Knowledge Graphs | 2022S

- Thomas Weber | 01553755
- Marvin Seidl | 11777747
- Kirill Medovshchikov | 12144024
- Giuseppe Tripodi | 12135199

The Condos Graph of Vienna

- Representation
  - Logical Knowledge about condos, map & public transport data
    - explicit relations (has/is/serves)
    - implicit relations (near - inferred based on location)
  - Clustering with traditional methods (DBSCAN, KMEANS) to create a baseline
  - Future Work:
    - MAGNN - Metapath Aggregated Graph Neural Network (for Heterogeneous Graph Embedding)
- System
  - MongoDB & Neo4j
  - Nodejs/Typescript, Python, Docker & Web
- Application
  - Filtering of condo offers
  - Visualization on map (eg. along public transport lines)
  - show similar condos based on selection
  - Future Work:
    - predict prices of condos for specific areas with predefined features

Graph Architecture

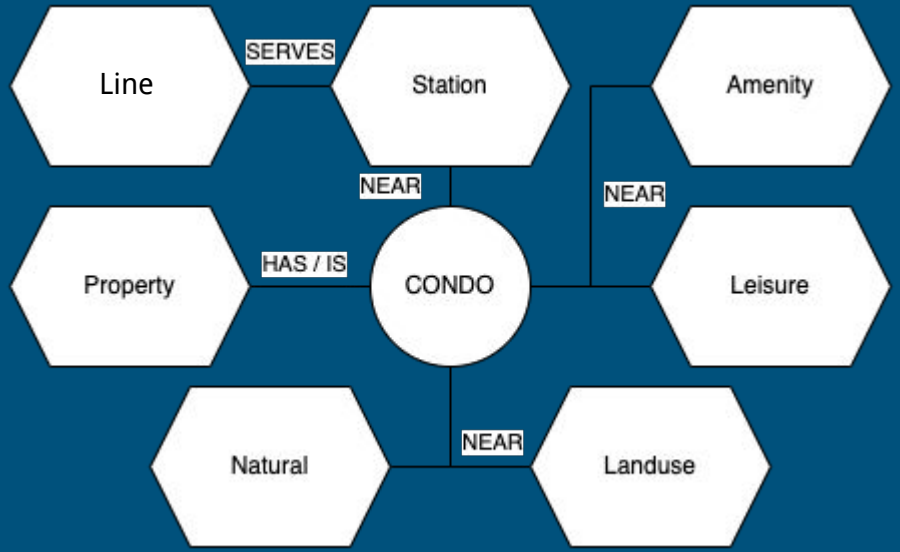
Neo4j Details

Condo:

- additional_cost
- gross_size
- living_size
- mongo_id
- price
- price_suggestions
- usable_size
- uuid

Properties:

- AVAILABLE_DATE_FREETEXT
- BUILDING_CONDITION
- BUILDING_TYPE
- COORDINATES
- ENERGY_FGEE_CLASS
- ENERGY_HWB
- ENERGY_HWB_CLASS
- FLOOR_SURFACE
- HEATING
- NO_OF_ROOMS
- OWNAGETYPE
- PROPERTY_TYPE
- UNIT_TITLE



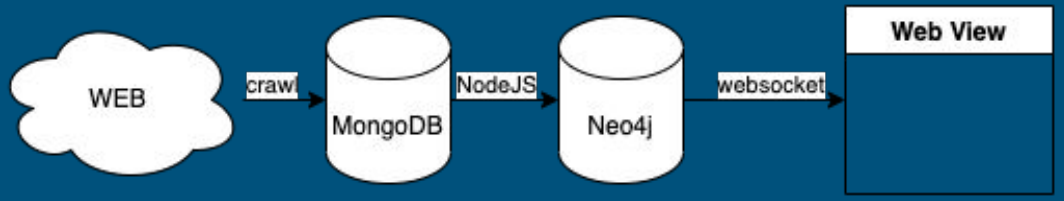
Open Street Map (OSM) Features:

- Amenities
- Leisure
- Landuse
- Natural

Stations then serve Lines

Tools & Process

1. Crawling
   a. Willhaben / OSM
   b. JSON Lines (jsonl)
2. Import data to MongoDB
3. Clustering of Condos
4. Import graph to Neo4j
   a. Preprocessing / Filtering
   b. Creating relations
   c. Typescript / NodeJS
5. Web View
   a. HTML + JS
6. docker & docker-compose



Data Sources

Open Street Map

- Download regional dump XML (840MB)
- Filter Nodes based on Tags
  - Leisure
  - Landuse (eg. recreation_ground)
  - Natural
  - Amenity
- Create json record
- Location valid geojson
  - Point → Shops, Gyms, Doctors, …
  - Polygon → Parks, Natural Areas, …
- 194k entries

Willhaben

- Webscraping
- JSON payload
- One request / ~ 1-2 seconds
- For sale not for rent
- 1200 Apartments

Public Transport Stations

- GTFS data for Vienna
- Open Data
- 4765 Stations

https://www.data.gv.at/katalog/dataset/wiener-linien-fahrplandaten-gtfs-wien

Clustering

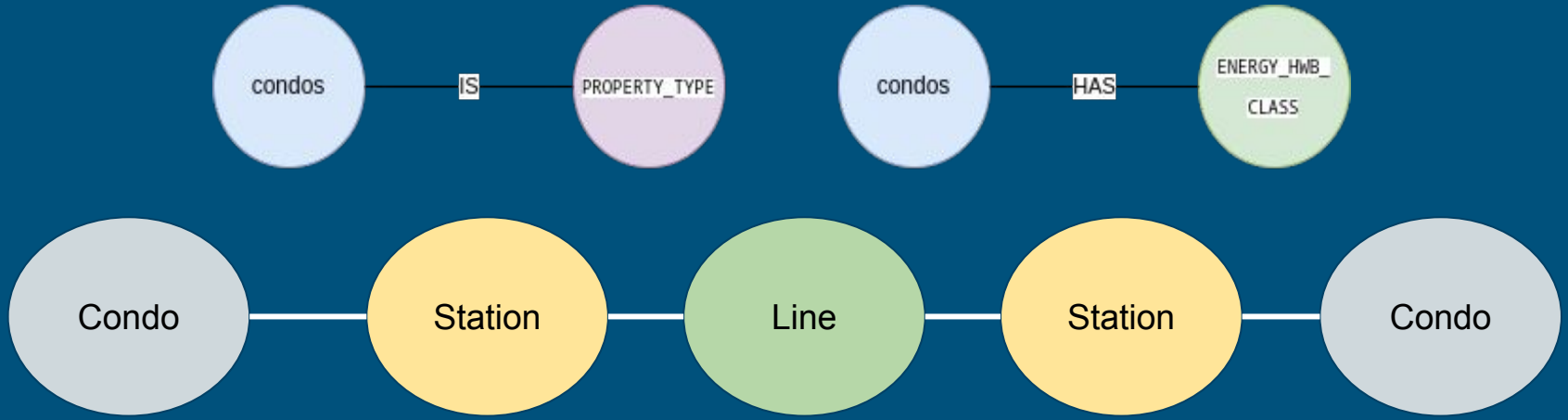
MAGNN (Metapath Aggregated Graph Neural Network for Heterogeneous Graph Embedding)

Three major components:

1. Node content transformation: transform node content features to address the heterogeneity of the graph.
2. Intra-metapath aggregation: capture the structural and semantic information of the graph from nodes and metapath context in between.
3. Inter-metapath aggregation: combine semantic information from multiple metapath. It gives different weights to different metapaths.

MAGNN (Metapath Aggregated Graph Neural Network for Heterogeneous Graph Embedding)

Metapath: path between nodes which describe composite relation

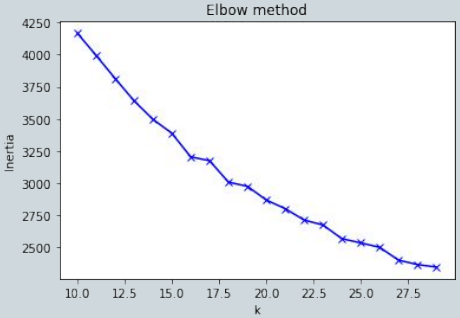


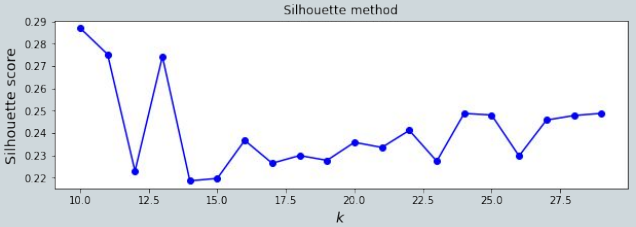
Clustering

Put together condos with similar characteristics

Steps:

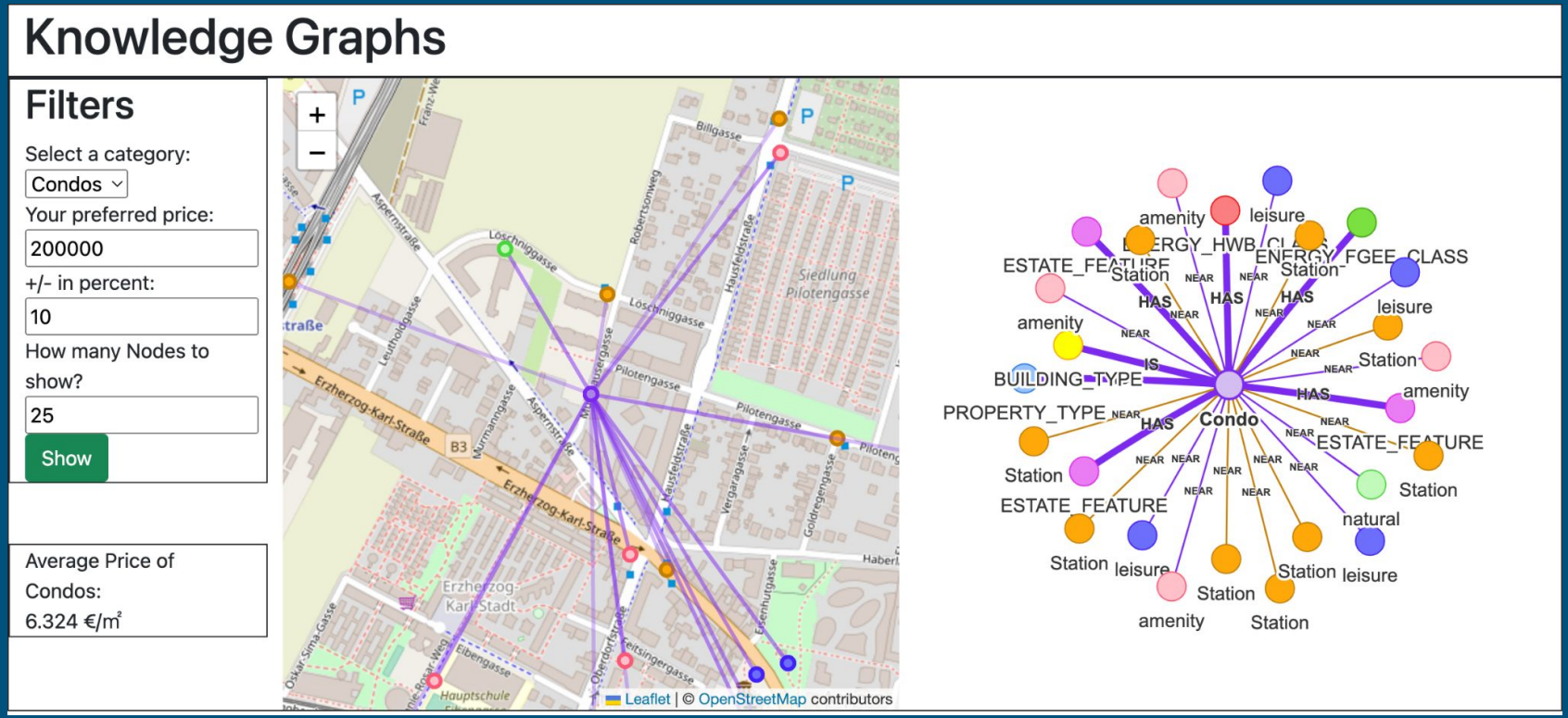
- Preprocessing (drop null value, delete useless features, delete high correlated features)
- Applied PCA to reduce the dimensionality
- Tried different basic clustering algorithms (*DBSCAN*, *KMEANS*)
- Parameter Tuning to find the best cluster's number
- Add the result on *Neo4j*



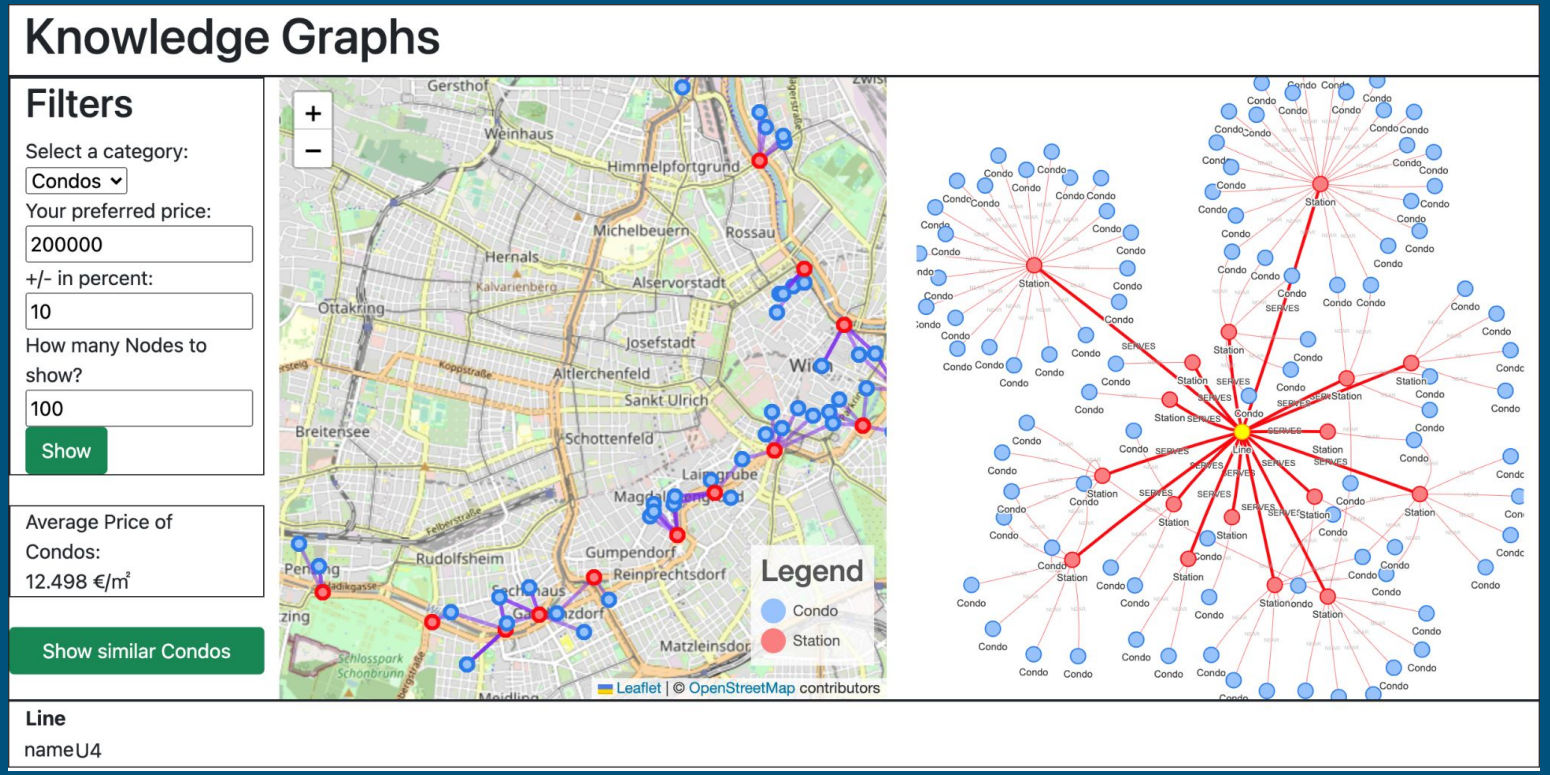


Demo

Web View


Knowledge Graphs
Filters
Select a category:
Condos
Your preferred price:
200000
+/- in percent:
10
How many Nodes to show?
25
Show
Average Price of Condos:
6.324 €/m²
Leaflet | © OpenStreetMap contributors
amenity
leisure
ESTATE_FEATURE
Station
ENERGY_FGEE_CLASS
HAS
NEAR
IS
BUILDING_TYPE
PROPERTY_TYPE
Condo
natural

Web View (Cont.)



Troubles & Issues

- Data integration takes a lot of time
  - We should have added the location data directly to neo4j
- Not enough knowledge beforehand
  - Not much knowledge about knowledge graphs
- High dimensionality (cluster could be difficult in this conditions)
  - Many different features, data is sparse, clustering becomes more difficult
- Many null values

Future Improvements

- Find more information about condos and cleanup data
- Create more sophisticated relations
  - More meaningful connections in between the nodes
- Use a more advanced clustering method (MAGNN)
  - Not well documented → Not implemented due to time constraints

Thank you for your attention!

We are happy to hear any feedback you might have!:)