Chapitre 1

Introduction to Graph Data Science

Overview of graph data science as a method to explore contextual relationships in data.

Our daily lives are full of graphs, from social media to the maps we use to drive to work, to the recommendations provided to us on our favorite TV streaming network.

We will analyze Python and Neo4j as the tools to learn and explore graphs. These tools offer extensive libraries as well as robust community support, which makes them a great choice for the journey of graph data science.

Structure

In this chapter, the following topics will be covered:

- Understanding Graphs, Graph Networks, and their Relevance
- Introduction to Neo4j Graph Database
- Overview of the Importance of Graph Visualizations
- Data Science and Machine Learning
- Introduction to Graph Data Science
- Introduction to the Python Programming Language

Data Science vs Machine Learning

Data Science is a multidisciplinary field that involves **extracting knowledge** and **insights from data through various techniques** such as **data mining, data visualization, and statistical analysis.** Data science involves the end-to-end process of acquiring, cleaning, transforming, and analyzing data to uncover patterns, make predictions, and drive better decision-making:

Machine Learning, on the other hand, is a subset of data science that focuses on developing algorithms and models that enable computers to learn from data and make predictions or take actions without being explicitly programmed. Machine learning algorithms learn from historical data to identify patterns, make predictions, and automate decision-making processes on new, never-before-seen data.

Data science provides the foundation and tools to explore, interpret, and gain insights from data, while machine learning leverages the data to build predictive models and make accurate predictions and/or automated decisions.

Together, the two form a powerful combination that drives innovation and enables data-driven solutions.

Defining Graph

In discrete mathematics and graph theory, a graph is a structure that consists of objects or nodes where pairs of objects or nodes are connected or related in some way.

These objects can be referred to as vertices, nodes, or points. We will refer to these objects as nodes. The connections between the vertices are referred to as edges, relationships, or links We will refer to the connections between nodes as relationships.

Neo4J - native graphe database

Neo4j is a highly scalable, native graph database designed to store and process graphs efficiently. In Neo4j, data can be stored on both nodes and relationships. We will refer to this data as properties of either the node or relationship:

Key features of Neo4j include:

- ☐ **Native Graph Database**: Neo4j is designed from the ground up as a graph database, meaning it's optimized for storing and querying graph data structures.
- ☐ **Cypher Query Language**: Neo4j uses Cypher, a declarative query language, to interact with the database. Cypher allows users to express graph patterns and operations concisely and intuitively.
- ☐ **ACID Compliance**: Neo4j ensures data integrity by providing ACID (Atomicity, Consistency, Isolation, Durability) compliance, making it suitable for applications that require strong transactional guarantees.
- ☐ **Scalability**: Neo4j is built to scale both vertically and horizontally, allowing users to handle large datasets and high transaction volumes.
- ☐ **Flexible Data Modeling**: Neo4j's graph model allows for flexible and expressive data modeling, enabling users to represent complex real-world relationships easily.
- ☐ **Community and Enterprise Editions**: Neo4j is available in both Community and Enterprise editions. The Community edition is free and open-source, while the Enterprise edition offers additional features, support, and scalability options.

Neo4j is widely used in various domains, including

- **social networking**,
- **recommendation systems**,
- network and IT operations,
- **fraud detection**, and
- **knowledge graphs**.

Its ability to represent and query highly connected data makes it a powerful tool for solving complex problems in these domains.

Consider your social circle or network of friends. Each person can be represented as a node, while the connections or friendships between individuals can be represented as a relationship. This forms a social graph that helps us understand how people are interconnected and how information flows within our social circles.

- **In the realm of internet** search, graphs are central to search engines like Google. When you enter a search query, Google analyzes the web graph to determine the relevance and importance of different web pages. This enables the search engine to provide you with the most relevant results, considering factors like page popularity, linking patterns, and the overall structure of the web graph.
- **In transportation and logistics**, graphs play a crucial role in route planning and optimization. Road networks can be represented as graphs, with intersections as nodes and roads as relationships. By analyzing this transportation graph, algorithms can calculate the shortest or fastest routes from one location to another, helping us navigate efficiently.
- **In banking**, graphs play a crucial role in detecting potentially suspicious and fraudulent behavior in transaction data. Banks analyze transactional data using graph-based techniques to identify patterns and anomalies that may indicate fraudulent activities. **By representing customer accounts, transactions, and their relationships as a graph, banks can detect unusual connections**, such as **multiple accounts linked to a single individual**, **unexpected money flows**, or **patterns resembling known fraud schemes**.

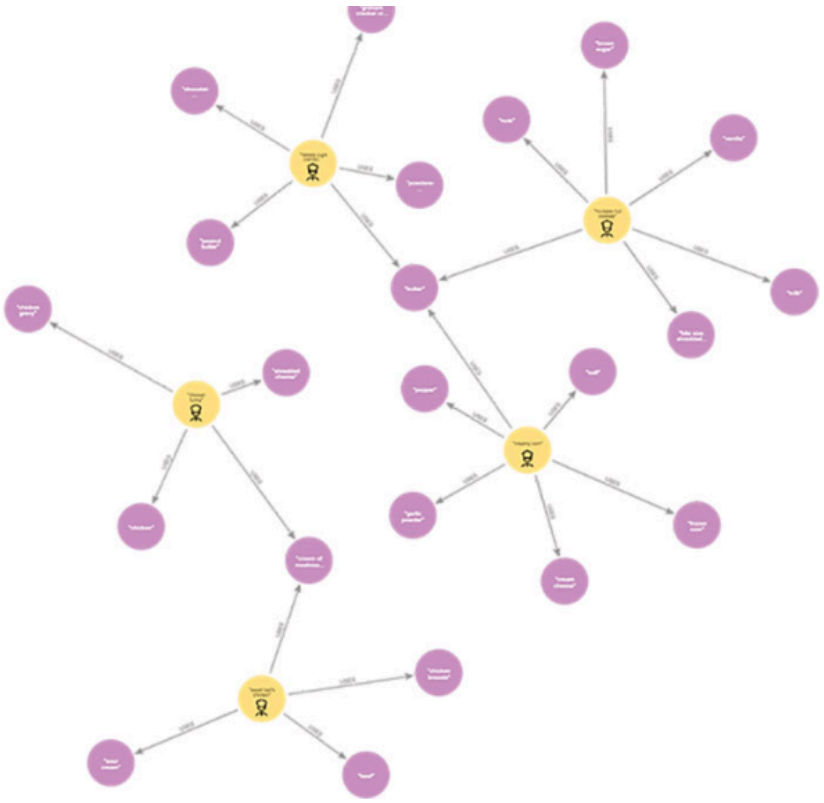
  The graph-based analysis enables banks to proactively monitor and mitigate risks, safeguarding the integrity of financial systems and protecting customers from fraudulent activities. Graphs provide a comprehensive view of transactional networks, empowering banks to stay one step ahead in the constant battle against financial crime.
- **In recommendation systems** : Graphs are also integral to recommendation systems that suggest products, movies, or music based on our preferences. By analyzing our past interactions, such as purchases, ratings, or clicks, recommendation engines create a personalized graph of our preferences. They then utilize graph-based algorithms to find similar users or items, enabling them to suggest relevant and personalized recommendations.

  vector databases have increased in popularity since the release of **ChatGPT** and other more recent **large language models (LLMs)**

Visualizing networks data - Tools

There are several vendor tools and open-source options available for visualizing networks as well, which are outside the scope of this book, but are mentioned here:

- **Linkurious**: It is widely used in various fields such as fraud detection, intelligence analysis, and cyber-security.
- **Hume by GraphAware**: Hume is built on top of Neo4j and aims to help organizations derive insights from connected data through visualization, analysis, and machine learning. Hume allows users to create visualizations that reflect real-world entities and relationships, making it easier to understand complex networks.
- **VizNetwork package in R:** VizNetwork is a package available in the R programming language. It is geared toward the visualization of networks and graphs. VizNetwork allows users to create interactive network graphs and is highly customizable, allowing for different types of nodes, edges, and graph layouts.
- **Gephi**: Gephi is an open-source network visualization and analysis tool. It allows users to explore data through various graphs, interactive
- **D3.js**: D3.js is a JavaScript library for manipulating documents based on data. It can be used to create intricate, interactive graph visualizations in web browsers.
- **Power BI**: Microsoft's Power BI has some network visualization plugins that can be used to incorporate graph data visualizations into broader BI dashboards and reports. However, as of the publication of this book, Power BI's network visualization capabilities are relatively limited. Users are restricted to specific visualizations without much ability to customize visualizations.



Graph visualization with Neo4j Bloom

Summary

In summary, this chapter lays the foundation for understanding graph data science and its relevance in real-world applications. In the next chapter, we will walk step-by-step through the installation and configuration of Python and Neo4j, so that you can follow along with code examples throughout the book.