Human Disease Network

Clustering and performance evaluation

Data Analytics 2018-2019

Adorni Giorgia 806787
Mammana Lorenzo 807391

Overview

- Network description
- Centrality measures analysis
- Clustering
- Performance evaluation
- Conclusions

1.
Network description

LiveSlides web content

To view

Download the add-in.
liveslides.com/download

Start the presentation.

Ear,Nose,Throat
Psychiatric
Endocrine
Hematological
Cancer
Muscular
Neurological
Ophthamological
Nutritional
Immunological
Bone
Cardiovascular
Metabolic
Multiple
Dermatological
Renal
Skeletal
Unclassified
Gastrointestinal
Respiratory
Developmental
Connective tissue disorder

Network description

1419 network **nodes**

Bipartite graph consisting of two disjoint sets of nodes: 516 **genetic disorders** and 903 **disease genes**.

No isolated nodes detected.

3926 network **edges**

A disorder and a gene are then connected by a link if mutations in that gene are implicated in that disorder.

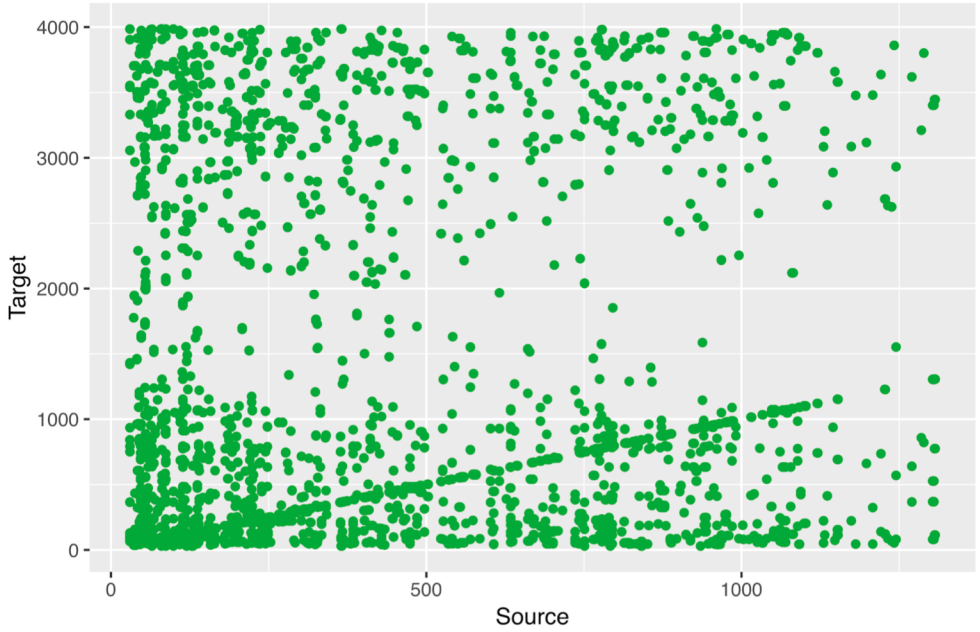
Two disorders are connected to each other if they share at least one gene in which mutations are associated with both disorders.

Network analysis

Node similarity

On the *x-axis* we have half of the nodes with respect to the *y-axis*, this is due to the fact that the gene nodes have only incoming arcs and are labeled with a value greater than 1300.

The lower part of the graph therefore represents all the links between the diseases, while the upper part are those between diseases and genes.



Network analysis

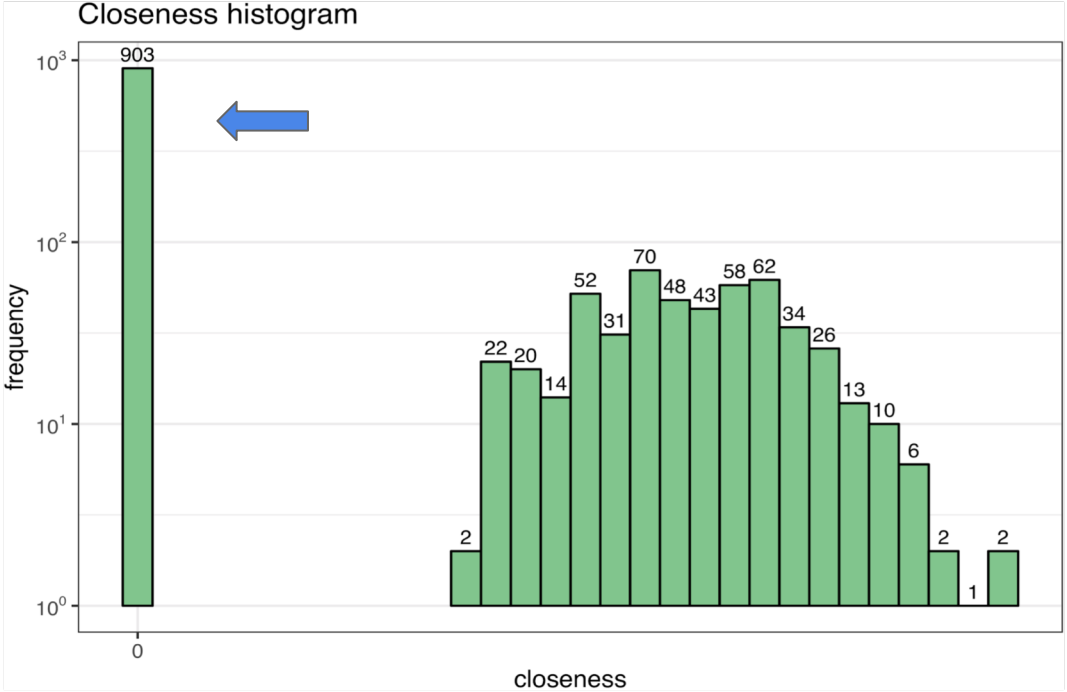
Local clustering coefficient measures the local density of each node.

Global clustering coefficient measures the clustering capability of the network.

	Value
Network Local Transitivity	0.313
Network Global Transitivity	0.251

Network analysis

Closeness



All the genes present closeness zero.

Network analysis

From HDN to analysed network

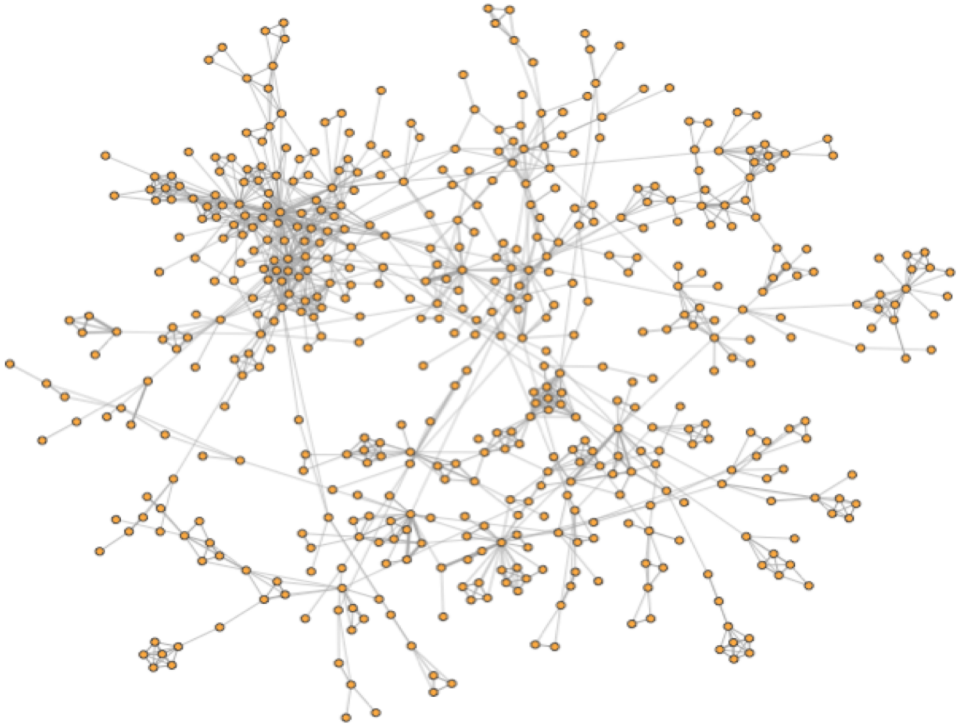
Removed genes from network and weighted edges.

Disoriented graph.

The weight of each edge is set as the number of genes that share the two diseases connected

516 network **nodes**

1188 network **edges**



Network analysis

Local clustering coefficient measures the local density of each node.

Global clustering coefficient measures the clustering capability of the network.

Weighted clustering coefficient measures the local density of each node weighted on neighbours.

	Value
Network Local Transitivity	0.313
Network Global Transitivity	0.251

	Value
Network Local Transitivity	0.636
Network Global Transitivity	0.430
Network Weighted Transitivity	0.641

3. Centrality measures analysis

Nogenes network

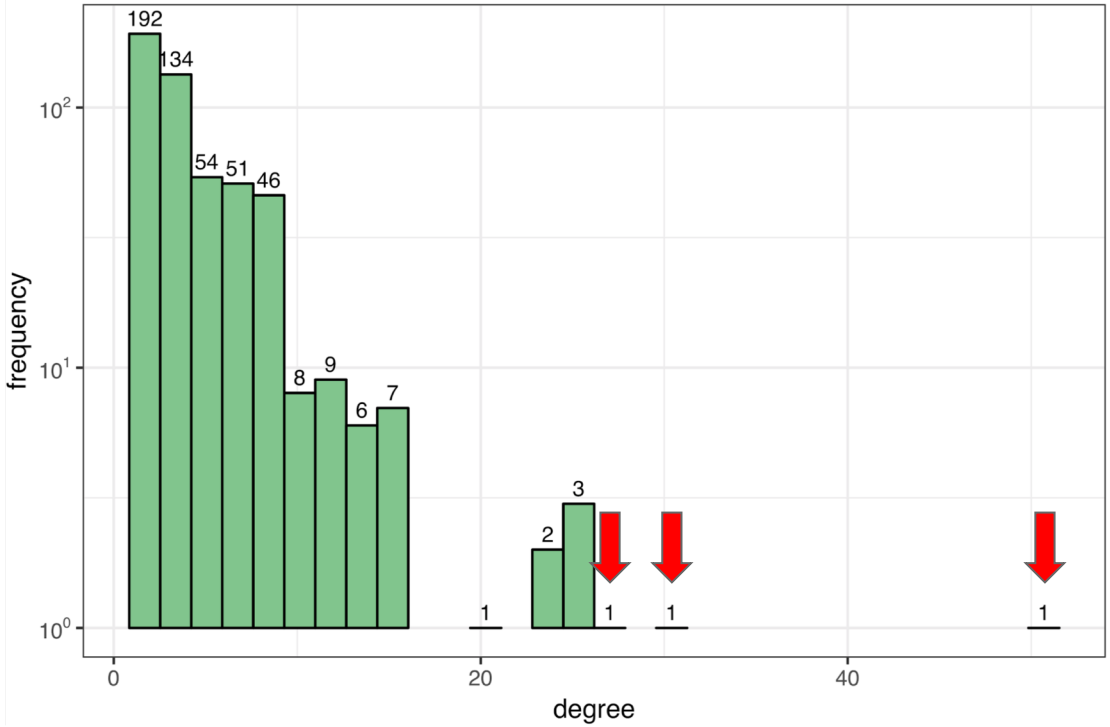
Centrality measures

Degree

$$C_D = \frac{\sum_{i=1}^{N} C_D(n^*) - C_D(i)}{(N-1)(N-2)}$$

Most disorders are linked to only a few other disorders.

Whereas a few phenotypes, such as **Colon cancer**, but also **Breast cancer**, **Gastric cancer**, **Leukemia**, and **Thyroid carcinoma**, represent *hubs*.

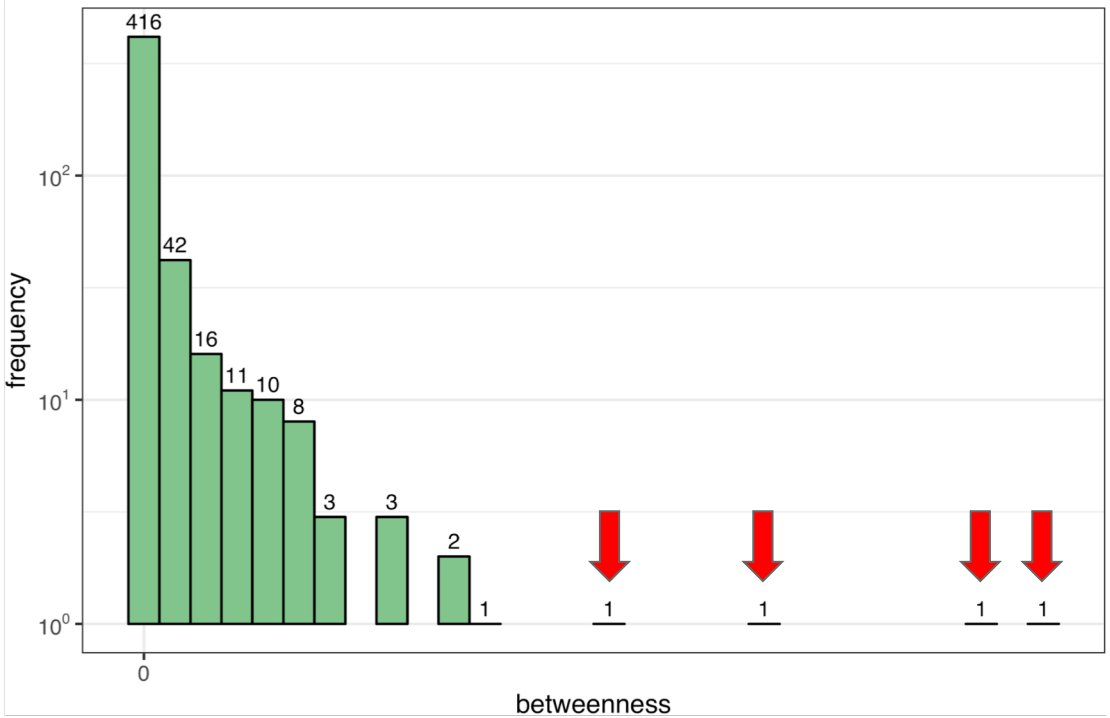


Centrality measures

Betweenness

$$C_B(i) = \sum_{j \neq k} \frac{g_{jk}(i)}{g_{jk}}$$

The nodes with the highest betweenness value, that are **Cardiomyopathy**, **Lipodystrophy**, **Diabetes mellitus**, **Glioblastoma** and **Myopathy**, represent *hubs*.

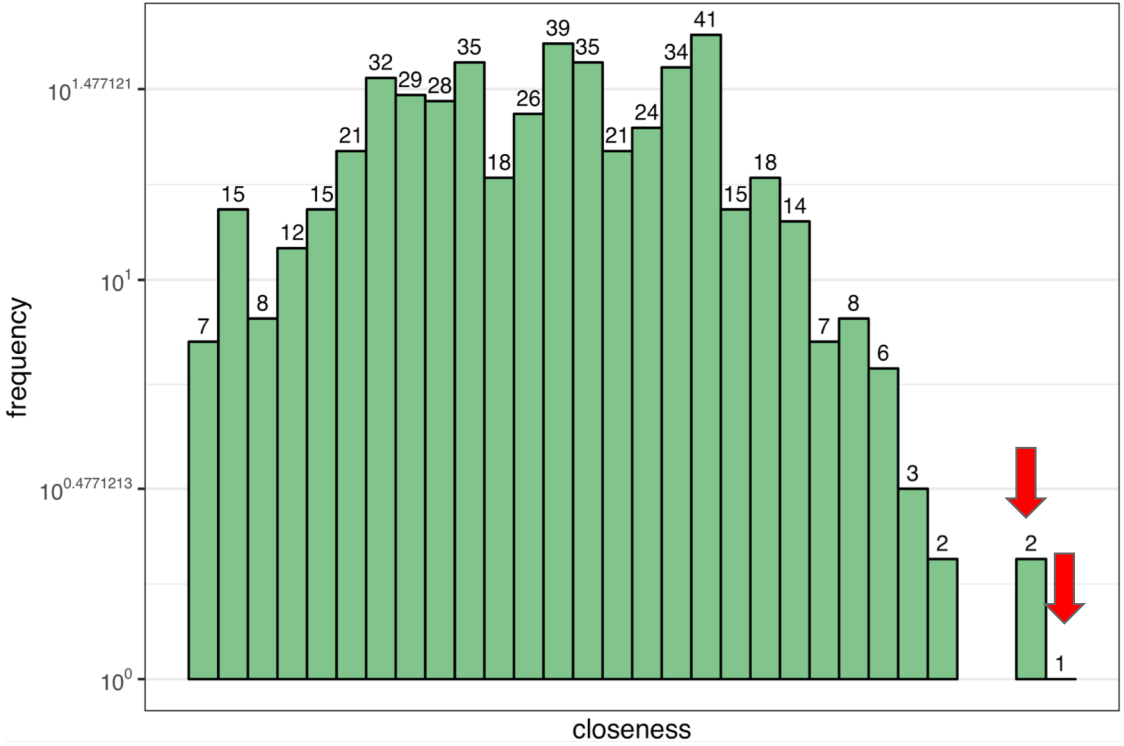


Centrality measures

Closeness

$$C_c(i) = \left[\sum_{j=1}^{n} d(i,j)\right]^{-1}$$

The nodes with the highest capacity to exchange information are **Diabetes mellitus**, **Lipodystrophy** and **Glioblastoma.**



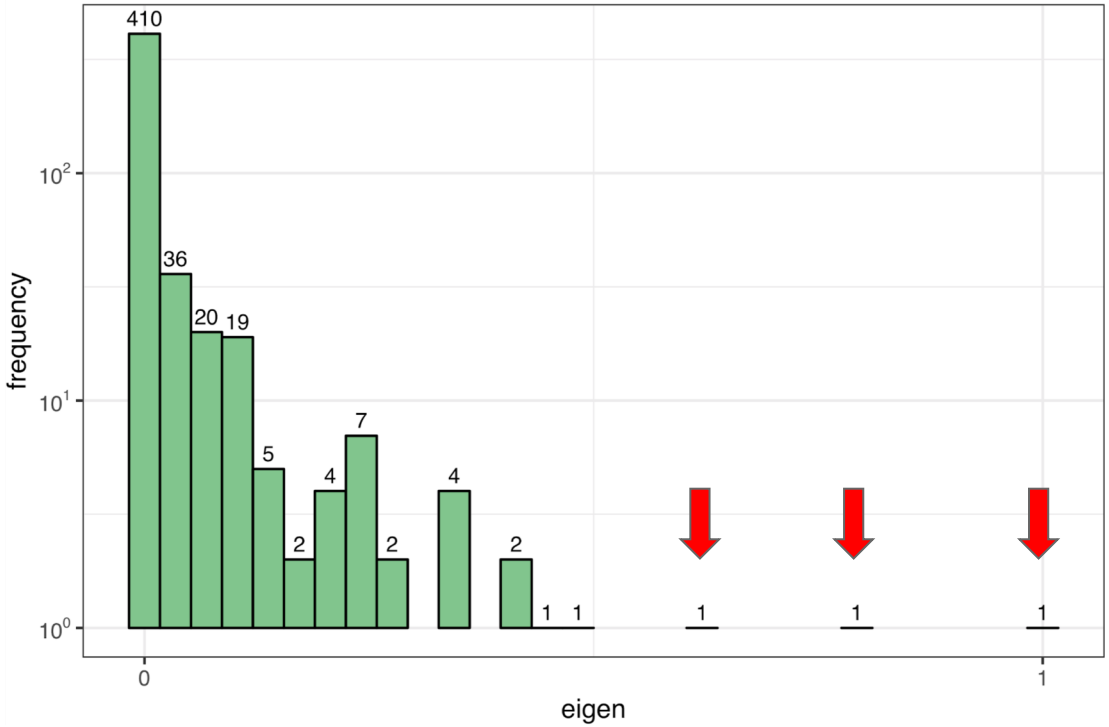
Centrality measures

Eigenvector

$$x_i = \frac{1}{\lambda} \sum_k a_{ki} x_k$$

«A node is important if it is connected to other important nodes».

Here the most important nodes are **Colon cancer**, **Breast cancer** and **Ovarian cancer.**



Centrality measures analysis

Nogenes network

Degree	Colon cancer, Breast cancer, Gastric cancer, Leukemia, Thyroid carcinoma
Betweenness	Cardiomyopathy, Lipodystrophy, Diabetes mellitus, Glioblastoma, Myopathy
Closeness	Diabetes mellitus, Lipodystrophy, Glioblastoma, Cardiomyopathy, Insulin resistance
Pagerank	Colon cancer, Deafness, Diabetes mellitus, Breast cancer, Leukemia
Eigenvector	Colon cancer, Breast cancer, Ovarian cancer, Lymphoma, Pancreatic cancer

4.
Clustering

LiveSlides web content

To view

Download the add-in.

liveslides.com/download

Start the presentation.

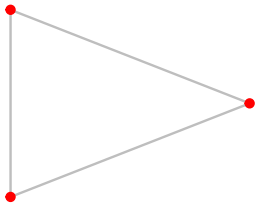
Ear,Nose,Throat | Psychiatric | Endocrine | Hematological | Cancer | Muscular | Neurological | Ophthamological | Nutritional | Immunological | Bone

Cardiovascular | Metabolic | Multiple | Dermatological | Renal | Skeletal | Unclassified | Gastrointestinal | Respiratory | Developmental | Connective tissue disorder

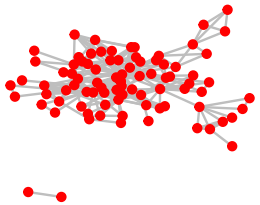
Clustering

Connected components

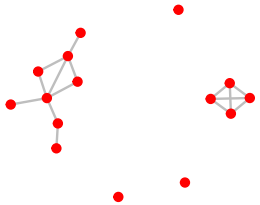
Ear,Nose,Throat



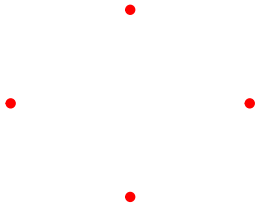
Cancer



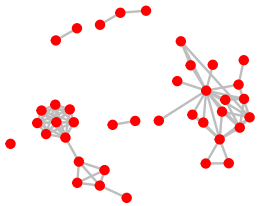
Muscular



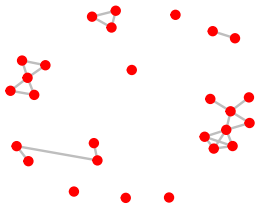
Respiratory



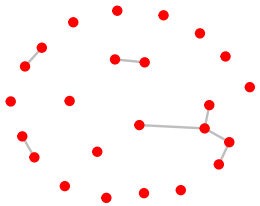
Ophthamological



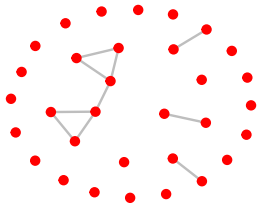
Endocrine



Immunological



Metabolic



Community detection

Clustering algorithm

The algorithms used are based on the concept of **community detection** and **graph partitioning**.

Since nodes' attributes are unavailable, it's impossible to operate with traditional clustering algorithms.

Clustering rule

Each cluster is labeled using the most common label within it.

Girvan–Newman algorithm

The **Girvan-Newman** community detection algorithm is based on the use of edge betweenness measure, in particular on the idea that arcs that connect separated modules of the network should have a high edge betweenness value.

It uses a divisive hierarchical approach aimed to maximise the so-called modularity:

$$\frac{1}{2E}\sum_{C}\sum_{i\in C, j\in C}\left(A_{ij}-\frac{d_i d_j}{2E}\right)z_{ij}$$

This algorithm distinguishes **29** different clusters within the network.

LiveSlides web content

To view

Download the add-in.
liveslides.com/download

Start the presentation.

Ear,Nose,Throat | Psychiatric | Endocrine | Hematological | Cancer | Muscular | Neurological | Ophthamological | Nutritional | Immunological | Bone

Cardiovascular | Metabolic | Multiple | Dermatological | Renal | Skeletal | Unclassified | Gastrointestinal | Respiratory | Developmental | Connective tissue disorder

Fastgreedy algorithm

This algorithm tries to find dense subgraph, or communities, in the network by maximising a **modularity score**, using a bottom-up approach, compared to the top-down one used by Girvan–Newman.

Initially each node represents a community and, iteratively, each community is joined in such a way that the union is locally optimal.

The algorithm stops when it’s no longer possible to increase modularity.

LiveSlides web content

To view

Download the add-in.
liveslides.com/download

Start the presentation.

Ear,Nose,Throat | Psychiatric | Endocrine | Hematological | Cancer | Muscular | Neurological | Ophthamological | Nutritional | Immunological | Bone

Cardiovascular | Metabolic | Multiple | Dermatological | Renal | Skeletal | Unclassified | Gastrointestinal | Respiratory | Developmental | Connective tissue disorder

Louvain algorithm

The **Louvain** algorithm can be seen as an evolution of the previous one.

Even in this case, the objective is to maximise the modularity.

The same steps of the previous algorithm are followed.

At the end a graph is created whose nodes are the communities discovered and the procedure is repeated until it’s no longer possible increase modularity.

LiveSlides web content

To view

Download the add-in.
liveslides.com/download

Start the presentation.

Ear,Nose,Throat | Psychiatric | Endocrine | Hematological | Cancer | Muscular | Neurological | Ophthamological | Nutritional | Immunological | Bone

Cardiovascular | Metabolic | Multiple | Dermatological | Renal | Skeletal | Unclassified | Gastrointestinal | Respiratory | Developmental | Connective tissue disorder

Spinglass algorithm

The **Spinglass** algorithm uses techniques derived from physical statistics to build communities.

In particular, it allows to specify the spins parameter, which ideally represents the number *k* of clusters.

The problem with this approach is that the algorithm tries to fill all the k clusters, but it's possible that it doesn't succeed and therefore the result will show a very small number of clusters.

LiveSlides web content

To view

Download the add-in.
liveslides.com/download

Start the presentation.

Ear,Nose,Throat | Psychiatric | Endocrine | Hematological | Cancer | Muscular | Neurological | Ophthamological | Nutritional | Immunological | Bone

Cardiovascular | Metabolic | Multiple | Dermatological | Renal | Skeletal | Unclassified | Gastrointestinal | Respiratory | Developmental | Connective tissue disorder

Markov Cluster algorithm

The **Markov Cluster** algorithm is based on the simulation of stochastic paths on the graph, exploiting the clustering paradigm according to which the communities have the following property:

> *"A random path on a graph G that visits a dense cluster, most likely will come out of the cluster only after passing through most of its vertices".*

The *inflation* parameter allows to increase the granularity, this makes it possible to allow the algorithm to detect smaller clusters and therefore increase the number of clusters detected by the algorithm.

The algorithm detects 169 communities.

LiveSlides web content

To view

Download the add-in.
liveslides.com/download

Start the presentation.

Ear,Nose,Throat | Psychiatric | Endocrine | Hematological | Cancer | Muscular | Neurological | Ophthamological | Nutritional | Immunological | Bone

Cardiovascular | Metabolic | Multiple | Dermatological | Renal | Skeletal | Unclassified | Gastrointestinal | Respiratory | Developmental | Connective tissue disorder

Leiden algorithm

The **Leiden** algorithm is a further evolution of the Louvain algorithm shown above.

The creators of the algorithm have explained that Louvain often generates communities that are non optimal. Nevertheless, this algorithm converges to an optimal solution much more quickly than Louvain.

In particular with this method it’s possible to decide in advance the number of clusters and for this reason it should be able to obtain very good performances on the graph in question.

LiveSlides web content

To view

Download the add-in.
liveslides.com/download

Start the presentation.

Ear,Nose,Throat | Psychiatric | Endocrine | Hematological | Cancer | Muscular | Neurological | Ophthamological | Nutritional | Immunological | Bone

Cardiovascular | Metabolic | Multiple | Dermatological | Renal | Skeletal | Unclassified | Gastrointestinal | Respiratory | Developmental | Connective tissue disorder

Label propagation algorithm

The **Label propagation** algorithm initialises all the nodes of the network with a random label and then iteratively updates the label of each node based on a majority vote among the neighbors' labels.

The result of the algorithm depends on the initialisation, so it’s necessary to iterate multiple times to obtain consisten values.

The algorithm is very simple and extremely fast.

LiveSlides web content

To view

Download the add-in.
liveslides.com/download

Start the presentation.

Ear,Nose,Throat | Psychiatric | Endocrine | Hematological | Cancer | Muscular | Neurological | Ophthamological | Nutritional | Immunological | Bone

Cardiovascular | Metabolic | Multiple | Dermatological | Renal | Skeletal | Unclassified | Gastrointestinal | Respiratory | Developmental | Connective tissue disorder

Label propagation algorithm

The algorithm is then tested using a semi-supervised technique, initializing about 20% of the labels and seeing how it behaves.

The number of clusters detected is reduced, but the performance should potentially improve as the algorithm has a better understanding of the structure of the graph.

LiveSlides web content

To view

Download the add-in.
liveslides.com/download

Start the presentation.

Ear,Nose,Throat | Psychiatric | Endocrine | Hematological | Cancer | Muscular | Neurological | Ophthamological | Nutritional | Immunological | Bone

Cardiovascular | Metabolic | Multiple | Dermatological | Renal | Skeletal | Unclassified | Gastrointestinal | Respiratory | Developmental | Connective tissue disorder

Leading eigenvector algorithm

The **Leading eigenvector** algorithm is based on the use of *divisive clustering* with the aim of *maximizing modularity*.

These method try to find densely connected subgraphs in a graph by calculating the leading non-negative eigenvector of the modularity matrix of the graph.

At each step it separates the graph into two components so that the separation goes to increase the modularity.

The separation is determined by evaluating the main eigenvector of the modularity matrix

B = A - P.

LiveSlides web content

To view

Download the add-in.
liveslides.com/download

Start the presentation.

Ear,Nose,Throat | Psychiatric | Endocrine | Hematological | Cancer | Muscular | Neurological | Ophthamological | Nutritional | Immunological | Bone

Cardiovascular | Metabolic | Multiple | Dermatological | Renal | Skeletal | Unclassified | Gastrointestinal | Respiratory | Developmental | Connective tissue disorder

5.
Performance evaluation

Performance evaluation

Performances are measured by evaluation with *groundtruth*.

We evaluate the algorithms in terms of classification accuracy of the nodes, using different metrics.

We decided to consider the nodes with the label ”Multiple” and ”Unclassified”, as "Jolly", making sure that they are always *true positives*.

Performance evaluation

Purity

$$\frac{1}{N}\sum_{i=1}^{k} \max_{j} |C_i \cap L_j|$$

	betweenness	fastgreedy	louvain	spinglass	markov	leiden	label prop	label prop init	lead eigenvector	average
Ear,Nose...	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.111
Cancer	0.964	0.963	0.934	0.933	0.951	1.000	0.972	0.991	0.941	0.959
Opthalmolog	0.867	0.915	0.867	0.860	0.902	0.902	0.800	0.902	0.918	0.881
Endocrine	0.296	0.296	0.296	0.464	0.781	0.897	0.429	0.667	0.259	0.482
Cardiovasc	0.656	0.357	0.414	0.433	0.750	0.794	0.645	0.724	0.393	0.577
Neurologic	0.689	0.689	0.689	0.765	0.695	0.750	0.688	0.861	0.786	0.734
Hematologic	0.486	0.568	0.568	0.568	0.714	0.865	0.703	0.703	0.605	0.642
Nutritional	0.000	0.000	0.000	0.000	0.333	1.000	0.000	0.000	0.000	0.148
Muscular	0.895	0.895	0.895	0.882	1.000	0.750	0.882	0.882	0.267	0.816
Respiratory	0.000	0.000	0.000	0.000	0.400	0.000	0.000	0.000	0.000	0.044
Immunologic	0.417	0.417	0.556	0.417	0.583	0.708	0.583	0.720	0.308	0.523
Dermatologic	0.760	0.760	0.760	0.750	0.696	0.852	0.821	0.833	0.750	0.775
Psychiatric	0.000	0.000	0.000	0.556	0.556	0.500	0.556	0.556	0.000	0.302
Metabolic	0.000	0.000	0.000	0.000	0.375	0.800	0.000	0.250	0.000	0.158
Gastrointest	0.235	0.286	0.000	0.541	0.639	0.706	0.605	0.531	0.286	0.428
Bone	0.000	0.000	0.000	0.000	0.600	0.400	0.000	0.000	0.000	0.111
Skeletal	0.800	0.667	0.920	0.632	0.917	0.765	0.312	0.647	0.000	0.628
Renal	0.217	0.462	0.481	0.667	0.458	0.667	0.750	0.821	0.571	0.570
Development	0.455	0.400	0.455	0.455	0.500	0.500	0.455	0.000	0.400	0.402
Connective tissue disorder	0.000	0.000	0.000	0.000	0.333	0.588	0.529	0.562	0.000	0.223
Purity	0.614	0.612	0.607	0.653	0.748	0.812	0.692	0.764	0.581	0.676

Performance evaluation

F-Measure

$$F_i = \frac{2\pi_i \rho_i}{\pi_i + \rho_i}, \quad F(\text{macro-averaged}) = \frac{\sum_{i=1}^{M} F_i}{M}$$

	F-Measure
betweenness	0.3391
fastgreedy	0.3429
louvain	0.3312
spinglass	0.4027
markov	0.6222
leiden	0.7564
label prop	0.4457
label prop init	0.4928
lead eigenvector	0.2939

Performance evaluation

Adjusted Rand Index

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2}\sum_i \binom{b_j}{2}\right]/\binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_i \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2}\sum_i \binom{b_j}{2}\right]/\binom{n}{2}}$$

	Adjusted Rand index
betweenness	0.451
fastgreedy	0.477
louvain	0.460
spinglass	0.498
markov	0.597
leiden	0.706
label prop	0.536
label prop init	0.628
lead eigenvector	0.414

Performance evaluation

Normalized Mutual Information

$$NMI(\Omega; C) = \frac{MI(\Omega; C)}{[H(\Omega) + H(C)]/2}$$

	Normalized Mutual Information
betweenness	0.458
fastgreedy	0.475
louvain	0.472
spinglass	0.515
markov	0.642
leiden	0.720
label prop	0.551
label prop init	0.631
lead eigenvector	0.423

Performance evaluation

Analysis on "Unclassified" Diseases

The disease **Aquaporin-1 deficiency** almost always belongs to the cluster **Hematological**.

The **Benzene toxicity** certainly belongs to the cluster **Cancer**.

The **Bannayan-Riley- Ruvalcaba syndrome** seems too to belong to the cluster **Cancer.**

	Beta-2-adrenoreceptor agonist, reduced response to	Aquaporin-1 deficiency	Aneurysm, familial arterial	Benzene toxicity
Betweenness	Neurological	Hematological	Bone	Cancer
Fastgreedy	Neurological	Hematological	Ophthamological	Cancer
Louvain	Neurological	Hematological	Bone	Cancer
Spinglass	Neurological	Hematological	Bone	Cancer
Markov	Endocrine	Hematological	Bone	Cancer
Leiden	Nutritional	Hematological	Connective tissue disorder	Cancer
Label_prop	Metabolic	Hematological	Connective tissue disorder	Cancer
Label_prop_init	Immunological	Hematological	Bone	Cancer

	Alcohol dependence	van Buchem disease	Placental abruption	Bannayan-Riley-Ruvalcaba syndrome	Carpal tunnel syndrome, familial
Betweenness	Cancer	Bone	Cardiovascular	Cancer	Metabolic
Fastgreedy	Cancer	Ophthamological	Metabolic	Cancer	Hematological
Louvain	Cancer	Bone	Immunological	Cancer	Hematological
Spinglass	Psychiatric	Bone	Metabolic	Cancer	Hematological
Markov	Psychiatric	Bone	Cardiovascular	Cancer	Metabolic
Leiden	Nutritional	Unclassified	Cardiovascular	Unclassified	Hematological
Label_prop	Psychiatric	Bone	Cardiovascular	Cancer	Hematological
Label_prop_init	Psychiatric	Bone	Cardiovascular	Cancer	Hematological

Performance evaluation

Analysis on "Multiple" Diseases

The disease **Walker-Warburg syndrome** certainly belongs to the cluster **Muscular**.

The **Fanconi anemia** and the **Rubenstein- Taybi syndrome** almost always belongs to the cluster **Cancer**.

The **Dejerine-Sottas disease** seems too to belong to the cluster **Neurological.**

	Fanconi anemia	Usher syndrome	Mitochondrial complex deciency	Dejerine-Sottas disease	Waardenburg syndrome
Betweenness	Cancer	Bone	Muscular	Neurological	Cancer
Fastgreedy	Cancer	Bone	Muscular	Neurological	Cancer
Louvain	Cancer	Bone	Muscular	Neurological	Cancer
Spinglass	Cancer	Neurological	Metabolic	Neurological	Cancer
Markov	Cancer	Neurological	Metabolic	Neurological	Multiple
Leiden	Multiple	Neurological	Metabolic	Neurological	Cancer
Label_prop	Cancer	Neurological	Metabolic	Neurological	Cancer
Label_prop_init	Cancer	Neurological	Neurological	Neurological	Cancer

	Stickler syndrome	Walker-Warburg syndrome	Rubenstein-Taybi syndrome	Waardenburg-Shah syndrome	Kallmann syndrome
Betweenness	Bone	Muscular	Cancer	Cancer	Cancer
Fastgreedy	Bone	Muscular	Cancer	Cancer	Skeletal
Louvain	Bone	Muscular	Cancer	Cancer	Skeletal
Spinglass	Bone	Muscular	Cancer	Cancer	Skeletal
Markov	Skeletal	Muscular	Cancer	Multiple	Cancer
Leiden	Bone	Muscular	Multiple	Neurological	Skeletal
Label_prop	Skeletal	Muscular	Cancer	Neurological	Skeletal
Label_prop_init	Skeletal	Muscular	Cancer	Neurological	Skeletal

5. Conclusions

Conclusions

Automatically clustering the diseases is an **arduous task**.

We are unable to apply standard clustering techniques due to **lack of features**.

Less than half of the initial clusters had *good-sized* connected components.

Most diseases didn't have *neighbors* that shared the same cluster, or had a particularly low number.

The most common algorithms of community detection **fail** because they are not able to detect small clusters.

Algorithms like *Leiden* or *MCL* have superior performance because they are able to detect a large number of communities thanks to the tuning of internal parameters.

Conclusions

The Leiden algorithm (2018) was the only recent algorithm and it is in fact the one that achieves decidedly superior performances compared to the others.

Future works

Implementation and evaluation of graph partitioning algorithms.

Integration with other data sources to obtain, for example, intra-genic relationships that can be used as a feature set for clustering.

Thanks for the attention!

Credits:

- GOH, Kwang-Il, et al. The human disease network. Proceedings of the National Academy of Sciences, 2007, 104.21: 8685-8690.
- CLAUSET, Aaron; NEWMAN, Mark EJ; MOORE, Cristopher. Finding community structure in very large networks. Physical review E, 2004, 70.6: 066111.
- NEWMAN, Mark EJ. Finding community structure in networks using the eigenvectors of matrices. Physical review E, 2006, 74.3: 036104.

Repository

https://github.com/GiorgiaAuroraAdorni/human-disease-network