Enabling Open Dataset relatedness

Oladipupo A. Sennaike

Outline

- Open Data Platforms
- Open Data Portals
- Objectives
- Dataset Relatedness
- Self Organising Maps （SOM）
- Dataset
- Model Development and Selection
- Topographic & Quantisation Errors
- Evaluation
- Dataset Recommender Service
- Related datasets for ‘Parks’ dataset
- Beyond Dataset Relatedness
- Knowledge Graphs
- Graph Schema
- Generating the Graph
- Generated Graph
- Centrality Measures
- Applying the Knowledge Graph
- Concluding Remarks

Open Data Platforms

- Provides access to available data
- Manage data catalogues
- Publish, explore, analyse, visualise and share datasets
- Over ten known open data platforms: CKAN, DKAN, Socrata, PublishMyData, Information Workbench, Enigma, Junar, OpenDataSoft, Callimachus, DataTank and Semantic Media Wiki

Open Data Portals

- Built on Open Data Platforms
- data.gov with over 195,000 datasets
- data.gov.uk with over 42,000 datasets

Objectives

- Determine the implicit semantic relatedness of datasets
- Provide recommendation features in open data platforms
- Discovery of different categories or themes that are implicit in the datasets
- Present data to the user in such a way that they have a good idea of what the portal has to offer.

Dataset Relatedness

- Relatedness defines an established or discoverable connection or association between two concepts
- Dataset relatedness is a measure of the proportion of shared concepts between two datasets in a catalog
- Explicitly methods
  - assigning Datasets with the same theme
  - tagging them with the same keywords
  - subjective, incomplete, sometimes absent
  - specifying dataset relatedness relationship manually is infeasible.

Self Organising Maps (SOM)

- An unsupervised, competitive, winner take all neural network
- Projects high dimensional data unto a low (usually two) dimensional space
- Preserves topological order
- Related data are close on the resulting map.

Dataset

- Extracted from the Dublin City Council (http://dublinked.ie/)
- 255 available datasets and associated metadata
- Features include
  - Title, Organization, Theme, Notes and Tag extracted from metadata
  - Resource Fields extracted from field names of tabular data
  - Location, Person, Organization extracted using named entity recognition (NER)

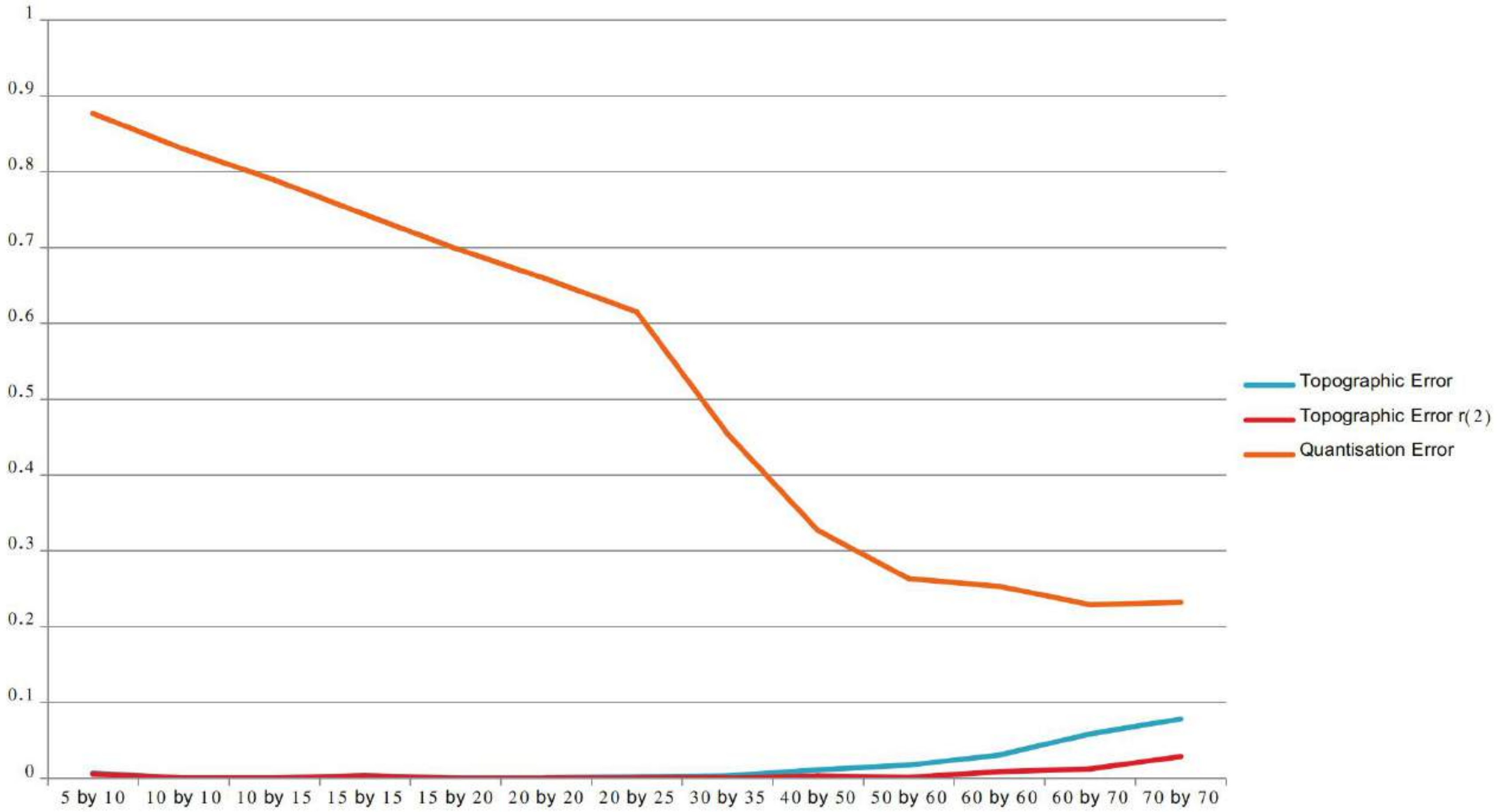
Model Development and Selection

- SOM was trained with different SOM sizes

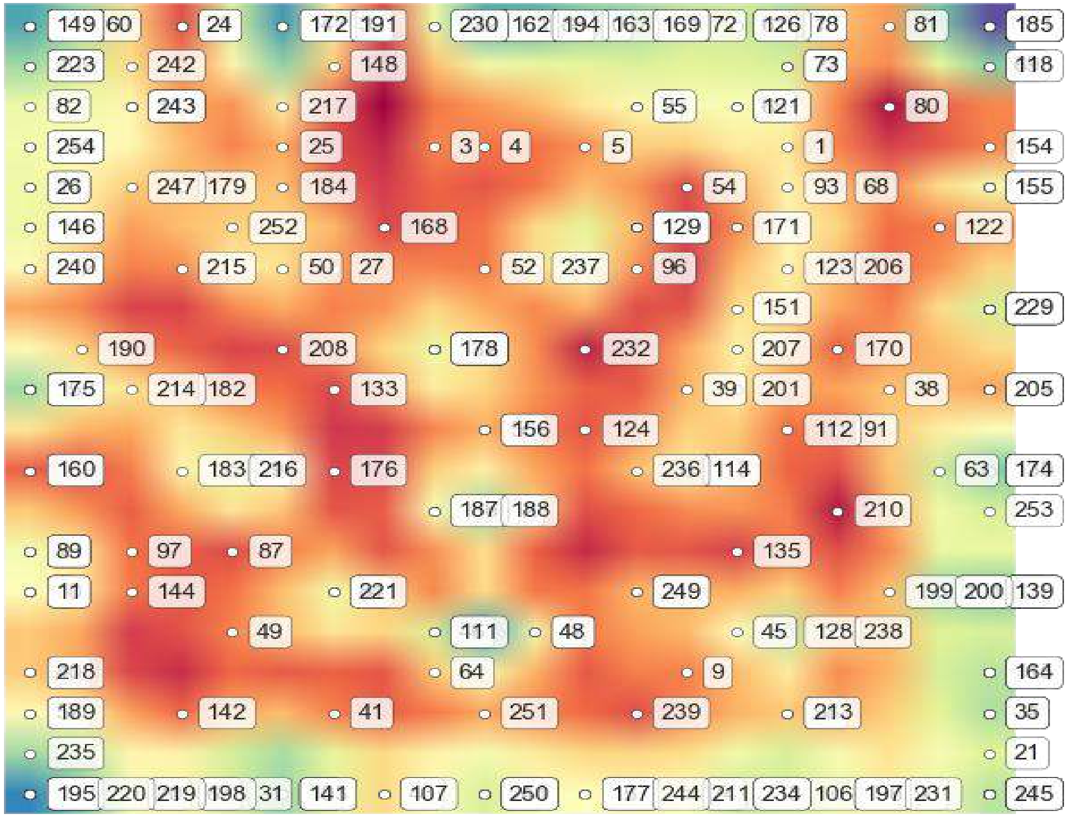
- The folowing measures were computed for each map instance
  - Topological error
  - Quantisation error

- A 20 by 20 map was selected

Topographic & Quantisation Errors



Topographic Map of Datasets



Evaluation

- Results were presented to domain experts for evaluation

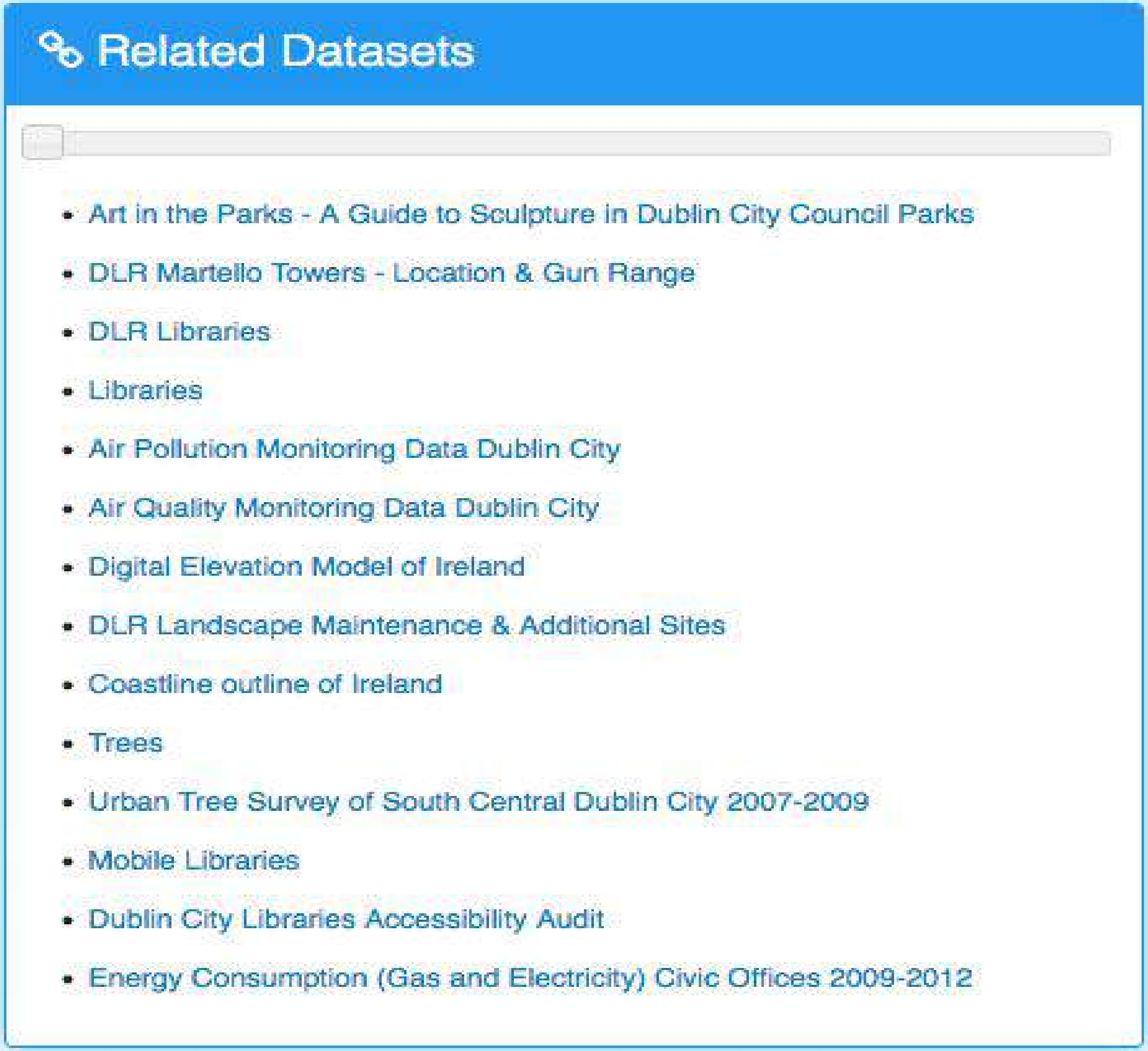
- Each node and their neighbours, usually up to a radius of 2, were examined

- The experts were able to identify the topics that relate each node and its neighbours in the datasets

Dataset Recommender Service

- Model was implemented in CKAN-based open data platform (Route-To-PA Platform)
- Results for “Parks” in Dublin City produced a list of datasets on other parks, libraries, air pollution and monitoring data, trees, landscape maintenance, energy consumption.
- Result relates recreation, sustainable environment and culture
- The model has been extended to the Dutch Language with equally good results. It has also been used as a basis for recommending datasets that can be merged.

Related datasets for ‘Parks’ dataset



Related Datasets

- Art in the Parks - A Guide to Sculpture in Dublin City Council Parks
- DLR Martello Towers - Location & Gun Range
- DLR Libraries
- Libraries
- Air Pollution Monitoring Data Dublin City
- Air Quality Monitoring Data Dublin City
- Digital Elevation Model of Ireland
- DLR Landscape Maintenance & Additional Sites
- Coastline outline of Ireland
- Trees
- Urban Tree Survey of South Central Dublin City 2007-2009
- Mobile Libraries
- Dublin City Libraries Accessibility Audit
- Energy Consumption (Gas and Electricity) Civic Offices 2009-2012

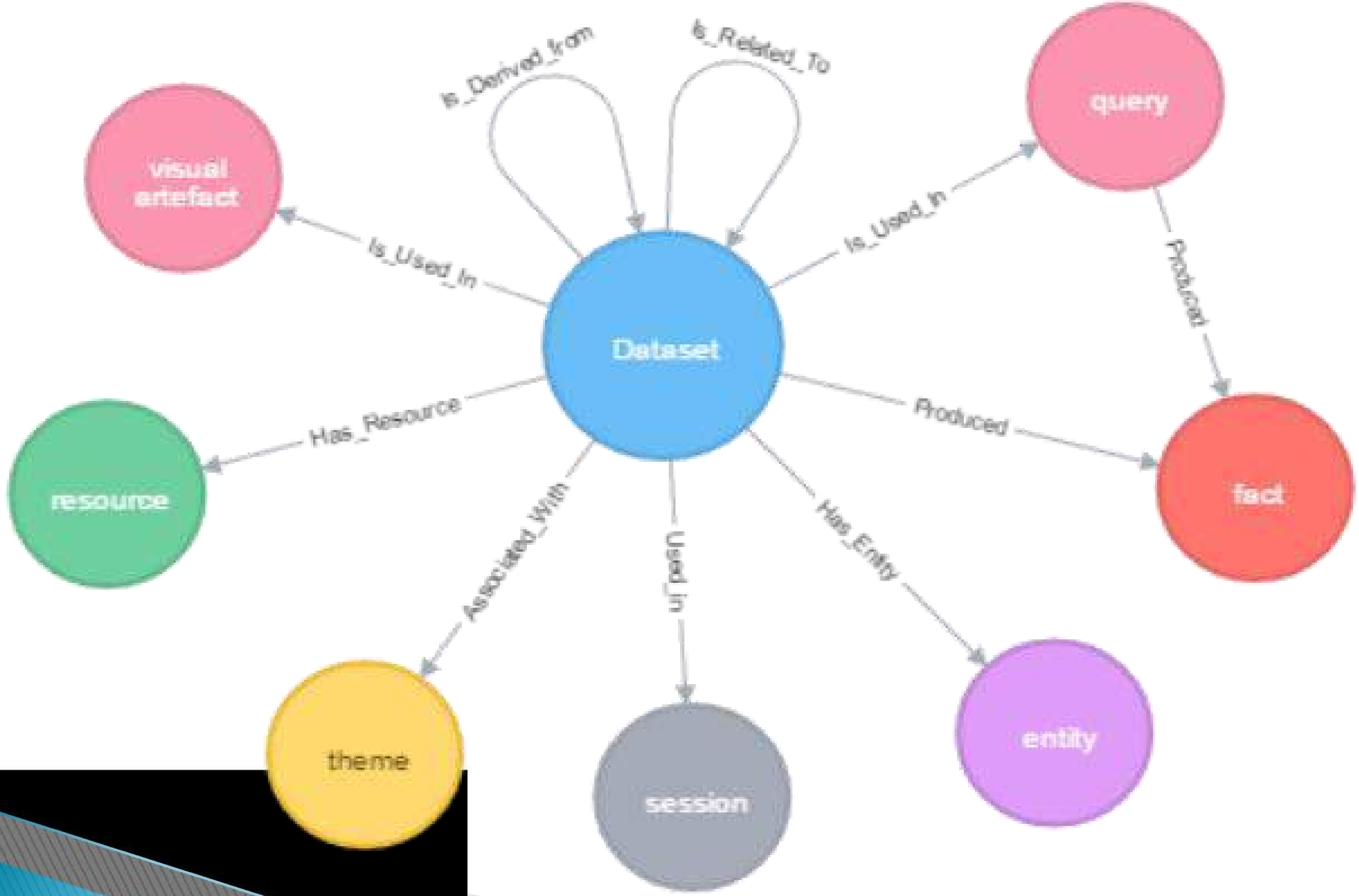
Beyond Dataset Relatedness

- Can we discover different categories or themes implicit in the datasets？
- Can we build and explore the social network of the datasets？
- Can we explore how these datasets are connected to one another？
- Can we discover centrality or isolation of datasets？

Knowledge Graphs

- Large networks of structured information about entities and their semantic relationships.
- Made up of entities as nodes and relationships between entities as edges
- based on the Resource Description Format (RDF) data model
- Querying the data in a KG is based on structured patterns, using query languages in the style of SPARQL.

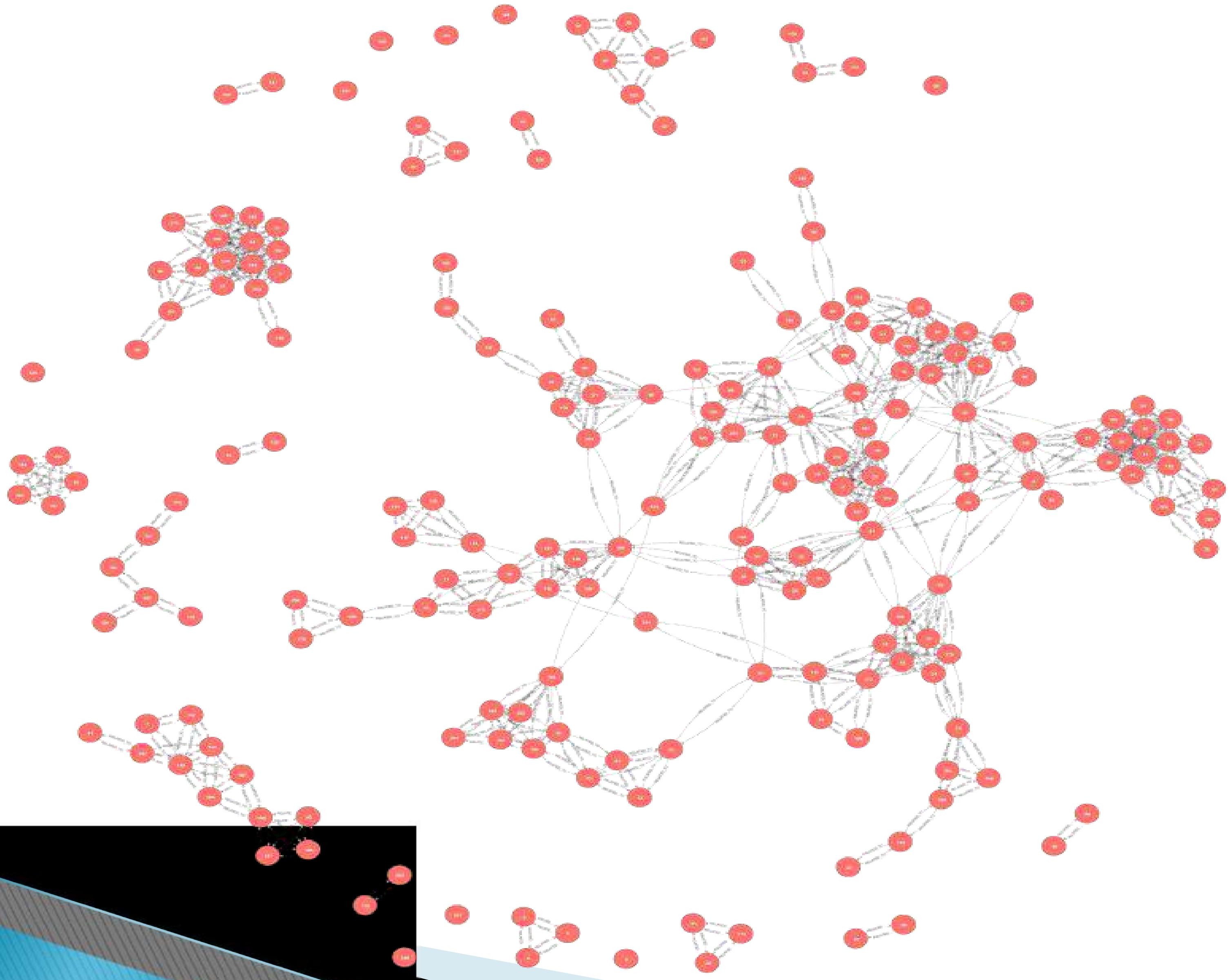
Graph Schema



Generating the Graph

- Focused only on dataset and the *is_related_to* relationship
- degree of 1 for the dataset relatedness
- 205 nodes and 956 edges
- Each node is labelled with the serial number of the dataset
- Each edge is labelled “RELATED_TO” and has the following properties: the distance between the datasets, and the common terms between the datasets

Generated Graph



Centrality Measures

- Degree Centrality
- Betweenness Centrality
- Closeness Centrality
- Clusters

Applying the Knowledge Graph

- Profiling
  - degree centrality for each cluster serves as entry point to the different clusters
  - datasets with the highest betweeness centrality are datasets that provides a bridge for two apparently different concepts
- Recommendation
  - content-based recommendation
  - collaborative recommendation (use user profiles)
  - hybrid approaches
- Integration

Concluding Remarks

- Our representation of relatedness is a simplistic view of the relationship in the dataset considering our proposed graph schema. Interestingly, this simplistic representation gives a lot of insight into the dataset, revealing very otherwise unknown and interesting properties in the dataset.