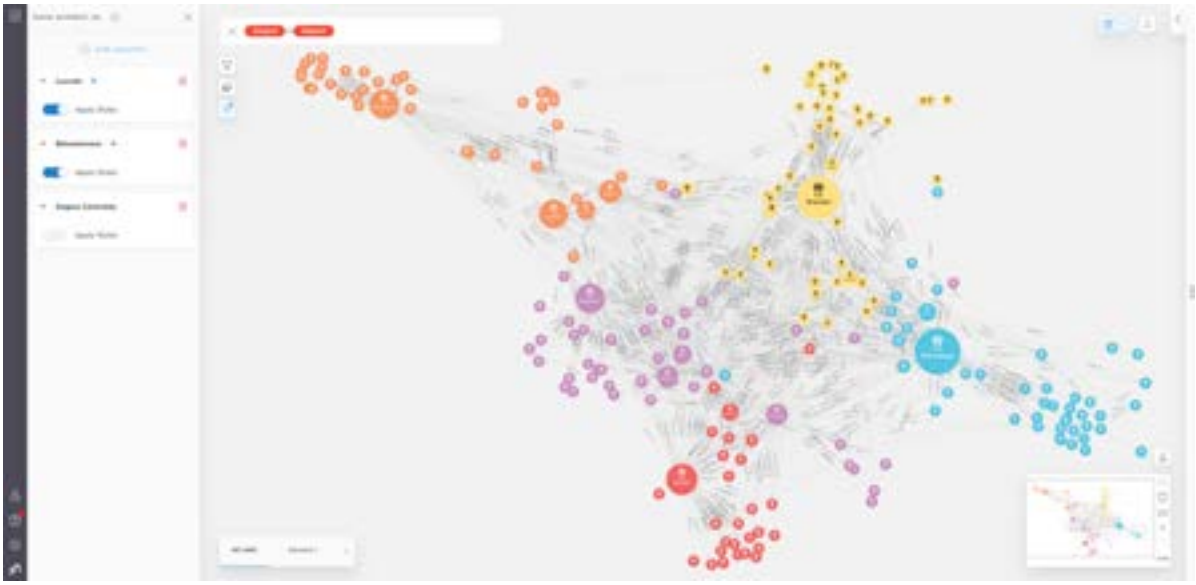
fuel customer satisfaction and business growth, gaining otherwise unattainable insights from the relationships that exist in the data you already have.

Supply Chain Optimization

Improving a supply chain leads to savings, not just in dollars, but also in carbon emissions. Every optimized route and prompt delivery mean happier customers, fewer emissions, and time savings. Graph Data Science helps you optimize supply chain routes by finding the best path, balancing cost and efficiency with customer satisfaction and sustainability.

Supply chains are inherently complex, involving multiple stages, inputs, outputs, and interconnectivity. Looking at supply chain data in raw form (tables) can be daunting, and extracting valuable insights can be difficult and unintuitive. Luckily, supply chains are intrinsically structured as a graph – a network of stages and interconnecting arcs, which in graph terminology we refer to as nodes and relationships, respectively. Visualizing supply chains in a graph shows how raw materials, manufacturers, distributors, retailers, and customers are all connected and better represents real world behavior.

By further coupling graph with a graph data science engine, you will be empowered to make more advanced inferences and gain insights that would otherwise remain hidden or challenging to uncover using other data models.



Visualize and extract insights using graph algorithms such as Community Detection, which finds interdependencies within a network, and Centrality, which shows what is most influential and well connected.

Learn More!
Read our Use Case Selection Guide

The Power of Graph Algorithms

Graph algorithms are a set of instructions that analyze relationships in connected data. Some algorithms are used to find a specific node or the path between two given nodes. Leverage pretuned graph algorithms to help you scale rapidly from proof of concept to production.

Community detection algorithms cluster your graph based on relationships to find communities where members have more significant interactions. Detecting communities helps predict similar behavior, find duplicate entities, or prepare data for other analyses.

> These include Louvain, Triangle Count, Strongly Connected Components, Label Propagation, and more.

Centrality algorithms reveal which nodes are important based on graph topology. They identify influential nodes based on their position in the network and are used to infer group dynamics such as credibility, rippling vulnerability, and bridges between groups.

> These include Page Rank, Betweenness Centrality, Closeness Centrality and others.

Similarity algorithms employ set comparisons to score how alike individual nodes are based on their neighbors or properties. This approach is used in applications such as personalized recommendations and developing categorical hierarchies.

> These include Node Similarity, Filtered Node Similarity, K-Nearest Neighbors, and others.

Pathfinding algorithms find the most efficient or shortest paths between nodes. Use them to analyze complex dependencies and evaluate routes for uses such as physical logistics and least-cost call or IP routing.

> These include Dijkstra and Yen's shortest path, Steiner and Spanning Trees, and Random Walk.

Node embedding algorithms transform the topology and features of your graph into fixed-length vectors that represent each node. They capture the complexity and structure of a graph and transform it for use in various ML tasks.

> These include HashGNN, Node2Vec, FastRP, and GraphSAGE.

Graph-native ML techniques like link prediction and node classification fill in the blanks in your data and predict changes in the structure of your graph. They enable use cases such as fraud detection, drug discovery, entity resolution, and more.

> These include Node Classification Pipelines, Link Prediction Pipelines, and Node Regression Pipelines.

Get Smarter Predictive Analytics with Graph Data Science

When you analyze data in tabular form, as you do in a relational database, you try to make sense of data points without a coherent way to analyze their connections. It's like trying to solve a Rubik's Cube by only looking at one side.

A graph gives you the ability to visualize, explore, understand, and analyze the connections between each data point. This adds context to the data that is nearly impossible to get from a tabular data model. In a graph, you can see how strong connections are, where groups of connections form, how important each connection is, and how connections influence one another.

Graph Data Science empowers you to see the relationships in your data to put data in context and answer pressing questions.

Neo4j Graph Data Science

Neo4j Graph Data Science is an analytics and ML engine that uses the relationships in your data to discover fast, actionable insights and plugs into enterprise data ecosystems so you can get more data science projects into production quickly.

Using a library of pretuned graph algorithms, you can explore billions of data points in seconds and generate high-impact ML models to identify hidden connections and create compelling visualizations that lead to better stakeholder decision making.

Neo4j Graph Data Science can be deployed on-premises or in the cloud as a self-managed and fully-managed cloud offering. It fits in easily with your existing data stack and pipelines, offering groundbreaking insights.

See how graph data science can answer your data questions.
Read our Use Case Selection Guide

Neo4j is the world's leading graph data platform. We help organizations – including Comcast, ICIJ, NASA, UBS, and Volvo Cars – capture the rich context of the real world that exists in their data to solve challenges of any size and scale. Our customers transform their industries by curbing financial fraud and cybercrime, optimizing global networks, accelerating breakthrough research, and providing better recommendations. Neo4j delivers real-time transaction processing, advanced AI/ML, intuitive data visualization, and more. Find us at neo4j.com and follow us at @Neo4j.

Questions about Neo4j? Contact us around the globe:

info@neo4j.com
neo4j.com/contact-us