Graph analysis and similarity detection: an application on Italian medical prescriptions

Paolo Mariani[1], Ilaria Giordani[1,2], Andrea Marletta[1], Mauro Mussini[1], Mariangela Zenga[1]

[1]University of Milano-Bicocca, [2]Consorzio Milano Ricerche

28-31 October 2019, University of Salerno

Seventh International Workshop on

Social Network Analysis

1. Brief introduction to graph databases
2. Methodology: An exploratory approach on migration of the database and graph modelling
3. Application and results
4. Summary and conclusions

Graph databases

Graph databases are data management systems allowing persistent representation of entity and relationship in a graph structure, implementing the Property Graph Model efficiently down to the storage level.

A graph $G =< E, V >$ is an abstract data type showing connections (edges E) between pairs of vertices (V). Nodes identify entities and their properties, while relationships are joining attributes between tables with eventual additional characteristics.

Unlike other databases, relationships take first priority. A graph database is purpose-built to handle highly connected data, providing great performance, flexibility and frictionless development.

Queries allow to match pattern of nodes and relationships in a graph, providing transaction compliance without specifying details on how to implement operations.

An overview on Neo4j

Neo4j is an online graph database management system with Create, Read, Update and Delete (CRUD) operations working on a graph data model. The data model for a graph database is also significantly simpler and more expressive than those of relational or other NoSQL databases.

In Neo4j, everything is stored in the form of an edge, node, or attribute. Each node and edge can have any number of attributes, and both nodes and edges can be labelled. Labels can be used to narrow searches, improving speed.

Queries are written using Cypher, a declarative graph query language that allows for expressive and efficient querying and updating of the graph.

Cypher is inspired by a number of different approaches and builds on established practices for expressive querying, with SQL-inspired keywords and high-level semantics.

Goals of the study

Goals of analytics through graphs is completion of antibiotic patterns changes and patient journey, providing a different point of view on those two important aspects altogether.

This part of the research aims to focus on:

- Co-prescriptions, understanding whether specified couples of drugs are often prescribed together
- Clustering in communities, to identify similar kinds of doctors according to their prescription history

An excessive usage of antibiotics causes death of microorganisms in the human body which provide to maintaining immune cells and killing certain oral infections. Lactic ferments are often taken together with antibiotics, so that new "good" bacteria can restore the probiotic action.

If this hypothesis is correct, the dataset will show antibiotic prescriptions paired with other drugs, on the same date - or it will highlight linkings between infections and other pathologies receiving a specific prescription.

Relational database structure

The available data comprehends patients, general practitioners, and their prescriptions in the time span from 2000 to 2018, located in Campania.

Summarising the amount of records for each entity:

- $888,219$ patients
- $2,486$ doctors
- $118,716,403$ prescriptions
- $33,523$ drugs

Due to the amount and veracity of data, identifying a subset of records is useful to have detailed and targeted results, removing dispersive information and leaving a restricted pool of prescriptions, setting acceptability conditions.

Reduction of the database

Since analytics are aimed to identify antibiotic prescription patterns, similarly to past approaches, a new dataset has been extracted, imposing the following constraints:

1. AIC corresponding to an antibiotic
2. Prescription date between 2008-01-01 and 2017-12-31
3. Active general practitioners
4. Patients with usable information about sex, date of birth and location

This leads to obtaining a new model, composed by:

- $670,634$ patients
- $1,377$ doctors
- $8,386,057$ prescriptions
- $2,802$ antibiotics

Final database

To allow analytics on patient journey and co-prescriptions, it is necessary to access all the prescriptions assigned to all patients belonging in the subset. A major extraction is performed from the main table, comprehending:

- Identifier of patients who received at least one antibiotic prescription on the same date as a generic prescription
- Prescription date between 2008-01-01 and 2017-12-31

The resulting structure is an unweighted directed graph, with a data composition of:

- $670,634$ patients
- $1,377$ doctors
- $8,328,272$ prescriptions of antibiotics
- $2,465$ antibiotics
- $7,587,009$ other prescriptions
- $21,248$ other medicines

Migration of the database and graph modelling (1)

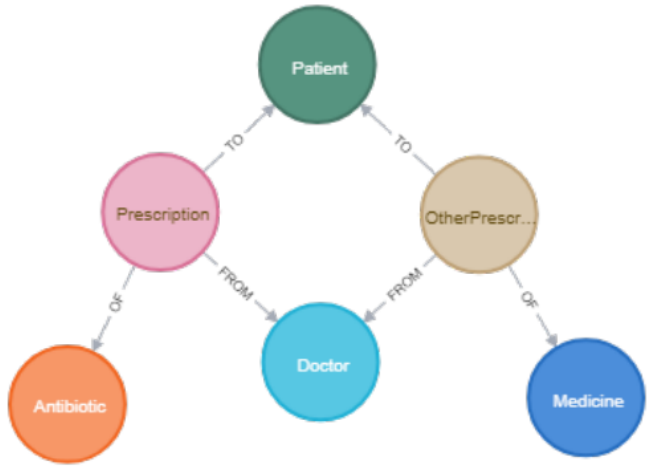
The database has to be structured following the SQL to Cypher practices and guidelines, assigning nodes and relationships in an appropriate way considering the existing dataset and the related goals. The entity-relationship model translates with the following nodes and attributes:

Nodes	**Attribute1**	**Attribute2**	**Attribute3**	**Attribute4**
Patient	ID	birthdate	sex	
Doctor	ID			
Antibiotic	AIC code	ATC code		
Medicine	AIC code	ATC code		
Prescription	patient	doctor	date	drug
OtherPrescription	patient	doctor	date	drug

Migration of the database and graph modelling (2)

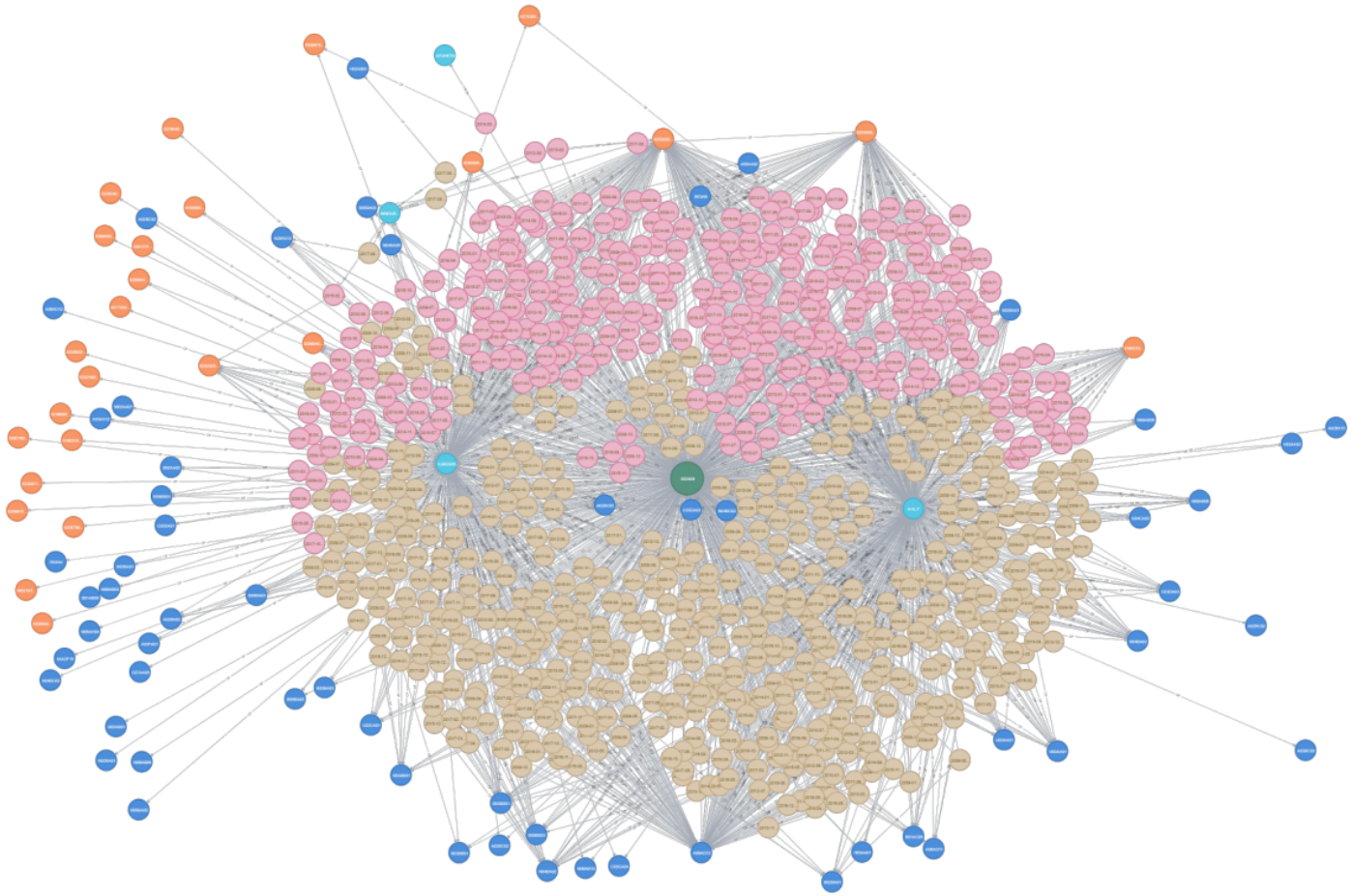
All nodes are imported, and main indexes are created for optimisation of queries speed. Relationships are then created according to IDs and AIC codes:

- Prescription - TO → Patient
- Prescription - FROM → Doctor
- Prescription - OF → Antibiotic
- OtherPrescription - TO → Patient
- OtherPrescription - FROM → Doctor
- OtherPrescription - OF → Medicine



A patient-centred graph

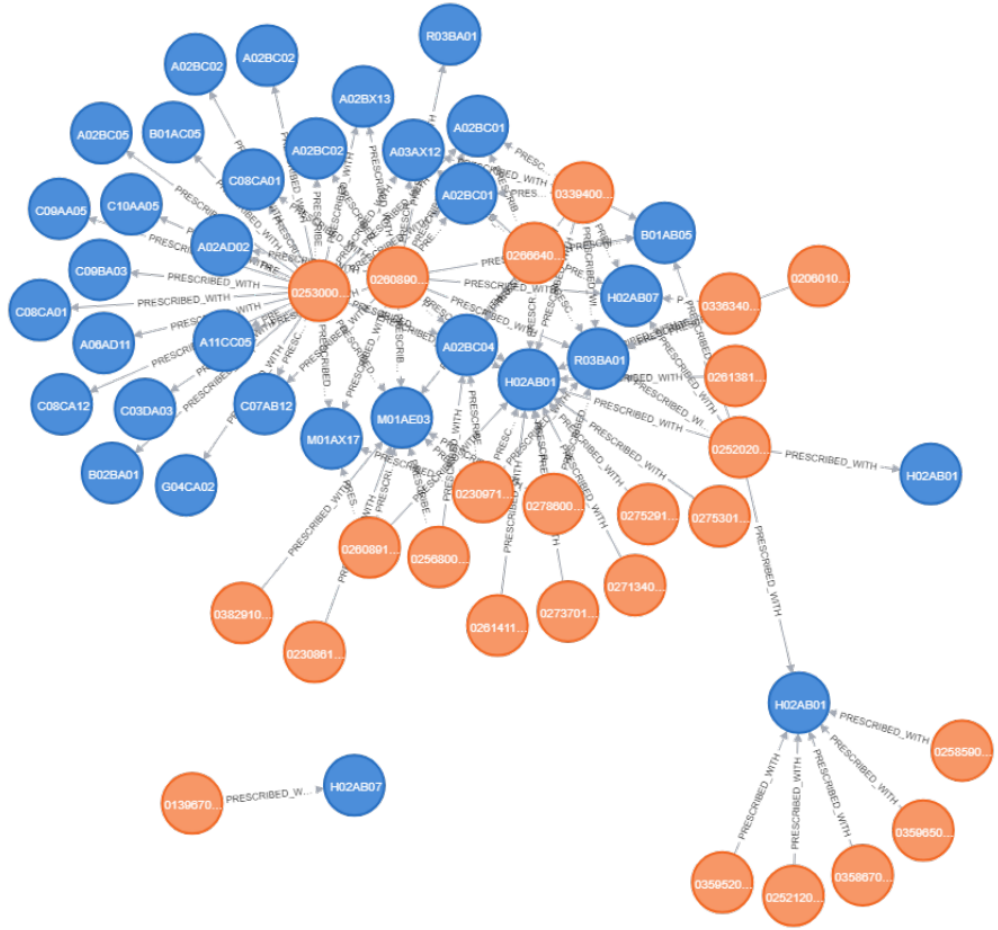
An example of graph subset can be obtained extracting one of the individuals with the most antibiotic prescriptions and his associated drugs and doctors. The prescriptions history in 10 years is displayed below.



Projecting a co-prescription graph

Since the restriction on generic prescriptions involves having the same date and patient of another antibiotic prescription, couples are analysed adding a relationship between Antibiotic (orange) and Medicines (blue).

The first 100 most popular ones are used to couple nodes, with the amount as property of the relationship PRESCRIBED-WITH.



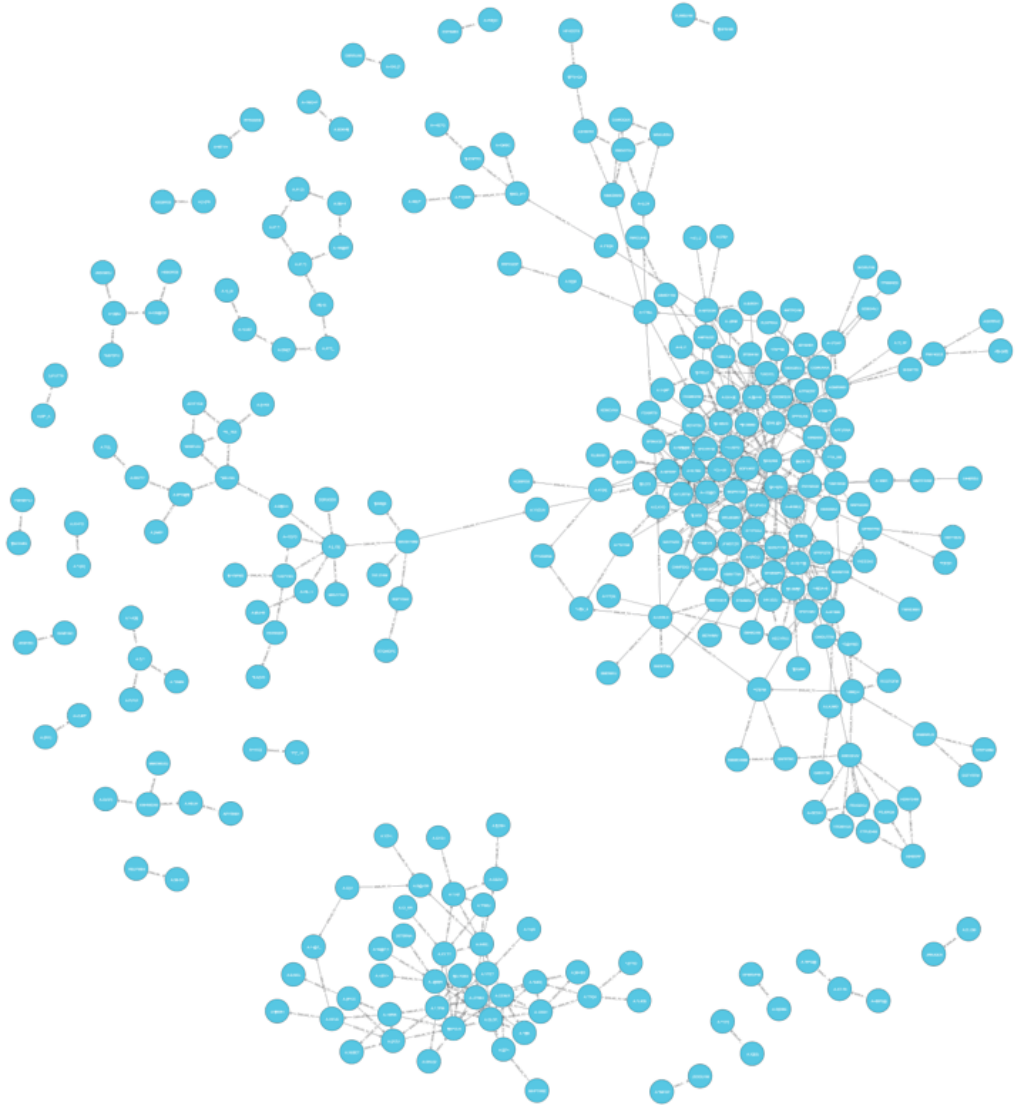
The 5 most popular co-prescriptions are:

1. Augmentin - Oki; 2. Rocefin - Bentelan; 3. Augmentin - Bentelan;
4. Normix - Cardioaspirin; 5. Augmentin - Aulin.

Similarity detection among doctors

Selecting all data from the year 2017 of patients having arbitrarily at least 10 antibiotic prescriptions (775 doctors), Jaccard similarity is computed among doctors.

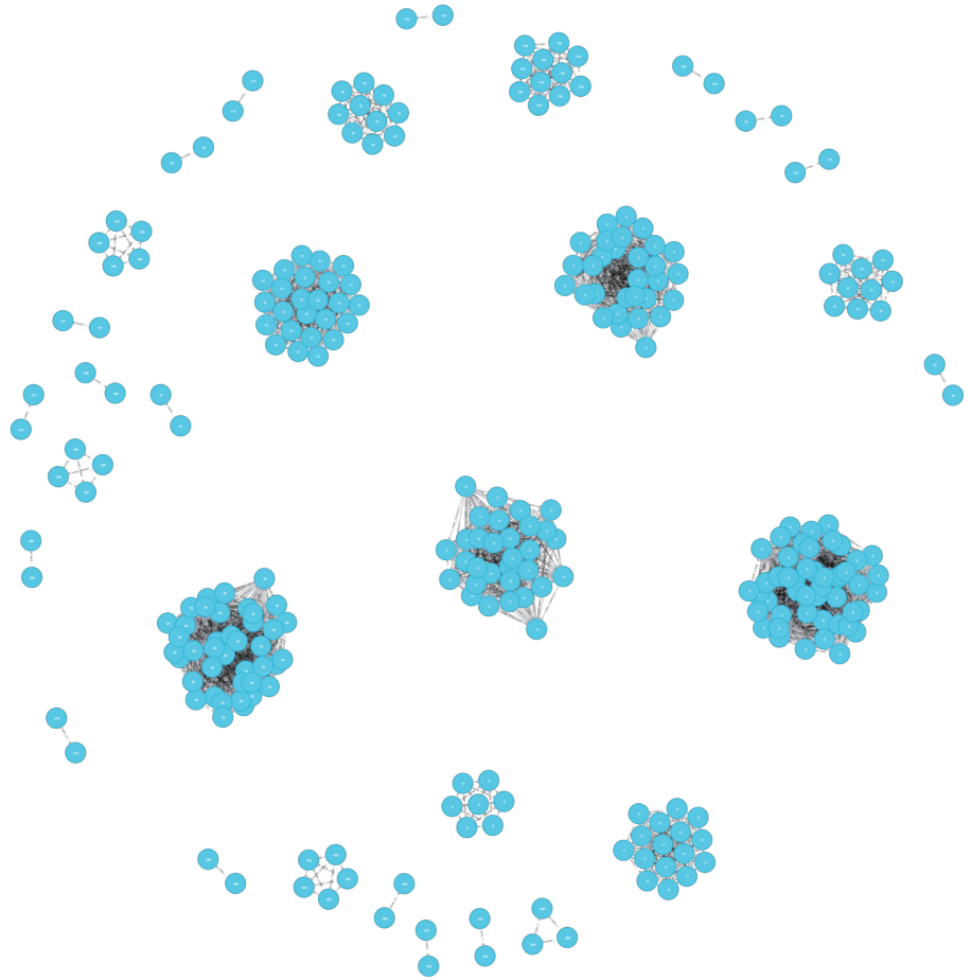
The considered parameter to determine similarity is antibiotic prescribing, therefore linked nodes have the same habits.



Community detection among doctors

Communities are identified between doctors according to their similarity, based on antibiotic prescriptions and resulting in 526 communities, 495 of which are composed by a singlet.

The presence of clusters implies that groups of doctors have the same prescriptive habits, while others do not fall in any specific category.



The 5 doctors groups are:

1. Large Normix prescriptions; 2. Rare prescribers; 3. Large all antibiotics;
4. Medium all antibiotics; 5. Large Augmentin prescribers.

Conclusions

- Graph databases offer a completely different perspective compared to relational ones, allowing to understand behaviour of nodes and linkage between them
- Despite not having a relationship between each pair of node, the whole graph can be efficiently crossed through paths, displaying different views focused on nodes linking them to the whole structure
- Building targeted datasets and extracting features is immediate, adding relationships and filtering according to properties
- Three applications on graph algorithms showed an example of patient-centred graph, a major prevalence of a smaller set of products mainly prescribed for common diseases in co-prescription graph and similarity and community detection among generic practitioners