

# Graph Algorithms for Healthcare Analytics

## Advanced Databases

Ilaria Battiston

Academic year 2018-2019

Graph analysis

## 1 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  $G = \langle E, V \rangle$  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 ACID transaction compliance without specifying details on how to implement operations. Graph-crossing and related algorithms are highly efficient.

### 1.1 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.<sup>?</sup>

## 2 Prescription coupling

An excessive usage of antibiotics causes death of microorganisms in the human body which provide to maintaining immune cells and killing certain oral infections.<sup>?</sup>

To equilibrate the intestinal flora, 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 potential linkings between infections and other pathologies receiving a specific prescription.

## 3 Goals

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;

- Centrality measures of nodes, to highlight particularly important entities in the graph.

## 4 Practical approach

### 4.1 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.

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.

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:

1. Identifier of patients who received at least one other antibiotic prescription;
2. Prescription date between 2008-01-01 and 2017-12-31.

This reduces the number of other prescriptions, adding drugs not belonging to the antibiotic class. Duplicates, mistakes and empty fields are removed.

Information loss is displayed in the figure below:

Values of prescriptions and medicines are aggregated in the plot, including both antibiotics and other drugs. The amount of removed prescriptions is most likely caused by the strict constraint of couples having the same date.

The final detailed composition of data is:

- 670 634 patients;
- 1 377 doctors;
- 8 328 272 prescriptions of antibiotics;
- 2 465 antibiotics;
- 7 587 009 other prescriptions;

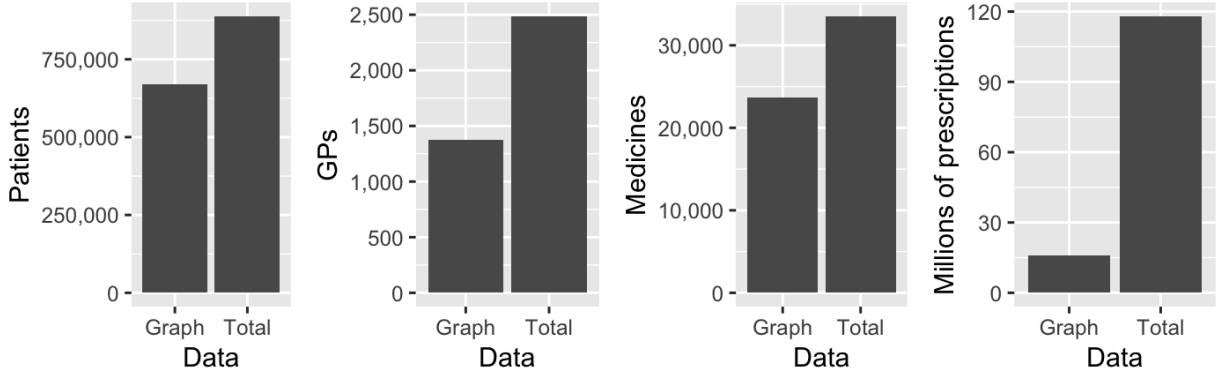


Figure 1: Information loss barplot, Campania Millennium database

- 21 248 other medicines.

The resulting structure is an unweighted directed graph.

## 4.2 Migration of the database and graph modelling

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.

After having a final version of the data to import, the entity-relationship model translates with the following nodes and attributes:

- Patient;
  - ID, birthdate, sex;
- Doctor;
  - ID;
- Antibiotic;
  - AIC code, ATC code;
- Medicine (anything not Antibiotic);
  - AIC code, ATC code;
- Prescription;
  - patient, doctor, date, drug;
- OtherPrescription (not Antibiotic Prescription);
  - patient, doctor, date, drug.

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.

## 5 Visualisation and analytics

### 5.1 Sample graph

A sample graph is obtained using the `apoc` functions.

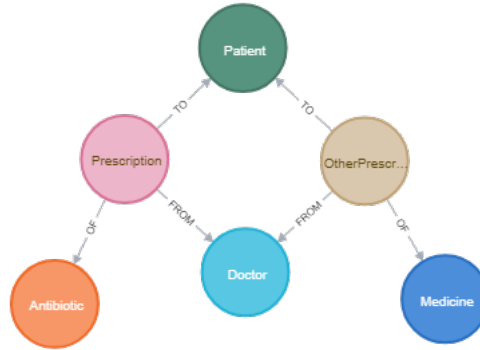


Figure 2: Sample graph, Campania Millennium database

### 5.2 Examples

An example of graph subset can be obtained extracting one of the individuals with the most antibiotic prescriptions (the 10th in descending order), a male patient born in 1943, and his associated drugs and doctors.


/images/patient.png

Figure 3: View focussed on patient, Campania Millennium database

The graph displays:

1. 1 patient, the green node in the centre;
2. 360 antibiotic prescriptions, the pink nodes;
3. 524 other prescriptions, the brown nodes;
4. 4 doctors, the light blue nodes;
5. 25 antibiotics, the orange nodes;
6. 59 other medicines, the blue nodes.

From this first patient-centred visualisation of the graph is possible to identify different behaviours of general practitioners: each one of them is linked to specific antibiotics and medicines.

To have a view focussed on prescriptions, the same procedure is applied extracting the 500th antibiotic in descending order according to number of prescriptions, corresponding to Locabiotol spray bottle 15 ml (50 mg / 5 ml).


/images/antibiotic.png

Figure 4: View focussed on antibiotic, Campania Millennium database

The previous graph displays:

1. 1 antibiotic, the orange node in the centre;
2. 896 antibiotic prescriptions, the pink nodes;
3. 107 doctors, the light blue nodes;

4. 817 patients, the green nodes.

The same antibiotic is prescribed by several doctors, although only 7% of total. There are nodes single-handedly taking a relevant slice of prescriptions (left center and right center), yet most of them is not a habitual prescriber.

Those two examples allow to have a general idea of how nodes interact with each other, grouping in clusters.

### 5.3 Graph statistics

A first set of global and local statistics are used to get the first insight on the graph and its components.

Number of nodes	16 611 005
Number of relationships	47 745 841
Average prescriptions per doctor	11 557,93
Standard deviation	15 457,6
Maximum per doctor	96 841
Minimum per doctor	1
Average prescriptions per patient	23,73
Standard deviation	40,46
Maximum per patient	1 567
Minimum per patient	1

### 5.4 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 and Medicines in the main graph.

After counting the number of repetitions for each couple Antibiotic-Medicine, the first 100 most popular ones are used to couple nodes, with the amount as property of the relationship PRESCRIBED\_WITH.

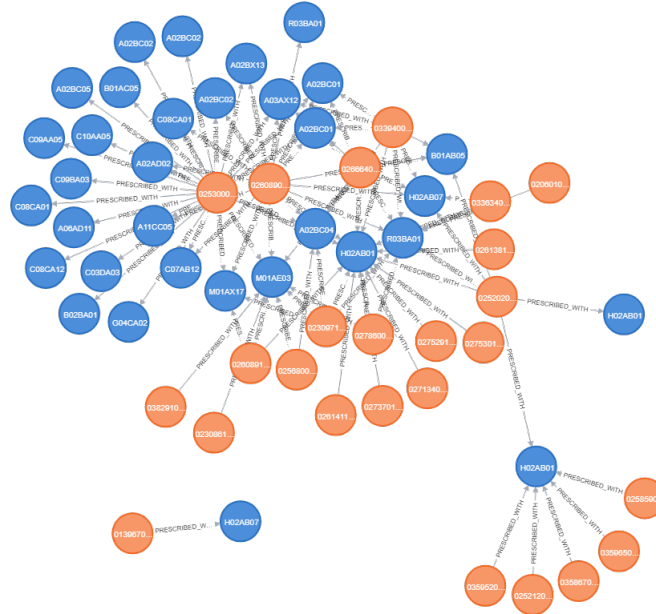


Figure 5: Co-prescriptions graph, Campania Millennium database

Visualising the newly created relationships, two connected components are highlighted (orange antibiotics, blue other medicines).

The component with only one couple corresponds to Plaquenil - Deltacortene, prescribed together approximately 5 000 times. Plaquenil can be used as antibiotic (antimalarial), but it is mostly given to treat arthritis, while Deltacortene is a corticosteroid for rheumatisms.

The range of values varies from 4 145 to 38 153 in the timespan of 10 years. The 5 most popular co-prescriptions are:

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

Bentelan is a corticosteroid which cannot be used without an antibiotic in presence of systemic (concerning the whole organism) infections,<sup>7</sup> since it is an immunosuppressive drug, and this would explain the frequent co-prescriptions.

All the antibiotics are among the most prescribed ones, which justifies their presence in the co-prescriptions as well.

## 5.5 Projecting prescriptive habits

Having a detailed view of doctors' most common prescriptions gives another insight on how antibiotics are related.

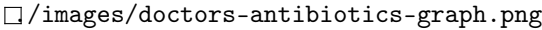


Figure 6: Most prescribed antibiotics graph, Campania Millennium database

The three most prescribed antibiotics for each doctor are taken, along with their count, and a new relationship `OFTEN_PRESCRIBED` between Doctor and Antibiotic is created. Only amounts of prescriptions greater or equal to 50 are taken into account, for consistency of results, resulting in 2 353 links.

The obtained graph displays:

1. 809 doctors, the light blue nodes;
2. 95 antibiotics, the orange nodes.

There are a few antibiotics in the center, linked to a large amount of doctors, and external antibiotic nodes having rare prevalence.

Comparing the quantity of antibiotics and doctors and their relationships, it can be seen that most doctors prescribe a very restricted set of antibiotics.

To have a better understanding of the popularity of specific drugs, an auxiliary relationship `PRESCRIBED_BY_SAME_DOCTOR` is created between Antibiotics.

Antibiotics are interlinked with new relationships:

95 nodes are connected by 294 relationships.

The previous graph confirms the discrepancy between popularity of antibiotics. Nodes on the top left have a higher number of connections, while ones on the bottom right tend to have fewer prescriptions.

One hypothesis consists in having a subset of common antibiotics shared between most general practitioners, and another one of individual preferences.

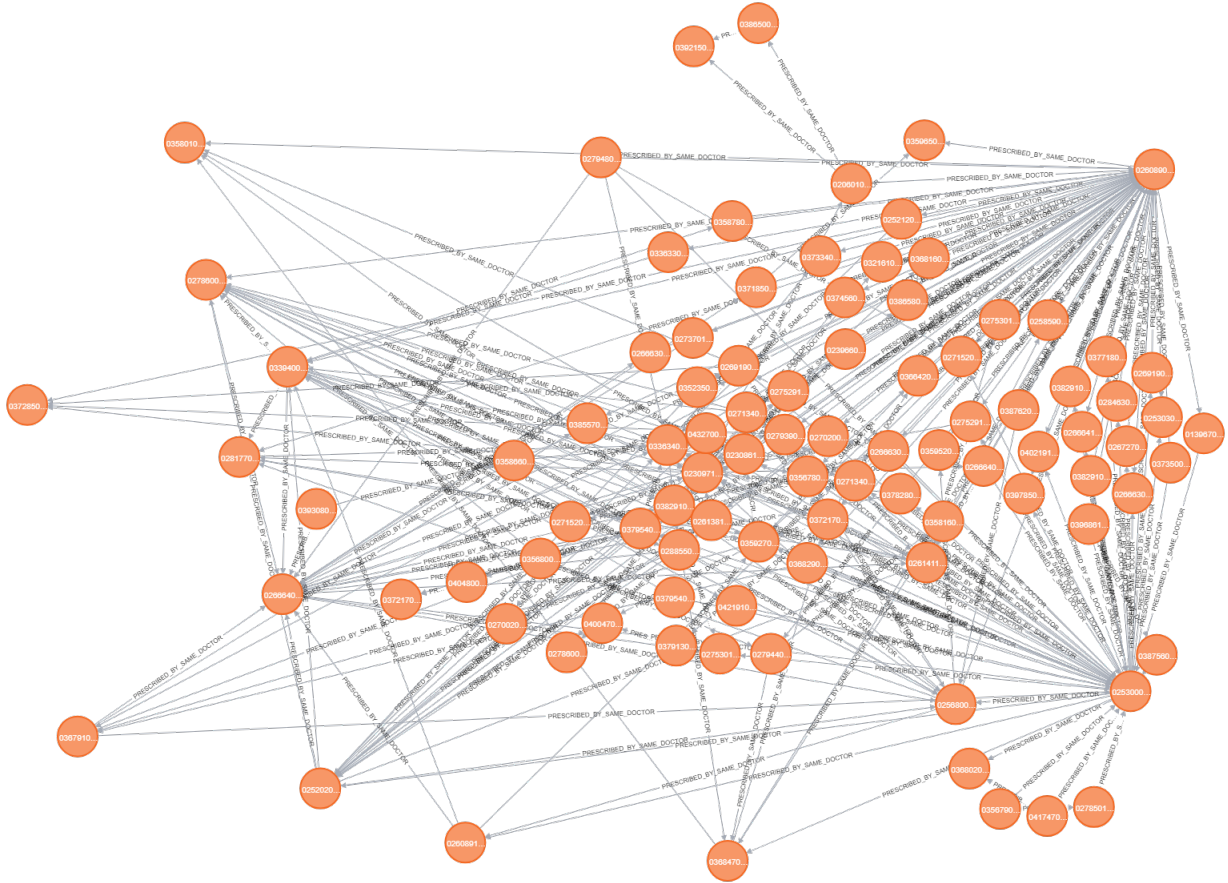


Figure 7: Antibiotics prescribed by same doctors, Campania Millennium database

## 6 Graph algorithms

Graph algorithms are the powerhouse behind analytics for connected systems. These algorithms use the connections between data to evaluate and infer the organization and dynamics of real-world systems.<sup>?</sup>

### 6.1 Centrality algorithms

#### 6.1.1 Betweenness

The Betweenness Centrality algorithm calculates the shortest (weighted) path between every pair of nodes in a connected graph, using the breadth-first search algorithm.

Each node receives a score, based on the number of these shortest paths that pass through the node. Nodes that most frequently lie on these shortest paths will have a higher betweenness centrality score.

Betweenness among the top 5 antibiotics:

Antibiotic	Betweenness
Augmentin (tablets)	2 562
Normix	1 852, 74
Velamox	683, 78
Ciproxin	562, 71
Augmentin (bottles)	489,93

As expected, the most prescribed antibiotics have the highest betweenness scores: since they often get prescribed, the number of relationships involving the nodes is high, increasing the probability of a shortest path crossing them.



### 6.1.2 Degree

Degree Centrality is the simplest of all the centrality algorithms. It measures the number of incoming and outgoing relationships from a node, analysing its influence.

Degree among the top 5 antibiotics, according to co-prescriptions:

Antibiotic	Degree
Augmentin (tablets)	32
Normix	22
Ciproxin	14
Augmentin (bottles)	12
Velamox	10

Results are similar to Betweenness Centrality, as expected: most prescribed products have the largest number of paths.

Another approach of Degree Centrality analysis involves general practitioners, to understand their influence as prescribers.

Calculating degree among doctors is useful to determine whether the top antibiotics prescribers are also the top prescribers, using the top 5 doctors.

Doctor	Degree (antibiotics)	Degree (medicine)	Total degree	Patients
1	45 625	51 221	96 846	4 466
2	42 554	41 594	84 148	2 022
3	35 645	22 383	58 028	1 816
4	34 587	40 315	74 902	2 028
5	34 380	28 763	63 143	1 874

Seeing the obtained results, the overprescribing of some general practitioners is clear: most of them makes a number of antibiotic prescriptions equal to others having double the amount of patients.

Doctors prescribing too much and not strictly when needed is one of the main causes of antibiotic resistance in Italy, therefore further analysis is required to assess changes in patients numbers and break down results in shorter time spans.

## 6.2 Graph sampling

Machine learning algorithms need a heavy amount of resources and computational time, therefore running clustering on a graph with nodes in the magnitude order of millions is not the best practice to extract insightful information in limited time.

Imposing small time ranges and only relevant features (lower bound on prescriptions) improves not only machine usage, but also accuracy of data.

A new subset is extracted, selecting all data from the year 2017 of patients having at least 10 antibiotic prescriptions during that year.

The new database is composed by:

- 8 228 patients;
- 115 443 antibiotic prescriptions;
- 775 doctors;
- 946 antibiotics;
- 181 016 other prescriptions;
- 4 654 other medicines.

Having a time range of one year and only considering patients getting antibiotic prescriptions rather often does not offer additional information on antibiotic resistance, yet it allows to obtain relevant patterns of prescribing habits.

The output of graph processing will be used to apply computationally expensive graph algorithms.

### 6.3 Similarity detection

Jaccard similarity is computed among doctors, considering the Cartesian product of nodes ( $775 \times 775$ ) with a threshold of 0.4 as coefficient. The result consists in about 500 values, which show the presence of clusters. The considered parameter to determine similarity is antibiotic prescribing (relationship between doctors and antibiotics), therefore linked nodes have the same habits.

The algorithm computes for each node, the most similar other node, according to the maximum Jaccard coefficient.

Doctors are then assigned a link, connecting similar pairs. Since a single doctor can be the most similar to several ones, nodes tend to group in clusters: connected components have a high probability of having common prescription patterns.

A large cluster on the right groups most of general practitioners, confirming the hypothesis of popular antibiotics prescribed by the majority of individuals.

Going into a higher level of detail, Augmentin and Normix are the most common prescriptions, yet a single doctor often leans towards one of the two instead of giving them both to patients.

There also is a pair of doctors connecting different components of the graph, acting as a “bridge” between prescribers of different antibiotics.

### 6.4 Community detection

The concept of community refers to the structure (topological and relational aspects) of dense sub-components of a graph. Node attributes are considered as well, to obtain additional inferences on their membership. Data is linked with discrete attributes, and applications are based on graphs properties.

Identifying communities allows to understand the global functioning and relationships between individuals with the same features. Different groups interact between one another, and can be related, giving information on the global schema.

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.

Results are similar to the  $k$ -means approach, although the considered dataset is significantly smaller: each community is distinguished by relevant characteristics, taking into account additional factors such as the amount of different products.

The first 5 communities in terms of numerosity are:

Community	Doctors	Characteristic
9	48	Large Normix prescriptions
26	41	Rare prescribers
11	31	All antibiotics prescribers, large amount
$0 \cup 51$	$30 + 25$	All antibiotics prescribers, medium amount
17	16	Large Augmentin prescribers

As expected, Normix and Augmentin prescribers are different classes of doctors: since both antibiotics are wide spectrum, the preference is most likely influenced by promotional factors or personal bias.

## 7 Considerations

Graph databases offer a completely different perspective compared to relational ones, allowing to understand behaviour of nodes and linkage between them.

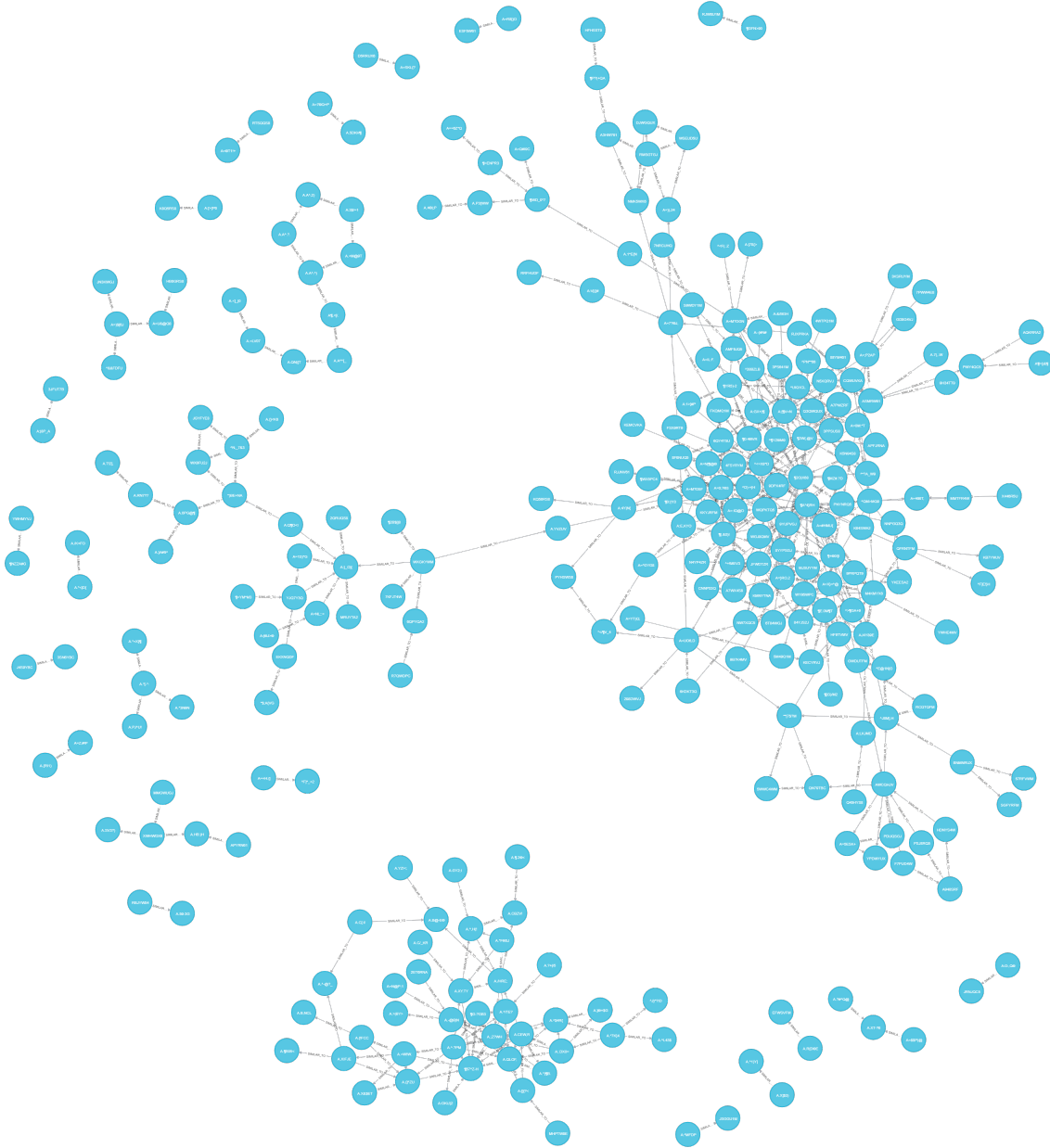


Figure 8: Similarity on doctors, Campania Millennium database

Despite not having a relationship between each pair of node, the whole graph can be efficiently crossed through paths, displaying different views focussed on nodes or their type while linking them to the whole structure.

Building targeted datasets and extracting features is immediate, adding relationships with eventual weight and filtering according to properties.

Those datasets or query results are fed to graph algorithms, which automatically stream or add resulting properties to then make further considerations without having to export data in a format readable by programming languages.

The main issue is encountered while handling large amount of data: performance with 200k nodes is optimal, yet scaling in the order of millions makes computationally expensive algorithms impossible to run in an acceptable amount of time.

Results are consistent with other machine learning approaches and exploration analytics, providing different methodologies to improve speed and comprehension.

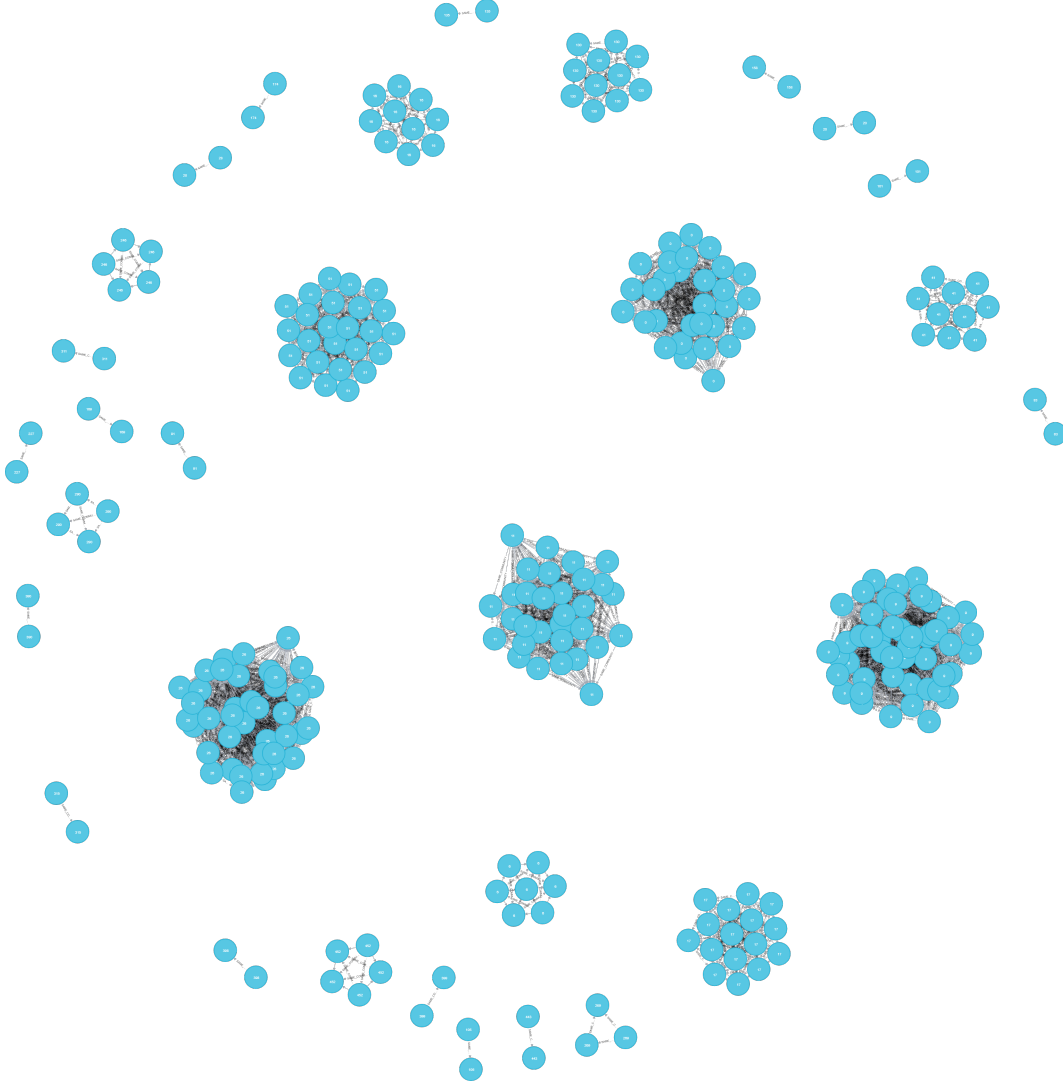


Figure 9: Communities on doctors, Campania Millennium database

Concrete outcomes after an approach with graphs consist in a further confirmation of the issue of antibiotic resistance, comparing doctors' degree to their number of patients and analyzing their prescriptive habits which, although clear, are concerning.

Co-prescriptions also highlight a major prevalence of a smaller set of products, influenced by the pressure of pharmaceutical companies and majorly prescribed for common diseases.

Future work requires a detailed analysis on variations through time, understanding the relationship between number of patients and prescriptions.