Università degli studi di
Milano-Bicocca

Relazione di progetto

Data Analytics

Human Disease Network

Authors:
Ruggieri Andrea - 806808 - a.ruggieri4@campus.unimib.it
Stranieri Francesco - 816551 - f.stranieri1@campus.unimib.it

20 Giugno, 2020

Indice

1 Introduzione **1**

2 Obiettivi **2**

3 Network Analysis **3**
3.1 Degree Centrality 4

4 Trasformazione del Grafo **5**

5 Network Analysis **6**
5.1 Degree Distribution 6
5.2 Eigenvector Centrality 7

6 Community Detection **8**
6.1 Selezione degli Algoritmi 8
6.1.1 Label Propagation 9
6.1.2 Vertex Partition 9
6.1.3 Infomap 9
6.2 Risultati 10
6.2.1 Label Propagation 10
6.2.2 Vertex Partition 12
6.2.3 Infomap 13
6.3 Considerazioni 14

7 Ulteriori Analisi **14**
7.1 Cliques 15
7.2 Malattie 'Unclassified' 16

8 Conclusioni **18**

A Risultati per Community **19**
A.1 Label Propagation 19
A.2 Vertex Partition 20
A.3 Infomap 21

Bibliografia **22**

Elenco delle figure

1 Visualizzazione della Human Disease Network. 3
2 Il grafo dopo le operazioni di trasformazione. 5
3 Degree Distribution del grafo trasformato. 6
4 Community individuate da Label Propagation. 11
5 Community individuate da Vertex Partition. 12
6 Community individuate da Infomap. 13
7 Cliques individuate con la tecnica del pruning con $k = 5$. 15

Elenco delle tabelle

1 Geni con il più alto indegree. 4
2 Coefficiente di clustering prima e dopo la trasformazione. 6
3 Malattie con il più alto degree. 7
4 Malattie con la più alta eigenvector. 8
5 Metriche per Label Propagation. 11
6 Metriche per Vertex Partition. 12
7 Metriche per Infomap. 13
8 Metriche per i tre algoritmi confrontati. 14
9 Metriche per la tecnica del Pruning. 16
10 Categorie predette dai tre algoritmi 16

Sommario

Gli algoritmi di Community Detection su una rete biologica svolgono un ruolo essenziale, in quanto permettono di comprenderne la struttura e le funzioni, oltre ad essere utili per predirne alcuni comportamenti e per poterla analizzare. Con questo studio si analizzerà una nota rete biologica, la *"Human Disease Network"*, al fine di coglierne le proprietà più importanti, attraverso un'attività di Network Analysis, e di valutare le connessione tra i diversi nodi, attraverso tecniche di Community Detection.

1 Introduzione

La rete Human Disease Network (HDN) è stata proposta nel 2007 da *Goh*, usando le informazioni disponibili sul database OMIM [1]. La rete in questione rappresenta diverse malattie umane, con riferimento alle loro origini genetiche. In particolare, si è deciso di analizzarla poiché risulta essere di rilievo all'interno della comunità scientifica. Tale rete ha difatti permesso di approfondire e di rappresentare in maniera chiara ed efficiente tutte quelle connessioni presenti tra le malattie e i disordini genetici.

All'interno di questo progetto si cercherà di rispondere ad alcune domande relative alla HDN (Sezione 2). Si descriveranno quindi le proprietà più importanti di questa rete, attraverso l'attività di Network Analysis (Sezione 3). Successivamente si condurrà una trasformazione della rete, in modo da renderla più compatta e facile da studiare, rimuovendo tutte le informazioni ritenute ridondanti (Sezione 4). Questo porterà ad avviare una nuova attività di Network Analysis (Sezione 5), sulla rete trasformata. Si procederà quindi con un'attività di Community Detection (Sezione 6) che, assieme ad ulteriori analisi (Sezione 7), permetterà di trarre importanti conclusioni (Sezione 8).

Stato dell'Arte

Negli ultimi anni la ricerca scientifica ha compiuto molti passi in avanti nell'individuazione delle componenti genetiche alla base delle malattie umane. Una malattia genetica è un disordine causato in tutto, o in parte, da un cambiamento nella sequenza del DNA che lo allontana dalla normale sequenza. Queste malattie possono essere causate dalla mutazione di un singolo gene (disordine monogenico), da mutazioni di più geni (disturbo ereditario multifattoriale) o da una combinazione di mutazioni di geni e fattori ambientali (come l'esposizione a sostanze inquinanti, fumo di sigarette, etc.). Dal punto di vista dell'ereditarietà, alcune malattie sono causate da mutazioni ereditate dai genitori e quindi presenti in un individuo fin dalla nascita (come la malattia falciforme, ad esempio). Altre malattie sono invece causate da mutazioni acquisite in un gene, o gruppo di geni, che possono verificarsi durante l'arco di vita di una persona [Ins18].

Ad oggi, esistono in letteratura diverse reti che offrono descrizioni sempre più dettagliate circa le relazioni tra i diversi geni che provocano specifiche malattie. La maggior parte degli studi basati su

[1] https://omim.org/

questi approcci sono concentrati sugli effetti che la mutazione di un gene comporta sulla presenza di una determinata malattia e vengono utilizzate, tra le altre, anche tecniche basate sulla Network Analysis. Questi studi hanno permesso di ottenere una maggiore comprensione circa le relazioni tra i geni implicati in una determinata patologia, come ad esempio nel caso di alcuni tumori noti [Goh+07].

Limiti

Le ricerche sull'HDN hanno come limite principale quello di basarsi esclusivamente sulle informazioni genetiche. Questo è dovuto al fatto che la maggior parte delle malattie, se non tutte, sono solo parzialmente molecolari. Difatti, anche le esposizioni ambientali condivise e diversi fattori di rischio, come quello socioeconomico, possono portare a correlazioni tra la presenza di alcune malattie. Altri limiti sono causati dalla ricerca ancora limitata, nonostante nei prossimi anni verranno verosimilmente scoperti nuovi geni e quindi nuovi legami con altrettante patologie. Anche il numero di malattie attualmente presenti nell'Human Disease Network rimane comunque circoscritto. Nonostante queste significative limitazioni, la HDN è ritenuta comunque un pilastro ed ha permesso a vari ricercatori di scoprire preziose informazioni su molte patologie [Jia+18].

2 Obiettivi

In questo progetto, si è deciso di analizzare nel dettaglio la rete Human Disease Network, concentrandosi maggiormente sui seguenti punti:

- Quali geni causano la maggior parte delle malattie? E quali tra queste sono associate al gene "più rilevante"?
- Come cambia la rete dopo averla trasformata? E cosa comporta questa trasformazione?
- Le malattie tendono ad "essere legate" tra di loro?
- Esistono malattie "più rilevanti" di altre?
- Quale algoritmo di Community Detection partiziona meglio la rete? Considerare questa come pesata ha effetto sulle partizioni?
- Esistono gruppi di malattie "ben distinte" tra di loro?
- E' possibile assegnare una categoria alle malattie classificate come *'Unclassified'*?

Una definizione su cosa si intende per *"più rilevante"*, *"essere legate"* e *"ben distinte"* verrà data nelle sezioni successive.

Implementazione

L'intero lavoro è stato svolto utilizzando il linguaggio Python. In particolare, la librerie di riferimento sono `iGraph` [2] per quanto riguarda la parte di Network Analysis, affiancata a `Louvain` [3] per la parte di Community Detection. Per la parte di *Visualization*, è stata inoltre utilizzata la piattaforma `Cytoscape` [4], la quale, in alcuni casi, permette di ottenere rappresentazioni grafiche molto significative.

3 Network Analysis

La HDN si presenta come un **grafo bipartito** dove sono presenti due diversi tipologie di nodi, ossia i geni e le malattie. Tramite la Figura 1, è possibile distinguerli: all'esterno sono presenti i geni, mentre all'interno le diverse malattie.

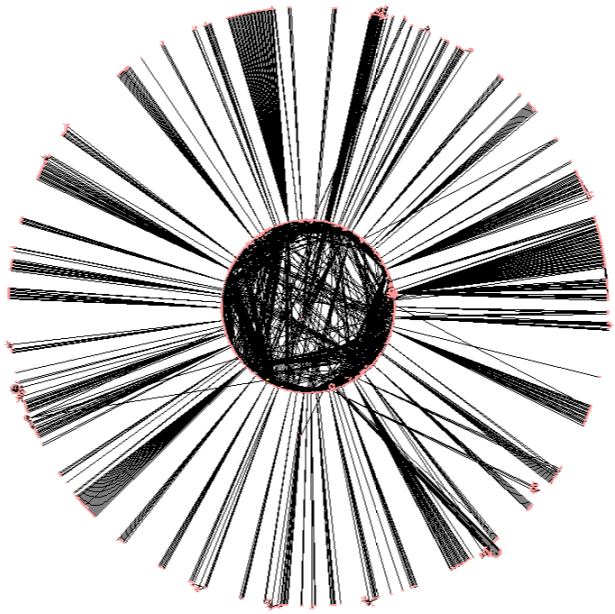

Figura 1: Visualizzazione della Human Disease Network.

In totale sono presenti 1419 nodi e 3926 archi. Questi sono orientati e hanno tutti associato un peso pari a 1. Un arco può andare da una malattia verso un gene e questo indica che la malattia è provocata dalla mutazione di quel determinato gene. Esistono inoltre degli archi che collegano due diverse malattie. Questo indica che le due malattie condividono un gene in comune. Ad ogni patologia e ad ogni gene è anche associata una label, riguardante il corrispettivo nome. Infine, ad ogni malattia è associata una categoria, la quale è stata assegnata dai ricercatori tra 22 diverse categorie,

[2] https://igraph.org/python/doc/igraph.Graph-class.html
[3] https://louvain-igraph.readthedocs.io/en/latest/
[4] https://cytoscape.org/

sulla base del sistema fisiologico colpito. Ad esempio, la malattia *Colon Cancer* ha associata la categoria *Cancer*, mentre *Alzheimer* ha assegnata la categoria *Neurological*, in quanto colpisce aree del cervello.

3.1 Degree Centrality

Attraverso il calcolo della *Degree Centrality*, è possibile rispondere alla prima domanda prefissata, ossia:

> Quali geni causano la maggior parte delle malattie?
> E quali tra queste sono associate al gene più rilevante?

Al fine di ottenere una risposta, è stato necessario computare, per tutti i nodi rappresentanti un gene, il relativo *indegree*. E' possibile quindi definire il **gene più rilevante**, come quel gene che presenta il più alto numero di link entranti, corrispondenti al numero di malattie associate.

Dalla Tabella 1 è possibile notare come il nodo *TP53* risulta essere associato a ben 11 diverse malattie.

Disease	**Degree**
TP53	11
PAX6	10
FGFR2	9
PTEN	9
FGFR3	8
MEN1	8
MSH2	8

Tabella 1: Geni con il più alto indegree.

E' possibile esaminare con maggiore attenzione il primo nodo, estraendo la lista di tutte le malattie in relazione con esso. In particolare, questo gene è associato alle seguenti patologie: *Breast Cancer, Colon Cancer, Hepatic Adenoma, Histiocytoma, Li-Fraumeni Syndrome, Nasopharyngeal Carninoma, Osteosarcoma, Pancreatic Cancer, Thyroid Carcinoma, Adnrenal Corical Carcinoma, Multiple Malignancy Syndrome*. Notiamo come tutte queste malattie risultano essere associate alla categoria *Cancer*. Come ulteriore supporto a questa nostra analisi, é possibile sottolineare quanto segue:

> Il *TP53* è un gene che istruisce la cellula a produrre la proteina del tumore (p53). Le mutazioni ereditate o somatiche in *TP53* possono provocare la perdita di controllo del ciclo cellulare. Approssimativamente, il 40% dei tumori al seno hanno mutazioni somatiche *TP53* [Sim18].

4 Trasformazione del Grafo

La struttura della rete incontrata finora può essere tuttavia migliorata. Prima di tutto, è possibile aggiungere un peso (diverso dal valore 1 presente inizialmente) per ogni arco, raffigurante il numero dei geni in comune tra due malattie. A questo punto è possibile rimuovere i nodi relativi ai geni, dal momento che risultano riportare un'informazione ridondante. Infine, è possibile rimuovere la direzione degli archi, in modo da ottenere un grafo non orientato. Questo poiché è necessario un singolo arco non orientato per indicare la relazione tra due malattie.

Queste operazioni permettono di ottenere il grafo sul quale lavoreremo e che permetterà di rispondere alle restanti domande, poste inizialmente. In Figura 2 è possibile osservare il nuovo grafo, il quale presenta ora 516 nodi (raffiguranti solo le malattie) e 2376 archi, rispetto ai 1419 nodi e 3926 archi presenti inizialmente. Ogni nodo risulta avere associata una categoria di riferimento e viene quindi visualizzato con un apposito colore, in accordo alla categoria di appartenenza. Ad esempio, il colore azzurro è riferito a tutte le malattie appartenenti alla categoria *Cancer*.

Figura 2: Il grafo dopo le operazioni di trasformazione.

E' possibile rispondere ora alla domanda:

Come cambia la rete dopo averla trasformata? E cosa comporta questa trasformazione?

Per verificare se le operazioni di trasformazione del grafo lo hanno in qualche misura "migliorato", si è deciso di calcolare la relativa transitività, meglio conosciuta come **coefficiente di clustering**. Questo coefficiente misura la probabilità con cui i vicini di un determinato nodo risultano essere connessi. Più questo valore risulta essere alto, più la rete risulterà facilmente partizionabile. I risultati sono mostrati tramite la Tabella 2 ed evidenziano come la trasformazione effettuata abbia portato dei miglioramenti, essendo il coefficiente di clustering quasi raddoppiato. Tuttavia, un valore pari a 0.43 indica che sarà comunque possibile incontrare eventuali difficoltà nel valutare gli algoritmi di Community Detection.

Clustering Coeff.	**Value**
Before Transformation	0.25
After Transformation	0.43

Tabella 2: Coefficiente di clustering prima e dopo la trasformazione.

5 Network Analysis

Tramite il grafo trasformato, è possibile condurre delle nuove analisi a livello di centralità.

5.1 Degree Distribution

La *Degree Distribution* permette di trarre alcune considerazioni rispetto alla rete presa in esame. Tramite la Figura 3, è possibile notare che la maggior parte dei nodi esibisce un grado minore di 22. Nel dettaglio, 214 nodi, corrispondenti al 41% dei nodi totali, esibiscono un grado inferiore o uguale a 5.

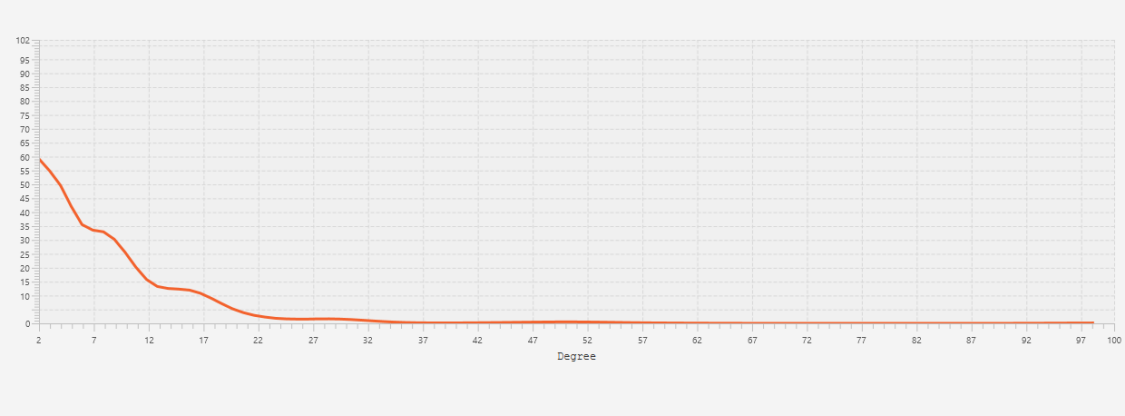

Figura 3: Degree Distribution del grafo trasformato.

Tramite queste osservazioni è possibile rispondere alla terza domanda:

Le malattie tendono ad "essere legate" tra di loro?

La Degree Distribution suggerisce che poche malattie risultano essere legate tra di loro. Più precisamente, poche malattie tendono a condividere un alto numero di geni con altre malattie. La patologia che risulta avere un *degree* maggiore risulta essere *Colon Cancer*, con un valori pari a 50. Tale risultato suggerisce che le mutazioni genetiche che causano *Colon Cancer* possono essere associate ad altre 50 diverse malattie.

Questa analisi permette anche di ottenere informazioni circa la struttura della rete. In questo caso la maggior parte dei nodi esibisce lo stesso grado, tendenzialmente con un valore compreso tra 2 e 22. Viceversa, alcuni nodi risultano essere caratterizzati da un alto grado, tra cui *Colon Cancer*, e sono mostrati tramite la Tabella 3.

Disease	**Degree**
Colon Cancer	50
Breast Cancer	30
Gastric Cancer	27
Thyroid Carcinoma	26
Leukemia	26

Tabella 3: Malattie con il più alto degree.

Questi svolgono la funzione di **hub** e la loro rimozione può implicare il partizionamento della rete in diverse componenti, con una conseguente perdita di informazioni potenzialmente utili [Ric10].

5.2 Eigenvector Centrality

Dopo aver condotto le prime analisi sul nuovo grafo trasformato, possiamo rispondere alla quarta domanda:

Esistono malattie "più rilevanti" di altre?

Si è quindi deciso di utilizzare l'*Eigenvector Centrality*, nella sua forma weighted, per definire il concetto di nodo più rilevante. In particolare, grazie all'interpretazione fornita da questa misura, è possibile considerare alcuni nodi all'interno della rete come più importanti rispetto ad altri, se questi risultano essere collegati ad altri nodi ritenuti a loro volta importanti. Possiamo quindi definire i nodi più rilevanti, tenendo in considerazione il corrispettivo valore dell'*eigenvector*.

Dalla Tabella 4 è possibile notare che le malattie con la maggiore eigenvector, risultano appartenere tutte alla categoria *Cancer*.

Disease	Eigenvector
Colon Cancer	1
Breast Cancer	0.80
Ovarian Cancer	0.61
Lymphoma	0.47
Pancreatic Cancer	0.44

Tabella 4: Malattie con la più alta eigenvector.

Questo risultato è in accordo con quello ottenuto con la *Degree Centrality*, dove *Colon Cancer* risultava essere il nodo con il più grado più alto. E' possibile quindi concludere che questo particolare nodo risulta essere il **più rilevante** all'interno della rete.

6 Community Detection

Le tecniche di *Community Detection* mirano a scoprire la struttura, il comportamento, le dinamiche e l'organizzazione di una rete complessa, trovando gruppi coesi all'interno dei quali i nodi risultano essere più simili tra loro rispetto ai nodi presenti in altri gruppi. Nel corso degli ultimi anni sono stati proposti un alto numero di algoritmi di Community Detection. Quelli da noi selezionati sono *Label Propagation*, *Vertex Partition* e *Infomap*. Sono stati scelti questi particolari algoritmi poiché risultano essere efficienti, facili da studiare ed appartenenti a tre diverse categorie. Questi metodi risultano essere inoltre *non supervisionati*; questo implica che vengono utilizzate soltanto le informazioni topologiche della rete, senza considerare la reale classe di appartenenza della malattia [Fan+18].

Il nostro studio non vuole quindi concentrarsi nell'implementazione di svariati algoritmi presenti in letteratura, ma bensì su una più ristretta e dettagliata analisi che coinvolge i tre algoritmi presentati precedentemente e appartenenti a tre diverse categorie.

6.1 Selezione degli Algoritmi

Le analisi condotte fino a questo momento, hanno permesso di tratte le prime rilevanti considerazioni. Tuttavia, ora si è interessati a studiare la rete HDN non più a livello di singoli nodi ma a livello di comunità. L'obiettivo consiste quindi nell'individuare gli algoritmi che meglio riescono a partizionare la rete, in modo tale che i diversi gruppi di malattie individuati abbiano associati la categoria corretta. E' dunque disponibile una *ground truth*, la quale permette di avere una corrispondenza con la reale *category* di ciascuna malattia. La misurazione delle performance, prende in considerazione alcune tra le principali metriche, le quali verrano descritte successivamente: *Precision*, *Recall*, *F1 Measure* e *Rand Index*. La malattia *Colon Cancer* dovrà quindi appartenere alla community *Cancer*, se classificata correttamente.

Si vuole ora rispondere alla seguente domanda:

Quale algoritmo di Community Detection partiziona meglio la rete? Considerare questa come pesata ha effetto sulle partizioni?

L'algoritmo che meglio partiziona la rete, secondo la nostra interpretazione, è quello che meglio massimizza le metriche di performance.

6.1.1 Label Propagation

Label Propagation risulta essere l'algoritmo più semplice ed efficiente dal punto di vista del tempo di esecuzione, il che lo rende ideale per reti di medie dimensioni, come quella da noi presa in esame. Questo algoritmo inizializza ogni nodo con una label univoca e successivamente, in maniera iterativa, aggiorna le label dei diversi nodi in modo tale che, al termine della computazione, ogni nodo risulta aver assegnato la label maggiormente condivisa con i suoi vicini. L'assunzione alla base è che alla fine di questo processo iterativo, i gruppi di nodi raggiungono un consenso su una label univoca, definendo così una certa comunità di malattie [Zit14].

6.1.2 Vertex Partition

L'algoritmo Vertex Partition, presente nelle libreria Louvain, cerca di suddividere la rete in diverse comunità in modo tale da ottenere, per ognuna di esse, il più alto score di modularità. Risulta essere ampiamente utilizzato su reti di medie dimensioni, superando alcuni limiti di altri algoritmi come l'*Optimal Modularity*, il quale risulta essere computazionalmente più oneroso [NG04].

6.1.3 Infomap

Infomap è il terzo e ultimo algoritmo di Community Detection preso in considerazione. Esso risolve il problema del trovare le comunità ottimali massimizzando una funzione obiettivo, chiamata *Minimum Description Length*. Questo algoritmo, basato sui principi della teoria dell'informazione, risulta essere particolarmente noto all'interno della comunità scientifica [Emm+16].

Metriche di Valutazione

La metrica Precision è stata presa in considerazione poiché si vuole valutare l'abilità di un algoritmo di non etichettare un'istanza positiva che è in realtà negativa. La Recall permette invece di valutare la capacità di un algoritmo di trovare tutte le istanze positive. La F1 è una media armonica ponderata delle metriche Precision e Recall che verrà usata per confrontare i diversi algoritmi scelti. La Rand Index permette di studiare la qualità, la correttezza e l'affidabilità dei metodi di Community Detection. Infine, la Partition Modularity valuta la bontà del partizionamento in diverse comunità; un buon partizionamento è riferito quindi ad alti valori della modularità, dal momento che la densità all'interno di una specifica comunità risulterà essere, per l'appunto, alta.

Al fine di assegnare una label ad una comunità, è stato necessario identificare la categoria di malattie più frequente all'interno della stessa; tale categoria è stata successivamente assegnata a tutti

i nodi presenti all'interno della specifica community. Queste operazioni sono state ripetute per ogni community individuata. Infine sono state valutate, per ogni algoritmo, le rispettive performance, confrontando i valori predetti con quelli presenti nella ground truth.

6.2 Risultati

In questa sezione verranno presentati, e commentati, i risultati ottenuti dai tre diversi algoritmi di Community Detection. Per ciascuno di essi verrà mostrata la struttura della rete, con evidenziate le diverse community identificate. Inoltre, verranno riportate le metriche di valutazione in modo tale da poter misurare le performance di ciascun algoritmo.

E' importante notare come Label Propagation e Vertex Partition non permettono di impostare un numero prefissato di community da identificare, in quanto sono gli algoritmi stessi a determinarne il numero ottimale.

6.2.1 Label Propagation

Le community individuate da Label Propagation, nella sua forma weighted, sono mostrate in Figura 4.

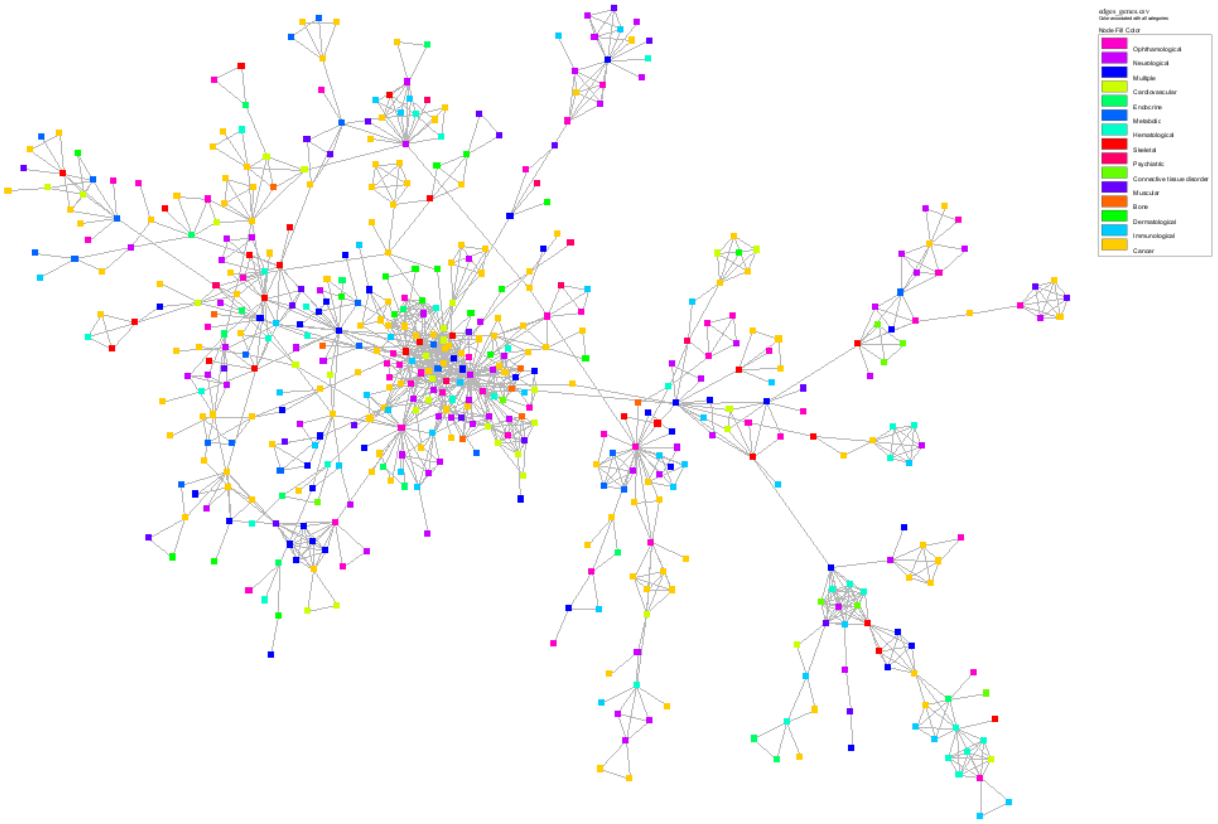

Figura 4: Community individuate da Label Propagation.

Attraverso la Tabella 6.2.1, è possibile notare che l'algoritmo individua in media un totale di 57 community, più del doppio rispetto alle reali categorie. La Rand Index mostra un valore prossimo all'1, il che indica che l'algoritmo risulta essere particolarmente affidabile. In questo specifico caso, un valore medio pari a 0.89 indica che selezionate due malattie in maniera random, se queste appartengono alla stessa categoria reale allora con buona probabilità appartengono anche alla stessa community. I valori di Precision e Recall mostrano invece dei limiti per quanto riguarda la classificazione delle singole malattie, rispetto alla reale categoria di appartenenza.

Label Propagation				
Description	**N° of communities**	**Rand Index**	**Precision**	**Recall**
Weighted	57 [±3] (Min 49, Max 64)	0.89 [±0.003]	0.54 [±0.01]	0.56 [±0.01]
Not Weighted	57 [±3] (Min 48, Max 64)	0.88 [±0.003]	0.53 [±0.002]	0.55 [±0.01]

Tabella 5: Metriche per Label Propagation.

E' importante notare che *Label Propagation* aggiorna i vertici in maniera random ad ogni iterazione. Questo non fornisce alcuna garanzia sul poter ottenere le stesse partizioni al termine di

ogni esecuzione. Per ovviare a questo problema l'algoritmo è stato eseguito per un totale di 100 esecuzioni, riportando alla fine i valori medi ottenuti con la corrispettiva deviazione standard.

6.2.2 Vertex Partition

In Figura 5 sono mostrate le community individuate da Vertex Partition, nella sua forma weighted.

Figura 5: Community individuate da Vertex Partition.

Attraverso la Tabella 6, si evince che Vertex Partition individua solamente 14 community. Questo giustifica i valori relativi a Precision e Recall, i quali risultano essere non particolarmente alti. La Rand Index mostra invece un valore pari di circa 0.85 che sottolinea comunque una buona affidabilità generale.

Vertex Partition				
Description	**N° of communities**	**Rand Index**	**Precision**	**Recall**
Weighted	14	0.85	0.4	0.47
Not Weighted	14	0.84	0.41	0.47

Tabella 6: Metriche per Vertex Partition.

6.2.3 Infomap

Le community individuate da Infomap, anche in quest'ultimo caso considerando la forma weighted, sono mostrate in Figura 6.

Figura 6: Community individuate da Infomap.

Dalla Tabella 7, è possibile notare come Infomap mostra dei risultati molto simili a quelli di Label Propagation. Il numero di comunità individuate è pari a 55, rispetto alle 57 individuate in media da Label Propagation, mentre entrambi esibiscono una Rand Index uguale a 0.89. Anche per quanto riguarda la Precision e Recall i due algoritmi mostrano risultati molto simili.

Infomap				
Description	**N° of communities**	**Rand Index**	**Precision**	**Recall**
Weighted	56	0.89	0.55	0.57
Not Weighted	55	0.89	0.55	0.57

Tabella 7: Metriche per Infomap.

6.3 Considerazioni

Label Propagation e Infomap risultano essere gli algoritmi che riportano le migliori prestazioni in termini di Precision, Recall ed F1, come visibile dalla Tabella 6.3.

Label Propagation		**Vertex Partition**		**Infomap**	
F1	**Partition Mod.**	**F1**	**Partition Mod.**	**F1**	**Partition Mod.**
0.53 [±0.01]	0.76 [±0.008]	0.41	0.83	0.55	0.78

Tabella 8: Metriche per i tre algoritmi confrontati.

Questi due algoritmi riescono quindi a classificare in maniera più precisa le malattie all'interno delle diverse comunità. Osservando la Rand Index, è possibile notare come tutti gli algoritmi presentano valori molto alti, tendenti a 1.

Ampliando ulteriormente l'analisi, si è deciso di considerare anche la *Partition Modularity*. Questa misura valuta la qualità del partizionamento: un valore alto corrisponderà quindi ad una buona suddivisione dei nodi, dal momento che la densità all'interno delle community sarà elevata. Tuttavia, è stato dimostrato un limite riguardante la Partition Modularity, ossia che tale misura subisce un limite di risoluzione e, pertanto, non è in grado di considerare le comunità di piccole dimensioni [DLT15]. Osservando i valori della Partition Modularity, l'algoritmo Vertex Partition risulta offrire le migliori prestazioni. Tale risultato risulta essere comunque poco attendibile, dal momento che Vertex Partition identifica solamente 12 comunità, rispetto a Label Propagation e Infomap che ne identificano più di 50, e risente quindi meno del limite di risoluzione.

I risultati a livello di singola community sono disponibili in Appendice A.1, A.2 e A.3. Questi mostrano come la partizione di dimensione maggiore risulta contenere per lo più le malattie appartenenti alla categoria *Cancer* e questo vale per tutti e tre gli algoritmi. In riferimento a questa partizione, Label Propagation contiene 92 nodi, mentre Infomap 45, considerando per entrambi la versione weighted. Inoltre, per questi due algoritmi è possibile notare un alto numero di partizioni con solamente 3 nodi. Questo conferma la loro difficoltà nel suddividere la rete nelle giuste community, come già previsto nella sezione 4.

Alla luce di queste considerazioni, è possibile concludere che gli algoritmi Label Propagation e Infomap risultato essere quelli maggiormente affidabili. Inoltre, è stato osservato che non esistono differenze significative, in termini prestazionali, tra le versioni pesate e non pesate.

7 Ulteriori Analisi

Per completare lo studio riguardo alla rete HDN, si è deciso di condurre due ulteriori analisi. La prima volta a identificare le cliques all'interno del grafo, mentre la seconda volta a classificare le malattie che attualmente risultano essere *'Unclassified'*.

7.1 Cliques

La domanda a cui si vuole rispondere ora è la seguente:

> Esistono gruppi di malattie "ben distinte" tra di loro?

Al fine di ottenere una risposta, è necessario scomodare il concetto di *clique* di un grafo ed, in particolare, bisogna far riferimento alla famiglia degli algoritmi di Community Detection conosciuti come *Node-Centric*. L'obiettivo consiste nel determinare le cliques massimali di malattie, ossia quei particolari sottografi, completi e massimali, che risultano essere composti da almeno 3 nodi e dove ogni nodo risulta essere adiacente agli altri. In questo modo è possibile ottenere le community che più si distinguono tra di loro.

Purtroppo, gli algoritmi *Node-Centric* risultano essere molto onerosi dal punto di vista computazionale, pertanto sono generalmente usati su reti di piccole dimensioni, generalmente con al più 50 nodi. Per ovviare a questo problema risulta necessario ricorrere ad approcci *greedy*. Tra i vari approcci, l'attenzione è stata focalizzata sulla tecnica di **pruning**, la quale permette di ridurre lo spazio di ricerca, eliminando tutti quei nodi con un grado inferiore ad un certo parametro k, fissato a priori. Tale tecnica risulta essere particolarmente utilizzata, ma può presentare dei limiti se all'interno del grafo sono presenti dei gruppi di nodi considerati come rari.

Osservando la Figura 7, è possibile distinguere diverse cliques di piccole dimensioni e una cliques di dimensione maggiore, la quale risulta essere tra l'altro molto densa.

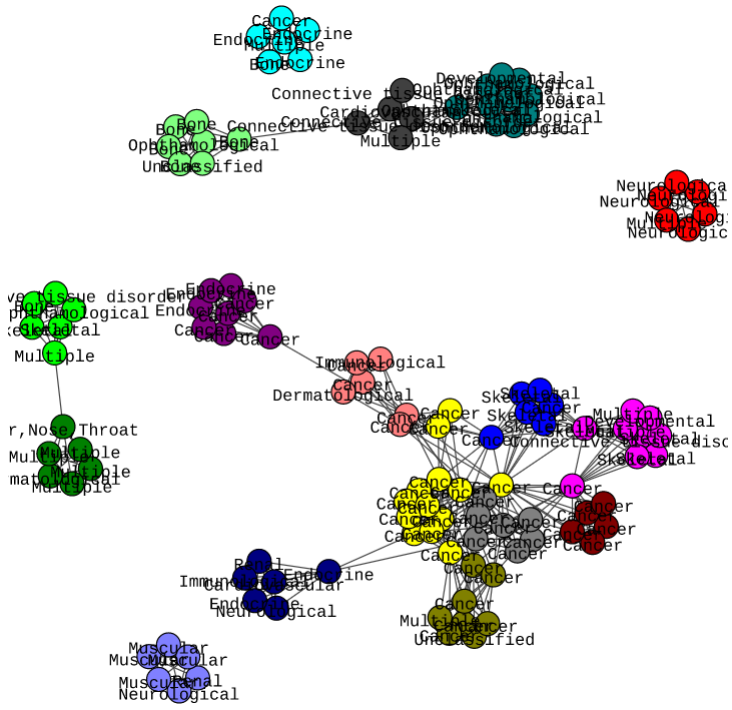

Figura 7: Cliques individuate con la tecnica del pruning con $k = 5$.

Il colore dei nodi è stato attribuito sulla base delle diverse cliques individuate, in maniera tale da contraddistinguerle meglio. E' possibile notare come nella clique di dimensione maggiore, caratterizzata dal colore giallo, la maggior parte delle malattie risultano essere associate alla categoria *Cancer*. In questa specifica clique compare anche la malattia ritenuta come più rilevante nella Sezione 5.2, ossia *Colon Cancer*. Inoltre, osservando le altre cliques, è possibile notare che le malattie appartenenti alle categorie *Neurological*, di colore rosso, *Endocrine*, di colore azzurro, e *Muscolar*, di colore violetto, risultano essere **ben distinte** tra loro.

E' possibile ora valutare le performance della tecnica del pruning, attraverso due metriche definite precedentemente: Precision e Recall. Queste riportano rispettivamente un valori pari a 0.65 e 0.70,

Pruning		
Description	**Precision**	**Recall**
Not Weighted	0.65	0.7

Tabella 9: Metriche per la tecnica del Pruning.

e permettono così di affermare che l'approccio greedy adottato, nonostante le diverse limitazioni, ha permesso di ottenere comunque risultati soddisfacenti. Possiamo quindi concludere che esistono gruppi di malattie che **ben si distinguono** dalle altre.

7.2 Malattie 'Unclassified'

Durante le analisi condotte, è stato possibile notare che la rete HDN presenta al suo interno 9 malattie appartenenti categoria *'Unclassified'*. Utilizzando gli algoritmi di Community Detection, studiati nella Sezione 6, si vuole quindi classificarle, andando al contempo a rispondere alla domanda:

E' possibile assegnare una categoria alle malattie classificate come 'Unclassified'??

Attraverso la tabella 10, è possibile osservare la specifica categoria che ciascun algoritmo ha assegnato per ognuna di queste malattie.

Disease	**Categories**		
	Label Propagation	**Infomap**	**Vertex Partition**
Alcohol Dependence	Psychiatric	Psychiatric	Psychiatric
Placental Abruption	Cardiovascular	Cardiovascular	Bone
Carpal Tunnel Syndrome	Metabolic	Metabolic	Metabolic
Aquaporin-1 Deficiency	Hematological	Hematological	Hematological
Beta-2-adrenoreceptor Agonist	Immunological	Neurological	Immunological
Bannayan-Riley-Ruvalcaba Syndrome	Cancer	Cancer	Cancer
Benzene Toxicity	Cancer	Cancer	Cancer
Van Buchem Disease	Bone	Bone	Bone
Aneurysm, Familial Arterial	Connective tissue disorder	Bone	Bone

Tabella 10: Categorie predette dai tre algoritmi

Su un totale di 9 diverse malattie, è possibile notare che ben 6 volte i tre algoritmi hanno raggiunto un'unanimità. Nel dettaglio, queste malattie risultano essere:

- *Alcohol dependence*: caratterizzata da un comportamento di ricerca compulsiva di bevande alcoliche. Tale disturbo viene trattato in ambito psichiatrico [Min17].

- *Carpal Tunnel Syndrome*: chiamata anche compressione del nervo mediano, è una condizione che causa intorpidimento, formicolio o debolezza della mano. Tali sintomi accadono a causa della pressione esercitata sul nervo mediano. Questo disturbo viene spesso classificato come *Neurological*, anche se può essere assegnata a più categorie [BU07].

- *Aquaporin-1 Deficiency*: tale disturbo viene classificato come una malattia del sangue [19a].

- *Bannayan-Riley-Ruvalcaba Syndrome*: è una malattia congenita rara con poliposi amartomatosa intestinale, lipomi, macrocefalia e lentigginosi genitale. Tale malattia viene classificata come *Cancer*, in quanto presenta caratteristiche simili al Lipoma provocate dalla mutazione del gene PTEN (già incontrato precedentemente attraverso la Tabella 1) [Gon+13].

- *Benzene Toxicity*: si tratta di una patologia correlata con il tumore al seno e alla Leucemia. Attuamente viene associato al *Cancer* [19b].

- *Ban Buchem Disease*: l'iperostosi corticale generalizzata, nota anche come malattia di van Buchem, è un'iperostosi cranio-tubolare rara, caratterizzata da iperostosi del cranio, della mandibola, della clavicole, delle costole, delle diafisi delle ossa lunghe e delle ossa tubolari delle mani e dei piedi. Tale malattia può quindi essere classificata in *Bone* [17].

Attraverso le fonti mediche a nostra disposizione, è possibile affermare che Vertex Partition, Label Propagation e Infomap riescono a classificare sensatamente queste malattie rispetto alla reale categoria di appartenenza. L'unico dubbio coinvolge il disturbo *Carpal Tunnel Syndrome*, il quale viene considerato come appartenere alla categoria *Metabolic* invece che alla più plausibile *Neurological*.

Per concludere, è possibile considerare le restanti malattie, per le quali i tre algoritmi non hanno sono riusciti a raggiungere un'unanimità:

- *Placental Abruption*: è una complicazione rara ma grave che può manifestarsi durante la gravidanza e che può ridurre, o bloccare, l'apporto di ossigeno e nutrienti da parte del bambino e causare gravi emorragie nella madre. Questa malattia risulta appartenere alla famiglia delle patologie cardiovascolari [And17].

- *Beta-2-adrenoreceptor*: si tratta di un disturbo che coinvolge le cellule e in particolare la produzione di amminoacidi. Tale disturbo viene classificato solitamente come *Neurological* [07].

- *Aneurysm, Familial Arterial*: è una malattia ereditaria rara che colpisce prevalentemente l'aorta e il tessuto connettivo. Questa malattia è classificata come *Cardiovascular*, ma può avere altre classificazioni tra cui *Connective tissue disorder* [19c].

Queste risultano essere più complicate da classificare anche per la comunità scientifica, in quanto possono appartenere a più categorie. Nonostante questo, Infomap riesce comunque a classificare sensatamente 2 malattie su 3, Label Propagation 1 su 3, mentre Vertex Partition nessuna.

8 Conclusioni

All'interno di questo studio si è analizzata la rete Human Disease Network, la quale risulta essere di rilievo all'interno della comunità scientifica e che mostra le connessioni presenti tra le malattie e i disordini genetici. Un primo studio di Network Analysis sulla rete originale ha permesso di individuare il gene *TP53* come quello causante la maggior parte delle malattie e, conseguentemente, si sono potute individuare tutte quelle malattie associate a tale gene. Successivamente si è cercato di migliorare la struttura della rete, rimuovendo le informazioni ritenute superflue e andando a studiare gli effetti causati da tale trasformazione. Per verificarne l'effettivo miglioramento, è stato preso in considerazione il coefficiente di clustering, il quale ha mostrato un valore quasi raddoppiato al termine della trasformazione. Su questo nuovo grafo è stato possibile condurre una nuova attività di Network Analysis, la quale ha permesso di verificare, attraverso la Degree Distribution, come poche malattie tendono ad essere legate tra di loro. Questa attività ha permesso di definire anche la malattia *Colon Cancer* come essere la più rilevante, sulla base del valore dell'Eigenvector Centrality. In seguito, sono stati studiati tre algoritmi di Community Detection appartenenti a tre diverse categorie, arrivando alla conclusione che Label Propagation ed Infomap risultano partizionare meglio la rete, sulla base di specifiche metriche. Inoltre, considerare il grafo come non pesato non ha portato a differenze sostanziali, per nessuno degli algoritmi. Infine, sono state condotte due ulteriori analisi. Attraverso la tecnica del pruning è stato possibile verificare che esistono gruppi di malattie che ben si distinguono dalle altre. Attraverso gli algoritmi di Community Detection studiati in precedenza, è stato invece possibile assegnare, con buoni risultati, una categoria a tutte quelle malattie che ne erano inizialmente sprovviste.

A Risultati per Community

A.1 Label Propagation

Label Propagation (Weighted)				Label Propagation (Not Weighted			
Community ID	**Size**	**Most Frequent Disease (MFD)**	**MFD Frequency**	**Community ID**	**Size**	**Most Frequent Disease (MFD)**	**MFD Frequency**
8	92	Cancer	0,64	7	125	Cancer	0,54
22	18	Neurological	0,39	27	22	Muscular	0,41
41	16	Ophthamological	1,00	20	19	Cardiovascular	0,32
23	16	Cardiovascular	0,38	35	18	Ophthamological	0,89
24	16	Metabolic	0,25	13	15	Hematological	0,60
1	15	Ophthamological	0,67	19	13	Immunological	0,38
15	15	Hematological	0,60	21	13	Metabolic	0,31
34	15	Muscular	0,53	23	12	Cancer	0,58
18	13	Neurological	0,31	3	12	Multiple	0,33
27	12	Cancer	0,58	16	12	Neurological	0,25
3	12	Multiple	0,33	24	11	Ophthamological	0,73
5	11	Bone	0,27	15	11	Metabolic	0,27
9	10	Skeletal	0,50	34	10	Bone	0,50
2	10	Neurological	0,40	2	10	Neurological	0,40
31	9	Dermatological	0,89	28	9	Dermatological	0,89
35	9	Cancer	0,78	5	9	Skeletal	0,33
56	8	Endocrine	1,00	9	8	Cardiovascular	0,88
11	8	Cardiovascular	0,88	6	8	Skeletal	0,50
6	8	Skeletal	0,50	43	8	Muscular	0,50
39	8	Bone	0,50	25	7	Ophthamological	0,71
40	8	Connective tissue disorder	0,50	31	7	Immunological	0,71
48	8	Muscular	0,50	17	7	Psychiatric	0,57
17	8	Metabolic	0,38	49	7	Hematological	0,57
28	7	Ophthamological	0,71	42	7	Skeletal	0,29
20	7	Psychiatric	0,57	50	6	Neurological	0,83
53	7	Hematological	0,57	32	6	Multiple	0,67
12	7	Endocrine	0,43	37	6	Connective tissue disorder	0,67
47	7	Skeletal	0,29	39	6	Neurological	0,67
54	6	Neurological	0,83	10	6	Endocrine	0,50
36	6	Immunological	0,67	18	6	Metabolic	0,50
44	6	Neurological	0,67	36	6	Multiple	0,50
21	6	Metabolic	0,50	33	6	Multiple	0,33
42	6	Multiple	0,50	11	5	Immunological	1,00
13	5	Immunological	1,00	22	5	Neurological	1,00
26	5	Neurological	1,00	52	5	Endocrine	1,00
7	5	Skeletal	0,80	14	5	Hematological	0,80
16	5	Hematological	0,80	44	5	Neurological	0,80
37	5	Multiple	0,80	26	5	Connective tissue disorder	0,60
49	5	Neurological	0,80	29	5	Neurological	0,60
29	5	Connective tissue disorder	0,60	8	5	Ophthamological	0,40
32	5	Neurological	0,60	12	5	Multiple	0,40
10	5	Ophthamological	0,40	40	5	Metabolic	0,40
14	5	Multiple	0,40	41	5	Multiple	0,40
30	5	Multiple	0,40	45	5	Multiple	0,40
38	5	Multiple	0,40	51	5	Cancer	0,40
46	5	Multiple	0,40	46	4	Hematological	0,75
50	5	Multiple	0,40	30	4	Skeletal	0,50
55	5	Cancer	0,40	4	3	Metabolic	1,00
43	4	Cancer	1,00	38	3	Cancer	1,00
19	4	Multiple	0,75	47	3	Neurological	1,00
51	4	Hematological	0,75	48	3	Dermatological	1,00
33	4	Skeletal	0,50	1	3	Ophthamological	0,67
45	4	Metabolic	0,50				
4	3	Metabolic	1,00				
52	3	Dermatological	1,00				
25	3	Neurological	0,33				
57	2	Endocrine	0,50				

A.2 Vertex Partition

Vertex Partition (Weighted)				Vertex Partition (Not Weighted)			
Community ID	**Size**	**Most Frequent Disease (MFD)**	**MFD Frequency**	**Community ID**	**Size**	**Most Frequent Disease (MFD)**	**MFD Frequency**
1	105	Cancer	0,590476	1	90	Cancer	0,644444
4	60	Cancer	0,383333	9	58	Neurological	0,5
8	52	Neurological	0,519231	5	54	Muscular	0,462963
6	51	Ophthamological	0,54902	4	48	Cancer	0,458333
2	39	Multiple	0,205128	2	44	Metabolic	0,181818
3	37	Metabolic	0,189189	7	41	Hematological	0,439024
12	28	Cardiovascular	0,678571	3	37	Multiple	0,189189
10	28	Bone	0,464286	6	28	Ophthamological	0,642857
5	26	Hematological	0,346154	12	24	Immunological	0,708333
7	24	Dermatological	0,541667	13	24	Skeletal	0,625
9	21	Muscular	0,380952	8	24	Dermatological	0,541667
13	16	Endocrine	0,75	14	16	Multiple	0,5
11	15	Hematological	0,6	10	16	Bone	0,375
14	14	Skeletal	0,5	11	14	Cardiovascular	0,714286

A.3 Infomap

Infomap (Weighted)				Infomap (Not Weighted)			
Community ID	**Size**	**Most Frequent Disease (MFD)**	**MFD Frequency**	**Community ID**	**Size**	**Most Frequent Disease (MFD)**	**MFD Frequency**
8	45	Cancer	0,67	8	50	Cancer	0,78
23	38	Metabolic	0,18	9	20	Cancer	0,40
19	22	Multiple	0,27	23	19	Cardiovascular	0,32
39	17	Cancer	0,41	39	18	Ophthamological	0,89
37	16	Muscular	0,56	24	18	Metabolic	0,33
7	15	Skeletal	0,60	26	17	Cancer	0,59
16	15	Hematological	0,60	1	16	Ophthamological	0,63
12	13	Cancer	0,77	33	16	Muscular	0,56
29	13	Ophthamological	0,62	18	16	Multiple	0,31
14	12	Immunological	0,83	15	15	Hematological	0,60
28	12	Cancer	0,58	36	14	Cancer	0,29
40	12	Cancer	0,58	16	12	Hematological	0,67
22	12	Immunological	0,42	27	12	Cancer	0,58
27	11	Cancer	0,82	37	12	Cancer	0,58
3	11	Multiple	0,36	22	12	Immunological	0,42
43	10	Ophthamological	0,80	3	12	Multiple	0,33
6	10	Skeletal	0,50	17	11	Metabolic	0,27
41	10	Bone	0,50	30	10	Dermatological	0,60
2	10	Neurological	0,40	6	10	Skeletal	0,50
13	9	Cancer	1,00	38	10	Bone	0,50
34	9	Dermatological	0,89	2	10	Neurological	0,40
42	8	Ophthamological	1,00	5	9	Multiple	0,22
10	8	Cardiovascular	0,88	11	8	Cardiovascular	0,88
20	8	Psychiatric	0,50	49	8	Dermatological	0,88
50	8	Muscular	0,50	46	8	Muscular	0,50
18	8	Metabolic	0,38	54	7	Endocrine	1,00
55	7	Endocrine	1,00	28	7	Ophthamological	0,71
17	7	Hematological	0,71	34	7	Immunological	0,71
30	7	Ophthamological	0,71	20	7	Psychiatric	0,57
11	7	Endocrine	0,43	45	7	Skeletal	0,29
5	7	Multiple	0,29	52	6	Neurological	0,83
49	7	Skeletal	0,29	41	6	Connective tissue disorder	0,67
54	6	Neurological	0,83	42	6	Neurological	0,67
45	6	Connective tissue disorder	0,67	12	6	Endocrine	0,50
46	6	Neurological	0,67	21	6	Metabolic	0,50
21	6	Metabolic	0,50	40	6	Multiple	0,50
44	6	Multiple	0,50	13	5	Immunological	1,00
25	5	Neurological	1,00	25	5	Neurological	1,00
33	5	Dermatological	0,80	7	5	Skeletal	0,80
51	5	Neurological	0,80	47	5	Neurological	0,80
31	5	Connective tissue disorder	0,60	29	5	Connective tissue disorder	0,60
35	5	Neurological	0,60	31	5	Neurological	0,60
56	5	Hematological	0,60	10	5	Ophthamological	0,40
9	5	Ophthamological	0,40	14	5	Multiple	0,40
15	5	Multiple	0,40	43	5	Metabolic	0,40
47	5	Metabolic	0,40	44	5	Multiple	0,40
48	5	Multiple	0,40	48	5	Multiple	0,40
52	5	Multiple	0,40	50	4	Hematological	0,75
26	4	Endocrine	0,50	32	4	Skeletal	0,50
36	4	Skeletal	0,50	35	4	Cancer	0,50
38	4	Cancer	0,50	4	3	Metabolic	1,00
4	3	Metabolic	1,00	51	3	Neurological	1,00
53	3	Neurological	1,00	55	3	Cancer	1,00
1	3	Ophthamological	0,67	19	3	Multiple	0,67
24	3	Neurological	0,33	53	3	Cancer	0,33
32	3	Multiple	0,33				

Bibliografia

[Ric10] Laura Ricci. *Analisi di reti complesse, Università degli Studi di Pisa*. Mar 19, 2010. URL: http://pages.di.unipi.it/ricci/19-03-10-NetworkAnalysis.pdf.

[Sim18] Hannah Simmons. *Che cosa è TP53?* Aug 23, 2018. URL: https://www.news-medical.net/life-sciences/What-is-TP53-(Italian).aspx.

[Ins18] National Human Genome Research Institute. *Genetic Disorders*. May 18, 2018. URL: https://www.genome.gov/For-Patients-and-Families/Genetic-Disorders.

[NG04] M. E. J. Newman e M. Girvan. «Finding and evaluating community structure in networks». In: *Phys. Rev. E* 69 (2 feb. 2004), p. 026113. DOI: 10.1103/PhysRevE.69.026113. URL: https://link.aps.org/doi/10.1103/PhysRevE.69.026113.

[BU07] K. Balci e U. Utku. «Carpal tunnel syndrome and metabolic syndrome». In: *Acta Neurologica Scandinavica* 116.2 (2007), pp. 113–117. DOI: 10.1111/j.1600-0404.2007.00797.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1600-0404.2007.00797.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1600-0404.2007.00797.x.

[Goh+07] Kwang-Il Goh et al. «The human disease network». In: *Proceedings of the National Academy of Sciences* 104.21 (2007), pp. 8685–8690. ISSN: 0027-8424. DOI: 10.1073/pnas.0701361104. eprint: https://www.pnas.org/content/104/21/8685.full.pdf. URL: https://www.pnas.org/content/104/21/8685.

[07] *I.P. Hall, in Encyclopedia of Respiratory Medicine, 2006*. 2007. URL: https://www.sciencedirect.com/topics/neuroscience/beta-2-adrenergic-receptor.

[Gon+13] Gabriela Maria Abreu Gontijo et al. «Bannayan-Riley-Ruvalcaba syndrome with deforming lipomatous hamartomas in infant - Case report». en. In: *Anais Brasileiros de Dermatologia* 88 (dic. 2013), pp. 982–985. ISSN: 0365-0596. URL: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0365-05962013000600982&nrm=iso.

[Zit14] Marinka Zitnik. «The Anatomy of a Human Disease Network». In: *XRDS* 21.2 (dic. 2014), pp. 58–60. ISSN: 1528-4972. DOI: 10.1145/2684197. URL: https://doi.org/10.1145/2684197.

[DLT15] T. N. Dinh, X. Li e M. T. Thai. «Network Clustering via Maximizing Modularity: Approximation Algorithms and Theoretical Limits». In: *2015 IEEE International Conference on Data Mining*. 2015, pp. 101–110.

[Emm+16] Scott Emmons et al. «Analysis of network clustering algorithms and cluster quality metrics at scale». In: *PloS one* 11.7 (2016).

[And17] Ananth CV. Hansen AV Williams MA. Nybo Andersen. «Cardiovascular Disease in Relation to Placental Abruption: A Population-Based Cohort Study from Denmark». In: *Paediatr Perinat Epidemiol.* Vol. 31(3):209-218. 2017. DOI: 10.1111/ppe.12347.

[17] *Hsu, Shang-Fu, and Chen-Chun Lin. "Van Buchem disease: First case report in Taiwan." Medicine vol.* 2017. DOI: "doi:10.1097/MD.0000000000009209".

[Min17] State of Mind. *Alcol.* 2017. URL: https://www.stateofmind.it/tag/alcool/#:~:text=La%20dipendenza%20alcolica%2C%20o%20alcolismo,sempre%20maggiori%20di%20bevande%20alcoliche)..

[Fan+18] L. Fan et al. «Semi-Supervised Community Detection Based on Distance Dynamics». In: *IEEE Access* 6 (2018), pp. 37261–37271.

[Jia+18] Yefei Jiang et al. «An epidemiological human disease network derived from disease Co-occurrence in Taiwan». English. In: *Scientific Reports* 8.1 (dic. 2018). ISSN: 2045-2322. DOI: 10.1038/s41598-018-21779-y.

[19a] *OMIM.* 2019. URL: https://omim.org/entry/107776.

[19b] *OMIM.* 2019. URL: https://omim.org/entry/125860.

[19c] *OMIM.* 2019. URL: https://omim.org/entry/607086.