POLITECNICO DI MILANO

Scuola di Ingegneria Industriale e dell'Informazione
Laurea Magistrale in Mathematical Engineering

Analisi di comunitá in una rete bipartita pazienti/diagnosi di ricoveri ospedalieri in Regione Lombardia

Relatore: Prof. Carlo PICCARDI
Correlatore: Prof. Francesca IEVA

Elaborato di tesi di
Monica GIORDANO
Matr. 854561

Anno Accademico 2017-2018

Indice

Introduzione **1**

1 Presentazione del Dataset **3**
1.1 Il Dateset originale 3
1.2 Il Dataset ridotto 5

2 Introduzione alle reti complesse **11**
2.1 Caratteristiche generali 11
2.2 Proprietà delle reti 14
2.3 Misure di centralità 17
2.4 Reti Bipartite 21

3 Definizione delle Reti analizzate **25**
3.1 Costruzione della Rete Ricoveri 25
3.2 Proiezione della Rete Ricoveri: Rete Diagnosi 29
3.3 Proiezione della Rete Ricoveri: Rete Pazienti 49

4 Metodi di Community Detection **54**
4.1 Introduzione al problema e ai metodi 54
4.2 Metodo di Louvain 58
4.3 Modello Stocastico a Blocchi 61
4.4 Valutazione delle partizioni 70
4.5 Classificazione dei ruoli dei nodi 74

5 Analisi di comunità della Rete Diagnosi e della Rete Pazienti **79**
5.1 Analisi di comunità della Rete Diagnosi:
Metodo di Louvain 80
5.2 Analisi di comunità della Rete Diagnosi:
Modello stocastico a blocchi 87
5.3 Classificazione dei nodi della Rete Diagnosi 101

5.4 Analisi di comunità della Rete Pazienti: Metodo di Louvain 107

Conclusioni e sviluppi futuri **109**

Appendici **112**

A Analisi delle comunità individuate con il metodo SBM **113**

B Classificazione dei nodi **127**

C Pacchetto igraph **133**

Bibliografia **141**

Elenco delle figure

2.1 Rete della scuola di karatè di Zachary. 12
2.2 Esempio di rete pesata e di rete diretta. 13
2.3 Esempio di rete multigrafo. 14
2.4 Esempio di rete con componente gigante e nodi isolati. 14
2.5 Esempio di reti assortativa e disassortativa. 17
2.6 Esempio di rete con nodo di grado e strength massimi. 18
2.7 Esempio di rete con un nodo con alto valore di betweenness. . 19
2.8 Esempio di rete con un nodo con alto valore di betweenness ma basso grado. 20
2.9 Rete bipartita delle Donne del Sud. 22
2.10 Esempio di rete bipartita e delle proiezioni sui due insiemi di nodi. 23
2.11 Esempio di rete bipartita e di proiezione pesata. 23

3.1 Rete bipartita dei ricoveri in versione ridotta. 27
3.2 Distribuzione di grado dei nodi Paziente e Diagnosi. 27
3.3 Diametro, Rete Diagnosi. 33
3.4 Distribuzione di grado, Rete Diagnosi. 34
3.5 Distribuzione cumulativa di grado, Rete Diagnosi. 34
3.6 Distribuzione di grado nelle varie MDC. 36
3.7 Funzione grado medio dei nodi vicini, Rete Diagnosi. 37
3.8 Distribuzione di strength, Rete Diagnosi. 38
3.9 Distribuzione cumulativa di strength, Rete Diagnosi. 38
3.10 Distribuzione di strength nelle varie MDC. 40
3.11 Confronto tra grado e strength dei nodi, Rete Diagnosi. 40
3.12 Confronto tra numero di occorrenze e grado e tra numero di occorrenze e strength dei nodi, Rete Diagnosi. 41
3.13 Distribuzione cumulativa di betweenness, Rete Diagnosi. . . . 42
3.14 Distribuzione di betweenness nelle varie MDC. 43
3.15 Confronto tra grado e betweenness dei nodi, Rete Diagnosi. . . 45
3.16 Confronto tra strength e betweenness dei nodi, Rete Diagnosi. 47

4.1 Esempio di rete con una partizione in 3 comunità. 55
4.2 Visualizzazione della procedura dell'algoritmo di Louvain. . . . 60
4.3 Metodo SBM standard vs Metodo SBM Degree-Corrected. . . 69
4.4 Esempio di ruoli dei nodi in una rete. 75
4.5 Esempio di classificazione dei nodi in una rete. 76
4.6 Regioni di assegnazione dei ruoli dei nodi nello spazio dei parametri P-z. 77

5.1 Numero di ricoveri e strength, Louvain. 83
5.2 Matrici di Precision e Recall. MDC vs comunità Louvain. . . . 86
5.3 Numero di ricoveri e di strength, SBM $K = 5$. 94
5.4 Matrici di Precision e Recall. MDC vs comunità SBM $K = 5$. 94
5.5 Numero di ricoveri e di strength, SBM $K = 6$. 98
5.6 Matrici di Precision e Recall. MDC vs comunità SBM $K = 6$. 98
5.7 Matrici di Precision e Recall. Comunità Louvain vs comunità SBM $K = 6$. 100
5.8 Ruolo dei nodi, Louvain. 104
5.9 Ruolo dei nodi, SBM $K = 5$. 105
5.10 Ruolo dei nodi, SBM $K = 6$. 106

A.1 Numero di ricoveri e di strength, SBM $K = 3$. 116
A.2 Matrici di Precision e Recall. MDC vs comunità SBM $K = 3$. 116
A.3 Numero di ricoveri e di strength, SBM $K = 4$. 118
A.4 Matrici di Precision e Recall. MDC vs comunità SBM $K = 4$. 118
A.5 Numero di ricoveri e di strength, SBM $K = 7$. 120
A.6 Matrici di Precision e Recall. MDC vs comunità SBM $K = 7$. 120
A.7 Numero di ricoveri e di strength, SBM $K = 8$. 122
A.8 Matrici di Precision e Recall. MDC vs comunità SBM $K = 8$. 122
A.9 Numero di ricoveri e di strength, SBM $K = 9$. 124
A.10 Matrici di Precision e Recall. MDC vs comunità SBM $K = 9$. 124

B.1 Ruolo dei nodi, SBM $K = 3$. 128
B.2 Ruolo dei nodi, SBM $K = 4$. 129
B.3 Ruolo dei nodi, SBM $K = 7$. 130
B.4 Ruolo dei nodi, SBM $K = 8$. 131
B.5 Ruolo dei nodi, SBM $K = 9$. 132

Elenco delle tabelle

1.1 Estratto dall'elenco di classificazione dei DRG. 5
1.2 Elenco dei codici MDC. 6
1.3 Caratteristiche del dataset ridotto. 8
1.4 Riassunto delle frequenze di ricovero. 8
1.5 Numero di ricoveri per anno di riferimento. 9
1.6 Numero di ricoveri, DRG e diagnosi per ciascuna MDC. 9
1.7 Numero di pazienti per genere. 10
1.8 Numero di pazienti per fascia di età e genere. 10

3.1 Frequenza del valore di grado per i nodi Paziente e i nodi Diagnosi. 28
3.2 Informazioni generali sulla Rete Ricoveri. 28
3.3 Quadro delle diagnosi nella massima componente connessa. . . 31
3.4 Quadro del numero di ricoveri per diagnosi, classificate secondo MDC. 31
3.5 Diagnosi di con numero di occorrenze massimo per ogni MDC. 32
3.6 Quadro del valore del grado dei nodi diagnosi, classificate secondo MDC. 35
3.7 Diagnosi di grado massimo per ogni MDC. 35
3.8 Quadro del valore di strength dei nodi diagnosi, classificate secondo MDC. 39
3.9 Diagnosi di strength massima per ogni MDC. 39
3.10 Quadro del valore di betweenness dei nodi diagnosi, classificate secondo MDC. 43
3.11 Diagnosi di betweenness massima per ogni MDC. 43
3.12 Quadro del valore di betweenness pesata dei nodi diagnosi, classificate secondo MDC. 44
3.13 Diagnosi di betweenness pesata massima per ogni MDC. . . . 44
3.14 Esempi di diagnosi con valore di betweenness sopra la media e valore di grado sotto la media. 46

3.15 Esempi di diagnosi con valore di betweenness sotto la media e valore di grado sopra la media. 46
3.16 Esempi di diagnosi con valore di betweenness (pesata) sopra la media e valore di strength sotto la media. 48
3.17 Esempi di diagnosi con valore di betweenness (pesata) sotto la media e valore di strength sopra la media. 48
3.18 Numero di ricoveri nella rete campionata, per genere e fasce di età. 50
3.19 Dimensioni della Rete Pazienti completa e ridotta. 50
3.20 Quadro dei valori di grado dei nodi paziente. 52
3.21 Quadro dei valori di strength dei nodi paziente. 52

4.1 Performance dell'algoritmo di Louvain su reti test. 61

5.1 Informazioni generali sulla Rete Diagnosi. 80
5.2 Cardinalità delle comunità, Louvain. 82
5.3 Probabilità di persistenza, Louvain. 82
5.4 NMI e VI, MDC vs Comunità Louvain. 84
5.5 Valori di log-verosimiglianza per le partizioni SBM. 90
5.6 Valori di modularità per le partizioni SBM. 90
5.7 Cardinalità delle comunità, SBM $K = 5$. 93
5.8 Probabilità di persistenza, SBM $K = 5$. 93
5.9 NMI e VI, MDC vs Comunità SBM $K = 5$. 93
5.10 Matrice ω della partizione SBM con $K = 5$. 95
5.11 Cardinalità delle comunità, SBM $K = 6$. 97
5.12 Probabilità di persistenza, SBM $K = 6$. 97
5.13 NMI e VI, MDC vs Comunità SBM $K = 6$ 97
5.14 Matrice ω della partizione SBM con $K = 6$. 99
5.15 NMI e VI, Comunità Louvain vs Comunità SBM $K = 6$. . . . 99
5.16 Quadro dei ruoli dei nodi nella partizione di Louvain. 103
5.17 Quadro dei ruoli dei nodi nella partizione SBM $K = 5$. 103
5.18 Quadro dei ruoli dei nodi nella partizione SBM $K = 6$. 103
5.19 Elenco diagnosi *hub* per Louvain. 104
5.20 Elenco diagnosi *hub* per SBM con $K = 5$. 105
5.21 Elenco diagnosi *hub* per SBM con $K = 6$. 106
5.22 Informazioni generali sulla Rete Pazienti. 107

A.1 Cardinalità delle comunità, SBM $K = 3$. 115
A.2 Probabilità di persistenza, SBM $K = 3$. 115
A.3 NMI e VI, MDC vs Comunità SBM $K = 3$. 115
A.4 Cardinalità delle comunità, SBM $K = 4$. 117

A.5 Probabilità di persistenza, SBM $K = 4$. 117
A.6 NMI e VI, MDC vs Comunità SBM $K = 4$. 117
A.7 Cardinalità delle comunità, SBM $K = 7$. 119
A.8 Probabilità di persistenza, SBM $K = 7$. 119
A.9 NMI e VI, MDC vs Comunità SBM $K = 7$. 119
A.10 Cardinalità delle comunità, SBM $K = 8$. 121
A.11 Probabilità di persistenza, SBM $K = 8$. 121
A.12 NMI e VI, MDC vs Comunità SBM $K = 8$. 121
A.13 Cardinalità delle comunità, SBM $K = 9$. 123
A.14 Probabilità di persistenza, SBM $K = 9$. 123
A.15 NMI e VI, MDC vs Comunità SBM $K = 9$. 123
A.16 Matrice ω della partizione SBM con $K = 3$. 125
A.17 Matrice ω della partizione SBM con $K = 4$. 125
A.18 Matrice ω della partizione SBM con $K = 7$. 125
A.19 Matrice ω della partizione SBM con $K = 8$. 126
A.20 Matrice ω della partizione SBM con $K = 9$. 126

B.1 Elenco diagnosi *hub* per SBM con $K = 3$. 128
B.2 Elenco diagnosi *hub* per SBM con $K = 4$. 129
B.3 Elenco diagnosi *hub* per SBM con $K = 7$. 130
B.4 Elenco diagnosi *hub* per SBM con $K = 8$. 131
B.5 Elenco diagnosi *hub* per SBM con $K = 9$. 132

Sommario

In questo lavoro analizziamo un dataset amministrativo, fornito dalla Regione Lombardia, riguardante le ospedalizzazioni nelle strutture sanitarie regionali tra il 2013 e il 2015, tramite metodi sviluppati nell'ambito di studio delle cosiddette delle *reti complesse*.
Ci concentreremo sullo studio della rete bipartita costituita da nodi che rappresentano rispettivamente pazienti e diagnosi, connessi da un arco se un certo paziente risulta ricoverato per una certa diagnosi.
Un'analisi approfondita è stata dedicata alla rete unipartita ottenuta tramite proiezione della rete bipartita sui nodi diagnosi. Su questa rete abbiamo applicato metodi di *community detection*, quali il metodo di Louvain e il modello stocastico a blocchi (SBM), con lo scopo di individuare gruppi di patologie strettamente correlate.
L'approccio della *network analysis* consente di analizzare dati amministrativi dal punto di vista delle *relazioni* esistenti tra essi e si rivela un interessante strumento a supporto della loro gestione.

Le analisi sulle reti considerate sono state svolte utilizzando il software R, in particolare il pacchetto *igraph* [15], e una versione del metodo SBM implementata nel codice C++ fornito in [3].

Parole chiave: Ospedalizzazioni, Dati amministrativi, Analisi di reti complesse, Metodi di identificazione di comunità.

Abstract

In this work we analyse an administrative dataset, provided by Regione Lombardia, about hospedalizations in regional healthy facilities from 2013 to 2015. We study these data using methods developed in the context of the so-called *complex networks*.
We focus on the analysis of a bipartite network where the nodes identify patients and diseases respectively and they are connected by a link if a certain patient is hospedalized for a certain disease.
In particular, we give an in-depth analysis of the unipartite network obtained by the projection of the bipartite network on the set of nodes of type "disease". On this network we apply tecniques of *community detection*, such as the Louvain method and the Stochastic Blockmodel (SBM), in order to identify groups of strongly correlated diseases.
The *network analysis* approach allows to analyse administrative data from the point of view of the *relations* between them and it proves to be an interesting tool to support their management.

The analysis on the networks of interest are carried out using sofware R, in particular the *igraph* package [15], and a C++ implementation of the SBM method provided in [3].

Keywords: Hospedalizations, Administrative Data, Network Analysis, Community detection methods.

Introduzione

I dati amministrativi relativi ai ricoveri ospedalieri costituiscono un'importante fonte di informazioni per la gestione del sistema sanitario.
Lo strumento di raccolta di questi dati, i quali comprendono, tra le altre, informazioni anagrafiche sul paziente ricoverato, informazioni di tipo clinico sulle diagnosi e informazioni legate ai costi del ricovero, è la Scheda di Dimissione Ospedaliera (SDO). Questo strumento, un tempo esclusivamente legato a finalità amministrative, si è rivelato un mezzo fondamentale su cui basare analisi ed elaborazioni a supporto dell'attività di programmazione sanitaria, sia dal punto di vista clinico sia da quello economico.
Il nostro lavoro si inserisce in questo contesto: analizzeremo i dati relativi ai ricoveri ospedalieri registrati nelle SDO in Lombardia tra il 2013 e il 2015, concentrandoci in particolare sulle ospedalizzazioni dovute ad alcune categorie di patologie.
L'approccio utilizzato per lo studio di questi dati è quello della *network analysis*. Le reti complesse sono, da circa due decenni, un ambito di studio in forte sviluppo. La modellizzazione matematica tramite strutture a rete risulta particolarmente utile ed efficace nell'analisi dei dati relazionali, diffusi in molti contesti, come ad esempio le reti sociali e i sistemi di trasporto.
L'ipotesi modellistica che consente di approcciare l'analisi del dataset amministrativo con gli strumenti della *network analysis* è la seguente: ogni ricovero registrato può essere interpretato come *relazione* tra il paziente ospedalizzato e la diagnosi causa del ricovero. Sulla base di questa considerazione, il dataset originario verrà inserito in una struttura di rete, in cui le diagnosi e i pazienti presenti costituiranno i *nodi* e le loro relazioni saranno espresse tramite *archi* tra questi nodi.
L'utilizzo di reti complesse nello studio di dati amministrativi, dunque, è interessante in quanto consente di indagare la struttura di un insieme di dati focalizzandosi sulle *relazioni* esistenti tra essi. Obiettivo finale del lavoro sarà quello di descrivere la struttura del sistema sanitario della Regione Lombardia, modellizzato tramite una rete di pazienti e diagnosi, e di metterne in evidenza le proprietà. Interessante sarà poi studiare le relazioni esistenti

tra le diagnosi e tra i pazienti costruendo delle opportune reti *proiezione* a partire dalla rete di partenza.

La tesi è articolata come segue:

- Il **Capitolo 1** costituisce un'introduzione al dataset analizzato. In particolare vengono fornite informazioni relative alla tipologia di dati amministrativi di cui ci siamo occupati e vengono presentate le scelte di ridimensionamento effettuate per ottenere il dataset oggetto delle analisi nei capitoli successivi.

- Nel **Capitolo 2** presentiamo l'argomento dell'analisi delle reti complesse. Vengono descritte dal punto di vista teorico le principali proprietà delle reti e gli strumenti di analisi che utilizzeremo nel nostro lavoro.

- Nel **Capitolo 3** si illustra la costruzione delle reti oggetto di studio. Verrà descritta la costruzione della rete bipartita a partire dai dati dei ricoveri e la costruzione, tramite la proiezione di questa rete, della Rete Diagnosi e della Rete Pazienti. Vengono poi analizzate le caratteristiche principali delle tre reti.

- Il **Capitolo 4** fornisce un quadro teorico dei metodi di community detection. All'introduzione del problema e alla definizione del concetto di comunità, seguono l'esposizione dettagliata dei due metodi di individuazione di comunità utilizzati nel nostro lavoro: il metodo di Louvain e il modello stocastico a blocchi. Vengono inoltre introdotti strumenti per la validazione delle partizioni prodotte dagli algoritmi di community detection e per la classificazione dei nodi di una rete in base al ruolo assunto nelle comunità a cui risultano assegnati.

- Nel **Capitolo 5** mostriamo i risultati ottenuti applicando i metodi di analisi descritti nel Capitolo 4 alle Rete Diagnosi e alla Rete Pazienti. Verrà fornita la valutazione della qualità delle partizioni ottenute e una loro analisi a posteriori, sulla base del confronto con la classificazione nota delle diagnosi in categorie, allo scopo di evidenziare le caratteristiche qualitative delle comunità ottenute e di interpretare i risultati.

- Infine, dettaglieremo le conclusioni che possiamo trarre dal nostro lavoro e segnaleremo problemi riscontrati e punti di interesse per possibili sviluppi futuri.

Capitolo 1

Presentazione del Dataset

In questo capitolo presentiamo il dataset analizzato durante il lavoro, fornito dalla Regione Lombardia. Nella prima sezione si dà una panoramica sul dataset completo e le variabili presenti, nella seconda si espone nel dettaglio il dataset ridotto utilizzato nelle analisi e le motivazioni delle scelte effettuate per ottenere il campione di interesse.

1.1 Il Dateset originale

Il dataset originale preso in considerazione consiste nell'elenco dei ricoveri nelle strutture sanitarie in Lombardia negli anni 2013, 2014 e 2015. Le ospedalizzazioni totali nel corso dei tre anni di riferimento sono 4450677 e riguardano 2652408 pazienti distinti. Ogni record del dataset corrisponde a un singolo ricovero ed è costituito da 142 campi che riportano le informazioni presenti nella Scheda di Dimissione Ospedaliera (SDO) relativa al ricovero. Si tenga conto che nel dataset considerato le informazioni vengono in molti casi riportate in forma di codici associati a una seconda variabile di tipo descrittivo.

La SDO è lo strumento con cui ordinariamente vengono raccolte le caratteristiche essenziali dei ricoveri dei pazienti in tutti gli istituti pubblici e privati del territorio nazionale [6].

Il set di dati rilevati dalla SDO, definito nel 1991 da un Decreto Ministeriale, è ampio e dettagliato ed è costituito da dati anagrafici del paziente, da caratteristiche del ricovero e da informazioni di tipo clinico.

Si riportano nell'elenco alcuni dati significativi presenti nella SDO per ciascuna tipologia.

- **dati anagrafici del paziente:** codice fiscale, data di nascita, età, genere, residenza, professione, stato civile;

- **caratteristiche del ricovero:** struttura di ricovero, regime di ricovero, modalità di dimissione, data di ricovero, date di eventuali interventi o decesso, classe di priorità del ricovero, giorni di degenza, provenienza del paziente, costo dell'ospedalizzazione e degli eventuali interventi;

- **informazioni di tipo clinico:** diagnosi principale, diagnosi concomitanti, procedure diagnostiche o interventi eseguiti, codici DRG e MDC (si veda più avanti per le loro definizioni), elenco delle comorbidità presenti.

Le informazioni presenti nella SDO, raccolte in primis per finalità amministrative, sono interessanti, per la loro completezza e la copertura che offrono, anche dal punto di vista clinico e sono utilizzate per analisi a supporto della programmazione sanitaria e di altre attività, come il monitoraggio dell'erogazione dei Livelli Essenziali di Assistenza, la valutazione del rischio clinico ospedaliero, il calcolo di indicatori di appropriatezza e qualità dell'assistenza, oltre che per valutazioni di impatto economico.

É utile descrivere con più precisione alcune informazioni di tipo clinico, che saranno poi utilizzate nel nostro lavoro.

Diagnosi principali e secondarie In primo luogo, sottolineiamo la possibilità di associare un ricovero a più diagnosi diverse, indicate nella SDO in modo distinto per la *diagnosi principale* e le eventuali *diagnosi secondarie o concomitanti*.
Per *diagnosi principale* si intende la condizione che risulta essere la principale responsabile del bisogno del ricovero ospedaliero al termine del quale il paziente viene dimesso [7]. Nel caso si presentino più condizioni analoghe o non sia evidente un'unica causa di ricovero, il criterio con cui si identifica la diagnosi principale è di natura non più clinica ma economica: la diagnosi principale è quella che contribuisce al maggiore consumo di risorse da parte della struttura ospedaliera.
Nella SDO è possibile che vengano riportate fino a un massimo di cinque altre *diagnosi secondarie* per fornire un quadro clinico completo del ricovero. Sono queste condizioni che coesistono alla diagnosi principale al momento del ricovero, o si sviluppano nel corso della degenza, influenzando il tipo di trattamento o la durata del ricovero stesso, senza essere identificate come causa principale dell'ospedalizzazione.

Diagnosis Related Group, DRG Un altro campo presente nei record del dataset è l'indice DRG, Diagnosis Related Group o, in italiano, ROD, *Raggruppamento omogeneo di diagnosi*. L'impiego di risorse associato a un

ricovero è legato al tipo di diagnosi principale, alle eventuali diagnosi che completano il quadro clinico o alla presenza di complicanze, ma anche alle caratteristiche del paziente, come ad esempio l'età. Il sistema DRG, tenendo conto di questi aspetti, consente di classificare le diagnosi in gruppi omogenei (identificati da un codice) per impiego di risorse, favorendo la quantificazione dei costi associati al ricovero e la remunerazione adeguata delle strutture ospedaliere.
Nel sistema DRG utilizzato in Italia si hanno 538 codici DRG contraddistinti da un numero a tre cifre da 001 a 579. In [4] è riportata la classificazione DRG completa, in cui i codici sono associati alla descrizione del gruppo di diagnosi.
A titolo di esempio, riportiamo in Tabella 1.1 alcune righe dell'elenco citato, dove una stessa tipologia di diagnosi viene classificata in tre DRG diversi a seconda dell'età del paziente ricoverato e della presenza o meno di complicanze (CC).

DRG	**Descrizione**
46	Altre malattie dell'occhio, età superiore a 17 anni con CC
47	Altre malattie dell'occhio, età superiore a 17 anni senza CC
48	Altre malattie dell'occhio, età inferiore a 18 anni

Tabella 1.1: Estratto dall'elenco di classificazione dei DRG.

Major Diagnostic Criteria, MDC I codici DRG vengono ulteriormente aggregati secondo la classificazione in Major Diagnostic Category (MDC) [5]. Questo livello di classificazione è basato su criteri anatomici: l'attribuzione della MDC è stabilita in base all'apparato coinvolto o, in generale, dalla specialità medica che si occupa della data patologia. A ogni diagnosi è associato un codice MDC univoco da 1 a 25. In Tabella 1.2 è riportato l'elenco completo dei codici MDC. Il codice MDC assegnato al ricovero nella SDO coincide con il codice relativo alla diagnosi indicata come principale.

Nel dataset completo sono presenti tutti i codici DRG e MDC indicati negli elenchi e 10159 codici distinti per le diagnosi principali.

1.2 Il Dataset ridotto

Durante il lavoro di tesi è stata presa in considerazione per le analisi solo una parte del dataset a disposizione. In particolare, l'interesse si è focalizza-

MDC	Descrizione
1	Malattie e disturbi del sistema nervoso
2	Malattie e disturbi dell'occhio
3	Malattie e disturbi dell'orecchio della bocca e della gola
4	Malattie e disturbi dell'apparato respiratorio
5	Malattie e disturbi dell'apparato cardiocircolatorio
6	Malattie e disturbi dell'apparato digerente
7	Malattie e disturbi epatobiliari e del pancreas
8	Malattie e disturbi del sistema muscolo-scheletrico e del tessuto connettivo
9	Malattie e disturbi della pelle
10	Malattie e disturbi endocrini
11	Malattie e disturbi del rene e delle vie urinarie
12	Malattie e disturbi dell'apparato riproduttivo maschile
13	Malattie e disturbi dell'apparato riproduttivo femminile
14	Gravidanza parto e puerperio
15	Malattie e disturbi del periodo neonatale
16	Malattie e disturbi del sangue
17	Malattie e disturbi mieloproliferativi e neoplasie scarsamente differenziate
18	Malattie infettive e parassitarie (sistemiche o di sedi non specificate)
19	Malattie e disturbi mentali
20	Abuso di alcol / droghe e disturbi mentali organici indotti
21	Traumatismi
22	Ustioni
23	Fattori che influenzano lo stato di salute ed il ricorso ai servizi sanitari
24	Traumatismi multipli rilevanti
25	Infezioni da H.I.V.

Tabella 1.2: Elenco dei codici MDC.

to sui pazienti ricoverati e le diagnosi per cui questi vengono ricoverati, non considerando gli aspetti legati alle caratteristiche cliniche e amministrative del ricovero. Basandoci sulla connessione tra paziente e diagnosi stabilita mediante il ricovero, infatti, verrà costruita la rete oggetto delle nostre analisi.
Per quanto riguarda i pazienti, si è scelto di utilizzare il campo "ASSISTITO KEY", che rappresenta un codice numerico che identifica in modo univoco e anonimo i pazienti nel dataset.
É possibile, inoltre, che in uno stesso ricovero siano indicate da 1 fino a 6 diagnosi. La scelta fatta è stata quella di considerare solo la diagnosi principale e di trascurare le eventuali diagnosi secondarie. In questo modo, possiamo

interpretare ogni ricovero come una coppia paziente-diagnosi facilmente modellizzabile con una struttura di rete bipartita, evitando una disomogeneità nel numero di diagnosi associate a un paziente nel medesimo ricovero. Inoltre, poichè i codici MDC sono assegnati ai ricoveri sulla base della diagnosi principale, possiamo lavorare con diagnosi di cui conosciamo questo attributo, utile in fase di interpretazione dei risultati. Concentrando l'attenzione solo sulle diagnosi principali, ci allineiamo, da ultimo, alla logica sottesa alla compilazione delle SDO dove è proprio la diagnosi principale ad essere significativa per le valutazioni cliniche ed economiche legate al ricovero. Data questa scelta di modellizzazione, si precisa che il caso in cui, nella rete costruita, un paziente risulti associato a più di una diagnosi è dovuto a ricoveri multipli del paziente per diagnosi principali diverse nel periodo di riferimento.
Un'ulteriore scelta è stata quella di restringere l'analisi ai ricoveri relativi a diagnosi attribuibili esclusivamente ad alcune MDC. In particolare, abbiamo deciso di focalizzarci sulle diagnosi delle seguenti categorie:

- **MDC 4**, *Malattie e disturbi dell'apparato respiratorio*,
- **MDC 5**, *Malattie e disturbi dell'apparato cardiocircolatorio*,
- **MDC 8**, *Malattie e disturbi del sistema muscolo-scheletrico e del tessuto connettivo*,
- **MDC 11**, *Malattie e disturbi del rene e delle vie urinarie*,
- **MDC 14**, *Gravidanza, parto e puerperio.*

La riduzione del dataset a ricoveri relativi ai cinque MDC indicati, dovuta a necessità tecniche, garantisce comunque un'ampia copertura dell'originale, sia per quanto riguarda il numero di ricoveri sia per le caratteristiche dei pazienti. La scelta di queste categorie, inoltre, consente di avere una stima "a priori" delle proprietà strutturali della rete con cui confrontare i risultati. Delle cinque categorie MDC selezionate, infatti, tre (4, 5, 11) contengono diagnosi che, proprio per gli apparati coinvolti (rispettivamente cardiocircolatorio, respiratorio e urinario), è possibile supporre, almeno sulla base delle nostre conoscenze a priori, correlate o copresenti in ricoveri dello stesso paziente, due (8 e 14) riguardano invece diagnosi, rispettivamente relative al sistema muscolo-scheletrico e alla gravidanza, per loro natura più difficilmente legate a diagnosi di altri apparati. Ricoveri per diagnosi attribuibili alla MDC 8, come ad esempio una frattura o la sostituzione di una protesi, si suppone siano poco frequentemente legati a ricoveri per patologie cardiache

o respiratorie; è invece più ragionevole immaginare, facendo considerazioni di natura clinica o basate sui trattamenti necessari per curare determinate patologie, che ricoveri classificati nella MDC 5 siano correlati a ricoveri classificati nella MDC 4. Un esempio è dato dalla diagnosi SCOMPENSO CARDIACO CONGESTIZIO (MDC 5): è noto che pazienti ricoverati per questo tipo di patologia abbiano come complicazione problematiche a livello di apparato respiratorio.

Presentiamo nelle tabelle riportate di seguito le caratteristiche del dataset ridotto utilizzato nel corso del lavoro.
In Tabella 1.3 si ha un quadro delle numerosità dei dati considerati. Si noti che il numero di ricoveri nel dataset ridotto, considerando le 5 categorie MDC descritte sopra, copre il 47% dei ricoveri del dataset completo. É interessante sottolineare, inoltre, che il numero totale di ricoveri è 1,5 volte il numero totale dei pazienti.

Ricoveri	2092516
Pazienti	1351224
Diagnosi	3279
DRG	193
MDC	5
Anni di riferimento	3

Tabella 1.3: Caratteristiche del dataset ridotto.

In Tabella 1.4 è riportato in dettaglio il numero di ricoveri per paziente, sempre nel periodo di riferimento di tre anni.

Ricoveri per paziente	**Pazienti**
1	909863
2	288369
3	83877
4	3638
da 5 a 10	31143
da 11 a 20	1516
da 21 a 30	68
superiore a 30	8

Tabella 1.4: Riassunto delle frequenze di ricovero.

Il numero di ricoveri nei singoli anni 2013, 2014 e 2015 è riportato in Tabella 1.5: si nota una sostanziale omogeneità nel numero di ricoveri durante il periodo di riferimento.

Anno di riferimento	**Numero di ricoveri**	**%**
2013	710394	33,95%
2014	691451	33,04%
2015	690671	33,01%
totale	2092516	

Tabella 1.5: Numero di ricoveri per anno di riferimento.

Data l'importanza che daremo alla classificazione in categorie MDC nel corso del lavoro, riportiamo in Tabella 1.6 il numero di ricoveri, DRG e diagnosi, ripartiti per ciascuna MDC considerata.

MDC	**4**	**5**	**8**	**11**	**14**	**totale**
numero di ricoveri	313103	550835	643304	211907	373367	2092516
%	15,0%	26,3%	30,7%	10,1%	17,8%	
numero di DRG	30	56	59	33	15	193
%	15,5%	29,0%	30,6%	17,1%	7,8%	
numero di diagnosi	392	443	1455	299	690	3279
%	12,0%	13,5%	44,4%	9,1%	21,0%	

Tabella 1.6: Numero di ricoveri, DRG e diagnosi per ciascuna categoria MDC.

Forniamo, infine, in Tabella 1.7 e Tabella 1.8 un quadro sulle caratteristiche principali dei pazienti presenti nel dataset, in particolare la ripartizione in base al genere e alle fasce di età.

	Maschi	**Femmine**	**Totale**
numero di pazienti	562952	788272	1351224
%	41,7%	58,3%	

Tabella 1.7: Numero di pazienti per genere.

Età	**Maschi**	**%**	**Femmine**	**%**	**Complessivi**	**%**
0	10636	1,9%	8035	1,0%	18671	1,4%
1-5	11677	2,1%	9524	1,2%	21201	1,6%
6-10	7765	1,4%	5982	0,8%	13747	1,0%
11-15	11975	2,1%	8937	1,1%	20912	1,5%
16-20	12710	2,3%	15994	2,0%	28704	2,1%
21-30	25348	4,5%	113925	14,5%	139273	10,3%
31-40	34506	6,1%	186348	23,6%	220854	16,3%
41-50	57907	10,3%	58252	7,4%	116159	8,6%
51-60	76480	13,6%	56451	7,2%	132931	9,8%
61-70	109486	19,4%	84959	10,8%	194445	14,4%
71-80	124931	22,2%	115744	14,7%	240675	17,8%
81-90	70289	12,5%	100048	12,7%	170337	12,6%
91-100	9113	1,6%	23558	3,0%	32671	2,4%
>100	129	0,0%	515	0,1%	644	0,0%
totale	562952		788272		1351224	

Tabella 1.8: Numero di pazienti per fascia di età e genere.

Capitolo 2

Introduzione alle reti complesse

In questo capitolo forniamo un quadro del contesto di riferimento del nostro lavoro, presentando le proprietà principali delle reti complesse. Introduciamo l'argomento fornendo alcune informazioni di tipo generale, per poi dettagliare dal punto di vista teorico i più comuni indici descrittivi delle reti. Da ultimo, presentiamo le caratteristiche delle reti bipartite e il metodo di proiezione utilizzato per ottenere da esse delle reti unipartite. Questo capitolo costituisce un'introduzione teorica ai metodi di analisi della rete che costruiremo e analizzeremo nel Capitolo 3. Per ulteriori approfondimenti rimandiamo a testi dedicati all'argomento, ad esempio [24].

2.1 Caratteristiche generali

Si definisce *rete complessa* un sistema costituito da elementi (detti agenti, individui ecc...) connessi tra loro da relazioni a coppie, che danno luogo a strutture con proprietà non banali, cioè non presenti in reti "regolari". Lo studio delle reti complesse, in grande sviluppo negli ultimi due decenni, fornisce strumenti per modellizzare e analizzare le proprietà di tali sistemi.

Nel nostro lavoro, come vedremo, utilizziamo una struttura di rete complessa per mettere in evidenza le connessioni tra diagnosi e pazienti, elementi che risultano in relazione se si verifica un ricovero di un certo paziente per una data diagnosi. Altri contesti di applicazione sono le reti dei sistemi di trasporto, come gli aeroporti [19], o le reti sociali, che rappresentano relazioni tra individui. Un esempio di rete sociale è la Rete di Zachary [31] riportata in Figura 2.1, che viene utilizzata come riferimento nella ricerca in questo ambito e che descrive le relazioni sociali tra i membri di una scuola di karatè.

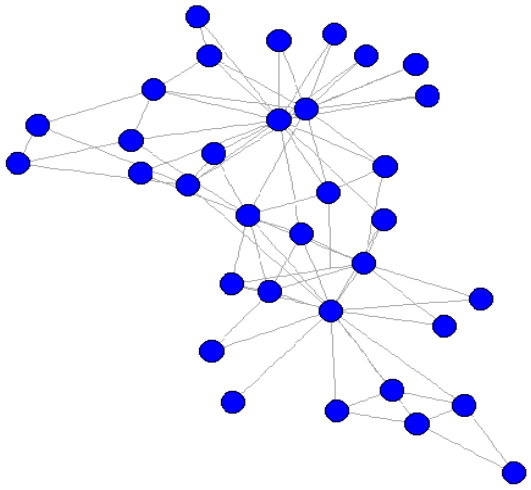

Figura 2.1: Rete della scuola di karatè di Zachary [31], formata da 34 nodi e 78 archi. Ogni nodo rappresenta un membro della scuola, mentre un link indica l'esistenza di una relazione sociale.

Una rete complessa è rappresentata da un grafo, cioè un insieme di N nodi, detti anche vertici, connessi tra loro da L archi, o links. In termini matematici definiamo rete o grafo l'oggetto $G = (V, E)$, dove V è l'insieme dei nodi e E l'insieme degli archi.
La rete rappresentata in Figura 2.1 è *non pesata* e *non diretta*: gli archi rappresentano una connessione bidirezionale (o, più in generale, in cui non è significativa la direzione) e in cui non è specificato il peso. Una rete non diretta e non pesata può essere descritta dalla sua matrice di adiacenza: la matrice A, di dimensione $N \times N$, è simmetrica e tale che l'elemento A_{ij} è uguale a 1 se il nodo i è connesso da un arco al nodo j, uguale a 0 altrimenti. Le reti complesse reali hanno tipicamente matrici di adiacenza sparse, in quanto la proporzione di archi esistenti è molto minore rispetto al numero delle possibili connessioni tra nodi o, in altri termini, la densità ρ degli archi della rete, pari a $\frac{L}{N(N-1)/2}$, è bassa.
Nelle reti pesate, agli archi è assegnato un valore numerico, detto *peso*, che porta un'informazione quantitativa sulla connessione esistente tra i nodi. La matrice di adiacenza per reti pesate è sostituita da una matrice i cui elementi w_{ij} sono dati dal peso dell'arco che connette i e j. Il peso di ogni arco può essere un numero intero o reale, positivo se esiste l'arco tra i e j. Se i nodi i e j non sono connessi, il peso w_{ij} risulta pari a zero.
Le reti dirette, che non utilizziamo nel nostro lavoro, sono invece caratterizzate dalla direzione univoca assegnata agli archi, che possono dunque essere

distinti in archi *uscenti* o *entranti* in un nodo. Nelle reti dirette la matrice di adiacenza e la matrice dei pesi, se si tratta di reti dirette pesate, risultano non simmetriche. Per reti dirette la densità ρ è pari a $\frac{L}{N(N-1)}$.
In Figura 2.2 mostriamo un esempio di rete pesata e un esempio di rete diretta.

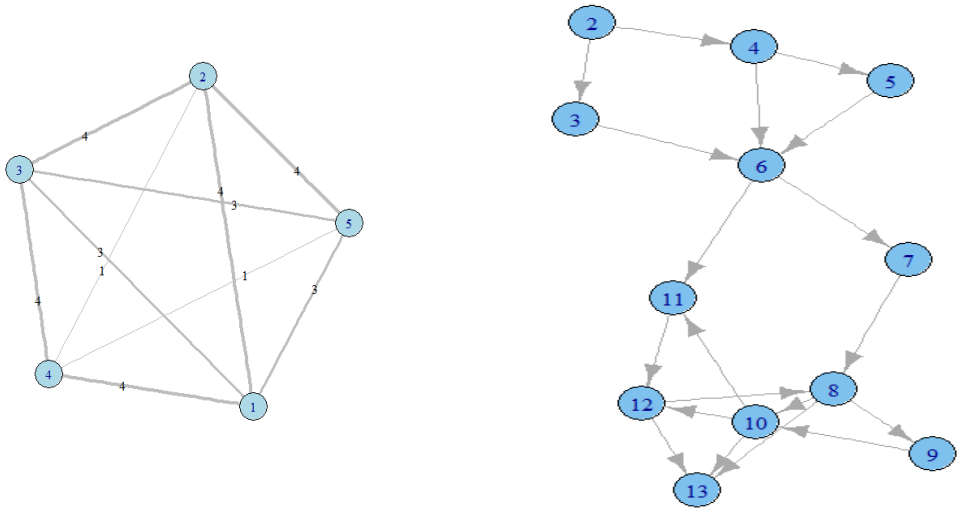

Figura 2.2: Esempio di rete pesata (a sinistra) e di rete diretta (a destra).

É possibile che in una rete esistano connessioni multiple tra coppie di nodi, detti *multi-archi*, e archi che connettono un nodo a se stesso, detti *self-loop*. Una rete che presenta queste caratteristiche è detta *multigrafo* (altrimenti si tratta di una rete *semplice*). In Figura 2.3 mostriamo un esempio di multigrafo. Nel proseguio del lavoro utilizzeremo sia reti semplici sia multigrafi: in base ai metodi utilizzati per le analisi sarà preferibile in alcuni casi lavorare con reti semplici dotate di pesi e in altri tradurre il valore numerico intero dei pesi degli archi in archi multipli tra la stessa coppia di nodi.

Una rete si dice *connessa* se per ogni coppia di nodi $\{i, j\}$ esiste un percorso o *cammino*, cioè un insieme di archi consecutivi che connette i a j. Nel caso di reti non connesse, è possibile suddividere il grafo in *componenti*, cioè sottografi connessi contenenti il massimo numero di nodi possibile. Le componenti di una rete possono essere formate da un solo nodo, in questo caso si parla di nodi *isolati*, o da gruppi di nodi connessi. Si definisce componente *gigante* una componente in cui la cardinalità dell'insieme dei suoi nodi è dell'ordine di N, numero complessivo dei nodi della rete. Nello studio delle reti complesse è a volte necessario lavorare con reti connesse: per questo motivo tipicamente si estrae la componente gigante della rete e si svolgono le analisi

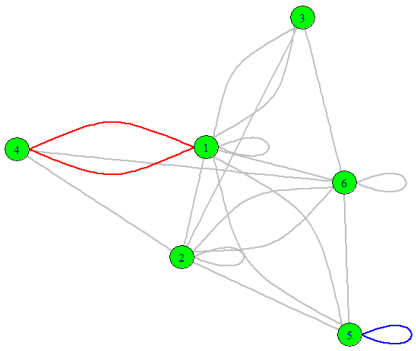

Figura 2.3: Esempio di rete multigrafo: in rosso è evidenziato uno degli archi multipli, in blu uno dei self-loop.

su di essa. In Figura 2.4 è rappresentato un esempio di componente gigante di una rete.

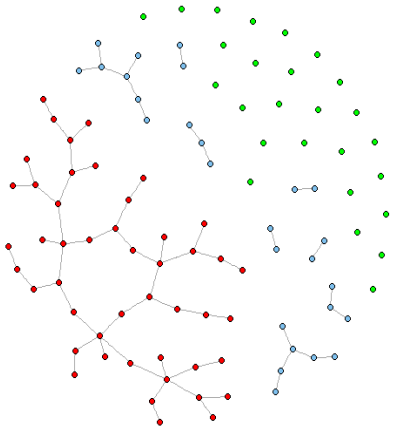

Figura 2.4: Rete in cui si evidenzia la componente gigante (in rosso), le componenti composte da gruppi di nodi (in azzurro) e le componenti formate da nodi isolati (in verde).

2.2 Proprietà delle reti

Introduciamo in questo paragrafo alcuni indici che utilizzeremo per descrivere le proprietà delle reti oggetto delle analisi.
Due valori interessanti da valutare sono la **distanza media** tra i nodi e il **diametro** della rete. Per una rete connessa, consideriamo la coppia di nodi $\{i, j\}$ e la lunghezza d_{ij} del cammino minimo tra i due nodi. La quantità

d_{ij}, che possiamo definire *distanza* tra i e j, è data dal minimo numero di archi che è necessario percorrere per collegare i nodi. Il diametro è invece la lunghezza del massimo cammino minimo. Le definizioni di distanza media e diametro sono, dunque, date dalle formule (2.1) e (2.2):

$$d = \frac{\sum_{i,j:i\neq j} d_{ij}}{N(N-1)} \tag{2.1}$$

$$D = \max_{i,j:\ i\neq j} d_{ij} \tag{2.2}$$

Per reti pesate, il concetto di distanza tra coppie di nodi deve essere modificato in base al significato che si attribuisce al peso degli archi. Nell'ambito dell'analisi della rete delle collaborazioni scientifiche tra autori, in [25] viene proposto di calcolare la distanza tra due nodi adiacenti come inverso del peso dell'arco. Si suppone, infatti, che due autori connessi da un arco di peso elevato (con il peso da interpretare come, ad esempio, numero di articoli scientifici pubblicati insieme dagli autori) abbiano una collaborazione molto stretta e, di conseguenza, possano essere considerati "vicini".
La distanza pesata tra due nodi qualsiasi i e j è definita in [26] come

$$d_{ij}^{W} = min\big(\frac{1}{w_{ih}} + \cdots + \frac{1}{w_{hj}}\big) \tag{2.3}$$

dove la somma è estesa a tutti gli archi del cammino che collega la coppia di nodi $\{i, j\}$.
Un'altra proprietà da evidenziare è il **grado** dei nodi della rete, pari al numero di nodi direttamente connessi al nodo in esame. Nel caso di reti pesate si introduce invece una generalizzazione del concetto di grado: la **strength**. La strength del nodo i è pari alla somma dei pesi degli archi ad esso connessi. In caso di reti dirette è possibile un'ulteriore distizione in grado (o strength) entrante o uscente ai nodi, pari alla somma degli archi (o dei pesi degli archi) entranti o uscenti.
Di seguito riportiamo, nel caso di rete non diretta, le definizioni di grado (2.4) e strength (2.5) per il nodo i, dove a_{ij} e w_{ij} sono gli elementi corrispondenti all'arco $\{i, j\}$, rispettivamente della matrice di adiacenza e della matrice dei pesi:

$$k_i = \sum_{j:j\neq i} a_{ij} \tag{2.4}$$

$$s_i = \sum_{j:j\neq i} w_{ij} \tag{2.5}$$

Una volta definito il grado dei nodi, è utile valutare la **distribuzione di grado** P dell'intera rete, cioè la frazione $P(k)$ di nodi di grado k per ogni valore fissato di k. P è assimilabile a una distribuzione di probabilità discreta di cui è possibile definire la corrispondente distribuzione di grado cumulativa $\bar{P}$ e valutare i momenti. I momenti di ordine r, con $r \geq 1$, di P sono così definiti:

$$\langle k^r \rangle \ = \sum_k k^r \ P(k)$$

Il momento di ordine 1 di P è il **grado medio**, utile da valutare per avere un'informazione complessiva sulle caratteristiche della rete. Il grado medio è così definito:

$$\langle k \rangle \ = \frac{1}{N} \ \sum_{i \in V} k_i = \frac{2L}{N}$$

Sottolineiamo che le reti che più frequentemente si analizzano presentano una distribuzione di grado fortemente eterogenea, con nodi di grado basso e nodi, al contrario, molto connessi. Le reti con questa proprietà si definiscono *scale-free*, in quanto non hanno una dimensione di scala caratteristica. L'informazione legata al grado medio risulta quindi poco significativa. Se invece la rete è *omogenea*, il grado medio è rappresentativo del grado di ciascun nodo. Per quanto riguarda le reti pesate, è possibile calcolare in modo analogo la distribuzione di strength e la corrispondente distribuzione cumulata, così come valutare il suo valore medio.
É utile, infine, considerare la correlazione che può sussistere tra i nodi della rete. In particolare, una rete si dice *correlata* se la probabilità $Q(h|k)$ che un nodo di grado k abbia un vicino di grado h dipende effettivamente dal grado k. Per valutare il tipo di correlazione esistente si calcola, per ogni valore k di grado dei nodi, la funzione **grado medio dei nodi vicini**, così definita:

$$k_{nn}(k) \ = \sum_h h \ Q(h|k) \tag{2.6}$$

Una rete che presenta un andamento monotono crescente di $k_{nn}(k)$ è definita *assortativa*: nodi di grado alto tendono a connettersi con nodi di grado alto. Se, invece, si ha un andamento decrescente, la rete è *disassortativa*: nodi di grado basso tendono a connettersi a nodi di grado alto. In Figura 2.5 sono mostrate una rete assortativa e una rete disassortativa. Come vedremo nel Capitolo 3, la rete costituita dalle diagnosi del nostro dataset è una

rete disassortativa. Un esempio di rete assortativa è invece dato dal social network Facebook, dove utenti con molti contatti tendono a connettersi con altri utenti popolari.

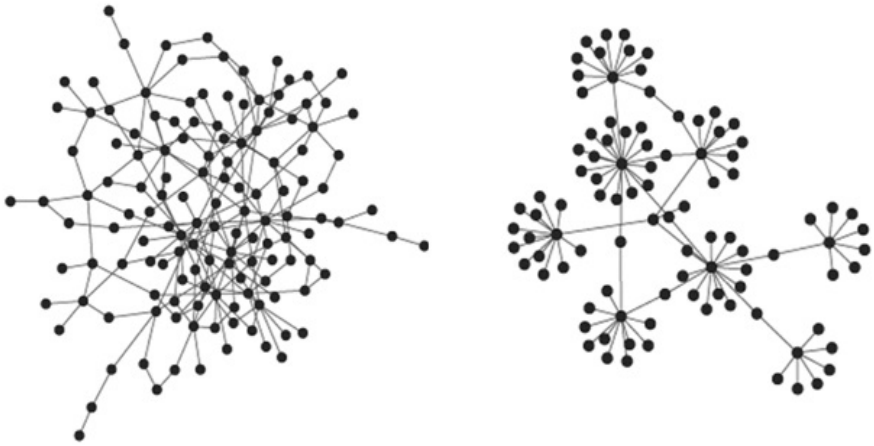

Figura 2.5: Esempio di rete assortativa (a sinistra) e disassortativa (a destra).

Nel caso di reti pesate è possibile generalizzare l'indice k_{nn} valutando per ogni nodo i il **grado medio pesato dei nodi vicini**, così definito:

$$k^{w}_{nn,i} = \frac{1}{s_i} \sum_{j=1}^{N} w_{ij} \; k_j$$

2.3 Misure di centralità

Nel contesto dell'analisi delle rete complesse, uno dei problemi di maggiore interesse è l'individuazione dei nodi più importanti all'interno della rete. In letteratura sono presenti svariati indici utili a quantificare l'importanza dei nodi e che valutano la loro *centralità* considerando aspetti diversi. Si noti che le misure di centralità sono quantità riferite ai singoli nodi. É possibile quindi, una volta calcolato l'indice di interesse, ottenere una sorta di classifica dei nodi che ben mette in evidenza la loro importanza in termini assoluti o in riferimento agli altri.
In questa sezione descriveremo dal punto di vista teorico le misure di centralità utilizzate nel nostro lavoro. Precisiamo che presenteremo solamente quantità riferite a reti non dirette, senza soffermarci sulle loro possibili estensioni per reti dirette. Il pacchetto *igraph* di R [15] fornisce funzioni specifiche dedicate al calcolo di questi valori di utilizzo molto diffuso in letteratura e di

cui riportiamo la documentazione in Appendice C.

Il modo più immediato di valutare la centralità di un nodo è il calcolo del suo **grado**. Si può infatti considerare un nodo come importante all'interno della rete se risulta connesso a un numero elevato di altri nodi. In una rete sociale, ad esempio, un individuo centrale è legato da rapporti con molti altri individui, cioè è un nodo con molte connessioni, o ancora, per quanto riguarda la rete dei ricoveri che analizzeremo, una diagnosi può essere valutata tanto più importante quanti più pazienti sono ricoverati per essa, e quindi connessi al nodo che la rappresenta.
Nel caso di reti pesate, è possibile valutare l'importanza di un nodo anche calcolando la sua **strength**. In questo contesto, infatti, può essere utile non solo tenere conto del numero di connessioni di un nodo, ma anche del loro peso. Tornando all'esempio della rete sociale, l'importanza di un individuo viene valutata considerando l'entità delle connessioni con gli altri individui. Si può notare che le due misure di centralità, se applicate alla stessa rete pesata, forniscono ranking dei nodi tipicamente simili: nodi con grado alto sono spesso anche nodi con alto valore di strength. É utile, tuttavia, valutare entrambe le misure anche in questo caso in quanto, in particolare se i pesi della rete sono distribuiti su valori molto diversi da 1, possono essere messi in evidenza nodi "importanti" per aspetti differenti. A titolo di esempio mostriamo in Figura 2.6 un esempio di nodo di alto grado, connesso a 6 altri nodi, e di nodo di alta strength, connesso solo a 3 nodi, ma con strength alta grazie al peso elevato degli archi connessi ad esso.

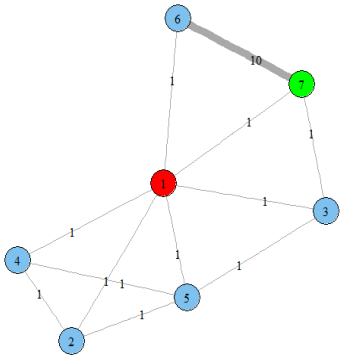

Figura 2.6: Esempio di nodo di grado massimo (in rosso), pari a 6 e di nodo di strength massima (in verde), pari a 12.

Tra le altre misure di centralità presenti in letteratura, introduciamo di seguito le centralità betweenness e closeness.

La centralità **betweenness** assegna a ciascun nodo i un valore pari alla frazione di cammini minimi esistenti tra ogni coppia di nodi diversi da i che passa attraverso il nodo i. La definizione è dunque la seguente:

$$b_i = \sum_{\substack{j,k: \\ j,k \neq i}} \frac{n_{j,k}(i)}{n_{j,k}}$$

dove $n_{jk}(i)$ è il numero di cammini minimi tra j e k passanti per i e n_{jk} è il numero di cammini minimi tra j e k. Questa definizione di betweenness è data per reti non pesate e non dirette.
Un nodo con alto valore di betweenness è centrale in quanto nodo che svolge un ruolo di "ponte" e di collegamento tra i nodi della rete. In Figura 2.7 mostriamo un esempio di rete in cui è evidenziato il nodo con valore di betweenness massimo.

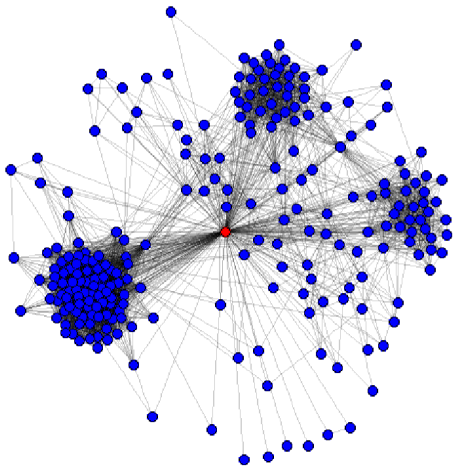

Figura 2.7: Esempio di rete con un nodo con alto valore di betweenness.

Nodi con alto valore di betwenness svolgono un ruolo importante, ad esempio, nelle reti di telecomunicazione e di trasporto, in quanto la presenza di questi nodi centrali garantisce la connessione della rete stessa. In una rete random il grado di un nodo e la sua betweenness risultano fortemente correlati: i nodi con il più alto numero di connessioni sono anche i nodi più centrali dal punto di vista della betweenness. Nelle reti reali è possibile, invece, evidenziare nodi con alto valore di betweenness ma di grado non elevato. In Figura 2.8 è rappresentato un esempio di questo tipo: pur essendo

di grado basso, il nodo al centro svolge un ruolo importante nella rete, in quanto 'ponte' tra due gruppi di nodi molto connessi tra loro.

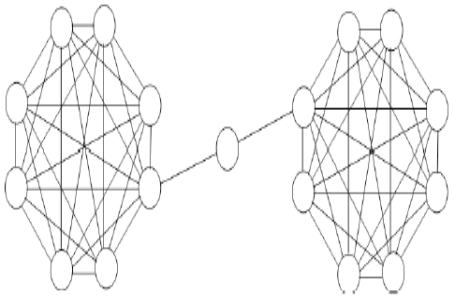

Figura 2.8: Esempio di rete con un nodo (al centro) con alto valore di betweenness ma basso grado.

La relazione tra grado e betweenness dei nodi è stata indagata in [19], dove si analizza la rete del trasporto aereo sottolineando l'importanza per la struttura del sistema aeroportuale internazionale di alcuni nodi "anomali" perchè di basso grado ma di alto valore di betweenness.

La definizione di betweenness proposta è adatta a reti non pesate. É possibile introdurre una versione "pesata" di betweenness utilizzando, per valutare le distanze tra le coppie di nodi necessarie per calcolare il numero di cammini minimi, la definizione di distanza "pesata", descritta in (2.3). Questo tipo di scelta influisce chiaramente sui cammini minimi individuati e ha quindi conseguenze anche sul valore di betweenness calcolato. In [13] sono discusse varianti della misura di betweenness adatte a varie situazioni. Per il calcolo della betweenness dei nodi, sia nella versione base sia nell'estensione per reti pesate, è disponibile una funzione del pacchetto *igraph* [15] di R. Si tenga conto che nella versione "pesata" della funzione `betweenness` implementata in *igraph* i cammini minimi vengono individuati sommando i pesi degli archi percorsi (che risultano quindi interpretati come "costi" e non come "strength" della connessione). Per ottenere i valori di betweenness secondo la definizione indicata è dunque necessario invertire i pesi della rete in una fase precedente al calcolo. Per ulteriori dettagli si rimanda alla descrizione della funzione in Appendice C.

La centralità **closeness**, o di vicinanza, quantifica, invece, l'importanza di un nodo considerando la sua distanza media dagli altri nodi della rete. Possiamo dire, quindi, che è un nodo è tanto più centrale quanto più risulta vicino, in media, agli altri. Se si considera una rete reale, un nodo vicino agli altri elementi svolge un ruolo importante perchè può avere accesso alle infor-

mazioni presenti negli altri nodi o avere un'influenza su di essi. Si definisce closeness il valore inverso della distanza media:

$$cl_i = \frac{N-1}{\sum_{j \in V} d_{ij}}.$$

Nella formula, d_{ij} è la distanza tra i nodi i e j e N il numero di nodi della rete. La definizione è, anche in questo caso, riferita a reti non pesate e non dirette. In caso di reti pesate, è possibile introdurre un'estensione della centralità di closeness, definita come segue:

$$cl_i^w = \frac{N-1}{\sum_{j \in V} d_{ij}^w}.$$

dove d_{ij}^w è la distanza "pesata" tra i nodi i e j, definita in (2.3). Si precisa inoltre che la funzione `closeness` nel pacchetto *igraph* [15] di R consente di calcolare il valore di closeness dei nodi secondo queste definizioni, in entrambe le versioni descritte, a meno di un fattore pari a $N-1$.

2.4 Reti Bipartite

La rete fondamentale che analizzeremo nel nostro lavoro ha la caratteristica di essere **bipartita**. Le reti bipartite sono costituite da due insieme di nodi disgiunti $I1$ e $I2$ e da archi che connettono solo nodi appartenenti a insiemi diversi. Esempi di reti bipartite sono la rete sociale cosidetta delle "Donne del Sud" [16], i cui nodi rappresentano rispettivamente donne ed eventi connessi da archi se vi è stata la partecipazione di una donna ad un certo evento. La rete, rappresentata in Figura 2.9, costituisce un riferimento importante in letteratura per lo studio di metodi per l'analisi di reti bipartite. Altri esempi di reti bipartite in letteratura sono la rete film-attori [22], la rete aziende-direttori [29], la rete piante-insetti impollinatori [11]. Come si può notare dagli esempi, le reti bipartite mettono in evidenza le relazioni esistenti tra categorie di elementi distinte. Nella rete che costruiremo nel nostro lavoro i nodi sono dati rispettivamente dall'insieme delle diagnosi e dall'insieme dei pazienti e gli archi rappresentano il ricovero di un certo paziente per una data diagnosi. Dal punto di vista matematico, una rete bipartita è descritta dalla sua matrice di incidenza B. La matrice ha dimensioni $P \times Q$, dove P e Q sono le cardinalità rispettivamente degli insiemi $I1$ e $I2$, ed è tale che l'elemento B_{ij} è uguale a 1 se esiste la connessione tra il nodo i e il nodo j, altrimenti B_{ij} è pari a 0.

In una rete di questo tipo può essere utile valutare la distribuzione di grado (o di strength) e rispettivi valori medi per i due insiemi di nodi se-

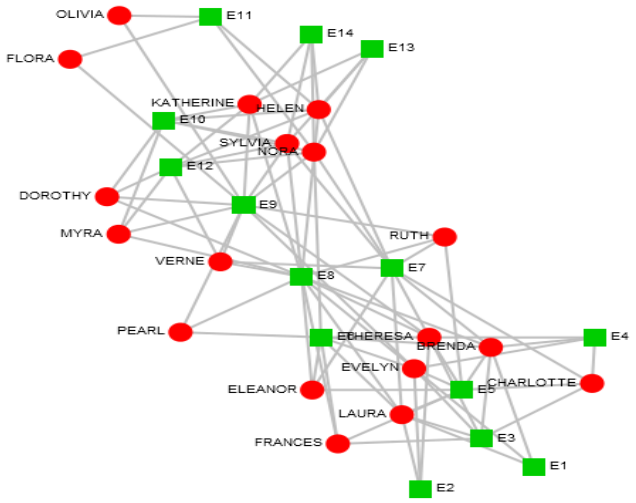

Figura 2.9: Rete bipartita delle Donne del Sud, in rosso i 18 nodi "donna" e in verde i 14 nodi "evento".

paratamente. La densità della rete è definita come $\frac{L}{|I1||I2|}$, con $|I1|$ e $|I2|$ cardinalità degli insiemi dei nodi.
Ogni rete bipartita può essere interpretata, se si ignora la distinzione esistente tra i due gruppi di nodi, come una rete unipartita e su di essa si possono applicare i consueti metodi di analisi. L'approccio tipico alle reti bipartite è, tuttavia, quello di costruire una rete unipartita *proiettando* la rete originale su uno dei due insiemi di nodi. Formalmente, data una rete G bipartita formata dagli insiemi disgiunti di nodi $I1$ e $I2$ e dall'insieme degli archi E, con $E \subseteq I1 \times I2$, si definisce **proiezione** one-mode di G sull'insieme $I1$ la rete P costituita dai nodi dell'insieme $I1$ e dall'insieme degli archi E' tale che $E' = \{(i1, i1') : \exists\ i2 \in I2 \text{ e } (i1, i2), (i1', i2) \in E\}$. I nodi nella rete proiettata appartenenti all'insieme $I1$ risultano quindi connessi se e solo se nella rete bipartita hanno almeno un vicino, appartenente all'insieme $I2$, in comune. Si noti che per ogni rete bipartita è possibile ottenere due proiezioni distinte a seconda dell'insieme di nodi di interesse: in Figura 2.10 mostriamo un esempio delle due proiezioni di una rete.

É possibile ottenere rete proiettate pesate, che risultano più informative e più rappresentative dell'insieme dei link presenti nella rete bipartita. La proiezione pesata si ottiene associando agli archi che connettono due nodi nella rete proiettata un peso w_{ij} pari al numero di vicini in comune tra i e j. In questo modo si tiene conto del fatto che due nodi dello stesso insieme hanno una connessione significativa tra di loro se condividono un numero significativo di vicini nell'altro insieme di nodi. Per costruire la proiezione

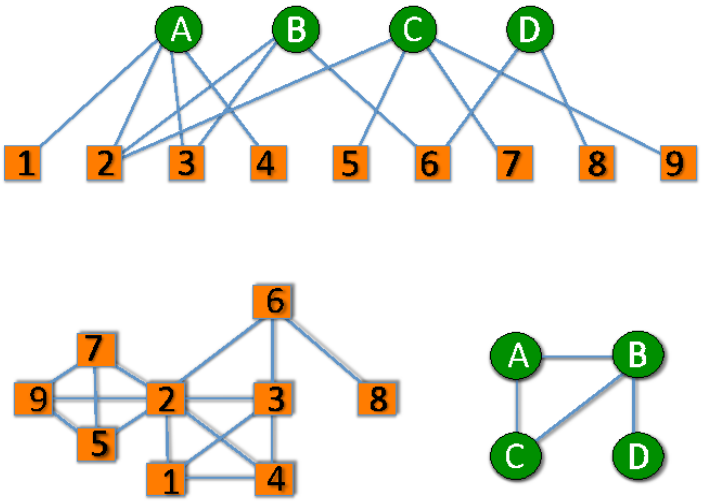

Figura 2.10: Esempio di rete bipartita e delle proiezioni sui due insiemi di nodi.

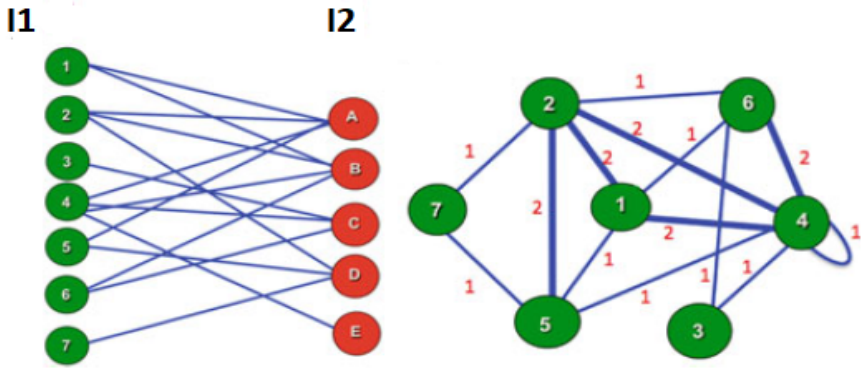

Figura 2.11: Esempio di rete bipartita e di proiezione pesata sull'insieme dei nodi $I1$, con i pesi degli archi indicati.

pesata si può moltiplicare la matrice di incidenza B della rete bipartita per la sua trasposta B^T: i termini della matrice dei pesi della rete proiettata risultano pari ai pesi degli archi e gli elementi diagonali vengono posti a 0. La proiezione sul secondo insieme di nodi si ottiene invertendo l'ordine del prodotto delle matrici B e B^T. In Figura 2.11 si ha un esempio di proiezione pesata su uno dei due insiemi di nodi.
In [10] viene proposto un algoritmo per ottenere proiezioni pesate e il pacchetto *igraph* [15] di R fornisce una funzione per questo scopo, di cui riportiamo

la documentazione in Appendice C.
Nel nostro lavoro costruiremo e analizzaremo le proiezioni della rete bipartita dei ricoveri sui nodi diagnosi e sui nodi pazienti. La necessità di utilizzare reti proiettate deriva dal fatto che molti metodi di analisi, in particolare i metodi di community detection, sono maggiormente sviluppati per reti unipartite.

Capitolo 3

Definizione delle Reti analizzate

L'obiettivo del nostro lavoro è quello di studiare il dataset dei ricoveri introdotto nel Capitolo 1 con metodi di analisi delle reti complesse. Il primo passo è, dunque, costruire la rete oggetto di studio inserendo i dati a disposizione in una opportuna struttura di rete bipartita, che verrà descritta nella sezione 3.1. Di seguito decideremo di analizzare le proiezioni della rete bipartita sui due gruppi di nodi della rete di partenza, i nodi diagnosi e i nodi pazienti. Nelle sezioni 3.2 e 3.3 verranno illustrate le caratteristiche delle due reti unipartite, oggetto delle analisi presentate nei capitoli successivi. Per la costruzione e le analisi della rete bipartita a partire dal dataset dei ricoveri e delle sue proiezioni abbiamo utilizzato il software R e il pacchetto *igraph* [15], che fornisce strumenti adeguati per l'importazione dei dati, l'impostazione della struttura di network, il calcolo delle quantità significative e la loro visualizzazione. In Appendice C riportiamo la documentazione relativa alle funzioni utilizzate.

3.1 Costruzione della Rete Ricoveri

La costruzione della rete oggetto dell'analisi si basa sulla scelta di rappresentare ogni ricovero nel dataset come una coppia paziente-diagnosi, dove la diagnosi considerata è quella indicata nel campo diagnosi principale della SDO riferita all'ospedalizzazione considerata. Le coppie paziente-diagnosi sono espresse nella struttura della rete come archi che collegano due nodi. La Rete Ricoveri risulta essere bipartita: i nodi sono distinti in due tipologie, nodi pazienti e nodi diagnosi, e sono presenti collegamenti solo tra nodi di tipo diverso. La rete bipartita è stata costruita con l'ausilio del pacchetto *igraph* [15] di R, che fornisce una funzione che traduce una lista di archi (nel nostro caso la lista di ricoveri costituita dalle righe del dataset) in un oggetto

grafo e che consente di assegnare ai nodi della rete il loro tipo, rendendo il grafo bipartito.
Si precisa che i codici identificativi con cui sono registrati pazienti e diagnosi consentono di individuarli in modo univoco e di assegnare l'attributo *nome* ai nodi della Rete Ricoveri. Esigenze di praticità, di gestione e di analisi hanno reso oppurtuno effettuare una rietichettatura dei nodi della rete: nel nostro caso i nodi diagnosi sono individuati da numeri interi progressivi da 1 a 3279 e i nodi pazienti con numeri da 3280 a 1354503.
La Rete Ricoveri risulta formata da 2092516 archi, pari al numero di ricoveri nel dataset, e da 1354503 nodi, suddivisi in 3279 nodi di tipo diagnosi e 1351224 nodi paziente. Il peso degli archi è unitario e il caso in cui un paziente risulti ricoverato più volte per la stessa diagnosi viene rappresentato da archi multipli tra la stessa coppia di nodi: la struttura della rete bipartita è quella di un *multigrafo*. La rete è non diretta, in quanto le connessioni tra paziente e diagnosi non necessitano di essere espresse con archi direzionati. Le caratteristiche proprie delle diagnosi e dei pazienti, quali la categoria MDC, il genere e la fascia di età di appartenenza possono essere inseriti nella struttura della rete come attributi dei nodi.
La Rete Ricoveri, inoltre, non è totalmente connessa: è possibile individuare una massima componente connessa costituita da 1353659 nodi e un totale di altre 265 componenti. Le componenti secondarie della rete, composte da un numero esiguo (da 2 a un massimo di 23) di nodi di entrambe le tipologie, sono dovute a ricoveri per diagnosi poco frequenti di pazienti ospedalizzati solo per esse.
Rappresentiamo in Figura 3.1 una rete bipartita che esemplifica, in versione ridotta, la struttura delle Rete Ricoveri che abbiamo descritto.

Proponiamo due diversi layout della stessa rete per metterne in evidenza le caratteristiche peculiari. Un aspetto importante da tenere in considerazione, espresso anche nell'immagine della rete esempio, è il forte sbilanciamento dei due gruppi di nodi: le diagnosi (nodi azzurri) sono in numero molto minore rispetto ai pazienti (nodi rossi). Allo stesso modo, anche il grado dei nodi presenta forte eterogeneità: alcuni nodi diagnosi risultano connessi a un grande numero di nodi pazienti, mentre i nodi pazienti hanno in generale un grado più basso. La distribuzione di grado dei due tipi di nodi della rete è rappresentata in Figura 3.2: sottolineiamo che il massimo grado dei nodi pazienti è 16, mentre per i nodi diagnosi è 100749. Si precisa che per grado di un nodo intendiamo, in questa rete, il numero totali di archi che collegano il dato nodo a nodi distinti, trascurando il conteggio degli eventuali archi multipli. Significativo è sottolineare che la percentuale di nodi pazienti di grado unitario è del $71,4\%$ (905548 su 1351224): si ha che la maggior parte dei pazienti viene ricoverata un'unica volta e per un'unica diagnosi

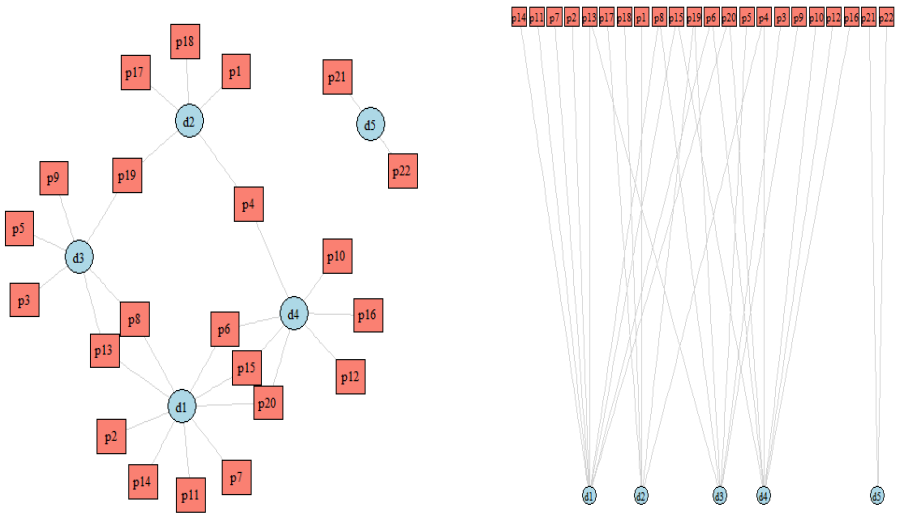

Figura 3.1: Rete bipartita dei ricoveri in versione ridotta, in rosso i nodi paziente, in azzurro i nodi diagnosi.

nell'intervallo di tempo dei tre anni considerati. Le diagnosi per cui si registra un unico ricovero, che sono connesse perciò ad un unico nodo paziente, sono invece 292 ($8,9\%$ del totale dei nodi diagnosi). In Tabella 3.1 riportiamo in dettaglio la frequenza dei valori di grado per i nodi paziente e diagnosi.

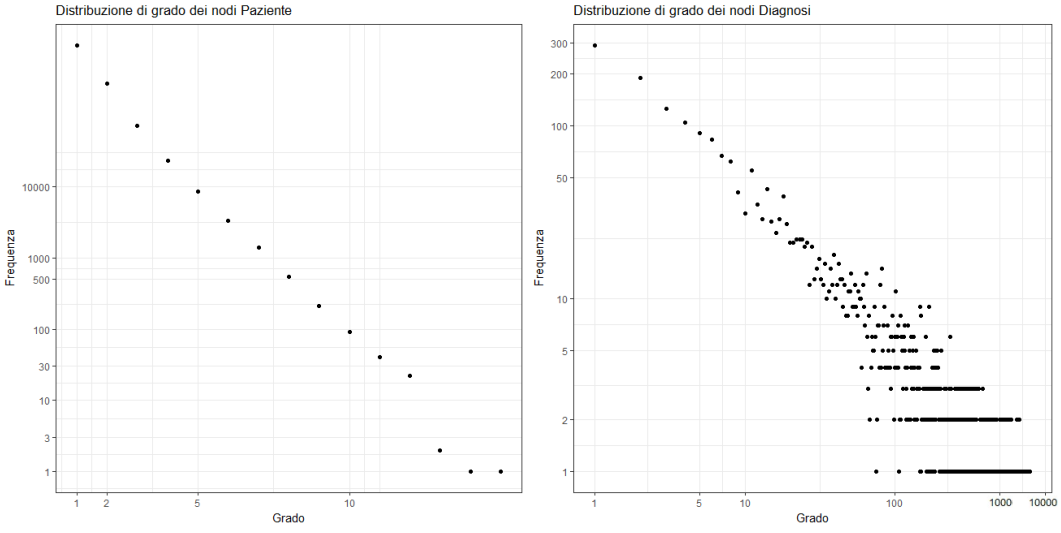

Figura 3.2: Distribuzione di grado dei nodi Paziente (a sinistra) e Diagnosi (a destra, in scala logaritmica su entrambi gli assi) nella Rete Ricoveri.

Grado	Frequenza
1	965548
2	277161
3	70829
4	23468
5	8583
6	3328
7	1398
8	541
9	211
10	91
11	40
12	22
13	2
15	1
16	1

Grado	Frequenza
1	292
2	190
da 3 a 10	605
da 11 a 100	1102
da 101 a 1000	813
da 1001 a 10000	239
più di 10000	38

Tabella 3.1: Frequenza del valore di grado per i nodi Paziente (a sinistra) e i nodi Diagnosi (a destra) nella Rete Ricoveri.

In Tabella 3.2 riassumiamo, infine, le informazioni sulla Rete Ricoveri.

Rete Ricoveri	
Numero di nodi	3279 diagnosi + 1351224 pazienti
Numero di archi	2092516
Densità	$4,72 \times 10^{-4}$
Grado medio nodi diagnosi	581,5
Grado medio nodi paziente	1,41

Tabella 3.2: Informazioni generali sulla Rete Ricoveri.

3.2 Proiezione della Rete Ricoveri: Rete Diagnosi

Per poter analizzare la rete con metodi adatti a reti unipartite, decidiamo di proiettare la Rete Ricoveri sui nodi di tipo diagnosi. Ci siamo riferiti al metodo di proiezione pesata descritto nel Capitolo 2 e abbiamo utilizzato la funzione `bipartite_projection` implementata nel pacchetto *igraph* [15] del software R (si veda, per dettagli, la documentazione in Appendice C). Di seguito ci riferiremo a questa rete come **Rete Diagnosi**.
La rete unipartita ottenuta dalla proiezione è costituita da 3279 nodi e 103593 archi. I nodi di questa rete sono le diagnosi principali presenti nella Rete Ricoveri e gli archi rappresentano la presenza, nella rete bipartita, di pazienti ricoverati per entrambe le diagnosi che vengono connesse. Si sottolinea che, per la scelta di considerare solo le diagnosi principali, ogni ricovero è rappresentato nella rete bipartita da un solo arco tra un nodo paziente e un nodo diagnosi. Questa struttura della rete bipartita fa sì che nella proiezione sui nodi diagnosi un arco tra due nodi esista solo se uno stesso paziente *in due ricoveri distinti* viene ricoverato per le due diagnosi connesse dall'arco. Precisiamo da subito che il tipo di legame esistente tra le diagnosi espresso tramite gli archi della rete proiettata non è interpretabile direttamente come causa-effetto: ciò che si registra è solamente la presenza di ricoveri per diagnosi diverse in momenti successivi. Tuttavia, poichè il dataset copre un intervallo di tempo di soli tre anni, è indubbio che un ricovero ripetuto di uno stesso paziente per diagnosi diverse, rappresentato da un arco nella rete proiettata, possa essere dovuto a motivazioni di tipo clinico. In questo lavoro non ci soffermiamo a indagare la natura di queste motivazioni e delle conseguenti correlazioni tra le diagnosi, ma ci limitiamo ad analizzare la struttura della rete.
La rete proiettata che abbiamo ottenuto risulta *pesata*: nel metodo di proiezione si tiene conto della molteplicità dei nodi pazienti connessi alla stessa coppia di diagnosi. Il peso dell'arco nella Rete Diagnosi risulta, quindi, pari al numero di pazienti ricoverati, in ricoveri distinti, per entrambe le diagnosi.
La Rete Diagnosi è, inoltre, una rete *semplice*: non sono presenti nè archi multipli (la molteplicità delle connessioni viene espressa tramite i pesi degli archi) nè self-loop, cioè archi che connettono un nodo diagnosi a se stesso. Ai fini della proiezione, infatti, non vengono considerati ricoveri multipli di uno stesso paziente per la stessa diagnosi che, secondo la logica della rete proiettata appena esposta, darebbero luogo a nodi diagnosi connessi con un arco a se stessi. Questo tipo di connessione, oltre a rendere tecnicamente più difficile la gestione della rete, non risulta interessante per l'analisi che

svolgeremo, volta a individuare le relazioni tra diagnosi diverse e gruppi di esse. Precisiamo che, per la gestione della rete ai fini dell'individuazione delle comunità, aspetto che dettaglieremo nel Capitolo 5, è stata utilizzata la struttura della Rete Diagnosi più adatta al metodo scelto: per le analisi con il metodo di Louvain si è adottata una struttura di rete semplice pesata e, invece, una struttura multigrafo con archi multipli di peso unitario per il metodo basato sul modello stocastico a blocchi.
Una volta ottenuta la proiezione della rete bipartita, individuiamo le componenti della Rete Diagnosi tramite la funzione `components`, disponibile nel pacchetto *igraph* [15] di R e documentata in Appendice C. I 3279 nodi diagnosi risultano appartenenti a 266 componenti distinte: una composta da 3013 nodi, una formata da una coppia di nodi e le restanti 264 costituite da nodi singoli. Questi ultimi sono i nodi isolati (cioè di grado zero) nella rete proiettata che corrispondono, quindi, a diagnosi per cui non esiste un paziente ricoverato per queste e, in un altro ricovero, per una delle altre diagnosi. La componente formata dalla coppia di nodi è, invece, dovuta a due diagnosi, precisamente due forme particolari e distinte di tubercolosi ossea [1], per cui esiste un unico paziente ricoverato per entrambe e nessun altro paziente ricoverato per una di esse e per un'altra diagnosi.
Nel proseguio della trattazione, ci concentreremo sulla massima componente connessa individuata, non svolgendo analisi sui 266 nodi non appartenenti ad essa. La scelta di escludere questi nodi è dettata dall'esigenza di operare con reti connesse nello svolgimento di analisi di comunità. Sottolineiamo, inoltre, che l'esclusione di queste diagnosi, che costituiscono $8,11\%$ del totale delle diagnosi, comporta l'esclusione dall'analisi di soli 586 ricoveri, pari allo $0,02\%$ del totale dei ricoveri: questo gruppo infatti, è formato da diagnosi per lo più molto rare e che, in media, hanno 2,2 occorrenze nel dataset.

Forniamo in Tabella 3.3 un quadro delle diagnosi considerate e escluse dalla massima componente connessa per ogni MDC.

La nuova rete di interesse risulta dunque costituita da 3013 nodi e 103592 archi e in questo capitolo ne presentiamo le caratteristiche principali.

Numero di ricoveri La prima caratteristica che è interessante considerare per descrivere la Rete Diagnosi è il numero di occorrenze assolute delle diagnosi nel dataset dei ricoveri. Esistono, infatti, diagnosi molto diffuse per cui si registrano molti ricoveri e altre più rare. Il numero di ricoveri per

[1] Si tratta delle diagnosi TUBERCOLOSI DI ALTRE OSSA, NON SPECIFICATA e TUBERCOLOSI DI ALTRE OSSA,ESAMI BATTERIOLOGICI O ISTOLOGICI NEGATIVI, MA TUBERCOLOSI CONFERMATA IN LABORATORIO CON ALTRI METODI

MDC	Diagnosi nella MCC	%	Escluse	%	totali
4	361	91,9%	32	8,1%	393
5	429	96,8%	14	3,2%	443
8	1309	89,9%	146	10,1%	1455
11	288	96,6%	10	3,4%	298
14	626	90,7%	64	9,3%	690
totale	3013	91,9%	266	8,1%	3279

Tabella 3.3: Quadro delle diagnosi nella massima componente connessa (MCC) ed escluse per ogni MDC.

ciascuna diagnosi costituisce un primo indice che ci consente di individuare, in una prima analisi, quali siano più impattanti per il sistema sanitario e da considerare con interesse nel corso dell'analisi. Ricordiamo che le diagnosi considerate sono esclusivamente quelle indicate come principali nella SDO dei ricoveri. Stiamo quindi escludendo le diagnosi secondarie, benchè, per la loro natura di complicanze o patologie insorgenti in concomitanza ad altre, possano avere un numero di occorrenze elevato e possano rivelare, in un'analisi incentrata su di esse, caratteristiche interessanti.
In Tabella 3.4 riportiamo un quadro riassuntivo del numero di ricoveri per le diagnosi suddivise in MDC. Considerando i valori complessivi, si evidenzia una forte eterogeneità nel numero di occorrenze della diagnosi, con valore mediano di 44 ricoveri e valore massimo di 104500. Le diagnosi per cui si ha un unico ricovero risultano in tutto 152 $(5,04\%)$ e sono presenti in tutte le categorie: 11 per MDC 4 $(3,05\%)$, 4 per MDC 5 $(0,93\%)$, 59 per MDC 8 $(4,51\%)$, 10 per MDC 11 $(3,47\%)$ e 68 per MDC 14 $(10,86\%)$.

MDC	media	dev.std.	mediana	massimo
4	865	3387	44	29910
5	1284	4368	125	45760
8	491	2633	42	43260
11	733	2739	71	22620
14	596	4752	16	104500
totale	694	3545	44	104500

Tabella 3.4: Quadro del numero di ricoveri per diagnosi, classificate secondo MDC.

Le diagnosi con il massimo numero di ricoveri per ogni MDC sono riportate in Tabella 3.5.

MDC	Diagnosi	Ricoveri
4	Insufficienza respiratoria	29910
5	Insufficienza cardiaca congestizia (scompenso cardiaco)	45760
8	Sostituzione di articolazione dell'anca	43260
11	Tumori maligni di parte non specificata della vescica	22620
14	Parto normale	104500

Tabella 3.5: Diagnosi di con numero di occorrenze massimo per ogni MDC e relativi numero di ricoveri.

Densità La **densità** della Rete Diagnosi è pari a 0,02. Sottolineiamo che nelle reti proiettate, poichè la costruzione della proiezione introduce nuove connessioni tra i nodi, il valore della densità risulta spesso significativamente più alto rispetto a quello della rete bipartita di partenza. Anche in questo caso possiamo notare che il valore di densità è superiore di due ordini di grandezza rispetto a quello della Rete Ricoveri.
É inoltre possibile valutare la **distanza media** tra i nodi della Rete Diagnosi, che risulta pari a 2,45. La media della lunghezza dei cammini minimi (in termini di numero di archi percorsi) tra ogni coppia di vertici è dunque piuttosto bassa.
Il valore del **diametro** della rete fornisce un'informazione coerente con quella data dalla distanza media. Il diametro è, infatti, pari a 10. In questo caso nel calcolo della lunghezza dei cammini minimi vengono considerati anche i pesi. É però da notare che, nell'ottica dell'individuazione del cammino più breve, non vengono mai percorsi archi con peso alto, in quanto la struttura della rete consente "percorsi alternativi" formati da archi di peso unitario o molto basso con valori complessivi inferiori.
Riportiamo come esempio un cammino composto da 5 nodi che presenta lunghezza complessiva pari al diametro. L'immagine in Figura 3.3 mette in evidenza una caratteristica della struttura della rete, che vedremo di seguito più in dettaglio: i nodi più distanti all'interno della rete hanno grado 1 ma risultano connessi tramite un cammino che passa da nodi di grado più alto.

Grado dei nodi Come dettagliato nel Capitolo 2, l'indice che rappresenta il numero di connessioni di un nodo è il suo **grado**. Nella rete proiettata che abbiamo costruito, il grado di un nodo quantifica il numero di diagnosi distinte per cui i pazienti ospedalizzati per la diagnosi rappresentata dal nodo risultano ricoverati in una diversa occasione. Un nodo con grado alto rappresenta una diagnosi che risulta legata a molte altre diagnosi. Secondo la logica con cui è stata costruita la Rete Diagnosi, la molteplicità di con-

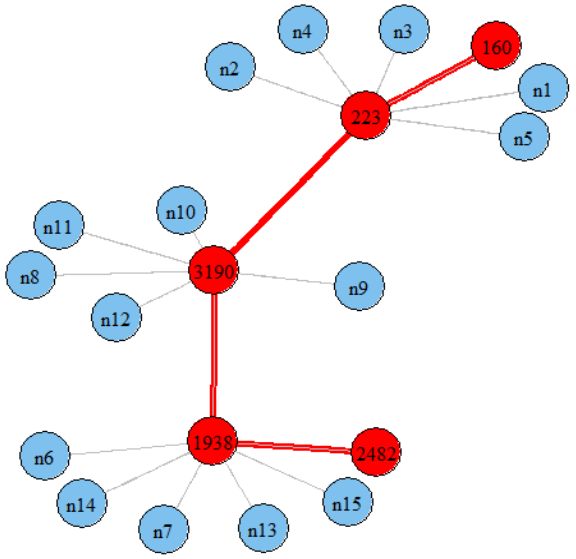

Figura 3.3: Diametro della Rete Diagnosi (in rosso), messo in evidenza in una versione ridotta della rete originale, Le diagnosi rappresentate dai nodi sono: 160-CISTITE GONOCOCCICA CRONICA, 223-TUMORI MALIGNI DI ALTRE SPECIFICATE SEDI DELLA VESCICA, 3190-ALTRE COMPLICAZIONI DA ALTRE PROTESI, IMPIANTI E INNESTI INTERNI ORTOPEDICI, 1938-LUSSAZIONE ARTICOLARE EVOLUTIVA DELL'ANCA, 2482-GENU RECURVATUM.

nessioni esprime l'eterogenità delle storie cliniche dei pazienti ricoverati per la diagnosi considerata. Un esempio di diagnosi di questo tipo è l'INSUFFICIENZA RESPIRATORIA, di grado 1043: esistono 1043 altre diagnosi che compaiono in un ricovero che coinvolge un paziente ricoverato per insufficienza respiratoria.
Nodi con basso grado denotano, invece, patologie che risultano presenti in ricoveri precedenti o successivi a ricoveri dovuti a un numero ristretto di altre diagnosi. Si tratta, in questo caso, di diagnosi con un basso numero di occorrenze o di natura molto specifica. A titolo rappresentativo, possiamo considerare la diagnosi DISTORSIONE E DISTRAZIONE DI SITO NON SPECIFICATO DEL GOMITO E DELL'AVAMBRACCIO, di grado 1: i pazienti ricoverati per essa, in caso di ricovero multiplo, risultano ospedalizzati solo per un'altra diagnosi, la POLMONITE (con AGENTE NON SPECIFICATO).
Nei grafici in Figura 3.4 e 3.5 riportiamo la distribuzione di grado in scala logaritmica e la corrispondente distribuzione cumulativa. Ciò che emerge in modo evidente è la forte eterogeneità della rete in termini di grado dei nodi: un'importante percentuale di nodi presenta un grado basso (il 75% dei nodi

ha grado inferiore a 100) e si ha un range di valori ampio, da 1 a 1067.

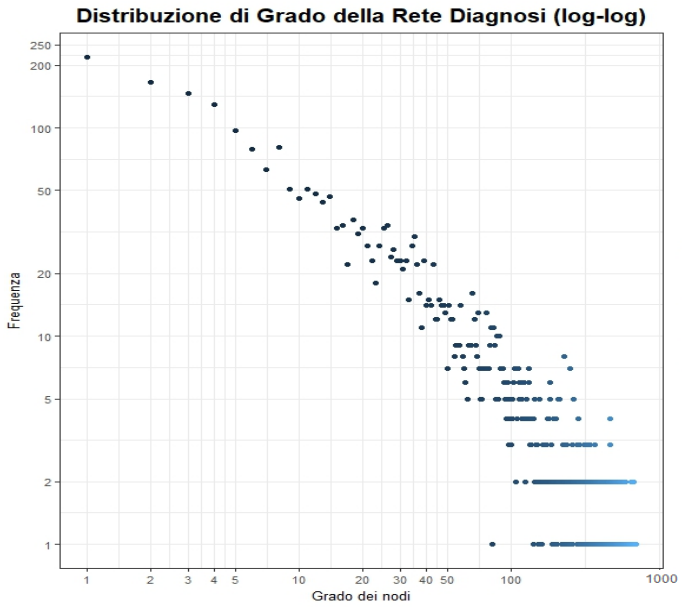

Figura 3.4: Distribuzione di grado in scala logaritmica.

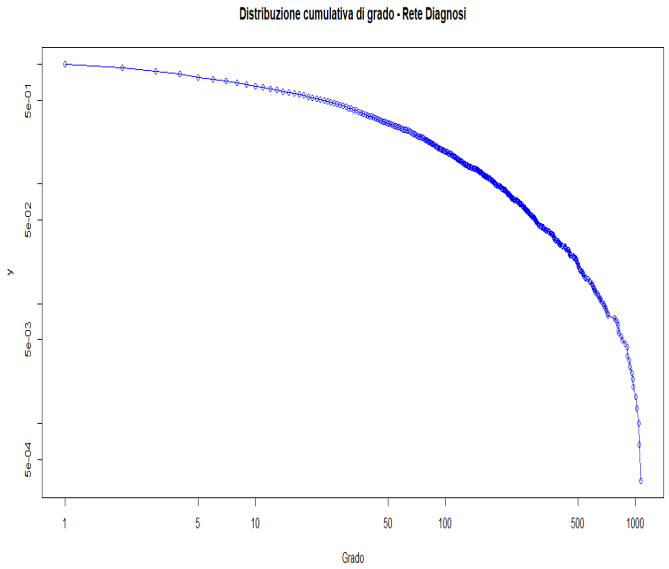

Figura 3.5: Distribuzione cumulativa di grado nella Rete Diagnosi.

Un quadro riassuntivo dei valori di grado dell'intera rete e per ogni categoria di diagnosi è riportato in Tabella 3.6. Il valore minimo di grado registrato per tutte le categorie è 1, pertanto in ogni MDC è presente almeno una diagnosi connessa ad una sola altra diagnosi. Per ogni MDC riportiamo la diagnosi di grado massimo (con il relativo valore di grado) in Tabella 3.7.

	Grado dei nodi			
MDC	**media**	**dev.std.**	**mediana**	**massimo**
4	89,2	162,3	26	1067
5	130,9	173,7	62	1050
8	53,9	103,1	18	977
11	90,8	144,0	39	906
14	35,1	72,6	10	801
totale	68,7	126,9	22	1067

Tabella 3.6: Quadro del valore dei grado dei nodi diagnosi, classificate secondo MDC.

MDC	Diagnosi	Grado
4	Polmonite, con Agente non specificato	1067
5	Insufficienza cardiaca congestizia (scompenso cardiaco)	1050
8	Sostituzione di articolazione dell'anca	977
11	Infezione del sistema urinario (in sito non specificato)	906
14	Parto normale	801

Tabella 3.7: Diagnosi di grado massimo per ogni MDC e relativi valori di grado.

Riprendendo il significato dei codici MDC presentato nel Capitolo 1, possiamo notare la coerenza delle diagnosi di grado massimo messe in evidenza in Tabella 3.7: la diagnosi di MDC 4 è una patologia cardiaca, la diagnosi di MDC 5 è, invece, legata ai polmoni e così via.

In Figura 3.6 vengono rappresentate le distribuzioni di grado nelle varie MDC. Si può notare che le diagnosi della categoria 5 e 11 presentano una distribuzione di grado concentrata su valori superiori al valore mediano della rete, pari a 22. Al contrario, il grado mediano delle diagnosi della MDC 14 risulta significativamente più basso del valore mediano. Questo risultato è ragionevole e coerente con le nostre conoscenze sulle tipologie di diagnosi. Le patologie cardio-polmonari e le patologie legate al sistema genito-urinario sono, in generale, più complesse e quindi sono più connesse ad altre co-patologie. É dunque facile che diagnosi di queste categorie inducano ricoveri

per un numero elevato di altre patologie. Diagnosi della categoria 8 e 14, che sono legate a problematiche di tipo ortopedico o riguardanti la gravidanza e il parto, inducono per loro natura a ricoveri "episodici" che difficilmente, una volta conclusi, portano a un altro ricovero per una patologia ad esse connessa.

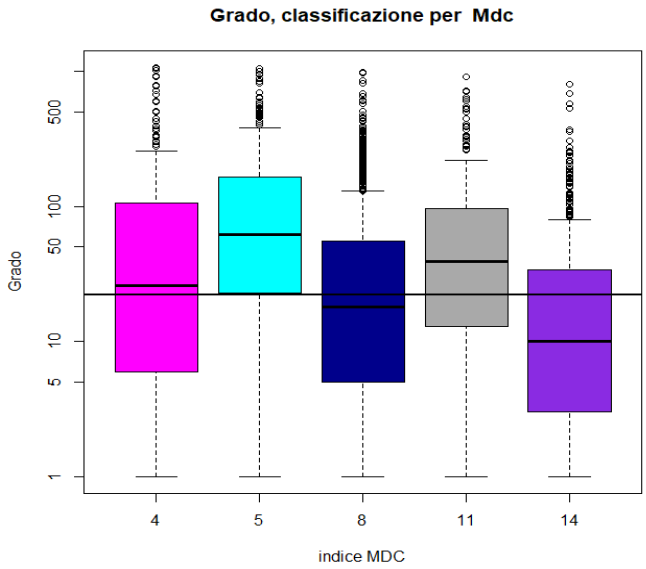

Figura 3.6: Distribuzione di grado nelle varie MDC. La linea orizzontale indica il valore mediano di grado su tutti i nodi della rete, pari a 22.

Da ultimo, è possibile evidenziare una correlazione tra il grado dei nodi e il grado medio dei loro vicini. L'andamento della funzione grado medio dei nodi vicini, definita in (2.6), è decrescente con il valore del grado: possiamo quindi definire, in base a quanto illustrato nel Capitolo 2, la Rete Diagnosi come *disassortativa*. Nodi di grado basso tendono a connettersi con nodi di grado alto: diagnosi presenti in pochi ricoveri o, più in generale, con poca varietà di altre diagnosi collegate, sono connesse preferenzialmente a diagnosi molto diffuse e per questo presenti in un numero significativo di storie cliniche differenti.

Strength dei nodi La Rete Diagnosi costruita proiettando la Rete Ricoveri è dotata di archi pesati, il cui valore corrisponde al numero effettivo di pazienti "in comune" tra una coppia di diagnosi. L'informazione portata dagli archi delle Rete Diagnosi è, dunque, duplice: la presenza di un arco $\{i, j\}$ segnala l'esistenza di almeno un paziente ricoverato in diverse ospedalizza-

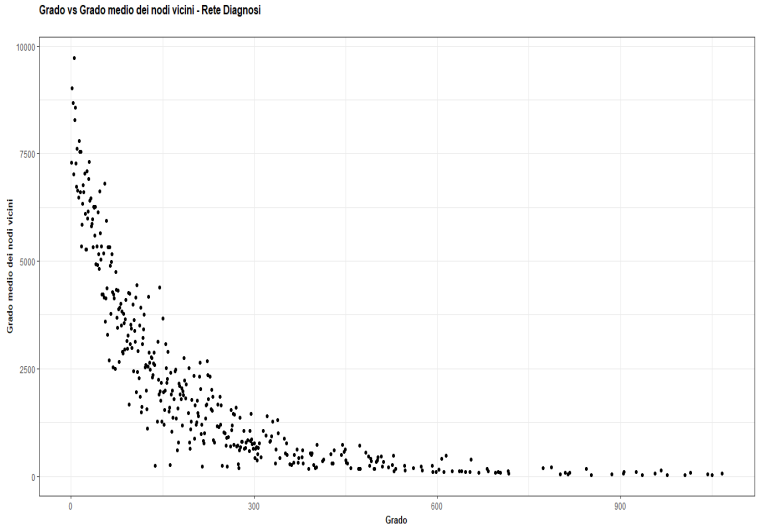

Figura 3.7: Funzione grado medio dei nodi vicini per i nodi della Rete Diagnosi, con andamento decrescente con il valore del grado.

zioni per le diagnosi i e j e il suo peso indica per quanti pazienti si verifica questa situazione. Per tenere conto di questo aspetto nell'analisi della rete, è opportuno valutare la **strength** dei nodi, cioè la somma dei pesi degli archi adiacenti a ciascun vertice. Il valore della strength dei nodi diagnosi è pari al numero complessivo di pazienti ricoverati per diagnosi rappresentata dal nodo e per tutte le altre diagnosi ad essa connesse. Calcolare la strength dei nodi consente dunque di valutare la loro importanza all'interno della rete non solo in termini di connessione con le altre diagnosi, ma anche in termini di numero di pazienti complessivi coinvolti nelle storie cliniche in cui è presente la diagnosi considerata.
Come il grado dei nodi analizzato nel paragrafo precedente, anche per la strength si registra una significativa eterogeneità nella rete. La distribuzione di strength e la distribuzione cumulativa sono rappresentate in Figura 3.8 e 3.9: si può notare un andamento simile a quello della distribuzione ma con un range di valori più ampio.

In Tabella 3.8 è riassunto un quadro dei valori di strength per l'intera rete e per ogni MDC. In tutte le MDC si registrano diagnosi di strength di valore minimo, pari a 1. Per queste diagnosi si ha un'unica altra diagnosi connessa e un unico paziente che le presenta nei suoi ricoveri. Riportiamo in Tabella 3.9 la diagnosi di strength massima per ogni categoria.
In Tabella 3.9 sono presenti diagnosi già segnalate per avere il massimo grado della loro categoria di appartenenza. Nelle categorie MDC 4 e 11, invece, non

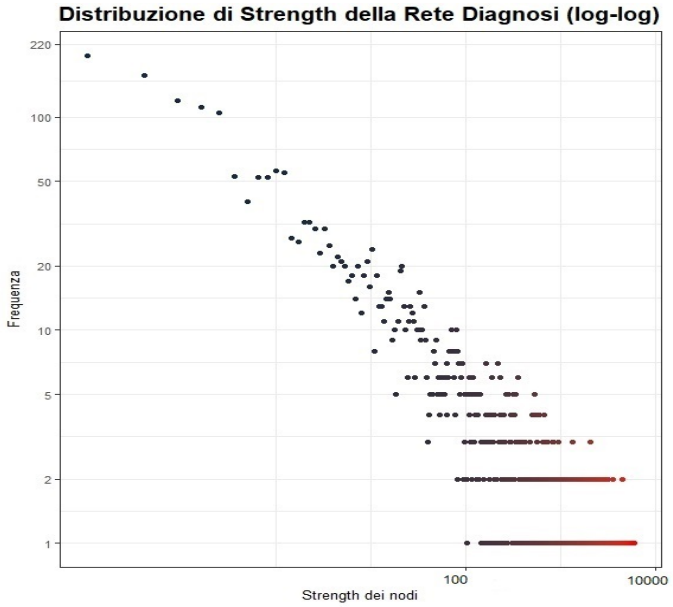

Figura 3.8: Distribuzione di strength in scala logaritmica.

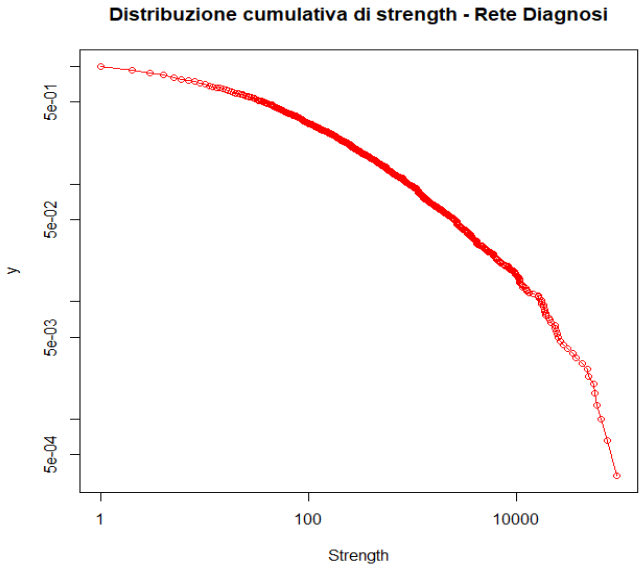

Figura 3.9: Distribuzione cumulativa di strength nella Rete Diagnosi.

vi è corrispondenza tra diagnosi di grado massimo e di strength massima ed è quindi possibile individuare in ciascuna MDC due diagnosi leader per due aspetti distinti.

Strength dei nodi				
MDC	**media**	**dev.std.**	**mediana**	**massimo**
4	1166,4	5275,3	44,5	54870
5	1903,6	7357,3	153	92973
8	431,0	3203,6	27	65270
11	972,7	3255,2	79	24460
14	256,8	1322,5	13	23980
totale	744,3	4141,4	36	92973

Tabella 3.8: Quadro del valore di strength dei nodi diagnosi, classificate secondo MDC.

MDC	Diagnosi	Strength
4	Bronchite cronica ostruttiva	54870
5	Insufficienza cardiaca congestizia (scompenso cardiaco)	92973
8	Sostituzione di articolazione dell'anca	65270
11	Tumori maligni di parte non specificata della vescica	24460
14	Parto normale	23980

Tabella 3.9: Diagnosi di strength massima per ogni MDC e relativi valori di strength.

La distribuzione dei valori di strength all'interno delle categorie è presentata nei grafici in Figura 3.10. La posizione dei grafici rispetto al valore mediano di strength (pari a 36) è analoga a quella evidenziata nei grafici in Figura 3.6 con le MDC 5 e 11 significativamente al di sopra di esso e la MDC 14 al di sotto.

Risulta chiaro, come si può vedere anche in Figura 3.11, che le distribuzioni di grado e di strength siano fortemente correlate in questa rete. La motivazione è data dalla definizione stessa di strength: i nodi ricevono per ogni nodo adiacente ad essi un contributo maggiore o uguale a 1 alla loro strength. Per il significato dei nodi della nostra rete, inoltre, non sorprende che siano le diagnosi più connesse ad avere il maggior numero complessivo di pazienti coinvolti in un ricovero per la diagnosi stessa e per un'altra ad essa collegata: maggiore è il numero di diagnosi "vicine", più è facile che il bacino di pazienti coinvolti, nel senso precisato all'inizio del paragrafo, sia grande.

Per completare il quadro fornito in questa sezione, mostriamo in Figura 3.12 la correlazione esistente tra il numero di occorrenze nel dataset di una

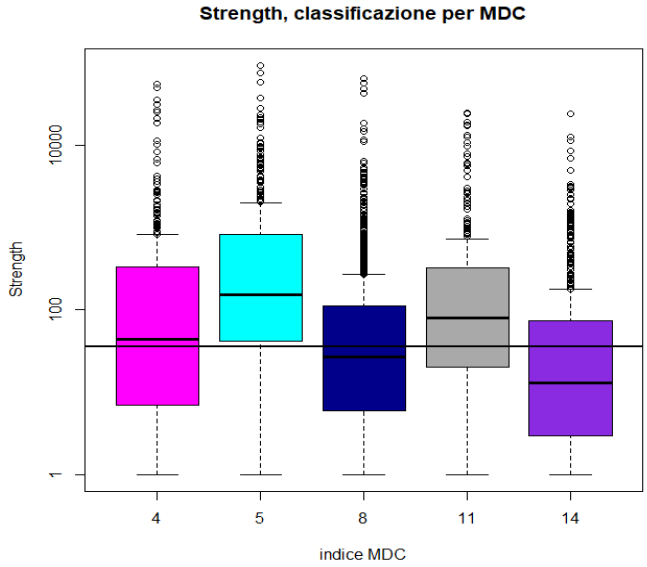

Figura 3.10: Distribuzione di strength nelle varie MDC, La linea orizzontale indica il valore mediano di strength su tutti i nodi della rete, pari a 36.

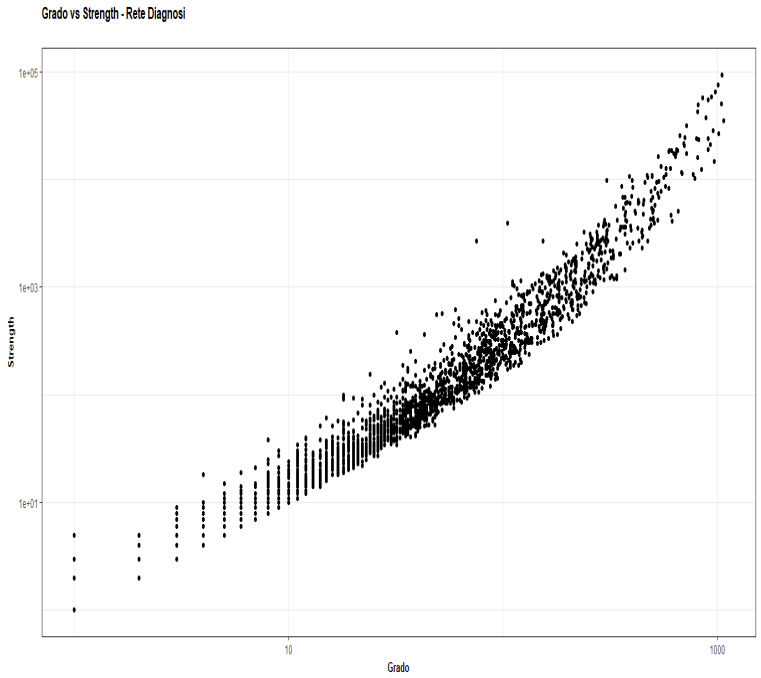

Figura 3.11: Confronto tra grado e strength dei nodi, Rete Diagnosi.

diagnosi e il suoi valori di grado e strength nella rete proiettata. Come ci si può aspettare, le diagnosi più frequenti e diffuse tendono ad essere connesse a un numero elevato di altre diagnosi, presenti in altri ricoveri degli stessi pazienti che vengono ricoverati per esse e, di conseguenza, ad essere presente in storie cliniche che coinvolgono un numero elevato di pazienti. Sottolineiamo, tuttavia, che il dato relativo al numero di ricoveri riguarda la frequenza in termini assoluti delle diagnosi nel dataset originale, mentre i valori di grado e strength sono ricavati dalla struttura della rete proiettata che abbiamo costruito.

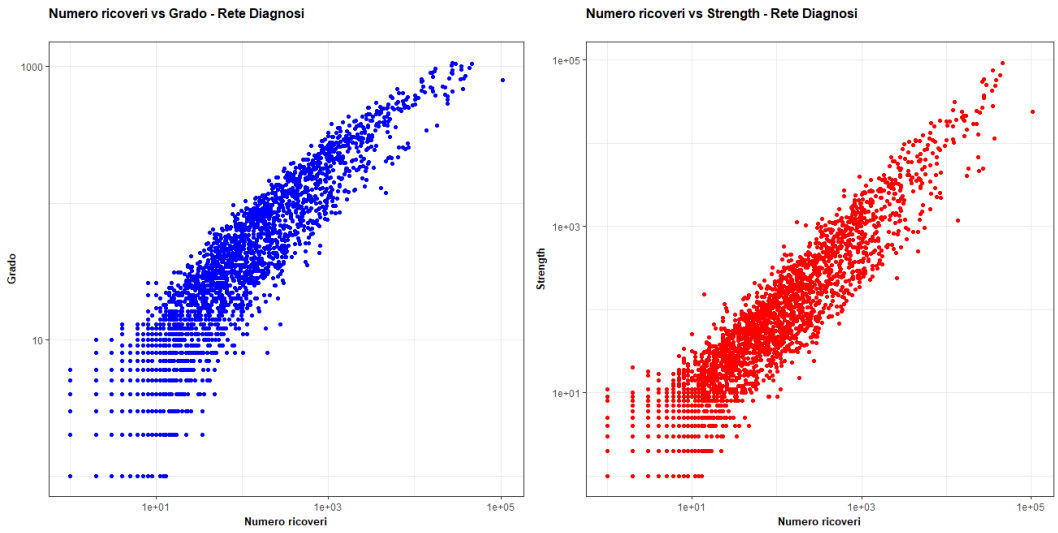

Figura 3.12: Grafici del grado (a sinistra) e della strength dei nodi diagnosi in funzione del loro numero di occorrenze (numero di ricoveri per diagnosi).

Betweenness Un ulteriore valore che può essere calcolato per analizzare i nodi della rete è la loro **betweenness**. Come è stato illustrato nel Capitolo 2, questa misura di centralità consente di individuare i nodi importanti all'interno della struttura perchè attraversati da una frazione significativa dei cammini minimi che collegano le coppie di vertici. Individuare nodi diagnosi con significativi valori di betweenness consente di mettere in evidenza patologie o, in generale, situazioni che portano a un ricovero con un ruolo centrale e di "ponte" tra le diagnosi della rete.
Ricordiamo che la Rete Diagnosi è una rete pesata ed è quindi possibile valutare la betweenness dei nodi tenendo conto di questa caratteristica. Decidiamo di presentare la betweenness dei nodi della Rete Diagnosi calcolata *senza* considerare i pesi degli archi e una sua seconda versione "pesata". Il calcolo della betweenness pesata richiede, come precisato nel Capitolo 2,

un'inversione dei pesi degli archi della rete prima dell'esecuzione.
Per i valori di betweenness dei nodi, nella versione non pesata, proponiamo in Figura 3.13 la distribuzione cumulativa e in Tabella 3.10 un quadro riassuntivo per la rete complessiva e per ogni MDC. Si può notare, anche in questo caso, una forte eterogeneità di valori: a una betweenness mediana piuttosto contenuta (pari a 26) si associa un valore massimo di 330250. Si ha dunque conferma di una struttura eterogenea della rete anche dal punto di vista della betweenness. Un numero consistente di nodi presenta valori bassi e sono solo pochi i nodi che svolgono, in modo predominante, un ruolo centrale nella rete. Sottolineaiamo che, in questa versione "non pesata" della betweenness, la centralità dei nodi è valutata attribuendo lo stesso peso a ogni arco.

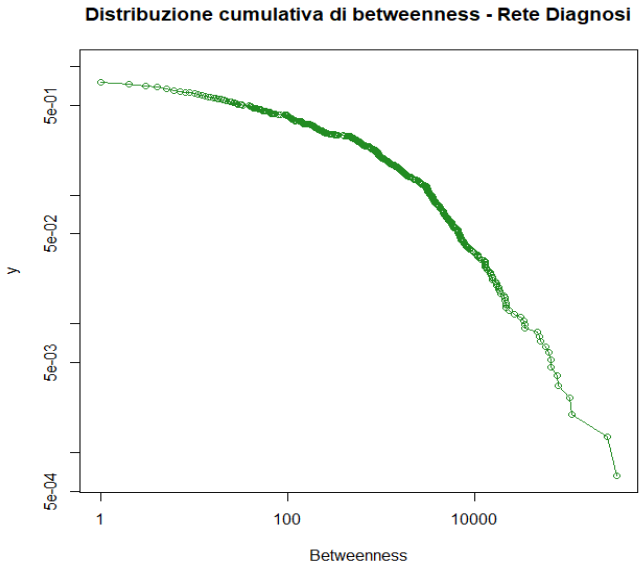

Figura 3.13: Distribuzione cumulativa di betweenness nella Rete Diagnosi.

In Tabella 3.11 riportiamo l'elenco delle diagnosi che presentano il valore di betweenness più alto per ogni MDC. Nei grafici in Figura 3.14 emergono, analogamente a quanto osservato per la distribuzione di grado e di strength, MDC con valori distribuiti al di sopra del valore mediano di betweenness della rete (5 e 11), e una MDC (14) con valori concentrati su un livello inferiore.

Nelle Tabelle 3.12 e 3.13 sono riportati il quadro riassuntivo dei valori di betweenness pesata e l'elenco dei valori massimi di betweenness pesata per ogni MDC. Dal confronto con le betweenness presentate sopra emergono valori massimi decisamente più alti, ma anche valori mediani nulli per

Betweenness dei nodi				
MDC	**media**	**dev.std.**	**mediana**	**massimo**
4	3239,5	15319,2	38	147249
5	2707,5	10956,3	97	103117
8	1588,1	9731,4	23	257840
11	2397,1	9634,8	62	85297
14	2427,7	17424,9	3	330250
totale	2196,9	12584,8	26	330250

Tabella 3.10: Quadro del valore di betweenness dei nodi diagnosi, classificate secondo MDC.

MDC	Diagnosi	Betweenness
4	Polmonite con agente non specificato	147249
5	Insufficienza cardiaca congestizia (scompenso cardiaco)	103117
8	Trattamento per rimozione di dispositivo di fissazione interna	257840
11	Infezione del sistema urinario (in sito non specificato)	85297
14	Parto Normale	330250

Tabella 3.11: Diagnosi di betweenness massima per ogni MDC e relativi valori di betweenness.

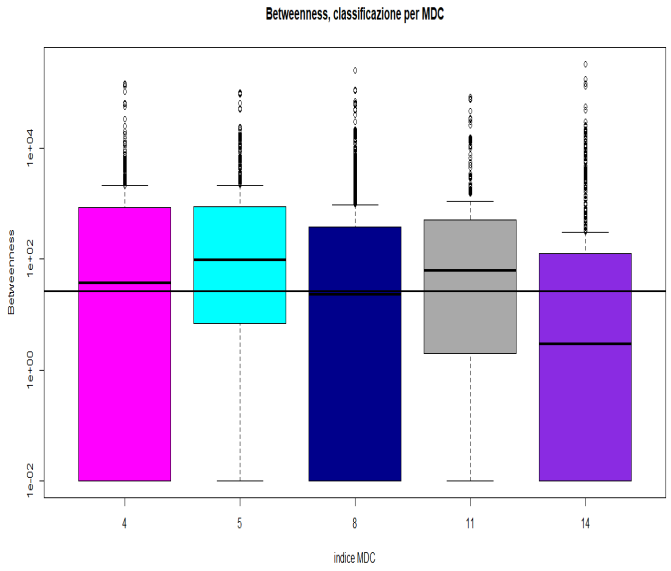

Figura 3.14: Distribuzione di betweenness nelle varie MDC, La linea orizzontale indica il valore mediano di betweenness su tutti i nodi della rete, pari a 26.

ogni MDC. Inoltre, sottolineiamo che solo 641 nodi su 3013 hanno valore di betweenness pesata diverso da 0. Tenere conto dei pesi della rete consente dunque di distinguere in modo più forte i nodi con un ruolo centrale, in quanto i collegamenti tra i nodi che costituiscono i cammini minimi non risultano tutti equivalenti, ma tanto più significativi (in questo caso "brevi") quanto più grande è il numero di pazienti coinvolti in ricoveri per entrambe le diagnosi rappresentate.

Betweenness pesata dei nodi				
MDC	**media**	**dev.std.**	**mediana**	**massimo**
4	6501,8	37175,4	0	483015
5	12354,3	160204,6	0	3248308
8	5245,6	56598,1	0	1155167
11	24470,1	170067,1	0	1477218
14	4831,9	68155,5	0	1633186
totale	8165,9	94692,2	0	3248308

Tabella 3.12: Quadro del valore di betweenness pesata dei nodi diagnosi, classificate secondo MDC.

MDC	**Diagnosi**	**Betweenness pesata**
4	Bronchite cronica ostruttiva	483015
5	Insufficienza cardiaca congestizia (scompenso cardiaco)	3248308
8	Postumi di fratture agli arti inferiori	1155167
11	Insufficienza renale acuta	1477218
14	Parto normale	1633186
totale	Insufficienza cardiaca congestizia (scompenso cardiaco)	3248308

Tabella 3.13: Diagnosi di betweenness pesata massima per ogni MDC e relativi valori di betweenness.

É interessante confrontare i valori di betweenness (non pesata) dei nodi con il loro grado. Nel grafico in Figura 3.15 emerge una forte correlazione tra grado e betweenness dei nodi. Questa situazione è tipica delle reti random e si osserva molto frequentemente anche nelle reti reali: i nodi centrali, dal punto di vista della betweenness, sono anche i nodi più connessi. É interessante allora mettere in evidenza i nodi diagnosi che si discostano da questo comportamento: l'anomalia dei loro valori di grado e betweenness può essere utile a far emergere un ruolo particolare all'interno della rete. Questo tipo

di approccio è stato introdotto in [19] per lo studio di anomalie negli indici di centralità dei nodi della rete del trasporto aereo. In Figura 3.16 sono evidenziate quattro zone delimitate dai valori medi di grado e betweenness, pari a 68 e 2196.

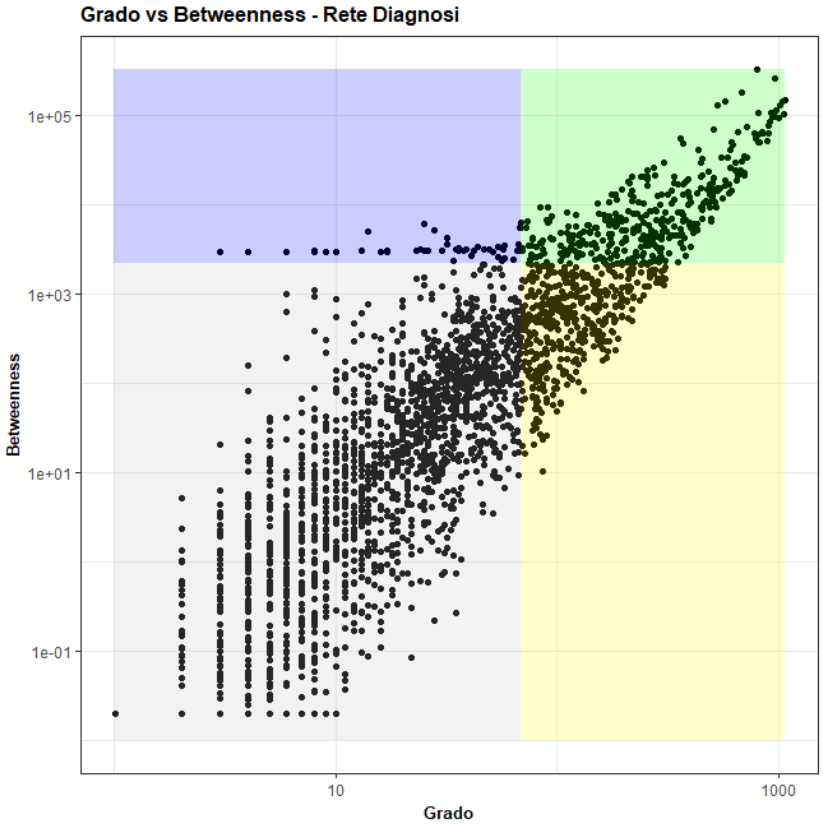

Figura 3.15: Confronto tra grado e betweenness dei nodi nella Rete Diagnosi. In evidenza le zone di grado e betweenness al di sotto della media (grigio), grado al di sotto della media e betweenness al di sopra (blu), grado al di sopra della media e betweenness al di sotto (giallo), grado e betweenness al di sopra della media (verde).

Risulta innanzitutto evidente che un numero significativo di nodi (2184 , il 72,4%) presenta valori al di sotto della media per entrambi gli indici e si trova dunque nella zona grigia (in basso a sinistra). La zona verde, in alto a destra, contiene, invece, i nodi con entrambi i valori di grado e betweenness sopra la media. I nodi con queste caratteristiche sono 351 (11,6%). Nelle zone blu e gialla sono presenti i nodi che possiamo considerare anomali: nella prima vi sono diagnosi con betweenness sopra la media ma grado al di sotto e nella seconda si ha la situazione opposta. Nelle Tabelle 3.14 e 3.15

sono elencate, con i rispettivi valori, un campione di 8 diagnosi appartenenti a questi gruppi. I nodi con alto valore di betweenness e basso grado sono in tutto 48 (1,6%), più numerosi, 490 (16,2%), sono i nodi con alto valore di grado e bassa betweenness.

Diagnosi	**Betweenness**	**Grado**
Distorsione e distrazione metacarpofalangea	2442,6	63
Patologia non specificata dell'uretra	2335,8	56
Artralgia,tibia e perone	2564,2	58
Frattura chiusa di sette costole	3092,3	67
Tumori maligni di connettivo	3076,7	66
Coartazione aortica	2625,0	54
Tumori maligni dell'arto superiore	3646,2	67
Postumi di distorsioni e distrazioni	3278,4	57

Tabella 3.14: Esempi di diagnosi con valore di betweenness sopra la media e valore di grado sotto la media.

Diagnosi	**Betweenness**	**Grado**
Occlusione coronarica acuta	10,2	86
Anomalie respiratorie non specificate	16,5	71
Sindrome post infartuale	20,6	78
Infarto miocardico acuto	23,3	86
Infarto strettamente posteriore	20,0	70
Aneurisma aortico	26,4	90
Insufficienza cardiaca diastolica	27,1	84
Embolia e trombosi dell'aorta addominale	29,7	77

Tabella 3.15: Esempi di diagnosi con valore di betweenness sotto la media e valore di grado sopra la media.

Questo approccio consente di mettere in evidenza le diagnosi "anomale" con alta betweenness e basso grado, che difficilmente sarebbero emerse come significative. Le diagnosi con basso valore di grado, infatti, risulterebbero trascurabili in una prima analisi, ma la loro betweenness suggerisce che abbiano un ruolo importante all'interno della rete. La rilevanza a livello clinico di queste patologie può essere indagata solo dispondendo conoscenze specifiche. Dal punto di vista dell'analisi della rete, invece, possiamo affermare

che l'importanza di queste diagnosi è data dal fatto che, pur essendo poco connesse, sono nodi "ponte" che garantiscono la connessione della rete. Considerando le patologie con queste caratteristiche, riportate parzialmente in Tabella 3.10, possiamo riconoscere che si tratta per lo più di diagnosi specifiche che non compromettono nello stesso momento tanti organi o apparati. Seguendo lo stesso metodo, è possibile confrontare i valori di betweenness pesata con la strength dei nodi, sostituendo all'analisi appena svolta alla betweenness e al grado i valori corrispondenti che tengono conto del peso degli archi. In Figura 3.16 viene riportato il grafico strength-betweenness pesata e le quattro zone descritte sopra delimitate dai valori medi di strength e betweenness pesata, rispettivamente pari a 744 e a 8165.

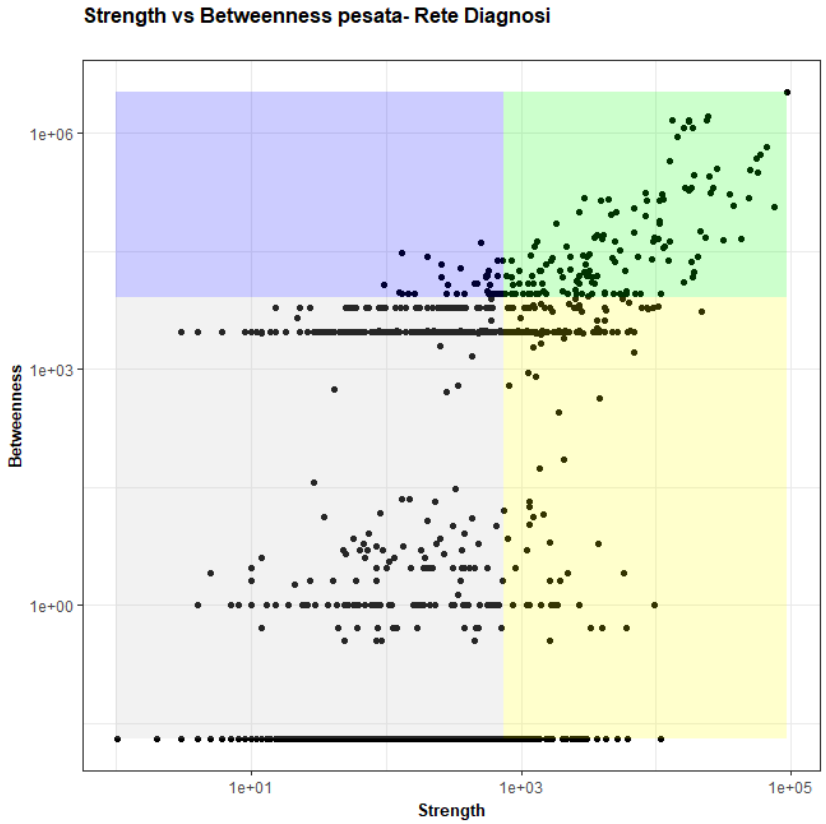

Figura 3.16: Confronto tra strength e betweenness dei nodi nella Rete Diagnosi. In evidenza le zone di strength e betweenness al di sotto della media (grigio), strength al di sotto della media e betweenness al di sopra (blu), strength al di sopra della media e betweenness al di sotto (giallo), strength e betweenness al di sopra della media (verde).

Anche in questo caso emerge chiaramente che una percentuale consistente di nodi, l' 87,4%, presenta valori al di sotto della media per entrambi gli

indici. I nodi con valori di strength e betweenness pesata sopra la media sono invece 138 (4.5%). Per quanto riguarda le zone in cui sono messe in evidenza le anomalie, si registrano 33 nodi (1%) con betweenness pesata alta e strength bassa e 207 nodi (6.8%) con strength alta e betweenness pesata bassa.
Nelle Tabelle 3.16 e 3.17 riportiamo alcune diagnosi rappresentative dei gruppi in cui si registrano le anomalie, rispettivamente perchè si ha betweenness (pesata) sopra la media e strength al di sotto (zona blu) e strength sopra la media e betweenness (pesata) al di sotto (zona gialla).

Diagnosi	**Betweenness pesata**	**Strength**
Entesopatia non specificata della caviglia e del tarso	9011	727
Difetto del setto atriale	9030	703
Frattura chiusa dell'omero	9030	689
Sostituzione di articolazione della spalla	9043,3	675
Diabete con complicazione antepartum	9031	654
Tumori maligni dell'arto inferiore	9029	610
Ipertensione venosa cronica con ulcera	9029	576
Lussazione chiusa,acromioclavicolare	9875,3	557

Tabella 3.16: Esempi di diagnosi con valore di betweenness (pesata) sopra la media e valore di strength sotto la media.

Diagnosi	**Betweenness pesata**	**Strength**
Presentazione podalica o complicazione antepartum	6018,5	793
Perdite ematiche antepartum	6021	797
Arterite non specificata	6036	801
Malattie dell'apparato respiratorio	6021	801
Degenerazione del disco intervertebrale lombare	6043	826
Pleurite con versamento	6026,5	869
Sviluppo fetale insufficiente	5996	1064
Frattura chiusa dell'olecrano	6020	1070

Tabella 3.17: Esempi di diagnosi con valore di betweenness (pesata) sotto la media e valore di strength sopra la media.

3.3 Proiezione della Rete Ricoveri: Rete Pazienti

In questa sezione presenteremo la proiezione della Rete Ricoveri sui nodi di tipo paziente. L'analisi della **Rete Pazienti** consente di studiare le relazioni all'interno del dataset dei ricoveri da un punto di vista diverso rispetto a quello considerato per la Rete Diagnosi. Le connessioni tra due nodi in questa rete, infatti, rappresentano il fatto che i pazienti in questione siano ricoverati, in uno dei loro ricoveri, per la stessa diagnosi.
La proiezione della Rete Ricoveri sui nodi paziente è costituita da 1351224 nodi e circa 180 milioni di archi. La dimensione particolarmente elevata di questa rete rende impossibile la sua effettiva costruzione e la sua gestione tramite il software R e gli strumenti di calcolo a nostra disposizione.
Per poter disporre di una proiezione su questo insieme di nodi ai fini di un'analisi di base, che potrà poi essere approfondita in altri lavori, decidiamo di campionare il dataset dei ricoveri che costituisce la rete bipartita. La selezione dei ricoveri è stata effettuata tramite un campionamento stratificato che conserva le proporzioni di ricoveri per pazienti maschi e femmine presenti nel dataset originale e, in un secondo livello, le proporzioni dei ricoveri relative a 13 fasce di età dei pazienti. I dettagli del campionamento sono riportati in Tabella 3.18.

La rete campionata costruita riducendo il dataset è costituita da 62773 ricoveri, pari al 3% dei ricoveri totali. I pazienti distinti coinvolti nei ricoveri selezionati dal nostro campionamento random sono 61551 (il 4,5% dei pazienti) e si registrano in totale 1916 diagnosi (il 58% delle diagnosi). Questo tipo di campionamento consente di preservare alcune caratteristiche fondamentali della rete originale e ci aspettiamo, per questo, di ottenere risultati che possano valere verosimilmente per l'intera rete.
Avendo a disposizione una rete bipartita di dimensioni ridotte, possiamo ottenere la proiezione sui nodi pazienti tramite la funzione `bipartite_projection` del pacchetto *igraph* [15] di R, analogamente a quanto fatto per la proiezione sui nodi diagnosi. Riportiamo in Tabella 3.19 le dimensioni della rete proiettata costruita, messe a confronto con la rete che si otterrebbe proiettando la rete non campionata.

Descriviamo di seguito alcune proprietà della particolare Rete Pazienti ottenuta dalla proiezione della Rete Ricoveri campionata che abbiamo descritto. Si tenga conto che la rete analizzata nel nostro lavoro dipende fortemente dal campionamento casuale effettuato sulla Rete Ricoveri: la selezione di determinati ricoveri nel dataset di partenza induce una determinata struttura della rete proiettata. Se, infatti, il campionamento è costruito in modo

CAMPIONAMENTO RETE RICOVERI				
Età	**maschi**		**femmine**	
	totali	**rete campionata**	**totali**	**rete campionata**
0	12909	387	9586	288
1-5	14870	446	11863	356
6-10	9943	298	7838	235
11-15	15636	469	12023	361
16-20	15936	478	20385	612
21-30	31181	935	145371	4361
31-40	44939	1348	233952	7019
41-50	81180	2435	76473	2294
51-60	119235	3577	82678	2480
61-70	191886	5757	138938	4168
71-80	238913	7167	209217	6276
81-90	134342	4030	179013	5370
90-100	15429	463	3793	1137
>100	178	5	709	21
totale	926577	27795	1131839	34978

Tabella 3.18: Numero di ricoveri nella rete campionata, per genere e fasce di età.

tale da conservare, in linea di principio, anche le proporzioni esistenti tra il numero di occorrenze delle diagnosi, è possibile che l'esclusione di una percentuale alta (il 42%) delle diagnosi totali porti a una rete proiettata priva di connessioni significative tra alcuni dei nodi paziente. Sottolineiamo, quindi, l'importanza di utilizzare l'intera rete bipartita, se gli strumenti di calcolo lo consentono, o di analizzare reti proiettate ottenute da diversi campionamenti della rete originale e di confrontare i relativi risultati.

Rete Pazienti	nodi paziente	archi
Proiezione della Rete Ricoveri originale	1351224	180 M
Proiezione della Rete Ricoveri campionata	61551	17,5 M

Tabella 3.19: Dimensioni della Rete Pazienti completa (proiezione della Rete Ricoveri campionata) e della Rete Pazienti ridotta (ottenuta proiettando la Rete Ricoveri campionata).

In primo luogo, possiamo dire che la Rete Pazienti considerata è una rete *semplice*, in quanto non sono presenti archi multipli e self-loop, e dotata di pesi. Il peso di un arco in questa rete rappresenta il numero di diagnosi in comune tra una coppia di pazienti. Ad esempio, l'arco $\{i, j\}$ ha peso pari a 3 se il paziente i e il paziente j risultano entrambi ricoverati, in ricoveri distinti, per le stesse 3 diagnosi. Precisiamo, tuttavia, che, avendo campionato solo una percentuale molto piccola della rete originale, l'informazione relativa ai ricoveri multipli di uno stesso paziente può risultare fortemente distorta nei nostri risultati. Il peso degli archi può essere, dunque, trascurato in una prima analisi e il legame tra nodi pazienti può essere considerato solo dal punto di vista qualitativo.
La Rete Pazienti risulta, inoltre, non connessa: si individuano, infatti, in questo caso, 1497 componenti distinte. Possiamo notare che il quadro delle componenti di questa rete risulta molto diverso da quello della Rete Diagnosi, dove si distingue una massima componente connessa e un certo numero di nodi isolati che costituiscono componenti a sè stanti. In questo caso, invece, possiamo individuare una *componente gigante*, formata da 49011 nodi, e 451 nodi isolati. I restanti nodi sono, invece, raggruppati in componenti connesse di dimensione intermedia (da 2 a 338 nodi). Come abbiamo fatto per la Rete Diagnosi, anche in questo caso ci concentreremo sull'analisi della sola massima componente connessa. É, tuttavia, interessante notare che le componenti di dimensione intermedia, che contengono in totale circa il 20% dei nodi pazienti, costituiscono dei gruppi di nodi connessi tra loro e isolati dal resto della rete che possono essere analizzati in modo separato. I nodi appartenenti a queste componenti connesse della rete rappresentano gruppi ristretti di pazienti che "condividono" tra loro una o più diagnosi ma che non hanno diagnosi in comune con il resto dei pazienti nella rete. I nodi isolati, invece, rappresentano pazienti ricoverati per una diagnosi con un'unica occorrenza nel dataset campionato. Possiamo notare, quindi, che la struttura della Rete Pazienti, benchè ottenuta da un campionamento random, riflette l'eterogeneità delle diagnosi nel dataset dal punto di vista del numero di occorrenze. Diagnosi rare inducono, infatti, nodi isolati nella Rete Pazienti, mentre diagnosi con alta frequenza portano a numerose connessioni tra i nodi paziente.
La densità della massima componente connessa della rete analizzata è pari circa all'1%: si ha che, in media, ogni paziente è connesso, cioè presenta una diagnosi in comune, con un paziente su 100 del dataset. La distanza media calcolata tra i nodi della rete è di 3,95 (sempre in riferimento alla massima componente connessa). Questo fatto indica che, in media, il cammino minimo che connette una coppia di pazienti è costituito da circa 4 archi, in cui

ogni arco rappresenta una diagnosi in comune tra una coppia di pazienti. Questo valore risulta particolarmente basso in quanto, come abbiamo potuto osservare in precedenza, le diagnosi presenti nel dataset sono in numero nettamente inferiore ai pazienti.
Per i nodi della massima componente connessa della Rete Pazienti è stata calcolata la distribuzione di grado e di strength, di cui riportiamo un quadro riassuntivo in Tabella 3.20 e Tabella 3.21.

Grado dei nodi			
media	**dev.std.**	**mediana**	**max**
707	759,7	505	4281

Tabella 3.20: Quadro dei valori di grado dei nodi paziente.

Strength dei nodi			
media	**dev.std.**	**mediana**	**max**
711	763,6	505	6326

Tabella 3.21: Quadro dei valori di strength dei nodi paziente.

Il grado dei nodi indica il numero di pazienti che presentano una diagnosi in comune con il paziente del nodo in questione. Si ha dunque che il grado di tutti i nodi pazienti ricoverati per una certa diagnosi è pari al numero di occorrenze nel dataset considerato di questa diagnosi. Il valore del grado dei nodi dipende dalla dimensione della rete ridotta: in questo caso abbiamo che, in media, i pazienti sono connessi a 707 altri nodi pazienti (circa all'1% dei pazienti, coerentemente con il dato relativo alla densità della rete). Notiamo, inoltre, che il valore mediano di grado è piuttosto alto. Il massimo grado registrato è di 4281. I nodi paziente che presentano grado massimo e i gradi più elevati risultano ricoverati per PARTO e per CALCOLOSI RENALE, che sono due cause di ricovero con alta frequenza di occorrenza.
Per quanto riguarda la strength dei nodi paziente, notiamo che i valori risultano simili a quelli relativi al grado. Questo fatto indica che il numero di diagnosi in comune tra una coppia di pazienti (cioè il peso degli archi della rete) è 1 in quasi tutti i casi. Da una parte, possiamo ricondurre questo risultato al campionamento dei ricoveri fatto, dall'altra possiamo ritenere, vista la distribuzione del numero di ricoveri per paziente illustrata in Tabella 1.4, che anche nella Rete Pazienti completa si avrebbe una situazione simile.

Si registra, infatti, che la maggior parte dei pazienti viene ricoverato per una sola volta nel corso del periodo di riferimento e risulta, dunque, associabile ad un'unica diagnosi.
I valori di betweenness dei nodi di questa rete non sono stati calcolati. Un'analisi che mette in evidenza la centralità dei nodi nel senso dato dalla betweenness, infatti, porterebbe a risultati difficilmente interpretabili e poco utilizzabili.
Più interessante, invece, sarebbe valutare per ogni diagnosi, o categoria di diagnosi, la similarità dei nodi pazienti che risultano connessi perchè la presentano entrambi in un loro ricovero. L'analisi della struttura delle sottoreti indotte da singole diagnosi e delle caratteristiche dei nodi pazienti che le costituiscono può rivelarsi un approccio utile per la profilazione del rischio clinico dei pazienti. É possibile, infatti, che per una data diagnosi siano ricoverati pazienti con caratteristiche simili che una struttura di rete può contribuire, rappresentatando la loro connessione tramite archi, a mettere in evidenza. Al contrario, il ricovero per una stessa diagnosi può costituire un elemento che mette in relazione pazienti con caratteristiche diverse (per genere, per fascia di età o per altri aspetti più specifici). Analizzare i dati relativi alle ospedalizzazioni con questo approccio può costituire un supporto alla gestione del ricovero del singolo paziente che tenga conto delle caratteristiche dei pazienti ad esso correlati per motivazioni cliniche e non solo simili per caratteristiche anagrafiche.

Capitolo 4

Metodi di Community Detection

Un aspetto importante dell'analisi delle reti è l'individuazione di comunità costituite dai loro nodi (*community detection*). La ricerca della migliore partizione dei nodi di una rete secondo un determinato criterio è un problema ampio e affrontato in letteratura con molti approcci diversi. In [17] è fornito un quadro generale dell'argomento. In questo capitolo ne presentiamo alcuni aspetti, introducendo il concetto di comunità, che si presta a diverse interpretazioni, e i metodi di clustering (metodo di Louvain e modello stocastico a blocchi) che saranno poi applicati alle reti analizzate nel nostro lavoro. Seguirà poi l'esposizione del problema della validazione della partizione ottenuta e di alcuni indici utili a questo scopo. Viene infine presentato un metodo di classificazione dei nodi di una rete basata sui loro ruoli, definiti sulla base delle comunità a cui vengono assegnati.

4.1 Introduzione al problema e ai metodi

Nel contesto delle reti complesse è comune che i nodi presentino una struttura di comunità e che tendano a formare gruppi (detti anche "blocchi" o "moduli") con determinate caratteristiche. Identificare la struttura la rete consente di mettere in evidenza l'organizzazione interna dei nodi e le relazioni esistenti tra di essi. É possibile, per esempio, individuare e focalizzare l'analisi su regioni con una certa autonomia rispetto all'intera rete o classificare i nodi sulla base del loro ruolo rispetto ai nodi della comunità a cui sono assegnati. La definizione di comunità, tuttavia, non è univoca e le numerose interpretazioni diverse di questo concetto portano ad altrettanto numerosi metodi per identificarle. In generale, si definisce *comunità* un sottoinsieme C dell'insieme dei nodi che presenti una densità di connessioni al suo interno

significativamente più alta rispetto alla densità di connessioni esistenti tra i nodi di C e quelli non appartenti ad esso. Possiamo dire, quindi, che le comunità sono sottoreti densamente connesse ben distinte dal resto dei nodi della rete. Per definire in modo appropriato una comunità è necessario dunque tenere conto sia della coesione interna della sottorete sia della sua separazione. Un esempio di rete partizionata in 3 comunità ben distinte è riportato in Figura 4.1.
Una definizione di comunità coerente con questo approccio e molto diffusa è quella di *sottorete tale che il numero di archi interni (rivolti verso nodi appartenenti al gruppo) è maggiore del numero di archi esterni (rivolti verso nodi non appartenenti al gruppo)*. Da questa definizione derivano quelle di *strong community* e *weak community*: la prima è una sottorete tale che per ogni nodo ad essa appartenente il grado interno è maggiore del grado esterno e la seconda, in cui si rilassa la condizione, è una sottorete tale che il grado interno complessivo tra tra tutti i nodi della sottorete è maggiore del grado esterno complessivo. Si precisa che per grado interno e esterno di una sottorete si intende la somma rispettivamente dei gradi interni e esterni dei nodi appartenenti alla sottorete C.
Una definizione alternativa di comunità tiene conto anche della relazione esistente tra i vari blocchi in cui una rete risulta suddivisa: *una sottorete C è una comunità se ognuno dei nodi appartenenti ad essa è più fortemente legato ai nodi di C rispetto ai nodi di ogni altra comunità*. Secondo questa interpretazione, allora, C è una strong community se il grado interno di ogni nodo rivolto a C è maggiore del grado del nodo rivolto a ogni altra sottorete, è invece una weak community se il grado interno di ogni nodo rivolto a C è maggiore del grado complessivo del nodo rivolto a tutte le altre sottoreti.

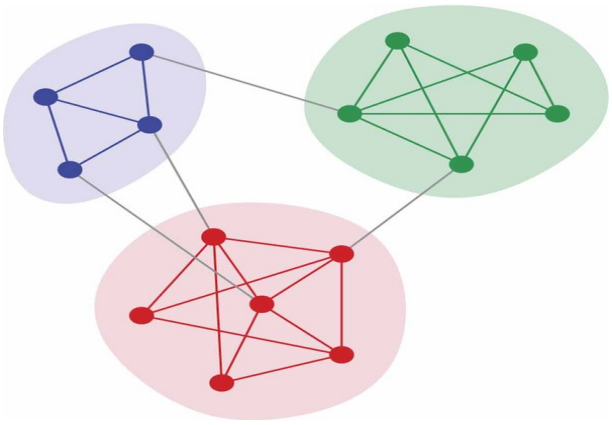

Figura 4.1: Esempio di rete in cui è individuata una partizione in 3 comunità

Nelle interpretazioni presentate finora la definizione di comunità risulta sempre legata al numero effettivo di archi presenti tra i nodi della rete. In un approccio più moderno al concetto di comunità, su cui si basano i metodi di community detection sviluppati più recentemente, ci si concentra invece sulla *probabilità* che i nodi siano connessi da un arco in una rete in cui è assegnata una partizione: nodi appartenti alla stessa comunità hanno maggiore probabilità di essere connessi tra loro rispetto a nodi appartenenti a comunità diverse. Da questa formulazione in termini probabilistici del concetto di comunità seguono nuove definizioni di strong e weak community. Si definisce strong community una sottorete C i cui nodi hanno probabilità più alta di essere connessi a ogni altro nodo appartenente a C rispetto a ogni altro nodo non appartenente a C. Una weak community è, invece, una sottorete C in cui la probabilità media che esista un arco tra ciascuno dei nodi e gli altri nodi di C è maggiore della probabilità media che esista un arco tra ciascuno dei nodi e i nodi di ognuno degli altri gruppi. La differenza tra le due definizioni è data dal fatto che nel concetto di strong community la probabilità di esistenza degli archi è confrontata a livello di ciascuna coppia di nodi, mentre in quello di weak community si valutano disuguaglianze di probabilità di esistenza degli archi mediate su tutte i gruppi.
Il valore della probabilità di un arco tra le coppie di nodi può essere calcolato basandosi sulle ipotesi che defiscono il modello generativo della rete. Ad esempio, si può ipotizzare che in una rete sociale la probabilità di connessione tra due nodi decresca in media con la distanza geografica tra gli individui rappresentati dai nodi.
Il più noto modello generativo, basato sulle probabilità di esistenza degli archi che descriveremo in dettaglio nel prossimo paragrafo, è il **Modello stocastico a blocchi**. In questo caso, l'ipotesi è che la probabilità di connessione tramite un arco dei nodi i e j dipenda esclusivamente dalla comunità di appartenenza dei due nodi ed è dunque uguale per ogni coppia di nodi appartenenti alle stesse comunità di i e j.

Una volta definito il concetto di comunità, è opportuno evidenziare i problemi legati all'individuazione della partizione ottima di una rete. Lo spazio $\mathcal{P}_N$ delle partizioni di un insieme di N nodi ha cardinalità pari a B_N, dove B_N è il N-esimo numero di Bell (p.e. $B_{10} = 21147$, $B_{20} > 5 \times 10^{12}$). Il numero delle partizioni da valutare, enorme già per valori di N contenuti, obbliga a utilizzare metodi efficienti per individuare la migliore partizione.
I metodi più diffusi per questo scopo prevedono l'ottimizzazione di una opportuna funzione obiettivo, costruita in modo da essere espressione della qualità della partizione. La funzione che più comunemente viene utilizzata per valutare la qualità di una partizione è la *modularità*, così definita per reti

non dirette e non pesate:

$$Q = \frac{1}{2L} \sum_{i,j} \left[A_{ij} - P_{ij} \right] \delta(c_i, c_j). \tag{4.1}$$

dove A_{ij} è l'elemento ij della matrice di adiacenza della rete, L il numero di archi complessivo. La funzione $\delta(c_i, c_j)$ è la funzione indicatrice delle comunità dei nodi: vale 1 se $c_i = c_j$, cioè se i nodi i e j sono assegnati alla stessa comunità, 0 altrimenti. Il termine P_{ij} è il cosidetto *modello nullo*: indica la matrice di adiacenza media calcolata su un campione di reti ottenute randomizzando la rete originale in modo da preservare determinate caratteristiche. Una scelta comune del modello nullo è $P_{ij} = \frac{k_i k_j}{2L}$, con k_i e k_j pari al grado dei nodi i e j, e corrisponde alla probabilità che esista l'arco tra i nodi i e j se gli archi fossero randomizzati in modo da preservare, nella media di tutti i campioni random delle reti, il grado di tutti i nodi. Da questa scelta deriva la definizione classica di modularità:

$$Q = \frac{1}{2L} \sum_{i,j} \left[A_{ij} - \frac{k_i k_j}{2L} \right] \delta(c_i, c_j). \tag{4.2}$$

Questa definizione può essere estesa, con semplici modifiche, alle reti pesate, come vedremo nella sezione successiva. Il valore di modularità calcolato per una determinata partizione di una rete quantifica la differenza tra la frazione di archi interni alle comunità e il valore atteso della frazione di tali archi, cioè la frazione di archi interni alla comunità in questione che si troverebbero nella rete data dal modello nullo. Si noti che il valore atteso dell'arco $\{i, j\}$ è dato dal prodotto dei gradi dei nodi i e j diviso per il doppio del numero di archi complessivi della rete. In altri termini, la modularità esprime il risultato del confronto tra la partizione applicata alla rete di interesse e la stessa partizione applicata a una rete random con la stessa distribuzione di grado dei nodi. Valori di modularità alti sono pertanto indicatori di una partizione buona e di significatività delle comunità individuate: il numero di archi effettivamente presenti all'interno dei gruppi è più alto rispetto al valore atteso di tali archi in una rete random.
Come vedremo nei prossimi paragrafi, l'algoritmo di Louvain è basato sulla ricerca della partizione che fornisce la massima modularità della rete. Nel modello stocastico a blocchi, invece, si definisce e si ottimizza una diversa funzione obiettivo, che rappresenta la verosimiglianza di osservare una certa struttura a blocchi nella rete.

4.2 Metodo di Louvain

Il metodo di Louvain è un metodo semplice, efficiente e facile da implementare per identificare comunità nelle reti. Il metodo, sviluppato da V. Blondel, J-L. Guillame e R. Lambiotte presso l'Università di Louvain e pubblicato nel 2008 in [12], è uno dei più diffusi per il clustering su reti di grandi dimensioni e il suo successso è dovuto alla sua efficienza e alla possibilità che offre di identificare comunità in modo gerarchico.

Il metodo consiste in un'ottimizzazione "greedy" che ha lo scopo di massimizzare la modularità della partizione di una rete. La procedura è adatta all'utilizzo su reti unipartite, che possono essere anche pesate.

La modularità funge da funzione obiettivo da ottimizzare, oltre che da indicatore di valutazione della qualità della partizione ottenuta ed è definita, nel caso di reti pesate non dirette, dalla formula:

$$Q = \frac{1}{2m} \sum_{i,j} \left[w_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j). \tag{4.3}$$

dove w_{ij} rappresenta il peso dell'arco tra il nodo i e il nodo j, $k_i = \sum_j w_{ij}$ è la somma dei pesi degli archi incidenti nel nodo i, c_i è la comunità a cui il vertice i è assegnato e $m = \frac{1}{2} \sum_{ij} w_{ij}$ è il somma dei pesi degli archi nella rete. La funzione $\delta(c_i, c_j)$ è la funzione indicatrice delle comunità dei nodi. Si noti che la formula (4.3) è l'estensione per reti pesate della modularità definita in (4.2).
L'algoritmo è diviso in due fasi che vengono ripetute iterativamente fino a raggiungere il valore massimo di modularità.

Prima fase La prima fase è composta dai seguenti step:

- Step 0) Inizializzazione: ognuno degli N nodi viene assegnato a una comunità diversa. L'algoritmo inizia pertanto con una partizione in N comunità, una per ogni nodo.

- Step k) Per ogni nodo i si considera l'insieme dei suoi nodi adiacenti j e si valuta il guadagno di modularità ΔQ_{ij} che avrebbe luogo se si rimuovesse il nodo i dalla comunità in cui si trova e lo si assegnasse alla comunità del nodo j. Il guadagno viene valutato per ogni nodo j nell'insieme dei vicini del nodo i.

- Step $k+1$) Si sposta il nodo i nella comunità per cui il guadagno di modularità è massimo se e solo se il guadagno è positivo (gestendo

l'eventuale caso di parità con una regola opportuna) e si ritorna allo step k. Se non è possibile ottenere nessun guadagno positivo, il nodo rimane nella comunità corrente. Si valuta lo spostamento ottimale per ogni nodo i e la procedura viene ripetuta iterativamente finchè non è possibile ottenere miglioramenti nel valore di modularità e si raggiunge quindi un massimo locale.

Si precisa che in questa fase ciascun nodo può essere considerato e spostato più volte. La procedura che consente di raggiungere il massimo locale che segna la fine della prima fase dell'algoritmo risulta essere fortemente condizionata, per quanto riguarda il valore ottimo ottenibile e il tempo di esecuzione, dall'ordine con cui vengono considerati i nodi. L'efficienza dell'algoritmo anche su reti di grandi dimensione è tuttavia garantita dall'espressione del guadagno di modularità ΔQ_{ij} che risulta facilmente calcolabile in modo esplicito. Fissato il nodo i e preso in considerazione il nodo j appartente all'insieme dei nodi adiacenti a i, il guadagno di modularità ΔQ_{ij} ottenibile spostando i nella comunità C del nodo j è esprimibile come:

$$\Delta Q_{ij} = Q_{new} - Q_{old} \tag{4.4}$$

dove

$$Q_{new} = \frac{\sum_{in} + k_{i,in}}{2m} - \left(\frac{\sum_{tot} + k_i}{2m}\right)^2 \tag{4.5}$$

$$Q_{old} = \frac{\sum_{in}}{2m} - \left(\frac{\sum_{tot}}{2m}\right)^2 - \left(\frac{k_i}{2m}\right)^2 \tag{4.6}$$

Q_{new} rappresenta la modularità della partizione in cui il nodo i è assegnato alla nuova comunità C, Q_{old} il valore di modularità prima dell'eventuale spostamento di i. Nella formula riportata $\sum_{in}$ è la somma dei pesi degli archi all'interno della comunità C, $\sum_{tot}$ è la somma dei pesi degli archi incidenti a nodi della comunità C, k_i è la somma dei pesi degli archi incidenti al nodo i, $k_{i,in}$ è la somma dei pesi degli archi che collegano i ai nodi nella comunità C e m è la somma dei pesi degli archi della rete.

Seconda fase La seconda fase dell'algoritmo consiste nella costruzione di una nuova rete in cui i nodi, che indicheremo come "super-nodi", sono costituiti dalle comunità individuate nella prima fase. Gli archi presenti tra i nodi della nuova rete sono dati dai collegamenti tra i nodi assegnati a comunità

diverse e i loro pesi sono determinati dalla somma dei pesi degli archi tra le comunità considerate. Gli archi tra nodi assegnati alla stessa comunità costituiscono, invece, self-loop per i nodi della nuova rete, cioè archi che connettono il nodo con se stesso. Una volta costruita la nuova rete aggregata, si procede con la riapplicazione della prima fase, iterando fino al raggiungimento di un nuovo massimo locale di modularità.

La combinazione della prima fase di ricerca di massima modularità e della seconda fase di aggregazione costituisce un passo dell'algoritmo, ripetuto iterativamente finchè viene raggiunto il massimo valore di modularità.

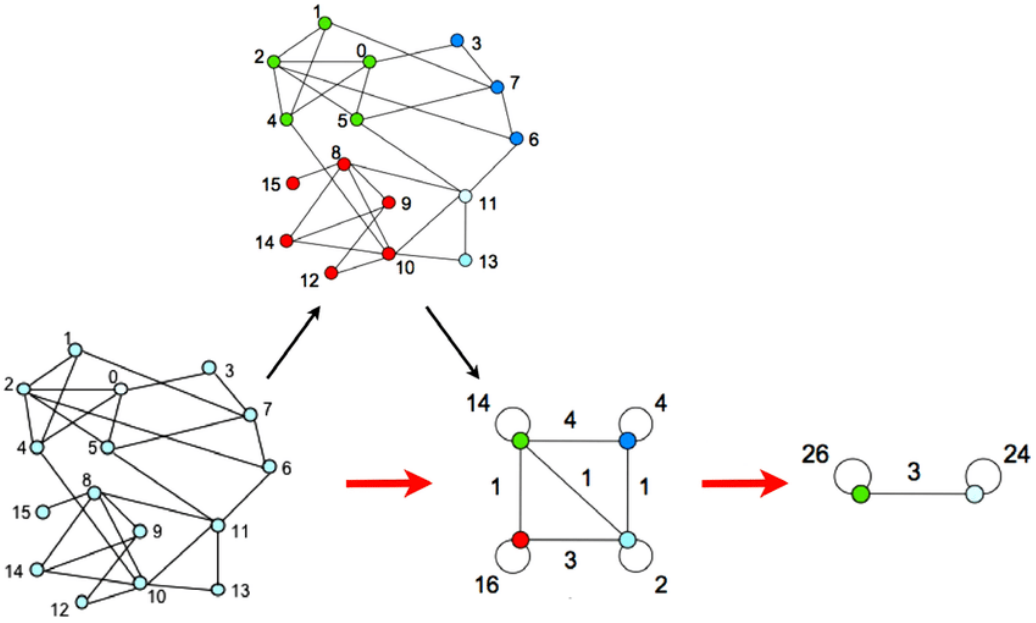

Figura 4.2: Visualizzazione della procedura dell'algoritmo di Louvain (da [12]).

In Figura 4.2 sono mostrati gli step dell'algoritmo su una rete esempio. Da sinistra verso destra si hanno il primo step di ricerca di massima modularità locale tramite lo spostamento dei singoli nodi e il secondo di aggregazione dei nodi assegnati alle stesse comunità in "super-nodi" (primo passo dell'algoritmo), seguiti dal secondo passo, in cui la procedura del primo passo è applicata alla rete costituita dai "super-nodi" individuati. Si può notare che i pesi degli archi delle reti aggregate rappresentano la somma delle connessioni all'interno della comunità, se riferiti ai self-loop, e la somma delle connessioni tra le comunità, se riferiti a archi tra nodi diversi.

Il costo computazionale dell'algoritmo è di $O(Nlog(N))$, con N numero di nodi delle rete, se viene implementata una buona inizializzazione e un ordine di analisi dei nodi opportuno. É da notare, inoltre, che il maggior costo computazionale richiesto per i passi dell'algoritmo decresce con l'avanzamento della procedura che, grazie alla aggregazione iterativa in "super-nodi", consente di gestire un numero di nodi sempre minore, con evidenti vantaggi dal punto di vista dell'esecuzione.

La necessità di utilizzare un algoritmo che garantisca buone performance in termini di tempi di esecuzione è dovuta alla sempre più grande dimensione delle reti studiate oggi. Gli autori in [12] hanno testato l'algoritmo sulla rete degli utenti di telefoni cellulari in Belgio formata da 2.6 milioni di nodi, ottenendo buoni risultati in termini di tempo di esecuzione e di bontà della partizione. In Tabella 4.1 citiamo i risultati presentati in [12] sulle reti testate dagli autori.

Rete considerata	Karate Club	Pagine Web	Cellulari in Belgio
Nodi / Archi	34/77	70K/351K	2.6 M/6.3 M
Tempo di esecuzione	0 s	1s	134 s
Modularità	0,42	0,781	0,769

Tabella 4.1: Performance dell'algoritmo di Louvain su reti test di diversa grandezza, si veda [12] per dettagli.

Come verrà esposto nel Capitolo 5, l'algoritmo di Louvain fornisce buoni risultati in termini di modularità e tempi di esecuzione molto rapidi anche sulle reti analizzate in questo lavoro.
In rete sono disponibili un'implementazione del metodo in linguaggio C++, ottenibile da [2], e una versione Matlab. Nel nostro lavoro abbiamo utilizzato l'implementazione della funzione `cluster_louvain` nel pacchetto *igraph* ([15]) del software R. In Appendice C è riportata la documentazione relativa a questa funzione.

4.3 Modello Stocastico a Blocchi

Il modello stocastico a blocchi, come anticipato in precedenza, può essere utilizzato come metodo per identificare la struttura di comunità nelle reti. Presenteremo in primo luogo il metodo standard e la versione degree-

corrected introdotta in [21] da B. Karrer e M. Newman dal punto di vista teorico, per poi illustrare il procedimento algoritmico in dettaglio. L'implementazione in linguaggio C++ fornita dagli autori in [3] e da noi utilizzata nel lavoro è disponibile online in [3].
Nel più semplice modello stocastico a blocchi, abbreviato in seguito con SBM (Stochastic block-model), ogni nodo i di una rete costituita da N nodi viene assegnato a una delle K comunità (o "blocchi") supponendo che la probabilità di esistenza di un arco tra una coppia di nodi sia indipendente dalle altre coppie di nodi e funzione solo del blocco di appartenenza dei nodi considerati. Una volta assegnati i nodi a uno dei K blocchi e avendo quindi costruito il vettore g_i degli assegnamenti, si definisce la matrice ψ delle probabilità di connessione di dimensione $K \times K$, i cui elementi $\psi_{g_i g_j}$ sono le probabilità indipendenti di connessione tra i nodi i e j. Come si può notare dall'espressione di ψ, la probabilità dell'arco $\{i, j\}$ dipende solo dal gruppo a cui appartengono i e j: si può dire quindi che i nodi appartenenti alla stessa comunità siano statisticamente equivalenti dal punto di vista della struttura della rete. Il modello SBM può quindi essere visto come una generalizzazione "a blocchi" del modello generativo di Erdös-Rènyi, che produce reti random.
Prima di introdurre il modello stocastico a blocchi standard precisiamo che le reti utilizzate per la formulazione del metodo sono non dirette e non pesate. Il modello presentato in [21] consente di gestire, inoltre, i multigrafi, cioè le reti che contengono archi multipli o archi che connettono un nodo a se stesso (i cosidetti self-loop). Se si utilizzano reti con archi multipli, la matrice ψ delle probabilità di archi tra nodi appartenenti a determinate comunità deve essere sostituita dalla matrice dei valori attesi del numero di questi archi: l'elemento ψ_{rs} rappresenta il numero atteso di archi tra coppie di nodi appartenenti a r e a s. Il numero effettivo di archi presenti tra la coppia di nodi delle date comunità è estratto da una distribuzione di Poisson con media ψ_{rs}. Si tenga conto che le differenze di interpretazione tra l'estensione del modello per multigrafi e il modello SBM base per reti semplici risultano del tutto trascurabili se si considerano reti sparse e di grande dimensione, in quanto la probabilità di un arco e il valore atteso del numero di archi tendono in questi casi a coincidere.
L'estensione del modello a reti con archi multipli (multigrafi) consente di avere un modello consistente con la versione modificata "degree-corrected", che verrà esposta nel paragrafo seguente, e di gestire anche reti con pesi degli archi interi. È possibile, infatti, costruire una rete multigrafo a partire da una rete semplice pesata, connettendo due nodi con un numero di archi pari al peso dell'arco corrispondente nella rete pesata. Come vedremo nel capitolo seguente, questo è stato il metodo utilizzato per analizzare con i modelli a blocchi stocastici la rete oggetto del lavoro, inizialmente dotata di pesi interi.

Nel caso di reti pesate con pesi reali non è possibile riferirsi a un modello multigrafo e agli SBM standard, tuttavia sono presenti in letteratura delle estensioni del modello stocastico a blocchi adatte a reti pesate. Approfondimenti di questo argomento si possono trovare in [8] e [9].
Da ultimo, ci soffermiamo sulla notazione utilizzata nell'esposizione del metodo. La matrice di adiacenza A di una rete con archi multipli è definita i in modo tale che i suoi elementi A_{ij} siano pari al numero di archi tra i nodi i e j e gli elementi diagonali A_{ii} siano invece pari a due volte il numero di self-loop del nodo i.

Il modello stocastico a blocchi standard Il modello stocastico a blocchi standard si basa su una rete non diretta con archi multipli, che denotiamo con G, con matrice di adiacenza A. Ipotizziamo, come già detto, che il numero di archi tra ogni coppia di nodi sia distribuito in modo indipendente come una Poisson di parametro ω_{rs}, che rappresenta il valore atteso dell'elemento A_{ij} della matrice di adiacenza per nodi appartenenti rispettivamente alle comunità r e s. Il valore atteso di self-loop per i nodi della comunità r è invece, per la definizione data sopra degli elementi diagonali di A, pari a $1/2\ \omega_{rr}$. Una volta impostato questo modello, l'espessione della probabilità di osservare la rete G, dati i parametri e gli assegnamenti dei nodi ai blocchi, è:

$$\begin{aligned} P(G \mid \omega, g) = &\prod_{i<j} \frac{(\omega_{g_i g_j})^{A_{ij}}}{A_{ij}!} \exp(-\omega_{g_i g_j}) \\ &\times \prod_i \frac{(\frac{1}{2}\omega_{g_i g_i})^{A_{ii}/2}}{(A_{ii}/2)!} \exp(\frac{1}{2}\omega_{g_i g_i}) \end{aligned} \tag{4.7}$$

Sfruttando la simmetria della matrice di adiacenza e della matrice ω e introducendo i parametri n_r, numero di nodi nella comunità r, e m_{rs}, numero totale di archi tra i blocchi r e s, si ottiene la seguente formula semplificata dell'espressione (4.8):

$$\begin{aligned} P(G \mid \omega, g) = &\frac{1}{\prod_{i<j} A_{ij}! \prod_i 2^{A_{ii}/2}(A_{ii}/2)!} \\ &\times \prod_{rs} \omega_{rs}^{m_{rs}/2}\ \exp(-\frac{1}{2} n_r n_s\ \omega_{rs}) \end{aligned} \tag{4.8}$$

L'obiettivo del metodo SBM è quello di massimizzare l'espressione (4.7) o (4.8) rispetto ai parametri ω_{rs} del modello, da stimare in modo opportuno,

e rispetto alle possibili partizioni della rete. Il primo passaggio da effettuare è ottenere il logaritmo dell'espressione (4.7), che consente di gestire, una volta semplificate le costanti e trascurati i termini indipendenti dai parametri e dagli assegnamenti ai blocchi, la seguente espressione, più compatta e interpretabile:

$$logP(G \mid \omega, g) = \sum_{rs}(m_{rs}\ log\ \omega_{rs} - n_r n_s \omega_{rs}) \tag{4.9}$$

La massimizzazione dell'espressione della log-verosimiglianza di G viene ottenuta in due passaggi. Nel primo si calcolano le stime di massima verosimiglianza (ML) dei parametri ω, cioè i valori ω_{rs} che garantiscono, fissato un certo assegnamento, il valore ottimo di $logP$. Con un semplice calcolo si ricavano le stime ML per ogni elemento della matrice ω:

$$\hat{\omega}_{rs} = \frac{m_{rs}}{n_r n_s} \tag{4.10}$$

Sostituendo tali stime nell'espressione (4.9) si ottiene un'espressione della verosimiglianza di G dipendente solo dalla partizione.

$$\mathcal{L}(G \mid g) = \sum_{rs}\ m_{rs}\ log\frac{m_{rs}}{n_r n_s} \tag{4.11}$$

In (4.11) abbiamo dunque ottenuto la funzione obiettivo del metodo: la funzione assume valore ottimo in corrispondenza della partizione più verosimile e, più in generale, assume valori alti per assegnamenti "buoni" dei nodi e valori bassi per assegnamenti poco significativi. Nel paragrafo ad esso dedicato, illustreremo l'algoritmo che consente di individuare la partizione ottimale. L'aspetto importante da considerare rispetto alla natura di questa funzione obiettivo è che viene costruita a partire dalla struttura stessa della rete e, essendo basata sulla distribuzione di probabilità degli archi, ha anche un'interpretazione di tipo statistico.

Il modello stocastico a blocchi Degree-Corrected Il modello a blocchi standard descritto nel paragrafo precedente risulta in molti casi inadatto a descrivere le strutture di comunità presenti nelle reti reali. Il limite principale all'applicazione del modello è la sua scarsa flessibilità nel generare e nell'individuare strutture caratterizzate da eterogeneità nella distribuzione dei gradi dei nodi. In [21] viene proposta un'estensione, semplice ma di grande efficacia, al metodo standard che consente di tenere in considerazione nel modello la caratteristica di eterogeneità dei gradi, tipica di molte reti reali, e di ottenere partizioni di qualità migliore.

Nel modello degree-corrected viene introdotta una correzione nello schema standard: la distribuzione di probabilità sul multigrafo G dipende non solo dai parametri introdotti nel modello base, ma anche da un nuovo insieme di parametri θ che controllano il valore atteso del grado dei nodi.
I parametri θ introdotti sono arbitrari rispetto a una costante moltiplicativa che viene assorbita nei parametri della matrice ω. É possibile normalizzare i parametri fissando per ogni gruppo r il vincolo $\sum_i \theta_i \delta_{g_i,r} = 1$. Grazie a questa normalizzazione, è possibile interpretare il parametro θ riferito al nodo i (θ_i) come probabilità che un arco connesso alla comunità a cui il nodo i appartiene sia incidente al nodo i. Riprendendo lo schema del modello standard e basandoci su una partizione in K comunità di una rete di N nodi, introduciamo la matrice simmetrica ω_{rs}, di dimensione $K \times K$, contenente i parametri di controllo degli archi tra le comunità r e s e il vettore g, di dimensione N, delle comunità assegnate ai nodi. Il numero di archi tra i nodi è estratto da una distribuzione di Poisson, come nel modello non-degree-corrected, ma in questo caso il valore atteso di ciascun arco $\{i, j\}$ tiene conto del nuovo parametro θ ed è posto uguale a $\theta_i \theta_j \omega_{g_i g_j}$. Si sottolinea nuovamente che, in questo modello, il valore atteso dell'arco $\{i, j\}$ coincide con il parametro della distribuzione di Poisson da cui sono estratti gli archi e che è utile interpretare il valore atteso di ciascun arco $\{i, j\}$ (cioè il numero di connessioni tra i nodi i e j) come il valore atteso dell'elemento A_{ij} della matrice di adiacenza della rete.
La probabilità di osservare la rete G è dunque, secondo questo modello :

$$\begin{aligned} P(G \mid \theta, \omega, g) &= \prod_{i<j} \frac{(\theta_i \theta_j \omega_{g_i g_j})^{A_{ij}}}{A_{ij}!} \exp(-\theta_i \theta_j \omega_{g_i g_j}) \\ &= \prod_i \frac{(\frac{1}{2}\theta_i^2 \omega_{g_i g_i})^{A_{ii}/2}}{(A_{ii}/2)!} \exp(-\frac{1}{2}\theta_i^2 \omega_{g_i g_i}) \end{aligned} \tag{4.12}$$

La costante di normalizzazione introdotta consente poi di semplificare l'espressione nel modo seguente:

$$\begin{aligned} P(G \mid \theta, \omega, g) = &\frac{1}{\prod_{i<j} A_{ij}! \prod_i 2^{A_{ii}/2}(A_{ii}/2)!} \\ &\times \prod_i \theta_i^{k_i} \prod_{rs} \omega_{rs}^{m_{rs}/2} \exp(-\frac{1}{2}\omega_{rs}) \end{aligned} \tag{4.13}$$

dove k_i indica il grado del nodo i e m_{rs} il numero totale di archi tra la comunità r e s. Anche in questo caso, la massimizzazione della probabilità di osservare la rete G è più facilmente ottenibile impostando la massimizzazione

del suo logaritmo, esprimibile, trascurando i termini costanti, come:

$$logP(G \mid \theta, \omega, g) = 2\sum_i k_i \ log \ \theta_i + \sum_{rs}(m_{rs} log \ \omega_{rs} - \omega_{rs}) \tag{4.14}$$

Seguendo il procedimento illustrato nel caso standard, la massimizzazione del valore nell'equazione (4.14) si ottiene calcolando le stime di massima verosimiglianza dei parametri θ_i e ω_{rs}, pari a

$$\hat{\theta}_i = \frac{k_i}{\kappa_{g_i}}, \qquad \hat{\omega}_{rs} = m_{rs} \tag{4.15}$$

e, in un secondo passaggio, la partizione che fornisce la massima verosimiglianza della funzione obiettivo P. In (4.15) si introduce il parametro κ_{g_i} pari alla somma dei gradi nella comunità g a cui appartiene il nodo i.

La proprietà importante delle stime di massima verosimiglianza dei parametri è quella di preservare il numero atteso di archi tra i gruppi e, allo stesso tempo, anche il grado atteso per la sequenza dei nodi. Nel modello standard quest'ultima caratteristica non viene conservata e si ottiene che tutti i nodi assegnati alla stessa comunità r hanno lo stesso grado atteso, pari a $\frac{\kappa_r}{n_r}$, con κ_r pari alla somma dei gradi nella comunità r e n_r pari al numero di nodi assegnati a r.

La dimostrazione di questa proprietà è riportata di seguito: nell'espressione (4.16) si ha il numero atteso di archi tra i blocchi r e s e in (4.17) si ha il valore atteso del grado del nodo i. Si noti che con $\langle x \rangle$ si indica il valore medio di x nell'insieme delle reti con i parametri dati dalle stime di massima verosimiglianza calcolate in (4.15).

$$\sum_{ij}\langle A_{ij}\rangle \delta_{g_i,r}\delta_{g_j,s} = \sum_{ij}\frac{k_i k_j \ m_{g_i g_j}}{\kappa_{g_i}\kappa_{g_j}} = m_{rs} \tag{4.16}$$

$$\begin{aligned}\sum_j \langle A_{ij}\rangle &= \sum_j \hat{\theta}_i \ \hat{\theta}_j \ \hat{\omega}_{g_i g_j} = \frac{k_i}{\kappa_{g_i}}\sum_j \frac{k_j}{\kappa_{g_j}} m_{g_i g_j} \\ &= \frac{k_i}{\kappa_{g_i}}\sum_j\sum_r \frac{k_j}{\kappa_r} m_{g_i,r}\delta_{g_j,r} \\ &= \frac{k_i}{\kappa_{g_i}}\sum_r m_{g_i,r} = k_i\end{aligned} \tag{4.17}$$

Sostituendo i valori di massima verosimiglianza in (4.14) si ottiene una nuova espessione di $logP(G|\theta, \omega, g)$:

$$logP(G \mid \theta, \omega, g) = 2\sum_i k_i \ log\frac{k_i}{\kappa_{g_i}} + \sum_{rs} m_{rs} \ log \ m_{rs} - 2m \tag{4.18}$$

Con le opportune sostituzioni e le semplificazioni dei termini costanti, da (4.14) si ottiene la funzione obiettivo (4.19), dipendente solo dalla partizione g.

$$\mathcal{L}(G \mid g) = \sum_{rs} m_{rs}\ log\frac{m_{rs}}{\kappa_r\ \kappa_s} \tag{4.19}$$

L'algoritmo ricerca iterativamente la partizione della rete che fornisce il valore massimo di questa espressione.
Considerando la formula (4.19) si può notare che l'unica differenza con la funzione obiettivo (4.11) del modello standard è costituita dalla sostituzione del termine n_r, numero totale dei nodi nella comunità r, con κ_r, somma dei gradi nella comunità r. La semplicità di questa modifica consente un facile confronto con il modello standard sia in termini dei risultati ottenibili applicando le due versioni, sia dal punto di vista interpretativo.

Algoritmo di ottimizzazione Una volta indivuata un'espressione per la log-verosimiglianza delle reti in esame secondo i modelli standard (4.11) e degree-corrected (4.19), è necessario descrivere il procedimento che consente di individuare la partizione ottimale. La struttura delle espressioni individuate consente di calcolare in modo efficiente le modifiche del valore di log-verosimiglianza causate dallo spostamento, nel corso della ricerca dell'assegnamento ottimo, dei nodi da un blocco a un altro. In particolare, si può notare che quando un nodo si sposta dal blocco r al blocco s gli unici valori dell'espressione della log-verosimiglianza che vengono coinvolti sono κ_r, κ_s, m_{rt} e m_{st} per ogni comunità t. Nel calcolo della variazione di log-verosimiglianza $\Delta\mathcal{L}$ molti termini possono quindi essere trascurati. L'espressione di $\Delta\mathcal{L}$ per lo spostamento del nodo i dal blocco r al blocco s risulta pari a:

$$\begin{aligned}\Delta\mathcal{L} = \sum_{t\neq r,s}[a(m_{rt}-k_{it}) - a(m_{rt}) + a(m_{st}+k_{it}) - a(m_{st})]+ \\ + a(m_{rs}+k_{ir}-k_{is}) - a(m_{rs}) \\ + b[m_{rr}-2(k_{ir}+u_i)] - b(m_{rr}) \\ + b[m_{ss}+2(k_{is}+u_i)] - b(m_{ss}) \\ - a(\kappa_r - k_i) + a(\kappa_r) \\ - a(\kappa_s + k_i) + a(\kappa_s)\end{aligned} \tag{4.20}$$

dove si sono introdotte le notazioni $a(x) = 2x\log x$ e $b(x) = x\ logx$ con la convenzione $a(0) = 0$ e $b(0) = 0$. Si noti inoltre che, per ogni possibile

spostamento di i, si conservano le quantità k_{it}, numero degli archi da i ai nodi nel gruppo t (esclusi i self-loop), e u_i, numero di self-loop del nodo i. La valutazione di questa quantità può essere eseguita, se il numero di blocchi K è contenuto, in modo rapido ed è quindi possibile calcolare, per ogni nodo, il $\Delta\mathcal{L}$ corrispondente al suo spostamento in una nuova comunità. In particolare, il tempo di calcolo di $\Delta\mathcal{L}$ per uno spostamento di un singolo nodo è dell'ordine di $O(K + \langle k \rangle)$, dove $\langle k \rangle$ è il grado medio della rete. Per individuare la comunità ottima in cui spostare il singolo nodo i è quindi necessario un tempo $O(K(K + \langle k \rangle)$ e di $O(N\ K(K + \langle k \rangle)$ per individuare la partizione ottima per tutti i nodi. Riportiamo di seguito i passi dell'algoritmo greedy di ricerca locale utilizzato in [21]:

- Step 0) Inizializzazione: ogni nodo viene assegnato a una delle K comunità iniziali in modo random.
- Step 1) Per ogni nodo e ogni blocco si calcola il valore $\Delta\mathcal{L}$, selezionando lo spostamento che garantisce il maggiore incremento, o, se non è possibile un incremento, il minore decremento, della funzione obiettivo. In questa fase, una volta che un nodo viene spostato, non può più essere soggetto a spostamenti.
- Step 2) Una volta spostati una e una sola volta tutti i nodi, si valutano tutti gli stati attraversati dal sistema nello step 1, cioè tutte le partizioni individuate con gli spostamenti successivi dei nodi, e si seleziona lo stato che fornisce il valore maggiore della funzione obiettivo;
- Step 3) Si utilizza la partizione selezionata come nuova partizione di partenza per una nuova iterazione dello step 1, registrando il valore ottimo ottenuto.
- Fine: una volta che si attraversano gli step 1, 2 e 3 senza registrare un incremento della funzione obiettivo, l'algoritmo termina: si ottiene il valore ottimo della log-verosimiglianza della rete e ogni nodo risulta assegnato al blocco della partizione ottima.

Un aspetto critico della procedura è costituito dall'inizializzazione random: come segnalano gli autori, è utile avviare l'algoritmo più volte con diverse partizioni random iniziali e considerare il risultato migliore tra i diversi run. In questo modo si minimizza la probabilità di ottenere risultati condizionati dall'assegnazione iniziale.

Un altro limite significativo è costituito dalla necessità di indicare come input dell'algoritmo il numero di blocchi in cui partizionare la rete. In molti casi il valore di K più opportuno non è noto a priori ed è, anzi, spesso il primo valore di interesse nell'analisi di una rete.

Risultati e esempi di applicazione Presentiamo ora un esempio di applicazione del metodo SBM per evidenziare le differenze di performance su una rete reale delle due versioni dell'algoritmo. In [21] vengono illustrate le partizioni ottenute, riportate in Figura 4.3, applicando il metodo SBM standard e il metodo SBM Degree corrected sulla rete del club di karatè di Zachary, introdotta nel Capitolo 2, imponendo la suddivisione dei 34 nodi in 2 comunità. Come si può notare dalla figura, il metodo standard fornisce una partizione che suddivide i nodi in due gruppi: i nodi di grado più alto (in blu) e i nodi periferici (in giallo). La partizione corretta, rappresentata in figura dalla linea tratteggiata, è invece ricostruita fedelmente (a eccezione di un solo nodo) dal metodo degree-corrected. La partizione non corretta data dal metodo standard è dovuta al fatto che l'algoritmo non tiene conto della sequenza dei gradi dei nodi e per questo la partizione che suddivide i nodi in base al loro grado risulta essere ottima. La probabilità di esistenza di un arco tra nodi di grado alto risulta, infatti, molto alta (è pari al prodotto dei gradi dei nodi) e dunque l'algoritmo assegna questi nodi alla stessa comunità. Analogamente, i nodi di grado basso sono assegnati ad una stessa comunità perchè scarsamente connessi. Il metodo degree-corrected, includendo nella formula della verosimiglianza anche il grado dei nodi, consente di individuare la partizione che rispecchia in modo corretto la struttura della rete.

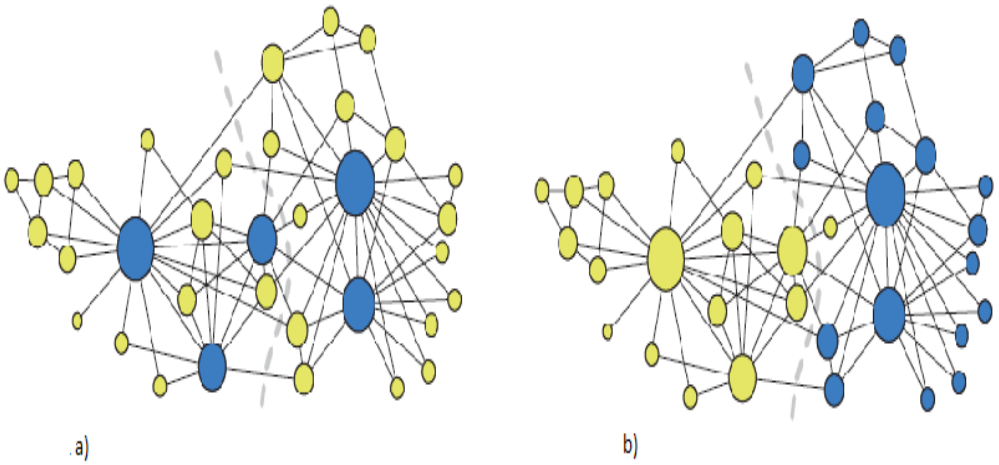

Figura 4.3: Metodo SBM standard (a) vs Metodo SBM Degree-Corrected (b) sulla rete Karate Club (da [21]).

4.4 Valutazione delle partizioni

Dopo aver ottenuto la partizione dei nodi della rete con un algoritmo di community detection, è utile valutare la qualità del risultato. Un limite degli algoritmi di ottimizzazione è, infatti, la loro capacità di individuare in ogni rete una partizione di massima modularità o che, in generale, fornisca il valore ottimo di una data funzione obiettivo. É possibile, allora, che per favorire l'ottimizzazione globale metodi di questo tipo forzino partizioni dei nodi che possono risultare difficilmente interpretabili e che non mettano in evidenza in una comunità nodi particolarmente connessi. Un altro problema che tipicamente si riscontra nell'analisi di reti grandi è il limite di risoluzione: l'algoritmo non è in grado di individuare comunità, benchè significative, che abbiano cardinalità al di sotto di una certa soglia. Si rende necessaria, dunque, una verifica a posteriori che valuti la significatività della partizione e la qualità delle singole comunità ottenute.
Nei prossimi paragrafi descriveremo i metodi utilizzati nel lavoro per valutare la qualità delle comunità individuate (probabilità di persistenza), e confrontare diverse partizioni (matrici di precision e recall e indici di similarità).
Si precisa che il confronto può avvenire tra una partizione fissata e basata su caratteristiche note dei nodi e una partizione indotta da un algoritmo, o tra partizioni ottenute con due metodi diversi. Nel primo caso il confronto ha come scopo la valutazione della capacità dell'algoritmo di ricostruire la struttura nota della rete, nel secondo è possibile mettere in evidenza eventuali differenze di performance e di risultati a supporto della scelta del metodo più efficace e adatto alla rete in analisi.

Probabilità di Persistenza Data una partizione dei nodi in K comunità, è possibile valutare la loro coesione calcolando la *probabilità di persistenza* α. In una rete non diretta pesata, la probabilità di persistenza della comunità C_k è definita dalla seguente espressione:

$$\alpha_k = \frac{\sum_{i \in C_k} \sum_{j \in C_k} w_{ij}}{\sum_{i \in C_k} \sum_{j \in \{1,2,\dots K\}} w_{ij}} \tag{4.21}$$

La probabilità di persistenza è, dunque, pari alla frazione di strength dei nodi di C_k che rimane all'interno di C_k. In letteratura questa quantità è indicata anche come *embeddedness*. Il termine probabilità di persistenza deriva dal fatto che α_k coincide con la probabilità che un random walker che è in C_k al passo t sia ancora in C_k al passo $t+1$ [28].

Nel nostro lavoro seguiremo l'approccio utilizzato in [14], dove si valuta la probabilità di persistenza per la rete, non diretta e pesata, dei partecipanti a organizzazioni criminali. Per ogni comunità individuata viene calcolato il valore di α_k: tale comunità è tanto più coesa quanto più è grande la sua probabilità di persistenza. In questo contesto, un raggruppamento di nodi si può definire comunità se la probabilità di persistenza risulta $> 0,5$. É da sottolineare il fatto che il valore di α_k tende a aumentare con la cardinalità N_k della comunità C_k, fino a risultare pari a 1 se si considera la partizione (banale) che assegna tutti i nodi della rete alla stessa comunità. É opportuno valutare, per questo, oltre al valore della probabilità di persistenza della comunità C_k, la sua significatività dal punto di vista statistico mediante lo *z-score* così definito:

$$z_k = \frac{\alpha_k - \mu(\bar{\alpha}_k)}{\sigma(\bar{\alpha}_k)} \tag{4.22}$$

dove α_k è la probabilità di persistenza della comunità C_k e $\bar{\alpha}_k$ è la probabilità di persistenza empirica di una sottorete connessa di cardinalità N_k, pari a quella di C_k. I termini $\mu(\bar{\alpha}_k)$ e $\sigma(\bar{\alpha}_k)$ rappresentano rispettivamente la media e la deviazione standard delle probabilità di persistenza empiriche calcolate su un campione casuale di sottoreti connesse di cardinalità N_k. Una comunità che presenta valori alti di α_k e di z risulta, quindi, coesa non per motivazioni legate alla sua cardinalità, ma in quanto presenta valori di probabilità di persistenza significativamente più grandi rispetto a altre sottoreti di pari cardinalità. Un valore di z che, tipicamente, è da ritenere significativamente alto è 3.

Indici di similarità delle partizioni Date due diverse partizioni $\mathcal{X} = \{X_1, X_2 \dots X_H\}$ e $\mathcal{Y} = \{Y_1, Y_2 \dots Y_K\}$ della stessa rete, rispettivamente in H e K comunità, è possibile valutare la loro similarità calcolando opportuni indici. Per un quadro completo sull'argomento rimandiamo a [17], mentre descriveremo di seguito in dettaglio le quantità utilizzate durante il lavoro. Introduciamo in primo luogo la *matrice di confusione* $N_{\mathcal{X}\mathcal{Y}}$, di dimensione $H \times K$, i cui elementi n_{hk} indicano il numero di nodi appartenenti alla comunità X_h nella partizione $\mathcal{X}$ e alla comunità Y_k nella partizione $\mathcal{Y}$.
Due indici di similarità utilizzati diffusamente in letteratura per valutare le performance degli algoritmi di community detection sono la *Informazione Mutua Normalizzata* (NMI, Normalized Mutual Information) e la *Variazione di Informazione* (VI). Entrambi gli indici si basano sul calcolo, data una partizione, della quantità di informazione aggiuntiva necessaria per inferire l'altra partizione. L'entità dell'informazione richiesta fornisce una misura della similarità o dissimilarità delle partizioni.

Per definire l'indice, si introducono gli assegnamenti $\{x_i\}$ e $\{y_i\}$, dove x_i e y_i rappresentano le comunità a cui il vertice i risulta assegnato rispettivamente nelle partizioni $\mathcal{X}$ e $\mathcal{Y}$, e si indicano con x e y i valori di due variabili aleatorie X e Y che rappresentano due partizioni della rete. É possibile a questo punto definire una distribuzione di probabilità congiunta $P(x,y) = P(X = x, Y = y) = n_{xy}/N$, con N numero totale dei nodi della rete e n_{xy} numero dei nodi appartenenti alle comunità x e y rispettivamente delle partizioni $\mathcal{X}$ e $\mathcal{Y}$. Le distribuzioni marginali di X e Y sono invece definite come $P(x) = P(X = x) = n_x^X/N$ e $P(y) = P(Y = y) = n_y^Y/N$: la probabilità della comunità x della partizione $\mathcal{X}$ è data dalla frazione di nodi assegnati alla comunità x. Da ultimo si introducono anche le quantità $H(X)$, detta *entropia di Shannon* della variabile X, pari a $-\sum_x P(x)\ logP(x)$, e l'*entropia condizionale* di X dato Y, pari a $H(X|Y) = -\sum_{x,y} P(x,y)\ logP(x|y)$.
La misura di similarità *Informazione mutua* di due variabili aleatorie è, infine, data dall'espressione:

$$I(X,Y) = H(X) - H(X|Y) \tag{4.23}$$

Il significato dell'informazione mutua è, intuitivamente, legato all'ammontare di informazione che la conoscenza di una delle due variabili, nel nostro caso, di una delle partizioni, fornisce riguardo all'altra. L'entropia $H(X)$ è, infatti, una misura dell'incertezza riguardo a X e $H(X|Y)$ quantifica l'incertezza riguardo a X nota la variabile Y. L'espressione di $I(X,Y)$ può essere dunque interpretata come l'ammontare di incertezza nella variabile X meno l'ammontare di incertezza in X che rimane data Y, equivalente all'ammontare di incertezza in X eliminata conoscendo Y.
L'informazione mutua, tuttavia, risulta poco adatta come misura di similarità tra partizioni: infatti, data una partizione $\mathcal{X}$, tutte le partizioni ottenute dividendo alcune delle comunità di $\mathcal{X}$ presentano, pur essendo partizioni diverse, lo stesso valore di informazione mutua, a causa del valore nullo dell'entropia condizionale. Per avere a disposizione una misura che tenga conto esplicitamente della eventuale dipendenza presente tra le partizioni, si introduce l'*Informazione mutua normalizzata* (NMI), così definita:

$$NMI(\mathcal{X},\mathcal{Y}) = \frac{2I(X,Y)}{H(X) + H(Y)} \tag{4.24}$$

L'Informazione mutua normalizzata può assumere valori tra 0 e 1. In particolare, la NMI è pari a 1 se e solo se le due partizioni sono identiche ed

è pari a 0 se solo totalmente indipendenti.

Un altro indice, introdotto in [23], è la *Variazione di Informazione* (VI), così definita:

$$V(\mathcal{X}, \mathcal{Y}) = H(X|Y) + H(Y|X) \tag{4.25}$$

dove $H(X|Y)$ e $H(Y|X)$ sono le entropie condizionate definite sopra. L'indice è costituito dalla somma di due termini che quantificano rispettivamente l'ammontare di informazione che si perde riguardo a X e l'ammontare di informazione che si guadagna riguardo a Y nel passare dalla partizione X alla partizione Y. L'importanza di questo indice è dato dal fatto che costituisce una metrica nello spazio delle partizioni e possiede le proprietà di una distanza (non negatività, simmetria e disuguaglianza triangolare). Sottolineiamo che due partizioni risultano tanto più simili quanto più l'indice Variazione di Informazione assume valori vicini a 0, cioè quanto più la distanza tra le due partizioni è piccola.

Matrici di Precision e Recall Un strumento immediato per valutare la similarità di due partizioni è dato inoltre dalle *matrici di Precision e Recall*. Date due partizioni $\mathcal{X}$ e $\mathcal{Y}$ della stessa rete costituite rispettivamente da H e K comunità, sia n_{hk} l'elemento $\{hk\}$ della matrice di confusione tra le due partizioni, pari al numero di nodi appartenenti alla comunità h nella prima partizione e alla comunità k nella seconda.
L'elemento p_{hk} della matrice di Precision, di dimensioni $H \times K$, è:

$$p_{hk} = \frac{n_{hk}}{|Y_k|} \tag{4.26}$$

con $|Y_k|$ cardinalità della comunità Y_k, ed è pari alla frazione di elementi di Y_k che appartengono a X_h.
L'elemento r_{hk} della matrice di Recall, anch'essa di dimensioni $H \times K$, è, invece:

$$r_{hk} = \frac{n_{hk}}{|X_h|} \tag{4.27}$$

con $|X_h|$ cardinalità della comunità X_h, ed è pari alla frazione di elementi di X_h che appartengono a Y_k.

Il metodo risulta efficace perchè consente un confronto a due a due tra le comunità individuate dalle due partizioni. In particolare, è interessante notare i casi in cui $p_{hk} = r_{hk} = 1$, dove si ha perfetta corrispondenza tra le

comunità X_h e Y_k. Risulta, invece, $p_{hk} \to 1$ se un numero grande di nodi della comunità Y_k appartiene anche a X_h, e $r_{hk} \to 1$ se, viceversa, un numero grande di nodi della comunità X_h appartiene anche a Y_k. Tipicamente la partizione $\mathcal{X}$ è la partizione nota della rete (nel nostro lavoro, come vedremo nel Capitolo 5, sarà data dalla classificazione nelle 5 categorie MDC) e la partizione $\mathcal{Y}$ quella ottenuta con un algoritmo di community detection.

4.5 Classificazione dei ruoli dei nodi

Il concetto di *ruolo* è stato sviluppato in modo sistematico da R. Guimerà e da L.A.N. Amaral in [18], dove viene proposto un metodo per determinare, sulla base delle caratteristiche topologiche e della partizione assegnata alla rete, il ruolo di ciascun nodo. L'approccio del metodo è basato sull'idea che nodi con lo stesso ruolo abbiano proprietà simili, individuabili indagando in modo approfondito gli assegnamenti dei nodi ai blocchi della partizione della rete. In Figura 4.4 è rappresentato un esempio di rete in cui è possibile individuare una partizione basata sull'equivalenza strutturale dei nodi: il blocco dei nodi bianchi, strutturalmente equivalenti perchè connessi solo a nodi neri, e il blocco dei nodi neri, strutturalmente equivalenti perchè connessi solo a nodi bianchi. Questa suddivisione in blocchi non mette in evidenza, tuttavia, il ruolo dei nodi nella rete: i nodi A e B hanno grado alto e sono al centro delle due comunità distinte a sinistra e a destra, i nodi C e D connettono le due comunità mentre gli altri nodi sono periferici. Sono queste caratterizzazioni di ruolo che si vogliono evidenziare in una qualunque rete.

La prima proprietà individuata come significativa è il *grado intra-comunità* (di seguito indicato come *Within-community degree* o *z-score*). La definizione di z-score per il nodo i è la seguente:

$$z_i = \frac{k_i^{c(i)} - \mu(k_i^{c(i)})}{\sigma(k_i^{c(i)})} \tag{4.28}$$

dove $k_i^{c(i)}$ è il numero di archi che collegano il nodo i agli altri nodi della comunità $c(i)$ a cui appartiene il nodo i, $\mu(k_i^{c(i)})$ è la media del grado interno dei nodi nella comunità $c(i)$ e $\sigma(k_i^{c(i)})$ la sua deviazione standard. Lo z-score di un nodo è dunque un indice standardizzato del grado interno alla comunità a cui è assegnato e misura quanto il nodo sia "ben connesso" agli altri nodi del suo blocco.
Un'estensione adatta a reti pesate viene utilizzata in [14]: nella definizione

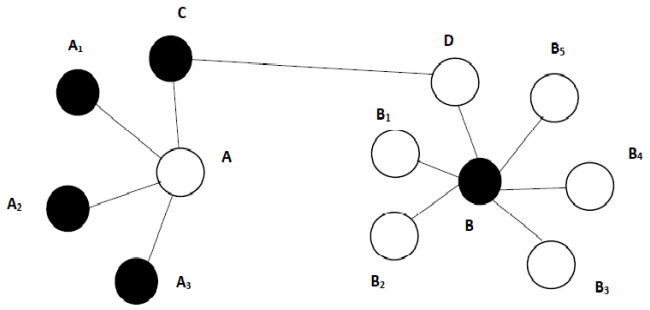

Figura 4.4: Rete di esempio, formata da due comunità distinte e in cui sono identificati due blocchi di nodi strutturalmente equivalenti: nodi connessi solo a nodi neri (A, D, B_{1-5}) e nodi connessi solo a nodi bianchi (B, D, A_{1-3}).

(4.28) il grado interno $k_i^{c(i)}$ viene sostituito con la strength $s_i^{c(i)}$ del nodo i diretta verso nodi di $c(i)$. L'espressione della strength interna è $s_i^{c(i)} = \sum_{j \in c(i)} w_{ij}$, con w_{ij} peso dell'arco $\{i, j\}$, e lo z-score è definito in questo caso come:

$$z_i = \frac{s_i^{c(i)} - \mu(s_i^{c(i)})}{\sigma(s_i^{c(i)})} \tag{4.29}$$

dove valgono le notazioni della formula precedente. In caso di reti pesate l'indice z si interpreta come *strength intra-comunità*.
La seconda proprietà da considerare è il *coefficiente di partecipazione* (*P-score*), indice che quantifica la diversificazione delle connessioni del nodo i verso le varie comunità. La definizione data in [18] e l'estensione per reti pesate introdotta in [14] per il nodo i sono riportate in seguito rispettivamente in (4.30) e (4.31):

$$P_i = 1 - \sum_{c=1}^{K} \left(\frac{k_i^c}{k_i} \right)^2 \tag{4.30}$$

$$P_i = 1 - \sum_{c=1}^{K} \left(\frac{s_i^c}{s_i} \right)^2 \tag{4.31}$$

dove K è il numero di blocchi della partizione assegnata, k_i^c indica il grado del nodo i verso nodi della comunità c e s_i^c, pari a $\sum_{j \in c} w_{ij}$, la corrispondente strength.
Si noti che il coefficiente di partecipazione di un nodo assume valori compresi tra 0 e 1, dove al valore 1 si tende se gli archi del nodo sono uniformemente distribuiti tra tutte le comunità e 0 si ottiene se tutti gli archi collegano il nodo con nodi della sua comunità. Valori vicini a 1 sono attribuiti, quindi, a nodi che tendono a "partecipare" e a connettersi alle altre comunità, al contrario dei nodi con valori di P vicini a 0, connessi in modo preferenziale con i nodi della comunità di appartenenza.
Nota la partizione della rete, data a priori o assegnata tramite un metodo di community-detection, i due indici descritti sono facilmente calcolabili e consentono di individuare il ruolo dei nodi.
Facendo riferimento all'esempio illustrato in Figura 4.4, i nodi della rete risultano, tramite questo approccio, classificabili come in Figura 4.5.

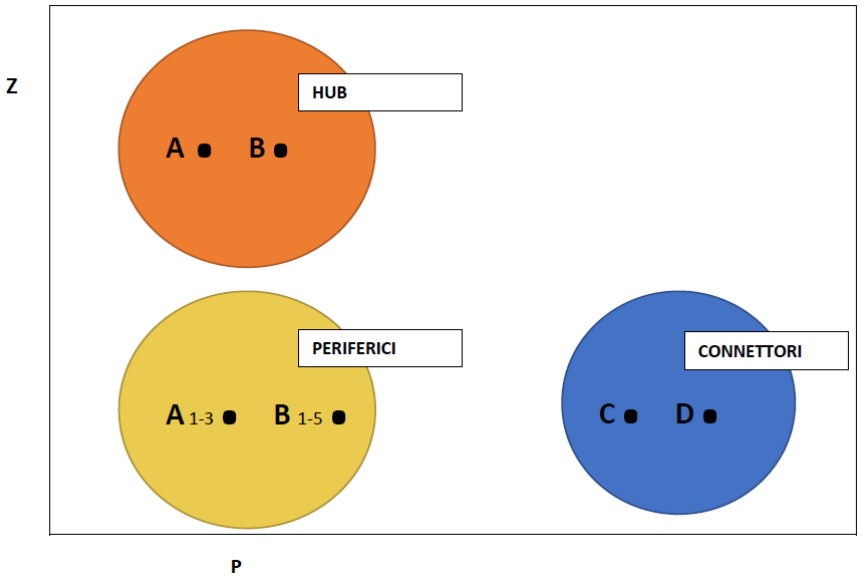

Figura 4.5: Classificazione dei nodi della rete in 4.4 secondo l'approccio basato sugli indici z e P

In [18] viene proposta una classificazione sistematica dei ruoli basata sui valori del within-community degree e del coefficiente di partecipazione. In primo luogo, il within-community degree consente di distinguere i nodi *centrali*, o *hub*, aventi indice $z \geq 2,5$, e i nodi *periferici* o *non-hub*, con $z < 2,5$. Viene poi introdotta una classificazione più fine dei nodi *hub* e *periferici* basata sul coefficiente di partecipazione: nodi aventi indice $P < 0,625$ sono

detti *provinciali* e i nodi con P compreso tra $0,625$ e $0,8$ sono detti *connettori*. Nodi con valori di P superiori a $0,8$ sono talmente connessi con nodi di altre comunità che l'assegnazione della loro comunità risulta non chiara e non supportata dall'evidenza: sono detti, per questo motivo, nodi *kinless* o, traducendo letteralmente, *senza parenti*. Considerando congiuntamente le due classificazioni introdotte è possibile individuare sette ruoli attribuibili ai nodi di una rete, descritti nel dettaglio nell'elenco e illustrati in Figura 4.6.

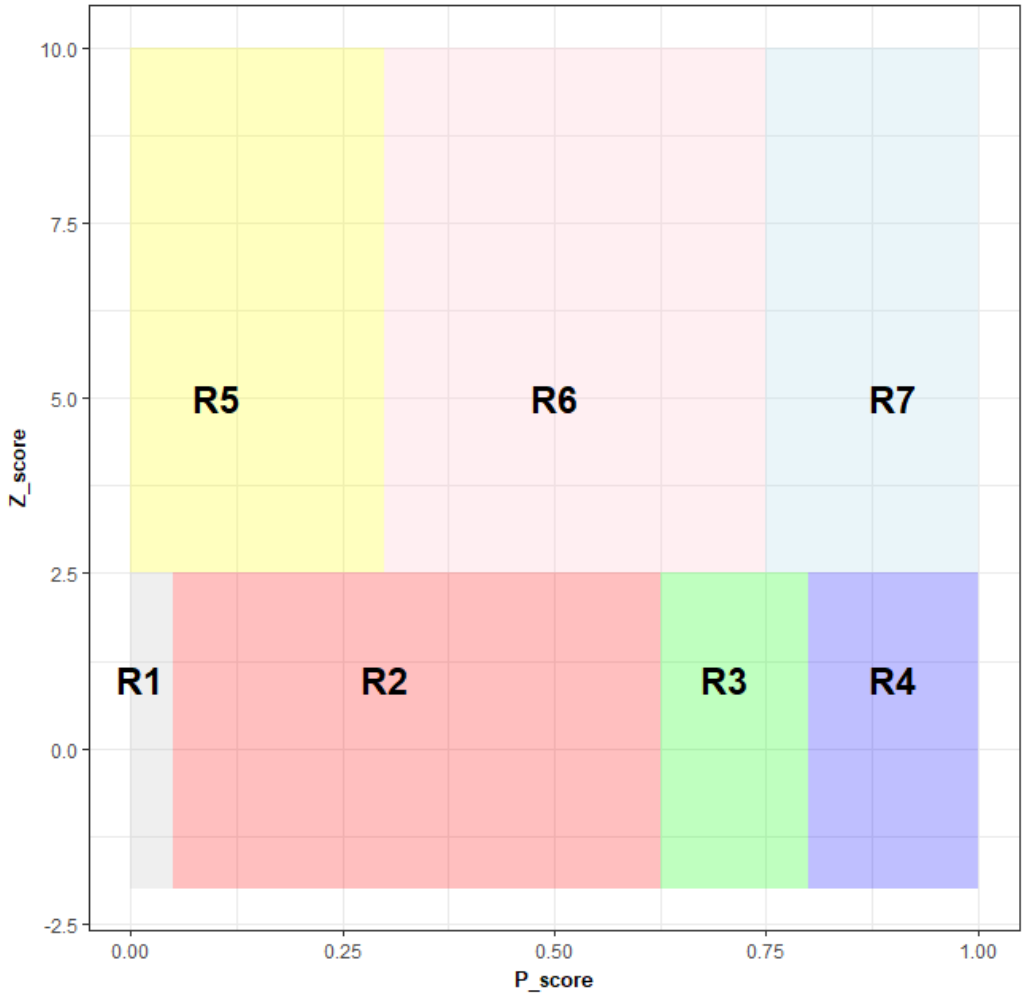

Figura 4.6: Regioni di assegnazione dei ruoli dei nodi nello spazio dei parametri P-z.

- Nodi ultra periferici, **Ruolo 1**: nodi con basso grado interno e coefficiente di partecipazione con valore vicino a zero: tutti gli archi connet-

tono il nodo alla sua comunità di appartenenza;

- Nodi periferici, **Ruolo 2**: nodi con basso grado interno con almeno il 60% dei propri archi rivolti all'interno della comunità di appartenenza ($P < 0,625$);
- Nodi periferici connettori, **Ruolo 3**: nodi con basso grado interno con almeno il 60% dei propri archi rivolti verso altre comunità ($P > 0,625$), ma con una percentuale significativa (almeno il 35%) degli archi rivolti verso la comunità di appartenenza ($P < 0,8$);
- Nodi periferici *kinless*, **Ruolo 4**: nodi con basso grado interno con una percentuale dei propri archi rivolti verso la comunità di appartenenza minore del 35% ($P > 0,8$);
- Nodi hub provinciali, **Ruolo 5**: nodi con alto grado interno e almeno $\frac{5}{6}$ dei propri archi rivolti verso la comunità di appartenenza ($P < 0,3$);
- Nodi hub connettori, **Ruolo 6**: nodi con alto grado interno e un'alta percentuale dei propri archi rivolti verso altre comunità ma almeno il 50% rivolti verso la comunità di appartenenza (P compreso tra $0,3$ e $0,75$);
- Nodi hub *kinless*, **Ruolo 7**: nodi con alto grado interno con meno del 50% dei propri archi rivolti verso la comunità di appartenenza ($P > 0,75$).

La scelta dei valori di z e P e delle percentuali di archi rivolti verso la comunità di appartenenza o le altre comunità che delimitano le regioni di assegnazione dei ruoli sono ottenute tramite argomenti illustrati nel dettaglio e validati in [18].
Nel nostro lavoro utilizzeremo la classificazione dei ruoli dei nodi descritta, sostituendo agli indici z e P la loro versione pesata (4.29) e (4.31).

Capitolo 5

Analisi di comunità della Rete Diagnosi e della Rete Pazienti

In questo capitolo presentiamo i risultati ottenuti applicando i metodi di community detection, descritti nel Capitolo 4, alle reti costruite proiettando la Rete Ricoveri sui nodi diagnosi e i nodi pazienti.
Nella prima parte del capitolo ci concentreremo sulla Rete Diagnosi: verranno analizzate le partizioni ottenute su questa rete con il metodo di Louvain e con il modello stocastico a blocchi (SBM), presentandone il valore di modularità complessivo, la cardinalità e la qualità, in termini di probabilità di persistenza, delle comunità individuate. Ricordiamo che con il metodo di Louvain il numero di comunità ottimale è un risultato del metodo di ottimizzazione, mentre nel modello stocastico a blocchi è necessario fornire questo valore come input all'algoritmo. Nel nostro lavoro sceglieremo di calcolare con il metodo SBM partizioni formate da 3 a 9 comunità.
Uno spazio importante verrà inoltre dato al confronto delle partizioni calcolate dagli algoritmi con la classificazione nota delle diagnosi nelle 5 categorie MDC. Tramite le matrici di Precision e Recall e il calcolo degli indici di similarità delle partizioni, presentati dal punto di vista teorico nel Capitolo 4, sarà possibile mettere in evidenza le corrispondenze esistenti tra le categorie basate sulla tipologia di diagnosi e le comunità, formate da gruppi di nodi densamente connessi, individuate dagli algoritmi. In generale, sarà interessante notare sia situazioni in cui le partizioni calcolate ricostruiscono la classificazione MDC nota a priori sia situazioni in cui non vi è una perfetta corrispondenza. In quest'ultimo caso, la partizione prodotta dall'algoritmo può fornire elementi utili per evidenziare la struttura della rete e su cui eventualmente basare una classificazione nuova delle diagnosi considerate, interessante anche dal punto di vista della gestione clinica dei ricoveri.

A seguito dell'analisi delle partizioni ricavate, verrà proposta una classificazione dei nodi diagnosi basata sul loro ruolo nelle comunità a cui risultano assegnati. La classificazione, come vedremo, sarà ottenuta calcolando il coefficiente di partecipazione e la strength intra-comunità, introdotti nel Capitolo 4, per ciascun nodo e potrà fornire informazioni interessanti sull'importanza e sul tipo di ruolo attribuibile alle diagnosi presenti nella rete.
Concluderemo, infine, illustrando la partizione ottenuta con il metodo di Louvain sulla rete costruita campionando, come descritto nel Capitolo 3, il 3% degli archi della Rete Ricoveri e proiettando tale rete campionata sui nodi paziente.

5.1 Analisi di comunità della Rete Diagnosi: Metodo di Louvain

Presentiamo in questa sezione i risultati ottenuti dall'analisi di comunità svolta sulla Rete Diagnosi con il metodo di Louvain. In Tabella 5.1 riportiamo le caratteristiche principali della rete di interesse, che abbiamo descritto in modo dettagliato nel Capitolo 3.

Rete Diagnosi	
Numero di nodi	3279
Numero di nodi (massima componente connessa)	3013
Numero di archi	103593
Numero di archi (massima componente connessa)	103592
Densità	0,02
Grado medio dei nodi	68,7
Strength media dei nodi	744,3

Tabella 5.1: Informazioni generali sulla Rete Diagnosi.

La scelta di individuare una partizione della rete con l'algoritmo di Louvain, presentato dal punto di vista teorico nel Capitolo 4, è stata dettata dall'efficienza e dalla velocità del metodo, che ha consentito di ottenere in modo immediato una suddivisione dei nodi della rete. Ricordiamo che il metodo di Louvain individua la partizione della rete che fornisce il valore massimo di modularità. La modularità della partizione ottenuta con questo metodo, oltre a costituire un indice per valutarne la sua qualità, può essere

utilizzata anche come valore di riferimento per partizioni risultanti dall'applicazione di altri algoritmi di community detection. Sottolineiamo, inoltre, un aspetto importante del metodo: il numero di comunità in cui suddividere i nodi è ottenuto come risultato dell'algoritmo e non deve essere stabilito a priori. Questo fatto consente di analizzare la rete in via preliminare pur non avendo a disposizione conoscenze a priori sulla sua struttura e, in seguito, di avere un numero di comunità indicativo in vista di analisi successive con metodi in cui sarà necessario fissare la cardinalità della partizione da individuare.
Nel nostro lavoro abbiamo applicato il metodo di Louvain alla Rete Diagnosi utilizzando la funzione `cluster_louvain` presente nel pacchetto *igraph* [15] del software R e documentata in Appendice C. La funzione riceve in input la rete e restituisce per ciascun nodo la comunità a cui risulta assegnato nella partizione di massima modularità. Come visto nel Capitolo 4, l'implementazione di *igraph* del metodo di Louvain è in grado di gestire i pesi della rete e massimizza la funzione di modularità pesata definita in (4.3). Sottolineiamo, inoltre, che in questa fase abbiamo utilizzato la massima componente connessa della Rete Diagnosi, costituita da 3013 nodi, poichè il metodo di Louvain richiede che la rete analizzata dall'algoritmo sia completamente connessa. Ricordiamo, infatti, che gli spostamenti dei nodi tra le comunità, implementati nelle varie fasi dell'algoritmo, sono basati sulle comunità assegnate ai nodi adiacenti a quelli di volta in volta presi in considerazione. I 264 nodi diagnosi isolati, se fossero considerati nella rete data in input all'algoritmo, verrebbero assegnati a una comunità iniziale, ma non sarebbe possibile la loro aggregazione nelle fasi successive, passaggio chiave per ottenere la partizione di massima modularità. I nodi esclusi dalla massima componente connessa (i 264 nodi isolati e i 2 appartenenti a una componente isolata) non risultano dunque assegnati ad alcuna comunità e vengono perciò indicati con NA nelle Tabelle di questo capitolo.

La partizione ottima per la Rete Diagnosi è costituita da 6 comunità e ha valore di modularità pari a 0,46. In Tabella 5.2 sono riportate le cardinalità delle 6 comunità inviduate dall'algoritmo di Louvain e la percentuale di nodi assegnati ad ognuna di esse.

Si può notare, come primo aspetto, che la partizione di Louvain assegna alla stessa comunità (la comunità 5) un terzo dei nodi totali e distribuisce, invece, la restante parte sulle altre 5 comunità in modo piuttosto equilibrato. Per valutare il risultato dato dall'algoritmo, calcoliamo la probabilità di persistenza delle comunità individuate. I valori ottenuti, calcolati tramite il software R e riportati nella Tabella 5.3, ci consentono di affermare che le comunità individuate dall'algoritmo sono significative: tutte le probabilità

Partizione LOUVAIN		
Comunità	**Cardinalità**	**%**
NA	266	8,11%
1	347	10,58%
2	471	14,36%
3	643	19,61%
4	225	6,86%
5	1102	33,61%
6	225	6,86%

Tabella 5.2: Cardinalità delle 6 comunità individuate con il metodo di Louvain e percentuale dei nodi attribuiti a ciascuna comunità.

di persistenza risultano maggiori di 0,5. Possiamo notare la presenza di una comunità particolarmente coesa: la comunità 3 ha probabilità di persistenza pari a 0,95. Fatta eccezione per la comunità 4, che presenta un valore di persistenza pari a 0,51, gli altri gruppi possono essere considerate a tutti gli effetti comunità significative, in quanto una percentuale molto superiore al 50% della strength dei nodi ad esse attribuita è rivolta verso nodi della comunità stessa.

Partizione LOUVAIN		
Comunità	**Probabilità di persistenza**	**z**
1	0,69	14,09
2	0,60	8,66
3	0,95	16,54
4	0,51	11,52
5	0,78	8,91
6	0,64	25,14

Tabella 5.3: Probabilità di persistenza delle 6 comunità individuate con il metodo di Louvain e relativo z-score. Tutte le probabilità di persistenza sono evidenziate in quanto maggiori di 0,5.

In Tabella 5.3 sono riportati anche gli z-score corrispondenti alla probabilità di persistenza. Seguendo l'approccio presentato in [14], infatti, una volta ottenuti i valori di interesse, è importante valutare la loro significatività statistica che, in questo caso, risulta verificata. Gli z-score sono, infatti,

sufficientemente elevati. Ricordiamo che l'importanza di questa fase di verifica è data dal fatto che, tramite un algoritmo di ottimizzazione, è sempre possibile ottenere una partizione di massima modularità ed è dunque fondamentale valutare a posteriori se i gruppi individuati presentino caratteristiche che consentano di definirli comunità.

Valutate la qualità della partizione e la significatività delle comunità, ci concentreremo ora sull'interpretazione a posteriori dei raggruppamenti individuati dall'algoritmo.
Presentiamo in primo luogo nei grafici in Figura 5.1 le distribuzioni di numero di ricoveri (a sinistra) e strength (a destra) dei nodi nei 6 gruppi. Questo tipo di informazioni sui nodi diagnosi consente di effettuare una prima analisi a posteriori delle comunità individuate dall'algoritmo. In particolare, pur dovendo tenere conto della forte eterogeneità nelle caratteristiche dei nodi della rete complessiva (che si evidenzia anche all'interno delle singole comunità), è possibile cogliere alcuni aspetti peculiari dei raggruppamenti sulla base di questi attributi noti.

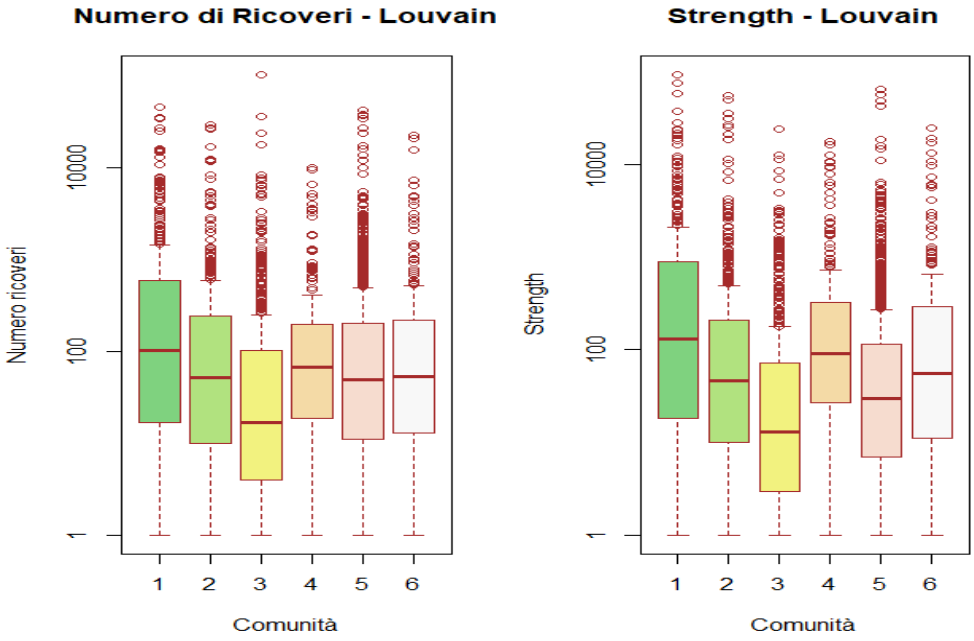

Figura 5.1: Distribuzione del numero di ricoveri e strength nelle 6 comunità individuate con il metodo di Louvain.

Sottolineiamo che le informazioni relative al numero di ricoveri complessivi per una certa diagnosi e la strength (così come il grado) del nodo che la

rappresenta nella rete sono strettamente correlate, come abbiamo illustrato nel Capitolo 3 nel grafico in Figura 3.11, ma non coincidenti dal punto di vista concettuale. Il numero di occorrenze nel dataset associato a una diagnosi descrive, infatti, la sua diffusione in termini assoluti e consente di distinguere le patologie più "frequenti" dalle patologie "rare". Il grado e la strength di un nodo diagnosi sono, invece, legate alla struttura della rete e alle relazioni esistenti tra le varie patologie nella rete proiettata. Risulta quindi sensato mettere in evidenza in modo distinto i dati relativi alla distribuzione nelle comunità del numero di ricoveri e della strength benchè, come si può vedere dalla Figura 5.1, le peculiarità più evidenti delle comunità legate a uno di essi si ritrovino spesso considerando l'altra informazione. Ad esempio, la comunità 1 presenta un valore mediano al di sopra del valore mediano della rete sia per quanto riguarda la distribuzione del numero di ricoveri sia della strength. Al contrario, si può notare che la comunità 3 riunisce nodi diagnosi con valori mediani di entrambi i termini inferiori ai rispettivi valori mediani della rete.

Un'altra informazione nota relativa ai nodi diagnosi e utilizzabile per l'interpretazione delle comunità della partizione è la categoria MDC di appartenenza. Obiettivo di questa fase del lavoro sarà quello di confrontare la classificazione qualitativa delle diagnosi sulla base del criterio clinico con la partizione prodotta dall'algoritmo.
In primo luogo, riportiamo in Tabella 5.4 i valori di Informazione mutua normalizzata (NMI) e Variazione di Informazione (VI) calcolati per le due partizioni, con i relativi z-score. I due indici, introdotti nel Capitolo 4, forniscono un'informazione sintetica della similarità tra le due partizioni della rete.

MDC vs Partizione Louvain		
	Valore calcolato	**z**
NMI	0,615	868,26
VI	0,147	-868,26

Tabella 5.4: Valori di NMI e VI e relativi z-score per il confronto tra la classificazione MDC e la partizione di Louvain.

Sottolineiamo che in questo caso, come nelle altre partizioni proposte nel capitolo, il numero di comunità individuate nei due raggruppamenti confrontati non è coincidente. Considerando i valori calcolati, possiamo concludere che la similarità tra la classificazione nota dei nodi in 5 MDC e la partizione ottenuta con Louvain, espressa dal valore di NMI, è piuttosto alta e, coerente-

mente, la distanza tra le due, rappresentata dalla Variazione di Informazione, è bassa. I valori di z, calcolati mediante randomizzazione dell'assegnazione dei nodi alle comunità, confermano la significatività statistica degli indici.

Per valutare in modo più approfondito la partizione prodotta dall'algoritmo di Louvain, utilizzeremo le matrici di Precision e Recall introdotte nel Capitolo 4. Con questo tipo di analisi è possibile confrontare a livello di comunità e in modo immediato i gruppi di nodi individuati dall'algoritmo con la loro classificazione in MDC. Scopo di questa fase del lavoro è mettere in evidenza, se sono presenti in modo chiaro, le caratteristiche qualitative che accomunano i nodi attribuiti alla stessa comunità.

Questo approccio si rivela particolarmente utile in questo contesto in primo luogo grazie al fatto che il numero di categorie MDC e il numero di comunità individuate dal metodo di Louvain risultano confrontabili. Con questo metodo sarebbe possibile, in linea di principio, confrontare anche la partizione ottenuta con la classificazione delle diagnosi data dai codici DRG, ma, in questo caso, le numerosità dei due raggruppamenti sarebbero fortemente sbilanciate (193 DRG contro 6 comunità) e le corrispondenti matrici di Precision e Recall risulterebbero più difficilmente interpretabili. Si noti, infine, che, utilizzando questo approccio ma avendo a disposizione altri tipi di informazioni legate alle diagnosi, si potrebbero mettere in evidenza altre caratteristiche delle comunità individuate e cogliere, se esistono, altre logiche di raggruppamento che in questo lavoro non vengono indagate.

Prima di esporre le considerazioni che si possono ricavare dall'analisi delle matrici di Precision e Recall, presentiamo le modalità di interpretazione delle figure presenti nel capitolo. Per ciascun confronto tra due partizioni, vengono proposte le matrici di Precision (in alto) e di Recall (in basso) ponendo sulle righe le 5 categorie MDC e sulle colonne le comunità individuate dagli algoritmi (in Figura 5.2 le comunità da 1 a 6, per le altre figure del capitolo le comunità da 1 alla cardinalità K stabilita per la partizione). A destra delle matrici è riportata la scala cromatica che consente di interpretare più facilmente i valori numerici degli elementi delle matrici. I valori alti, cioè vicini a 1, sono indicati dal colore giallo e sono presenti, nella matrice di Precision, se le diagnosi di una data comunità risultano quasi interamente contenute in una certa MDC e, nella matrice di Recall, se le diagnosi di una data MDC risultano quasi tutte appartenenti a una certa comunità. Ricordiamo infatti che gli elementi p_{hk} della matrice di Precision corrispondono alla frazione di nodi della comunità C_k che appartengono alla MDC h e che gli elementi r_{hk} della matrice di Recall corrispondono, invece, alla frazione di nodi della MDC h che vengono assegnati alla comunità k. Se gli elementi corrispondenti a

una coppia MDC-comunità risultano alti in entrambe le matrici si può concludere che, nelle partizioni considerate, i due insieme siano essenzialmente coincidenti. I valori più bassi, vicini a 0, sono rappresentati da colori tendenti al blu e, nella matrice di Precision, denotano che le diagnosi assegnate a una data comunità non appartengono a una certa MDC e, nella matrici di Recall, che diagnosi di una data MDC non vengono assegnati a una certa comunità. É possibile, inoltre, che nella matrice di Precision si abbiano valori alti per più di una comunità in corrispondenza di una stessa MDC: in questo caso si può affermare che quel MDC riunisce le diagnosi assegnate a quella comunità. Analogamente, se in corrispondenza di una certa comunità nella matrice di Recall sono presenti valori alti per più di una MDC si ha che la data comunità raggruppa un numero significativo di diagnosi appartenenti alle MDC coinvolte. L'analisi congiunta delle matrici di Precision e Recall corrispondenti alle varie partizioni di volta in volta considerate consente, svolgendo osservazioni analoghe a quelle esposte sopra, di evidenziare aspetti significativi delle comunità dei nodi della rete.
Entrando nel merito del confronto tra la partizione di Louvain e la classificazione in MDC, riportiamo in Figura 5.2 le matrici di Precision e Recall.

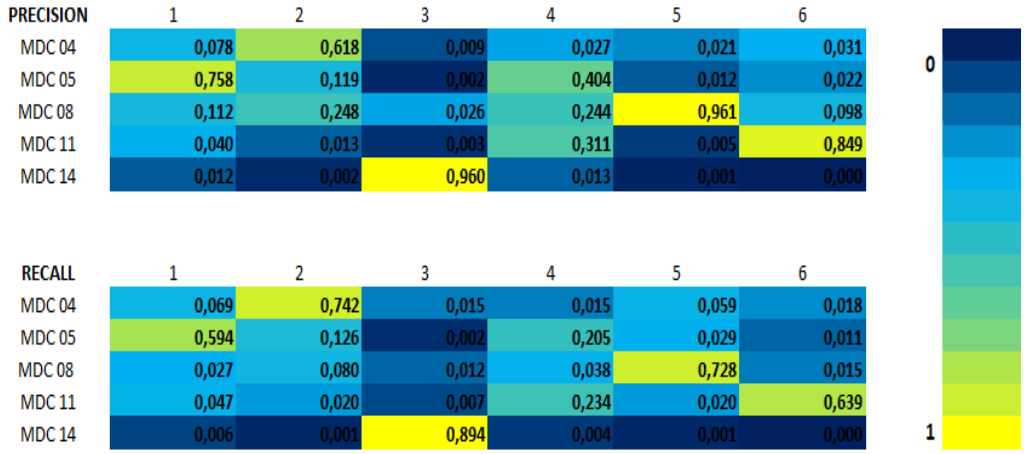

Figura 5.2: Matrici di Precision e Recall. Classificazione MDC vs Partizione di Louvain (6 comunità).

Gli elementi più intessanti che possono essere messi in evidenza da questi risultati sono le corrispondenze, con valori di Precision e Recall particolarmente elevati, tra la comunità 3 e la MDC 14 e tra la comunità 5 e la MDC 8. Per quanto riguarda la prima coppia ricordiamo che la probabilità di persistenza della comunità 3 risulta pari al 95%, come riportato in 5.3: in questa partizione emerge in modo chiaro una comunità coesa e ben distin-

ta formata da diagnosi legate alla gravidanza e al parto. Possiamo notare inoltre un valore molto vicino a 1 per la Precision tra la comunità 5 e la MDC 8 che indica che la categoria di diagnosi legate all'ortopedia è quasi interamente contenuta in una comunità. Queste due osservazioni confermano la nostra ipotesi a priori sulla struttura della rete: le patologie riguardanti la gravidanza e l'ortopedia risultano slegate dalle altre e presentano una tendenza evidente a formare gruppi distinti. É interessante notare anche le corrispondenze significative tra la comunità 6 e la categoria 11 (patologie del sistema urinario), la comunità 1 e la categoria 5 (patologie del sistema cardiocircolatorio) e la comunità 2 e la categoria 4 (patologie del sistema respiratorio). Possiamo dunque affermare che le 5 categorie MDC vengano ricostruite con 5 comunità della partizione ottenuta con il metodo di Louvain. La sesta comunità inviduata dall'algoritmo, che viene etichettata con il numero 4, presenta valori significativi per quanto riguarda gli MDC 5, 8 e 11, ma presenta il valore di probabilità di persistenza meno significativo (si veda ancora la Tabella 5.3). I nodi attribuiti a questa comunità "mista" risultano quindi poco coesi tra loro, ma non sufficientemente connessi ai nodi di altri gruppi per essere assegnati alla comunità corrispondente alla categoria MDC a cui appartengono.

5.2 Analisi di comunità della Rete Diagnosi: Modello stocastico a blocchi

In questa sezione presenteremo i risultati ottenuti applicando il modello stocastico a blocchi alla Rete Diagnosi.
L'algoritmo utilizzato, che abbiamo descritto in dettaglio nel Capitolo 4, è implementato in un codice C++ fornito dagli autori dell'articolo di riferimento [21] e disponibile on-line (si veda [3]). La versione del codice con cui abbiamo svolto le analisi prevede di ricevere in input la lista degli archi della rete, in cui ogni arco è identificato da una coppia di codici diagnosi. L'algoritmo, analogamente a quello relativo al metodo di Louvain, restituisce per ogni nodo l'etichetta corrispondente alla comunità assegnata nella partizione finale calcolata. Inoltre, il codice fornisce come output la matrice ω, definita nel Capitolo 4, i cui elementi ω_{rs} esprimono il valore atteso del numero di archi tra le comunità r e s. La matrice ω sarà utile per identificare la struttura delle comunità individuate, mettendo in evidenza le relazioni dei nodi all'interno di esse e tra i nodi appartenenti a gruppi diversi.
Il codice fornito dagli autori consente di analizzare la rete sia con il metodo standard sia con la sua estensione degree-corrected. Nel nostro lavoro

abbiamo utilizzato quest'ultima versione, che si è rivelata, nel confronto dei risultati ottenuti rispetto a quelli forniti dal metodo standard, decisamente più adeguata a individuare comunità significative nella Rete Diagnosi, che abbiamo identificato come fortemente eterogenea.
Per utilizzare correttamente il codice che implementa il metodo, sono necessari alcuni accorgimenti tecnici, che di seguito precisiamo.
In primo luogo, sottolineiamo l'importanza di disporre di etichette identificative dei nodi con valori progressivi (nel nostro caso i nodi diagnosi sono numerati da 1 a 3013). Nelle fasi preliminari dell'algoritmo, infatti, viene implementata, a seguito della lettura della lista fornita in input, l'allocazione di un vettore contenente i gradi dei nodi presenti nella rete indicizzato tramite il codice identificativo del nodo stesso. L'accorgimento di rietichettare i nodi è fondamentale per una gestione corretta di una rete ottenuta, come nel nostro caso, considerando solo la massima componente connessa della rete originale e in cui si trascurano alcuni nodi. Un altro aspetto da considerare, inoltre, è proprio che, per l'applicazione del metodo SBM, così come per il metodo di Louvain, risulta sensato analizzare solo reti completamente connesse, escludendo preliminarmente i nodi isolati o non appartenenti alla massima componente connessa. Il motivo è dato dalla procedura implementata: in fase di inizializzazione tutti i nodi (compresi i nodi isolati) risultano assegnati a una comunità in modo casuale ma, se per i nodi connessi è possibile avere uno spostamento in un'altra comunità dettato dall'ottimizzazione della funzione obiettivo, i nodi isolati rimangono assegnati in ogni step al gruppo random di partenza. La partizione finale risulterebbe dunque fortemente dipendente dall'assegnazione casuale di questi nodi.
L'inizializzazione casuale delle comunità, primo passaggio dell'algoritmo, costituisce un aspetto critico del metodo. Se, infatti, l'effetto dell'inizializzazione casuale nel metodo di Louvain viene mitigato dalla successiva aggregazione delle comunità, nel metodo SBM ha importanti conseguenze sul risultato finale, in quanto le comunità possono contenere nodi assegnati ad esse per l'assegnamento casuale iniziale (nel caso non avvenga alcuno spostamento nel corso delle fasi della procedura). Come abbiamo descritto nel Capitolo 4, è possibile fissare il numero di inizializzazioni da effettuare sulla stessa rete per ogni partizione che si intende individuare: l'algoritmo restituirà la partizione ottima tra tutte quelle ottenute nei vari run. Nel nostro lavoro abbiamo eseguito 100 inizializzazioni per ogni valore fissato del numero di comunità. Abbiamo notato, nelle varie fasi del lavoro, come i risultati ottenuti siano tanto migliori in termini di qualità quante più run vengono eseguite. É da sottolineare, tuttavia, che la necessità di eseguire l'algoritmo per un numero elevato di run riduce drasticamente la velocità e l'efficienza del metodo. Per ottenere la partizione ottima in 3 comunità per la rete in analisi, ad esem-

pio, sono necessarie circa 3 ore, che diventano 25 se si vogliono individuare 9 comunità. I tempi riportati si riferiscono all'esecuzione del codice fornito in [3] su una CPU Intel(R) Core (TM) i5-5200U con frequenza di 2.20 GHz e con 8 GB di RAM.
Un importante aspetto da evidenziare è la gestione dei pesi degli archi da parte dell'algoritmo SBM. Il codice utilizzato non considera direttamente i pesi della rete, come avviene nel metodo di Louvain, ma è in grado di analizzare reti multigrafo. É necessaria, quindi, la costruzione preliminare del multigrafo: ogni arco dotato di peso maggiore di uno viene ripetuto nella lista un numero di volte pari al peso corrispondente. Il numero di archi totali presenti nella lista gestita dall'algoritmo risulta così pari a 1121282. Possiamo notare che questo valore è, coerentemente, uguale alla metà del valore della somma delle strength di tutti i nodi della rete (2242564).
Il metodo SBM richiede di imporre all'algoritmo il numero di comunità da individuare. Questo fatto può costituire un limite, poichè corrisponde a formulare ipotesi a priori sulla struttura della rete, ma offre grande flessibilità nell'indagine. Nel nostro caso, avendo a disposizione la classificazione nelle 5 categorie MDC e la partizione di massima modularità di Louvain in 6 comunità, abbiamo cercato partizioni di cardinalità K con valori simili, per consentire un confronto dei risultati e favorire la loro interpretabilità. Un limite importante che è stato preso in considerazione è il tempo di esecuzione dell'algoritmo, che, come abbiamo sottolineato sopra, aumenta in modo considerevole con l'aumento del numero di comunità da individuare. La scelta di limitare a 9 la cardinalità della partizione massima è stata dettata, quindi, nel nostro caso, anche da necessità tecniche. Il nostro scopo nell'utilizzo del metodo basato sul modello stocastico a blocchi non è quello di ottenere una partizione con il numero ottimale di comunità, ma di testare le performance dell'algoritmo e di mettere in evidenza, con un'analisi a posteriori, eventuali caratteristiche dei gruppi individuati.
Ricordiamo, inoltre, che l'algoritmo utilizzato si basa sull'ottimizzazione della funzione obiettivo definita in (4.14). In Tabella 5.5 riportiamo i valori ottimi ottenuti nelle diverse partizioni. Per ogni valore di cardinalità K delle partizioni, i raggruppamenti ottimi forniscono il valore di log-verosimiglianza riportato.

Possiamo notare che il valore ottimo della funzione obiettivo aumenta (diventa meno negativo) con il valore di K. Questo fatto non è sorprendente ed è messo in evidenza in letteratura, ad esempio in [27]. La scelta della cardinalità della partizione che massimizzerebbe la verosimiglianza sarebbe quella banale con $K = N$, con N numero di nodi della rete. Questo tipo di partizione non risulta utile e mette in evidenza l'importanza di una scelta opportuna di K, in modo da descrivere efficacemente la struttura della rete

Partizioni SBM	
Numero di comunità	**Log-verosimiglianza**
$K = 3$	$-1,595 \times 10^7$
$K = 4$	$-1,579 \times 10^7$
$K = 5$	$-1,575 \times 10^7$
$K = 6$	$-1,568 \times 10^7$
$K = 7$	$-1,560 \times 10^7$
$K = 8$	$-1,555 \times 10^7$
$K = 9$	$-1,551 \times 10^7$

Tabella 5.5: Valori di log-verosimiglianza per le partizioni SBM.

senza incorrere nell'*overfitting* dei dati.
Per ogni partizione è stata poi calcolata la modularità corrispondente. Pur non essendo la funzione obiettivo massimizzata dall'algoritmo, essa costituisce, infatti, un utile indice per valutare la qualità della partizione e per il confronto con il risultato del metodo di Louvain, che, invece, ricordiamo essere un metodo di massimizzazione di modularità. In Tabella 5.6 riportiamo i valori ottenuti per tutte le partizioni calcolate.

Partizioni SBM	
Numero di comunità	**Modularità**
$K = 3$	0,377
$K = 4$	0,426
$K = 5$	0,451
$K = 6$	0,432
$K = 7$	0,460
$K = 8$	0,386
$K = 9$	0,459
Louvain $K = 6$	0,460

Tabella 5.6: Valori di modularità per le partizioni SBM e, per confronto, per la partizione di Louvain.

Le modularità in tabella hanno valori inferiori ma molto vicini al valore di massima modularità calcolato con il metodo di Louvain. Questo fatto

fornisce un'ulteriore conferma della buona capacità del metodo SBM di identificare correttamente la struttura della rete, una volta imposto il vincolo sulla cardinalità della partizione da individuare.

Di seguito presenteremo in dettaglio le partizioni ottenute con il metodo SBM per $K = 5$ e $K = 6$ riportando le loro caratteristiche dal punto di vista dell'analisi della rete e fornendo un'ipotesi di interpretazione delle comunità. I risultati ottenuti per le partizioni con $K = 3, 4, 7, 8, 9$, invece, sono riportati nell'Appendice A.

Partizione SBM con $K = 5$

Analizziamo nel dettaglio i risultati relativi alla partizione in 5 comunità, interessante in quanto la sua cardinalità è uguale a quella della classificazione MDC ed quindi consente di effettuare un confronto diretto.
Tutti i gruppi individuati presentano dei valori alti di probabilità di persistenza (si veda la Tabella 5.8) e possiamo quindi considerare la partizione come significativa. Sottolineiamo l'importanza di valutare a posteriori i risultati in termini di coesione delle comunità, soprattutto in metodi in cui, come nel modello a blocchi stocastico, si impone a priori la cardinalità della partizione.
Le cardinalità delle 5 comunità sono riportate in Tabella 5.7: possiamo notare che i gruppi non sono perfettamente bilanciati. Questo fatto ci consente di constatare la capacità dell'algoritmo di caratterizzare la struttura della rete, costituita anche da gruppi di nodi particolarmente connessi ma ristretti, come le comunità 2 e 5 (con cardinalità rispettivamente pari a 385 e 224).
Le matrici di Precision e Recall in Figura 5.4, unitamente al valore significativo di NMI riportato in Tabella 5.9, evidenziano una buona corrispondenza tra le seguenti categorie MDC e le comunità identificate: MDC 8 e comunità 4, MDC 14 e comunità 3, MDC 5 e comunità 2 e MDC 11 e comunità 5. Più incerta, anche se riscontrabile, è la corrispondenza tra la comunità 1 e la categoria MDC 4. Le diagnosi legate al sistema respiratorio (appartenenti alla categoria MDC 4) sono, infatti, assegnate alla comunità 1 insieme a una frazione non trascurabile di diagnosi appartenenti alle categorie MDC 5, 8 e 11. Questa partizione, dunque, individua un gruppo di patologie riguardanti per lo più il sistema respiratorio unite a diagnosi di altro tipo ad esse connesse. Un esempio di diagnosi non appartenente alla categoria MDC 4 ma assegnata in questa partizione alla comunità 1 è l'insufficienza renale. É ragionevole affermare che esista una connessione, che il metodo di community detection in questo caso riesce a cogliere, tra questa patologia e problematiche a livello

respiratorio. Notiamo che le diagnosi appartenenti alla categoria MDC 11 vengono assegnate, in questa partizione, alla comunità 5, ben distinta e con alta corrispondenza con questa MDC, ma anche alla comunità 1, che abbiamo visto essere "mista". Questo fatto indica che l'algoritmo distingue all'interno della stessa categoria le patologie riguardanti in modo specifico il sistema urinario (ad esempio, vi troviamo la calcolosi renale), che costituiscono una comunità distinta, e le patologie più complesse (come, appunto, l'insufficienza renale) e, per questo, connesse a diagnosi riguardanti altri apparati ed assegnate ad altre comunità.
La matrice ω, riportata in Tabella 5.10, fornisce informazioni sul tipo di connessione esistente tra le varie comunità. Possiamo notare che le comunità 5 e 2 presentano un valore atteso alto di archi rivolti internamente alla comunità e che le comunità 1 e 2, pur essendo ben distinte, presentano un valore atteso significativo per un arco che connette nodi appartenenti ai due gruppi. In particolare, si ha un valore ω_{12} pari a 0,58: in altri termini, circa un nodo su 2 appartenente alla comunità 1 ci si aspetta che sia connesso a un nodo della comunità 2.
Dai grafici in Figura 5.3 si possono evidenziare due comunità con distribuzione, sia per quanto riguarda la frequenza delle occorrenze sia la strength, rispettivamente concentrata su valori superiori e inferiore ai valori mediani della rete. Si tratta nel primo caso della comunità 2, che nelle precendenti osservazioni abbiamo messo in forte corrispondenza alla categoria MDC 5 (patologie cardiache). La comunità con valori mediani più bassi è, invece, la comunità 3, strettamente associata alla categoria MDC 14 (patologie legate alla gravidanza e al parto). Questo tipo di distribuzioni, una volta evidenziata la corrispondenza esistente tra le comunità e le categorie, non sorprende: nel grafico in Figura 3.10 possiamo notare che la distribuzioni di strength della MDC 5 e 14 rispecchianno quelle delle comunità 2 e 3.

Partizione SBM $K = 5$		
Comunità	**Cardinalità**	**%**
NA	266	8,11%
1	661	21,94%
2	385	12,78%
3	656	21,77%
4	1087	36,08%
5	224	7,43%

Tabella 5.7: Cardinalità delle comunità individuate con il metodo SBM con $K = 5$ e percentuali di nodi attribuiti a ciascuna comunità.

Partizione SBM $K = 5$		
Comunità	**Probabilità di persistenza**	**z**
1	0,61	7,71
2	0,72	11,02
3	0,86	13,34
4	0,70	7,54
5	0,64	14,51

Tabella 5.8: Probabilità di persistenza delle comunità individuate con il metodo SBM con $K = 5$ e relativo z-score. Tutte le probabilità di persistenza sono evidenziate in quanto maggiori di 0,5.

MDC vs Partizione Sbm $K = 5$		
	Valore calcolato	**z**
NMI	0,623	974,42
VI	0,137	-974,42

Tabella 5.9: Valori di NMI e VI e relativi z-score per il confronto tra la classificazione MDC e la partizione ottenuta con il metodo SBM con $K = 5$.

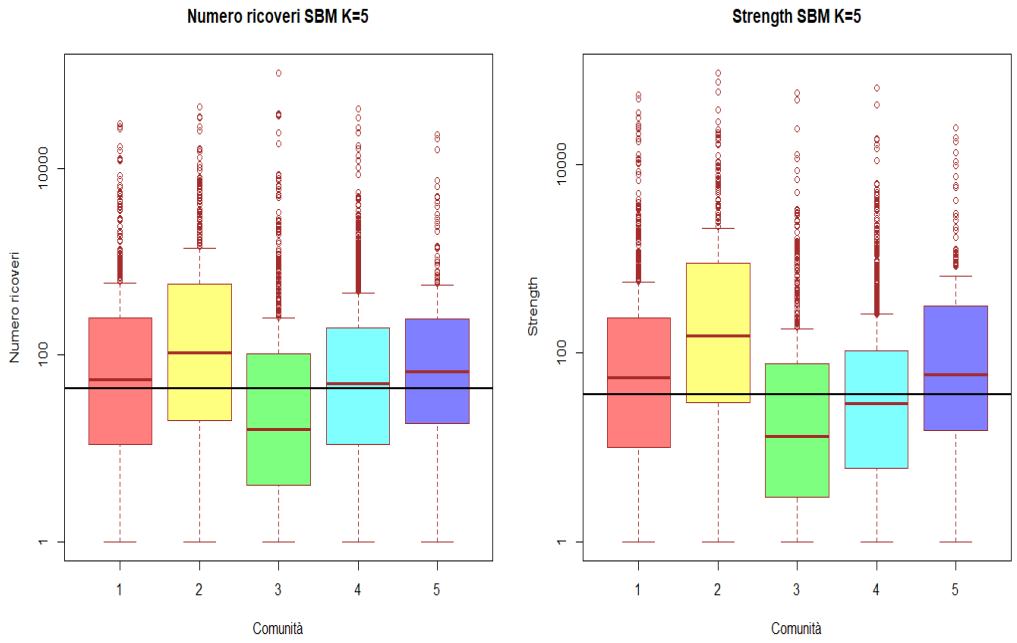

Figura 5.3: Distribuzione del numero di ricoveri e strength nelle comunità individuate con il metodo di SBM con $K = 5$.

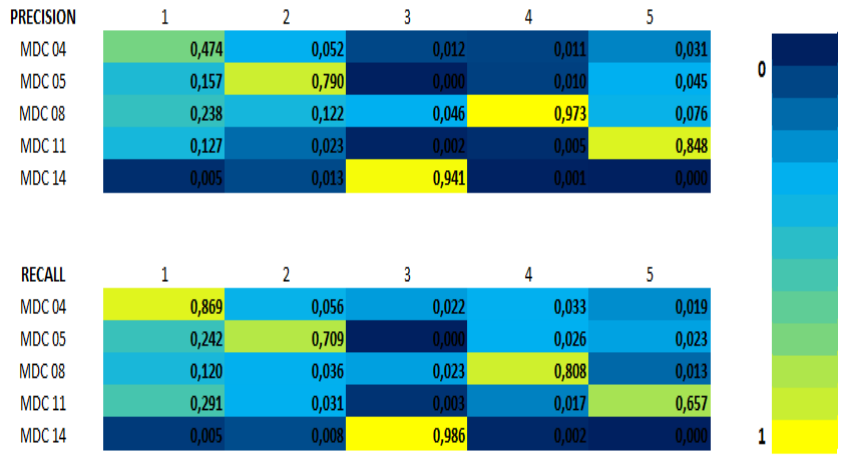

Figura 5.4: Matrici di Precision e Recall. Classificazione MDC vs Partizione con metodo SBM (5 comunità)

matrice ω					
	com 1	com 2	com 3	com 4	com 5
com 1	**0,794**	**0,581**	**0,011**	**0,058**	**0,196**
com 2	**0,581**	**3,822**	**0,026**	**0,110**	**0,290**
com 3	**0,011**	**0,026**	**0,541**	**0,034**	**0,028**
com 4	**0,058**	**0,110**	**0,034**	**0,246**	**0,046**
com 5	**0,196**	**0,290**	**0,028**	**0,046**	**2,483**

Tabella 5.10: Matrice ω della partizione SBM con $K = 5$: l'elemento ω_{rs} rappresenta il numero atteso di archi tra un nodo della comunità r e un nodo della comunità s. Il colore della cella è indicativo del valore in essa contenuto.

Partizione SBM con $K = 6$

Presentiamo ora i risultati ottenuti dal metodo SBM imponendo la ricerca di 6 comunità. In questa fase del lavoro, scegliamo di concentrare le analisi sulla partizione SBM con $K = 6$ perchè presenta la stessa cardinalità della partizione di massima modularità individuata con il metodo di Louvain. Il confronto dei risultati dell'applicazione dei due metodi sulla stessa rete ci consente di valutare in modo interessante le performance dell'algoritmo SBM, avendo un riferimento che abbiamo valutato essere significativo.
Prima di effettuare il confronto, illustriamo i risultati relativi alla sola partizione SBM.
Possiamo notare, in primo luogo, che, come nella partizione in 5 comunità, esiste uno sbilanciamento evidente tra le numerosità delle comunità: in Tabella 5.11 è presente un gruppo (comunità 3) formato da 1057 nodi e un altro (comunità 4) che ne contiene solo 104. Anche in questo caso, dunque, l'algoritmo è in grado di cogliere gruppi di nodi ristretti ma ben distinti dalle altre comunità della rete.
Le probabilità di persistenza in Tabella 5.12 sono tutte di valore alto e confermano, quindi, la significatività delle comunità individuate. In particolare, evidenziamo la presenza di una comunità (comunità 6) con probabilità di persistenza pari al 95%. Questa comunità si può evidenziare anche nei grafici in Figura 5.5, in quanto presenta distribuzioni di numero di ricoveri e strength al di sotto del valore mediano della rete. L'analisi delle matrici di Precision e Recall in Figura 5.6 rivela che essa presenta una corrispondenza quasi perfetta con la categoria MDC 14, che identifica le diagnosi legate alla gravidanza e al parto. Tutte le osservazioni fatte su questa comunità risultano perfettamente coerenti con le ipotesi a priori sulle patologie della categoria MDC 14: ci aspettiamo, infatti, che questo tipo di diagnosi siano

causa di ricoveri per lo più dovuti a un evento isolato nel tempo e che, a meno di complicanze, siano legate solo ad altre diagnosi simili. Un'ulteriore conferma di questo è data dalla matrice ω in Tabella 5.14: il valore atteso di connessioni della comunità 6 con se stessa, riportato nell'elemento ω_{66}, è l'unico significativo nella riga (e nella colonna) relativa a questa comunità.
Dalle matrici di Precision e Recall si possono ricavare altre informazioni interessanti. É possibile notare, innanzittutto, che la categoria MDC 8, formata dalle diagnosi di tipo ortopedico, è suddivisa quasi interamente nelle comunità 3 e 4, che sono rispettivamente il gruppo più numeroso e il gruppo più ristretto. Possiamo concludere dunque che l'algoritmo divide la categoria MDC 8, che ricordiamo essere formata da un numero elevato di diagnosi, sulla base di criteri di connessione interna. Notiamo, infatti, che il valore atteso di un arco all'interno della comunità 4 è decisamente elevato (l'elemento ω_{44} della matrice in Tabella 5.14 è pari a 18,5). D'altra parte, la stessa matrice ω suggerisce che la comunità 3, più numerosa, sia poco connessa internamente e connessa quasi esclusivamente con la comunità 4.
Una corrispondenza abbastanza precisa è riscontrabile tra la comunità 5 e la categoria MDC 4 (diagnosi del sistema respiratorio). Nella stessa comunità sono tuttavia presenti quantità non del tutto trascurabili di diagnosi appartenenti alle categorie MDC 5 (diagnosi del sistema cardiocircolatorio) e MDC 11 (diagnosi del sistema urinario). Un'indagine più approfondita, tramite altri strumenti o che disponga di ulteriori informazioni sulle diagnosi coinvolte, consentirebbe un'interpretazione più chiara dei legami esistenti tra le diagnosi assegnate alla stessa comunità pur essendo di categorie diverse.
Un'altra situazione interessante si nota valutando la comunità 2, che riunisce un certo numero di diagnosi appartenenti alle categorie MDC 5 e MDC 11. Il legame che emerge tra queste due categorie non sorprende: patologie complesse ai reni possono provocare scompensi anche al sistema cardiocircolatorio. É noto, inoltre, che alcuni tipi di trattamenti farmacologici utilizzati per curare patologie cardiache possano provocare problemi renali.
Da ultimo, evidenziamo come le diagnosi della categoria MDC 5 siano assegnate alle comunità 1 e 2. La comunità 1, composta da solo 204 elementi, presenta un elevato valore atteso di archi interni (l'elemento ω_{11} in Tabella 5.14 è pari a 4,6) ma una probabilità di persistenza non particolarmente elevata (pari a 0,55). La distribuzione di strength di questa comunità, che nel grafico in Figura 5.5 vediamo concentrata su valori superiori ai valori mediani della rete, suggerisce che l'elevato valore atteso di archi interni sia dovuto più ad essa che alla coesione della comunità.

Partizione SBM $K = 6$		
Comunità	**Cardinalità**	**%**
NA	266	8,11%
1	204	6,77%
2	496	16,46%
3	1057	35,08%
4	104	3,45%
5	520	17,26%
6	632	20,98%

Tabella 5.11: Cardinalità delle comunità individuate con il metodo SBM con $K = 6$ e percentuali di nodi attribuiti a ciascuna comunità.

Partizione SBM $K = 6$		
comunità	**probabilità di persistenza**	**z**
1	0,55	17,71
2	0,60	8,30
3	0,57	4,71
4	0,77	42,02
5	0,61	12,70
6	0,95	14,31

Tabella 5.12: Probabilità di persistenza delle comunità individuate con il metodo SBM con $K = 6$ e relativo z-score. Tutte le probabilità di persistenza sono evidenziate in quanto maggiori di 0,5.

MDC vs Partizione Sbm $K = 6$		
	Valore calcolato	**z**
NMI	0,577	833,35
VI	0,160	-833,35

Tabella 5.13: Valori di NMI e VI e relativi z-score per il confronto tra la classificazione MDC e la partizione ottenuta con il metodo SBM con $K = 6$.

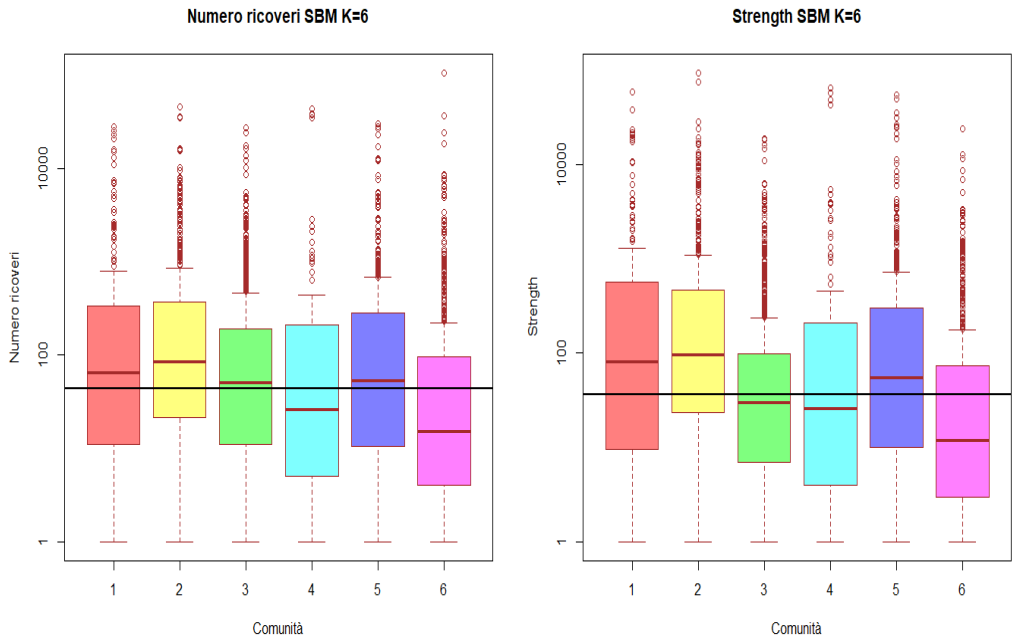

Figura 5.5: Distribuzione del numero di ricoveri e strength nelle comunità individuate con il metodo di SBM con $K = 6$.

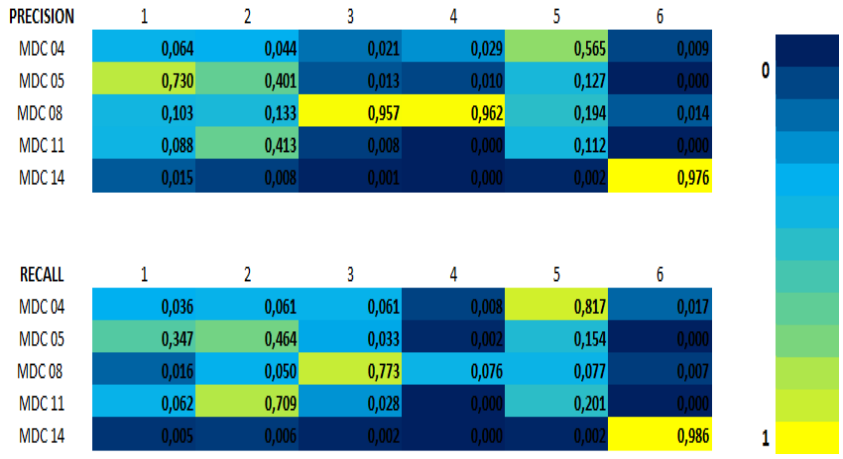

Figura 5.6: Matrici di Precision e Recall. Classificazione MDC vs Partizione con metodo SBM (6 comunità).

matrice ω						
	com 1	com 2	com 3	com 4	com 5	com 6
com1	**4,650**	**1,004**	**0,066**	**0,336**	**0,297**	**0,014**
com 2	**1,004**	**1,697**	**0,075**	**0,240**	**0,482**	**0,006**
com 3	**0,066**	**0,075**	**0,141**	**0,278**	**0,055**	**0,005**
com 4	**0,336**	**0,240**	**0,278**	**18,519**	**0,158**	**0,002**
com 5	**0,297**	**0,482**	**0,055**	**0,158**	**1,128**	**0,005**
com 6	**0,014**	**0,006**	**0,005**	**0,002**	**0,005**	**0,381**

Tabella 5.14: Matrice ω della partizione SBM con $K = 6$: l'elemento ω_{rs} rappresenta il numero atteso di archi tra un nodo della comunità r e un nodo della comunità s. Il colore della cella è indicativo del valore in essa contenuto.

Mettiamo ora a confronto i risultati ottenuti con quelli relativi al metodo di Louvain, descritti nella sezione precedente. Una prima informazione sintetica per valutare le due partizioni in modo congiunto è data dagli indici NMI e VI riportati in Tabella 5.15. Notiamo che il valore significativamente alto di NMI e basso di VI ci consentono di concludere che i due raggruppamenti non sono particolarmente "lontani" (nello spazio delle partizioni).

Partizioni Louvain vs Sbm $K = 6$		
	Valore calcolato	**z**
NMI	0,701	353,70
VI	0,119	-353,70

Tabella 5.15: Valori di NMI e VI e relativi z-score per il confronto tra la la partizioni ottenute con il metodo SBM con $K = 6$ e il metodo di Louvain.

Per analizzare in modo più preciso gli assegnamenti delle diagnosi ottenuti dai due metodi, utilizziamo le matrici di Precision e Recall in Figura 5.7. In primo luogo possiamo notare che esiste una quasi perfetta corrispondenza tra la comunità SBM 6 e la comunità 3 di Louvain: riprendendo osservazioni fatte in precedenza, si conclude che entrambi i metodi raggruppano in una comunità ben distinta le diagnosi della categoria MDC 14 (patologie legate alla gravidanza e al parto).
Un'altra evidenza è data dalla sostanziale coincidenza anche della comunità SBM 5 con la comunità 2 di Louvain: entrambe sono composte da diagnosi della categoria MDC 4 (patologie respiratorie).
Si può notare, inoltre, che la comunità 5 di Louvain riunisce le comunità

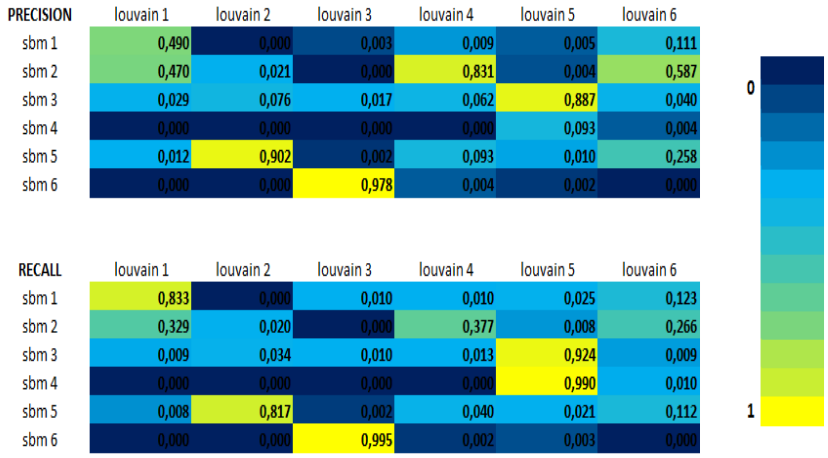

Figura 5.7: Matrici di Precision e Recall. Partizione con metodo SBM (6 comunità) vs partizione con Louvain (6 comunità).

SBM 3 e 4. Abbiamo osservato in precedenza che queste comunità hanno forte corrispondenza con la MDC 8: riconosciamo, dunque, che se il metodo di Louvain assegna alla stessa comunità quasi tutte le diagnosi ortopediche, il metodo SBM le suddivide su due gruppi.
Corrispondenze meno nette si registrano tra le altre comunità, che abbiamo visto essere costituite, in entrambe le partizioni, in prevalenza da diagnosi appartenenti alle categorie MDC 4, 5 e 11 che, per la natura degli organi coinvolti (cuore, polmoni e reni), tendono a presentare connessioni significative. Possiamo evidenziare, ad esempio, che la comunità SBM 2 riunisce una frazione significativa di diagnosi che nella partizione di Louvain sono assegnate alla comunità 1, 4 e 6 e che la comunità 1 di Louvain riunisce, invece, le comunità SBM 1 e 2.
Per interpretare in modo più preciso il significato dei raggruppamenti delle diagnosi individuati dai due metodi dal punto di vista clinico, è necessario entrare nel merito delle patologie coinvolte. É possibile, infatti, che i due metodi mettano in evidenza legami tra certe diagnosi che potrebbero risultare interpretabili prendendo in considerazione ulteriori livelli di analisi (ad esempio, i trattamenti utilizzati per la cura, l'età e il sesso dei pazienti coinvolti, il decorso e le eventuali complicanze delle patologie).

5.3 Classificazione dei nodi della Rete Diagnosi

In questa sezione applichiamo il metodo di classificazione dei nodi basato sull'analisi del coefficiente di partecipazione P e della stregth intra-comunità z, che abbiamo descritto nel Capitolo 4, alla Rete Diagnosi. Questo tipo di analisi consente di individuare il ruoli dei nodi in una partizione data. Ricordiamo che i 7 ruoli proposti in questa classificazione sono:

- Nodi *ultraperiferici*, Ruolo 1;
- Nodi *periferici*, Ruolo 2;
- Nodi *periferici connettori*, Ruolo 3;
- Nodi *periferici kinless*, Ruolo 4;
- Nodi *hub provinciali*, Ruolo 5;
- Nodi *hub connettori*, Ruolo 6;
- Nodi *hub kinless*, Ruolo 7;

Nel nostro caso, il metodo ci permette di approfondire l'analisi delle comunità effettuata nelle sezioni precedenti, distinguendo i ruoli delle diagnosi all'interno dei gruppi a cui sono assegnate. In particolare, il nostro interesse è volto a individuare i nodi che svolgono un ruolo di *hub*, cioè i nodi più centrali nella rete, e evidenziare se siano *provinciali*, cioè connessi per lo più con nodi appartenenti alla propria comunità, o *connettori* tra comunità diverse. Questa analisi completa quella svolta nelle precedenti sezioni e fornisce ulteriori elementi utili per identificare la struttura della rete e delle comunità individuate dai metodi utilizzati.
Di seguito presenteremo e commenteremo i risultati ottenuti per le partizioni di Louvain e SBM per $K = 5$ e 6, che abbiamo analizzato in modo dettagliato in questo capitolo, rimandando all'Appendice B per quelli relativi alle partizioni SBM per $K = 3, 4, 7, 8, 9$.
Per ogni partizione mostreremo il grafico $P - z$ in cui sono evidenziate le zone corrispondenti ai 7 ruoli della classificazione riportando i valori degli indici dei nodi. A ogni nodo rappresentato nel grafico è attribuito un colore, riportato in legenda, corrispondente alla comunità a cui appartiene, e una forma, che indica la classificazione in *hub* e *non-hub* (nodo periferico). Inoltre, riportiamo nelle Tabelle 5.19, 5.20 e 5.21 i nomi di alcune diagnosi che svolgono il ruolo di *hub* nelle comunità a cui appartengono nelle varie partizioni. In particolare, vengono proposte 5 diagnosi *hub connettori* e 5 diagnosi *hub provinciali*, unitamente alla loro categoria MDC, in modo da facilitare

l'interpretazione. Nelle Tabelle riassuntive 5.16,5.17 e 5.18 è presentato un quadro sintetico dei ruoli delle diagnosi nelle partizioni considerate.
A partire dalle Tabelle riassuntive, possiamo notare come primo elemento, evidente anche nei grafici 5.8, 5.9 e 5.10 , che i nodi *non hub*, siano essi connettori o provinciali (Ruolo 2 o Ruolo 3), sono in netta maggioranza in tutte le partizioni considerate, mentre i nodi *hub* costituiscono una percentuale molto piccola all'interno della rete. Questo fatto è dato dalla forte eterogeneità della Rete Diagnosi, costituita da molte diagnosi con frequenza e strength bassa. Le loro connessioni con altre diagnosi sono rare e, di conseguenza, lo sono anche le connessioni all'interno della comunità a cui sono assegnate, sulla base delle quali si calcola la within-community strength. L'indice z risulta, dunque, molto basso per questi nodi diagnosi.
É interessante, per completare la nostra analisi, mettere in evidenza quali siano i nodi più centrali nella rete e se le partizioni inviduate consentano di identificare diagnosi rilevanti all'interno delle comunità. Il ruolo di *hub provinciale* viene assegnato, infatti, alle diagnosi *leader* all'interno di una comunità, ed è utile per l'interpretazione della partizione in quanto consente di associare un gruppo a una o più diagnosi rappresentative. Il ruolo di *hub connettore*, invece, è dato a diagnosi assegnate a una data comunità ma che, allo stesso tempo, sono fortemente connesse a nodi di altre comunità. Queste diagnosi sono interessanti da mettere in evidenza in quanto risultano, spesso, patologie complesse e che coinvolgono organi o apparati diversi e, per questo, è utile valutare in modo approfondito da tipologia di connessioni di questi nodi.
Nella partizione ottenuta con il metodo di Louvain possiamo notare che le diagnosi a cui viene attribuito il ruolo di *hub provinciale* sono per lo più diagnosi appartenenti alle categorie MDC 14 e 8, legate rispettivamente alla gravidanza e all'ortopedia. Questo tipo di diagnosi risulta particolarmente frequente e, coerentemente con quanto osservato in precedenza, tende a essere connesso solo ad altre diagnosi della stessa categoria e, di conseguenza, date le caratteristiche di questa partizione, anche della stessa comunità. In Tabella 5.19 sono riportate alcune significative diagnosi con il ruolo di *hub connettore* per la propria comunità. Notiamo che diagnosi con elevata frequenza di occorrenza, come la polmonite, l'infarto e la fibrillazione atriale, svolgono un ruolo rilevante all'interno della rete fungendo da connettori tra le comunità, pur avendo un valore alto di strength interna al proprio gruppo. In riferimento alla partizione SBM con $K = 5$ si possono riscontrare aspetti simili a quelli evidenziati per la partizione di Louvain.
Nella partizione SBM con $K = 6$ vediamo, invece, che nessuna delle diagnosi relative alla MDC 8 svolge il ruolo di *hub provinciale*, che risulta assegnato solo ad alcune diagnosi, riportate in Tabella 5.21, relative alla MDC 14.

Questo fatto è coerente con la divisione delle diagnosi legate all'ortopedia in due comunità distinte che abbiamo in evidenza in precedenza.
É, infine, interessante sottolineare che in nessuna partizione vi sono nodi che il nostro metodo di classificazione attribuisce ai ruoli 4 e 7 che, come visto nel Capitolo 4, sono dati a nodi *kinless*, cioè nodi, periferici o hub, non assegnabili chiaramente ad alcuna comunità. Questa osservazione ci consente di confermare le buone performance del metodo di Louuvain e del metodo SBM, con questi valori di K.

LOUVAIN								
Ruolo	**1**	**2**	**3**	**4**	**5**	**6**	**totale**	**%**
hub connettori	6	8	-	6	2	5	27	0,90%
hub provinciali	1	-	6	-	6	1	14	0,46%
non hub connettori	245	317	43	207	597	170	1579	52,41%
non hub provinciali	95	146	594	12	497	49	1393	46,23%

Tabella 5.16: Quadro dei ruoli dei nodi nella partizione di Louvain.

SBM $K = 5$							
Ruolo	**1**	**2**	**3**	**4**	**5**	**totale**	**%**
hub connettori	11	6	2	4	5	28	0,93%
hub provinciali	-	3	4	3	1	11	0,37%
non hub connettori	461	265	53	688	170	1637	54,33%
non hub provinciali	189	111	597	392	48	1337	44,37%

Tabella 5.17: Quadro dei ruoli dei nodi nella partizione SBM $K = 5$.

SBM $K = 6$								
Ruolo	**1**	**2**	**3**	**4**	**5**	**6**	**totale**	**%**
hub connettori	7	8	8	1	9	-	33	1,10%
hub provinciali	-	-	4	3	-	6	13	0,43%
non hub connettori	182	400	765	79	353	34	1813	60,17%
non hub provinciali	15	88	280	21	158	592	1154	38,30%

Tabella 5.18: Quadro dei ruoli dei nodi nella partizione SBM $K = 6$.

Partizione Louvain con $K = 6$

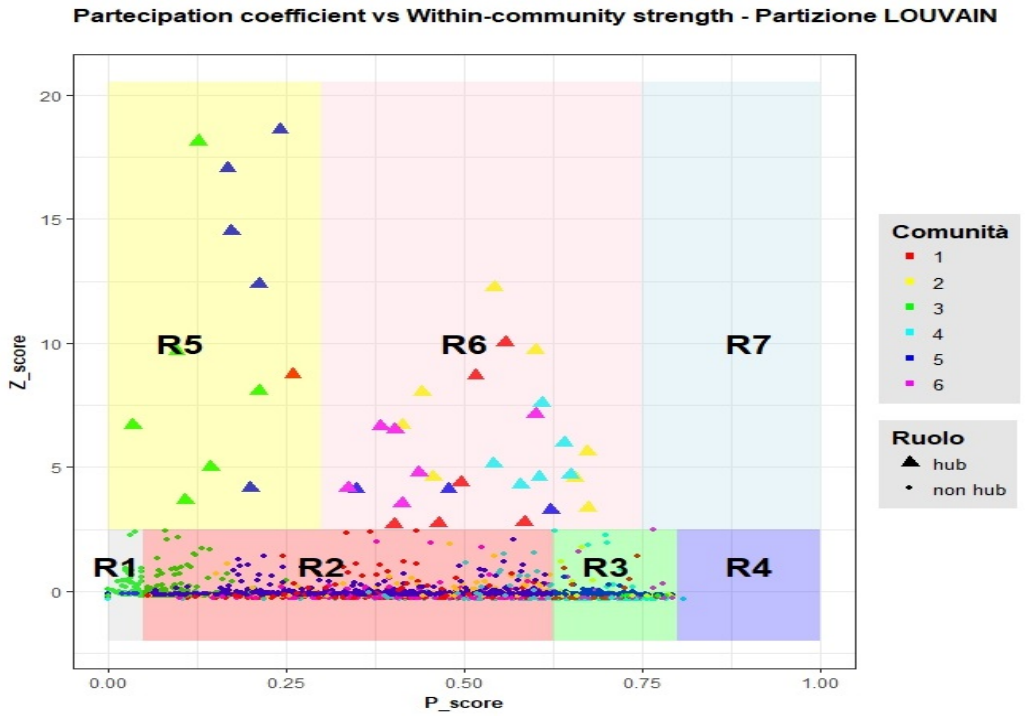

Figura 5.8: Grafico dei ruoli per i nodi assegnati alle 6 comunità individuate con il metodo Louvain.

Diagnosi	MDC	Comunità	Ruolo
Polmonite batterica	4	**2**	hub.conn
Broncopolmonite	4	**2**	hub.conn
Infarto Subendocardico	5	**1**	hub.conn
Fibrillazione Atriale	5	**1**	hub.conn
Malattia renale cronica	11	**4**	hub.conn
Postumi di fratture agli arti inferiori	8	**5**	hub.prov
Sostituzione di articolazione dell'anca	8	**5**	hub.prov
Aborto indotto	14	**3**	hub.prov
Minaccia di travaglio prematuro	14	**3**	hub.prov
Parto Normale	14	**3**	hub.prov

Tabella 5.19: Elenco di alcune diagnosi che svolgono ruolo di *hub provinciali* (Ruolo 5) e *hub connettori* (Ruolo 6) nelle comunità inviduate dal metodo di Louvain con $K = 6$.

Partizione SBM con $K = 5$

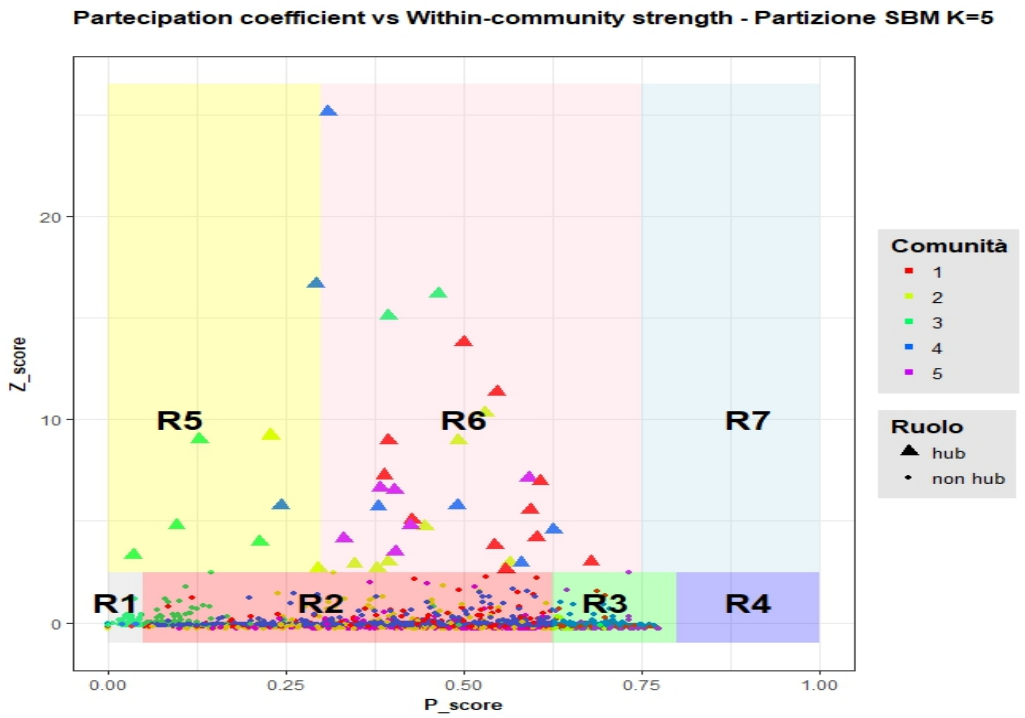

Figura 5.9: Grafico dei ruoli per i nodi assegnati alle 5 comunità individuate con il metodo SBM.

Diagnosi	MDC	Comunità	Ruolo
Infarto subendocardico	5	**2**	hub.conn
Insufficienza cardiaca congestizia	5	**2**	hub.conn
Insufficienza respiratoria	4	**1**	hub.conn
Tumori maligni vescica	11	**5**	hub.conn
Artrosi al ginocchio	8	**3**	hub.conn
Aterosclerosi coronarica	5	**2**	hub.prov
Aborto indotto	14	**3**	hub.prov
Minaccia di travaglio prematuraa	14	**3**	hub.prov
Parto normale	14	**3**	hub.prov
Sostituzione di articolazione dell'anca	8	**4**	hub.prov

Tabella 5.20: Elenco di alcune diagnosi che svolgono ruolo di *hub provinciali* (Ruolo 5) e *hub connettori* (Ruolo 6) nelle comunità inviduate dal metodo SBM con $K = 5$.

Partizione SBM con $K = 6$

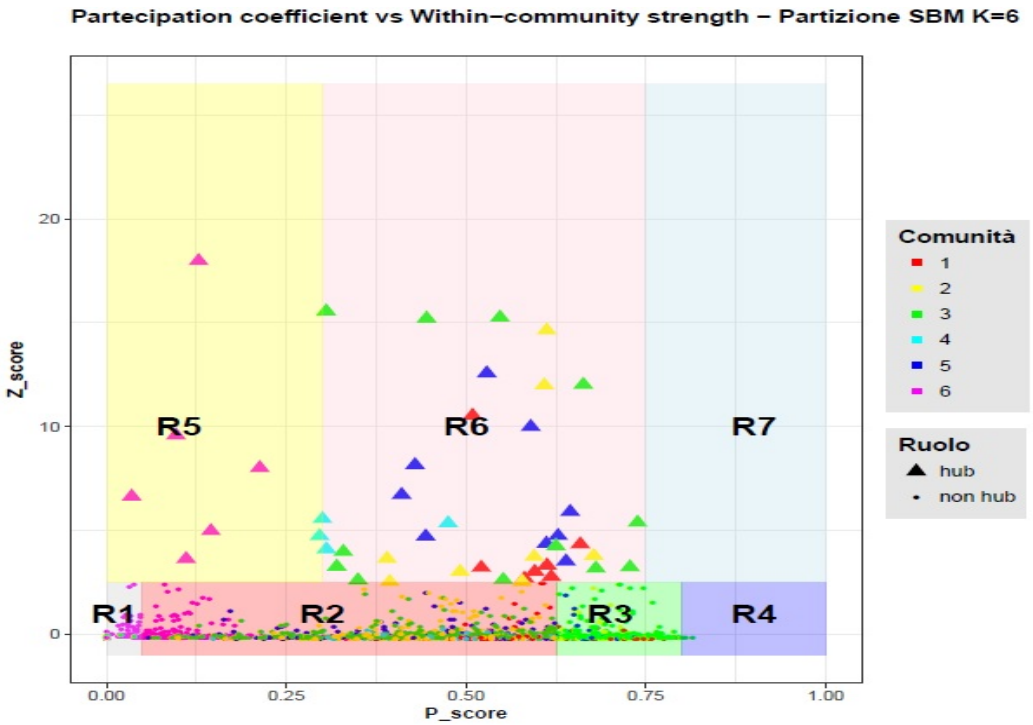

Figura 5.10: Grafico dei ruoli per i nodi assegnati alle 6 comunità individuate con il metodo SBM.

Diagnosi	MDC	Comunità	Ruolo
Polmonite	4	**5**	hub.conn
Tumori maligni della vescica	11	**5**	hub.conn
Fibrillazione atriale	5	**2**	hub.conn
Insufficienza cardiaca congestizia	5	**2**	hub.conn
Postumi di fratture degli arti inferiori	8	**3**	hub.conn
Aborto ritenuto	14	**6**	hub.prov
Aborto indotto	14	**6**	hub.prov
Minaccia di travaglio prematuro	14	**6**	hub.prov
Parto normale	14	**6**	hub.prov
Parto cesareo pregresso complicante la gravidanza	14	**6**	hub.prov

Tabella 5.21: Elenco di alcune diagnosi che svolgono ruolo di *hub provinciali* (Ruolo 5) e *hub connettori* (Ruolo 6) nelle comunità inviduate dal metodo SBM con $K = 6$.

5.4 Analisi di comunità della Rete Pazienti: Metodo di Louvain

In questa sezione presentiamo brevemente l'analisi di comunità effettuata sulla Rete Pazienti costruita nel Capitolo 3. In Tabella 5.22 riprendiamo le caratteristiche evidenziate per la particolare Rete Pazienti ottenuta dalla proiezione della rete bipartita dei ricoveri, campionata al 3%.

Rete Pazienti	
Numero di nodi	61551
Numero di nodi (massima componente connessa)	49011
Numero di archi	17595055
Numero di archi (massima componente connessa)	17328314
Densità	0,01
Grado medio dei nodi	707
Strength media dei nodi	711

Tabella 5.22: Informazioni generali sulla Rete Pazienti.

Come abbiamo evidenziato nel Capitolo 3, la Rete Pazienti presenta un numero elevato di componenti costituite da gruppi di nodi (da 2 a 338) connessi tra loro ma separati dal resto della rete. Questi gruppi, benchè nell'analisi tramite metodi di community detection ci si concentri solo sulla massima componente connessa, possono ritenersi a tutti gli effetti delle comunità.
Sulla massima componente connessa della rete applichiamo il metodo di Louvain, anche in questo caso utilizzando la funzione `cluster_louvain` disponibile nel pacchetto *igraph* di R [15] e documentata in Appendice C.
L'algoritmo raggruppa i 49011 nodi in 105 comunità. La comunità più piccola è costituita da 17 nodi e la più grande da 3161 e si registra un valore di cardinalità mediano dei gruppi di 338 nodi. Notiamo che l'algoritmo presenta una risoluzione elevata, in quanto è in grado di distinguere gruppi di nodi molto esigui anche a fronte di una cardinalità della rete alta.
Il valore di modularità della partizione ottenuta è di 0,88. Ricordiamo che il metodo di Louvain individua la partizione di massima modularità e sottolineiamo che, in questo caso, il valore ottenuto è decisamente alto, a indicare una struttura di comunità significativa e ben evidente all'interno della Rete Pazienti. Questo fatto è riconducibile alla logica con cui è stata costruita la rete proiezione. Ogni nodo paziente risulta, infatti, connesso a tutti i nodi paziente che in un loro ricovero presentano la stessa diagnosi. Queste con-

nessioni tra pazienti accomunati dallo stesso motivo di ricovero costuiscono i link interni alle comunità, che risultano per questo ben individuabili.
Il numero elevato di comunità individuate rende, tuttavia, difficile un'interpretazione efficace dei risultati: non disponiamo, infatti, di una classificazione nota dei pazienti che risulti confrontabile con la partizione ottenuta. Inoltre, all'interno di ogni comunità sono presenti pazienti con caratteristiche disomogenee in termini di età e genere, e che risultano ricoverati per un certo numero di diagnosi diverse. Non è possibile, quindi, attribuire alle comunità delle caratteristiche distintive con le informazioni a nostra disposizione. Un'analisi condotta con i metodi di valutazione a posteriori della partizione applicati per la Rete Diagnosi non risulta praticabile ed è quindi necessario mettere a punto tecniche diverse per interpretare utilmente questi risultati.
Un aspetto importante da sottolineare è l'efficienza del metodo di Louvain, anche su una rete di dimensioni consistenti come quella in esame, che lo rende un interessante strumento su cui basare analisi più approfondite con altri metodi.
L'applicazione del modello stocastico a blocchi sulla Rete Pazienti non è stato possibile a causa dell'elevato costo computazionale richiesto dalle dimensioni della rete considerata. Inoltre, sottolineiamo che per effettuare un confronto tra i risultati ottenuti dal metodo di Louvain e dal metodo SBM, come fatto per la Rete Diagnosi, è necessario calcolare una partizione con più di 100 comunità. Questo aspetto, come abbiamo precisato nella sezione precedente, risulta essere critico per l'efficienza del metodo.

Conclusioni e sviluppi futuri

In questa tesi abbiamo presentato l'analisi di un dataset amministrativo relativo alle ospedalizzazioni nella Regione Lombardia, sviluppata utilizzando metodi relativi alle reti complesse.
Questo tipo di approccio allo studio di dati relativi a ricoveri ospedalieri ha consentito, in primo luogo, di analizzare la struttura e le caratteristiche del sistema sanitario regionale dal punto di vista delle *relazioni* tra pazienti e diagnosi, modellizzate tramite una rete complessa. In particolare, il focus principale del lavoro è stata l'analisi della Rete Diagnosi, ottenuta per proiezione dalla rete originale, con l'obiettivo di identificare comunità significative date da connessioni rilevanti tra le patologie.

Un primo obiettivo del nostro lavoro è stato quello di testare due metodi di community detection, il metodo di Louvain e il metodo SBM, sulla Rete Diagnosi. I risultati hanno, in generale, confermato la capacità dei metodi di inviduare comunità significative e, dunque, possiamo valutare positivamente le performance di entrambi. Sottolineiamo che il metodo di Louvain offre un'efficienza di gran lunga maggiore rispetto al metodo SBM che, tuttavia, si rivela molto vantaggioso per la sua flessibilità, in quanto ci consente di stabilire, a seconda delle esigenze di analisi, la cardinalità della partizione da individuare.
Con gli strumenti di analisi utilizzati, abbiamo ottenuto partizioni significative che ricostruiscono in modo piuttosto preciso la classificazione delle diagnosi, nota e basata sul criterio clinico delle categorie MDC. L'analisi di questo dataset tramite i metodi di community detection applicati alla rete consente, tuttavia, anche di ottenere risultati meno scontati. É possibile, infatti, mettere in evidenza il tipo di connessioni esistenti tra le diagnosi classificate in una stessa categoria e le relazioni esistenti tra questi gruppi. Questi strumenti di analisi, non supervisionati e non vincolati alle classificazioni esterne, possono essere utilizzati nella profilazione dei pazienti ricoverati e per una gestione delle patologie non esclusivamente basata su criteri di tipo anatomico. Come abbiamo avuto modo di evidenziare, infatti, gli algoritmi

individuano, in alcuni casi, comunità formate da diagnosi appartenenti a categorie diverse. L'interpretazione a posteriori di queste comunità "miste", che in questo lavoro abbiamo individuato e per cui abbiamo impostato un metodo di analisi, può essere approfondita a vari livelli, a seconda dell'interesse clinico e delle informazioni di cui si dispone. Alcuni aspetti che possono utilmente essere presi in considerazione per interpretare i gruppi di diagnosi sono l'età dei pazienti ricoverati per le diagnosi in una certa comunità, informazioni relative ai trattamenti di cura e ad eventuali complicanze registrate in modo frequente.
Il metodo di classificazione dei nodi che abbiamo utilizzato, infine, fornisce uno strumento per focalizzare l'analisi sulle diagnosi più importanti a livello dell'intera rete e all'interno delle varie comunità. Questo approccio si rivela molto utile, soprattutto se, come nel nostro caso, la rete è fortemente eterogenea.

In generale, possiamo concludere che il nostro lavoro costituisce un punto di partenza per analisi future volte a caratterizzare, sulla base del loro status clinico, profili di rischio non già noti per i pazienti ospedalizzati. Sottolineiamo l'importanza di un confronto e della supervisione da parte di clinici per favorire un'impostazione adeguata delle analisi e un'interpretazione corretta dei risultati.
Evidenziamo di seguito alcuni possibili sviluppi del lavoro.
In primo luogo, è possibile prendere in considerazione un dataset che comprenda ricoveri relativi a diagnosi di altre categorie MDC di interesse. Un ampliamento dei dati a disposizione consentirebbe di mettere in evidenza relazioni interessanti tra altri tipi di patologie e fornirebbe come risultato un quadro più completo del problema.
Un'interessante estensione del nostro lavoro dal punto di vista modellistico può essere ottenuta utilizzando le informazioni relative alle diagnosi secondarie, che in questa tesi sono state escluse dall'analisi. In questo caso, la Rete Ricoveri sarebbe costituita da nodi pazienti connessi, con archi pesati in modo proporzionale all'importanza della diagnosi nel ricovero, a nodi diagnosi di tipo diverso. Questo tipo di impostazione, più complessa di quella proposta nel nostro lavoro, consentirebbe di caratterizzare le relazioni tra le diagnosi all'interno dello stesso ricovero e di considerare quadri clinici che comprendano anche le complicanze legate alle patologie principali.
In questo lavoro l'analisi delle Rete Pazienti è stata svolta, per problemi legati ai costi computazionali degli algoritmi, solo su una sottorete campionata. Disponendo di strumenti di calcolo più adeguati, sarebbe interessante approfondire le relazioni tra i pazienti presenti nell'intero dateset, in modo da poter interpretare in modo più preciso i gruppi individuati.

Un ulteriore sviluppo di questo lavoro può essere l'analisi della rete dei ricoveri nella sua forma bipartita originale utilizzando metodi SBM sviluppati *ad hoc* per questo tipo di strutture. A questo scopo, è possibile fare riferimento all'articolo di D. Larremore, A. Clauset e A. Jacobs [22], che sviluppa un'estensione dei metodi SBM in grado di gestire l'analisi di comunità su reti bipartite. L'algoritmo proposto, implementato in un codice disponibile online in [1], è in grado di individuare comunità nei due insiemi di nodi, considerati in modo separato. Nella Rete Ricoveri sarebbe possibile, dunque, ottenere simultaneamente gruppi di soli nodi diagnosi e gruppi di soli nodi pazienti. I vantaggi dell'utilizzo di questo metodo sono la maggiore efficienza computazionale, dovuta alla separazione del problema di community detection in due problemi distinti sui due insiemi di nodi, e, soprattutto, la possibilità di analizzare la rete senza costruire la sua proiezione.

Appendici

Appendice A

Analisi delle comunità individuate con il metodo SBM

Riportiamo in questa Appendice i risultati ottenuti dall'analisi delle comunità individuate con il metodo SBM per le partizioni di cardinalità K pari a 3, 4, 7, 8, 9.
Come abbiamo fatto nel Capitolo 5, presentiamo per ogni partizione le tabelle relative alla numerosità e alla probabilità di persistenza dei gruppi, i valori di NMI e VI relativi alla partizione in categorie MDC, i grafici delle distribuzioni del numero di ricoveri e della strength e le matrici di Precision e Recall, anch'esse relative alla partizione in categorie MDC. Le matrici ω dei valori attesi del numero di archi tra i nodi delle comunità sono riportate alla fine dell'Appendice.

Sottolineiamo che la scelta del numero di comunità da individuare dipende fortemente dal problema considerato. Nel nostro caso, abbiamo valutato partizioni che garantissero un costo computazionale dell'algoritmo accettabile e che avessero cardinalità confrontabile con quella di altre suddivisioni della rete note (la classificazione MDC e la partizione ottenuta con il metodo di Louvain).
Con le informazioni a nostra disposizione, le partizioni di seguito riportate risultano poco interpretabili in un'analisi a posteriori.
Forniamo, quindi, solo alcune sintetiche considerazioni su questi risultati, utili a evidenziare le problematicità dovute alla scelta di K e le caratteristiche più rilevanti delle comunità individuate.

Nella prima partizione, di cardinalità $K = 3$, i risultati evidenziano che il numero di comunità è troppo basso per descrivere in modo efficace la struttura della rete. Si può notare, infatti, che le matrici di Precision e Recall in

Figura A.2 non presentano alcun valore significativamente alto. É possibile, tuttavia, basare un'ipotesi di interpretazione delle comunità individuate sui risultati mostrati nei grafici in Figura A.1. Si può notare, infatti, che le 3 comunità presentano distribuzioni di numero di ricoveri e di strength differenti. La prima comunità, la più numerosa, è formata da nodi diagnosi con numero di occorrenze e di strength concentrati su valori inferiori al valore mediano della rete, al contrario, nella comunità 2 le due distribuzioni presentano valori mediani più alti. La comunità 3, che notiamo essere meno coesa, presenta una distribuzione centrata sul valore mediano dell'intera rete. Possiamo quindi interpretare le tre comunità come gruppi di diagnosi con caratteristiche simili per frequenza di occorrenza e strength.
Un problema di interpretazione delle comunità emerge anche per le partizioni con $K = 4$ e $K = 7$. Nel primo caso si hanno gruppi in numero inferiore rispetto a quello delle categorie MDC con cui le confrontiamo, nel secondo si ha la situazione opposta. In entrambi i casi, l'analisi delle matrici di Precision e Recall (in Figura A.4 e A.6) può suggerire interpretazioni dei raggruppamenti individuati, che risultano, come si può osservare considerando le Tabelle delle probabilità di persistenza A.5 e A.8, tutti significativi. In queste partizioni è possibile evidenziare corrispondenze precise tra alcune comunità e alcune categorie MDC (ad esempio, considerando la Figura A.4, vediamo che la comunità 2 e la categoria MDC 14 coincidono). É utile analizzare questi risultati mettendo in evidenza l'aggregazione da parte dell'algoritmo di nodi appartenenti a due categorie MDC nella stessa comunità (per la partizione con $K = 4$) e viceversa, nella partizione con $K = 7$ come una categoria MDC risulti distribuita su due o più gruppi. Ad esempio, possiamo notare in Figura A.4 che la comunità 4 è costuita quasi interamente diagnosi della categoria MDC 4 e 5 e in Figura A.6 che la categoria MDC 8 sia suddivisa nei gruppi 1,6 e 7.
Possiamo evidenziare, invece, che per le partizioni con $K = 8$ e $K = 9$ vengono individuate alcune comunità con valore di persistenza basso (si vedano le Tabelle A.11 e A.14) e che, per questo, la suddivisione in 8 e 9 gruppi imposta dall'algoritmo risulta poco significativa.
Le matrici ω, infine, consentono di valutare il tipo di connessioni esistenti all'interno delle comunità e tra di esse e sono riportate nelle Tabelle A.16, A.17, A.18, A.19 e A.20.

Partizione SBM $K = 3$		
Comunità	**Cardinalità**	**%**
NA	266	8,11%
1	1210	40,16%
2	847	28,11%
3	956	31,73%

Tabella A.1: Cardinalità delle comunità individuate con il metodo SBM con $K = 3$ e percentuali di nodi attribuiti a ciascuna comunità.

Partizione SBM $K = 3$		
Comunità	**Probabilità di persistenza**	**z**
1	0,83	6,24
2	0,88	11,03
3	0,57	7,33

Tabella A.2: Probabilità di persistenza delle comunità individuate con il metodo SBM con $K = 3$ e relativo z-score. Tutte le probabilità di persistenza sono evidenziate in quanto maggiori di 0,5.

MDC vs Partizione Sbm $K = 3$		
	Valore calcolato	**z**
NMI	0,375	663,16
VI	0,197	-663,16

Tabella A.3: Valori di NMI e VI e relativi z-score per il confronto tra la classificazione MDC e la partizione ottenuta con il metodo SBM con $K = 3$.

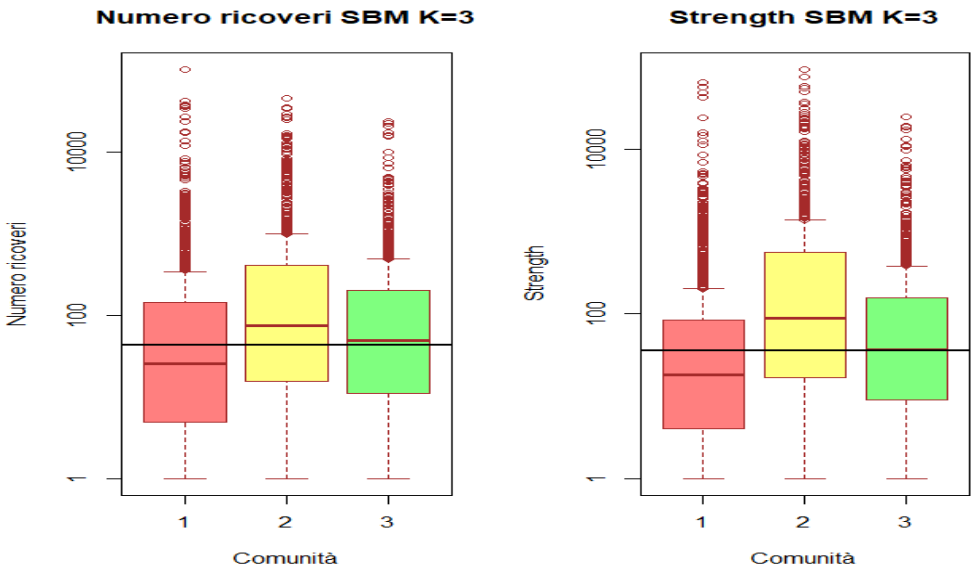

Figura A.1: Distribuzione del numero di ricoveri e strength nelle comunità individuate con il metodo di SBM con $K = 3$.

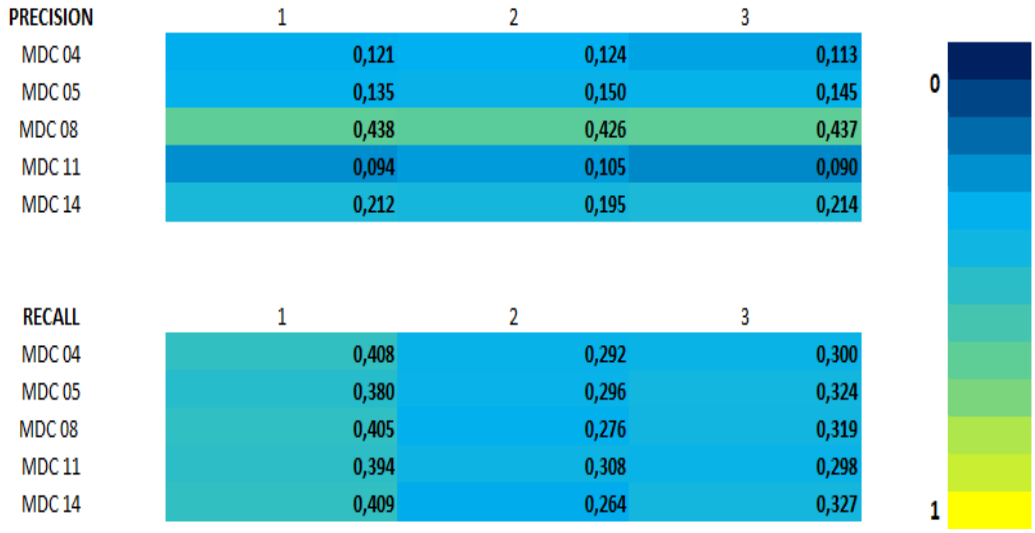

Figura A.2: Matrici di Precision e Recall. Classificazione MDC vs Partizione con metodo SBM (3 comunità).

Partizione SBM $K = 4$		
Comunità	**Cardinalità**	**%**
NA	266	8,11%
1	497	16,50%
2	646	21,44%
3	994	32,99%
4	876	29,07%

Tabella A.4: Cardinalità delle comunità individuate con il metodo SBM con $K = 4$ e percentuali di nodi attribuiti a ciascuna comunità.

Partizione SBM $K = 4$		
Comunità	**Probabilità di persistenza**	**z**
1	0,62	7,66
2	0,95	14,78
3	0,79	8,06
4	0,82	10,45

Tabella A.5: Probabilità di persistenza delle comunità individuate con il metodo SBM con $K = 4$ e relativo z-score. Tutte le probabilità di persistenza sono evidenziate in quanto maggiori di 0,5.

MDC vs Partizione Sbm $K = 4$		
	Valore calcolato	**z**
NMI	0,592	969,95
VI	0,142	-969,95

Tabella A.6: Valori di NMI e VI e relativi z-score per il confronto tra la classificazione MDC e la partizione ottenuta con il metodo SBM con $K = 4$.

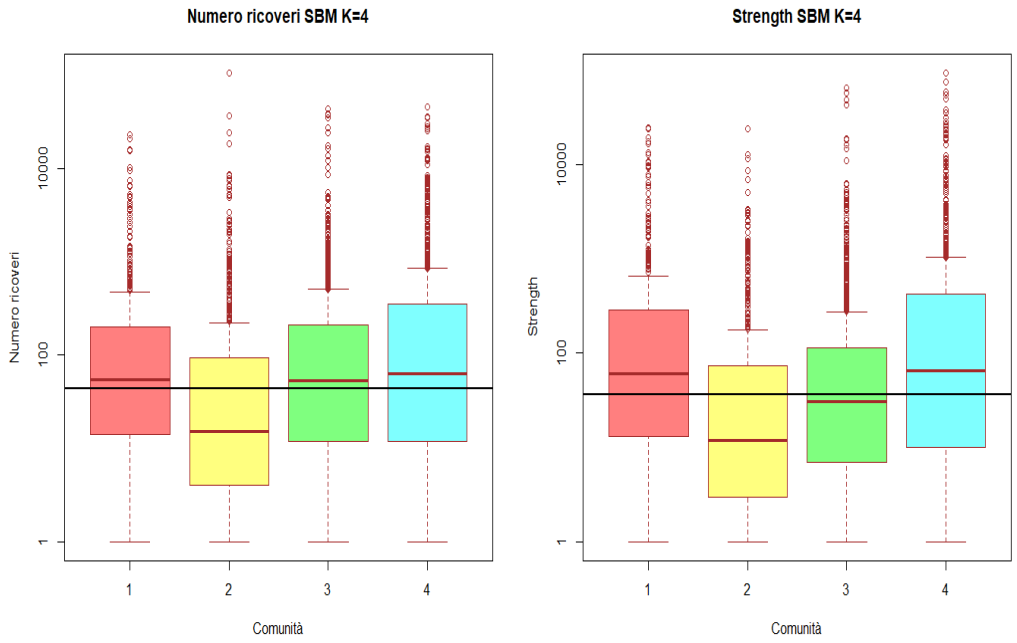

Figura A.3: Distribuzione del numero di ricoveri e strength nelle comunità individuate con il metodo di SBM con $K = 4$.

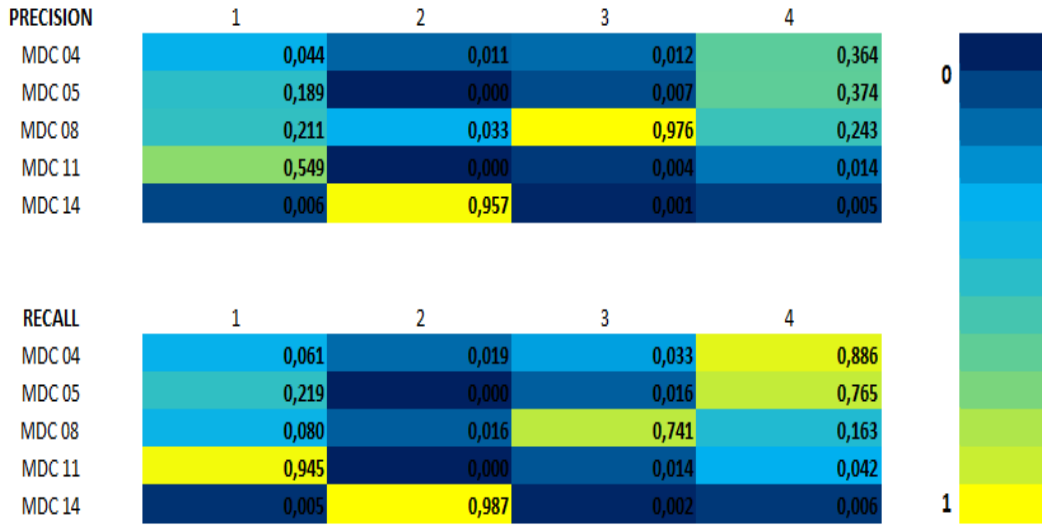

Figura A.4: Matrici di Precision e Recall. Classificazione MDC vs Partizione con metodo SBM (4 comunità).

Partizione SBM $K = 7$		
Comunità	**Cardinalità**	**%**
NA	266	8,11%
1	994	30,31%
2	523	15,95%
3	631	19,24%
4	222	6,77%
5	480	14,64%
6	64	1,95%
7	99	3,02%

Tabella A.7: Cardinalità delle comunità individuate con il metodo SBM con $K = 7$ e percentuali di nodi attribuiti a ciascuna comunità.

Partizione SBM $K = 7$		
Comunità	**Probabilità di persistenza**	**z**
1	0,57	4,73
2	0,76	13,50
3	0,95	15,50
4	0,64	12,64
5	0,60	12,06
6	0,71	35,87
7	0,66	42,01

Tabella A.8: Probabilità di persistenza delle comunità individuate con il metodo SBM con $K = 7$ e relativo z-score. Tutte le probabilità di persistenza sono evidenziate in quanto maggiori di 0,5.

MDC vs Partizione Sbm $K = 7$		
	Valore calcolato	**z**
NMI	0,618	-802,41
VI	0,148	-802,41

Tabella A.9: Valori di NMI e VI e relativi z-score per il confronto tra la classificazione MDC e la partizione ottenuta con il metodo SBM con $K = 7$.

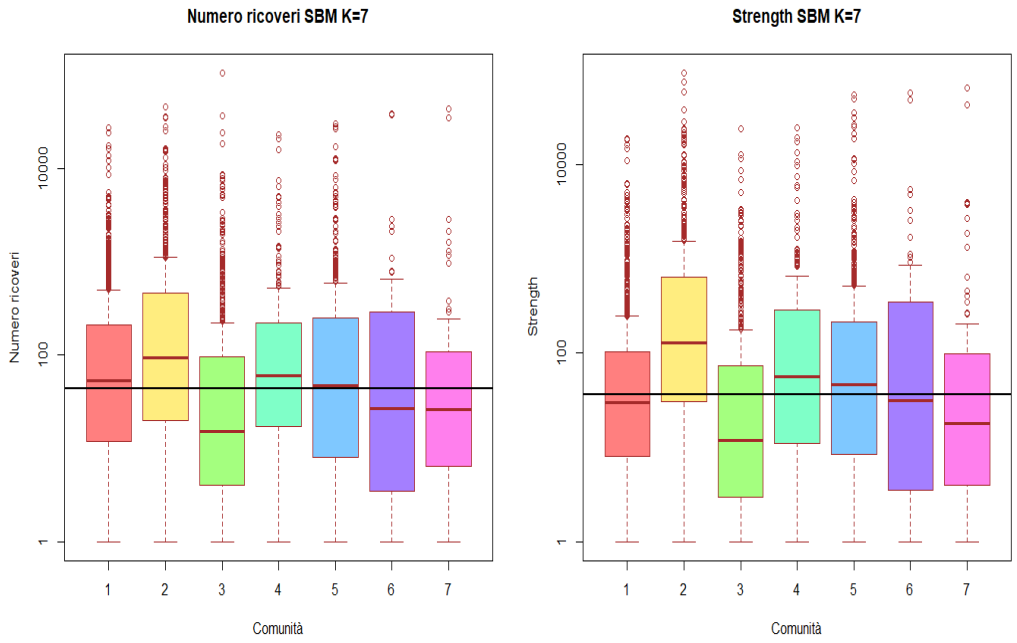

Figura A.5: Distribuzione del numero di ricoveri e strength nelle comunità individuate con il metodo di SBM con $K = 7$.

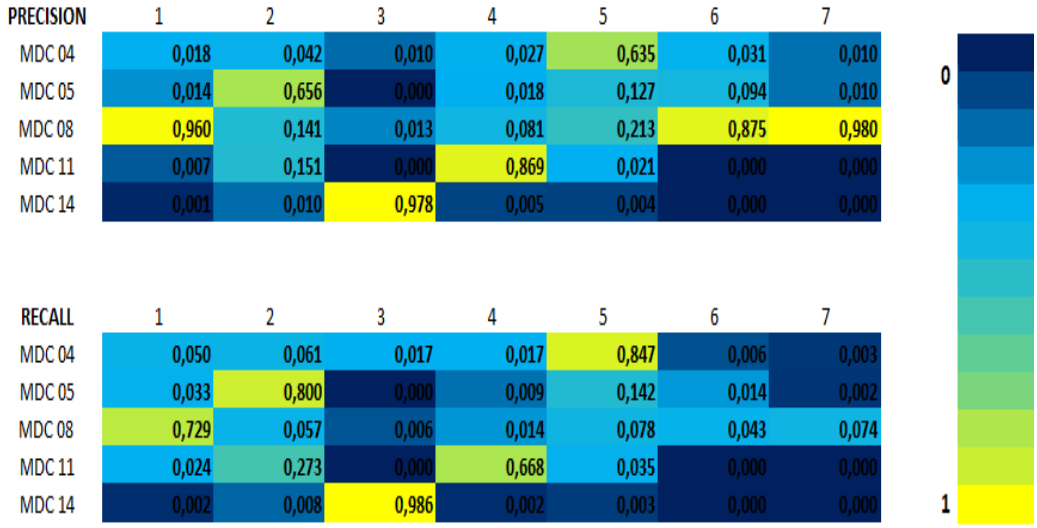

Figura A.6: Matrici di Precision e Recall. Classificazione MDC vs Partizione con metodo SBM (7 comunità).

Partizione SBM $K = 8$		
Comunità	**Cardinalità**	**%**
NA	266	8,11%
1	28	0,93%
2	220	7,30%
3	1072	35,58%
4	631	20,94%
5	446	14,80%
6	223	7,40%
7	62	2,06%
8	331	10,99%

Tabella A.10: Cardinalità delle comunità individuate con il metodo SBM con $K = 8$ e percentuali di nodi attribuiti a ciascuna comunità.

Partizione SBM $K = 8$		
Comunità	**Probabilità di persistenza**	**z**
1	0,02	3,51
2	0,64	13,86
3	0,58	4,56
4	0,95	14,23
5	0,60	10,44
6	0,58	21,99
7	0,09	1,80
8	0,53	9,85

Tabella A.11: Probabilità di persistenza delle comunità individuate con il metodo SBM con $K = 8$ e relativo z-score. Le probabilità di persistenza sono evidenziate se maggiori di 0,5.

MDC vs Partizione Sbm $K = 8$		
	Valore calcolato	**z**
NMI	0,608	739,05
VI	0,155	739,05

Tabella A.12: Valori di NMI e VI e relativi z-score per il confronto tra la classificazione MDC e la partizione ottenuta con il metodo SBM con $K = 8$.

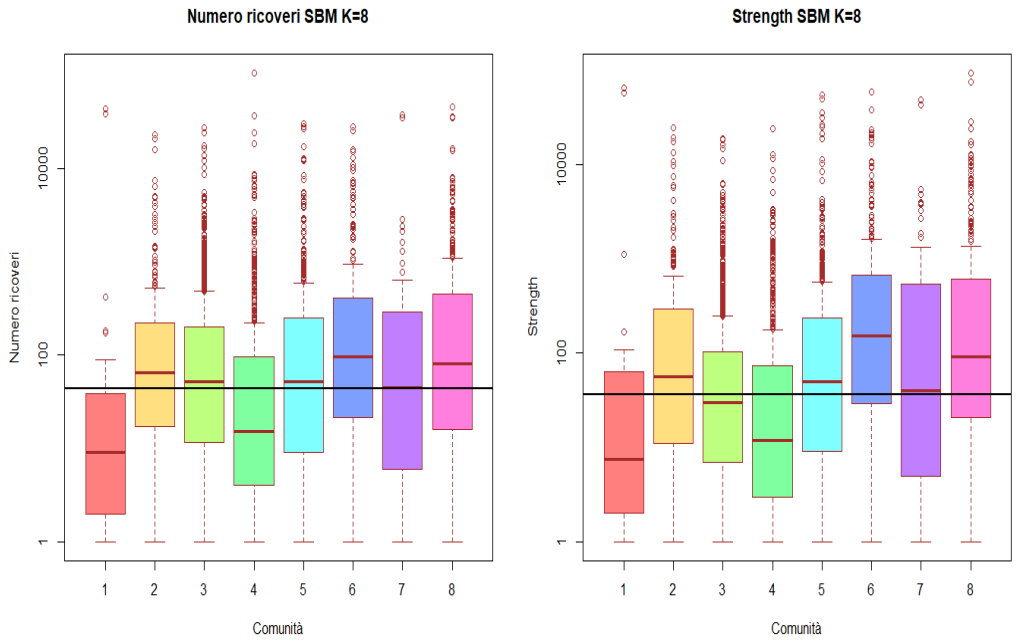

Figura A.7: Distribuzione del numero di ricoveri e strength nelle comunità individuate con il metodo di SBM con $K = 8$.

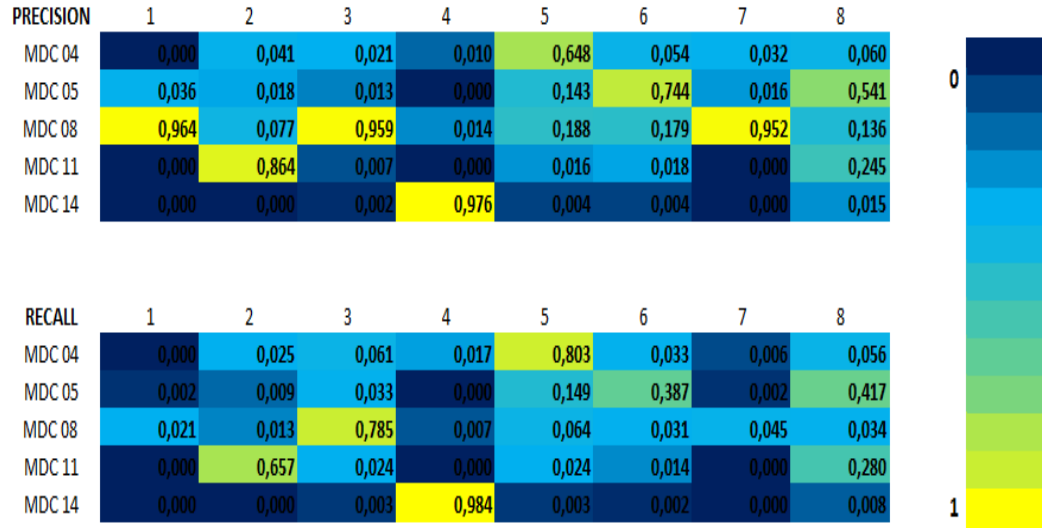

Figura A.8: Matrici di Precision e Recall. Classificazione MDC vs Partizione con metodo SBM (8 comunità).

Partizione SBM $K = 9$		
Comunità	**Cardinalità**	**%**
NA	266	8,11%
1	34	1,13%
2	1001	33,22%
3	233	7,73%
4	231	7,67%
5	632	20,98%
6	224	7,43%
7	83	2,75%
8	158	5,24%
9	417	13,84%

Tabella A.13: Cardinalità delle comunità individuate con il metodo SBM con $K = 9$ e percentuali di nodi attribuiti a ciascuna comunità.

Partizione SBM $K = 9$		
Comunità	**Probabilità di persistenza**	**z**
1	0,72	64,37
2	0,57	4,40
3	0,50	13,94
4	0,48	15,04
5	0,95	15,98
6	0,64	11,09
7	0,65	44,84
8	0,55	21,38
9	0,60	9,21

Tabella A.14: Probabilità di persistenza delle comunità individuate con il metodo SBM con $K = 9$ e relativo z-score. Le probabilità di persistenza sono evidenziate se maggiori di 0,5.

MDC vs Partizione Sbm $K = 9$		
	Valore calcolato	**z**
NMI	0,586	702,91
VI	0,170	702,91

Tabella A.15: Valori di NMI e VI e relativi z-score per il confronto tra la classificazione MDC e la partizione ottenuta con il metodo SBM con $K = 9$.

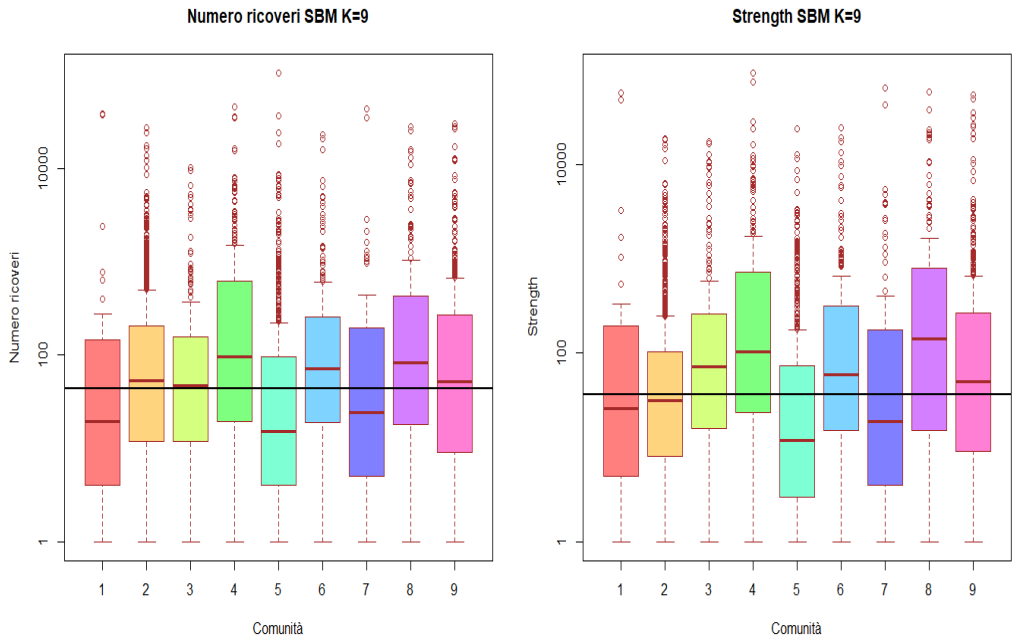

Figura A.9: Distribuzione del numero di ricoveri e strength nelle comunità individuate con il metodo di SBM con $K = 9$.

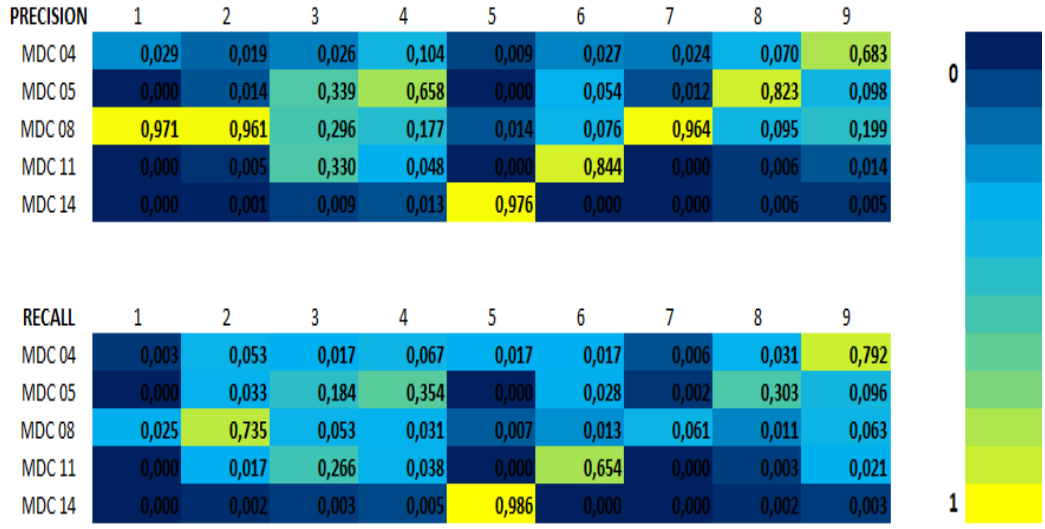

Figura A.10: Matrici di Precision e Recall. Classificazione MDC vs Partizione con metodo SBM (9 comunità).

matrice ω			
	com 1	com 2	com 3
com 1	**0,292**	**0,037**	**0,043**
com 2	**0,037**	**1,627**	**0,150**
com 3	**0,043**	**0,150**	**0,246**

Tabella A.16: Matrice ω della partizione SBM con $K = 3$: l'elemento ω_{rs} rappresenta il numero atteso di archi tra un nodo della comunità r e un nodo della comunità s. Il colore della cella è indicativo del valore in essa contenuto.

matrice ω				
	com 1	com 2	com 3	com 4
com 1	**1,013**	**0,009**	**0,048**	**0,285**
com 2	**0,009**	**0,366**	**0,005**	**0,004**
com 3	**0,048**	**0,005**	**0,406**	**0,094**
com 4	**0,285**	**0,004**	**0,094**	**1,253**

Tabella A.17: Matrice ω della partizione SBM con $K = 4$: l'elemento ω_{rs} rappresenta il numero atteso di archi tra un nodo della comunità r e un nodo della comunità s. Il colore della cella è indicativo del valore in essa contenuto.

matrice ω							
	com 1	com 2	com 3	com 4	com 5	com 6	com 7
com 1	**0,157**	**0,084**	**0,005**	**0,042**	**0,062**	**0,215**	**0,174**
com 2	**0,084**	**2,510**	**0,004**	**0,300**	**0,493**	**0,225**	**0,181**
com 3	**0,005**	**0,004**	**0,382**	**0,018**	**0,005**	**0,002**	**0,002**
com 4	**0,042**	**0,300**	**0,018**	**2,492**	**0,170**	**0,125**	**0,097**
com 5	**0,062**	**0,493**	**0,005**	**0,170**	**1,159**	**0,117**	**0,100**
com 6	**0,215**	**0,225**	**0,002**	**0,125**	**0,117**	**22,865**	**1,831**
com 7	**0,174**	**0,181**	**0,002**	**0,097**	**0,100**	**1,831**	**8,755**

Tabella A.18: Matrice ω della partizione SBM con $K = 7$: l'elemento ω_{rs} rappresenta il numero atteso di archi tra un nodo della comunità r e un nodo della comunità s. Il colore della cella è indicativo del valore in essa contenuto.

matrice ω								
	com 1	com 2	com 3	com 4	com 5	com 6	com 7	com 8
com 1	**3,722**	**0,293**	**0,572**	**0,002**	**0,294**	**0,553**	**51,911**	**0,525**
com 2	**0,293**	**2,537**	**0,041**	**0,018**	**0,175**	**0,252**	**0,143**	**0,321**
com 3	**0,572**	**0,041**	**0,141**	**0,005**	**0,057**	**0,072**	**0,208**	**0,086**
com 4	**0,002**	**0,018**	**0,005**	**0,382**	**0,005**	**0,002**	**0,002**	**0,004**
com 5	**0,294**	**0,175**	**0,057**	**0,005**	**1,297**	**0,345**	**0,117**	**0,585**
com 6	**0,553**	**0,252**	**0,072**	**0,002**	**0,345**	**4,641**	**0,263**	**1,304**
com 7	**51,911**	**0,143**	**0,208**	**0,002**	**0,117**	**0,263**	**3,178**	**0,229**
com 8	**0,525**	**0,321**	**0,086**	**0,004**	**0,585**	**1,304**	**0,229**	**2,543**

Tabella A.19: Matrice ω della partizione SBM con $K = 8$: l'elemento ω_{rs} rappresenta il numero atteso di archi tra un nodo della comunità r e un nodo della comunità s. Il colore della cella è indicativo del valore in essa contenuto.

matrice ω									
	com 1	com 2	com 3	com 4	com 5	com 6	com 7	com 8	com 9
com 1	**71,1427**	**0,2812**	**0,0781**	**0,4506**	**0,0007**	**0,2025**	**4,0726**	**0,4227**	**0,1695**
com 2	**0,2812**	**0,1547**	**0,0336**	**0,1189**	**0,0051**	**0,0423**	**0,2503**	**0,0741**	**0,0633**
com 3	**0,0781**	**0,0336**	**1,5594**	**0,5729**	**0,0015**	**0,1640**	**0,0704**	**0,4904**	**0,1719**
com 4	**0,4506**	**0,1189**	**0,5729**	**4,0030**	**0,0050**	**0,3608**	**0,3032**	**1,9048**	**0,8224**
com 5	**0,0007**	**0,0051**	**0,0015**	**0,0050**	**0,3809**	**0,0183**	**0,0024**	**0,0033**	**0,0050**
com 6	**0,2025**	**0,0423**	**0,1640**	**0,3608**	**0,0183**	**2,4838**	**0,1237**	**0,2586**	**0,1848**
com 7	**4,0726**	**0,2503**	**0,0704**	**0,3032**	**0,0024**	**0,1237**	**13,8723**	**0,2849**	**0,1375**
com 8	**0,4227**	**0,0741**	**0,4904**	**1,9048**	**0,0033**	**0,2586**	**0,2849**	**6,8260**	**0,3765**
com 9	**0,1695**	**0,0633**	**0,1719**	**0,8224**	**0,0050**	**0,1848**	**0,1375**	**0,3765**	**1,4680**

Tabella A.20: Matrice ω della partizione SBM con $K = 9$: l'elemento ω_{rs} rappresenta il numero atteso di archi tra un nodo della comunità r e un nodo della comunità s. Il colore della cella è indicativo del valore in essa contenuto.

Appendice B

Classificazione dei nodi nelle comunità individuate con il metodo SBM

In questa Appendice riportiamo i grafici dei ruoli dei nodi nelle partizioni con $K = 3, 4, 7, 8, 9$. Per ogni partizione possiamo individuare le diagnosi che svolgono un ruolo centrale nella rete e per ogni comunità evidenziare, se esistono, quali sono le diagnosi con tale ruolo. Notiamo che, come abbiamo evidenziato per le classificazioni illustrate nel dettaglio nel Capitolo 5, la maggior parte dei nodi della rete ha un ruolo periferico, mentre meno numerosi sono i nodi *hub*.
A titolo di esempio, nelle Tabelle nelle pagine seguenti vengono elencate 5 diagnosi che svolgono il ruolo di *hub connettori* (Ruolo 6) e altrettante che svolgono il ruolo di *hub provinciali* (Ruolo 5), con la relativa comunità a cui sono assegnati nelle corrispondenti partizioni e la categoria MDC di appartenenza. La selezione delle diagnosi da presentare ha favorito le patologie più conosciute e che consentissero di mettere in evidenza alcuni aspetti interessanti. Un aspetto che si può cogliere, senza entrare nel merito di questioni strettamente cliniche in modo approfondito, è che i nodi *hub provinciali*, cioè i nodi centrali strettamente legati solo a nodi della comunità di appartenenza, rappresentano soprattutto diagnosi appartenenti alle comunità che corrispondono agli MDC 8 e 11, come vediamo dai grafici e dalle tabelle relative alle partizioni con $K = 4, 8, 9$ (Figure B.2, B.4, B.5 e Tabelle B.2, B.4 e B.5). Notiamo, infine, che in queste partizioni, come abbiamo visto per quelle analizzate nel Capitolo 5, non si assegnano nodi ai ruoli 4 e 7.

Partizione SBM con $K = 3$

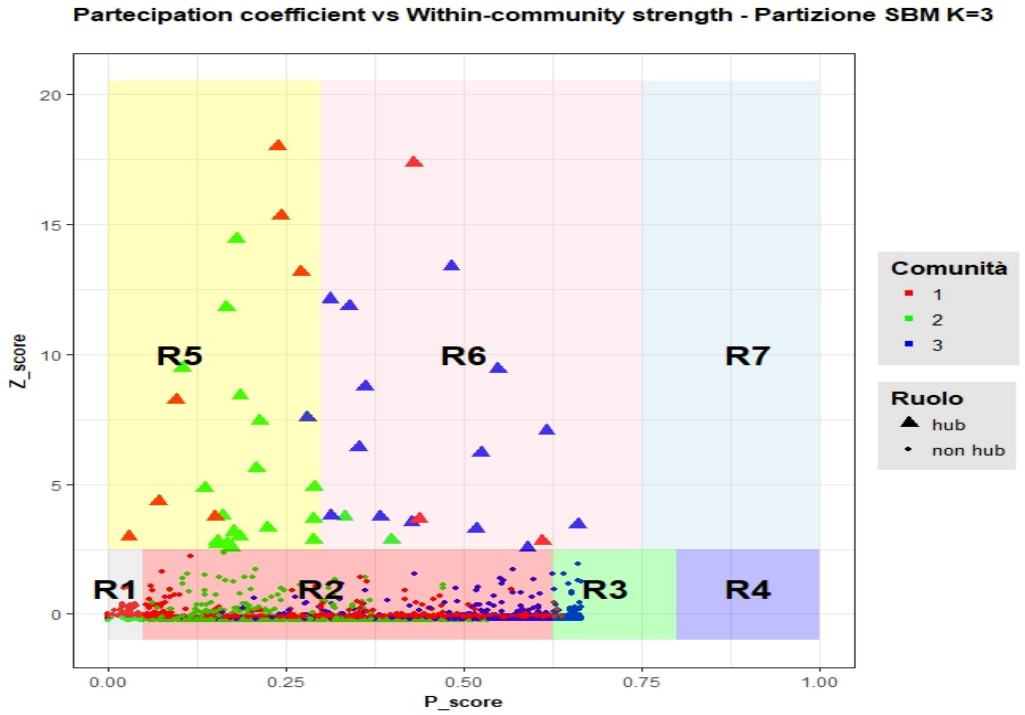

Figura B.1: Grafico dei ruoli per i nodi assegnati alle 3 comunità individuate con il metodo SBM.

Diagnosi	MDC	Comunità	Ruolo
Emangioma di sede non specificata	5	**3**	hub.conn
Cure post partum	14	**1**	hub.conn
Artropatia al ginocchio	8	**2**	hub.conn
Artrite allergica	8	**3**	hub.conn
Lussazione di articolazioni	8	**3**	hub.conn
Insufficienza cardiaca diastolica	5	**2**	hub.prov
Postumi di infarto miocardico	5	**2**	hub.prov
Aterosclerosi delle arterie renali	11	**2**	hub.prov
Dissezione dell'aorta	5	**2**	hub.prov
Embolia e trombosi delle arterie	5	**2**	hub.prov

Tabella B.1: Elenco di alcune diagnosi che svolgono ruolo di *hub provinciali* (R5) e *hub connettori* (R6) nelle comunità inviduate dal metodo SBM con $K = 3$.

Partizione SBM con $K = 4$

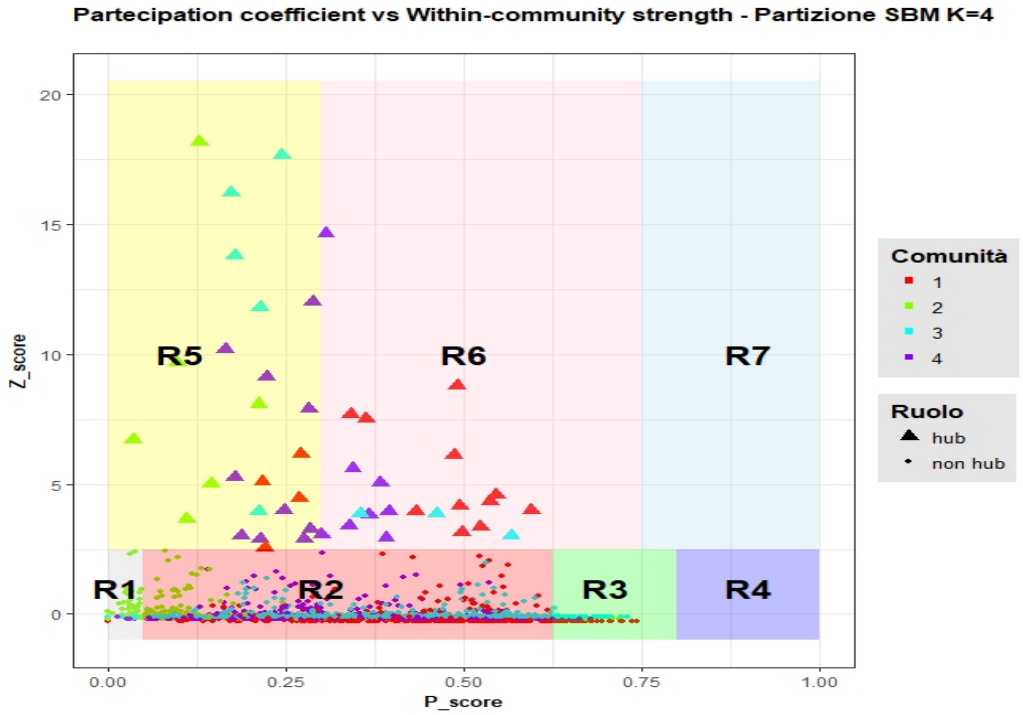

Figura B.2: Grafico dei ruoli per i nodi assegnati alle 4 comunità individuate con il metodo SBM.

Diagnosi	MDC	Comunità	Ruolo
Tumori maligni alla vescica	11	**1**	hub.conn
Diabete tipo II	5	**1**	hub.conn
Fibrillazione atriale	5	**4**	hub.conn
Aterosclerosi delle arterie	5	**1**	hub.conn
Aterosclerosi delle arterie con gangrena	5	**1**	hub.conn
Parto normale	14	**2**	hub.prov
Parto cesareo pregresso complicante la gravidanza	14	**2**	hub.prov
Artrosi all'anca	8	**3**	hub.prov
Artrosi al ginocchio	8	**3**	hub.prov
Trattamento per rimozione di dispositivo	8	**3**	hub.prov

Tabella B.2: Elenco di alcune diagnosi che svolgono ruolo di *hub provinciali* (R5) e *hub connettori* (R6) nelle comunità inviduate dal metodo SBM con $K = 4$.

Partizione SBM con $K = 7$

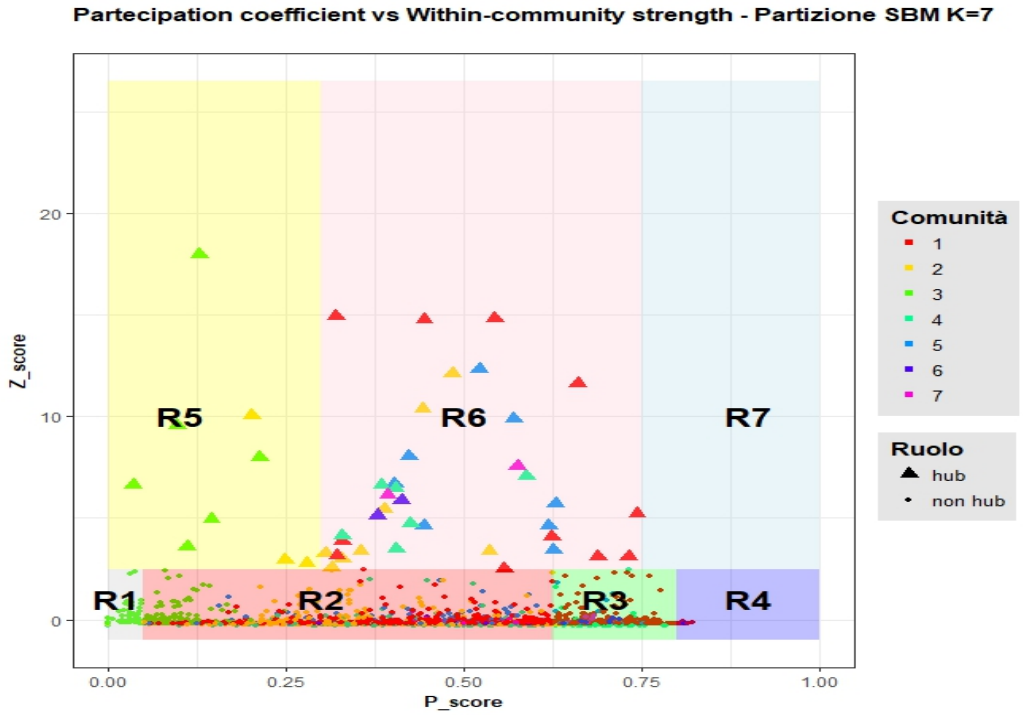

Figura B.3: Grafico dei ruoli per i nodi assegnati alle 7 comunità individuate con il metodo SBM.

Diagnosi	MDC	Comunità	Ruolo
Infarto subendocardico	5	**2**	hub.conn
Polmonite batterica,non specificata	4	**5**	hub.conn
Calcolosi renale	11	**4**	hub.conn
Artrosi all'anca	8	**7**	hub.conn
Sostituzione di articolazione dell'anca	8	**7**	hub.conn
Parto normale	14	**3**	hub.prov
Aborto indotto	14	**3**	hub.prov
Minaccia di travaglio prematuro	14	**3**	hub.prov
Sindrome coronarica	5	**2**	hub.prov
Aterosclerosi coronarica	5	**2**	hub.prov

Tabella B.3: Elenco di alcune diagnosi che svolgono ruolo di *hub provinciali* (R5) e *hub connettori* (R6) nelle comunità inviduate dal metodo SBM con $K = 7$.

Partizione SBM con $K = 8$

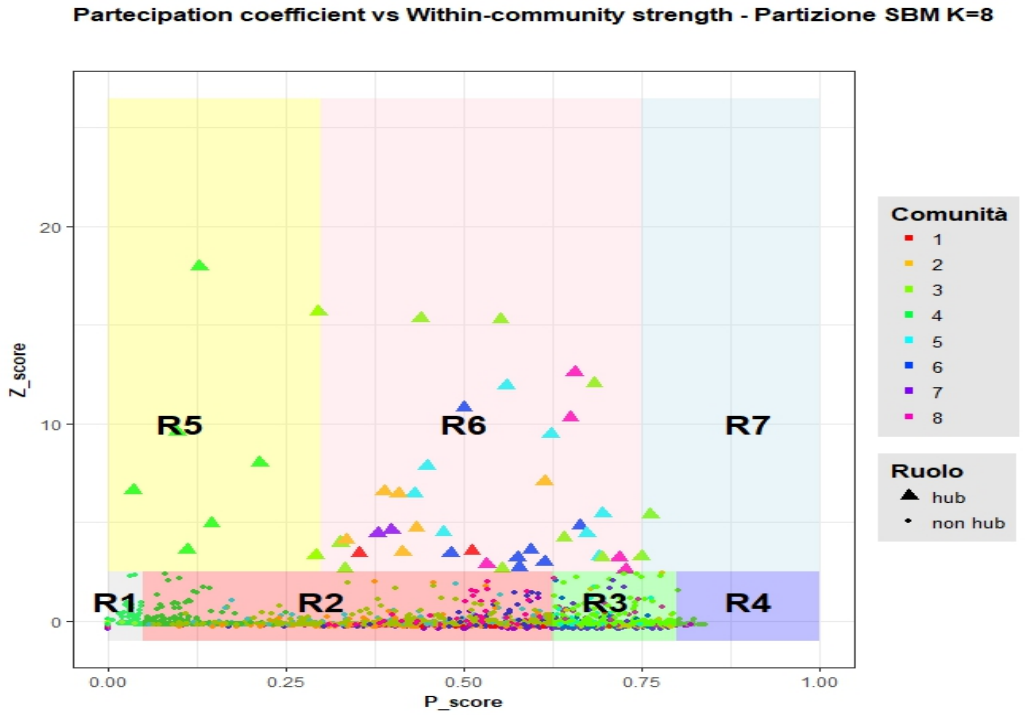

Figura B.4: Grafico dei ruoli per i nodi assegnati alle 8 comunità individuate con il metodo SBM.

Diagnosi	MDC	Comunità	Ruolo
Tumori maligni alla vescica	11	2	hub.conn
Infarto subendocardico	5	6	hub.conn
Polmonite batterica	4	5	hub.conn
Insufficienza cardiaca congestizia	5	8	hub.conn
Artrosi al ginocchio	8	7	hub.conn
Aborto ritenuto	14	4	hub.prov
Aborto indotto	14	4	hub.prov
Minaccia di travaglio prematuro	14	4	hub.prov
Parto normale	14	4	hub.prov
Parto cesareo pregresso complicante la gravidanza	14	4	hub.prov

Tabella B.4: Elenco di alcune diagnosi che svolgono ruolo di *hub provinciali* (R5) e *hub connettori* (R6) nelle comunità inviduate dal metodo SBM con $K = 8$.

Partizione SBM con $K = 9$

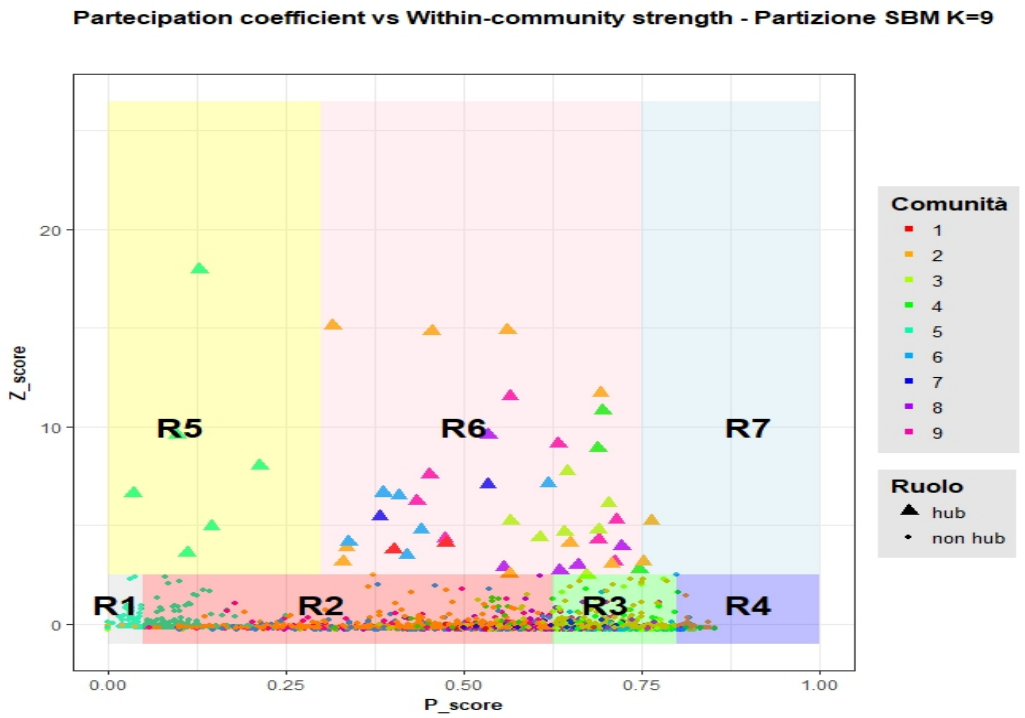

Figura B.5: Grafico dei ruoli per i nodi assegnati alle 9 comunità individuate con il metodo SBM.

Diagnosi	MDC	Comunità	Ruolo
Infarto subendocardico	5	8	hub.conn
Broncopolmonite	4	9	hub.conn
Malattia renale cronica	11	3	hub.conn
Artrosi all'anca	8	7	hub.conn
Postumi di fratture degli arti inferiori	8	2	hub.conn
Aborto ritenuto	14	5	hub.prov
Aborto indotto	14	5	hub.prov
Minaccia di travaglio prematuro	14	5	hub.prov
Parto normale	14	5	hub.prov
Parto cesareo pregresso complicante la gravidanza	14	5	hub.prov

Tabella B.5: Elenco di alcune diagnosi che svolgono ruolo di *hub provinciali* (R5) e *hub connettori* (R6) nelle comunità inviduate dal metodo SBM con $K = 9$.

Appendice C

Pacchetto igraph

Riportiamo di seguito la documentazione relativa alle funzioni del pacchetto *igraph* utilizzate nel nostro lavoro. Per la documentazione completa si rimanda a [15].

- `bipartite_projection` *Project a bipartite graph.*

  Description
  A bipartite graph is projected into two one-mode networks

  Usage
  ```
  bipartite_projection(graph, types = NULL, multiplicity = TRUE,
  probe1 = NULL, which = c("both", "true", "false"), remove.type
  = TRUE)
  ```

  Arguments
  `graph` The input graph. It can be directed, but edge directions are ignored during the computation.

  `types` An optional vertex type vector to use instead of the 'type' vertex attribute. You must supply this argument if the graph has no 'type' vertex attribute.

  `multiplicity` If TRUE, then igraph keeps the multiplicity of the edges as an edge attribute. E.g. if there is an A-C-B and also an A-D-B triple in the bipartite graph (but no more X, such that A-X-B is also in the graph), then the multiplicity of the A-B edge in the projection will be 2.

`probe1` This argument can be used to specify the order of the projections in the resulting list. If given, then it is considered as a vertex id (or a symbolic vertex name); the projection containing this vertex will be the first one in the result list. This argument is ignored if only one projection is requested in argument which.

`which` A character scalar to specify which projection(s) to calculate. The default is to calculate both.

`remove.type` Logical scalar, whether to remove the type vertex attribute from the projections. This makes sense because these graphs are not bipartite any more. However if you want to combine them with each other (or other bipartite graphs), then it is worth keeping this attribute. By default it will be removed.

Details
Bipartite graphs have a type vertex attribute in igraph, this is boolean and FALSE for the vertices of the first kind and TRUE for vertices of the second kind.

`bipartite_projection_size` calculates the number of vertices and edges in the two projections of the bipartite graphs, without calculating the projections themselves. This is useful to check how much memory the projections would need if you have a large bipartite graph.

`bipartite_projection` calculates the actual projections. You can use the probe1 argument to specify the order of the projections in the result. By default vertex type FALSE is the first and TRUE is the second.

`bipartite_projection` keeps vertex attributes.

Value
A list of two undirected graphs. See details above.

- `degree` *Degree and degree distribution of the vertices*

Description
The degree of a vertex is its most basic structural property, the number of its adjacent edges.

Usage
```
degree(graph, v = V(graph), mode = c("all", "out", "in", "total"),
loops = TRUE, normalized = FALSE)

degree_distribution(graph, cumulative = FALSE, ...)
```

Arguments
`graph` The graph to analyze.

`v` The ids of vertices of which the degree will be calculated. mode Character string, "out" for out-degree, "in" for in-degree or "total" for the sum of the two. For undirected graphs this argument is ignored. "all" is a synonym of "total".

`loops` Logical; whether the loop edges are also counted.

`normalized` Logical scalar, whether to normalize the degree. If TRUE then the result is divided by $n - 1$, where n is the number of vertices in the graph.

`cumulative` Logical; whether the cumulative degree distribution is to be calculated;

Value
For `degree` a numeric vector of the same length as argument `v`.

For `degree_distribution` a numeric vector of the same length as the maximum degree plus one. The first element is the relative frequency zero degree vertices, the second vertices with degree one, etc.

- `strength` *Strength or weighted vertex degree*

  Description
  Summing up the edge weights of the adjacent edges for each vertex.

  Usage
  ```
  strength(graph, vids = V(graph), mode = c("all", "out", "in",
  "total"), loops = TRUE, weights = NULL)
  ```

  Arguments
  `graph` The input graph.

  `vids` The vertices for which the strength will be calculated. mode Character string, "out" for out-degree, "in" for in-degree or "all" for the sum of the two. For undirected graphs this argument is ignored.

  `loops` Logical; whether the loop edges are also counted.

  `weights` Weight vector. If the graph has a weight edge attribute, then this is used by default. If the graph does not have a weight edge attribute and this argument is NULL, then a warning is given and degree is called.

  Value
  A numeric vector giving the strength of the vertices.

- `betweenness` *Vertex and edge betweenness centrality*

  Description
  The vertex and edge betweenness are defined by the number of geodesics (shortest paths) going through a vertex or an edge.

  Usage
  ```
  betweenness(graph, v = V(graph), directed = TRUE, weights =
  NULL, normalized = FALSE)
  ```

Arguments
`graph` The graph to analyze.

`directed` Logical, whether directed paths should be considered while determining the shortest paths.

`weights` Optional positive weight vector for calculating weighted betweenness. If the graph has a weight edge attribute, then this is used by default.

`v` The vertices for which the vertex betweenness will be calculated.

`normalized` Logical scalar, whether to normalize the betweenness scores. If TRUE, then the results are normalized according to $B_n = \frac{2B}{n^2-3n+2}$, where B_n is the normalized, B the raw betweenness, and n is the number of vertices in the graph.

Details
The vertex betweenness of vertex `v` is defined by $\sum_{i \neq j, i \neq v, j \neq v} g_{ivj}/g_{ij}$.

Value
A numeric vector with the betweenness score for each vertex in `v` for `betweenness`.

- `components` *Connected components of a graph*

Description
Calculate the maximal (weakly or strongly) connected components of a graph

Usage
```
component_distribution(graph, cumulative = FALSE, mul.size =
FALSE, ...)

components(graph, mode = c("weak", "strong"))
```

Arguments
`graph` The graph to analyze.

`cumulative` Logical, if TRUE the cumulative distirubution (relative frequency) is calculated.

`mul.size` Logical. If TRUE the relative frequencies will be multiplied by the cluster sizes.

`mode` Character string, either "weak" or "strong". For directed graphs "weak" implies weakly, "strong" strongly connected components to search. It is ignored for undirected graphs.

Details
`is_connected` decides whether the graph is weakly or strongly connected.

`components` finds the maximal (weakly or strongly) connected components of a graph.

`count_components` does almost the same as components but returns only the number of clusters found instead of returning the actual clusters.

`component_distribution` creates a histogram for the maximal connected component sizes. The weakly connected components are found by a simple breadth-first search. The strongly connected components are implemented by two consecutive depth-first searches.

Value
For `is_connected` a logical constant.

For `components` a named list with three components:
`membership` numeric vector giving the cluster id to which each vertex belongs.

`csize` numeric vector giving the sizes of the clusters.

`no` numeric constant, the number of clusters.

For `count_components` an integer constant is returned.

For `component_distribution` a numeric vector with the relative frequencies. The length of the vector is the size of the largest component plus one. Note that (for currently unknown reasons) the first element of the vector is the number of clusters of size zero, so this is always zero value

For `is_connected` a logical constant.

For `components` a named list with three components:
`membership` numeric vector giving the cluster id to which each vertex belongs.
`csize` numeric vector giving the sizes of the clusters.
`no` numeric constant, the number of clusters.

For `count_components` an integer constant is returned.

For `component_distribution` a numeric vector with the relative frequencies. The length of the vector is the size of the largest component plus one. Note that (for currently unknown reasons) the first element of the vector is the number of clusters of size zero, so this is always zero.

- `cluster_louvain` *Finding community structure by multi-level optimization of modularity.*

Description
This function implements the multi-level modularity optimization algorithm for finding community structure, see references below. It is based on the modularity measure and a hierarchial approach.

Usage
```
cluster_louvain(graph, weights = NULL)
```

Arguments
`graph` The input graph.

`weights` Optional positive weight vector. If the graph has a weight edge attribute, then this is used by default. Supply NA here if the graph has a weight edge attribute, but you want to ignore it.

Details
This function implements the multi-level modularity optimization algorithm for finding community structure, see VD Blondel, J-L Guillaume, R Lambiotte and E Lefebvre: Fast unfolding of community hierarchies in large networks, ([12]). It is based on the modularity measure and a hierarchial approach. Initially, each vertex is assigned to a community on its own. In every step, vertices are re-assigned to communities in a local, greedy way: each vertex is moved to the community with which it achieves the highest contribution to modularity. When no vertices can be reassigned, each community is considered a vertex on its own, and the process starts again with the merged communities. The process stops when there is only a single vertex left or when the modularity cannot be increased any more in a step.

Value
`cluster_louvain` returns a `communities` object.

Bibliografia

[1] Implementazione del metodo biSBM. http://danlarremore.com/bipartiteSBM/index.html/. Accessed: 2018-06-08.

[2] Implementazione del metodo di Louvain. https://sites.google.com/site/findcommunities//. Accessed: 2018-06-10.

[3] Implementazione del metodo SBM. http://www-personal.umich.edu/~mejn/dcsbm/KLOptimization.cpp/. Accessed: 2018-06-08.

[4] La classificazione DRG. http://www.salute.gov.it/imgs/C_17_pubblicazioni_1094_allegato.pdf. Accessed: 2018-06-08.

[5] La classificazione MDC. https://www.cheatography.com/major-diagnostic-category-mdc-to-ms-drg-mapping/. Accessed: 2018-06-08.

[6] La scheda di dimissione ospedaliera. http://www.salute.gov.it/portale/temi/p2_6.jsp?id=1232&area=ricoveriOspedalieri&menu=vuot. Accessed: 2018-06-08.

[7] Le diagnosi principali. http://www.salute.gov.it/portale/temi/p2_6.jsp?lingua=italiano&id=1280&area=ricoveriOspedalieri&menu=classificazione. Accessed: 2018-06-08.

[8] Christopher Aicher, Abigail Z Jacobs, and Aaron Clauset. Adapting the stochastic block model to edge-weighted networks. *arXiv preprint arXiv:1305.5782*, 2013.

[9] Christopher Aicher, Abigail Z Jacobs, and Aaron Clauset. Learning latent block structure in weighted networks. *Journal of Complex Networks*, 3(2):221–248, 2014.

[10] Taher Alzahrani and KJ Horadam. Community detection in bipartite networks: Algorithms and case studies. In *Complex Systems and Networks*, pages 25–50. Springer, 2016.

[11] Stephen J Beckett. Improved community detection in weighted bipartite networks. *Royal Society Open Science*, 3(1):140536, 2016.

[12] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.

[13] Ulrik Brandes. On variants of shortest-path betweenness centrality and their generic computation. *Social Networks*, 30(2):136–145, 2008.

[14] Francesco Calderoni, Domenico Brunetto, and Carlo Piccardi. Communities in criminal networks: A case study. *Social Networks*, 48:116–125, 2017.

[15] Gabor Csardi and Tamas Nepusz. The igraph software package for complex network research. *InterJournal*, Complex Systems:1695, 2006.

[16] Allison Davis, Burleigh Bradford Gardner, and Mary R Gardner. *Deep South: A social anthropological study of caste and class*. Univ of South Carolina Press, 2009.

[17] Santo Fortunato and Darko Hric. Community detection in networks: A user guide. *Physics Reports*, 659:1–44, 2016.

[18] Roger Guimera and Luís A Nunes Amaral. Cartography of complex networks: modules and universal roles. *Journal of Statistical Mechanics: Theory and Experiment*, 2005(02):P02001, 2005.

[19] Roger Guimera, Stefano Mossa, Adrian Turtschi, and LA Nunes Amaral. The worldwide air transportation network: Anomalous centrality, community structure, and cities' global roles. *Proceedings of the National Academy of Sciences*, 102(22):7794–7799, 2005.

[20] Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137, 1983.

[21] Brian Karrer and Mark EJ Newman. Stochastic blockmodels and community structure in networks. *Physical review E*, 83(1):016107, 2011.

[22] Daniel B. Larremore, Aaron Clauset, and Abigail Z. Jacobs. Efficiently inferring community structure in bipartite networks. *Physical review E*, 90:012805, Jul 2014.

[23] Marina Meilă. Comparing clusterings by the variation of information. In *Learning theory and kernel machines*, pages 173–187. Springer, 2003.

[24] Mark Newman. *Networks: An introduction*. Oxford university press, 2010.

[25] Mark EJ Newman. Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. *Physical review E*, 64(1):016132, 2001.

[26] Tore Opsahl, Filip Agneessens, and John Skvoretz. Node centrality in weighted networks: Generalizing degree and shortest paths. *Social networks*, 32(3):245–251, 2010.

[27] Tiago P Peixoto. Parsimonious module inference in large networks. *Physical review letters*, 110(14):148701, 2013.

[28] Carlo Piccardi. Finding and testing network communities by lumped Markov chains. *PloS one*, 6(11):e27028, 2011.

[29] Carlo Piccardi, Lisa Calatroni, and Fabio Bertoni. Communities in Italian corporate networks. *Physica A: Statistical Mechanics and its Applications*, 389(22):5247–5258, 2010.

[30] Tom AB Snijders and Krzysztof Nowicki. Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *Journal of classification*, 14(1):75–100, 1997.

[31] Wayne W Zachary. An information flow model for conflict and fission in small groups. *Journal of anthropological research*, 33(4):452–473, 1977.