

A brief introduction to Neo4j

Neo4j (<https://neo4j.com/>) is an interactive, native graph database, purpose-built to leverage data relationships. The objects studied in each analysis are the typical components of a graph, such as nodes, relationships (edges), paths between the nodes, node centrality, and the type of relationships that connect nodes. In Neo4j, data is stored in the form of nodes, relationships, and properties. Each node and relationship can have any number of properties. Both nodes and relationships can be labeled. Labels are used to group nodes and to narrow searches.

Neo4j allows us to run queries written in the **Cypher** query language over a data model expressed as a graph, and to work with the results in both graph visualization and tabular forms. If a Cypher query returns graph entities (nodes, relationships, paths, etc.), Neo4j Browser renders a graph visualization that can be explored interactively. However, many queries return tabular data, for instance as the result of aggregations. It is also possible to return simple rows of values; in this case, Neo4j Browser renders the output as a table.

Introductory tutorials are available at <https://neo4j.com/docs/getting-started/>

Introduction to Cypher is available at <https://neo4j.com/docs/getting-started/cypher-intro/>

Neo4j Sandbox

The **Neo4j Sandbox** enables you to get started with Neo4j, with built-in guides and sample datasets for popular use cases. The sandbox comes pre-loaded with sample data and a step-by-step guide with queries and explanations. You can start a free sandbox from <https://neo4j.com/sandbox/>. Besides the Movies sandbox that we already used during class, and which you are encouraged to practice with, the sandboxes available for your presentation, pre-loaded with data (classified as 'For Developers' and 'For Data Scientists'—but don't worry about this classification), are:

- **Network and IT Management:** Dependency and root cause analysis plus more for network and IT management.
- **Russian Twitter Trolls:** Explore data released by NBC News from their investigation into Russian Twitter Trolls around the 2016 US election.
- **Crime Investigation:** Explore connections in crime data using the POLE (Person, Object, Location, Event) model in a public dataset from Manchester, U.K.
- **Paradise Papers by ICIJ:** Explore the Paradise Papers dataset from the International Consortium of Investigative Journalists (ICIJ).
- **ICIJ FinCEN Files Investigation:** Explore large volumes of Suspicious Activity Report (SAR) filings between entities around the globe using the FinCEN Files investigation.
- **ICIJ Offshore Leaks:** Explore ICIJ Offshore Leaks database, which contains information on offshore entities that are part of the Pandora, Paradise, Panama Papers and other investigations.
- **Graph Data Science:** Leverage Neo4j Graph Data Science library to explore graph algorithms for analytics and feature engineering; the dataset shows the connections between different airports across the world.
- **Women's World Cup 2019:** Explore the data behind the Women's World Cup with the World Cup Graph.
- **OpenStreetMap:** Identify points of interest and routing with Neo4j using the OpenStreetMap dataset of Central Park in New York City.
- **Recommendations:** Generate personalized real-time recommendations using a dataset of movie reviews.
- **Contact Tracing:** Explore contact tracing using a synthetic dataset of places, persons, and visits.
- **Fraud Detection:** Identify fraud detection with the Paysim financial dataset and Neo4j Graph Data Science.
- **Healthcare Analytics:** Load and analyze FDA Adverse Event Reporting System data with Neo4j. The FDA Adverse Event Reporting System is an information database designed to support the U.S. Food and Drug Administration's post marketing safety surveillance program for all approved drug and therapeutic biologic products.
- **Twitch:** Explore data related to Twitch social network, an online platform that allows users to share their content via live stream.
- **Yelp Public Data:** Explore real-world data related to businesses including reviews, photos, check-ins, and attributes like hours, parking availability, and ambience.
- **Stack Overflow:** Explore data related to Stack Overflow questions, answers, tags, and comments and the relationships between them.
- **Entity Resolution:** Entity resolution is the process of disambiguating data to determine if multiple digital records represent the same real-world entity such as a person, organization,

place, or other type of object. This use case is about entity resolution for an online movie streaming platform.

- **Cybersecurity:** This sandbox is based on the data and themes from the BloodHound project, which is a tool for auditing an Active Directory environment. Active Directory helps IT teams monitor various network resources and users and allows to grant and revoke different user permissions. Using Neo4J and Neo4j Graph Data Science library, you can analyze possible attack paths based on access.

By default, each sandbox you create is available for **3 days**, but you can extend this for 7 additional days before expiration (making it a total of 10 days). Once the sandbox expires, you will lose any data you have created and you need to create a new sandbox.

To display the graph model in terms of node types and relationship types, use:

```
CALL db.schema.visualization()
```

To learn how many nodes there are in the graph, use the `apoc.meta.stats` procedure:

```
CALL apoc.meta.stats() YIELD labels
```

In the following, there is a more detailed overview of some of the sandboxes available at <https://neo4j.com/sandbox/>

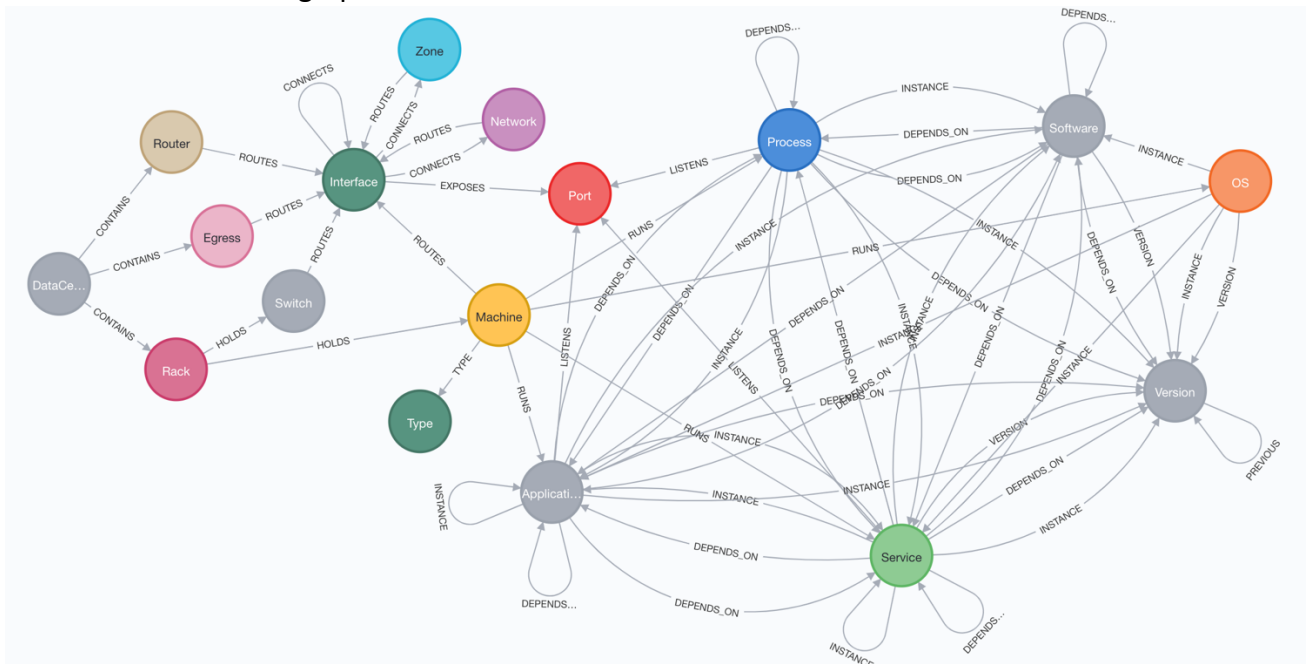
Network and IT Management

Computer networks span all layers of the stack, from physical connections to mobile and web applications that connect networks of users. A graph database provides a natural way to model, store, and query all types of computer networks. A graph database like Neo4j can be used for:

- Configuration management;
- Impact analysis;
- Planning;
- Security and hardening of networks;
- Intrusion detection;
- Network traffic analytics;
- User behavior analytics.

In this sandbox, the focus is on network management and impact analysis from the level of routing (TCP/IP network protocols) upwards to managing applications and tracing their dependencies.

The data model of the graph is shown below.



Russian Twitter Trolls

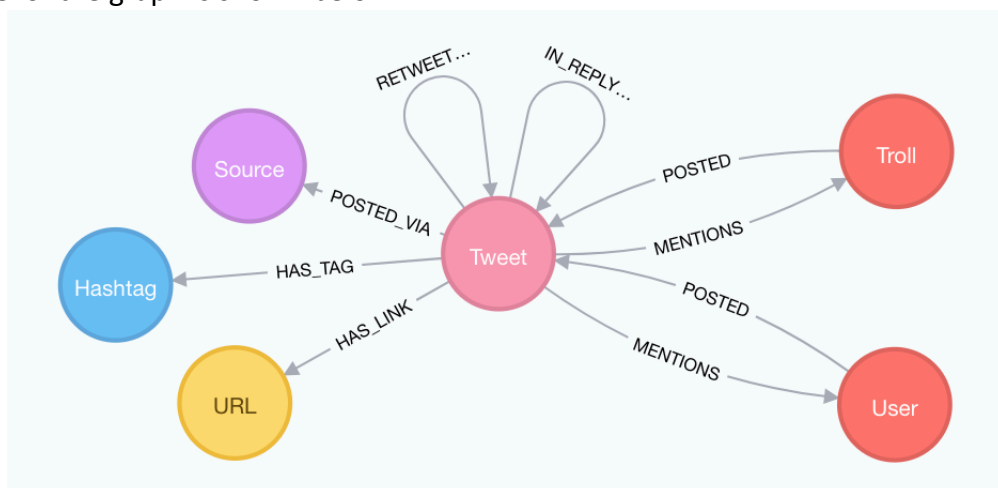
As part of the House Intelligence Committee investigation into how Russia may have influenced the 2016 US Election, Twitter released the screen names of almost 3000 Twitter accounts believed to be connected to Russia's Internet Research Agency, a company known for operating social media troll accounts. Twitter immediately suspended these accounts, deleting their data from Twitter.com and the Twitter API.

A team at NBC News was able to reconstruct a dataset consisting of a subset of the deleted data for their investigation and using Neo4j were able to show how these troll accounts went on attack during key election moments (see <https://www.nbcnews.com/tech/social-media/russian-trolls-went-attack-during-key-election-moments-n827176>). NBC News open-sourced the reconstructed dataset and released it as this Neo4j database.

This Neo4j sandbox will help you to explore the dataset of Russian Troll tweets by guiding you through:

- an overview of the data model;
- how to explore the data using Cypher;
- some of the investigative queries used to make sense of the dataset.

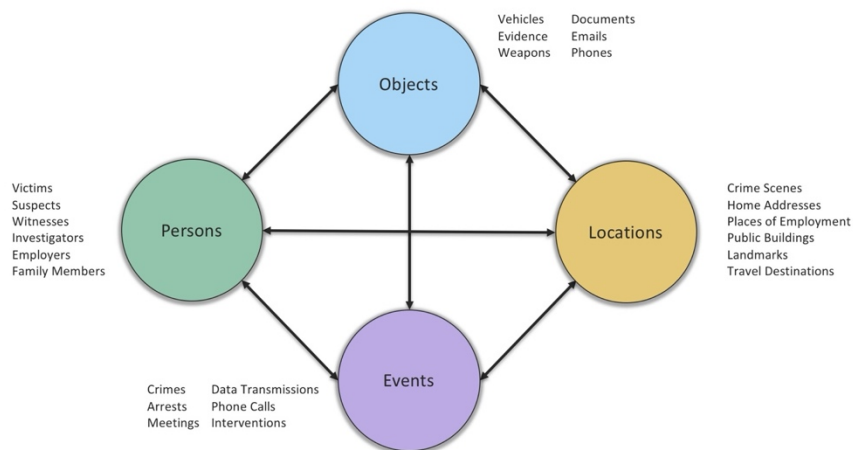
The data model of the graph is shown below.



Crime Investigation

This sandbox will demonstrate how Neo4j can be used with a POLE data model in the context of police and child protection investigations.

The POLE data model focuses on four basic types of entities and the relationships between them: **P**ersons, **O**bjects, **L**ocations, and **E**vents, as shown in the figure below. It is a standard approach used in policing, investigative, and security use cases. It can also, however, be applied in other areas. Typical POLE use cases include: policing, counter terrorism, border control/immigration, child protection/social services.



Using this sandbox, you will learn:

- How crime data can be modelled in a graph;
- How to query the graph and answer questions using Cypher;
- How to use spatial and aggregation functions in Cypher;
- How to use the built-in Cypher shortest path algorithm.

Crime data for this sandbox was downloaded from a public open data repository (<https://data.gov.uk/>) and is related to the city Manchester in UK.

The data model of the graph is shown below.



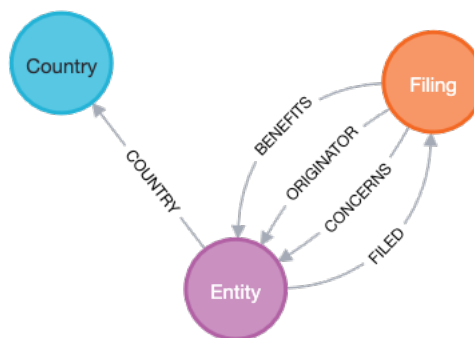
ICIJ FinCEN Files Investigation

The International Consortium of Investigative Journalists (ICIJ) has published a new investigation, the FinCEN Files, which exposed a vast network of industrial-scale money laundering running through Western banks and generally ignored by U.S. regulators – and they used Neo4j to help crack the case wide open.

The results draw from more than 2100 suspicious activity reports (SARs) between 1997 to 2017, which accounted for transactions of more than \$2 trillion USD in dirty money. These reports were filed by banks and financial firms with the U.S. Department of Treasury's Financial Crimes Enforcement Network (FinCEN) but were largely ignored or overlooked.

Using this sandbox, you will explore the dataset and also apply graph algorithms such as the Louvain algorithm.

The graph data model is shown below.



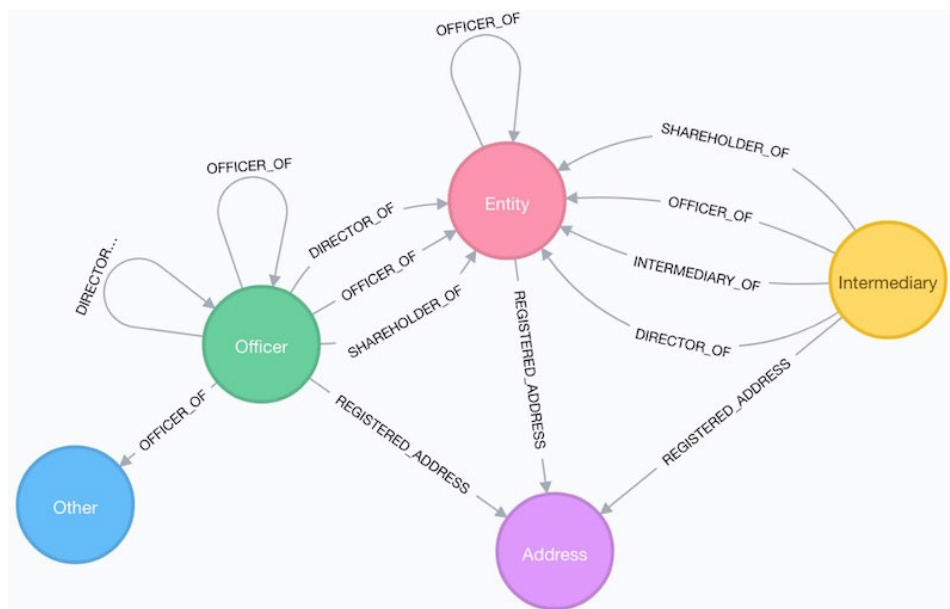
ICIJ Offshore Leaks

ICIJ Offshore Leaks database (<https://offshoreleaks.icij.org/>) contains information on more than 800,000 offshore entities that are part of the Pandora, Paradise, Panama Papers and the other Offshore Leaks investigations. The data covers a long time of activities – and links to people and companies in more than 200 countries and territories.

The real value of the database is that it strips away the secrecy that cloaks companies and trusts incorporated in tax havens and exposes the people behind them. This includes, when available, the names of the real owners of those opaque structures. In all, it reveals more than 500,000 names of people and companies behind secret offshore structures. They come from leaked records and not a standardized corporate registry, so there may be duplicates.

The Offshore Leaks Database was imported into Neo4j to be used by journalists and researchers to take advantage of the connections in the data.

Below you can see a simplified diagram how the nodes connect to each other.



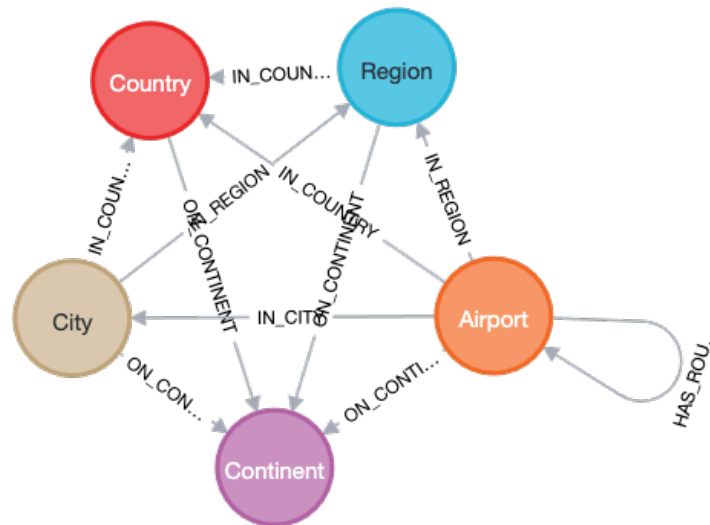
Graph Data Science

The Neo4j Graph Data Science (GDS) library contains a set of graph algorithms, exposed through Cypher procedures. Graph algorithms provide insights into the graph structure and elements, for example, by computing centrality and similarity scores, and detecting communities.

The example dataset used to demonstrate the GDS library considers the connections between different airports across the world.

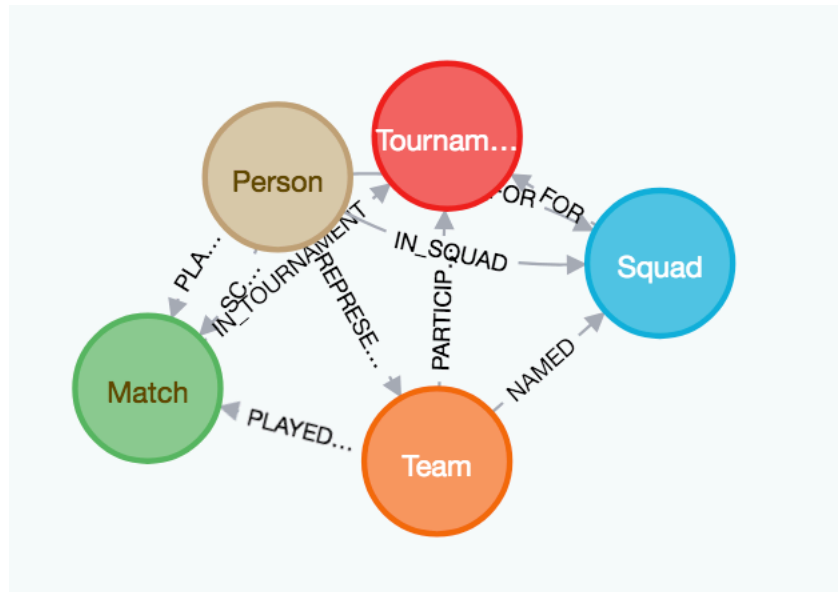
Using this dataset, you will first conduct some Exploratory Data Analysis (EDA) and calculate some summary statistics on the data. Then, you will explore the application of centrality metrics via PageRank and community detection via the Louvain algorithm.

The data model of the graph is shown below.



Women's World Cup 2019

This sandbox allows you to play around with the Women's World Cup 2019. All the matches, squads, lineups, and scorers from all the World Cups between 1991 and 2019 have been loaded into the dataset.



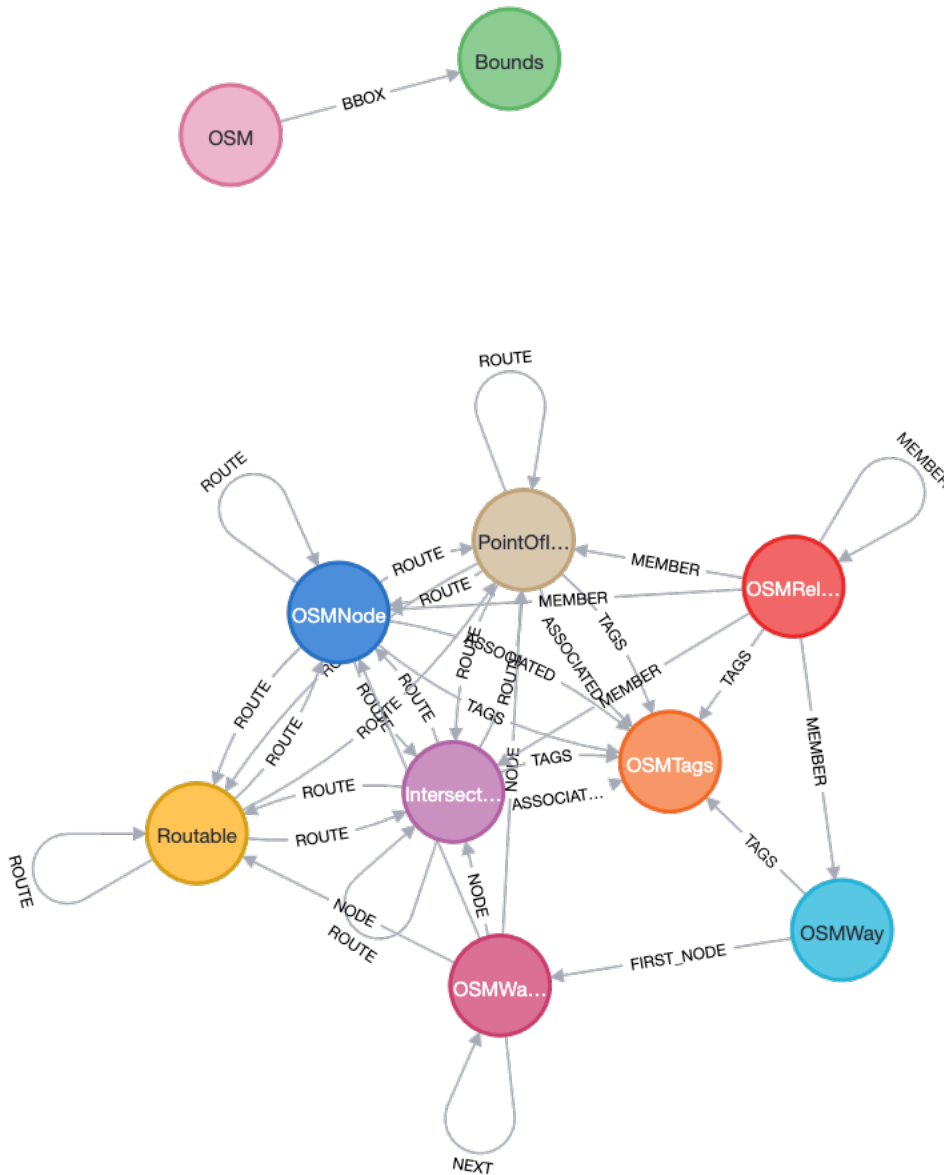
OpenStreetMap

This sandbox is based on global geodata provided by OpenStreetMap (OSM, see <https://www.openstreetmap.org>) and allows you to explore routes and tagged Points of Interest for Central Park in New York City.

In this sandbox, you will also find shortest paths.

See also <https://medium.com/neo4j/new-sandbox-in-town-e126246d2605>

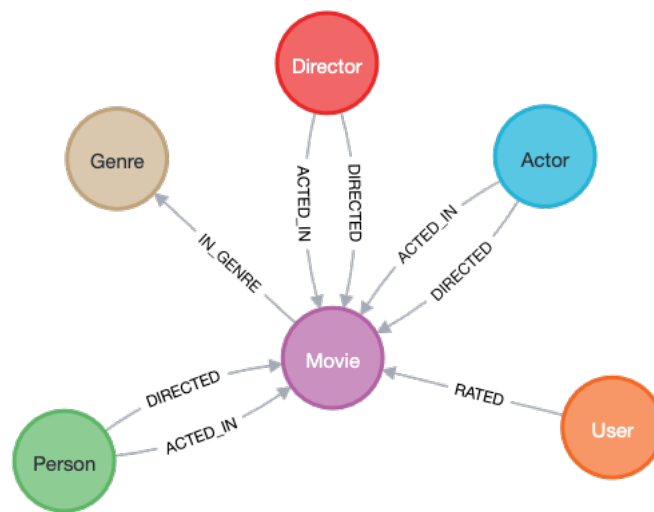
The graph data model is shown below.



Recommendations

Personalized product recommendations can increase conversions, improve sales rates and provide a better experience for users. In this Neo4j sandbox, you will take a look at how you can generate graph-based real-time personalized product recommendations using a dataset of movies and movie ratings. The dataset is from Open Movie Database (<http://www.omdbapi.com>). The sandbox allows you to apply content-based filtering and collaborative filtering. This sandbox also uses GraphQL (<https://graphql.org>), another query language.

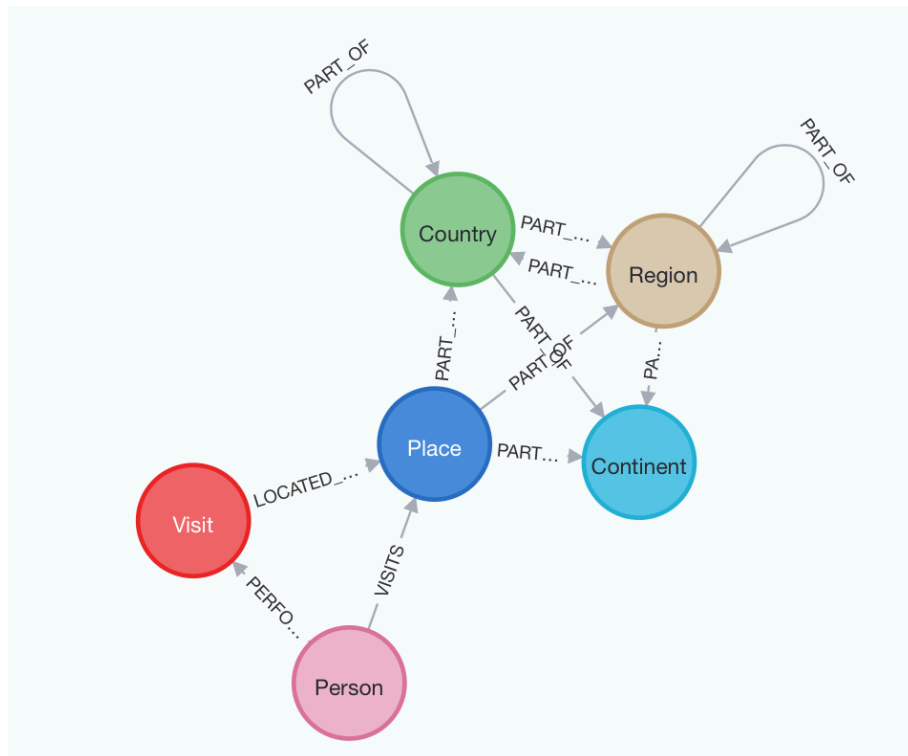
The graph data model is shown below.



Contact Tracing

This sandbox is based on a series of blog posts about contact tracing with Neo4j. It uses a synthetic dataset imported into Neo4j, on which you can run some interesting queries (i.e., potential infection spread, tracking visits by sick persons, measuring infection risk). You will also use Neo4j Graph Data Science (GDS) library on this dataset, and understand some of the predictive metrics like PageRank, Betweenness Centrality and use community detection to direct policies.

The graph data model is shown below.

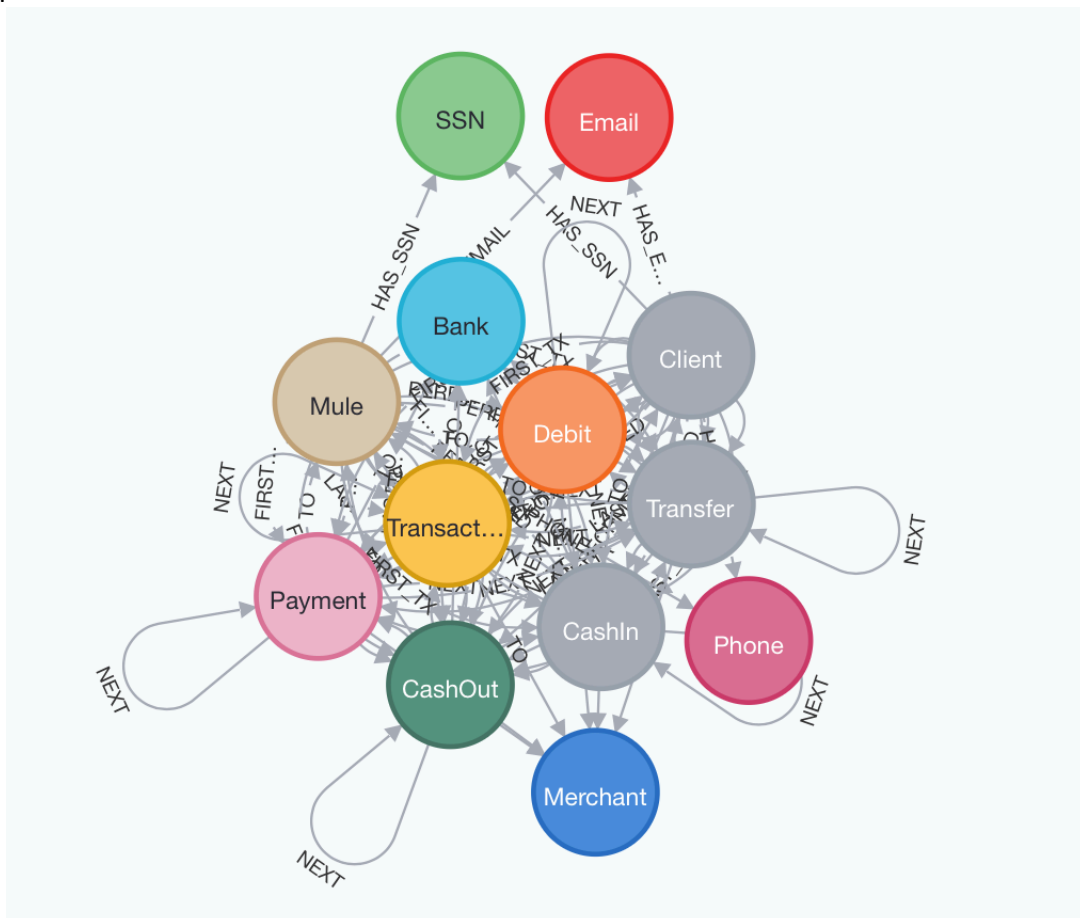


Fraud Detection

This sandbox allows you to apply the Graph Data Science (GDS) library in the financial domain using Paysim, a synthetic identity fraud detection and investigation dataset that mimics real world mobile money transfer network.

You will use it to explore first-party fraud (that occurs when an individual, or group of individuals, misrepresent their identity or give false information when applying for a product or services to receive more favorable rates or when have no intention of repayment), to identify clusters of clients sharing personally identifiable information (PII), and to compute and assign fraud score (to clients in the clusters) using a centrality algorithm.

The graph data model is shown below.



Healthcare Analytics

Health care analytics is the analysis activities that can be undertaken as a result of data collected from claims and cost data, pharmaceutical and research and development (R&D) data, clinical data (collected from electronic medical records (EHRs)), and patient behavior and sentiment data.

This sandbox uses the FDA Adverse Event Reporting System (FAERS or AERS) data. FAERS is the computerized information database designed to support the US Food and Drug Administration's (FDA) post marketing safety surveillance program for all approved drug and therapeutic biologic products.

Using this sandbox, you will perform an analysis of drug adverse events, including some statistical data analysis.

The graph data model is shown below.

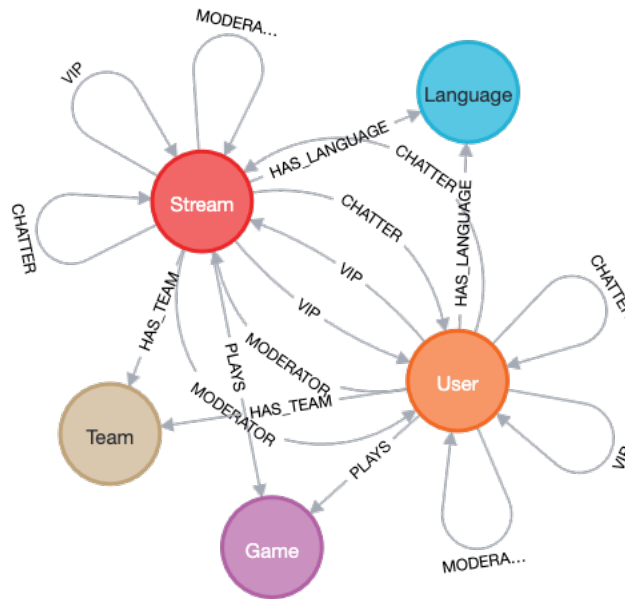


Twitch Network Analysis

Twitch (<https://www.twitch.tv/>) is an online platform that allows users to share their content via live stream. Twitch streamers broadcast their gameplay or activity by sharing their screen with fans who can hear and watch them live.

Using this sandbox, you will perform exploratory graph analysis, analyze the network of users, and find communities of streamers with shared audience.

The graph data model is shown below.



Yelp Dataset

The Yelp dataset (<https://business.yelp.com/data/resources/open-dataset>) provides real-world data related to businesses including reviews, photos, check-ins, and attributes like hours, parking availability, and ambience.

In this sandbox, you will explore graph algorithms, such as similarity algorithms as well as centrality algorithms, for example to return relevant reviews for a business based on the user's context.

The graph data model is shown below.

